drones

MDPI

Article

Multiscale Object Detection from Drone Imagery Using Ensemble Transfer Learning

Rahee Walambe [1,2], Aboli Marathe [1,3] and Ketan Kotecha [1,2,*]

[1] Symbiosis Centre for Applied Artificial Intelligence (SCAAI), Symbiosis International Deemed University (SIU), Pune 412115, India; rahee.walambe@scaai.siu.edu.in (R.W.); C2K18105958@pictsctr.onmicrosoft.com (A.M.)
[2] Symbiosis Institute of Technology, Symbiosis International Deemed University (SIU), Pune 412115, India
[3] Pune Institute of Computer Technology, Affiliated to Savitribai Phule Pune University, Pune 411043, India
* Correspondence: director@sitpune.edu.in

Abstract: Object detection in uncrewed aerial vehicle (UAV) images has been a longstanding challenge in the field of computer vision. Specifically, object detection in drone images is a complex task due to objects of various scales such as humans, buildings, water bodies, and hills. In this paper, we present an implementation of ensemble transfer learning to enhance the performance of the base models for multiscale object detection in drone imagery. Combined with a test-time augmentation pipeline, the algorithm combines different models and applies voting strategies to detect objects of various scales in UAV images. The data augmentation also presents a solution to the deficiency of drone image datasets. We experimented with two specific datasets in the open domain: the VisDrone dataset and the AU-AIR Dataset. Our approach is more practical and efficient due to the use of transfer learning and two-level voting strategy ensemble instead of training custom models on entire datasets. The experimentation shows significant improvement in the mAP for both VisDrone and AU-AIR datasets by employing the ensemble transfer learning method. Furthermore, the utilization of voting strategies further increases the 3reliability of the ensemble as the end-user can select and trace the effects of the mechanism for bounding box predictions.

Keywords: drone imagery; 2D object detection; ensemble techniques; voting strategies

Citation: Walambe, R.; Marathe, A.; Kotecha, K. Multiscale Object Detection from Drone Imagery Using Ensemble Transfer Learning. *Drones* **2021**, *5*, 66. https://doi.org/10.3390/drones5030066

Academic Editor: Anastasios Dimou

Received: 15 June 2021
Accepted: 21 July 2021
Published: 23 July 2021

1. Introduction

The number of computer vision (CV) tasks such as object detection and image segmentation have gained extreme popularity in the last few decades. Object detection (OD) is challenging and useful for detecting the various visual objects of a specific class (such as cars, pedestrians, animals, terrains, etc.) in the images. OD deals with the development of computational models and techniques and is one of the fundamental problems of computer vision. Moreover, it is a basis of other tasks such as segmentation [1–4], image captioning [5–7], and object tracking [8,9], etc. Thus, OD finds its usage in multiple domains such as face detection, pedestrian detection, and remote satellite detection, etc. In this work, we focus on OD from drone images of two separate datasets: the VisDrone 2019 test dev set [10,11] and the AU-AIR dataset [12] using our novel framework based on the algorithm proposed in [13].

OD tasks can be grouped into two: firstly, the general OD (e.g., different types of objects to simulate human vision) and secondly, the detection applications (e.g., face detection, pedestrian detection, etc.). Prior to 2014, traditional OD methods were based on handcrafted features and lacked an effective representation of images. Limited computational sources were also a particular challenge at that time. Viola and Jones [14] were the first to detect human faces and achieved faster computation with comparable accuracy by using the sliding windows approach. Integral images [15–17], feature selection [18] and

detection cascading techniques were incorporated. In 2005, Dalal et al. [19] proposed the Histogram of Oriented Gradients (HOG) feature descriptor. The HOG detector has long been part of many OD algorithms [20–22] and several CV applications. Another traditional method proposed in 2008 was the Deformable Part-Based Model (DPM) [20,23], which is essentially the extension of the HOG detector. A variety of improvements to DPM are reported in [21,24,25], which featured increased speed in detection and state-of-the-art results in the PASCAL object detection challenges.

After 2010, several deep learning-based algorithms have been implemented for OD tasks [26] based on the Convolutional Neural Networks (CNN) [27]. Deep CNN can learn the feature representations and extraction of images. With further modifications, it was applied successfully to the OD tasks. The seminal work in this was reported by Girshick et al. [28,29] by proposing the Regions with CNN features (RCNN) method. The OD tasks can be grouped into single-stage and two-stage detection methods based on the deep learning methods [9,30]. The standard evaluation parameter of these models is the mean Average Precision (mAP) [31]. Each of these models is reported with the corresponding mAP values for the specific OD task and domains. The following subsections survey the progress in single-stage and two-stage detectors, of which a few models have been selected for the experimentation in this study.

1.1. Two-Stage Methods

RCNN uses selective search to extract the object proposals or object candidate boxes [28]. Then, rescaling is applied to each proposal to convert to images of fixed size. This is fed into the CNN model trained on ImageNet (e.g., AlexNet [27]), which performs extraction of features, followed by a linear SVM classifier to predict and classify the object in each of these proposal regions.

Although RCNN outperforms the DPM to a considerable extent [32], its primary disadvantage is the redundant feature computation over overlapped proposal regions leading to the extremely slow processing speed. SPPNet [33] overcomes this problem by implementing the Spatial Pyramid Pooling (SPP) layer. This helps the convolutional network create fixed-length representation irrespective of the region's size in the image. This avoids rescaling and, subsequently, the repeated computations of the features in the convolutional layer. As a result, SPPNet is approximately 20 times faster than the R-CNN while maintaining the mAP; however, it still exhibits a few limitations, the most important being the two-stage process of OD and classification. Fast RCNN [34] was introduced to remedy these issues with SPPNet [28,33] and R-CNN. Fast RCNN simultaneously trains a detector and a bounding box regressor and has reported higher accuracy than SPPNet and RCNN while being approximately 200 times faster than RCNN.

Although the Fast RCNN offers advantages over the previous models, the detection speed is still constrained by the region proposal detection. Later, Faster R-CNN [35] proposed dealing with this issue by introducing the Region Proposal Network (RPN), combining various stages of the OD task. However, the computational redundancies of the Faster RCNN remain an issue. Later, various improvements such as RFCN [36] and Light head RCNN [37] Were proposed. Before FPN [38], most deep learning-based methods carried out the detection task in the top layer of the network. These methods did not consider the features of the deep layers of the CNN, which may be useful for object classification. Also, these features are not typically useful for localizing objects. To that end, FPN was developed for building semantics at all scales of resolution. As a result, FPN works very well for detecting objects with a wide variety of scales and has now become a basic building block of many latest detectors.

1.2. Single-Stage Methods

Proposed by R. Joseph et al. in 2015 [39], the YOLO (You Only Look Once) algorithm is faster and has higher mAP than earlier models. As opposed to its predecessors, YOLO does not employ the two-stage proposal detection and verification strategy. Instead, it

uses a singular neural network on the entire image, dividing its image into regions, then predicting bounding boxes and probabilities simultaneously for each region, making the entire process one stage and faster. However, the limitation of YOLO is observed in terms of low localization accuracy, especially for small objects. Subsequent improvements on YOLO have been proposed in [40,41], improving detection accuracy while increasing the detection speed. These versions [40,41] and the Single Shot Multi-Box Detector (SSD) [42] have provided solutions to this issue. SSD introduced multi-reference and multi-resolution detection techniques improving the detection accuracy for smaller objects. The primary difference between previous detectors and the SSD is that SSD can detect objects of different scales on different layers of the network instead of the previous ones that run detection on their top layers.

Although the single-stage methods have improved speed and simplicity, they typically have lesser accuracy as compared to the two-stage detectors. Lin et al. have identified class imbalance as the primary obstacle preventing one-stage object detectors and proposed RetinaNet [43] to combat this shortcoming. A new loss function termed 'focal loss,' a variation of standard cross-entropy loss, was introduced in RetinaNet, which eliminates the problem of class imbalance typically encountered in the training of the dense detectors. This enables the detector to focus more on the misclassified samples during training and achieves comparable accuracy to two-stage detectors while maintaining a higher detection speed.

There are multiple applications of OD in various domains, most notably in autonomous driving and surveillance. Various OD applications, including face detection, text detection, traffic light detection, sign detection, pedestrian detection, and remote sensing target detection, are attempted in the last two decades. For pedestrian detection, which is useful in autonomous driving, surveillance, and security applications, etc., has used multiple methods, including HOG detector [19] and ICF detector [44–46], with some improvement and variations. In recent years, Faster RCNN [35,47] is proposed. Small size and dense population are typical challenges in this OD application area [48–50]. Analyzing the progress of detectors in pedestrian detection have been studied from multiple angles [48,51–62], including lessons learned about pedestrian detection over time, benchmarks in detection, the progress of vision algorithms in this task, learning multilayer channel features and experimenting with deep CNNs. The second application of OD is in the detection of faces [63]. From smile detection in cameras to facial makeup in mobile apps has employed face detection. VJ detector [14] was one of the first methods employed for face detection. Rule-based methods [64,65], subspace analysis-based methods [66,67], learning-based methods [67,68] and SVM [69] have also been proposed.

Text detection also uses some of the aforementioned techniques, including [70,71] for applications such as assisting the visually impaired to read street signs and currencies [72,73], for detection of house numbers/car numbers [74–77], etc. Traffic Sign and Traffic Light Detection is also a variant of OD and has attracted attention over many years [78]. Colour detection methods are usually based on colour thresholding [78–80], visual saliency detection [81], morphological filtering [82], and edge/contour analysis [83,84], etc., are proposed. In addition to these broad applications, Remote-Sensing Detection is one of the recent application areas of OD. It includes OD from satellite or drone data, and employs remote-sensing target detection. Remote sensing has many applications, such as safety and security, surveillance, disaster rescue, and urban traffic management [85–87]. One of the primary works that forms the basis of Remote Sensing Detection is OD using drone imagery.

The detection of everyday objects in drone-captured images is a fascinating challenge, with diverse autonomous driving, navigation, and tracking applications. With a surge in demand for autonomous vehicles (ground, aerial, or water), computer vision is pushing its records to equip these vehicles with the power to identify the components of the diverse environments they are exposed to. Special embedded hardware and software are being introduced regularly for faster processing that can be integrated easily with

UAV systems. For autonomous navigation, basic UAV systems need to contain at least an airframe and a computer system combining sensors, GPS, servos, and CPUs. The embedded computing systems include Commercial-Off-The-Shelf (COTS) computing platforms, Single Board Computers (SBCs) and processing modules [88].

Specifically, for image processing, newer devices support video compression and streaming, object detection and GPS/INS or AHRS systems. Some state-of-the-art image processing platforms for drones include Raspberry PI, Jetson TK1 Development Kit and Mini-ITX platforms. Equipped with special features like Nvidia GPUs for vision algorithms, support for OpenCV, SSD hard drives and other features, these platforms can continuously process images and videos at high speed with low power consumption.

The improvement of current UAV systems in terms of accuracy, efficiency, tracking and deployment [89] has been of great interest with the rapid integration of UAV systems with diverse applications ranging from drone-based delivery services to critical military software. Several software and patents integrating these object detection models with embedded hardware and computing resources have also been making progress in aiding autonomous vehicles for navigation, surveillance, and more applications. The Berkeley UAV platform [90] was one of such significant contributions towards enabling autonomously operating UAV teams. It presents a solution that demonstrates autonomous vision-based navigation and obstacle avoidance using many different sensor and controller processes. An important open-source hardware and software-based system for aerial surveying [91], implemented using a commercial, remote-controlled model airplane and an open-source GPS/IMU system, was able to capture images of desired spatial resolution in field testing. The increasing utilization of UAVs has naturally led to the development of patents like air traffic control for avoiding collisions between drones and crewed aircraft. Some patents have been successful in tackling this challenge including a robust, decentralized UTM system [92], an automated UAV traffic control system [93] and an airfield conflict resolution patent proposed for commercial airliners.[94] The rapid development of UAV-based systems also raised concerns for the ethical usage of UAVs and the discussion of benefits and misapplications led to the development and suggestion of several guidelines for their usage by several authorities and researchers [95–97].

