Drone Based RGBT Vehicle Detection and Counting: A Challenge

Pengfei Zhu, Yiming Sun, Longyin Wen, Yu Feng, Qinghua Hu

***Abstract*—Camera-equipped drones can capture targets on the ground from a wider field of view than static cameras or moving sensors over the ground. In this paper we present a large-scale vehicle detection and counting benchmark, named DroneVehicle, aiming at advancing visual analysis tasks on the drone platform. The images in the benchmark were captured over various urban areas, which include different types of urban roads, residential areas, parking lots, highways, etc., from day to night. Specifically, DroneVehicle consists of 15,532 pairs of images, *i.e.*, RGB images and infrared images with rich annotations, including oriented object bounding boxes, object categories, etc. With intensive amount of effort, our benchmark has 441,642 annotated instances in 31,064 images. As a large-scale dataset with both RGB and thermal infrared (RGBT) images, the benchmark enables extensive evaluation and investigation of visual analysis algorithms on the drone platform. In particular, we design two popular tasks with the benchmark, including object detection and object counting. All these tasks are extremely challenging in the proposed dataset due to factors such as illumination, occlusion, and scale variations. We hope the benchmark largely boost the research and development in visual analysis on drone platforms. The DroneVehicle dataset can be download from https://github.com/VisDrone/DroneVehicle.**

***Index Terms*—Drone, Vehicle Detection, Vehicle Counting, RGBT.**

I. Introduction

Due to the wide application of computer vision and the latest breakthroughs in many important problems, computer vision has attracted more and more attention in recent years. As two core problems in computer vision, object detection and object counting are under extensive investigation in both academia and real world applications, $e.g.$, transportation surveillance, and smart city. Among many factors and efforts that have led to the rapid development of computer vision technology, a noteworthy contribution should be attributed to the invention or organization of numerous benchmarks. For example, in the field of object detection, representative benchmarks are Pascal VOC [1], KITTI [2], ImageNet [3], and MS COCO [4]. And in the field of object counting, typical datasets are UCSD [5], Mall [6], CBSR [7], NWPU-Crowd [8], and Shanghaitech [9].

Object detection and counting based on drones is becoming increasingly important in various intelligent applications. Therefore, drones (or UAVs) equipped with cameras have been widely used in various industries, but how to intelligently analyze and understand the visual data collected from these platforms is becoming a concern for researchers. In the current field of computer vision, with the development of deep learning, the related algorithms of object detection and object counting have made great progress. However, due to various challenges in the images collected by the drone (such as different lighting, occlusion and different object scales), the existing algorithms cannot usually process the images acquired by the drone very well. Consequently, developing and evaluating new vision algorithms for drone generated visual data is a key problem in drone-based applications [10]. It is gratifying that some valuable work [11], [12], [10], [13] has been proposed in recent years. They are committed to building datasets or benchmarks focused on object detection, object tracking, and object counting through drone platforms, which has strongly promoted the research of computer vision technology on drone platforms. However, due to the limitations of the hardware platform and the difficulty of data collection, these datasets are mostly single-modal images, or the annotation without using oriented bounding boxes, which have limited the comprehensive evaluation of computer vision algorithms on the datasets collected by drones. Therefore, a more general and comprehensive benchmark is needed to further boosting visual analysis research on drone platforms.

Thus motivated, we present a multi-modal benchmark, named DroneVehicle, with carefully annotated groundtruth for various important computer vision tasks. The benchmark dataset consists of 15,532 pairs of images, *i.e.*, RGB images and infrared images, captured by drone-mounted dual cameras, covering a wide range of aspects including Scenarios (different types of urban roads, residential areas, parking lots, highways, etc. from Tianjin, China), objects (car, bus, truck, van, feright car, etc.), and density (sparse and crowded scenes), etc. With thorough annotations of 441,642 object instances, the benchmark focuses on two tasks:

- **Task 1: object detection in images.** Given a predefined set of object classes ($e.g.$, car, bus, and truck), the task aims to detect objects of these classes from individual images taken from drones.
- **Task 2: object counting in images.** The task aims to estimate the number of vehicles from individual images in DroneVehicle.

In this challenge we select five categories of objects of frequent interests in drone applications, such as car, bus, truck, van, and feright car. Altogether we carefully annotated 441,642 oriented bounding boxes of object instances from these categories. The detailed comparison of the provided drone datasets with other related benchmark datasets in object detection and object counting are presented in Table. I and Table.II.

