Scale-Invariant Feature Disentanglement via Adversarial Learning for UAV-based Object Detection

Fan Liu[1*], **Liang Yao**[1*], **Chuanyi Zhang**[1†], **Ting Wu**[1], **Xinlei Zhang**[1], **Jun Zhou**[2]

[1]Hohai University

[2]Griffith University

{fanliu, liangyao, 20231104, tingwu, xinleizhang}@hhu.edu.cn, jun.zhou@griffith.edu.au

Abstract

Detecting objects from Unmanned Aerial Vehicles (UAV) is often hindered by a large number of small objects, resulting in low detection accuracy. To address this issue, mainstream approaches typically utilize multi-stage inferences. Despite their remarkable detecting accuracies, real-time efficiency is sacrificed, making them less practical to handle real applications. To this end, we propose to improve the single-stage inference accuracy through learning scale-invariant features. Specifically, a Scale-Invariant Feature Disentangling module is designed to disentangle scale-related and scale-invariant features. Then an Adversarial Feature Learning scheme is employed to enhance disentanglement. Finally, scale-invariant features are leveraged for robust UAV-based object detection. Furthermore, we construct a multi-modal UAV object detection dataset, State-Air, which incorporates annotated UAV state parameters. We apply our approach to three state-of-the-art lightweight detection frameworks on three benchmark datasets, including State-Air. Extensive experiments demonstrate that our approach can effectively improve model accuracy. Our code and dataset are provided in Supplementary Materials and will be publicly available once the paper is accepted.

1 Introduction

With the rapid development of the Unmanned Aerial Vehicles (UAV) industry, UAV technology has been widely applied in various fields such as agriculture, logistics, and rescue [Qian *et al.*, 2022; Rejeb *et al.*, 2022; Srivastava and Prakash, 2023; Su *et al.*, 2023]. As one of the fundamental tasks in UAV applications, UAV-based object detection (UAV-OD) has attracted wide attention from the research community [Mittal *et al.*, 2020; Wu *et al.*, 2021; Zitar *et al.*, 2023]. As illustrated in Fig. 1, a significant difference between general and UAV-based object detection is the viewing angle and the object scale. Specifically, UAV tends to have a top-down view at a high altitude and most of the objects in the field of view are small-scale. Furthermore, the UAV altitude may change in the flight, resulting in a varying object scale. If the extracted features are scale-related, they may interfere with robust small object detection. In addition, the UAV computing platforms tend to have limited computational capability, they can hardly apply large object detection models [Han *et al.*, 2023; Biswas *et al.*, 2022; Suh *et al.*, 2023]. Owing to the above drawbacks (small objects and restricted computational resources), UAV-OD is a challenging task.

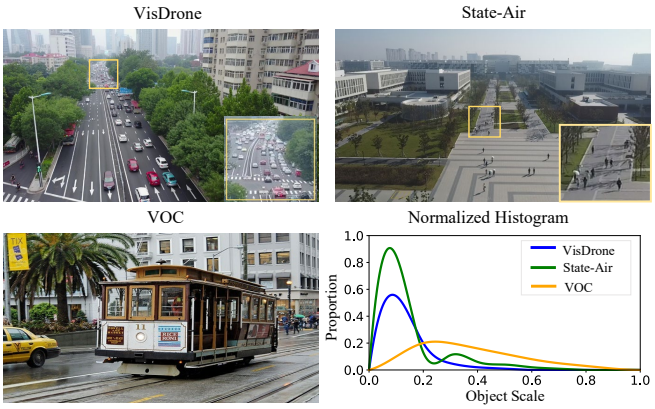

Figure 1: Comparison of general (VOC) and UAV (Visdrone, State-Air) datasets. The object scale is normalized by the ratio of the object's actual area to the source image. Proportion represents the percentage of objects of each scale in the overall dataset.

To overcome the challenge of small and varying scale objects in UAV-OD, researchers typically employ multi-stage coarse-to-fine reasoning methods [Huang *et al.*, 2022]. Initially, a detector is utilized to roughly localize regions containing small objects. Subsequently, the area resolution is enlarged for refined small object detection. Although many of these methods have achieved remarkable accuracy, multiple inferences on a single image tend to be time-consuming. Therefore they are still not suitable for deployment on UAVs due to real-time requirements.

Another approach to alleviating the problem of small target detection is to leverage scale-invariant features [Park *et al.*, 2023; Wang *et al.*, 2020]. Scale-invariant features, such as shape features [Lowe, 1999], remain unchanged regardless of variations in the object scale. If a model can effec-

[*]Equal contribution

[†]Corresponding author

tively learn scale-invariant features, the varying object scale issue can be mitigated and the capability to detect small objects can be enhanced. However, during the process of extracting deep features, there is often an ambiguity between Scale-Related and Scale-Invariant features [Van Noord and Postma, 2017]. Consequently, it is difficult to extract specific scale-invariant features in an unsupervised manner. [Park *et al.*, 2023] proposed ssFPN to use a convolutional structure to extract scale-invariant features. However, the capability of extracted features to handle the scale-invariant properties of objects in UAV-captured images is uncertain.

The widespread adoption of UAVs leads to the creation of a large number of UAV-OD datasets [Zhu *et al.*, 2018; Du *et al.*, 2018; Robicquet *et al.*, 2016]. Although these datasets significantly advance the research of UAV-OD, the majority of them neglect flight status data such as UAV-specific parameters and altitudes. The flight status data can be potentially beneficial for UAV-OD research, and a minority of datasets capture this ancillary data, *e.g.*, AU-AIR [Bozcan and Kayacan, 2020] and SynDrone [Rizzoli *et al.*, 2023]. Nevertheless, they tend to be plagued by issues such as imprecise annotations or a lack of diversity in environments.

