Id-Free Person Similarity Learning

Bing Shuai, Xinyu Li, Kaustav Kundu, Joseph Tighe
AWS AI Labs
{bshuai, xxnl, kaustavk, tighej}@amazon.com

Abstract

Learning a unified person detection and re-identification model is a key component of modern trackers. However, training such models usually relies on the availability of training images / videos that are manually labeled with both person boxes and their identities. In this work, we explore training such a model by only using person box annotations, thus removing the necessity of manually labeling a training dataset with additional person identity annotation as these are expensive to collect. To this end, we present a contrastive learning framework to learn person similarity without using manually labeled identity annotations. First, we apply image-level augmentation to images on public person detection datasets, based on which we learn a strong model for general person detection as well as for short-term person re-identification. To learn a model capable of longer-term re-identification, we leverage the natural appearance evolution of each person in videos to serve as instance-level appearance augmentation in our contrastive loss formulation. Without access to the target dataset or person identity annotation, our model achieves competitive results compared to existing fully-supervised state-of-the-art methods on both person search and person tracking tasks. Our model also shows promising results for saving the annotation cost that is needed to achieve a certain level of performance on the person search task.

1. Introduction

Detecting and re-identifying people in images and videos underpins many person understanding tasks [33, 49, 50, 55, 57, 60]. Recent works [42, 47, 50, 55, 60] have converged on the concept of detecting and re-identifying people simultaneously with a single model due to its higher end-to-end efficiency during inference. Typical learning setups for such models rely on a set of training images or videos that are annotated with both person boxes and their identities. Unfortunately, the cost of curating such a training set is extremely high not least because sourcing the right images / videos for both tasks is costly. More importantly, annotating person identity induces a significant cost overhead, as annotators have to check against a set of known persons from memory. Thus, existing training datasets for joint person detection and re-identification [33, 49, 57] are limited both in scope and in size.

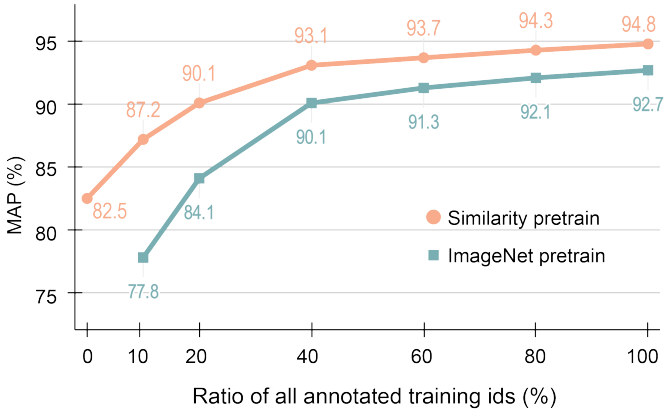


Figure 1. By using only 40% of id annotations on CUHK-SYSU dataset [49], our similarity pre-trained model outperforms the standard state-of-the-art model trained on all id annotations. Manually labeled id annotation is not used in similarity pre-training.

In this work, we explore learning such a model by using person bounding boxes as our only supervision, so that we are able to leverage the existing large-scale training sets for person detection [30, 41]. The core question hinges on how to learn a good person embedding model from images only annotated with person boxes. To this end, we adopt the concept of instance discrimination tasks [8, 19, 22, 48] that pull together identity embeddings of the same person box under different image transformations while pushing away those from other person boxes. This formulation enables us to jointly train a person detection and identity embedding model. While image-level transformations are able to capture many expected changes to a person's appearance such as scale or lighting changes, they are not able to render more drastic pose and viewpoint changes which are common in re-identification applications and as such this embedding model is not robust to such appearance variations.

To address this shortcoming, we add unlabeled videos to provide examples of natural appearance changes in pose and viewpoint. We utilize our image trained model to extract person boxes and their embeddings, based on which we produce pseudo person identity labels. While the appearance of the person at the start and end of the video can be drastically different, it changes smoothly over time. We use density-based clustering [17] to exploit such temporal continuity so the embeddings of the same person are clustered together. Thus we are able to extract disparate views of the same person, from which we fine-tune our model with our instance discrimination learning framework.

We perform ablation studies on the person search task, in particular on CUHK-SYSU [49]. By training on images annotated with person boxes alone, our model works quite well (76.5% MAP / 77.8% Top-1), even for re-identifying a person that goes through modest appearance changes. Introducing unlabeled videos during training brings a substantial performance benefit (+6.0% MAP / +6.6% Top-1), which indicates the significance of the appearance diversity to similarity learning. In addition, as shown in Fig. 1, by only using 40% of the identity annotation randomly sampled from the target CUHK-SYSU dataset [49], our model is able to achieve state-of-the-art results.

Furthermore, we apply our model to the multi-person tracking task. We adopt the same solver [2] as that in FairMOT [55] that takes as input the detected person boxes and their embedding. On MOT17 [33], without using any manual identity annotation or frames from the target dataset during training, our model achieves the same level of performance as the fully-supervised FairMOT [55].

2. Related Work

2.1. Person detection and re-identification

Person search first detects and then queries the person of interest among a set of gallery images. The same person could look substantially different between query and gallery images as they can be captured at different time / location / cameras. Therefore, the quality of person embedding is the key towards accurate person search. Two-step approaches [14, 21, 44] adopted two separate models for person detection and person re-identification respectively, but recent unified models [5, 13, 34, 49, 50, 58] are gaining popularity as they are significantly faster and achieve state-of-the-art performance. In order to train those models, manually labeled identity annotations are indispensable. In this work, we explore training such a model without relying on expensive manual identity annotations. Importantly, it achieves fairly competitive results compared to existing fully-supervised models.

Multi-person tracking detects all persons in each frame and links those detected persons across time to produce person trajectories. Recent end-to-end trackers [37, 38, 42, 47, 55, 60] jointly model person detection and identity association in a unified formulation, and they push the performance to new highs on MOTChallenge [33]. Although some methods [15, 42, 47, 55] can be technically trained on image datasets and some recent works start to exploit unlabeled videos to improve multi-object tracking [25, 29, 32], it's essential for them to be fine-tuned on annotated video sequences to achieve competitive performance. In contrast, our model achieves state-of-the-art on MOT17 [33] without fine-tuning on annotated video sequences.

