Neurocomputing 481 (2022) 91–101

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

SegDQ: Segmentation assisted multi-object tracking with dynamic query-based transformers

Yating Liu [a,b], Tianxiang Bai [a,b], Yonglin Tian [c], Yutong Wang [a], Jiangong Wang [a,b], Xiao Wang [a,d], Fei-Yue Wang [a,*]

[a] The State Key Laboratory for Management and Control of Complex Systems, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China
[b] School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing 100049, China
[c] Department of Automation, University of Science and Technology of China, Hefei 230027, China
[d] Qingdao Academy of Intelligent Industries, Qingdao 266000, China

ARTICLE INFO

Article history:
Received 8 November 2021
Revised 7 January 2022
Accepted 18 January 2022
Available online 21 January 2022
Communicated by Zidong Wang

Keywords:
Multi-object tracking
Transformer
Semantic task
Dynamic query

ABSTRACT

Multi-Object Tracking (MOT) has been one of the most important topics in computer vision. The traditional tracking-by-detection framework of MOT is severely suffered from the poor detection results. In this paper, based on Transformer, we introduce the tracking-by-query MOT framework, and propose to apply semantic segmentation as an auxiliary task to optimize the training of MOT trackers, which addresses more on extracted foreground features. In addition, a feature-dependent dynamic object query (DOQ), instead of a fixed-learned object query (LOQ), is put forward to retrieve the new detections, improving the flexibility and constringency of the framework. We tested our SegDQ method on various scenarios including MOTChallenge 15, 16 and 17 datasets. The experimental results show that it obviously improves the MOTA and IDF1 indexes of tracking results.

1. Introduction

Multi-Object Tracking (MOT) has become an important topic in the field of computer vision since the 1960s [1–5]. It has been widely used in various fields including autonomous driving [6,7], safety monitoring [8,9], sports statistics, and behavior analysis [10,11], thus is of vital importance for improving machine intelligence. In brief, the goal of MOT is to simultaneously locate multiple targets in continuous video sequences, which is developed on the basis of object detection techniques. However, in daily scenarios, effects like occlusion, ambiguity, vibration and shading can lead to poor detecting results and therefore pose great challenges to MOT tasks.

To deal with these challenges, existing MOT methods mainly fall into the tracking-by-detection paradigm, which isolate the detection steps from matching steps. This paradigm has close relations to the re-identification (ReID) [12,13] problem. Although improvements are seen in literature that use feature engineering to obtain robust features for accurate associations, these methods fail to take advantage of present video information and therefore fail to adjust dynamically as scenario condition change.

In 2020, Zhu et al. introduced Transformer [14] from the field of Natural Language Processing (NLP) into object detection, called DEtection TRansformer (DETR) [15]. The method uses a query to get the potential location and individual representation of a candidate object. Besides, it also calculates the attention weights between a given position and all other pixels of the image to obtain the final foreground regions. Based on DETR, Sun et al. [16] introduce Transformer into MOT field named TransTrack. They use two parallel branches to get the objects in the last frame and correlate them with objects in the current frame. The track query is obtained to get the existing tracks while the learned object query is used to get the new detections. This query-key pattern falls out of the tracking-by-detection paradigm and integrates detection and matching steps into each other. The method brings benefits to global correlation that brought by attention mechanism and allows for end-to-end learning.

In this paper, based on TransTrack, we propose a segmentation aided auxiliary task that improves feature representation and generate dynamic object queries for detections. To be specific, we stretch a semantic mask branch out of the tracking model and pose a mask prediction task to guide Transformer to acquire the attention of pedestrian fore-ground features. This process is supervised by our modifications on the MOTS dataset [17]. In DETR, learned object queries (LOQ) are responsible for assigning potential regions

* Corresponding author.
E-mail address: feiyue@gmail.com (F.-Y. Wang).

https://doi.org/10.1016/j.neucom.2022.01.073
0925-2312/

to candidate detections. However, in TransTrack, the LOQs are learned from random initialization and keep fixed in the stage of reasoning, which means that learned LOQs designate fixed positions for candidate detections throughout every single frame in videos. Thus, we name LOQs the static queries. This ignorance of content differences between frames can lead to insufficiently accurate predictions. Instead, we use a feature-based query generation approach that takes into account frame information to train feature dependent dynamic object queries (DOQ). The learned DOQs help to improve Transformer in predicting more accurate foreground candidates.

In summary, our main contributions are listed as follows:

- First, we propose a Transformer-based multi-object tracking framework with feature dependent queries and assisted segmentation task to help solve the MOT problem.
- Second, we extend the TransTrack method to a multi-task MOT method by posing semantic segmentation as an auxiliary task in addition to the original MOT problem to enhance foreground extraction. The semantic segmentation task is trained on the MOTS dataset and results in a positive effect on both detection and feature extraction.
- Third, on the basis of DETR, we propose to use feature-dependent dynamic object queries (DOQ) over fixed queries, which combines the image features with static queries. As a result, the feature-dependent DOQs can provide flexible and dynamic candidate regions for pedestrian detection.
- Forth, we experiment with our method on the commonly used multi-object tracking dataset, and compare it with current end-to-end MOT methods. We also conduct a series of validity analysis to verify the effectiveness of proposed method. The experimental results show that our method improves IDF1 by 2.7%.

The remainder of this paper is organized as following: In Section 2 we provide a literature review on relevant topics in the MOT field and the applications of Transformer in this field. Section 3 illustrates our proposed method, including the semantic segmentation aided training and feature-dependent dynamic object queries. In Section 4, we present our experimental results on MOT17 dataset and conduct ablation analysis on proposed method. Section 5 draws the conclusion of our proposed method and puts forward the future work in the MOT field.

2. Related Works

In this section, we first introduce MOT methods that fall in classic tracking-by-detection paradigm. Afterwards, multi-task aided MOT methods and Transformer-based MOT methods are detailed therewith.

2.1. Tracking by detection methods

In tracking-by-detection paradigm, detections are always independently obtained before the association process comes in. The pipeline can be regarded as a cascade of feature extraction, motion prediction and object association processes.

Kalal et al. [18,19] first integrate tracking, detection and learning to solve long-term tracking tasks. After that, [20,21] make full use of detection information, and focus on improving the speed of tracking. This framework is used in tracking wildlife by Wang et al. [22]. Bergmann et al. [23] obtain the tracklets of the current frame by adjusting the previous frame's detections as the initial bounding box positions. Besides, the detector is used to calibrate new trajectories in the current frame.

In order to improve feature robustness, Zhang et al. [24] and Bae et al. [25,26] conduct the correlation algorithm based on confidence to deal with the occlusion problem. Wojke et al. [27] extract and save the appearance features of the objects extracted from the deep network, so as to reduce the ID switch of targets caused by occlusion and re-entering the field of view.

To speed up, Bewley et al. [28] simply adopt Kalman filter and Hungarian algorithm for multi-object correlation, which realize high speed and state-of-art performance. Wang et al. [29] get both object detection and appearance information from a shared model. The positions and embedding characteristics of bounding boxes are obtained by superimposed Feature Pyramid Network (FPN) results of different scales and different prediction heads of the image. This method avoids repetitive calculations in feature extractions thereby increasing frame rate from 22 to 40.

Despite the improvements, those methods cannot transcend the limitations of the detection performances. Existing tracks may serve as priors to contribute to object detection process to improve the accuracy and reduce false negatives. In addition, the feature extraction process inevitably overlaps with the object detection process to some extent, which affects computational efficiency.

2.2. MOT in multi-task learning

MOT is closely related to other computer vision methods. By combining tracking with other tasks, performance on both tasks can be improved due to shared foundations. Among them, object detection and object segmentation received much attention in literature.

Object detection and ReID tasks [30] are combined to improve the speed of MOT ID inferences. Two branches of object detection and ReID feature are established on the same encoder-decoder structure, and the algorithm maintains high accuracy when running 30 frames faster than on public datasets.

