BirdsEyeView: Aerial View Dataset for Object Classification and Detection

Yunlong Qi
Dept. of Electrical Engineering
University of North Texas
Denton, TX, 76207
yunlongqi@my.unt.edu

Dong Wang
Google Cloud
Google Inc.
New York, NY, 10011
wdzc@google.com

Junfei Xie
Dept. of Computing Sciences
Texas A&M University-Corpus Christi
Corpus Christi, TX, 78412
junfei.xie@tamucc.edu

Kejie Lu
Dept. of Computer Science and Engineering
University of Puerto Rico at Mayagüez
Mayagüez, Puerto Rico, 00681
kejie.lu@upr.edu

Yan Wan
Dept. of Electrical Engineering
University of Texas at Arlington
Arlington, TX, 76010
yan.wan@uta.edu

Shengli Fu
Dept. of Electrical Engineering
University of North Texas
Denton, TX, 76207
shengli.fu@unt.edu

***Abstract*—In recent years, deep learning based computer vision technology has progressed rapidly thanks to the significant increases in computing power and high-quality datasets. In this article, we present an aerial view image and video dataset dedicated to facilitating vision applications on the drone platform, such as object detection, classification and tracking. The dataset consists of 5,000 images, each of which is carefully annotated according to the guidelines of the PASCAL VOC. The dataset is designed to cover diverse real-life scenes with aerial view angles which is different from other datasets. Such kind of specific dataset will be of great importance in developing and testing deep learning algorithms for UAV applications. Moreover, the dataset can serve as a benchmark to evaluate UAV visual solutions.**

***Index Terms*—drone, image dataset, benchmark, object detection, object tracking**

I. Introduction

In the past few years, deep learning based computer vision technology has attracted tremendous attention due to its state-of-the-art performance on a wide range of visual applications. Both academia and industry have made significant progress in several core techniques such as object detection and object tracking. From these techniques various applications could be derived such as surveillance, resource monitoring, and wilderness search and rescue.

Hardware platforms designed specifically for deep learning, such as GPU and TPU, make it practical for deploying powerful and sophisticated deep learning models in real-world applications. For example, faster RCNN [1], YOLO [2] and SSD [3] are used for object detection and classification, while generative adversarial networks [4] and cascaded refinement networks [5] are used for image synthesis. The implementation of all these models is based on powerful computing power.

In addition to the powerful computing hardware platform, the high-quality datasets also play an important role in driving the development of these models. Many organizations, as pioneers in the field of artificial intelligence, have made a significant contribution to the development of high-quality datasets for computer vision applications. ImageNet [6], Pascal VOC [7] and MS COCO [8] provide the cornerstone for the preliminary progress of recognition algorithms. Supported by other specific datasets and benchmarks, such as VIVID [9], OTB [10], MOT Challeng [11] and so on, tracking algorithms have achieved great success in recent years. However, most of the currently available visual datasets are collected from ground view angles, and therefore, are not suitable for aerial applications, which is a field of great potential.

As Unmanned Aerial Vehicles (UAVs) become mature and affordable, they have been used in many applications. Equipped with WiFi devices, aerial networks can be quickly deployed in certain emergency scenarios. Furthermore, if directional antennas are installed, airborne WiFi network can provide more reliable and larger coverage that is of great importance in disaster relief practices [12] [13]. With the support of the UAV system, a high-performance UAV system was studied to make the most efficient use of limited computing resources to achieve computation extensive tasks, such as positioning biometric objects and outdoor casualty searches [14]. Equipped with cameras and integrated with computer vision algorithms, UAVs have tremendous potential in real disaster relief applications. Therefore there is great demand for aerial view visual datasets that can be used for developing visual applications on UAVs. There are some efforts [15], [16] that have been devoted to constructing datasets focusing on object detection or tracking on drone platform. However, the availability of large-scale and high-quality aerial visual datasets are very limited due to the limitations of flying drones in public areas and the difficulty in aerial data collection and annotation.

In this paper, we present a large-scale aerial visual benchmark with carefully annotated that can be used for various

This work was partially supported by National Science Foundation (NSF) under Grants CI-1730589/1730675/1730570/1730325 and CAREER-1714519.

kinds of computer vision tasks on drone platform. The dataset consists of images from different sources and with different view angles. It covers many real-life scenes, such as parking lots, street views, social parties, travelling and so on. Furthermore, the image annotation files are saved as XML files in the same way as ImageNet and Pascal VOC do, which is a standard and more representative way of data annotation. This dataset can be used to develop, optimize, and validate object detection and tracking algorithms for drone applications.