2.2. Unsupervised person re-identification

Our work is also related to unsupervised person re-identification [6, 7, 18], in which contrastive formulation is usually adopted to learn person embedding. On this particular issue, our work has 3 technical differences: 1), we are optimizing a multi-task model, in which similarity learning is only one of the tasks; 2), conventional person re-id works [6, 7, 18] deal with person crops, while we work with images that have multiple persons. Thus, we develop dense contrastive formulation instead of using the basic form of contrastive formulation [8, 22]; 3), we mine person tracks from videos, from which two views of the same person are sampled during training. From the perspective of contrastive formulation, such person track mining can be understood as a novel augmentation method. Without using identity annotations, we achieve competitive or state-of-the-art results on both person search and tracking tasks.

2.3. Self-supervised learning

Self-supervised learning uses a proxy task that guides the model to learn meaningful features for downstream tasks. Those proxy tasks should not require human labels in any form, some examples are predicting the rotation of a rotated image w.r.t the original one [27], gray-scale image colorization [53], solving jigsaw puzzles [36], predicting the relative position between patches [11], etc. Recently, instance discrimination based contrastive learning [4, 8, 19, 22, 48] has achieved a level of performance that is comparable to models trained fully-supervised on the ImageNet classification task [10]. As well, there are works that exploit temporal cycle consistency to learn feature correspondence that can be used for general object tracking [16, 45, 46]. Although our model can be applied in person tracking, our work is more related to instance discrimination based contrastive learning [4, 8, 19, 22, 48]. To remedy the limitation of image-level augmentation in our contrastive loss formulation, we leverage the natural appearance evolution of each person in videos to serve as instance-level appearance augmentation, which is shown to be significant in improving person embeddings that can be used to re-identify the person when their appearance changes drastically.

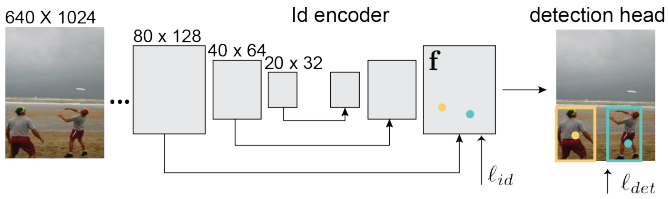


Figure 2. We use a single-stage model for joint person detection and re-identification. It is a fully convolutional network that includes an identity feature encoder and a detection head. The skip layer encodes multi-scale features into the point-based embedding.

3. Joint Person Detection and Re-identification

In this work, we use a unified person detection and re-identification model that follows a design similar to AlignPS [50], where the identity feature map is extracted from an intermediate layer of a single-stage detector framework. As shown in Fig. 2, the network includes an identity feature encoder and a detection head. The identity feature encoder follows the UNet [31,35,40] design, widely used in the dense prediction task. It includes skip layers to encode multi-scale features into a single point-based embedding. This embedding is directly used as the identity features, based on which an id loss (ℓ_{id}) [24,49,52] can be applied to train the identity feature encoder. Following FCOS [43], our detection head is a standard 4-layer fully convolutional head that is added on top of the identity feature encoder. It is trained with the same detection loss (ℓ_{det}) as that in FCOS [43]. The model is jointly optimized with detection and id losses in the following form $\ell_{total} = \ell_{det} + \lambda\ell_{id}$, in which λ is the modulated weight for the id loss. For easy reference, we refer to this network as PointID.

4. Unsupervised Person Similarity Learning

The most direct way to train PointID would be with a large dataset of images or videos that include both person bounding box and identity annotations. Unfortunately, collecting and annotating such a dataset is prohibitively expensive, and thus most works train their models on the smaller datasets available with such annotations [33,49,57]. In this work we instead focus on how to train our model without manual identity annotations. First we present our dense contrastive similarity loss (ℓ_{id}) which we use to train our identity feature encoder (Sec. 4.1). This loss still requires person identity information, for which we do not have annotations. Next we demonstrate how we train our model on images labeled with only person boxes (Sec. 4.2) and on unlabeled videos (Sec. 4.3).

4.1. Dense Contrastive similarity loss

We propose to use a contrastive loss for person similarity learning, which pushes the embeddings of people with the same identity together while pulling the embeddings of people with different identities apart. Formally, given a point (x, y) that corresponds to person identity i (i.e. $\texttt{ID}(x, y) = i$) and its feature vector $\mathbf{v} = \mathbf{f}^{(x,y)}$, we calculate the similarity loss for point (x, y) as follows.

$$\ell_{id}^{(x,y)} = \frac{1}{|\mathbf{P}_i|}\sum_{\mathbf{p}\in\mathbf{P}_i} -\texttt{log}\left(\frac{\texttt{exp}(\mathbf{v}\cdot\mathbf{p}/\tau)}{\texttt{exp}(\mathbf{v}\cdot\mathbf{p}/\tau) + \sum_{\mathbf{n}\in\mathbf{N}_i}\texttt{exp}(\mathbf{v}\cdot\mathbf{n}/\tau)}\right) \tag{1}$$

in which $\mathbf{P}_i = \cup_{\{\texttt{ID}(\tilde{x},\tilde{y})=i\}}\mathbf{f}^{(\tilde{x},\tilde{y})}$ is the set of point feature vectors $\mathbf{p}$ that corresponds to person i in current training batch, $\mathbf{N}_i = \cup_{\{\texttt{ID}(\tilde{x},\tilde{y})\neq i\}}\mathbf{f}^{(\tilde{x},\tilde{y})}$ are the point features of persons other than i, and τ is the temperature. Note that all feature vectors (i.e. $\mathbf{v}$, $\mathbf{p}$, $\mathbf{n}$) are $\texttt{L}_2$ normalized. In general, $\ell_{id}^{(x,y)}$ summarizes the average contrastive loss between point feature vector at (x, y) and the remaining point feature vectors that also corresponds to person i. As the example in Fig. 3 shows, it's important to point out that each person is sampled from multiple point feature vectors in the same image. Thus, each point feature vector corresponding to the same person identity captures a slightly different "view" of that particular person, which provides natural jitter in the feature level. This non-sparse sampling is also shown to be important in recent object tracking literature [38].

