

# Integrated Object Detection and Tracking with Tracklet-Conditioned Detection

Zheng Zhang<sup>1\*</sup> Dazhi Cheng<sup>1,2\*†</sup> Xizhou Zhu<sup>1,3\*†</sup> Stephen Lin<sup>1</sup> Jifeng Dai<sup>1</sup>

<sup>1</sup>Microsoft Research Asia

<sup>2</sup>Beijing Institute of Technology

<sup>3</sup>University of Science and Technology of China

{zhez, v-dachen, stevelin, jifdai}@microsoft.com

ezra0408@mail.ustc.edu.cn

## Abstract

Accurate detection and tracking of objects is vital for effective video understanding. In previous work, the two tasks have been combined in a way that tracking is based heavily on detection, but the detection benefits marginally from the tracking. To increase synergy, we propose to more tightly integrate the tasks by **conditioning the object detection in the current frame on tracklets computed in prior frames**. With this approach, the object detection results not only have high detection responses, but also improved coherence with the existing tracklets. This greater coherence leads to estimated object trajectories that are smoother and more stable than the jittered paths obtained without tracklet-conditioned detection. Over extensive experiments, this approach is shown to achieve state-of-the-art performance in terms of both detection and tracking accuracy, as well as noticeable improvements in tracking stability.

## 1. Introduction

Detection and tracking of moving objects is an essential element of many video understanding tasks, such as visual surveillance, autonomous navigation, and video captioning. Different from the more commonly addressed problem of object detection in still images, the additional temporal dimension in the video case introduces challenges that arise from scene dynamics. As an object moves, its appearance can vary due to occlusions, pose changes, and illumination differences. Imaging-related degradations such as motion blur and video defocus may affect object appearance as well. These factors collectively complicate the task of discovering objects and following their trajectories in a scene.

A common practice among existing methods for object detection and tracking is to detect objects in each frame

independently and then link the detected objects across frames to form tracklets [58, 25, 10, 17]. Applying detection and tracking in this sequential manner is of appealing simplicity. But unlike how detection assists tracking in this approach, there are no means for tracking to aid detection. Some methods attempt to address this issue by using tracklets to propagate detection bounding boxes from previous frames to the current frame, and then add these boxes to those produced by the detector [58, 25, 10]. However, with this *late integration* of tracking into the detection process, the tracking has no effect on the object detector itself. Rather, tracking exerts its influence only *after* the object detector has computed its bounding box results.

The disjoint design can be partially attributed to the relatively independent development of video object detection and multi-object tracking techniques. In the research of video object detection, the focus is on improving the per-frame object detection accuracy, while employing off-the-shelf trackers for post-processing [25, 71]. Meanwhile, for research on multi-object tracking, the detection results are usually assumed to be given by external object detectors applied on individual frames [22, 67, 4]. Such decoupling simplifies research for each task, but misses the benefit of integrating detection and tracking.

In this paper, we present an approach in which detection and tracking are more closely intertwined through an *early integration* of the two tasks. Instead of simply aggregating two sets of bounding boxes that are estimated separately by the detector and tracker, a single set of boxes is generated jointly by the two processes by conditioning the outputs of the object detector on the tracklets computed over the prior frames. In this way, the resulting detection boxes are both consistent with the tracklets and have high detection responses, instead of often having just one or the other in late integration techniques.

This advantage is illustrated in Fig. 1, which shows an example of detection boxes obtained with late integration

\*Equal contribution. †This work is done when Dazhi Cheng and Xizhou Zhu are interns at Microsoft Research Asia.

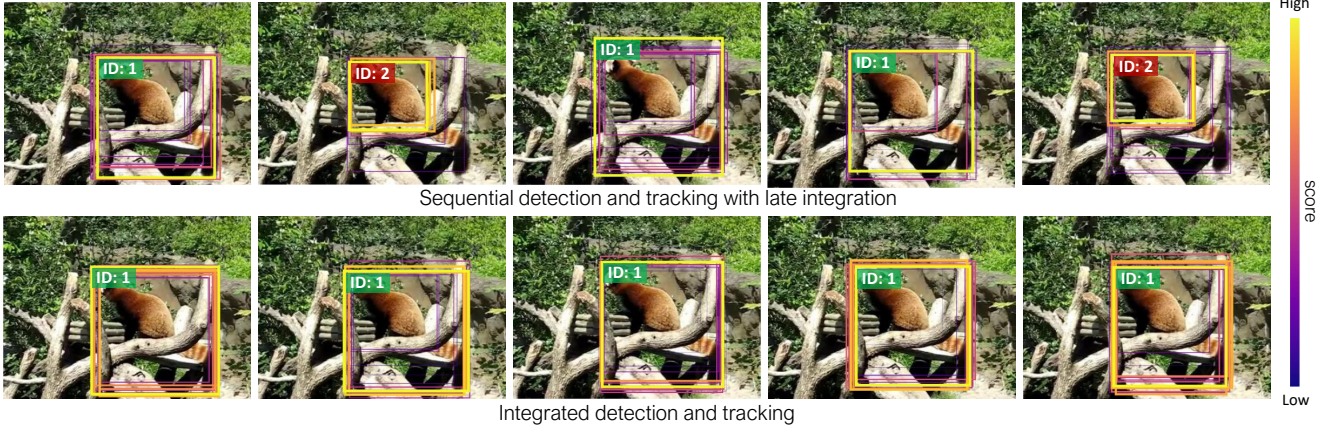


Figure 1. Visualized detection and tracking results by the previous late integration (top row) and the proposed early integration (bottom row) approaches. Shown are scored bounding boxes prior to non-maximum suppression (NMS), where the boxes are colored according to the corresponding category scores. The highest scored bounding box at each image is kept after NMS as the detection result, and is associated to existing tracklets (or initiates a new tracklet if association fails). The tracklet ID is indicated at the top-left corner of the detection box. More accurate and stable results are generated by the proposed approach of integrating detection and tracking early.

---

**Algorithm 1** Online sequential detection and tracking.

---

**input:** video frames  $\{\mathbf{I}_t\}_{t=0}^T$   
 $\mathbf{B}_0 := \text{DetectOnImage}(\mathbf{I}_0)$   
 initialize the tracklets  $\mathbf{D}_0$  from  $\mathbf{B}_0$   
**for**  $t = 1$  **to**  $T$  **do**  
    $\mathbf{B}_t := \text{DetectOnImage}(\mathbf{I}_t)$   
    $\mathbf{B}'_t := \text{PropagateBox}(\mathbf{D}_{t-1})$    Optional  
    $\mathbf{B}_t := [\mathbf{B}_t, \mathbf{B}'_t]$    Optional  
    $\mathbf{B}_t := \text{NMS}(\mathbf{B}_t)$   
    $\mathbf{D}_t := \text{AssociateTracklet}(\mathbf{D}_{t-1}, \mathbf{B}_t)$   
    $\mathbf{B}_t := \text{RescoreBox}(\mathbf{D}_t)$    Optional  
**end for**  
**output:** all tracklets  $\mathbf{D}_T$  and all boxes  $\{\mathbf{B}_t\}_{t=0}^T$

---

as done in [25], and with early integration via our tracklet-conditioned detection. Due in part to the aforementioned challenges of object detection in video, the boxes that have the highest detection scores without consideration of tracking may lie at various locations that deviate from the corresponding tracklet. Including boxes from late integration provides additional candidates, but they may not coincide closely with the actual object location due to errors in optical flow. With our tracklet-conditioned detection, temporal cues compiled over multiple frames can robustly guide the detector in a way that can compensate for variabilities in the detection of moving objects.

A natural outcome of tracklet-conditioned detection is increased stability in tracking. Besides generating detection boxes that more closely adhere to the target object, the conditioning also results in smoother trajectories where the detection boxes overlap the moving object in a consistent manner, as shown in Fig. 1 and detailed in Sec. 2.5. This

property is beneficial for applications such as live compositing of virtual makeup on faces, where a lack of stability would produce unwanted jittering of the makeup relative to the face.

We show how tracklet-based conditioning can be applied within a modern two-stage detector, employing it in both region proposal generation and classification. Through comprehensive evaluation on the Image VID [47] and MOT [30] datasets, it is shown that this provides state-of-the-art performance on both object detection and tracking. Noticeable gains in tracking stability are achieved as well. The code for this technique will be released.

