Attention-Enabled Object Detection to Improve One-Stage Tracker

Madan, Neelu; Nasrollahi, Kamal; Moeslund, Thomas B.

Published in:
Intelligent Systems and Applications

DOI (link to publication from Publisher):
10.1007/978-3-030-82196-8_55

Publication date:
2022

Citation for published version (APA):
Madan, N., Nasrollahi, K., & Moeslund, T. B. (2022). Attention-Enabled Object Detection to Improve One-Stage Tracker. In K. Arai (Ed.), *Intelligent Systems and Applications: Proceedings of the 2021 Intelligent Systems Conference (IntelliSys) Volume 1* (Vol. 2, pp. 736-754). Springer. https://doi.org/10.1007/978-3-030-82196-8_55

Attention-Enabled Object Detection to Improve One-Stage Tracker

Neelu Madan[1], Kamal Nasrollahi[1,2], and Thomas B. Moeslund[1]

[1] Visual Analysis and Perception, Aalborg University, Rendsburggade 14, Aalborg, Denmark,
[2] Research Department, Milestone Systems A/S, Brøndby, Denmark
nema@create.aau.dk

Abstract. State-of-the-art (SoTA) detection-based tracking methods mostly accomplish the detection and the identification feature learning tasks separately. Only a few efforts include the joint learning of detection and identification features. This work proposes two novel one-stage trackers by introducing implicit and explicit attention to the tracking research topic. For our tracking system based on implicit attention, we further introduce a novel fusion of feature maps combining information from different abstraction levels. For our tracking system based on explicit attention, we introduce utilization of an additional auxiliary function. These systems outperform the SoTA tracking systems in terms of MOTP (Multi-Object Tracking Precision) and IDF1 score when evaluated on public benchmark datasets including MOT15, MOT16, and MOT17. High MOTP score indicates precise detection of bounding boxes of objects, while high IDF1 score indicates accurate ID detections, which is very crucial for surveillance and security systems. Therefore, proposed systems are good choice for event-detections in surveillance feeds as we are capable of detecting correct ID and precise location.

Keywords: One-stage Trackers, Deep Neural Networks, Tracking-by-Detection, Computer Vision

1 Introduction

Multi Object Tracking (MOT) is one of the most widely used and yet challenging applications of computer vision. The aim is to predict the object trajectories across the video frames. The predicted trajectories could further be used for the analysis of the sports videos [20], anomaly detection in crowded scenes [1], Automatic Driving Assistance Systems [17], to name a few.

The challenges related to visual object tracking such as occlusion, motion blur, social interaction, and low-resolution images make it a complicated task. Myriads of approaches are proposed overtime to resolve the multiple challenges. Some of the earliest efforts in this field used correlation-filter [9], and mean-shift algorithm [11]. Later on, deep learning-based tracking approaches become popular due to an increase in GPU computation power. Some of the widely used deep learning based tracking approaches include Siamese Neural Network (SNN) [4] which learns similarity function between target and search region, Recurrent Neural Networks (RNN) [32] which incorporate temporal features in addition to appearance based features, and Generative Adversarial Network (GAN) [39] based tracker which works well even if training dataset is small.

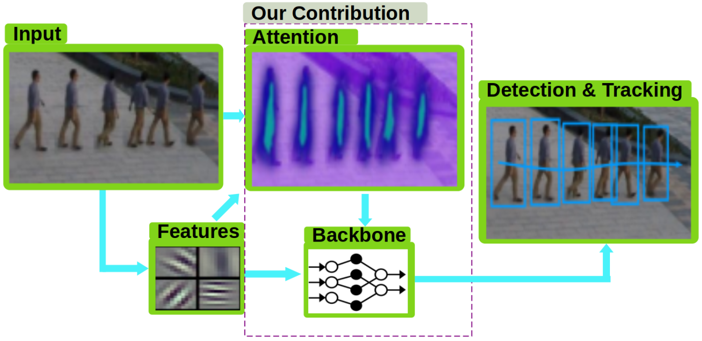

Fig. 1. Illustrating the proposed tracking system.

The success of deep neural networks in object detection paved the way for developing trackers using detectors in their backbone also know as tracking-by-detection. Some research in this area includes two-stage and one-stage trackers. In a two-stage trackers, target objects are detected as Bounding Boxes (BBoxes) by an object detector in the first stage. In the second stage, a unique identification (ID) is assigned to each BBox using a different network. Then, usually an affinity/cost matrix is constructed by combining Intersection over Union (IoU) of detection BBox and the assigned ID. Once the affinity matrix is created, an algorithm like Hungarian [27] is used to associate the target object by minimizing the total cost function. Reliable tracking accuracy can be achieved by using the best detection and identification networks in a two-stage tracker. However, this also increases the training complexity of the network and its inference time. To reduce the training effort, one-stage trackers [44] performing both the detection and the identification task in a single network using multi-task learning are proposed. In multi-task learning [26], a shared network is used to learn multiple tasks, here detection and re-identification (re-id). However, this joint learning reduces the network's generalization capabilities by sharing the same low-level and high-level features for both tasks. Therefore, the accuracy of the one-stage trackers is lower as compared to their two-stage counterparts. On further analysis, the unstable detection at small scale seems to be one of the major causes of decreased accuracy of one-stage trackers. This paper improves the detection accuracy of one-stage trackers adding different attention modules to the backbone architecture. Additionally, the proposed work also improved the identity detection by enhancing the feature propagation.

The accuracy of detection-based one-stage trackers can be improved by either improving the detection task or association-task. State-of-The-Art (SoTA) approaches such as Chained-Tracker [34] improves the data-association by incorporating task-specific attentions. In this paper, we focused on improving the detection task and modified existing backbone architectures of a known network that has been used mostly for object detection, HRNet [43], by introducing different attention modules to propose a novel one-stage tracker as shown in Figure 1. The concept of attention networks [38] is motivated by the human visual system, which learns to focus on a relevant region of the image while discarding the irrelevant part. In the past few years, attention-based networks have proved to be very successful in improving the performance of multiple computer vision tasks [23] and [15].

Attention mechanisms can be broadly classified into two categories, i.e., implicit [21] and explicit attention [29]. Implicit attention doesn't enforce any additional constraints on the attention unit. On the contrary, learning attention units explicitly with additional supervision from ground-truth (GT) enforces the network to improve on certain tasks. In this paper, we have introduced attention (in both forms of implicit [16] and explicit [28] attention) to one-stage trackers and shown that this improves the accuracy of the trackers.