Next,we derive our dense similarity loss by averaging the above loss to all point feature vectors within an image that corresponds to a person instance. Mathematically,

$$\ell_{id} = \frac{1}{\texttt{N}_{\texttt{pos}}}\sum_{x,y}\mathbb{1}(\mathbf{c}^{(x,y)})\ell_{id}^{(x,y)} \tag{2}$$

where $\mathbb{1}(\mathbf{c}^{(x,y)})$ indicates whether the point (x, y) belongs to the center region of a person box, which is the same as that used in the detection head [43] to supervise the classification map.

In the ideal case where annotated person identities are available [49, 50], we can manage a memory bank $\mathbf{m} \in \mathbb{R}^{d\times\texttt{M}}$ ($\texttt{M}$ being the number of person ids, and d the channels of feature vectors) that stores a feature vector for each person identity, each of which conceptually summarizes all views of a particular person. Therefore, $\mathbf{P}_i$ and $\mathbf{N}_i$ can be constructed by simply retrieving the corresponding slices of memory, e.g. $\mathbf{P}_i = \{\mathbf{m}_i\}$, $\mathbf{N}_i = \{\ldots\mathbf{m}_{i-1}, \mathbf{m}_{i+1}, \ldots\}$. In the following sections, we explore a different and more challenging setting in which annotated person identities are not provided during training and show how to construct $\mathbf{P}_i$ and $\mathbf{N}_i$ in those cases.

4.2. Self-Supervised Image Training

Given a public image dataset annotated with person bounding boxes but no identity information, we assume that every image is sourced independently in the wild and thus

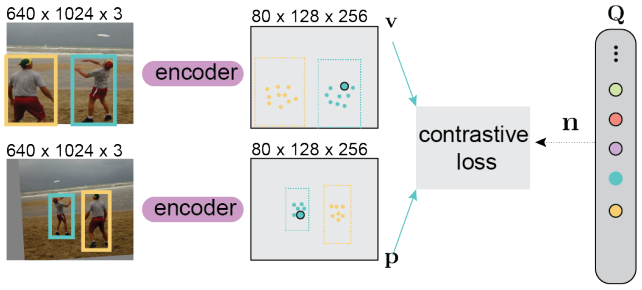


Figure 3. In contrastive similarity learning, feature vectors corresponding to the same person (e.g. all blue points) would be pulled together, while they are pushed away from the feature vectors for other people (e.g. yellow points).

every person in each image has a unique identity. [1] In order to generate meaningful $\mathbf{P}_i$ for person i, we could use the same memory strategy presented in Sec.4.1, but as each person i appears only once per training epoch, the memory feature $\mathbf{m}_i$ would be out of date by the time person i is seen again during training. Instead, we adopt the concept of instance discrimination [8,19,22,48] to construct positive and negative feature sets $\mathbf{P}_i$, $\mathbf{N}_i$ for each person box. Specifically, we apply image-level transformations [8, 19, 22] to synthesize two views of the same image and sample all positive feature vectors for person i from both images, giving rise to $\mathbf{P}_i$. An example is shown in Fig. 3. To construct $\mathbf{N}_i$, we manage a fixed-size queue $\mathbf{Q}$ that is updated with the feature vector of the latest encountered person embeddings, so all features in $\mathbf{Q}$ are relatively up-to-date. Therefore, we can easily get the negative feature set $\mathbf{N}_i$ for person i by retrieving feature vectors in $\mathbf{Q}$ that do not belong to identity i. As the ground truth person boxes are available during training, the full PointID model – id feature encoder and detection head – is jointly optimized with ℓ_{det} and ℓ_{id}.

4.3. Unsupervised Video Training

The above image based training generates a reasonably diverse set of views of the same person (i.e. $\mathbf{P}_i$) that is able to guide the model to focus on informative visual features (i.e. clothing or distinct accessories, etc) rather than distractors (i.e. lightning, scale, etc). However, a shortcut and non-trivial solution is to associate the person's identity with distinct features that are not related to person identity, for example a person's unique pose. This is due to the fact that the image-level augmentation is limited in its ability to produce instance-level transformations that cover the full range of expected movement for a person.

To address this limitation, we exploit the temporal continuity of videos to automatically generate multiple views of the same person with large pose and viewpoint changes that are not present in our image-level augmentation. We achieve this by linking a person across time, as the embeddings between consecutive frames are very similar for the same person. This is reminiscent of object tracking [1, 20, 28, 42, 55]. However, their embeddings could be very different if they are extracted from distant timestamps, and thus they can be naturally used as hard training examples in our similarity learning formulation. Visual examples are provided in supplementary materials. Analogous to the presumption underlying the person instance discrimination task, we assume that the same person does not appear in a different video. [2]

Specifically, we extract person bounding boxes and their embeddings by using our PointID network trained in Sec. 4.2. We cluster all detected people in a video by using the density-based clustering method DBSCAN [17] on all detected person identity embeddings. As DBSCAN is able to identify outliers in the embedding space, each detected person instance would be assigned to either a unique identity or an invalid identity that would not be used during training. This video-level embedding clustering serves as an offline solver to produce person trajectories in a video, within which a person's appearance can change substantially. This approach to generating pseduo-labels resembles other unsupervised learning methods in image classification [3] and person re-identification [6, 18].

To compute the id loss ℓ_{id}, we sample two different frames $\mathbf{I}_1, \mathbf{I}_2$ from a video. For a particular person i, we construct its positive feature set $\mathbf{P}_i$ by sampling all feature vectors for person i on both frames where it appears. We construct the negative feature set $\mathbf{N}_i$ using the same queue strategy presented in Sec. 4.2. As ground truth person boxes are not available, we only fine-tune the weights of the identity feature encoder with id loss ℓ_{id} and freeze the weights of detection head. This ensures that the noise in detection boxes does not derail the model training. Finally, we freeze the identity feature encoder and fine-tune the detection head with ℓ_{det} on ground truth person boxes provided by the image dataset in Sec. 4.2.