The extension of this work can be seen in the remote sensing data for the orthomosaic generation where an interactive drone map is generated from many drones derived images called orthophotos and are stitched together [98,99]. An orthophoto derived from the drone captured images needs to be modified to remove the problems such as lens distortion, tilt, perspective due to heights and earth's elevation and lighting conditions. Due to these issues, the central part of the orthomosaic has typically better positional precision as compared to the areas along the edges [100]. The work proposed in this research for drone-based object detection can be employed for generating better quality orthomosaics by multiscale object detection, especially for the objects which are present around the edges of the orthophoto. With the specific techniques described in this work, such as color augmentation and ensembling, the object detection around the edges of an orthophoto can be improved for various height and lighting conditions. This is of high relevance and importance for detecting the change over time based on the drone-based orthophotos. The proposed methods can be adopted for processing, colour balancing and optimizing the drone-based images processing tasks for orthomosaic generation [101].

In particular, the challenge of OD in UAV images is difficult due to the lack of drone datasets and the large orientation and scale variance in drone images. Aerial images captured by drones and other autonomous vehicles have challenged computer vision models for many reasons. The size of the objects in the captured images varies from very large objects like trucks and planes to small objects like traffic cones and plants which are difficult to detect with high accuracy. The images contain a wide variety of terrains, backgrounds, and irregular objects like garbage bags. Furthermore, the resolution of UAV images is often low, and poor resolution causes objects to appear blurred, which makes the object detection difficult. Focusing on the detection classes, which are imbalanced with

certain types of objects poorly represented, the size of the objects varies greatly from small to large objects depending on the angle, elevation, and view of the drone. Various state-of-the-art methods have been introduced to the field to conquer these issues, capturing the objects with increasing precision with time.

For the first dataset used in this study—the VisDrone dataset—custom models such as DPNet-ensemble, RRNet, and ACM-OD [10] when trained and tested on the UAV dataset produced the best results. However, these approaches do not utilize the concept of transfer learning directly from the baseline models independent of the target data. They need to be trained on the target data, which is more time-consuming and less efficient. To that end, in this work, to tackle these challenges in UAV image datasets and to solve the general problem of deficient UAV detection sets, we propose and demonstrate an implementation of an ensemble algorithm by using the transfer learning from baseline OD models and data augmentation technique on the VisDrone 2019 dataset [10,11] and AU-AIR dataset [12].

In this paper, we explore both data augmentation and ensemble techniques for OD in drone images, which is an important challenge in computer vision due to the deficiency of UAV image datasets. The goals of the work presented in this paper can be summarized below.

A. We have experimented with several OD algorithms specified above and carried out extensive research to identify their suitability for detecting various scaled objects.
B. To solve the lack of UAV datasets, we applied the test-time augmentation on the drone images datasets to boost the accuracy of OD and ensemble models. A comprehensive study of the performance of the ensemble method and voting strategies was conducted, and we have demonstrated the effects of using test-time augmentation.
C. A framework combining multiple OD algorithms (both single-stage and two-stage) using the ensemble approach is proposed and demonstrated. We have implemented a general method for ensembling OD models independently of the underlying algorithm. This multi-technique ensemble algorithm is designed to choose any three approaches and experiment with OD, enabling multiscale OD by applying several voting strategies. This method effectively detects objects over a range of scales, from small-scale objects like humans, plants, and bikes to medium-like cars and larger objects like cargo trucks. Due to the change in camera angles in the images in the VisDrone and AU-AIR datasets, ordinary objects appear smaller or larger than actual, and this detection has been handled as well. The performance of the ensemble and augmentation techniques was better than the baseline models when tested on the test-dev set of the VisDrone dataset [10,11] and the AU-AIR dataset [12].

The paper is organized as follows: Section 2 discusses the datasets along with the individual OD methods implemented in this work in detail and introduces the novel framework that employs the two-level ensemble technique. Results and experimental analysis are presented in Section 3, followed by the discussion in Section 4. Finally, the paper concludes with Section 5 and the future prospects for this work in Section 6.

2. Materials and Methods

2.1. Datasets

In this study, we performed experiments on two UAV image datasets that contain images from diverse environments with multiple objects of different types and sizes and from different locations. There are several available datasets for satellite images, however, UAV imagery datasets with varying resolutions, multiscale objects, different lighting conditions and multiple objects present are limited. We selected these UAV datasets due to the diversity of objects present in them, multi-angle views of the scenes, diversity in lighting conditions, location, and the quantity of data.

2.1.1. VisDrone Dataset

Identifying objects in UAV images has been a topic of interest for computer vision researchers for drone-based applications and autonomous navigation. To facilitate research in this domain, the VisDrone datasets [10,11] were created. Presented as a challenge in OD and tracking, this dataset was tackled with a combination of state-of-the-art models and ensemble detection techniques. The top three detectors were DPNet-ensemble, RRNet, and ACM-OD, achieving 29.62%, 29.13%, and 29.13% Aps, respectively. However, the best detector DPNet-ensemble achieved less than a 30% AP score, demonstrating the need for improvement in this area when we think about usage in real applications. The VisDrone-DET2019 Dataset, captured by drone platforms, contains 8599 images in different places at different heights, which is the same data as the VisDrone-DET2018 Dataset [10,11]. The annotations cover 10 predefined categories and contain 540k bounding boxes of target objects. These categories are van, bus, person, truck, motor, awning-tricycle, bicycle, pedestrian, car, and tricycle. The dataset is divided into 6471 images for the training subset, 548 images for validation, and 1610 for the testing subset collected from 14 different cities in China with different environments. The input size of the images used is 1360 × 765. The maximum resolution of the images in the dataset is 2000 × 1500 pixels. For this analysis, the test dev set of 1610 images has been selected to test the detectors' performance.

2.1.2. AU-AIR Dataset

This dataset is a multi-modal UAV dataset containing UAV images from videos of 2 h (8 video streams) of traffic surveillance recorded at Skejby Nordlandsvej and P.O Pedersensvej roads (Aarhus, Denmark). [12] For the UAVs used to capture video for the dataset, the flight altitude changes between 10 m to 30 m in the videos, and the camera angle is adjusted from 45 degrees to 90 degrees. The input size of images used is 1920 × 1080. The maximum resolution of the images in the dataset is 1920 × 1080 pixels. The dataset covers images taken over a broad range of lighting conditions and in different weather conditions, including sunny, cloudy, and partly cloudy. The entire dataset covers eight object categories for detection: person, car, bus, van, truck, bike, motorbike, and trailer but the annotated bounding boxes contain mostly three vehicle types (car, van, and truck). The baseline networks YOLOv3 and MobileNetV2-SSDLite achieved 30.22 and 19.50 mAP, respectively. For this analysis, 1000 images were selected from this dataset, which resulted in 4000 images after augmentation, on which the results were tested.

2.1.3. Handling the Dataset Challenges

Object detection in UAV images is difficult due to the limited drone datasets and the large orientation and scale variance in drone images. To conquer these issues, data augmentation and ensembling procedures are suitable.

2.2. Methods

2.2.1. Data Augmentation

For biodiversity detection in the wild, researchers proposed innovative data augmentation techniques like multiple rotated copies of the original image [102,103], horizontal and vertical flipping [104], mirroring (horizontal and vertical), rotations, shifting (horizontal and vertical) [105] to obtain the best results. For other UAV dataset applications like vehicle and OD, researchers popularly use scaling, rotation, and blurring [106], rotation over 0 to 360 degrees [107], histogram equalization, Gaussian blur, random translation, scaling, cutout, and rotation [108]. The complete list of data augmentation techniques used in this study is included in Table 1.

Table 1. Data augmentation techniques used in the study.

Average blurring	None
Bilateral blurring	Raising the blue channel
Blurring	Raising the green channel
Changing to HSV color space	Raising the hue
Blurring the image	Raising the red channel
Cropping the image	Raising the saturation
Dropout	Raising the value
Elastic deformation	Resizing the image
Equalize histogram technique	Rotation by 10°
Flipping the image vertically	Rotation by 90°
Flipping the image horizontally	Rotation by 180°
Flipping the image vertically and horizontally	Rotation by 270°
Applying Gamma correction	Adding salt and pepper noise
Gaussian blurring	Sharpen the image
Adding Gaussian noise	Shifting the channel
Inverting the image	Shearing image

2.2.2. Object Detection Models

RetinaNet (Resnet50 Backbone)

This is a single-stage detector [43] that uses focal loss and is very fast and accurate. It is a medium-sized model often used to speed over single-stage detectors and greater accuracy over two-stage detectors. It utilizes a backbone of ResNet+FPN, which extracts the features and contains two task-specific subnetworks that perform bounding box regression and classification. This feature is particularly useful for the multiscale classification that we are attempting. This model was pre-trained on the COCO dataset [109].

YOLO (v3)

This model [41] is an adaptation of the Darknet architecture, with 53 more layers stacked onto the usual 53 layers, giving a total of 106 layers building the fully convolutional underlying architecture for YOLO v3. Known for its speed and accuracy, this improves OD by supporting the detection of small objects. This model was pretrained on the VOC dataset. [110]

SSD (Resnet Backbone)

SSD stands for Single-Shot Detector and is a single-stage OD model [42]. A single feed-forward convolutional network produces a fixed-size collection of bounding boxes and scores in this model, with a final non-maximum suppression step to produce the final detections. The Resnet backbone acts as a feature extractor, and thus, combined provides a model that preserves the spatial and semantic context of the input image. This model was trained on the VOC dataset. [110]

2.2.3. Ensembling Procedure

Ensemble techniques are very popular for OD in common datasets like Pascal VOC and COCO [109,110]. In 2011, an ensemble Exemplar SVM framework [22] achieved comparable SOTA performance to the complex latent part-based model of Felzenszwalb et al. [24]. In 2015 an ensemble of deep CNN models with different architectures outperformed the SOTAs on the VOC dataset. [111] Another interesting approach was the use of NMS Ensembling and Feature Ensembling, which achieved great results on the MS COCO set and the PASCAL VOC set [112].

For detecting cars, people (pedestrians), buildings, terrains in the urban and rural backgrounds of UAV images of the VisDrone dataset [10,11] and AU-AIR dataset [12], we

used different combinations of the models and train test augmentation to deliver the results. Our implementation of ensemble methods for OD is based on the algorithm proposed by Garcia et al. al. [13], who created this algorithm for regular OD on datasets like VOC [110] and achieved great performance. We apply a two-level voting strategy ensemble, as shown in Figure 1, both on the single-model and meta-model levels, which differs from the single-level ensemble in the original paper.

To obtain the final predictions, we applied various voting strategies to obtain the results:

- Affirmative: In this strategy, the initial predictions are considered, and when one of the ensemble methods detects an object, it is considered valid.
- Consensus: In this strategy, an agreement of the majority of the ensemble methods must be met for successful detection. It mirrors the majority voting strategy that is usually applied in ensemble methods for OD and image classification.
- Unanimous: In this strategy, all the ensemble methods need to agree upon detection to be deemed valid. This is the most stringent of voting strategies.