To this end, we propose a novel approach named **SIFDAL** (Scale-Invariant Feature Disentanglement via Adversarial Learning), a new plug-and-play module for effective single-stage UAV-OD. It is composed of a Scale-Invariant Feature Disentangling (**SIFD**) module and an Adversarial Feature Learning (**AFL**) training scheme. Specifically, we first analyze the effect of various resolution layers of the feature pyramid network (FPN) [Lin *et al.*, 2017a; Liu *et al.*, 2018] on UAV-OD. The results indicate that the high-resolution layer of FPN plays a more vital role in small object detection. Then the SIFD module is developed to disentangle scale-invariant features from the high-resolution feature map. Next, we utilize the AFL training scheme to realize the maximal disentanglement. Finally, discriminative scale-invariant features can boost the detection accuracy. Our SIFDAL can be easily extended to feature-pyramid-based object detectors (*e.g.*, YOLOv7) to improve the accuracy. In addition, we propose a real-world multi-modal UAV-based object detection dataset named **State-Air**, which records UAV IMU (Inertial Measurement Unit) parameters and flight altitudes. Our contributions are summarized as follows:

- We propose a scale-invariant feature disentangling module, which can be applied to any FPN-based object detector. To the best of our knowledge, this is the first method to improve UAV-based object detection accuracy by disentangling scale-invariant features.
- We introduce a training scheme with adversarial feature learning to enhance feature disentanglement. It significantly improves the disentanglement effect as well as detection accuracy.
- We construct a multi-scene and multi-modal UAV-based object detection dataset, **State-Air**. It incorporates UAV IMU parameters as well as flight altitudes and covers multiple scenes and weather conditions.
- We validate our proposed approach with extensive experiments on three UAV benchmarks by integrating SIFDAL into various base detectors with FPN. The results demonstrate the superiority of our method in performance improvement.

2 Related Work

2.1 UAV-based Object Detection

Different from general object detection tasks [Girshick *et al.*, 2014; Ren *et al.*, 2015; Redmon *et al.*, 2016; Bochkovskiy *et al.*, 2020; Lv *et al.*, 2023], UAV-OD typically encounters the challenge of small targets and the limited computing power of edge equipment. A typical solution to alleviate the small object detection problem is to adopt a coarse-to-fine strategy [Duan *et al.*, 2021; Li *et al.*, 2020; Yang *et al.*, 2019a; Li *et al.*, 2017]. Initially, large targets are detected, and small target-dense subregions are located. Then subregions are employed as model input to obtain further detection results. In this step, a Gaussian mixture model could be used to supervise the detector in generating target clusters composed of focusing regions [Koyun *et al.*, 2022]. Alternatively, a CZ detector [Meethal *et al.*, 2023] utilized a density crop labeling algorithm to label the crowded object regions and then upscaled those regions to augment the training data.

To alleviate the problem of restricted computing resources, some methods were proposed to balance the accuracy and the efficiency. For example, sparse convolution [Liu *et al.*, 2015] was used to design lightweight network architectures that significantly reduce computing costs [Figurnov *et al.*, 2017; Yan *et al.*, 2018]. Typical examples include Querydet [Yang *et al.*, 2022] and CEASC [Du *et al.*, 2023]. The former utilized a sparse detection head to enable fast and accurate small object detection. It introduced a novel query mechanism to accelerate the inference speed of feature-pyramid-based object detectors. The latter adopted a plug-and-play detection head optimization method with context-enhanced sparse convolution and an adaptive multi-layer mask scheme.

2.2 Scale-Invariant Feature Extraction

Scale-invariant features remain unchanged even when the object scale varies, so it is widely used to address the multi-scale issue in computer vision. For example, SIFT algorithm [Lowe, 2004] conducted Gaussian difference operations at various scales and directions to detect local key feature points. Its strong scale and rotation invariance property enabled effective image feature matching. Later, SURF [Bay *et al.*, 2008] was proposed to enhance SIFT with faster calculation and improved robustness.

A more common approach is to consider scale dependencies in feature pyramids. For example, a Trident-FPN backbone network [Lin *et al.*, 2021] was designed to address the multi-scale problem in aerial images. It introduced a novel attention and anchor generation algorithm to enhance target detection performance. [Wang *et al.*, 2022] employed scale-invariant features to transform visible and infrared images for time series alignment and matching. [Behera *et al.*, 2023] utilized super-pixel images with key context information to extract scale-invariant features for predicting the object class of each pixel. To prevent information loss of targets in deep structures, [Park *et al.*, 2023] proposed a scale-

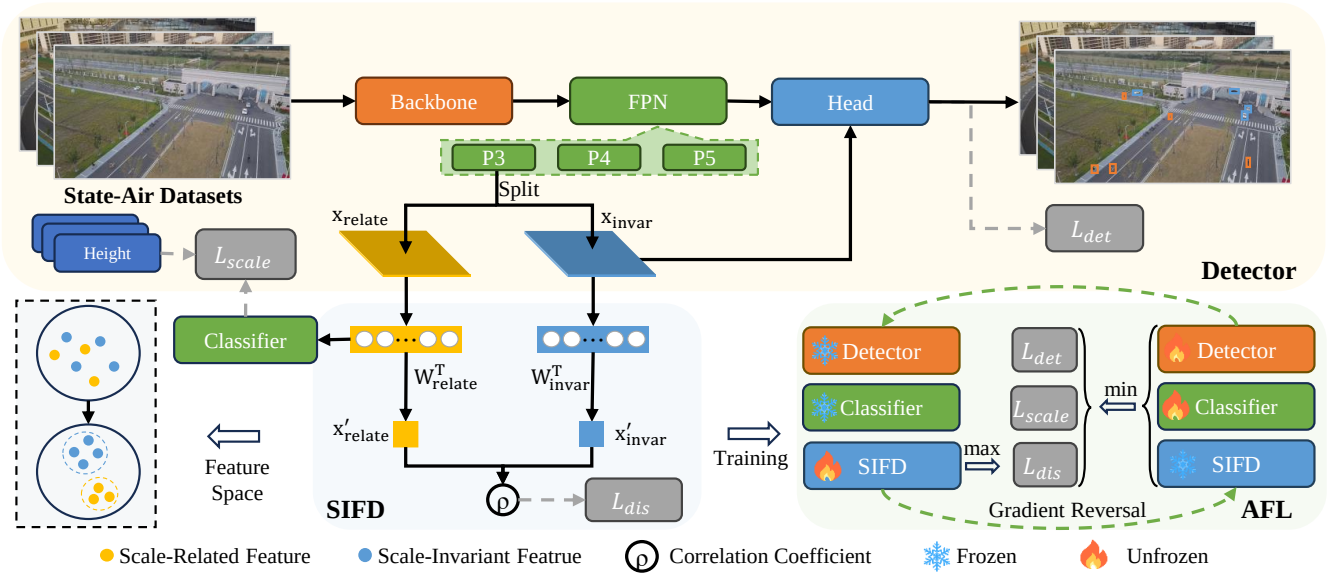

Figure 2: Overview of our proposed SIFDAL method. State-Air is a multi-scene and multi-modal UAV-based object detection dataset that incorporates UAV state parameters. We propose a SIFD module by utilizing correlation coefficients to disentangle Scale-Related and Scale-Invariant features. Subsequently, we introduce an AFL method to enhance feature disentanglement and ensure that the disentanglement is attributed to the detector and feature splitting.

sequence-based feature extraction method for FPN. The FPN structure was viewed as a scale space, and the scale sequence features were extracted through three-dimensional convolution as scale-invariant features.