Hurtado et al. [31] combine semantic segmentation and instance segmentation together with MOT task to create a new perceptual task called Multi-Object Panoramic Tracking (MOPT). The authors use pixel-level fine-grained information and overall information to solve sub-problems, which are mutually promoted through their complementarity between each other. The proposed model is called the Panoptic TrackNet, which constructs multi-task heads and learns multiple sub-tasks simultaneously. To evaluate, a comprehensive SPTO measurement method is also proposed to quantitatively analysis the MOPT results.

In addition, there are methods that learn detection and instance segmentation simultaneously. Voigtlaender et al. [17] annotate the MOT and KITTI dataset, and label the foreground object masks by semi-automatic annotation method, including 65,213 pixel masks and 977 different objects. Specifically, they use the same network to jointly obtain detection, track and segmentation results. Three consecutive frames of images are stacked together as model input, and 3D convolution with sharing weights is used for feature extraction. Then the detections and the features are obtained through the Proposal Network (RPN). Tracking results are also obtained in the meantime.

Porzi et al. [32] use GPS and image information to automatically obtain the mask of the video annotation, and adopt optical flow to obtain the current track fragments. Furthermore, they also propose Multi-Object Tracking and Segmentation network (MOTSNet) to obtain MOT masks. In addition to obtaining the regions of interest of a single image through the backbone network, the region proposal heads and tracking heads are also achieved. Associated features of detection mask and tracking are obtained as well.

Cai et al. [33] utilize the given detections of the dataset as RPN and put it into the Mask R-CNN module to extract detection mask and image information. In addition, the spatial attention module is

integrated with the Mask R-CNN tracking feature head to obtain the appearance representations of tracks. Current detection and tracking results are correlated by Intersection over Union (IoU) and cosine similarity. ReID model is used to correlate the remaining uncorrelated detection and short time tracking fragments.

The above methods still adopt CNN features to obtain relationships and foreground pixels, and the RPN structure is always built to compute the candidate positions. However, CNN with local receptive fields is one of the restrictions for modeling global relationships, which is vital in MOT tasks. Another constraint in above multi-task methods is that they treat MOT equally or less to other tasks, thus neglecting some important factors in MOT, such as occlusion, shading, etc.

2.3. Transformer-based MOT methods

Transformer has received strong attention since its release in 2017. After DETR was proposed in 2020, it has become popular in the MOT field in the past two years. Zeng et al. [34] propose an end-to-end Transformer-based tracking framework. Track query is introduced to model the track information of an object, and automatic update is continuously realized during the time period when the target exists, so as to automatically complete the tracking association of different frames. The paper also proposes the Temporal Aggregation Network (TAN) to realize end-to-end tracking.

Meinhardt et al. [35] introduce Transformer structure to the field of MOT. After obtaining features through CNN backbone, trajectory features in the previous frame are obtained by constructing track query, and later associated with detection results obtained by object query in the current frame. Herein, the track query plays an information transmission role in this loop. Additionally, the authors use multi-head self-attention mechanism to transform the track query in adaptation to the object query space. Thus, bounding boxes are regressed and classifications are determined for new objects in the current frame. Zhu et al. [36] propose an end-to-end method to realize the track initialization and removal without object association process. As the authors suggest, the method comes from the idea that the "query-key" mechanism can be regarded as a naturally joint-detection-and-tracking paradigm.

Sun et al. [16] carry out in-depth feature extraction of two consecutive frames through backbone CNN network and combine the two features as the input of Transformer structure. All the detections of the current frame and the matching results of the previous frame are obtained through two parallel branches. Through a feed-forward network, the output of Transformer's decoder is used to predict detection bounding boxes. The detection results are correlated to the predicted tracking locations by Kuhn-Munkres (KM) algorithm [37], so as to obtain the final tracking results. The tracking process is performed iteratively between consecutive frames. After extracting features through backbone network, they are fed into Transformer encoders, to be queried by a global learned object query and a featured previous object query (POQ) from previous frame respectively, to result in a current and a previous set of detection bounding boxes. Then an IoU matching method is used to associate detections and tracks in adjacent frames. As can be seen, the global LOQ serves to find all potential detections in the current frame, while the featured POQ is responsible to propagate previous tracks in the current frame.

In above methods, the queries that used for generating new objects are trained on entire dataset and kept fixed during inference. This is not sufficient for time-varying scenarios, e.g. camera-moving scenarios, resulting in much missed detections and computational redundancies.

3. Proposed Methods

3.1. Overview

We adopt a similar architecture of the TransTrack method, including a CNN backbone, a Transformer with encoder and decoder, and a Feed Forward Neural Network (FFN) for bounding box regression. According to Fig. 1, we also stretch out two branches from the encoded features. One of them is used for auxiliary task, and another one used for object query.

To be specific, we use ResNet as our CNN backbone to obtain the feature information of the image and combine the appearance features of the previous frame with the features of the current frame. The features are then fed in three separate passages, which are the Transformer module, the semantic segmentation module, and the learned object query module.

The mainstream is the Transformer module, similar to DETR [15]. Extracted features are fed into self-attention modules to obtain multi-scale feature vectors in encoder layers. The candidate foreground positions are selected by the attention weights calculated by the dot product of keys and queries. In terms of the decoder module, the input queries are the output of the previous decoder or the learned object queries, while the keys and values are the outputs of the encoder. Because of multiple stacking of encoding and decoding, the module acquires multi-level and abundant features. After decoding, the detection bounding boxes are calculated through the fully connected layer and associated to tracks by performing a Hungarian matching method.

For the segmentation module, our method learns complementary features to obtain the pixel-level information. The semantic mask is acquired by a series of operations shown in Fig. 2. For the object query module, the image features are utilized to generate the bias of the learned queries to effectively obtain all the object candidates appearing in the current frame. Those modules will be elaborated in Section 3.3 and 3.4 respectively.

3.2. Revisiting Transformer

In this part, we briefly introduce the architecture of Transformer and DETR. For the Transformer, the key step is the attention mechanism, which mainly includes three parts: self-attention mechanism, interactive attention mechanism, and feed-forward neural network. It mainly realizes information acquisition by constructing encoder-decoder modules, and is widely used in detection and classification areas. The basic Transformer encoder-decoder structure totally includes 12 convolutional layers. Among them, the encoder includes six layers, which is consistent with the number of decoders. Compared with the Recurrent Neural Network (RNN) time series prediction model, it can have better parallelism, and the inputs do not need to be strictly time dependent. Besides, it uses trigonometric function to encode the positions and time information of inputs, and builds the attention mechanism to encode any two positions as a constant.

In the encoder of the Transformer, the data first pass through the self-attention module. Transformer's self-attention structure captures the relevant information of the target, and establishes three vectors of Q, K and V which represent Query, Key and Value to obtain different information of each object. The attention formula of Q, K and V is shown in Eq. 1.

$$Attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V, \tag{1}$$

where Q represents the query vector, K represents the key vector, and V represents the value vector. They are obtained by multiplying