5. Experiments

Image-based person detection dataset. We use the COCO [30] and CrowdHuman [41] datasets, which have been widely used for general object and person detection. We discard all images in which no person bounding boxes are annotated. That leaves in total 64,115 images with

[1] Under such assumption, several public data sets for person detection are not appropriate for the proposed learning framework, such as CityPerson [54], Caltech Pedestrian [12], to name a few.

[2] In general, this is a fair assumption for two random in-the-wild videos. However, this assumption would not hold in multi-camera surveillance scenario where the same person can appear in different videos. We would like to come back to this issue during discussion and explore this specific situation in the future.

257,249 person boxes for COCO and 15,000 images with 339,563 person boxes for CrowdHuman.

Unlabeled video dataset. We use the Kinetics-700 dataset [26] that is primarily used for action recognition. We anticipate that the person's pose and viewpoint change drastically in those videos due to camera motion and people performing different actions. We randomly sample 150,000 videos in our study, from which 45,108 (roughly 30%) videos are identified to include at least 1 valid person trajectory. In total, that includes 57,200 unique person identities and more than 1 million person boxes. Specifically, we name this dataset Kinetics-150K.

Person search dataset. We validate the performance of our person detection and re-identification model on person search. To this end, we adopt both CUHK-SYSU [49] and PRW [57] in our study. In detail, CUHK-SYSU includes 18,184 images (11,206 train / 6,978 val) with 8,432 unique ids (5,532 train / 2,900 val), and PRW includes 11,816 (5,704 train / 6,112 val) images with 932 unique ids (482 train / 450 val). Following other literature [13,34,49,50,58], we report both MAP and Top-1 metrics by using the default gallery size of 100.

Network configuration and training. We take ResNet-50 [23] as the backbone of our id feature encoder unless specified otherwise. As shown in Fig. 2, we upsample the feature maps from previous layers and concatenate it with the output of the corresponding skip layer, which is parametrized as a single deformable convolution kernel [9] with 256 output channels. The scaling temperature used in Eq. 1 is set to be 0.07 and the id loss modulating factor λ is empirically set to 0.2 (the effect of λ is studied in supplementary materials). We first train the model on COCO and CrowdHuman with 32 unique image pairs per iteration (64 actual images due to augmentation). We optimize the model with momentum-based SGD starting with the initial learning rate of 0.01, which is modulated with 0.1 after the model completes 60% and 80% of all iterations (25,000 in total). The size of the latest encountered person pool $\mathbf{Q}$ is set to be 32,768 unless specified otherwise. Next, we fine-tune the identity feature encoder on Kinetics-150K for 20,000 total iterations. Finally, we fine-tune the detection head on COCO and CrowdHuman fro another 10,000 iterations. During training, each image is resized to have a shorter size that is randomly drawn from $\{480, 560, 640, 720, 800\}$ while constraining it with larger size to be less than 1024 pixels. During inference, images are resized to have 640×1024 pixels.

Image-level augmentation. We apply the following image-level transformation in image training: `rotation`, `occlusion` (random patch erasing [59]), `video jitter` (e.g. motion blur, JPEG compression), image `mirror`, `zoom-in` motion transformation. In the supplementary materials, we define those transformations and also show the concrete visual examples of the effect for those transformations.

Train data	Person Id	Loss	MAP	Top-1
CUHK	✓	$\ell_{det} + \ell_{id}$	92.7%	93.7%
n/a	×	n/a	28.0%	27.8%
COCO+CH	×	ℓ_{det}	17.0%	15.8%
COCO+CH	×	ℓ_{det}	14.3%	12.5%
COCO+CH	×	$\ell_{det} + \ell_{id}$	76.5%	78.0%
COCO+CH+Kinetics	×	$\ell_{det} + \ell_{id}$	82.5%	84.6%

Table 1. Results on CUHK-SYSU dataset. Results in gray indicate that the models use oracle boxes during inference. CH is CrowdHuman dataset.

5.1. Main result

We firstly train PointID on CUHK-SYSU with full supervision. We adopt the id loss in Eq. 1 and manage the memory for all ids as that elaborated in Sec. 4.1. The implementation details are included in the supplementary materials. In Tab. 1 and Tab. 3, we show that our model achieves 92.7% MAP and 93.7% Top-1 accuracy, which performs competitively well against state-of-the-art ResNet-50 based models [13,34,49,50,58]. The model runs at 29.7 FPS on a single V100 GPU. We don't dig into analysing different network components of PointID, as this is not the focus of this work and much of its design philosophies are derived from AlignPS [50] and dense prediction networks [31,35,40].

To understand the difficulty of unsupervised person similarity learning, we setup two baseline models: the first one being the model whose weights are initialized from an ImageNet pre-trained model [10]; the second one being the model whose weights are optimized for person detection on COCO [30] and CrowdHuman [41]. As the detection head is not trained in the first model, we introduce oracle detection based inference in which feature vectors corresponding to the center points of ground truth boxes are used for matching. In this case, the result of the model would be better than the same model under default inference. We report their results in Tab. 1, which shows that both models perform abysmally. Interestingly, the feature trained for the person detection task does even worse, suggesting that features used for person detection and re-identification tasks are not naturally compatible.

In Tab. 1, we further report the results of the same model trained on person bounding boxes from COCO and CrowdHuman but with the proposed contrastive similarity loss, as well as a model that is further fine-tuned on unlabeled videos from Kinetics-150K [26]. Their benefits over baselines are substantial. Even without training our model on the target dataset, it does not trail significantly behind the fully-

Loss	data	MAP	Top-1
Cross Entropy	part	54.4%	55.7%
Contrastive	part	75.5%	77.8%
Contrastive	full	76.5%	78.0%

(a) Result comparison of models trained with different losses on COCO (**CO**) and CrowdHuman (**CH**).

Frame sampling	MAP	Top-1
n/a (memory)	84.3%	86.2%
Random	84.4%	85.4%
Biased	**85.5%**	**86.7%**

(b) Result comparison of models trained with different frame sampling strategies on Kinetics-100K.

Augment	data	MAP	Top-1
Image	CO+CH	78.9%	79.8%
Image	CO+CH+K	73.2%	74.4%
Instance	CO+CH+K	85.5%	86.7%

(c) Result comparison of models trained with different augmentation on Kinetics-100K (**K**).