3 Method

In this section, we introduce the proposed Scale-Invariant Feature Disentanglement via Adversarial Learning (SIFDAL) method. It consists of two components: Scale-Invariant Feature Disentangling (SIFD) and Adversarial Feature Learning (AFL). The overall framework of our approach with YOLOv7-L as the base detector is illustrated in Fig. 2.

3.1 Revisiting FPN in UAV-OD

When the convolutional neural network (CNN) [Li *et al.*, 2021] is applied for deep feature extraction, the network gradually decreases the resolution of the feature map. Consequently, small objects eventually vanish in deep layers. In most object detection methods, FPN is commonly utilized as the model's "neck". One of its functions is detecting objects of varying scales by entering features from each layer to the corresponding detection head. The detection head for a low-resolution layer is used to detect large objects, while one for a high-resolution layer is employed for small objects.

We visualize the heat map of different FPN layers in YOLOv7-L. As illustrated in Fig. 3, the detection head with the high-resolution layer (P3) is responsible for the majority of objects in the UAV's visual field. With a large number of small objects, the high-resolution detection head frequently plays an important role in the UAV-based object detection tasks. To boost the accuracy of small object detection, we aim to guide the high-resolution detection head to leverage scale-invariant features.

3.2 Scale-Invariant Feature Disentangling

After multi-scale feature fusion of FPN, object features often contain both scale-related and scale-invariant ones. Since scale-invariant features are not easily affected by varying object scales, they can be more conducive to UAV-OD than scale-related ones. To enable the model to learn scale-invariant features, we design a SIFD module that can be utilized in any detection model with FPN.

Feature Splitting

We directly apply channel splitting F_{split} to the high-resolution layer in FPN. Then the feature map is segmented into two groups. The formula can be expressed as:

$$x_{relate}, x_{invar} = F_{split}(x), \tag{1}$$

where $x \in \mathbb{R}^{H\times W\times 2L}$ represents the high-resolution feature map, $x_{relate} \in \mathbb{R}^{H\times W\times L}$ and $x_{invar} \in \mathbb{R}^{H\times W\times L}$ are the two feature maps after splitting. H, W, and L represent the height, width, and number of channels of the feature map, respectively. At this point, two features do not have corresponding meanings to their symbols. Next, they are disentangled to make x_{relate} and x_{invar} learn scale-related and scale-invariant features, respectively. To accomplish this objective, we propose a scale-related loss and adopt a feature disentangling loss [Liu *et al.*, 2022].

Scale-Related Loss

Intuitively, the object scale in the view is in connection with the UAV's altitude. As the UAV flies higher, the object scale becomes smaller. If x_{relate} is trained to correctly perform the height estimation task, it can be regarded as scale-related features. In other words, we can utilize the height estimation task for scale-related feature learning.

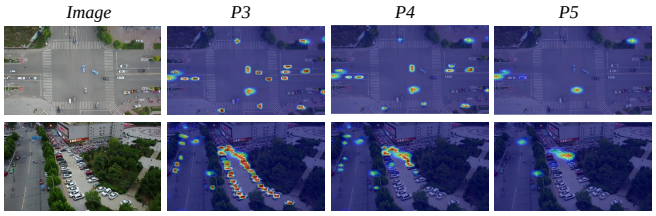

Figure 3: Visualization of heat maps in different FPN layers of YOLOv7-L. The resolution of the feature map decreases sequentially from P3 to P5. The detection head for the P3 layer corresponds to the majority of targets in the drone's field of view.

However, without relying on external knowledge such as laser ranging, it is difficult to predict specific heights solely from images. Our strategy is to replace the height regression with an altitude classification task. Specifically, we accurately group the heights and label the images with a height level. The height grouping is accomplished by the k-means clustering algorithm [Duda *et al.*, 1973], where the group number k is determined according to the characteristics of each dataset. Then, we leverage x_{relate} to perform the height classification task for learning scale-related information.

To be specific, we utilize global average pooling to reduce the dimensionality of x_{relate}, then input it into the classifier to obtain height prediction results. The final scale-related loss can be expressed as follows:

$$\mathcal{L}_{relate} = -\frac{1}{N}\sum_{i=1}^{N} y^i_{relate} \log\Big(F_{fc}\left(P_{avg}(x^i_{relate})\right)\Big), \quad (2)$$

where y^i_{relate} stands for the i-th image scale label, $P_{avg}(\cdot)$ denotes the channel-based average pooling, and $F_{fc}(\cdot)$ denotes the classifier. By optimizing $\mathcal{L}_{relate}$, x_{relate} can be scale-related through the height classification task.

Feature Disentangling Loss

Since x_{relate} becomes scale-related, we can intuitively make x_{invar} scale-invariant by disentangling them. Specifically, we employ the correlation coefficient analysis [Yang *et al.*, 2019b] to quantify the degree of the disentanglement. To facilitate the calculation of correlation coefficients, x_{invar} and x_{relate} are projected as scale-related vector x'_{relate} and scale-invariant vector x'_{invar}. Each vector is in the size of $batch\_size \times 1 \times 1$. The projection is expressed as:

$$x'_j = W_j^T\, x_j, \quad j = \{invar, relate\}, \quad (3)$$

where j represents $invar$ or $relate$, W_{invar} and W_{relate} are the two linear layers in the above projection process.

Then, we calculate the correlation coefficient between the above two vectors. The formula is given as follows:

$$\rho\left(x'_{invar}, x'_{relate}\right) = \frac{cov\left(x'_{invar}, x'_{relate}\right)}{\sqrt{D\left(x'_{invar}\right)}\sqrt{D\left(x'_{relate}\right)}}, \quad (4)$$

where $cov(X, Y)$ is the covariance to measure the correlation between X and Y, $D[X]$ represents the variance. Since the covariance between two independent random variables should be close to 0, ρ can be utilized as a correlation loss, which can reduce the correlation between scale-related and scale-invariant features.