II. Related Work

A number of benchmarks and datasets have been created that lay a foundation for computer vision algorithm development. Some well-known datasets, such as PASCAL VOC [7], ImageNet [6], and MS COCO [8], are used for general object classification and detection. There are also datasets created for target tracking, such as [10], [17] for single object tracking and [11], [18] multiple object tracking. There are also dedicated datasets, such as Cityscapes Dataset [19], including Pascal VOC [7], for semantic and segmentation analysis, a very popular area of computer vision.

A. Visual Datasets for General Purpose Computer Vision

Several large image benchmarks have been created for object classification and detection. The PASCAL VOC [7] provides a competition platform since 2005 for object recognition and includes the competitions for classification, detection and segmentation until now. It provides a large image dataset with 20 classes and annotated on 11,530 images. Furthermore, it provides a standardized evaluation for recognition algorithms. The ImageNet [6] is also a well-known benchmark in object classification and detection and it starts from 2010 and runs annually. It follows the PASCAL VOC but greatly expands the number of classes and images. It also provides a way to track progress and learn from innovative models. The Microsoft COCO [8] is another widely used visual recognition dataset which is also designed for studying object recognition. It focuses on natural scenes in daily life and 2.5 million instances have been labelled on 328k images in this dataset. These databases have been widely used in the field of deep learning for object recognition and spurred the emergence of some well-known deep learning models.

Except for these well-known datasets or benchmarks for general purposes, there also have some important dataset used in some specific areas. Enzweiler and Gavrila [20] provide a survey and experiments on pedestrian detection, which is a hot and rapidly evolving branch in the computer vision and has great potential in recent applications, such as auto driving vehicles and advanced robotics. They have created a large-scale dataset with 15,660 training examples obtained from 3,915 rectangular positions by means of mirroring and randomly shifting and comprising 21,790 images for testing. Piotr et al [21] present the Caltech Pedestrian Dataset, which has a larger scale than previous existing datasets, for pedestrian detection. Approximately 10 hours video has been taken from a driving vehicle and 350,000 bounding boxes have been labelled on 250,000 frames. And they also provides an improved evaluation metrics.

As for object tracking, there are also datasets created for developing and testing tracking algorithms. In [9], a testbed is provided for evaluating the performance of tracking algorithms. On this evaluation website, ground-truth datasets are provided for tracking experiments and corresponding testbed software is also provided. In [22], 26,500 labelled frames have been extracted from 28 video sequences following the representation model of CAVIAR. These frames are classified into 6 activity scenaria, and for a bounding box and related description are provided for each individual person. In [23], a online benchmark for object tracking is provided. In order to evaluate different tracking algorithms, they created a uniform and representative dataset which contains 50 fully annotated sequences. And they also created a code library including 29 tracking algorithms for performance comparison. Visual Object Tracking challenge 2015 (VOT2015) [17] provides a testing platform for short-term visual trackers, and 62 trackers have been tested using this benchmark. Compared to the previous version, VOT2015's data set is twice that of VOT2014 and introduces new performance testing method. In [24], they focus on tracking model for deformation and occlusion and provide an evaluation dataset for deformable object tracking. In [25], they focus more on the real-time performance of trackers, therefore a video dataset of high frame rate and extensive evaluation were provided for this purpose. In [26], they studied tracking algorithms with addition information, depth. Therefore, a dataset of RGBD videos has been created to compare performance of different tracking algorithms with RGB and RGBD input.

Even though lots of datasets have been created for visual tasks, all of these datasets are from a normal perspective, which limited the application and research area. There is demand for datasets for the aerial perspective.

B. Drone-based Datasets

With the proliferation of drone based applications and the popularity of deep learning algorithms, there is a great need for aerial visual datasets that can be used for computer vision algorithm development. However, there are very limited drone based datasets available in the field of computer vision. Robicquet et al. [16] studied the impact of social common sense rules on trajectory prediction and provided a new multi-object dataset containing various goals. Hsieh et al. [27] provides a method to count and localize objects simultaneously. Correspondingly, a large-scale dataset of parking lots has been created to evaluate their methods and nearly 90,000 cars from aerial view have been recorded for counting. In [15], they collected a video dataset from aerial view for target tracking. 123 video sequences from aerial perspective are captured and annotated. And they tried to make this a benchmark and compared different trackers. In [28], the authors focused on a motion model which is less explored compared to observation model. Their algorithm is used for camera motion estimation and a corresponding benchmark dataset that contains 70 videos from the aerial view was created and used as a drone tracking benchmark to evaluate this algorithm. In [29], they presented a benchmark named VisDrone2018. This dataset contains 179,264 frames and 10,209 static images acquired from the aerial perspective. It has been used for both object recognition and tracking algorithm development. However, the annotations files of this dataset are saved as text files, which are different from the common XML format that was used in Pascal VOC and ImageNet.

