Boosting End-to-end Multi-Object Tracking and Person Search via Knowledge Distillation

Wei Zhang,[1,3,*] Lingxiao He,[2] Peng Cheng,[2] Xingyu Liao,[2] Wu Liu,[2] Qi Li[1] and Zhenan Sun[1,★]
[1]CRIPAC & NLPR, CASIA; [2]JD AI Research, Beijing, China
[3]School of Artificial Intelligence, University of Chinese Academy of Sciences
wei.zhang@cripac.ia.ac.cn,{helingxiao3,liaoxingyu5,liuwu1}@jd.com,peng_c@bupt.edu.cn,{qli,znsun}@nlpr.ia.ac.cn

ABSTRACT

Multi-Object Tracking (MOT) and Person Search both demand to localize and identify specific targets from raw image frames. Existing methods can be classified into two categories, namely two-step strategy and end-to-end strategy. Two-step approaches have high accuracy but suffer from costly computations, while end-to-end methods show greater efficiency with limited performance. In this paper, we dissect the gap between two-step and end-to-end strategy and propose a simple yet effective end-to-end framework with knowledge distillation. Our proposed framework is simple in concept and easy to benefit from external datasets. Experimental results demonstrate that our model performs competitively with other sophisticated two-step and end-to-end methods in multi-object tracking and person search.

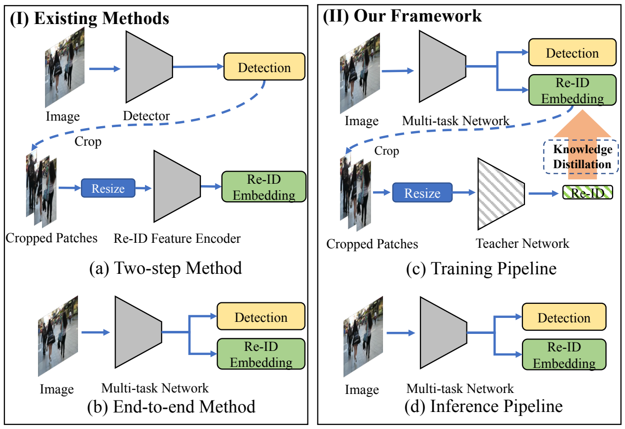
CCS CONCEPTS

• **Computing methodologies → Computer vision tasks**; *Computer vision representations.*

KEYWORDS

Multi-object tracking, Person search, End-to-end strategy, Knowledge distillation

ACM Reference Format:
Wei Zhang,[1,3,*] Lingxiao He,[2] Peng Cheng,[2] Xingyu Liao,[2] Wu Liu,[2] Qi Li[1] and Zhenan Sun[1,★]. 2021. Boosting End-to-end Multi-Object Tracking and Person Search via Knowledge Distillation. In *Proceedings of the 29th ACM International Conference on Multimedia (MM '21), October 20–24, 2021, Virtual Event, China.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3474085.3481546



(a) Two-step Method

(b) End-to-end Method

(c) Training Pipeline

(d) Inference Pipeline

Figure 1: Comparison of existing methods and proposed framework. (I) Existing methods can be summarized into two types. a) Two-step approach: it uses off-the-shelf detectors to generate candidate object regions and then uses a Re-ID network for identity feature extraction. b) End-to-end approach: it predicts detection and identity embedding by a multi-task network. (II) Our framework combines the structure of two-step approach and end-to-end approach, applying an individual teacher network to generate soften Re-ID labels in a) training stage while maintaining the simple structure in b) inference stage.

1 INTRODUCTION

Multi-object tracking (MOT) [50] and person search [53] have received considerable attention in the computer vision community. The former targets to estimate the trajectories for objects of interest in videos, while the latter aims to find a target person in a gallery of scene images. In order to track multiple objects in long time range, many MOT approaches follow the tracking-by-detection framework, i.e., tracking is applied as a follow-up association approach with given detection results as well as identity features. Person search demands to localize query person in source images, hence involves both person detection and re-identification (Re-ID). The communal objective of tracking-by-detection MOT and person search is to localize and identify a specific target person from raw image frames.

Existing methods can be classified into two categories, namely two-step strategy [50, 55] and end-to-end strategy [28, 47, 53]. Two-step approaches tackle detection and Re-ID tasks [61] independently via separate networks. Firstly, the detection model crops candidates from raw images, then the identity model extracts Re-ID features for each candidate. Although two-step approaches show high accuracy both in MOT [50, 55] and person search [6, 22, 63], they still have limited applications in industrial systems, mainly because of costly computations in both detector and identity model

[*] This work was done when Wei Zhang was an intern at JD AI Research.
[★] Corresponding author.

MM '21, October 20–24, 2021, Virtual Event, China
© 2021 Association for Computing Machinery.
ACM ISBN 978-1-4503-8651-7/21/10...$15.00
https://doi.org/10.1145/3474085.3481546

in inference. In order for real-time inference, end-to-end methods [28, 47, 53, 59] deal with detection and Re-ID feature extraction by a single backbone-shared multi-task network, which shows greater efficiency. However, their performances are limited compared with two-step methods. We argue that the deterioration mainly comes from three aspects:

1) Conflicted objectives. Detection head aims to find a common feature space for all pedestrians while Re-ID head struggles to distinguish them. The conflicts between two heads hurt network training.

2) Scale variance. Objects' size varies in a large range across frames in MOT and person search. Two-step methods crop and resize person candidate regions generated by detection model, then feed these fixed sized image patches into identity model, as illustrated in Fig. 1. While in end-to-end approaches, the "crop and resize" process is omitted. This gap leads to the failure of optimal Re-ID feature extraction.

3) Marginalized Re-ID task. Most end-to-end methods are extended on top of anchor-based detectors, e.g., Faster RCNN [37], which firstly generate region proposals and then extract corresponding Re-ID features. Thus the quality of proposals influences Re-ID embedding extraction heavily. In consequence, the model is biased to estimate good proposals while marginalizes Re-ID feature extraction.

Since end-to-end strategy adopt a backbone-shared multi-task network for both detection and Re-ID feature extraction, it needs data with both location and identity annotations in training stage. Compared with large detection and Re-ID datasets, public datasets with simultaneous location and identity annotations are relatively small, which limits training powerful models for industrial usages. Some end-to-end methods [47, 59] propose to use extrinsic datasets with location labels alone for only detection heads updating. This biased training strategy may lead to biased results. In order to alleviate the afore-mentioned marginalization problem, Zhang *et al.* [59] proposes to fairly balance detection and Re-ID tasks by treating detection and Re-ID feature extraction in a parallel style. Although the parallel structure in [59] alleviates the marginalization problem, it still has to do training on large-scaled pedestrian detection datasets with no identity annotations for better detection performance. This imbalanced training process introduces unfairness for detection and Re-ID task as well.

