Computers, Materials & Continua

Tech Science Press

DOI: 10.32604/cmc.2024.052725

ARTICLE

Robust and Discriminative Feature Learning via Mutual Information Maximization for Object Detection in Aerial Images

Xu Sun, Yinhui Yu* and Qing Cheng

School of Communication Engineering, Jilin University, Changchun, 130012, China

*Corresponding Author: Yinhui Yu. Email: yuyh@jlu.edu.cn

Received: 12 April 2024 Accepted: 06 August 2024 Published: 12 September 2024

ABSTRACT

Object detection in unmanned aerial vehicle (UAV) aerial images has become increasingly important in military and civil applications. General object detection models are not robust enough against interclass similarity and intraclass variability of small objects, and UAV-specific nuisances such as uncontrolled weather conditions. Unlike previous approaches focusing on high-level semantic information, we report the importance of underlying features to improve detection accuracy and robustness from the information-theoretic perspective. Specifically, we propose a robust and discriminative feature learning approach through mutual information maximization (RD-MIM), which can be integrated into numerous object detection methods for aerial images. Firstly, we present the rank sample mining method to reduce underlying feature differences between the natural image domain and the aerial image domain. Then, we design a momentum contrast learning strategy to make object features similar to the same category and dissimilar to different categories. Finally, we construct a transformer-based global attention mechanism to boost object location semantics by leveraging the high interrelation of different receptive fields. We conduct extensive experiments on the VisDrone and Unmanned Aerial Vehicle Benchmark Object Detection and Tracking (UAVDT) datasets to prove the effectiveness of the proposed method. The experimental results show that our approach brings considerable robustness gains to basic detectors and advanced detection methods, achieving relative growth rates of 51.0% and 39.4% in corruption robustness, respectively. Our code is available at https://github.com/cq100/RD-MIM (accessed on 2 August 2024).

KEYWORDS

Aerial images; object detection; mutual information; contrast learning; attention mechanism

1 Introduction

Unmanned aerial vehicles (UAVs) collecting high-resolution images have a wide range of applications in many fields, such as traffic management, disaster response, and infrastructure inspection [1,2]. Using object detection approaches to understand and analyze aerial images has attracted significant attention [3]. In recent years, deep neural networks (DNNs) have developed rapidly in object detection, which achieved superior performance on natural images [4].

Directly applying these detection models to aerial images still has two problems. On the one hand, interclass similarity and intraclass variability are more severe on aerial image datasets, impairing the accuracy of object detection [5,6]. Fig. 1a intuitively illustrates that some objects indeed have confusing appearances and low distinctiveness. On the other hand, UAVs have complex operating environments, such as varying light, motion blur, and uncontrolled weather conditions, which make object detection even harder [7]. As depicted in Fig. 1b, changing illumination drastically affects object visibility. To deal with these problems, a great amount of effort has been carried out [8,9]. Nevertheless, these efforts often ignore the importance of underlying features. For example, the dual sampler and head detection network (DSHNet) [8] designed class-biased samplers and bilateral box heads from the perspective of high-level semantic information to process semantically similar objects. Although it allowed detectors to perform better for each category on clean aerial images, the robustness was poor in highly variable capturing conditions.

(a) (b)

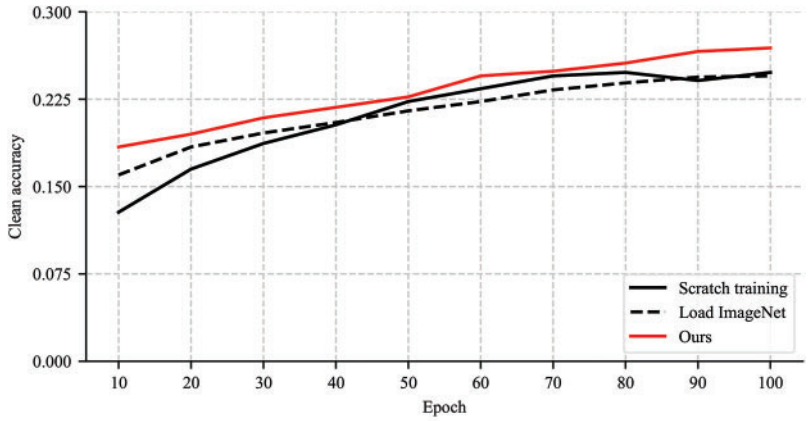
Figure 1: Intuitive explanation of the two problems in object detection for aerial images. (a) Interclass similarity and intraclass variability. (b) UAV complex operating environments (daytime *vs*. nighttime)

Transfer learning available everywhere is an effective scheme to learn underlying features. It not only accelerates model convergence but also tremendously improves model robustness [10]. Common transfer learning methods use the mature backbone trained on the ImageNet dataset [11] as initialization and freeze part of the layers to retain more robust underlying features [12]. When we consider transfer learning from the ImageNet backbone to object detection models in aerial images, the obtained results are not as desirable as described in the leading papers [10,13]. As shown in Fig. 2, in the case of excluding model overfitting, the accuracy of the loading ImageNet method is similar to training from scratch, and robustness performance is even worse with the increase of iterations. We assume the reason for this phenomenon is that the underlying features are poorly learned due to the significant differences between natural images and aerial images. The differences may stem from the different heights and angles captured by UAV cameras, leading to smaller and denser objects in aerial images.

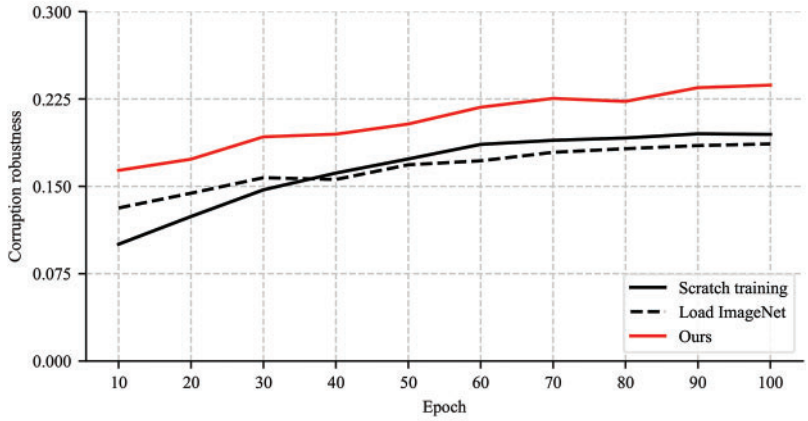
To this end, we propose a robust and discriminative underlying feature learning approach according to the mutual information maximization principle, which is able to assist object detectors for aerial images to further improve robustness rather than achieve state-of-the-art performance. Similar to the robustness measure method in natural images [13], we also add various corruption types to aerial image datasets according to the UAV operation environments. The experiments are conducted on clean and corrupted VisDrone [14] and Unmanned Aerial Vehicle Benchmark Object Detection and Tracking (UAVDT) [15] datasets. Fig. 2 intuitively shows the great performance gains of our approach, especially on corrupted aerial images. Our code is available at https://github.com/cq100/RD-MIM (accessed on 2 August 2024). The main contributions of this paper are as follows:

- We report from the information-theoretic perspective that effective underlying feature learning for aerial image object detection resists pervasive nuisances and reduces interclass similarity and intraclass variability.
- We propose a robust and discriminative feature learning approach based on mutual information maximization (RD-MIM). Our method consists of the rank sample mining, the momentum contrast learning strategy, and a flexible global attention mechanism based on transformer encoder.
- Our RD-MIM is a generic method that can be easily coupled with various backbones and detectors. The RD-MIM brings considerable performance gains in terms of clean accuracy and corruption robustness to advanced aerial object detection methods.

The remaining sections of this paper are organized as follows. In Section 2, the related work is reviewed. In Section 3, the proposed method is illustrated in detail. In Section 4, experimental datasets and implementation details are presented, and related results are analyzed. Finally, in Section 5, the work of this paper is summarized.

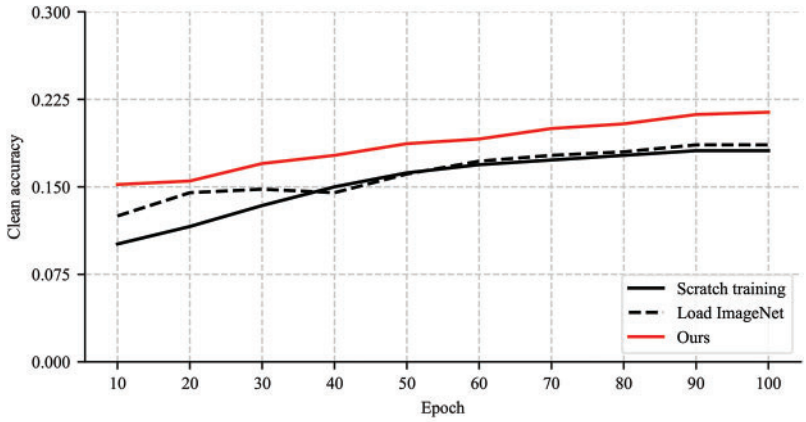


(a)



(b)

Figure 2: (Continued)



(c)

Figure 2: Performance comparison of training from scratch, loading ImageNet, and our method. (a) Accuracy measure on the clean VisDrone validation dataset. (b) Robustness measure on the corrupted VisDrone validation dataset. (c) Accuracy measure on the clean VisDrone testing dataset

2 Related Work

2.1 General Object Detection

With the introduction of region-based convolutional neural network (R-CNN) model, DNN-based networks have played a central role in general object detection tasks [16]. There are primarily two streams of object detection approaches: anchor-free and anchor-based detectors [17]. Anchor-free detectors use the key points to estimate the predicted bounding box, such as fully convolutional one-stage object detection (FCOS) [18]. Anchor-based detectors locate object positions by generating anchor candidates, which can be further divided into two-stage and one-stage detectors [19]. For example, compared with the representative one-stage method RetinaNet [20], Faster R-CNN [21] involves a region proposal network (RPN) to search for possible object locations. We measured the proposed method on three different types of detectors.

