CVF

Unmanned Aerial Vehicle Visual Detection and Tracking using Deep Neural Networks: A Performance Benchmark

Brian K. S. Isaac-Medina
Department of Computer Science
Durham University
Durham, UK
brian.k.isaac-medina@durham.ac.uk

Matt Poyser
Department of Computer Science
Durham University
Durham, UK
matthew.poyser@durham.ac.uk

Daniel Organisciak
Department of Computer and Information Sciences
Northumbria University
Newcastle upon Tyne, UK
daniel.organisciak@northumbria.ac.uk

Chris G. Willcocks
Department of Computer Science
Durham University
Durham, UK
christopher.g.willcocks@durham.ac.uk

Toby P. Breckon
Department of Computer Science
Durham University
Durham, UK
toby.breckon@durham.ac.uk

Hubert P. H. Shum
Department of Computer Science
Durham University
Durham, UK
hubert.shum@durham.ac.uk

Abstract

Unmanned Aerial Vehicles (UAV) can pose a major risk for aviation safety, due to both negligent and malicious use. For this reason, the automated detection and tracking of UAV is a fundamental task in aerial security systems. Common technologies for UAV detection include visible-band and thermal infrared imaging, radio frequency and radar. Recent advances in deep neural networks (DNNs) for image-based object detection open the possibility to use visual information for this detection and tracking task. Furthermore, these detection architectures can be implemented as backbones for visual tracking systems, thereby enabling persistent tracking of UAV incursions. To date, no comprehensive performance benchmark exists that applies DNNs to visible-band imagery for UAV detection and tracking. To this end, three datasets with varied environmental conditions for UAV detection and tracking, comprising a total of 241 videos (331,486 images), are assessed using four detection architectures and three tracking frameworks. The best performing detector architecture obtains an mAP of 98.6% and the best performing tracking framework obtains a MOTA of 98.7%. Cross-modality evaluation is carried out between visible and infrared spectrums, achieving a maximal 82.8% mAP on visible images when
training in the infrared modality. These results provide the first public multi-approach benchmark for state-of-the-art deep learning-based methods and give insight into which detection and tracking architectures are effective in the UAV domain.

1. Introduction

Over the past decade, Unmanned Aerial Vehicles (UAV) have become more accessible for both professional and casual use, resulting in an increased risk of UAV-caused disturbances. Whether a pilot has malicious intent or is negligent, UAVs can impact aviation safety, travel into restricted airspace, or capture sensitive data. UAV disturbances can also have serious security and economic consequences [17]. These disturbances are more likely to occur as UAV usage grows.

As such, a major focus of counter-UAV systems is to detect and track UAVs that are small and fast-moving. Traditional technologies used in counter-UAV systems include radio frequency [50], radar [25] and acoustic sensors [51]. These sensors, however, perform sub-optimally to locate objects at long distances [46, 22, 23] or in settings with significant sources of noise such as airports, where UAV

threats are prevalent [35, 42, 6]. While the use of standard cameras is attractive due to the visual cues they provide to security personnel, a key challenge is the relatively small image size of UAVs, which impacts the detection performance. With the recent advances in deep neural networks, image-based models have become widely used for video surveillance [27]. Furthermore, these models can be used in infrared (IR) imagery, overcoming common problems when using optical images, such as lighting or weather conditions, although the data is more limited. To this end, we focus our research on vision-based systems.

In the last decade, deep convolutional neural networks (DCNN) have become the base model of state-of-the-art architectures for visual tasks such as object detection and tracking [54, 44]. These learning-based methods rely on extracting feature maps from the input images to provide a probability distribution over a set of categories or to regress a set of real-valued variables. In the context of object detection, DCNNs are used to regress a parametric model of bounding boxes enclosing the objects of interest [15, 33, 37], while more recent models use encoder-decoder architectures for the same task [5]. When operating within the tracking domain, objects must be assigned an additional identification feature. Subsequently, the tracking task involves further identification of the same instance at different locations in the temporal dimension. Most object tracking architectures are based on *tracking-by-detection*, which consists of associating the predicted objects from standard detection architectures across the temporal dimension (i.e., consecutive video frames) [10].

The different settings in which counter-UAV systems may be encountered imply that highly robust tracking and detection models are required. According to Rozantsev *et al*. (2015) [41], there are three key challenges for such models. Firstly, given the speed of UAVs, they must be detected at a long distance, when they appear very small within a camera image. This problem is compounded by the fact that secondly, the background scene is often highly complex and diverse; landscapes can vary dramatically and also be highly cluttered. Thirdly, UAV are capable of relatively complex movement, and tracking them often involves cameras mounted on a high-speed system themselves. Therefore, the evaluation of detection and tracking models over different datasets, demonstrating their robustness over such challenges, is an imperative task.

Although detecting and tracking UAV has been previously studied [47, 36], to the best of our knowledge, there has been no wide-ranging performance benchmark study evaluated across multiple datasets for these tasks. Concurrent to this work, Jiang *et al*. [24] introduce the Anti-UAV dataset and benchmark tracking performance. However, they do not compare different detectors nor datasets, which can greatly impact tracking performance. As a result,

it is difficult to evaluate the effectiveness of the proposed frameworks.

Given this motivation, this work compares the performance of several vision-based detection and tracking models under three different datasets, subject to varying setups. Our main contributions are as follow:

– a benchmark performance study is presented across leading state of the art object detection (Faster-RCNN [38], YOLOv3 [37], SSD [33], and DETR [5]) and tracking architectures (SORT [4], DeepSORT [48], and Tracktor [3]).

– state-of-the-art detection performance (0.986 mAP) is achieved for visual detection and tracking within the counter-UAV domain, compared to previous leading performance (0.952 mAP [18]) .

– collation and assessment of three UAV datasets (MAV-VID [40], Drone-vs-Bird [11], and Anti-UAV [24]), yielding a plethora of varied environment and environmental condition detection challenges. Images are captured by both ground and UAV-mounted cameras, under highly dynamic scenes, at short and long distances and by both optical and IR cameras.

– novel evaluation of cross-modality training and testing for UAV. Training on IR data is demonstrated to be sufficient for inference on RGB imagery (0.828 mAP).