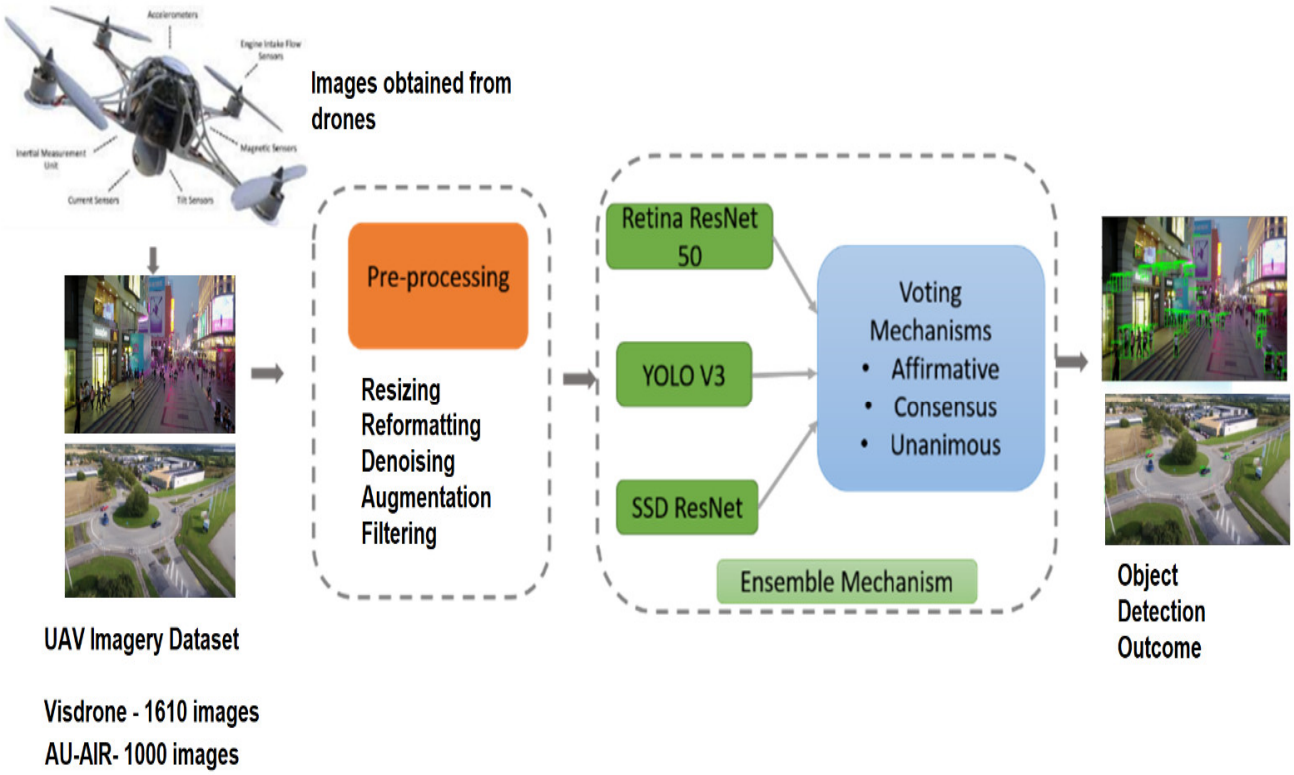

Figure 1. Complete pipeline for the developed approach.

The working of the two-level ensemble procedure works as described below:

1. Before the procedure is initiated, the following parameters are required to be set according to the user's intended application.
   a. Dataset—The target dataset on which predictions are to be made must be provided to the pipeline.
   b. Models—The list and weights of pretrained models to be ensembled together must be provided.
   c. Augmentation techniques- Any combination of augmentation techniques can be selected from Table 1 and provided to the model.

   d. Number of levels- The ensembling can be one level or two-level. One level model will only perform data augmentation for a single model. Two-level models will integrate augmentation and multiple models.
   e. Voting strategy- Out of affirmative, unanimous and consensus, the voting strategies for both levels need to be provided to the model.
   f. Ground truth—For measuring the performance, the ground truth labels with the list of classes must be provided.
   g. Performance Metric—According to the user preference, a performance metric needs to be selected. For this study setup, the VisDrone metrics, AP and mAP were configured.
2. Each model starts detecting objects in parallel to the other models and uses the voting strategy to combine the detection results into single xml files. At this level, the data augmentation is used to improve the model performance.
3. The results of all individual models are now processed together using a single selected voting strategy and the final bounding boxes are determined.
4. The results are compared with the ground truth labels and measured using the selected performance metric.
5. Visualization of the bounding boxes, percentage of wrongly assigned labels, predicted label distribution and overall results are generated for comprehensive result analysis.

In this study, we experimented extensively with the structure of the ensemble framework, input parameters and have presented the best framework after observing the results. The affirmative and consensus strategies support the detection of multiscale objects by combining the results of detectors that work well on smaller and larger objects. A sample working demonstration of this pipeline can be seen in Figure 2, in which an UAV image from the VisDrone dataset [10,11] is passed through the two-level framework and the object detection predictions are obtained in the results.

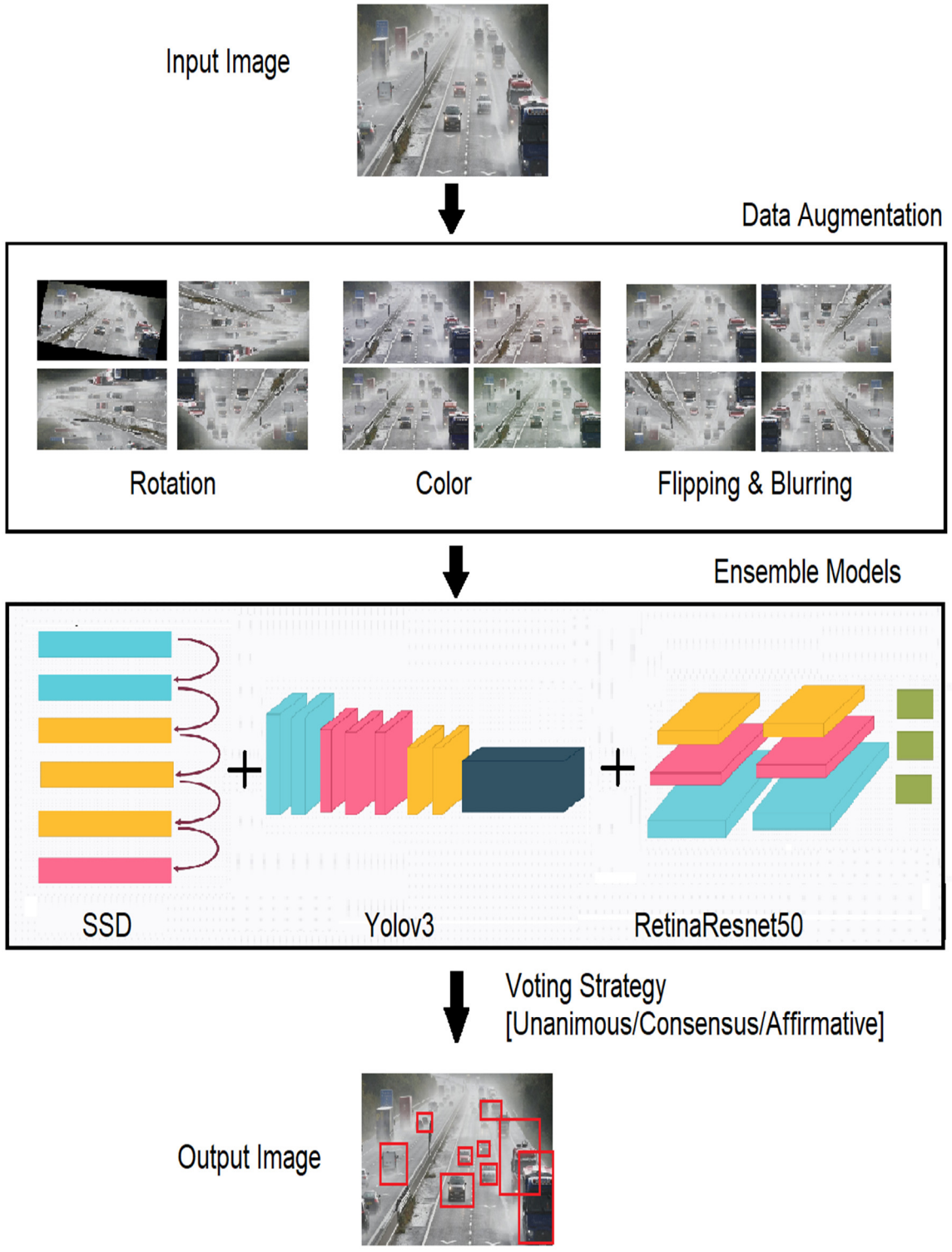

Figure 2. Ensemble framework architecture.

2.2.4. Model Performance Assessment

VisDrone Results Assessment

To evaluate the performance of the methods, we have used the evaluation protocol in MS COCO [109], AP, AP50, AP75, AR1, AR10, AR100, and AR50 metrics. AP is calculated by averaging over all 10 Intersection over Union (IoU) thresholds of all categories, which is used as the primary metric for ranking. AP50 and AP75 are computed at the single IoU thresholds 0.5 and 0.75 overall categories. The AR1, AR10, AR100, and AR500 scores are the maximum recalls given 1, 10, 100, and 500 detections per image, averaged over all categories and IoU thresholds.

AU-AIR Results Assessment

For evaluating the performance of the methods, the evaluation protocol of the AP metric is used. AP is calculated for each separately, and mAP is average overall categories, which is used as the primary metric for ranking.

3. Experiments and Results

To validate the ensemble methods and data augmentation techniques and evaluate their performance, we conducted a series of experiments on the UAV images. First, the three baseline models (RetinaNet, YOLO Darknet, and SSD) and three voting strategies (consensus, unanimous and affirmative) were tested to predict objects on the VisDrone 2019 Test-dev set [10,11] and AU-AIR dataset [12] using different augmentation techniques.

Initially, the experiments were carried out using one baseline model and different data augmentations to determine the best augmentation for the UAV images for both datasets. Then, the second set of experiments took the best augmentation selected from the first experiment and tried the various deep learning OD algorithms.

3.1. VisDrone Dataset Results

For the VisDrone dataset, Table 2 contains the AP and recall metrics showing the performance of the SSD model before and after augmentation for the VisDrone dataset. Table 3 reports the AP and recall metrics showing the performance of the ensemble of the three models, namely RetinaNet, SSD, and YOLOV3, using the affirmative voting strategy. The experimentation consisted of all the voting strategies. However, in this paper, we have reported the best performance provided by the affirmative voting strategy. Table 4 contains the class-wise AP scores showing the performance of the ensemble models with color augmentation. Figures 3 and 4 highlight the comparative performance of the best models in bar plots.

Table 2. Performance before and after application of augmentation for the SSD baseline model.

Algorithm	Augmentation	AP	AP50	AP75	AR1	AR10	AR100	AR500
SSD	No Augmentation	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
SSD	Best Augmentation (Color)	**0.0%**	**0.02%**	**0.00%**	**0.01%**	**0.06%**	**0.07%**	**0.07%**

Table 3. Top overall performance of ensemble and baseline models on Test-Dev set with color [RaiseBlue, RaiseGreen, RaiseRed, RaiseHue, RaiseSatu] augmentation and affirmative voting strategy for VisDrone dataset.

Algorithms	Voting Strategy	AP	AP50	AP75	AR1	AR10	AR100	AR500
RetinaNet	Affirmative	**1.66%**	**3.13%**	**1.60%**	**0.40%**	**2.30%**	3.27%	3.27%
YOLO + RetinaNet	Affirmative	1.07%	2.07%	1.02%	0.40%	2.29%	**3.29%**	**3.29%**
YOLO +SSD+RetinaNet	Affirmative	1.01%	1.95%	0.95%	0.07%	1.37%	**3.29%**	**3.29%**

Table 4. Top class-wise performance of ensemble and baseline models on Test-Dev set with color [RaiseBlue, RaiseGreen, RaiseRed, RaiseHue, RaiseSatu] augmentation. The results show the class-wise best AP obtained by individual baseline model and the combination of models using ensembling by various voting strategies.

