MOR-UAV: A Benchmark Dataset and Baselines for Moving Object Recognition in UAV Videos

Murari Mandal, Lav Kush Kumar, Santosh Kumar Vipparthi
murarimandal.cv@gmail.com,lavkushkumarmnit@gmail.com,skvipparthi@mnit.ac.in
Vision Intelligence Lab, Malaviya National Institute of Technology Jaipur, INDIA

ABSTRACT

Visual data collected from Unmanned Aerial Vehicles (UAVs) has opened a new frontier of computer vision that requires automated analysis of aerial images/videos. However, the existing UAV datasets primarily focus on object detection. An object detector does not differentiate between the moving and non-moving objects. Given a real-time UAV video stream, how can we both localize and classify the moving objects, i.e. perform moving object recognition (MOR)? The MOR is one of the essential tasks to support various UAV vision-based applications including aerial surveillance, search and rescue, event recognition, urban and rural scene understanding.To the best of our knowledge, no labeled dataset is available for MOR evaluation in UAV videos. Therefore, in this paper, we introduce MOR-UAV, a large-scale video dataset for MOR in aerial videos. We achieve this by labeling axis-aligned bounding boxes for moving objects which requires less computational resources than producing pixel-level estimates. We annotate 89,783 moving object instances collected from 30 UAV videos, consisting of 10,948 frames in various scenarios such as weather conditions, occlusion, changing flying altitude and multiple camera views. We assigned the labels for two categories of vehicles (car and heavy vehicle). Furthermore, we propose a deep unified framework MOR-UAVNet for MOR in UAV videos. Since, this is a first attempt for MOR in UAV videos, we present 16 baseline results based on the proposed framework over the MOR-UAV dataset through quantitative and qualitative experiments. We also analyze the motion-salient regions in the network through multiple layer visualizations. The MOR-UAVNet works online at inference as it requires only few past frames. Moreover, it doesnâĂŹt require predefined target initialization from user. Experiments also demonstrate that the MOR-UAV dataset is quite challenging.

CCS CONCEPTS

• **Applied computing** → **Surveillance mechanisms**; • **Computing methodologies** → **Motion capture**; **Object recognition**.

KEYWORDS

Remote sensing, UAV, aerial vehicular vision, moving object recognition, deep learning

MM '20, October 12–16, 2020, Seattle, WA, USA

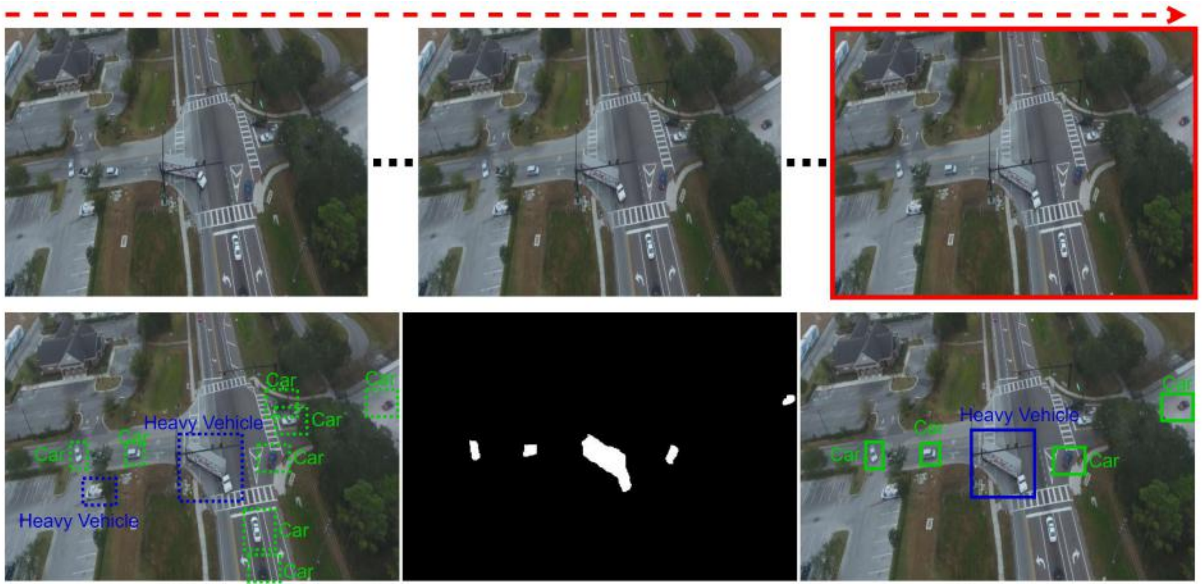
ACM ISBN 978-1-4503-7988-5/20/10…$15.00
https://doi.org/10.1145/3394171.3413934

ACM Reference Format:
Murari Mandal, Lav Kush Kumar, Santosh Kumar Vipparthi. 2020. MOR-UAV: A Benchmark Dataset and Baselines for Moving Object Recognition in UAV Videos. In *Proceedings of the 28th ACM International Conference on Multimedia (MM '20), October 12–16, 2020, Seattle, WA, USA.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3394171.3413934

1 INTRODUCTION

Visual data captured from unmanned aerial vehicles (UAVs) is used in numerous applications for military, industry and consumer market space. It has opened a new frontier of computer vision, i.e., aerial vision, which requires analysis and interpretation of images and videos gathered from UAVs either in real-time or in post-event study. Some of the applications include aerial surveillance, search and rescue, exploration, urban planning, industrial and agricultural monitoring, action/event recognition, sports analysis and scene understanding [68, 72].

In addition to the existing datasets and algorithms [3, 4, 7, 8, 10, 13, 14, 17, 19–22, 24–27, 34, 36, 40, 44–46, 48, 50, 51, 55–59, 62, 69] for analysis of images/videos captured in regular view, researchers in the literature have also developed numerous UAV datasets for some of the fundamental tasks such as detection and tracking. The Campus [54] and CARPK [23] videos are captured in specific locations such as campus or parking lot. The UAV123[47] was developed for low-altitude object tracking in both real and simulated scenes. Similarly, the Okutama [4] provides annotations for human action recognition in aerial views. More recently, several datasets were presented for object detection [4, 35, 52, 68, 70, 72, 74] and tracking [12, 23, 72, 75] in unconstrained aerial scenarios. These datasets have accelerated deep learning research in UAV vision-based applications. However, no labeled dataset is available in the literature for one of the low-level tasks of moving object recognition (MOR), i.e. simultaneous localization and classification of moving objects in a video frame. The task of MOR is different from both object detection and visual tracking. In both these tasks, the methods do not differentiate between the moving and non-moving objects. Furthermore, MOR is even different from moving object detection (MOD) which performs pixel-wise binary segmentation in each frame. The difference between object detection (OD), MOD and MOR are demonstrated in Figure 1. MOR has widespread applications in intelligent visual surveillance, intrusion detection, agricultural monitoring, industrial site monitoring, detection-based tracking, autonomous vehicles, etc. More specifically, the MOR algorithm can recognize the movements of different classes of objects such as people, vehicles, etc. in real-world scenarios and this information can further be used in high-level decision making such as anomaly detection and selective target tracking. Thus, there is a



(a) OD (b) MOD (c) MOR

Figure 1: Difference between the three tasks: object detection (OD), moving object detection (MOD) and MOR is depicted. The OD algorithm detects all (moving and non-moving) object instances. Whereas, in MOD, only the pixel-wise changes are identified in a class-agnostic manner. The proposed MOR method simultaneously detects and classifies the moving objects in the video. The same is shown in this figure.

need for a comprehensive UAV benchmark video dataset for MOR in unconstrained scenarios.