In this paper, we propose a simple yet effective knowledge distillation framework for boosting end-to-end multi-object tracking and person search. The improvements mainly come from two aspects. Firstly, knowledge distillation provides stable and informative supervision. In two-step strategy, identity model takes image patches cropped by ground truth bounding boxes as input. This independent training process guarantees the stability of Re-ID embeddings. Inspired by knowledge distillation [17] (KD), we employ a powerful pretrained teacher model to distill the capacity of general identification to Re-ID head. Since teacher model is separately trained with ground truth image patches, it provides reliable supervision unaffected by objective confliction and scale variance. Secondly, inspired by [59], we implement detection and Re-ID in a homogeneous way. Following anchor-free detector CenterNet [65], we utilize keypoint estimation to predict the bounding box centers. Sizes and center offsets estimations are represented as pixel-wise maps regression. In addition, a Re-ID head is added to represent identity features for each pixel, in which corresponding feature embeddings in object centers indicate targets' identity embeddings. This anchor-free structure performs detection and Re-ID in a parallel style, mitigates the marginalization of Re-ID feature extraction. In inference stage, only the end-to-end model is reserved, which makes our method simple and light-weighted.

Moreover, our framework is general to abundant detection and Re-ID datasets. The identity supervisions are completely from pretrained teacher model, which means our approach requires no ID annotations or sequential connection in training domain. In other words, our framework can be trained or fine-tuned on static autonomous images. Generally, the major contributions of our work are summarized as follows:

- We introduce a novel end-to-end knowledge distillation framework that joints object detection and re-identification tasks into an multi-task framework.
- Our framework not only mitigates the conflicts between detection and Re-ID task, but also keeps a real-time speed in inference stage. It achieves a well balance between high performance and efficiency. Experimental results have shown the efficiency and efficacy of our framework.
- Our framework requires no ID annotations or sequential connections in training domain and can be trained on static autonomous images.

2 RELATED WORK

Multiple Objects Tracking. Recent MOT approaches mainly follow the tracking-by-detection paradigm [5, 18, 41], which separate objects tracking into two sequential pipelines: objects detection and bounding box association. Traditional works have been committed to the improvement of association by proposing new models [18], new cost metrics [5] or new optimization strategies [41] based on available detection results and appearance embeddings. With the development of deep learning [9, 23, 26, 27, 45], an increasing number of works [2, 3, 28, 44, 47, 49, 55] put their efforts on exploring better features for simple matching algorithms (e.g. Hungarian matching). According to the structures, they can be classified into two types: two-step and end-to-end MOT approaches.

Two-step MOT methods [50, 55] treat detection and Re-ID feature extraction as two separate tasks. Without loss of generality, [55] employs off-the-shelf detector Faster-RCNN [37] for bounding boxes generation. After that, all candidate objects are cropped from the input images according to the detection results and then resized to a fixed size. Then the resized image patches are fed into a standard Re-ID network which are trained separately on person Re-ID datasets [12, 24, 62, 63] with softmax and triplet loss. This "detection then Re-ID" procedure enables both detector and Re-ID encoder to optimize independently with different training setting, hence ensuring good performance. However, two-step approaches usually suffer from heavy computations and long inference time because of large parameters and spitted structure, which limits their applications.

End-to-end methods aim at combining both detection and identity features extraction into a single multi-task network. Most existing end-to-end trackers follow the same insight which is adding identify embedding heads on top of existing detection networks [28, 44, 47, 59]. For example, [44] adds a identify embedding head on top of Mask-RCNN [15] to regress Re-ID feature for each proposal. [47] follows the same paradigm while replaces the detector to YOLOv3 [36]. [28] extends RetinaNet [25] detector and predicts Re-ID features for each anchor by a lightweight head. [59] replaces previous anchor-based detector into anchor-free detector CenterNet [65]. In this paper, we also follow the end-to-end paradigm. Instead of supervising Re-ID head with hard ID labels like previous works, we propose to draw support from strong Re-ID models for stable and informative identity supervision. We show that our model achieves improved performance even without ID annotations in training domain.

For further simplification in box association, some methods attempt to predicting the spatial offsets of targets directly, which is called tracking-by-regression paradigm. For instance, Bergmann *et al.* [2] applies Faster R-CNN for detection and regresses offsets in adjacent frames for each objects, while Zhou *et al.* [64] proposes add offsets head on CenterNet [65] to predict objects' motion. Although these methods are simple and fast, their accuracy is far from satisfactory especially in situations where large motions or heavy occlusions exist.

Person Search. Person Search is a practically relevant task which can be regarded as a multi-task problem that combines pedestrian detection and person Re-ID. Similar to MOT, they can also be classified into two-step and end-to-end approaches. The pioneering work of Xu *et al.* [54] provides a two-step person search approach that first detects pedestrians from a scene image and then inputs these detected pedestrians into the Re-ID network. However, it is hard to achieve satisfactory speed as previously mentioned. Different from approaches in MOT, end-to-end person search approaches pay more attention to promoting feature discrimination capability. Xiao *et al.* [53] are the first to propose an end-to-end person search model by employing Faster R-CNN. The Re-ID network is directly connected to Faster R-CNN with base layers shared. Remarkably, an online instance matching (OIM) loss is introduced to address the joint training of detector and Re-ID network. Bharti *et al.* [33] introduce a novel query-guided end-to-end person network with extensive use of the full query image instead of the cropped image. Han *et al.* [14] propose a localization refinement framework for person search, which uses the Re-ID loss instead of regression loss in Faster R-CNN to supervise the model training to obtain more reliable bounding boxes. [7] proposes to disentangle the person embedding into norm and angle, for detection and Re-ID respectively, which achieves comparable performances to two-step methods. Currently, advanced approaches [6, 7] inevitably use the inefficient Faster R-CNN framework, which severely restricts the use of person search in large-scale video monitoring scenarios. Instead, our proposed method applies the anchor-free structure which simply predicts a single center per pedestrian without the requirement of post-processing. The simplicity of structure keeps the inference procedure fast.

3 APPROACH

In this section, we elaborate on the proposed end-to-end knowledge distillation framework. Firstly, we introduce the backbone network. And then, we respectively introduce the object detection and identity embedding branches. Finally, we explain the strategy for training and inference.

3.1 Backbone Network

We employ the architecture of ResNet34 based feature extractor named DLA34 as backbone [65]. Iterative deep aggregation (IDA) is a structure for Deep Layer Aggregation (DLA) [56], focusing on iteratively merging the feature hierarchy to improve the recognition, resolution and speed of the network. We modify the original IDA as CenterNet [65] does. Specifically, first, the fully convolutional up-sampling version of IDA is leveraged for dense prediction, and the original convolution is replaced by 3×3 deformable convolution at every up-sampling layer. In this manner, the reception field can be dynamically adapted, which helps treat objects with different scales. Secondly, a $3 \times 3, 256$ channels convolution layer, followed by a 1×1 convolution are added before each output head. Let $I \in \mathbb{R}^{w \times h \times 3}$ be an input image of width w and height h. And then, the backbone outputs feature maps $F \in \mathbb{R}^{\frac{w}{r} \times \frac{h}{r} \times d}$ of width $\frac{w}{r}$, height $\frac{h}{r}$ and channel d, where r is the output stride. We use the predefined output stride of $r = 4$ in literature [59].