To facilitate calculation, ρ^2 is taken as the feature disentangling loss, which can be described as:

$$\mathcal{L}_\rho = \rho^2\left(x'_{invar}, x'_{relate}\right). \quad (5)$$

Algorithm 1 SIFDAL

Require: $x \in \mathbb{R}^{H\times W\times 2L}$: Feature map of P_3, F_{split}: Channel splitting, ρ: Correlation coefficient, D: Detector, $W = \{W_{relate}, W_{invar}\}$: Linear layers, $\mathcal{L}_{det}, \mathcal{L}_{relate}, \mathcal{L}_{dis}$: Object detection loss, scale-related loss and feature disentangling loss.

for each training epoch **do**
  for each training sample feature x_i **do**
    $x_{relate}, x_{invar} = F_{split}(x_i)$ Eq. (1)
    $x'_{relate} = W^T_{relate} x_{relate}$
    $x'_{invar} = W^T_{invar} x_{invar}$ Eq. (3)
    $\mathcal{L}_\rho = \rho^2(x'_{invar}, x'_{relate})$ Eq. (5)
    if in the first 30 of 80 iterations **then**
      $Freeze(D)$, $Unfreeze(W)$
      $\mathcal{L}_{dis} \leftarrow \max_W \mathcal{L}_\rho$
    else
      $Unfreeze(D)$, $Freeze(W)$
      $\mathcal{L}_{dis} \leftarrow \min_D \mathcal{L}_\rho$ Eq. (6)
    end if
    $\mathcal{L} = \mathcal{L}_{det} + \lambda_1\mathcal{L}_{relate} + \lambda_2\mathcal{L}_{dis}$ Eq. (7)
    Optimize $\mathcal{L}$ to find optimal x_{relate} and x_{invar}
  end for
end for
return Disentangled features x_{relate} and x_{invar}

3.3 Adversarial Feature Learning

Due to the relatively small parameter size of the SIFD module, it is prone to insufficient training, resulting in deficient disentangling. Furthermore, the introduction of additional W_{invar} and W_{relate} may also cause a reduction in the value of ρ^2 in their training process. We need to ensure that the decrease in ρ^2 is attributed to the feature disentanglement rather than to projections W_{invar} and W_{relate}.

Inspired by Age-Invariant Adversarial Feature [Liu *et al.*, 2022] for kinship verification, we employ an Adversarial Feature Learning method to alleviate the above issues for UAV-OD. Specifically, we freeze detector $\mathcal{D}$ and perform gradient reversal [Ganin *et al.*, 2016], train W_{invar} and W_{relate} to maximize ρ^2. When ρ^2 reaches its maximum, we unfreeze $\mathcal{D}$, and freeze W_{invar} and W_{relate} to minimize ρ^2. Two operations are alternated until ρ^2 converges to a minimum.

This training process can be viewed as an adversarial game, where one side aims to maximize ρ^2, while the other side seeks to minimize it. The entire training process involves iteratively minimizing the maximum correlation coefficient ρ^2. For example, for every 80 iterations, we can take maximum in the first 30 steps and then minimize $\mathcal{L}_\rho$ in the next 50 ones. The training scheme can be formulated as follows:

$$\mathcal{L}_{dis} = \min_D \max_{W_{invar}, W_{relate}} \mathcal{L}_\rho. \quad (6)$$

Figure 4: Annotation comparison among State-Air, AU-AIR, and Syndrone. Green: labels given by the datasets; Yellow: revision of incorrect labels; Red: missed labels; Blue: negative labels.

Finally, by integrating scale-related and disentangling loss functions, our overall training target function becomes:

$$\mathcal{L} = \mathcal{L}_{det} + \lambda_1 \mathcal{L}_{relate} + \lambda_2 \mathcal{L}_{dis}, \tag{7}$$

where $\mathcal{L}_{det}$ represents the object detection loss, $\mathcal{L}_{relate}$ and $\mathcal{L}_{dis}$ denote the scale-related loss and the feature disentangling loss, respectively. λ_1 and λ_2 are balancing parameters.

To detect small-scale objects, we utilize the disentangled scale-invariant features x_{invar} as input to the high-resolution detection head, as shown in Fig. 2. Regarding scale-related features x_{relate}, we opt to discard them as they have the potential to be misleading and detrimental to the detection process. Furthermore, discarding redundant features can reduce computational complexity and slightly improve detection efficiency. We will substantiate this assertion through experimentation. Formally, our SIFDAL approach is illustrated in Algorithm 1.

4 State-Air Dataset

Existing UAV-OD datasets typically exclude additional data modalities related to the flight, such as those captured by internal sensors. Nevertheless, flight data (e.g., altitude) has the potential to be valuable and could contribute to UAV-OD research. To this end, we propose an aerial dataset (State-Air) with multi-modal sensor data collected in real-world outdoor environments. The multi-modal information incorporates aerial images, altitude, and UAV flight status from IMU (height, roll, pitch, and yaw).

The State-Air dataset was collected using DJI Mini2, a micro-UAV [Stanković *et al.*, 2021]. We designed an Android APP to obtain and store real-time images containing UAV flight status information through the DJI Mobile SDK. Finally, we gathered 2864 aerial images, including 2246 images of sunny days and 616 instances of snowy ones. Each image has a size of 1280 * 720 pixels and contains objects covering a wide variety of scales and shapes.

Dataset	S/R	Height[m]	View Angle	Weather
AU-AIR	R	5-30	45 to 90	no
SynDrone	S	20	30, 60, 90	no
State-Air	R	**5-75**	**0 to 90**	**yes**

Table 1: Comparison of UAV datasets with multi-modal sensor data. 'Weather' indicates involving multiple weather conditions. 'S/R' denotes Synthetic or Real-world data.

These aerial images were then annotated for four common object categories: person, car, bus, and van. Specifically, objects were initially annotated by Grounding DINO [Liu *et al.*, 2023] and subsequently manually proofread. This annotation method is expected to be more accurate and convenient compared to the traditional crowd-sourced image annotations.