From the component perspective, DNN-based object detectors generally consist of three parts: the backbone, the neck, and the head [22]. The backbone plays a key role, which has strong feature extraction capability, such as the traditional ResNet50 [23], the popular RegNet [24], and the latest PVTv2 (pyramid vision transformer version 2) [25]. RegNet can balance accuracy and computation amount well, which is conducive to designing lightweight networks. PVTv2 improves the current promising pyramid vision transformer to establish a more robust and flexible baseline. We use the three networks with different structures as the backbone of detectors.

2.2 Object Detection in Aerial Images

Compared with natural images, object detection for aerial images brings greater challenges. Interclass similarity and intraclass variability degrade the performance of general detectors. Huang et al. [26] constructed a multi-proxy detection network by analyzing object distribution for aerial images to deal with serious category confusion. Wang et al. [27] used a class-aware strategy to maintain a balance in the number of objects from different categories, which facilitated the distinction of confused objects. Tang et al. [5] designed the points estimated network based on a coarse-to-fine

framework to predict the positions of similar objects. These customized approaches are limited due to non-universality, which cannot be effectively applied to common object detection methods.

Object detection methods for aerial images require detectors with strong robustness due to UAV complex working conditions [28]. Existing potential solutions generally can be divided into three aspects [29,30]. First, adversarial training frameworks are utilized to obtain robustness gains. Wu et al. [31] used UAV meta-data in conjunction with the detector based on adversarial learning to extract domain-robust features. Second, the expert head is designed on the top of the model for each specific domain, such as capturing time and weather [29]. As more domains are involved, model ensemble becomes difficult, and it cannot be generalized to unseen domains. Last, data augmentation is a typical strategy to enhance the robustness of detectors. Dense and small object detection (DSDet) [9] used data augmentation based on crop functionality to increase training data, achieving robust detection in aerial images.

Besides the aforementioned methods, there exist other advanced techniques for aerial object detection. Blas et al. [32] designed a swimming pool detection platform by utilizing geographic information tools and common machine learning methods. Xie et al. [33] presented a plug-and-play object activation network (OAN) to help detectors focus on patches containing objects in aerial images. Task-wise sampling convolutions (TS-Conv) [34] guided dynamic label assignment by adaptively sampling features from sensitive areas for localization and classification, boosting detector adaptation and resistance.

2.3 Transfer Learning in Object Detection for Aerial Images

Using transfer learning to improve the performance of DNN-based models is a common practice, because underlying features are close to each other when the input domains of the tasks are similar [35,36]. Specifically, models transfer the structure or parameters from the pre-trained network trained on the ImageNet dataset available publicly [37]. Generally speaking, there exist three transfer learning methods. The first is full-network transfer only using the pre-trained network as initialization, which easily erases the robustness property during fine-tuning. The second is fixed-feature transfer, which freezes the whole backbone and only trains the detection head. It performs slightly worse, but facilitates robustness transfer [13]. The third is a common approach that fixes part of the pre-trained backbone and loads parameters as initialization, which is usually applied for object detection models in aerial images [22,38].

We find that directly transferring ImageNet backbone to detectors for aerial images yields poor performance gains. The possible reason is that the complex capturing conditions of UAVs make the differences between natural images and aerial images larger, which leads to fewer similar underlying features.

3 Methodology

3.1 Theoretical Analysis

Many approaches have successfully learned the representations through the mutual information maximization principle [39]. Inspired by this, we attempt to solve the poor robustness of detectors to object categorical similarity and UAV nuisances from the information-theoretic perspective. To be specific, we compute the mutual information between the underlying features extracted from different backbones and corresponding categories. In practice, it is intractable to maximize mutual information, so this problem is usually translated into estimating a lower bound.

Assuming that B_0 is the backbone trained on the ImageNet dataset, B_1 is the enhanced backbone using the momentum contrast learning strategy and rank sample mining method. Let (x_1, y_1) denote an example pair, where x_1 is an object or background area on the aerial image, and y_1 is the corresponding category. According to the InfoNCE [39] lower bound, we can estimate the mutual information I as follows:

$$I\left(B_0\left(x_1\right), y_1\right) \geq \underset{\zeta}{E}\left[\log \frac{f_\theta\left(B_0\left(x_1\right), y_1\right)}{\sum\limits_{y' \in \zeta} f_\theta\left(B_0\left(x_1\right), y'\right)}\right] + \log \zeta \tag{1}$$

$$I\left(B_1\left(x_1\right), y_1\right) \geq \underset{\zeta}{E}\left[\log \frac{f_\theta\left(B_1\left(x_1\right), y_1\right)}{\sum\limits_{y' \in \zeta} f_\theta\left(B_1\left(x_1\right), y'\right)}\right] + \log \zeta \tag{2}$$

where f_θ is a function used to score the correlation of $B(x_1)$ and y_1, and ζ is a category collection of objects and backgrounds on the aerial image dataset.

The backbone B_0 trained on the ImageNet dataset uses cross-entropy loss L_{in}, which is defined as

$$L_{\text{in}} = -\log \frac{\exp\left(\varphi_\theta\left[B_0\left(u_1\right)\right]^{\mathrm{T}} v_1\right)}{\sum\limits_{v' \in V} \exp\left(\varphi_\theta\left[B_0\left(u_1\right)\right]^{\mathrm{T}} v'\right)} \tag{3}$$

where (u_1, v_1) is an example pair from the ImageNet dataset, V is a category collection, and φ_θ is a mapping function. By minimizing the loss function L_{in}, the backbone B_0 only can indirectly increase the correlation of $B_0(x_1)$ and y_1 due to the differences between natural images and aerial images.

The enhanced backbone B_1 with the momentum contrast learning is trained on the set of object and background chips from the aerial image through the proposed rank sample mining method. Given category y_1, B_1 learns to minimize the L_{ce} and L_{mc} losses of the sample x_1

$$L_{\text{ce}} = -\log \frac{\exp\left(g_\theta\left[B_1\left(x_1\right)\right]^{\mathrm{T}} y_1\right)}{\sum\limits_{y' \in \zeta} \exp\left(g_\theta\left[B_1\left(x_1\right)\right]^{\mathrm{T}} y'\right)} \tag{4}$$

$$L_{\text{mc}} = -\log \frac{\exp\left(B_1\left(x_1\right)^{\mathrm{T}} B_1\left(x_+\right)/\tau\right)}{\sum\limits_{y' \in \zeta} \exp\left(B_1\left(x_1\right)^{\mathrm{T}} B_1\left(x_{y'}\right)/\tau\right)} \tag{5}$$

where g_θ is a mapping function, $B_1(x_+)$ is the positive representation, and τ is a temperature parameter.

By optimizing Eqs. (4) and (5), the enhanced backbone B_1 can learn more similar features to the size and viewing angle of aerial objects, which directly maximizes the lower bound of $I(B_1(x_1), y_1)$. Considering that the backbone B_0 can learn generalized underlying features from a vast number of natural images, our enhanced backbone B_1 adopts the backbone B_0 parameters as initialization. Fig. 2 also verifies the fact that the loading ImageNet method has better accuracy and robustness compared to the training from scratch in the early stage of network training.

Based on the above analysis, it can be concluded that $I(B_1(x_1), y_1)$ has a larger lower bound on mutual information than $I(B_0(x_1), y_1)$. It demonstrates that the enhanced backbone B_1 learns more robust and discriminative underlying features than B_0. That is, the enhanced backbone B_1 can provide more benefits for object detection in aerial images. This conclusion also motivates us to design an efficient robust and discriminative feature learning approach.

3.2 Overview Framework

The proposed RD-MIM approach can efficiently learn robust and discriminative features, which is available for almost all detectors. The framework of the proposed method based on Faster R-CNN with feature pyramid network (FPN) [40] is depicted in Fig. 3. Different from the original Faster R-CNN, we design the rank sample mining method, the momentum contrast learning strategy, and a GATM module. Firstly, the priorities of object categories are ranked on aerial image datasets. Upon the priorities, we utilize them and the positions of the ground truth anchors to mine object chips and background chips. Then, these chips are used to fine-tune the backbone that loads the parameters of a pre-trained ImageNet network. During fine-tuning, the backbone utilizes the momentum contrast learning to improve the capability to distinguish object features. Finally, we directly transfer the enhanced backbone to detectors and integrate the GATM module to efficiently extract global feature information, which flows to each feature map through the FPN. The RPN and detection head are the same as Faster R-CNN.