Our benchmark toolkit is open for further research and development. It can be downloaded on the project website: https://github.com/KostadinovShalon/UAVDetectionTrackingBenchmark

2. UAV Detection and Tracking Architectures

Overall, prior work within the UAV domain is limited in both scope and complexity whereas multiple comprehensive surveys for detection and tracking have already been published more broadly [54] [44]. We provide an overview of popular architectures and literature that addresses generic detection and tracking challenges.

2.1. Object Detection

Object detection is one of the most fundamental and well-studied tasks in computer vision. Objects from desired classes must be located with a bounding box, which should enclose the object as tightly as possible. Deep learning has strongly influenced object detectors because of the variety of expressive features that deep models can learn, leading towards a focus on generic object detectors that perform well on a number of distinct classes and datasets, rather than designing detectors for specific objects [32].

Deep-learning architectures are mostly composed of two categories: two-stage detectors and one-stage detectors.

The two-stage detectors, such as RCNN [16], Faster RCNN [38] and RFCN [13], consist of one stage to compose a set of candidate regions of interest that could contain desired objects and a second stage to classify the proposed regions and regress its bounding box parameters. These detectors typically outperform one-stage detection architectures in terms of detection accuracy at the expense of computational efficiency.

One-stage detectors such as OverFeat [43], YOLO (You Only Look Once) [37], SSD (Single Shot Detector) [33] and CornerNet [28] drop the region proposal stage, favouring the use of global image features to determine bounding box locations. In particular, YOLO and SSD split the image into a grid and regress the parameters of the bounding box with respect to a set of anchor boxes on each grid cell. CornerNet instead focuses on detecting pairs of keypoints at the top-left and bottom-right of each object. DETR (Detection Transformer) [5] is a more recent approach that uses a transformer encoder-decoder architecture with a bipartite matching loss to alternatively propose bounding boxes.

Despite the considerable progress on object detection, a major challenge is that UAV often appear at dramatically different sizes in the image. Small objects are composed of fewer pixels and subsequently the deepest layers of the convolutional backbone within such detection architectures can struggle to extract detailed object information. Furthermore, successive max-pooling layers may suppress such detection responses for small scene objects within the penultimate stages of the architecture. To combat this, feature pyramid networks [30] fuse features from several layers to perform detection at multiple scales. An additional problem is differentiating UAV from other small, similarly-coloured objects that appear in the sky, such as birds. This is epitomised by the Drones vs Birds dataset [11], where the winner team implemented a two-stage detector having temporal-aware input channels and a standard tracking algorithm to filter out false positives [12].

UAV can also be difficult to detect due to poor visibility in adverse weather conditions, poor lighting, low-quality cameras (on other UAV) and buildings with a similar colour profile. Yang *et al*. [53] survey methods to improve visibility for detection, including de-hazing, de-raining and low-light enhancement. These methods improve detection results on well-researched objects such as faces, pedestrians and vehicles, but it is unclear whether they transfer to UAV.

2.2. Tracking

Single-object tracking (SOT) introduces the key challenge of distinguishing objects from the background, for which relatively established traditional algorithms can perform very well [21, 20, 1, 29]. Multi-object tracking (MOT), however, within which lies the scope of this paper, introduces further challenges. In particular, we must be able to track multiple objects that occlude each other over a short or long time frame. To this end, algorithms commonly employ an appearance descriptor [2, 52, 49] or exploit motion information [8, 9] to differentiate between nearby moving objects.

Geiger *et al*. [14] employ the Hungarian Algorithm [26] for object tracking. They are able to precisely track vehicles within urban scenes by associating predicted locations from motion, with detected locations and object geometry from appearance cues. Bewley *et al*. [4] demonstrate that good performance can be achieved (but not limited to vehicle tracking) without the need of expensive geometric calculations. Additionally, they demonstrate that their online framework is quicker than other methods, without sacrificing accuracy. Wojke *et al*. (DeepSORT) [48] further introduce a deep auxiliary network within the SORT framework to better capture an appearance descriptor of the objects. This re-identification network, robust to object pose and camera viewpoint changes, is used within the association step to dramatically reduce the number of times an object is identified as new rather than associated with one that has been previously identified. Bergmann *et al*. (Tracktor) [3] indicate that detectors are in fact enough for a tracking system, and that there is no need for explicit motion prediction and association. They show that motion models can be sufficiently captured within the bounding box regression of detectors, and achieve state of the art accuracy and frame throughput.

In almost all cases however, the detection model architecture exists as a backbone within the tracking framework. Therefore, evaluating tracking performance is a natural step in determining the general performance of a detector, which we include in this benchmark.

3. Experimental Setup

3.1. UAV Datasets

In order to evaluate the performance of DCNNs based architectures for object detection and consequently tracking, three datasets are assessed in this work. Each dataset comprises a set of varying length videos subject to different constraints and setups. These datasets have only one class, labelled as "drone". Figure 1 shows the datasets statistics.