## 2. Integrated Object Detection and Tracking

### 2.1. Background

Given a video of multiple frames  $\mathbf{I}_t, t = 0, \dots, T$ , our goal is to detect and to track all the object instances within it up to time  $t$ , which we denote as  $\mathbf{D}_t$ .  $\mathbf{D}_t = \{ \langle d_j^t, c_j^t \rangle \}, j = 1, \dots, m$ , where  $d_j^t$  denotes the  $j$ -th tracklet, and  $c_j^t$  denotes its corresponding category. For a tracklet  $d^t$ , it is composed of a set of bounding boxes detected on individual frames up to time  $t$ , as  $d^t = [b_k^{t_k}]$ , where  $b_k^{t_k}$  is the  $k$ -th bounding box in  $d^t$  at frame  $t_k$ , where  $t_k \leq t$ .

A scheme widely adopted in previous work [58, 25, 10, 17] is sequential detection and tracking, outlined in Algorithm 1. Here we describe an online variant of the algorithm. Given a new video frame  $\mathbf{I}_t$ , an object detector for individual images is first applied to produce per-frame detection results  $\mathbf{B}_t := \text{DetectOnImage}(\mathbf{I}_t)$ , where  $\mathbf{B}_t$  denotes a set of bounding boxes together with their corresponding category scores. Non-maximum suppression is then applied to remove redundant bounding boxes, result-

ing in  $\mathbf{B}_t := \text{NMS}(\mathbf{B}_t)$ . Then the tracking algorithm associates the existing tracklets  $\mathbf{D}_{t-1}$  to the detection results  $\mathbf{B}_t$ , producing tracklets up to frame  $\mathbf{I}_t$  as  $\mathbf{D}_t := \text{AssociateTracklet}(\mathbf{D}_{t-1}, \mathbf{B}_t)$ . Finally, the algorithm outputs all the tracklets  $\mathbf{D}_T$  up to time  $T$ .

To improve performance, two optional techniques are widely used as adds-on to better exploit tracklet information: (1) Box propagation, where detected boxes in the existing tracklets  $\mathbf{D}_{t-1}$  are propagated to the current frame  $\mathbf{I}_t$ , usually with the aid of flow information, to get boxes  $\mathbf{B}'_t := \text{PropagateBox}(\mathbf{D}_{t-1})$ . The propagated boxes are concatenated with the per-image detected boxes as  $\mathbf{B}_t := [\mathbf{B}_t, \mathbf{B}'_t]$ . The concatenated boxes undergo non-maximum suppression and are associated to the existing tracklets. This technique can be helpful when bounding boxes are not reliably detected in the new frame  $\mathbf{I}_t$ . (2) Box rescoring, a post-processing step to obtain more accurate classification scores for the detected boxes. For a bounding box newly associated to a tracklet, its score is set to the average score of all the bounding boxes that compose the tracklet. Here we denote this operation as  $\mathbf{B}_t := \text{RescoreBox}(\mathbf{D}_t)$ .

Including box propagation and/or box rescoring leads to better integration of detection and tracking. However, these techniques allow tracking to impact detection at only a late stage, after the per-image detection boxes are fixed. As a result, the detector cannot take full advantage of the tracking information.

## 2.2. Tracklet-Conditioned Detection Formulation

We aim at improving per-frame detection results through early integration of object detection and tracking. Our goal is for detection to exploit not only the image appearance of the current frame, but also information from tracklets recovered in the previous frames. We refer to this approach as tracklet-conditioned detection.

The problem can be formulated as: Given a set of candidate boxes  $\{b_i^t\}^1$  on frame  $\mathbf{I}_t$ , where  $b_i^t$  specifies the 4-D coordinates of the  $i$ -th box, together with the tracklets  $\{d_j^{t-1}\}_{j=1}^m$  up to frame  $\mathbf{I}_{t-1}$ , classify each box to different categories (including background) by estimating the score  $P(c|b_i^t, \{d_j^{t-1}\})$ . Based on the intuition that a candidate box should more likely take labels consistent with tracklets it is more likely to be associated with, the score is further decomposed to be conditioned on each tracklet, as

$$P(c|b_i^t, \{d_j^{t-1}\}) = \sum_{j=0}^m w(b_i^t, d_j^{t-1}) P(c|b_i^t, d_j^{t-1}), \quad (1)$$

where  $w(b_i^t, d_j^{t-1})$  specifies the association weight between box  $b_i^t$  and tracklet  $d_j^{t-1}$ . To account for newly detected

<sup>1</sup>The candidate boxes can be either dense sliding windows / anchor boxes in the first stage of two-stage object detectors, or sparse region proposals in the second stage.

objects that do not appear in existing tracklets, we include a null tracklet  $d_0^{t-1}$ , as detailed at the end of this subsection.

The score  $P(c|b_i^t, d_j^{t-1})$  is estimated based on both the appearance of the current frame and information from previous tracklets, as

$$P(c|b_i^t, d_j^{t-1}) \propto \exp(\log P_{\text{det}}(c|b_i^t) + \alpha \log P_{\text{tr}}(c|d_j^{t-1})), \quad (2)$$

where  $P_{\text{det}}(c|b_i^t)$  is predicted by the per-image object detector on  $\mathbf{I}_t$ ,  $P_{\text{tr}}(c|d_j^{t-1})$  is the classification probability for tracklet  $d_j^{t-1}$ , and the hyper-parameter  $\alpha$  balances the two log-likelihood terms ( $\alpha = 1$  by default).  $P(c|b_i^t, d_j^{t-1})$  is normalized over all the categories, by  $\sum_{c=0}^C P(c|b_i^t, d_j^{t-1}) = 1$ , where  $C$  denotes the number of foreground object categories, plus one for the background ( $c = 0$ ).  $P_{\text{tr}}(c|d_j^{t-1})$  is defined on the classification scores of all the bounding boxes assigned to tracklet  $d_j^{t-1}$ , in a running average fashion. Suppose tracklet  $d_j^{t-1}$  ( $j > 0$ ) is composed of box  $b_k^t$  and tracklet  $d_j^{t-1}$ , then  $P_{\text{tr}}(c|d_j^{t-1})$  is computed as

$$P_{\text{tr}}(c|d_j^{t-1}) = \frac{P(c|b_k^t, \{d_j^{t-1}\}) + \beta P_{\text{tr}}(c|d_j^{t-1}) \text{len}(d_j^{t-1})}{1 + \beta \text{len}(d_j^{t-1})} \quad (3)$$

where  $\beta$  is an exponential decay parameter ( $\beta = 0.99$  by default), and  $\text{len}(d_j^{t-1})$  denotes the trajectory length of  $d_j^{t-1}$ .

The association weight  $w(b_i^t, d_j^{t-1})$  is defined based on the intuition that box  $b_i^t$  is more likely to be associated to a tracklet that is visually similar:

$$w(b_i^t, d_j^{t-1}) := \exp(\gamma \cos(\mathcal{E}(b_i^t), \mathcal{E}(d_j^{t-1}))) \quad j > 0, \quad (4)$$

where  $\mathcal{E}(b_i^t)$  and  $\mathcal{E}(d_j^{t-1})$  are embedding features (128-D in our work) that encode the visual appearance of box  $b_i^t$  and tracklet  $d_j^{t-1}$  respectively, which are generated as described in Section 2.3. The cosine similarity between the embedding features is calculated and modulated by hyper-parameter  $\gamma$  (set to 8 in this paper) to be the log-likelihood of the association weight.

It is worth noting that new objects may appear in a video frame, and these objects will not be associated with any existing tracklets. To handle these cases, a null tracklet  $d_0^{t-1}$  is introduced. For every candidate box, its association weight with  $d_0^{t-1}$  is set to a constant, as

$$w(b_i^t, d_0^{t-1}) := \exp(R), \quad (5)$$

where  $R = 0.3$  in this paper. The association weights defined in Eq. (4) and Eq. (5) are further normalized over all the tracklets, as

$$w(b_i^t, d_j^{t-1}) := \frac{w(b_i^t, d_j^{t-1})}{\sum_{k=0}^m w(b_i^t, d_k^{t-1})}. \quad (6)$$



Thus, when a candidate box has low association weights with all the existing tracklets, its normalized association weight with the null tracklet will be high. For the null tracklet, its classification probability is set to a uniform distribution over all the categories, as

$$P_{\text{tr}}(c|d_0^t) = \frac{1}{C+1}. \quad (7)$$

### 2.3. Tracklet-Conditioned Two-stage Detectors

The proposed tracklet-conditioned detection algorithm can be readily applied in state-of-the-art object detectors. In this paper, we incorporate it into the two-stage Faster R-CNN [46] + ResNet-101 [20] detector, with OHM [50]. For this baseline, following the practice in [12], all the convolutional layers in ResNet-101 are applied on the whole input image. The effective stride in the conv5 blocks is reduced from 32 to 16 pixels to increase feature map resolution. The RPN [46] head is added on top of the conv4 features of ResNet-101. The Fast R-CNN [18] head is added on top of the conv5 features, and is composed of RoIpooling and two fully-connected (fc) layers of 1024-D, followed by the classification and the bounding box regression branches.