In this paper, we selectively collected and create large-scale image dataset from the aerial perspective which covered scenes of high diversity, and annotated carefully following the standalized way like PASCAL VOC and ImageNet.

III. Drone-based Benchmark dataset

Large-scale and well-annotated datasets are essential for developing a well-performed deep learning model. Ideally, the datasets should cover as many as possible scenarios from various view angles so that the model could learn more representative features of the same objects that can be adopted in different applications with less generalization errors.

A. Multi sources

Drones have already been used for different purposes in realty. For personal life, drone can be used to record travel or parties for memory; in scientific research, drones can collect data and monitor objects; patrol supervision can also be done by drones, and so on. Although drones are used in those applications, the characteristics of visual data captured from drone are quite different. Therefore, a large-scale dataset that can represent all different scenarios is vital and to provide relatively comprehensive coverage of diverse scenarios, we collect visual data from different sources, as following.

- In [30], a video dataset has been created for monitoring. The main scenario is the parking lot in which the behavior of many people were recorded to detect possible criminal activities. Cars were recorded from different angles and at different distances and people were on the move and in different forms.
- We have also used two datasets, UCF Aerial Action and PNNL Parking [31] [32], from Center For Research In Computer Vision of the University of Central Florida. The former covers various actions of a person recorded from various angles, while the latter monitored the crowd and people in the scene have various forms and actions.
- DJI is a leading company in the field of civil drones. They have a community for their drone users, skypixel, and many drone enthusiasts upload their own videos that record their life activities. These resources contain many scenes recorded in their lives and travels. The resource is very diverse, so we also selected a lot of useful data from here.

Besides these datasets, we also selected some data from other sources to enhance the diversity. Moreover, we also collect lots of data using our own drones.

B. Diverse scenes

Since drones can be used in a wide range of applications in various fields, we tried to include as many scenes as possible in our dataset, and some examples are shown in Fig 1. We also have a basic statistic of scenes as shown in Table II. In parking lots, the combination of cars and people are typical scenes and the corresponding actions of people usually are to open a door or enter into a car. In travelling, the background varies in a wide range, from a monochrome background to a colourful background. This kind of feature may affect the recognition of the background. In parties, the crowd is the representative feature, people were recorded from various angles and they overlapped with each other. There are also lots of other scenes and our dataset provides reasonable samples to represent them.

We collect image data from various dataset, and carefully select typical images to cover as many scenes as possible to increase the representative.

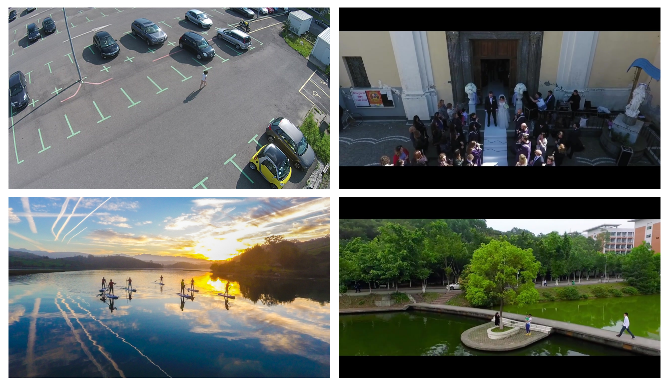

Fig. 1: Various scenes has been covered in our dataset. Here are representative samples from a parking lot, social party, travelling and routine life.

C. Multi resolutions

Usually, the camera devices mounted on drones could have multiple models with various kinds of resolutions. Therefore, if the dataset covers different resolutions, it can simulate more scenes and simulate multiple different devices. Our dataset includes various resolutions to cover various types of device as shown in Fig 2. Video material and images from skypixel are usually a record of daily life and entertainment for drone enthusiasts, and these images usually have higher resolution. So we carefully collected and sorted out some of the data from skypixel. While the data from UCF Aerial Action and PNNL Parking contains different resolutions. This diversity facilitates the training of more practical deep learning models.

