Machine Vision and Applications (2024) 35:92
https://doi.org/10.1007/s00138-024-01580-w

RESEARCH

GOA-net: generic occlusion aware networks for visual tracking

Mohana Murali Dasari[1] · **Rama Krishna Gorthi**[1]

Received: 27 August 2023 / Revised: 23 June 2024 / Accepted: 26 June 2024 / Published online: 7 July 2024

Abstract

Occlusion is a frequent phenomenon that hinders the task of visual object tracking. Since occlusion can be from any object and in any shape, data augmentation techniques will not greatly help identify or mitigate the tracker loss. Some of the existing works deal with occlusion only in an unsupervised manner. This paper proposes a generic deep learning framework for identifying occlusion in a given frame by formulating it as a supervised classification task for the first time. The proposed architecture introduces an "occlusion classification" branch into supervised trackers. This branch helps in the effective learning of features and also provides occlusion status for each frame. A metric is proposed to measure the performance of trackers under occlusion at frame level. The efficacy of the proposed framework is demonstrated on two supervised tracking paradigms: One is from the most commonly used Siamese region proposal class of trackers, and another from the emerging transformer-based trackers. This framework is tested on six diverse datasets (GOT-10k, LaSOT, OTB2015, TrackingNet, UAV123, and VOT2018), and it achieved significant improvements in performance over the corresponding baselines while performing on par with the state-of-the-art trackers. The contributions in this work are more generic, as any supervised tracker can easily adopt them.

Keywords Occlusion · Siamese · RPN · Transformers · Visual object tracking

1 Introduction

Visual object tracking is a well-known task in computer vision, having applications in machine vision, surveillance, autonomous driving, and various other fields. Recently, the performance of visual trackers has increased significantly with Siamese region proposal networks (RPN) [1–3] and transformer encoder decoder (TED) networks [4–6]. However, it is still challenging to track objects under real challenging scenarios such as occlusion, background clutter, and deformation, to name a few [7]. When tracking an object of interest, it may often go behind the visible scene due to interference from other objects. This phenomenon is called occlusion, which causes the object's track loss, hindering its tracking. According to GOT-10k [7], occlusion is one of the most challenging factors that can easily cause tracker failures. During occlusion, the tracker is uncertain about the object's position and may be misguided. Identifying the frames in which the object of interest is occluded helps us provide effective feedback to improve the development of the trackers significantly.

Existing state-of-the-art trackers are broadly classified into Correlation filter based and Deep learning based [7]. Advanced Correlation filters employ deep learning network features for correlation (frequency-domain multiplication) to improve the performance [8] substantially. Given the difficulty in modeling uncountable natural objects that cause the target's occlusion, state-of-the-art correlation filter-based methods cannot directly address the occlusion challenge. On the other hand, the need for more annotated data hinders the development of deep learning based trackers to tackle this challenge.

Few works in the literature address occlusion's effect on the tracker's performance. However, to our knowledge, none of the existing works used annotated occlusion label information that is available in some of the tracking datasets [7, 9, 10]. Thus, occlusion awareness is not posed as a supervised learning task. Also, no separate metrics exist to quantify the occlusion identification accuracy at the frame level during the inference. In this work, we proposed a framework to

✉ Mohana Murali Dasari
ee18d001@iittp.ac.in

Rama Krishna Gorthi
rkg@iittp.ac.in

[1] Department of Electrical Engineering, Indian Institute of Technology, Yerpedu, Tirupati, Andhra Pradesh 517619, India

embed frame level occlusion information in the supervised training process with popular deep learning based tracking frameworks. It is achieved by formulating the frame level identification of occlusion as a binary classification problem. The proposed framework integrates the occlusion awareness learning branch along with the deep trackers' bounding box regression and classification branches. We demonstrate that this occlusion aware training also improves the performance of the trackers, along with inferring the occlusion status at frame level. We further propose a metric to quantify the frame level performance of the occlusion status prediction in the deep trackers. The performance is occlusion information available in a few testing datasets [9, 11],

The main contributions of this work are:

- To foster the development of frame level occlusion identification in the state-of-the-art visual object tracking frameworks.
- To infuse the occlusion awareness through the occlusion classification branch in visual tracking frameworks to effectively learn features during training.
- To improve the overall performance of the trackers by using occlusion status information in locating the object of interest during online tracking.
- To design an evaluation methodology for analyzing the occlusion status identification at the frame level, while tracking.

This work proposes a novel tracking framework, referred to as "*GOA-Net: Generic Occlusion Aware Network*". This supervised learning framework provides occlusion status at the frame level and improves overall tracking performance. The contributions are more generic as any supervised tracker can adapt the proposed strategy for occlusion aware learning. Two popular deep learning-based supervised tracking frameworks are considered for demonstration purposes. The first one is SiamRPN++ [1] from established Siamese Region Proposal Networks (RPN) based trackers, while the second one is TransT [4] from recent Transformer based trackers. These two baselines in the proposed framework are GOA-Net (RPN) and GOA-Net (TED). Information on visibility and occlusion available in the GOT-10k [7], LaSOT [9], and VID dataset [10] is effectively used for the proposition of this supervised occlusion aware learning strategy. Efficacy of the proposed methods is demonstrated on test datasets from GOT-10k [7], LaSOT [9], OTB2015 [12], TrackingNet [13] UAV123 [14] and VOT2018 [11]. An instance of the visible results of the baseline and proposed framework, under occlusion, are shown in Fig. 1.

2 Literature review

There exist various frameworks to track an object in short term. We can broadly classify them into a few categories, viz Correlation filter based trackers, Siamese trackers, deep reinforcement learning based trackers, transformer based trackers and diffusion based trackers.

2.1 Correlation filter based trackers

Given an object of interest in a video frame, correlation trackers employ correlation operation to find its location in the next frame. Correlation filters achieve higher computational performance in establishing this through well-formulated optimizations. Correlation filter-based approaches are most popular in visual object tracking. Henriques et al. [15] estimated the location of an object using frequency response maps of image and learned filter. They exploited circulant matrix properties to correlate various shifted search regions with the filter to track higher frames per second (fps). Danelljan et al. [16] improved the performance further by efficient feature representations and continuous correlation operation. Bhat et al. [8] achieved state-of-the-art performance by analyzing and evaluating effective shallow and deep feature combinations. Lu et al. [17] work consists of region proposals with channel regularisation within correlation filter learning. In Ref. [18], Fu et al. learned a unique correlation filter for dealing with latent distractions such as similar objects and clutter by carefully reducing its target response values. Parts based tracking to better deal with the partial occlusion are proposed in Refs. [19–21].

2.2 Siamese trackers

Advancements in deep learning attracted the tracking community a lot. To find the object's location in a frame, trackers search around its location in the previous frame and calculate correlation with a template. The spatial correlation was performed between the feature maps obtained from a learned network instead of pixels in the image in Siamese trackers. A famous work from Bertinetto et al. in SiamFC [22] used a fully convolutional Siamese Network for object tracking and inspired many works later. Going ahead with SiamFC, region proposal network (RPN) block consists of classification and regression branches, brought into tracking by Li et al. in SiamRPN [2]. Other works such as DaSiamRPN [3] and SiamRPN++ [1] also employed RPN block for enhancement in the performance. Some of the recent works like ATOM [23], IOU-SiamTrack [24] and SiamBAN [25], focused on maximising intersection over union (IOU). All of these can be called as Siamese RPN class of trackers. Li et al. [26] incorporated Correlation Filter and Siamese network into a single tracking framework to complement each other.

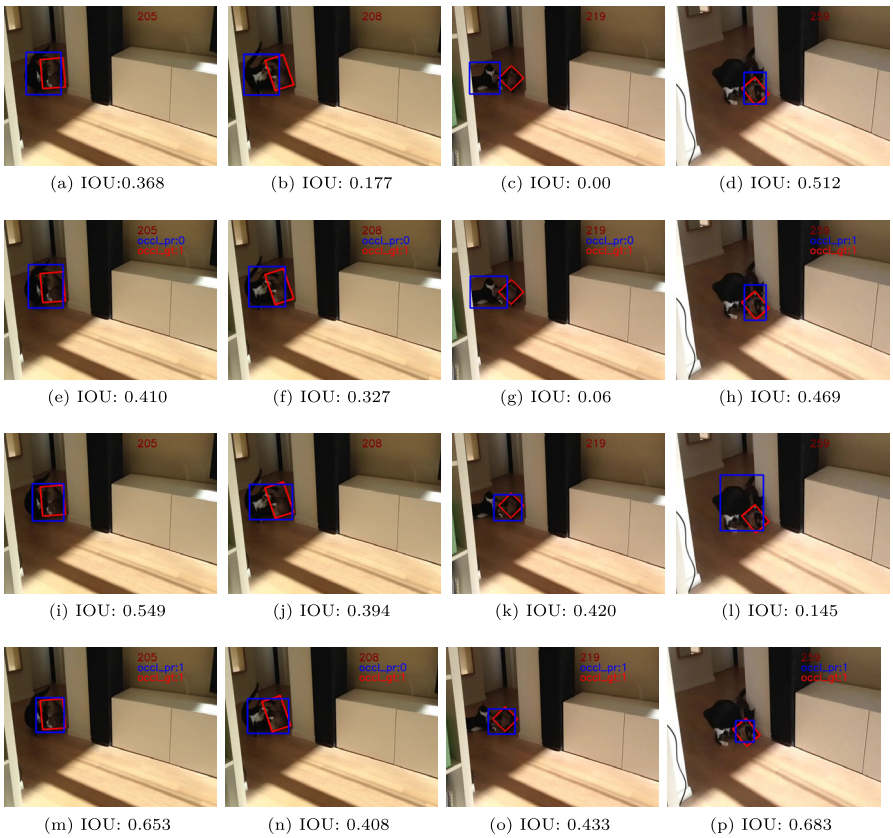

Fig. 1 Visual results on the Fernando sequence from the VOT2018 dataset. SiamRPN++, GOA-Net (RPN), TransT, and GOA-Net (TED) are in row 1, row 2, row 3, and row 4, respectively. Observe the successful prediction of occlusion status and increased overlap using GOA-Net models. Note that ground truth and predictions are in red and blue, respectively. Here occl_pr and occl_gt are the predicted and the ground-truth occlusion statuses, respectively. 1 and 0 correspond to the presence and absence of occlusion