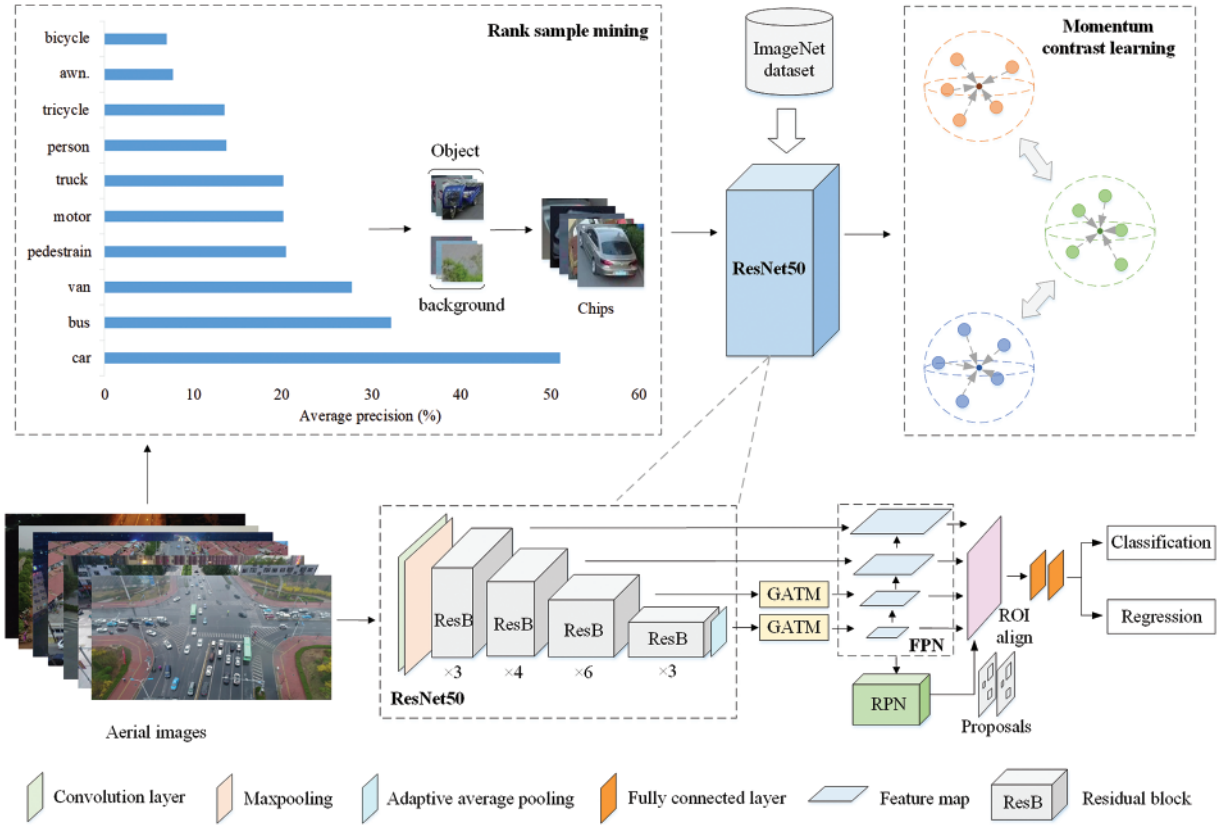


Figure 3: Framework of the proposed method based on Faster R-CNN with FPN. The rank sample mining method and momentum contrast learning strategy mainly learn the robust and discriminative underlying features between different image domains. The global attention mechanism based on transformer encoder (GATM) captures global information between different regions of feature maps to enhance object location semantics

3.3 Rank Sample Mining Method

Unlike natural images, aerial images usually contain small and dense objects and more complex backgrounds, since UAVs capture these images at different altitudes, angles, lighting, and weather conditions. The significant differences impair robust feature learning. To maximize the mutual information between natural image domain features and aerial image domain features, we design a rank sample mining method to sample image chips from high resolution aerial images.

The designed approach includes object chip mining and background chip mining. Object chips are predefined regions containing annotated objects within the aerial image. We state the detailed object chip mining pipeline in Algorithm 1. Firstly, we rank each category in descending order based on a predefined category priority. This priority is represented by CP as a hyperparameter, which is specifically discussed in Section 4.4. For clarity, let $B = \{b_1, \ldots, b_N\}$ be the collection of ground truth anchors, where each b_i corresponds to an annotated object in the aerial image. T_o is defined as the intersection over union (IoU) threshold. Then, we cherry-pick the ground truth anchors of the highest category priority from B, denoted as $M \leftarrow argmax(B, CP)$. The selected anchors M are added to collection D that saves the required ground truth anchors. To avoid duplicate processing, we sequentially remove M from B. Finally, we calculate the IoU of the ground truth anchors M and those of other categories. If the IoU value of M_i and b_i is greater than the threshold T_o, the anchor b_i is removed from B. The collection D is mapped to the original aerial image, obtaining object chips O. Moreover, we perform random rotation, horizontal, and vertical flipping operations for data augmentation to further increase the number of object chips with high priority, which facilitates alleviating model bias.

Algorithm 1: Object chip mining

Input: $B = \{b_1, \ldots, b_N\}$, $CP = [c_1, \ldots, c_P]$, T_o, where B is the collection of ground truth anchors, CP is the category priority, and T_o is the IoU threshold.

Output: Object chips O

Initialize: $D = \{\}$, which is used to save the required ground truth anchors.
while $B \neq \{\}$ **do**
  $M \leftarrow argmax(B, CP)$
  $D \leftarrow D \cup M$, $B \leftarrow B - M$
  for $b_i \in$ B **do**
    if $IoU\ (M, b_i) \geq T_o$ **then**
      $B \leftarrow B - b_i$
    end if
  end for
end while
$O \leftarrow cut(img, D)$

The background chip mining aims to select regions that do not contain any annotated objects of interest, promoting the detector to better understand and separate the foreground objects from the complex backgrounds. One principle is the appropriate chip size. The width w and height h of the chip are calculated as

$$w = h = \sqrt{\frac{1}{N}\sum_{i=1}^{N} \mathrm{S}\,(b_i)} \tag{6}$$

where b is the ground truth anchors, S is an area calculation operation, and N is the number of ground truth anchors on the current aerial image. Another principle is the positions of background chips, which do not overlap with the ground truth anchor of the object as much as possible. We divide the aerial image into a grid of candidate regions. The center point of each grid is used as the center of the candidate background chip. We calculate the IoU between these candidate chips and the ground truth anchors. When the IoU value is lower than the threshold T_b, the candidate chip is retained as a background chip. More details of the thresholds T_o and T_b are discussed in Section 4.4.

After mining the object chips and background chips, we uniform them to a fixed size. These resized chips are used to enhance the backbone by fine-tuning. The rank sample mining method lays the foundation for detectors to distinguish objects from complex backgrounds, and confused objects from different categories.

3.4 Momentum Contrast Learning Strategy

In the aerial image domain, object categories often exhibit high semantic similarity, which makes sample detection more difficult. To mitigate this, we propose a momentum contrast learning strategy to enhance the backbone ability to learn more discriminative features. This approach enables the backbone to encode features of the same category similarly and those of different categories dissimilarly [41].

During the initialization phase, we use a matrix E with dimension $[C, D_{\text{im}}]$ to store the average encoded features for each category, where C denotes the number of categories and D_{im} is the length of encoded features. We also initialize a queue Q with dimension $[C, N_{\text{um}}, D_{\text{im}}]$ as a temporary storage for encoded features during the learning process. The N_{um} is the maximum number of encoded features for each category that the queue can store, which is discussed in Section 4.4.

In a concrete implementation, the feature map output from the backbone undergoes an average pooling operation to reduce its dimensionality and retain important features. This operation results in a feature map F with dimension $[B, D_{\text{im}}]$, where B denotes the batch size, which is less than the N_{um}. We use momentum contrast loss function L_{mc} to fine-tune the backbone pre-trained on the ImageNet dataset.

$$L_{\text{mc}} = -\log \frac{\exp\left(x_{\text{q}} \text{E}\left(c_{+}\right)/\tau\right)}{\sum_{i=1}^{C} \exp\left(x_{\text{q}} \text{E}\left(c_i\right)/\tau\right)} \tag{7}$$

where x_{q} is the encoded feature of a sample in F, $\text{E}(c_{+})$ is the expectation of encoded features that are the same category as x_{q}, and $\text{E}(c_i)$ is the expectation of encoded features for category i in the matrix E.

The feature map F is enqueued into queue Q according to the category labels of input samples, maintaining a dynamic and updated set of features for each category. When the queue reaches its maximum capacity N_{um}, we use the momentum method to update matrix E.

$$\text{E}\left(c_i\right) \leftarrow m\text{E}\left(c_i\right) + (1-m)\,\text{E}_{\text{q}}\left(c_i\right) \tag{8}$$

where m is a momentum factor, and $\text{E}_{\text{q}}(c_i)$ is the expectation of encoded features for category i in the updated queue Q. This momentum update mechanism ensures that the encoded features in the matrix E are continuously and stably updated, facilitating the backbone to learn consistent feature representations more easily.

Fig. 4 depicts the encoded feature visualization from ResNet50. After applying the proposed momentum contrast learning strategy, we observe that the distances between features from the same

category are closer, while those of different categories are more distant. The stark contrast in feature clustering highlights the effectiveness of our proposed momentum contrast learning strategy.

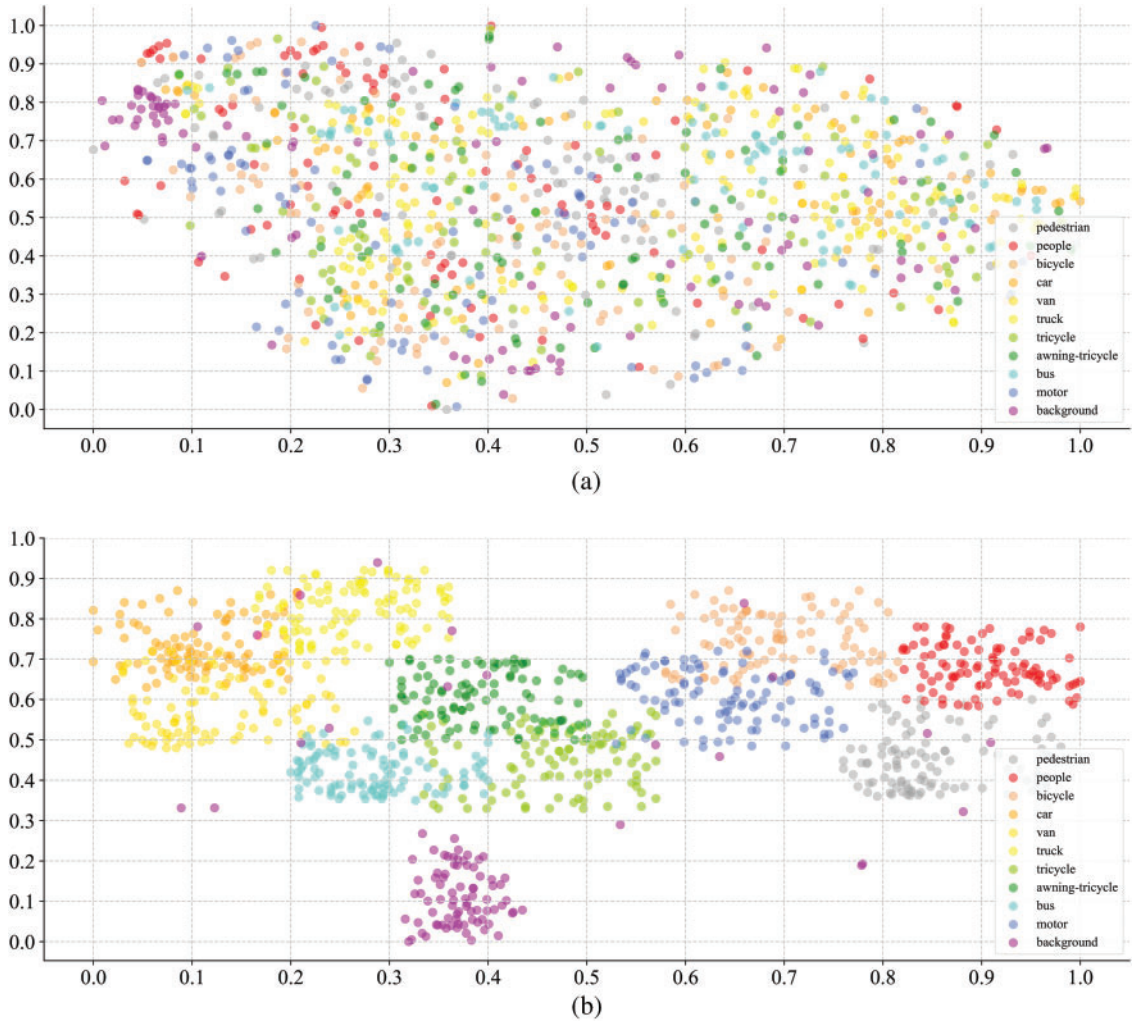


Figure 4: Visualization results of the encoded feature distribution from ResNet50. (a) Without momentum contrast learning. (b) With momentum contrast learning

