# VisDrone-MOT2021: The Vision Meets Drone Multiple Object Tracking Challenge Results

Guanlin Chen[1], Wenguan Wang[3], Zhijian He[2], Lujia Wang[2], Yixuan Yuan[4], Dingwen Zhang[5], Jinglin Zhang[6], Pengfei Zhu[1], Luc Van Gool[3], Junwei Han[5], Steven Hoi[7], Qinghua Hu[1], Ming Liu[2], Andrea Sciarrone[13], Chao Sun[10], Chiara Garibotto[13], Duong Nguyen-Ngoc Tran[11], Fabio Lavagetto[13], Halar Haleem[13], Hakkı Motorcu[15], Hasan F. Ateş[16], Huy-Hung Nguyen[11], Hyung-Joon Jeon[11], Igor Bisio[13], Jae Wook Jeon[11], Jiahao Li[10], Long Hoang Pham[11], Moongu Jeon[12], Qianyu Feng[14], Shengwen Li[9], Tai Huu-Phuong Tran[11], Xiao Pan[10], Young-min Song[12], Yuehan Yao[8], Yunhao Du[9], Zhenyu Xu[8], Zhipeng Luo[8],

[1]Tianjin University, Tianjin, China.
[2]The Hong Kong University of Science and Technology, Hong Kong, China.
[3]ETH Zurich, Zurich, Switzerland.
[4]City University of Hong Kong, Hong Kong, China.
[5]Northwestern Polytechnical University, Xian, China.
[6]Nanjing University of Information Science and Technology, Nanjing, China.
[7]Singapore Management University, Singapore.
[8]DeepBlue Technology(Shanghai) Co., Ltd, Shanghai, China.
[9]Beijing University of Posts and Telecommunications, Beijing, China.
[10]Zhejiang University, Hangzhou, China.
[11]Sungkyunkwan University, Suwon, South Korea.
[12]Gwangju Institute of Science and Technology, Gwangju, Korea.
[13]University of Genova, Genova, Italy.
[14]University of Technology Sydney, Sydney, Australia.
[15]Özyeğin University, İstanbul, Turkey.
[16]Medipol University, İstanbul, Turkey.

## Abstract

*Vision Meets Drone: Multiple Object Tracking (VisDrone-MOT2021) challenge – the forth annual activity organized by the VisDrone team – focuses on benchmarking UAV MOT algorithms in realistic challenging environments. It is held in conjunction with ICCV 2021. VisDrone-MOT2021 contains 96 video sequences in total, including 56 sequences (∼24K frames) for training, 7 sequences (∼3K frames) for validation and 33 sequences (∼13K frames) for testing. Bounding-box annotations for novel object categories are provided every frame and temporally consistent instance IDs are also given. Additionally, occlusion ratio and truncation ratio are provided as extra useful annotations. The results of eight state-of-the-art MOT algorithms are reported and discussed. We hope that our VisDrone-MOT2021 challenge will facilitate future research and applications in the field of UAV vision. The website of our challenge can be found at* `http://www.aiskyeye.com/`.

***Key words: VisDrone, multi-object tracking, drone, challenge, benchmark***

# 1. Introduction

In recent years, UAV swarm has raised a lot of research interests due to its wide applications, as well as challenges and characteristics in system complexity, flexibility and scalability, and robustness [21]. As a crucial step for drones to emerge intelligence, smart perception of the environment heavily relies on UAV vision. Multi-object tracking (MOT) – identify and track object instances in video sequences – is one of the most critical functions of UAV vision.

Benchmark dataset serves as a main driver of MOT. Although many benchmark datasets, such as UA-DETRAC [25, 26], KITTI [12], MOT20 [8, 18], TAO [7] and GMOT-40 [1], have been proposed and greatly advanced the field of MOT, most of them are only aware of tracking pedestrians and vehicles, captured by surveillance or hand-hold cameras. However, videos captured by UAVs yield quite different challenges, such as fast camera moving, dramatic viewpoint change, and large scale variation. Although a few recent datasets were collected upon UAV platform [13, 28, 32], their scales are typically small and the scenes are quite limited, due to the difficulty of data collection and annotation. More importantly, they rarely pay attention to MOT. We therefore introduce a large-scale benchmark dataset [39], VisDrone, which is specifically collected for drone-camera based MOT and fully addresses the above issues.

Based on our VisDrone dataset, we organize the "Vision Meets Drone: Multiple Object Tracking (VisDrone-MOT2021)" challenge, as a part of our "Vision Meets Drone: A Challenge" workshop held on August, 2021, in conjunction with ICCV 2021. Our challenge dataset provides ID-consistent bounding box ground-truth as well as occlusion ratio and truncation ratio annotations. There are 29 teams participate in VisDrone-MOT2021 and the top-leading teams were invited to share their algorithms in our workshop. Performance rankings and detailed information of our challenge can also be found at `http://www.aiskyeye.com/`.