3.2 Object Detector

3.2.1 *Detect Objects as Points.*

Following [65], we formulate object detection as keypoints estimation. We present objects by multiple single points at their bounding box center instead of region proposals. Object size (height and width) is then regressed directly from feature maps at each center location. With a single input image $I \in \mathbb{R}^{h \times w \times 3}$, the network produces a down-sampled heatmap $H \in \mathbb{R}^{\frac{w}{r} \times \frac{h}{r} \times 1}$ and a size map $S \in \mathbb{R}^{\frac{w}{r} \times \frac{h}{r} \times 2}$. The responsive peaks in the heatmap H represent the center of the objects, and corresponding position of each center in size map S indicates object's width and height. In addition, since final feature maps are down-sampled by r, there are quantuzation errors between coordination in heatmap H and image I. To this end, we apply the offset head to estimate a continuous offset for eliminating the influence of output stride. As a result, the network produces a additional offset map $O \in \mathbb{R}^{\frac{w}{r} \times \frac{h}{r} \times 2}$. Let $(x_1^c, y_1^c, x_2^c, y_2^c)$ be the bounding box of c-th objects in the image. The ground truth c-th objects center point p^c, size s^c, and offset o^c are respectively formulated as:

$$\begin{aligned} p^c &= (p_x^c, p_y^c) = (\frac{\widetilde{x}_1^c + \widetilde{x}_2^c}{2}, \frac{\widetilde{y}_1^c + \widetilde{y}_2^c}{2}) \\ s^c &= (h^c, w^c) = (x_2^c - x_1^c, y_2^c - y_1^c) \\ o^c &= (\frac{x_1^c + x_2^c}{2r}, \frac{y_1^c + y_2^c}{2r}) - p^c \end{aligned} \tag{1}$$

where $(\widetilde{x}_1^c, \widetilde{x}_2^c, \widetilde{y}_1^c, \widetilde{y}_2^c) = (\lfloor \frac{x_1^c}{r} \rfloor, \lfloor \frac{x_2^c}{r} \rfloor, \lfloor \frac{y_1^c}{r} \rfloor, \lfloor \frac{y_2^c}{r} \rfloor)$. After that, we generate a heatmap $H \in [0,1]^{\frac{w}{r} \times \frac{h}{r}}$ using a Gaussian kernel $H_{xy}^c = \exp(-\frac{(x-p_x^c)^2 + (y-p_y^c)^2}{2\sigma_p^2})$, where σ_p is a standard deviation that adapts the bounding box size.

3.2.2 *Losses for Detection Heads.*

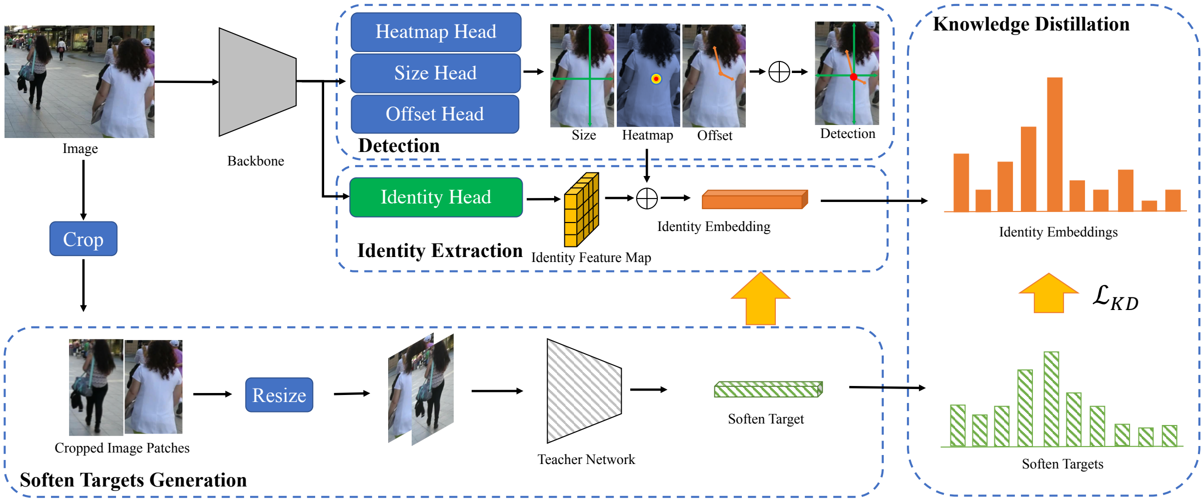


Figure 2: Overview of our framework. In training stage, a fixed Re-ID model performs as teacher, providing embedding supervision for identity branch. In inference stage, the predicted identity embedding along with detection results are extracted in parallel for further temporal bounding box linking.

In order to predict accurate object's center locations, we want the predicted center points to be infinitely close to the ground truth points. So the training objective is a pixel-wise logistic regression with focal loss [25]:

$$\mathcal{L}_c = \frac{-1}{C}\sum_{xy}\begin{cases}(1-\hat{H}_{xy})^{\alpha}\log(\hat{H}_{xy}), & \text{if } H_{xy}=1\\ (1-H_{xy})^{\beta}(\hat{H}_{xy})^{\alpha}\log(1-\hat{H}_{xy}), & \\ \text{otherwise.} & \end{cases} \tag{2}$$

where α and β are hyper-parameters of the focal loss, and C is the number of objects in the image I. A prediction $\hat{H}_{xy}=1$ represents a detected object center, and $\hat{H}_{xy}=0$ represents background.

We regress the bounding box width and height for each object. The optimization objective is trained with an L1 loss:

$$\mathcal{L}_s = \frac{1}{C}\sum_{c=1}^{C}|\hat{s}^c - s^c|. \tag{3}$$

where s^c is the ground truth size of the c-th object and $\hat{s}^c$ is the predicted c-th object size. The operation is processed on the raw pixel coordinates.

Similar, the training objective for offset head is formulated as:

$$\mathcal{L}_o = \frac{1}{C}\sum_{c=1}^{C}|\hat{o}^c - o^c|. \tag{4}$$

where o_c is the ground truth center offset of the c-th object and $\hat{o}_c$ is the predicted center point offset.

The total detection loss can be written as a weighted linear sum of them:

$$\mathcal{L}_{det} = w_c\mathcal{L}_c + w_s\mathcal{L}_s + w_o\mathcal{L}_o. \tag{5}$$

w_c, w_s and w_o are loss weights.

3.3 Identity Embedding Extensions

3.3.1 *Re-ID Branch.*

In order to maintain real-time inference rate, we follow the end-to-end strategies by sharing most of the parameters with detector. We adopt the same output head structure as detection heads for simplify, which consists a 3 × 3 conv and a 1 × 1 conv specifically. Given input image I, the identity embedding branch outputs embedding feature maps $E \in \mathbb{R}^{\frac{w}{s}\times\frac{h}{s}\times k}$ with channel k. $\hat{E}_{(x^c,y^c)}$ represents corresponding estimated identity feature vector at location (x^c, y^c). We use $k = 2048$ in all experiments.

3.3.2 *Soften Targets Generation.*

The key idea behind knowledge distillation is that the soften probabilities outputted by a pre-trained teacher network, denoted as soften labels, containing a lot more information than hard label (e.g. one-hot encoding for class label) [17]. In knowledge distillation setting, the pretrained Re-ID model can be considered as a teacher, while the identity embedding branch acts as a student. Previous works [47, 59] treat Re-ID branch training as classification task by utilizing cross-entropy loss for ID supervision. Although straightforward, the inherent differences between detection and Re-ID are overlooked. Detection tends to treat different objects in same category equally with a bunch of regression-based losses for locating, which naturally conflicts with ID-based Re-ID task. In other words, objective of ID classification has contradiction with location regression, especially in anchor-free structures where detection and Re-ID task are of comparable importance. In this work we address this problem by replacing ID classification into feature regression. We adopt identity features that generated by powerful Re-ID model as the soften labels. Compared with hard ID labels, this formulation not only relieves the competition between detection and Re-ID but also takes the advantage of well-designed teacher's representation capacity.