To advance the MOR research in aerial videos, this paper introduces a largescale challenging UAV moving object recognition benchmark dataset named MOR-UAV. The MOR-UAV consists of 89, 783 moving object instances annotated in 30 videos comprising of 10,948 frames. The videos are captured in various challenging scenarios such as night time, occlusion, camera motion, weather conditions, camera views, etc. Moreover, no constraints are enforced to maintain the same camera sensor and platform for all videos. This puts the onus on the MOR algorithm to robustly recognize the moving objects in the wild. The altitude and camera views also vary in different video sequences. All the moving objects are manually annotated with bounding boxes along with the object category. In this paper, the objects of interest are two types of vehicles: *cars* and *heavy vehicles*. A sample set of video frames from MOR-UAV dataset is depicted in Figure 2. The complete dataset with annotations will be made publicly available in the future.

We also present an online deep unified framework named MOR-UAVNet for MOR evaluation in the newly constructed dataset. The proposed framework retains the property of online inference by using only the recent history frames in live streaming videos. This paper makes the following contributions:

- We introduce a fully annotated dataset, MOR-UAV for the fundamental task of MOR in unconstrained UAV videos. To the best the author's knowledge, this is the first aerial/UAV video dataset for simultaneous localization and classification of moving objects.
- We propose a novel deep learning framework MOR-UAVNet to simultaneously localize and classify the moving objects in the UAV videos. The visualizations of different convolutional layers in MOR-UAVNet is qualitatively analyzed for a better understanding of the network functionality for MOR.
- Based on the new MOR-UAVNet framework, we evaluate 16 baseline models for MOR evaluation over MOR-UAV. We also depict the qualitative results for MOR in UAV videos.

2 MOR-UAV DATASET

2.1 Comparison with Existing UAV Datasets

Advancement in deep learning algorithms and the availability of largescale labeled datasets has fueled the progress in important applications in several domains. Some of the low-level tasks in computer vision include image classification [10, 20, 21, 55], object detection [31–33, 36, 53, 73], semantic segmentation [7, 19], video object segmentation [44, 51, 69], motion detection [1, 38, 39, 41, 48, 62, 71] and visual tracking [2, 15, 61]. Although many challenging applications are presented to the researchers in UAV based computer vision community. However, limited labeled datasets [4, 12, 35, 47, 52, 68, 70, 72, 74, 75] are available for accelerating the improvement and evaluation of various vision tasks. Recently, numerous datasets have been constructed with annotations for object detection and visual tracking. Muellerl [47] presented a tracking dataset recorded from drone cameras to evaluate the ability of single object trackers to tackle camera movements, illumination variations, and object scale changes. Moreover, several video clips were recorded for pedestrian behavior analysis from an aerial view in [54]. Hsieh et al. [23]

Figure 2: Sample video frames taken from MOR-UAV. The videos are collected from different scenarios including night time, occlusion, camera motion, weather conditions, camera views, etc. The red and green boxes denote two object classes car and heavy-vehicle respectively.

augmented a dataset for vehicle counting in parking lots. Similarly, for object detection in aerial views, several datasets have been presented over time. Some of the most comprehensive and challenging datasets are the DOTA [68], UAVDet [72], VisDrone [74, 75] and Dac-sdc [70]. It has led to rapid advancement in the development of specialized object detectors [5, 11, 18, 30, 37, 42, 43, 66, 67] for aerial images. Similarly, VisDrone [12], UAVDet [72] and UAV123 [47] have provided annotated data for bench-marking visual tracking in UAV videos. Several tracking algorithms [49, 63] have been evaluated over these datasets to advance research in aerial object tracking. Furthermore, an extension of DOTA was recently made available in [64] to also facilitate instance segmentation in aerial images as well.

Few researchers [6, 28] have utilized the videos from VIVID [9] and UAV123 to self-annotate a few frames for moving object detection (MOD). Others [29] have collected UAV videos for specialized purposes. However, to the best of our knowledge, no labeled dataset is available in the literature for MOR. The proposed MOR-UAV is a first large-scale dataset with bounding-box labels which can be used as the benchmark for both MOR and MOD in UAV videos. The dataset consists of videos from numerous unconstrained scenarios, resulting in better generalization to unseen videos. The detailed comparison of the proposed MOR-UAV with other datasets in the literature is summarized in Table 1.

2.2 Data Collection and Annotation

The MOR-UAV dataset comprises of 30 videos collected from multiple video recordings captured with a UAV. Locations in highways, flyovers, traffic intersections, urban areas and agricultural regions are collected for analysis. These videos represent various scenarios including occlusion, nighttime, weather changes, camera motion, changing altitudes, different camera views, and angles. The videos are recorded at 30 frames per second (fps) and the resolution varies from 1280×720 to 1920×1080. The average, min, max lengths of the sequences are 364.93, 64 and 1,146 respectively.

Table 1: Comparison of MOR-UAV with other largescale UAV datasets. Det: Detection, T: Visual tracking, Act: Action recognition, MOR: Moving object recognition

Dataset	Tasks	Labeled Moving Objects
VisDrone [75]	Det, T	No
DOTA [68]	Det	No
UAV123 [47]	T	No
UAVDT [72]	Det, T	No
Okutama [4]	Det, Act	No
Dac-sdc [70]	Det	No
MOR-UAV	**MOR**	**Yes**

The moving objects are labeled using the Yolo-mark[1] tool. The bounding boxes are described with $(x1, y1, x2, y2, c)$, where $(x1, y1)$ and $(x2, y2)$ are the top-left and bottom-right locations of the bounding box respectively. The object class is represented with c. About 10,948 frames in the MOR-UAV dataset are annotated with approximately 89,783 bounding boxes representing moving vehicles. These vehicles are categorized into two classes: *car* (80,340 *bounding boxes*) and *heavy vehicles* (9,443 *bounding boxes*). Figure 2 shows some sample frames with annotations in the dataset. The *average*, *min*, *max* lengths of the bounding box heights are 29.011, 6, 181, respectively. Similarly, *average*, *min*, *max* lengths of the bounding box widths are 17.641, 6, 106, respectively. The complete dataset details are depicted in Figure 3.

