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, 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.

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 Fair-MOT [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. These 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], grayscale 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.
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 ($\ell_{id}$) [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. Following FCOS [43], it is trained with the same detection loss ($\ell_{det}$) 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 $\lambda$ 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 ($\ell_{id}$) 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 images (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. $ID(x, y) = i$) and its feature vector $v = f(x,y)$, we calculate the similarity loss for point $(x, y)$ as follows.

$$\ell_{id}^{(x,y)} = \frac{1}{|P_i|} \sum_{p \in P_i} -\log \left( \frac{\exp(v \cdot p / \tau)}{\sum_{n \in N_i} \exp(v \cdot n / \tau)} \right)$$

(1)

in which $P_i = \{ID(x, y) = i\} f(x, y)$ is the set of point feature vectors $p$ that corresponds to person $i$ in current training batch, $N_i = \{ID(x, y) \neq i\} f(x, y)$ are the point features of persons other than $i$, and $\tau$ is the temperature. Note that all feature vectors (i.e. $v$, $p$, $n$) are $L_2$ 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,

$$\ell_{id} = \frac{1}{N_{pos}} \sum_{x,y} \mathbb{1}(c(x,y)) \ell_{id}^{(x,y)}$$

(2)

where $\mathbb{1}(c(x,y))$ 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 $m \in \mathbb{R}^{d \times N}$ ($N$ 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, $P_i$ and $N_i$ can be constructed by simply retrieving the corresponding slices of memory, e.g. $P_i = \{m_i\}$, $N_i = \{\ldots m_i-1, m_{i+1}, \ldots\}$. 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 $P_i$ and $N_i$ 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
every person in each image has a unique identity.  

In order to generate meaningful $P_i$ 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 $m_i$ 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 $P_i, N_i$ 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 $P_i$. An example is shown in Fig. 3. To construct $N_i$, we manage a fixed-size queue $Q$ that is updated with the feature vector of the latest encountered person embeddings, so all features in $Q$ are relatively up-to-date. Therefore, we can easily get the negative feature set $N_i$ for person $i$ by retrieving feature vectors in $Q$ 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 $\ell_{det}$ and $\ell_{id}$.

### 4.3. Unsupervised Video Training

The above image based training generates a reasonably diverse set of views of the same person (i.e. $P_i$) 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.

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 pseudo-labels resembles other unsupervised learning methods in image classification [3] and person re-identification [6, 18].

To compute the id loss $\ell_{id}$, we sample two different frames $I_1, I_2$ from a video. For a particular person $i$, we construct its positive feature set $P_i$ by sampling all feature vectors for person $i$ on both frames where it appears. We construct the negative feature set $N_i$ 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 $\ell_{id}$ 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 $\ell_{det}$ 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 $\lambda$ is empirically set to 0.2 (the effect of $\lambda$ 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 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 finetune the detection head on COCO and CrowdHuman for 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 $640 \times 1024$ pixels.

**Image-level augmentation.** We apply the following image-level transformation in image training: rotation, occlusion (random patch erasing [39]), video jitter (e.g. motion blur, JPEG compression), image

<table>
<thead>
<tr>
<th>Train data</th>
<th>Person Id</th>
<th>Loss</th>
<th>MAP</th>
<th>Top-1</th>
</tr>
</thead>
<tbody>
<tr>
<td>COCO+CH</td>
<td>√</td>
<td>$\ell_{id}$</td>
<td>92.7%</td>
<td>93.7%</td>
</tr>
<tr>
<td>n/a</td>
<td>×</td>
<td>n/a</td>
<td>28.0%</td>
<td>27.8%</td>
</tr>
<tr>
<td>COCO+CH</td>
<td>×</td>
<td>$\ell_{id}$</td>
<td>17.0%</td>
<td>15.8%</td>
</tr>
<tr>
<td>COCO+CH</td>
<td>×</td>
<td>$\ell_{det}$ + $\ell_{id}$</td>
<td>76.5%</td>
<td>78.0%</td>
</tr>
<tr>
<td>COCO+CH+Kinetics</td>
<td>×</td>
<td>$\ell_{det}$ + $\ell_{id}$</td>
<td>82.5%</td>
<td>84.6%</td>
</tr>
</tbody>
</table>

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

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.

**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-
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 $\ell_{det}$ and $\ell_{id}$ 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 influence the result 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. $\mathbb{R}^{256 \times N}$ and $N$ 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 $N = 600,000$ boxes.