2.3 Deep reinforcement learning based trackers

Deep reinforcement learning (DRL) is an area of machine learning concerned with how intelligent agents ought to take actions in an environment to maximize the notion of cumulative reward. In this direction, Sangdoo et al. used controlled sequential actions in their work [27] using deep reinforcement learning. Choi et al. [28] used a template selection strategy constructed by deep reinforcement learning methods for real-time tracking. Luo et al. [29] proposed an end-to-end tracking and camera control by using DRL. Q-Learning improved the tracking performance of systems with faults in Ref. [30]. In Ref. [31], deep reinforcement learning is exploited to deal with the localization delay in the action steps effectively and explore the long-term information in videos efficiently.

2.4 Transformer based trackers

Recent state-of-the-art trackers are based on transformers as they are designed to handle sequential input data with attention. Some works from this category include STARK (learning spatio-temporal transformer for visual tracking) [5], keep track (learning target candidate association to keep track of what not to track) [32]. Other works on transformers are TrDimp (transformer meets tracker: exploiting temporal context for robust visual tracking) [6], TransT (transformer tracking) [4] and latest MixFormer (end to end tracking with mixed attention). In most of these works, transformers with encoder and decoder blocks (TED) having attention layers have replaced RPN (of Siamese trackers).

2.5 Diffusion based tracker

Latest trends in reaction diffusion neural networks (RDNNs) learning methodologies extended for tracking. Inspiration is drawn from works such as cooperative-competitive neural networks with reaction-diffusion [33] and anti-disturbance state estimation for PDT-switched RDNNs utilizing time-sampling and space-splitting measurements [34]. Authors have used diffusion model to denoise the target under tracking in their work [35]. It models the diffusion process using a point set representation, which can better handle appearance variations for more precise localization.

2.5.1 Existing works dealing occlusion

Occlusion has been a long-standing challenge in visual object tracking [7]. Given the dynamic nature of the scenes in the real world, it is nearly impossible to model any visual feature of occlusions like shape or color. Except for the availability of target information to be tracked, no prior information about

the nature of occlusion is available. So, most existing works only devise strategies to mitigate the occlusion effect based on the target response score. Performance of single object trackers has raised many-folds through several variations of the well-known Siamese trackers [1, 2, 36], and boosted further by the introduction of transformers based trackers [4–6].

However, most of these state-of-the-art trackers either need to pay more attention to the specific challenge of occlusion or need a dedicated mechanism to deal with the occlusion.

Here, we review a few existing works that specifically deal with the occlusion. The content-adaptive progressive occlusion analysis (CAPOA) algorithm by Jiyan et al. [37] distinguished the target and outliers by combining the information provided by spatiotemporal context, reference target, and motion constraint. Accurate tracking of an occluded target was achieved by refining the target location using variant mask template matching (VMTM). To deal with the template drift, they have proposed a drift-inhibiting masked Kalman appearance filter (DIMKAF), which accurately evaluates the influence of template drift when updating the masked template. Finally, a local best match authentication (LBMA) algorithm was used to handle complete occlusions.

Deepak et al. [38], in their work SiamFC-SD, used structured dropouts in feature maps to mimic the changes under occlusion. Wu et al. [39] used a hard example discrimination method to estimate occlusion occurrence. Wenil et al. [40] used depth information to predict occlusion and precise object location. Fan et al. [41] proposed predefined masks at different locations and took these masks as the conditions to guide occlusion-aware feature learning. Parts-based tracking was proposed to deal with the partial occlusion in a series of works with scale adaptation (OAPT) [42], using geometry constraint and attention selection[19], applying discriminative correlation filters [20], online latent structured learning [21]. These works reflect the importance of occlusion information in tracking a visual object.

However, the existing methods for addressing occlusion [38–42] are unsupervised, and there was no direct metric to quantify their performance on occlusion identification ability as well. This work focuses on proposing supervised occlusion aware networks and highlights their ability to identify occlusion for effective tracking.

3 Proposed method

This section presents the proposed supervised learning framework for frame level occlusion awareness. It also elaborates on loss functions and implementation details for the same. The work is different from the existing visual object tracking works that deal with occlusion in the following ways:

- Existing methods deal with occlusion in an unsupervised manner. The proposed method is based on supervised learning of occlusion, leading to an effective representation of features under occlusion.
- GOA-Net identifies occlusion at the frame level and uses this information to update the object location.
- Existing methods are specific to a tracking framework and demonstrated only on one or two datasets. However, the proposed method is generic to integrate with many tracking frameworks and is demonstrated on six diverse datasets using two popular and emerging frameworks.
- This method established an evaluation methodology using annotated labels to analyze the tracker's performance under occlusion.

The proposed idea of supervised learning for occlusion awareness is demonstrated on the Siamese and transformer based trackers, as they are supervised and well-known state-of-the-art frameworks.

In the Siamese region proposal class of trackers such as SiamRPN [2], DaSiamRPN [3] and SiamRPN++ [1], the framework has a region proposal network (RPN) module trained by two branches, namely classification, and regression. Keeping the recent works like ATOM [23], IOU-SiamTrack [24], and SiamBAN [25], the regression branch in RPN is replaced with intersection over union (IOU) guided bounding box regression to make the framework more generic. From these variants, SiamRPN++ [1], one of the state-of-the-art RPN-based Siamese trackers, is used as a baseline method for demonstrating the effectiveness of the propositions. The proposed framework introduces additional "occlusion classification" to the RPN module to effectively learn and guide the tracking process. The proposed framework, with SiamRPN++ baseline, is referred to as "GOA-Net (RPN)".

The recent transformer based trackers (such as discriminative transformer tracking [43], transformer tracking (TransT) [4], STARK [5]) have attention-based transformer encoder and decoder (TED) modules trained by two branches similar to classification and regression in RPN. TransT [4], one of the state-of-the-art frameworks in the transformers-based tracker, is considered the baseline method in this category. The proposed framework introduces a new "occlusion classification" branch to the TED module to effectively learn and guide the tracking process. This proposed framework, with TransT as the baseline, is referred to as "GOA-Net (TED)".

The overall block diagram of the proposed generic occlusion aware network for tracking framework, referred to as GOA-Net, is shown in Fig. 2. From the first frame, a fixed patch (z) of 127×127, referred to as the "target image," centered around the object is cropped. A fixed patch ($\times$) of 255×255, referred to as the "search image," centered around the previous frame bounding box is cropped from the next

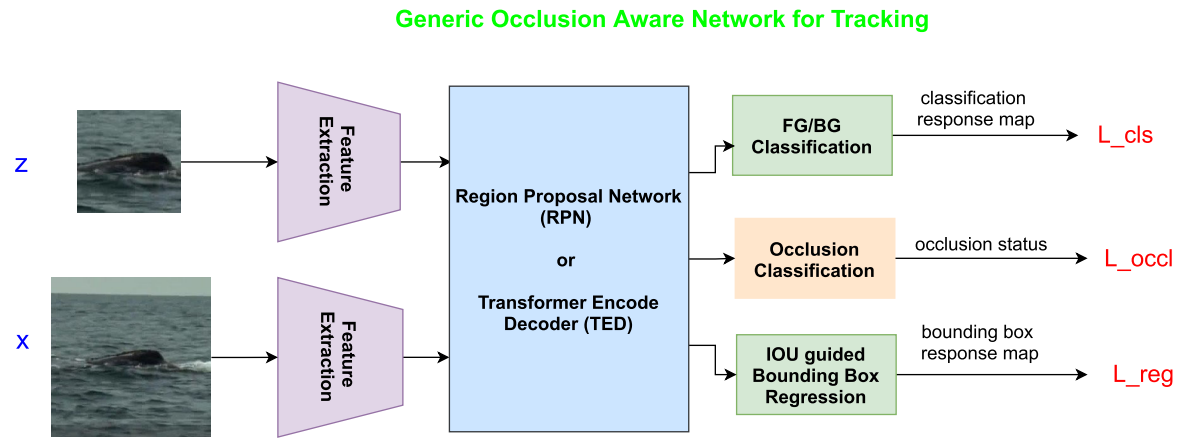

Fig. 2 Block diagram of generic occlusion aware network for tracking (GOA-Net)

frame onward. These patches are separately passed through a feature extraction network to obtain their embedded feature maps, as shown in Fig. 2. These feature maps are then fed to the RPN or TED to produce "occlusion status" along with "classification response map" and "bounding box response map," based on which the tracking is performed.