Algorithms	Voting Strategy	Person	Bicycle	Car	Truck	Bus
RetinaNet	Consensus	2.10%	**0.43%**	5.88%	0.67%	0.89%
YOLO + RetinaNet	Consensus	1.29%	**0.43%**	3.72%	0.67%	0.69%
RetinaNet+SSD	Consensus	1.47%	**0.43%**	4.53%	0.67%	0.63%
RetinaNet	Affirmative	2.05%	0.32%	**6.09%**	**0.69%**	**0.90%**
YOLO + RetinaNet	Affirmative	1.26%	0.32%	3.79%	**0.69%**	0.70%
RetinaNet+SSD	Affirmative	1.42%	0.32%	4.67%	**0.69%**	0.66%
YOLO+SSD+RetinaNet	Affirmative	1.16%	0.32%	3.58%	**0.69%**	0.58%
RetinaNet	Unanimous	**2.12%**	0.38%	5.81%	0.63%	0.81%

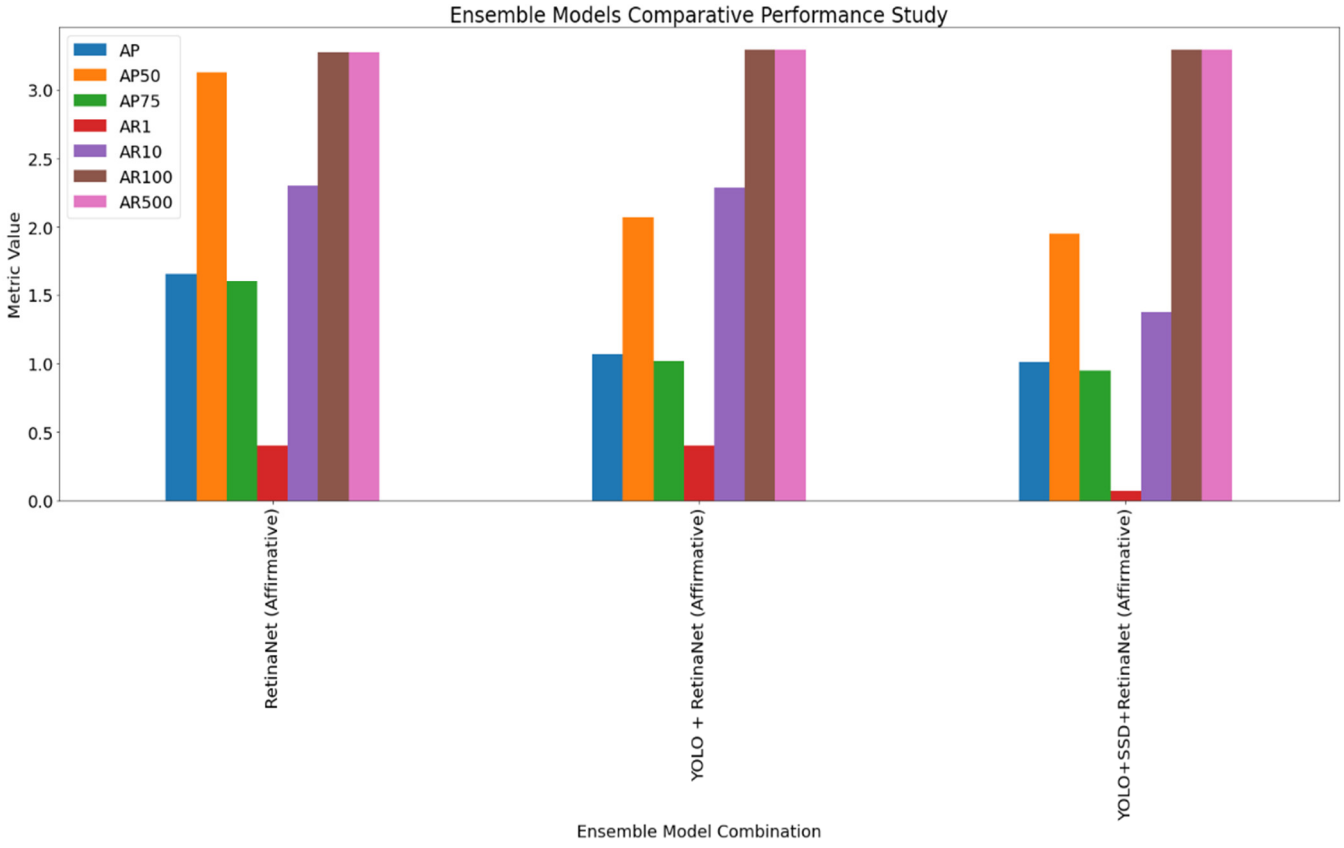

Figure 3. Top-performing ensemble models on the VisDrone dataset.

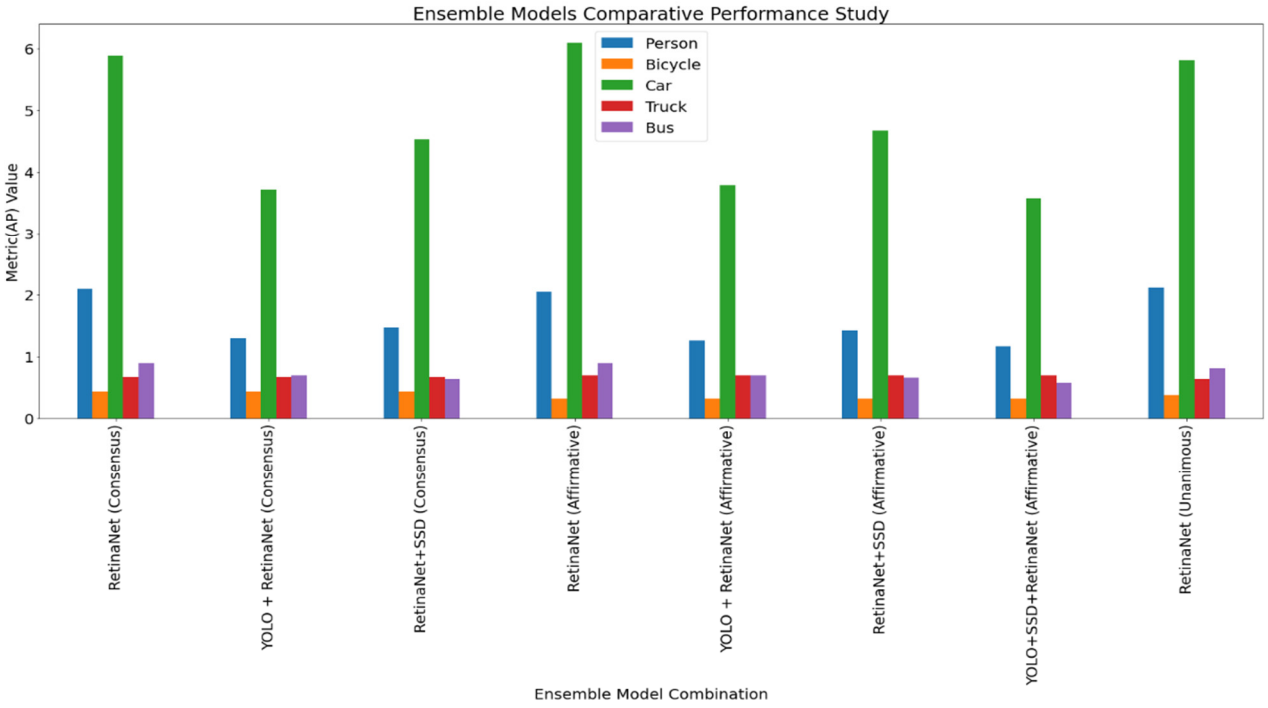

Figure 4. Top class-wise performing ensemble models for the VisDrone dataset.

Figures 5–7 demonstrate the raw and annotated images with the results of ensemble models, augmentation, and voting strategies, to visualize the working of this ensemble pipeline experiments. In Figure 5, we observe the effectiveness of the color augmentation techniques in detecting pedestrians and background objects.

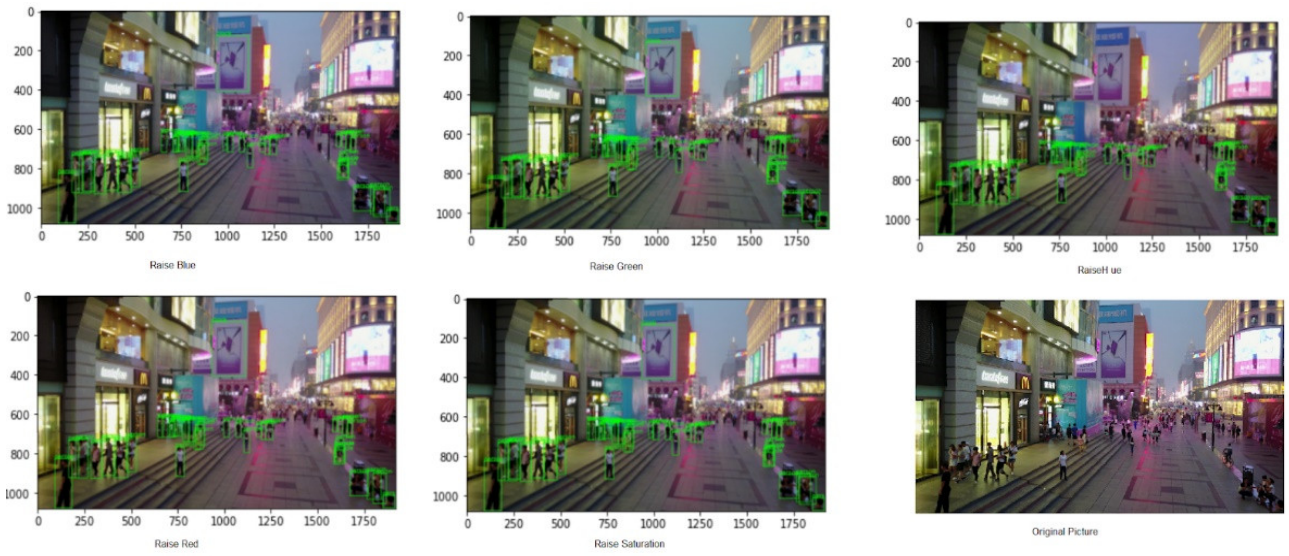

Figure 5. Comparison of data augmentation techniques when tested with the SSD Resnet detector.

In Figure 6, the number of detections per voting strategy has been shown, helping us understand how the strategies' predictions appear when applied to the real data. As expected, unanimous strategy which expects all collaborating models to agree upon all predictions, has the minimum number of detections and an affirmative strategy that accepts any one set of predictions has the maximum number of detections.

Figure 6. Comparison of voting strategies for all three detectors ensembled together with color augmentation.

In Figure 7, the multiscale object detection results have been presented. The ensemble model is able to detect objects of multiple scales ranging from pedestrians to large cargo trucks.

Figure 7. Multiscale object detection.

The purpose of these experiments was to determine the best augmentation techniques for the VisDrone images. For this experiment, the best results were achieved by combining [raiseBlue, raiseGreen, raiseHue, raiseRed, raiseSatu] techniques, with

consensus, unanimous and affirmative voting strategies indicating that color-based augmentation proved to be most effective for both overall performance and class-wise performance. In addition, the class-wise results showed superior performance for the same color augmentation techniques for detecting bicycles than with hvflip and Gaussian blur. Choosing [raiseBlue, raiseGreen, raiseHue, raiseRed, raiseSatu] as the technique, we then examined the results of the ensemble models and baseline models. For the class-wise performance, the RetinaNet detector with unanimous strategy worked best for detecting people. YOLO + RetinaNet, RetinaNet + SSD, and RetinaNet using consensus strategy worked well for detecting bicycles. On the other hand, RetinaNet, with affirmative strategy, was able to detect cars, trucks, and buses. RetinaNet, YOLO + RetinaNet, RetinaNet+SSD, YOLO+SSD+RetinaNet worked best for detecting vans using affirmative strategy. Among all the techniques and detectors, the best class prediction was for detecting bicycles by the RetinaNet detection using affirmative strategy and Color [raiseBlue, raiseGreen, raiseHue, raiseRed, raiseSatu] augmentation. Observing the overall performance, the RetinaNet detector with affirmative strategy performed best across the dataset for all metrics and all classes.

3.2. AU-AIR Dataset Results

For the AU-AIR Dataset, Table 5 contains the results of the baseline model before and after augmentation, and Table 6 contains the top class-wise AP results for the AU-AIR dataset Figure 8 highlights the comparative performance of the models in bar plots. . Figures 9–12 demonstrate the raw and annotated images to visualize the working of this ensemble pipeline experiments.

Table 5. Class wise performance of best baseline model on AU-AIR dataset before and after augmentation using RetinaNet.

Model	Augmentation	Person	Car	Truck	Bus	mAP
RetinaNet	Without Augmentation	3.41%	5.59%	5.24%	21.05%	4.41%
RetinaNet	With Color Augmentation	**8.57%**	**6.61%**	**6.89%**	**30.95%**	**6.63%**

Table 6. Top class-wise performance of ensemble and baseline models on AU-AIR dataset with [RaiseBlue, RaiseGreen, RaiseRed, RaiseHue, RaiseSatu] augmentation.