Multirotor Aerial Vehicle VID (MAV-VID) [40]. This dataset consists on videos at different setups of single UAV. It contains videos captured from other drones, ground based surveillance cameras and handheld mobile devices. It comprises 53 videos (29,500 images) for training and 11 videos (10,732 images) for validation, with an average object size of 136×77 pixels (0.66% of the image size). Although this dataset can be considered as the most straightforward among the three, the differences between the videos make this dataset a good benchmark to evaluate the ability of de-
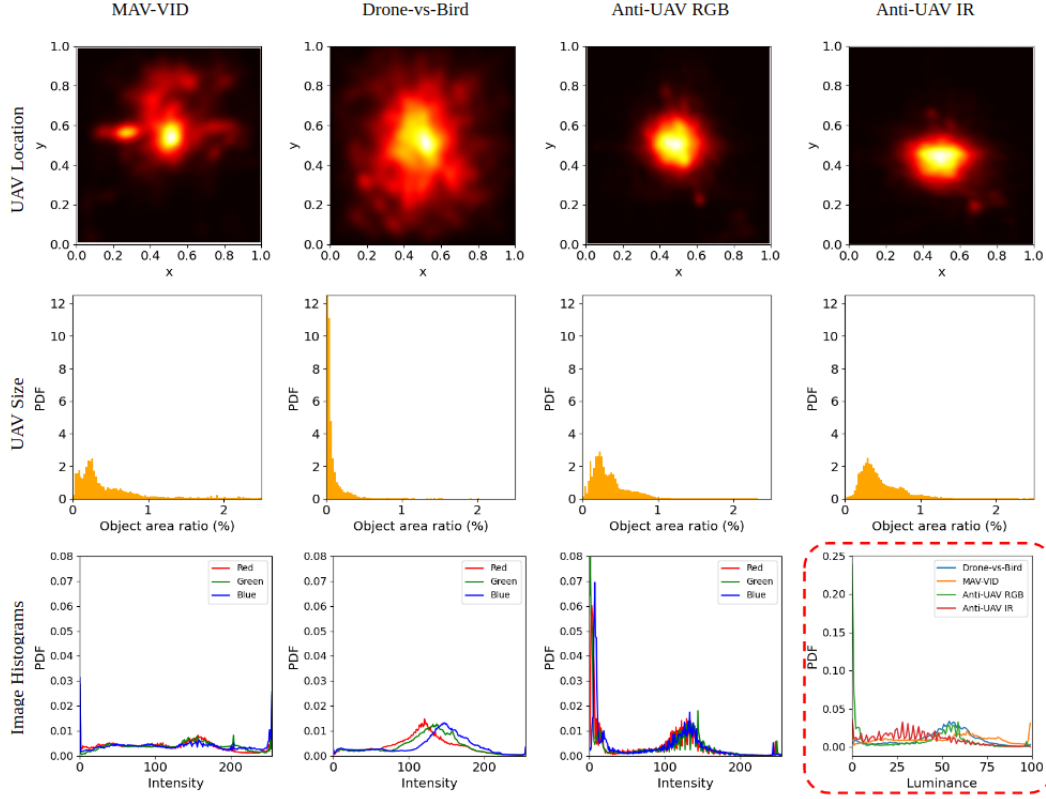
Figure 1: Dataset statistics. UAV location refers to the bounding box centre of the objects in the image. UAV size is the ratio of the object size with the image size. The colour histograms are generated using all images from the dataset. The image in the red dashed square shows the luminance component from the L*a*b* colour space of each dataset.

tectors and trackers for generalisation. As seen in Figure 1, UAVs usually move across the x axis and are recorded from the bottom.

Drone-vs-Bird Detection Challenge [11]. As part of the International Workshop on Small-Drone Surveillance, Detection and Counteraction techniques of IEEE AVSS 2020, the main goal of this challenge is to reduce the high false positive rates that vision-based methods usually suffer. This dataset comprises videos of UAV captured at long distances and often surrounded by small objects, such as birds, with an average UAV in image size of 34×23 pixels (0.10% of the image size). This dataset is formed by 77 videos (104,760 images), which we divided in 61 videos (85,904 images) for training and 16 videos (18,856 images) for validation. Figure 1 reveals that UAVs in this dataset are smaller and appear with more variation across the image plane.

Anti-UAV [24]. This multi-modal dataset comprises 100 fully-annotated RGB and IR unaligned videos, giving a total of 186,494 images including both modalities. Anti-UAV dataset is intended to provide a real-case benchmark for evaluating object tracker algorithms in the context of UAV.

It contains recordings of 6 UAV models flying at different lightning and background conditions. The average object size for the RGB modality is 125×59 pixels (0.40% of the image size) and 52×29 pixels (0.50% of the image size). We divided this dataset in 60 videos (149,478 images) for training and 40 videos (37,016) for validation. In our experiments, we refer to *Anti-UAV Full* to the dataset containing all images regarding their modality while *Anti-UAV RGB* and *Anti-UAV IR* refer to each separate modality. It is apparent from Figure 1 that UAVs are slightly bigger relative to the image size. Similar intensities of the three colour channels indicate the poor contrast in the RGB videos. Furthermore, the shape of the histogram indicates a possible pre-processing carried out during the acquisition process.

The plot with the red dashed square in Figure 1 shows the luminance (*L* from the L*a*b* colour space) of each dataset. Anti-UAV IR has the lowest luminance across the datasets, which is expected from an IR dataset. On the other hand, MAV-VID shows a higher luminance, indicating it is brighter than the other datasets.
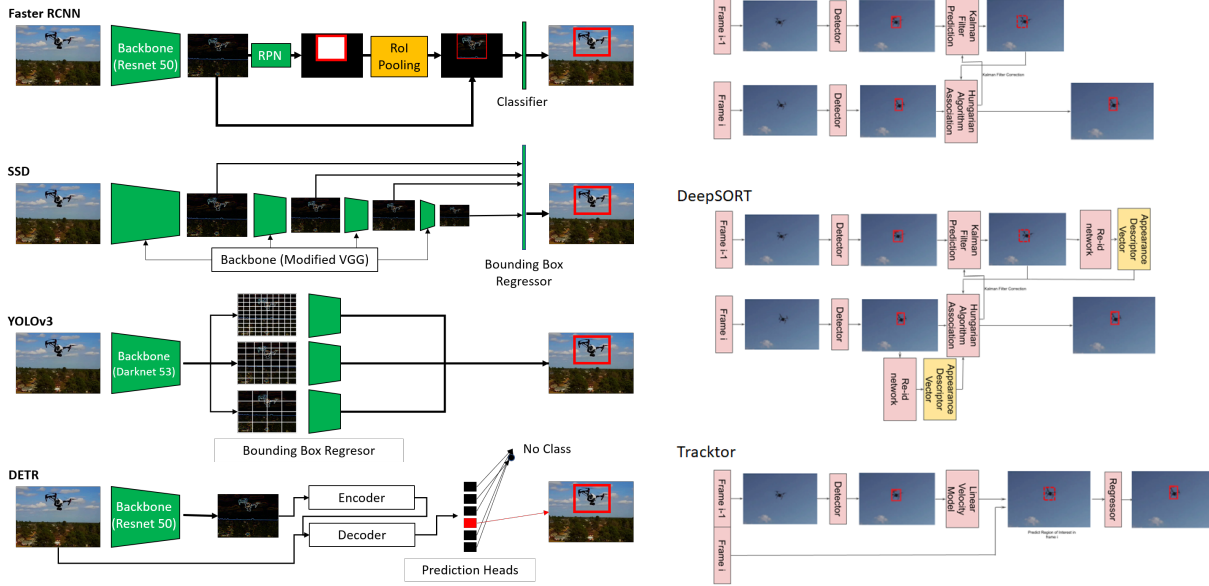
Figure 2: Architectures evaluated in this work. Detectors (left): Faster RCNN, SSD, YOLOv3 and DETR. Tracking frameworks (right): SORT, DeppSORT and Tracktor.