Table 2. Ablation experiments on CUHK-SYSU [49]. Results in gray indicate that ground truth person boxes are used during inference.

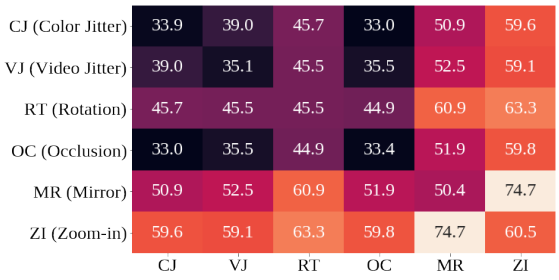
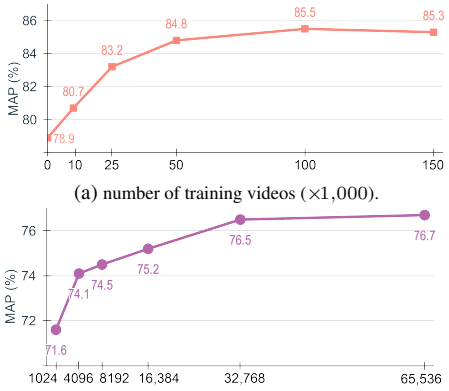


Figure 4. Results (MAP) of the model trained with pairwise combinations of transformations on CUHK-SYSU dataset [49].



(a) number of training videos (×1,000).

(b) Size of **Q**.

Figure 5. Ablation studies on CUHK-SYSU dataset [49].

supervised model. We compare those models with qualitative visual examples in the supplementary materials. In the following, we conduct careful ablation experiments to understand how this model achieves this level of performance.

5.2. Image Training

In this section, all models are trained with ℓ_{det} and ℓ_{id} on the COCO [30] and CrowdHuman [41] dataset.

How does image-level augmentation affect similarity learning? To answer this question, we enumerate all pairwise combination of transformations and summarize their results in Fig. 4. As shown, `zoom-in` motion transformation and image `mirror` are the two most significant augmentations that heavily influences the results of the model, as the model trained with these two transformations trails behind the full model by only 1.8%. This is not surprising as `zoom-in` motion transformation introduces large scale variation and image `mirror` augmentation breaks shortcut solutions that simply memorizing the person's pose.

Cross entropy loss vs contrastive loss. We compare the proposed contrastive similarity loss with cross entropy loss, which is widely used in similarity learning [52, 55]. However the cross entropy loss is not scalable to a large number of person identities, in which the model needs to learn a gigantic projection matrix (e.g. $\mathbb{R}^{256\times \mathtt{N}}$ and $\mathtt{N}$ being the number of unique person identities) before feature vectors are optimized for re-identification. We found the model struggles to converge on the full COCO and CrowdHuman datasets that includes approximately $\mathtt{N} = 600{,}000$ boxes (ids). To make the training feasible, we sample a subset of images that includes roughly 50,000 boxes (8.3% of all boxes) for this particular study. Results are summarized in Tab. 2 (a). As shown, the model trained with contrastive similarity loss substantially outperforms the one trained on cross entropy loss, and its results continue to improve when all person boxes are used during training.

Size of latest encountered person embeddings Q. We further show how the size of **Q** affects the model training in Fig. 5b. We observe that the model's performance keeps improving when more person identities are cached into **Q**. This is expected as more hard examples can be retrieved from **Q**, which benefits the embedding learning in our similarity loss formulation (Eq. 1). We identify $|\mathbf{Q}| = 32{,}768$ to be a good trade-off between memory consumption and performance optimization.

5.3. Video Fine-tuning

In this section, all models are fine-tuned with ℓ_{id} alone, so we use the ground truth boxes during inference.

Does frame sampling affect model training? As shown in Fig. 4, it becomes clear that having a diverse set of views of the same person is one key aspect to contrastive similarity learning. To test how frame sampling influences the model training, we compare two models trained with the following strategies: 1) sample two random frames from a video; and 2) sample one random frame from the first half

of the video and one from the second, which we call biased sampling. In the latter case, the appearance of the same person is more likely to be different, which is expected to produce a harder positive feature set $\mathbf{P}_i$ for that particular person. As shown in Tab. 2 (b), the model trained with biased sampling does have a clear edge over the one trained with random sampling. In addition, we compare against the model trained with memory-based contrastive loss, as the memory conceptually aggregates the feature of a particular person from the full video. In Tab. 2 (b), we also find it is outperformed by the model trained with non-memory based contrastive loss. As discussed earlier, this is largely due to the difficulty of maintaining an up-to-date memory. In practice the memory degenerates to only summarizing a random view of the person in the video.

Does the number of training person identities affect the model's performance? To answer this question, we randomly sample a subset of videos from Kinetics-150K. For example, Kinetics-10K is a subset of Kinetics-150K that includes $10{,}000$ videos, in which there are $3{,}060$ videos that include at least 1 valid person trajectory. We include the statistics of different subsets in the supplementary materials. We train the model on those subsets with the same training configurations, and their results are summarized in Fig. 5a. Overall, we observe a clear upward trend when more videos (or person identities) are used during training. For example, adding $9{,}594$ unique person instances from the Kinetics-25K subset to the model training brings in 4.27% MAP improvement. This result suggests that videos introduce instance-level appearance transformations that are possibly missing in the image training. The benefit saturates beyond the Kinetics-100K subset that includes $38{,}266$ unique person identities and covers as many as $765{,}000$ person bounding boxes. That amount of data is sufficient for the model to exploit the instance-level appearance diversity for robust similarity learning.

