Neural Computing and Applications (2022) 34:18171–18182
https://doi.org/10.1007/s00521-022-07456-2

ORIGINAL ARTICLE

Similarity based person re-identification for multi-object tracking using deep Siamese network

Harun Suljagic[1] · Ertugrul Bayraktar[2] · Numan Celebi[1]

Received: 30 October 2021 / Accepted: 22 May 2022 / Published online: 14 June 2022

Abstract
The process of object tracking involves consistently identifying each instance across frames depending on initial set of object detection(s). Moreover, in multiple object tracking (MOT), the process through tracking-by-detection paradigm consists of performing two common steps consecutively, which are detection and data association. In MOT, it is targeted to associate detections across frames by localizing and identifying all objects of interest. MOT algorithms further keep tracking even the most challenging issues such as revisiting the same view, missing detections, occlusion and temporarily unseen objects, same-appearance objects coexisting in the same frame occur. Hence, re-identification (re-id) appears to be the most powerful tool for assigning the correct identities to each individual instance when aforementioned issues arise. In this work, we propose a similarity-based person re-id framework, called SAT, using a Siamese neural network via shared weights. Once detections are obtained from the backbone SAT applies a Siamese feature extraction model and then we introduce a similarity array for assessing tracklet(s) and detection(s). We examine the performance of SAT on several benchmarks with extensive experiments and statistical tests, where we improve the current state-of-the-art according to commonly used performance metrics with higher accuracy, less ID switches, less false positive and negative rates.

Keywords Multiple object tracking · Deep Siamese neural network · Similarity array · Re-identification

1 Introduction

One of the biggest challenges in video processing is scene understanding. It is a critical problem in computer vision because it can play an important role in many applications (e.g., in automatic driving, surveillance video processing, sports analysis, and robot navigation) [1–3]. Majority of these systems, pedestrians can be the main focus of the scene, which makes it to lead the problems of detecting and tracking them using different algorithms [4]. It is usually assumed that the tracking-by-detection is the principal paradigm in MOT [2, 5–7]. Most recently proposed MOT methods have impressive performance results by building data association modules on top of CNN-based object detectors [7, 8]. Methods using this paradigm focus on the association of object detection across video frames [9]. MOT algorithm takes consecutive frames of a video as inputs, then detects the objects and mark each one with its characteristics like bounding-box (bbx) and tracking ID as shown in Fig. 1 focused on input and output frame.

Data association approaches can be considered in 2 groups; (1) online methods, where only the data from current and previous frames are used, and (2) offline methods, where learning is done before the actual track [6]. In this paper, we primarily focus on pedestrian tracking in consecutive frames in an online manner. Pedestrians in MOT scenarios usually challenge in heavy occlusion, mainly in crowded scenes and this makes them challenging for detection [10]. Also, pedestrian detection and tracking can suffer from different pose variations [11]. Another

✉ Harun Suljagic
harun.suljagic@ogr.sakarya.edu.tr

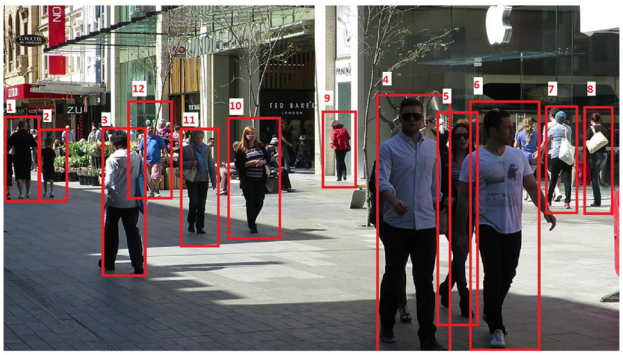
Ertugrul Bayraktar
eb@yildiz.edu.tr

Numan Celebi
ncelebi@sakarya.edu.tr

[1] Department of Information Systems Engineering, Institute of Natural Sciences, Sakarya University, 54187 Serdivan, Sakarya, Turkey

[2] Department of Mechatronics Engineering, Yildiz Technical University, 34349 Besiktas, Istanbul, Turkey

(a) Input frame - raw image, namely pedestrians



(b) Output frame - detections and tracks

Fig. 1 A general view of multi-object tracking across sequential frames for the same kind of object instances, i.e. pedestrians in this sample video frames

challenge is to combine tracks over a long period in online scenarios [12]. In online approaches, the frame has multiple instances then the association process can only use information from previous frames. Detected instances that go out of the camera view and another one enters or re-enters make MOT more challenging in addition to coexistence of same-looking instances in the same frame and revisiting the same view multiple times.

The appearance of an object is important information for data association. However, only appearance is insufficient for assigning newly detected objects to tracklets, especially for tracking objects with similar appearances (e.g., vehicles or pedestrian) [13]. Moreover, main challenge in MOT applications is data association of objects across frames in online methods where objects are assigned to detections with solely the information from the former frame [1, 7, 14, 15]. To address the listed challenges for data association, we introduce a similarity-based re-id technique to exploit the information between detections, which are then processed by the similarity array. In this study, we further propose a new data association method for manipulating the data between detected objects in consecutive frames for online MOT, which takes unexpected object motions into account as well as uncertainties due to misleading detections (e.g., false negatives and positives). The proposed SAT framework involves only a single association process as new detections are being received at every frame, which yields two sets in similarity array; a set of detection and a set of existing tracks to organize an input to the network. Then, the input of our network is formed by the association of detection with an initiated track.

Our main contributions are as follows: (1) We propose a framework, called SAT, using deep Siamese network to evaluate the similarity between objects in each frame; (2) We propose the use of a similarity array that can solve the association problem between detected objects in consecutive frames; and (3) we conducted extensive experiments on MOT16 [16, 17], MOT17 [16, 17], and MOT20 [18] separately, where we also examined the performance of the model when the detector is fine-tuned on MOT16, MOT17, and MOT20 separately and combined. We justified the results via statistical tests by examining the evaluation metrics and their relations.

The remaining of this paper is organized as follows: In Sect. 2, we review relevant previous works. In Sect. 3, the complete explanation of the proposed method is given. In Sect. 4, we provide the implementation details and report the experimental results. Finally, in Sect. 5, we conclude the study by discussing the findings.

2 Motivation and related works

One of the most popular areas in computer vision is object tracking [16, 18–25]. This technology is critical not only in security-related software but also in the improved interface between humans and computers. For this reason, we developed SAT to improve existing techniques according to the widely used metrics by exploiting the similarity information of detections in consecutive frames. In this section, we give place two main modules, data association and person re-id in addition to their relations with MOT, being relevant to our framework.

2.1 Data association

An ideal MOT system once detects all instances, then assigns consistent and robust IDs to each instance depending on the discriminant representations across frames. However, instance representations in the embedding space expose to disturbing factors due to aforementioned challenges. Therefore, the data association, the researchers pay a significant attention to which, plays a critical role in assigning tracks to new detections at each frame.