3.5 Global Attention Mechanism Based on Transformer Encoder

Aerial images always contain complex geographical elements, which easily confuse objects and backgrounds. In our work, we not only focus on the object regions but also consider the background regions that provide abundant contextual information for objects and facilitate object localization. Inspired by the transformer [42], we propose an efficient and simple GATM module to extract relevant information from different receptive fields in robust and discriminative feature maps, which is able to enhance the object location semantics. Our module maintains original input sizes and can be extended as an additional component for various typical object detectors. As presented in Fig. 5, we only add the GATM module in the last two stages of the backbone, which allows us to decrease the expensive computation and memory costs. The enhanced information can flow to the shallow feature maps

through FPN. Related experiments on the arrangement of the proposed module are discussed in Section 4.4.

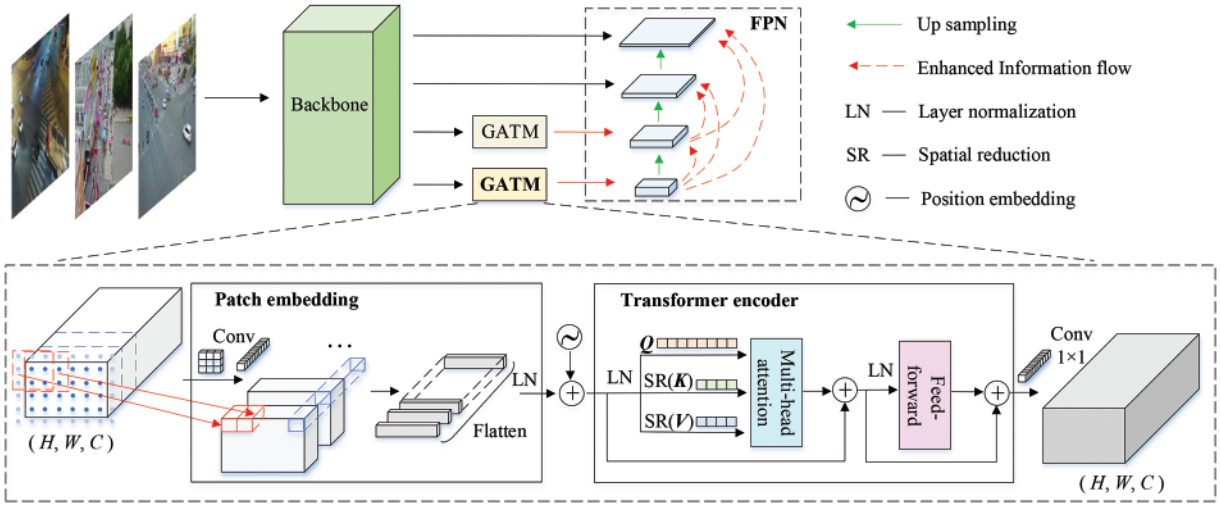


Figure 5: The architecture of GATM. Our module is embedded in the last two stages of the backbone network, which primarily consists of the patch embedding and a transformer encoder

In GATM module, given an input feature map with dimension $[H, W, C]$, we use the zero padding and depthwise separable convolution to obtain a series of patches. They are flattened and projected to the fixed dimensional embedding. After that, we feed the patch embedding and position embedding into a transformer encoder. The output is reshaped to its original size through a 1×1 convolution. For the transformer encoder, it contains the multi-head attention and feed-forward layers, which achieves better performance for occluded and dense objects in aerial images. To further decrease computation costs, we use spatial reduction (SR) to reduce input sequences for key $\boldsymbol{K}$ and value $\boldsymbol{V}$. The self-attention operation A can be formulated as

$$A\left(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}\right) = \text{Softmax}\left(\frac{\boldsymbol{Q} \times \text{SR}\left(\boldsymbol{K}\right)^{\text{T}}}{\sqrt{d}}\right) \text{SR}\left(\boldsymbol{V}\right) \tag{9}$$

where $\boldsymbol{Q}$ is the query, and d is a scaling factor.

4 Experimental Results and Analysis

4.1 Datasets

We used two popular datasets VisDrone and UAVDT to verify the effectiveness of the proposed method for object detection in aerial images. In addition, to evaluate the compatibility of our method with oriented object detection methods, we introduced the DroneVehicle dataset [43] with oriented bounding box (OBB) annotations.

- *VisDrone dataset*: The dataset has 10 object categories and 10,209 images, including 6471 for training, 548 for validation, and 3190 for testing. The image size is about 2000×1500 pixels, and the majority of objects are small, dense, and occluded from each other. We used the validation set to evaluate our method, following the same as previous methods [8].
- *UAVDT dataset*: The dataset consists of 23,258 training images and 15,069 testing images. The images with a 1080×540 size include three kinds of vehicles: car, bus and truck.

- *DroneVehicle dataset*: It includes 28439 visible-infrared image pairs with OBB annotations and has five categories: car, truck, bus, van, and freight car. The input image size is 840 × 712. We only adopted RGB images of the dataset.

To further measure robustness for detectors in aerial images, inspired by Yamada et al. [13] work on natural images, we introduced 10 synthetic image corruption types to the two VisDrone and UAVDT datasets. Considering the complex working conditions of UAVs, we adopted the following corruption types: Gaussian, shot, defocus, motion, snow, frost, fog, brightness, contrast, and compression [44]. The corrupted images are algorithmically synthesized and only used to test detector robustness. We applied each type to the VisDrone validation dataset and UAVDT testing dataset, obtaining 5480 and 150,690 corrupted images, respectively. Fig. 6 presents the 10 corruption types on the VisDrone dataset.

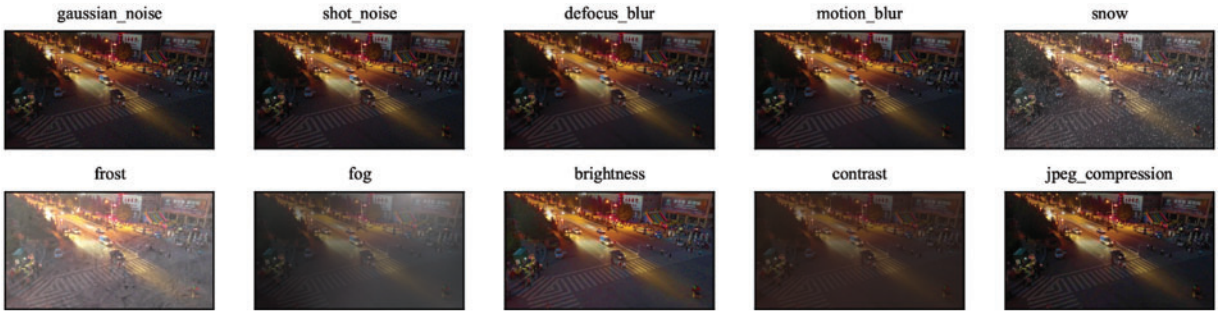


Figure 6: Corruption types on the VisDrone dataset

4.2 Implementation Details

Our experiments were implemented on the open-source MMDetection toolbox [45]. The comparative method YOLOv8 [46] used the MMYOLO [47] framework. We adopted Faster R-CNN, RetinaNet, and FCOS with three different backbones of ResNet50, RegNetX-1.6GF, and PVTv2-B0. These models were trained and tested on two NVIDIA Tesla P40 GPUs with 24 GB memory.

In the training phase, we set the batch size to 4, the initial learning rate to 1.0×10^{-4}, and the number of the epochs to 18. The testing phase used the same input image size as the training phase. We used the primary AP (average precision), AP50, and AP75 as evaluation metrics. AP50 and AP75 are the average precision with the IoU threshold of 0.5 and 0.75, respectively. Clean accuracy was evaluated on images without corruption types. The mean detection accuracy of the detector for each corruption type is defined as corruption robustness. The confusion matrix was also introduced to evaluate and analyze errors and confusions between different object categories.

4.3 Comparison Experiments

To compare the detection performance of our approach and the original method that transfers ImageNet network, we implemented three representative detectors, including Faster R-CNN (FRCNN), RetinaNet (Retina), and FCOS as basic models. We chose ResNet50, RegNetX-1.6GF, and PVTv2-B0 as backbones in Tables 1–3, respectively. RD-MIM achieves obvious improvements on almost all basic models, demonstrating the importance of underlying features. It can be seen that our method has higher AP score increases on the VisDrone dataset compared with the UAVDT dataset. Also, the improvement in robustness is greater than that of accuracy on the UAVDT dataset. The

reason is that the UAVDT dataset contains only three categories, so the categorical similarity is not severe. In Table 1, original FCOS obtains higher clean accuracy than Retina, but it has poor corruption robustness. It means that accuracy and robustness are not always correlated, indicating the necessity of studying corruption robustness on aerial images. RD-MIM achieves the relative growth rates of 51.0% and 39.4% on the corrupted VisDrone dataset using FCOS and on the corrupted UAVDT dataset using FRCNN, respectively. It suggests that the proposed approach has considerable benefits in enhancing robustness. In Table 3, although the PVTv2-B0 composed of transformer architecture performs well in handling dense and small object detection problems, it still obtains gains from our method. The training loss of FRCNN with ResNet50 on the VisDrone dataset is presented in Fig. 7. With the increase of training steps, our approach achieves faster convergence and lower loss values.