2.3 Dataset Attributes

Some of the challenging attributes of the MOR-UAV dataset are as follows:

[1] https://github.com/AlexeyAB/Yolo_mark

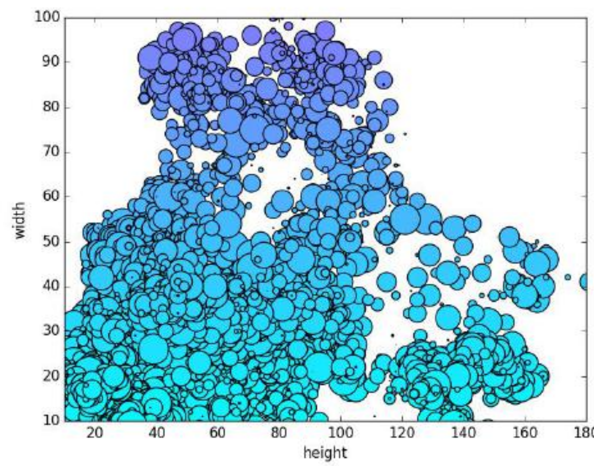
Variable object density. The UAV videos are captured in both dense and sparsely populated regions. For example, a large number of vehicles are present in flyovers, parking lot and traffic signal intersections. Whereas, very few objects are present in forest, agricultural and other remote areas with complex backgrounds. Such diverse scenarios in terms of object density make it challenging for the MOR algorithms to obtain robust performance.

Small and large object shapes. Due to high altitude of the UAVs, the objects appear very small. Thus, it is very difficult to accurately detect the motion features. Moreover, in some low altitude UAV videos, the objects appear reasonably large due to closer view. Thus, for the same class, both large, medium and small shapes of objects are present in the dataset. These multiscale object appearances challenge the researchers to design more generalizable algorithms.

Sporadic camera motion. In addition to the variable speed of object movements, the UAV camera speed and rotation is also unconstrained in the dataset. Sometimes the camera itself moves faster or rotates drastically. This makes it very difficult to differentiate between moving and non-moving objects. Moreover, sometimes it creates confusion between camera motion and actual object motion.

Changes in the aerial view. Changes in the camera orientations in UAV sometimes make the object appear from different side view angles or corner view in the video. Thus, multiple views of the objects appear in different videos or sometimes even in the same video.

Since the videos are collected from different open-source platforms, the exact details about the UAV altitude, camera angle, UAV speed, etc. are not available. Such unconstrained data collection makes the algorithm design more challenging and ensures robust performance in real-world UAV videos. Thus, the algorithms designed for MOR in this UAV video dataset should consider these practical factors. Moreover, the weather changes and nighttime videos present further challenges for accurate moving object recognition.



(a) Avg. BB height = 29.01, avg. BB width = 17.64, min. BB height = 6, min BB width = 6, max. BB height = 181, max. BB width = 106
(b) Avg. video sequence length = 364.93, min. video sequence length = 64, max. video sequence length = 1,146

Figure 3: The bounding-box (BB) height-width scatter-plot of all the object instances in MOR-UAV along with the complete dataset description. All the videos are normalized to the shape of $608 \times 608 \times 3$ for uniform analysis of the complete dataset.

3 MOR-UAVNET FRAMEWORK AND BASELINE MODELS

We propose a deep unified framework MOR-UAVNet for simultaneous localization and classification of moving objects. The overall architecture of MOR-UAVNet is shown in Figure 4. We discuss the functionality of the proposed network along with visualizations in the following subsections.

3.1 Parallel Feature Encoding for Motion and Object Saliency Estimation

To encode the motion-salient regions in the current frame, we compute optical flow at cascaded positions along the temporal dimension. The optical flow between the current and previous frames at multiple distances are computed. We compute the optical flow map between the current frame and frames from temporal history with distance 1, 3 and 5, respectively. If the cascaded optical flow and the current frame are denoted with C_{OF} and I, then the assimilated feature map is computed as given in Eq. 1.

$$AsOF = [C_{OF}, I] \tag{1}$$

It provides crucial encoding to construct coarse motion saliency maps. We then extract deep features from $AsOF$ for higher-level abstractions. The features are extracted from three layers ($C3, C4, C5$) of the ResNet residual stage as in [21, 32, 33]. We use resnet50 pre-trained over the ImageNet dataset as the backbone model for feature extraction. However, any other model can also be used as a backbone model. Moreover, in order to reinforce the semantic features of the salient foreground/moving objects more accurately, backbone features are also parallelly extracted from the current frame. These two sets of backbone feature maps are combined at matching scales for both temporal and spatial saliency aware feature representation. The motion-salient features are encoded as given in Eq. 2.

$$MSF = [resnet(AsOF), resnet(I)] \tag{2}$$

where $resnet(x)$ returns the features from ResNet50 backbone.

3.2 Baseline Models

Since this is the first attempt for MOR in UAV videos, we compute the baseline results using the proposed MOR-UAVNet framework. In addition to the proposed MOR-UAVNetv1, we also designed MOR-UAVNetv2 by removing the parallel feature extraction (computed separately for the current frame) part from the original network. Thus, we could also evaluate the effect of directly using AsOF without the reinforcements of base features extracted from the current frame. We created MOR-UAVNetv1, MOR-UAVNetv2 by using resnet50 [21] and MOR-UAVNetv3, MOR-UAVNetv4 by using mobileNetv2 [55] as backbone respectively.