D. View angles

The most important difference between a typical visual dataset and an aerial view dataset is the view angle. The form of a character, car or most other thing appears very differently from the top and front views as shown in Fig 3. For example, it is easy to distinguish different parts of human body, such as the faces, arms or legs, from the front view, while from the

TABLE I: The basic statistic of UNT_Aerail_Dataset

UNT_Aerial_Dataset	Person	boar	car	bicycle	truck	bus
No. of Images	4394	90	3324	296	267	26
No. of Object	26921	797	14225	583	272	27

TABLE II: Statistic of Dataset

Scenes	No. of Images	**No. of Objects per Image**	No. of Images	**Angle of View**	No. of Images
Parking Lot	552	$1 \sim 10$	3967	Vertical View	1335
Action Test	2108	$11 \sim 20$	608	Side View	3665
Routine Life	414	$21 \sim 50$	378		
Outdoor Living	626	$51 \sim 100$	25		
Harbour	50	$101 \sim 150$	8		
Social Party	1251	$151 \sim 200$	14		

Fig. 2: Examples with different resolutions. The first row shows images with low resolution (850x480). The second row shows images with high resolution (1920x1080).

top view, only the crown of the head and shoulders can be seen, and these are either invisible or looks different from the front perspective. If it is overlooking, it is a totally different scene. The same is true for cars and most other things. Usually the videos are recorded continuously when the drone moves, therefore, images are taken at different distances and from various view angles.

Fig. 3: Example images from different angles. The first row shows the different forms of people from overlooking and the top view. The second row shows the different shapes of cars from different view angle.

E. Different heights and distances

An important difference between images or videos of a drone view and a head-up view is that the image and video captured by the drone are top-down or at an angled top view. The view seen from this angle is quite different from the view of the head-up. For safety reasons, the drone cannot be too close to the subject, so the drone must have a certain distance from the object being photographed. Accordingly, this result in a smaller target and makes object recognition more challenging.

F. Summary

The datasets of images from aerial view are indeed necessary because they show clearly different features with normal image datasets.

- The angle of sight is different from that observed from the ground, which indicates that the image seen from the perspective of the drone is quite different from the usual angle of view. The angle will have a great influence on the final model. Therefore, our dataset managed to cover as many angles as possible.
- For safety reasons, the drone must be at a certain distance from the objects, so the distance is farther than the usual angle of view, so the target object will be smaller or even difficult to observe.
- Due to the distance, the range of the field of view will become larger and the number of objects captured by the image will be larger. Therefore, dense crowds or groups of vehicles are common in aerial view, and overlap and different forms are typical features that should be considered. Our dataset contains many of these scenarios.

To create a dataset that accomplishes these three goals, we collect photos from a variety of scenes, including various resolutions. Photographs in different scenarios make the database more inclusive and more representative. Our database contains pictures of various scenes, such as parking lots, crowd activities, travel activities, and scenes on the highway. Images of different resolutions are more representative, imitating input in a variety of situations, and are beneficial for training more robust and more adaptable models. The labelled images are shown in Fig 4. From the examples, different scenes, angles and object densities are shown.

Our benchmark is of high diversity. The source of our dataset consists of frames captured from more than 70 videos and also images from different scenes. DJI is the world's leader in commercial and civilian drone industry, and some of our sources are from SKYPIXEL, the community supported by DJI.

Fig. 4: Example of images with label boxes. Sample images from different angles and different object density are included.

IV. TEST AND EVALUATION

Convolution Neural Networks (CNNs) [33] has been proved to be a great success in the two core problems of computer vision, object recognition and detection. CNNs apply 3D kernel filters to extract different features from original input image or feature maps of previous layers and occasionally interspersed with several pooling layers, which aims to reduce the size of the feature map, thus greatly reducing the parameters of the model. Ended with fully-connected layers, the network can be used to predict detections and classifications. The typical object detection network SSD is a well-known network model based on CNN. It has a base convolution network to extract features and with additional progressively smaller convolution layers corresponding to different receptive fields, and then follows fully connected layers to do detection and classification.

A. Models

There are three kinds of well-known deep learning models which would be used for training and evaluating on our dataset.

