Id-Free Person Similarity Learning

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

1. Image-level transformation

Here we detail the specific image-level transformations used in our similarity learning framework.

1. Image-level rotation (`Rotate`). We rotate the image with a degree that is randomly drawn from $[-10°, 10°]$
2. Random patch erase for each person box (`Occlusion`). We randomly sample a image patches within a person box that is larger than 64×64, and substitute the value of pixels in corresponding image regions with global mean pixel value. The maximum region of the random image patch is 40% of the region for the corresponding person box.
3. Motion blur and JPEG compression (`Video jitter`). We apply augmentations of motion blur and JPEG compression to synthesize video-level artifacts. We use the implementation from the open-source toolbox in `imgaug` [2].
4. Color augmentation (`Color jitter`). We randomly shift the values of each pixel by 10% for brightness, contrast and saturation channels. We use the implementation `ColorJitter` class from torchvision [1].
5. Image mirror (`Mirror`). We create the mirror of an image by flipping it w.r.t its horizontal axis.
6. Camera Zoom-in motion (`Zoom-in`). We synthesize the camera zoom-in motion effect by firstly cropping an image region and then resizing it to the resolution of the original image. In this process, we randomly sample the top-left (x_0, y_0) and bottom-right (x_1, y_1) coordinates of the cropped image region with the following equations: $x_0 = w * r_0, y_0 = y * r_1, x1 = w * (1 - r_2), y1 = h * (1 - r_3)$, in which w, h are the width and height of original image respectively, and r_i is a scalar that is randomly drawn from $[0, 0.3]$.

The visual effects of each transformation are shown in Fig. 1. The source codes would be released soon.

[1] https://github.com/pytorch/vision

Loss	data	MAP	Top-1
Cross Entropy	part	54.4%	55.7%
Contrastive (Memory)	part	73.2%	75.0%
Contrastive	part	75.5%	77.8%

Table 1. Result comparison of models trained with different losses on a subset of COCO and CrowdHuman datasets, which includes 50,000 person boxes in total. The results are based on CUHK-SYSU [8] dataset.

2. Details of Model Training with memory-based contrastive loss

In the case that the identities of each person boxes are available, we elaborated in Sec.4.1 of the main paper that a memory bank $\mathbf{m} \in \mathbb{R}^{d \times \mathtt{M}}$ can be used to construct the positive / negative feature set $\mathbf{P}_i$, $\mathbf{N}_i$ for person i. In order to manage such a memory bank, we update it with momentum during network backward pass with the following equation: $\mathbf{m}_i^{new} = \eta \cdot \mathbf{m}_i^{old} + (1 - \eta) \cdot \frac{1}{|\mathbf{f}_i|} \sum_{\mathbf{f} \in \mathbf{f}_i} \mathbf{f}$, in which $\eta = 0.5$ is the momentum and $\mathbf{f}_i$ represents all feature vectors that correspond to person i in the current training batch.

Training on Person Search Dataset. In order to train a PointID network on CUHK-SYSU [8] / PRW [10] with full supervision, we adopt the contrastive similarity loss and manage the memory for all ids as above. In addition, as there are bounding boxes that are not annotated with identities, they are however useful in similarity learning [8]. To this end, we manage a fixed-size queue $\mathbf{f}^- \in \mathbb{R}^{256 \times 50{,}000}$ that is made up with the latest encountered $50,000$ feature vectors that corresponds to those person boxes. We append $\mathbf{f}^-$ to existing $\mathbf{N}_i$ that is constructed from $\mathbf{m}$.

3. Loss comparison

In Sec. 5.2 and Tab. 2(a) of the main paper, we compare how the adopted loss (cross entropy loss vs contrastive loss) affects the person embedding learning. We also compare memory-based contrastive loss with the above two alternatives. This is interesting as the memory-based contrastive loss is widely used in person search dataset where manual

(a) Original image. (b) **Rotate**. (c) **Occlusion**. (d) **Video jitter**. (e) **Color jitter**. (f) **Mirror**. (g) **Zoom-in**.

Figure 1. Visual effects of each image transformation.

person identities are available and the same person appears in different images. Similar to cross entropy loss, memory-based contrastive loss is not scalable to large number of identities in our training framework. This is because each person only appears a single time in a training epoch, thus managing an up-to-date memory bank is almost impossible when the number of iterations is large to go through a training epoch. [2]

As shown in Tab. 1, the model trained with memory-based contrastive loss outperforms the model trained with cross entropy loss, underpinning the effectiveness of contrastive loss. However, it lags behind the model trained with non-memory based contrastive loss by a noticeable margin. We conjecture it is the out-of-sync memory that largely contributes to the performance gap.