The tracklet-conditioned two-stage detector is exhibited in Figure 2. The tracklet conditioning in the second stage is relatively straightforward, with the equations in Section 2.2 applied on sparse region proposals.  $P_{\text{det}}(c|b_i^t)$  is predicted by the classification branch of the Fast R-CNN detection head. The box embedding  $\mathcal{E}_{s2}(b_i^t)$  (of the second stage) is computed by attaching a branch (consisting of a fully-connected layer) to the Fast R-CNN head, sibling to the classification and bounding box regression branches. The tracklet embedding  $\mathcal{E}_{s2}(d_j^t)$  ( $j > 0$ ) is updated based on the embedding features of the boxes associated to it, as

$$\mathcal{E}_{s2}(d_j^t) = \begin{cases} \eta \mathcal{E}_{s2}(b_k^t) + (1 - \eta) \mathcal{E}_{s2}(d_j^{t-1}) & \text{if } t > 0, \\ \mathcal{E}_{s2}(b_k^0) & \text{otherwise,} \end{cases} \quad (8)$$

where  $b_k^t$  denotes the detection box associated to tracklet  $d_j^t$  at time  $t$ , and  $\eta$  is the update weight parameter ( $\eta = 0.8$  by default). The box embedding features  $\mathcal{E}_{s2}(b_i^t)$  are compared to the tracklet embedding features  $\mathcal{E}_{s2}(d_j^{t-1})$  by Eq. (4) to obtain the association weights for the second stage.

We further apply the tracklet-conditioned detection in the first stage, to make use of tracklet information for improving region proposal quality. Compared to the application in the second stage, the key differences are that the candidate boxes are dense anchor boxes, and only two categories are involved, namely foreground and background. Given an anchor box  $b_i^t$ , its foreground probability  $P(\text{fg}|b_i^t, \{d_j^{t-1}\})$  is estimated by Eq. (1), with  $P_{\text{det}}(\text{fg}|b_i^t)$  predicted by the RPN classification branch and  $P_{\text{tr}}(\text{fg}|d_j^{t-1})$

---

#### Algorithm 2 Online integrated detection and tracking.

---

**input:** video frames  $\{\mathbf{I}_t\}_{t=0}^T$   
 $\mathbf{B}_0 := \text{DetectOnImage}(\mathbf{I}_0)$   
initialize the tracklets  $\mathbf{D}_0$  from  $\mathbf{B}_0$   
**for**  $t = 1$  **to**  $T$  **do**  
 $\mathbf{B}_t := \text{TrackletCondDetect}(\mathbf{I}_t, \mathbf{D}_{t-1})$   
 $\mathbf{B}_t := \text{NMS}(\mathbf{B}_t)$   
 $\mathbf{D}_t := \text{AssociateTracklet}(\mathbf{D}_{t-1}, \mathbf{B}_t)$   
 $\mathbf{D}_t := \text{RescoreTracklet}(\mathbf{D}_t)$   
**end for**  
**output:** all tracklets  $\mathbf{D}_T$  and all boxes  $\{\mathbf{B}_t\}_{t=0}^T$

---

computed as

$$P_{\text{tr}}(\text{fg}|d_j^{t-1}) = \sum_{c=1}^C P_{\text{tr}}(c|d_j^{t-1}), \quad (9)$$

which is the summation of the probability  $P_{\text{tr}}(c|d_j^{t-1})$  over all the foreground categories ( $c > 0$ ).

To derive the association weights for the first stage, two additional branches are added for producing the embedding features. The embedding features  $\mathcal{E}_{\text{anchor}}(b_i^t)$  for the dense anchor boxes are computed following the design in [35], via a sibling branch (consisting of a  $1 \times 1$  convolution) added to the RPN classification branch. Supposing there are  $K$  anchors at each location and the embedding features are 128-D, the output of the embedding branch is of dimension  $128 \times K$ . The tracklet embedding features  $\mathcal{E}_{s1}(d_j^t)$  of the first stage are computed in a manner similar to those of the second stage. An additional branch is added to the Fast R-CNN head to produce  $\mathcal{E}_{s1}(b_i^t)$ . After the RoIpooling layer, two additional fc layers of 1024-D are added (sibling to the existing two fc layers), followed by one fc layer to produce  $\mathcal{E}_{s1}(b_i^t)$ . Here, we tried different network designs for producing  $\mathcal{E}_{s1}(b_i^t)$ , as shown in Table 2. We find that adding two fc layers to reduce correlation between the embedding features of the two stages is beneficial for accuracy. Given  $\mathcal{E}_{s1}(b_i^t)$ ,  $\mathcal{E}_{s1}(d_j^t)$  is obtained by applying Eq. (8) (replacing the subscript s2 by s1 in the equation). Finally, the anchor box embedding features  $\mathcal{E}_{\text{anchor}}(b_i^t)$  are compared to the tracklet embedding features  $\mathcal{E}_{s1}(d_j^{t-1})$  by Eq. (4) to obtain the association weights for the first stage.

### 2.4. Training and Inference

**Inference.** Algorithm 2 presents the inference procedure for our integrated object detection and tracking with tracklet-conditioned detection. Given the input video frames  $\{\mathbf{I}_t\}_{t=0}^T$ , the per-image object detector is applied on the first frame  $\mathbf{I}_0$  to produce detection boxes,  $\mathbf{B}_0 := \text{DetectOnImage}(\mathbf{I}_0)$ . With these boxes, the tracklets  $\mathbf{D}_0$  are initialized (one tracklet per box). Then for each subsequent frame  $\mathbf{I}_t$ , tracklet-conditioned detection is applied

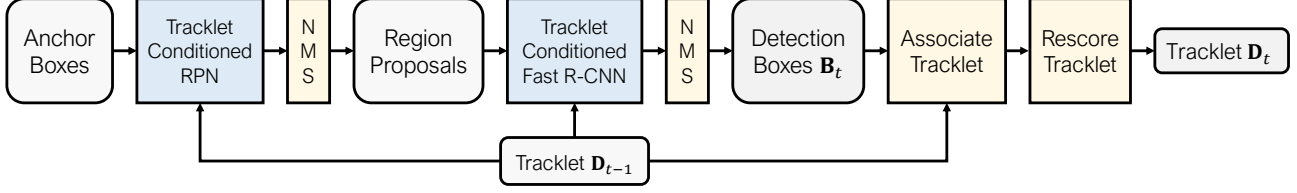


Figure 2. Tracklet-conditioned two-stage detectors.

and followed by non-maximum suppression, as  $\mathbf{B}_t := \text{TrackletCondDetect}(\mathbf{I}_t, \mathbf{D}_{t-1})$  and  $\mathbf{B}_t := \text{NMS}(\mathbf{B}_t)$ . As done in Algorithm 1, the detected bounding boxes  $\mathbf{B}_t$  are associated with the existing tracklets  $\mathbf{D}_{t-1}$  by  $\mathbf{D}_t := \text{AssociateTracklet}(\mathbf{D}_{t-1}, \mathbf{B}_t)$ . Then the obtained tracklets  $\mathbf{D}_t$  are rescored as  $\mathbf{D}_t := \text{RescoreTracklet}(\mathbf{D}_t)$ , by applying Eq. 3. Finally, the algorithm outputs all the tracklets  $\mathbf{D}_T$  and all the boxes  $\{\mathbf{B}_t\}_{t=0}^T$ .

**Training.** The network is trained to better detect objects based on image content and to better associate them across frames. Due to memory constraints, the forward pass in SGD training cannot be kept identical to that in inference. In each mini-batch, two consecutive frames from the same video,  $\mathbf{I}_{t-1}$  and  $\mathbf{I}_t$ , are randomly sampled. In the forward pass, bounding boxes are detected on  $\mathbf{I}_{t-1}$  based on image content only, as  $\mathbf{B}_{t-1} := \text{DetectOnImage}(\mathbf{I}_{t-1})$ . The detected boxes are matched to the ground-truth annotations  $\mathbf{B}_{t-1}^{\text{gt}}$  and  $\mathbf{B}_t^{\text{gt}}$ , on  $\mathbf{I}_{t-1}$  and  $\mathbf{I}_t$  respectively, to inject object detection loss and tracking loss.