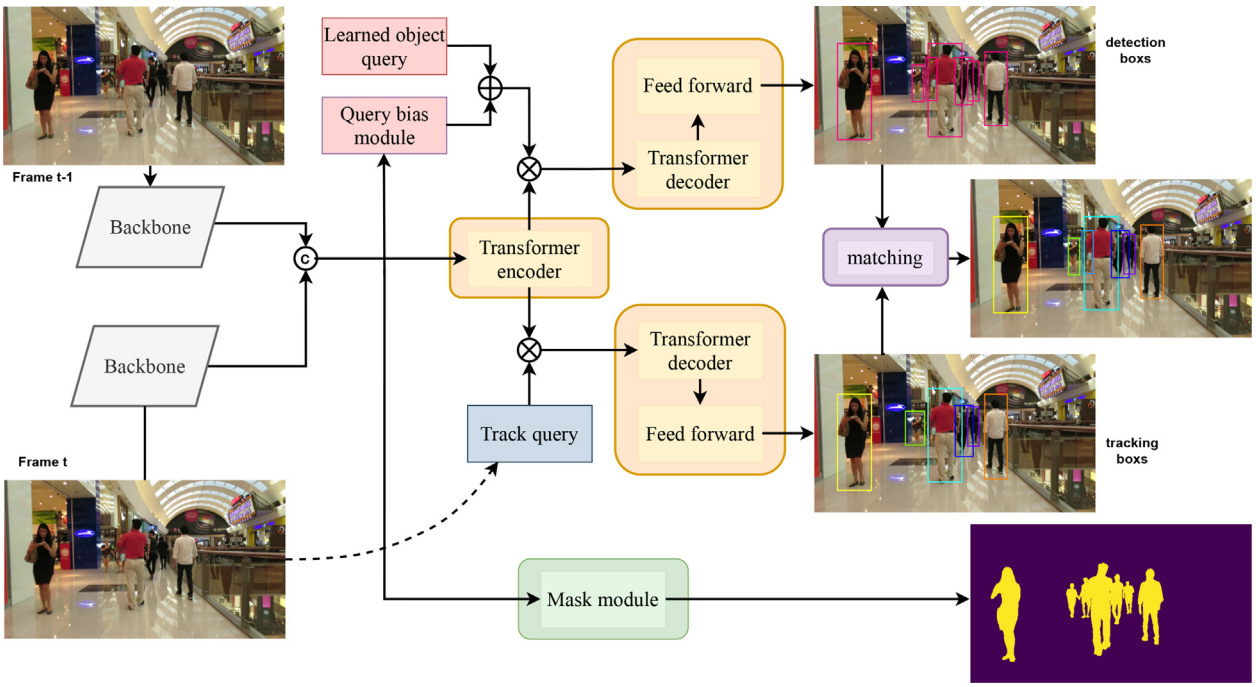

Fig. 1. Overall framework of our segmentation assisted MOT with dynamic query-based Transformer (SegDQ) method.

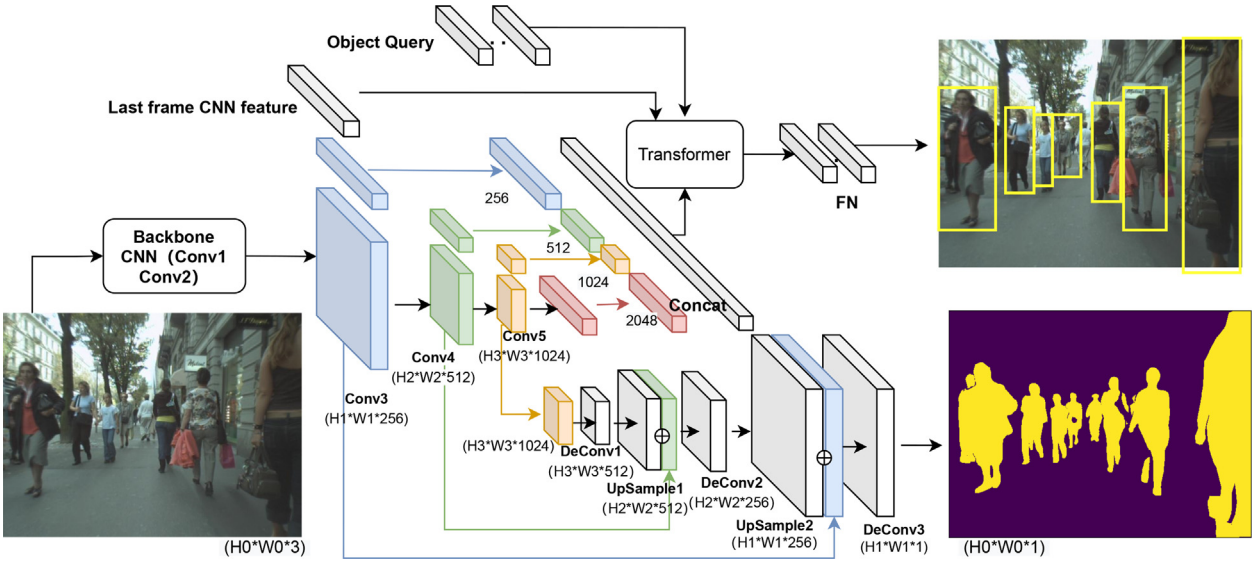

Fig. 2. The structure of auxiliary semantic segmentation branch. The last three feature maps of ResNet50 are selected to construct the segmentation results.

the current target vectors with different weight matrices, and the three vectors have the same sizes.

In the above formula, query and key perform dot product to obtain the similarity of the current two targets. Simultaneously, in order to solve the degradation problem of the model, self-attention adopts the residual network structure of ResNet [38], whose structure is shown in Fig. 3.

After obtaining the self-attention results, the Transformer adds a multi-head attention structure to focus on different parts of images, thereby increasing the model capacity; Besides, the parallel operation of the model is realized to increase the running speed and save the running time.

For the feed forward neural network part, it mainly consists of a fully connected layer and an activation function. Its main purpose is to map the model to a large space, thereby increasing the capacity and improving the representation ability of the model.

3.3. Semantic Segmentation as an Auxiliary Task

In specific MOT area, accurate semantic segmentation results can facilitate the acquisition of accurate detection regress on target objects. Meanwhile, association by features is also commonly used in tracking-by-detection paradigm, whose qualities are also heavily affected by foreground extraction. Therefore, as Transformer is built upon attention mechanism, we propose to address more attention on foreground extraction. This is done by posing an auxiliary semantic segmentation alongside the main end-to-end learning task.

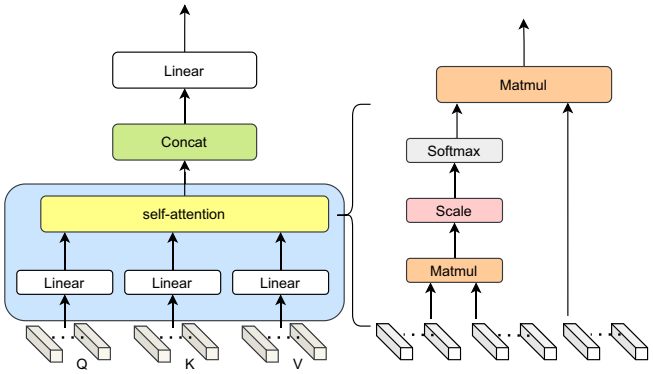

Fig. 3. The structure of self-attention module.

As shown in Fig. 2, we branch out from the Transformer and connect it with a U-Net to predict the mask of all objects in the current frame. By introducing this task, the model can be trained to better extract foreground features.

The essence of Transformer is to use the attention mechanism to discover the relationships between pixels.

According to DETR, the distribution of the heat map of the attention mechanism calculated from the Transformer's encoder's output contains only a rough outline, which can be improved to more accurate attention results. Notably, the main task and the auxiliary task are mutually aided. On one hand, the semantic segmentation task can be used as a supplement to tracking, which improves the accuracy of tracking foreground acquisition by adding sufficient mask information. On the other hand, since the MOT field contains foreground targets of different scales, it also has a guiding role for semantic segmentation task. The semantic mask extraction task also introduces noise interference to the MOT task, which can improve the generalization effect of our proposed method. The multi-task approach can also make it easy for the model to break away from the local minimum point, and learn the difficult and potential features of the current task by auxiliary one.

As mentioned above, the task is trained in a supervised manner. The supervised data is acquired from the MOTS [17] dataset, which complements the original MOT dataset with additional semantic mask data. In MOT datasets, target shapes vary largely from each other. Several small-size objects are eliminated because they are hardly to be identified, and we fuse object masks of a certain range of sizes together to formulate semantic masks for supervised data.

We use the last 3 layers of backbone ResNet, and use a series of convolution and upsampling operations to reconstruct the foreground mask of given images. ReLU activation is used for every convolutional layers. As shown in Fig. 2, the last layer of backbone network is firstly convoluted and upsampling to the size of the second last layer, and the add operation is used to jointly get the mixed feature of the two layers. Besides, the Group normalization and ReLU operations are followed behind. The operations are repeated to the third last layer. The kernel sizes of the three convolutional layers are set to 3. Then, the output layer reduces the channel of features and semantic segmentation results are obtained and the kernel size is 1.