# 2. Related Work

## 2.1. MOT Benchmark Datasets

The goal of MOT is accurately identifying and stably tracking objects in video sequences. MOT is a fundamental and challenging problem in computer vision due to its critical role in a wide range of real-world applications, such as intelligent monitoring, autonomous driving, *etc*. Compared with single object tracking [11, 9], it is more difficult to collect and annotate large-scale MOT datasets, as more object instances and categories should be involved. Thus relative few benchmark datasets are proposed for MOT, but they significantly boost the advance of this field even so.

MOT challenge series, with the focus of multiple people tracking and detection, are influential in MOT. In its MOT20 version, eight video sequence, collected from three very crowded scenes, are provided. Another famous dataset, KITTI [12], provides MOT annotations in autonomous driving video sequences of five classes, *i.e.*, Road, City, Residential, Campus and Person. UA-DETRAC [25, 26] was captured by traffic cameras and contains multiple kinds of attribute annotations including vehicle category, weather, scale, occlusion ratio, and truncation ratio. GMOT-40 [1] targets at generic MOT benchmarking, as many existing MOT solutions can only handle targets within predefined categories, hard to be generalized to unseen categories. Moreover, most MOT datasets pay attention to a handful of categories such as people and vehicle while ignoring the vast majority of objects in nature. To solve this problem, TAO [7] was introduced. It contains 2907 high resolution videos and marks all kinds of objects that move in the video. Different from the above mentioned datasets, our VisDrone-MOT2021 is specifically collected for benchmarking MOT over drone data.

## 2.2. MOT Algorithms

According to the initialization mode of target objects, MOT algorithms can be broadly classified into two sets [30, 31]: detection-based tracking (DBT) and detection-free tracking (DFT). DFT approaches MOT with first-frame target object initialization. For examples, [15] conducts bi-layer inference of spatio-temporal grouping to comprehensively exploit visual cues in the sequence. To better distinguish similar targets, [34] applies online structured SVM to learn object detectors accounting for spatial constraints. Though promising, DFT algorithms can only handle the objects labeled in the first frame. This inspires the emergence of detection-based tracking algorithms.

Nowadays most MOT algorithms are detection-based. In DBT, object hypotheses are first obtained by applying type-specific object detectors or motion-based detection and then object hypotheses in different frames are linked into trajectories. DBT algorithms do not need manual initialization but the tracking performance highly depends on the quality of the detection results [17]. Zhang *et al.*[33] establish a graph model based MOT solutions, where nodes are the detections over the whole vide sequence, and then search for the global optimal by using the minimum cost flow algorithm to give the corresponding correlations and trajectories. The minimum cost flow algorithm is further improved in [23], by accounting for special structure and properties of the MOT graphs. In [14], a bilinear long-short term memory model is proposed to facilitate the learning of long-term appearance models of objects. Zhu *et al.*[38] make use of a single object tracker in the data association stage to deal with the impact of noise and frequent interactions between objects. Wen *et al.*[27] propose a non-uniform hy-

pergraph to explore the different degrees of dependencies among tracklets. Chen *et al.*[6] designs an appearance guidance attention module for filtering out background noise from appearance embedding.

## 3. VisDrone-MOT2021 Challenge

### 3.1. VisDrone-MOT2021 Challenge Setup

Participants are allowed to make use of extra data. The trackers are required to approach MOT without first-frame target-object initialization. Top-leading algorithms are presented in ICCV 2021 workshop proceedings.

### 3.2. The VisDrone-MOT2021 Dataset

VisDrone-MOT2021 is built upon VisDrone-MOT2020 [10], augmented with a few sequences. Specifically, VisDrone-MOT2021 contains 96 challenging video sequences, including 56 videos for training (24,201 frames in total), 7 sequences for validation (2,819 frames in total) and 33 sequences for testing (12,968 frames in total). For each frame, tight bounding box annotations with temporally consistent IDs and object categories are labeled. Occlusion ratio and truncation ratio are also annotated. Notably, in the VisDrone-MOT2021 Challenge, we only consider five object categories in evaluation, *i.e.*, *car*, *bus*, *truck*, *pedestrian*, and *van*. Some annotated samples are shown in Figure 1.

### 3.3. VisDrone-MOT2021 Evaluation Protocol

As VisDrone-MOT2021 doesn't provide generic detection results for tracker initialization, contestants can use their own detection methods if needed. Therefore, we evaluate the overall performance of the tracking system with or without detector embedded. We use the protocol in [20] to evaluate the tracking performance. Specifically, each algorithm is required to output a list of bounding box with confidence scores and the corresponding identities. We sort the tracklets (formed by the bounding box detections with the same identity) according to the average confidence of their bounding box detections. A tracklet is considered correct if the intersection over union (IoU) overlap with ground truth tracklet is larger than a threshold. Similar to [20], we use three thresholds in evaluation, *i.e.*, 0.25, 0.50, and 0.75. The performance of an algorithm is evaluated by averaging the mean average precision (mAP) across object classes over different thresholds. The evaluation code is available at `https://github.com/VisDrone`.