3.1 Foreground-background classification

The classification response map results from a depth-wise correlation between embedded feature maps of the "target image" and "search image" [1]. All Siamese region proposal frameworks use fixed bounding boxes (called Anchors) of size 8x8 with five aspect ratio variations (1/3,1/2,1,2,3) throughout the feature map following baseline [1]. Then, calculate the overlap of Anchor boxes with ground-truth bounding boxes and prepare labels for the classification branch. Label '1' for foreground class is assigned if $overlap \geq th1$. Label '0' for background class is assigned if $overlap \leq th2$. We note that the Anchor boxes with other overlaps are of no interest for training the classification branch. Following the same convention as in Refs. [1, 2, 22], we set $th1 = 0.7$ and $th2 = 0.3$ for classification label preparation.

A similar classification response map to label preparation is adopted in transformer trackers using baseline. A point to note here is that while GOA-Net(RPN) produces five classification response maps corresponding to five aspect ratios, GOA-Net (TED) produces only one classification response map. This difference in Transformer trackers (like TransT) is due to token-based query predictions of bounding box coordinates for all scales and aspect ratios. Hence, while training TED, label preparation is adopted from baseline [4]. Here, feature vector prediction corresponds to pixels in the ground-truth bounding box as positive samples; the rest are negative samples. Binary cross-entropy is employed for calculating classification loss (L_{cls}) as defined in the Eq. (1) for both works.

$$L_{cls} = -\{\sum_{j=1}^{M*N*k} g_j * log(p_j) + (1 - g_j) * log(1 - p_j)\} \tag{1}$$

Where g_j is the ground-truth label, $g_j = 1$ denotes the foreground, p_j is the predicted probability (classification response map) belonging to the foreground of the j^{th} box. Here, M, N, and k represent the width, height of the feature maps and number of feature maps, respectively (applicable for RPN and TED).

3.2 IOU guided bounding box regression

Inspired by the recent developments in SiamRPN variants [24, 25]. Other trackers such as Refs. [23, 44], we do incorporate IOU loss term L_{iou} as in Eq. (2) into the proposed framework. In the baseline, Siamese RPN++ tracker learning relies only on the regression bounding box, in addition to the classification scores, but does not explicitly employ the bbox IOU information as in the other state-of-the-art trackers [23–25]. We note that trackers learning using IOU of the ground truth and the predicted boxes had effective localization capability and better tracking performance. Hence, a similar IOU-guided bbox regression strategy is proposed in this framework, and this loss term makes the framework more generic for effective bounding box regression (refer to Sect. 5 for the performance analysis).

$$L_{iou}(j) = |pred_j - gt_j| \tag{2}$$

where $pred_j = IOU(anchor\_bb_j, pred\_bb_j)$, $gt_j = IOU(anchor\_bb_j, gt\_bb)$. Like other region proposal frameworks, instead of directly regressing the object bounding box (bbox), the center and size of the bbox are rela-

tively regressed using Anchor boxes. Consider j^{th} Anchor box $(A_{x,j}, A_{y,j}, A_{w,j}, A_{h,j})$ and ground-truth bounding box (T_x, T_y, T_w, T_h), we define

$$\delta[0, j] = \frac{T_x - A_{x,j}}{A_{w,j}}, \qquad \delta[1, j] = \frac{T_y - A_{y,j}}{A_{h,j}}$$
$$\delta[2, j] = ln\frac{T_w}{A_{w,j}}, \qquad \delta[3, j] = ln\frac{T_h}{A_{h,j}}$$

L_{l1} is l_1-norm on relative regression bbox values $\delta[i, j], i = 0:3$ as in Eq. (3).

$$L_{l1}(j) = \sum_{i=0}^{3} L_1(\delta[i, j]) \tag{3}$$

Here, the IOU-guided bounding box regression loss is a linear combination of two losses, L_{l1} and L_{IoU}. Thus, overall regression loss L_{reg1} is defined as in Eq. (4) for training GOA-Net (RPN). Here, $\lambda_d = 1.2$ (taken from baseline [1]), $\lambda_i = 1$ are set as regularisation parameters.

$$L_{reg1} = \sum_{j=1}^{M*N*k} \mathbf{1}_{g_j=1}\{\lambda_d * L_{l1}(j) + \lambda_i * L_{IoU}(j)\} \tag{4}$$

For training GOA-Net (TED), L_{reg2} defined in Eq. (5) is used and set $\lambda_l = 2, \lambda_g = 5$ following the baseline [4]. Here, L_{l1} is the same as in Eq. (3), and L_{giou} is Generalised IOU loss (GIOU) introduced in Ref. [45] and used in baseline [4]. Since TransT does not use Anchor boxes, the predicted regression may have zero overlaps with the ground truth bounding box. GIOU loss solves vanishing gradients for non-overlapping cases; hence, it is considered over IOU loss here. On the other hand, GIOU loss has slow convergence and inaccurate regression, especially for the boxes with extreme aspect ratios, thus not applied in the GOA-Net (RPN).

$$L_{reg2} = \sum_{j=1}^{M*N} \mathbf{1}_{g_j=1}\{\lambda_l * L_{l1}(j) + \lambda_g * L_{giou}(j)\} \tag{5}$$

Only positive samples contribute to the regression loss in the two equations above.

3.3 Occlusion classification

As discussed in the introduction, occlusion awareness is crucial in understanding the tracking status and designing trackers. To learn and use the occlusion status (presence/absence) in tracking, a binary occlusion classification branch is proposed to the existing tracker (as a part of RPN or TED) frameworks. This occlusion branch follows the same architecture as the RPN or TED classification branch and generates an occlusion response map. Occlusion label data is used for training these frameworks in a supervised learning strategy. Since the occlusion label is a single value for the entire frame, the predicted occlusion status is calculated as the average of the occlusion response map. The required label information is obtained from various datasets (refer Sect. 3.3.1). This work is the first deep neural network tracker framework trained and tested using annotated occlusion labels, paving a new direction for supervised occlusion identification in tracking.

The proposed model incorporated an additional branch for occlusion classification at the frame level. Supervised learning of occlusion classification differentiates the current work from the existing unsupervised occlusion identification frameworks [39]. Further, differing from the predefined masks used in Ref. [41] to mimic the occlusion, this framework brings occlusion awareness to features through supervised learning using annotated occlusion labels. While learning the features, occlusion awareness is incorporated into the network through a binary classifier and trained with occlusion loss (L_{occl}) as defined in Eq. (6).

$$L_{occl} = -\{g * log(os) + (1 - g) * log(1 - os)\} \tag{6}$$

where 'os' is the occlusion status, computed as the mean value of the occlusion response map (being '1' for full occlusion, '0' for no occlusion), and g is the ground-truth label for the status of occlusion. The mean value is considered because of the availability of occlusion labels only at the frame level. Occlusion response map size is on par with the classification response map to ensure occlusion-aware effective learning of features. Since the occlusion branch is trained and tested from target and search features, this proposition is generic and applicable to many supervised trackers.

3.3.1 Occlusion labels preparation

Since occlusion can cover the object of interest partially or fully and the labeling for occlusion is subjective, the available information on occlusion status is not uniform across these datasets. Customized approaches are followed for each dataset, as described below.

1. *VID* [10] this is a detection dataset where the occlusion label is available at the bounding box level in an XML file [10], and was set to '1' for occlusion.
2. *LaSOT* [9] this dataset has a supporting file with the separate label '1' corresponding to occlusion at the frame level.
3. *GOT-10k* [7] this dataset has 'cover label' files for each video at the frame level. 'cover label' corresponding to partial visibility (in terms of the visibility ratio, 0 being full occlusion, seven being full visible) encoded into eight

levels. Hence, the occlusion label '1' is considered for zero visibility (value 0).

3.4 Training and implementation

We have trained the models using NVIDIA Ge Force GTX 1080 Ti, and it took one week for GOA-Net(RPN) and three weeks for GOA-Net (TED) for every single experiment. The number of learnable parameters and operational speed of the tracker in terms of frames per second are now included in Table 1 (refer to Section 4.1, GOT-10k dataset results). However, the key advantage of these trained models is the ability to work at very low inference times, which are of the order of milliseconds/frame for the proposed trackers.