Do the benefits really come from instance-level diversity? This question emerges naturally when videos are used during model fine-tuning, as encountering extra video frames could be the major cause that lifts the model's performance. To this end, we fine-tune the model by adopting the same techniques as that in Sec. 4.2: a frame $\mathbf{I}$ is firstly sampled for each video and then the other frame $\mathbf{I}'$ is generated by applying image-level transformation to $\mathbf{I}$. We compare the results of different models in Tab. 2 (c). As shown, fine-tuning the model on video frames alone in fact harms the model's performance. This is probably due to the fact that the hard-to-detect people are excluded during training, so the learning is relatively straightforward to discriminate a set of prominent people in easy background. In this case, it's clear that it's the instance-level appearance diversity that largely contributes to improving the model's performance.

methods	CUHK-SYSU MAP	CUHK-SYSU Top-1	PRW MAP	PRW Top-1
QEEPS [34]	88.9%	89.1%	37.1%	76.7%
APNet [58]	88.9%	89.3%	41.2%	81.4%
BiNet [13]	90.0%	90.7%	45.3%	81.7%
NAE [5]	91.5%	92.4%	43.3%	80.9%
AlignPS [50]	93.1%	93.4%	45.9%	81.9%
Ours (baseline)	92.7%	93.7%	41.3%	81.1%
Ours (detection pretrain)	93.2%	94.1%	42.4%	83.1%
Ours (similarity pretrain)	**94.8%**	**95.5%**	**47.4%**	**84.1%**

Table 3. Result comparison on person search datasets.

5.4. Similarity pre-training

In this section, we further investigate the benefits of using the proposed learning framework as a pre-training step.

It helps to reduce annotation cost. We first study how similarity pre-training helps reduce the need for manual identity annotation. For this purpose, we randomly sample a fraction of annotated person identities and mask out the rest as if they are not annotated. Then we compare two models, one being trained as normal (ImageNet pre-training), the other one being pre-trained with our proposed id loss. Their results are summarized in Fig. 1. As clearly shown: 1) the similarity pretrained model consistently outperforms its counterpart, especially when the number of identity annotations are limited and 2) with only 40% of identity annotation, the similarity pretrained model achieves the same level of performance of the standard model that is trained on the full dataset. Those results suggest that the similarity pre-training has the potential to reduce the annotation cost to achieve a certain level of re-identification performance.

It improves the model's performance. Next, we train the above two models on the full target datasets. Tab. 3 lists the results on both CUHK-SYSU [49] and PRW [57] datasets. As shown, our similarity pre-trained model achieves significantly better results than baseline, and the benefit is more pronounced ($+6.1\%$ MAP) on the smaller PRW. We also compare the model that is pre-trained with ℓ_{det} on COCO [30] and CrowdHuman [41]. Although detection pre-training boosts the model's performance, it significantly lags behind the similarity pre-trained one. This result indicates that it is the proposed similarity loss that makes the difference.

6. Multi-person tracking

In order to validate the greater potential of our model, we further apply it to multi-person tracking, specifically to MOT17 [33]. We find that person re-identification is less challenging in person tracking compared to that in person search. This is due to the fact that the person's appearance changes slightly from frame to frame, and that there are

Methods	Train data	Person Id	MOTA↑	IDF1↑	IDsw↓
CenterTrackV1 [60]	CH	×	58.1	53.3	946
SiamMOT [42]	CH	×	63.4	60.8	616
FairMOT [55]	CH	×	62.9	63.2	813
Ours	CH	×	**66.9**	**70.9**	**372**

Table 4. Result comparison on full MOT17 train set, and CH is for CrowdHuman dataset [41].

Methods	Train data	Person Id	MOTA↑	IDF1↑	MT↑	ML↓	IDsw↓
TubeTK [37]	MOT17	✓	63.0	58.6	31.2	19.9	4137
CTrackerV1 [39]	MOT17	✓	66.6	57.4	32.2	24.4	5529
QuasiDense [38]	MOT17	✓	68.7	66.3	40.6	21.9	3378
CenterTrackV1 [60]	CH+MOT17	✓	67.3	59.9	34.9	24.8	2898
CenterTrackV2 [60]	CH+MOT17	✓	67.8	67.1	34.6	24.6	**2583**
SOTMOT [56]	MIX	✓	71.0	70.9	42.7	15.3	5184
FairMOT [55]	CH+MIX	✓	73.7	72.3	44.7	15.9	3303
Ours	CH	×	**74.2**	**72.4**	**46.6**	**12.2**	2748

Table 5. Result comparison on MOT17 test set with "private detection" protocol. MIX [47,55] is a combined dataset that includes MOT17 and 5 other datasets, all of which are manually annotated with person identities.

very few long occlusions of a person (e.g. longer than 10 seconds) that induces large appearance changes. Therefore, we train our model only on CrowdHuman [41] (amodal bounding box annotation) with the techniques elaborated in Sec. 4.2. We adopt the online solver implemented by FairMOT [55] to generate temporal trajectory for each person, which takes as input the detected person boxes and their embedding per frame. Following the common standard, we report the following metrics: MOTA, IDF1, IDsw (ID switches), MT (Mostly Tracked), ML (Mostly lost).

Comparison with other self-supervised tracking models. There are a few tracking models that can also be trained without person identity annotation. For example, CenterTrack [60] and SiamMOT [42] train a person motion model with image pairs generated from the same image. Similar to ours, FairMOT [55] learns a person re-identification model but with cross entropy loss. We compare their end-to-end results on the full MOT17 train set in Tab. 4. As shown, our model has an impressive performance edge over all other models on every key metric. Note that all models use the standard DLA-34 [51] as the feature backbone, so their results are comparable.[3] Particularly, it's interesting to compare our model with FairMOT [55], as both models use the same online solver during inference. The substantial performance benefits of our model over FairMOT can be largely attributed to a better person re-identification model, which is consistent with the observation in Tab. 2 (a).

[3]We generate the results by using the official implementation of the above methods and by using the provided model weights pretrained on the CrowdHuman dataset. Then we use the same evaluation code to generate their corresponding evaluation metrics.

Comparison with state-of-the-arts. We generate the results on the MOT17 test sequence [33] by submitting the output of our model to the evaluation server. As shown in Tab. 5, our model outperforms recent state-of-the-art methods even though it has not been trained either on the target dataset or with person identity annotations. This promising result suggests that we are able to develop a generalized and well-performing person tracking model without first curating a multi-person tracking dataset for training purposes, whose cost is prohibitively high in reality.

7. Discussion and Conclusion

In this work, we aim to learn a generalized model for joint person detection and re-identification. Unfortunately, the existing datasets that can be used to train such a model are limited both in scope and scale. To this end, we develop a contrastive similarity learning framework such that the person embedding model can be jointly optimized with the person detection model on images / videos without manual id annotations. We show that our model generalizes to both person search and person tracking datasets, where it achieves promising results compared to fully-supervised state-of-the-art methods. In the following, we first talk about the limitation of our assumption underlying the learning framework, and we further discuss the potential of our work to reduce the annotation cost to a larger degree.