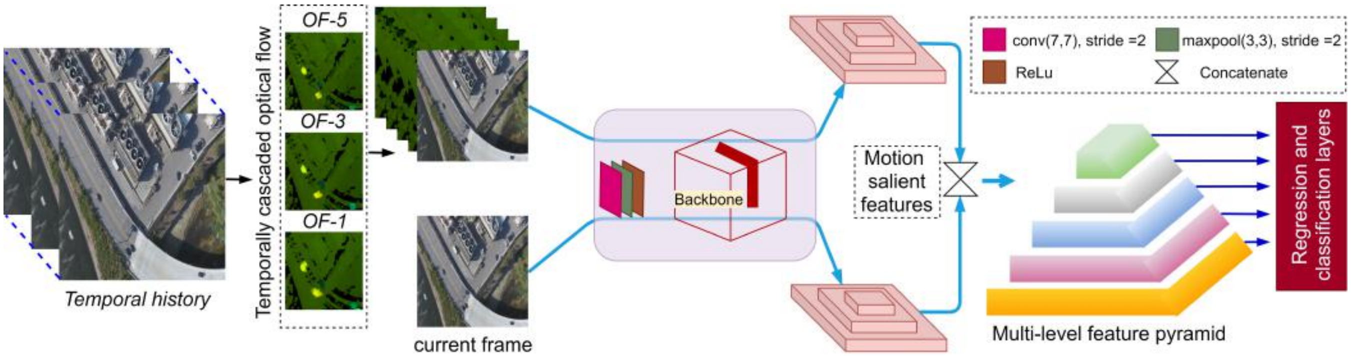
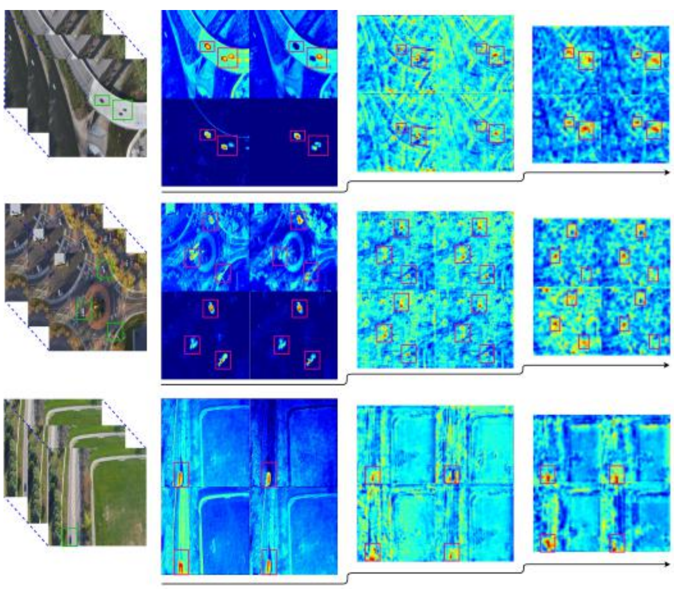


Figure 4: Schematic illustration of the proposed MOR-UAVNet framework for MOR in UAV videos. The motion saliency is estimated through cascaded optical flow computation at multiple stages in the temporal history frames. In this figure, optical flow between the current frame and the last ($OF-1$), third last ($OF-3$), fifth last frame ($OF-5$) is computed respectively. We then assimilate the salient motion features with the current frame. These assimilated features are forwarded through the ResNet backbone to extract spatial and temporal dimension aware features. Moreover, the base features from the current frame are also extracted to reinforce the semantic context of the object instances. These two feature maps are concatenated at matching scales to produce a feature map for motion encoding. Afterward, multi-level feature pyramids are generated. The dense bounding box and category scores are generated at each level of the pyramid. We use 5 pyramid levels in our experiments. *This figure shows the MOR-UAVNetv1 model architecture*



(a) (b) (c)

Figure 5: Visualization of different layers in MOR-UAVNet. (a) convolutional layer before the backbone feature extraction, (b) pyramid levelâĂŞP3, (c) pyramid levelâĂŞP4. The relevant motion saliencies of moving objects are highlighted using red boxes.

Thus, we have four different networks to compute baseline results on MOR-UAV dataset. Each network is trained four times for $T = 1(C_{OF} = 1), T = 2(C_{OF} = 1, 3), T = 2(C_{OF} = 1, 5)$ and $T = 3(C_{OF} = 1, 3, 5)$, respectively. Thus, overall, we evaluate the quantitative results for 16 models over MOR-UAV dataset.

3.3 Training, Inference & Visualization

Network Configurations. We resize all the video frames in MOR-UAV dataset to $608 \times 608 \times 3$ for a uniform setting in training and evaluation. The MOR-UAVNet takes two tensors of shape $608 \times 608 \times T$ (number of cascaded optical flow maps) and $608 \times 608 \times 3$ (current frame) as input and returns the spatial coordinates with class labels for moving object instances. We compute the dense optical flow using the algorithm given in [16]. The following values of T is used in our experiments: $T = 3(C_{OF} = 1-3-5), T = 2(C_{OF} = 1-3), T = 2(C_{OF} = 1-5) and T = 1(C_{OF} = 1)$ in our experiments.

Training. The one-stage MOR-UAVNet network is trained end-to-end with multiple input layers. The complete framework is implemented in Keras with Tensorflow backend. Training is performed with $batchsize = 1$ over 11 GB Nvidia RTX 2080 Ti GPU. Due to the memory limitation and large image-size, we use $batchsize = 1$ in our experiments. The network is optimized with Adam optimizer and initial learning rate of 10^{-5}. All models are trained for approximately $250K - 300K$ iterations. For regression and classification, $smooth\ell_1$ and focal loss functions are used respectively. The training loss is computed as the sum of the above-mentioned two losses.

Inference. Similar to training, inference involves simply giving the current frame and T cascaded optical flow maps computed from past history frames as input to the network. Only a few optical flow maps ($T = 1/2/3$) are required, enabling online moving object recognition for real-time analysis. Top 1000 prediction scores per pyramid level are considered after thresholding detector confidence at 0.05. The final detections are collected by combining top predictions from all levels and non-maximum suppression with a threshold of 0.5.

Table 2: Summary description of the training and testing sets used in our experiments for evaluation

Videos	#Frames	#Car	#Heavy Vehicle	#Objects
video1	299	6,026	600	6,626
video2	499	6,584	493	7,077
video3	225	1,575	0	1,575
video4	139	695	138	833
video5	351	1,049	610	1,659
video6	218	1,902	156	2,058
video7	64	65	42	107
video8	118	187	0	187
video9	299	4,571	251	4,822
video10	477	12,579	2,527	15,106
video11	225	8,699	2,617	11,316
video12	550	940	78	1,018
video13	285	0	286	286
video13	285	0	286	286
Training Set	*3,749*	*44,872*	*7,798*	*52,670*
video14	210	430	0	430
video15	199	626	0	626
video16	200	457	0	457
video17	70	427	71	498
Testing Set	*679*	*1,940*	*71*	*2,011*

Visualization. We depict the intermediate layer visualizations of MOR-UAVNet in Figure 5. Here, we can see that the salient-motion regions are robustly localized from recent temporal history. The pyramid levels are able to represent the moving objects at multiple scales. Further processing through detection and classification sub-networks results in accurate MOR.

4 EXPERIMENTS AND DISCUSSIONS

In this section, we present the training and testing data description, evaluation metrics and the baseline results computed with the proposed MOR-UAVNet framework. We give a detailed quantitative and qualitative analysis of the proposed baseline methods to report the effectiveness of our novel MOR framework.

Dataset. From the MOR-UAV dataset, we select a subset of videos for training and evaluation. The training set consists of 13 video sequences having $3,749$ frames and $52,670$ objects ($44,872$ cars and $7,798$ heavy vehicles). The testing set consists of 4 video sequences with 679 frames and $2,011$ objects ($1,940$ cars and 71 heavy vehicles). The detailed description of train and test sets used for qualitative and quantitative evaluation is given in Table 2. The proposed labels can also be used to design and evaluate class-agnostic moving object detection algorithms.