The baseline trackers and the corresponding GOA-Net variants are trained and evaluated under same conditions such as learning rate, number of epochs, and processor (Ge Force GTX 1080 Ti in our case). This is to ensure fair comparison and quantify the the performance improvements from the proposed learning strategy with occlusion information in the respective networks.

3.4.1 GOA-Net (RPN) tracker

This tracker is trained using four datasets as in baseline [1] (COCO [46], DET [10], VID [10], and YOUTUBE-BB [47]). Here, occlusion information is available in the VID dataset only and utilized during training. ResNet50 [48] architecture with pre-trained weights is used for feature extraction. We have initialised the common parameters as in baseline, while other perimeters are identified by empirical experiments. From the literature and following regular practices in the deep learning models training, we used Xavier and He initializations for other model parameters. Total training loss (L_{total}) in Eq. (7) is a combination of three losses as defined in Eqs. (1), (4) and (6).

$$L_{total} = L_{cls} + L_{reg1} + L_{occl} \tag{7}$$

This network is trained for 20 epochs. Adam optimizer is used with an exponentially decaying learning rate from 0.005 to 0.0005. After ten epochs, the feature extraction network is also fine-tuned in the training processes as in baseline [1].

3.4.2 GOA-Net (TED) tracker

This tracker is trained using four datasets as in baseline [4] (COCO [46], GOT-10k [7], LaSOT [9] and TrackingNet [13]). Occlusion labels from LaSOT and GOT-10k datasets are employed in training. ResNet50 [48] architecture with pre-trained weights is used for feature extraction. Total training loss (L_{total}) in Eq. (8) is a combination of three losses as defined in Eqs. (1), (5) and (6).

$$L_{total} = L_{cls} + L_{reg2} + L_{occl} \tag{8}$$

This network is trained for 1000 epochs with 1000 iterations per epoch. Adam optimizer is used, and the learning rate is set to 1e-4 (1e-5 after 500 epochs) with a weight decay of 1e-4. The feature extraction network is included in the training with a learning rate of 1e-5.

To understand the contribution of the proposed supervised occlusion status identification, we present visual results on a few occluded frames and their class activation maps (CAM) in Fig. 3. Columns one to four of Fig. 3 shows "target image" and "search image" pairs and corresponding gradient class activation (grad-cam) of "classification response map" (denoted with CR_CAM) and "occlusion status" (denoted with OR_CAM), respectively. In the first case, objects are directly visible in row-1 and row-2, so both activation maps (CR_CAM and OR_CAM) respond somewhat similarly. However, in row-3 and row-4, where objects are partially occluded cases, the corresponding CR_CAM falsely represents the object localization. At the same time, the OR_CAM is pointing at the object of interest, hidden behind an occlusion. In summary, occlusion status provides complementary information in tracking and, in particular, identifies the object's approximate location when occluded.

3.5 Online tracking

In addition to the supervised learning for occlusion identification, as discussed in the last section, we propose employing this occlusion status information as explicit feedback for guiding the bbox regression. It is incorporated as follows. In each frame, the mean value of the occlusion response map is used as the occlusion status (os). Along with methodology from baseline works in updating the bounding box, using this os information, GOA-Net (RPN and TED variants) additionally uses occlusion status information as in Eq. (9).

$$\begin{aligned} bbox[t] &= (1-\alpha) * bbox[t-1] + (\alpha) * bbox_{pred}[t]) \\ \alpha &= exp(-os) \\ here,\ os &= occlusion\ status \\ &= mean\ of\ occlusion\ response\ map \end{aligned} \tag{9}$$

Note that in full occlusion scenarios, the mean os value will be high (close to 1), and thus this update encourages relying more on the previous bbox information. While for no occlusion case $os = 0$, the current regressed bbox information is used. Hence, the proposed framework facilitates the frame-level identification of occlusion and its use in improved estimation of object track even in occlusion. This

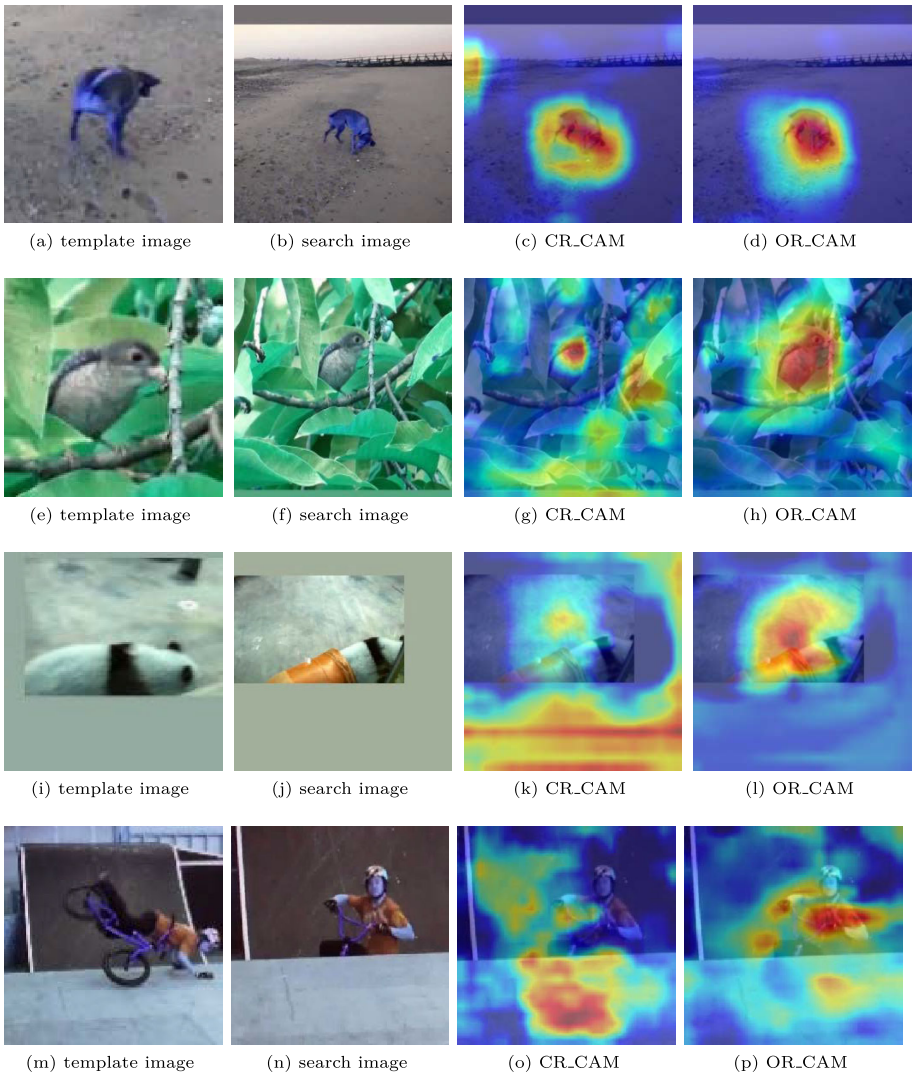

Fig. 3 Visualising gradient class activation maps of classification (CR_CAM) and occlusion branches (OR_CAM) in GOA-Net. Rows 1 and 2 depict visible objects, while rows 3 and 4 depict occluded objects [rows 1 and 3 for GOA-Net (RPN) and rows 2 and 4 for GOA-Net (TED)]. Observe the role of occlusion status (column 4) in providing complementary information for identifying objects in all scenarios

ability of the GOA-Net tracking framework offers robustness to tracking failures due to occlusion.

3.5.1 Methodology for evaluation of the model performance for occlusion status identification

To our knowledge, there is no specific existing metric to quantify the performance of trackers under occlusion at frame level. Identifying occlusion presence/absence can be considered as a classification task, and hence, accuracy is the preferred metric for its measure. However, this classification is a highly imbalanced task because the occlusion is absent in most cases compared to its presence. In such cases, regular accuracy cannot capture the real performance. Hence, we proposed the adoption of balanced accuracy as an appropriate metric for occlusion performance analysis in this work.

The following definitions are considered in devising the methodology for occlusion status evaluation. The performance of occlusion classification is calculated by considering occlusion presence as positive and absence as negative.

(a) True positive (TP): actual occlusion and the network predicts it (hit).
(b) False positive (FP): actual occlusion is not there, but the network predicts it to be there (false alarm).
(c) False negative (FN): actual occlusion is there, but the network misses it (miss).
(d) True negative (TN): actual occlusion is not there, and the network also agrees with it.
(e) We define positives, Pve = (TP+FN) and negatives, Nve = (TN+FP)

Based on the above notations, balanced accuracy (BA) is defined in Eq. (10). Since the occlusion classification task is highly unbalanced, balanced accuracy is adopted to quantify and reflect the network's performance in occlusion prediction.

$$\begin{aligned} Sensitivity\ or\ True\ Positive\ Rate\ (TPR) &= TP/Pve \\ Selectivity\ or\ True\ Negative\ Rate\ (TNR) &= TN/Nve \\ Balanced\ Accuracy\ (BA) &= [TPR \\ &\quad + TNR]/2 \end{aligned} \tag{10}$$

This BA metric measures how well a tracker could perform occlusion status identification at frame level. The higher the BA, the better the tracker's ability to identify the occlusion status information.