### 3.4. Submitted MOT Algorithms

There are totally 29 different MOT algorithms submitted to our VisDrone-MOT2021 challenge. We present the best nine algorithm in this paper. Some key characteristics of the piked algorithms are summarized in Table 1, and detailed descriptions can be found in Appendix A. All of top-leading algorithms in this challenge are detection based. Many advanced technologies are adopted in the leading solutions, such as Cascade R-CNN [5] and CenterNet [37] for detection, and DeepSORT [29], IOU [2] and FairMOT [35] for tracking and re-identification methods [16, 36] for feature extraction and detection association.

Among these submissions, SOMOT (A.1) and Deep IoU Tracker (A.4) uses detection model based on Cascade RCNN [5]. SOMOT (A.1), GIAOTracker-Fusion (A.2), MMDS (A.3) and Yolo-Deepsort-Visdrone (A.5) constructs their tracking model based on DeepSort [29]. In particular, SOMOT (A.1) embeds MGN [24] in the detector for handling detection associations to improve model performance. GIAOTracker-Fusion (A.2) uses the time-softnms approach to fuse the results of three trackers: base, global and post. The base tracker uses DetectoRS [19] as the detector and the joint mechanism of DeepSORT [29] and FairMOT [35] as the tracker. global tracker uses the VideoReID model [16] to extract features based on the base tracker. Post tracker applies the postprocessing method on top of global tracker. MMDS (A.3) uses the DetectoRS [19] + DeepSORT [29] framework, which introduces enhanced correlation coefficient maximization for aligning frames and calculating homography, and uses OSNet [36] to extract trajectory's appearance feature. Deep IoU Tracker (A.4) proposes to replace the commonly used cosine distance with robust jaccard distance for deep feature similarity computation, and tackle the association process as a retrieval task. Yolo-DeepSort-VisDrone (A.5) utilizes scaled-Yolov4 [4] as the backbone of the detector and applies ReID [16] for linking the bounding boxes and tracks. CenterPointCF (A.6) proposes to use center position and score information from CenterNet [37] and utilizes GMPHD filter to build a light object state model based on the center point (2-D vector). MIYoT (A.7) uses a combination of IOU and visul tracker [3], and dynamically switches between the two trackers depending on the detection and tracking match. Motorcu and Ateş proposes a new architecture, HNet (A.8), which conducts center-based detection and movement offset prediction. It introduces a heatmap feedback process using selective center reconstruction method. The detailed architectures of the above-mentioned MOT algorithms can be found in the appendix (A).

## 4. Results and Analysis

In this section, we provide detailed analyses for the benchmarking results on VisDrone-MOT2021 Challenge. Some open questions in drone MOT are also discussed to shed light on future directions.

Table 1. The summary of the picked MOT algorithms in the VisDrone-MOT2021 Challenge. GPUs for training. Implementation details (P for python). Framework of the proposed method. Pre-trained datasets.

| Algorithm | GPU | Code | Framework | Pre-trained |
|---|---|---|---|---|
| SOMOT (A.1) | Tesla V100 | P | Cascade RCNN [5]+MGN [24]+FairMOT [35] | COCO |
| GIAOTracker-Fusion (A.2) | × | P | DetectoRs [19]+DeepSORT [29]+FairMOT [35] | COCO |
| MMDS (A.3) | GTX1080Ti | P | DetectoRS [19]+DeepSORT [29] | COCO |
| Deep Iou Tracker (A.4) | Tesla V100 | P | Cascade RCNN [5]+IOU [2] | Market1501+COCO |
| Yolo-Deepsort-VisDrone (A.5) | Tesla V100 | P | DeepSORT [29]+Yolov4 [4] | COCO |
| CenterPointCF (A.6) | × | C++ | CenterNet [37] | × |
| MIYoT (A.7) | GTX1660Ti | P | Yolov5 [22]+IOU [2] | × |
| HNet (A.8) | RTX 2080 | P | HNet | × |

### 4.1. Overall Results

The evaluation results of top nine algorithms are present in Table 2 and SOMOT (A.1), GIAOTracker-Fusion (A.2) and MMDS (A.3) win the top three scores, *i.e.*, 58.61, 54.18 and 52.68. At the same time, we set three thresholds, 0.25, 0.5 and 0.75, under which SOMOT achieved the best performance. We also find that all of the top three trackers are built upon DeepSORT [29] with some modification for feature processing. Therefore, we believe that similarity calculation based re-identification is crucial for developing advanced MOT algorithms.

Since the performance of DBT algorithms are largely affected by the detector, it is important to build a good detector. SOMOT uses a Cascade RCNN [5] pre-trained on the COCO dataset and embeds MGN [24] to improve the performance. GIAOTracker-Fusion and MMDS utilize DetectoRS [19]. Deep IoU Tracker (A.4) also refers to Cascade RCNN [5] detector with IOU [2] tracker which achieves the forth place in the benchmarking.