Table 1: Performance comparison of basic models based on ResNet50 backbone

Method	VisDrone						UAVDT					
	Clean accuracy			Corruption robustness			Clean accuracy			Corruption robustness		
	AP (%)	AP50 (%)	AP75 (%)	AP (%)	AP50 (%)	AP75 (%)	AP (%)	AP50 (%)	AP75 (%)	AP (%)	AP50 (%)	AP75 (%)
FRCNN	21.4	37.4	21.5	16.6	29.9	16.3	17.1	29.2	18.6	10.4	18.9	9.2
Retina	15.7	29.8	14.6	11.1	23.6	10.2	15.8	30.2	15.3	12.8	24.8	11.7
FCOS	16.9	29.8	17.2	9.8	19.2	9.3	16.9	29.9	17.8	10.5	19.8	10.0
FRCNN+RD-MIM	23.8	41.0	24.4	20.2	34.1	20.1	19.1	32.3	20.4	14.5	26.2	14.4
Retina+RD-MIM	18.0	34.1	16.6	15.1	27.4	13.8	16.1	29.2	16.6	13.5	25.3	13.1
FCOS+RD-MIM	18.7	32.6	18.8	14.8	26.5	14.6	19.0	32.1	20.9	14.4	25.4	15.2

Table 2: Performance comparison of basic models based on RegNetX-1.6GF backbone

Method	VisDrone						UAVDT					
	Clean accuracy			Corruption robustness			Clean accuracy			Corruption robustness		
	AP (%)	AP50 (%)	AP75 (%)	AP (%)	AP50 (%)	AP75 (%)	AP (%)	AP50 (%)	AP75 (%)	AP (%)	AP50 (%)	AP75 (%)
FRCNN	19.9	35.5	19.4	15.1	27.5	14.7	16.2	27.8	17.6	11.8	20.8	12.3
Retina	13.4	26.4	12.0	9.3	19.1	8.0	13.9	25.5	14.0	9.5	18.3	8.7
FCOS	14.3	26.0	14.2	10.9	20.2	10.6	14.1	25.7	14.6	10.8	20.3	10.7
FRCNN+RD-MIM	22.9	39.8	22.8	17.8	31.9	17.5	19.2	32.4	21.3	15.5	26.6	16.9
Retina+RD-MIM	16.1	31.2	14.6	11.9	23.8	10.6	15.7	28.8	16.1	12.0	22.7	11.6
FCOS+RD-MIM	18.1	31.6	17.8	13.9	24.8	13.7	15.6	29.2	15.8	13.3	25.5	13.1

Table 3: Performance comparison of basic models based on PVTv2-B0 backbone

Method	VisDrone						UAVDT					
	Clean accuracy			Corruption robustness			Clean accuracy			Corruption robustness		
	AP (%)	AP50 (%)	AP75 (%)	AP (%)	AP50 (%)	AP75 (%)	AP (%)	AP50 (%)	AP75 (%)	AP (%)	AP50 (%)	AP75 (%)
FRCNN	23.7	41.9	23.3	19.9	36.2	19.2	18.9	34.4	18.8	14.7	27.4	14.1
Retina	19.6	37.7	18.4	16.4	32.3	15.0	15.1	27.7	15.2	12.3	22.9	11.9
FCOS	17.2	30.7	17.3	14.5	26.6	14.3	16.7	31.3	16.6	14.1	26.9	13.5
FRCNN+RD-MIM	25.9	44.9	26.5	21.7	38.7	21.4	20.7	34.3	22.7	17.0	28.9	18.2
Retina+RD-MIM	21.1	39.9	19.8	17.7	34.5	16.3	15.1	27.4	15.1	12.9	23.9	12.7
FCOS+RD-MIM	18.8	32.6	18.8	16.0	28.4	15.9	17.6	32.3	17.9	16.2	29.8	16.4

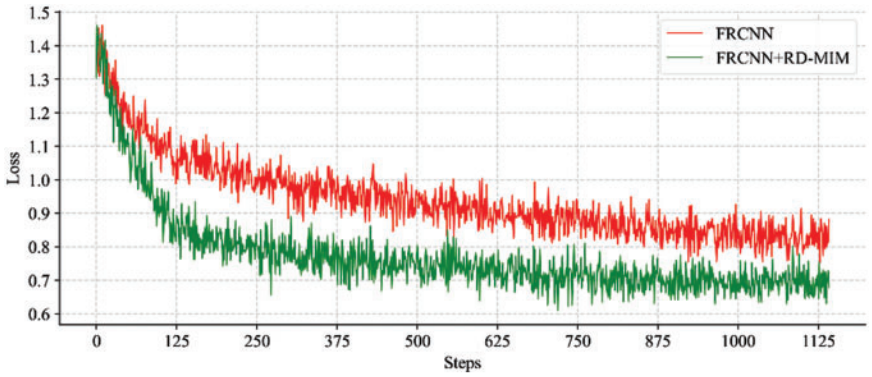


Figure 7: The training loss of FRCNN with ResNet50 on the VisDrone dataset

Table 4 presents the detection results on each class by using FRCNN with ResNet50 on the clean VisDrone dataset. Our method improves AP scores for all classes, especially those categories with hard samples. For example, the detection accuracy for the class awning-tricycle and bicycle is increased by 2.3% and 2.0%, respectively. We further explored the confusion matrices of AP scores for different categories in Fig. 8. From the observation, our approach allows detectors to better distinguish objects and complicated backgrounds, and reduces the detection errors of objects from confused categories. The results reveal that the designed rank sample mining method and momentum contrast learning strategy can effectively alleviate the problems of interclass similarity and intraclass variability.

Table 4: Detection results for each class on the clean VisDrone dataset

Method	AP (%)	Car	Bus	Van	Ped.	Motor	Truck	Person	Tricycle	Awn.	Bicycle
FRCNN	21.4	51.2	32.2	27.8	20.4	20.1	20.1	13.7	13.5	7.7	7.0
FRCNN+RD-MIM	23.8	53.7	35.5	31.5	23.3	22.8	22.2	15.8	14.5	10.0	9.0

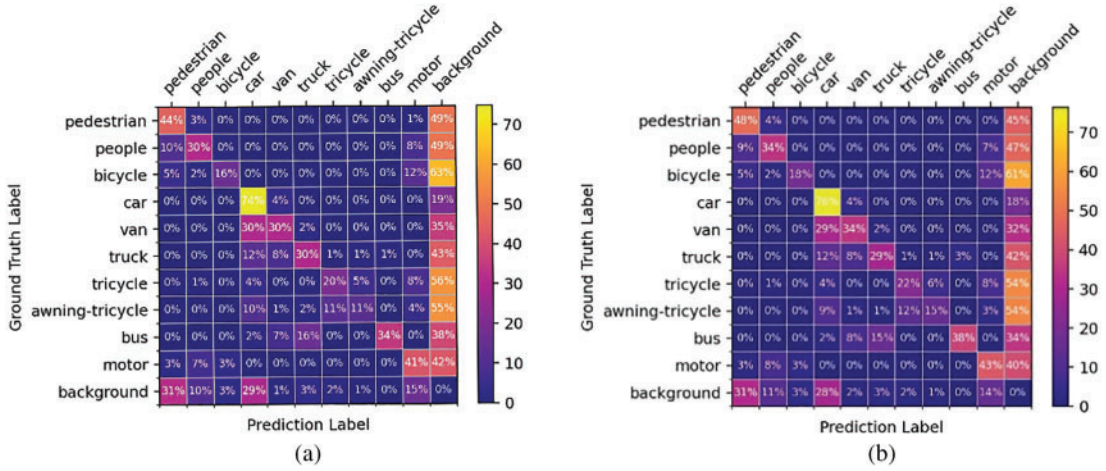


Figure 8: Confusion matrices on the clean VisDrone dataset using (a) original FRCNN and (b) FRCNN with our RD-MIM

To further explore the robustness performance of the RD-MIM, we evaluated the AP scores of FRCNN with ResNet50 under different corruption types in Table 5. The shot noise and snow severely deteriorate the accuracy of the original detector on the VisDrone and UAVDT datasets, resulting in 8.1% and 9.3% AP drops compared to clean accuracy. After applying the RD-MIM, the AP scores increase by 5.2% and 6.0%. Moreover, we also used the models trained on the VisDrone dataset to evaluate the unseen UAVDT dataset by using AP scores in Table 6. Even though the training and testing sets have completely different data distributions, our implementation still outperforms the original models, yielding 4.5%, 4.2%, and 4.8% average improvements. These results reveal that our approach brings great robustness gains to the detector.

Table 5: Robustness evaluation under different corruption types in AP

Dataset	Method	Gaussian	Shot	Defocus	Motion	Snow	Frost	Fog	Brightness	Contrast	Compression
VisDrone	FRCNN	13.4	13.3	15.4	16.3	13.8	18.0	17.8	21.1	17.8	19.1
	FRCNN+RD-MIM	18.6	18.5	17.9	19.2	19.7	21.2	20.9	23.3	20.0	22.7
UAVDT	FRCNN	8.5	8.5	9.7	10.2	7.8	11.4	11.2	13.3	11.4	12.2
	FRCNN+RD-MIM	12.2	12.3	12.6	13.6	13.8	15.5	15.0	18.5	15.0	16.8

Table 6: Evaluation results of robustness on the unseen UAVDT dataset by using models trained on the VisDrone dataset

Method	Car	Truck	Bus	Average (%)
FRCNN	15.8	3.8	13.7	11.1
Retina	16.7	3.6	13.3	11.2
FCOS	17.2	2.8	14.2	11.4
FRCNN+RD-MIM	20.9	6.8	19.0	15.6
Retina+RD-MIM	21.4	5.8	18.9	15.4
FCOS+RD-MIM	22.1	7.2	19.4	16.2

The comparison of existing approaches on the VisDrone dataset is shown in Table 7. The proposed approach outperforms the first two methods [9,27], resulting in 3.8% and 1.2% AP score improvements respectively. Compared to the DSHNet [8] with FRCNN, RD-MIM obtains slightly lower accuracy on clean aerial images but increases by 2.2% AP score on corrupted aerial images. QueryDet [48] is specifically designed for aerial images, which performs well for small objects. After combining the RD-MIM, the AP scores increase by 1.9% and 3.8% on clean and corrupted images, respectively. We also evaluated the performance gains of the recently released YOLOv8 [46] and the advanced DINO (DETR with Improved denoising anchor boxes) [49]. As can be seen, our RD-MIM still improves the performance of advanced detectors, especially in terms of corruption robustness.