4 Evaluation results and analysis

This section compares the proposed GOA-Net trackers (RPN and TED variants) with respective baseline (SiamRPN++ [1], and TransT [4]) trackers and other state-of-the-art tracking networks. Given more data for training, deep learning-based models are likely to improve performance. In the original SiamRPN++ [1] and DaSiamRPN [3] works, the networks were trained with four datasets (VID [10], YouTube-BB [47], DET [10] and COCO [46]). Further, ATOM [23] and SiamBAN [25] were trained with the above four datasets and two additional datasets. SiamRCNN [36] also trained under different datasets ([7, 9, 10, 47]). Other factors, such as the number of epochs, batch size, and graphics cards, influence performance. *Hence, for fair evaluation, baseline work and its corresponding GOA-Net variant are trained and evaluated under the same conditions*. The proposed generic framework GOA-Net's (RPN and TED variants) performance is evaluated on six diverse datasets.[1] However, note that the ground truth for occlusion status is available only in LaSOT and VOT2018. Thus, performance evaluation in predicting occlusion status is restricted to these two datasets only.

4.1 GOT-10k dataset

GOT-10K [7] is the large-scale benchmark dataset covering many common object-tracking challenges. It uses average overlap (AO) and success rate$_T$ (% of frames with $overlap > T$) across all frames. Results from online server evaluation (i.e., AO and SR$_T$) are reported in Table 1. Here GOA-Net (TED) has shown improvements of 1.7%, 1.5%, and 2.5% in AO, SR$_{0.5}$ and SR$_{0.75}$ respectively over TransT. While the GOA-Net (RPN), has shown 4.4% and 6.5% improvement in AO and SR$_{0.5}$ respectively. Hence the proposed GOA-Net models show substantial improvements over the baselines, demonstrating the generality of occlusion aware learning in boosting the tracker performance. Computational complexity and processing speed of each tracker are tabulated for a fair understanding cost involved in improving performance.

[1] The top one, two, and three performers in each metric are indicated by red, blue, and green colors for better comparisons.

Table 1 Performance comparison of state-of-the-art trackers on GOT-10k

Tracker	AO↑	SR$_{0.5}$ ↑	SR$_{0.75}$ ↑	FPS ↑	LP
ECO [16]	0.316	0.309	0.111	48	$\mathcal{O}(10^4)$
SiamFC [22]	0.355	0.395	0.118	58	$\mathcal{O}(10^4)$
SiamFC-SD [38]	0.361	0.402	0.129	55	$\mathcal{O}(10^4)$
SiamRPN++ [1]	0.474	0.565	0.285	35	$\mathcal{O}(10^7)$
GOA-Net (RPN)	0.495	0.602	0.285	33	$\mathcal{O}(10^7)$
ATOM [23]	0.556	0.634	0.402	30	$\mathcal{O}(10^7)$
PrDimp [44]	0.634	0.738	0.543	40	$\mathcal{O}(10^7)$
TransT [4]	0.646	0.752	0.575	45	$\mathcal{O}(10^6)$
SiamRCNN [36]	0.649	0.728	0.597	5	$\mathcal{O}(10^8)$
GOA-Net (TED)	0.657	***0.763***	0.589	44	$\mathcal{O}(10^6)$
STARK-S50 [5]	0.672	0.761	0.612	50	$\mathcal{O}(10^6)$

AO: average overlap, SR$_{0.5}$: % of frames with overlap > 0.5, SR$_{0.75}$: % of frames with overlap > 0.75, FPS: frames Per second, and LP: learnable parameters

Table 2 Performance comparison of state-of-the-art trackers on LaSOT

Tracker	AUC↑	P$_{norm}$ ↑
ECO [16]	32.4	33.8
SiamFC [22]	33.6	42.0
TLFF [49]	49.6	57.7
SiamRPN++ [1]	49.9	58.6
GOA-Net (RPN)	49.9	59.4
ATOM [23]	51.5	57.6
PrDimp [44]	59.8	68.8
TransT [4]	60.3	69.8
GOA-Net (TED)	63.1	72.6
STARK-S50 [5]	63.3	72.9
SiamRCNN [36]	64.8	72.2

AUC: area under the curve, P$_{norm}$: normalised precision

4.2 LaSOT dataset

LaSOT [9] is a large-scale tracking dataset with high-quality annotations containing 1400 challenging videos: 1120 for training and 280 for testing. The one-pass evaluation (Success and Precision) method compares different tracking algorithms on the LaSOT test set. Here, Success measures overlap between predicted and ground-truth boxes, and Precision measures the center distance between predicted and ground-truth boxes. Note that the Precision (P) metric is sensitive to the target size and image resolution. Hence, Normalised Precision (P$_{norm}$) is introduced [9]. Compar-

Table 3 Performance metrics for occlusion evaluation on LaSOT

Tracker	TPR	TNR	Balanced accuracy
GOA-Net (RPN)	37.48	66.94	52.21
GOA-Net (TED)	49.02	85.64	67.33

ison of results using the metrics Success (also known as Area Under the Curve (AUC)) and Normalised Precision (P_{norm}) scores are noted in Table 2. This table shows that the GOA-Net (TED) performance is improved by 4.64% and 4% in terms of AUC and P_{norm}, respectively. This significant gain of the proposed framework is attributed to training with occlusion data, which implicitly enabled the network to learn only the true representations of the object features at non-occluded regions (refer Fig. 3). Also, compared to baseline SiamRPN++, GOA-Net (RPN) has shown improvement in P_{norm} with the same AUC. The success metric quantifies the overall success of the target track. Thus, the consistent improvements in metrics reflect the importance of training with occlusion labels.

The results for TransT and GOA-Net (TED) on the zebra-14 sequence are shown visually in Fig. 4. Note that the resumption after occlusion is improved by occlusion-aware training of GOA-Net (TED) compared to the base work (TransT).

4.2.1 Performance of occlusion status identification

The performance of the proposed framework under occlusion [using the metric defined in Eq. (10)] is reported in Table 3. Note that GOA-Net (TED) is trained on the LaSOT training dataset, while GOA-Net (RPN) is not trained on the same, following baseline [1]. This dataset has rich background context variations for each target and varied foreground class target objects. Hence, GOA-Net (TED), which learns with attention, performs well over GOA-Net (RPN) in occlusion status prediction on the LaSOT testing dataset. While occlusion presence is accurately identified by GOA-Net (TED) in 50% frames (TPR), occlusion absence is precisely declared in 85% of frames (TNR). In a nutshell, GOA-Net (TED) could identify the occlusion status (present or absent) correctly over 67.33% of frames, bringing the occlusion status information into the base tracker (TransT).

4.3 OTB2015 dataset

The OTB2015 dataset, one of the standard benchmarks for evaluating short-term single object trackers, consists of 100 videos with different challenges. It evaluates the tracker performance using Success (a measure of overlap between predicted and ground-truth boxes) and Precision (a measure of center distance between predicted and ground-truth bounding boxes).

Table 4 Performance comparison on OTB2015

Tracker	Success↑	Precision↑
SiamFC [22]	0.597	0.809
SiamFC-SD [38]	0.610	0.808
SiamMFA [50]	0.618	0.818
OAPT [42]	0.624	0.851
SiamBAN [25]	0.624	0.853
IOU-SiamTrack [24]	0.635	0.862
SiamON [41]	0.644	0.854
SiamRPN++ [1]	0.647	0.855
GOA-Net (RPN)	0.664	0.874
ATOM [23]	0.669	0.883
TransT [4]	0.675	0.879
GOA-Net (TED)	0.679	0.883
ECO [16]	0.693	0.913
SiamRCNN [36]	0.703	0.894

A comparison of different trackers with the proposed GOA-Net on the OTB2015 dataset is reported in Table 4. It can be noted that GOA-Net (RPN) has a 2% improvement in both Success and Precision metrics compared to the base tracker. Compared with baseline TransT, GOA-Net (TED) has shown only 0.5% improvement in Precision and Success values. A possible reason behind this difference in performance improvements is due to low-resolution images in the dataset and a deeper backbone network for TransT. The same reason may be attributed to the better performance of correlation filters like ECO [16] as compared to the deep trackers on this dataset. However, ECO tracker performance is not on par with the state-of-the-art on other datasets.

4.4 TrackingNet dataset

TrackingNet [13] is a large-scale tracking dataset that covers diverse object classes and has the same evaluation metrics as LaSOT (refer 4.2). Its test set contains 511 sequences with ground truth. Based on the official online evaluation provided for this dataset, the Success (AUC) and Normalised Precision (P_{norm}) results are reported in Table 5. GOA-Net (TED) surpassed the baseline method TransT in all metrics by achieving 80.6% and 86.2% as AUC and P_{norm}, respectively. It achieved state-of-the-art results in P_{norm}. Also, GOA-Net (RPN) has shown improved performance over SiamRPN++ in both AUC and P_{norm}.