II. Related Work

With the development of computer vision technology, for the evaluation of typical computer vision algorithms such

Fig. 1: Some example images of the DroneVehicle dataset.

as object detection and object counting, many representative benchmarks [1], [2], [8], [9] have been proposed, which has effectively promoted the progress of computer vision research. In this section, we review the most relevant drone-based benchmarks and other benchmarks in object detection and object counting fields.

A. Drone based Datasets

In recent years, some drone-based datasets have been proposed in computer vision field. Mundhenk $etal.$ [14] proposed a dataset named COWC (The Cars Overhead With Context) collected by drones, which includes 32,716 unique annotated cars and 58,247 unique negative examples. Besides, a baseline is established for car detection and counting tasks. Hsieh $etal.$ [11] present a dataset named CARPK for detection and counting tasks, which consists of 1,448 images and contains 89,777 annotated cars captured by the drone platform from different parking lots. In [12], a dataset named DOTA based on aerial images was proposed, which includes 16 different categories and contains 0.4 million annotated object instances. The latest DOTA-v1.5 also adds many annotations to the small object instances about or below 10 pixels that were missed previously. Li $etal.$ [15] proposed a benchmark dataset of high diversity, consisting of 70 videos captured by drone cameras. Among them, the annotation of the dataset was completed manually for the evaluation of object tracking. Zhu $etal.$ [10] proposed a a large scale benchmark, named VisDrone2018, which consists of 263 video clips and 10,209 images (no overlap with video clips) with rich annotations, including object bounding boxes, object categories, occlusion, truncation ratios, $etc.$ VisDrone2018 has more than 2.5 million annotated instances in 179,264 images/video frames. Mueller$etal.$ [16] proposed a high-resolution dataset named UAV123. It consists of 123 aerial video sequences and contains $110k$ ($1k = 1,000$) annotated frames, and the annotation includes the bounding boxes of person and their corresponding action labels.

In contrast to the above-mentioned datasets for object detection or object counting, which is only obtained in the visible light scene, the DroneVehicle dataset includes both RGB images and infrared images, collected in various typical urban environments, focusing on two core problems in computer vision fields, $i.e.$, object detection and object counting.

B. Object Detection Datasets

Object detection is a very important task in computer vision, and many excellent benchmarks or datasets for evaluation of this task are established. Caltech [17] is a traffic scenarios dataset collected by cameras installed on the cars in the urban environment. It consists of approximately 10 hours of video, and 250,000 frames with a total of 350,000 annotated bounding boxes of 2,300 unique pedestrians. KITTI [2] is a well-known benchmark used in autonomous driving scenarios, which is designed to evaluate environment perception algorithms, and contains 7,481 training and 7,518 testing images. UA-DETRAC [19] is a large-scale dataset for vehicle detection and tracking. It is mainly shot on road crossing bridges in Beijing and Tianjin, China. It has been manually labeled with 8,250 vehicles and 1.12 million object labels. The PASCAL VOC [1] is a benchmark for object classification and detection. As a well-known benchmark, from 2005 to 2012, a challenge named PASCAL VOC was held every year to attract many relevant researchers from all over the world to participate,

TABLE I: Comparison of the state-of-the-art benchmarks and datasets. Note that, the resolution indicates the maximum resolution of videos/images included in the benchmarks and datasets, and the BB is short for bounding box. ($1k = 1,000$)

Image object detection	Scenario	Modality	#Images	Categories	Avg. #labels/categories	Resolution	Oriented BB	Year
Caltech Pedestrian [17]	driving	RGB	$249k$	1	$347k$	640×480		2012
KITTI [2]	driving	RGB	$15.4k$	2	$80k$	1241×376		2012
PASCAL VOC2012 [1]	life	RGB	$22.5k$	20	$1,373$	469×387		2012
ImageNet [3]	life	RGB	$456.2k$	200	$2,007$	482×415		2013
MS COCO [4]	life	RGB	$328.0k$	91	$27.5k$	640×640		2014
VEDAI [18]	satellite	RGB	$1.2k$	9	733	1024×1024	√	2015
UA-DETRAC [19]	surveillance	RGB	$140.1k$	4	$302.5k$	960×540		2015
COWC [14]	aerial	RGB	$32.7k$	1	$32.7k$	2048×2048		2016
CARPK [11]	drone	RGB	$1,448$	1	$89.8k$	1280×720		2017
DOTA [12]	aerial	RGB	$2,806$	14	$13.4k$	12029×5014	√	2018
UAVDT [13]	drone	RGB	$80k$	3	$280.5k$	1080×540		2018
VisDrone [10]	drone	RGB	$10,209$	10	$54.2k$	2000×1500		2018
DroneVehicle(ours)	drone	RGB + Infrared	$31,064$	5	$88.3k$	840×712	√	2020

TABLE II: Comparison of existing crowd counting datasets. "–" indicates different resolutions in the dataset.