Compared with the winners in VisDroneMOT2020 [10], the top three methods obtained similar scores. However, there were more different approaches presented in this competition. Besides, the average score of the top nine algorithms presented in this competition is 45.33, which is much higher than that in VisDroneMOT2020 [10] (*i.e.*, 40.39). This demonstrates the significant advance of this field.

### 4.2. Performance Analyzed by Categories

To provide more comprehensive evaluation, we reported the AP scores of each algorithm under each category (*i.e.*, $AP_{car}$, $AP_{bus}$, $AP_{trk}$, $AP_{ped}$ and $AP_{van}$) in Table 2. From the results, we can conclude that SOMOT (A.1) performs best over most of the categories, *i.e.*, 63.46 for bus, 55.64 for pedestrian, and 56.34 for van. MMDS (A.3) achieves the best scores in car and truck, which are 70.20 and 51.94, respectively. It is worth noting that Deep IoU Tracker (A.4) wins the second place in both the bus and van categories with the scores of 60.05 and 55.94, but the performance on other categories are not satisfied. GIAOTracker-Fusion (A.2) shows relatively good results over all the categories. One possible reason is that, GIAOTracker-Fusion fuses Base tracker, Global tracker and Post tracker, handling different categories well.

### 4.3. Discussion

Recently, the community has witnessed the astonishing developments in tracking a single target, belonging to novel classes, such as pedestrian and vehicle, in normal video sequences captured by stationary, surveillance, vehicle, or mobile built-in cameras. However, as our benchmarking results suggested, in the field of drone-camera based MOT, there is still large room for improvement, and many open questions. Below we list two future research directions that would be interesting to pursue.

- **Performance Improvement:** Our benchmarking results reveal the critical role of two major modules in performance improvements, namely detector design and feature enhancement. First, a robust detector can significantly push forward the SOAT in drone-camera based MOT. For example, all the top-leading solutions in our challenge adopt some advanced detectors, such as MGN [24], Cascade RCNN [5], and DetectoRs [19]. Second, the cross-frame object instance association leans on reliable and highly-representative features. With regard to this point, DeepSORT [29] and FairMOT [35] are some good examples.
- **Efficiency Enhancement:** It is clear that the efficiency is critical in application scenarios of drones. However, computation efficiency is still a main bottleneck for many drone camera based MOT algorithms. Fortunately, there already have some scholars to address this issue: CenterPointCF (A.6) builds a high-speed online MOT model that achieves 107 FPS without GPU acceleration, while also achieving good performance.

## 5. Conclusions

This paper concludes the VisDrone-MOT2021 challenge. There are totally 29 different methods submitted and

Table 2. Multi-object tracking results on VisDrone-MOT2021 Challenge. The best three results for each evaluation mode are bolded and highlighted in red, green and blue.

| Algorithm | $AP$ | $AP_{0.25}$ | $AP_{0.5}$ | $AP_{0.75}$ | $AP_{car}$ | $AP_{bus}$ | $AP_{trk}$ | $AP_{ped}$ | $AP_{van}$ |
|---|---|---|---|---|---|---|---|---|---|
| SOMOT | **58.61** | **70.75** | **61.26** | **43.84** | **69.18** | **63.46** | **48.45** | **55.64** | **56.34** |
| GIAOTracker-Fusion | **54.18** | **63.41** | **55.35** | **43.78** | **69.33** | **51.05** | **43.20** | **55.06** | 52.26 |
| MMDS | **52.68** | 62.92 | **53.42** | **41.69** | **70.20** | 40.68 | **51.94** | **50.27** | 50.29 |
| Deep IoU Tracker | 48.54 | **63.16** | 48.11 | 34.33 | 51.97 | **60.05** | 37.66 | 37.06 | **55.94** |
| Yolo-Deepsort -VisDrone | 46.70 | 57.43 | 48.92 | 33.75 | 60.32 | 43.61 | 36.22 | 40.73 | **52.62** |
| CenterPointCF | 44.03 | 56.91 | 44.09 | 31.09 | 65.65 | 39.08 | 41.47 | 28.34 | 45.61 |
| MIYoT | 39.35 | 50.72 | 39.25 | 28.10 | 62.05 | 30.95 | 36.10 | 29.79 | 37.88 |
| HNet | 24.71 | 33.88 | 24.35 | 15.89 | 56.78 | 11.90 | 10.99 | 27.35 | 16.50 |

eight of them are reviewed in this paper. From our benchmarking results, we find many new techniques in object detection, tracking, and reidentification are adopted in the top-leading methods. Hence, we also highlight some future directions. We expect our challenge to speed up the development of this exciting research field.