For general identity feature guidance, the teacher network must be general and effective. We formulate our teacher baseline based on a simple and effective Re-ID model [29]. Moreover, we apply the teacher training procedure on a bunch of datasets with FastReID [16] framework in order to further promote the discrimination and generalization of feature embedding. For each ground truth object bounding box $b^c = (x_1^c, y_1^c, x_2^c, y_2^c)$, we crop the image patch on input image $I \in \mathbb{R}^{W \times H \times 3}$ and then resize it to fixed size 256×128 before feeding it into teacher model. The output feature vector E_{x^c, y^c} represents the identity embedding for $c-$th object in location (x^c, y^c).

3.3.3 *Distilling Knowledge from Softening Targets.*

In common knowledge distillation setting, soften target supervision is often combined with hard labels supervision for better results:

$$\mathcal{L}_{KD} = \eta\mathcal{L}_{soft} + (1-\eta)\mathcal{L}_{hard} \tag{6}$$

$$\mathcal{L}_{soft} = \frac{1}{C}\sum_{c=1}^{C} \begin{cases} 0.5(\hat{E_c} - E_c)^2, & |\hat{E_c} - E_c| < 1 \\ |\hat{E_c} - E_c| - 0.5, & \text{otherwise.} \end{cases} \tag{7}$$

$$\mathcal{L}_{hard} = \frac{1}{C}\sum_{c=1}^{C}\sum_{i=1}^{M} l_i^c log(p_i) \tag{8}$$

where $\mathcal{L}_{soft}$ and $\mathcal{L}_{hard}$ are loss of soften labels (teacher's outputs) supervision and hard labels (ground truth labels) supervision, η is the hyper-parameter for loss balance. $\hat{E_c}$ is the predicted c-th identity embedding, E_c is the c-th identity embedding produced by frozen Re-ID teacher. p_i denotes the mapped probability for c-th item belongs to the i-th category in the class distribution vector $P = \{p_i, i \in [1, M]\}$. l_i^c is the c-th one-hot representation of the ground truth class label.

However, in experiments we found that soften labels alone has already sufficient for confident identity information. Besides, comparing with slight performance gain, getting rid of the demand for identity annotations is more appealing. We set $\alpha = 1$ in our framework. In a word, our identity supervision completely comes from pre-trained teacher, no identity annotations involved in training.

We apply a smooth L1 loss for knowledge transformation for soften labels. Specifically, the identity embedding can be trained with the following objective function:

$$\mathcal{L}_{id} = \frac{1}{C}\sum_{c=1}^{C} \begin{cases} 0.5(\hat{E_c} - E_c)^2, & |\hat{E_c} - E_c| < 1 \\ |\hat{E_c} - E_c| - 0.5, & \text{otherwise.} \end{cases} \tag{9}$$

where $\hat{E_c}$ is the predicted c-th identity embedding, E_c is the c-th identity embedding produced by frozen Re-ID teacher.

3.4 Training and Inference

3.4.1 *Train.*

We train the whole networks in an end-to-end way. During the whole training period, weights in teacher network are kept frozen. For better balancing detection and Re-ID task, we adopt the uncertainty loss balance strategy proposed in [20] by using task-independent uncertainty for automatic weights learning. The joint objective can be written as

$$\mathcal{L}_{total} = \frac{1}{2}(\frac{1}{e^{\omega_1}}\mathcal{L}_{det} + \frac{1}{e^{\omega_2}}\mathcal{L}_{id} + \omega_1 + \omega_2). \tag{10}$$

where ω_1 and ω_2 are task-dependent uncertainties for detection loss and identity loss and are learnable during training.

3.4.2 *Inference.*

In inference period, our model proceed detection and identity embedding learning simultaneously. On top of the predicted heatmap, we perform non-maximum suppression based on the heatmap scores to extract the peak keypoints. Then, we compute the corresponding bounding boxes based on the estimated offsets and box sizes, the identity features are extracted from the identity heatmaps on corresponding location of centers in the same manner.

3.4.3 *Postponed Operations.*

Multi-Object Tracking. We follow the classic online tracking algorithm for box association [49]. In the first frame, we perform initialization based on detected boxes. For following frames, linked boxes are produced according to cosine distances of corresponding Re-ID features and their boxes' overlaps by bipartite matching. Kalman Filter is also applied to predict tracklets in current frame in order to filter out large motion linking. The apperance features updates in each step. We refer readers to [49] for more detailed discussion.

Person Search. Given a query person image, it simply passes the teacher network to obtain query feature. For gallery, our network takes scene images as inputs and generates detected locations as well as Re-ID embeddings simultaneously by a simple forward propagation. Finally, we use the query person to retrieve the similar persons from gallery images by cosine distances in feature embedding space.

4 EXPERIMENTS

4.1 Implementation Details

Our implementation is built with PyTorch. We train our models on Tesla P40 GPUs, and conduct inference on GeForce RTX 2080Ti. As for the optimization, we train our model with the Adam optimizer for 30 epochs with the starting learning rate of 1e-4. The learning rate drops at the 20th epoch by a factor of 10. During training, all training samples are all re-scaled to size 1088×608, and the output feature map resolution is 272×152. The teacher network takes resized image patch with resolution of 256×128 as input. One batch with 16 scene images are used to train the overall network.

4.2 Multi-Object Tracking

4.2.1 *Datasets and Metrics.*

We evaluate our end-to-end knowledge distillation framework in multi-pedestrian tracking on two multi-object tracking benchmarks: MOT17 and MOT16 [32]. Our ablation studies are mainly conducted on MOT17. MOT17 contains 7 training sequences and 7 test sequences with location and identity annotations on pedestrians. Note that MOT17 provides no validation split officially, we use first half of frames in all training sequences as training set

Table 1: Datasets for teacher training and testing.

	Dataset	#Identities	#Cameras	#Images
Train	VIPeR [12]	632	2	1264
	SAIVT-Softbio [4]	152	8	64472
	CUHK03 [24]	1467	10	13164
	PKU-Reid [30]	114	2	1824
	Airport [11]	9651	6	39902
	ThermalWorld [21]	516	20	15118
	iLIDS-VID [46]	300	2	42495
	3DPeS [1]	192	8	1011
	CAVIAR4ReID [8]	72	2	1220
	Shinpuhkan [19]	24	16	video
	LPW [42]	2731	11	592438
	PRAI-1581 [58]	1581	mobile	39461
	SenseReID [60]	1717	/	4438
	SYSU-MM01 [51]	491	6	287628
Test	Market1501 [62]	1501	6	32217
	MSMT17 [48]	4101	15	126441
	DukeMTMC [38]	1812	8	36441

Table 2: Ablation study on feature effectiveness on MOT17 validation set. ↑ means the larger the better and ↓ means the opposite.

	MOTA↑	IDF1↑	IDs↓	FPS↑
Box	71.0	71.6	801	21.7
Emb	70.0	71.8	449	22.1
Box+Emb	72.3	74.8	332	21.4
Two-step	72.5	74.9	318	7.1

Table 3: Ablation study on the effectiveness of different knowledge distillation schemes. ↑ means the larger the better and ↓ means the opposite.