The study in [26] organizes tracks and associates them through long trajectories that improve performance via modeling and reducing interfaces from noisy or confusing object detection results. Furthermore, Chen et al. [6] introduces a multi-task CNN to associate tracks for MOT, which precisely focuses on the visual appearances. Likewise, Chu et al. [14] modifies the appearance models and searches for the target objects in the next frame using spatial-temporal attention mechanism reduced the drift caused by occlusion between targets. Moreover, there are also frameworks [1, 4, 10], which are based on graph models improving the precision of data association and robustness of the similarity model for multiple object tracking. Our framework, SAT, is also based on the paradigm of tracking-by-detection in which we once run a detector across the video sequences, and then, we link the detections to create tracks. We propose to focus on the data association problem by learning a similarity model to calculate whether two consecutive detections belong to the same track or not.

2.2 Person re-ID

Person re-id is the problem of recognizing and matching people across the sequential frames [27]. Usually, re-id is limited to a short period of time and low number of frames in a video sequence. This task is challenging due to various viewpoints, complex environment, unrestricted poses, occlusions, non-rigid and deformable shapes, and more [28–30]. Typically, there are 3 main learning strategies beneath a standard person re-id algorithm, which are; (1) feature representation learning, which is focused on developing consistent feature extraction approaches that are robust to disruptive effects, (2) deep metric learning, for planning the training objectives with various loss functions or sampling strategies employed to discriminate the query person from other people, and (3) ranking optimization, to improve the retrieved ranking list. Herein, we also follow these processes to accomplish person re-id task for MOT as explained in Sect. 3.1 in detail.

Sequential video frames provide spatio-temporal visual cues for the aforementioned learning strategies that have yielded satisfactory results so far in person re-id [31]. Moreover, Li et al. [32] propose a feature aggregation method to improve the feature extraction ability of the network in person re-id, whilst Nousi et al. [28] introduce a long-term tracking framework for classification-based re-detection and tracking, that includes object re-id of tracking and detection results. Additionally, Lin et al. [33] focus on unsupervised re-id problem according to the Euclidean distance between feature embeddings. Authors in [34] claim to improve the re-id performance by introducing attribute recognition model employing similarity and k-reciprocal re-ranking algorithm, called K-RNNA, which obtains competitive results on two primitive datasets compared to MOT17, Market-1501 [35] and Duke-MTMC-reID [36] datasets, w.r.t. mAP ($\approx$ 72.3), and Rank1 (82.3). The study in [37] proposes a half-precision compressive sensing technique measuring different image blocks to sample and reconstruct where a trade-off between speed and accuracy occur and hence they obtain almost twice a faster method while mAP only becomes 55% on PRW dataset [38]. Similar to our approach [39] uses cross-view similarity to optimize the the relationship between images by combining style-transferred samples, which achieves an mAP of 56.1% on DukeMTMC-reID dataset. On the another hand, Zhu et al. [40] encodes visual-appearance based features to learn the representations in terms of visual-appearance-level dictionary and a spatial-temporal-level dictionary, where they get Rank1 results as 59.4% and79.5% on iLIDS-VID [20] and PRID-2011 [41] datasets, respectively.

2.3 Multi-object tracking

The interest in MOT increased because of its academic and commercial potential. To overcome the aforesaid challenges, researchers also adopt assigning the last detection to the track, [42, 43] or combining temporal information into the track history [5], which examines the similarity between detections and tracks in addition to mostly used tracking-bay-detection mode [44]. Suhai et al. [45] illustrate the impact of SNNs in MOT setup and achieve a state-of-the-art performance on the MOT challenges.

Suhai et al. [45] illustrate the impacts SNNs in MOT setup by achieving a state-of-the-art performance on the MOT challenge [16]. The work in [46] exploits a modified version of DeepSORT by claiming a modification in track initialization and its rationale considering the tracking of object in previous frames. One way to measure the similarity between frames is by extracting the identity information from the similarity of objects across frames. The other way is by extracting the same information from the interaction between objects. Similar to our work, Yang et al. [47] uses a SNN that combines various attention mechanisms yielding features from different layers, which are tested on the selected sequences of OTB100 [48] with a result achieving an AUC score of 0.628.

3 Similarity-based re-ID and tracking

In this section, we describe the proposed similarity-based person re-id framework for MOT using a Siamese DenseNet [49, 50] in detail. We explain how our framework combine these procedures in the sequential video frames.

Beginning from how detection and track initiation is performed within the concept of object detection, we then detail the re-id by including the similarity calculation between each detection across consecutive frames. Our proposed framework, SAT, is given in Fig. 2 in which we illustrate the problem of person re-id via the proposed similarity array.

As the details of object detector are given in the next section, SAT framework exploits a YOLO-based [51] object detector to obtain bbxes identifying the semantic and location information of each instance at consecutive frames even if they are deformable. Afterward, the cropped regions of paired detections are employed to extract features and form representations through DenseNet-based SNN of which the weight sharing is also utilized for enabling similarity check. The main purpose is to find the most similar instances in consecutive frames. The similarity score, d, given in the similarity array as the comparisons of detections and tracks become 1 or very close to 1 if the track and detection are the same or very similar-looking, whereas d converges to zero if the track and detection are not the same or similar. The tracking module further considers 3 consecutive frames to establish a reliable track among tracklets, which are detections from the current frame or at most 2 consecutive frames and still await initialization as a track. Following each similarity examination, instance IDs are updated depending on the visual cues due to appearance changes or occlusions. Visual appearance information of tracks is stored in the tracking module, which thus enables to obtain the similarity scores by comparing them with tracklets. Track memory is refreshed by eliminating unnecessary tracks due to invisible objects and expired tracks, however, inactive tracks can be resurrected if a new detection is matched with one of the previous detections. Eventually, a track, namely ID and its corresponding representation in the embedding space, is deleted from the memory if the object is no longer visible for 3 consecutive frames or the representation of it is modified to an unidentifiable form.

3.1 Similarity-based data association

To associate the detections from the current frame with the existing tracks, we propose a new similarity array where we store the comparisons between each detection and track. Herein, we exploit the features that are extracted by Siamese DenseNet121s to check if the newly detected person is the same with any of the tracks or not. Once the detections from current frame flow through the SNNs, person re-id module moves into play over similarity simultaneously and compares the query with the so-called gallery images that are among tracks.