3.2. Object Detection

With the aim of assessing the detection performance, we evaluate four detectors corresponding to different object detection paradigms: Faster RCNN [38], SSD512 [33], YOLOv3 and DETR [5]. All networks are pretrained on the MS COCO dataset [31], a general-purpose dataset containing more than 330k images and 1.5M instances over 80 categories. Regarding the optimisation algorithm, all models were trained for 24 epochs, decreasing the learning rate by a factor of 0.1 after epochs 16 and 22. A brief diagram for each detector architecture used in this work is shown in Figure 2.

Faster RCNN [38] is trained using a ResNet-50 [19] as backbone and a Feature Pyramid Network [30] at the end of each convolutional block (conv2_3, conv3_4, conv4_6 and conv5_3) with 256 output channels on each level. Batch normalization is used for regularisation and stochastic gradient descent with a learning rate of 0.001.

SSD512 [33] is implemented with the same configuration details in the original work and with the same optimisation settings as in Faster RCNN.

YOLOv3 [37], from the YOLO family of detectors, is trained using a DarkNet-53 backbone. Input images were resized and square padded to have a final size of 608×608. The same optimisation technique as in Faster RCNN and SSD512 was used.

DETR [5], one of the most recent architectures for detection, is trained with a ResNet-50 backbone and the orig-

inal implementation details (AdamW [34] as the optimisation algorithm with an initial learning rate of 10^{-4}). However, the learning rate was decreased by a factor of 10 for the Drone-vs-Bird dataset to achieve convergence.

Detection performance is evaluated using MS COCO metrics. In order to compare our results with other works, we will refer the COCO $AP_{0.5}$ simply as *mAP*. All models were trained using the MMDetection framework [7] with an Nvidia 2080 Ti GPU.

3.3. Object Tracking

To evaluate object tracking within the counter-UAV domain, we employ each of the detectors detailed in section 3.2, within three MOT frameworks: SORT [4], DeepSORT [48], and Tracktor [3]. While several of the datasets consist of only a single UAV within each frame, and thus SOT trackers would suffice, we consider only MOT trackers that are more generalizable to counter-UAV domains.

SORT was implemented using the same configuration as the original paper from Bewley *et al.* [4]. As such, a Kalman filter is employed to capture the linear velocity motion model of the UAV, with the Hungarian algorithm to associate detected and predicted tracks.

Our **DeepSORT** framework builds upon SORT as per the original work by Wojke *et al.* [48]. Thus we incorporate a pre-trained deep re-identification network (that can be found within the MMTracking toolbox) to generate appearance descriptors. This network has been trained by the

Tracktor authors and is equivalently usable within the Deep-SORT framework. Additional DeepSORT tracking details can be found within the original paper. Otherwise, implementation details remain the same as SORT.

Tracktor utilizes a camera motion compensation model to generate track predictions, alongside the same re-identification network employed within DeepSORT. Other implementation details are unchanged from the original paper from Bergmann *et al*. [3].

We test the performance of each tracking framework under the UAV datasets illustrated in 2, with respect to multi object tracking accuracy (MOTA [45]), the ratio of correctly identified detections over the average number of ground-truth and computed detections (IDF1 [39]), and how often a detected object is assigned an incorrect ID ('ID Sw.'). Each tracker was implemented using the MMTracking toolbox[1].

4. Results

In this section we review the results for both detection and tracking tasks. These results correspond to the best performing detectors when evaluated over the validation sets; the same maximally performing weights are used within the tracking evaluation.

4.1. Detection

Comparative UAV detection performance among the datasets and detectors is shown in Table 1. It can be observed that detection performance varies significantly across the datasets with the best performance achieved in the Anti-UAV RGB dataset (0.986 mAP). This dataset has the most uniform set of image sequences since all videos are recorded in similar settings, with only 6 types of UAV present within it. The mAP across both modalities in the Anti-UAV dataset remains similar (mAP ranges from 0.978 to 0.986 for RGB and from 0.975 to 0.980 for IR), implying that the detectors are learning to capture the shape rather than the colour information. This is further supported by the cross-modality experiments, discussed later. However, the IR dataset obtains a dramatic increase of the AP for small objects (maximal AP_S: 0.533 via YOLOv3). This can be explained by the higher contrast for images captured with harsh weather conditions or poor lighting. Detection performance for the MAV-VID dataset is similar to Anti-UAV (0.978 mAP with Faster RCNN). Although MAV-VID dataset comprises bigger and clearer UAV images, it has less input data than the other two datasets. Additionally, the low precision for small objects can be explained by the low number of visually small UAV in the training set. This reveals the need of annotated datasets with UAV at a long distance. Finally, the Drone-vs-Bird dataset obtains a maximal

mAP of 0.667 via DETR and COCO AP of 0.283 via Faster RCNN. This indicates that Faster RCNN is better at predicting tighter bounding boxes. The relatively poorer performance in the Drone-vs-Bird dataset can be attributed to the significantly smaller size of the UAVs (Figure 1).

Faster RCNN and DETR generally outperform the one-stage detectors. However, it is important to note that both Faster RCNN and DETR have a minimum image height of 800 pixels which allows for a better performance at detecting small UAV. On the other hand, the inference time for SSD512 and YOLOv3 models are faster with 32.4 and 36.0 fps, as seen in Table 2. Although YOLOv3 is the fastest architecture, SSD512 is the lightest model with only 24M parameters ($\sim$192 MB). As a result, in settings where memory and computational power is a constraint (such as an embedded system), lightweight networks such as SSD or YOLOv3 may be preferred.

