INTERNATIONAL JOURNAL OF COMPUTERS COMMUNICATIONS & CONTROL
Online ISSN 1841-9844, ISSN-L 1841-9836, Volume: 21, Issue: 5, Month: October, Year: 2026
Article Number: 7542, https://doi.org/10.15837/ijccc.2026.5.7542

UAVCity-YOLO: An Improved YOLOv13 Architecture for Robust UAV Detection Using Infrared Data in Noisy Urban Environments

P.-T. Nguyen, V.-P. Hoang

Phat T. Nguyen, Van-Phuc Hoang*
Le Quy Don Technical University, Hanoi, Vietnam
236 Hoang Quoc Viet Street, Hanoi, Vietnam

*Corresponding author: phuchv@lqdtu.edu.vn

Abstract

The rapid growth of unmanned aerial vehicles (UAVs), especially drones, in urban airspace raises significant security and safety concerns, creating an urgent demand for accurate real-time detection. While infrared imaging enables continuous surveillance, urban infrared UAV detection remains challenging due to the small target size, weak thermal contrast, and severe background thermal noise, which limit the robustness of existing detectors. This paper proposes UAVCity-YOLO, a lightweight and robust single-stage detector based on the YOLOv13n (You Only Look One version 13n) architecture, specifically designed for UAV detection using infrared data in thermally complex urban environments. The proposed approach integrates a coordinate-aware backbone for enhanced small-target representation, an improved multi-scale feature fusion Neck structure, and a novel infrared-aware regression loss to improve localization stability under thermal noise. Extensive experiments on urban infrared UAV datasets demonstrate that UAVCity-YOLO consistently outperforms state-of-the-art models. In particular, on the thermally noisy dataset, UAVCity-YOLO achieves an $mAP_{0.5}$ of 89.1% and $mAP_{0.5:0.95}$ of 57.9%, while maintaining fast inference at 1.5 ms and a compact model size of only 4.1 MB, confirming its effectiveness for real-time UAV surveillance.
Keywords: UAV detection, anti-drone, infrared imaging, YOLO.

1 Introduction

Technological developments in recent decades have led to the widespread use of unmanned aerial vehicles (UAVs), especially drones, in a variety of civilian applications, from transportation, precision agriculture to surveillance and rescue. However, this proliferation has also increased security threats. Unauthorized intrusion of UAVs, especially drones, into sensitive areas such as airports and critical infrastructure, has become a global problem [1, 2]. Serious incidents at Gatwick Airport (UK) or the appearance of drones near Frankfurt Airport, Germany, in May 2019, threatened aviation safety [2]; due to a drone attack in September 2019, the largest oil refining facilities of Aramco, the national oil company of Saudi Arabia, were burned down and had to stop operations [2]. The appearance of illegal drones on October 2, 2025 has greatly affected the operations of Munich Airport [3]. Therefore, the research and development of effective UAV detection systems has become a strategic priority for public security of many countries.

Traditional UAV detection methods often rely on sensors such as radar, radio frequency (RF), and audio and video [4]. Although radars are capable of operating at long ranges and in all weather conditions, they often have difficulty detecting low-flying, small, and slow-flying targets due to their low radar cross section (RCS) and are easily confused with background noise. To distinguish UAVs from other objects, advanced radar signal processing algorithms focus on extracting micro-Doppler (MDS) features from the rotational motion of the propellers. The MDS data, often represented as spectrograms, are then fed into Convolutional Neural Network (CNN) or Long Short Term Memory (LSTM) models for classification [5]. Most commercial UAVs use RF signals to communicate with the controller. This approach focuses on capturing and analyzing the RF characteristics of these signals. DL models such as Deep Neural Networks (DNNs) and CNNs have achieved high accuracy in classifying UAVs based on RF signal characteristics, working effectively regardless of weather or lighting conditions, but are ineffective against UAVs operating in fully autonomous mode [6, 7]. Acoustic-based UAV detection uses the characteristic sounds from propellers or UAV engines [8, 9]. Systems use microphone arrays to capture sound, then extract acoustic features such as Mel-Frequency Cepstral Coefficients (MFCC). These features are fed into deep learning classifiers such as LSTMs or CNNs to classify UAVs [8, 10]. This is a low-cost, effective solution for short-range applications but has a very limited range and performance degrades sharply in noisy urban environments. UAV detection methods based on image data (optical, infrared) are prominent methods [11]. The use of IR data is an effective approach, especially in low-light conditions, due to the ability to detect the heat of the UAV's engine. Despite the lower resolution, IR cameras still perform on par with optical cameras in many cases [12, 13, 14].

Sensor fusion approaches, which combine multiple sensing modalities, have been widely investigated for UAV detection [2, 15, 16, 17]. Although these methods improve robustness by exploiting complementary sensor information, they suffer from complex sensor synchronization, cross-modal feature fusion, high computational cost, and expensive system integration. These limitations motivate the development of cost-effective single-sensor solutions. Among them, infrared image-based detection is particularly attractive for continuous day-and-night surveillance, as it remains effective under low-light conditions while avoiding the limitations of radar, RF, and acoustic sensing in complex urban environments.

In recent years, deep learning–based object detection methods have achieved significant progress in visual perception tasks, including both 2D object detection from images and 3D object detection using depth data or point cloud representations [18, 19]. While 2D detection approaches typically exploit visual features extracted from RGB or infrared images to localize and classify objects, 3D detection methods further incorporate geometric and spatial structural information to improve scene understanding in complex environments. For the purpose of UAV monitoring, 2D image-based detectors, particularly those belonging to the YOLO family, have demonstrated remarkable effectiveness due to their strong balance between detection accuracy and real-time inference speed. Consequently, many recent studies have focused on improving YOLO-based architectures to enhance small-object detection capability, increase robustness against background interference, and maintain computational efficiency for real-time UAV surveillance applications. For instance, lightweight architectures such as LCE-YOLO have been proposed to improve feature sensitivity and reduce computational complexity for detecting UAVs in complex backgrounds [20]. Similarly, enhanced YOLOv5 models integrating Ghost convolution modules and multiscale feature extraction strategies have shown improved performance in detecting UAVs and other flying objects under low-light and infrared imaging conditions [21]. In addition, recent work incorporating bidirectional feature pyramid networks and attention mechanisms has further improved the detection of small UAV targets in complex scenes by strengthening multiscale feature fusion capabilities [22].

In studies [23, 24, 25, 26], many good results have been achieved for the problem of detecting small sized UAVs on infrared data. Improvements focus on optimizing lightweight architectures, adding small target detection modules, and using GANs to enhance data. However, a weakness in these studies is the limited scope of testing, and there is no assessment of the impact of thermal noise on the performance of the model. To our knowledge, comprehensive studies addressing UAV detection in urban infrared imagery specifically with thermal noise remain limited. The complexity of the urban thermal environment, where artificial heat sources are easily confused with the weak thermal signals of UAVs, is posing a serious challenge. This significantly reduces the generalizability and reliability of current models in practical deployment. The detection of UAVs in urban infrared images faces three major challenges that current deep learning models. Firstly, UAV targets are small and blurry. Due to the long distance, UAVs occupy only a small proportion of pixels in IR images. Moreover, the temperature difference between the engine and the sky is usually small, making them blurry and easily eliminated by pooling layers in deep neural networks. Secondly, the urban background is complex and thermally noisy. Urban environments contain many different heat sources (buildings, vehicles, lighting), creating a complex thermal noise background. This leads to a high false alarm rate, as the model easily confuses small heat spots on the ground or building edges with the thermal signature of UAVs. The third challenge is the lack of high-quality training data. Collecting quality IR image datasets of UAVs flying over urban areas with different weather conditions and observation angles is difficult and expensive. Existing data are often biased towards simple skies, which do not reflect the actual complexity.

To address the aforementioned challenges, this paper proposes an improved model based on the YOLOv13n architecture, specifically designed for the problem of detecting UAVs in infrared images in urban areas with thermal noise. The goal is to improve the ability to extract and amplify UAV features, while enhancing the ability to resist noise from complex backgrounds. The main contributions of this work include:

- Redesign the Backbone of the YOLOv13n model to improve the ability to extract features of infrared small targets. Adding the Coordinate attention mechanism to the DS-C3k2 module, denoted as the AMC-C3k2 module; replacing the A2C2f module with the AMC-C3k2 module; adding the Spatial Pyramid Pooling - Fast (SPPF) module.

- Improve the Neck part by removing modules that can cause interference to the detectors, and rebuild the connections of the feature maps to reduce the complexity of the model structure.

- Propose a novel infrared aware bounding box regression loss function, termed IR-NCIoU, to replace the original CIoU loss in YOLOv13n. The proposed IR-NCIoU loss is specifically designed to address the limitations of CIoU in detecting small, blurry UAV targets under thermally noisy infrared conditions. By introducing a normalized IoU formulation to amplify gradients for low IoU samples, incorporating object scale aware center distance normalization, and employing a log-scale shape consistency constraint for improved robustness against thermal blur, IR-NCIoU significantly enhances localization accuracy and convergence stability for small UAVs. This loss level improvement complements the architectural enhancements of the Backbone and Neck, forming a unified optimization framework tailored for infrared UAV detection in complex urban environments.

- Build an infrared image dataset on an urban background with added thermal noise. Conduct comprehensive experiments on a thermally noisy infrared UAV dataset to evaluate the performance of the proposed model and compare it with other modern methods, demonstrating the improvement of the solution.

The rest of the paper is structured as follows. Section 2 presents the detailed architecture and improved components of the proposed model. Section 3 presents the experimental setup, results analysis, and performance comparison. Finally, Section 4 provides conclusions and future research directions.

2 Methodology

2.1 Original YOLOv13n model

YOLOv13n was developed to address the limitations of previous YOLO versions in capturing complex semantic relationships, particularly in scenarios involving small, dense objects and highly dynamic environments [27, 28, 29]. Although traditional convolution-based architectures have achieved strong performance in object detection, their ability to model long-range dependencies and high-order feature interactions remains limited. To overcome these challenges, YOLOv13n introduces several architectural enhancements that improve both feature representation capability and computational efficiency.

The overall architecture of YOLOv13n is illustrated in Figure 1. The model follows the typical single-stage detection paradigm and consists of five main components: the input module, backbone network, Hypergraph-based Adaptive Correlation Enhancement module (HyperACE), neck network, and the detection head. The input image is first processed by the backbone network, which extracts hierarchical feature representations from low-level spatial information to high-level semantic features. These features are then enhanced through the HyperACE module, which aims to capture high-order correlations between spatial regions. Subsequently, the refined feature maps are aggregated by the neck for multi-scale feature fusion before being passed to the detection head to produce the final detection results.

The detailed architecture and internal structure of each component are illustrated in Figure 2. In the backbone, YOLOv13n employs an efficient feature extraction design that replaces traditional large-kernel convolutions with Depthwise Separable Convolutions, significantly reducing the number of parameters and computational complexity while preserving strong representation capacity. This design enables the model to maintain high efficiency while processing high-resolution inputs. A key innovation of YOLOv13n lies in the proposed Hypergraph-based Adaptive Correlation Enhancement (HyperACE) module. Unlike conventional convolutional operations that focus on local neighborhood interactions, HyperACE models complex relationships between multiple feature nodes through hypergraph structures. Specifically, pixels or feature regions are mapped into hypergraph vertices, and hyperedges are dynamically constructed to capture high-order dependencies among spatially distributed features. Through this mechanism, the model is able to learn richer contextual representations and strengthen feature interactions across distant regions in the image. To further facilitate information propagation across the entire network, YOLOv13n introduces the Full-process Aggregation and Distribution (FullPAD) framework. FullPAD enhances global feature synergy by injecting correlation-enhanced representations generated by HyperACE into different stages of the network, including the backbone, neck, and detection head. This design allows the model to maintain consistent feature coordination across multiple levels, improving gradient propagation and stabilizing the training process.