Why not use non-memory based contrastive loss on Person search dataset? This is a reasonable question to ask after observing the consistent benefits of non-memory based contrastive loss in both image training (Tab. 1) and video training (Tab. 2(a) in the main paper). In our training framework, we can easily generate an image pair that include the same set of people either from image-level transformation or from video-level frame sampling, which is the core of non-memory based contrastive loss. However, it's not straightforward to generate such image pairs during training on either CUHK-SYSU [8] or PRW [10] dataset. Moreover, as the number of person identities is relatively small (5532 in CUHK-SYSU and 483 in PRW) and more importantly each person appear in multiple images. Therefore the memory slice $\mathbf{m}_i$ can be updated multiple times in a training epoch, so the memory is relatively update-to-date when person i is encountered during training. This explains the effectiveness of memory-based contrastive learning on person search dataset.

4. Kinetics-150K

Pseudo label generation. As elaborated in the main text of the paper, we first extract person boxes and their embeddings by using the person detection and embedding model trained on COCO and CrowdHuman. Next, we use the density-based clustering method DBSCAN [1] to cluster all the embeddings for each video, based on which we derive the unique identity of each detected person box. In this case, each cluster corresponds to a unique person. Although DBSCAN is able to identify "abnormal" embeddings as outliers, we further adopt the following heuristics to automatically filter out low-quality person boxes or clusters: 1), for each cluster, we take the person box that has the highest confidence score for each frame, which ensures that two different person boxes do not have the same identity in the same frame. After this, each cluster corresponds to a unique person trajectory; 2), we further filter out person trajectories that are short (shorter than 50% of temporal length of the video) and low-confident (the mean confidence score over all corresponding person boxes is lower than 0.5). These post-processing steps are important to remove the noise in pseudo labels.

In the supplementary folder, we include a few demo videos that viusalize the pseudo labels (person boxes and their unique identities). Note that only valid person trajectories are visualized in the demo videos, and people without labels in the videos are not used for model fine-tuning.

Subset	Number of videos	Number of unique ids
Kinetics-10K	$3,060$	$3,933$
Kinetics-25K	$7,553$	$9,594$
Kinetics-50K	$15,212$	$19,302$
Kinetics-100K	$30,175$	$38,266$
Kinetics-150K	$45,108$	$57,200$

Table 2. Statistics of different subsets of Kinetics-150K that are used in our study.

Statistics of different subsets. In Tab. 2, we provide the detailed statistics of different subsets of Kinetics-150K, which consists of 150,000 videos randomly sampled from Kinetics-700 dataset [3]. In general, around 30% of videos include at least 1 valid person trajectory in each subset. To make sure that other researchers have the access to this data, we will release the list of all videos in Kinetics-150K and

[2] As the number of person boxes (i.e. ids) is usually correlated with the number of training images, it takes more iterations to go through a training epoch when there are more person boxes (or ids) during training.

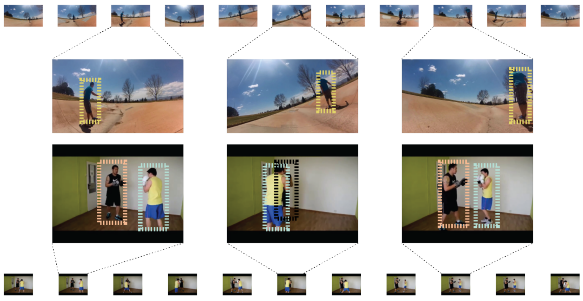

Figure 2. Pseudo-labels are visualized for three random frames of each video, in which the same color-coded boxes correspond to the same person, and the black boxes correspond to the person with "unsure" identity (i.e. outlier during clustering).

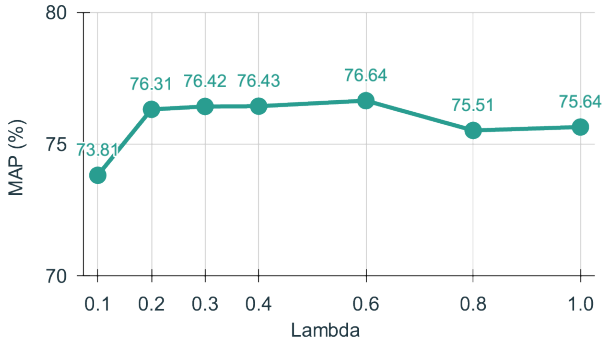

Figure 3. Results of the image-trained model trained with the same multi-task loss with varying λ. The results are evaluated on CUHK-SYSU dataset [8].

their pseudo labels upon the acceptance of this work.

Visual examples. In Fig. 2, we show visual examples of un-labeled videos and their corresponding pseudo-labels (person bounding boxes and their unique identities) that we use to train our model.