SNN-based trackers attracted significant attention because of their balance of precision and speed [52]. In [53], the proposed SNN with a reduction method for long-term tracking for classification using cross-entropy loss

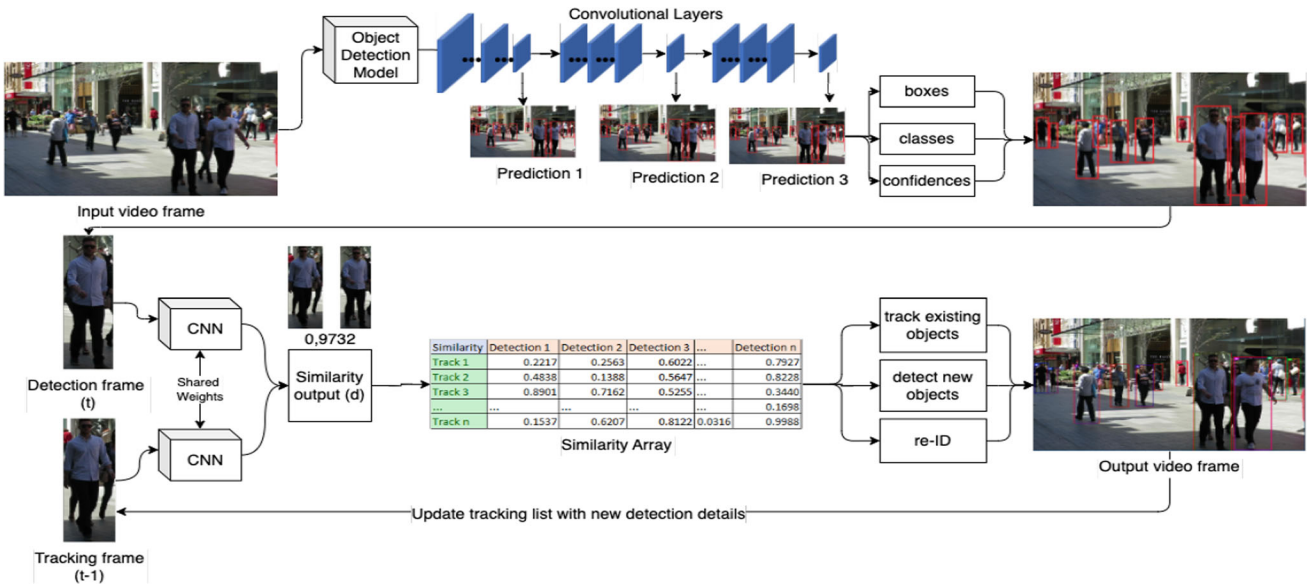


Fig. 2 Overview of our online multi-object tracking framework, SAT. The input video frames are used for object detection results, in which bbx information is integrated to complete the re-id proposal process. For classification and localising objects in each frame, YOLO-based object detectors are employed. Then, a data association procedure is performed using DenseNet-based Siamese Neural Network to get the final object tracks. According to similarity array output, existing objects are kept on being as the tracks, while new objects from next frames are being detected, and re-id action is being taken. As output, each tracklet is being displayed in a bbx with its unique ID number

function. Likewise, in our SAT framework, we use a supervised training methodology to learn features of object based on the training data, then it makes predictions between unknown objects. The function in the similarity checking part of SAT is set to be Euclidean distance $d = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$. Recent intend follows the procedure of using contrastive loss for comparing the encoded representations of SNN outputs. Contrastive loss takes the output of the network and calculates its distance to positive and negative examples for which the equation is given in Eq. 1.

$$L = y * d^2 + (1 - y) * max(margin - d, 0)^2, \quad (1)$$

where y is the actual label that is 1 if the image pairs are of the same instance and 0 otherwise, d is the Euclidean distance and *max* is the function that takes the maximum of its parameters. In other words, the loss is low if the representations are similar (closer) for positive samples and different (farther) for negative samples. During SNN training, we set the images of the same person to be the positive pairs, whilst the images of other people to be the negative pairs. A basic SNN architecture, which is illustrated in Fig. 3, takes two input images have identical sub-networks for each input with each sub-network ending in a fully connected layer computes the Euclidean distance between the fully connected layer outputs, and then passes the distance through a sigmoid activation function to determine the similarity between inputs.

Since we present a similarity array where the similarity of detected and tracked objects will be stored for the association of the objects, we show its details as illustrated in Fig. 4. After detection is performed, data association and tracking are accomplished by utilizing the highest similarity score. A person detected in the current frame is associated with a track(let) and tracking remains by preserving the same ID. If the tracked object disappears for 3 consecutive frames the ID becomes a terminated track and we delete it from the array keeping the tracked IDs.

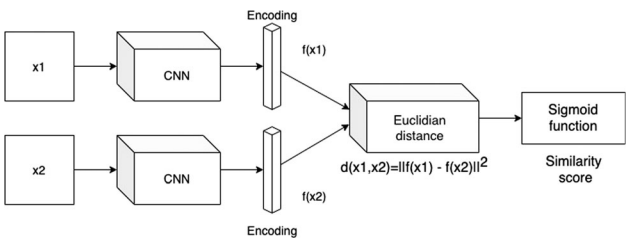


Fig. 3 One-shot learning scheme of the SNN which has the DenseNet121 architecture to decide similarity or dissimilarity between input images through two similar sub-networks sharing the same weights

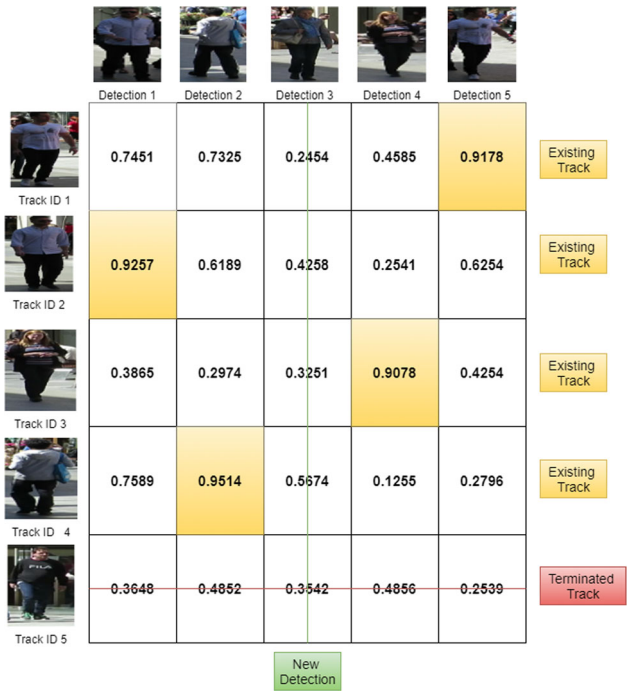


Fig. 4 The principle of similarity array based re-id on randomly selected cropped images from MOT-17 dataset. In this example the similarity between each detected object and tracked object is calculated, Track ID 1 has got the best similarity score with Detection 5 and that track will continue as existing track. The same logic is used for Track ID 2, 3 and 4. Detection 3 does not have a matching track, it is marked as a new track and will be assigned with new ID. If the track is not associated with any detection of more than three frames, that track is denoted as terminated, as shown for the Track ID 5

4 Experiments

In this section, we explain the implementation details of the proposed framework. Since the detection plays a vital role in our framework, we first analyzed performance of classical object detection methods and their combinations [54] in addition to effects of geometrical transformations on the detection [55]. We concluded that classical methods are not suitable due to their speed and accuracy rates, we then employed different YOLO-based methods, namely YoloV3 [51], YoloV4 [56], and YoloV5 [57], for transfer learning to achieve a better person detector than the original versions of these models. We accordingly used MOT16 [16, 17], MOT17 [16, 17], and MOT20 [18] datasets for fine-tuning these models where the details are given in the Ablation Study.