Lastly, we evaluate the detector performance on RGB and IR modalities when testing on a different modality to the one that the network was trained on (Table 3). The results show that features learned in the IR domain can be used as a good approximation for detecting drones in the visual spectrum (0.828 mAP with Faster RCNN). This may indicate that the network is learning to represent a UAV based on its shape and not in its colour or texture. On the other hand, the Anti-UAV IR mAP dropped to 0.644 with Faster RCNN trained on the Anti-UAV RGB. This bigger may be caused by the loss of the colour information in IR imagery. One-stage detectors performed poorer, indicating the difficulty of getting the proper features on each grid cell defined by these architectures.

4.2. Tracking

We compare the performance of each detector within each tracking framework in Table 4. It can be observed that tracking performance varies dramatically, but largely depends upon the detection architecture employed. Evidently, Faster-RCNN once again clearly outperforms other detection models for the easier MAV-VID challenge (MOTA: 0.955), while SSD512 performs best overall for the Drone-vs-Bird (MOTA: 0.525) and Yolov3 for Anti-UAV (up to MOTA: 0.985) challenges. However, DETR still yields highest performance towards tracking small objects (better shown in the detection results in Table 4.1). The Tracktor tracking framework overall yields best performance, owing to its vastly better ability to correctly attribute ids.

Interestingly, tracking accuracy is generally similar across all three tracking frameworks, as shown by the MOTA results in Table 4; differences in accuracy between trackers is less apparent than the difference afforded by different detection architectures. We can attribute this behaviour to two primary causes. Firstly, the Kalman filter within the SORT variants, and the motion model within

---

[1]Marginal changes are required to the MMTracking toolbox to enable single stage detectors within Tracktor (as per 05/03/2021)

Table 1: Detection Performance Benchmark

Dataset	Model	AP	$AP_{0.5}$	$AP_{0.75}$	AP_S	AP_M	AP_L	AR	AR_S	AR_M	AR_L
MAV-VID	Faster RCNN	**0.592**	**0.978**	**0.672**	**0.154**	**0.541**	**0.656**	0.659	0.369	0.621	0.721
	SSD512	0.535	0.967	0.536	0.083	0.499	0.587	0.612	**0.377**	0.578	0.666
	YOLOv3	0.537	0.963	0.542	0.066	0.471	0.636	0.612	0.208	0.559	0.696
	DETR	0.545	0.971	0.560	0.044	0.490	0.612	**0.692**	0.346	**0.661**	**0.742**
Drone-vs-Bird	Faster RCNN	**0.283**	0.632	**0.197**	**0.218**	**0.473**	0.506	0.356	0.298	0.546	0.512
	SSD512	0.250	0.629	0.134	0.199	0.422	0.052	0.379	0.327	0.549	0.556
	YOLOv3	0.210	0.546	0.105	0.158	0.395	0.356	0.302	0.238	0.512	**0.637**
	DETR	0.251	**0.667**	0.123	0.190	0.444	**0.533**	**0.473**	**0.425**	**0.631**	0.550
Anti-UAV Full	Faster RCNN	0.612	0.974	**0.701**	0.517	**0.619**	**0.737**	0.666	0.601	0.670	0.778
	SSD512	**0.613**	**0.982**	0.697	0.527	**0.619**	0.712	**0.678**	0.616	**0.682**	**0.780**
	YOLOv3	0.604	0.977	0.676	0.529	**0.619**	0.708	0.667	**0.618**	0.668	0.760
	DETR	0.586	0.977	0.648	0.509	0.589	0.692	0.649	0.598	0.649	0.752
Anti-UAV RGB	Faster RCNN	**0.642**	0.982	**0.770**	0.134	**0.615**	0.718	0.694	0.135	0.677	0.760
	SSD512	0.627	0.979	0.747	0.124	0.593	0.718	**0.703**	0.156	**0.682**	0.785
	YOLOv3	0.617	**0.986**	0.717	**0.143**	0.595	0.702	0.684	**0.181**	0.664	0.758
	DETR	0.628	0.978	0.740	0.129	0.590	**0.734**	0.700	0.144	0.675	**0.794**
Anti-UAV IR	Faster RCNN	0.581	0.977	0.641	0.523	0.623	-	0.636	0.602	0.663	-
	SSD512	0.590	0.975	0.639	0.518	0.636	-	0.649	0.609	0.681	-
	YOLOv3	0.591	0.976	0.643	**0.533**	0.638	-	0.651	0.620	0.675	-
	DETR	**0.599**	**0.980**	**0.655**	0.525	**0.642**	-	**0.671**	**0.633**	**0.701**	-

(a) MAV-VID   (b) Drone-vs-Bird   (c) Anti-UAV (RGB)   (d) Anti-UAV (Infrared)
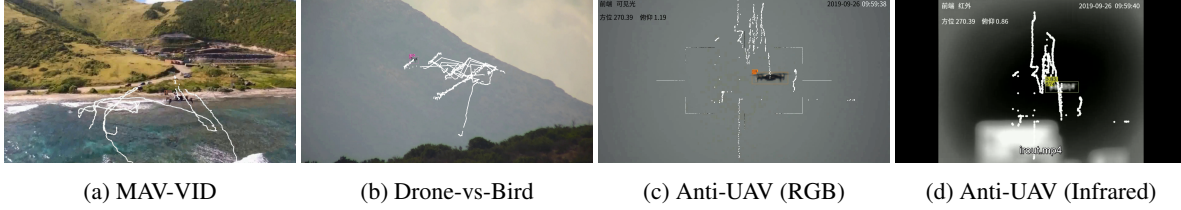
Figure 3: Examples from each dataset of visualization of drone position history

Table 2: Detection inference time

Model	GFLOPS	FPS	Params
Faster RCNN	207	18.0	41M
SSD512	88	32.4	24M
YOLOv3	70	36.0	61M
DETR	86	21.4	41M

Tracktor are unable to account for the sudden camera movements within the videos in the dataset. As such, we suspect the differences in trackers are ameliorated when they all fail to reidentify UAV when the camera stops moving. This poses a particular problem for Tracktor, where the regions of interest proposed by the detector and linear velocity model (even after camera motion compensation), are very far from where the UAV will be in the next view. Therefore, the regressor would be unable to locate the new position of the UAV. The challenge can be visualized within Figure 3b, where the long, rigid lines white lines indicate where the tracker lost the UAV after the sharp movement of the camera. On the other hand, the light camera shake (the jagged lines within the figure) does not present a problem, through which the UAV is successfully tracked.