Model	Voting Strategy	Person	Car	Truck	Motorbike	Bus	mAP
RetinaNet	Unanimous	**8.57%**	**6.61%**	**6.89%**	0.00%	**30.95%**	**6.63%**
RetinaNet	Consensus	3.88%	6.21%	5.73%	**1.14%**	20.86%	4.73%
YOLO + RetinaNet	Consensus	3.45%	3.95%	5.73%	**1.14%**	15.23%	3.69%
RetinaNet + SSD	Consensus	3.45%	4.62%	5.73%	**1.14%**	17.31%	4.03%

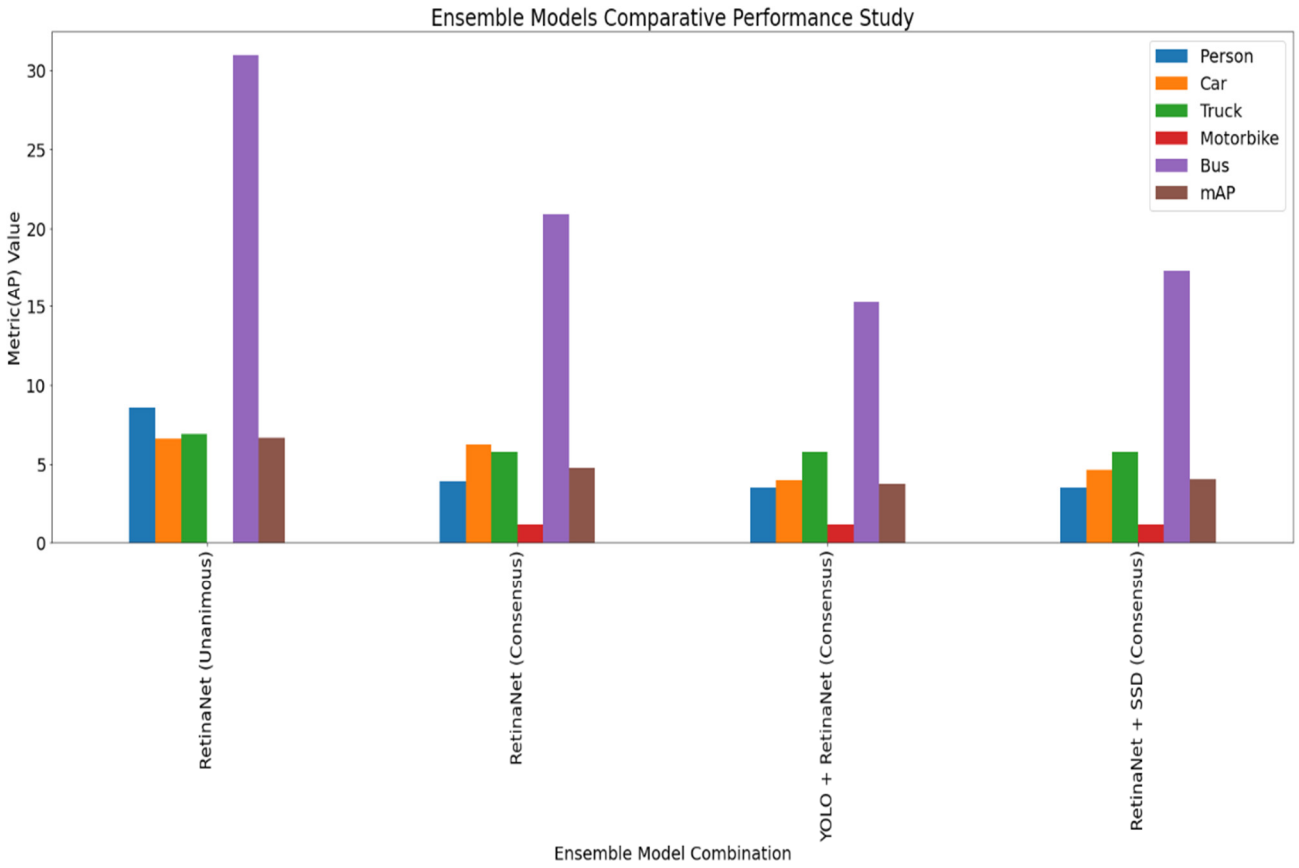

Figure 8. Top class-wise performing ensemble models on AU-AIR dataset.

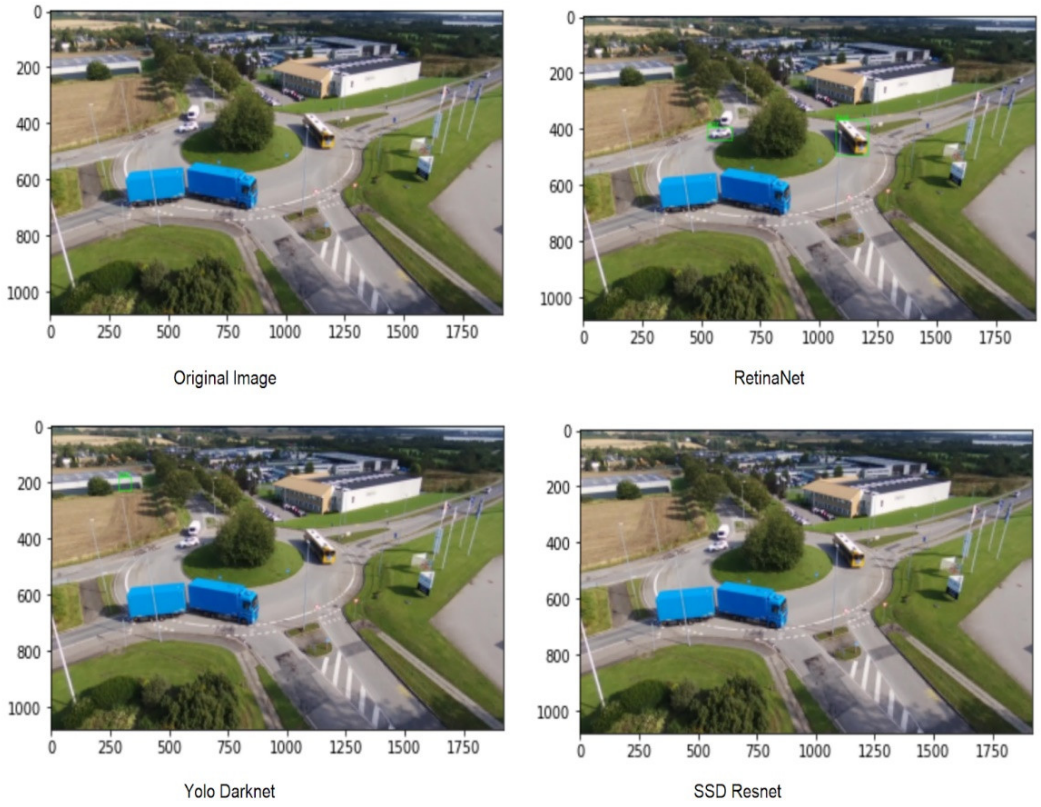

Figure 9. Comparison of baseline models on the dataset without any augmentation.

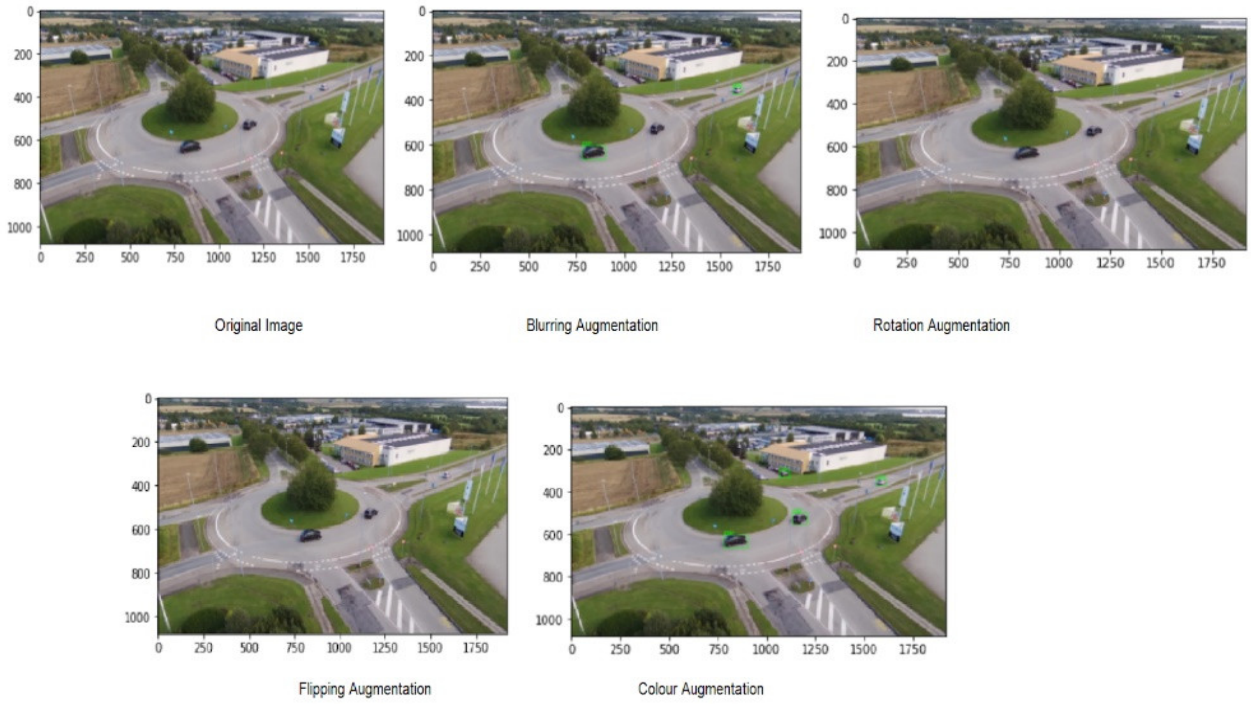

Figure 10. Comparison of RetinaNet detector using different data augmentation techniques.

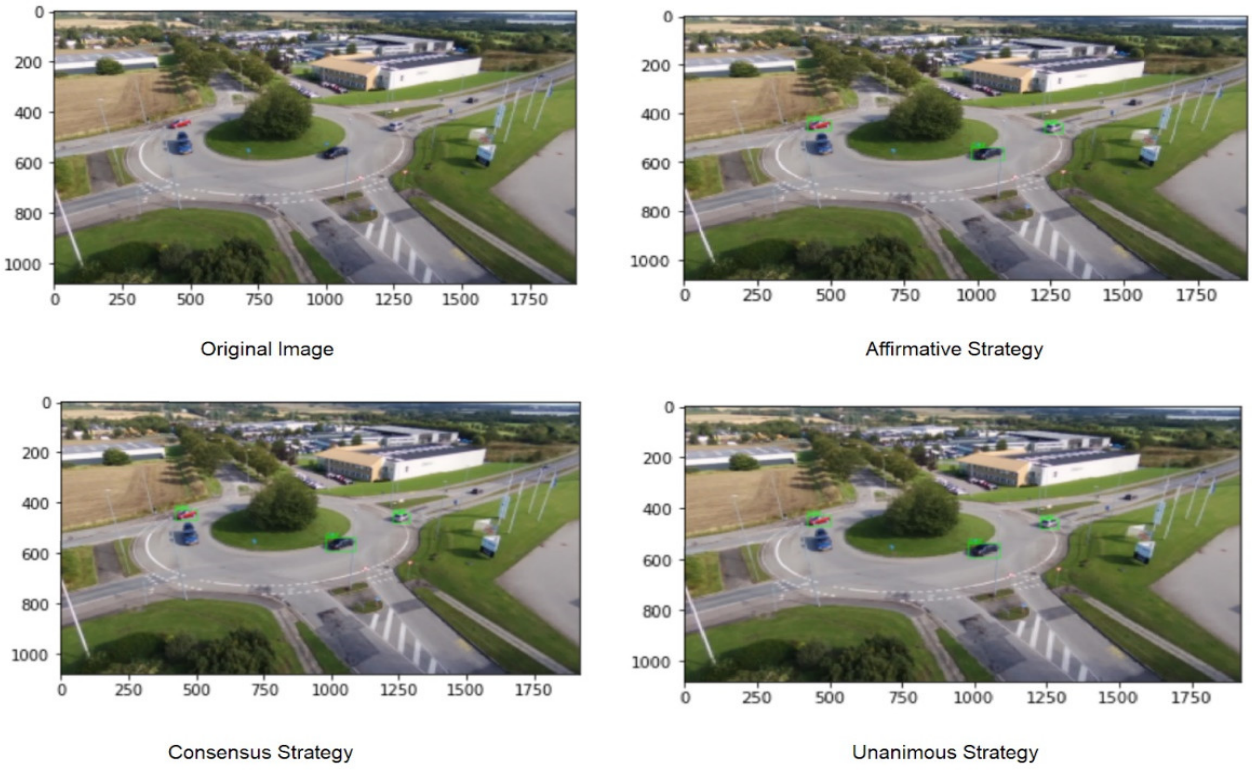

Figure 11. Comparison of RetinaNet detector on all voting strategies and color augmentation.