On the other hand, we also fine-tuned twin DenseNet121s [49, 50] by combining 3 datasets from different domains, which are MOT16 [17] containing consecutive frames recorded with a static and moving camera, Market-1501 [58] that is collected in front of a supermarket from six different cameras, and CUHK03 [59], which is composed of cropped images of persons from different cameras

and angles. For training the SNN, all input images are resized to be 221 × 221 pixels. The reason behind exploiting DenseNet-based feature extractor is to get a good balance between accuracy and speed, because DenseNet is a type of CNN that utilizes dense connections between layers, through blocks. In DenseNet, each layer obtains additional inputs from all preceding layers and they relate to matching feature-map sizes directly with each other. An output of the previous layer is used as an input of the second layer. We conduct all experiments using a PC, which is equipped with a GPU having 16 GB VRAM, an i7 CPU running @3.3 GHz and a RAM of 32 GB.

The performance metrics have a fundamental role in the assessment of algorithms in MOT tasks. A proper evaluation is only possible if particular metrics are employed and the model performance thereby can be measured in an appropriate manner. As the benchmark datasets ensure, we use multiple performance metrics to measure the impact of various components and factors on total performance, which provide ability for justification of results through quantitative comparison of various approaches. We list the most common evaluation metrics as follows that we also employed them for assessing and comparing the performance of SAT framework:

- Multiple object tracking accuracy (MOTA); accounts for all object configuration mistakes produced by the tracker, including false positives, false negatives, and mismatches.
- ID F1 Score (IDF1); is the proportion of accurately identified detections over the normal number of ground-truth and computed location.
- Mostly tracked trajectories (MT); if an object is effectively tracked for at least 80% of its life cycle, it is considered mostly tracked.
- Mostly lost trajectories (ML), if object can only be tracked for less than 20% of its overall length, it is considered mostly lost.
- Identity switches (IDS); counts the number of times a track switches from one ground-truth item to another.
- False positives (FP); is the number of items identified incorrectly by the tracker but not present in the ground truth
- False negatives (FN); is the number of ground truth objects that were not detected.

4.1 Ablation study

We experimentally demonstrate the effects of each component of our SAT framework as a result of various experiments with different configurations. We consider the ablative experiments in three consecutive phases on which we first analyze the effects of fine-tuning various YOLO-based object detectors using MOT-related datasets, then we also evaluate the performance of fine-tuning Siamese DenseNet121s using datasets from different domains as of MOT16, Market-1501, and CUHK03, finally we assess the whole SAT framework performance with different detector and filtering configurations. The common hyperparameter values are set to be as follows; 100 as the epoch number, 0.001 as the learning rate, exponential weight decay with a rate of 0.00005 at each epoch, and 0.92 as the momentum rate.

Detecting each individual person across frames is a crucial step for our framework. Therefore, we fine-tuned the most recent YOLO-based detection methods to achieve not only an accurate and precise human detector but also a robust and reliable person detection backbone for the overall SAT framework. We present the mean average precision (mAP) rates of the fine-tuned models YoloV3, YoloV4, and YoloV5 in Table 1 for which MOT16, MOT17, and MOT20 are employed as the datasets. In addition, the combination consisting of all these datasets is also used as another dataset for fine-tuning. We consequently see that YoloV5 yields higher rates for the same datasets than its antecedent. However, the dataset also plays an indispensable role, because performance of the same model decreases from MOT16 that contains less complicated scenes to MOT20 that contains most complicated scenes whilst MOT17 contains more complex than MOT16 but less complex than MOT20 scenes. The combination of all datasets case further results not the highest mAP rate in but the most robust and reliable since test sets are also composed of particular dataset samples while training and test splits (%80-%20) of the combined dataset includes images from each dataset. We exploit the YoloV5, which is fine-tuned on the combined dataset for SAT framework because it is more robust and reliable. The precision rate of fine-tuning YoloV5 on the combined dataset is 0.93 and the confidence rate is 0.91, while the recall rate is 0.863 and mAP@.5 becomes 98.54. It is noteworthy that all the results provided in tables hereafter are obtained by running the algorithms on test parts of relevant datasets, which are neither seen before testing nor used during training/fine-tuning.

Once the fine-tuning process of object detector justified that re-training on comprehensive data enables DNN to be more reliable and robust against variations, we thus fine-tuned the Siamese DenseNet121s on the combination of MOT16, Market-1501, and CUHK03. Since DenseNet is an object recognition model originally, we can evaluate the performance of fine-tuned model via accuracy rate. We modified its output layer to discriminate if an object is a person or not in addition to freezing the weights of the earlier layers. After running the training for 100 epoch, we achieved an accuracy of %98.54. However, we only

Table 1 mAP@.5:.95 of the YOLO-based models, which are fine-tuned on the datasets MOT16, MOT17, and MOT20 separately and combined

Models	MOT16	MOT17	MOT20	MOT16 + MOT17 + MOT20 combined
YoloV3	80.12	78.03	73.34	74.85
YoloV4	82.07	80.22	74.71	76.54
YoloV5	87.41	84.90	82.41	85.61

employ twin DenseNet121 models to extract as rich, robust, strong, and discriminative as possible features via its bottleneck layer from the bbxes belongs to humans that are already provided by the fine-tuned YoloV5 model.

Following the comparison of similarities, SAT framework has 3 sub-modules for tracklets, tracks, and re-id in which a filter steps in to sustain successive detection and tracking processes simultaneously. To achieve better tracking performance due to filters, we compared widely used Kalman and particle filters. The basic principle behind a Kalman filter is to use the bbx coordinates of existing detections and prior predictions to arrive at the best estimation of the present state as the bbx coordinate and shape while allowing for the likelihood of errors. In this paper, we have a reasonably accurate object detector that detects humans. However, it is not perfectly precise and occasionally misses detections, for example approximately 9 out of every 100 frames for the first 5 estimations. We, therefore, assume a constant/linear velocity model to successfully track and anticipate the next state by means of bbx and its coordinates of the person. In other words, after we specify the simple model according to physical principles, we can make a good prediction as to where the person will be in the next frame based on the present detection. In an ideal world, we detect then track each individual without any errors, but there is always a noise component such as process noise and measurement noise. In Kalman filter, we utilize current readings to estimate the current state, then use measurements to update our expectations. It subsequently all comes down to determining a new distribution (the predictions) from the prior state distribution and the measurement distribution. On the other hand, particle filter, namely sequential Monte Carlo, can follow nonlinear motion models opposite to the Kalman filters. A typical particle filter, which is also used as a generic optimization technique, calculates posterior state distribution $p(x_k \mid Z_k)$ (the probability of having sample x_k for given an overall distribution Z_k that summarizes knowledge following the observation of all data) at each time step, which is the bbx coordinates of detected humans in our case. The discrete particles approximate the posterior probability in two steps such as (1) prediction and (2) update, which are repeated recursively. Each particle includes tests to determine how likely it is that the samples (bbxes in this study) are at the position where the particles are. After the particles have been assessed, weights are distributed depending on how good the particles are located, which are then multiplied while the poor particles are deleted via the re-sampling procedure. The next particle generation then estimates where the sample object, namely bbx of person, could be. The benefits of these filters on MOT tasks are discussed in [60] in detail. The works in [61–63] further investigates the contributions of Kalman filter, while the studies in [64, 65] report the performance improvements thanks to particle filter. We integrate both of these filters to each of the SAT configurations separately that are employing different person detection backbones as the results are given in Table 2. Although, there is a remarkable difference when YoloV3 is used as the detector, the effect of the filter decreases. However, Kalman filter yields better results for each case according to all metrics except IDS when YoloV4 is the detector. In essence, the most powerful configuration of SAT framework appears to involve YoloV5 as the person detector and Kalman filter to track the feature dependent positions across frames in addition to employing our similarity array.