Secondly, the reidentification network used within Deep-SORT and Tracktor is unable to provide an effective appearance descriptor for UAV, particularly for infrared images. The descriptor is unable to reduce id-switching within tracking (the DeepSORT 'ID Sw.' column in table 4 exceeds that of SORT), nor can it contribute to reducing occlusion introduced by other objects within the dynamic environment. As such, the difference in accuracy between tracking frameworks is minimized. In many cases in fact, the re-id network serves to reduce accuracy, where the SORT algorithm resultantly exceeds the DeepSORT algorithm with regards to MOTA. This phenomenon also partially explains how Tracktor yields superior IDF1 performance, which only uses the weak reidentification model to associate new, and previously killed tracks. DeepSORT uses the appearance descriptor at each and every time step however, as part of the distance metric within the Hungarian algorithm. Nevertheless, these results serve to demonstrate the importance of the re-identification network within the tracking pipeline. Investigation of improved re-identification for counter-UAV systems is subsequently identified as an area for future work.

Table 3: Cross modality detection performance

Dataset	Model	AP	AP$_{0.5}$	AP$_{0.75}$	AP$_S$	AP$_M$	AP$_L$	AR	AR$_S$	AR$_M$	AR$_L$
Anti-UAV RGB	IR Faster RCNN	**0.394**	**0.828**	**0.481**	0.322	0.000	**0.372**	**0.481**	**0.484**	0.000	**0.453**	**0.598**
	IR SSD512	0.161	0.342	0.124	0.000	0.126	0.291	0.204	0.016	0.165	0.341
	IR YOLOv3	0.113	0.299	0.061	0.002	**0.082**	0.240	0.156	**0.019**	0.119	0.286
	IR DETR	0.284	0.728	0.148	0.000	0.241	0.409	0.390	0.000	0.354	0.519
Anti-UAV IR	RGB Faster RCNN	**0.321**	**0.644**	**0.281**	**0.241**	**0.371**	-	**0.488**	**0.453**	**0.515**	-
	RGB SSD512	0.169	0.452	0.098	0.133	0.213	-	0.344	0.312	0.370	-
	RGB YOLOv3	0.215	0.475	0.164	0.138	0.303	-	0.303	0.228	0.363	-
	RGB DETR	0.149	0.333	0.109	0.125	0.170	-	0.216	0.190	0.236	-

Table 4: Object Tracking Performance Benchmark

Dataset	Model	MOTA ↑	IDF1 ↑	IDSw. ↓
		SORT / DeepSORT / Tracktor		
MAV-VID	Faster RCNN	0.952 / 0.951 / **0.955**	0.547 / 0.591 / 0.680	54 / 69 / **15**
	SSD512	0.865 / 0.851 / 0.888	0.402 / 0.430 / 0.944	283 / 435 / 35
	YOLOv3	0.881 / 0.877 / 0.934	0.260 / 0.419 / 0.970	269 / 309 / 26
	DETR	0.943 / 0.944 / 0.947	0.600 / 0.603 / **0.975**	61 / 50 / 16
Drone-vs-Bird	Faster RCNN	0.420 / 0.415 / 0.459	0.377 / 0.391 / 0.439	35183 / 35854 / 21952
	SSD512	0.364 / 0.328 / **0.525**	0.304 / 0.335 / **0.872**	35421 / 39532 / 16517
	YOLOv3	0.400 / 0.370 / 0.523	0.287 / 0.359 / 0.868	38451 / 41887 / 19623
	DETR	0.012 / 0.090 / 0.130	0.020 / 0.303 / 0.377	29126 / 29407 / **15316**
Anti-UAV Full	Faster RCNN	0.878 / 0.835 / 0.950	0.445 / 0.478 / 0.790	2620 / 4094 / 179
	SSD512	0.856 / 0.806 / 0.936	0.373 / 0.406 / 0.966	2831 / 4509 / 93
	YOLOv3	0.767 / 0.693 / **0.959**	0.193 / 0.292 / **0.979**	3582 / 6097 / 98
	DETR	0.749 / 0.708 / 0.830	0.405 / 0.451 / 0.920	2814 / 4214 / **40**
Anti-UAV RGB	Faster RCNN	0.791 / 0.717 / 0.912	0.295 / 0.290 / 0.702	2042 / 3163 / 163
	SSD512	0.771 / 0.696 / 0.900	0.271 / 0.306 / 0.946	2096 / 3239 / 126
	YOLOv3	0.714 / 0.633 / 0.897	0.150 / 0.227 / 0.948	2390 / 3617 / 87
	DETR	0.803 / 0.753 / **0.940**	0.323 / 0.327 / **0.969**	2101 / 2857 / **19**
Anti-UAV IR	Faster RCNN	0.954 / 0.919 / **0.987**	0.528 / 0.571 / 0.873	560 / 1154 // 22
	SSD512	0.922 / 0.904 / 0.963	0.442 / 0.454 / 0.981	716 / 1020 / 20
	YOLOv3	0.780 / 0.718 / 0.985	0.180 / 0.327 / **0.992**	1216 / 2259 / **10**
	DETR	0.951 / 0.930 / 0.984	0.552 / 0.628 / 0.991	575 / 927 / 20

5. Conclusion

Computer vision for UAV is an exciting but underdeveloped research field, relative to the necessity of aviation safety and the potential threat (advertent or inadvertent) that UAV carry. To accelerate research in this field, we have conducted a benchmark study on UAV detection and tracking. We have performed an exhaustive evaluation over well-established generic object detectors on three UAV datasets. To the best of our knowledge, we are the first to evaluate a comprehensive array of datasets, across environments, and conditions, captured by both ground and UAV-mounted cameras, optical and IR cameras, at near and far viewpoints. As such, we are able to show that common neural network architectures are capable of detecting small, fast moving objects. Furthermore, YOLOv3 overall yields the best precision (as high as 0.986 mAP), but Faster RCNN is best suited towards detecting small UAV, for which category it consistently achieves the highest mAP (up to 0.770). This is a critical observation when we consider that detecting UAV early is often a stronger priority than detecting precisely.