To demonstrate the compatibility of our method, we tested different types of oriented object detection methods, including Faster R-CNN (OBB) [50], Retina (OBB) [20], ROI Transformer [51], OAN [33], and recent TS-Conv [34] on the DroneVehicle dataset in Table 8. We adopted AP50 to evaluate each subcategory. As can be observed, our RD-MIM help the original detection methods achieve higher AP50 scores. It illustrates that our approach can be easily integrated into oriented object detection methods to further improve performance.

Table 7: Performance comparison of existing approaches on the VisDrone dataset

Method	Detector	Clean accuracy			Corruption robustness		
		AP (%)	AP50 (%)	AP75 (%)	AP (%)	AP50 (%)	AP75 (%)
SimCal [27]	FRCNN	20.0	35.8	19.6	–	–	–
DSDet [9]		22.6	41.7	21.6	–	–	–
DSHNet [8]		24.6	44.4	24.1	18.0	34.0	18.3
DSHNet [8]+RD-MIM		23.8	41.0	24.4	20.2	34.1	20.1
DSHNet [8]	Retina	16.1	30.2	15.5	11.6	25.3	10.3
DSHNet [8]+RD-MIM		18.0	34.1	16.6	15.1	27.4	13.8
QueryDet [48]		26.3	46.5	26.6	22.6	43.1	22.0
QueryDet [48]+RD-MIM		28.2	52.7	26.7	26.4	44.6	25.7
DINO [49]	Advanced	23.1	41.8	22.0	19.7	36.8	18.1
DINO [49]+RD-MIM		26.2	42.8	26.9	23.5	38.9	23.9
YOLOv8 [46]		26.4	43.1	27.1	21.2	35.4	21.5
YOLOv8 [46]+RD-MIM		28.3	45.6	29.3	24.9	42.4	26.6

Table 8: Compatibility evaluation of our RD-MIM and oriented object detection methods

Method	Car	Freight car	Truck	Bus	Van	AP50 (%)
FRCNN(OBB) [50]	67.9	26.3	38.6	67.0	23.2	44.6
Retina(OBB) [20]	67.5	13.7	28.2	62.0	19.3	38.1
ROI Transformer [51]	68.1	29.1	44.2	70.6	27.6	47.9
OAN(OBB) [33]	68.6	26.1	38.4	68.2	24.3	45.1
TS-Conv(OBB) [34]	90.1	48.1	64.4	91.7	46.5	68.2
FRCNN(OBB) [50]+RD-MIM	68.2	33.5	43.3	67.7	29.8	48.5
Retina(OBB) [20]+RD-MIM	67.9	15.4	35.2	65.1	23.5	41.4
ROI Transformer [51]+RD-MIM	68.3	35.2	46.7	71.8	32.2	50.8
OAN(OBB) [33]+RD-MIM	69.5	29.2	43.8	69.8	27.1	47.9
TS-Conv(OBB) [34]+RD-MIM	90.0	50.7	66.5	92.0	49.3	69.7

4.4 Ablation Study

We conducted ablation experiments on the VisDrone dataset to verify the contributions of major components in RD-MIM, as depicted in Table 9. Rank sample mining method (RSM) shows an obvious increase over the original FRCNN, because it is conducive to better learning underlying features of aerial images. After employing momentum contrast learning (MCL), the AP score increases by 1.2% on the clean VisDrone dataset. The promising result indicates that our momentum contrast learning can help the detector distinguish object features more effectively. The detection accuracy is further improved with the GATM module, illustrating that it can enhance feature representations. The performance trend of corruption robustness is similar to clean accuracy and has a higher gain. It suggests that each component of RD-MIM facilitates remarkable robustness enhancement.

Table 9: Results of the ablation study for RD-MIM on the VisDrone dataset

Method	Clean accuracy			Corruption robustness		
	AP (%)	AP50 (%)	AP75 (%)	AP (%)	AP50 (%)	AP75 (%)
FRCNN	21.4	37.4	21.5	16.6	29.9	16.3
FRCNN+RSM	22.0	38.7	22.1	18.1	32.4	18.1
FRCNN+RSM+MCL	23.2	40.1	23.6	19.5	33.9	19.0
FRCNN+RD-MIM	23.8	41.0	24.4	20.2	34.1	20.1

In Table 10, we explored the impact of relevant parameters on the AP scores of the rank sample mining method. We conducted three ways of ranking category priorities, including random order, quantity order, and performance order. The quantity order is based on long-tail distribution where the fewer the sample number, the higher the category priority. The performance order is determined by the detection results of the original FRCNN, where the lower the accuracy, the higher the category priority. It can be seen from Table 10 that using performance order obtains higher AP scores. We also observe that data augmentation of high priority object chips can enhance clean accuracy and corruption robustness. In addition, we explored the impact of object clip and background chip IoU thresholds on the classification accuracy for the enhanced backbone with the momentum contrast learning in Fig. 9. When the object clip IoU threshold T_o is set to 0.1 and the background chip IoU threshold T_b is set to 0.01, the highest classification accuracy can be reached.

Table 10: The impact of relevant parameters on the AP scores of the rank sample mining method

Random order	Quantity order	Performance order	Data augmentation	Background chips	Clean accuracy	Corruption robustness
✓					21.2	17.0
	✓				21.5	16.9
		✓			22.3	17.4
		✓	✓		22.8	18.9
		✓	✓	✓	23.8	20.2

We measured the maximum number N_{um} of each category stored in the queue Q for momentum contrast learning in Fig. 10. When the N_{um} value is 512, the model achieves better accuracy and robustness. A higher or lower value results in poorer performance. This is attributed to the fact that the larger the queue capacity, the fewer update steps. When the capacity is small, it is difficult to approach the expectation of encoded features. Therefore, we chose 512 as the N_{um} value.

Table 11 shows the results of ablation study on GATM module. Using same the setting as TPH-YOLOv5 [22], we added the attention module between the backbone and the neck. We explored the impact of the GATM module position and the spatial reduction ratio. It can be observed that the GATM module is only applied to the last two stages of the backbone and spatial reduction ratio c is halved, which can achieve higher accuracy and robustness with fewer parameters and giga floating point of operations (GFLOPs).

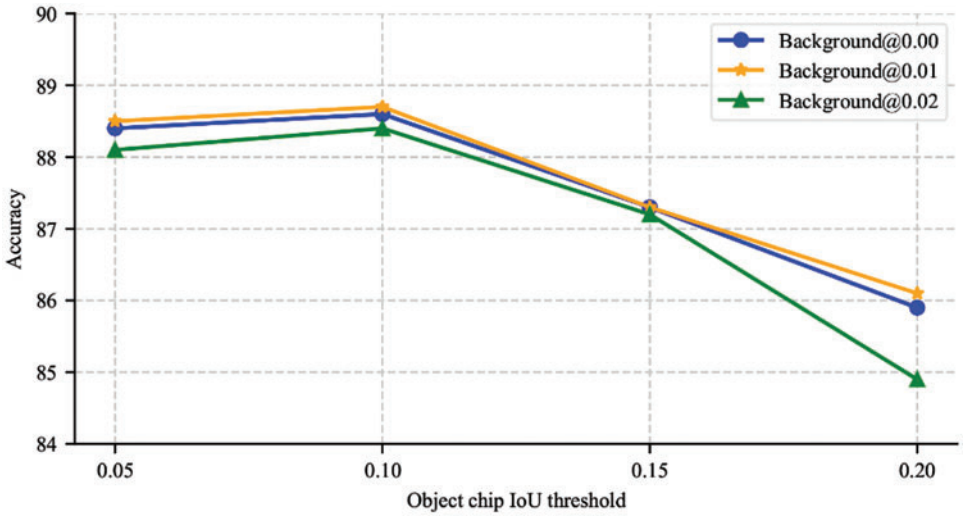


Figure 9: The impact of object clip and background chip IoU thresholds on the classification accuracy. Background@0.01 means that background chip IoU threshold T_b is set to 0.01

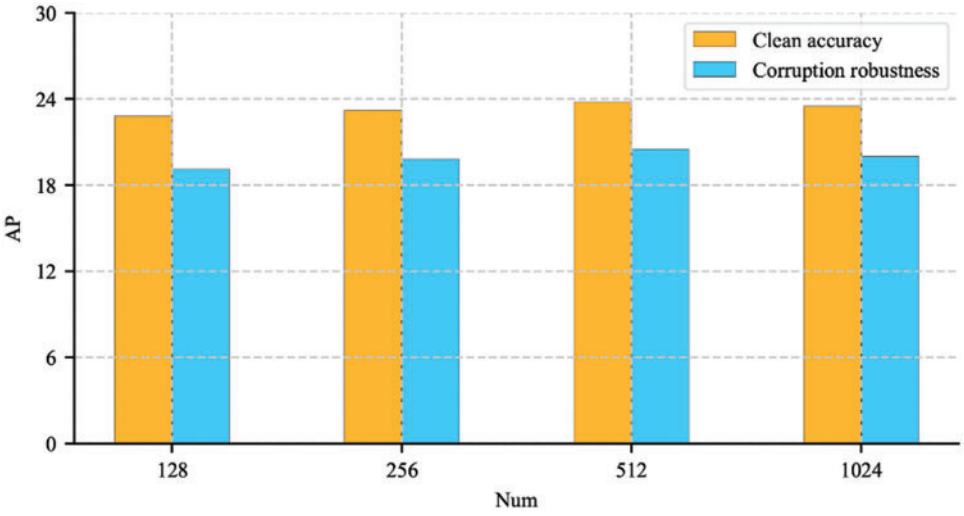


Figure 10: Evaluation of the hyperparameter N_{um} value for momentum contrast learning

Table 11: Results of the ablation study on the GATM module by AP scores

Last one stage	Last two stages	All stages	$c/2$	$c/1$	Clean accuracy	Corruption robustness	No. of parameters	GFLOPs
✓			✓		23.3	18.2	6.4×10^6	3.9
	✓		✓		23.8	20.2	9.1×10^6	7.4
		✓	✓		23.9	20.0	14.5×10^6	9.9
	✓			✓	23.5	19.5	30.5×10^6	24.1

Performance comparison of different attention modules is shown in Table 12. The squeeze-and-excitation (SE) [52], convolutional block attention module (CBAM) [53], and efficient channel

attention (ECA) [54] have fewer parameters than the non-local neural network (NL) [55], but their performances are worse. Compared to NL, our GATM module improves AP scores in both clean accuracy and corruption robustness. It not only utilizes the advantages of the transformer architecture but also has acceptable parameters and computational complexity. This suggests that the design of the attention module is scientific and reasonable.