The object detection loss is defined on  $\mathbf{B}_{t-1}$  and  $\mathbf{B}_{t-1}^{\text{gt}}$  in the same way as in conventional two-stage object detectors [46, 11]. It is composed of a foreground / background Softmax cross-entropy loss for region proposal scoring, an L1 regression loss for regressing proposal boxes, a  $(C+1)$ -way Softmax cross-entropy loss for detection scoring, and an L1 regression loss for regressing detected boxes.

The tracking loss is defined on  $\mathbf{B}_{t-1}$  and the  $\mathbf{B}_t^{\text{gt}}$  associated to  $\mathbf{B}_{t-1}^{\text{gt}}$ . For a detected box  $b_{t-1} \in \mathbf{B}_{t-1}$ , it is assigned to the ground-truth box  $b_{t-1}^{\text{gt}} \in \mathbf{B}_{t-1}^{\text{gt}}$  having the highest IoU overlap with it. Let  $b_t^{\text{gt}}$  be the ground-truth bounding box on the next frame that corresponds to the same object as  $b_{t-1}^{\text{gt}}$ . The tracking loss on  $b_{t-1}$  is then defined as

$$L_{\text{track\_box}}(b_{t-1}, b_{t-1}^{\text{gt}}, b_t^{\text{gt}}) = \begin{cases} (1 - \cos(\mathcal{E}(b_{t-1}), \mathcal{E}(b_t^{\text{gt}})))^2 & \text{if } \text{IoU}(b_{t-1}, b_{t-1}^{\text{gt}}) \geq 0.5 \\ \max(0, \cos(\mathcal{E}(b_{t-1}), \mathcal{E}(b_t^{\text{gt}})))^2 & \text{otherwise} \end{cases} \quad (10)$$

which encourages the cosine similarity  $\cos(\mathcal{E}(b_{t-1}), \mathcal{E}(b_t^{\text{gt}}))$  to be close to 1 if  $b_{t-1}$  captures the same object as  $b_t^{\text{gt}}$  ( $\text{IoU}(b_{t-1}, b_{t-1}^{\text{gt}}) \geq 0.5$ ), and to be no more than 0 otherwise. The overall tracking loss is the summation of the loss values on all the detected boxes.

## 2.5. Discussion

**Accuracy and robustness** In the proposed approach, detection is enhanced by accounting for temporal information when determining the classification probabilities of the bounding boxes. In previous techniques, these probabilities  $P_{\text{det}}(c|b_t^i)$  are obtained from the per-image object detector based solely on the appearance of frame  $\mathbf{I}_t$ . Object appearance variations and visual degradations in  $\mathbf{I}_t$  can lead to significant distortions in the predicted probabilities of its detection boxes  $\mathbf{B}_t$ . To counteract these complications, our method takes advantage of the tracklets  $\mathbf{D}_{t-1}$  from the previous frame, which model the visual appearance  $\mathcal{E}(d_j^{t-1})$  and classification probabilities  $P_{\text{tr}}(c|d_j^{t-1})$  of each object. As these tracklet attributes are computed over the full existing trajectory of an object, they provide a representation that is relatively robust to the appearance changes that may occur during object motion, while placing greater weight on more recent frames. The classification probabilities of a bounding box are influenced by the tracklets most similar to it, as determined from association weights. By taking advantage of tracklet information in this way, the classification scores of bounding boxes are more robustly obtained, leading to more accurate final boxes.

By comparison, late integration techniques typically incorporate temporal information by adding bounding boxes propagated from preceding frames by optical flow. These boxes are aggregated with the detector’s bounding boxes just prior to NMS. As illustrated in Fig. 1, this is less ideal because the resulting boxes either have distorted classification probabilities (boxes from the detector) or rely on optical flow (boxes from tracklets) which can be inaccurate especially over the course of multiple frames or when there is background movement. Furthermore, since propagated boxes inherit the high classification scores of their corresponding tracklets, they may suppress more accurate boxes from the detector in the NMS. The difference in performance is examined in Sec. 4.2.

**Stability** Another key advantage of the proposed approach is the improvement of box localization stability across frames, as illustrated in Figure 1. Unstable localization is a commonly observed problem in video object detection and tracking, and such instability can be attributed to appearance change in different frames. For example, in Figure 1, suppose  $b_i^{t-1}$  and  $b_k^{t-1}$  are two candidate boxes properly

covering the object ‘squirrel’ on frame  $\mathbf{I}_{t-1}$ , but with slight shifts respectively. For both boxes, the corresponding per-frame recognition scores  $P_{\text{det}}(c|b_i^{t-1})$  and  $P_{\text{det}}(c|b_k^{t-1})$  are high. But after NMS, only one of  $b_i^{t-1}$  and  $b_k^{t-1}$  would be kept. Suppose  $P_{\text{det}}(c|b_k^{t-1})$  is slightly higher, so box  $b_k^{t-1}$  is kept and associated with the existing tracklet to form  $d_j^{t-1}$ . On frame  $\mathbf{I}_t$ , suppose  $b_{i'}^t$  and  $b_{k'}^t$  are the highest overlapped candidate boxes with  $b_i^{t-1}$  and  $b_k^{t-1}$ , respectively. Due to slight appearance changes on frame  $\mathbf{I}_t$ ,  $P_{\text{det}}(c|b_{i'}^t)$  is slightly higher than  $P_{\text{det}}(c|b_{k'}^t)$ , and thus  $b_{i'}^t$  is kept after NMS. As a result, there is jitter between box  $b_k^{t-1}$  and box  $b_{i'}^t$ , because of the sudden shift in box position relative to the actual object. If the jitter is large enough, the embedding features  $\mathcal{E}(b_{i'}^t)$  can be quite different from those of the tracklet  $\mathcal{E}(d_j^{t-1})$  that bounding box  $b_k^{t-1}$  is associated with. Thus a mismatch occurs even though frames  $\mathbf{I}_{t-1}$  and  $\mathbf{I}_t$  are of high visual quality. Although box  $b_{k'}^t$  could have been well associated with tracklet  $d_j^{t-1}$  to generate a stable trajectory, it was already suppressed by NMS, thus becoming unavailable for consideration in tracklet association.

Tracklet-conditioned detection can effectively remedy this issue. The candidate boxes on a new frame would be scored not only based on the per-frame appearance, but also based on their association weights with the existing tracklets. In the example of Figure 1, when scoring candidate boxes  $b_{i'}^t$  and  $b_{k'}^t$ , the association weight  $w(b_{k'}^t, d_j^{t-1})$  is high, and tracklet  $d_j^{t-1}$  would cast a large vote on box  $b_{k'}^t$ . Thus the tracklet-conditioned score  $P(c|b_{k'}^t, \{d_j^{t-1}\})$  would be higher than those of the other boxes, and box  $b_{k'}^t$  would be kept after NMS, generating a stable trajectory.

## 2.6. Implementation Details

In the procedure *AssociateTracklet*, we employ a modified version of maximum bipartite graph matching, an algorithm widely used in multi-object tracking systems [1, 41, 60]. Given tracklets  $\mathbf{D}_{t-1}$  and bounding boxes  $\mathbf{B}_t$ , a bipartite graph is generated in which nodes corresponding to  $d_j^{t-1} \in \mathbf{D}_{t-1}$  and  $b_i^t \in \mathbf{B}_t$  are on the two sides of the graph, respectively. An edge is added between  $d_j^{t-1}$  and  $b_i^t$  if there is overlap between the last box in  $d_j^{t-1}$  and  $b_i^t$ , with their connection weight set to  $\cos(\mathcal{E}(b_i^t), \mathcal{E}(d_j^{t-1}))$ . There are no edges between non-overlapping tracklets and bounding boxes, so they will not be associated. To account for newly detected boxes which are not associated with any existing tracklet, a pseudo tracklet  $d_i^{\text{pseudo}}$  is initialized for each such bounding box  $b_i^t$ . The nodes of the pseudo tracklets are added on the side of the existing tracklets in the bipartite graph. An edge is added between pseudo tracklet  $d_i^{\text{pseudo}}$  and its corresponding bounding box  $b_i^t$ , with the connection weight set to 0. If the cosine similarity values between  $b_i^t$  and existing tracklets are all low (less than 0),  $b_i^t$  is likely to be associated with  $d_i^{\text{pseudo}}$  and a new tracklet is

formed. Finally, the standard Hungarian maximum matching algorithm [28] is applied on the constructed bipartite graph to associate the bounding boxes to the tracklets.

For the procedure *PropagateBox*, we follow the implementation in [25], but replace the OpenCV flow estimator [6, 38] with the more recent FlowNet\_v2 [23] for more accurate correspondence estimation between frames.