Evaluation. As the results are computed in terms of spatial coordinates and class labels of moving object instances in every frame. Thus, to measure the MOR performance, we use the standard average precision (AP) [32] metrics. Evaluation is performed with IoU threshold 0.50, i.e. predicted an object instance is considered true positive if it has at least 50% intersection-over-union (IoU) with the corresponding ground truth object instance. The mean average precision $mAP50$ computes the means of APs across two classes: car and heavy vehicle. It is to be noted that the mAP is computed only for object instances with movements. The remaining objects (non-moving) are part of the background according to the definition of MOR.

Table 3: Quantitative results (mAP) of the proposed MOR-UAVNet based baseline models over the MOR-UAV dataset. The best results are highlighted in bold

Method	C_{OF}	Vid14	Vid15	Vid16	Vid17	Avg
MOR-UAVNetv1	1-3-5	82.70	40.35	53.06	17.76	48.47
	1-3	**86.94**	32.48	85.18	29.64	**58.56**
	1-5	53.43	32.45	**91.49**	6.04	45.85
	1	85.65	56.41	81.35	4.98	57.09
MOR-UAVNetv2	1-3-5	71.12	19.68	69.46	29.79	47.51
	1-3	79.31	40.51	59.06	31.75	52.65
	1-5	83.18	38.02	80.53	30.66	58.09
	1	85.57	23.34	19.17	39.02	41.77
MOR-UAVNetv3	1-3-5	39.04	35.73	16.59	**49.14**	35.13
	1-3	60.54	25.91	05.25	17.54	27.31
	1-5	71.61	33.94	61.46	8.06	43.77
	1	79.04	44.09	72.85	19.27	53.81
MOR-UAVNetv4	1-3-5	60.90	32.65	79.52	14.72	46.95
	1-3	65.41	48.82	38.28	29.01	45.38
	1-5	80.59	41.14	62.47	19.48	50.92
	1	58.47	**57.62**	41.67	5.71	40.87

4.1 Performance Analysis

Quantitative analysis. From Table 3, it can be noticed that MOR-UAVNetv1 performs better than MOR-UAVNetv2 in vid14, vid15, and vid16. The MOR-UAVNetv2 performs better in vid17. However, MOR-UAVNetv1 outperforms MOR-UAVNetv2 in overall results. The models MOR-UAVNetv3, MOR-UAVNetv4 with mobileNetv2 backbone also perform reasonably well and can be considered for the resource-constrained environment. The best performing model achieves mAP of 58.56 which highlights the challenging nature of the MOR-UAV dataset. The mAP at different IoUs for MOR-UAVNetv1 is further analyzed through Figure 6. The IoU vs mAP graph for MOR across different C_{OF} depths for each test video is depicted in Figure 6 (a), Figure 6 (b), Figure 6(c) and Figure 6 (d) respectively. The mAP is computed at IoU thresholds $0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8$. We also show the overall average performance through the graph in Figure 6 (e). It is clear that if we could lower the IoU threshold, we can recognize a higher number of moving objects. However, it might also increase the number of false detections. The decision for the same can be taken according to the demands of real-world applications.

Qualitative analysis. We show the qualitative results of the proposed method on five completely unseen videos in Figure 7. The MOR-UAVNetv1 obtains reasonable performance in diverse objects and camera movements. For example, in the first, second and fifth rows, in addition to the object movements, the UAV camera itself is moving. Similarly, in the fourth row, the vehicles are moving nearby an industrial area with complex structures. In the third row, the moving and non-moving objects are adjacently located at some point. All these scenarios are handled quite well.

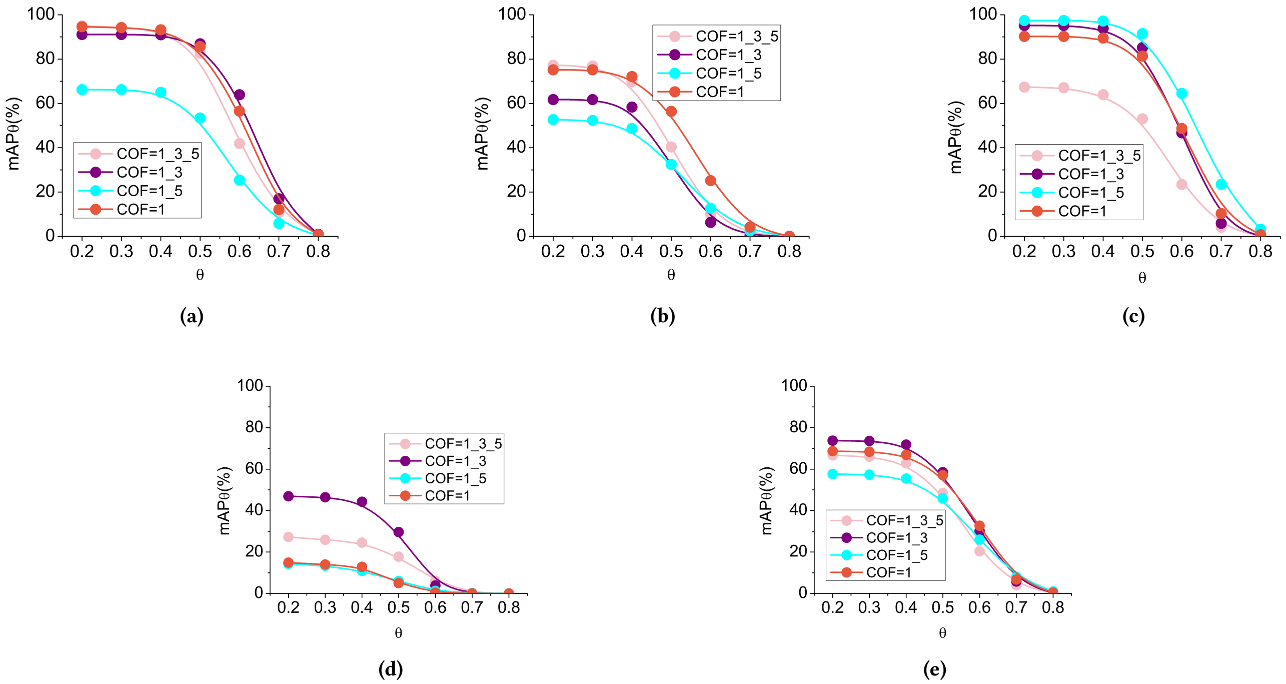


Figure 6: MOR mAP of MOR-UAVNetv1 across different IoU thresholds over (a) vid14, (b) vid15, (c) vid16, (d) vid17 videos and (e) average across all video sequences. C_{OF} is the cascaded optical flow maps used in the input layer

Table 4: Inference speed, number of parameters and model size comparison of the baseline models.