The Neck module is responsible for multi-scale feature fusion. By integrating feature maps from different levels of the backbone, the neck enhances the model's ability to detect objects of various sizes. This multi-scale aggregation mechanism ensures that both fine-grained details and high-level semantic information are preserved during feature propagation. Finally, the detection head performs object classification and bounding box regression. YOLOv13n adopts an anchor-free detection mechanism and integrates reparameterization techniques together with a Non-Maximum Suppression (NMS)-free design, which eliminates the need for conventional post-processing steps while maintaining high detection accuracy. This design improves inference efficiency and reduces latency, making the model suitable for real-time applications.

Overall, the architecture of YOLOv13n combines efficient convolutional operations, hypergraph-based relational modeling, and global feature aggregation mechanisms. As illustrated in Figure 1 and Figure 2, these components work together to improve detection accuracy, strengthen contextual feature representation, and maintain the lightweight and real-time characteristics that are fundamental to the YOLO family of detectors.

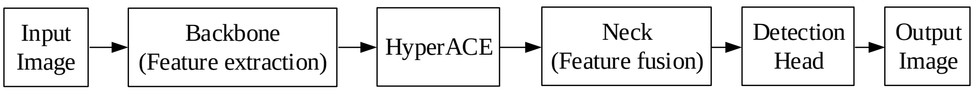

Figure 1: Overall pipeline diagram of the YOLOv13n model.

2.2 Improved YOLOv13n

Although YOLOv13n has many outstanding improvements compared to previous YOLO versions in the problem of detecting objects in large data sets such as COCO [30]. However, for specific

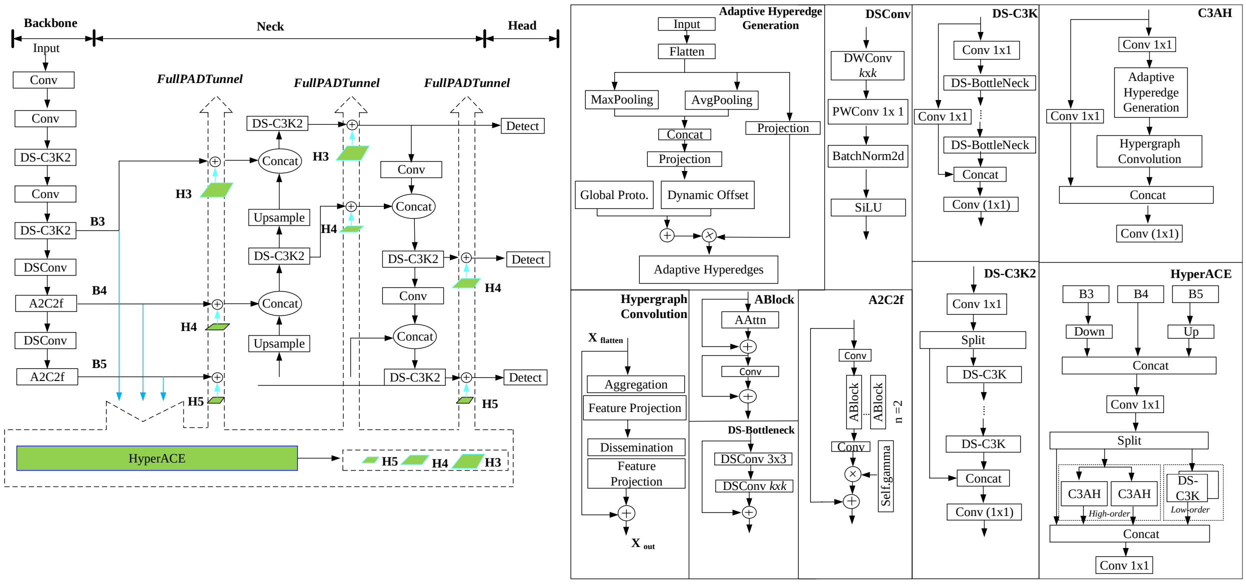

Figure 2: Structure of the original YOLOv13n model.

tasks, YOLOv13n still has certain limitations that need to be improved. Specifically, in the problem of detecting small-sized UAV (or drone) targets in infrared image data on urban backgrounds. In this case, the YOLOv13n model reveals disadvantages that affect the performance of detecting UAV targets. In the structure of the YOLOv13n model, there are 3 parts: Backbone; Neck and Head [27, 28]. In which, small targets are greatly affected by the Backbone and Neck parts related to feature extraction and multi-scale association of feature maps, often causing noise or loss of small object feature information. Therefore, the structure of the proposed model is improved from YOLOv13n shown in Figure 3, to improve the efficiency of UAV target detection.

2.2.1 Improved Backbone

In YOLO models, the Backbone module plays an important role of extracting the target features, directly and greatly affects the efficiency of object detection. The first improvement is to add the Attention mechanism - Coordinate (AMC) block with the C3K2 block, called the AMC-C3k2 module. The second one is to replace two A2C2f modules by AMC-C3k2 modules. The third improvement is the SPPF module added at the end of the Backbone. The improvements are described specifically as follows:

a. Improved DS-C3k2 module (AMC-C3K2)

Traditional attention mechanisms such as Squeeze-and-Excitation (SE) or Channel Attention (CA) only focus on modeling channel correlations while ignoring spatial location information, which is disadvantageous for locating small targets. The AMC module structure is depicted in Figure 4 and Equations (1-5) to solve this problem by embedding coordinate information into channel attention maps through two independent horizontal and vertical pooling operations [30]. The AMC-C3k2 module structure is depicted in Figure 5.

Given the input tensor $X \in^{RC\times H\times W}$, the transformation process is defined as follows. Adaptive Average Pooling operations are applied along the horizontal (z) and vertical (r) directions to obtain the feature vectors $u^z \in R^{C\times H\times 1}$ and $u^r \in R^{C\times 1\times W}$

$$u_c^z = \frac{1}{W}\sum_{0\leq j<W} x_c(z,j) \tag{1}$$

$$u_c^r = \frac{1}{H}\sum_{0\leq i<H} x_c(i,r) \tag{2}$$

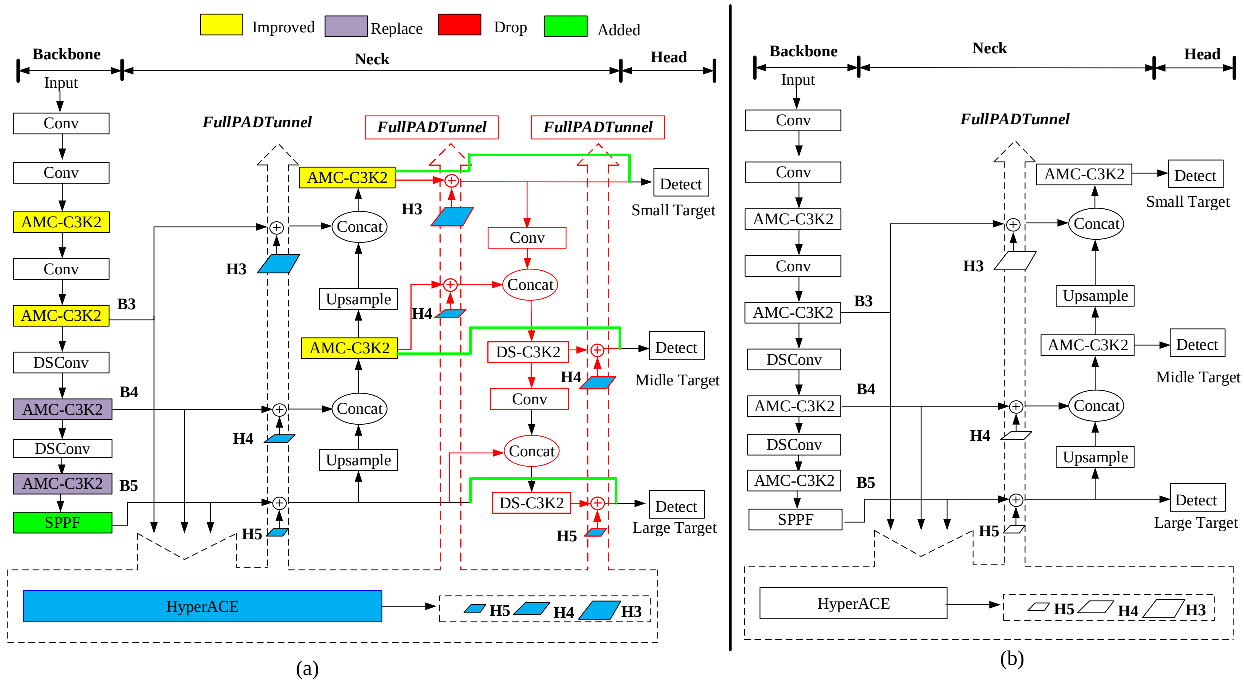

Figure 3: Proposed model: (a) Locations of the improved, replaced, removed, and added modules; (b) Structure of proposed model (UAVCity-YOLO).

These two feature vectors are concatenated ($u = [u^z; u^r]$), then transformed by a 1×1 convolution ($F_{1\times1}$) to reduce the channel dimensionality:

$$f = \delta(\mathbf{B}(F_{1x1}([u^z, u^r]))) \tag{3}$$

where $f \in R^{C/g\times(H+W)\times1}$ is the intermediate attention map, δ is the Hardswish activation function and $\mathbf{B}$ is Batch Normalization (BN), g is the channel reduction factor. Then f_B is decomposed back into $f^z \in R^{C/g\times H\times1}$ and $f^r \in R^{C/g\times W\times1}$. Two further 1×1 convolutions (F_z, F_r) are applied to restore the channel dimension C and generate the vertical and horizontal spatial attention weights:

$$t^z = \sigma(\mathbf{F}_z(f^z)),\ t^r = \sigma(\mathbf{F}_r(f^r)), \tag{4}$$

where σ is the Sigmoid activation function. The output X_{out} is computed by multiplying the original feature map X_{input} with the learned attention weights:

$$X_{out} = X_{int} \otimes t^z \otimes t^r, \tag{5}$$

where $\otimes$ is the element-wise multiplication.

b. Replacing the A2C2f module by the AMC-C3k2 module

Although the A2C2f module leverages Self-Attention to model long-term dependencies, this mechanism is inefficient in handling small-sized objects [27, 28]. Self-Attention is computationally expensive and is easily distracted by prominent structural thermal noise sources in urban environments, reducing the signal-to-noise ratio (SNR) of UAV targets. In contrast, the improved DS-Bottleneck module combines the computational efficiency of Depthwise Separable Convolution (DSConv) with the location-specific attention of AMC. The AMC module requires the consensus of two independent attention vectors, z (horizontal) and r (vertical), to enhance the feature map. By suppressing thermal noise signals that span only one direction (vertical or horizontal), AMC focuses network resources on important (z, r) coordinate intersections, thereby significantly improving the ability to distinguish between targets and noise.

c. Adding the SPPF module

The SPPF module was first introduced in the YOLOv5 architecture [31] as one of the powerful improvements of the real-time object detection model. Although the original SPP module has been