## 3. Related Work

**Object Detection in Images** Current leading object detectors are built on deep Convolutional Neural Networks (CNNs). They can be mainly divided into two families, namely, region-based two-stage detectors (e.g., R-CNN [19], Fast(er) R-CNN [18, 46], and R-FCN [11]) and one-stage detectors that directly predict boxes (e.g., YOLO [45], SSD [37], and CornerNets [29]).

We build our approach on Faster R-CNN with ResNet-101 and OHEM, which is a state-of-the-art object detector.

**Multiple Object Tracking (MOT)** Research on MOT primarily follow the “sequential detection and tracking” paradigm, often under the setting where detection results are given by an external object detector and the focus is on correctly associating the detection boxes across frames. Various approaches to the association problem have been proposed, including but not limited to min-cost flow [67, 33], energy models [43], Markov decision processes [58], node labeling [34], graph matching [1], and Graph Cut [63, 52]. In addition, recent works [48, 51, 42, 53, 68, 26, 10, 49] explore utilizing deep networks to better solve the association problem. In [32, 61, 57, 3], the authors seek to refine the detection and tracking results by various optimization formulations. Improved accuracies are reported, but the refinement is performed at a late stage on the results produced by off-the-shelf object detectors and trackers.

In contrast to the previous research on multi-object tracking, we advocate a new paradigm of “integrated object detection and tracking”, which aims to improve detection by considering tracking information and in turn further enhance tracking performance. The integration is at an early stage within the object detector. In this paper, tracking is performed simply by maximum bipartite graph matching, and we note that advances in MOT can benefit our method.

**Single Object Tracking** In this classic vision problem, a single object is annotated in the first frame of an input video and the tracking algorithm aims to follow the specified object throughout subsequent frames. The major challenge lies in distinguishing the object from background clutter and occluding objects. To address these issues, recent works [65, 35, 54, 39, 44, 5, 21, 13] leverage the strong representation power of deep networks, via either Siamese networks [5, 35, 21] or correlation filters [54, 14, 13] based on network features. Such trackers are usually employed

in MOT to provide the raw association weights of possible tracklet-box pairs. In this paper, we utilize a simple Siamese-network-based tracker to obtain the association weights, while also noting the benefits our method can reap from improvements in single object tracking.

**Video Object Detection** Research on video object detection gained renewed interest with the introduction of the ImageNet VID benchmark [47], which evaluates detection performance on individual frames. Numerous algorithms [25, 64, 31, 71, 17] and systems [59, 56, 15] have been developed on it, with the main focus of improving per-frame object detection results by exploiting temporal information. In large, these works can be classified into box-level methods and feature-level techniques.

Box-level methods [25, 64, 17] primarily follow the “sequential detection and tracking” approach. Bounding boxes are detected based on features from individual frames, and then are associated and rescored across frames. Prior to [17], box-level techniques associate boxes across frames by employing external tracking modules. In [17], for the first time, object detection and tracking modules share backbone features and are trained end-to-end. The network architecture design in our work follows [17] in sharing features. However, our inference procedure diverges from [17], whose tracking module associates the detected boxes on individual frames at a late stage, like other “sequential detection and tracking” techniques.

Feature-level techniques [72, 71, 69] enhance the quality of per-frame features by integrating temporal information, via flow-guided feature propagation from previous frames. This early exploitation of temporal cues leads to improved detection accuracy. We found that this technique can work in tandem with ours, by utilizing it to obtain the per-image detection scores in our method. Our experiments demonstrate the complementarity of these two approaches.

## 4. Experiments

### 4.1. Experimental Settings

We evaluate our method on two popular datasets. The first is ImageNet VID [47], a large-scale video set where the object instances are fully annotated. Following the protocol in [31, 25], we train our model on the union of the ImageNet VID and ImageNet DET training sets, and test our method on the ImageNet VID validation set. Evaluation is based on the ImageNet VID competition metrics. For detection, it is the mean average precision ( $mAP^{det}$ ) score under a box-level IoU threshold of 0.5. For tracking, it is the  $mAP^{track}$  score, under a box-level IoU threshold of 0.5 and temporal-level thresholds of [0.25, 0.5, 0.75]. All the ablations in this paper are performed on ImageNet VID.

The other dataset is 2D MOT 2015 [30], consisting of 11 training videos and 11 test videos with fully annotated

| method                |                       | $mAP^{det}$ | $mAP^{track}$ | $mAP^{track}_{slow}$ | $mAP^{track}_{med}$ | $mAP^{track}_{fast}$ |
|-----------------------|-----------------------|-------------|---------------|----------------------|---------------------|----------------------|
| sequential baseline   |                       | 74.6        | 65.2          | 79.3                 | 54.8                | 33.5                 |
| with late integration | + <i>PropagateBox</i> | 75.2        | 65.9          | 80.3                 | 53.3                | 38.9                 |
|                       | ++ <i>RescoreBox</i>  | 75.8        | 65.9          | 80.3                 | 53.3                | 38.9                 |
| integrated            | first stage only      | 75.4        | 67.1          | 79.9                 | 55.2                | 37.9                 |
|                       | second stage only     | 76.8        | 66.8          | 80.4                 | 58.0                | 36.9                 |
|                       | both stages           | <b>78.1</b> | <b>67.9</b>   | <b>80.9</b>          | <b>58.1</b>         | <b>41.9</b>          |

Table 1. Ablation of key components of our integrated detection and tracking, and of sequential detection and tracking with late integration, on the ImageNet VID validation set.

object instances. On this benchmark, submission entries are split into public / private tracks, depending on whether they use the provided set of detection boxes or their own detector. As our work proposes a new detector, we compare it to entries in the private track. Due to the limited training samples, a common practice is to finetune the model on MOT train after training on large-scale external datasets [48, 51, 2]. We note that since our approach integrates detection and tracking, we cannot train our detector on datasets consisting of cropped image patches (for person re-ID) as done in some sequential detection and tracking methods [48, 51, 2]. So instead, we train our network on the COCO [36] training and validation sets for object detection only, and finetune the whole network on the MOT training set for integrated detection and tracking. The standard evaluation metric of this dataset is MOTA, which combines false positives (FP) and false negatives (FN) for object detection, together with ID switch (IDSw) for tracking.

The hyper-parameters in training and inference on both datasets are presented in the Appendix.

### 4.2. Ablation Study

To examine the impact of the key components in our integrated object detection and tracking, we perform ablations in an online setting. The results are shown in Table 1. The baseline method excludes tracklet-conditioned detection in Algorithm 2, which is equivalent to a basic version of sequential detection and tracking where box propagation and rescoring are removed in Algorithm 1. This baseline obtains  $mAP^{det}$  and  $mAP^{track}$  scores of 74.6% and 65.2%, respectively. Applying tracklet-conditioned object detection on either the first or second stages of the object detector leads to improvements in  $mAP^{det}$  and  $mAP^{track}$  of 0.8% and 1.9%, and 2.2% and 1.6%, respectively. With tracklet-conditioned detection on both stages,  $mAP^{det}$  and  $mAP^{track}$  become 78.1% and 67.9%, respectively, which are improvements of 3.5% and 2.7% over the baseline.

In the sequential counterpart with late integration, applying box propagation improves the  $mAP^{det}$  and  $mAP^{track}$  scores to 75.2% and 65.9%, respectively. Additionally applying online box rescoring as a postprocess improves the  $mAP^{det}$  score to 75.8%. The  $mAP^{track}$  score remains the



| design               | parameters (default marked by *) |      |      |         |       |      |        |      |      |          |      |      |      |      |      | box embedding features |        |          |
|----------------------|----------------------------------|------|------|---------|-------|------|--------|------|------|----------|------|------|------|------|------|------------------------|--------|----------|
|                      | $\alpha$                         |      |      | $\beta$ |       |      | $\eta$ |      |      | $\gamma$ |      |      | $R$  |      |      | fully                  | shared | separate |
|                      | 0.5                              | 1.0* | 2.0  | 0.95    | 0.99* | 1.0  | 0.7    | 0.8* | 0.9  | 4        | 8*   | 16   | 0.2  | 0.3* | 0.4  | shared                 | 2fc    | 2fc      |
| mAP <sup>det</sup>   | 78.4                             | 78.1 | 76.7 | 78.1    | 78.1  | 78.1 | 78.0   | 78.1 | 78.1 | 76.9     | 78.1 | 78.3 | 78.1 | 78.1 | 78.1 | 77.2                   | 77.4   | 78.1     |
| mAP <sup>track</sup> | 66.7                             | 67.9 | 67.9 | 67.5    | 67.9  | 67.4 | 66.8   | 67.9 | 67.4 | 67.5     | 67.9 | 67.8 | 67.4 | 67.9 | 67.7 | 66.5                   | 67.2   | 67.9     |

Table 2. Ablation study of hyper-parameter settings and choices for producing the box embedding features in the proposed approach.

same, because tracklets are not changed by box rescoring. To sum up, our full version of integrated detection and tracking outperforms that of sequential detection and tracking with late integration by 2.3% and 2.0% in mAP<sup>det</sup> and mAP<sup>track</sup>, respectively.