Method	C_{OF}	FPS	#Param	Model Size
MOR-UAVNetv1	1-3-5	9.59	~65.4M	~263.6MB
	1-3/1-5	9.59		
	1	11.11		
MOR-UAVNetv2	1-3-5	8.44	~36.3M	~146.3MB
	1-3/1-5	8.79		
	1	10.05		
MOR-UAVNetv3	1-3-5	7.81	~19.3M	~77.8MB
	1-3/1-5	8.79		
	1	9.59		
MOR-UAVNetv4	1-3-5	9.17	~13.2M	~53.3MB
	1-3/1-5	9.59		
	1	10.55		

Run-time analysis. We also tabulate the inference speed, compute and memory requirements of the proposed methods in Table 4. The MOR-UAVNetv4 has the lowest number of trainable parameters and model size as compared to the other three versions. It also demonstrates the fastest inference at ~11 FPS in RTX 2080Ti which is quite reasonable. Since MOR-UAVNetv1 achieves better mAP, overall, it outperforms the remaining 3 models when all the performance measures are taken into consideration. However, more work needs to be done to develop faster and more accurate MOR algorithms for real-time applications.

Failure cases. We discuss three scenarios in which MOR-UAVNet fails as shown in Figure 8. The first case (Figure 8(a)) arises due to motion in a background object which does not belong to car or heavy vehicle category. In the second case (Figure 8 (b)), the UAV camera is moving in the same direction as the objects making it difficult for the algorithm to identify the object motion accurately. In the last case (Figure 8(c)), the slowly moving objects are sometimes not detected by the proposed method. These failure cases can be overcome by including even more diversified scenarios in the training dataset and develop better algorithms for robust MOR performance in the future.

4.2 Discussions

Our benchmark caters to real-world demands with vivid samples collected from numerous unconstrained circumstances. Although the proposed MOR-UAVNet algorithms perform reasonably well on the test set, there is a lot of scopes to further improve the performance. We feel this benchmark dataset can support promising research trends in vehicular technology. We mention some of the future research directions for exploration.

Realtime challenges. Inference speed is one of the essential requirement for practical UAV based applications. The commonly deployed UAVs for aerial scene analysis are highly resource-constrained in nature. Although the proposed method achieves ~11 fps, even better speed is desired. Thus, more MOR algorithms are needed that can robustly operate with better accuracy and speed in resource-constrained environment. Some recent works [60, 65] have shown

Figure 7: Qualitative results of MOR-UAVNetv1 over completely unseen video sequences in MOR-UAV dataset.

(a) (b) (c)

Figure 8: Failure cases in MOR prediction with the MOR-UAVNetv1 model.

promising directions to increase efficiency through neural architectural search, network pruning and compression. We expect future works to develop solutions to address both accuracy and real-time constraints.

Locating motion clues. The motion appearance varies among different video sequences in the dataset. For example, variable speed (fast, medium, slow) of moving vehicles, unconstrained camera motion and rotation resulting in changing backgrounds make it difficult to accurately identify the motion clues. Moreover, the object scales also fluctuate between small, medium and large shapes due to changes in the altitude of UAV device. Thus, to obtain robust performance, these demanding factors need to be addressed in future methods.

5 CONCLUSIONS

This paper introduces a challenging unconstrained benchmark MOR-UAV dataset for the task of moving object recognition (MOR) in UAV videos. A deep learning framework MOR-UAVNet is presented along with 16 different models for baseline results. As a first largescale labeled dataset dedicated to MOR in UAV videos, the proposed work contributes to vehicular vision community by establishing a new benchmark. The proposed MOR-UAVNet framework may also be used to design new algorithms to advance MOR research in the future.

6 ACKNOWLEDGEMENTS

The authors are highly grateful to IBM for providing with online GPU grant. The work was also supported by the DST-SERB project #EEQ/2017/000673.

REFERENCES

[1] Thangarajah Akilan and QM Jonathan Wu. 2019. sEnDec: An improved image to image CNN for foreground localization. *IEEE Transactions on Intelligent Transportation Systems* (2019).

[2] Jhonghyun An, Baehoon Choi, Hyunju Kim, and Euntai Kim. 2019. A New Contour-Based Approach to Moving Object Detection and Tracking Using a Low-End Three-Dimensional Laser Scanner. *IEEE Transactions on Vehicular Technology* 68, 8 (2019), 7392–7405.

[3] Linchao Bao, Baoyuan Wu, and Wei Liu. 2018. CNN in MRF: Video object segmentation via inference in a CNN-based higher-order spatio-temporal MRF. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 5977–5986.

[4] Mohammadamin Barekatain, Miquel Martí, Hsueh-Fu Shih, Samuel Murray, Kotaro Nakayama, Yutaka Matsuo, and Helmut Prendinger. 2017. Okutama-action: An aerial view video dataset for concurrent human action detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*. 28–35.

[5] Yakoub Bazi and Farid Melgani. 2018. Convolutional SVM networks for object detection in UAV imagery. *Ieee transactions on geoscience and remote sensing* 56, 6 (2018), 3107–3118.

[6] K Berker Logoglu, Hazal Lezki, M Kerim Yucel, Ahu Ozturk, Alper Kucukkomurler, Batuhan Karagoz, Erkut Erdem, and Aykut Erdem. 2017. Feature-based efficient moving object detection for low-altitude aerial platforms. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*. 2119–2128.

[7] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. 2018. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European conference on computer vision (ECCV)*. 801–818.

[8] Jingchun Cheng, Yi-Hsuan Tsai, Shengjin Wang, and Ming-Hsuan Yang. 2017. Segflow: Joint learning for video object segmentation and optical flow. In *Proceedings of the IEEE international conference on computer vision*. 686–695.

[9] Robert Collins, Xuhui Zhou, and Seng Keat Teh. 2005. An open source tracking testbed and evaluation web site. In *IEEE International Workshop on Performance Evaluation of Tracking and Surveillance*, Vol. 35.

[10] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 248–255.

[11] Jian Ding, Nan Xue, Yang Long, Gui-Song Xia, and Qikai Lu. 2019. Learning roi transformer for oriented object detection in aerial images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2849–2858.

[12] Dawei Du, Pengfei Zhu, Longyin Wen, Xiao Bian, Haibin Ling, Qinghua Hu, Jiayu Zheng, Tao Peng, Xinyao Wang, Yue Zhang, et al. 2019. VisDrone-SOT2019: The Vision Meets Drone Single Object Tracking Challenge Results. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*. 0–0.

[13] Shivangi Dwivedi, Murari Mandal, Shekhar Yadav, and Santosh Kumar Vipparthi. 2020. 3D CNN with Localized Residual Connections for Hyperspectral Image Classification. In *Computer Vision and Image Processing: 4th International Conference, CVIP 2019, Jaipur, India, September 27–29, 2019, Revised Selected Papers, Part II 4*. Springer, 354–363.

[14] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. 2010. The pascal visual object classes (voc) challenge. *International journal of computer vision* 88, 2 (2010), 303–338.

[15] Heng Fan and Haibin Ling. 2019. Siamese cascaded region proposal networks for real-time visual tracking. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 7952–7961.