# A. Descriptions of Submitted MOT Algorithms

In the appendix, we present 8 state-of-the-art MOT algorithms that got good results in the VisDrone-MOT2021 Challenge.

## A.1. Simple Online Multi-Object Tracker (SOMOT)

*Zhipeng Luo, Yuehan Yao and Zhenyu Xu*
*{luozp, yaoyh, xuzy}@deepblueai.com*

Following Separate Detection and Embedding model, Luo et al. build a strong detector based on Cascade RCNN [5] and embedding model based on Multiple Granularity Network (MGN) [24]. For association step, they build simple online multi-object tracker browsing ideas from DeepSORT [29] and FairMOT [35]. For detector, Cascade RCNN pretrained on COCO is applied. For embedding model, bag of tricks are used to improve the performance of MGN. For association step, they initialize a number of tracklets based on the estimated boxes in the first frame. In the subsequent frames, they associate the boxes to the existing tracklets (all activated tracklets) according to their distances measured by embedding features. They update the appearance features of the trackers in each time step to handle appearance variations. Then, unmatched activated tracklets and estimated boxes are associated by their distance of Intersection over Union(IoU). Also, inactivated tracklets and estimated boxes are associated by their distance of IoU.

## A.2. GIAOTracker-Fusion

*Yunhao Du*
*dyh_bupt@163.com*

GIAOTracker-Fusion is a fusion tracker of base tracker, global tracker and post tracker. GIAOTracker-Base uses DetectoRS [19] as the detector, which is pretrained on COCO and finetuned on VisDrone2019MOT-train+val. Then DeepSORT [29] algorithm is used as tracker. Considering the effects of camera motion, ORB+RANSAC is used for image alignment. It combine the feature bank mechanism in the SORT/DeepSORT and the feature updating mechanism in the JDE/FairMOT and optimize the Kalman Algorithm by using the bbox confidence to define the noise scale. It also uses the stronger features extractor OSNet [36] to extract appearance features from bboxes, which is trained on the VisDrone2019MOT dataset. "Rough2Fine" tracking strategy is also utilized, which tracks all "person" objects and all "vehicle" objects separately, then determine the tracklet's class by "SoftVote" mechanism. When tracking "vehicle" objects, the Unscented Kalman is used to replace the Linear Kalman Algorithm, which is more robust to the nonlinear motion. GIAOTracker-Global: Based on the tracklets from the GIAOTracker-Base Algorithm, VideoReID Model is used to extract the features from these tracklets, and then associate them using appearance feature cost, time cost and motion cost. GIAOTracker-Post: Based on the results from the GIAOTracker-Global Algorithm, it uses some postprocessing method like denoising, interpolation, rescoring. Finally, it uses a stronger detector to produce better bboxes, then abtain a better tracking results and then fuses it with the previous tracking results using time-softnms as the fusion tracker.

## A.3. An improved multi-object tracking approach based on DeepSort (MMDS)

*Shengwen Li*
*2019140337@bupt.edu.cn*

MMDS uses DetectoRS [19] as the base detector ,which is a two-stage method and DeepSORT [29] without appearance features as framework of objects tracking. Its improvements include: (1) To reduce the impact of drone motion, we adopt ECC(Enhanced Correlation Coefficient Maximization) to align frames and calculate homography matrix between consecutive frames. Then we map the object location in previous frame to the current frame by homography matrix. (2) UKF(Unscented Kalman Filter) is used instead of the Linear Kalman Filter in DeepSORT [29] to estimate the motion state of objects more accurately. (3) The tracked objects which don't match within k frames are not allowed to associate with objects detected in current frame. And the value of k is not fixed, it will change according to the length and confidence of tracklets. (4) Instead of performing non-maximum suppression on objects at first, it tracks all the objects and delete overlapping trajectories finally. (5) OSNet [36] is used to extract each trajectory's appearance feature, measure their distance from others and we simply merge two trajectories if their distance is close enough.

### A.4. Deep IoU Tracker

*Xiao Pan, Qianyu Feng, Chao Sun, Jiahao Li*
{*xiaopan, c_sun, xljh*}*@zju.edu.cn,*
*qianyu.feng@student.uts.edu.au*

Deep IoU Tracker uses the tracking-by-detection method, which first detect bounding boxes for each frame and then perform association between two adjacent frames. The detector is the ensemble of Cascade-RCNN [5] trained under different configurations. It split the image during inference to improve the detection performance of small objects. The association is solved by Hungary algorithm, and the key lies in the design of a robust similarity for the cost matrix. IoU distance of bounding boxes is used together with the cosine similarity of deep re-id features, which are weighted summed after softmax operation. Enhanced Correlation Coefficient Maximization is applied to compensate for the camera motion. It conduct tracking for each class separately to reduce the interference from another class.