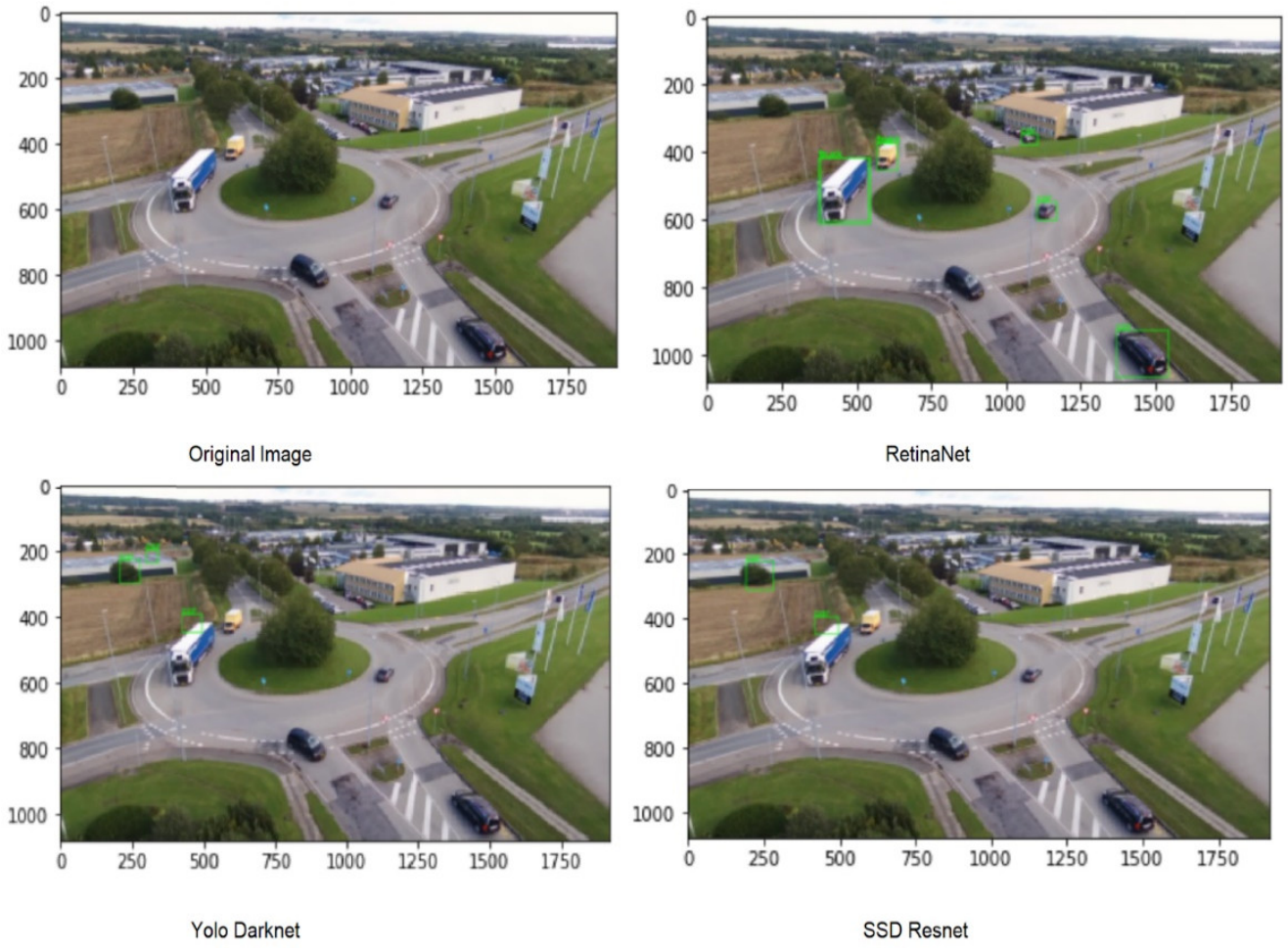

Figure 12. Comparison of detectors using affirmative strategy and color augmentation.

The baseline models have a moderate precision in detecting the object initially before augmentation, and the maximum AP is 21.05% for bus and mAP is 4.41% for RetinaNet. Using the RetinaNet as the baseline model, the color augmentation technique showed the best results with 6.63% mAP, over 2.2% more than blurring, flipping, and rotations augmentation. After selecting the color augmentation techniques, all the ensemble models and voting strategies were tested, among which RetinaNet with the Unanimous strategy provided the best results with 6.63% mAP and 30.95% AP for detecting buses. Overall, the affirmative strategy for all ensemble models worked better observing the average mAP, but the best performer was the unanimous strategy combined with the RetinaNet detector. The models are best at detecting buses, pedestrians, carstrucks and showing poor performance in detecting vans and trailers.

4. Discussion

The Average Recall metric values obtained by our method consisting of ensemble models are comparable with the VisDrone 2019 Challenge Leaderboard results, [113,114] as seen in Table 7. Three of our ensemble models (RetinaNet, RetinaNet+ YOLO, YOLO+SSD+RetinaNet with Affirmative Strategy) that have not been trained on the actual VisDrone data were able to achieve 1.09% greater Average Recall (maxDets = 10) as compared to the RetinaNet model that was trained on the VisDrone dataset. They were also able to attain 0.19% greater Average Recall (maxDets = 1) than the RetinaNet model trained on the VisDrone dataset. For the AU-AIR dataset, we compared the performance of the ensemble algorithm with the original baseline models provided with the original

dataset, as seen in Table 8 [12]. An ensemble shows a 6.63% increase in mAP for AU-AIR dataset over miscellaneous baseline models trained on the COCO dataset in the original survey, as shown in Table 8. The ensemble algorithm (RetinaNet with Unanimous Strategy) performs better than the provided models for all classes, with an 8.56%, 6.61%, 6.89%, 30.95%, 6.63% increase in AP in the Person, Car, Truck, Bus, mAP classes, respectively.

Table 7. Comparison of results with previous SOTAs for the VisDrone dataset. .

Algorithms	Voting Strategy	AR1	AR10
Previous Models			
DPNet-ensemble [1st Rank in VisDrone Challenge, Trained on VisDrone Data] [10,11]	n/a	0.58%	3.69%
RRNet (A.28) [2nd Rank in VisDrone Challenge, Trained on VisDrone Data] [10,11]	n/a	1.02%	8.50%
ACM-OD (A.1) [3rd Rank in VisDrone Challenge, Trained on VisDrone Data] [10,11]	n/a	0.32%	1.48%
RetinaNet [Trained on VisDrone Data] [10,11]	n/a	0.21%	1.21%
Our Models			
RetinaNet	Affirmative	0.40%	2.30%
YOLO + RetinaNet	Affirmative	0.40%	2.29%
YOLO+SSD+RetinaNet	Affirmative	0.07%	1.37%

Table 8. Comparison of results with previous SOTAs and papers for the AU-AIR dataset..

Model	Training Dataset	Voting Strategy	Person	Car	Truck	Motor-bike	Bus	mAP
Previous Models								
YOLO V3-Tiny [12]	COCO	n/a	0.01%	0%	0%	**0**%	0%	n/a
MobileNetV2 -SSDLite [12]	COCO	n/a	0%	0%	0%	0%	0%	n/a
Our Models								
RetinaNet	COCO	Unani-mous	**8.57%**	**6.61%**	**6.89%**	0.00%	**30.95%**	**6.63%**
RetinaNet	COCO	Consen-sus	3.88%	6.21%	5.73%	**1.14%**	20.86%	4.73%
YOLO + RetinaNet	VOC +COCO	Consen-sus	3.45%	3.95%	5.73%	**1.14%**	15.23%	3.69%
RetinaNet + SSD	COCO + VOC	Consen-sus	3.45%	4.62%	5.73%	**1.14%**	17.31%	4.03%

5. Conclusions

We carried out extensive experimentation and analysis of the ensembling methods and augmentation techniques for the drone-based image datasets. The performance of the ensemble techniques on the chosen datasets shows promise for ensemble algorithms and augmentation techniques in UAV object detection. The performance of several test train augmentation techniques indicates the potential for a solution for the deficiency of UAV image datasets. For example, we see the 6% AP performance for detection of bicycles by the RetinaNet detector using color augmentation techniques for the VisDrone dataset and 30.95% AP to detect buses by the RetinaNet detector using color augmentation for the AU-AIR dataset. Furthermore, the voting strategies employed present a causal explanation for

the precision and can be used to render task-specific results. The key insight is that the performance of these ensemble techniques is based on detectors pretrained on non-UAV datasets like COCO [109] and VOC [110], but can still perform OD satisfactorily on the VisDrone and AU-AIR data with the help of data augmentation.

6. Future Scope

We have observed the limitations of this methodology in detecting new objects like the awning tricycle and tricycle absent from their training datasets and hope to improve them in the future. Future work will include testing the algorithm on other drone-based datasets and including more models in the ensembling process. Additionally, the work proposed in this research for drone-based object detection can be employed for generating better quality orthomosaics by multiscale object detection, especially for the objects present around the edges of the orthophoto. With the specific techniques described in this work, such as color augmentation and ensembling, the object detection around the edges of an orthophoto can be improved for various height and lighting conditions.

Author Contributions: Conceptualization, K.K. and R.W.; methodology, K.K., R.W., A.M.; software, A.M.; validation, A.M., R.W., K.K.; formal analysis, A.M., R.W.; investigation, A.M.; resources, A.M.; writing—original draft preparation, A.M., R.W.; writing—review and editing, R.W. and K.K.; visualization, A.M.; supervision, R.W.; project administration, K.K.; funding acquisition, K.K. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by Symbiosis International University (SIU) under the Research Support Fund.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Data is available from public datasets. The links for which are provided in the references section. The source code can be made available upon request from the corresponding author.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Hariharan, B.; Arbel'aez, P.; Girshick, R.; Malik, J. Simultaneous detection and segmentation. In *European Conference on Computer Vision*; Springer: Berlin/Heidelberg, Germany, 2014; pp. 297–312.
2. Hariharan, B.; Arbel'aez, P.; Girshick, R.; Malik, J. Hypercolumns for object segmentation and finegrained localization. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, San Diego, CA, USA, 20–25 June 2015; pp. 447–456.
3. Dai, J.; He, K.; Sun, J. Instance-aware semantic segmentation via multi-task network cascades. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 3150–3158.
4. He, K.; Gkioxari, G.; Doll'ar, P.; Girshick, R. Mask rcnn. In Proceedings of the Computer Vision (ICCV), 2017 IEEE International Conference on, Venice, Italy, 22–29 October 2017; pp. 2980–2988.
5. Karpathy, A.; Fei-Fei, L. Deep visual-semantic alignments for generating image descriptions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, San Diego, CA, USA, 20–25 June 2015; pp. 3128–3137.
6. Xu, K.; Ba, J.; Kiros, R.; Cho, K.; Courville, A.; Salakhudinov, R.; Zemel, R.; Bengio, Y. Show, attend and tell: Neural image caption generation with visual attention. In Proceedings of the International Conference on Machine Learning, Lille, France, 6–11 July 2015; pp. 2048–2057.
7. Wu, Q.; Shen, C.; Wang, P.; Dick, A.; van den Hengel, A. Image captioning and visual question answering based on attributes and external knowledge. *IEEE Trans. Pattern Anal. Mach. Intell.* **2018**, *40*, 1367–1381.
8. Kang, K.; Li, H.; Yan, J.; Zeng, X.; Yang, B.; Xiao, T.; Zhang, C.; Wang, Z.; Wang, R.; Wang, X.; et al. T-cnn: Tubelets with convolutional neural networks for object detection from videos. *IEEE Trans. Circuits Syst. Video Technol.* **2018**, *28*, 2896–2907.
9. Zou, Z.; Shi, Z.; Guo, Y.; Ye, J. Object Detection in 20 Years: A Survey. *arXiv* **2019**, arXiv:1905.05055.
10. Zhu, P.; Wen, L.; Bian, X.; Ling, H.; Hu, Q. Vision meets drones: A challenge. *arXiv* **2018**, arXiv:1804.07437.
11. Zhu, P.; Wen, L.; Du, D.; Bian, X.; Hu, Q.; Ling, H. Vision meets drones: Past, present and future. *arXiv* **2020**, arXiv:2001.06303.
12. Ilker, B.; Kayacan, E. Au-air: A multi-modal unmanned aerial vehicle dataset for low altitude traffic surveillance. In Proceedings of the 2020 IEEE International Conference on Robotics and Automation (ICRA), Paris, France, 31 May–31 August 2020; IEEE: Toulouse, France, 2020.