Faster RCNN [1] is an improved version based on RCNN and Fast RCNN, which has higher accuracy and faster recognition than previous versions. The Faster RCNN actually has two subnets. One is a small CNN network called the Regional Proposal Network (RPN) for generating regional proposals. The other CNN network is used for predicting categories and detecting locations from the proposals of RPN.

You Only Look Once (YOLO) [2] is an object detection model designed for real-time detection. YOLO is much faster than the Faster RCNN, but it is less accurate than the Faster RCNN, which is a balance between speed and accuracy. In YOLO, the author considers the detection problem as a regression problem. The YOLO model takes the image as input, and then the input image is divided into grids of $S \times S$, each grid has N bounding boxes and predicts $N \times C$ confidence, and C is the total number of categories. Confidence reflects this bounding box containing objects and possibilities as well as object categories.

While the accuracy of Faster RCNN is higher than that of YOLO, YOLO is much faster than the speed of Faster RCNN. The SSD (Single Shot Detector) network strikes a good balance between speed and accuracy. The basic idea of SSD network is similar to YOLO, then the input image and different feature maps is divided into grids of different sizes, and bounding boxes are generated from grids and used for detection and classification. The final detector and classifier are based on feature maps with different sizes.

We used these three recognition models to train and test in our dataset.

B. Dataset

In order to make the dataset more representative and the model trained on this dataset more robust, we have further expanded the dataset during the training process by randomly shifting and cropping. While this data extension does not cover all angles and the effect of the extension is limited, it will also facilitate the training of powerful models.

C. Test Result

To demonstrate the usefulness of our dataset in training models, we conducted a controlled trial. We used a typical detection model, SSD detection network. We split our dataset into two parts, the training set and testing set. The model that was not trained on our dataset had a very low accuracy of 0.03 mAP during the test, but the model trained on our dataset achieved an accuracy of 0.399 mAP, although this It does not reach the optimal data of SSD because SSD still has certain defects in the recognition of small targets. As for the YOLOv3, without training and directly test on our dataset, the accuracy is only 0.05 mAP and can get 0.63 mAP after training on our dataset.

V. CONCLUSION

We introduce a new image dataset for the research of object detection and classification in everyday life in their natural environments on the drone platform. This dataset consists of 5000 images from various of videos and some high resolution images, about 10k instances are gathered, annotated and organized to promote the development of classification and detection algorithms. Our dataset covers various of scenarios and has different resolutions, and it more representative and more practical.

There are still several promising directions for future annotations on our dataset. We will scale up our dataset and annotate more kinds of objects in the future work. Due to the difference between the UAV view and the normal view, we will also consider and include more features into our dataset.

References

[1] R. Girshick, “Fast r-cnn,” in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1440–1448.

[2] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, “You only look once: Unified, real-time object detection,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 779–788.

[3] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, “Ssd: Single shot multibox detector,” in *European conference on computer vision*. Springer, 2016, pp. 21–37.

[4] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, “Generative adversarial nets,” in *Advances in neural information processing systems*, 2014, pp. 2672–2680.

[5] Q. Chen and V. Koltun, “Photographic image synthesis with cascaded refinement networks,” in *IEEE International Conference on Computer Vision (ICCV)*, vol. 1, no. 2, 2017, p. 3.

[6] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein *et al.*, “Imagenet large scale visual recognition challenge,” *International Journal of Computer Vision*, vol. 115, no. 3, pp. 211–252, 2015.

[7] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, “The pascal visual object classes (voc) challenge,” *International Journal of Computer Vision*, vol. 88, no. 2, pp. 303–338, Jun. 2010.

[8] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, “Microsoft coco: Common objects in context,” in *European conference on computer vision*. Springer, 2014, pp. 740–755.

[9] R. Collins, X. Zhou, and S. K. Teh, “An open source tracking testbed and evaluation web site,” in *IEEE International Workshop on Performance Evaluation of Tracking and Surveillance (PETS 2005)*, vol. 2, 2005, p. 35.

[10] Y. Wu, J. Lim, and M.-H. Yang, “Object tracking benchmark,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 37, no. 9, pp. 1834–1848, 2015.

[11] L. Leal-Taixé, A. Milan, I. Reid, S. Roth, and K. Schindler, “Motchallenge 2015: Towards a benchmark for multi-target tracking,” *arXiv preprint arXiv:1504.01942*, 2015.

[12] J. Chen, J. Xie, Y. Gu, S. Li, S. Fu, Y. Wan, and K. Lu, “Long-range and broadband aerial communication using directional antennas (acda): design and implementation,” *IEEE Transactions on Vehicular Technology*, vol. 66, no. 12, pp. 10 793–10 805, 2017.