We also show that cross-modality detection is possible for optical and infrared images, wherein training on IR imagery for inference upon visible band spectrum data is not only sufficient but yields excellent performance (0.828 mAP). Furthermore, our tracking results corroborate the suitability towards employing DETR within counter-UAV systems; DETR is best suited for a detection backbone to tracking systems for small objects, and performs adequately upon cross-modal videos. We further demonstrate that the Tracktor framework is best able to track UAV (MOTA: 0.987).

This study serves as a baseline for the community to work towards UAV-specific detectors and trackers. We identify four future research directions to improve counter-UAV systems: 1) improving long-distance detection ability for when UAV appear small; 2) cross modality training to enforce learning UAV shape 3) bespoke re-identification networks for UAV; 4) improving available datasets. Nevertheless, we have demonstrated the efficacy of generic object detectors within both detection and tracking frameworks, upon which bespoke architectures can only improve.

References

[1] S. Avidan. Ensemble tracking. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(2):261–271, 2007. 3

[2] S. Bae and K. Yoon. Robust online multi-object tracking based on tracklet confidence and online discriminative appearance learning. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1218–1225, 2014. 3

[3] P. Bergmann, T. Meinhardt, and L. Leal-Taixe. Tracking without bells and whistles. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 941–951, Los Alamitos, CA, USA, nov 2019. IEEE Computer Society. 2, 3, 5, 6

[4] A. Bewley, Z. Ge, L. Ott, F. Ramos, and B. Upcroft. Simple online and realtime tracking. In *2016 IEEE International Conference on Image Processing (ICIP)*, pages 3464–3468, 2016. 2, 3, 5

[5] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-End Object Detection with Transformers. *arXiv e-prints*, page arXiv:2005.12872, May 2020. 2, 3, 5

[6] Joint Air Power Competence Centre. A comprehensive approach to countering unmanned aircraft systems, Aug 2019. 2

[7] Kai Chen, Jiaqi Wang, Jiangmiao Pang, Yuhang Cao, Yu Xiong, Xiaoxiao Li, Shuyang Sun, Wansen Feng, Ziwei Liu, Jiarui Xu, Zheng Zhang, Dazhi Cheng, Chenchen Zhu, Tianheng Cheng, Qijie Zhao, Buyu Li, Xin Lu, Rui Zhu, Yue Wu, Jifeng Dai, Jingdong Wang, Jianping Shi, Wanli Ouyang, Chen Change Loy, and Dahua Lin. MMDetection: Open mmlab detection toolbox and benchmark. *arXiv preprint arXiv:1906.07155*, 2019. 5

[8] Wongun Choi. Near-online multi-target tracking with aggregated local flow descriptor. In *Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV)*, ICCV '15, page 3029–3037, USA, 2015. IEEE Computer Society. 3

[9] Wongun Choi. Near-online multi-target tracking with aggregated local flow descriptor. In *Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV)*, ICCV '15, page 3029–3037, USA, 2015. IEEE Computer Society. 3

[10] Gioele Ciaparrone, Francisco Luque Sánchez, Siham Tabik, Luigi Troiano, Roberto Tagliaferri, and Francisco Herrera. Deep learning in video multi-object tracking: A survey. *Neurocomputing*, 381:61–88, Mar 2020. 2

[11] Angelo Coluccia, Alessio Fascista, Arne Schumann, Lars Sommer, Marian Ghenescu, Tomas Piatrik, Geert De Cubber, Mrunalini Nalamati, Ankit Kapoor, Muhammad Saqib, et al. Drone-vs-bird detection challenge at ieee avss2019. In *2019 16th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, pages 1–7. IEEE, 2019. 2, 3, 4

[12] Celine Craye and Salem Ardjoune. Spatio-temporal semantic segmentation for drone detection. In *2019 16th IEEE International conference on advanced video and signal based surveillance (AVSS)*, pages 1–5. IEEE, 2019. 3

[13] Jifeng Dai, Yi Li, Kaiming He, and Jian Sun. R-fcn: Object detection via region-based fully convolutional networks. In *Advances in neural information processing systems*, pages 379–387, 2016. 3

[14] A. Geiger, M. Lauer, C. Wojek, C. Stiller, and R. Urtasun. 3d traffic scene understanding from movable platforms. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(5):1012–1025, 2014. 3

[15] R. Girshick. Fast r-cnn. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 1440–1448, 2015. 2

[16] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pages 580–587, 2014. 3

[17] HM Government. Counter-unmanned aircraft strategy, ISBN: 978-1-5286-1554-9, 2019. 1

[18] S. A. Hassan, T. Rahim, and S. Y. Shin. Real-time uav detection based on deep learning network. In *2019 International Conference on Information and Communication Technology Convergence (ICTC)*, pages 630–632, 2019. 2

[19] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition, 2015. 5

[20] J. Ho, Kuang-Chih Lee, Ming-Hsuan Yang, and D. Kriegman. Visual tracking using learned linear subspaces. In *Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2004. CVPR 2004.*, volume 1, pages I–I, 2004. 3

[21] Michael Isard and Andrew Blake. Condensation—conditional density propagation for visual tracking. *International Journal of Computer Vision*, 29(1):5–28, Aug 1998. 3

[22] M. Jahangir and C. Baker. Robust detection of micro-uas drones with l-band 3-d holographic radar. In *2016 Sensor Signal Processing for Defence (SSPD)*, pages 1–5, 2016. 1

[23] Sungho Jeon, Jong-Woo Shin, Young-Jun Lee, Woong-Hee Kim, YoungHyoun Kwon, and Hae-Yong Yang. Empirical study of drone sound detection in real-life environment with deep neural networks, 2017. 1

[24] Nan Jiang, Kuiran Wang, Xiaoke Peng, Xuehui Yu, Qiang Wang, Junliang Xing, Guorong Li, Jian Zhao, Guodong Guo, and Zhenjun Han. Anti-uav: A large multi-modal benchmark for uav tracking, 2021. 2, 4