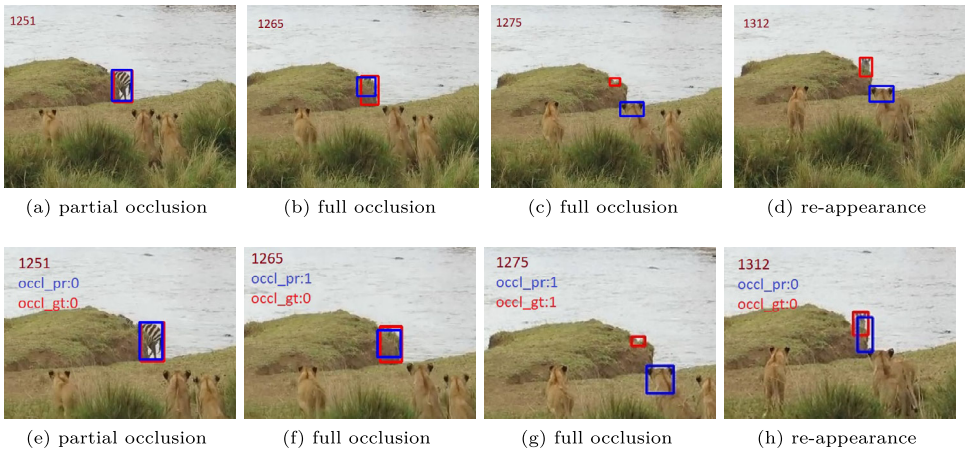

Fig. 4 TransT (row-1) and GOA-Net (TED) (row-2) on zebra-14 sequence from LaSOT dataset. Observer that TransT failed to resume after occlusion. At the same time, GOA-Net (TED) successfully predicted occlusion status and resumed tracking after occlusion

Table 5 Performance comparison of state-of-the-art trackers on TrackingNet

Tracker	AUC↑	P_{norm} ↑
ECO [16]	55.4	61.8
SiamFC [22]	57.1	66.3
SiamRPN++ [1]	68.9	76.5
GOA-Net (RPN)	69.5	77.3
ATOM [23]	70.3	77.1
PrDimp [44]	75.8	81.6
TMSDA [51]	78.1	83.3
STARK-S50 [5]	80.3	85.1
TransT [4]	80.5	85.8
GOA-Net (TED)	80.6	86.2
SiamRCNN [36]	81.2	85.4

AUC: area under the curve, P_{norm}: normalised precision

Table 6 Performance comparison on UAV123

Tracker	Success↑	Precision↑
ECO [16]	0.525	0.741
SiamFC [22]	0.523	0.731
SiamFC-SD [38]	0.535	0.736
IOU-SiamTrack [24]	0.549	0.748
SiamON [41]	0.568	0.774
SiamBAN [25]	0.584	0.772
SiamRPN++ [1]	0.584	0.780
DaSiamRPN [3]	0.585	0.795
GOA-Net (RPN)	0.598	0.791
SiamRCNN [36]	0.649	0.834
TransT [4]	0.663	0.860
TMSDA [51]	0.676	–
GOA-Net (TED)	0.678	0.875

4.5 UAV123 dataset

UAV123 dataset consists of 123 videos for evaluating short-term single object trackers. Like OTB2015, it also evaluates the tracker performance using success (a measure of overlap between predicted and ground-truth boxes) and Precision (a measure of center distance between predicted and ground-truth). This dataset is quite challenging given the similar objects and their small sizes. A comparison of the proposed GOA-Net with various trackers on the UAV123 dataset is reported in Table 6 with one pass evaluation (OPE). From Table 6, infer that GOA-Net (RPN) has a 0.6% improvement in Precision, with the same success over base work. Compared with baseline TransT, GOA-Net (TED) has shown 2.2% and 1.7% improvement in Success and Precision, respectively. This table reiterates the substantial performance improvement of both GOA-Net frameworks over the existing Siamese occlusion network (SiamON) [41] framework.

4.6 VOT2018 dataset

The VOT2018 dataset, a popular benchmark for evaluating short-term single object trackers, consists of 60 videos under varying conditions such as shape change, illumination, and occlusion. It uses metrics, namely accuracy (average overlap between predicted and ground-truth bounding boxes), robustness (measure for track failures), and expected average overlap (a statistical measure combining both accuracy and robustness) [11]. According to VOT evaluation protocol, EAO metric is considered for ranking the tracker performance. The comparison results of different trackers with the proposed GOA-Net framework on the VOT2018 dataset are reported in Table 7. GOA-Net (TED) has improved performance in all metrics over its baseline TransT. It can also be observed that GOA-Net (RPN) achieved a significant 6.1% improvement in Expected Average Overlap (EAO) over the base tracker SiamRPN++ due to a reduced number of tracker losses (measured by robustness). Visual results depicted in

Table 7 Performance comparison of state-of-the-art trackers on VOT2018

Tracker	EAO↑	Accuracy↑	Robustness ↓
SiamFC [22]	0.188	0.503	0.588
SiamMFA [50]	0.221	0.531	0.542
ECO [16]	0.280	0.484	0.276
IOU-SiamTrack [24]	0.301	0.565	0.328
SiamBAN [25]	0.319	0.582	0.347
SiamRPN++ [1]	0.321	0.601	0.337
TransT [4]	0.324	0.593	0.290
SiamON [41]	0.326	0.508	0.200
ATOM [23]	0.329	0.591	0.204
GOA-Net (RPN)	0.333	0.576	0.267
GOA-Net (TED)	0.350	0.602	0.267
SiamRCNN [36]	0.405	0.612	0.220

EAO expected average overlap

Table 8 Performance metrics for occlusion evaluation on VOT2018

Tracker	TPR	TNR	Balanced accuracy
GOA-Net (RPN)	22.97	73.89	48.43
GOA-Net (TED)	35.58	84.19	59.89

Fig. 5 confirm the same on an example sequence from the dataset.

In VOT evaluations, once the tracker losses its track completely, the track will be reinitialized to the ground truth, accounting for the tracker's failure. As shown in Fig. 5, the track is lost in Fig. 5c and reinitialized in Fig. 5d, while the proposed GOA-Net (RPN) continued the track, without the need for the reinitialization. Note that the loss of tracker is indicated by the tracker_loss flag of '0' in Fig. 5h against tracker_loss flag equal to '1' for baseline in Fig. 5d. These results are attributed to the occlusion awareness incorporated into the learned features.

4.6.1 Performance of occlusion status identification

Using metrics defined in Eq. (10), the performance of the proposed framework in identifying occlusion is measured and reported in Table 8. From Table 8, it is observed that GOA-Net (TED) has shown dominant performance over GOA-Net (RPN) in all metrics. Overall, the proposed GOA-Net achieved accurate occlusion status identification in 50–60% of frames, as inferred by the BA metric.

4.7 Comparison with the unsupervised occlusion aware networks

In this subsection, we explicitly compare the proposed tracker's performances with two existing state-of-the-art unsupervised occlusion aware networks [38, 41]. Regarding absolute performance, GOA-Net(TED) has exceeded all these trackers under consideration. Note that, SiamFC-SD used SiamFC as a baseline, while SiamON and GOA-Net (RPN) used SiamRPN++ as baselines, while GOA-Net(TED) used the TransT tracker (one of the state-of-the-art) as a baseline. Hence, the performance improvements of these trackers over their corresponding baselines are reported in Table 9. All the entries represent percentage gain regarding average overlap (or similar equivalent metric). SiamFC-SD and GOA-Net (RPN) have around 2% gain on OTB2015 and UAV123 datasets, while GOA-Net (RPN) significantly improves GOT-10k. An interesting comparison would be GOA-Net (RPN) versus SiamON, which has the same baseline. From Table 9, it is evident that in all datasets for which results are available, GOA-Net(RPN) outperformed SiamON. Moreover, GOA-Net also quantifies occlusion performance at frame level, clearly reflecting the importance of supervised learning for occlusion status identification as in the present work.