Dataset	Object	Modality	Resolution	Frames	Max	Min	Ave	Total	Year
UCSD [5]	people	RGB	158×238	$2,000$	46	11	24.9	$49,885$	2008
Mall [6]	people	RGB	640×480	$2,000$	53	13	31.2	$62,315$	2012
CBSR Dataset1 [7]	people	Depth	240×320	$2,834$	7	0	1.6	$4,541$	2012
CBSR Dataset2 [7]	people	RGB + Depth	240×320	$1,500$	7	0	1	$1,553$	2012
UCF_CC_50 [20]	people	RGB	-	50	$4,543$	94	$1,279.5$	$63,974$	2013
MICC [21]	people	RGB + Depth	480×640	$3,358$	11	0	5.32	$17,630$	2014
WorldExpo2010 [22]	people	RGB	576×720	$3,980$	253	1	50.2	$199,923$	2016
Shanghaitech A [9]	people	RGB	589×868	482	$3,139$	33	501.4	$241,677$	2016
Shanghaitech B [9]	people	RGB	768×1024	716	578	9	123.6	$88,488$	2016
CARPK [11]	vehicle	RGB	1280×720	$1,448$	188	1	62.0	$89,777$	2017
UCF-QNRF [23]	people	RGB	2013×2902	$1,535$	$12,865$	49	815.4	$1,251,642$	2018
SmartCity [24]	people	RGB	1920×1080	50	14	1	7.4	369	2018
FDST [25]	people	RGB	1080×1920	$15,000$	57	9	26.7	$394,081$	2019
GCC [26]	people	RGB	1080×1920	$15,212$	$3,995$	0	501	$7,625,843$	2019
Crowd Surveillance [27]	people	RGB	840×1342	$13,945$	$1,420$	2	35	$386,513$	2019
DLR-ACD [28]	people	RGB	3619×5226	33	$24,368$	285	$6,857$	$226,291$	2019
ShanghaiTechRGBD [29]	people	RGB + Depth	1920×1080	$33,600$	455	25	144.8	$4,864,280$	2019
DroneCrowd [30]	people	RGB	1920×1080	$33,600$	455	25	144.8	$4,864,280$	2019
NWPU-Crowd [8]	people	RGB	2311×3383	5109	$20,033$	0	418	$2,133,238$	2020
DroneVehicle(ours)	vehicle	RGB + Infrared	840×712	$31,064$	206	0	14.2	$441,642$	2020

and a new datasets was also released. For example, VOC 2012 contains 11,540 images and annotation files for a total of 20 objects, providing a standard image annotation dataset and a standard evaluation system for detecting algorithm performance. ImageNet [3] is an dataset organized according to the WordNet hierarchy. WordNet contains about 100,000 phrases, and ImageNet provides an average of about 1,000 description images for each phrase. The total number of images in the ImageNet dataset is approximately 1.5 million, and each image has multiple bounding boxes and individual category labels. MS COCO dataset [4] is a large dataset mainly used for object detection and semantic segmentation tasks, which contains more than 328,000 images with 2.5 million manually segmented object instances. It has 91 object categories with $27k$ instances on average per category. UAVDT [13] is a dataset used for object detection and object tracking. It mainly contains about 80,000 representative frames from 10 hours raw videos, and annotated 14 kinds of attributes ($e.g.$, weather condition, fying altitude, camera view, vehicle category, and occlusion) with bounding boxes.

C. *Object Counting Datasets*

From the perspective of city management and security, it is very important to automatically count the object in the image based on advanced computer vision algorithms. The Mall [6] is a dataset for crowd counting and profiling research. Its images are collected from publicly accessible webcam. It mainly includes 2,000 video frames, and the head position of every pedestrian in all frames is annotated. A total of more than 60,000 pedestrians are annotated in this dataset. WorldExpo'10 Crowd Counting Dataset [22] is derived from 108 surveillance cameras from Shanghai 2010 WorldExpo. Through various cameras with different fields of view, a total of 1,132 annotated video sequences were collected. It covers a variety of scenarios and contains a total of 199,923 annotations. UCF-QNRF [23] is a dataset for training and evaluating crowd counting and localization. It contains 1,535 images with 1,251,642 annotations. It is collected from the web, and images in the dataset come from all parts of the world. FDST dataset [25] collected 100 videos captured from 13 different scenes, and FDST dataset contains 150,000 frames, with a total of 394,081 annotated heads. The NWPU-Crowd Dataset [8] is a crowd counting dataset that contains 5,109 images crawled from the Internet, with a total of 2,133,238 annotated instances.