### A.5. Implementation of DeepSORT with Scaled-YOLOv4 for Visual Drone Multi-Object Tracking (Yolo-Deepsort-VisDrone)

*Duong Nguyen-Ngoc Tran, Long Hoang Pham, Huy-Hung Nguyen, Tai Huu-Phuong Tran, Hyung-Joon Jeon and Jae Wook Jeon*
{*duongtran, phlong, huyhung91, taithp, joonjeon, jwjeon*}*@skku.edu*

a) Detector – Scaled-yolov4 [4] Information about the training set for Detection Backbone: VisDrone 2021-DET dataset (train and val sets; NOT using testdev) and pre-trained scaled-yolov4 model on COCO dataset. b) Tracker – DeepSORT ReID [16] is mainly used when linking bounding boxes and tracks. The distance between the feature vectors computed by ReID from the object image of the current tracking target (tracks) and the feature vectors also calculated by ReID from the object image out by the bounding box (detections) in scaled-yolov4.

### A.6. High-speed online multi-class multi-object tracking with Center Point based Cascaded Filtering (CenterPointCF)

*Young-min Song and Moongu Jeon*
{*sym, mgjeon*}*@gist.ac.kr*

CenterPointCF is a high-speed online multi-class multi-object tracking method based on the tracking-by-detection paradigm. In order to achieve the high speed and track all multi-class objects in parallel simultaneously, we just utilize the center position and score information from the CenterNet [37] detector trained in VisDrone2021 Object Detection train set. In addition, to build a light object state model based on the center point (2-D vector), the Gaussian mixture probability hypothesis density (GMPHD) filer [3] is exploited which presents a closed-form solution for recursive Bayesian filtering. Finally, tracking process propagates with three cascaded stages that consist of (1) initialization (birth) by reliable detections, (2) update by reliable detections-to-track association, and (3) track-wise association. The GMPHD filter is used to calculate the center point based probabilistic distances in both association steps.

### A.7. Medianflow-Iou-Yolo-Tracker (MIYoT)

*Halar Haleem, Igor Bisio, Chiara Garibotto, Fabio Lavagetto, Andrea Sciarrone*
{*halar.haleem, chiara.garibotto*}*@edu.unige.it,*
{*igor.bisio, fabio.lavagetto, andrea.sciarrone*}*@unige.it*

Performance of the detection has significant impact on the tracking outcome. In this context, latest deep learning model [22] is used along with the combination of IoU and visual tracker [3]. The tracker work as follows: Initially, IoU based matching is performed and visual tracker is activated to track the object from its previous position for 'x' number of frames in case it is not matched with any of the detections. During this process, the tracking is shifted back to the IoU tracker from visual tracker if a new detection matches with the tracked object.

## A.8. HNet

*Hakkı Motorcu, Hasan F. Ateş*
*hakki.motorcu@ozu.edu.tr, hfates@medipol.edu.tr*

Our deep neural network based joint tracking and detection approach HNet uses center-based detection and movement offset prediction Our method is inspired from CenterTrack [2]. However, we used our own design backbone architecture, and we introduced a different Heatmap feedback process which uses our selective center reconstruction method. HNet takes current and previous video frames

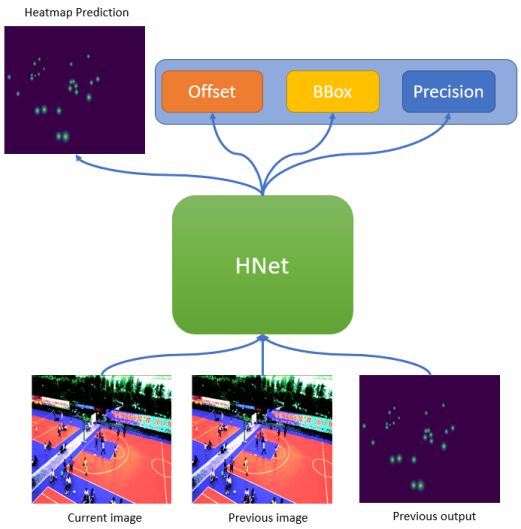


Figure 1. HNet Input Output Diagram.

and takes postprocessed version of previous heatmap output as input. After feed forward procedure HNet outputs one heatmap consists of 2D Gaussian peaks which are representing center locations, alongside the heatmap model predicts bounding box size, and a movement vector (offset) for each detection. By matching those detections with the prior ones with Hungarian matching algorithm it obtain trajectories of our detections in real time. HNet just uses current and prior outputs, and do not employ any major post processing algorithm on trajectories. Shown on (Fig.2), it can be seen that the input output diagram and the detailed architecture diagram of our model HNet.

## References

[1] Hexin Bai, Wensheng Cheng, Peng Chu, Juehuan Liu, Kai Zhang, and Haibin Ling. GMOT-40: A benchmark for generic multiple object tracking. *CoRR*, abs/2011.11858, 2020.

[2] Erik Bochinski, Volker Eiselein, and Thomas Sikora. High-speed tracking-by-detection without using image information. In *AVSS*, pages 1–6. IEEE Computer Society, 2017.