We use focal loss as the loss function for supervised learning in this auxiliary task as defined in Eq. 2, where p is 0.25 and γ is 2. Generally speaking, the focal loss is computed according to different foreground information, highlighting the degree of misdivision of the foreground information. As previously, we also neglect those out-group and small objects.

$$FL = \begin{cases} -(1-p)^{\gamma} log(p), & if \ \ y = 1, \\ -p^{\gamma} log(1-p), & if \ \ y = 0, \end{cases} \tag{2}$$

where p represents the probability that the corresponding pixel belongs to the foreground information. γ is a constant, and it is used to reduce the loss of easy parts and pay attention to the more difficult misclassification parts. When it is 0, FL is consistent with the cross entropy loss. Here we set p to 0.25 and γ to 2.

3.4. Feature-dependent dynamic object query

In original TransTrack, when taking new images as input, the learned object queries are used to indicate candidate locations for new detections.

Normally, in classic Transformer, after random initialization, the locations of learned object queries are adjusted according to the training videos. However, the query has nothing to do with current input image in the process of inference, which is in fact trained according to the general distribution of the foreground under the entire training dataset and results several positions for all circumstances. However, as shown in Fig. 5, the distributions of these images vary considerably in object sizes and locations, especially in camera-moving scenarios which is similar with Tian et al. [39] proposed. Thus, fixed candidate positions of the input Transformer are not suitable for searching object detections in changing scenes.

To solve the problem, we introduce image feature information into the query generation process, so that the queries depend on incoming image features, thereby adapt the candidate position predictions and the Transformer's attention calculations as well.

In addition to vanilla DETR model, we draw image features from backbone CNN and put them into a bias module. The module outputs a biased object query, which is combined with the original learned object query and then fed into Transformer. In practice, we splice the last three layers of backbone ResNet model, and use max-pooling layer and a feed-forward network to construct the bias module. We use Tanh as activation for the module output layer, and ReLU activation for other intermediate layers. The output shape of the bias module is 300, equivalent to the shape of original learned object query. The total structure of feature-dependent dynamic object query is shown in Fig. 4.

Worth mentioning, as trained by the semantic segmentation auxiliary task, features drawn from backbone CNN is considered to hold both mask and appearance information. These features are not only used to generate DOQ for current frame, but also used in Transformer's decoder to generate object query in up-coming frame for association step. While TransTrack uses image features only for detection propagate, we also use image features for current querying.

3.5. Association process

The association step between new detections and existing tracks close the problem of tracking. Here, let $D^k = \left\{ d_i^k \right\}$ denotes the bounding box collection of i-th detection in the current frame k, and t_j^{k-1} represents the existing j-th track inthe previous frame, and $\overline{T}^k = \left\{ \bar{t}_j^k \right\}$ denotes the collection of bounding box of the prediction of j-th track in the current frame. The problem is to find the best match between D^k and $\overline{T}^k$.

Unlike the tracking-by-detection methods, we neglect the object feature matching score which is always calculated by cosine similarity. In fact, following the tracking-by-query paradigm, we only use IoU to measure the overlap between detection d_i^k and $\bar{t}_j^k$.

Fig. 5. various distributions of pedestrians in different datasets. 5(a) is from MOT17-04 and 5(b) is selected from MOT17-05.

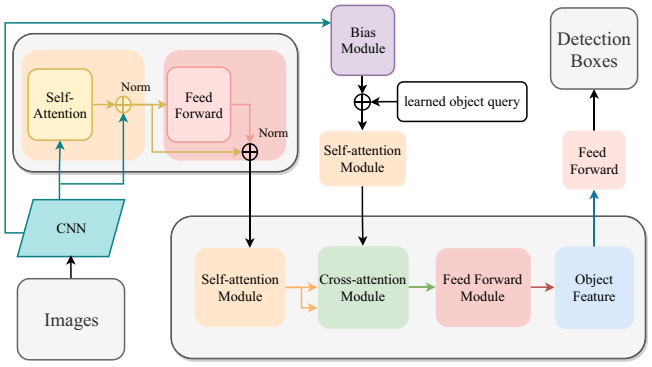

Fig. 4. The structure of dynamic query generation.

Let $g\left(d_i^k, \overline{t}_j^k\right)$ denotes the IoU similarity between detection and track, the total matching cost between D^k and $\overline{T}^k$ can be calculated and the Hungarian algorithm [37] is used to get correlation results with minimum cost.

Therefore, after obtaining new detections D^k and propagating existing tracks $\overline{T}^k$, we first find best match between D^k and $\overline{T}^k$. Afterwards, we calculate the best matching IoU loss L_{giou} in training the tracking model. The total detection and tracking training loss is calculated through Eq. 3.

$$L = \lambda_{cls} * L_{cls} + \lambda_{L1} * L_{L1} + \lambda_{giou} * L_{giou} + \lambda_{msk} * L_{msk}, \tag{3}$$

where L_{cls} represents the classification and category loss, L_{L1} and L_{giou} loss are used to calculate the $L1$ and generalized IoU [40] losses of predicted bounding boxes compared with ground truth bounding boxes. L_{msk} is the semantic segmentation loss from Eq. 2. $\lambda_{cls}, \lambda_{L1}, \lambda_{giou}$ and λ_{msk} are the hyper-parameters to balance these losses. Each loss is scaled between 0 and 1 to guarantee that every loss is with the same scale. Details of our configuration is discussed in Section 4.

4. Experimental Results

Parameter settings and results on MOT datasets are shown and discussed in this section. In addition, the ablation study results and comparisons with current competitive methods are demonstrated to verify the effectiveness of our modifications.

4.1. Experimental setup

The ResNet50 [38] is used as the backbone network of our proposed method. Besides, we adopt the Adam optimizer to update the parameters of the proposed model. The initial learning rate of the Transformer is set to 2×10^{-4}, and the weight decay is set to 1×10^{-4} to dealing with overfitting. The batch size of training is set to 2 on the device of a 16G Titan Xp, which is relatively small due to the limitation of GPU memory. Besides, data augmentation steps, which are commonly known as random resizing and crop, are utilized to increase the diversity of training images. The model is trained for 100 epochs with learning rate scheduled from 2×10^{-4} to 2×10^{-5}. Part of our model weights are initialized with the pre-trained weights obtained in [16] for reducing the training time. For the training loss, $\lambda_{cls}, \lambda_{L1}, \lambda_{giou}, \lambda_{msk}$ are set to 2, 5, 2 and 3, respectively. $\lambda_{cls}, \lambda_{L1}$ and λ_{giou} are equal to the TransTrack settings. Since Transtrack sets parameters' range from 0 to 5, we select 0, 1, 3, and 5 as the values of λ_{msk} for analysis. The average precisions with $IoU \geqslant 0.95$ of these hyperparameters are 0.5461, 0.5507, 0.5512 and 0.5480, respectively. Therefore, we set λ_{msk} to be 3 to ensure better performance of our model.

4.2. Detection results

The detection performance indexes and their meanings are listed in Table 1. We focus on the average precision and recall rate of detections over ground truth under certain IoU indicators. The dataset used for comparison with TransTrack is half of the MOT17 [41] training set, and the results are listed in Table 2. Here, we evaluate with $IoU \geqslant 0.95, 0.75$ and 0.50 circumstances, and the average precision of our method when $IoU \geqslant 0.95$ is 0.554, which

Table 1
The detection evaluation indexes. (↑ denotes the higher the better, and ↓ denotes the lower the better.)