[13] Y. Gu, M. Zhou, S. Fu, and Y. Wan, “Airborne wifi networks through directional antennae: An experimental study,” in *Wireless Communications and Networking Conference (WCNC), 2015 IEEE*. IEEE, 2015, pp. 1314–1319.

[14] B. Wang, J. Xie, S. Li, Y. Wan, S. Fu, and K. Lu, “Enabling high-performance onboard computing with virtualization for unmanned aerial systems,” in *2018 International Conference on Unmanned Aircraft Systems (ICUAS)*. IEEE, 2018, pp. 202–211.

[15] 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.

[16] A. Robicquet, A. Sadeghian, A. Alahi, and S. Savarese, “Learning social etiquette: Human trajectory understanding in crowded scenes,” in *European conference on computer vision*. Springer, 2016, pp. 549–565.

[17] M. Kristan, J. Matas, A. Leonardis, M. Felsberg, L. Cehovin, G. Fernandez, T. Vojir, G. Hager, G. Nebehay, and R. Pflugfelder, “The visual object tracking vot2015 challenge results,” in *Proceedings of the IEEE international conference on computer vision workshops*, 2015, pp. 1–23.

[18] 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,” *arXiv preprint arXiv:1511.04136*, 2015.

[19] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, “The cityscapes dataset for semantic urban scene understanding,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 3213–3223.

[20] M. Enzweiler and D. M. Gavrila, “Monocular pedestrian detection: Survey and experiments,” *IEEE Transactions on Pattern Analysis & Machine Intelligence*, no. 12, pp. 2179–2195, 2008.

[21] P. Dollár, C. Wojek, B. Schiele, and P. Perona, “Pedestrian detection: A benchmark,” in *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*. IEEE, 2009, pp. 304–311.

[22] R. B. Fisher, “The pets04 surveillance ground-truth data sets,” in *Proc. 6th IEEE international workshop on performance evaluation of tracking and surveillance*, 2004, pp. 1–5.

[23] Y. Wu, J. Lim, and M.-H. Yang, “Online object tracking: A benchmark,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2013, pp. 2411–2418.

[24] D. Du, H. Qi, W. Li, L. Wen, Q. Huang, and S. Lyu, “Online deformable object tracking based on structure-aware hyper-graph.” *IEEE Trans. Image Processing*, vol. 25, no. 8, pp. 3572–3584, 2016.

[25] H. K. Galoogahi, A. Fagg, C. Huang, D. Ramanan, and S. Lucey, “Need for speed: A benchmark for higher frame rate object tracking,” in *Computer Vision (ICCV), 2017 IEEE International Conference on*. IEEE, 2017, pp. 1134–1143.

[26] S. Song and J. Xiao, “Tracking revisited using rgbd camera: Unified benchmark and baselines,” in *The IEEE International Conference on Computer Vision (ICCV)*, December 2013.

[27] M.-R. Hsieh, Y.-L. Lin, and W. H. Hsu, “Drone-based object counting by spatially regularized regional proposal network,” in *The IEEE International Conference on Computer Vision (ICCV)*, vol. 1, 2017.

[28] S. Li and D.-Y. Yeung, “Visual object tracking for unmanned aerial vehicles: A benchmark and new motion models.” in *AAAI*, 2017, pp. 4140–4146.

[29] P. Zhu, L. Wen, X. Bian, L. Haibin, and Q. Hu, “Vision meets drones: A challenge,” *arXiv preprint arXiv:1804.07437*, 2018.

[30] M. Bonetto, P. Korshunov, G. Ramponi, and T. Ebrahimi, “Privacy in mini-drone based video surveillance,” in *Automatic Face and Gesture Recognition (FG), 2015 11th IEEE International Conference and Workshops on*, vol. 4. IEEE, 2015, pp. 1–6.

[31] G. Shu, A. Dehghan, O. Oreifej, E. Hand, and M. Shah, “Part-based multiple-person tracking with partial occlusion handling,” in *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*. IEEE, 2012, pp. 1815–1821.

[32] A. Dehghan, S. Modiri Assari, and M. Shah, “Gmmcp tracker: Globally optimal generalized maximum multi clique problem for multiple object tracking,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 4091–4099.

[33] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, “Gradient-based learning applied to document recognition,” *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.