However, the above object counting datasets are mostly used for crowd counting, and they are all RGB images. The scenes are mostly daylight. The object counting task in the DroneVehicle dataset is based on vehicles, which includes RGB images and thermal infrared images. At the same time, the distribution of object in the DroneVehicle dataset is relatively diverse. And the DroneVehicle dataset contains images of both day and night scenes. Hence, our DroneVehicle dataset poses a higher challenge to the object counting algorithm.

Fig. 2: Some example annotated images of the DroneVehicle dataset.

III. DroneVehicle Dataset

A. Data Collection

The DroneVehicle dataset is collected by the Lab of Machine Learning and Data Mining , Tianjin University, China. The benchmark dataset consists of 31,064 images, including 15,532 RGB images and 15,532 infrared images, captured by drone-mounted dual cameras, covering a wide range of aspects including scenarios (different types of urban roads, residential areas, parking lots, highways, etc.), objects (car, bus, truck, van, feright car, etc.), and density (sparse and crowded scenes). Note that, the dataset was collected using drone equipped with dual cameras (visible camera and thermal infrared camera), in different scenarios, and under various weather and lighting conditions. These frames are manually annotated with 441642 bounding boxes of targets of vehicles.

B. Data Pruning and Annotation

1) Data Pruning: Data pruning is an important step in making a dataset, so the raw data collected by the drone needs to be pre-processed. A part of the images with poor imaging quality are discarded, including images with blurred imaging and target ghosting. Then we manually check all the image data and uniformly convert the resolution of the image to 840×712. Finally we get the data without annotation. Some example images are shown in Fig.1.

2) Data Annotation: In computer vision, the typical method of annotation mainly uses rectangular bounding boxes to annotate the objects on an image. The bounding boxes are annotated with (x_c, y_c, w, h), where (x_c, y_c) is the center location, and w, h are the width and height of the bounding box. Of course, this method of annotation is sufficient for many scenarios, such as autonomous driving scenarios, traffic surveillance scenarios, *etc.*. However, for the aerial images based on drones, this annotation manner may bring certain inaccuracies. Because objects in the aerial images often have different orientations, in order to more accurately and compactly represent the outline of the object, we need to use the oriented bounding box for annotation.

As mentioned in [12], θ-based oriented bounding box is used by some text detection benchmarks, and that is (x_c, y_c, w, h, θ), where θ denotes the angle from the horizontal

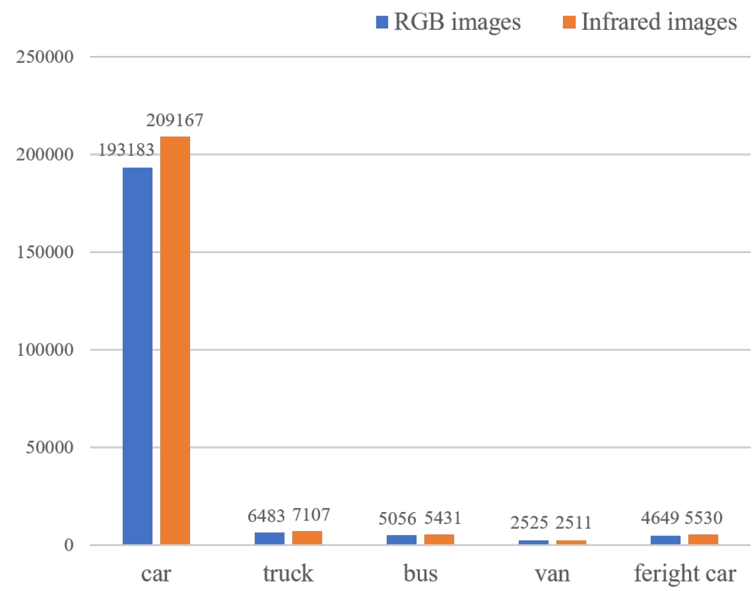

Fig. 3: The distribution of categories of both RGB and Infrared modality in the DroneVehicle dataset.

direction of the standard bounding box. But the disadvantage of this method is that it cannot compactly enclose the oriented objects with large deformation among different parts. Note that the main objects are several different kinds of vehicles. Therefore, due to the diversity and complexity of the scenarios involved in the DroneVehicle dataset, we choose a method for annotating oriented objects which named four-point bounding boxes. In details, this method of annotation can be denoted as $\{(x_i, y_i), i = 1, 2, 3, 4\}$, where (x_i, y_i) denotes the positions of the oriented bounding boxes vertices in the image. Some samples of annotated images in our dataset are shown in Fig.2.