The contributions of this paper are as follows:

1. We, for the first time, introduce attention, in its two forms of implicit and explicit attentions, to one-stage trackers based on multi-task learning, resulting in two one-stage trackers (one based on implicit attention and one based on explicit attention) which outperform SoTA trackers in terms of MOTP and IDF1 score on public benchmark datasets, while achieving competitive results on MOTA score. The proposed attention-gates improves the feature propagation across the later stages of the network and hence the ID detection.
2. For our tracker based on implicit attention, we propose a novel fusion of feature maps combining information coming from different abstraction levels, that improves the accuracy of the tracker even further (Section 5, Table 4).
3. For our tracker based on explicit attention, we propose a novel modification to our employed backbone architecture via an auxiliary loss function, that further improves the tracking accuracy (Section 5, Table 5).

The rest of this paper is organized as follows: The related work in the literature is reviewed in the next section. Then, Section 3 explains the details of the proposed idea. Section 4, gives the details of the experimental results comparing the performance of the system with the SoTA systems. Section 5 presents the ablation study, and finally the paper is concluded in Section 6.

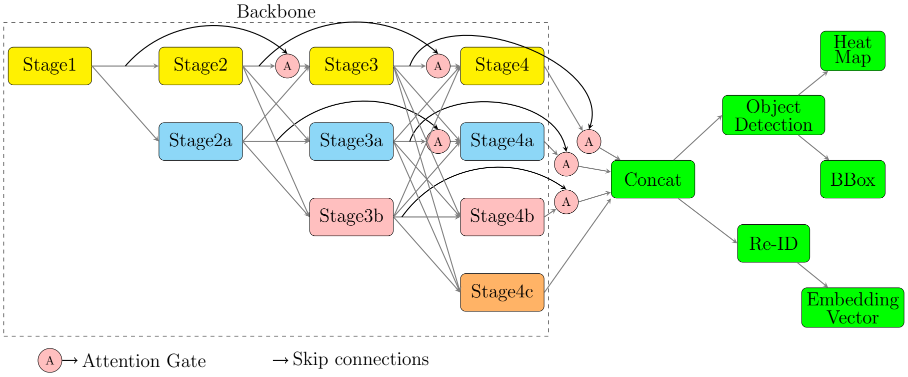

Fig. 2. Introducing implicit attention to a one-stage tracker based on HRNet [43] backbone.

2 Related Work

This section is divided into two parts. First, we review SoTA MOT approaches focusing on detection-based trackers. Then, we review the reported incorporation of implicit and explicit attention for improving the performance of deep learning-based computer vision methods.

2.1 Deep Learning-based Tracking

Most of the recent research works in MOT are using deep-learning based architectures. The recent major contributions include Graph Convolutional Network (GCN) [12], multi-task learning [41], articulated tracking [24], and tracking-by-detection [49]. SoTA GCN-based trackers consist of two different GCNs, i.e., spatial-temporal GCN represents the structure and contextual GCN models the context. Multi-task learning-based architectures jointly optimizing on different related tasks such as: Multi-Object Tracking and Segmentation (MOTS) [41] optimize detection, tracking and segmentation tasks, JDE [44] optimizing detection and re-id tasks, and FAMNet [5] which is based on joint optimization of feature extraction, affinity estimation, and multi-dimension assignment. Articulated trackers track the key-points across the video frames, e.g., architecture proposed in [24] contains two networks SpatialNet and TemporalNet. SpatialNet detects the body parts and group them in a single frame and TemporalNet converts those key-points into trajectories. Tracking-by-detection uses detected BBox to compute the affinity matrix, e.g., POI [50] incorporated an accurate object detector based on Faster-RCNN for pedestrian tracking. CenterNet [54] and Chained Tracker [34] proposed an integrated system performing simultaneous detection and association using consecutive frames as input.

Large scale image recognition challenges such as ImageNet [6] are bringing up accurate and efficient object detectors, which oriented the research direction towards detector-based trackers [49]. There are multiple tracking approaches based on different object detection architecture such as YOLO [35], Faster-RCNN [36], and SSD [30]. Different association techniques discussed further are used to establish the temporal relationship and predict the trajectories of target objects. Deep-association is proposed in [22], where a separate deep neural network is used to accomplish the association task. A tubelet proposal network [25] incorporated temporal features using Long Short-Term Memory (LSTM) networks. Detection-based tracking approaches mentioned above are computationally expensive and suffer from increased inference time. To meet the real-time constraint, Tracktor++ [2] transformed object detector to a tracker by simply using regression and classification differently. This simple conversion improved the inference time but failed to improve accuracy due to the missing temporal relationships. To further improve the idea, one-stage trackers [44] which jointly learn detection for spatial and feature embedding for the temporal association are proposed.

2.2 Attention Gates

Attention gates have were initially primarily used in natural language processing (NLP), e.g., hand-writing synthesis [13], and machine translation [14] to improve the contextual information in a text. More recently attention networks became a popular means to improve various computer-vision tasks such as image classification [15], and segmentation [10]. Not every feature generated by a deep neural network is equally important. Therefore, channel attention providing weights to different output channels has been applied in such cases. For example, squeeze-and-excitation networks proposed in [18] learn the channel weights by squeezing the spatial dimension. Additionally, few locations in the feature-maps are more relevant than the other. Therefore, spatial attention [42] generates the attention map by utilizing the inter-spatial relationship of features. Some networks such as Convolutional Block Attention Module (CBAM) [47] have incorporated both channel and spatial attention in a single network.

Attention maps can be learned implicitly using self-guidance from the network itself or explicitly using external supervision. Guided inference network [28] learns attention explicitly by adding an auxiliary loss function. In general, it has been observed that the attention generated by additional supervision usually performs better as also mentioned in [28]. To the best of our knowledge, attention mechanisms either implicit or explicit have not previously been used for accuracy improvement of one-stage tracker.