Limitation. The essential assumption underlying our contrastive similarity learning framework is that each image / video is independent, which means that the same person does not appear in different images or videos. In the cases in which the same person is captured by multiple different images or videos, the proposed learning framework would not work properly. To leverage those precious images/videos for training, we can adopt similar techniques used in unsupervised person re-identification work [6, 18]. We would like to explore using a wider range of images / videos for Id-Free person similarity learning in the future.

Annotation cost reduction. As we have shown in Tab. 3 and Tab. 5, our model can achieve very competitive results on both person search and person tracking tasks even if it is not trained on the target data set. As also shown in Fig. 1, adding as little as 10% of person identity annotations significantly lifts the model's performance on the target dataset. In the same spirit of active learning, we are interested in exploring how to automatically identify a few hard training examples with our model that are needed for human annotation. This is an important step to achieve the best trade-off between annotation cost and performance maximization.

Acknowledgment. The authors would like to thank Uta Buechler and Andrew Berneshawi for idea brainstorming and paper proofreading.

References

[1] Luca Bertinetto, Jack Valmadre, Joao F Henriques, Andrea Vedaldi, and Philip HS Torr. Fully-convolutional siamese networks for object tracking. In *European conference on computer vision*, pages 850–865. Springer, 2016. 4

[2] Alex Bewley, Zongyuan Ge, Lionel Ott, Fabio Ramos, and Ben Upcroft. Simple online and realtime tracking. In *2016 IEEE international conference on image processing (ICIP)*, pages 3464–3468. IEEE, 2016. 2

[3] Mathilde Caron, Piotr Bojanowski, Armand Joulin, and Matthijs Douze. Deep clustering for unsupervised learning of visual features. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 132–149, 2018. 4

[4] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. *arXiv preprint arXiv:2006.09882*, 2020. 2

[5] Di Chen, Shanshan Zhang, Jian Yang, and Bernt Schiele. Norm-aware embedding for efficient person search. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12615–12624, 2020. 2, 7

[6] Hao Chen, Benoit Lagadec, and Francois Bremond. Ice: Inter-instance contrastive encoding for unsupervised person re-identification. *ICCV*, 2021. 2, 4, 8

[7] Hao Chen, Yaohui Wang, Benoit Lagadec, Antitza Dantcheva, and Francois Bremond. Joint generative and contrastive learning for unsupervised person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2004–2013, 2021. 2

[8] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020. 1, 2, 4

[9] Jifeng Dai, Haozhi Qi, Yuwen Xiong, Yi Li, Guodong Zhang, Han Hu, and Yichen Wei. Deformable convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 764–773, 2017. 5

[10] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 2, 5

[11] Carl Doersch, Abhinav Gupta, and Alexei A Efros. Unsupervised visual representation learning by context prediction. In *Proceedings of the IEEE international conference on computer vision*, pages 1422–1430, 2015. 2

[12] Piotr Dollar, Christian Wojek, Bernt Schiele, and Pietro Perona. Pedestrian detection: An evaluation of the state of the art. *IEEE transactions on pattern analysis and machine intelligence*, 34(4):743–761, 2011. 4

[13] Wenkai Dong, Zhaoxiang Zhang, Chunfeng Song, and Tieniu Tan. Bi-directional interaction network for person search. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2839–2848, 2020. 2, 5, 7

[14] Wenkai Dong, Zhaoxiang Zhang, Chunfeng Song, and Tieniu Tan. Instance guided proposal network for person search. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2585–2594, 2020. 2

[15] Kaiwen Duan, Song Bai, Lingxi Xie, Honggang Qi, Qingming Huang, and Qi Tian. Centernet: Keypoint triplets for object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6569–6578, 2019. 2

[16] Debidatta Dwibedi, Yusuf Aytar, Jonathan Tompson, Pierre Sermanet, and Andrew Zisserman. Temporal cycle-consistency learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1801–1810, 2019. 2

[17] Martin Ester, Hans-Peter Kriegel, Jörg Sander, Xiaowei Xu, et al. A density-based algorithm for discovering clusters in large spatial databases with noise. In *kdd*, 1996. 2, 4

[18] Yixiao Ge, Feng Zhu, Dapeng Chen, Rui Zhao, and Hongsheng Li. Self-paced contrastive learning with hybrid memory for domain adaptive object re-id. *NIPS*, 2020. 2, 4, 8

[19] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre H Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Daniel Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent: A new approach to self-supervised learning. *NIPS*, 2020. 1, 2, 4

[20] Dongyan Guo, Jun Wang, Ying Cui, Zhenhua Wang, and Shengyong Chen. Siamcar: Siamese fully convolutional classification and regression for visual tracking. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6269–6277, 2020. 4

[21] Chuchu Han, Jiacheng Ye, Yunshan Zhong, Xin Tan, Chi Zhang, Changxin Gao, and Nong Sang. Re-id driven localization refinement for person search. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9814–9823, 2019. 2

[22] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9729–9738, 2020. 1, 2, 4

[23] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 5

[24] Alexander Hermans, Lucas Beyer, and Bastian Leibe. In defense of the triplet loss for person re-identification. *arXiv preprint arXiv:1703.07737*, 2017. 3

[25] Shyamgopal Karthik, Ameya Prabhu, and Vineet Gandhi. Simple unsupervised multi-object tracking. *arXiv preprint arXiv:2006.02609*, 2020. 2

[26] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*, 2017. 5

[27] Nikos Komodakis and Spyros Gidaris. Unsupervised representation learning by predicting image rotations. In *International Conference on Learning Representations (ICLR)*, 2018. 2

[28] Bo Li, Junjie Yan, Wei Wu, Zheng Zhu, and Xiaolin Hu. High performance visual tracking with siamese region proposal network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8971–8980, 2018. 4

[29] Wei Li, Yuanjun Xiong, Shuo Yang, Mingze Xu, Yongxin Wang, and Wei Xia. Semi-tcl: Semi-supervised track contrastive representation learning. *arXiv preprint arXiv:2107.02396*, 2021. 2