C. Statistics and Attributes

Five categories are chosen and annotated in our DroneVehicle dataset, including car, truck, bus, van and freight car. The distribution of these five categories of both RGB and Infrared modality in the DroneVehicle dataset is shown in the Fig.3.

The DroneVehicle dataset consists of a total of 31,064 images collected by the drone, including 15,532 RBG images, and 15,532 infrared images. We have made rich annotations with oriented bounding boxes for the five categories. Among them, the category of car has 193183 annotations in RGB images, 209167 annotations in infrared images. The category of truck has 6483 annotations in RGB images, and 7107 annotations in thermal infrared images. The category of bus has 5056 annotations in RGB images, and 5431 annotations in thermal infrared images. The category of van has 2525 annotations in RGB images, and 2511 annotations in thermal infrared images. The category of freight car has 4649 annotations in RGB images, and 5530 annotations in thermal infrared image.

It can be seen that most thermal infrared images have more annotations than RGB images. The main reason is that there are many night scenes in our dataset. In this scenario, many vehicle targets in the RGB image are difficult for human eyes to distinguish, so it can't be labeled. However, in the thermal infrared image, various targets can still be displayed very well, which also shows that in various practical applications based on computer vision technology, thermal infrared images are often a good complement to vision understanding tasks in RGB images. On the one hand, our DroneVehicle dataset contains a large number of RGB images and their corresponding thermal infrared images, which is a very advanced benchmark. On the other hand, for the object counting task, the average number of vehicle targets per image in our DroneVehicle dataset is 14.2, of which the maximum number of vehicle targets per image is 206.

IV. Conclusions

We introduce a new drone based RGBT vehicle detection and counting benchmark, DroneVehicle, to facilitate the research of object detection and counting on the drone platform. Notably, the dataset is recorded over different urban areas with visible-thermal infrared cameras equipped drones. DroneVehicle covers different types of urban roads, residential areas, parking lots, highways, *etc.* from day to night. We provide a rich set of annotations including more than 441,642 annotated object instances along with several important attributes. The DroneVehicle benchmark will be made available to the research community. We expect the benchmark to largely boost the research and development in visual analysis on drone platforms.

References

[1] M. Everingham, S. M. A. Eslami, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes challenge: A retrospective," *International Journal of Computer Vision*, vol. 111, no. 1, pp. 98–136.

[2] A. Geiger, "Are we ready for autonomous driving? the kitti vision benchmark suite," in *Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012.

[3] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, and M. Bernstein, "Imagenet large scale visual recognition challenge," *International Journal of Computer Vision*, vol. 115, no. 3, pp. 211–252.

[4] T. Y. Lin, M. Maire, S. Belongie, L. Bourdev, R. Girshick, J. Hays, P. Perona, D. Ramanan, C. L. Zitnick, and P. Dollr, "Microsoft coco: Common objects in context," 2014.

[5] A. B. Chan, Z.-S. J. Liang, and N. Vasconcelos, "Privacy preserving crowd monitoring: Counting people without people models or tracking," in *2008 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2008, pp. 1–7.

[6] K. Chen, C. C. Loy, S. Gong, and T. Xiang, "Feature mining for localised crowd counting." in *BMVC*, vol. 1, no. 2, 2012, p. 3.

[7] X. Zhang, J. Yan, S. Feng, Z. Lei, D. Yi, and S. Z. Li, "Water filling: Unsupervised people counting via vertical kinect sensor," in *2012 IEEE ninth international conference on advanced video and signal-based surveillance*. IEEE, 2012, pp. 215–220.

[8] Q. Wang, J. Gao, W. Lin, and X. Li, "Nwpu-crowd: A large-scale benchmark for crowd counting," *arXiv preprint arXiv:2001.03360*, 2020.

[9] Y. Zhang, D. Zhou, S. Chen, S. Gao, and Y. Ma, "Single-image crowd counting via multi-column convolutional neural network," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 589–597.

[10] P. Zhu, L. Wen, X. Bian, H. Ling, and Q. Hu, "Vision meets drones: A challenge," *arXiv preprint arXiv:1804.07437*, 2018.