Fig. 4 presents sample annotation results for comparing our proposed State-Air with existing two multi-modal UAV datasets, AU-AIR [Bozcan and Kayacan, 2020] and Syndrone [Rizzoli *et al.*, 2023]. It can be observed from Fig. 4 that AU-AIR's annotations tend to be coarse and incorrect, *e.g.*, missing labeling or annotating multiple cars with one box. Regarding Syndrone, several negative objects blocked by walls or trees are incorrectly labeled. Furthermore, it is a synthetic drone imagery dataset, which weakens its applicability to complicated natural scenes. Detailed statistical comparisons of three datasets are demonstrated in Table 1.

Given the above comparisons and analyses, our State-Air demonstrates the following superiorities over existing multi-modal UAV-OD datasets:

- State-Air provides more precise and detailed image annotations paired with state parameters and flight altitude.
- State-Air is captured in a real-world outdoor setting with a wide variety of scenes, including courts, buildings, squares, and roads.
- State-Air encompasses diverse challenging weather conditions including snowfall and rainfall, as well as a greater variety of heights and view angles.

5 Experiments

5.1 Datasets and Evaluation Metrics

We adopted three datasets that have UAV altitude information for experiments (AU-Air [Bozcan and Kayacan, 2020], SynDrone [Rizzoli *et al.*, 2023], and our State-Air). We employed the mean Average Precision (mAP) [Everingham *et al.*, 2010] as the evaluation metrics on accuracy, as well as GFLOPs on efficiency.

5.2 Experimental Setup

We adopted YOLOv7-L [Wang *et al.*, 2023], EfficientDet-d2 [Tan *et al.*, 2020] and RetinaNet [Lin *et al.*, 2017b] with ResNet50 [He *et al.*, 2016] as the baseline models. All experiments were conducted in Pytorch with three NVIDIA RTX 3090 GPUs. We trained the framework for 200 epochs with a batch size of 128. All detectors were trained using an Adam optimizer [Kingma and Ba, 2014] with a momentum of 0.937, and the learning rate was initialized as 0.001 with a cosine

Dataset		AU-AIR		SynDrone		State-Air		GFLOPS	
		AP_{50}	AP_{75}	AP_{50}	AP_{75}	AP_{50}	AP_{75}	Training	Test
YOLOv7-L	baseline	34.60%	7.4%	65.18%	34.3%	80.96%	16.5%	106.472G	106.472G
	ours	**36.92%**	**8.4%**	**68.32%**	**40.2%**	**89.98%**	**37.2%**	106.698G	103.136G
	improvements	**+2.32%**	**+1.0%**	**+3.14%**	**+5.9%**	**+9.02%**	**+20.7%**	-	-
EfficientDet-d2	baseline	29.08%	5.4%	31.90%	22.9%	63.23%	25.9%	22.394G	22.394G
	ours	**30.12%**	**7.2%**	**32.60%**	**23.7%**	**66.26%**	**26.5%**	22.746G	21.756G
	improvements	**+1.04%**	**+1.8%**	**+0.70%**	**+0.8%**	**+3.03%**	**+0.6%**	-	-
RetinaNet-R50	baseline	25.38%	5.9%	22.03%	15.6%	30.48%	10.6%	191.423G	191.423G
	ours	**27.93%**	**6.7%**	**24.14%**	**19.4%**	**31.91%**	**12.0%**	192.295G	185.217G
	improvements	**+2.55%**	**+0.8%**	**+2.11%**	**+3.8%**	**+1.43%**	**+1.4%**	-	-

Table 2: Comparison of AP and GFLOPs on three benchmark datasets by utilizing our approach with various base detectors. Our approach effectively improves detection accuracy and mildly reduces test inference costs on multiple detectors.

decay [Loshchilov and Hutter, 2017a]. During AFL, SIFD was optimized using AdamW [Loshchilov and Hutter, 2017b] with the same setting as Adam optimizer.

The number of height categories for the three datasets was 5, 3, and 8, respectively. The height of SynDrone was fixed in three classes: 20m, 50m, and 80m. The height categories of the other two datasets were obtained through k-means clustering. The value of k was determined by Silhouette Coefficient [Rousseeuw, 1987; Pelleg *et al.*, 2000].

5.3 Experimental Results and Analyses

As demonstrated in Table 2, our method outperforms all baselines on three benchmarks under a similar model size. In the majority of cases, our method leads to an improvement of more than 1.0% in the mAP_{50} and mAP_{75} scores. Furthermore, it improves detection efficiency by decreasing test computational complexity.

Results on AU-AIR. In this experiment, SIFDAL increased the mAP_{50} of YOLOv7-L, EfficientDet-d2, and RetinaNet-R50 by 2.32%, 1.04%, and 2.55%, respectively. It can be noted that the accuracy of all detectors is relatively low and the differences between them are marginal. The reason may be that the annotations of AU-AIR are not precise enough. In other words, the results on AU-AIR may not accurately reflect the model's performance.

Results on SynDrone. Our SIFDAL can achieve remarkable performance with mAP_{50} improvements by 3.14%, 0.70% and 2.11% on three detectors, respectively. YOLOv7 and EfficientDet both attain a moderate score on SynDrone among the three benchmarks. The reason may be that the abundance of negative samples in SynDrone misguides the training and test processes. Furthermore, as a synthetic dataset, SynDrone lacks authenticity to some extent and its practical application tends to be limited.

Datasets	SIFD	AFL	mAP_{50}	mAP_{75}
AU-AIR			34.60%	7.4%
	✓		35.72%	7.9%
	✓	✓	**36.92%**	**8.4%**
State-Air			80.96%	16.5%
	✓		87.62%	26.3%
	✓	✓	**89.98%**	**37.2%**

Table 3: Effectiveness analysis on our proposed SIFD and AFL with YOLOv7-L as the base detector.

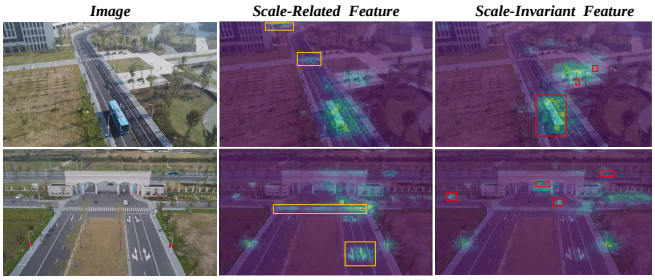

Figure 5: Visualization of scale-related and scale-invariant features. Scale-related features are not only focused on the foreground but also exist in the background (yellow boxes). While scale-invariant features exhibit a stronger concentration on objects (red boxes).