For a more detailed look at our algorithm’s performance, following [71], we break down the results into different motion speeds, based on whether the ground-truth tracklet is slow (the mean IoU overlap between boxes in consecutive frames is more than 0.8), medium ( $0.6 \leq \text{mean IoU} \leq 0.8$ ), or fast (mean IoU < 0.6). As shown in Table 1, the gain in mAP<sup>track</sup> over the sequential baseline and late integration grows larger with medium and faster object motion. For these more challenging cases, object detection benefits even more from tracklet information, which in turn leads to improved tracking performance.

We additionally ablate choices for producing the box embedding features used in determining the association weights. The results, displayed in Table 2, show that adding a separate 2fc head to produce the first stage embedding features leads to better performance. Table 2 also shows ablations over hyper-parameter values. The performance was found to be relatively stable with respect to these values, and the best combination was chosen as the default setting.

### 4.3. Tracklet Stability

We also analyze the tracklet stability of different approaches. In [66], the stability of detection boxes in videos was first studied, using proposed metrics that account for temporal stability (fragment error) and spatial stability (box center and aspect ratio errors). Here, we employ slight modifications of these metrics that instead measure tracklet stability. Details on these metrics are given in the Appendix.

Table 3 compares the difference in stability of our approach to those of sequential detection and tracking with late integration. Our approach reduces the fragment, center and aspect ratio errors by a relative 4%, 6% and 7%, respectively. Improvements in stability are found to be more obvious for objects with fast motion. These numerical results verify the discussion in Section 2.5.

### 4.4. Results on Stronger Baselines and Comparison to State-of-the-art Approaches

In Table 4, our approach is compared to sequential detection and tracking with late integration on stronger baselines. The network features are enhanced by applying combina-

| motion split | frag ( $\times 10^{-3}$ )   | center ( $\times 10^{-3}$ )   | aspect ( $\times 10^{-3}$ )  |
|--------------|-----------------------------|-------------------------------|------------------------------|
| all          | 26 $\xrightarrow{-4\%}$ 25  | 134 $\xrightarrow{-6\%}$ 126  | 236 $\xrightarrow{-7\%}$ 219 |
| slow         | 11 $\xrightarrow{+27\%}$ 14 | 88 $\xrightarrow{-1\%}$ 87    | 184 $\xrightarrow{-8\%}$ 170 |
| median       | 37 $\xrightarrow{+11\%}$ 41 | 173 $\xrightarrow{-3\%}$ 168  | 304 $\xrightarrow{-7\%}$ 282 |
| fast         | 63 $\xrightarrow{-24\%}$ 48 | 227 $\xrightarrow{-10\%}$ 205 | 336 $\xrightarrow{-6\%}$ 317 |

Table 3. Tracking stability change from “sequential detection and tracking with late integration” to “integrated detection and tracking”. ‘Frag’, ‘center’, and ‘aspect’ denote fragment, box center, and aspect ratio errors, respectively. The error numbers are shown in the format of “sequential with late integration  $\xrightarrow{\text{relative change}}$  integrated”.

| DCNv2 | FGFA | mAP <sup>det</sup>      | mAP <sup>track</sup>    |
|-------|------|-------------------------|-------------------------|
|       |      | 75.8 $\rightarrow$ 78.1 | 65.9 $\rightarrow$ 67.9 |
| ✓     |      | 79.4 $\rightarrow$ 82.0 | 68.4 $\rightarrow$ 70.8 |
| ✓     | ✓    | 81.5 $\rightarrow$ 83.5 | 70.2 $\rightarrow$ 72.6 |

Table 4. Improvement on stronger baselines with FGFA [71], and Deformable ConvNets v2 (DCNv2) [70]. The scores are reported in the format of “sequential with late integration  $\rightarrow$  integrated”.

| method              | inference | backbone    | mAP <sup>det</sup> | mAP <sup>track</sup> |
|---------------------|-----------|-------------|--------------------|----------------------|
| NUIST [24]          | off-line  | ensemble    | 81.2               | N.A.                 |
| NUS-Qihoo-UIUC [55] | off-line  | DPN-131 [8] | 83.1               | 70.3                 |
| FGFA [71]           | off-line  | ResNet-101  | 78.4               | N.A.                 |
| THP [69]            | on-line   | ResNet-101  | 78.6               | N.A.                 |
| D&T [17]            | on-line   | ResNet-101  | 78.7               | -                    |
|                     | off-line  |             | 79.8               | -                    |
| D&T (reproduced)    | off-line  | ResNet-101  | 79.0               | 60.5                 |
| Ours                | on-line   | ResNet-101  | <b>83.5</b>        | <b>72.6</b>          |

Table 5. Comparison to state-of-the-art systems on the ImageNet VID validation set. In the paper of D&T [17], the mAP<sup>track</sup> score is not reported, so we reproduced the approach and report the results.

tions of FGFA [71] and Deformable ConvNets v2 [70]. On these high baselines, the integrated detection and tracking approach still outperforms the sequential counterpart by a clear margin in both detection and tracking.

We further compare the proposed approach implemented on the highest baseline to the state-of-the-art methods at the system level. Table 5 and Table 6 present the results. We note that due to system complexity and missing implementation details, direct and fair comparison among different works is difficult. Our system of “integrated detection and tracking” achieves accuracy that is very competitive with all the other systems. And we note that the idea of early integration and tracklet-conditioned detection should be appli-



| method           | inference | pre-trained | MOTA        | FP          | FN           | IDS <sub>w</sub> |
|------------------|-----------|-------------|-------------|-------------|--------------|------------------|
| H1_SJTUZZTE [62] | off-line  | unknown     | <b>56.6</b> | 7198        | 18926        | 533              |
| RAR15 [16]       | on-line   | D+T         | 56.5        | 9386        | <b>16921</b> | 428              |
| TRID [40]        | off-line  | D+T         | 55.7        | 6273        | 20611        | <b>351</b>       |
| NOMTwSDP [9]     | off-line  | unknown     | 55.5        | 5594        | 21322        | 427              |
| AP_HWDPL [7]     | on-line   | D+T         | 53.0        | <b>5159</b> | 22984        | 708              |
| CDA_DDAL [2]     | on-line   | D+T         | 51.3        | 7110        | 22271        | 544              |
| MDP_SubCNN [48]  | on-line   | D           | 47.5        | 8632        | 22969        | 628              |
| DMT [27]         | off-line  | D           | 44.5        | 8088        | 25336        | 684              |
| Ours             | on-line   | D           | 56.1        | 5717        | 20460        | 788              |

Table 6. Comparison to state-of-the-art systems on the 2D MOT15 test set. ‘D’ and ‘T’ indicate pre-training for the object detection task and the tracking task, respectively.

cable to these other detection and tracking systems as well.

## 5. Conclusion

Both object detection and tracking are fundamental tasks in video understanding that are closely coupled by nature. However, in the previous approaches, the object detection and tracking modules are applied in a sequential manner, and are optionally integrated at a late stage. In this paper, we propose the first approach to tightly integrate the tasks by conditioning object detection on the current frame by tracklets from the previous frames. The object detection results are not only more accurate, but also more coherent with the existing tracklets, which further improves tracking results. Extensive experiments on the ImageNet VID and the 2D MOT 2015 benchmarks demonstrate the effectiveness of the proposed approach. The idea of early integration and tracklet-conditioned detection can also be applied to other video understanding tasks which involve both recognition and temporal association, such as jointly estimating and tracking human pose.

## A1. Experimental Setting Details

**ImageNet VID dataset [47]** This dataset is a commonly used large-scale benchmark for video object detection and tracking. The training, validation, and test sets contain 3862, 555, and 937 video snippets, respectively. The frame rate is 25 or 30 fps for most snippets. All the object instances are fully annotated with bounding boxes and instance IDs, providing a good benchmark for joint object detection and tracking. There are 30 object categories, which are a subset of the categories in the ImageNet DET dataset.