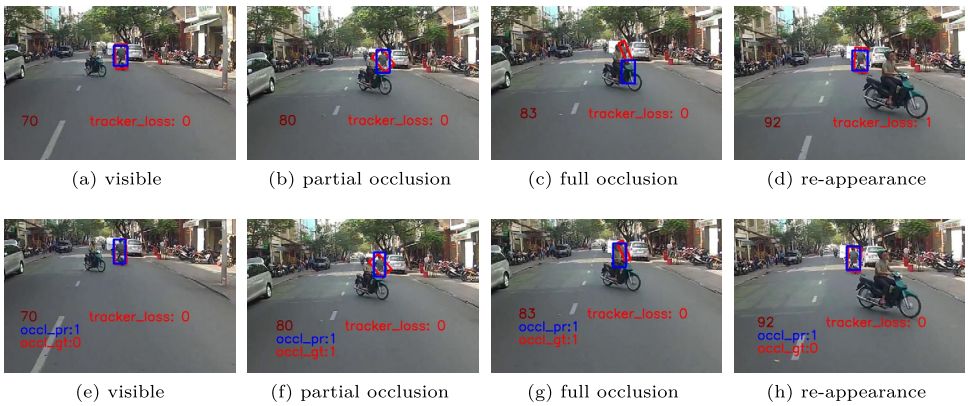

Fig. 5 SiamRPN++ (row-1) and GOA-Net (RPN) (row-2) on traffic sequence from VOT2018 dataset. While the SiamRPN++ method failed in the presence of occlusion, GOA-Net (RPN) tracker has continued tracking without the tracker loss. In the VOT2018 baseline evaluation, tracker loss is declared when IOU between predicted and ground-truth bounding boxes is zero. In such times, a tracker is reinitialized after five frames with a bounding box near the ground-truth [11]

Table 9 Percentage improvement of unsupervised trackers versus GOA-Net over corresponding baselines on various datasets

Data set	SiamFC-SD	SiamON	GOA-Net (RPN)	GOA-Net (TED)
GOT-10k	1.6	–	4.4	1.7
OTB2015	2.1	0.0	2.6	0.5
UAV123	2.2	−2.7	2.3	2.2
VOT2018	–	1.5	6.6	8.1

Table 10 Ablation study on OTB2015 dataset

Tracker	Success↑	Precision↑
Model-1 (SiamRPN++)	0.647	0.855
Model-2 (Model-1 with L_{iou})	0.649	0.857
Model-3 (Model-1 with L_{occl})	0.650	0.862
Model-4 (Model-1 with L_{iou} and L_{occl})	0.658	0.868
Model-5 (TransT)	0.675	0.879
Model-6 (Model-2 with L_{occl})	0.679	0.883

5 Ablation study

This section studies the significance of each proposed loss term in the GOA-Net framework. Here Model-1 is the baseline (SiamRPN++ [1]) trained with only L_{cls} and L_{l1}, without L_{iou}, Model-2 is Mode-1 additionally trained with L_{iou}, Model-3 is Model-1 trained additionally with L_{occl}. While Model-4 is the proposed GOA-Net (RPN) trained with both additional loss terms L_{occl} and L_{iou}. Further, Model-5 is the baseline (TransT [4]), replacing the region proposal network with an attention-based feature fusion network. Model-6 is Model-5 trained with L_{occl}. This ablation study results in Table 10 reflect the improvement of the model by including these additional losses.

6 Conclusion and future work

In this work, a generic framework for visual object tracking (GOA-Net) is proposed by introducing the "occlusion classification" branch to deep trackers for learning occlusion status in a supervised manner. This branch helps in the effective learning of features and also provides occlusion status at the frame level. Two kinds of deep trackers with the "occlusion classification" branch (in SiamRPN++ and TransT) referred to as GOA-Net(RPN) and GOA-Net (TED) are proposed. Six diverse tracking video datasets including thousands of videos comprising various challenges are considered for evaluation of the performance of the proposed supervised occlusion-guided tracking and occlusion status identification. Experimental results of GOA-Net has shown 1–6% improvement in average overlap (AO) metrics across different datasets. Occlusion performance at frame level is exclusively quantified with balanced accuracy. This metric and results of GOA-Net serve as a baseline and can be used for a measure of performance under occlusion. Since the proposed changes are generic for any supervised learning tracker, this occlusion-aware framework acts as the new direction for the further evolution of the long-term trackers.

Author Contributions All authors contributed equally.

Funding No funds, grants, or other support was received.

Data availability All datasets used are openly accessible.

Code availability On demand.

Declarations

Conflict of interest The authors have no Conflict of interest to declare that are relevant to the content of this article.

Ethical approval All the principles of ethical and professional conduct have been followed.

Consent to participate Not applicable.

Consent for publication All authors have given their consent to publish this article.

References

1. Li, B., Wu, W., Wang, Q., Zhang, F., Xing, J., Yan, J.: SiamRPN++: Evolution of Siamese visual tracking with very deep networks. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 4277–4286 (2019). https://doi.org/10.1109/CVPR.2019.00441
2. Li, B., Yan, J., Wu, W., Zhu, Z., Hu, X.: High performance visual tracking with Siamese region proposal network. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 8971–8980 (2018). https://doi.org/10.1109/CVPR.2018.00935
3. Zhu, Z., Wang, Q., Li, B., Wu, W., Yan, J., Hu, W.: Distractor-aware Siamese networks for visual object tracking. In: European

Conference on Computer Vision (ECCV), pp. 103–119 (2018). https://doi.org/10.1007/978-3-030-01240-3_7
4. Chen, X., Yan, B., Zhu, J., Wang, D., Yang, X., Lu, H.: Transformer tracking. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 8122–8131 (2021). https://doi.org/10.1109/CVPR46437.2021.00803
5. Yan, B., Peng, H., Fu, J., Wang, D., Lu, H.: Learning spatio-temporal transformer for visual tracking. In: IEEE International Conference on Computer Vision (ICCV), pp. 10448–10457 (2021). https://doi.org/10.1109/ICCV48922.2021.01028
6. Wang, N., Zhou, W., Wang, J., Li, H.: Transformer meets tracker: exploiting temporal context for robust visual tracking. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1571–1580 (2021). https://doi.org/10.1109/CVPR46437.2021.00162
7. Huang, L., Zhao, X., Huang, K.: GOT-10k: a large high-diversity benchmark for generic object tracking in the wild. IEEE Trans. Pattern Anal. Mach. Intell. (TPAMI) **43**(5), 1562–1577 (2019). https://doi.org/10.1109/TPAMI.2019.2957464
8. Bhat, G., Johnander, J., Danelljan, M., Shahbaz Khan, F., Felsberg, M.: Unveiling the power of deep tracking. In: European Conference on Computer Vision (ECCV), pp. 493–509 (2018). https://doi.org/10.1007/978-3-030-01216-8_30
9. Fan, H., Lin, L., Yang, F., Chu, P., Deng, G., Yu, S., Bai, H., Xu, Y., Liao, C., Ling, H.: LaSOT: a high-quality benchmark for large-scale single object tracking. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 5369–5378 (2019). https://doi.org/10.1109/CVPR.2019.00552
10. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A.C., Fei-Fei, L.: ImageNet large scale visual recognition challenge. Int. J. Comput. Vis. (IJCV) **115**, 211–252 (2015). https://doi.org/10.1007/s11263-015-0816-y
11. Kristan, M., Matas, J., Leonardis, A., Vojir, T., Pflugfelder, R., Fernandez, G., Nebehay, G., Porikli, F., Čehovin, L.: A novel performance evaluation methodology for single-target trackers. IEEE Trans. Pattern Anal. Mach. Intell. (TPAMI) **38**(11), 2137–2155 (2016). https://doi.org/10.1109/TPAMI.2016.2516982
12. Wu, Y., Lim, J., Yang, M.-H.: Object tracking benchmark. IEEE Trans. Pattern Anal. Mach. Intell. (TPAMI) **37**(9), 1834–1848 (2015). https://doi.org/10.1109/TPAMI.2014.2388226
13. Muller, M., Bibi, A., Giancola, S., Alsubaihi, S., Ghanem, B.: Trackingnet: A large-scale dataset and benchmark for object tracking in the wild. In: European Conference on Computer Vision (ECCV), pp. 310–327 (2018). https://doi.org/10.1007/978-3-030-01246-5_19
14. Mueller, M., Smith, N., Ghanem, B.: A benchmark and simulator for UAV tracking. In: European Conference on Computer Vision (ECCV), pp. 445–461 (2016). https://doi.org/10.1007/978-3-319-46448-0_27
15. Henriques, J.F., Caseiro, R., Martins, P., Batista, J.: High-speed tracking with kernelized correlation filters. IEEE Trans. Pattern Anal. Mach. Intell. (TPAMI) **37**, 583–596 (2015). https://doi.org/10.1109/TPAMI.2014.2345390
16. Danelljan, M., Bhat, G., Shahbaz Khan, F., Felsberg, M.: ECO: Efficient convolution operators for tracking. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 6931–6939 (2017). https://doi.org/10.1109/CVPR.2017.733
17. Lu, X., Ma, C., Ni, B., Yang, X.: Adaptive region proposal with channel regularization for robust object tracking. IEEE Trans. Circuits Syst. Video Technol. (TCSVT) **31**(4), 1268–1282 (2021). https://doi.org/10.1109/TCSVT.2019.2944654
18. Fu, C., Ding, F., Li, Y., Jin, J., Feng, C.: Learning dynamic regression with automatic distractor repression for real-time UAV tracking. Eng. Appl. Artif. Intell. (EAAI). (2021). https://doi.org/10.1016/j.engappai.2020.104116
19. Fang, J., Wang, Q., Yuan, Y.: Part-based online tracking with geometry constraint and attention selection. IEEE Trans. Circuits Syst. Video Technol. (TCSVT) **24**(5), 854–864 (2014). https://doi.org/10.1109/TCSVT.2013.2283646
20. Liu, T., Wang, G., Yang, Q., Wang, L.: Part-based tracking via discriminative correlation filters. IEEE Trans. Circuits Syst. Video Technol. (TCSVT) (2016). https://doi.org/10.1109/TCSVT.2016.2637798
21. Yao, R., Shi, Q., Shen, C., Zhang, Y., Hengel, A.: Part-based robust tracking using online latent structured learning. IEEE Trans. Circuits Syst. Video Technol. (TCSVT) **27**(6), 1235–1248 (2017). https://doi.org/10.1109/TCSVT.2016.2527358
22. Bertinetto, L., Valmadre, J., Henriques, J.F., Vedaldi, A., Torr, P.H.S.: Fully-convolutional Siamese networks for object tracking. Lect. Notes Comput. Sci. (LNCS) **9914**, 850–865 (2016). https://doi.org/10.1007/978-3-319-48881-3_56
23. Danelljan, M., Bhat, G., Shahbaz Khan, F., Felsberg, M.: ATOM: accurate tracking by overlap maximization. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 4655–4664 (2019). https://doi.org/10.1109/CVPR.2019.00479
24. Dasari, M.M., Gorthi, R.K.S.S.: IOU - Siamtrack: IOU guided Siamese network for visual object tracking. In: IEEE Conference on Image Processing (ICIP), pp. 2061–2065 (2020). https://doi.org/10.1109/ICIP40778.2020.9191188
25. Chen, Z., Zhong, B., Li, G., Zhang, S., Ji, R.: Siamese box adaptive network for visual tracking. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 6667–6676 (2020). https://doi.org/10.1109/CVPR42600.2020.00670
26. Li, X., Huang, L., Wei, G., Wei, Z.: Online parallel framework for real-time visual tracking. Eng. Appl. Artif. Intell. (EAAI) **102**, 104266 (2021). https://doi.org/10.1016/j.engappai.2021.104266
27. Yun, S., Choi, J., Yoo, Y., Yun, K., Choi, J.Y.: Action-decision networks for visual tracking with deep reinforcement learning. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1349–1358 (2017). https://doi.org/10.1109/CVPR.2017.148
28. Choi, J., Kwon, J., Lee, K.M.: Real-time visual tracking by deep reinforced decision making. Comput. Vis. Image Underst. (CVIU) **171**, 10–19 (2018). https://doi.org/10.1016/j.cviu.2018.05.009
29. Luo, W., Sun, P., Zhong, F., Liu, W., Zhang, T., Wang, Y.: End-to-end active object tracking and its real-world deployment via reinforcement learning. IEEE Trans. Pattern Anal. Mach. Intell. (TPAMI) **42**, 107188 (2020). https://doi.org/10.1016/j.patcog.2019.107188
30. Wang, R., Zhuang, Z., Tao, H., Paszke, W., Stojanovic, V.: Q-learning based fault estimation and fault tolerant iterative learning control for MIMO systems. ISA Trans. **142**, 123–135 (2023). https://doi.org/10.1016/j.isatra.2023.07.043
31. Teng, Z., Zhang, B., Fan, J.: Three-step action search networks with deep Q-learning for real-time object tracking. Pattern Recognit. **101**, 107188 (2020). https://doi.org/10.1016/j.patcog.2019.107188
32. Mayer, C., Danelljan, M., Paudel, D.P., Van Gool, L.: Learning target candidate association to keep track of what not to track. In: IEEE International Conference on Computer Vision (ICCV), pp. 13444–13454 (2021). https://doi.org/10.1109/ICCV48922.2021.01319
33. Song, X., Wu, N., Song, S., Zhang, Y., Stojanovic, V.: Bipartite synchronization for cooperative-competitive neural networks with reaction-diffusion terms via dual event-triggered mechanism. Neurocomputing **550**, 126498 (2023). https://doi.org/10.1016/j.neucom.2023.126498
34. Song, X., Peng, Z., Song, S., Stojanovic, V.: Anti-disturbance state estimation for PDT-switched RDNNs utilizing time-sampling and space-splitting measurements. Commun. Nonlinear Sci. Numer. Simul. **132**, 107945 (2024). https://doi.org/10.1016/j.cnsns.2024.107945