	MOTA↑	IDF1↑	IDs↓
Mix-Teacher	72.3	74.8	332
MOT17-Teacher	72.2	73.4	394
MOT17-Label	68.8	70.2	443
Mix-Teacher-Hard	71.8	75.8	359
Mot17-Teacher-Hard	71.6	75.4	350

for network training and second half of frames as validation set in our ablation experiments following [64]. We adopt the CrowdHuman dataset [40] for pre-training in all ablation experiments unless otherwise specified. For fair comparison, we follow common practice of previous works [47] that initialize weights with model pretrained on external data when comparing with other State-of-the art MOT methods in Table 5. Following [47], we adopt multiple datasets including CrowdHuman [40], ETH [10] and CityPerson [57] which contains detection annotations only, as well as CalTech [13], CUHK-SYSU [53] and PRW [63] which provides both detection and identity annotations. More discussion of pre-training strategies could be found in Section 4.2.2.

To obtain feature embedding with sufficient discrimination and generalization, we apply a bunch of Re-ID datasets as listed in Table 1. for teacher network training unless otherwise specified. We use Market-1501 [62], MSMT17 [48] and DukeMTMC [38] for evaluation. We use CLEAR metrics and ID metrics [39] for evaluation. We mainly adopt comprehensive index $MOTA$ and $IDF1$ for comparison. Comparing to $MOTA$, $IDF1$ measures the consistency of ID matching better. More discussions are displayed in Section 4.2.2. We also report some specific indexes including $MOTP$, FP, FN, $Recall$, $Precision$, IDs, MT and ML to reveals the capability of object coverage and identity.

4.2.2 Ablation Study.

Effectiveness of Identity Embedding. Recall that we use second half of frames in MOT17 as validation set in our ablation study. Firstly, we ablate the effectiveness of identity embedding in Table 2. *Box* utilizes only detection results of our framework for association. It performs boxes matching with distances calculated by detected boxes’ IOU. *Emb* uses only feature embedding results of our framework for association. We use cosine distances of predicted feature embeddings of detected objects as similarity scores for matching. *Box+Emb* considers both boxes IOU distances and embedding cosine distances during linking. *Two-step* replaces embeddings with features outputs of teacher model. In this setting, our network severs as a detector and provides bounding boxes results to downstream Re-ID networks for feature extraction, which is a typical two-step tracking paradigm. Kalman Filter is applied to all four boxes as linking strategy. Results are presented in Table 2. From second line, we can see that with only feature embeddings, our network achieves reasonable scores for all indexes especially in IDs. On all scores, *Emb* results are comparable with *Box* results. It means that our learned feature is of good discrimination. By combining detection and embedding, *Box+Embed* achieves significantly improvement in all scores compared with *Box*, which proves that the identity head successfully learns the discriminative ability. Crowd scenes exists plenty of targets overlapping, which causes serious ID switches with only boxes matching are performed. Our learned identity feature provides extra cues to this situation. Moreover, *Box+Emb* achieves approximate results compared with the combination of teacher embedding and detection in *Two-step*, with only sightly decrease by 0.2 in MOTA, while kept much higher FPS. This comparison shows that the knowledge-distillation framework retains most identity information provided by heavy-weighted teacher model with a light head.

Effectiveness of Different Knowledge Distillation Schemes. Next, we validate the effectiveness of different knowledge distillation schemes in Table 3. *Mix-Teacher* is the standard pipeline of our framework in which teacher model is trained on a bunch of Re-ID datasets listed in Table 1. Identity supervision is completely from soften labels. *Mot17-Teacher* also uses the soften labels as supervision, but with soften labels generated from a weak Re-ID teacher model which trained only on image patches cropped from MOT17. *Mot17-Label* treats Re-ID branch as a classification task by utilizing cross entropy loss to supervised identity head’s outputs with one-hot representation of the ground truth class labels. From the comparison of *Mot17-Teacher*, *Mix-Teacher* and *Mot17-Label*, we can observe that separate teacher model provides reliable identity information. Even if *Mix-Teacher*’s teacher model has no access to

Table 4: Ablation study on knowledge distillation effectiveness on MOT17 validation set. ↑ means the larger the better and ↓ means the opposite.

	MOTA↑	IDF1↑	IDs↓	MT↑	ML↓	FP↓	FN↓	Recall↑	Precision↑
CH	64.9	69.7	318	140	65	2364	16308	69.8	94.1
CH w/o Emb	65.4	62.4	475	143	65	2526	15693	71.0	93.8
MOT17	67.0	68.5	605	139	55	3006	14215	73.7	93.0
CH+MOT17	72.3	74.8	332	173	44	2760	11891	78.0	93.9

Table 5: Comparison to published state-of-the-art trackers on MOT17 test and MOT16 test under private detector protocol. ↑ means the larger the better and ↓ means the opposite. End-to-end methods are labeled by *. The best results are in bold.

Dataset	Method	MOTA↑	IDF1↑	MT↑	ML↓	IDs↓	FPS↑
MOT16	TubeTK (CVPR 2020)* [34]	64.0	59.4	33.5%	19.4%	1117	1.0
	JDE (ECCV 2020)* [47]	61.4	62.2	32.8%	18.2%	**781**	16.7
	TAP (ICPR 2018) [66]	64.8	73.5	38.5%	21.6%	571	7.1
	CNNMTT (MM 2019) [31]	65.2	62.2	32.4%	21.3%	946	5.5
	POI (ECCV 2016) [55]	66.1	65.1	34.0%	20.8%	805	5.3
	CTrackerV1 (ECCV 2020)* [35]	67.6	57.2	32.9%	23.1%	1897	6.4
	FairMOT (Arxiv 2020)* [59]	**74.9**	72.8	**44.7%**	**15.9%**	1074	**21.7**
	KDMOT (Ours)*	74.3	**74.7**	40.4%	17.6%	797	21.4
MOT17	SST (TPAMI 2019) [43]	52.4	49.5	21.4%	30.7%	8431	<3.9
	TubeTK (CVPR 2020)* [34]	63.0	58.6	31.2%	19.9%	4137	2.6
	CTrackerV1 (ECCV 2020)* [35]	66.6	57.4	32.2%	24.2%	5529	6.4
	CenterTrack (ECCV 2020)* [64]	67.8	64.7	34.6%	24.6%	3039	**24.0**
	FairMOT (Arxiv 2020)* [59]	**73.7**	72.3	**43.2%**	17.3%	3303	21.7
	KDMOT (Ours)*	73.4	**73.8**	41.4%	**16.7%**	**2673**	21.4

identity annotations in training domain, it still outperforms *Mot17-Teacher*. Moreover, *Mot17-Teacher* outperforms *Mot17-Label* by a large margin, indicating that our framework with feature embedding distillation is much more effective than ID-based classification. We believe that it is because the way of using teacher outputs as supervision not only well captures reliable identity information, but also mitigates the marginalization of Re-ID feature extraction.

As we stated that soften labels alone is already sufficient for confident identity information. Here we provide more experimental analysis in Table 3. *Mix-Teacher* and *Mot17-Teacher* denotes training only with identity features produced by teacher network pre-trained on the mix datasets collection (Table 1.) and on MOT17 [32]. *Mix-Teacher-Hard* and *Mot17-Teacher-Hard* supervises identity branch with the combination of identity features (soften labels) and ID labels (hard labels) as noted in eq 6. with $\eta = 0.5$. From the results, we found that soften labels alone is already sufficient for confident identity information. Additional hard label promotes *IDF1* but get worse *MOTA*. We believe this deterioration mainly comes from the conflicts between regression based soften loss as well as detection losses and classification based hard label loss.