3 The Proposed Idea

This section discusses the details of the proposed idea of introducing attention to one-stage trackers, resulting in two systems, one using implicit and the other one explicit attention. This section also includes the details about similarities and differences between proposed explicit and implicit attention-based networks. We

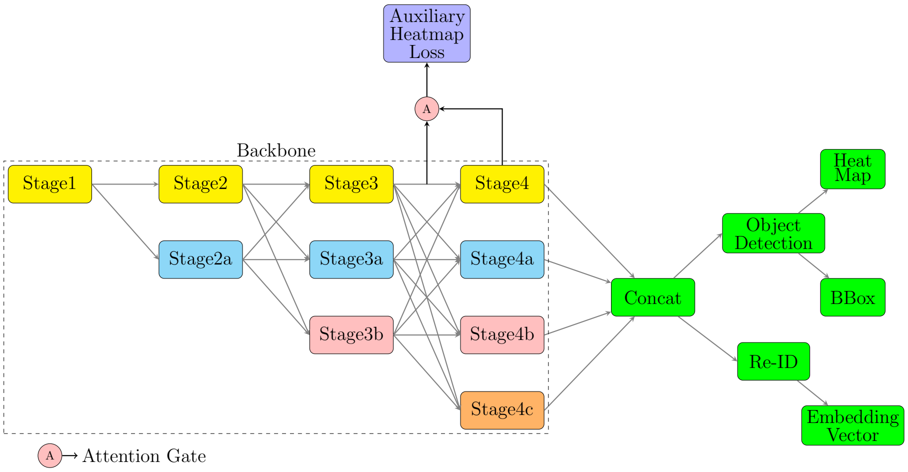

Fig. 3. Introducing explicit attention to a one-stage tracker based on HRNet [43] backbone.

have introduced our attention idea to the HRNet [43] architecture as it maintains high-resolution throughout the network unlike Feature Pyramid Net (FPN) [35] based architecture, which consist of an encoder branch to reduce the resolution of feature-maps from high to low and a decoder branch to recover the high-resolution information. This property of HRNet [43] provided us with the scope for improving the detection accuracy for small objects by adding different attention modules. The following subsections present the proposed building blocks, network architectures, objective functions, and inference details of the two systems based on the different mentioned attentions.

3.1 Components

The basic building blocks for both implicit and explicit attention networks are the same. This section contains the details for the common blocks of the two networks as show in Figure 2 (implicit-based) and Figure 3 (explicit-based).

Attention-gate A basic function of any gate is to allow some information to pass through it while blocking the rest. The attention gates used in this paper are similar to the one proposed in [33]. The idea of the proposed attention gates is to filter the irrelevant features and prevent them from propagating across the network, which in turn improves the gradient flow of the network.

Backbone HRNet [43] typically contains four stages, where a lower resolution is added to the network at the end of each stage. The basic structure of each stage remains unchanged as in original HRNet [43], i.e., each stage contains parallel multi-resolution convolution followed by multi-resolution fusion. For example, Stage 2 of HRNet [43] is shown in Figure 4, other stages are designed in a similar way. This repeated addition of new resolution and multi-resolution fusion increases the amount of global information after each stage [43]. The multi-resolution feature maps are concatenated and provided to the predictions heads, i.e., object detection and BBox estimation, as shown in Figure 2 and 3. These figures are based on generic block diagram of one-stage trackers [44].

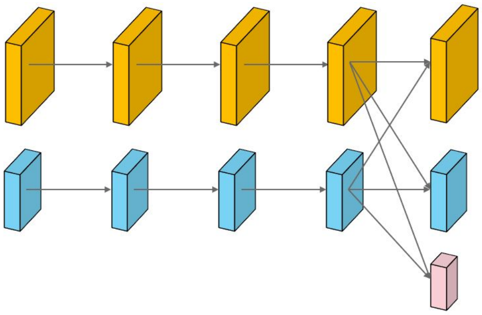

Fig. 4. Stage 2 in the backbone of proposed architecture contains parallel convolution for two different resolution(yellow and blue) followed by the addition of new resolution (pink) and multi-resolution fusion represented by cross-connections.

Object Detection The object detection branch predicts two different outputs, i.e., heatmap and BBoxes' size and center offset. These are explained here:

Heatmap Estimation Heatmaps are mostly used in the context of key-point estimation [40]. We employ them to get the estimated position of the object's center and the most probable detection areas. The size of the predicted heatmap is the same as the input image but it contains only one channel. The output response of the predicted heatmap is expected to be one at the object center and decays exponentially with increase in distance from the center.

BBox Estimation BBox estimation predicts the size and offset for the target objects. The size of a BBox corresponds to its height and width around the probable regions proposed by the predicted heatmaps and offset to the object's center. The precise localization of the object has a significant role in a tracking system. The reason is, re-id features are extracted on the basis of the object's center. Therefore, accuracy in locating object's center needs to be high in order to improve the tracking system.

Re-Identification The re-id branch generates the embedding vector, which distinguishes among different target objects and helps in predicting their corresponding tracks. The re-id branch learns a metric such that instances of the same identity are close to each other, and instances of different identities are far apart. To achieve the same, a convolutional layer predicting a 1-dimensional embedding vector of size 128 is introduced, which shares the same features as used for object detection. The embedding feature vector, corresponding to an object with center at (x,y) and is extracted from the feature map, which is finally used for association across the subsequent frames.

3.2 Network Architectures

Until now, we have discussed the structure of all the basic components, which are the same in both explicit and implicit attention networks. However, there are still some basic differences between their architectures which are described in this sub-section.

Implicit Attention The architecture of the proposed implicit attention-based network is shown in Figure 2. Higher resolution in the HRNet [43] goes through a series of convolutional layers and keeps on adding more global information after each stage. Due to this long series of convolutions, the feature-map towards the end of the network tends to lose the information about the local context. To recover this information, attention-gated skip connections between the consecutive stages are proposed in the current research work. The reason for fusing the different feature-maps between the consecutive stages is discussed along with some experiments in Section 5. This attention-gated fusion of features weighs the lower-level features from the previous stage to complement the global features at the current stage and finally combines the information. Such connections are repeated for each resolution in the proposed architecture. This type of attention is called implicit because the attention-gate automatically learns to weigh the local context based on the existing global context.

Explicit Attention The network architecture of the proposed explicit attention-based network is shown in Figure 3. We have introduced an auxiliary heatmap loss to our proposed explicit attention type which provides additional supervision. In this attention-type, a heatmap is predicted from the attention-gated feature at an intermediate stage. The attention gates here filter the feature maps at an intermediate stage based on the global information. An auxiliary loss between the predicted and ground-truth (GT) heatmap is calculated using the attention-gated feature-maps and later added to the objective function which is used during the network's training. The choice of the intermediate stage to add an auxiliary loss is decided experimentally and discussed in Section 5.

3.3 Loss Functions

This section contains the details of the overall objective which is obtained by combining multiple loss functions. The implicit and explicit attentions combine three and four types of different losses, respectively. The first three losses in explicit architecture are the same as that for implicit and acquired from [52]. Further details of the objective function are mentioned below.