Index	Description
$AP_x \uparrow$	Average precision of $IoU \geqslant x$
$AP_S \uparrow$	Average precision of small objects
$AP_M \uparrow$	Average precision of medium objects
$AP_L \uparrow$	Average precision of large objects
$Recall_x \uparrow$	Recall rate of $IoU \geqslant x$
$Recall_S \uparrow$	Recall rate of small objects
$Recall_M \uparrow$	Recall rate of medium objects
$Recall_L \uparrow$	Recall rate of large objects

Table 2
Detection performance comparisons of our method with TransTrack on MOT17 half dataset.

methods	AP_{95}	AP_{50}	AP_{75}	AP_S	AP_M	AP_L	$Recall_{95}$	$Recall_{50}$	$Recall_{75}$	$Recall_S$	$Recall_M$	$Recall_L$
TransTrack	0.543	0.880	0.602	0.078	0.444	0.640	0.050	0.349	0.636	0.197	0.555	0.727
Trans + query	0.551	0.880	0.616	0.072	0.456	0.644	0.050	0.350	0.633	0.167	0.555	0.722
Trans + mask + query	**0.554**	**0.881**	**0.620**	**0.072**	**0.457**	**0.646**	**0.051**	**0.351**	**0.635**	**0.164**	**0.558**	**0.725**

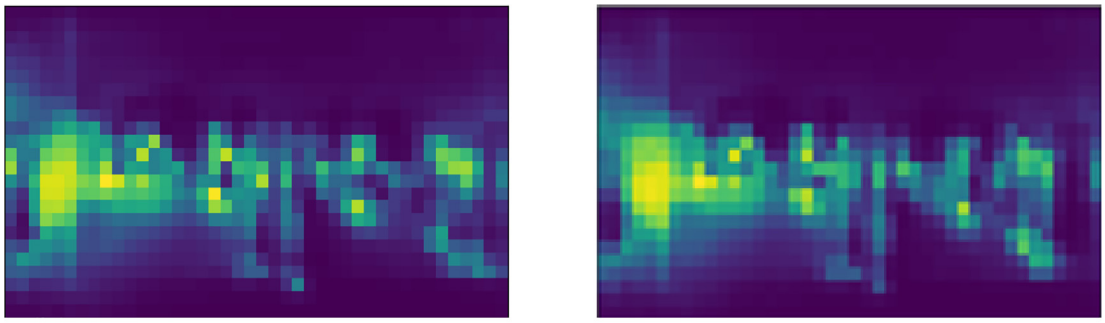

Fig. 7. The different feature maps of (a) SegDQ and (b) Transtrack..

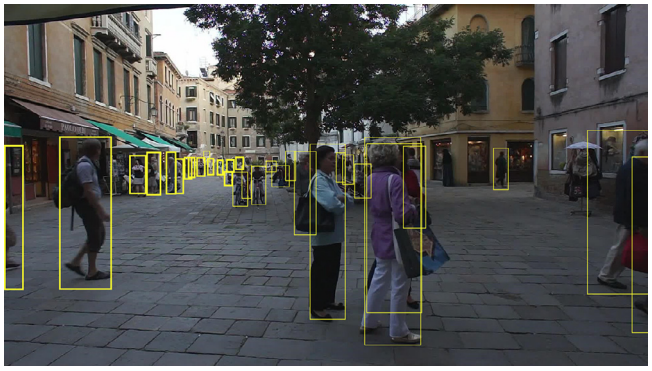

Fig. 6. The original image from MOT17-02 dataset..

Table 3
The tracking evaluation indexes.

Index	Description
MOTA ↑	Multiple Object Tracking Accuracy
MOTP ↑	Multiple Object Tracking Precision
ML ↓	Mostly lost
MT ↑	Mostly tracked
Frag ↓	Number of track fragmentations
FP ↓	False positive output
FN ↓	False negative output
IDs ↓	ID switch

is 1.1% higher than TransTrack. Besides, in $IoU \geqslant 0.75, IoU \geqslant 0.50$ circumstances, our average precisions have achieved an increase of 0.1% and 1.8%, respectively. Our method is valid in the recall indexes of all three IoU proportions as well. Besides, the mAP and recall rate of medium object detections improve 1.3% and 0.4%.

Fig. 7 shows the Transformer encoding attention results of the Fig. 6 based on SegDQ and Transtrack, respectively. Compared with Transtrack, more than only effectively detecting the pedestrians in the center of view, our method is also capable of detecting the pedestrians on the edge of view.

4.3. MOT datasets and Evaluation indicators

The tracking performance indexes and their meanings are listed in Table 3. In our experiments, we use MOT15, 16 and 17 training and test datasets to evaluate the tracking results. The MOT17 dataset is a widely used MOT benchmark focusing on pedestrian tracking in daily traffic and street scenes. There are totally 7 video sequences in MOT17 dataset, and each is evenly split into two halves for training and testing respectively. The training sets provide both ground truth detection and tracking results which can be used for custom evaluation, while the testing sets provide with original videos only. Worth mentioning, the dataset contains a diversity of videos taken from both high and low angles, day and night.

Among the indicators, the MOTA is calculated by the following Eq. 4:

$$MOTA = 1 - \frac{\sum_t (FP_t + FN_t + ID_{s_t})}{\sum_t GT_t} \tag{4}$$

where FP and FN represent the counts of false positive and false negative respectively. IDs stands for the counts of ID switches, and GT stands for the total counts of actual bounding boxes. The IDF1 is proposed to evaluate the MOT accuracy and recall rate based on the trajectory IDs, which can be calculated through Eq. 5,

$$IDF1 = \frac{2IDTP}{2IDTP + IDFP + IDFN} \tag{5}$$

where IDTP, IDFP and IDFN represent ID based true positive, false positive and false negative counts respectively.

4.4. Segmentation assisted MOT tracking results

In terms of training the segmentation assisted model, the hyper-parameter λ_{msk} is used to balance the importances of MOT and semantic segmentation tasks.

Examples of the semantic segmentation results are shown in Fig. 8. As can be seen, the predicted segmentation is close to the ground truth mask. The images are selected from different tracking datasets, which contain pedestrians in multiple sizes. Besides,

Fig. 8. The semantic mask generation results on MOT dataset. The first row shows the original challenging images on MOT17-02, MOT17-05, MOT17-09 and MOT17-11 respectively. The second row shows our auxiliary mask branch results of the corresponding results, and the last row displays the semantic segmentation groundtruth. Our segmentation mask can predict more accurate and fine-grained information, so the coarse-grained information such as bounding boxes can also predict more accurate.

Table 4
Tracking performance on the test set of the MOT17 Benchmark.

Tracker	Mode	MOTA	IDF1	MT	ML	FP	FN	IDs
				public detection				
eHAF [42]	Offline	51.8%	54.7%	551	893	33212	336772	1834
FWT [43]	Offline	51.3%	47.6%	505	830	24101	247921	2648
eTC17 [44]	Offline	51.9%	58.1%	544	836	36164	232783	2288
MOTDT17 [45]	Online	50.9%	52.7%	413	841	24069	250768	2474
DMAN [46]	Online	51.9%	55.7%	454	902	26218	263608	2194
JBNOT [47]	Online	52.6%	50.8%	465	844	31572	232659	3050
LSST17O [48]	Online	52.7%	57.9%	421	863	22512	241936	2167
				private detection				
Tracktor + CTdet	Online	54.4 %	78.1%	605	702	44109	210774	2574
TubeTK	Online	63.0%	58.6%	735	469	27060	177483	4137
DeepSORT	Online	60.3%	61.2 %	742	478	36111	185301	2442
ChainedTracker	Online	66.6%	78.2%	759	570	22284	160491	5529
TransTrack	Online	65.8%	56.9%	759	514	24000	163683	5355
Ours	**Online**	**65.0%**	**59.4%**	**690**	**609**	**20574**	**172902**	**4023**

Table 5
Tracking performance on the test set of the MOT16 Benchmark.

Tracker	Mode	MOTA	IDF1	MT	ML	FP	FN	IDs
eHAF [42]	Offline	47.2%	52.4%	141	325	12586	83107	542
TPM [49]	Offline	51.3%	47.9%	142	310	2701	85504	569
MLT [50]	Offline	52.8%	62.6%	160	322	5362	80444	299
LMCNN [51]	Online	67.4%	61.2%	290	146	10109	48435	931
TubeTK [52]	Online	64.0%	59.4%	254	147	10962	53626	1117
DeepSORT	Online	61.4%	62.2%	249	138	12852	56668	781
HTA [53]	Online	62.4%	64.2%	285	92	19071	47839	1619
ChainedTracker [54]	Online	67.6%	57.2%	250	175	8934	48305	1897
Ours	**Online**	**65.7%**	**59.9%**	**230**	**187**	**8125**	**53100**	**1317**

there also exist several failure segmentation cases. Because our MOT task mainly deals with bounding boxes, tracking results are robust to several inaccurate pixels within pedestrian contours.