4.2 Comparison with the state-of-the-art

We compared the performance of our framework, SAT, with relevant methods, which show state-of-the-art performance on the benchmarking datasets, in fact, the performance measures also taken from the benchmark. Both MOT16 and MOT17 contain 7 sequences while MOT20 contains 4 more complicated sequences all of which capture images from public areas via moving and static cameras from various aspects. As the results given in Table 3

Table 2 The effects of object detectors and tracking filters for our similarity-based framework, SAT, using MOT16, MOT17, and MOT20 datasets combined

Method	MOTA ↑	FP ↓	FN ↓	IDS ↓
YOLOv3 + particle filter	51.9	9749	202,987	989
YOLOv3 + Kalman	52.6	9050	202,343	961
YOLOv4 + particle filter	52.1	10,113	229,653	912
YOLOv4 + Kalman	52.4	10,532	201,735	991
YOLOv5 + particle filter	58.1	10,113	201,576	912
YOLOv5 + Kalman	58.9	11,125	201,343	905

This study shows multiple aspects of the performance of SAT. ↑ shows the upper is better, ↓ indicates lower is better

suggest, SAT framework appears to be a very competitive method for achieving the rates regarding MT and FN in addition to being runner-up for MOTA, IDF1, and FP among all other state-of-the-art MOT techniques. It is in particular significant that FN rate of SAT is much better then the second-best output. Additionally, the results verify the success of SAT, which is amid the best two methods for 5 metrics of 7.

When we conduct the experiments on MOT17, we get the results as given in Table 4. SAT achieves the best rates at 4 metrics as of ML, IDS, FP, and FN with %32.2, 905, 11125, and 201343, respectively. Moreover, SAT is the runner-up at 2 of the remaining 3 metrics of MOTA and IDF1 with %58.9 and %58.1, respectively. We essentially see that when the data contains more complex sample scenes, the performance of SAT improved significantly.

Our last tests are performed on MOT20 and the relevant results are provided in Table 5. Similar to previous results, SAT yields best scores for ML, IDS, and FN with %24.9, 1521, and 192343, respectively. Even if SAT gets second best results for MOTA and IDF1, there is a big gap between SAT and the best method, LPC MOT, while there are slight differences between SAT and remaining methods. By keeping in mind that MOT20 is the most complicated dataset in our experiments, SAT still generates better results than the rest when we consider the overall outputs.

We further examined the performance of SAT and other methods qualitatively as given in Fig. 5 where tracks are specified for 32nd, 49th, and 52nd frames of 09th sequence of MOT16 dataset. SAT keeps the track IDs consistently but there are less people tracked. From Fig. 5a–c, scenes are dynamic while the camera is fixed and occluded humans cannot be tracked if occlusion lasts longer than 5 consecutive frames. However person with the ID5 becomes visible again and we keep its track in our array after tracking seems to be stopped in Fig. 5b. If reappearance occurs as in Fig. 5d–f, new track IDs are assigned to people even if they already had former IDs due to interruptions in tracking, which also leads to ID switches.

We also investigated the performance outputs through statistical tests following the principles reported in [73, 74]. For this reason, we determined MOTA to be our base metric for testing because it seems more challenging due to the performance of SAT as the runner-up for each dataset regarding MOTA. The correlation values between the MOTA and F-score for all datasets lie between 0.92 and 0.97, which remark MOTA and F-score evaluate the same characteristics of the algorithms as stated by [73, 74]. Since the test sets contain various number of sample images in each sequence, we perform the arithmetic during the survival curve (as a result of Kaplan-Meier estimator [75]) calculation of relevant algorithms considering the sequence that has the least number of images (450 frames) as shown in Fig. 6, where the F-scores on the vertical axes represent the average values for all the datasets. The progress demonstrates the survival curve restrictions by neighboring curves due to exploiting average F-scores and there are also sudden drops in the F-score gradually, which cause intersections across curves belong to various algorithms. The results unveil that SAT algorithm outperforms all other methods after approximately 50th frame, which are even appear to be slightly better regarding MOTA. Therefore, it is remarkably vital that assessing the performance results even they are widely used on which the continual research also verifies this fact [76].

Moreover, we computed the correlation between evaluation metrics used thus far as illustrated in Fig. 7 to analyze their effects and influences on each other. As expected MOTA and IDF1 have close relation while MT and ML, which has almost no relation with MOTA, IDF1, and IDS have a strong negative correlation, namely they conduct almost the opposite. In addition, IDS has a notable relation with both FP and FN. In essence, the results reveal that using only 4 of the performance metrics, namely, MOTA, MT or ML, FP, and FN, are sufficient instead of using all of them because the rest of them have very strong relations in the same or opposite direction.

5 Discussion and conclusion

The primary goal of a typical MOT task is to assign IDs to object of interest and keep IDs consistent across frames even if challenges such as occlusion, revisiting the same

Table 3 Comparison of state-of-the-art MOT methods with our framework **SAT** on the **MOT16** dataset, where **SAT** shows a competitive performance for all metrics whilst yielding the best results in terms of MT and FN in addition to being runner-up for MOTA, IDF1, and FP