Heatmap Loss Object center (c_x^k, c_y^k) is computed as $c_x^k = \frac{x_1^k + x_2^k}{2}$, and $c_y^k = \frac{y_1^k + y_2^k}{2}$ for each GT BBox $b^k = (x_1^k, y_1^k, x_2^k, y_2^k)$. The location of center on the feature map is obtained by dividing the stride, i.e., $(\tilde{c}_x^k, \tilde{c}_y^k) = \left(\left\lfloor \frac{c_x^k}{4} \right\rfloor, \left\lfloor \frac{c_y^k}{4} \right\rfloor \right)$. Heatmap response at location (x,y) is computed by using Gaussian distribution ,i.e. , H_{xy} $= \sum_{k=1}^{N} \exp^{\frac{(x-\tilde{c}_x^k)^2 + (y-\tilde{c}_y^k)^2}{2\sigma_c^2}}$, which shows that response is decreasing with increase in distance from the center. The loss function used for the regression of heatmap is pixel-wise logistic regression with focal loss. It is represented via Equation 1 [52]:

$$\mathcal{L}_{hm} = -\frac{1}{N} \sum_{xy} \begin{cases} (1 - \hat{H_{xy}})^{\alpha} \log(\hat{H_{xy}}), & \text{if } H_{xy} = 1; \\ (1 - H_{xy})^{\beta} (\hat{H_{xy}})^{\alpha} \log(1 - \hat{H_{xy}}) & \text{otherwise,} \end{cases} \tag{1}$$

Offset and Size Loss The predictions from the offset and size head is denoted as $C_p \in \mathcal{R}^{W \times H \times 2}$ and $S_p \in \mathcal{R}^{W \times H \times 2}$ respectively, the size of single GT BBox with coordinates as $(x_1^k, y_1^k, x_2^k, y_2^k)$ is computed as $S = (x_2^k - x_1^k, y_2^k - y_1^k)$ and center offset is calculated as C = $C/4$. The detection loss of a single BBox is the sum of deviations in size and center offset. Total loss is obtained by the summation of K detected boxes as represented by equation 2 [52]:

$$\mathcal{L}_{bbox} = \sum_{k=1}^{K} |C^k - C_p^k| + |S^k - S_P^k| \tag{2}$$

Re-ID Loss The ID prediction is considered a classification task where the number of classes corresponds to the number of IDs in the dataset, i.e., objects with the same ID are treated as one class. To calculate the Re-ID loss for each BBox (B_k), the softmax function is applied to the predicted embedding vector to get the class distribution $P(\mathcal{N})$. GT can be represented as one-hot vector $GT^k(\mathcal{N})$. The loss function used for this case is a cross-entropy loss for multi-class classification as shown in Equation 3 [52].

$$\mathcal{L}_{id} = -\sum_{k=1}^{K} \sum_{n=1}^{\mathcal{N}} GT^k(\mathcal{N}) log(P(\mathcal{N})) \tag{3}$$

where $\mathcal{N}$ and K are total number of classes and detected BBoxes respectively.

Auxiliary Loss In case of explicit attention, a heatmap is predicted from intermediate feature-maps. An auxiliary loss is added to optimize the heatmap prediction from an intermediate layer. The calculation of this auxiliary loss is similar to that of Equation 1.

Loss Balancing based on Uncertainty of Tasks The overall objective of the networks is a weighted sum of the above-mentioned losses. However, the manual tuning of those weights is computationally expensive and difficult. Therefore, we used **automatic loss balancing** as proposed in [26] which uses task uncertainty to weigh the different losses. The total loss therefore can be represented via Equation 4 [52]. Weights $\mathcal{W}_{hm}, \mathcal{W}_{bbox}, \mathcal{W}_{id}$ are learned automatically as part of neural network learning process, as follows:

$$\mathcal{L}_{total}^{Implicit} = \frac{1}{2}\bigg(\frac{1}{e^{\mathcal{W}_{hm}}}\mathcal{L}_{hm} + \frac{1}{e^{\mathcal{W}_{bbox}}}\mathcal{L}_{bbox} + \frac{1}{e^{\mathcal{W}_{id}}}\mathcal{L}_{id} + s_{hm} + s_{bbox} + s_{id}\bigg) \tag{4}$$

where $\mathcal{L}_{hm}$, $\mathcal{L}_{bbox}$, and $\mathcal{L}_{id}$ are heatmap, BBox, ID loss respectively and s_{hm}, s_{bbox}, and s_{id} are task dependent uncertainties. In case of explicit attention-based architecture, total loss will be modified by adding an auxiliary loss ($\mathcal{L}_{aux}$) balanced via weight ($\mathcal{W}_{aux}$) and having task-dependent uncertainty of s_{aux} as in Equation 5:

$$\mathcal{L}_{total}^{Explicit} = \frac{1}{2}\bigg(\frac{1}{e^{\mathcal{W}_{hm}}}\mathcal{L}_{hm} + \frac{1}{e^{\mathcal{W}_{bbox}}}\mathcal{L}_{bbox} + \frac{1}{e^{\mathcal{W}_{id}}}\mathcal{L}_{id} + \frac{1}{e^{\mathcal{W}_{aux}}}\mathcal{L}_{aux} + s_{hm} + s_{bbox} + s_{id} + s_{aux}\bigg) \tag{5}$$

3.4 Inference and Online Association

Giving an input image to either of the two systems illustrated in Figure 2 and 3, they generate an output that can be categorized into two parts, i.e., detection and re-id. The detection part is represented via heatmap, BBox size, and offset for the detection task. The re-id part is depicted by the embedding vector. A non-maximum suppression (NMS) is performed based on heatmap scores on top of the predicted heatmap, which provides the most probable detection locations. The locations with score greater than a certain threshold are kept and the rest are discarded, which is finally followed by the estimation of BBox's size and offset. The embedding vectors corresponding to the detections are also extracted. The next step is to associate the detected BBox and identity embeddings across the subsequent video frames.

Association is based on a standard online tracking algorithm as also discussed in [44]. Tracklets are first initialized based on the appearance features extracted from the first video frames and added to the tracklet pool. In the subsequent frames, pairwise motion and appearance similarity is calculated between the observations and tracklets from the pool. Metrics used for the association of appearance and motion-based features are Cosine similarity and Mahalanobis distance respectively. Finally, the assignment problem is resolved by using the Hungarian algorithm [27]. The appearance features are updated using a weighted combination of BBox IoU and embedding vector. However, motion cues are updated by using Kalman Filter [45]. The observations which are not assigned to any of the existing tracklets from the pool are marked new. On the contrary, tracklets are marked as terminated if no observation is found for a few subsequent frames.