Comparison of Different Pre-training Models. Finally, we investigate effectiveness of pretraining and flexibility of our framework. The results are shown in Table 4. *CH* trains model for 60 epoches only on CrowdHuman with detection branches (supervised by detection labels) as well as Identity branch (supervised by soften labels). *CH w/o Embed* also uses CrowdHuman dataset alone but sets the weight of identity branch to 0. *MOT17* conduct training on MOT17 training split without external data. *CH+MOT17* follows the common-used pretraining practice with model pretrained in *CH* as initial weights and finetunes on MOT17 for 30 epoches.

Without CrowHuman pretraining, the performance drops by about 8% in $MOTA$ on validation set. Here we analyze the decrease in detail. Revisit the formulation of $MOTA = 1 - \frac{\sum_t (FN_t + FP_t + IDs_t)}{\sum_t GT_t}$, where GT_t denotes the number of ground truth objects in frame t. In the comparison between *MOT17* with *CH+MOT17*, we can see that there are obvious drops in FN. It indicates that pre-training on CrowdHuman helps a lot to detection. In the same manner, *CH+MOT17* outperforms *CH* by a large margin with the help of fine-tuning. It is worth noting that *CH* results are comparable with *MOT17* results in all scores, with slightly lower $MOTA$ and better $IDF1$. It indicates that our model is of great generality to unseen domains. Without embedding training, $MOTA$ in *CH w/o Emb* further improves a little by 0.5. This gain mainly comes from benefits of undisturbed detection training with FN decreases from 16308 to 15693. Although it promotes detection, the absence of identity branch training hurts to identity capability in turn which can be discerned from $IDF1$ and IDs. This result shows that our framework manages to achieve better balance between detection and identity instead of biases to estimate accurate detection.

4.2.3 Compare with State-of-the-art MOT Methods.

Results on MOT Challenges. We report our result on MOT16 and MOT17 in private detector protocol. Comparisons are made between our methods and published state-of-the-art approaches of both two-step methods and end-to-end methods in Table 5. Recall that we initialize weights with model pre-trained on external data

Table 6: Comparison of cross-domain performance with other methods on MOT17 validation. All the methods are trained with only static images on CrowdHuman. ↑ means the larger the better and ↓ means the opposite. The best results are in bold.

MOT17 validation set			
Method	MOTA↑	IDF1↑	IDs↓
CenterTrack(CH static) [64]	52.2	53.8	-
KDMOT (CH static)	**65.4**	**62.4**	**475**
CenterTrack(CH + MOT17) [64]	66.1	64.2	-
KDMOT (CH + MOT17)	**72.3**	**74.8**	**332**

for fair comparison. We can see that our approach outperforms most methods on both MOT16 and MOT17. Considering that our approach are trained without identity labels in training datasets, the results are strong and convincing. In comparison with FairMOT [59], our approach achieves much better performance on identity-related metrics including *IDF*1 and *IDs*, while the detection-related scores are slightly lower. We argue that it is because FairMOT mainly biased to obtain accurate detection. Although the bias leads to better detection results, it also causes deterioration in identity metrics *IDF1* and *IDs*. In addition, our approach preserves a simple structure in inference stage, which enables real-time tracking. Our model achieves comparable speed with FairMOT [59], and is only slightly slower than tracking-by-regression based method CenterTrack [64], mainly because it directly employs object offsets for association. This association is fast but lack of accuracy especially in occluded situation. To summarize, our method achieves a well balance between performance and speed.

Results on Static images. Further, we investigate the generalization of our model by conducting evaluations with models trained with only static images and detection annotations. Different from previous tracking models which need to adopt extra self-supervised learning training strategy, our framework is naturally tolerant to static images. We use CrowdHuman [40] dataset for training and MOT17 validation set for test. We choose CenterTrack [64] for comparison, which originally provides static images training strategy. Results are shown in Table 6. We can see that although there are obvious drops of performances for both methods, our approach suffers less from domain gap especially in *IDs*, *MT* and *ML*. This is mainly because our model requires no id annotations in train datasets, which helps the model generalize to detection only datasets. We believe that this characteristic will be helpful when ID annotations are inaccessible in practical industrial applications.

4.3 Person Search

Datasets and Metrics. We evaluate our end-to-end knowledge distillation framework on person search task with widely used benchmark Person re-identification in the Wild (PRW) [63]. PRW is transfered from 10-hours total length videos collected in Tsinghua university. Five 1080 × 1920 HD and a 576 × 720 SD cameras are used for collection. It consists of 11,816 images and 43,110 pedestrian bounding boxes, among which 34,304 pedestrians are labeled. Note that we use no ID annotations of PRW in whole training procedure for both teacher and student. We finetune our framework on PRW for 30 epoch with only detection label adopted. For performance evaluation, we employ the standard metrics as in most person search literature,namely the cumulative matching cure (CMC top-K) and the mean Average Precision (mAP). Only when candidates’ IoU to the ground truth bounding boxes is above 0.5 will be count as correct.

Table 7: Comparison to the state-of-the-art Person Search methods on PRW.

	Methods	mAP	Top-1	Time
two-step	DPM+IDE (CVPR 2017) [63]	20.5	48.3	-
	CNN+MGTS (ECCV 2018) [6]	32.6	72.1	-
	CNN+CLSA (ECCV 2018) [22]	38.7	65.0	-
	FPN+RDLR (ICCV 2019) [14]	42.9	70.2	-
end-to-end	OIM (CVPR 2017) [53]	21.3	49.9	-
	IAN (Arxiv 2017) [52]	23.0	61.9	-
	QEEPS (CVPR 2019) [33]	37.1	76.7	-
	NAE (CVPR 2020) [7]	43.3	80.9	158
	KDMOT (Ours)	38.7	73.2	138

Compare with State-of-the-arts. Comparisons are made between our approach and state-of-the-art approaches of two categories: two-step models and end-to-end models. The results are shown in Table 7. Our method achieves 38.7 in mAP and 73.2 in top-1, which outperforms most end-to-end and two-step methods. Our performance is only lower than NAE [7], and with shorter inference time. Considering that our model trained without any ID annotations in PRW dataset, this result is strong and convincing, which demonstrate that the designed framework effectively drew accurate identity information without ID annotations.

5 CONCLUSION

In this paper, we propose an end-to-end knowledge distillation framework for jointly learning object location and identification. Well-designed knowledge distillation framework helps to reach a good balance between high performance and real-time processing in inference stage. The proposed approach is very simple, fast and accurate without demands for simultaneous location and identity annotations in training, and can be trained in an end-to-end way on external detection datasets. Experimental results on three datasets validate the effectiveness of the framework for multi-objects tracking and person search. We hope this framework could inspire more ideas in real-time and high-performance multi objects tracking, person search and other tasks.

ACKNOWLEDGEMENTS

This work was supported in part by the National Key R&D Program of China (Grant No. 2020AAA0103800) and the National Natural Science Foundation of China (Grant No. U1836217, No. 62076240 and No. 61721004).