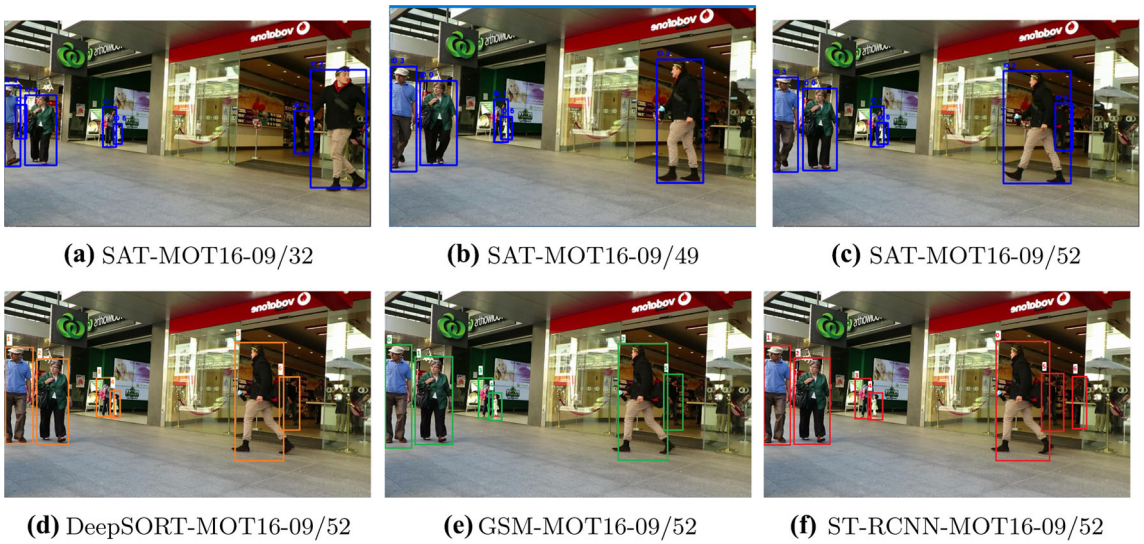
Method	MOTA↑ (%)	IDF1↑	MT↑ (%)	ML↓ (%)	IDS↓	FP↓	FN↓
HDTR [66]	53.6	53.4	21.2	37.0	618	4714	79,353
Tractor [29]	54.4	52.5	19.0	36.9	682	**3280**	79,149
MTDF [67]	45.7	40.1	14.1	36.4	1987	12,018	84,970
DeepMOT [68]	54.8	53.4	19.1	37.0	645	4389	68,376
GSM (Tractor) [10]	57.0	58.2	22.0	**34.5**	**457**	4332	73,573
Siamese Track-RCNN [45]	**59.8**	**60.8**	22.2	34.5	556	4389	68,376
SAT (Ours)	59.1	60.2	**22.5**	34.7	842	4032	**41,253**

Table 4 Comparison of state-of-the-art MOT methods with our framework **SAT** on the **MOT17** dataset, where **SAT** shows a competitive performance for all metrics whilst yielding the best results in terms of ML, IDS, FP, and FN in addition to being runner-up for MOTA and IDF1

Method	MOTA↑ (%)	IDF1↑	MT↑ (%)	ML↓ (%)	IDS↓	FP↓	FN↓
Tractor [29]	53.5	52.3	19.5	36.6	4611	12,201	230,174
MTDF [67]	49.6	45.2	18.9	33.1	5567	37,124	241,768
DeepMOT [68]	53.7	53.8	19.4	36.6	1947	11,731	247,447
GSM (Tractor) [10]	56.4	57.8	22.2	34.5	1485	14,379	230,174
Siamese Track-RCNN [45]	**59.6**	**60.1**	**23.9**	33.9	2068	15,532	210,519
SAT (Ours)	58.9	58.1	22.1	**32.2**	**905**	**11,125**	**201,343**

Table 5 Comparison of state-of-the-art MOT methods with our framework **SAT** on the **MOT20** dataset, where **SAT** shows a competitive performance for all metrics whilst yielding the best results in terms of ML, IDS, and FN in addition to being runner-up for MOTA and IDF1

Method	MOTA↑ (%)	IDF1↑	MT↑ (%)	ML↓ (%)	IDS↓	FP↓	FN↓
Tracktor [29]	52.6	52.7	29.4	26.7	1648	**6930**	236,680
MOT20 TBC [69]	54.5	50.1	33.4	19.7	2449	37,937	195,242
GNNMatch [70]	54.5	49.0	32.8	25.5	2038	9522	223,611
SP CON [71]	54.6	53.4	32.8	25.5	1674	9486	223,607
LPC MOT [72]	**56.3**	**62.5**	**34.1**	25.2	1562	11,726	213,056
SAT (Ours)	55.2	54.3	32.1	**24.9**	**1521**	11,125	**192,343**



(a) SAT-MOT16-09/32 **(b)** SAT-MOT16-09/49 **(c)** SAT-MOT16-09/52

(d) DeepSORT-MOT16-09/52 **(e)** GSM-MOT16-09/52 **(f)** ST-RCNN-MOT16-09/52

Fig. 5 **SAT** tracker performance, which successfully tracks when occlusion occurs due to crossing behind different person or another object, while the rest of the methods shift the IDs of the detected pedestrians across frames

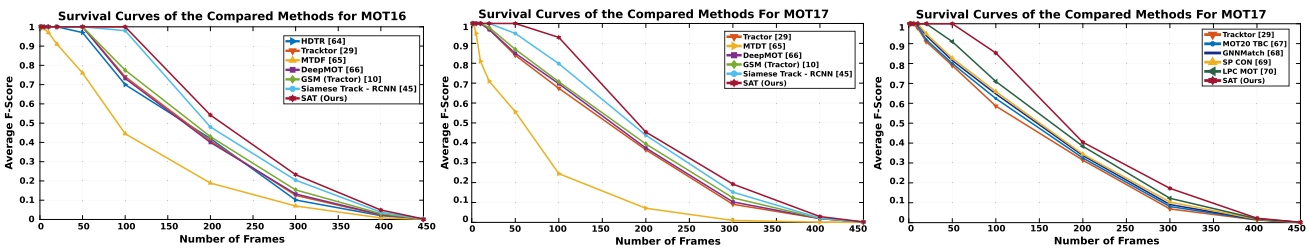


Fig. 6 Survival curves with respect to average F-scores regarding the consecutive number of frames from the datasets that we employed, namely MOT16, MOT17, and MOT20 on which the compared methods are executed. SAT algorithm outperforms all other methods after approximately 50th frame at each case

area, coexistence of same-looking objects, clutter in the scene, disappearing from the view for a few frames and then re-entering occur. Therefore, the main motivation behind SAT /framework is to overcome at least noteworthy

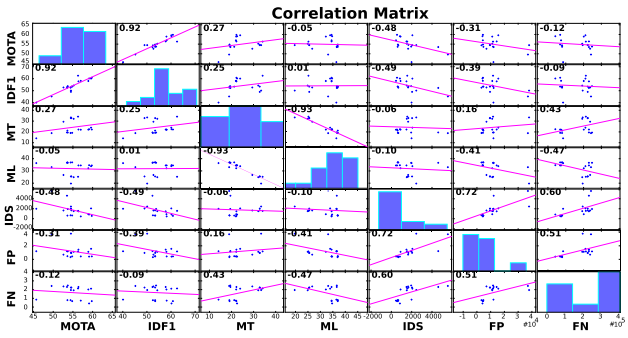


Fig. 7 The relation between the performance metrics used in this study for the results obtained from all datasets combined. The values on top-left represent the correlation ratio (The closer to 1 means a stronger relation, whilst 0 means no influence, and the closer to − 1 means the opposite behavior)

parts of the aforementioned challenges by analyzing the similarity of people in a given frame. We obtain person detections from the fine-tuned YoloV5 backbone and fine-tuned Siamese DenseNet121s extract enriched features for the similarity checking module of the framework. Thereafter, the information flowing from the sources such as similarity array, re-id module, and Kalman filtering is fused to establish tracks as the final process of the SAT framework. We conducted comprehensive experiments revealing the performance of SAT compared to state-of-the-art on different benchmarks, which are MOT16, MOT17, and MOT20. From MOT16 to MOT20, the performance of all methods compared in this study decreased due to the complexity of the scenes included in the datasets. However, SAT framework achieved satisfactory results for each dataset according to the most common metrics used in MOT task. We also justified our final configuration for SAT concerning vast number of experiments and statistical tests on different data sequences by examining possible contributions for each module of the framework. The outputs demonstrate that SAT is able to manage occlusion, revisiting the same area, disappearing from the view for a few frames and then re-entering with its robust structure by resulting in best or second best performance rates in the majority of experiment cases.