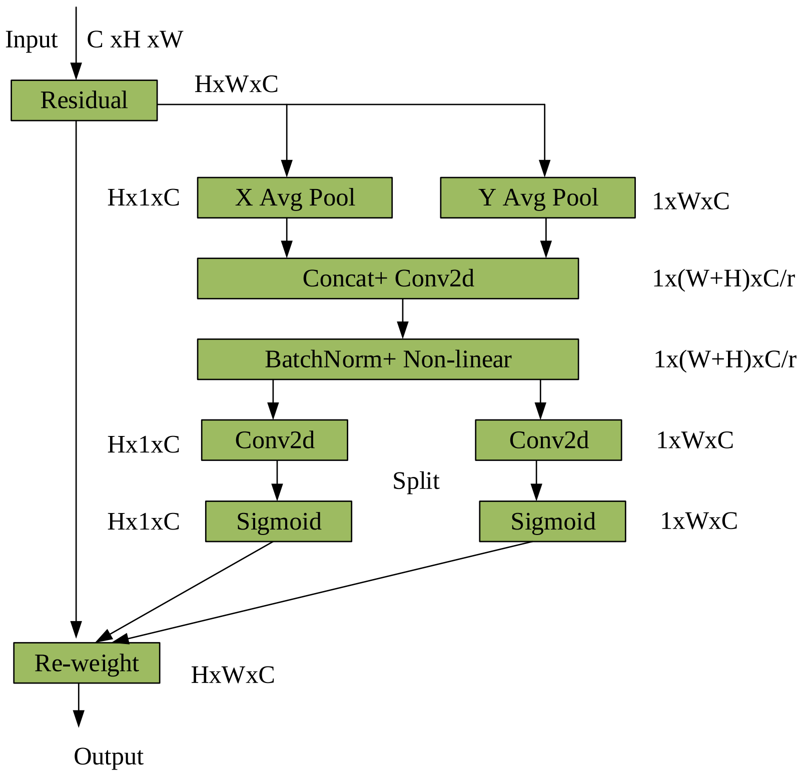

Figure 4: Structure of AMC module.

proven effective in fusing multi-scale contextual information, SPPF is introduced to provide the same capability with significantly faster processing speed by using serial Max Pooling layers instead of parallel ones. For the problem of detecting small UAV or drone targets in urban infrared images, the model often faces the challenge of losing detail information and feature discrimination due to the small object size and complex background noise. To improve the detection accuracy and enhance the ability to extract multi-scale context features, we propose to add the SPPF module at the end of the Backbone module of the UAVCity-YOLOv13n model.

As illustrated in Figure 6, the SPPF module takes the input from the previous Conv layer. Instead of using three Max Pooling layers with kernel size of $K \times K$ in parallel and then concatenating as in traditional SPP, SPPF performs three Max Pooling operations with the same kernel size ($K = 5 \times 5$) applied consecutively. The outputs from these three Max Pooling layers, together with the original output from the original Conv layer (before Max Pooling), are merged channel-wise. Finally, another Conv layer is used to reduce the number of channels and normalize the features. This serial arrangement allows for reuse of computation in the deep neural network architecture, significantly reducing the processing time compared to SPP, especially on modern hardware. Based on the improved Backbone architecture diagram as shown in Figure 3, the SPPF module is placed at the end of the Backbone network, right after the last AM-C3K2 block. This strategic placement ensures that the SPPF module can apply the multi-scale spatial pooling mechanism on the top-level features extracted by Backbone.

2.2.2 An improved Neck structure

The Neck part of the YOLOv13n model architecture is improved to optimize the multi-level feature fusion process [27, 28], which is especially important for the problem of detecting small UAV targets by infrared imaging with complex urban noise backgrounds. In infrared UAV imagery, targets typically occupy only a few pixels and exhibit weak thermal contrast against cluttered urban environments; therefore, preserving spatial details during feature fusion becomes critically important. The structure

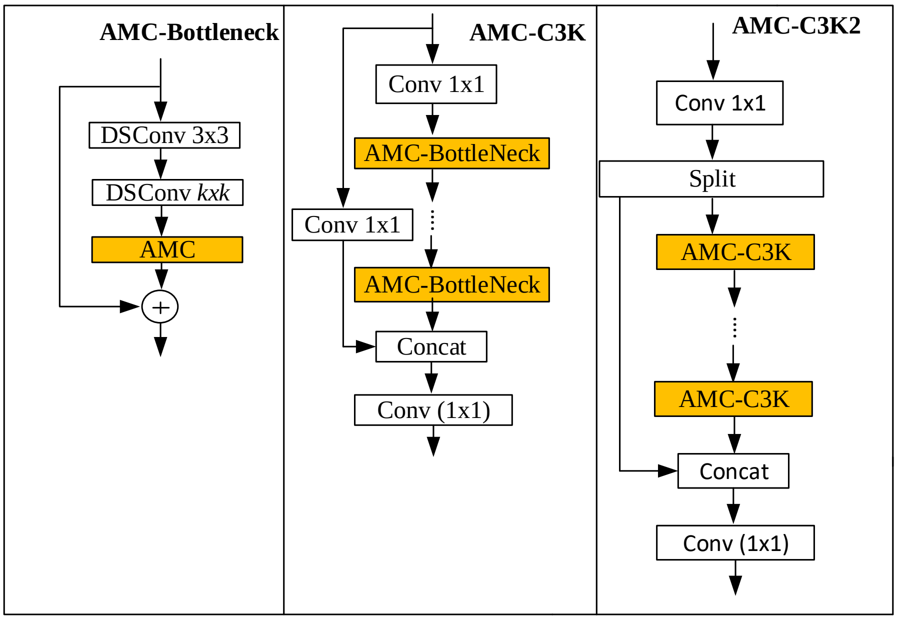

Figure 5: Structure of the AMC-C3k2 module.

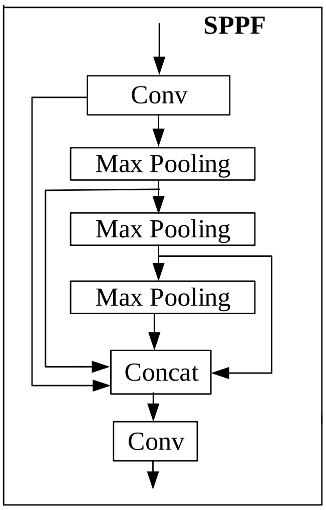

Figure 6: Structure of the SPPF module.

of the original Neck and the improved Neck is clearly depicted in Figure 3. The original Neck architecture is based on the traditional two-dimensional Feature Pyramid Network - Path Aggregation Network (FPN-PAN) model, using AM-C3K2 blocks and Concat operation to combine features from different levels (B3, B4, B5) of the Backbone. However, for small targets, the depth of the original FPN-PAN network can cause loss of location information and details when the features have to go through many transformation layers and combinations. In particular, the details of the small target (which is inherently weak) are easily "buried" by high-level contextual information, or degraded by the blurring effect of hidden pooling layers.

To minimize the loss of important information of the small target and enhance the efficiency of feature transfer, we propose to drop a part of the FPN-PAN connection flow and add a direct shortcut path. The improved Neck structure (Figure 2b) shows a drastic simplification. Specifically, the proposed model drops the complex DS-C3K2 and Conv blocks in the connection flow between FPN/PAN levels to reduce the depth of the network, minimize the computational cost and speed up the inference. More importantly, it minimizes the blurring or loss of details caused by multiple consecutive convolutional layers, helping to preserve the location features of the small target. The improved Neck structure shows a drastic simplification of the PAN (Bottom-up) flow. Instead of using a complex processing chain, the features from levels B3, B4, B5 are transferred in a shorter path, combining AM-C3k2 blocks with Upsample and Concat operations according to the standard FPN-PAN process but with a streamlined structure. This design helps maintain weak spatial cues of small infrared UAV targets while suppressing background thermal noise during feature propagation. The improved Neck structure is a balance between architecture depth and feature preservation, a key factor to achieve the improved performance in detecting small UAV targets in infrared environments.

2.2.3 Proposed IR-NCIoU loss function

In YOLOv13n, the CIoU (Complete Intersection over Union) loss function has been proven to be one of the most effective bounding-box regression loss functions in general object detection problems. However, in the specific scenario of detecting small UAVs with blurred edges in heavily noisy infrared image data in urban environments, CIoU still has inherent limitations. Specifically, the loss function CIoU is defined as in Equation (6):

$$L_{\mathrm{CIoU}} = 1 - \mathrm{IoU} + \frac{\rho^2(b, b^{gt})}{c^2} + \alpha v \tag{6}$$

where IoU denotes the Intersection over Union between the predicted bounding box B and the ground-truth bounding box B^{gt}:

$$\mathrm{IoU} = \frac{|B \cap B^{gt}|}{|B \cup B^{gt}|},$$

$\rho^2(b, b^{gt})$ represents the Euclidean distance between the centers of the two bounding boxes; c is the diagonal length of the smallest enclosing box covering both B and B^{gt} and v is the aspect ratio penalty term defined as:

$$v = \frac{4}{\pi^2}\left(\arctan\frac{\omega^{gt}}{h^{gt}} - \arctan\frac{\omega}{h}\right)^2$$

with the corresponding balancing factor of

$$\alpha = \frac{v}{v + 1 - IoU}$$

In the context of small UAV detection in infrared imagery, the first major limitation of CIoU arises from its linear IoU component. When the UAV occupies only a very limited number of pixels in the image, even substantial improvements in localization accuracy may lead to only marginal increases in IoU values. Another limitation lies in the center distance penalty term $\frac{\rho^2}{c^2}$, which is normalized solely by the size of the enclosing box and does not account for the absolute scale of the target. Consequently, for small UAVs, relatively severe center deviations are insufficiently penalized. The third limitation is associated with the shape penalty term v, where the use of $\arctan(\omega/h)$ becomes numerically unstable when the bounding box dimensions ω and h are small and corrupted by thermal noise, a common characteristic of infrared images in urban environments. These inherent drawbacks render CIoU suboptimal for effectively optimizing the weak and noisy features of small UAV targets, even when the Backbone and Neck architectures of YOLOv13n are carefully redesigned to preserve spatial information.

To overcome these issues, we propose to replace CIoU with a novel loss function termed IR-NCIoU (Infrared-aware Normalized Complete IoU), which is specifically tailored for small UAV detection in infrared imagery. Firstly, IR-NCIoU replaces the conventional IoU with a Normalized IoU (NIoU), defined as in Equation (7):

$$\mathrm{NIoU} = \frac{(1+n)\,|B \cap B^{gt}|}{|B \cup B^{gt}| + n\,|B \cap B^{gt}|}, \tag{7}$$

where n is a tunable parameter ($n > 0$). Next, the center distance penalty is re-normalized in a scale-aware manner to enhance localization sensitivity for small targets:

$$\mathrm{L_{center}} = \frac{\rho^2(b, b^{gt})}{c^2 + \beta \min{(\omega, h)}^2},$$

where ω and h denote the width and height of the predicted bounding box, respectively, and β is a weighting factor.

$$\upsilon_{ir} = \frac{4}{\pi^2}\left(\log\frac{\omega}{h} - \log\frac{\omega^{gt}}{h^{gt}}\right)^2$$

with the corresponding balancing coefficient defined as

$$\alpha = \frac{\upsilon_{ir}}{\upsilon_{ir} + 1 - NIoU}$$

By integrating these components, the complete IR-NCIoU loss function is formulated as Equation (8):

$$L_{\mathrm{IR-NCIoU}} = 1 - \mathrm{NIoU} + \frac{\rho^2(b, b^{gt})}{c^2 + \beta \min{(\omega, h)}^2} + \alpha\upsilon_{ir} \tag{8}$$

Owing to this design, IR-NCIoU effectively overcomes the inherent limitations of CIoU in small UAV detection tasks, while simultaneously strengthening gradient responses for samples that are difficult to optimize and exhibit poor localization, thereby improving localization accuracy and stabilizing the convergence under thermally noisy urban infrared environments.