REFERENCES

[1] Davide Baltieri, Roberto Vezzani, and Rita Cucchiara. 2011. 3dpes: 3d people dataset for surveillance and forensics. In *Proceedings of the 2011 joint ACM workshop on Human gesture and behavior understanding*. 59–64.

[2] Philipp Bergmann, Tim Meinhardt, and Laura Leal-Taixe. 2019. Tracking without bells and whistles. In *IEEE International Conference on Computer Vision*. 941–951.

[3] Alex Bewley, Zongyuan Ge, Lionel Ott, Fabio Ramos, and Ben Upcroft. 2016. Simple online and realtime tracking. In *IEEE International Conference on Image Processing*. 3464–3468.

[4] Alina Bialkowski, Simon Denman, Sridha Sridharan, Clinton Fookes, and Patrick Lucey. 2012. A database for person re-identification in multi-camera surveillance networks. In *2012 International Conference on Digital Image Computing Techniques and Applications (DICTA)*. IEEE, 1–8.

[5] Visesh Chari, Simon Lacoste-Julien, Ivan Laptev, and Josef Sivic. 2015. On pairwise costs for network flow multi-object tracking. In *IEEE Conference on Computer Vision and Pattern Recognition*. 5537–5545.

[6] Di Chen, Shanshan Zhang, Wanli Ouyang, Jian Yang, and Ying Tai. 2018. Person search via a mask-guided two-stream cnn model. In *European Conference on Computer Vision*. 734–750.

[7] Di Chen, Shanshan Zhang, Jian Yang, and Bernt Schiele. 2020. Norm-aware embedding for efficient person search. In *IEEE Conference on Computer Vision and Pattern Recognition*. 12615–12624.

[8] Dong Seon Cheng, Marco Cristani, Michele Stoppa, Loris Bazzani, and Vittorio Murino. 2011. Custom pictorial structures for re-identification.. In *Bmvc*, Vol. 1. Citeseer, 6.

[9] Qiyao Deng, Jie Cao, Yunfan Liu, Zhenhua Chai, Qi Li, and Zhenan Sun. 2020. Reference Guided Face Component Editing. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence*, Vol. 1. 502–508.

[10] Andreas Ess, Bastian Leibe, Konrad Schindler, and Luc Van Gool. 2008. A mobile vision system for robust multi-person tracking. In *IEEE Conference on Computer Vision and Pattern Recognition*. 1–8.

[11] Mengran Gou, Ziyan Wu, Angels Rates-Borras, Octavia Camps, Richard J Radke, et al. 2018. A systematic evaluation and benchmark for person re-identification: Features, metrics, and datasets. *IEEE transactions on pattern analysis and machine intelligence* 41, 3 (2018), 523–536.

[12] Douglas Gray, Shane Brennan, and Hai Tao. 2007. Evaluating appearance models for recognition, reacquisition, and tracking. In *IEEE International Workshop on Performance Evaluation for Tracking and Surveillance*, Vol. 3. 1–7.

[13] Gregory Griffin, Alex Holub, and Pietro Perona. 2007. Caltech-256 object category dataset. (2007).

[14] Chuchu Han, Jiacheng Ye, Yunshan Zhong, Xin Tan, Chi Zhang, Changxin Gao, and Nong Sang. 2019. Re-id driven localization refinement for person search. In *IEEE International Conference on Computer Vision*. 9814–9823.

[15] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. 2017. Mask r-cnn. In *IEEE International Conference on Computer Vision*. 2961–2969.

[16] Lingxiao He, Xingyu Liao, Wu Liu, Xinchen Liu, Peng Cheng, and Tao Mei. 2020. FastReID: A pytorch toolbox for general instance re-identification. *arXiv preprint arXiv:2006.02631* (2020).

[17] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2015. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531* (2015).

[18] Hao Jiang, Sidney Fels, and James J Little. 2007. A linear programming approach for multiple object tracking. In *IEEE Conference on Computer Vision and Pattern Recognition*. 1–8.

[19] Yasutomo Kawanishi, Yang Wu, Masayuki Mukunoki, and Michihiko Minoh. 2014. Shinpuhkan2014: A multi-camera pedestrian dataset for tracking people across multiple cameras. In *20th Korea-Japan joint workshop on frontiers of computer vision*, Vol. 5. Citeseer.

[20] Alex Kendall, Yarin Gal, and Roberto Cipolla. 2018. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In *IEEE Conference on Computer Vision and Pattern Recognition*. 7482–7491.

[21] Vladimir V Kniaz, Vladimir A Knyaz, Jiri Hladuvka, Walter G Kropatsch, and Vladimir Mizginov. 2018. Thermalgan: Multimodal color-to-thermal image translation for person re-identification in multispectral dataset. In *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*. 0–0.

[22] Xu Lan, Xiatian Zhu, and Shaogang Gong. 2018. Person search by multi-scale matching. In *European conference on computer vision*. 536–552.

[23] Peipei Li, Yibo Hu, Ran He, and Zhenan Sun. 2019. Global and Local Consistent Wavelet-Domain Age Synthesis. *IEEE Transactions on Information Forensics and Security* 14, 11 (2019), 2943–2957.

[24] Wei Li, Rui Zhao, Tong Xiao, and Xiaogang Wang. 2014. Deepreid: Deep filter pairing neural network for person re-identification. In *IEEE Conference on Computer Vision and Pattern Recognition*. 152–159.

[25] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. 2017. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*. 2980–2988.

[26] Wu Liu, Qian Bao, Yu Sun, and Tao Mei. 2021. Recent Advances in Monocular 2D and 3D Human Pose Estimation: A Deep Learning Perspective. *arXiv preprint arXiv:2104.11536* (2021).

[27] Yunfan Liu, Qi Li, and Zhenan Sun. 2019. Attribute-Aware Face Aging With Wavelet-Based Generative Adversarial Networks. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 11877–11886.

[28] Zhichao Lu, Vivek Rathod, Ronny Votel, and Jonathan Huang. 2020. Retinatrack: Online single stage joint detection and tracking. In *IEEE Conference on Computer Vision and Pattern Recognition*. 14668–14678.

[29] Hao Luo, Youzhi Gu, Xingyu Liao, Shenqi Lai, and Wei Jiang. 2019. Bag of tricks and a strong baseline for deep person re-identification. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops*. 0–0.

[30] Liqian Ma, Hong Liu, Liang Hu, Can Wang, and Qianru Sun. 2016. Orientation driven bag of appearances for person re-identification. *arXiv preprint arXiv:1605.02464* (2016).

[31] Nima Mahmoudi, Seyed Mohammad Ahadi, and Mohammad Rahmati. 2019. Multi-target tracking using CNN-based features: CNNMTT. *Multimedia Tools and Applications* 78, 6 (2019), 7077–7096.

[32] Anton Milan, Laura Leal-Taixé, Ian Reid, Stefan Roth, and Konrad Schindler. 2016. MOT16: A benchmark for multi-object tracking. *arXiv preprint arXiv:1603.00831* (2016).

[33] Bharti Munjal, Sikandar Amin, Federico Tombari, and Fabio Galasso. 2019. Query-guided End-to-End Person Search. (2019).

[34] Bo Pang, Yizhuo Li, Yifan Zhang, Muchen Li, and Cewu Lu. 2020. TubeTK: Adopting tubes to track multi-object in a one-step training model. In *IEEE Conference on Computer Vision and Pattern Recognition*. 6308–6318.