Table 12: Performance comparison of different attention modules by AP scores

Module	Clean accuracy	Corruption robustness	No. of parameters	GFLOPs
SE [52]	22.3	17.2	0.7×10^6	0.02
CBAM [53]	22.5	18.0	0.7×10^6	0.02
ECA [54]	22.3	18.0	–	0.02
NL [55]	23.6	19.8	15.8×10^6	74.3
GATM	23.8	20.2	9.1×10^6	7.4

Figs. 11 and 12 show some visualization results on the corrupted VisDrone and UAVDT datasets. These images cover complex capturing backgrounds, such as fog, snow, defocus blur, motion blur, and low contrast. For comprehensive illustration, we employed the common FRCNN, FRCNN+RD-MIM, QueryDet specifically designed for aerial object detection, and QueryDet+RD-MIM. As can be seen, despite the presence of UAV-specific nuisances, our approach can still distinguish objects from complex backgrounds and predict more accurately between confused objects.

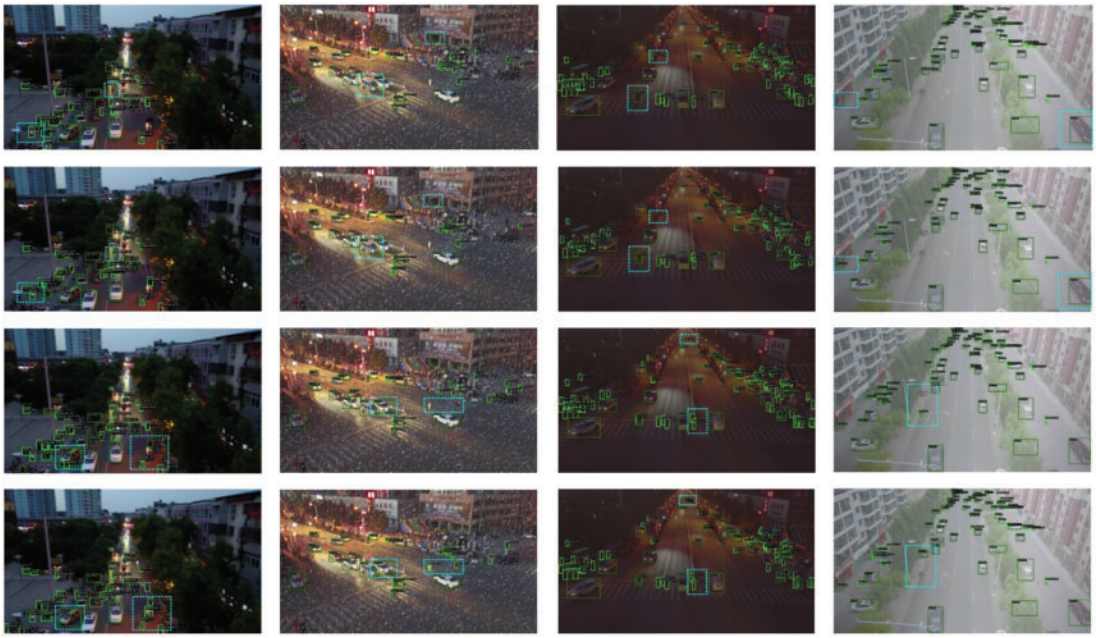


Figure 11: Some visualization results on the corrupted VisDrone dataset. From top to bottom, original FRCNN, FRCNN+RD-MIM, original QueryDet, and QueryDet+RD-MIM are given

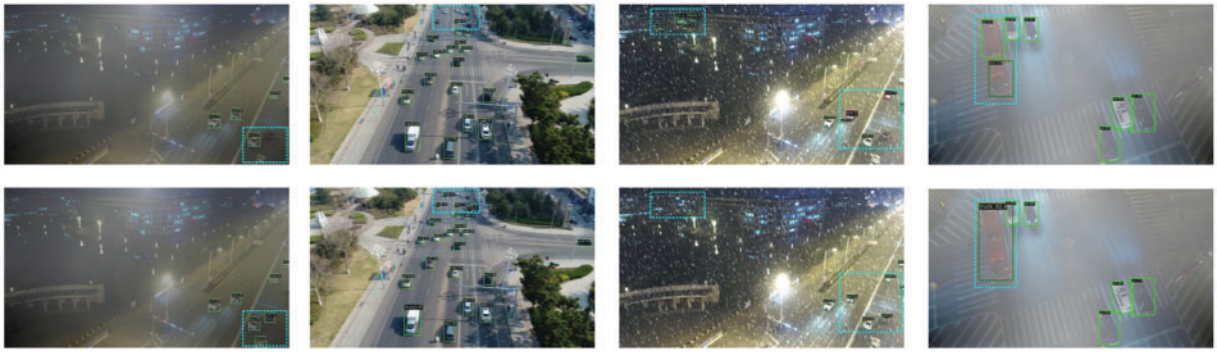


Figure 12: Some visualization results on the corrupted UAVDT dataset. From top to bottom, original FRCNN and FRCNN+RD-MIM are given

5 Conclusion

In this paper, we show that effective underlying features are beneficial for the detection performance improvement in aerial images by estimating the lower bound of mutual information. In light of this, we propose the RD-MIM approach to solve the problems of object categorical similarity and the poor robustness of object detection models under complex UAV working conditions. Our method can be applied to almost all available object detectors. The key components in RD-MIM include the rank sample mining method, the momentum contrast learning strategy, and the GATM module. Specifically, the first two components are able to learn underlying features by maximizing the mutual information of the robust and discriminative features between different image domains. They can also help the detectors distinguish objects from complex backgrounds, and confused objects from different categories. Furthermore, by using the GATM module, the global feature information from different receptive fields boosts the object location semantics. Theoretical analysis and experiments demonstrate that our approach improves the performance of basic detectors and advanced methods. Nevertheless, there still exists a limitation in our approach. The proposed rank sample mining method completely relies on supervised data with labels, and cannot be directly applied to unsupervised aerial images. In the future, we will explore the method's adaptability under various application conditions.

Acknowledgement: The authors would like to thank the support of the National Natural Science Foundation of China, and the authors sincerely appreciate the valuable comments of the editors and reviewers.

Funding Statement: This work was supported by the National Natural Science Foundation of China under Grant 61671219.

Author Contributions: The authors confirm contribution to the paper as follows: study conception and design: Xu Sun, Qing Cheng; data collection: Xu Sun, Qing Cheng; analysis and interpretation of results: Xu Sun, Yinhui Yu, Qing Cheng; draft manuscript preparation: Xu Sun. All authors reviewed the results and approved the final version of the manuscript.

Availability of Data and Materials: The code of the current study is available in the github repository, https://github.com/cq100/RD-MIM (accessed on 2 August 2024).

Ethics Approval: Not applicable.

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

[1] Z. Chen, H. Ji, Y. Zhang, Z. Zhu, and Y. Li, "High-resolution feature pyramid network for small object detection on drone view," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 34, no. 1, pp. 475–489, 2024. doi: 10.1109/TCSVT.2023.3286896.

[2] A. M. Qureshi, N. A. Mudawi, M. Alonazi, S. A. Chelloug, and P. Jeongmin, "Road traffic monitoring from aerial images using template matching and invariant features," *Comput. Mater. Contin.*, vol. 78, no. 3, pp. 3683–3701, 2024. doi: 10.32604/cmc.2024.043611.

[3] F. Fang, W. Liang, Y. Cheng, Q. Xu, and J. -H. Lim, "Enhancing representation learning with spatial transformation and early convolution for reinforcement learning-based small object detection," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 34, no. 1, pp. 315–328, 2023. doi: 10.1109/TCSVT.2023.3284453.

[4] C. Yan, X. Chang, M. Luo, H. Liu, X. Zhang and Q. Zheng, "Semantics-guided contrastive network for zero-shot object detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 46, no. 3, pp. 1530–1544, 2024. doi: 10.1109/TPAMI.2021.3140070.

[5] Z. Tang, X. Liu, and B. Yang, "PENet: Object detection using points estimation in high definition aerial images," in *IEEE Int. Conf. Mach. Learn. Appl. (ICMLA)*, Miami, FL, USA, Dec.14–17, 2020, pp. 392–398.

[6] C. Feng, C. Wang, D. Zhang, K. Renke, and Q. Fu, "Enhancing dense small object detection in UAV images based on hybrid transformer," *Comput. Mater. Contin.*, vol. 78, no. 3, pp. 3993–4013, 2024. doi: 10.32604/cmc.2024.048351.

[7] J. Ye, C. Fu, G. Zheng, D. P. Paudel, and G. Chen, "Unsupervised domain adaptation for nighttime aerial tracking," in *IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, New Orleans, LA, USA, Jun. 19–24, 2022, pp. 8886–8895.

[8] W. Yu, T. Yang, and C. Chen, "Towards resolving the challenge of long-tail distribution in UAV images for object detection," in *IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, USA, Jan. 5–9, 2021, pp. 3257–3266.

[9] X. Zhang, E. Izquierdo, and K. Chandramouli, "Dense and small object detection in UAV vision based on cascade network," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshop (ICCVW)*, Seoul, Republic of Korea, Oct. 27–28, 2019, pp. 118–126.