[25] J. Klare, O. Biallawons, and D. Cerutti-Maori. Uav detection with mimo radar. In *2017 18th International Radar Symposium (IRS)*, pages 1–8, 2017. 1

[26] H. Kuhn. The hungarian method for the assignment problem. *Naval Research Logistic Quarterly*, 2, 05 2012. 3

[27] S. L and C. S. Christopher. Video surveillance using deep learning - a review. In *2019 International Conference on Recent Advances in Energy-efficient Computing and Communication (ICRAECC)*, pages 1–5, 2019. 2

[28] Hei Law and Jia Deng. Cornernet: Detecting objects as paired keypoints. In *Proceedings of the European conference on computer vision (ECCV)*, pages 734–750, 2018. 3

[29] Bo Li, Wei Wu, Qiang Wang, Fangyi Zhang, Junliang Xing, and Junjie Yan. Siamrpn++: Evolution of siamese visual

tracking with very deep networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4282–4291, 2019. 3

[30] T. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie. Feature pyramid networks for object detection. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 936–944, 2017. 3, 5

[31] Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. Microsoft coco: Common objects in context, 2015. 5

[32] Li Liu, Wanli Ouyang, Xiaogang Wang, Paul Fieguth, Jie Chen, Xinwang Liu, and Matti Pietikäinen. Deep learning for generic object detection: A survey. *International journal of computer vision*, 128(2):261–318, 2020. 2

[33] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C. Berg. Ssd: Single shot multibox detector. In Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling, editors, *Computer Vision – ECCV 2016*, pages 21–37, Cham, 2016. Springer International Publishing. 2, 3, 5

[34] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization, 2019. 5

[35] Georgia Lykou, Dimitrios Moustakas, and Dimitris Gritzalis. Defending airports from uas: A survey on cyber-attacks and counter-drone sensing technologies. *Sensors*, 20(12), 2020. 2

[36] Mrunalini Nalamati, Ankit Kapoor, Muhammed Saqib, Nabin Sharma, and Michael Blumenstein. Drone detection in long-range surveillance videos. In *2019 16th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, pages 1–6. IEEE, 2019. 2

[37] Joseph Redmon and Ali Farhadi. Yolov3: An incremental improvement. *arXiv*, 2018. 2, 3, 5

[38] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks, 2016. 2, 3, 5

[39] Ergys Ristani, Francesco Solera, Roger Zou, Rita Cucchiara, and Carlo Tomasi. Performance measures and a data set for multi-target, multi-camera tracking. In Gang Hua and Hervé Jégou, editors, *Computer Vision – ECCV 2016 Workshops*, Cham, 2016. Springer International Publishing. 6

[40] A. Rodriguez-Ramos, J. Rodriguez-Vazquez, C. Sampedro, and P. Campoy. Adaptive inattentional framework for video object detection with reward-conditional training. *IEEE Access*, 8:124451–124466, 2020. 2, 3

[41] A. Rozantsev, V. Lepetit, and P. Fua. Flying objects detection from a single moving camera. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4128–4136, 2015. 2

[42] Waylon Dustin Scheller. Detecting drones using machine learning. Master's thesis, Iowa State University, 2017. 2

[43] Pierre Sermanet, D. Eigen, X. Zhang, Michaël Mathieu, R. Fergus, and Y. LeCun. Overfeat: Integrated recognition, localization and detection using convolutional networks. *CoRR*, abs/1312.6229, 2014. 3

[44] Z. Soleimanitaleb, M. A. Keyvanrad, and A. Jafari. Object tracking methods:a review. In *2019 9th International Conference on Computer and Knowledge Engineering (ICCKE)*, pages 282–288, 2019. 2

[45] Rainer Stiefelhagen, Keni Bernardin, Rachel Bowers, John Garofolo, Djamel Mostefa, and Padmanabhan Soundararajan. The clear 2006 evaluation. In Rainer Stiefelhagen and John Garofolo, editors, *Multimodal Technologies for Perception of Humans*, pages 1–44, Berlin, Heidelberg, 2007. Springer Berlin Heidelberg. 6

[46] B. Taha and A. Shoufan. Machine learning-based drone detection and classification: State-of-the-art in research. *IEEE Access*, 7:138669–138682, 2019. 1

[47] Eren Unlu, Emmanuel Zenou, Nicolas Riviere, and Paul-Edouard Dupouy. Deep learning-based strategies for the detection and tracking of drones using several cameras. *IPSJ Transactions on Computer Vision and Applications*, 11(1):1–13, 2019. 2

[48] N. Wojke, A. Bewley, and D. Paulus. Simple online and realtime tracking with a deep association metric. In *2017 IEEE International Conference on Image Processing (ICIP)*, pages 3645–3649, 2017. 2, 3, 5

[49] Yu Xiang, Alexandre Alahi, and Silvio Savarese. Learning to track: Online multi-object tracking by decision making. In *Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV)*, ICCV '15, page 4705–4713, USA, 2015. IEEE Computer Society. 3

[50] Y. Xiao and X. Zhang. Micro-uav detection and identification based on radio frequency signature. In *2019 6th International Conference on Systems and Informatics (ICSAI)*, pages 1056–1062, 2019. 1

[51] B. Yang, E. T. Matson, A. H. Smith, J. E. Dietz, and J. C. Gallagher. Uav detection system with multiple acoustic nodes using machine learning models. In *2019 Third IEEE International Conference on Robotic Computing (IRC)*, pages 493–498, 2019. 1

[52] Min Yang and Yunde Jia. Temporal dynamic appearance modeling for online multi-person tracking. *Comput. Vis. Image Underst.*, 153(C):16–28, Dec. 2016. 3

[53] Wenhan Yang, Ye Yuan, Wenqi Ren, Jiaying Liu, Walter J Scheirer, Zhangyang Wang, Taiheng Zhang, Qiaoyong Zhong, Di Xie, Shiliang Pu, et al. Advancing image understanding in poor visibility environments: A collective benchmark study. *IEEE Transactions on Image Processing*, 29:5737–5752, 2020. 3

[54] Z. Zhao, P. Zheng, S. Xu, and X. Wu. Object detection with deep learning: A review. *IEEE Transactions on Neural Networks and Learning Systems*, 30(11):3212–3232, 2019. 2