4 Experimental Results

This section includes information about the evaluation metrics, dataset, and training methodology. In addition, this section also discusses the qualitative analysis, as well as compares our results with SoTA methods which have reported their results on the same benchmark datasets used in our experiments.

4.1 Evaluation Metrics and Dataset

The proposed system developed in this paper are compared using Multi Object Tracking Accuracy (MOTA) [3], MOTP [3], IDF1 Score [37], and number of ID Switches from CLEAR metric [3] as described below:

- Multi-object tracking accuracy (MOTA): Computes the overall tracking accuracy from false positives, false negatives and identity switches [3].
- Multi-object tracking precision (MOTP): Depicts overall tracking precision in terms of BBox overlap between ground-truth and predicted location [3].
- Identification F1 (IDF1): Measures the extent that predicted identities confront the ground-truth [37].
- Identity switches (ID): Number of times the reported identity of a ground-truth track changes.

The system is evaluated on **MOT15 train**, **MOT17 test**, and **MOT16 test** data, where **MOT15 train** data is used as validation set in our experiments. We are using the private protocol of MOT 16 and 17 under which we are allowed to use additional training data. We have therefore collected a set of additional training datasets including ETH dataset [8], CityPersons (CP) dataset [51], CalTech (CT) dataset [7], MOT17 (M17) dataset [31], CUHK-SYSU (CS) dataset [48], and PRW dataset [53]. The same datasets are used in the training of JDE [44], tracker that we have compared our results against. The joint dataset can be divided into two categories, i.e., ETH and CP contain annotations for detection only while CT, MOT17, CS, and PRW contain annotations for both detection and ID. The overlapping videos between the testing and training dataset are removed from the training data.

4.2 Implementation Details

The training of the network on a huge dataset is a arduous task and requires a lot of computational power and time. To improve the convergence speed, weights of the proposed models are initialized using a network pre-trained on the COCO dataset. Our network with implicit attention is trained for 60 epochs with a

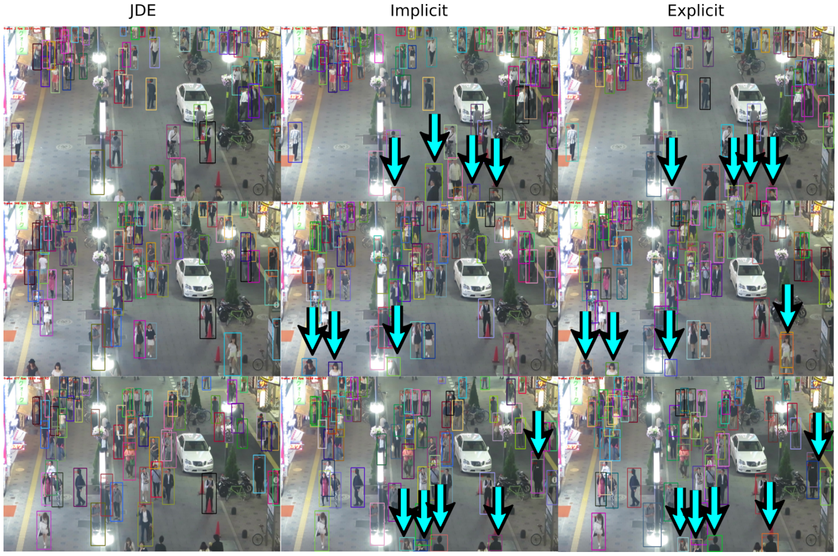

Fig. 5. Comparison among the tracklets detected and tracked in the proposed systems (Explicit and Implicit) and existing multi-task learning-based tracker, i.e., JDE [44]

Dataset	Tracker	MOTA↑	IDF1↑	IDs↓
MOT15 Train	JDE [44]	67.5	66.7	218
	Ours(Implicit)	73.2	73.5	**203**
	Ours(Explicit)	**73.8**	**75.9**	232

Table 1. Comparing the proposed trackers with the existing one-stage tracking approaches on MOT15 train dataset.

Tracker	Publication Year	MOT16				MOT17			
		MOTA↑	MOTP↑	IDF1↑	IDs↓	MOTA↑	MOTP↑	IDF1↑	IDs↓
DeepSORT [46]	ICIP 2017	61.4	79.1	62.2	**781**	60.3	79.1	61.2	**2442**
Tracktor+CTdet [2]	ICCV 2019	-	-	-	-	54.4	78.1	56.1	2574
JDE [44]	ICCV 2019	64.4	55.8	1881	-	-	-		
CenterNet [54]	ECCV 2020	-	-	-	-	**67.8**	-	64.7	3039
Chained-Tracker [34]	ECCV 2020	**67.6**	78.4	57.2	1897	66.6	78.2	57.4	5529
Ours(Implicit)	IntelliSys 2021	64.6	78.8	65.9	1234	63.2	78.7	64.8	3357
Ours(Explicit)	IntelliSys 2021	64.9	**79.7**	**66.4**	1489	63.7	**79.1**	**66.0**	3696

Table 2. Comparing the proposed systems against SoTA two-stage and one-stage trackers that have reported their results on MOT16 and MOT17.

batch size of 16 on 2 × Nvidia 1080 Ti GPUs and took ∼60 hrs to complete. However, our network with explicit attention is trained for 100 epochs with a batch size of 8 on 2 × Nvidia 1080 Ti GPUs and took ∼120 hrs. The collected training dataset is augmented by applying rotation, scaling and color jittering randomly to the input images similar to JDE [44].

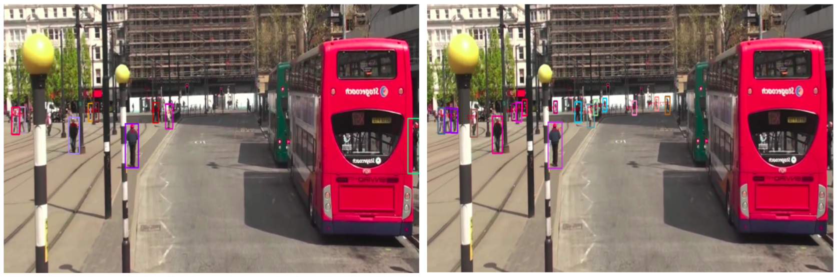

Fig. 6. left: results obtained from implicit attention-based tracker, **right**: results obtained via our explicit attention-based tracker. Small objects at the back are detected accurately using explicit attention (right), which are missing in left image.