[10] X. Chen, C. Xie, M. Tan, L. Zhang, C. -J. Hsieh and B. Gong, "Robust and accurate object detection via adversarial learning," in *IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Nashville, TN, USA, Jun. 19–25, 2021, pp. 16617–16626.

[11] J. Deng, W. Dong, R. Socher, L. -J. Li, K. Li and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Miami, FL, USA, Jun. 20–25, 2009, pp. 248–255.

[12] J. Ding *et al.*, "Deeply unsupervised patch re-identification for pre-training object detectors," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 46, no. 3, pp. 1348–1361, 2024. doi: 10.1109/TPAMI.2022.3164911.

[13] Y. Yamada and M. Otani, "Does robustness on ImageNet transfer to downstream tasks?" in *IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, New Orleans, LA, USA, Jun. 18–24, 2022, pp. 9205–9214.

[14] Y. Cao *et al.*, "VisDrone-DET2021: The vision meets drone object detection challenge results," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshop (ICCVW)*, Montreal, BC, Canada, Oct. 11–17, 2021, pp. 2847–2854.

[15] D. Du *et al.*, "The unmanned aerial vehicle benchmark: Object detection and tracking," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Munich, Germany, Sep. 8–14, 2018, pp. 375–391.

[16] L. Lu and F. Dai, "Accurate road user localization in aerial images captured by unmanned aerial vehicles," *Autom. Constr.*, vol. 158, no. 3, 2024, Art. no. 105257. doi: 10.1016/j.autcon.2023.105257.

[17] W. Lin, J. Chu, L. Leng, J. Miao, and L. Wang, "Feature disentanglement in one-stage object detection," *Pattern Recognit.*, vol. 145, no. 2, 2024, Art. no. 109878. doi: 10.1016/j.patcog.2023.109878.

[18] Z. Tian, C. Shen, H. Chen, and T. He, "FCOS: Fully convolutional one-stage object detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Seoul, Republic of Korea, Oct. 27–Nov. 2, 2019, pp. 9626–9635.

[19] Y. Yu, K. Zhang, X. Wang, N. Wang, and X. Gao, "An adaptive region proposal network with progressive attention propagation for tiny person detection from UAV images," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 34, no. 6, pp. 4392–4406, 2023. doi: 10.1109/TCSVT.2023.3335157.

[20] T. -Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollar, "Focal loss for dense object detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Venice, Italy, Oct. 22–29, 2017, pp. 2999–3007.

[21] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 6, pp. 1137–1149, 2017. doi: 10.1109/TPAMI.2016.2577031.

[22] X. Zhu, S. Lyu, X. Wang, and Q. Zhao, "TPH-YOLOv5: Improved YOLOv5 based on transformer prediction head for object detection on drone-captured scenarios," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshop (ICCVW)*, Montreal, BC, Canada, Oct. 11–17, 2021, pp. 2778–2788.

[23] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Las Vegas, NV, USA, Jun. 27–30, 2016, pp. 770–778.

[24] I. Radosavovic, R. P. Kosaraju, R. Girshick, K. He, and P. Dollar, "Designing network design spaces," in *IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, USA, Jun. 14–19, 2020, pp. 10425–10433.

[25] W. Wang *et al.*, "PVT v2: Improved baselines with pyramid vision transformer," *Comput. Vis. Media.*, vol. 8, no. 3, pp. 415–424, 2022. doi: 10.1007/s41095-022-0274-8.

[26] Y. Huang, J. Chen, and D. Huang, "UFPMP-Det: Toward accurate and efficient object detection on drone imagery," in *AAAI Conf. Artif. Intell*, Feb. 22–Mar. 1, 2022, pp. 852–860.

[27] T. Wang *et al.*, "The devil is in classification: A simple framework for long-tail instance segmentation," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Glasgow, UK, Aug. 23–28, 2020, pp. 728–744.

[28] H. Zhu, C. Xu, W. Yang, R. Zhang, Y. Zhang and G. -S. Xia, "Robust tiny object detection in aerial images amidst label noise," 2024, *arXiv:2401.08056*.

[29] B. Kiefer, M. Messmer, and A. Zell, "Diminishing domain bias by leveraging domain labels in object detection on uavs," in *Int. Conf. Adv. Robot. (ICAR)*, Ljubljana, Slovenia, Dec. 6–10, 2021, pp. 523–530.

[30] Y. -T. Shen, H. Lee, H. Kwon, and S. S. Bhattacharyya, "Progressive transformation learning for leveraging virtual images in training," in *IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Vancouver, BC, Canada, Jun. 17–24, 2023, pp. 835–844.

[31] Z. Wu, K. Suresh, P. Narayanan, H. Xu, H. Kwon and Z. Wang, "Delving into robust object detection from unmanned aerial vehicles: A deep nuisance disentanglement approach," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Seoul, Republic of Korea, Oct. 27–Nov. 2, 2019, pp. 1201–1210.

[32] H. S. S. Blas, A. C. Balea, A. S. Mendes, L. Augusto, and G. V. Gonzalez, "A platform for swimming pool detection and legal verification using a multi-agent system and remote image sensing," *Int. J. Interact. Multimed. Artif. Intell.*, vol. 8, no. 4, pp. 153, 2023. doi: 10.9781/ijimai.2023.01.002.

[33] X. Xie, G. Cheng, Q. Li, S. Miao, K. Li and J. Han, "Fewer is more: Efficient object detection in large aerial images," *Sci. Chin. Inf. Sci.*, vol. 67, no. 1, 2023.

[34] Z. Huang, W. Li, X. -G. Xia, H. Wang, and R. Tao, "Task-wise sampling convolutions for arbitrary-oriented object detection in aerial images," *IEEE Trans. Neural Netw. Learn. Syst.*, pp. 1–15, 2024. doi: 10.1109/TNNLS.2024.3367331.

[35] K. Sohn *et al.*, "Visual prompt tuning for generative transfer learning," in *IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Vancouver, BC, Canada, Jun. 17–24, 2023, pp. 19840–19851.

[36] Y. Tian, F. Suya, A. Suri, F. Xu, and D. Evans, "Manipulating transfer learning for property inference," in *IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Vancouver, BC, Canada, Jun. 17–24, 2023, pp. 15975–15984.

[37] A. Deng, X. Li, D. Hu, T. Wang, H. Xiong and C. -Z. Xu, "Towards inadequately pre-trained models in transfer learning," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Paris, France, Oct. 2–6, 2023, pp. 19340–19351.

[38] M. Noroozi and A. Shah, "Towards optimal foreign object debris detection in an airport environment," *Expert. Syst. Appl.*, vol. 213, no. 7, 2023, Art. no. 118829. doi: 10.1016/j.eswa.2022.118829.
[39] A. V. D. Oord, Y. Li, and O. Vinyals, "Representation learning with contrastive predictive coding," 2018, *arXiv:1807.03748*.
[40] T. -Y. Lin, P. Dollar, R. Girshick, K. He, B. Hariharan and S. Belongie, "Feature pyramid networks for object detection," in *IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Honolulu, HI, USA, Jul. 21–26, 2017, pp. 936–944.
[41] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum contrast for unsupervised visual representation learning," in *IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, USA, Jun. 13–19, 2020, pp. 9726–9735.
[42] A. Vaswani *et al.*, "Attention is all you need," *Adv Neural Inf. Process. Syst.*, vol. 30, pp. 5999–6009, 2017.
[43] Y. Sun, B. Cao, P. Zhu, and Q. Hu, "Drone-based RGB-infrared cross-modality vehicle detection via uncertainty-aware learning," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 10, pp. 6700–6713, 2022. doi: 10.1109/TCSVT.2022.3168279.
[44] D. Hendrycks and T. Dietterich, "Benchmarking neural network robustness to common corruptions and perturbations," in *Int. Conf. Learn. Represent. (ICLR)*, New Orleans, LA, USA, May 6–9, 2019.
[45] K. Chen *et al.*, "MMDetection: Open mmlab detection toolbox and benchmark," 2019, *arXiv:1906.07155*.
[46] G. Jocher, A. Chaurasia, and J. Qiu, "Ultralytics YOLOv8," *AGPL-3.0*, 2023. Accessed: Mar. 22, 2024. [Online]. Available: https://github.com/ultralytics/ultralytics
[47] M. Contributors, "MMYOLO: OpenMMLab YOLO series toolbox and benchmark," *AGPL-3.0*, 2022. Accessed: Mar. 22, 2024. [Online]. Available: https://github.com/open-mmlab/mmyolo
[48] C. Yang, Z. Huang, and N. Wang, "QueryDet: Cascaded sparse query for accelerating high-resolution small object detection," in *IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, New Orleans, LA, USA, Jun. 18–24, 2022, pp. 13658–13667.
[49] Z. Hao *et al.*, "DINO: Detr with improved denoising anchor boxes for end-to-end object detection," in *Int. Conf. Learn. Represent. (ICLR)*, Kigali, Rwanda, May 1–5, 2023. doi: 10.48550/arXiv.2203.03605.
[50] G. -S. Xia *et al.*, "DOTA: A large-scale dataset for object detection in aerial images," in *IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Salt Lake City, UT, USA, Jun. 18–22, 2018, pp. 3974–3983.
[51] J. Ding, N. Xue, Y. Long, G. -S. Xia, and Q. Lu, "Learning RoI transformer for oriented object detection in aerial images," in *IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Long Beach, CA, USA, Jun. 16–20, 2019, pp. 2844–2853.
[52] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *IEEE/CVF Conf. Comput. Vis. Pat-Tern Recognit. (CVPR)*, Salt Lake City, UT, USA, Jun. 18–22, 2018, pp. 7132–7141.
[53] S. Woo, J. Park, J. -Y. Lee, and I. S. Kweon, "CBAM: Convolutional block attention module," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Munich, Germany, Sep. 8–14, 2018, pp. 3–19.
[54] Q. Wang, B. Wu, P. Zhu, P. Li, W. Zuo and Q. Hu, "ECA-Net: Efficient channel attention for deep convolutional neural networks," in *IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Seattle, WA, USA, Jun. 14–19, 2020, pp. 11531–11539.
[55] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," in *IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Salt Lake City, UT, USA, Jun. 18–22, 2018, pp. 7794–7803.