Following the protocol in [31, 72, 71], in all our experiments, the models are trained on the union of the ImageNet VID training set and the ImageNet DET training set (only the same 30 category labels are used), and are evaluated on the ImageNet VID validation set. In both training and inference, the input images are resized to a shorter side of 600 pixels. In RPN, the anchors are of 3 aspect ratios {1:2, 1:1, 2:1} and 4 scales {64<sup>2</sup>, 128<sup>2</sup>, 256<sup>2</sup>, 512<sup>2</sup>}. 300 region

proposals are generated for each frame at an NMS threshold of 0.7 IoU. SGD training is performed, with one image at each mini-batch. 120k iterations are performed on 4 GPUs, with each GPU holding one mini-batch. The learning rates are 10<sup>-3</sup> and 10<sup>-4</sup> in the first 80k and last 40k iterations, respectively. In each mini-batch, images are sampled from ImageNet DET and ImageNet VID at a 1:1 ratio. The weight decay and the momentum parameters are set to 0.0001 and 0.9, respectively. In inference, detection boxes are generated at an NMS threshold of 0.3 IoU.

**2D MOT 2015 [30]** This dataset is a widely used benchmark for multiple object tracking. It contains a total of 22 videos collected under varying scenes, devices and angles. Only the pedestrians are annotated. These videos are divided into 11 training videos and 11 test videos. The training videos have 5500 frames, 500 tracklets, and 39905 boxes. The test videos have 5783 frames, 721 tracklets, and 61440 boxes. The average number of boxes for each frame is 7.3 and 10.6 in the training and test set, respectively. The frame rates of this dataset varies greatly, ranging from 2.5 fps to 30 fps. This dataset is very challenging for pedestrian detection and tracking, due to occlusions, high annotation density, high diversity of scenarios, etc.

In both training and inference, the input images are resized to a shorter side of 800 pixels. Anchors of 3 aspect ratios {1:2, 1:1, 2:1} and 5 scales {32<sup>2</sup>, 64<sup>2</sup>, 128<sup>2</sup>, 256<sup>2</sup>, 512<sup>2</sup>} are utilized in RPN. 512 and 2000 region proposals are generated on each frame during training and inference at an NMS threshold of 0.7, respectively. In SGD training on COCO for object detection, 120k iterations are performed on 8 GPUs with 2 images per GPU. The learning rate is initialized to 0.02 and is divided by 10 at the 75k and 100k iterations. In finetuning on 2D MOT 2015 for integrated detection and tracking, 110k iterations are performed on 4 GPUs, with each GPU holding one image. The learning rates are 10<sup>-3</sup> and 10<sup>-4</sup> in the first 70k and last 40k iterations, respectively. The weight decay and the momentum parameters are set to 0.0001 and 0.9, respectively. In inference, detection boxes are generated at an NMS threshold of 0.5 IoU. We also utilize common practices developed in previous works [60, 58] to better fit the MOTA metric: (1) To reduce FP error, detection boxes with confidence score less than 0.95 are removed, and tracklets with length less than 10 frames are removed; (2) To reduce IDS<sub>w</sub> error, in online processing, previous tracklets not associated with boxes for 10 consecutive frames are not allowed to be associated with any new boxes in the upcoming frames (but the tracklets are kept in the final results).

## A2. Tracklet Stability Metric

In [66], the authors first examined the problem of detection and tracking stability. Three metrics are proposed

for evaluating stability, namely, fragment error, center position error, and scale and aspect ratio error. The metrics are applied on the per-frame detection boxes, produced by video object detection algorithms. For stability evaluation, the detection boxes are assigned to pseudo tracklets, aided by the oracle ground-truth annotations. For each ground-truth tracklet, a pseudo tracklet is formed approximately by picking the detection box with the highest overlap with respect to the corresponding ground-truth at each frame<sup>2</sup>. The stability errors are averaged over all the pseudo tracklets. It is not specified in [66] how to extend their approach to tracklets produced by detection and tracking algorithms.

Here, we extend [66] for evaluating the stability of detection and tracking algorithms in a straightforward way. Similar to the approach in [66], we seek to find a “best-match” tracklet for each ground-truth tracklet. All the recognized tracklets are first classified into positive and negative tracklets, according to the box IoU and temporal IoU thresholds in the  $mAP^{\text{track}}$  metric. A positive tracklet is assigned to the ground-truth tracklet with the highest temporal IoU. For each ground-truth tracklet, the tracklet with the highest classification score among all its assigned tracklets is picked as its “best-match”. The resulting stability errors are the averaged errors over all the “best-match” tracklets (generated at various box and temporal IoU thresholds as done for  $mAP^{\text{track}}$ ).

## References

- [1] S.-H. Bae and K.-J. Yoon. Robust online multi-object tracking based on tracklet confidence and online discriminative appearance learning. In *CVPR*, 2014. 6
- [2] S.-H. Bae and K.-J. Yoon. Confidence-based data association and discriminative deep appearance learning for robust online multi-object tracking. *TPAMI*, 2018. 7, 9
- [3] A. Barbu, A. Michaux, S. Narayanaswamy, and J. M. Siskind. Simultaneous object detection, tracking, and event recognition. *Advances in Cognitive Systems*, 2012. 6
- [4] J. Berclaz, F. Fleuret, E. Turetken, and P. Fua. Multiple object tracking using k-shortest paths optimization. *TPAMI*, 2011. 1
- [5] L. Bertinetto, J. Valmadre, J. F. Henriques, A. Vedaldi, and P. H. Torr. Fully-convolutional siamese networks for object tracking. In *ECCV*, 2016. 6
- [6] G. Bradski and A. Kaehler. *Opencv. Dr. Dobbs journal of software tools*, 2000. 6
- [7] L. Chen, H. Ai, C. Shang, Z. Zhuang, and B. Bai. Online multi-object tracking with convolutional neural networks. In *ICIP*, 2017. 9
- [8] Y. Chen, J. Li, H. Xiao, X. Jin, S. Yan, and J. Feng. Dual path networks. In *NIPS*, 2017. 8
- [9] W. Choi. Near-online multi-target tracking with aggregated local flow descriptor. In *ICCV*, 2015. 9
- [10] Q. Chu, W. Ouyang, H. Li, X. Wang, B. Liu, and N. Yu. Online multi-object tracking using cnn-based single object tracker with spatial-temporal attention mechanism. In *ICCV*, 2017. 1, 2, 6
- [11] J. Dai, Y. Li, K. He, and J. Sun. R-fcn: Object detection via region-based fully convolutional networks. In *NIPS*, 2016. 5, 6
- [12] J. Dai, H. Qi, Y. Xiong, Y. Li, G. Zhang, H. Hu, and Y. Wei. Deformable convolutional networks. In *ICCV*, 2017. 4
- [13] M. Danelljan, G. Bhat, F. S. Khan, M. Felsberg, et al. Eco: Efficient convolution operators for tracking. In *CVPR*, 2017. 6
- [14] M. Danelljan, A. Robinson, F. Shahbaz Khan, and M. Felsberg. Beyond correlation filters: Learning continuous convolution operators for visual tracking. In *ECCV*, 2016. 6
- [15] J. Deng, Y. Zhou, B. Yu, Z. Chen, S. Zafeiriou, and D. Tao. Speed/accuracy tradeoffs for object detection from video. [http://image-net.org/challenges/talks\\_2017/Imagenet2017VID.pdf](http://image-net.org/challenges/talks_2017/Imagenet2017VID.pdf), 2017. 7
- [16] K. Fang, Y. Xiang, X. Li, and S. Savarese. Recurrent autoregressive networks for online multi-object tracking. In *WACV*, 2018. 9
- [17] C. Feichtenhofer, A. Pinz, and A. Zisserman. Detect to track and track to detect. In *CVPR*, 2017. 1, 2, 7, 8
- [18] R. Girshick. Fast r-cnn. In *ICCV*, 2015. 4, 6
- [19] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *CVPR*, 2014. 6
- [20] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 4
- [21] D. Held, S. Thrun, and S. Savarese. Learning to track at 100 fps with deep regression networks. In *ECCV*, 2016. 6
- [22] C. Huang, B. Wu, and R. Nevatia. Robust object tracking by hierarchical association of detection responses. In *ECCV*, 2008. 1
- [23] E. Ilg, N. Mayer, T. Saikia, M. Keuper, A. Dosovitskiy, and T. Brox. FlowNet 2.0: Evolution of optical flow estimation with deep networks. In *CVPR*, 2017. 6
- [24] Z. Y. R. F. Q. M. Q. L. J. D. Jing Yang, Hui Shuai. Efficient object detection from videos. <http://image-net.org/challenges/talks/2016/Imagenet%202016%20VID.pptx>, 2017. 8
- [25] K. Kang, W. Ouyang, H. Li, and X. Wang. Object detection from video tubelets with convolutional neural networks. In *CVPR*, 2016. 1, 2, 6, 7
- [26] C. Kim, F. Li, and J. M. Rehg. Multi-object tracking with neural gating using bilinear lstm. In *ECCV*, 2018. 6
- [27] H.-U. Kim and C.-S. Kim. Cdt: Cooperative detection and tracking for tracing multiple objects in video sequences. In *ECCV*, 2016. 9
- [28] H. W. Kuhn. The hungarian method for the assignment problem. *Naval research logistics quarterly*, 1955. 6
- [29] H. Law and J. Deng. Cornernet: Detecting objects as paired keypoints. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018. 6
- [30] L. Leal-Taixé, A. Milan, I. Reid, S. Roth, and K. Schindler. Motchallenge 2015: Towards a benchmark for multi-target tracking. *arXiv preprint arXiv:1504.01942*, 2015. 2, 7, 9

<sup>2</sup>The implementation in [66] performs maximum bipartite graph matching with the box IoUs as the weights of the bipartite graph.