4.3 Results and Discussion

Only a few published approaches, i.e., JDE [44], and Track-RCNN [41] are built on multi-task learning-based one-stage tracking. JDE [44] optimizes two tasks simultaneously, i.e., detection and re-id. However, Track-RCNN [41] jointly optimises detection, re-id, and segmentation task. The inclusion of extra task requires different training datasets containing annotations of segmentation in addition to detection, and re-id, which makes it incomparable with our tracker. We are comparing our method with trackers using private detections, e.g., JDE [44].

Quantitative Results Table 1 compares the existing one-stage tracker's accuracy with our systems using both explicit and implicit attention on MOT15 training dataset as also done in JDE [44]. It can be seen from the table, that our attention-based tracker outperforms the existing one-stage tracker, JDE [44] by a large margin both in terms of MOTA (6.3%) and IDF1 (9.2%).

Table 2 shows a comparison between the proposed approaches and other SoTA tracking methods reported on MOT challenge [31]. It can be observed from results that our tracker with explicit attention is performing the best in terms of MOTP and IDF1 scores on both MOT16 and MOT17 testing datasets. A high value of MOTP indicates precise localization of target object, while improvement in IDF1 score indicates better prediction of identities compared to Ground Truth (GT). Both ID preservation and precise localization are critical factors for designing a surveillance system as also stated by [37], which enables the proposed attention-gated tracker for security applications. On the other hand, the improvement in overall object detection in this framework increases the size of the affinity matrix, which in turn increases the ID switches (IDs) especially for the crowded sequences.

It can be observed from the quantitative results in Tables 1 and 2 that the tracking accuracy is improved by a large margin for the MOT15 validation dataset in comparison to MOT16 and MOT17 test data. The reason is that MOT15 validation sequences are less crowded in comparison to MOT16 and MOT17 test sequences. As discussed before, the proposed system improves the detection of small and partly occluded objects, which increases both the number of objects and the computational complexity of the association matrix. An increase in the size of the affinity matrix leads to multiple ambiguities and hence makes it difficult to optimize for the minimum cost. As a result, the MOTA of our proposed systems is comparatively lower on MOT16 and MOT17 test datasets than the MOT15 validation data containing fewer tracking objects.

Qualitative Results Figure 5 and 6 show the results of the proposed system using both explicit and implicit attentions. It can be clearly seen from the results in Figure 6 that the attention modules improve the detection even for small or partially visible objects. Furthermore, results in Figure 5 illustrate that the proposed attention-gated trackers are performing better in comparison to the existing one-stage tracker JDE [44] (also based on multi-task learning) for the objects approaching towards the boundaries of the frame. In the existing one-stage trackers like JDE [44], inconsistent detections for objects decreases the system's IDF1 score. The proposed systems improve object detection via incorporating attention-gates in the backbone network. However, it also increases the size of the affinity matrix, which further increases identity switches (IDs). One of the main benefits of using attention in the proposed system is that it also improves the embedding vector responsible for assigning an ID to a tracking object, which consequently improves the overall IDF1 of the proposed tracker.

5 Ablation Studies

This section contains the extended experiments required for adapting the base HRNet [43] to the proposed attention-gated architecture. The main experiments require a massive amount of computational time. In order to save computation time, all experiments in this section are trained on a small dataset, i.e., MOT17. The training on a small dataset provides a quick overview of different configurations setting required for explicit and implicit attention-based trackers. This section discusses the details about adding the attention-gates, feature fusion from different abstract levels, and novel modification of network architecture using explicit attention.

5.1 Attention-gates

The feature-maps at the highest resolution in HRNet [43] goes through a series of convolutional operations. As motivated by DenseNets [19], features from every stage are fused with each subsequent stage's features. This fusion can be a simple addition or attention-gated where low-level features are weighted based on high-level features. This helps in extracting only relevant information from the earlier stages. Experimental results in Table 3 compare the system's accuracy using HRNet [43], dense feature fusion without attention-gates, and attention-gated dense feature fusion. It can be observed from the results that dense feature fusion decreases the accuracy rather than showing any improvement. The main reason for this decrease in accuracy is that the combination of lowest and highest level features decreases the quality of features and hence the overall accuracy. The accuracy is improved by adding attention-gates to the dense-feature fusion but it is still lower than the HRNet [43]. The next section contains experiments to overcome this problem.

Backbone	MOTA↑	IDF1↑
HRNet	**79.2**	**72.9**
Ours(Implicit) + dense-connections + without attention-gates	76.5	72.8
Ours(Implicit) + dense-connections + with attention-gates	**79.2**	72.2

Table 3. The effect of adding attention-gates during the fusion of features.

5.2 Implicit Attention

The experiments in Table 3 depict that attention-gated fusion of features improves the accuracy, but dense connections decrease it. Therefore, further experiments are performed by reducing the dense skip connections to consecutive stages as also shown in Figure 2. In attention-gated dense fusion, features from the current stage are provided to all the subsequent stages at a single resolution. For example, feature maps from stage-1 are provided to stage-2, stage-3, and stage-4 at the highest resolution. However, the attention-gated fusion of features across the consecutive stages only combines the feature from current to the next subsequent stage, e.g., the connection between stage-1 and stage-2. It can be observed from the results obtained in Table 4 that attention-gated fusion of features across the consecutive stages improves the overall tracking accuracy.

Backbone	MOTA↑	IDF1↑
HRNet	79.2	72.9
Ours(Implicit) + dense-connections	79.2	72.2
Ours(Implicit) + consecutive-stages	**80.7**	**73.9**

Table 4. Comparing the tracking accuracy by adding attention gated dense-connections and consecutive-connection across the different stages.

5.3 Explicit Attention

The proposed explicit attention is guided by additional supervision by minimizing the auxiliary loss at an intermediate-stage as shown in Figure 3. This auxiliary loss minimizes the deviation between predicted and GT heatmaps. The heatmap is predicted using attention-gated features at different stages of the network. The results in Table 5 which shows adding an auxiliary loss after stage-1 has a negligible effect on overall tracking accuracy. As we shift the auxiliary loss towards the later stages of the network, the overall tracking accuracy starts increasing. This implies that the initial features of HRNet [43] do not contain any information about the global context and hence are incapable of making any predictions. The feature-maps get more relevant when the global context increases by an additional level after every stage. Additionally, this auxiliary loss is calculated only for the feature-maps generated from the high resolution because the lower resolutions are added at the later stages of the network. Furthermore, we would like to deal with inconsistencies in the detection of smaller objects which are detected at higher resolutions.