[3] Erik Bochinski, Tobias Senst, and Thomas Sikora. Extending IOU based multi-object tracking by visual information. In *AVSS*, pages 1–6. IEEE, 2018.

[4] Alexey Bochkovskiy, Chien-Yao Wang, and Hong-Yuan Mark Liao. Yolov4: Optimal speed and accuracy of object detection. *CoRR*, abs/2004.10934, 2020.

[5] Zhaowei Cai and Nuno Vasconcelos. Cascade R-CNN: delving into high quality object detection. In *CVPR*, pages 6154–6162. IEEE Computer Society, 2018.

[6] Yong Chen, Junjie Huang, Huanlin Liu, Meiyong Huang, and Zhibo Zou. Appearance guidance attention for multi-object tracking. *IEEE Access*, 9:103184–103193, 2021.

[7] Achal Dave, Tarasha Khurana, Pavel Tokmakov, Cordelia Schmid, and Deva Ramanan. TAO: A large-scale benchmark for tracking any object. In *ECCV (5)*, volume 12350 of *Lecture Notes in Computer Science*, pages 436–454. Springer, 2020.

[8] Patrick Dendorfer, Hamid Rezatofighi, Anton Milan, Javen Shi, Daniel Cremers, Ian D. Reid, Stefan Roth, Konrad Schindler, and Laura Leal-Taixé. MOT20: A benchmark for multi object tracking in crowded scenes. *CoRR*, abs/2003.09003, 2020.

[9] Xingping Dong, Jianbing Shen, Wenguan Wang, Ling Shao, Haibin Ling, and Fatih Porikli. Dynamical hyperparameter optimization via deep reinforcement learning in tracking. *IEEE Trans. Pattern Anal. Mach. Intell.*, 2019.

[10] Heng Fan, Dawei Du, Longyin Wen, Pengfei Zhu, Qinghua Hu, Haibin Ling, Mubarak Shah, Junwen Pan, Arne Schumann, Bin Dong, Daniel Stadler, Duo Xu, Filiz Bunyak, Guna Seetharaman, Guizhong Liu, V. Haritha, Hrishikesh P. S, Jie Han, Kannappan Palaniappan, Kaojin Zhu, Lars Wilko Sommer, Libo Zhang, Linu Shine, Min Yao, Noor M. Al-Shakarji, Shengwen Li, Ting Sun, Wang Sai, Wentao Yu, Xi Wu, Xiaopeng Hong, Xing Wei, Xingjie Zhao, Yanyun Zhao, Yihong Gong, Yuehan Yao, Yuhang He, Zhaoze Zhao, Zhen Xie, Zheng Yang, Zhenyu Xu, Zhipeng Luo, and Zhizhao Duan. Visdrone-mot2020: The vision meets drone multiple object tracking challenge results. In *ECCV Workshops (4)*, volume 12538 of *Lecture Notes in Computer Science*, pages 713–727. Springer, 2020.

[11] Heng Fan, Liting Lin, Fan Yang, Peng Chu, Ge Deng, Sijia Yu, Hexin Bai, Yong Xu, Chunyuan Liao, and Haibin Ling. Lasot: A high-quality benchmark for large-scale single object tracking. In *CVPR*, pages 5374–5383. Computer Vision Foundation / IEEE, 2019.

[12] Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. Vision meets robotics: The KITTI dataset. *Int. J. Robotics Res.*, 32(11):1231–1237, 2013.

[13] Nan Jiang, Kuiran Wang, Xiaoke Peng, Xuehui Yu, Qiang Wang, Junliang Xing, Guorong Li, Jian Zhao, Guodong Guo, and Zhenjun Han. Anti-uav: A large multi-modal benchmark for UAV tracking. *CoRR*, abs/2101.08466, 2021.

[14] Chanho Kim, Fuxin Li, and James M. Rehg. Multi-object tracking with neural gating using bilinear LSTM. In *ECCV*

*(8)*, volume 11212 of *Lecture Notes in Computer Science*, pages 208–224. Springer, 2018.

[15] Liang Lin, Yongyi Lu, Chenglong Li, Hui Cheng, and Wangmeng Zuo. Detection-free multiobject tracking by reconfigurable inference with bundle representations. *IEEE Trans. Cybern.*, 46(11):2447–2458, 2016.

[16] Hao Luo, Youzhi Gu, Xingyu Liao, Shenqi Lai, and Wei Jiang. Bag of tricks and a strong baseline for deep person re-identification. In *CVPR Workshops*, pages 1487–1495. Computer Vision Foundation / IEEE, 2019.

[17] Wenhan Luo, Junliang Xing, Anton Milan, Xiaoqin Zhang, Wei Liu, and Tae-Kyun Kim. Multiple object tracking: A literature review. *Artif. Intell.*, 293:103448, 2021.