Although our SAT framework makes improvements, MOT task still remains to be a field requiring notable advancements. Especially the aforesaid challenges related to this task make it harder than classical computer vision and deep neural network-based issues. However, the performance can be even increased by a reliable and precise object detector that eliminates missed detections. Obtaining enriched and robust feature representations is another promising subject for investigation in a future work for better results. Next, we also plan to integrate geometry information into the problem formulation without changing the domain to overcome the current limitations and then improving the results significantly.

Declarations

Conflict of interest The authors certify that there is no actual or potential conflict of interest in relation to this article.

References

1. Zhang Y et al (2020) Multiplex labeling graph for near-online tracking in crowded scenes. IEEE Internet Things J 7:7892–7902
2. Yoon Y, Kim D, Song Y, Yoon K, Jeon M (2021) Online multiple pedestrians tracking using deep temporal appearance matching association. Inf Sci 561:326–351
3. Cakir S, Cetin A (2021) Visual object tracking using Fourier domain phase information. Signal Image Video Process 16:119–126
4. Braso G, Lear-Taixe L (2020) Learning a neural solver for multiple object tracking. In: Proceedings of the IEEE computer society conference on computer vision and pattern recognition, pp 6246–6256
5. Wojke N, Bewley A, Paulus D (2018) Simple online and realtime tracking with a deep association metric. In: Proceedings of international conference on image processing, ICIP, pp 3645–3649
6. Chen L, Ai H, Chen R, Zhuang Z (2019) Aggregate tracklet appearance features for multi-object tracking. IEEE Signal Process. Lett. 26:1613–1617
7. Wu Y et al (2019) Instance-aware representation learning and association for online multi-person tracking. Pattern Recognit. 94:25–34
8. Ciaparrone G, Luque F, Sanchey L, Tabik S et al (2020) Deep learning in video multi-object tracking: a survey. Neurocomputing 381:61–88
9. Yang F, Chang X, Sakti S, Wu Y, Nakamura S (2021) Remot: a model-agnostic refinement for multiple object tracking. Image Vis Comput 106:104091
10. Liu Q, Chu Q, Liu B, Yu N (2020) Gsm: graph similarity model for multi-object tracking. In: Proceedings of the twenty-ninth international joint conference on artificial intelligence, pp 530–536
11. Xu Y, Cao Y, Zhang Z (2019) Spatial-temporal relation networks for multi-object tracking. In: Proceedings of the IEEE international conference on computer vision, pp 3987–3997
12. Sadeghian A, Alahi A, Saverse S (2017) Tracking the untrackable: learning to track multiple cues with long-term dependencies. In: Proceedings of the IEEE international conference on computer vision, pp 300–311
13. Xu Y, Osep A, Ban Y, Horaud R (2020) How to train your deep multi-object tracker. In: Proceedings of the IEEE computer society conference on computer vision and pattern recognition, pp 6786–6795
14. Chu Q et al (2017) Online multi-object tracking using cnn-based single object tracker with spatial-temporal attention mechanism. In: Proceedings of the IEEE international conference on computer vision, pp 4846–4855
15. Yang M, Wu Y, Jia Y (2017) A hybrid data association framework for robust online multi-object tracking. IEEE Trans Image Process 26:5667–5679