13. Casado-Garcıa; Angela; Heras, J. Ensemble Methods for Object Detection. In *ECAI 2020*; IOS Press: Amsterdam, The Netherlands, 2020; pp. 2688–2695.
14. Viola, P.; Jones, M. Rapid object detection using a boosted cascade of simple features. In Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Kauai, HI, USA, 8–14 December 2001; IEEE: Toulouse, France, 2001; Volume 1, p. I.
15. Papageorgiou, C.P.; Oren, M.; Poggio, T. A general framework for object detection. In Proceedings of the Sixth International Conference on Computer Vision, Bombay, India, 7 January 1998; IEEE: Toulouse, France, 1998; pp. 555–562.
16. Papageorgiou, C.; Poggio, T. A trainable system for object detection. *Int. J. Comput. Vis.* **2000**, *38*, 15–33.
17. Mohan, A.; Papageorgiou, C.; Poggio, T. Example based object detection in images by components. *IEEE Trans. Pattern Anal. Mach. Intell.* **2001**, *23*, 349–361.
18. Freund, Y.; Schapire, R.; Abe, N. A short introduction to boosting. *J. Jpn. Soc. Artif. Intell.* **1999**, *14*, 1612.
19. Dalal, N.; Triggs, B. Histograms of oriented gradients for human detection. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, San Diego, CA, USA, 20–25 June 2005; IEEE: Toulouse, France, 2005; Volume 1, pp. 886–893.
20. Felzenszwalb, P.; McAllester, D.; Ramanan, D. A discriminatively trained, multi-scale, deformable part model. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Anchorage, AK, USA, 23–28 June 2008; IEEE: Toulouse, France, 2008; pp. 1–8.
21. Felzenszwalb, P.F.; Girshick, R.B.; McAllester, D. Cascade object detection with deformable part models. In Proceedings of the 2010 IEEE Conference on Computer Vision and Pattern Recognition, San Francisco, CA, USA, 13–18 June 2010; IEEE: Toulouse, France, 2010; pp. 2241–2248.
22. Malisiewicz, T.; Gupta, A.; Efros, A.A. Ensemble of exemplar-svms for object detection and beyond. In Proceedings of the 2011 IEEE International Conference on Computer Vision, Barcelona, Spain, 6–13 November 2011; IEEE: Toulouse, France, 2011; pp. 89–96.
23. Dai, J., Qi, H., Xiong, Y., Li, Y., Zhang, G., Hu, H., Wei, Y. Deformable convolutional networks. In Proceedings of the IEEE International Conference on Computer Vision, Honolulu, HI, USA, 21–26 July 2017; pp. 764–773.
24. Felzenszwalb, P.F.; Girshick, R.B.; McAllester, D.; Ramanan, D. Object detection with discriminatively trained part-based models. *IEEE Trans. Pattern Anal. Mach. Intell.* **2010**, *32*, 1627–1645.
25. Girshick, R.B.; Felzenszwalb, P.F.; Mcallester, D.A. Object detection with grammar models. *Adv. Neural Inf. Process. Syst.* **2011**, *24*, 442–450.
26. LeCun, Y.; Bengio, Y.; Hinton, G. Deep learning. *Nature* **2015**, *521*, 436–444.
27. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. Imagenet classification with deep convolutional neural networks. *Adv. Neural Inf. Process. Syst.* **2012**, *25*, 1097–1105.
28. Girshick, R.; Donahue, J.; Darrell, T.; Malik, J. Rich feature hierarchies for accurate object detection and semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014; pp. 580–587.
29. Girshick, R.; Donahue, J.; Darrell, T.; Malik, J. Region based convolutional networks for accurate object detection and segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2016**, *38*, 142–158.
30. Jiao, L.; Zhang, F.; Liu, F.; Yang, S.; Li, L.; Feng, Z.; Qu, R. A survey of deep learning-based object detection. *IEEE Access* **2019**, *7*, 128837–128868.
31. Liu, L.; Özsu, M.T. Mean Average Precision. In *Encyclopedia of Database Systems*; Springer: Boston, MA, USA, 2009; doi:10.1007/978-0-387-39940-9_3032.
32. Girshick, R.B.; Felzenszwalb, P.F.; McAllester, D. Discriminatively Trained Deformable Part Models, Release 5. Available online: http://people.cs.uchicago.edu/ rbg/latentrelease5/ (accessed on 5th May 2021).
33. He, K.; Zhang, X.; Ren, S.; Sun, J. Spatial pyramid pooling in deep convolutional networks for visual recog-nition. In *European Conference on Computer Vision*; Springer: Berlin/Heidelberg, Germany, 2014; pp. 346–361.
34. Girshick, R. Fast r-cnn. In Proceedings of the IEEE international Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 1440–1448.
35. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster r-cnn: Towards real-time object detection with region proposal networks. *Adv. Neural Inf. Process. Syst.* **2015**, *28*, 91–99.
36. Dai, J.; Li, Y.; He, K.; Sun, J. R-fcn: Object detection via region-based fully convolutional networks. In *Advances in Neural Information Processing Systems*; Publisher: 2016; pp. 379–387.
37. Li, Z.; Peng, C.; Yu, G.; Zhang, X.; Deng, Y.; Sun, J. Light-head r-cnn: In defense of two-stage object detector. *arXiv* **2017**, arXiv:1711.07264.
38. Lin, T.-Y.; Doll'ar, P.; Girshick, R.B.; He, K.; Hariharan, B.; Belongie, S.J. Feature pyramid networks for object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 2117–2125.
39. Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, A. You only look once: Unified, real-time object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 779–788.
40. Redmon, J.; Farhadi, A. YOLO9000: Better, faster, stronger. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, USA, 21-26 July 2017.

41. Redmon, J.; Farhadi, A. YOLOv3: An incremental improvement. *arXiv* **2018**, arXiv:1804.02767.
42. Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.-Y.; Berg, A.C. Ssd: Single shot multibox detector. In *European Conference on Computer Vision*; Springer: Berlin/Heidelberg, Germany, 2016; pp. 21–37.
43. Lin, T.-Y.; Goyal, P.; Girshick, R.; He, K.; Doll'ar, P. Focal loss for dense object detection. *IEEE Trans. Pattern Anal. Mach. Intell.* **2018**, *42*.
44. Doll'ar, P.; Tu, Z.; Perona, P.; Belongie, S. *Integral Channel Features*; Proceedings of the British Machine Vision Conference , BMVA Press : 2009.
45. Maji, S.; Berg, A.C.; Malik, J. Classification using intersection kernel support vector machines is efficient. In Proceedings of the 2008 IEEE Conference on Computer Vision and Pattern Recognition, Anchorage, AK, USA, 23–28 June 2008; IEEE: Toulouse, France, 2008; pp. 1–8.
46. Zhu, Q.; Yeh, M.-C.; Cheng, K.-T.; Avidan, S. Fast human detection using a cascade of histograms of oriented gradients. In Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, New York, NY, USA, 17–22 June 2006; IEEE: Toulouse, France, 2006; Volume 2, pp. 1491–1498.
47. Zhang, L.; Lin, L.; Liang, X.; He, K. Is faster rcnn doing well for pedestrian detection? In *European Conference on Computer Vision*; Springer: Berlin/Heidelberg, Germany, 2016; pp. 443–457.
48. Dollar, P.; Wojek, C.; Schiele, B.; Perona, P. Pedestrian detection: An evaluation of the state of the art. *IEEE Trans. Pattern Anal. Mach. Intell.* **2012**, *34*, 743–761.
49. Enzweiler, M.; Gavrila, D.M. Monocular pedestrian detection: Survey and experiments. *IEEE Trans. Pattern Anal. Mach. Intell.* **2008**, *31*, 2179–2195.
50. Geronimo, D.; Lopez, A.M.; Sappa, A.D.; Graf, T. Survey of pedestrian detection for advanced driver assistance systems. *IEEE Trans. Pattern Anal. Mach. Intell.* **2010**, *32*, 1239–1258.
51. Dollar, P.; Wojek, C.; Schiele, B.; Perona, P. Pedestrian detection: A benchmark. In Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition, Miami, FL, USA, 20–25 June 2009; IEEE: Toulouse, France, 2009.
52. Benenson, R.; Omran, M.; Hosang, J.; Schiele, B. Ten years of pedestrian detection, what have we learned? In *European Conference on Computer Vision*; Springer: Berlin/Heidelberg, Germany, 2014; pp. 613–627.
53. Zhang, S.; Benenson, R.; Omran, M.; Hosang, J.; Schiele, B. How far are we from solving pedestrian detection? In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 1259–1267.
54. Zhang, S.; Benenson, R.; Omran, M.; Hosang, J.; Schiele, B. Towards reaching human performance in pedestrian detection. *IEEE Trans. Pattern Anal. Mach. Intell.* **2018**, *40*, 973–986.
55. Cao, J.; Pang, Y.; Li, X. Learning multilayer channel features for pedestrian detection. *IEEE Trans. Image Process.* **2017**, *26*, 3210–3220.
56. Mao, J.; Xiao, T.; Jiang, Y.; Cao, Z. What can help pedestrian detection? In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; IEEE: Toulouse, France, 2017; pp. 6034–6043.
57. Hu, Q.; Wang, P.; Shen, C.; van den Hengel, A.; Porikli, F. Pushing the limits of deep cnns for pedestrian detection. *IEEE Trans. Circuits Syst. Video Technol.* **2018**, *28*, 1358–1368.
58. Tian, Y.; Luo, P.; Wang, X.; Tang, X. Pedestrian detection aided by deep learning semantic tasks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 5079–5087.
59. Xu, D.; Ouyang, W.; Ricci, E.; Wang, X.; Sebe, N. Learning cross-modal deep representations for robust pe-destrian detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017.
60. Wang, X.; Xiao, T.; Jiang, Y.; Shao, S.; Sun, J.; Shen, C. Repulsion loss: Detecting pedestrians in a crowd. *arXiv* **2017**, arXiv:1711.07752.
61. Ouyang, W.; Zhou, H.; Li, H.; Li, Q.; Yan, J.; Wang, X. Jointly learning deep features, deformable parts, occlusion and classification for pedestrian detection. *IEEE Trans. Pattern Anal. Mach. Intell.* **2018**, *40*, 1874–1887.
62. Zhang, S.; Yang, J.; Schiele, B. Occluded pedestrian detection through guided attention in cnns. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 6995–7003.
63. Rowley, H.A.; Baluja, S.; Kanade, T. Human face detection in visual scenes. In *Advances in Neural Information Processing Systems*; Department of Computer Science, Carnegie-Mellon University: 1996; pp. 875–881.
64. Yang, G.; Huang, T.S. Human face detection in a complex background. *Pattern Recognit.* **1994**, *27*, 53–63.
65. Craw, I.; Tock, D.; Bennett, A. Finding face features. In *European Conference on Computer Vision*; Springer: Berlin/Heidelberg, Germany, 1992; pp. 92–96.
66. Turk, M.; Pentland, A. Eigenfaces for recognition. *J. Cogn. Neurosci.* **1991**, *3*, 71–86.
67. Pentl, A.; Moghaddam, B.; Starner, T. *View Based and Modular Eigenspaces for Face Recognition*; Proceedings of IEEE Conference on Computer Vision and Pattern Recognition,: 1994; pp. 84-91
68. Rowley, H.A.; Baluja, S.; Kanade, T. Neural network-based face detection. *IEEE Trans. Pattern Anal. Mach. Intell.* **1998**, *20*, 23–38.
69. Osuna, E.; Freund, R.; Girosit, F. Training support vector machines: An application to face detection. In Proceedings of the 1997 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, San Juan, PR, USA, 17–19 June 1997; IEEE: Toulouse, France, 1997; pp. 130–136.
70. Wu, Y.; Natarajan, P. Self-organized text detection with minimal post-processing via border learning. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 5000–5009.