[16] Gunnar Farnebäck. 2003. Two-frame motion estimation based on polynomial expansion. In *Scandinavian conference on Image analysis*. Springer, 363–370.

[17] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. 2014. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 580–587.

[18] Yiping Gong, Zhifeng Xiao, Xiaowei Tan, Haigang Sui, Chuan Xu, Haiwang Duan, and Deren Li. 2019. Context-Aware Convolutional Neural Network for Object Detection in VHR Remote Sensing Imagery. *IEEE Transactions on Geoscience and Remote Sensing* 58, 1 (2019), 34–44.

[19] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. 2017. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*. 2961–2969.

[20] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2015. Spatial pyramid pooling in deep convolutional networks for visual recognition. *IEEE transactions on pattern analysis and machine intelligence* 37, 9 (2015), 1904–1916.

[21] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 770–778.

[22] Jan Hosang, Rodrigo Benenson, and Bernt Schiele. 2017. Learning non-maximum suppression. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 4507–4515.

[23] Meng-Ru Hsieh, Yen-Liang Lin, and Winston H Hsu. 2017. Drone-based object counting by spatially regularized regional proposal network. In *Proceedings of the IEEE International Conference on Computer Vision*. 4145–4153.

[24] Suyog Dutt Jain, Bo Xiong, and Kristen Grauman. 2017. Fusionseg: Learning to combine motion and appearance for fully automatic segmentation of generic objects in videos. In *2017 IEEE conference on computer vision and pattern recognition (CVPR)*. IEEE, 2117–2126.

[25] Yeong Jun Koh and Chang-Su Kim. 2017. Primary object segmentation in videos based on region augmentation and reduction. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 7417–7425.

[26] Tao Kong, Fuchun Sun, Anbang Yao, Huaping Liu, Ming Lu, and Yurong Chen. 2017. Ron: Reverse connection with objectness prior networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 5936–5944.

[27] Hei Law and Jia Deng. 2018. Cornernet: Detecting objects as paired keypoints. In *Proceedings of the European Conference on Computer Vision (ECCV)*. 734–750.

[28] Hazal Lezki, I Ahu Ozturk, M Akif Akpinar, M Kerim Yucel, K Berker Logoglu, Aykut Erdem, and Erkut Erdem. 2018. Joint exploitation of features and optical flow for real-time moving object detection on drones. In *Proceedings of the European Conference on Computer Vision (ECCV)*. 0–0.

[29] Jing Li, Dong Hye Ye, Timothy Chung, Mathias Kolsch, Juan Wachs, and Charles Bouman. 2016. Multi-target detection and tracking from a single camera in Unmanned Aerial Vehicles (UAVs). In *2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 4992–4997.

[30] Xi Liang, Jing Zhang, Li Zhuo, Yuzhao Li, and Qi Tian. 2019. Small Object Detection in Unmanned Aerial Vehicle Images Using Feature Fusion and Scaling-Based Single Shot Detector with Spatial Context Analysis. *IEEE Transactions on Circuits and Systems for Video Technology* (2019).

[31] Che-Tsung Lin, Shu-Ping Chen, Patrisia Sherryl Santoso, Hung-Jin Lin, and Shang-Hong Lai. 2019. Real-time Single-Stage Vehicle Detector Optimized by Multi-Stage Image-based Online Hard Example Mining. *IEEE Transactions on Vehicular Technology* (2019).

[32] T. Lin, P. Goyal, R. Girshick, K. He, and P. DollÃąr. 2020. Focal Loss for Dense Object Detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 42, 2 (Feb 2020), 318–327. https://doi.org/10.1109/TPAMI.2018.2858826

[33] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. 2017. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2117–2125.

[34] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *European conference on computer vision*. Springer, 740–755.

[35] Kang Liu and Gellert Mattyus. 2015. Fast multiclass vehicle detection on aerial images. *IEEE Geoscience and Remote Sensing Letters* 12, 9 (2015), 1938–1942.

[36] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. 2016. Ssd: Single shot multibox detector. In *European conference on computer vision*. Springer, 21–37.

[37] Wenchao Liu, Long Ma, Jue Wang, et al. 2018. Detection of multiclass objects in optical remote sensing images. *IEEE Geoscience and Remote Sensing Letters* 16, 5 (2018), 791–795.

[38] Murari Mandal, Mallika Chaudhary, Santosh Kumar Vipparthi, Subrahmanyam Murala, Anil Balaji Gonde, and Shyam Krishna Nagar. 2018. ANTIC: ANTithetic isomeric cluster patterns for medical image retrieval and change detection. *IET Computer Vision* 13, 1 (2018), 31–43.

[39] Murari Mandal, Vansh Dhar, Abhishek Mishra, and Santosh Kumar Vipparthi. 2019. 3DFR: A Swift 3D Feature Reductionist Framework for Scene Independent Change Detection. *IEEE Signal Processing Letters* 26, 12 (2019), 1882–1886.

[40] Murari Mandal, Lav Kush Kumar, Mahipal Singh Saran, and Santosh Kumar vipparthi. 2020. MotionRec: A Unified Deep Framework for Moving Object Recognition. In *The IEEE Winter Conference on Applications of Computer Vision (WACV)*.

[41] Murari Mandal, Prafulla Saxena, Santosh Kumar Vipparthi, and Subrahmanyam Murala. 2018. CANDID: Robust change dynamics and deterministic update policy for dynamic background subtraction. In *2018 24th International Conference on Pattern Recognition (ICPR)*. IEEE, 2468–2473.

[42] Murari Mandal, Manal Shah, Prashant Meena, Sanhita Devi, and Santosh Kumar Vipparthi. 2019. AVDNet: A Small-Sized Vehicle Detection Network for Aerial Visual Data. *IEEE Geoscience and Remote Sensing Letters* (2019).

[43] Murari Mandal, Manal Shah, Prashant Meena, and Santosh Kumar Vipparthi. 2019. SSSDET: Simple Short and Shallow Network for Resource Efficient Vehicle Detection in Aerial Scenes. In *2019 IEEE International Conference on Image Processing (ICIP)*. IEEE, 3098–3102.

[44] K-K Maninis, Sergi Caelles, Yuhua Chen, Jordi Pont-Tuset, Laura Leal-Taixé, Daniel Cremers, and Luc Van Gool. 2018. Video object segmentation without temporal information. *IEEE transactions on pattern analysis and machine intelligence* 41, 6 (2018), 1515–1530.

[45] Kuldeep Marotirao Biradar, Ayushi Gupta, Murari Mandal, and Santosh Kumar Vipparthi. 2019. Challenges in Time-Stamp Aware Anomaly Detection in

Traffic Videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*. 13–20.

[46] Aditya Mehta, Harsh Sinha, Pratik Narang, and Murari Mandal. 2020. HIDeGan: A Hyperspectral-Guided Image Dehazing GAN. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*. 212–213.