[35] Jinlong Peng, Changan Wang, Fangbin Wan, Yang Wu, Yabiao Wang, Ying Tai, Chengjie Wang, Jilin Li, Feiyue Huang, and Yanwei Fu. 2020. Chained-tracker: Chaining paired attentive regression results for end-to-end joint multiple-object detection and tracking. In *European Conference on Computer Vision*. 145–161.

[36] Joseph Redmon and Ali Farhadi. 2018. Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767* (2018).

[37] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. *arXiv preprint arXiv:1506.01497* (2015).

[38] Ergys Ristani, Francesco Solera, Roger Zou, Rita Cucchiara, and Carlo Tomasi. 2016. Performance Measures and a Data Set for Multi-Target, Multi-Camera Tracking. In *European Conference on Computer Vision workshop on Benchmarking Multi-Target Tracking*.

[39] Ergys Ristani, Francesco Solera, Roger Zou, Rita Cucchiara, and Carlo Tomasi. 2016. Performance measures and a data set for multi-target, multi-camera tracking. In *European Conference on Computer Vision*. 17–35.

[40] Shuai Shao, Zijian Zhao, Boxun Li, Tete Xiao, Gang Yu, Xiangyu Zhang, and Jian Sun. 2018. Crowdhuman: A benchmark for detecting human in a crowd. *arXiv preprint arXiv:1805.00123* (2018).

[41] Francesco Solera, Simone Calderara, and Rita Cucchiara. 2015. Learning to divide and conquer for online multi-target tracking. In *IEEE International Conference on Computer Vision*. 4373–4381.

[42] Guanglu Song, Biao Leng, Yu Liu, Congrui Hetang, and Shaofan Cai. 2018. Region-based quality estimation network for large-scale person re-identification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 32.

[43] ShiJie Sun, Naveed Akhtar, HuanSheng Song, Ajmal Mian, and Mubarak Shah. 2019. Deep affinity network for multiple object tracking. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 43, 1 (2019), 104–119.

[44] Paul Voigtlaender, Michael Krause, Aljosa Osep, Jonathon Luiten, Berin Balachandar Gnana Sekar, Andreas Geiger, and Bastian Leibe. 2019. Mots: Multi-object tracking and segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition*. 7942–7951.

[45] Qi Wang, Xinchen Liu, Wu Liu, An-An Liu, Wenyin Liu, and Tao Mei. 2020. MetaSearch: Incremental Product Search via Deep Meta-Learning. *IEEE Transactions on Image Processing* 29 (2020), 7549–7564.

[46] Taiqing Wang, Shaogang Gong, Xiatian Zhu, and Shengjin Wang. 2016. Person re-identification by discriminative selection in video ranking. *IEEE transactions on pattern analysis and machine intelligence* 38, 12 (2016), 2501–2514.

[47] Zhongdao Wang, Liang Zheng, and Yixuan Liu. 2020. Towards real-time multi-object tracking. In *European Conference on Computer Vision*. 107–122.

[48] Longhui Wei, Shiliang Zhang, Wen Gao, and Qi Tian. 2018. Person transfer gan to bridge domain gap for person re-identification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 79–88.

[49] Nicolai Wojke and Alex Bewley. 2018. Deep cosine metric learning for person re-identification. In *IEEE Winter Conference on Applications of Computer Vision*. 748–756.

[50] Nicolai Wojke, Alex Bewley, and Dietrich Paulus. 2017. Simple online and realtime tracking with a deep association metric. In *IEEE International Conference on Image Processing*. 3645–3649.

[51] Ancong Wu, Wei-Shi Zheng, Hong-Xing Yu, Shaogang Gong, and Jianhuang Lai. 2017. RGB-infrared cross-modality person re-identification. In *Proceedings of the IEEE international conference on computer vision*. 5380–5389.

[52] Jimin Xiao, Yanchun Xie, Tammam Tillo, Kaizhu Huang, Yunchao Wei, and Jiashi Feng. 2019. IAN: the individual aggregation network for person search. *Pattern Recognition* 87 (2019).

[53] Tong Xiao, Shuang Li, Bochao Wang, Liang Lin, and Xiaogang Wang. 2017. Joint detection and identification feature learning for person search. In *IEEE Conference on Computer Vision and Pattern Recognition*. 3415–3424.

[54] Yuanlu Xu, Bingpeng Ma, Rui Huang, and Liang Lin. 2014. Person search in a scene by jointly modeling people commonness and person uniqueness. In *ACM International Conference on Multimedia*.

[55] Fengwei Yu, Wenbo Li, Quanquan Li, Yu Liu, Xiaohua Shi, and Junjie Yan. 2016. Poi: Multiple object tracking with high performance detection and appearance feature. In *European Conference on Computer Vision*. 36–42.

[56] Fisher Yu, Dequan Wang, Evan Shelhamer, and Trevor Darrell. 2018. Deep layer aggregation. In *IEEE Conference on Computer Vision and Pattern Recognition*.

[57] Shanshan Zhang, Rodrigo Benenson, and Bernt Schiele. 2017. Citypersons: A diverse dataset for pedestrian detection. In *IEEE Conference on Computer Vision and Pattern Recognition*. 3213–3221.

[58] Shizhou Zhang, Qi Zhang, Yifei Yang, Xing Wei, Peng Wang, Bingliang Jiao, and Yanning Zhang. 2020. Person Re-identification in Aerial imagery. *IEEE Transactions on Multimedia* 23 (2020), 281–291.

[59] Yifu Zhang, Chunyu Wang, Xinggang Wang, Wenjun Zeng, and Wenyu Liu. 2020. FairMOT: On the fairness of detection and re-identification in multiple object tracking. *arXiv preprint arXiv:2004.01888* (2020).

[60] Haiyu Zhao, Maoqing Tian, Shuyang Sun, Jing Shao, Junjie Yan, Shuai Yi, Xiaogang Wang, and Xiaoou Tang. 2017. Spindle net: Person re-identification with human body region guided feature decomposition and fusion. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 1077–1085.

[61] Kecheng Zheng, Wu Liu, Lingxiao He, Tao Mei, Jiebo Luo, and Zheng-Jun Zha. 2021. Group-aware Label Transfer for Domain Adaptive Person Re-identification. *arXiv preprint arXiv:2103.12366* (2021).

[62] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. 2015. Scalable person re-identification: A benchmark. In *IEEE International Conference on Computer Vision*. 1116–1124.

[63] Liang Zheng, Hengheng Zhang, Shaoyan Sun, Manmohan Chandraker, Yi Yang, and Qi Tian. 2017. Person re-identification in the wild. In *IEEE Conference on Computer Vision and Pattern Recognition*. 1367–1376.

[64] Xingyi Zhou, Vladlen Koltun, and Philipp Krähenbühl. 2020. Tracking objects as points. In *European Conference on Computer Vision*. 474–490.

[65] Xingyi Zhou, Dequan Wang, and Philipp Krähenbühl. 2019. Objects as points. *arXiv preprint arXiv:1904.07850* (2019).

[66] Zongwei Zhou, Junliang Xing, Mengdan Zhang, and Weiming Hu. 2018. Online multi-target tracking with tensor-based high-order graph matching. In *International Conference on Pattern Recognition*. 1809–1814.

,