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

<table>
<thead>
<tr>
<th>Loss</th>
<th>data</th>
<th>MAP</th>
<th>Top-1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cross Entropy</td>
<td>part</td>
<td>54.4%</td>
<td>55.7%</td>
</tr>
<tr>
<td>Contrastive</td>
<td>part</td>
<td>75.5%</td>
<td>77.8%</td>
</tr>
<tr>
<td></td>
<td>full</td>
<td>76.5%</td>
<td>78.0%</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Frame sampling</th>
<th>data</th>
<th>MAP</th>
<th>Top-1</th>
</tr>
</thead>
<tbody>
<tr>
<td>n/a (memory)</td>
<td></td>
<td>84.3%</td>
<td>86.2%</td>
</tr>
<tr>
<td>Random</td>
<td></td>
<td>84.4%</td>
<td>85.4%</td>
</tr>
<tr>
<td>Biased</td>
<td></td>
<td>85.5%</td>
<td>86.7%</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Augment</th>
<th>data</th>
<th>MAP</th>
<th>Top-1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Image CO+CH</td>
<td></td>
<td>78.9%</td>
<td>79.8%</td>
</tr>
<tr>
<td>Image CO+CH+K</td>
<td></td>
<td>73.2%</td>
<td>74.4%</td>
</tr>
<tr>
<td>Instance CO+CH</td>
<td></td>
<td>85.5%</td>
<td>86.7%</td>
</tr>
</tbody>
</table>

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

---

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

---

5.3. Video Fine-tuning

In this section, all models are fine-tuned with $\ell_{id}$ 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.
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 $P_i$ 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 3,000 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 $I$ is firstly sampled for each video and then the other frame $I'$ is generated by applying image-level transformation to $I$. 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.

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

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 $\ell_{det}$ 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
### Results on MOT17 Train Set

<table>
<thead>
<tr>
<th>Methods</th>
<th>Train data</th>
<th>Person Id</th>
<th>MOTA↑</th>
<th>IDF1↑</th>
<th>IDsw↓</th>
</tr>
</thead>
<tbody>
<tr>
<td>CenterTrackV1</td>
<td>CH</td>
<td>×</td>
<td>58.1</td>
<td>53.3</td>
<td>946</td>
</tr>
<tr>
<td>SiamMOT</td>
<td>CH</td>
<td>×</td>
<td>63.4</td>
<td>60.8</td>
<td>616</td>
</tr>
<tr>
<td>FairMOT</td>
<td>CH</td>
<td>×</td>
<td>62.9</td>
<td>63.2</td>
<td>813</td>
</tr>
<tr>
<td><strong>Ours</strong></td>
<td>CH</td>
<td>×</td>
<td>66.9</td>
<td>70.9</td>
<td>372</td>
</tr>
</tbody>
</table>

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

### Results on MOT17 Test Set

<table>
<thead>
<tr>
<th>Methods</th>
<th>Train data</th>
<th>Person Id</th>
<th>MOTA↑</th>
<th>IDF1↑</th>
<th>MT↑</th>
<th>ML↓</th>
<th>IDsw↓</th>
</tr>
</thead>
<tbody>
<tr>
<td>TubeTK</td>
<td>MOT17</td>
<td>✓</td>
<td>63.0</td>
<td>58.6</td>
<td>31.2</td>
<td>19.9</td>
<td>4137</td>
</tr>
<tr>
<td>CTrackerVI</td>
<td>MOT17</td>
<td>✓</td>
<td>66.6</td>
<td>57.4</td>
<td>32.2</td>
<td>24.4</td>
<td>5529</td>
</tr>
<tr>
<td>QuasiDense</td>
<td>MOT17</td>
<td>✓</td>
<td>68.7</td>
<td>66.3</td>
<td>40.6</td>
<td>21.9</td>
<td>3378</td>
</tr>
<tr>
<td>CenterTrackV1</td>
<td>CH+MOT17</td>
<td>✓</td>
<td>67.3</td>
<td>59.9</td>
<td>34.9</td>
<td>24.8</td>
<td>2898</td>
</tr>
<tr>
<td>CenterTrackV2</td>
<td>CH+MOT17</td>
<td>✓</td>
<td>67.8</td>
<td>67.1</td>
<td>34.6</td>
<td>24.6</td>
<td>2583</td>
</tr>
<tr>
<td>SOTMOT</td>
<td>MIX</td>
<td>✓</td>
<td>71.0</td>
<td>70.9</td>
<td>42.7</td>
<td>15.3</td>
<td>5184</td>
</tr>
<tr>
<td>FairMOT</td>
<td>CH+MIX</td>
<td>✓</td>
<td>73.7</td>
<td>72.3</td>
<td>44.7</td>
<td>15.9</td>
<td>3303</td>
</tr>
<tr>
<td><strong>Ours</strong></td>
<td>CH</td>
<td>×</td>
<td>74.2</td>
<td>72.4</td>
<td>46.6</td>
<td>12.2</td>
<td>2748</td>
</tr>
</tbody>
</table>

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.

### 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.

#### 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].

#### 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


[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