Results on State-Air. SIFDAL exhibits more significant improvements on State-Air than on the other two datasets. Specifically, it achieves remarkable mAP_{50} and mAP_{75} gains of 9.02% and 20.7% on YOLOv7, respectively. On RetinaNet and EfficientDet, the mAP_{50} improvements are 3.03% and 1.43%, respectively. All the detectors achieve their best performance on State-Air, which may be attributed to more precise and accurate labels.

5.4 Ablation Studies

We validated the effectiveness of each module, assessed the influence of SIFDAL at different layers, and examined the impact of scale-related and scale-invariant features on AU-AIR and State-Air datasets with YOLOv7-L as the base detector.

Effectiveness of SIFD and AFL

We conducted an ablation analysis of SIFD and AFL and assessed their impact on the final results in Table 3. By employing SIFD, the mAP_{50} on AU-AIR and State-Air increase by 1.22% and 6.66%, respectively. The performance gain can be attributed to employing scale-invariant features by SIFD. After adopting AFL, the detection accuracy further grows by 1.20% and 2.36% on two datasets, respectively. It can be concluded that our AFL method can achieve more thorough feature disentanglement, while enhanced scale-invariant features can further promote the model performance. The experimen-

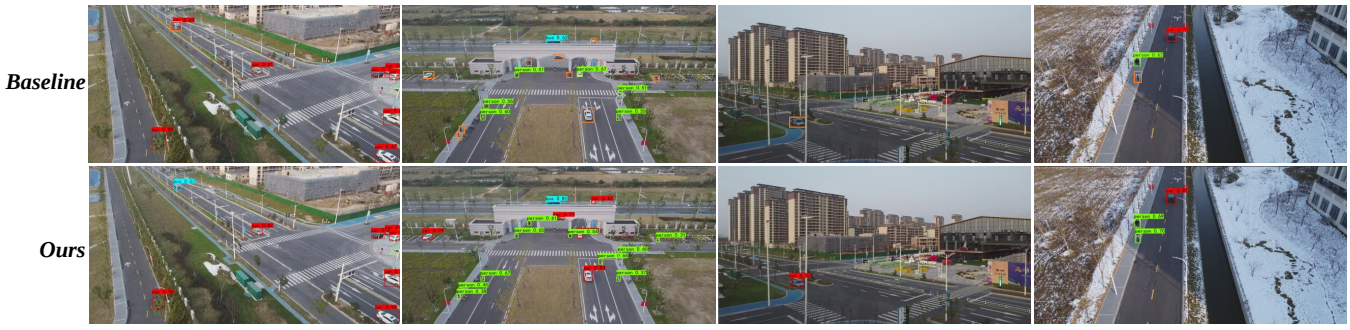

Figure 6: Visualization of detection results of our SIFDAL and baselines on State-Air. To be specific, red, green, and blue boxes represent car, person, and bus predictions, respectively. Orange boxes denote false or missing detection results.

Datasets	SIFDAL Location	mAP_{50}	mAP_{75}
AU-AIR	-	34.60%	7.4%
	P5	34.98%	7.4%
	P4	35.01%	8.0%
	P3	**36.92%**	**8.4%**
State-Air	-	80.96%	16.5%
	P5	84.79%	25.4%
	P4	87.81%	24.6%
	P3	**89.98%**	**37.2%**

Table 4: Influence of SIFDAL location with YOLOv7-L as the base detector. 'SIFDAL Location' represents different layers of FPN.

Datasets	x_{invar}	x_{relate}	mAP_{50}	mAP_{75}
AU-AIR			34.60%	7.4%
	✓		**36.92%**	**8.4%**
		✓	31.75%	6.9%
	✓	✓	35.36%	7.1%
State-Air			80.96%	16.5%
	✓		**89.98%**	**37.2%**
		✓	75.07%	14.7%
	✓	✓	83.59%	29.1%

Table 5: Impact of utilizing scale-related and scale-invariant features with YOLOv7-L as the base detector.

tal results demonstrate that both SIFD and AFL are effective and contribute to the final performance improvement.

SIFDAL at different FPN layers

We investigated the influence of the SIFDAL location on the effectiveness of disentanglement and conducted experiments at different resolution layers of FPN. The resolution of the feature map decreases sequentially from P3 to P5. The experimental results are demonstrated in Table 4. From the performance of the two datasets, we can observe that the detection accuracy gradually increases from layer P5 to P3, reaching its peak in the high-resolution layer (P3). This result indicates the significance of guiding high-resolution detection heads to leverage the scale-invariant features for UAV-OD tasks.

Impact of Scale-Related and Scale-Invariant Features

We leveraged the model without feature disentanglement as the baseline and conducted an experiment to verify the optimal features for UAV-OD, as reported in Table 5. The most notable improvement can be achieved by merely utilizing disentangled scale-invariant features x_{invar}, with an increase of 2.32% and 2.63% on AU-AIR and State-Air, respectively. Conversely, only using scale-related features x_{relate} results in the worst performance, even inferior to the baseline. Applying both two sets of features results in a 2.63% accuracy gain on State-Air, which is inferior to leveraging x_{invar} but outperforms the option to solely utilize x_{relate}. The reason can be that scale-invariant features contribute to the robust UAV-OD while scale-related features could interfere with the detection process. The experimental results suggest that scale-invariant features tend to be discriminative and can significantly benefit detection performance.

5.5 Visualization Analyses

Fig. 5 visualizes scale-invariant and scale-related features on two images. It can be observed that two types of features are visibly disentangled. Specifically, scale-related features exist in both foreground and background. For example, they exhibit evident activation on zebra crossings and surface marks, which tend to be inconducive to object detection. On the contrary, scale-invariant features mainly focus on objects (*e.g.*, car and bus). The visualization result demonstrates the disentangling effectiveness of our approach. Furthermore, it accounts for the reason that scale-invariant features can enhance the model performance.