[47] Matthias Mueller, Neil Smith, and Bernard Ghanem. 2016. A benchmark and simulator for uav tracking. In *European conference on computer vision*. Springer, 445–461.

[48] Peter Ochs, Jitendra Malik, and Thomas Brox. 2013. Segmentation of moving objects by long term video analysis. *IEEE transactions on pattern analysis and machine intelligence* 36, 6 (2013), 1187–1200.

[49] Siyang Pan, Zhihang Tong, Yanyun Zhao, Zhicheng Zhao, Fei Su, and Bojin Zhuang. 2019. Multi-Object Tracking Hierarchically in Visual Data Taken From Drones. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*. 0–0.

[50] Anestis Papazoglou and Vittorio Ferrari. 2013. Fast object segmentation in unconstrained video. In *Proceedings of the IEEE international conference on computer vision*. 1777–1784.

[51] Jordi Pont-Tuset, Federico Perazzi, Sergi Caelles, Pablo Arbeláez, Alex Sorkine-Hornung, and Luc Van Gool. 2017. The 2017 davis challenge on video object segmentation. *arXiv preprint arXiv:1704.00675* (2017).

[52] Sebastien Razakarivony and Frederic Jurie. 2016. Vehicle detection in aerial imagery: A small target detection benchmark. *Journal of Visual Communication and Image Representation* 34 (2016), 187–203.

[53] S. Ren, K. He, R. Girshick, and J. Sun. 2017. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 39, 6 (June 2017), 1137–1149. https://doi.org/10.1109/TPAMI.2016.2577031

[54] Alexandre Robicquet, Amir Sadeghian, Alexandre Alahi, and Silvio Savarese. 2016. Learning social etiquette: Human trajectory understanding in crowded scenes. In *European conference on computer vision*. Springer, 549–565.

[55] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. 2018. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 4510–4520.

[56] Pierre Sermanet, David Eigen, Xiang Zhang, Michaël Mathieu, Rob Fergus, and Yann LeCun. 2013. Overfeat: Integrated recognition, localization and detection using convolutional networks. *arXiv preprint arXiv:1312.6229* (2013).

[57] Abhinav Shrivastava, Abhinav Gupta, and Ross Girshick. 2016. Training region-based object detectors with online hard example mining. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 761–769.

[58] Bharat Singh, Mahyar Najibi, and Larry S Davis. 2018. SNIPER: Efficient multi-scale training. In *Advances in neural information processing systems*. 9310–9320.

[59] Hongmei Song, Wenguan Wang, Sanyuan Zhao, Jianbing Shen, and Kin-Man Lam. 2018. Pyramid dilated deeper convlstm for video salient object detection. In *Proceedings of the European Conference on Computer Vision (ECCV)*. 715–731.

[60] Mingxing Tan, Bo Chen, Ruoming Pang, Vijay Vasudevan, Mark Sandler, Andrew Howard, and Quoc V Le. 2019. Mnasnet: Platform-aware neural architecture search for mobile. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2820–2828.

[61] Jack Valmadre, Luca Bertinetto, Joao Henriques, Andrea Vedaldi, and Philip HS Torr. 2017. End-to-end representation learning for correlation filter based tracking. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2805–2813.

[62] Yi Wang, Pierre-Marc Jodoin, Fatih Porikli, Janusz Konrad, Yannick Benezeth, and Prakash Ishwar. 2014. CDnet 2014: An expanded change detection benchmark dataset. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*. 387–394.

[63] Yong Wang, Wei Shi, and Shandong Wu. 2017. Robust UAV-based tracking using hybrid classifiers. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*. 2129–2137.

[64] Syed Waqas Zamir, Aditya Arora, Akshita Gupta, Salman Khan, Guolei Sun, Fahad Shahbaz Khan, Fan Zhu, Ling Shao, Gui-Song Xia, and Xiang Bai. 2019. iSAID: A Large-scale Dataset for Instance Segmentation in Aerial Images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*. 28–37.

[65] Bichen Wu, Xiaoliang Dai, Peizhao Zhang, Yanghan Wang, Fei Sun, Yiming Wu, Yuandong Tian, Peter Vajda, Yangqing Jia, and Kurt Keutzer. 2019. Fbnet: Hardware-aware efficient convnet design via differentiable neural architecture search. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 10734–10742.

[66] Xin Wu, Danfeng Hong, Jiaojiao Tian, Jocelyn Chanussot, Wei Li, and Ran Tao. 2019. ORSIm Detector: A novel object detection framework in optical remote sensing imagery using spatial-frequency channel features. *IEEE Transactions on Geoscience and Remote Sensing* 57, 7 (2019), 5146–5158.

[67] Zhenyu Wu, Karthik Suresh, Priya Narayanan, Hongyu Xu, Heesung Kwon, and Zhangyang Wang. 2019. Delving into robust object detection from unmanned aerial vehicles: A deep nuisance disentanglement approach. In *Proceedings of the IEEE International Conference on Computer Vision*. 1201–1210.

[68] Gui-Song Xia, Xiang Bai, Jian Ding, Zhen Zhu, Serge Belongie, Jiebo Luo, Mihai Datcu, Marcello Pelillo, and Liangpei Zhang. 2018. DOTA: A large-scale dataset for object detection in aerial images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 3974–3983.

[69] Ning Xu, Linjie Yang, Yuchen Fan, Dingcheng Yue, Yuchen Liang, Jianchao Yang, and Thomas Huang. 2018. Youtube-vos: A large-scale video object segmentation benchmark. *arXiv preprint arXiv:1809.03327* (2018).

[70] Xiaowei Xu, Xinyi Zhang, Bei Yu, X Sharon Hu, Christopher Rowen, Jingtong Hu, and Yiyu Shi. 2019. Dac-sdc low power object detection challenge for uav applications. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2019).

[71] Yanchao Yang, Antonio Loquercio, Davide Scaramuzza, and Stefano Soatto. 2019. Unsupervised moving object detection via contextual information separation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 879–888.

[72] Hongyang Yu, Guorong Li, Weigang Zhang, Qingming Huang, Dawei Du, Qi Tian, and Nicu Sebe. 2019. The Unmanned Aerial Vehicle Benchmark: Object Detection, Tracking and Baseline. *International Journal of Computer Vision* (2019), 1–19.

[73] Hui Zhang, Kunfeng Wang, Yonglin Tian, Chao Gou, and Fei-Yue Wang. 2018. MFR-CNN: Incorporating multi-scale features and global information for traffic object detection. *IEEE Transactions on Vehicular Technology* 67, 9 (2018), 8019–8030.

[74] Pengfei Zhu, Dawei Du, Longyin Wen, Xiao Bian, Haibin Ling, Qinghua Hu, Tao Peng, Jiayu Zheng, Xinyao Wang, Yue Zhang, et al. 2019. VisDrone-VID2019: The Vision Meets Drone Object Detection in Video Challenge Results. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*. 0–0.

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