4.5. MOT Results

As Section 4.3 shows, MOT method performances can be evaluated by the corresponding indexes. We evaluate SegDQ method on MOT17, MOT16 and MOT15 datasets and the results are shown in Tables 4–6.

From Table 4, our end to end tracking method achieves 59.0% MOTA. Compared with the methods using public detections listed in the table, we reduce the FP and FN numbers simultaneously. Besides, the mostly tracked (MT) count is 492 and the mostly lost (ML) count is reduced to 783. The IDs count is slightly higher than others because we only use the IoU matching to get the final MOT tracking results.

Table 6
Tracking performance on the test set of the MOT15 Benchmark.

Tracker	Mode	MOTA	IDF1	MT	ML	FP	FN	IDs
CRFTrack [55]	Online	40.0%	49.6%	166	206	10295	25917	658
GNNMatch [56]	Online	46.7%	43.2%	157	203	6643	25311	820
Tracktor++ [23]	Online	44.1%	46.7%	130	189	6477	26577	1318
STRN [45]	Online	38.1%	46.6%	83	241	5451	31571	1033
AMIR15 [57]	Online	37.6%	46.0%	114	193	7933	29397	1026
Ours	**Online**	**40.3%**	**47.1%**	**319**	**88**	**18116**	**17550**	**1031**

Table 7
The ablation study of tracking performance on MOT17 training dataset. Our method using only dynamic query based transformer and with both dynamic query and asssisted segmentation task are shown in last two rows respectively.

Video number	Tracker	IDF1	MOTA	MT	ML	FP	FN	IDs	Frag
MOT17-02	TransTrack	40.2%	37.1%	11	23	310	5833	67	175
	ours	41.7%	37.3%	10	21	286	5842	62	160
MOT17-04	TransTrack	75.7%	80.1%	42	3	369	4294	141	420
	ours	80.5%	80.9%	42	1	374	4153	95	356
MOT17-05	TransTrack	56.9%	64.0%	23	15	61	1116	30	65
	ours	56.5%	63.3%	23	16	75	1544	17	44
MOT17-09	TransTrack	61.1%	65.9%	12	3	13	955	14	47
	ours	60.8%	65.5%	13	3	9	970	13	38
MOT17-10	TransTrack	63.4%	59.3%	14	2	686	1659	67	228
	ours	65.6%	58.2%	14	3	550	1874	52	202
MOT17-11	TransTrack	58.6%	62.1%	13	12	331	1362	20	64
	ours	60.9%	63.8%	10	17	75	1544	17	44
MOT17-13	TransTrack	65.4%	58.9%	24	3	367	781	150	122
	ours	64.5%	52.9%	18	7	435	945	108	113
Overall	TransTrack	64.9%	65.4%	139	61	2137	16000	489	1121
	Ours + query	**65.7%**	**64.3%**	**125**	**80**	**1271**	**17614**	**372**	**1031**
	Ours + query + mask	**67.8%**	**65.4%**	**130**	**68**	**1795**	**16463**	**378**	**979**

The results on MOT16 test dataset are shown in Table 5. MOT16 dataset is composed of 14 videos that are shot in various places with different scenes. Consequently, our method achieves 65.7% MOTA and 59.9% IDF1 as well. The MT count has been improved to 492 and ML count have been reduced to 783. The counts of FP and FN has also been decreased benefiting from effective Transformer based detection acquisition method.

Our method performs better MOTA on MOT15 test dataset as Table 6 shows. MOT15 consists of 22 video sequences that adds up to 996 s that include various challenging video sequences. Besides, the total IDF1 is 47.1% and the ML number has reduced to 88. The false negative number achieves 17550 and false positive number is 18116.

Table 7 shows the ablation results of our method. Half of the MOT17 training dataset is utilized to validate our method. As can be seen, the IDF1 of our method with dynamic query has increased by 0.8%, and the ID switch count is reduced by 117 as well. The number of fragments also reduces to 1031. Our method with dynamic query and auxiliary segmentation task on MOT17 training dataset is shown in the last row. The IDF1 has increased 2.9% and the IDs and fragment numbers both reduced. Besides, the FP numbers are reduced simultaneously. Worth mentioning, MOT17-04 is shot from a high angle and contains crowded pedestrians and the IDF1 of our method on MOT17-04 has improved 4.8%, which means our method is efficient in the crowd scenes.

Compared with Transtrack method, our method performs better on IDF1 and so on. IDF1 is more sensitive to the accuracy of ID information in MOT. The dynamic learned object queries can be acquired to obtain more accurate candidate foreground distributions, and our model can acquire more accurate detections as well.

Fig. 9. The performance of our algorithm on MOT17 test dataset. The corresponding observation objects are highlighted by yellow arrows and the corresponding appeared frame numbers are displayed in the right corner. Although the objects are occludied by their neighbors frequently in videos, the highlighted pedestrians are correctly tracked during the time gaps.

Because the queries obtained from the last frame are induced to the current frame to get the locations of the existing tracks, more accurate detection locations can lead to more stable correlation. In addition, after semantic segmentation task is added, fine-grained semantic segmentation information is added to the target on the basis of detection bounding box, which helps the model learn fine-grained knowledge, so as to better distinguish foreground targets.

Fig. 9 visualizes tracking performances of our method. As can be seen, the positions of pedestrians are correctly selected and the persons that highlighted by yellow arrows are tracked constantly in large frame gaps.

5. Conclusion

In this paper, based on Transformer, we propose a multi-task and dynamic feature based query acquisition method to improve multi-object tracking performance. A semantic segmentation branch is used to aid the learning of Transformer in predicting foreground masks. This auxiliary task can help the main MOT task with better foreground feature extracting and better bounding box regression. We also propose a dynamic query which use the deep features extracted from backbone network to generate biased object queries, thus to realize more robust and accurate prediction of new detections. Through our experiments on MOT15, MOT16 and MOT17, improvements on MOTA and IDF1 indexes can both be achieved, thus we can draw the conclusion that our proposed method is effective on multi-object tracking tasks.

CRediT authorship contribution statement

Yating Liu: Investigation, Conceptualization, Methodology, Software, Formal analysis, Visualization, Writing - original draft, Writing - review & editing. **Tianxiang Bai:** Conceptualization, Methodology, Validation, Formal analysis. **Yonglin Tian:** Conceptualization, Methodology, Validation, Formal analysis. **Yutong Wang:** Methodology, Validation, Formal analysis. **Jiangong Wang:** Conceptualization, Methodology, Validation, Formal analysis. **Xiao Wang:** Resources, Conceptualization. **Fei-Yue Wang:** Resources, Supervision.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgement

This work was supported by the National Natural Science Foundation of China (62173329) and Intel Collaborative Research Institute for Intelligent and Automated Connected Vehicles (“ICRI-IACV”).