3 Dataset and experimental results

3.1 Dataset

To evaluate the overall performance of the proposed model, we used an urban infrared image dataset augmented with thermal noise [23]. The complete dataset contains 1,093 infrared UAV images and was divided into 874 training images and 219 validation images (approximately 80:20). Owing to the relatively limited size of the available dataset, further partitioning it into separate training, validation, and independent test subsets would substantially reduce the number of training samples and the diversity of the data available for model learning. Therefore, this study adopts a two-subset training–validation protocol, and no additional independent test set was constructed. All quantitative results reported in this paper were obtained on the held-out validation set. The training and validation subsets were mutually exclusive, with no duplicated images shared between them, ensuring an objective and unbiased evaluation. To further assess the robustness and generalization capability of the proposed model, an additional five-fold cross-validation experiment was conducted, as described in Section 3.3. For performance evaluation, two experimental datasets were considered. Dataset_1 corresponds to the original infrared UAV dataset, whereas Dataset_2 was generated directly from Dataset_1 by applying the Gaussian additive thermal-noise model described in Equations (9)–(10). Consequently, Dataset_1 and Dataset_2 contain exactly the same number of images, identical UAV annotations, and the same training–validation partition. The only difference is that Dataset_2 includes synthetically generated thermal noise to simulate more challenging urban infrared sensing conditions, thereby enabling a controlled evaluation of the proposed model under both clean and thermally degraded environments.

Figure 7 illustrates the infrared images before and after adding the thermal noise where the red circle is the UAV target location. Thermal noise is added to the dataset as follows:

- All input infrared (IR) images are represented as an intensity matrix I. The thermal noise model applied here is Additive Noise, modeled by a Gaussian distribution [32].

- The thermal noise N at each pixel (x, y) is assumed to be an independent random variable following a Gaussian normal distribution $N(\mu, \sigma^2)$:

$$N(x, y) \sim N(0, \sigma_N^2), \tag{9}$$

where $N(0, \sigma_N^2)$ stands for Gaussian distribution and σ_N^2 is the variance which represents the noise intensity. The image after adding noise I_n is calculated as the sum of the original pixel intensity and the random noise component:

$$I_n(x, y) = I(x, y) + N(x, y) \tag{10}$$

In this study, the input infrared images were represented in the intensity range 1-255. Dataset_2 was generated by independently sampling zero-mean Gaussian noise with a fixed standard deviation of $\sigma_N = 10$ for each pixel of every image in Dataset_1. The same noise level was used for all images to maintain a controlled and reproducible experimental condition. After noise addition, the resulting pixel intensities were clipped to the valid image range. Consequently, Dataset_1 and Dataset_2 contain identical images in terms of scene content, UAV annotations, and training–validation partition, while differing only in the added Gaussian thermal-noise component. The additive Gaussian model was adopted as a first-order approximation of random thermal-sensor intensity fluctuations. Although this model does not fully reproduce all real urban infrared disturbances, it enables a controlled assessment of model robustness under image degradation. In the present study, a single noise level was considered; evaluating multiple noise intensities and more realistic thermal-degradation models is reserved for future work.

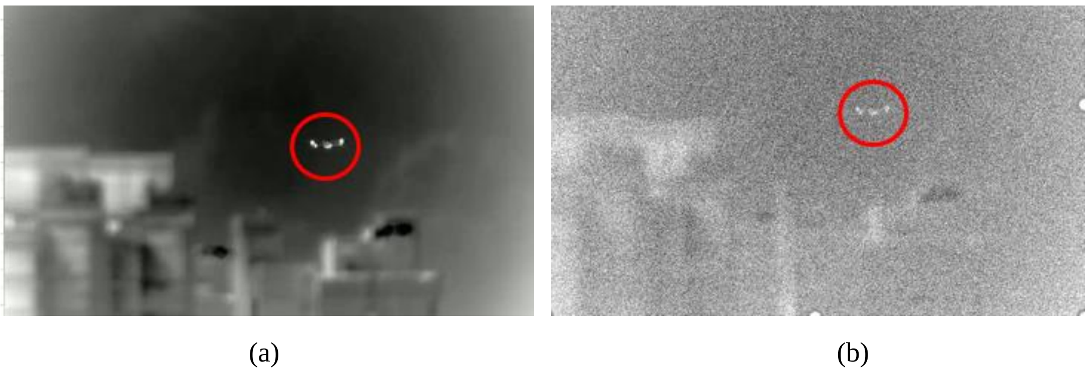

Figure 7: Examples of the data used: (a) original data, (b) data with added thermal noise.

3.2 Parameters for model evaluation

To quantitatively evaluate the performance of the proposed UAVCity-YOLOv13n model in detecting UAV targets on infrared images, we use standard evaluation parameters in object detection problems, including Precision (P), Recall (R), Average Precision (AP), and mean Average Precision (mAP). Since this problem focuses on only one object class, UAVs, the AP and mAP indices will have the same value [33]. P and R values are defined based on the classification results at the specified IoU threshold and are shown in Equations (11) and (12). P is the ratio of the number of correctly predicted UAV targets to the total number of positive predictions. Precision reflects the accuracy of the predictions. R is the ratio of the number of correctly predicted UAV targets to the total number of actual targets. Recall reflects the ability to find all actual targets.

$$\mathrm{P} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} \tag{11}$$

$$\mathrm{R} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \tag{12}$$

where True Positives (TP) is the number of UAV targets correctly detected; False Positives (FP) is the number of incorrect predictions that the model identifies as UAV targets; False Negatives (FN) is the number of actual UAV targets missed by the model. AP is calculated by integrating under the

Precision-Recall (P(R)) curve over the entire Recall range from 0 to 1.

$$\mathrm{AP} = \int_{0}^{1} \mathrm{P(R)dR} \tag{13}$$

Mean Average Precision (mAP) is calculated as the average AP value over all object classes. Since this problem has only one class of UAV (N = 1). The value of $\mathrm{mAP}_{0.5}$ is the mean Average Precision calculated with a fixed IoU threshold of 0.5 ($\mathrm{AP}_{0.5}$). Since N = 1, we have: $\mathrm{mAP}_{0.5} = \mathrm{AP}_{0.5}$. This index is an important standard measure to evaluate the performance of the model.

$$\mathrm{mAP}_{0.5}=\frac{1}{\mathrm{N}}\sum_{\mathrm{k=1}}^{\mathrm{N}}\mathrm{AP_i} \tag{14}$$

$\mathrm{mAP}_{0.5:0.95}$ is calculated by averaging AP over 10 IoU thresholds (Q = 10) increasing from 0.50 to 0.95 (in 0.05 steps).

$$\mathrm{mAP}_{0.5:0.95}=\frac{1}{\mathrm{N}}\sum_{\mathrm{k=1}}^{\mathrm{N}}\frac{1}{\mathrm{Q}}\sum_{\mathrm{q=1}}^{\mathrm{Q}}\mathrm{AP_{k,q}}. \tag{15}$$

Comprehensive evaluation of an object detection model requires analyzing computational performance metrics in addition to accuracy. We evaluate model complexity through the total number of parameters, which directly determines the model size and the amount of memory required. GFLOPs (Giga floating-point operations) accurately measure the computational load required to process an image. Finally, we measure the inference speed (ms), which are key factors to ensure the real-time performance of the model on resource-constrained embedded devices. In this work, the reported inference speed corresponds to the forward pass of the model only, measured per single input image under the hardware environment described in Section 3.3. The preprocessing and postprocessing operations are not included in this timing measurement.

3.3 Experimental environment settings and training model

Regarding the experimental environment, to ensure the reproducibility of our results, all experiments were conducted within a controlled environment. The models were trained and evaluated on a high-performance workstation equipped with an NVIDIA GeForce RTX 4070 Ti Super graphics processing unit (GPU) featuring 16 GB VRAM. The deep learning framework utilized was PyTorch (version 1.13.1), operating on 64-bit Windows 11 and CUDA version 11.8. Moreover, for the training protocol, all models were trained using a consistent set of hyperparameters, to ensure a fair and objective comparison. Key parameters include an input image resolution of 640x640 pixels. Models were trained for 150 epochs, a duration determined sufficient for convergence as illustrated in Figure 8. The initial learning rate was set to 0.01 with the Adam optimizer, the batch size of 16, the weight decay of 0.0005 and the momentum of 0.937.

The training progression and validation performance of the proposed UAVCity-YOLOv13 model are depicted in Figure 8. This figure plots the convergence trajectories for the primary loss functions: box regression (train/box_loss), classification (train/cls_loss), and Distribution Focal Loss (train/dfl_loss), alongside their corresponding validation counterparts. Concurrently, the evaluation metrics, including metrics/precision(B), metrics/recall(B), and mean Average Precision at various IoU thresholds (metrics/$\mathrm{mAP}_{0.5}$ (B), metrics/$\mathrm{mAP}_{0.75}$ (B), and metrics/ $\mathrm{mAP}_{0.5:0.95}$ (B)), demonstrate a stable increase, plateauing around the 150^{th} epoch. This behavior confirms that the model achieved robust convergence without significant overfitting, validating the efficacy and stability of our training strategy.

To further evaluate the robustness and generalization capability of the proposed UAVCity-YOLO model, a five-fold cross-validation (K = 5) experiment was conducted on the thermally noisy infrared UAV dataset. While the conventional train-validation split provides an initial assessment of model performance, the obtained results may be influenced by a particular partition of the dataset. Therefore, K-fold cross-validation was employed to investigate whether the proposed architecture can maintain

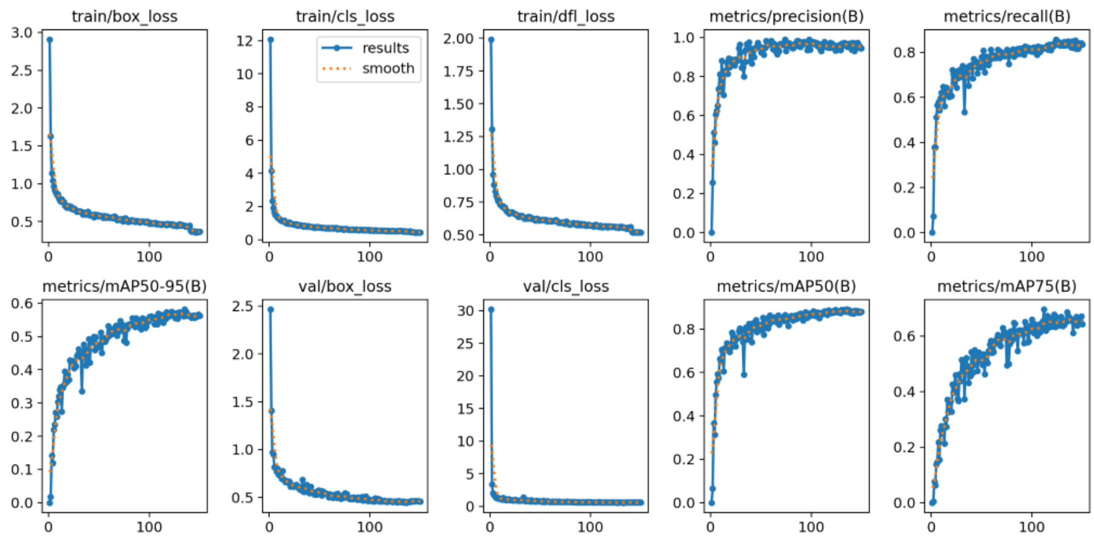

Figure 8: All specific evaluation metrics of UAVCity-YOLOv13n.

stable detection performance under different data distributions. Specifically, the entire dataset was randomly divided into five approximately equal subsets. For each fold, four subsets were used for training and the remaining subset was used for validation. This process was repeated five times, ensuring that every sample was used once for validation and four times for training. The experimental settings, training protocol, and hyperparameters were kept identical to those described in Section 3.3 to guarantee a fair and consistent evaluation. The corresponding results are summarized in Table 1.