5. Effect of λ in multi-task loss.

As elaborated in Sec 3 in the main paper, our model is trained with a multi-task loss $\ell_{total} = \ell_{det} + \lambda\ell_{id}$. In Fig. 3, we show how λ affects the model training. In general, the model achieves robustly well results when λ is within $[0.2, 0.6]$. We also adopt the loss similar to JDE [7] and FairMOT [9] that theoretically learns an adaptive λ by balancing those two tasks. Concretely, $\ell_{total} = \frac{1}{2}(\frac{1}{e^{w_1}}\ell_{det} + \frac{1}{e^{w_2}}\ell_{id}) + w_1 + w_2$. With this loss, the model achieves slightly lower performance (75.8% MAP) on CUHK-SYSU dataset.

6. Qualitative result comparison

In Fig. 4, we compare the qualitative results of three different models: the first one is trained only on COCO [4] and CrowdHuman dataset [6] (`image model`), the second one is further fine-tuned on Kinetics-150K (`video model`) and the third one is trained with full id annotation on the target dataset (`fully supervised model`). Note that the first two models do not have access to manual labeled identities during training. To recap, those three models achieve 78.0% , 84.6% and 93.7% top-1 matching accuracy on CUHK-SYSU [8] respectively. Overall, we have the following observations.

Both `image` and `video models` perform decently well. As shown in the first two rows in Fig. 4, the `image model` is able to match to the right person even if the appearance of the person changes modestly. Meanwhile, the `video model` is more likely to match to a different view of the same person, which indicates that the model is trained with examples (i.e. un-labeled videos) in which the same person exhibit large appearance diversity.

The `image model` fails on challenging scenarios. As shown in the third and fourth rows in Fig. 4, the `image model` fails to match to the right person when the appearance of the person changes relatively large. In particular, the failure case in the third example suggests that the `image model` encodes the distinctive human pose / viewpoint in the person embedding. Those features are not identity related but can be used as a shortcut in the person instance discrimination task, thus limiting the effectiveness of person embeddings. By fine-tuning the model on un-labeled videos in which the same person's appearance can be drastically different, the `video model` is more robust to such appearance changes.

The final example in Fig. 4 illustrates a hard case for person re-identification as the query person is occluded and moreover the viewpoint of the matched person changes significantly in the gallery image. As shown, both `image` and `video models` fail to re-identify the right person. We believe this is where the future work needs to focus on either through automatically identifying hard examples so that only those hard examples are manually annotated or through improving the existing id-free person similarity learning framework such that the person embedding model is more robust to those challenging cases.

7. Detailed Results on MOT17 [5]

Model Training and Inference. We train our model on CrowdHuman [6] dataset with the provided full body box annotations. All training configurations are the same as that in Sec.5. As there are multiple small-size people in MOT17 [5], we resize the video frame to have 900×1500 pixels during inference.

We use the online solver implemented by FairMOT [9] to generate trajectories for each person in the video. We use