References

[1] A. Guzmán, Decomposition of a visual scene into three-dimensional bodies, in: Proceedings of the Joint Computer Conference, 1968, pp. 291–304.
[2] N. Wax, Signal-to-noise improvement and the statistics of track populations, Journal of Applied physics 26 (5) (1955) 586–595.
[3] R.W. Sittler, An optimal data association problem in surveillance theory, IEEE Transactions on Military Electronics 8 (2) (1964) 125–139.
[4] Y. Bar-Shalom, Tracking methods in a multi-target environment, IEEE Transactions on Automatic Control 23 (4) (1978) 618–626.
[5] T. Benedict, G. Bordner, Synthesis of an optimal set of radar track-while-scan smoothing equations, IRE Transactions on Automatic Control 7 (4) (1962) 27–32.
[6] A. Rangesh, M.M. Trivedi, No blind spots: Full-surround multi-object tracking for autonomous vehicles using cameras and lidars, IEEE Transactions on Intelligent Vehicles 4 (4) (2019) 588–599.
[7] F. Camara, N. Bellotto, S. Cosar, D. Nathanael, M. Althoff, J. Wu, J. Ruenz, A. Dietrich, C. Fox, Pedestrian models for autonomous driving part I: Low-level models, from sensing to tracking, IEEE Transactions on Intelligent Transportation Systems (2020).
[8] D.B. Carr, P. Grover, The role of eye tracking technology in assessing older driver safety, Geriatrics 5 (2) (2020) 36.
[9] Y. Liu, J. Yin, D. Yu, S. Zhao, J. Shen, Multiple people tracking with articulation detection and stitching strategy, Neurocomputing 386 (2020) 18–29.
[10] H. Rabiee, H. Mousavi, M. Nabi, M. Ravanbakhsh, Detection and localization of crowd behavior using a novel tracklet-based model, International Journal of Machine Learning and Cybernetics 9 (12) (2018) 1999–2010.
[11] S. Wu, H. Yang, S. Zheng, H. Su, Y. Fan, M.-H. Yang, Crowd behavior analysis via curl and divergence of motion trajectories, International Journal of Computer Vision 123 (3) (2017) 499–519.
[12] A. Plantinga, Things and persons, The Review of Metaphysics (1961) 493–519.
[13] N. Gheissari, T.B. Sebastian, R. Hartley, Person reidentification using spatiotemporal appearance, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Vol. 2, 2006, pp. 1528–1535. .
[14] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, in: Advances in Neural Information Processing Systems, 2017, pp. 5998–6008.
[15] X. Zhu, W. Su, L. Lu, B. Li, X. Wang, J. Dai, Deformable DETR: Deformable Transformers for end-to-end object detection, arXiv preprint arXiv:2010.04159 (2020). .
[16] P. Sun, Y. Jiang, R. Zhang, E. Xie, J. Cao, X. Hu, T. Kong, Z. Yuan, C. Wang, P. Luo, TransTrack: Multiple-object tracking with Transformer, arXiv preprint arXiv:2012.15460 (2020). .
[17] P. Voigtlaender, M. Krause, A. Osep, J. Luiten, B.B.G. Sekar, A. Geiger, B. Leibe, MOTS, Multi-object tracking and segmentation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 7942–7951.
[18] Z. Kalal, K. Mikolajczyk, J. Matas, Face-TLD, Tracking-learning-detection applied to faces, in: Proceedings of the IEEE International Conference on Image Processing, 2010, pp. 3789–3792.
[19] Z. Kalal, K. Mikolajczyk, J. Matas, Tracking-learning-detection, IEEE Transactions on Pattern Analysis and Machine Intelligence 34 (7) (2011) 1409–1422.
[20] E. Bochinski, V. Eiselein, T. Sikora, High-speed tracking-by-detection without using image information, in: Proceedings of the IEEE International Conference on Advanced Video and Signal Based Surveillance, 2017, pp. 1–6.
[21] J. Shen, D. Yu, L. Deng, X. Dong, Fast online tracking with detection refinement, IEEE Transactions on Intelligent Transportation Systems 19 (1) (2017) 162–173.
[22] H. Wang, J. Zhu, W. Dai, J. Liu, A Re-ID and tracking-by-detection framework for multiple wildlife tracking with artiodactyla characteristics in ecological surveillance, in: Proceedings of the International Conference on Real-time Computing and Robotics, 2019, pp. 901–906, https://doi.org/10.1109/RCAR47638.2019.9043947.
[23] P. Bergmann, T. Meinhardt, L. Leal-Taixe, Tracking without bells and whistles, in: Proceedings of the IEEE International Conference on Computer Vision, 2019, pp. 941–951. .
[24] X. Zhang, X. Wang, C. Gu, Online multi-object tracking with pedestrian re-identification and occlusion processing, The Visual Computer 37 (2021) 1089–1099.
[25] S.-H. Bae, K.-J. Yoon, Robust online multi-object tracking based on tracklet confidence and online discriminative appearance learning, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 1218–1225.
[26] S.-H. Bae, K.-J. Yoon, Confidence-based data association and discriminative deep appearance learning for robust online multi-object tracking, IEEE Transactions on Pattern Analysis and Machine Intelligence 40 (3) (2017) 595–610.
[27] N. Wojke, A. Bewley, D. Paulus, Simple online and realtime tracking with a deep association metric, in: Proceedings of the IEEE International Conference on Image Processing, 2017, pp. 3645–3649.
[28] A. Bewley, Z. Ge, L. Ott, F. Ramos, B. Upcroft, Simple online and realtime tracking, in: Proceedings of the IEEE International Conference on Image Processing, 2016, pp. 3464–3468.
[29] Z. Wang, L. Zheng, Y. Liu, Y. Li, S. Wang, Towards real-time multi-object tracking, in: Proceedings of the European Conference on Computer Vision, 2020, pp. 107–122.
[30] Y. Zhang, C. Wang, X. Wang, W. Zeng, W. Liu, A simple baseline for multi-object tracking, arXiv preprint arXiv:2004.01888 (2020). .
[31] J.V. Hurtado, R. Mohan, W. Burgard, A. Valada, MOPT: Multi-object panoptic tracking, arXiv preprint arXiv:2004.08189 (2020). .
[32] L. Porzi, M. Hofinger, I. Ruiz, J. Serrat, S.R. Bulo, P. Kontschieder, Learning multi-object tracking and segmentation from automatic annotations, in: Proceedings