Table 1: Performance of UAVCity-YOLO under five-fold cross-validation ($K = 5$).

K-Fold	**mAP$_{0.5}$ (%)**	**mAP$_{0.5:0.95}$ (%)**	**Precision (%)**	**Recall (%)**
1	91.39	61.03	96.19	84.47
2	89.82	59.59	97.85	83.06
3	89.96	61.54	95.57	84.47
4	86.84	56.72	92.94	82.57
5	92.30	59.70	99.00	87.90
Average performance	**90.10**	**59.72**	**96.31**	**84.49**

The results demonstrate that the proposed UAVCity-YOLO model maintains consistently high detection performance across all folds. The mAP$_{0.5}$ values range from 86.84% to 92.30%, with an average value of 90.10%, while the mAP$_{0.5:0.95}$ values vary between 56.72% and 61.54%, yielding an average of 59.72%. Similarly, the Precision and Recall metrics remain stable, achieving average values of 96.31% and 84.49%, respectively. More importantly, the variation among folds remains relatively limited despite the challenging characteristics of the dataset, which contains small UAV targets embedded in complex urban infrared backgrounds with thermal noise. The best performance is achieved in Fold 5, with 92.30% mAP$_{0.5}$, 59.70% mAP$_{0.5:0.95}$, 99.00% Precision, and 87.90% Recall, whereas the lowest performance occurs in Fold 4, where mAP$_{0.5}$ reaches 86.84% and mAP$_{0.5:0.95}$ reaches 56.72%. Despite these variations, all folds consistently maintain mAP$_{0.5}$ values above 86% and Precision values above 92%, indicating that the proposed detector is not overly sensitive to changes in training and validation data partitions. The observed stability can be attributed to the complementary design of the proposed architecture. The AMC-C3k2 module enhances the representation of weak thermal signatures from small UAV targets, the improved Neck structure preserves multi-scale spatial information during feature propagation, the SPPF module enriches contextual feature aggregation, and the proposed IR-NCIoU loss function improves localization robustness under low signal-to-noise conditions. These components jointly contribute to reducing performance fluctuations caused by dataset

Table 2: Results of various ablation experiments.

Component	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6
Improved Neck	–	x	x	x	x	x
SPPF	–	–	x	–	x	x
AMC-C3k2	–	–	–	x	x	x
IR-NCIoU	–	–	–	–	–	x
$mAP_{0.5}$ (%)	84.5	86.6	86.5	86.5	88.4	89.1
$mAP_{0.75}$ (%)	62.5	65.9	66.3	63.2	68	68
$mAP_{0.5:0.95}$ (%)	54.6	56.6	56.7	55.1	57.2	57.9
Inference (ms)	2.6	2.1	1.8	2.3	2.2	1.5
Parameters (M)	2.448	1.677	1.841	1.695	1.859	1.859
Model size (Mb)	5.4	3.7	4.0	3.8	4.1	4.1
GFLOPS	6.2	6.5	6.7	6.6	6.7	6.7

variability and thermal interference. The K-fold cross-validation results provide strong evidence that UAVCity-YOLO exhibits excellent generalization capability and stable learning behavior. The consistently high detection accuracy across different folds confirms that the performance improvements reported in subsequent sections are not the consequence of a favorable data split, but rather originate from the effectiveness of the proposed architectural and optimization strategies. This robustness is particularly important for practical infrared UAV surveillance applications, where environmental conditions and target appearances may vary significantly across deployment scenarios.

3.4 Ablation experiments

3.4.1 Architectural ablation study

To systematically evaluate the contribution and effectiveness of each proposed module, a series of ablation experiments were conducted, and the corresponding quantitative results are summarized in Table 2. Specifically, Model 1 represents the baseline YOLOv13n and Model 2 introduces the improved Neck architecture. Moreover, Model 3 further incorporates the SPPF module, Model 4 replaces the standard C3 block with the proposed AMC-C3K2 module and Model 5 integrates all three structural improvements (improved Neck, SPPF, and AMC-C3k2). Finally, Model 6, which is the proposed model, extends Model 5 by introducing the IR-NCIoU loss function, aiming to further enhance localization accuracy and convergence behavior.

We first analyze the impact of the improved Neck design by comparing Model 2 with the baseline Model 1. The introduction of the redesigned Neck leads to a substantial performance improvement across all accuracy metrics. Specifically, $mAP_{0.5}$ increases from 84.5% to 86.6% (+2.1%), while $mAP_{0.75}$ shows a more pronounced improvement from 62.5% to 65.9% (+3.4%). Moreover, the comprehensive metric $mAP_{0.5:0.95}$ improves from 54.6% to 56.6% (+2.0%), indicating a consistent gain across multiple IoU thresholds. These results confirm that the improved Neck architecture, which emphasizes more effective multi-scale feature fusion, significantly enhances the representation of small and low-resolution UAV targets. Importantly, this accuracy gain is achieved alongside a remarkable reduction in model complexity, with the number of parameters decreasing from 2.45M to 1.68M, demonstrating improved parameter efficiency.

Building upon Model 2, Model 3 further incorporates the Spatial Pyramid Pooling–Fast (SPPF) module to enlarge the receptive field and enhance multi-scale contextual feature aggregation. As shown in Table 2, the introduction of SPPF results in an $mAP_{0.5}$ of 86.5%, which is only 0.1% lower than Model 2 (86.6%). This negligible variation indicates that the inclusion of SPPF does not degrade the overall detection capability at the standard IoU threshold. More importantly, the effect of SPPF becomes evident under stricter localization criteria. Specifically, $mAP_{0.75}$ increases from 65.9% to 66.3%, while $mAP_{0.5:0.95}$ improves from 56.6% to 56.7%. These improvements suggest that SPPF primarily enhances the quality of bounding box regression and spatial localization accuracy, rather than simply increasing detection recall at lower IoU thresholds. This behavior is consistent with the architectural design of SPPF, which aggregates contextual information from multiple receptive fields through sequential pooling operations, thereby allowing the network to capture richer spatial context around small UAV targets. Such contextual cues are particularly beneficial in infrared imagery, where UAVs often appear as small and low-contrast objects embedded within thermally noisy backgrounds. From a computational perspective, Model 3 achieves the fastest inference speed among Models 1–5 (1.8 ms), demonstrating that the SPPF module introduces minimal computational overhead while effectively enlarging the receptive field. This efficiency advantage stems from the serial pooling design of SPPF, which reuses intermediate computations and avoids the heavier parallel pooling operations used in the original SPP module. The resulting balance between improved localization capability and reduced inference latency makes the SPPF-enhanced architecture particularly suitable for real-time UAV detection scenarios, where both detection precision and low-latency processing are critical requirements.

To evaluate the contribution of the attention mechanism, Model 4 integrates the proposed AMC-C3K2 module into Model 2, replacing the conventional DS-C3k2 block and A2C2f block. It should be noted that the AMC attention mechanism is not introduced as a standalone component but is embedded within the DS-C3K2 structure to form the AMC-C3K2 module, as described in Section 2.2.1. Therefore, its contribution is evaluated through the performance change brought by replacing the original DS-C3K2 block and A2C2f block with the AMC-C3k2 module while keeping other architectural components unchanged. Compared with Model 3, Model 4 achieves a similar $mAP_{0.5}$ of 86.5%, but a lower $mAP_{0.5:0.95}$ of 55.1%, suggesting that attention enhancement alone, without complementary multi-scale context modeling, is insufficient for optimal performance. Nevertheless, relative to Model 2, Model 4 maintains stable accuracy while further reducing the number of parameters (1.70M vs. 1.84M in Model 3). This indicates that the AMC-C3K2 module effectively enhances feature discrimination for small UAV targets while preserving model compactness. These results demonstrate that the attention mechanism contributes to improving feature representation for small UAV targets, although its full potential becomes more evident when combined with other architectural improvements such as multi-scale context aggregation and optimized feature fusion, as shown in Model 5.

Model 5 integrates the improved Neck, SPPF, and AMC-C3K2 modules to evaluate their combined effect. As shown in Table 2, this configuration delivers the best overall performance among Models 1–5. Specifically, $mAP_{0.5}$ achieves 88.4%, $mAP_{0.75}$ achieves 68.0%, and $mAP_{0.5:0.95}$ achieves 57.2%, corresponding to improvements of +3.9%, +5.5%, and +2.6%, respectively, over the baseline Model 1. Despite a slight increase in parameters compared to Models 2 and 4, Model 5 still achieves a 24% reduction in parameters relative to the baseline, while maintaining a fast inference speed of 2.2 ms.

Model 6 (UAVCity-YOLO), which represents the proposed model, further incorporates the IR-NCIoU loss function on top of Model 5. This enhancement leads to additional and consistent gains in detection accuracy without increasing model complexity. As shown in Table 2, $mAP_{0.5}$ improves from 88.4% to 89.1%, and $mAP_{0.5:0.95}$ increases from 57.2% to 57.9%, while $mAP_{0.75}$ remains at a high level of 68.0%. Notably, the inference time is further reduced to 1.5 ms, indicating improved optimization and faster convergence during inference. Although the overall GFLOPs increase slightly compared with the baseline model (from 6.2 to 6.7), this increase remains relatively modest and mainly results from the enhanced feature representation mechanisms introduced in the network. In return, the proposed model achieves significant improvements in detection accuracy and localization precision for small UAV targets in thermally noisy infrared environments. From a practical perspective, the model still maintains a compact architecture with significantly fewer parameters (1.859M) and a smaller model size (4.1 MB) compared with the baseline YOLOv13n. Therefore, the proposed model can still be considered lightweight in terms of parameter efficiency and deployment cost. These results demonstrate that the IR-NCIoU loss function effectively enhances bounding box regression accuracy, particularly under higher IoU constraints, thereby complementing the structural improvements of the network. Overall, Model 6 achieves the best trade-off between accuracy, efficiency, and model compactness, validating it as the final and optimal configuration for real-time UAV detection in complex urban environments.

3.4.2 Sensitivity analysis of IR-NCIoU hyperparameters

The proposed IR-NCIoU loss function introduces two key hyperparameters, namely the normalization factor n in the NIoU formulation and the scale-aware center-distance weighting factor β, as defined in Section 2.2.3. Although theoretical analysis suggests that these parameters can enhance gradient responses for low-IoU samples and improve localization sensitivity for small targets, their actual effectiveness depends on the specific characteristics of infrared UAV imagery. Therefore, a sensitivity analysis was conducted to determine the most suitable parameter configuration for thermally noisy urban infrared environments. To ensure a fair evaluation, all experiments were performed using the complete UAVCity-YOLO architecture while varying only the values of n and β. The training protocol, dataset partition, optimizer settings, and all other hyperparameters remained unchanged. The corresponding results are summarized in Table 3. The experimental results indicate that the performance of UAVCity-YOLO is strongly influenced by the choice of the IR-NCIoU hyperparameters. Among all evaluated configurations, the combination of $n = 0.6$ and $\beta = 0.5$ achieves the best overall performance, yielding 89.1% $\mathrm{mAP}_{0.5}$, 57.9% $\mathrm{mAP}_{0.5:0.95}$, 96.5% Precision, and 83.6% Recall.