- [31] B. Lee, E. Erdenee, S. Jin, M. Y. Nam, Y. G. Jung, and P. K. Rhee. Multi-class multi-object tracking using changing point detection. In *ECCV*, 2016. 7, 9
- [32] B. Leibe, K. Schindler, and L. Van Gool. Coupled detection and trajectory estimation for multi-object tracking. In *ICCV*, 2007. 6
- [33] P. Lenz, A. Geiger, and R. Urtasun. Followme: Efficient on-line min-cost flow tracking with bounded memory and computation. In *ICCV*, 2015. 6
- [34] E. Levinkov, J. Uhrig, S. Tang, M. Omran, E. Insafutdinov, A. Kirillov, C. Rother, T. Brox, B. Schiele, and B. Andres. Joint graph decomposition & node labeling: Problem, algorithms, applications. In *CVPR*, 2017. 6
- [35] B. Li, J. Yan, W. Wu, Z. Zhu, and X. Hu. High performance visual tracking with siamese region proposal network. In *CVPR*, 2018. 4, 6
- [36] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft coco: Common objects in context. In *ECCV*, 2014. 7
- [37] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg. Ssd: Single shot multibox detector. In *ECCV*, 2016. 6
- [38] B. D. Lucas, T. Kanade, et al. An iterative image registration technique with an application to stereo vision. 1981. 6
- [39] C. Ma, J.-B. Huang, X. Yang, and M.-H. Yang. Hierarchical convolutional features for visual tracking. In *ICCV*, 2015. 6
- [40] S. Manen, M. Gygli, D. Dai, and L. Van Gool. Pathtrack: Fast trajectory annotation with path supervision. In *ICCV*, 2017. 9
- [41] A. Milan, L. Leal-Taixé, K. Schindler, and I. Reid. Joint tracking and segmentation of multiple targets. In *CVPR*, 2015. 6
- [42] A. Milan, S. H. Rezatofighi, A. R. Dick, I. D. Reid, and K. Schindler. Online multi-target tracking using recurrent neural networks. In *AAAI*, 2017. 6
- [43] A. Milan, S. Roth, and K. Schindler. Continuous energy minimization for multitarget tracking. *TPAMI*, 2014. 6
- [44] Y. Qi, S. Zhang, L. Qin, H. Yao, Q. Huang, J. Lim, and M.-H. Yang. Hedged deep tracking. In *CVPR*, 2016. 6
- [45] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi. You only look once: Unified, real-time object detection. In *CVPR*, 2016. 6
- [46] S. Ren, K. He, R. Girshick, and J. Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *NIPS*, 2015. 4, 5, 6
- [47] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *IJCV*, 2015. 2, 7, 9
- [48] A. Sadeghian, A. Alahi, and S. Savarese. Tracking the untrackable: Learning to track multiple cues with long-term dependencies. In *ICCV*, 2017. 6, 7, 9
- [49] S. Schuster, P. Vernaza, W. Choi, and M. Chandraker. Deep network flow for multi-object tracking. In *CVPR*, 2017. 6
- [50] A. Shrivastava, A. Gupta, and R. Girshick. Training region-based object detectors with online hard example mining. In *CVPR*, 2016. 4
- [51] J. Son, M. Baek, M. Cho, and B. Han. Multi-object tracking with quadruplet convolutional neural networks. In *CVPR*, 2017. 6, 7
- [52] S. Tang, B. Andres, M. Andriluka, and B. Schiele. Subgraph decomposition for multi-target tracking. In *CVPR*, 2015. 6
- [53] S. Tang, M. Andriluka, B. Andres, and B. Schiele. Multiple people tracking by lifted multicut and person reidentification. In *CVPR*, 2017. 6
- [54] J. Valmadre, L. Bertinetto, J. Henriques, A. Vedaldi, and P. H. Torr. End-to-end representation learning for correlation filter based tracking. In *CVPR*, 2017. 6
- [55] S. D. M. L. X. J. W. Lin, J. Peng and H. Xiong. Improving context modeling for video object detection and tracking. [http://image-net.org/challenges/talks\\_2017/ilsvrc2017\\_short\(posterior\).pdf](http://image-net.org/challenges/talks_2017/ilsvrc2017_short(posterior).pdf), 2017. 8
- [56] Y. Wei, M. Zhang, J. Li, Y. Chen, J. Feng, H. Shi, J. Dong, and S. Yan. Improving context modeling for video object detection and tracking. [http://http://image-net.org/challenges/talks\\_2017/ilsvrc2017\\_short\(posterior\).pdf](http://http://image-net.org/challenges/talks_2017/ilsvrc2017_short(posterior).pdf), 2017. 7
- [57] Z. Wu, A. Thangali, S. Sclaroff, and M. Betke. Coupling detection and data association for multiple object tracking. In *CVPR*, 2012. 6
- [58] Y. Xiang, A. Alahi, and S. Savarese. Learning to track: Online multi-object tracking by decision making. In *ICCV*, 2015. 1, 2, 6, 9
- [59] J. Yang, H. Shuai, Z. Yu, R. Fan, Q. Ma, q. Liu, and J. Deng. Ilsvrc2016 object detection from video: Team nuist. [http://image-net.org/challenges/talks/2016/Imagenet\\_202016%20VID.pptx](http://image-net.org/challenges/talks/2016/Imagenet_202016%20VID.pptx), 2016. 7
- [60] F. Yu, W. Li, Q. Li, Y. Liu, X. Shi, and J. Yan. Poi: Multiple object tracking with high performance detection and appearance feature. In *ECCV*, 2016. 6, 9
- [61] Q. Yu and G. Medioni. Integrated detection and tracking for multiple moving objects using data-driven mcmc data association. In *WMVC*, 2008. 6
- [62] J. L. Y. C. J. F. H. S. J. D. Yunchao Wei, Mengdan Zhang and S. Yan. 2d mot 2015 leader board. [https://motchallenge.net/tracker/H1\\_SJTUZTE](https://motchallenge.net/tracker/H1_SJTUZTE), 2017. 9
- [63] A. R. Zamir, A. Dehghan, and M. Shah. Gmcp-tracker: Global multi-object tracking using generalized minimum clique graphs. In *ECCV*. 2012. 6
- [64] X. Zeng, W. Ouyang, J. Yan, H. Li, T. Xiao, K. Wang, Y. Liu, Y. Zhou, B. Yang, Z. Wang, et al. Crafting gbd-net for object detection. *TPAMI*, 2018. 7
- [65] D. Zhang, H. Maei, X. Wang, and Y.-F. Wang. Deep reinforcement learning for visual object tracking in videos. *arXiv preprint arXiv:1701.08936*, 2017. 6
- [66] H. Zhang and N. Wang. On the stability of video detection and tracking. *arXiv preprint arXiv:1611.06467*, 2016. 8, 10
- [67] L. Zhang, Y. Li, and R. Nevatia. Global data association for multi-object tracking using network flows. In *CVPR*, 2008. 1, 6

- [68] J. Zhu, H. Yang, N. Liu, M. Kim, W. Zhang, and M.-H. Yang. Online multi-object tracking with dual matching attention networks. In *ECCV*, 2018. 6
- [69] X. Zhu, J. Dai, L. Yuan, and Y. Wei. Towards high performance video object detection. In *CVPR*, 2018. 7, 8
- [70] X. Zhu, H. Hu, S. Lin, and J. Dai. Deformable convnets v2: More deformable, better results. *Arxiv Tech Report*, 2018. 8
- [71] X. Zhu, Y. Wang, J. Dai, L. Yuan, and Y. Wei. Flow-guided feature aggregation for video object detection. In *ICCV*, 2017. 1, 7, 8, 9
- [72] X. Zhu, Y. Xiong, J. Dai, L. Yuan, and Y. Wei. Deep feature flow for video recognition. In *CVPR*, 2017. 7, 9