Backbone	MOTA↑	IDF1↑
HRNet	79.2	72.9
Ours(Explicit) + Stage 1	79.2	73.5
Ours(Explicit) + Stage 2	79.6	73.7
Ours(Explicit) + Stage 3	**81.8**	**74.8**

Table 5. Adding supervised attention and auxiliary losses after stage 1, 2 and 3 at highest resolution in HRNet [43].

6 Conclusion

In this paper, we proposed two one-stage trackers based on implicit and explicit attention mechanisms for multi-object tracking. The one based on implicit attention utilizes a novel fusion of feature maps for combining information extracted from different abstract levels, while the other based on explicit attention utilizes an auxiliary heatmap function that provides additional supervision for the attention mechanism. The latter tracker outperforms the former one, and both outperform state-of-the-art tracking systems in terms of MOTP and IDF1 scores when evaluated on public benchmark datasets. The main advantage of the proposed system over other SoTA approaches is that it detects small objects and the object IDs precisely, which makes it ideal for surveillance applications. We observed that our proposed attention-based architectures improve the tracking accuracy for less crowded scenes such as MOT15 sequences. For crowded sequences such as MOT16 and MOT17, the proposed attention-gated one-stage tracker improved the feature propagation and hence the object detection and identity-based measures such as IDF1 scores. However, to improve MOTA on MOT16 and MOT17, as a future work, we will work on developing an association block to reduce the identity switches.

7 Acknowledgements

This work was supported by the Milestone Research Programme at Aalborg University (MRPA).

References

1. Bera, A., Kim, S., Manocha, D.: Realtime anomaly detection using trajectory-level crowd behavior learning. In: CVPRW, pp. 1289–1296 (2016). DOI 10.1109/CVPRW.2016.163
2. Bergmann, P., Meinhardt, T., Leal-Taixe, L.: Tracking without bells and whistles. In: ICCV (2019)
3. Bernardin, K., Stiefelhagen, R.: Evaluating multiple object tracking performance: The CLEAR MOT metrics. EURASIP Journal on Image and Video Processing **2008**, 1–10 (2008). DOI 10.1155/2008/246309. URL https://doi.org/10.1155/2008/246309
4. Bertinetto, L., Valmadre, J., Henriques, F.J., Vedaldi, A., Torr, H.S.P.: Fully-convolutional siamese networks for object tracking. ECCV Workshops (2016)
5. Chu, P., Ling, H.: Famnet: Joint learning of feature, affinity and multi-dimensional assignment for online multiple object tracking. ICCV pp. 6171–6180 (2019)
6. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: ImageNet: A Large-Scale Hierarchical Image Database. In: CVPR (2009)
7. Dollár, P., Wojek, C., Schiele, B., Perona, P.: Pedestrian detection: A benchmark. In: CVPR (2009)
8. Ess, A., Leibe, B., Schindler, K., , van Gool, L.: A mobile vision system for robust multi-person tracking. In: CVPR. IEEE Press (2008)
9. Fitts, J.M.: Precision correlation tracking via optimal weighting functions. In: 1979 18th IEEE Conference on Decision and Control including the Symposium on Adaptive Processes, vol. 2, pp. 280–283 (1979)
10. Fu, J., Liu, J., Tian, H., Li, Y., Bao, Y., Fang, Z., Lu, H.: Dual attention network for scene segmentation. In: CVPR (2019)
11. Fukunaga, K., Hostetler, L.: The estimation of the gradient of a density function, with applications in pattern recognition. IEEE Transactions on Information Theory **21**(1), 32–40 (1975)
12. Gao, J., Zhang, T., Xu, C.: Graph convolutional tracking. In: CVPR, pp. 4644–4654 (2019)
13. Graves, A.: Generating sequences with recurrent neural networks. CoRR **abs/1308.0850** (2013). URL http://dblp.uni-trier.de/db/journals/corr/corr1308.html#Graves13
14. Graves, A., Wayne, G., Danihelka, I.: Neural turing machines. CoRR **abs/1410.5401** (2014). URL http://arxiv.org/abs/1410.5401
15. Guo, H., Zheng, K., Fan, X., Yu, H., Wang, S.: Visual attention consistency under image transforms for multi-label image classification. In: CVPR (2019)
16. He, C., Hu, H.: Image captioning with visual-semantic double attention. ACM Trans. Multimedia Comput. Commun. Appl. **15**(1) (2019). DOI 10.1145/3292058. URL https://doi.org/10.1145/3292058
17. Hoffmann, G.M., Tomlin, C.J., Montemerlo, M., Thrun, S.: Autonomous automobile trajectory tracking for off-road driving: Controller design, experimental validation and racing. In: American Control Conference, pp. 2296–2301 (2007). DOI 10.1109/ACC.2007.4282788