We also compare the prediction results of our approach and the baseline (vanilla YOLOv7-L) in Fig. 6. It can be easily observed that the baseline tends to produce a few false or missing detection results (highlighted by orange boxes). Conversely, after employing our method, false and missing detections are suppressed. Specifically, our approach can accurately detect the bus in Fig. 6 (a), cars in (b) and (c), as well as people in (d). This experiment intuitively illustrates the effectiveness of our approach in boosting detection accuracy.

6 Conclusions

In this paper, we introduced a Scale-invariant Feature Disentanglement via Adversarial Learning (SIFDAL) method to enhance the UAV-based object detection accuracy. Specifically, we designed a Scale-Invariant Feature Disentangling

module and introduced an Adversarial Feature Learning training scheme to obtain discriminative scale-invariant features. Our SIFDAL can be employed in any FPN-based object detector and experimental results demonstrated the superiority of our approach. Furthermore, we constructed a multi-scene and multi-modal UAV-based object detection dataset, State-Air. It was captured in a real-world outdoor setting with a wide variety of scenes and weather conditions. We are committed to further enhancing the scope and scale of State-Air, expanding both the coverage and depth of our data.

References

[Bay *et al.*, 2008] Herbert Bay, Andreas Ess, Tinne Tuytelaars, and Luc Van Gool. Speeded-up robust features (SURF). *CVIU*, 2008.

[Behera *et al.*, 2023] Tanmay Kumar Behera, Sambit Bakshi, Michele Nappi, and Pankaj Kumar Sa. Superpixel-based multiscale CNN approach toward multiclass object segmentation from UAV-captured aerial images. *IEEE J-STARS*, 2023.

[Biswas *et al.*, 2022] Debojyoti Biswas, MM Mahabubur Rahman, Ziliang Zong, and Jelena Tešić. Improving the energy efficiency of real-time dnn object detection via compression, transfer learning, and scale prediction. In *IEEE NAS*, 2022.

[Bochkovskiy *et al.*, 2020] Alexey Bochkovskiy, Chien-Yao Wang, and Hong-Yuan Mark Liao. YOLOv4: Optimal speed and accuracy of object detection. *arXiv preprint arXiv:2004.10934*, 2020.

[Bozcan and Kayacan, 2020] Ilker Bozcan and Erdal Kayacan. Au-air: A multi-modal unmanned aerial vehicle dataset for low altitude traffic surveillance. In *ICRA*, 2020.

[Du *et al.*, 2018] Dawei Du, Yuankai Qi, Hongyang Yu, Yifan Yang, Kaiwen Duan, Guorong Li, Weigang Zhang, Qingming Huang, and Qi Tian. The unmanned aerial vehicle benchmark: Object detection and tracking. In *ECCV*, 2018.

[Du *et al.*, 2023] Bowei Du, Yecheng Huang, Jiaxin Chen, and Di Huang. Adaptive sparse convolutional networks with global context enhancement for faster object detection on drone images. In *CVPR*, 2023.

[Duan *et al.*, 2021] Chengzhen Duan, Zhiwei Wei, Chi Zhang, Siying Qu, and Hongpeng Wang. Coarse-grained density map guided object detection in aerial images. In *ICCV*, 2021.

[Duda *et al.*, 1973] Richard O Duda, Peter E Hart, et al. *Pattern Classification and Scene Analysis*. Wiley New York, 1973.

[Everingham *et al.*, 2010] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The Pascal visual object classes (VOC) challenge. *IJCV*, 2010.

[Figurnov *et al.*, 2017] Michael Figurnov, Maxwell D Collins, Yukun Zhu, Li Zhang, Jonathan Huang, Dmitry Vetrov, and Ruslan Salakhutdinov. Spatially adaptive computation time for residual networks. In *CVPR*, 2017.

[Ganin *et al.*, 2016] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario March, and Victor Lempitsky. Domain-adversarial training of neural networks. *JMLR*, 2016.

[Girshick *et al.*, 2014] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *CVPR*, 2014.

[Han *et al.*, 2023] Gujing Han, Ruijie Wang, Qiwei Yuan, Liu Zhao, Saidian Li, Ming Zhang, Min He, and Liang Qin. Typical fault detection on drone images of transmission lines based on lightweight structure and feature-balanced network. *Drones*, 2023.

[He *et al.*, 2016] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016.

[Huang *et al.*, 2022] Yecheng Huang, Jiaxin Chen, and Di Huang. UFPMP-Det: Toward accurate and efficient object detection on drone imagery. In *AAAI*, 2022.

[Kingma and Ba, 2014] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[Koyun *et al.*, 2022] Onur Can Koyun, Reyhan Kevser Keser, Ibrahim Batuhan Akkaya, and Behçet Uğur Töreyin. Focus-and-detect: A small object detection framework for aerial images. *SPIC*, 2022.

[Li *et al.*, 2017] Jianan Li, Xiaodan Liang, Yunchao Wei, Tingfa Xu, Jiashi Feng, and Shuicheng Yan. Perceptual generative adversarial networks for small object detection. In *CVPR*, 2017.

[Li *et al.*, 2020] Changlin Li, Taojiannan Yang, Sijie Zhu, Chen Chen, and Shanyue Guan. Density map guided object detection in aerial images. In *CVPRW*, 2020.

[Li *et al.*, 2021] Zewen Li, Fan Liu, Wenjie Yang, Shouheng Peng, and Jun Zhou. A survey of convolutional neural networks: analysis, applications, and prospects. *TNNLS*, 2021.

[Lin *et al.*, 2017a] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *CVPR*, 2017.

[Lin *et al.*, 2017b] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *ICCV*, 2017.

[Lin *et al.*, 2021] Qizhang Lin, Yan Ding, Hong Xu, Wenxiang Lin, Jiaxin Li, and Xiaoxiao Xie. Ecascade-RCNN: Enhanced cascade RCNN for multi-scale object detection in UAV images. In *ICARA*, 2021.

[Liu *et al.*, 2015] Baoyuan Liu, Min Wang, Hassan Foroosh, Marshall Tappen, and Marianna Pensky. Sparse convolutional neural networks. In *CVPR*, 2015.

[Liu *et al.*, 2018] Shu Liu, Lu Qi, Haifang Qin, Jianping Shi, and Jiaya Jia. Path aggregation network for instance segmentation. In *CVPR*, 2018.

[Liu *et al.*, 2022] Fan Liu, Zewen Li, Wenjie Yang, and Feng Xu. Age-invariant adversarial feature learning for kinship verification. *Mathematics*, 2022.