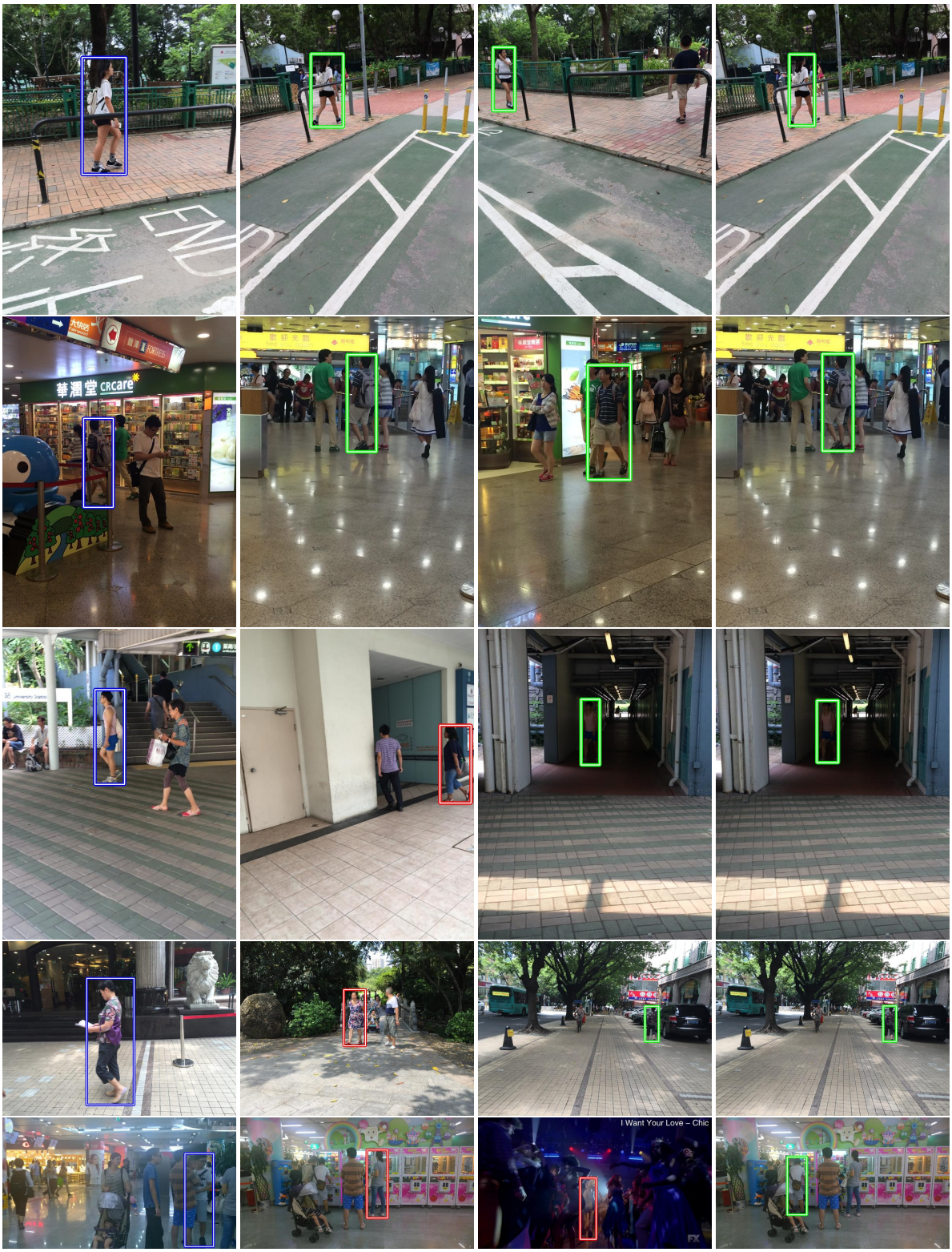

Figure 4. We visualize Top-1 matching results of different models. In detail, the 1st column visualizes the query person (blue box), and the 2nd, 3rd and 4th columns viusalize the top-1 matched person for `image model`, `video model` and `fully-supervised model` respectively. The green box denotes that the matched box is correct, and red box indicates otherwise. Note that both `image` and `video models` do not have access to manual labeled identities during training. Examples are from CUHK-SYSU dataset [8].

Sequence	MOTA↑	IDF1↑	MT↑	ML↓	FP↓	FN ↓	IDsw↓
MOT17-01	55.0	48.3	37.5%	29.2%	403	2463	37
MOT17-03	85.7	82.8	77.0%	6.80%	3296	11531	156
MOT17-06	61.7	62.4	42.8%	11.3%	1289	3063	163
MOT17-07	70.1	61.0	58.3%	5.00%	1103	3821	122
MOT17-08	54.8	45.4	29.0%	11.8%	1254	8015	283
MOT17-12	62.1	70.6	56.1%	7.70%	1159	2089	41
MOT17-14	55.3	63.7	24.4%	21.3%	740	7407	114
All	74.2	72.4	46.6%	12.2%	27732	115167	2748

Table 3. Detailed result summary on MOT17 test videos. Note that each video appears three times, so the values for accumulated metrics (FP, FN, IDsw) of **All** videos are $3\times$ of the accumulated values over 7 videos in the Table. Our model is only trained on CrowdHuman dataset [6].

the default setting except that a new trajectory is spawned when the confidence of an un-matched detection is larger than 0.6. We use "private detection" protocol during inference, and we report the detailed results on 7 MOT17 test sequence in Tab. 3.

References

[1] 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

[2] Alexander B. Jung, Kentaro Wada, Jon Crall, Satoshi Tanaka, Jake Graving, Christoph Reinders, Sarthak Yadav, Joy Banerjee, Gábor Vecsei, Adam Kraft, Zheng Rui, Jirka Borovec, Christian Vallentin, Semen Zhydenko, Kilian Pfeiffer, Ben Cook, Ismael Fernández, François-Michel De Rainville, Chi-Hung Weng, Abner Ayala-Acevedo, Raphael Meudec, Matias Laporte, et al. imgaug. https://github.com/aleju/imgaug, 2020. Online; accessed 01-Feb-2020. 1

[3] 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. 2

[4] 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. 3

[5] 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. 3

[6] 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. 3, 5

[7] 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. 3

[8] 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, 4

[9] 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. 3

[10] 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, 2