[30] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014. 1, 4, 5, 6, 7

[31] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440, 2015. 3, 5

[32] Daniel McKee, Bing Shuai, Andrew Berneshawi, Manchen Wang, Davide Modolo, Svetlana Lazebnik, and Joseph Tighe. Multi-object tracking with hallucinated and unlabeled videos. In *CVPR workshop*, 2021. 2

[33] Anton Milan, Laura Leal-Taixé, Ian Reid, Stefan Roth, and Konrad Schindler. Mot16: A benchmark for multi-object tracking. *arXiv preprint arXiv:1603.00831*, 2016. 1, 2, 3, 7, 8

[34] Bharti Munjal, Sikandar Amin, Federico Tombari, and Fabio Galasso. Query-guided end-to-end person search. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 811–820, 2019. 2, 5, 7

[35] Hyeonwoo Noh, Seunghoon Hong, and Bohyung Han. Learning deconvolution network for semantic segmentation. In *Proceedings of the IEEE international conference on computer vision*, pages 1520–1528, 2015. 3, 5

[36] Mehdi Noroozi and Paolo Favaro. Unsupervised learning of visual representations by solving jigsaw puzzles. In *European conference on computer vision*, pages 69–84. Springer, 2016. 2

[37] Bo Pang, Yizhuo Li, Yifan Zhang, Muchen Li, and Cewu Lu. Tubetk: Adopting tubes to track multi-object in a one-step training model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6308–6318, 2020. 2, 8

[38] Jiangmiao Pang, Linlu Qiu, Xia Li, Haofeng Chen, Qi Li, Trevor Darrell, and Fisher Yu. Quasi-dense similarity learning for multiple object tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 164–173, 2021. 2, 3, 8

[39] Jinlong Peng, Changan Wang, Fangbin Wan, Yang Wu, Yabiao Wang, Ying Tai, Chengjie Wang, Jilin Li, Feiyue Huang, and Yanwei Fu. Chained-tracker: Chaining paired attentive regression results for end-to-end joint multiple-object detection and tracking. In *European Conference on Computer Vision*, pages 145–161. Springer, 2020. 8

[40] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015. 3, 5

[41] Shuai Shao, Zijian Zhao, Boxun Li, Tete Xiao, Gang Yu, Xiangyu Zhang, and Jian Sun. Crowdhuman: A benchmark for detecting human in a crowd. *arXiv preprint arXiv:1805.00123*, 2018. 1, 4, 5, 6, 7, 8

[42] Bing Shuai, Andrew Berneshawi, Xinyu Li, Davide Modolo, and Joseph Tighe. Siammot: Siamese multi-object tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12372–12382, 2021. 1, 2, 4, 8

[43] Zhi Tian, Chunhua Shen, Hao Chen, and Tong He. Fcos: Fully convolutional one-stage object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 9627–9636, 2019. 3

[44] Cheng Wang, Bingpeng Ma, Hong Chang, Shiguang Shan, and Xilin Chen. Tcts: A task-consistent two-stage framework for person search. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11952–11961, 2020. 2

[45] Ning Wang, Yibing Song, Chao Ma, Wengang Zhou, Wei Liu, and Houqiang Li. Unsupervised deep tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1308–1317, 2019. 2

[46] Xiaolong Wang, Allan Jabri, and Alexei A Efros. Learning correspondence from the cycle-consistency of time. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2566–2576, 2019. 2

[47] Zhongdao Wang, Liang Zheng, Yixuan Liu, Yali Li, and Shengjin Wang. Towards real-time multi-object tracking. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XI 16*, pages 107–122. Springer, 2020. 1, 2, 8

[48] Zhirong Wu, Yuanjun Xiong, Stella X Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance discrimination. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3733–3742, 2018. 1, 2, 4

[49] Tong Xiao, Shuang Li, Bochao Wang, Liang Lin, and Xiaogang Wang. Joint detection and identification feature learning for person search. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3415–3424, 2017. 1, 2, 3, 5, 6, 7

[50] Yichao Yan, Jinpeng Li, Jie Qin, Song Bai, Shengcai Liao, Li Liu, Fan Zhu, and Ling Shao. Anchor-free person search. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 7690–7699, 2021. 1, 2, 3, 5, 7

[51] Fisher Yu, Dequan Wang, Evan Shelhamer, and Trevor Darrell. Deep layer aggregation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2403–2412, 2018. 8

[52] Yao Zhai, Xun Guo, Yan Lu, and Houqiang Li. In defense of the classification loss for person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–0, 2019. 3, 6

[53] Richard Zhang, Phillip Isola, and Alexei A Efros. Colorful image colorization. In *European conference on computer vision*, pages 649–666. Springer, 2016. 2

[54] Shanshan Zhang, Rodrigo Benenson, and Bernt Schiele. Citypersons: A diverse dataset for pedestrian detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3213–3221, 2017. 4

[55] Yifu Zhang, Chunyu Wang, Xinggang Wang, Wenjun Zeng, and Wenyu Liu. Fairmot: On the fairness of detection and re-identification in multiple object tracking. *International Journal of Computer Vision*, pages 1–19, 2021. 1, 2, 4, 6, 8

[56] Linyu Zheng, Ming Tang, Yingying Chen, Guibo Zhu, Jinqiao Wang, and Hanqing Lu. Improving multiple object tracking with single object tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2453–2462, 2021. 8

[57] Liang Zheng, Hengheng Zhang, Shaoyan Sun, Manmohan Chandraker, Yi Yang, and Qi Tian. Person re-identification in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1367–1376, 2017. 1, 3, 5, 7

[58] Yingji Zhong, Xiaoyu Wang, and Shiliang Zhang. Robust partial matching for person search in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6827–6835, 2020. 2, 5, 7

[59] Zhun Zhong, Liang Zheng, Guoliang Kang, Shaozi Li, and Yi Yang. Random erasing data augmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2020. 5

[60] Xingyi Zhou, Vladlen Koltun, and Philipp Krähenbühl. Tracking objects as points. In *European Conference on Computer Vision*, pages 474–490. Springer, 2020. 1, 2, 8