18. Hu, J., Shen, L., Sun, G.: Squeeze-and-excitation networks. In: CVPR (2018)
19. Huang, G., Liu, Z., van der Maaten, L., Weinberger, K.Q.: Densely connected convolutional networks. In: CVPR, pp. 2261–2269 (2017). URL http://dblp.uni-trier.de/db/conf/cvpr/cvpr2017.html#HuangLMW17
20. Huang, Y., Liao, I., Chen, C., İk, T., Peng, W.: Tracknet: A deep learning network for tracking high-speed and tiny objects in sports applications*. In: 16th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), pp. 1–8 (2019). DOI 10.1109/AVSS.2019.8909871
21. Huang, Z., Liang, S., Liang, M., Yang, H.: Dianet: Dense-and-implicit attention network. Proceedings of the AAAI Conference on Artificial Intelligence **34**(04), 4206–4214 (2020). DOI 10.1609/aaai.v34i04.5842. URL https://ojs.aaai.org/index.php/AAAI/article/view/5842
22. Jadhav, A., Mukherjee, P., Kaushik, V., Lall, B.: Aerial multi-object tracking by detection using deep association networks. In: 2020 National Conference on Communications (NCC), pp. 1–6 (2020). DOI 10.1109/NCC48643.2020.9056035
23. Jetley, S., Lord, N.A., Lee, N., Torr, P.: Learn to pay attention. In: International Conference on Learning Representations (2018). URL https://openreview.net/forum?id=HyzbhfWRW
24. Jin, S., Liu, W., Ouyang, W., Qian, C.: Multi-person articulated tracking with spatial and temporal embeddings. In: CVPR (2019)
25. Kang, K., Li, H., Xiao, T., Ouyang, W., Yan, J., Liu, X., Wang, X.: Object detection in videos with tubelet proposal networks. CVPR (2017). DOI 10.1109/cvpr.2017.101. URL http://dx.doi.org/10.1109/CVPR.2017.101
26. Kendall, A., Gal, Y., Cipolla, R.: Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In: CVPR (2018)
27. Kuhn, H.W., Yaw, B.: The hungarian method for the assignment problem. Naval Res. Logist. Quart pp. 83–97 (1955)
28. Li, K., Wu, Z., Peng, K.C., Ernst, J., Fu, Y.: Tell me where to look: Guided attention inference network. CVPR pp. 9215–9223 (2018)
29. Liu, C., Mao, J., Sha, F., Yuille, L.A.: Attention correctness in neural image captioning. AAAI (2017)
30. Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S.E., Fu, C., Berg, A.C.: SSD: single shot multibox detector. In: ECCV, vol. 9905, pp. 21–37 (2016). DOI 10.1007/978-3-319-46448-0_2. URL https://doi.org/10.1007/978-3-319-46448-0_2
31. Milan, A., Leal-Taixé, L., Reid, I., Roth, S., Schindler, K.: Mot16: A benchmark for multi-object tracking. arXiv preprint arXiv:1603.00831 (2016)
32. Ning, G., Zhang, Z., Huang, C., Ren, X., Wang, H., Cai, C., He, Z.: Spatially supervised recurrent convolutional neural networks for visual object tracking. IEEE International Symposium on Circuits and Systems (ISCAS) pp. 1–4 (2017)
33. Oktay, O., Schlemper, J., Folgoc, L.L., Lee, M.J., Heinrich, M., Misawa, K., Mori, K., McDonagh, S.G., Hammerla, N., Kainz, B., Glocker, B., Rueckert, D.: Attention u-net: Learning where to look for the pancreas. ArXiv **abs/1804.03999** (2018)
34. Peng, J., Wang, C., Wan, F., Wu, Y., Wang, Y., Tai, Y., Wang, C., Li, J., Huang, F., Fu, Y.: Chained-tracker: Chaining paired attentive regression results for end-to-end joint multiple-object detection and tracking. In: Proceedings of the European Conference on Computer Vision (2020)
35. Redmon, J., Divvala, S.K., Girshick, R.B., Farhadi, A.: You only look once: Unified, real-time object detection. CoRR **abs/1506.02640** (2015). URL http://arxiv.org/abs/1506.02640

36. Ren, S., He, K., Girshick, R., Sun, J.: Faster r-cnn: Towards real-time object detection with region proposal networks. In: International Conference on Neural Information Processing Systems, NIPS'15, p. 91–99 (2015)
37. Ristani, E., Solera, F., Zou, R.S., Cucchiara, R., Tomasi, C.: Performance measures and a data set for multi-target, multi-camera tracking. In: ECCV Workshops (2), pp. 17–35 (2016). URL https://doi.org/10.1007/978-3-319-48881-3_2
38. Schlemper, J., Oktay, O., Schaap, M., Heinrich, M., Kainz, B., Glocker, B., Rueckert, D.: Attention gated networks: Learning to leverage salient regions in medical images (2019)
39. Song, Y., Ma, C., Wu, X., Gong, L., Bao, L., Zuo, W., Shen, C., Lau, R.W.H., Yang, M.: Vital: Visual tracking via adversarial learning. In: CVPR, pp. 8990–8999 (2018). DOI 10.1109/CVPR.2018.00937
40. Sun, K., Geng, Z., Meng, D., Xiao, B., Liu, D., Zhang, Z., Wang, J.: Bottom-up human pose estimation by ranking heatmap-guided adaptive keypoint estimates (2020)
41. Voigtlaender, P., Krause, M., Osep, A., Luiten, J., Sekar, B.B.G., Geiger, A., Leibe, B.: Mots: Multi-object tracking and segmentation. In: CVPR, pp. 7934–7943 (2019). DOI 10.1109/CVPR.2019.00813
42. Wang, F., Jiang, M., Qian, C., Yang, S., Li, C., Zhang, H., Wang, X., Tang, X.: Residual attention network for image classification. In: CVPR (2017)
43. Wang, J., Sun, K., Cheng, T., Jiang, B., Deng, C., Zhao, Y., Liu, D., Mu, Y., Tan, M., Wang, X., Liu, W., Xiao, B.: Deep high-resolution representation learning for visual recognition. IEEE transactions on pattern analysis and machine intelligence (2020)
44. Wang, Z., Zheng, L., Liu, Y., Wang, S.: Towards real-time multi-object tracking. In: ECCV (2020)
45. Welch, G., Bishop, G.: An introduction to the kalman filter (1995)
46. Wojke, N., Bewley, A., Paulus, D.: Simple online and realtime tracking with a deep association metric. In: 2017 IEEE International Conference on Image Processing (ICIP), pp. 3645–3649 (2017). DOI 10.1109/ICIP.2017.8296962
47. Woo, S., Park, J., Lee, J.Y., Kweon, I.S.: Cbam: Convolutional block attention module. In: ECCV (2018)
48. Xiao, T., Li, S., Wang, B., Lin, L., Wang, X.: Joint detection and identification feature learning for person search. In: CVPR (2017)
49. Yu, F., Li, W., Li, Q., Liu, Y., Shi, X., Yan, J.: Poi: Multiple object tracking with high performance detection and appearance feature. ECCV Workshops (2016)
50. Yu, F., Li, W., Li, Q., Liu, Y., Shi, X., Yan, J.: Poi: Multiple object tracking with high performance detection and appearance feature. In: ECCV Workshops (2016)
51. Zhang, S., Benenson, R., Schiele, B.: Citypersons: A diverse dataset for pedestrian detection. In: CVPR (2017)
52. Zhang, Y., Wang, C., Wang, X., Zeng, W., Liu, W.: A simple baseline for multi-object tracking. arXiv preprint arXiv:2004.01888 (2020)
53. Zheng, L., Zhang, H., Sun, S., Chandraker, M., Yang, Y., Tian, Q.: Person re-identification in the wild. In: CVPR, pp. 1367–1376 (2017)
54. Zhou, X., Koltun, V., Krähenbühl, P.: Tracking objects as points. ECCV (2020)