[11] M. R. Hsieh, Y. L. Lin, and W. H. Hsu, "Drone-based object counting by spatially regularized regional proposal network," in *2017 IEEE International Conference on Computer Vision (ICCV)*, 2017.

[12] G.-S. Xia, X. Bai, J. Ding, Z. Zhu, S. Belongie, J. Luo, M. Datcu, M. Pelillo, and L. Zhang, "Dota: A large-scale dataset for object detection in aerial images," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[13] D. Du, Y. Qi, H. Yu, Y. Yang, K. Duan, G. Li, W. Zhang, Q. Huang, and Q. Tian, "The unmanned aerial vehicle benchmark: Object detection and tracking," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 370–386.

[14] T. N. Mundhenk, G. Konjevod, W. A. Sakla, and K. Boakye, "A large contextual dataset for classification, detection and counting of cars with deep learning," in *European Conference on Computer Vision*, 2016.

[15] S. Li and D.-Y. Yeung, "Visual object tracking for unmanned aerial vehicles: A benchmark and new motion models," in *Thirty-First AAAI Conference on Artificial Intelligence*, 2017.

[16] M. Mueller, N. Smith, and B. Ghanem, "A benchmark and simulator for uav tracking," in *European conference on computer vision*. Springer, 2016, pp. 445–461.

[17] C. Wojek, P. Dollar, B. Schiele, and P. Perona, "Pedestrian detection: An evaluation of the state of the art," *IEEE Transactions on Pattern Analysis & Machine Intelligence*, vol. 34, no. 4, pp. 743–761, 2012.

[18] S. Razakarivony and F. Jurie, "Vehicle detection in aerial imagery : A small target detection benchmark," *Journal of Visual Communication & Image Representation*, vol. 34, no. Jan., pp. 187–203.

[19] L. Wen, D. Du, Z. Cai, Z. Lei, M. C. Chang, H. Qi, J. Lim, M. H. Yang, and S. Lyu, "Ua-detrac: A new benchmark and protocol for multi-object detection and tracking," *Computer Science*, 2015.

[20] H. Idrees, I. Saleemi, C. Seibert, and M. Shah, "Multi-source multi-scale counting in extremely dense crowd images," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2013, pp. 2547–2554.

[21] E. Bondi, L. Seidenari, A. D. Bagdanov, and A. Del Bimbo, "Real-time people counting from depth imagery of crowded environments," in *2014 11th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*. IEEE, 2014, pp. 337–342.

[22] C. Zhang, K. Kang, H. Li, X. Wang, R. Xie, and X. Yang, "Data-driven crowd understanding: A baseline for a large-scale crowd dataset," *IEEE Transactions on Multimedia*, vol. 18, no. 6, pp. 1048–1061, 2016.

[23] H. Idrees, M. Tayyab, K. Athrey, D. Zhang, S. Al-Maadeed, N. Rajpoot, and M. Shah, "Composition loss for counting, density map estimation and localization in dense crowds," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 532–546.

[24] L. Zhang, M. Shi, and Q. Chen, "Crowd counting via scale-adaptive convolutional neural network," in *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 2018, pp. 1113–1121.

[25] Y. Fang, B. Zhan, W. Cai, S. Gao, and B. Hu, "Locality-constrained spatial transformer network for video crowd counting," in *2019 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, 2019, pp. 814–819.

[26] Q. Wang, J. Gao, W. Lin, and Y. Yuan, "Learning from synthetic data for crowd counting in the wild," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2019, pp. 8198–8207.

[27] Z. Yan, Y. Yuan, W. Zuo, X. Tan, Y. Wang, S. Wen, and E. Ding, "Perspective-guided convolution networks for crowd counting," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 952–961.

[28] R. Bahmanyar, E. Vig, and P. Reinartz, "Mrcnet: Crowd counting and density map estimation in aerial and ground imagery," *arXiv preprint arXiv:1909.12743*, 2019.

[29] D. Lian, J. Li, J. Zheng, W. Luo, and S. Gao, "Density map regression guided detection network for rgb-d crowd counting and localization," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 1821–1830.

[30] L. Wen, P. Zhu, D. Du, X. Bian, H. Ling, Q. Hu, J. Zheng, T. Peng, X. Wang, Y. Zhang *et al.*, "Visdrone-mot2019: The vision meets drone multiple object tracking challenge results," in *Proceedings of the IEEE International Conference on Computer Vision Workshops*, 2019, pp. 0–0.