of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 6846–6855.
[33] J. Cai, Y. Wang, H. Zhang, H.-M. Hsu, C. Ma, J.-N. Hwang, IA-MOT: Instance-aware multi-object tracking with motion consistency, arXiv preprint arXiv:2006.13458 (2020). .
[34] F. Zeng, B. Dong, T. Wang, C. Chen, X. Zhang, Y. Wei, MOTR: End-to-end multiple-object tracking with Transformer, arXiv preprint arXiv:2105.03247 (2021). .
[35] T. Meinhardt, A. Kirillov, L. Leal-Taixe, C. Feichtenhofer, TrackFormer: Multi-object tracking with Transformers, arXiv preprint arXiv:2101.02702 (2021). .
[36] T. Zhu, M. Hiller, M. Ehsanpour, R. Ma, T. Drummond, H. Rezatofighi, Looking beyond two frames: End-to-end multi-object tracking using spatial and temporal Transformers, arXiv preprint arXiv:2103.14829 (2021). .
[37] H.W. Kuhn, The Hungarian method for the assignment problem, Naval Research Logistics Quarterly 2 (1–2) (1955) 83–97.
[38] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 770–778, https://doi.org/10.1109/CVPR.2016.90.
[39] Y. Tian, K. Wang, Y. Wang, Y. Tian, Z. Wang, F.-Y. Wang, Adaptive and azimuth-aware fusion network of multimodal local features for 3D object detection, Neurocomputing 411 (2020) 32–44.
[40] H. Rezatofighi, N. Tsoi, J. Gwak, A. Sadeghian, I. Reid, S. Savarese, Generalized intersection over union: A metric and a loss for bounding box regression, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 658–666.
[41] A. Milan, L. Leal-Taixé, I. Reid, S. Roth, K. Schindler, MOT16: A benchmark for multi-object tracking, arXiv preprint arXiv:1603.00831 (2016). .
[42] H. Sheng, Y. Zhang, J. Chen, Z. Xiong, J. Zhang, Heterogeneous association graph fusion for target association in multiple object tracking, IEEE Transactions on Circuits and Systems for Video Technology 29 (11) (2018) 3269–3280.
[43] R. Henschel, L. Leal-Taixé, D. Cremers, B. Rosenhahn, Fusion of head and full-body detectors for multi-object tracking, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2018, pp. 1509–150909, https://doi.org/10.1109/CVPRW.2018.00192.
[44] G. Wang, Y. Wang, H. Zhang, R. Gu, J.-N. Hwang, Exploit the connectivity: Multi-object tracking with TrackletNet, in: Proceedings of the ACM International Conference on Multimedia, 2019, pp. 482–490.
[45] L. Chen, H. Ai, Z. Zhuang, C. Shang, Real-time multiple people tracking with deeply learned candidate selection and person re-identification, in: Proceedings of the IEEE International Conference on Multimedia and Expo, 2018, pp. 1–6, https://doi.org/10.1109/ICME.2018.8486597.
[46] J. Zhu, H. Yang, N. Liu, M. Kim, W. Zhang, M.-H. Yang, Online multi-object tracking with dual matching attention networks, in: Proceedings of the European Conference on Computer Vision, 2018, pp. 366–382.
[47] R. Henschel, Y. Zou, B. Rosenhahn, in: Multiple people tracking using body and joint detections, in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2019.
[48] W. Feng, Z. Hu, W. Wu, J. Yan, W. Ouyang, Multi-object tracking with multiple cues and switcher-aware classification, arXiv preprint arXiv:1901.06129 (2019). .
[49] J. Peng, T. Wang, W. Lin, J. Wang, J. See, S. Wen, E. Ding, TPM: Multiple object tracking with tracklet-plane matching, Pattern Recognition 107480 (2020).
[50] Y. Zhang, H. Sheng, Y. Wu, S. Wang, W. Ke, Z. Xiong, Multiplex labeling graph for near-online tracking in crowded scenes, IEEE Internet of Things Journal 7 (9) (2020) 7892–7902.
[51] M. Babaee, Z. Li, G. Rigoll, A dual CNN–RNN for multiple people tracking, Neurocomputing 368 (2019) 69–83.
[52] B. Pang, Y. Li, Y. Zhang, M. Li, C. Lu, Tubetk: Adopting tubes to track multi-object in a one-step training model, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 6308–6318.
[53] X. Lin, C.-T. Li, V. Sanchez, C. Maple, On the detection-to-track association for online multi-object tracking, Pattern Recognition Letters 146 (2021) 200–207.
[54] J. Peng, C. Wang, F. Wan, Y. Wu, Y. Wang, Y. Tai, C. Wang, J. Li, F. Huang, Y. Fu, Chained-tracker: Chaining paired attentive regression results for end-to-end joint multiple-object detection and tracking, in: European Conference on Computer Vision, Springer, 2020, pp. 145–161. .
[55] J. Xiang, G. Xu, C. Ma, J. Hou, End-to-end learning deep CRF models for multi-object tracking deep CRF models, IEEE Transactions on Circuits and Systems for Video Technology 31 (1) (2020) 275–288.
[56] I. Papakis, A. Sarkar, A. Karpatne, GCNNMatch: Graph convolutional neural networks for multi-object tracking via sinkhorn normalization, arXiv preprint arXiv:2010.00067 (2020). .
[57] A. Sadeghian, A. Alahi, S. Savarese, Tracking the untrackable: Learning to track multiple cues with long-term dependencies, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 300–311, https://doi.org/10.1109/ICCV.2017.41.

Yating Liu received her B.Eng. degree from the Civil Aviation University of hina in 2014. She is currently a Ph.D. student at the State Key Laboratory for Management and Control of Complex Systems, Institute of Automation, Chinese Academy of Sciences as well as University of Chinese Academy of Sciences. Her research interests include visual object tracking, machine learning, and intelligent transportation systems.

Tianxiang Bai received his bachelor's degree from the Zhejiang University in 2013. He is currently a Ph.D. student at the Department of Automation, University of Science and Technology of China as well as The State Key Laboratory for Management and Control of Complex Systems, Institute of Automation, Chinese Academy of Sciences. His research interests include robotics, reinforcement learning and multi-object tracking.

Yonglin Tian received his bachelor's degree from the University of Science and Technology of China in 2017. He is currently a Ph.D. student at the Department of Automation, University of Science and Technology of China as well as The State Key Laboratory for Management and Control of Complex Systems, Institute of Automation, Chinese Academy of Sciences. His research interests include computer vision and intelligent transportation systems.

Yutong Wang received her Ph.D. degree in control theory and control engineering from University of Chinese Academy of Sciences in 2021. After that, she joined the Institute of Automation, Chinese Academy of Sciences and became an Assistant Professor at the State Key Laboratory for Management and Control of Complex Systems. Her research interests include computer vision and adversarial attack.

Jiangong Wang received his bachelor of engineering degree in electronic information and engineering from Tongji University, Shanghai, China, in 2018. He is currently working toward the Ph.D. degree in pattern recognition and intelligent systems, Institute of Automation, Chinese Academy of Sciences, Beijing, China, and the University of Chinese Academy of Sciences, Beijing. His research interests include parallel vision, unsupervised learning, traffic scene understanding and medical image processing.

Xiao Wang received the bachelor's degree in network engineering from the Dalian University of Technology, Dalian, China, in 2011, and the Ph.D. degree in social computing from the University of Chinese Academy of Sciences, Beijing, China, in 2016. She is currently an Associate Professor with the State Key Laboratory for Management and Control of Complex Systems, Institute of Automation, Chinese Academy of Sciences. She has published more than a dozen SCI/EI articles and translated three technical books (English to Chinese). Her research interests include social transportation, cybermovement organizations, artificial intelligence, and social network analysis. Dr. Wang has served the IEEE Transactions on Intelligent Transportation Systems, the IEEE/CAA Journal of Automatica Sinica, and ACM Transactions on Intelligent Systems and Technology as a Peer Reviewer with a good reputation.

Fei-Yue Wang (S'87-M'89-SM'94-F'03) received his Ph.D. degree in computer and systems engineering from the Rensselaer Polytechnic Institute, Troy, NY, USA, in 1990. He joined The University of Arizona in 1990 and became a Professor and the Director of the Robotics and Automation Laboratory and the Program in Advanced Research for Complex Systems. In 1999, he founded the Intelligent Control and Systems Engineering Center at the Institute of Automation, Chinese Academy of Sciences (CAS), Beijing, China, under the support of the Outstanding Chinese Talents Program from the State Planning Council, and in 2002, was appointed as the Director of the Key Laboratory of Complex Systems and Intelligence Science, CAS. In 2011, he became the State Specially Appointed Expert and the Director of the State Key Laboratory for Management and Control of Complex Systems.

His current research focuses on methods and applications for parallel intelligence, social computing, and knowledge automation. He is a fellow of INCOSE, IFAC, ASME, and AAAS. In 2007, he received the National Prize in Natural Sciences of China and became an Outstanding Scientist of ACM for his work in intelligent control and social computing. He received the IEEE ITS Outstanding Application and Research Awards in 2009 and 2011, respectively. In 2014, he received the IEEE SMC Society Norbert Wiener Award. Since 1997, he has been serving as the General or Program Chair of over 30 IEEE, INFORMS, IFAC, ACM, and ASME conferences. He was the President of the IEEE ITS Society from 2005 to 2007, the Chinese Association for Science and Technology, USA, in 2005, the American Zhu Kezhen Education Foundation from 2007 to 2008, the Vice President of the ACM China Council from 2010 to 2011, the Vice President and the Secretary General of the Chinese Association of Automation from 2008-2018. He was the Founding Editor-in-Chief (EiC) of the International Journal of Intelligent Control and Systems from 1995 to 2000, the IEEE ITS Magazine from 2006 to 2007, the IEEE/CAA JOURNAL OF AUTOMATICA SINICA from 2014-2017, and the China's Journal of Command and Control from 2015-2020. He was the EiC of the IEEE Intelligent Systems from 2009 to 2012, the IEEE TRANSACTIONS ON Intelligent Transportation Systems from 2009 to 2016, and is the EiC of the IEEE TRANSACTIONS ON COMPUTATIONAL SOCIAL SYSTEMS since 2017, and the Founding EiC of China's Journal of Intelligent Science and Technology since 2019. Currently, he is the President of CAA's Supervision Council, IEEE Council on RFID, and Vice President of IEEE Systems, Man, and Cybernetics Society.