16. Leal-Taixé L, Milan A, Reid I, Roth S, Schindler K (2015) Motchallenge 2015: towards a benchmark for multi-target tracking. arXiv:1504.01942
17. Milan A, Leal-Taixé L, Reid I, Roth S, Schindler K (2016) Mot16: a benchmark for multi-object tracking. arXiv:1603.00831
18. Dendorfer P et al (2020) Mot20: a benchmark for multi object tracking in crowded scenes. arXiv:2003.09003
19. Geiger A, Lenz P, Urtasun R (2012) Are we ready for autonomous driving? The kitti vision benchmark suite
20. Wang T, Gong S, Zhu X, Wang S (2014) Person re-identification by video ranking. Springer, Berlin, pp 688–703
21. Milan A, Leal-Taixé L, Reid I, Roth S, Schindler K (2016) Mot16: a benchmark for multi-object tracking. arXiv:1603.00831
22. Chavdarova T et al (2018) Wildtrack: a multi-camera hd dataset for dense unscripted pedestrian detection, pp 5030–5039
23. Li M, Zhu X, Gong S (2019) Unsupervised tracklet person re-identification. IEEE Trans Pattern Anal Mach Intell 42(7):1770–1782
24. Luiten J et al (2020) Hota: a higher order metric for evaluating multi-object tracking. Int J Comput Vis: IJCV 129:548–578
25. Fabbri M et al (2021) Motsynth: how can synthetic data help pedestrian detection and tracking?, pp 10849–10859
26. Peng J et al (2020) Tpm: multiple object tracking with tracklet-plane matching. Pattern Recogn 107:107480
27. Wu Q, Dai P, Chen P et al (2021) Deep adversarial data augmentation with attribute guided for person re-identification. Signal Image Video Process 15:655–662. https://doi.org/10.1007/s11760-019-01523-3
28. Nousi P, Triantafyllidou D, Tefas A, Pitas I (2020) Re-identification framework for long term visual object tracking based on object detection and classification. Signal Process Image Commun 88:115969
29. Bergmann P, Meinhardt T, Leal-Taixé L (2019) Tracking without bells and whistles. CoRR arXiv:1903.05625
30. Yu T, Li D, Yang Y, Timothy H, Xiang T (2019) Robust person re-identification by modelling feature uncertainty. In: Proceedings of the IEEE international conference on computer vision, pp 552–561
31. Chen A, Biglari-Abhari M, Wang K (2019) Investigating fast re-identification for multi-camera indoor person tracking. Comput Electr Eng 77:273–288
32. Li Y, Liu L, Zhu L, Zhang H (2021) Person re-identification based on multi-scale feature learning. Knowl Based Syst 228:107281
33. Lin Y, Xie L, Wu Y, Yan C, Tian Q (2020) Unsupervised person re-identification via softened similarity learning. CoRR arXiv:2004.03547
34. Mansouri N, Ammar S, Kessentini Y (2021) Re-ranking person re-identification using attributes learning. Neural Comput Appl 33:12827–12843
35. Zheng L, Shen L, Tian L, Wang S, Wang J, Tian Q (2015) Scalable person re-identification: a benchmark. In: Proceedings of the IEEE international conference on computer vision, pp 1116–1124
36. Ristani E, Solera F, Zou RS, Cucchiara R, Tomasi C (2016) Performance measures and a data set for multi-target, multi-camera tracking. In: European conference on computer vision. Springer, Cham, pp 17–35
37. Liao L et al (2020) A half-precision compressive sensing framework for end-to-end person re-identification. Neural Comput Appl 32(4):1141–1155
38. Zheng L, Zhang H, Sun S, Chandraker M, Tian Q (2016) Person re-identification in the wild. arXiv:1604.02531
39. Zhou S, Wang Y, Zhang F, Wu J (2021) Cross-view similarity exploration for unsupervised cross-domain person re-identification. Neural Comput Appl 33(9):4001–4011
40. Zhu X, Jing X-Y, Ma F, Cheng L, Ren Y (2019) Simultaneous visual-appearance-level and spatial-temporal-level dictionary learning for video-based person re-identification. Neural Comput Appl 31(11):7303–7315
41. Hirzer M, Beleznai C, Roth PM, Bischof H (2011) Person re-identification by descriptive and discriminative classification. In: Scandinavian conference on image analysis. Springer, Berlin, Heidelberg, pp 91–102
42. Zhang J et al (2020) Multiple object tracking by flowing and fusing. *CoRR*arXiv:2001.11180
43. Wang Y, Weng X, Kitani K (2020) Joint detection and multi-object tracking with graph neural networks. *CoRR*arXiv:2006.13164
44. Meinhardt T, Kirillov A, Leal-Taixé L, Feichtenhofer C (2021) Trackformer: Multi-object tracking with transformers. CoRR arXiv:2101.02702
45. Shuai B, Berneshawi AG, Modolo D, Tighe J (2020) Multi-object tracking with siamese track-rcnn. CoRR arXiv:2004.07786
46. Meimetis D, Daramouskas I, Perikos I, Hatzilygeroudis I (2021) Real-time multiple object tracking using deep learning methods. Neural Comput Appl. https://doi.org/10.1007/s00521-021-06391-y
47. Yang K, Song H, Zhang K, Liu Q (2020) Hierarchical attentive siamese network for real-time visual tracking. Neural Comput Appl 32(18):14335–14346
48. Wu Y, Lim J, Yang MH (2013) Online object tracking: a benchmark. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 2411–2418
49. Huang G, Liu Z, Van Der Maaten L, Weinberger KQ (2017) Densely connected convolutional networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 4700–4708
50. Huang G, Liu Z, Pleiss G, Van Der Maaten L, Weinberger K (2019) Convolutional networks with dense connectivity. IEEE Trans Pattern Anal Mach Intll. https://doi.org/10.1109/TPAMI.2019.2918284
51. Redmon J, Farhadi A (2018) Yolov3: an incremental improvement. arXiv:1804.02767
52. Yu L, Zhao Y, Zheng X (2021) Towards real -time object tracking with deep siamese network and layerwise aggregation. Signal Image Video Process 15:1303–1311. https://doi.org/10.1007/s11760-021-01861-1
53. Li S, Zhao Z, Kou L, Zhou Z, Xia G-S (2020) Siamese networks with distractor-reduction method for long-term visual object tracking. Pattern Recogn 112:107698. https://doi.org/10.1016/j.patcog.2020.107698
54. Bayraktar E, Boyraz P (2017) Analysis of feature detector and descriptor combinations with a localization experiment for various performance metrics. Turki J Electr Eng Comput Sci 25(3):2444–2454
55. Bayraktar E, Basarkan ME, Celebi N (2020) A low-cost uav framework towards ornamental plant detection and counting in the wild. ISPRS J Photogramm Remote Sens 167:1–11
56. Bochkovskiy A, Wang C-Y, Liao H-YM (2020) Yolov4: optimal speed and accuracy of object detection. arXiv:2004.10934
57. Jocher G et al (2020) ultralytics/yolov5: v3.1—bug fixes and performance improvements. https://doi.org/10.5281/zenodo.4154370
58. Zheng L et al (2015) Scalable person re-identification: a benchmark, pp 1116–1124. https://doi.org/10.1109/ICCV.2015.133
59. Li W, Zhao R, Xiao T, Wang X (2014) Deepreid: deep filter pairing neural network for person re-identification, pp 152–159. https://doi.org/10.1109/CVPR.2014.27
60. Ciaparrone G et al (2020) Deep learning in video multi-object tracking: a survey. Neurocomputing 381:61–88

61. Khalkhali MB, Vahedian A, Yazdi HS (2019) Multi-target state estimation using interactive kalman filter for multi-vehicle tracking. IEEE Trans Intell Transp Syst 21(3):1131–1144
62. Li X, Wang K, Wang W, Li Y (2010) A multiple object tracking method using kalman filter. Piscataway, IEEE, pp 1862–1866
63. Arulampalam MS, Maskell S, Gordon N, Clapp T (2002) A tutorial on particle filters for online nonlinear/non-gaussian bayesian tracking. IEEE Trans Signal Process 50(2):174–188
64. Smal I, Draegestein K, Galjart N, Niessen W, Meijering E (2008) Particle filtering for multiple object tracking in dynamic fluorescence microscopy images: application to microtubule growth analysis. IEEE Trans Med Imaging 27(6):789–804
65. Cui Y, Zhang J, He Z, Hu J (2019) Multiple pedestrian tracking by combining particle filter and network flow model. Neurocomputing 351:217–227
66. Babaee M, Athar A, Rigoll G (2018) Multiple people tracking using hierarchical deep tracklet re-identification. arXiv:1811.04091
67. Fu Z, Angelini F, Chambers J, Naqvi S (2019) Multi-level cooperative fusion of gm-phd filters for online multiple human tracking. IEEE Trans Multimed 21:2277–2291. https://doi.org/10.1109/TMM.2019.2902480
68. Xu Y, Osep A, Ban Y, Horaud R, Leal-Taixé L, Alameda-Pineda X (2020) How to train your deep multi-object tracker. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 6787–6796
69. Ren W, Wang X, Tian J, Tang Y, Chan AB (2021) Tracking-by-counting: using network flows on crowd density maps for tracking multiple targets. IEEE Trans Image Process 30:1439–1452. https://doi.org/10.1109/TIP.2020.3044219
70. Papakis I, Sarkar A, Karpatne A (2020) Gcnnmatch: graph convolutional neural networks for multi-object tracking via sinkhorn normalization. CoRR arXiv:2010.00067
71. Wang G, Wang Y, Gu R, Hu W, Hwang J (2021) Split and connect: a universal tracklet booster for multi-object tracking. CoRR arXiv:2105.02426
72. Dai P et al (2021) Learning a proposal classifier for multiple object tracking. CoRR arXiv:2103.07889
73. Smeulders AW et al (2013) Visual tracking: an experimental survey. IEEE Trans Pattern Anal Mach Intell 36(7):1442–1468
74. Valmadre J et al (2021) Local metrics for multi-object tracking. arXiv:2104.02631
75. Kaplan EL, Meier P (1958) Nonparametric estimation from incomplete observations. J Am Stat Assoc 53(282):457–481
76. Luiten J et al (2021) Hota: a higher order metric for evaluating multi-object tracking. Int J Comput Vis 129(2):548–578