Table 3: Sensitivity analysis of IR-NCIoU hyperparameters.

Parameters	**$\mathrm{mAP}_{0.5}$ (%)**	**$\mathrm{mAP}_{0.5:0.95}$ (%)**	**Precision (%)**	**Recall (%)**
$n = 0.4;\ \beta = 0.3$	88.0	56.9	95.2	81.6
$n = 0.5;\ \beta = 0.4$	88.4	57.5	99.4	79.6
$n = 0.6;\ \beta = 0.5$	**89.1**	**57.9**	**96.5**	**83.6**
$n = 0.7;\ \beta = 0.6$	86.4	55.7	97.6	79.9
$n = 0.8;\ \beta = 0.7$	88.2	56.1	97.0	83.1

When smaller parameter values are used, such as $n = 0.4$ and $\beta = 0.3$, the detector achieves only 88.0% $\mathrm{mAP}_{0.5}$ and 56.9% $\mathrm{mAP}_{0.5:0.95}$. This behavior suggests that the gradient amplification effect introduced by the normalized IoU formulation is insufficient to effectively optimize low-IoU training samples, which frequently occur in infrared UAV detection due to blurred object boundaries and strong thermal interference. Consequently, the localization capability of the detector remains limited. Increasing the parameters to $n = 0.5$ and $\beta = 0.4$ improves the detection performance, resulting in 88.4% $\mathrm{mAP}_{0.5}$ and 57.5% $\mathrm{mAP}_{0.5:0.95}$. In particular, Precision reaches 99.4%, indicating a very low false-positive rate. However, Recall decreases to 79.6%, implying that the detector becomes relatively conservative and tends to miss a larger number of UAV targets. This trade-off suggests that the localization penalty is still not sufficiently balanced for small infrared targets under thermal noise conditions. The optimal configuration is obtained at $n = 0.6$ and $\beta = 0.5$. Under this setting, the model simultaneously achieves the highest $\mathrm{mAP}_{0.5}$ and $\mathrm{mAP}_{0.5:0.95}$ values while maintaining a balanced Precision–Recall relationship. This result confirms that the selected normalization strength effectively enhances regression gradients for difficult low-IoU samples, whereas the scale-aware center-distance penalty provides adequate localization sensitivity for small UAV targets without introducing optimization instability.

When the parameter values are further increased to $n = 0.7$, $\beta = 0.6$ and $n = 0.8$, $\beta = 0.7$, detection performance deteriorates noticeably. Specifically, $\mathrm{mAP}_{0.5}$ decreases to 86.4% and 88.2%, respectively, while $\mathrm{mAP}_{0.5:0.95}$ drops to 55.7% and 56.1%. Excessively large values of n strengthen the NIoU gradient response too aggressively, making the optimization process more sensitive to noisy samples and potentially leading to unstable regression updates. Meanwhile, larger values of β impose a stronger center-distance penalty, which may over-constrain the localization process when UAV targets occupy only a few pixels in infrared images. The results demonstrate that both hyperparameters have a significant influence on localization accuracy and detection robustness. The sensitivity analysis validates the empirical choice of $n = 0.6$ and $\beta = 0.5$ adopted throughout this paper, as this configuration provides the most favorable balance between localization precision, detection robustness, and optimization stability. Furthermore, the findings support the design motivation of IR-NCIoU, namely that moderate enhancement of low-IoU gradients and scale-aware center-distance normalization are particularly beneficial for detecting small UAV targets embedded in thermally noisy urban infrared backgrounds. The observed performance trend follows a bell-shaped distribution with respect to increasing values of n and β, indicating that both insufficient and excessive penalty strengths can negatively affect optimization. This observation further demonstrates the necessity of carefully balancing localization sensitivity and optimization stability when designing bounding-box regression

Table 4: Performance comparison of the proposed model and other models on Dataset_1.

Model	$\mathbf{mAP}_{0.5}$	$\mathbf{mAP}_{0.5:0.95}$	**Inference (ms)**	**Model size (Mb)**	**Parameters (M)**	**GFLOPS**	**No. of Layers**
RT-DETR [34]	98.5	74.7	5.9	86	41.936	125.6	440
Fast-RCNN [35]	79.8	39.7	11.7	153.9	41.3	133.9	113
SSD [36]	94.8	63.4	8.0	90.6	23.745	30.4	85
CenterNet [37]	89.8	52.3	4.1	53.7	14.041	17.8	76
RetinaNet [38]	82.7	49.6	21.4	123.2	32.169	126.9	121
LW-UAV-YOLOv10 [24]	95.5	72.5	1.8	3.5	1.505	19.5	310
YOLOv11n-UAV [26]	97.4	74.2	3.1	6.0	2.474	21.3	363
YOLOv12n [39]	97.0	73.6	1.6	5.4	2.508	5.8	376
YOLOv13n [28]	95.6	71.4	2.4	5.4	2.448	6.2	535
UAVCity-YOLO	97.4	75.3	1.5	4.1	1.859	6.7	449

losses for small infrared UAV targets.

3.5 Experimental results

To further validate the effectiveness and robustness of the proposed UAVCity-YOLO model, extensive comparative experiments were conducted against a wide range of representative object detection methods. The selected comparison models include classical two-stage detectors (Fast R-CNN), conventional one-stage detectors (SSD, RetinaNet), keypoint based methods (CenterNet), as well as recent lightweight and UAV-oriented YOLO variants. All experiments were carried out under identical hardware and software environments to ensure a fair comparison. Evaluations were performed on two infrared UAV detection datasets with urban backgrounds: Dataset_1 which contains infrared images without additional thermal noise, and Dataset_2 which introduces synthetic thermal noise to simulate more challenging real-world urban infrared sensing conditions. The quantitative results on Dataset_1 and Dataset_2 are reported in Table 4 and Table 5, respectively.

Firstly, for the performance comparison on Dataset_1, Table 4 presents the comparative results on Dataset_1, where the infrared images exhibit relatively clean backgrounds without explicit thermal noise interference. Classical detection frameworks, such as Fast R-CNN and RetinaNet, show limited suitability for infrared small UAV detection. Additionally, the transformer-based detector RT-DETR achieves competitive detection accuracy ($\mathrm{mAP}_{0.5} = 98.5\%$), but its extremely large model size (86 MB) and computational complexity (125.6 GFLOPs) lead to significantly slower inference (5.9 ms), which limits its practicality for lightweight UAV deployment. Although Fast R-CNN benefits from region-based refinement, its reliance on hand-crafted proposals and deep backbone networks leads to inferior localization performance ($\mathrm{mAP}_{0.5:0.95} = 39.7\%$) and prohibitively high computational cost. RetinaNet, despite achieving moderate accuracy, suffers from extremely high inference latency (21.35 ms), rendering it unsuitable for real-time UAV applications. Among traditional one stage detectors, SSD achieves relatively high $\mathrm{mAP}_{0.5}$(94.8%), but its performance drops significantly under stricter IoU thresholds ($\mathrm{mAP}_{0.5:0.95} = 63.37\%$), indicating limited localization precision for small infrared targets. CenterNet demonstrates improved efficiency but remains constrained by its keypoint-based representation, which is less effective for low contrast and small-scale UAV targets. Recent lightweight and UAV-specific YOLO variants exhibit significantly improved performance. Models such as LW-UAV-YOLOv10, YOLOv11n-UAV, YOLOv12n, and YOLOv13n demonstrate strong detection capability while maintaining compact model sizes and fast inference speeds. However, their performance is still bounded by limitations in multi-scale feature aggregation and attention modeling.

In contrast, the proposed UAVCity-YOLO achieves state-of-the-art performance on Dataset_1, with an $\mathrm{mAP}_{0.5}$ of 97.4% and the highest $\mathrm{mAP}_{0.5:0.95}$ of 75.3% among all compared models. Notably, UAVCity-YOLO matches the best $\mathrm{mAP}_{0.5}$ score achieved by YOLOv11n-UAV, while surpassing it by +1.1% in $\mathrm{mAP}_{0.5:0.95}$, demonstrating improved localization accuracy under stricter IoU constraints. From an efficiency perspective, UAVCity-YOLO achieves the fastest inference speed (1.5 ms) together with YOLOv12n, while maintaining a compact model size of only 4.1 MB and 1.86M parameters. Compared with the baseline YOLOv13n, UAVCity-YOLO reduces the number of parameters by ap-

Table 5: Performance comparison of the proposed model and other models on Dataset_2.

Model	$\mathbf{mAP}_{0.5}$	$\mathbf{mAP}_{0.5:0.95}$	Inference (ms)	Model size (Mb)	Parameters (M)	GFLOPS	No. of Layers
RT-DETR [34]	33.8	17.1	5.5	86	41.936	125.6	440
Fast-RCNN [35]	40.5	14.9	17.7	153.9	41.3	133.9	113
SSD [36]	74.6	43.1	15.7	90.6	23.745	30.4	85
CenterNet [37]	69.7	31.3	7.1	53.7	14.041	17.8	76
RetinaNet [38]	82.7	49.6	39.7	123.2	32.169	126.9	121
LW-UAV-YOLOv10 [24]	82.5	55.7	2.6	3.5	1.505	19.5	310
YOLOv11n-UAV [26]	87.1	57.2	3.2	6.0	2.474	21.3	363
YOLOv12n [39]	85.4	55.9	2.2	5.4	2.508	5.8	376
YOLOv13n [28]	84.5	54.6	2.6	5.4	2.448	6.2	535
UAVCity-YOLO (This paper)	89.1	57.9	1.5	4.1	1.859	6.7	449

proximately 24%, while improving $\mathrm{mAP}_{0.5:0.95}$ by +3.9%. These results confirm that the architectural improvements introduced in Section 4 translate into consistent performance gains on clean infrared urban scenes.

Moreover, we analyze the performance comparison on Dataset_2. This dataset represents a more challenging evaluation scenario, where additional thermal noise is introduced to simulate real-world infrared urban environments with cluttered backgrounds and low signal-to-noise ratios. As shown in Table 5, the performance of classical detectors degrades dramatically under thermal noise. The transformer-based detector RT-DETR exhibits particularly severe degradation, with $\mathrm{mAP}_{0.5}$ dropping to 33.8% and $\mathrm{mAP}_{0.5:0.95}$ to only 17.1%. This substantial performance drop contrasts sharply with its strong detection accuracy on Dataset_1, where the infrared images contain relatively clean backgrounds. The results indicate that large-scale transformer architectures struggle to maintain robust feature representations under strong infrared noise conditions. One possible reason lies in the global attention mechanism employed by transformer-based detectors. While global self-attention enables RT-DETR to capture long-range contextual dependencies and achieve high detection accuracy in clean environments, it also makes the model more sensitive to noise distributed across the entire feature map. When thermal noise is introduced, the attention mechanism may inadvertently propagate noisy responses across spatial regions, leading to degraded feature discrimination for small UAV targets. Furthermore, infrared UAV targets are typically characterized by weak contrast, small object scale, and blurred boundaries, which become further obscured under thermal interference. In such cases, the transformer queries in RT-DETR may fail to accurately focus on the true target regions, resulting in unstable object queries and reduced localization precision. Consequently, although RT-DETR achieves competitive performance in relatively clean infrared scenarios (Table 5), its detection robustness deteriorates significantly when confronted with noisy infrared environments, highlighting the limitations of large-scale transformer detectors for small-object UAV detection under thermal disturbances. Fast R-CNN exhibits severe performance deterioration, with $\mathrm{mAP}_{0.5}$ dropping to 40.5% and $\mathrm{mAP}_{0.5:0.95}$ to only 14.9%, highlighting the inability of two stage pipelines to robustly handle low contrast, small-sized infrared targets. SSD and CenterNet also experience significant accuracy degradation, particularly in localization precision. Although RetinaNet maintains a relatively high $\mathrm{mAP}_{0.5}$(82.7%), its inference time increases drastically to 39.7 ms, making it impractical for real-time deployment. Lightweight YOLO variants demonstrate better robustness to the thermal noise, with YOLOv11n-UAV and YOLOv12n achieving $\mathrm{mAP}_{0.5}$ values above 85%. However, their localization accuracy and inference efficiency remain limited.