35. Zhang, R., Cai, D., Qian, L., Du, Y., Lu, H., Zhang, Y.: DiffusionTracker: targets denoising based on diffusion model for visual tracking. Lect. Notes Comput. Sci. (LNCS) **14436**, 225–237 (2024). https://doi.org/10.1007/978-981-99-8555-5_18
36. Voigtlaender, P., Luiten, J., Torr, P.H.S., Leibe, B.: Siam R-CNN: visual tracking by re-detection. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 6577–6587 (2020). https://doi.org/10.1109/CVPR42600.2020.00661
37. Pan, J., Hu, B., Zhang, J.Q.: Robust and accurate object tracking under various types of occlusions. IEEE Trans. Circuits Syst. Video Technol. (TCSVT) **18**(2), 223–236 (2008). https://doi.org/10.1109/TCSVT.2007.913975
38. Gupta, D.K., Gavves, E., Smeulders, A.W.M.: Tackling occlusion in Siamese tracking with structured dropouts. In: International Conference on Pattern Recognition (ICPR), pp. 5804–5811 (2021). https://doi.org/10.1109/ICPR48806.2021.9412120
39. Wu, F., Zhang, J., Xu, Z.: Stably adaptive anti occlusion Siamese region proposal network for real time object tracking. IEEE Access **8**, 161349–161360 (2020). https://doi.org/10.1109/ACCESS.2020.3019206
40. Zhang, W., Yang, K., Xin, Y., Meng, R.: An occlusion-aware rgbd visual object tracking method based on Siamese network. In: IEEE International Conference on Signal Processing (ICSP), vol. 1, pp. 327–332 (2020). https://doi.org/10.1109/ICSP48669.2020.9320907
41. Fan, C., Yu, H., Huang, Y., Shan, C., Wang, L., Li, C.: Siamon: Siamese occlusion-aware network for visual tracking. IEEE Trans. Circuits Syst. Video Technol. (TCSVT) **33**(1), 186–199 (2023). https://doi.org/10.1109/TCSVT.2021.3102886
42. Wang, X., Hou, Z., Yu, W., Pu, L., Jin, Z., Qin, X.: Robust occlusion aware part based visual tracking with object scale adaptation. Pattern Recognit. **81**, 456–470 (2018). https://doi.org/10.1016/j.patcog.2018.04.011
43. Yu, B., Tang, M., Zheng, L., Zhu, G., Wang, J., Feng, H., Feng, X., Lu, H.: High-performance discriminative tracking with transformers. In: IEEE International Conference on Computer Vision (ICCV), pp. 9836–9845 (2021). https://doi.org/10.1109/ICCV48922.2021.00971
44. Danelljan, M., Gool, L.V., Timofte, R.: Probabilistic regression for visual tracking. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 7181–7190 (2020). https://doi.org/10.1109/CVPR42600.2020.00721
45. Rezatofighi, H., Tsoi, N., Gwak, J., Sadeghian, A., Reid, I., Savarese, S.: Generalized intersection over union: a metric and a loss for bounding box regression. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 658–666 (2019). https://doi.org/10.1109/CVPR.2019.00075
46. Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollar, P., Zitnick, C.L.: Microsoft COCO: common objects in context. In: European Conference on Computer Vision (ECCV), pp. 740–755 (2014). https://doi.org/10.1007/978-3-319-10602-1_48
47. Real, E., Shlens, J., Mazzocchi, S., Pan, X., Vanhoucke, V.: YouTube-BoundingBoxes: a large high-precision human-annotated data set for object detection in video. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 7464–7473 (2017). https://doi.org/10.1109/CVPR.2017.789
48. He, K., Zhang, X., Ren, S., Sun, J.: Deep Residual Learning for Image Recognition. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 770–778 (2016). https://doi.org/10.1109/CVPR.2016.90
49. Shi, Y., Wu, Z., Chen, Y., Dong, J.: Siamese tracker with temporal information based on transformer-like feature fusion mechanism. Mach. Vis. Appl. **34**(59), 59 (2023). https://doi.org/10.1007/s00138-023-01409-y
50. Pang, H., Han, L., Liu, C., Ma, R.: Siamese object tracking based on multi-frequency enhancement feature. Vis. Comput. **40**, 261–271 (2023). https://doi.org/10.1007/s00371-023-02779-0
51. Wang, J., Lai, C., Zhang, W., Wang, Y., Meng, C.: Transformer tracking with multi-scale dual-attention. Springer Complex Intell. Syst. **9**, 5793–5806 (2023). https://doi.org/10.1007/s40747-023-01043-1

Mohana Murail Dasari is currently pursuing his PhD with the Department of Electrical Engineering, Indian Institute of Technology, Tirupati, India (IITTP). His current research interest includes computer vision with special focus on visual object tracking.

Rama Krishna Sai Subramanyam Gorthi received his PhD from Indian Institute of Technology, Madras, India, in 2008. He was a postdoc fellow in the Fluminace group at INRIA, Rennes, France, from November 2009 to February 2011. He is currently working as an Professor, Electrical Engineering, Indian Institute of Technology, Tirupati, India. His research interests include visual object detection, object tracking, segmentation, 3-D shape extraction, and image superresolution.