[Liu *et al.*, 2023] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *arXiv preprint arXiv:2303.05499*, 2023.

[Loshchilov and Hutter, 2017a] I Loshchilov and F Hutter. Stochastic gradient descent with warm restarts. In *ICLR*, 2017.

[Loshchilov and Hutter, 2017b] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

[Lowe, 1999] David G Lowe. Object recognition from local scale-invariant features. In *ICCV*, 1999.

[Lowe, 2004] David G Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, 2004.

[Lv *et al.*, 2023] Wenyu Lv, Shangliang Xu, Yian Zhao, Guanzhong Wang, Jinman Wei, Cheng Cui, Yuning Du, Qingqing Dang, and Yi Liu. DETRs beat YOLOs on real-time object detection. *arXiv preprint arXiv:2304.08069*, 2023.

[Meethal *et al.*, 2023] Akhil Meethal, Eric Granger, and Marco Pedersoli. Cascaded zoom-in detector for high resolution aerial images. In *CVPR*, 2023.

[Mittal *et al.*, 2020] Payal Mittal, Raman Singh, and Akashdeep Sharma. Deep learning-based object detection in low-altitude UAV datasets: A survey. *IVC*, 2020.

[Park *et al.*, 2023] Hye-Jin Park, Ji-Woo Kang, and Byung-Gyu Kim. ssFPN: Scale sequence (s^2) feature-based feature pyramid network for object detection. *Sensors*, 2023.

[Pelleg *et al.*, 2000] Dan Pelleg, Andrew W Moore, et al. X-means: Extending k-means with efficient estimation of the number of clusters. In *ICML*, 2000.

[Qian *et al.*, 2022] Cheng Qian, Huanxing Wu, Qirui Zhang, Lvshun Yang, and Qi Jiang. Design and implementation of UAV formation cooperative system. In *ICAUS*, 2022.

[Redmon *et al.*, 2016] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *CVPR*, 2016.

[Rejeb *et al.*, 2022] Abderahman Rejeb, Alireza Abdollahi, Karim Rejeb, and Horst Treiblmaier. Drones in agriculture: A review and bibliometric analysis. *Comput Electron Agr*, 2022.

[Ren *et al.*, 2015] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *NeurIPS*, 2015.

[Rizzoli *et al.*, 2023] Giulia Rizzoli, Francesco Barbato, Matteo Caligiuri, and Pietro Zanuttigh. Syndrone-multimodal UAV dataset for urban scenarios. In *ICCV*, 2023.

[Robicquet *et al.*, 2016] Alexandre Robicquet, Amir Sadeghian, Alexandre Alahi, and Silvio Savarese. Learning social etiquette: Human trajectory understanding in crowded scenes. In *ECCV*. Springer, 2016.

[Rousseeuw, 1987] Peter J Rousseeuw. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *JACM*, 1987.

[Srivastava and Prakash, 2023] Ashish Srivastava and Jay Prakash. Techniques, answers, and real-world uav implementations for precision farming. *WPC*, 2023.

[Stanković *et al.*, 2021] Miloš Stanković, Mohammad Meraj Mirza, and Umit Karabiyik. UAV forensics: DJI mini 2 case study. *Drones*, 2021.

[Su *et al.*, 2023] Jinya Su, Xiaoyong Zhu, Shihua Li, and Wen-Hua Chen. Ai meets UAVs: A survey on AI empowered UAV perception systems for precision agriculture. *Neurocomputing*, 2023.

[Suh *et al.*, 2023] Han-sok Suh, Jian Meng, Ty Nguyen, Vijay Kumar, Yu Cao, and Jae-Sun Seo. Algorithm-hardware co-optimization for energy-efficient drone detection on resource-constrained FPGA. *TRETS*, 2023.

[Tan *et al.*, 2020] Mingxing Tan, Ruoming Pang, and Quoc V Le. Efficientdet: Scalable and efficient object detection. In *CVPR*, 2020.

[Van Noord and Postma, 2017] Nanne Van Noord and Eric Postma. Learning scale-variant and scale-invariant features for deep image classification. *PR*, 2017.

[Wang *et al.*, 2020] Xinjiang Wang, Shilong Zhang, Zhuoran Yu, Litong Feng, and Wayne Zhang. Scale-equalizing pyramid convolution for object detection. In *CVPR*, 2020.

[Wang *et al.*, 2022] Congqing Wang, Di Luo, Yang Liu, Bin Xu, and Yongjun Zhou. Near-surface pedestrian detection method based on deep learning for UAVs in low illumination environments. *Optical Engineering*, 2022.

[Wang *et al.*, 2023] Chien-Yao Wang, Alexey Bochkovskiy, and Hong-Yuan Mark Liao. Yolov7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. In *CVPR*, 2023.

[Wu *et al.*, 2021] Xin Wu, Wei Li, Danfeng Hong, Ran Tao, and Qian Du. Deep learning for unmanned aerial vehicle-based object detection and tracking: A survey. *IEEE GRSM*, 2021.

[Yan *et al.*, 2018] Yan Yan, Yuxing Mao, and Bo Li. Second: Sparsely embedded convolutional detection. *Sensors*, 2018.

[Yang *et al.*, 2019a] Fan Yang, Heng Fan, Peng Chu, Erik Blasch, and Haibin Ling. Clustered object detection in aerial images. In *ICCV*, 2019.

[Yang *et al.*, 2019b] Xinghao Yang, Weifeng Liu, Wei Liu, and Dacheng Tao. A survey on canonical correlation analysis. *TKDE*, 2019.

[Yang *et al.*, 2022] Chenhongyi Yang, Zehao Huang, and Naiyan Wang. Querydet: Cascaded sparse query for accelerating high-resolution small object detection. In *CVPR*, 2022.

[Zhu *et al.*, 2018] Pengfei Zhu, Longyin Wen, Xiao Bian, Haibin Ling, and Qinghua Hu. Vision meets drones: A challenge. *arXiv preprint arXiv:1804.07437*, 2018.

[Zitar *et al.*, 2023] Raed Abu Zitar, Mohammad Al-Betar, Mohamad Ryalat, and Sofian Kassaymehd. A review of UAV visual detection and tracking methods. *arXiv preprint arXiv:2306.05089*, 2023.