The proposed UAVCity-YOLO consistently outperforms all competing models on Dataset_2. It achieves the highest $\mathrm{mAP}_{0.5}$ of 89.1% and the highest $\mathrm{mAP}_{0.5:0.95}$ of 57.9%, surpassing the baseline YOLOv13n by +4.6% and +3.3%, respectively. Compared with the UAV-specific YOLOv11n-UAV, UAVCity-YOLO achieves a +2.0% improvement in $\mathrm{mAP}_{0.5}$, confirming its improved robustness against thermal noise. Importantly, UAVCity-YOLO maintains an inference time of only 1.5 ms, which is the fastest among all evaluated models, while preserving a compact model size and low computational complexity. This demonstrates that the proposed improvements namely the enhanced Neck architecture, SPPF-based receptive field expansion, AMC-C3k2 attention mechanism, and IR-

NCIoU loss jointly contribute to improved noise robustness and precise localization without sacrificing real-time performance.

The comparative experiments on both datasets clearly demonstrate that UAVCity-YOLO achieves the optimal trade-off between detection accuracy, robustness, and computational efficiency. The consistent performance gains observed across clean and thermally noisy infrared datasets validate the design choices motivated by the ablation studies. In particular, the proposed model effectively addresses the key challenges of infrared UAV detection in urban environments, including low contrast, limited target pixels, background clutter, and real-time deployment constraints. These results indicate that UAVCity-YOLO is a promising approach for infrared UAV detection, providing a favorable balance between detection accuracy, robustness, and computational efficiency.

For the visualization results, Figure 9 illustrates the performance of the proposed UAVCity-YOLO model. Here, Figure 9(a) represents the Ground Truth (labels), Figure 9(b) shows the results from the baseline YOLOv13n model, and Figure 9(c) is the results of the proposed UAVCity-YOLO model. From Figure 9, we can observe a significant difference in the performance of the two models when running through a sequence of 16 images. Specifically, the YOLOv13n model has obvious limitations: it misses two targets (marked with 'x') and makes three false detections. In contrast, the proposed UAVCity-YOLO model shows a remarkable improvement in accuracy and stability. Despite missing one targets (marked with an 'x'), the model completely eliminated false positives while maintaining high reliability and stability in target detection throughout the image sequence. This demonstrates the improved ability of UAVCity-YOLO to reduce noise and focus on the real features of UAVs in complex urban environments.

To further analyze the feature representation capability of the proposed model, heatmaps were generated for four challenging infrared UAV samples selected from the failure cases highlighted in Figure 9. These samples correspond to scenarios with weak thermal signatures, strong background clutter, or severe thermal noise. The corresponding activation maps are presented in Figure 10. As shown in Figure 10, the baseline YOLOv13n model exhibits relatively dispersed activation responses, with several high-response regions appearing on background structures rather than being concentrated on the UAV targets. This indicates that the extracted features are influenced by surrounding thermal clutter, making it more difficult to distinguish weak UAV signatures from complex urban backgrounds. Such behavior is consistent with the false detections and missed targets observed in Figure 9(b). In contrast, the heatmaps produced by UAVCity-YOLO show activation responses that are more concentrated around the actual UAV locations, while irrelevant background regions are largely suppressed. For all four challenging samples, the model maintains clear responses at the target positions, even when the UAV occupies only a small number of pixels and exhibits low thermal contrast. These results suggest that the proposed model is able to preserve target-related information more effectively and reduce the influence of thermal noise during feature extraction.

This behavior can be attributed to the combined effect of the proposed improvements. The AMC-C3k2 module enhances the representation of small infrared targets, the improved Neck structure helps preserve weak spatial cues during feature fusion, the SPPF module enriches contextual information, and the IR-NCIoU loss improves localization quality during training. Together, these components contribute to more target-focused feature responses. The heatmap analysis provides visual evidence supporting the quantitative results reported in Tables 4. Compared with YOLOv13n, UAVCity-YOLO demonstrates a stronger ability to focus on UAV-related thermal features while suppressing background interference, which contributes to more stable detection performance in noisy urban infrared environments.

4 Conclusion

This paper addresses the challenging task of UAV detection by infrared data in complex urban environments, where low target contrast, limited spatial resolution, background clutter, and strict real-time constraints severely degrade the detection reliability. To overcome these challenges, we proposed UAVCity-YOLO, a lightweight and high performance single-stage detection framework built upon the YOLOv13n architecture and specifically tailored for urban infrared UAV scenarios. The

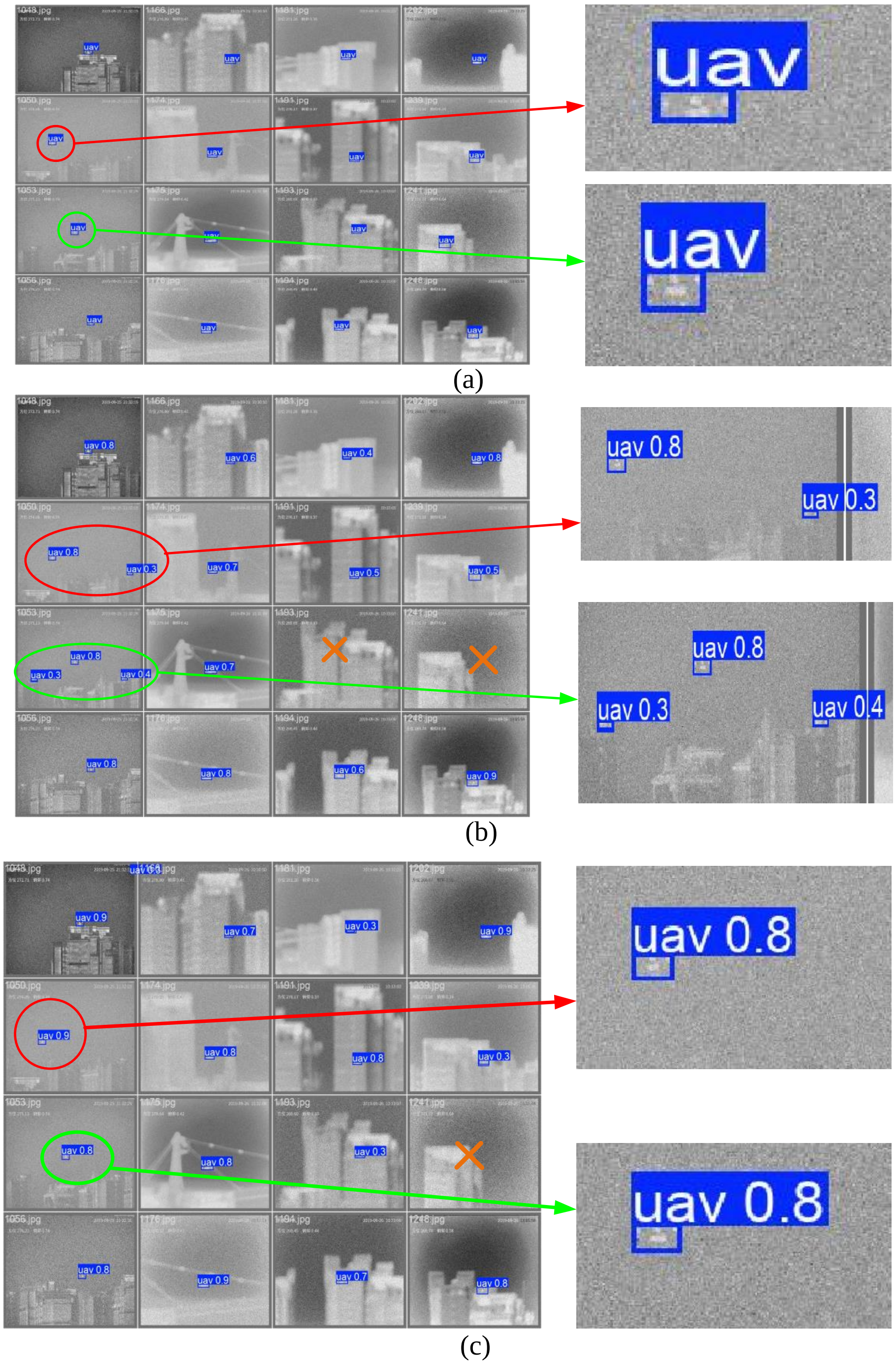

Figure 9: Results of models on 16 images: (a) Ground truth, (b) YOLOv13n model, (c) UAVCity-YOLO model.

proposed method integrates an optimized multi-scale feature fusion Neck structure, an SPPF-based receptive field enhancement module, an attention enhanced AMC-C3K2 block for robust feature discrimination, and a noise-aware IR-NCIoU loss function to improve localization accuracy under low signal-to-noise conditions. Extensive experiments on clean and thermally noisy urban infrared UAV datasets have demonstrated that UAVCity-YOLO consistently outperforms representative two-stage detectors, conventional one-stage methods, and recent UAV-oriented YOLO variants, achieving state-of-the-art $mAP_{0.5}$ and $mAP_{0.5:0.95}$ while maintaining fast inference speed (1.5 ms) and a compact model size of only 4.1 MB. Compared with the baseline YOLOv13n model, the proposed approach can reduce model parameters by approximately 24% while significantly improving localization precision and robustness to thermal noise. Qualitative results further confirmed its improved stability and

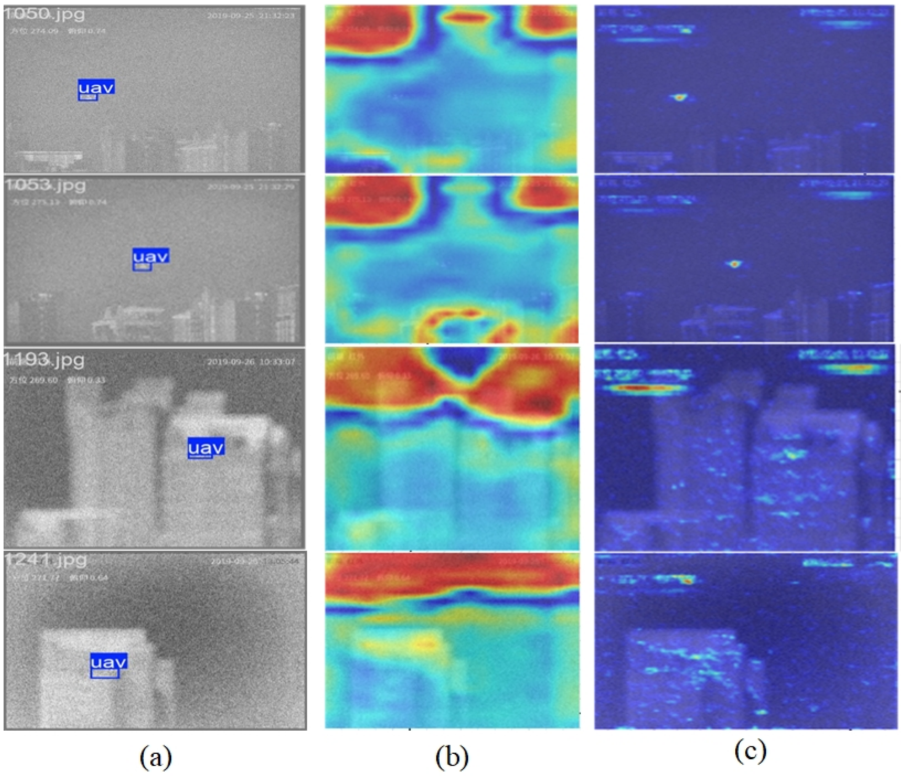

Figure 10: Heatmaps: (a) Ground truth, (b) YOLOv13n model, (c) UAVCity-YOLO model.

reliability in sequential infrared detection scenarios, with notably fewer missed detections and false positives.