71. Z. Zhu, Yingying, Cong Yao, and Xiang Bai. "Scene text detection and recognition: Recent advances and future trends." Frontiers of Computer Science 10.1 (2016): 19-36
72. Liu, X. A camera phone-based currency reader for the visually impaired. In Proceedings of the 10th International ACM SIGACCESS Conference on Computers and Accessibility, Halifax, NS, Canada, 13–15 October 2008; ACM: New York, NY, USA, 2008; pp. 305–306.
73. Ezaki, N.; Kiyota, K.; Minh, B.T.; Bulacu, M.; Schomaker, L. Improved text-detection methods for a cam-era-based text reading system for blind persons. In Proceedings of the Eighth International Conference on Document Analysis and Recognition, Seoul, Korea, 31 August–1 September 2005; IEEE: Toulouse, France, 2005; pp. 257–261.
74. Sermanet, P.; Chintala, S.; LeCun, Y. Convolutional neural networks applied to house numbers digit classification. In Proceedings of the 21st International Conference on Pattern Recognition (ICPR), Tsukuba, Japan, 11–15 November 2012; IEEE: Toulouse, France, 2012; pp. 3288–3291.
75. Wojna, Z.; Gorban, A.; Lee, D.-S.; Murphy, K.; Yu, Q.; Li, Y.; Ibarz, J. Attention-based extraction of structured information from street view imagery. *arXiv* **2017**, arXiv:1704.03549.
76. Zhu, Y.; Yao, C.; Bai, X. Scene text detection andrecognition: Recent advances and future trends. *Front. Comput. Sci.* **2016**, *10*, 19–36.
77. Ye, Q.; Doermann, D. Text detection and recognition in imagery: A survey. *IEEE Trans. Pattern Anal. Mach. Intell.* **2015**, *37*, 1480–1500.
78. Møgelmose, A.; Trivedi, M.M.; Moeslund, T.B. Vision-based traffic sign detection and analysis for intelligent driver assistance systems: Perspectives and survey. *IEEE Trans. Intell. Transp. Syst.* **2012**, *13*, 1484–1497.
79. Paulo, C.F.; Correia, P.L. Automatic detection and classification of traffic signs. In Proceedings of the Eighth International Workshop on Image Analysis for Multimedia Interactive Services, WIAMIS'07; IEEE: Toulouse, France, 2007; p. 11.
80. Omachi, M.; Omachi, S. Traffic light detection with color and edge information. In Proceedings of the 2009 2nd IEEE International Conference on Computer Science and Information Technology, Beijing, China, 8–11 August 2009; IEEE: Toulouse, France, 2009; pp. 284–287.
81. Xie, Y.; Liu, L.-f.; Li, C.-h.; Qu, Y.-y. Unifying visual saliency with hog feature learning for traffic sign detection. In Proceedings of the 2009 IEEE Intelligent Vehicles Symposium, Xi'an, Shaanxi, China, 3–5 June 2009; IEEE: Toulouse, France, 2009; pp. 24–29.
82. de Charette, R.; Nashashibi, F. Real time visual traffic lights recognition based on spotlight detection and adaptive traffic lights templates. In Proceedings of the Intelligent Vehicles Symposium, Xi'an, China, 3–5 June 2009; IEEE: Toulouse, France, 2009; pp. 358–363.
83. Houben, S. A single target voting scheme for traffic sign detection. In Proceedings of the Intelligent Vehicles Symposium (IV), 2011, Baden-Baden, Germany, 5–9 June 2011; IEEE: Toulouse, France, 2011; pp. 124–129.
84. Soetedjo, A.; Yamada, K. Fast and robust traffic sign detection. In Proceedings of the 2005 IEEE International Conference on Systems, Man and Cybernetics; IEEE: Toulouse, France, 2005; Volume 2, pp. 1341–1346.
85. Cheng, G.; Han, J. A survey on object detection in optical remote sensing images. *ISPRS J. Photogramm. Remote Sens.* **2016**, *117*, 11–28.
86. Zhang, L.; Zhang, L.; Du, B. Deep learning for remote sensing data: A technical tutorial on the state of the art. *IEEE Geosci. Remote Sens. Mag.* **2016**, *4*, 22–40.
87. Proia, N.; Pag'e, V. Characterization of a Bayesian ship detection method in optical satellite images. *IEEE Geosci. Remote Sens. Lett.* **2010**, *7*, 226–230.
88. Zhu, C.; Zhou, H.; Wang, R.; Guo, J. A novel hierarchical method of ship detection from spaceborne optical image based on shape and texture features. *IEEE Trans. Geosci. Remote Sens.* **2010**, *48*, 3446–3456.
89. Pastor, E.; Lopez, J.; Royo, P. A hardware/software architecture for UAV payload and mission control. In Proceedings of the 2006 IEEE/AIAA 25TH Digital Avionics Systems Conference; Portland, Oregon, 15–18 October 2006; IEEE: Toulouse, France, 2006; pp. 1–8.
90. Zeeshan, K.; Rehmani, M.H. Amateur drone monitoring: State-of-the-art architectures, key enabling technologies, and future research directions. *IEEE Wirel. Commun.* **2018**, *25*, 150–159.
91. Tisdale, J.; Ryan, A.; Zennaro, M.; Xiao, X.; Caveney, D.; Rathinam, S.; Hedrick, J.K.; Sengupta, R. The software architecture of the Berkeley UAV platform. In Proceedings of the 2006 IEEE Conference on Computer Aided Control System Design, Munich, Germany, 4–6 October 2006; In Proceedings of the 2006 IEEE International Conference on Control Applications, Munich, Germany, 4–6 October 2006; In Proceedings of the 2006 IEEE International Symposium on Intelligent Control, Munich, Germany, 4–6 October 2006; IEEE: Toulouse, France, 2006.
92. Mészarós, J. Aerial surveying UAV based on open-source hardware and software. *Int. Arch. Photogramm. Remote Sens. Spat. Inf. Sci.* **2011**, *37*, 555.
93. Rumba, R.; Nikitenko, A. Decentralized Air Traffic Management System for Unmanned Aerial Vehicles. U.S. Patent 9,997,080 B1, 12 June 2018.
94. Collins, T.J. Automated Unmanned Air Traffic Control System. U.S. Patent 2016/0196750 A1, 7 July 2016.
95. Jewett, S.P. Agent-Based Airfield Conflict Resolution. U.S. Patent US9153138B1, 6 October 2015.
96. Finn, R.L.; Wright, D. Privacy, data protection and ethics for civil drone practice: A survey of industry, regulators and civil society organisations. *Comput. Law Secur. Rev.* **2016**, *32*, 577–586.
97. Custers, B. *Future of Drone Use*; TMC Asser Press: The Hague, The Netherlands, 2016.

98. Rocci, L.; So, A. A technoethical review of commercial drone use in the context of governance, ethics, and privacy. *Technol. Soc.* **2016**, *46*, 109–119.
99. Doggett, S. What Is an Orthomosaic? Orthomosaic Maps & Orthophotos Explained. Dronegenuity, 23 November 2020. Available online: www.dronegenuity.com/orthomosaic-maps-explained (accessed on 7 July 2021).
100. Nordstrom, S. What Is an Orthomosaic Map and How Does Mapping Benefit My Property? Dronebase. Available online: blog.dronebase.com/what-is-an-orthomosaic-map-and-how-does-mapping-benefit-my-property (accessed on 7 July 2021).
101. Onishi, M.; Ise, T. Explainable identification and mapping of trees using UAV RGB image and deep learning. *Sci. Rep.* **2021**, *11*, 1–15.
102. Osco, L.P.; Junior, J.M.; Ramos, A.P.M.; Jorge, L.A. D.C.; Fatholahi, S.N.; Silva, J.D.A.; Li, J. A review on deep learning in UAV remote sensing. *arXiv* **2021**, arXiv:2101.10861.
103. Okafor, E.; Smit, R.; Schomaker, L.; Wiering, M. Operational data augmentation in classifying single aerial images of animals. In Proceedings of the 2017 IEEE International Conference on Innovations in Intelligent Systems and Applications (INISTA), Gdynia, Poland, 3–5 July 2017; IEEE: New York, NY, USA, 2017.
104. Castro, W.; Junior, J.M.; Polidoro, C.; Osco, L.P.; Gonçalves, W.; Rodrigues, L.; Santos, M.; Jank, L.; Barrios, S.; Valle, C.; et al. Deep Learning Applied to Phenotyping of Biomass in Forages with UAV-Based RGB Imagery. *Sensors* **2020**, *20*, 4802.
105. Kellenberger, B.; Volpi, M.; Tuia, D. Fast animal detection in UAV images using convolutional neural networks. In Proceedings of the 2017 IEEE International Geoscience and Remote Sensing Symposium (IGARSS), Fort Worth, TX, USA, 23–28 July 2017; IEEE: New York, NY, USA, 2017.
106. Sadykova, D.; Pernebayeva, D.; Bagheri, M.; James, A. IN-YOLO: Real-Time Detection of Outdoor High Voltage Insulators Using UAV Imaging. *IEEE Trans. Power Deliv.* **2020**, *35*, 1599–1601.
107. Tang, T.; Deng, Z.; Zhou, S.; Lei, L.; Zou, H. Fast vehicle detection in UAV images. In Proceedings of the 2017 IEEE International Workshop on Remote Sensing with Intelligent Processing (RSIP), Shanghai, China, 18–21 May 2017.
108. Song, C.; Xu, W.; Wang, Z.; Yu, S.; Zeng, P.; Ju, Z. Analysis on the Impact of Data Augmentation on Target Recognition for UAV-Based Transmission Line Inspection. *Complexity* **2020**, *2020*, 3107450, doi:10.1155/2020/3107450.
109. Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; Zitnick, C.L. Microsoft coco: Common objects in context. In Proceedings of the European Conference on Computer Vision, Zurich, Switzerland, 6–12 September 2014; Springer: Cham, Switzerland, 2014.
110. Everingham, M.; Van Gool, L.; Williams, C.K.I.; Winn, J.; Zisserman, A. The Pascal Visual Object Classes (VOC) Challenge. *Int. J. Comput. Vis.* **2010**, *88*, 303–338.
111. Jian, G.; Gould, S. Deep CNN ensemble with data augmentation for object detection. *arXiv* **2015**, arXiv:1506.07224.
112. Xu, J.; Wang, W.; Wang, H.; Guo, J. Multi-model ensemble with rich spatial information for object detection. *Pattern Recognit.* **2020**, *99*, 107098.
113. Reddy, D.R.; Du, D.; Zhu, P.; Wen, L.; Bian, X.; Ling, H.; Hu, Q.; Peng, T.; Zheng, J.; Wang, X.; et al. VisDrone-DET2019: The Vision Meets Drone Object Detection in Image Challenge Results. In Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops, 2019.
114. Zhu, P.; Wen, L.; Du, D.; Bian, X.; Ling, H.; Hu, Q.; Nie, Q.; Cheng, H.; Liu, C.; Liu, X.; et al. Visdrone-det2018: The vision meets drone object detection in image challenge results. In Proceedings of the European Conference on Computer Vision (ECCV) Workshops, Munich, Germany, 8–14 September 2018.