[18] Anton Milan, Laura Leal-Taixé, Ian D. Reid, Stefan Roth, and Konrad Schindler. MOT16: A benchmark for multi-object tracking. *CoRR*, abs/1603.00831, 2016.

[19] Siyuan Qiao, Liang-Chieh Chen, and Alan L. Yuille. Detectors: Detecting objects with recursive feature pyramid and switchable atrous convolution. *CoRR*, abs/2006.02334, 2020.

[20] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252, 2015.

[21] Erol Sahin. Swarm robotics: From sources of inspiration to domains of application. In *Swarm Robotics*, volume 3342 of *Lecture Notes in Computer Science*, pages 10–20. Springer, 2004.

[22] ultralytics. Yolov5. `https://github.com/ultralytics/yolov5`, 2020.

[23] Congchao Wang, Yizhi Wang, Yinxue Wang, Chiung-Ting Wu, and Guoqiang Yu. mussp: Efficient min-cost flow algorithm for multi-object tracking. In *NeurIPS*, pages 423–432, 2019.

[24] Guanshuo Wang, Yufeng Yuan, Xiong Chen, Jiwei Li, and Xi Zhou. Learning discriminative features with multiple granularities for person re-identification. In *ACM Multimedia*, pages 274–282. ACM, 2018.

[25] Longyin Wen, Dawei Du, Zhaowei Cai, Zhen Lei, Ming-Ching Chang, Honggang Qi, Jongwoo Lim, Ming-Hsuan Yang, and Siwei Lyu. DETRAC: A new benchmark and protocol for multi-object tracking. *CoRR*, abs/1511.04136, 2015.

[26] Longyin Wen, Dawei Du, Zhaowei Cai, Zhen Lei, Ming-Ching Chang, Honggang Qi, Jongwoo Lim, Ming-Hsuan Yang, and Siwei Lyu. UA-DETRAC: A new benchmark and protocol for multi-object detection and tracking. *Comput. Vis. Image Underst.*, 193:102907, 2020.

[27] Longyin Wen, Dawei Du, Shengkun Li, Xiao Bian, and Siwei Lyu. Learning non-uniform hypergraph for multi-object tracking. In *AAAI*, pages 8981–8988. AAAI Press, 2019.

[28] Longyin Wen, Dawei Du, Pengfei Zhu, Qinghua Hu, Qilong Wang, Liefeng Bo, and Siwei Lyu. Detection, tracking, and counting meets drones in crowds: A benchmark. *CoRR*, abs/2105.02440, 2021.

[29] Nicolai Wojke, Alex Bewley, and Dietrich Paulus. Simple online and realtime tracking with a deep association metric. In *ICIP*, pages 3645–3649. IEEE, 2017.

[30] Bo Yang and Ram Nevatia. Online learned discriminative part-based appearance models for multi-human tracking. In *ECCV (1)*, volume 7572 of *Lecture Notes in Computer Science*, pages 484–498. Springer, 2012.

[31] Junbo Yin, Wenguan Wang, Qinghao Meng, Ruigang Yang, and Jianbing Shen. A unified object motion and affinity model for online multi-object tracking. In *CVPR*, pages 6768–6777, 2020.

[32] Hongyang Yu, Guorong Li, Weigang Zhang, Qingming Huang, Dawei Du, Qi Tian, and Nicu Sebe. The unmanned aerial vehicle benchmark: Object detection, tracking and baseline. *Int. J. Comput. Vis.*, 128(5):1141–1159, 2020.

[33] Li Zhang, Yuan Li, and Ramakant Nevatia. Global data association for multi-object tracking using network flows. In *CVPR*. IEEE Computer Society, 2008.

[34] Lu Zhang and Laurens van der Maaten. Preserving structure in model-free tracking. *IEEE Trans. Pattern Anal. Mach. Intell.*, 36(4):756–769, 2014.

[35] Yifu Zhang, Chunyu Wang, Xinggang Wang, Wenjun Zeng, and Wenyu Liu. A simple baseline for multi-object tracking. *CoRR*, abs/2004.01888, 2020.

[36] Kaiyang Zhou, Yongxin Yang, Andrea Cavallaro, and Tao Xiang. Omni-scale feature learning for person re-identification. In *ICCV*, pages 3701–3711. IEEE, 2019.

[37] Xingyi Zhou, Vladlen Koltun, and Philipp Krähenbühl. Tracking objects as points. In *ECCV (4)*, volume 12349 of *Lecture Notes in Computer Science*, pages 474–490. Springer, 2020.

[38] Ji Zhu, Hua Yang, Nian Liu, Minyoung Kim, Wenjun Zhang, and Ming-Hsuan Yang. Online multi-object tracking with dual matching attention networks. *CoRR*, abs/1902.00749, 2019.

[39] Pengfei Zhu, Longyin Wen, Dawei Du, Xiao Bian, Qinghua Hu, and Haibin Ling. Vision meets drones: Past, present and future. *CoRR*, abs/2001.06303, 2020.