Future work will focus on improving model robustness and generalization under real-world infrared conditions through adaptive noise modeling and domain generalization techniques. More realistic thermal disturbance models, including sensor non-uniformity, atmospheric turbulence, dynamic heat sources, reflection effects, and structured background clutter, will be investigated to better approximate operational infrared environments. Furthermore, larger and more diverse infrared UAV datasets containing real sensor noise and varying environmental conditions will be developed, together with data-driven noise augmentation strategies, to bridge the gap between synthetic noise simulation and real-world infrared imaging. Finally, the proposed anti-drone detection framework is intended for legitimate security applications under appropriate human supervision. Its deployment should comply with applicable legal requirements, privacy regulations, and ethical principles to ensure responsible and transparent use of AI-assisted surveillance technologies.

Acknowledgment

This research is funded by Vietnam National Foundation for Science and Technology Development (NAFOSTED) under grant number 102.02-2023.06.

Conflict of interest

The authors declare no conflict of interest.

References

[1] S. O. De Macedo, M. Caetano, and R. M. da Costa, "Drone detection in airport environments: A literature review," *Array*, vol. 28, Art. no. 100511, 2025.

[2] S. Park, H. T. Kim, S. Lee, H. Joo, and H. Kim, "Survey on anti-drone systems: Components, designs, and challenges," *IEEE Access*, vol. 9, pp. 42635–42659, 2021.

[3] Flughafen München GmbH, "Drone sighting at Munich Airport," [Online]. Available: https://www.munich-airport.com/press-drone-sighting-at-munich-airport-35709068. Accessed on: Oct. 3, 2025.

[4] G. Ding, Y. Ren, Y. Liu, Q. Zhao, and S. Li, "Vision-Based Anti-Unmanned Aerial Technology: Opportunities and challenges," *IEEE Geosci. Remote Sens. Mag.*, vol. 13, no. 4, pp. 382–405, 2025.

[5] Y. Sun, S. Abeywickrama, L. Jayasinghe, C. Yuen, J. Chen, and M. Zhang, "Micro-Doppler signature-based detection, classification, and localization of small UAV with long short-term memory neural network," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 8, pp. 6285–6300, 2020.

[6] M. H. Rahman, M. A. S. Sejan, M. A. Aziz, R. Tabassum, J. I. Baik, and H. K. Song, "A comprehensive survey of unmanned aerial vehicles detection and classification using machine learning approach: Challenges, solutions, and future directions," *Remote Sens.*, vol. 16, no. 5, Art. no. 879, 2024.

[7] M. F. Al-Sa'd, A. Al-Ali, A. Mohamed, T. Khattab, and A. Erbad, "RF-based drone detection and identification using deep learning approaches: An initiative towards a large open source drone database," *Future Gener. Comput. Syst.*, vol. 100, pp. 86–97, 2019.

[8] M. Z. Anwar, Z. Kaleem, and A. Jamalipour, "Machine learning inspired sound-based amateur drone detection for public safety applications," *IEEE Trans. Veh. Technol.*, vol. 68, no. 3, pp. 2526–2534, 2019.

[9] O. H. Anidjar, A. Barak, B. Ben-Moshe, E. Hagai, and S. Tuvyahu, "A stethoscope for drones: Transformers-based methods for UAVs acoustic anomaly detection," *IEEE Access*, vol. 11, pp. 33336–33353, 2023.

[10] I. Aydin and E. Kizilay, "Development of a new light-weight convolutional neural network for acoustic-based amateur drone detection," *Appl. Acoust.*, vol. 193, Art. no. 108773, 2022.

[11] I. Garvanov, M. Garvanova, V. Ivanov, A. Lazarov, D. Borissova, and T. Kostadinov, "Detection of unmanned aerial vehicles based on image processing," in *Proc. Int. Conf. Telecommunications and Remote Sensing*, Cham, Switzerland: Springer Nature, 2022, pp. 37–50.

[12] F. Samadzadegan, F. Dadrass Javan, F. Ashtari Mahini, and M. Gholamshahi, "Detection and recognition of drones based on a deep convolutional neural network using visible imagery," *Aerospace*, vol. 9, no. 1, Art. no. 31, 2022.

[13] G. Zhou, X. Liu, and H. Bi, "Recognition of UAVs in infrared images based on YOLOv8," *IEEE Access*, vol. 13, pp. 1534–1545, 2025.

[14] Y. Chen, H. Sun, L. Tian, Y. Yang, S. Wang, and T. Wang, "Detecting infrared UAVs on edge devices through lightweight instance segmentation," *PLOS ONE*, vol. 20, no. 8, Art. no. e0330074, 2025.

[15] A. Hommes, A. Shoykhetbrod, D. Noetel, S. Stanko, M. Laurenzis, S. Hengy, and F. Christnacher, "Detection of acoustic, electro-optical and RADAR signatures of small unmanned aerial vehicles," in *SPIE Target and Background Signatures II*, vol. 9997, Art. no. 999701, 2016.

[16] X. Shi, C. Yang, W. Xie, C. Liang, Z. Shi, and J. Chen, "Anti-drone system with multiple surveillance technologies: Architecture, implementation, and challenges," *IEEE Commun. Mag.*, vol. 56, no. 4, pp. 68–74, 2018.

[17] D. Chauhan, H. Kagathara, H. Mewada, S. Patel, S. Kavaiya, and G. Barb, “Nation’s Defense: A Comprehensive Review of Anti-Drone Systems and Strategies,” *IEEE Access*, vol. 13, pp. 53476–53505, 2025.

[18] W. Chen, Y. Li, Z. Tian, and F. Zhang, “2D and 3D object detection algorithms from images: A survey,” *Array*, vol. 19, Art. no. 100305, 2023.

[19] H. Wang *et al.*, “Geometry-aware 3D point cloud learning for precise cutting-point detection in unstructured field environments,” *J. Field Robot.*, vol. 42, no. 7, pp. 3063–3076, 2025.

[20] Z. Kaleem, “Lightweight and Computationally Efficient YOLO for Rogue UAV Detection in Complex Backgrounds,” *IEEE Trans. Aerosp. Electron. Syst.*, vol. 61, no. 2, pp. 5362–5366, 2025.

[21] M. Misbah *et al.*, “MSF-GhostNet: Computationally Efficient YOLO for Detecting Drones in Low-Light Conditions,” *IEEE J. Sel. Topics Appl. Earth Obs. Remote Sens.*, vol. 18, pp. 3840–3851, 2025.

[22] M. Misbah, F. A. Orakazi, L. Tanveer, Z. Kaleem, and C. Yuen, “TF-BiFPN Improves YOLOv5: Enhancing Small-Scale Multiclass Drone Detection in Dark,” *IEEE Trans. Aerosp. Electron. Syst.*, vol. 61, no. 2, pp. 5354–5361, 2025.

[23] S. Yuan, B. Sun, Z. Zuo *et al.*, “IRSDD-YOLOv5: Focusing on the infrared detection of small drones,” *Drones*, vol. 7, no. 6, Art. no. 393, 2023.

[24] P. T. Nguyen, G. L. Nguyen, and D. D. Bui, “LW-UAV–YOLOv10: A lightweight model for small UAV detection on infrared data based on YOLOv10,” *Geomatica*, vol. 77, no. 1, Art. no. 100049, 2025.

[25] N. T. Phat, N. L. Giang, and B. D. Duy, “GAN-UAV-YOLOv10s: Improved YOLOv10s network for detecting small UAV targets in mountainous conditions based on infrared image data,” *Neural Comput. Appl.*, pp. 1–13, 2025.

[26] P. T. Nguyen and L. H. Nguyen, “YOLOv11n-UAV: Improved YOLOv11n model for detecting small UAVs using infrared images on complex backgrounds,” *Neural Comput. Appl.*, pp. 1–17, 2025.

[27] L. T. Ramos and A. D. Sappa, “A decade of You Only Look Once (YOLO) for object detection: A review,” *IEEE Access*, vol. 13, pp. 192747–192794, 2025.

[28] M. Lei, S. Li, Y. Wu, H. Hu, Y. Zhou, X. Zheng, *et al.*, “YOLOv13n: Real-time object detection with hypergraph-enhanced adaptive visual perception.,” arXiv preprint arXiv:2506.17733, 2025.

[29] https://github.com/iMoonLab/YOLOv13n

[30] Y. Li, M. Zhang, C. Zhang, H. Liang, P. Li, and W. Zhang, “YOLO-CCS: Vehicle detection algorithm based on coordinate attention mechanism,” *Digital Signal Process.*, vol. 153, Art. no. 104632, 2024.

[31] H. K. Jooshin, M. Nangir, and H. Seyedarabi, “Inception-YOLO: Computational cost and accuracy improvement of the YOLOv5 model based on employing modified CSP, SPPF, and inception modules,” *IET Image Process.*, vol. 18, no. 8, pp. 1985–1999, 2024.

[32] W. Liu and W. Lin, “Additive white Gaussian noise level estimation in SVD domain for images,” *IEEE Trans. Image Process.*, vol. 22, no. 3, pp. 872–883, 2013.

[33] J. Davis and M. Goadrich, “The relationship between Precision-Recall and ROC curves,” in *Proc. 23rd Int. Conf. Mach. Learn. (ICML)*, pp. 233–240, 2016.

[34] Y. Zhao *et al.*, "DETRs Beat YOLOs on Real-time Object Detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Seattle, WA, USA, 2024, pp. 16965–16974.

[35] X. Wang, A. Shrivastava, and A. Gupta, "A-Fast-RCNN: Hard positive generation via adversary for object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2017, pp. 2606–2615.

[36] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C. Y. Fu, and A. C. Berg, "SSD: Single shot multibox detector," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Cham, Switzerland: Springer, 2016, pp. 21–37.

[37] K. Duan, S. Bai, L. Xie, H. Qi, Q. Huang, and Q. Tian, "CenterNet: Keypoint triplets for object detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, 2019, pp. 6569–6578.

[38] Y. Li and F. Ren, "Light-weight RetinaNet for object detection," arXiv preprint arXiv:1905.10011, 2019.

[39] Y. Tian, Q. Ye, and D. Doermann, "YOLOv12: Attention-centric real-time object detectors," arXiv preprint arXiv:2502.12524, 2025.

Cite this paper as:

Nguyen, P.-T.; Hoang, V.-P. (2026). UAVCity-YOLO: An Improved YOLOv13 Architecture for Robust UAV Detection Using Infrared Data in Noisy Urban Environments, *International Journal of Computers Communications* & *Control*, 21(5), 7542, 2026.
https://doi.org/10.15837/ijccc.2026.5.7542