LEARNING FROM DANCES: POSE-INVARIANT RE-IDENTIFICATION FOR MULTI-PERSON TRACKING

Hsuan-I Ho[1,3] *Minho Shim*[2] *Dongyoon Wee*[3]

[1]Department of Computer Science, ETH Zürich
[2]*Seoul, South Korea*
[3]Clova AI, NAVER Corp.

ABSTRACT

Most existing multi-person tracking approaches rely on appearance based re-identification (re-ID) to resolve fragmented tracklets. However, simply using appearance information could be insufficient for videos containing severe pose changes, such as sports or dance videos. With the goal of learning pose-invariant representations, we propose an end-to-end deep learning framework *Sparse-Temporal ReID Network*. Our proposed network not only realizes human pose disentanglement in an image recovery manner, but also makes efficient linkages between the identical subjects via a unique *Sparse temporal identity sampling* technique across time steps. Experimental results demonstrate the effectiveness of our proposed method on both multi-view re-ID benchmarks and our newly collected dance video dataset *DanceReID*[1].

Index Terms— Pose-Invariant Features, Person Re-Identification, Multi-Person Tracking, Deep Learning

1. INTRODUCTION

Multi-person tracking [1] is a popular topic which benefits a variety of tasks like surveillance, autonomous vehicle, or content creation. In recent years, many attempts have been made to perform tracking on various video content. Among all, tracking fast moving subjects in a crowded scene, such as sports and dance videos, is considered important for improving user viewing experiences.

While the state-of-the-art tracking framework [2] has shown impressive performance on the benchmark [3] in light of convolutional neural network (CNN) based object detectors [4], it remains a challenging problem to tackle with losing track of occluded targets. A straightforward yet efficient strategy to handle occlusion is called *Buffer-and-Recovery* [5], which buffers observations when the tracking state becomes ambiguous and tries to re-identify targets by matching their appearance with buffers. However, this strategy might not be reliable for videos with severe pose changes, since the appearance of the subject might become very different due to movements or changing of poses (see Fig. 1).

[1]Our codes and dataset are now available at https://git.io/Jvcg4

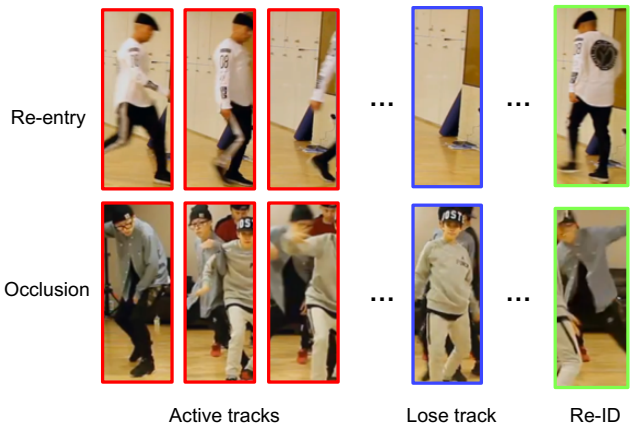

Fig. 1. Example of re-ID for multi-person tracking. Note that red bounding boxes indicate the active tracks to be stored in buffers while blue bounding boxes represent losing track due to occlusion or vanishing. Our goal is to match the new detections (in green bounding boxes) with the targets in buffers while their appearances could be very different.

On the other hand, multi-view person re-identification (re-ID) [6], which aims at matching the same identities among images captured from different viewpoints, is closely related to the aforementioned strategy. A vast number of approaches [7, 8, 9] have been proposed to tackle with occlusion or appearance changes for person re-ID. However, sports and dance videos would exhibit very unique content when comparing to those of pedestrian data, which capture only the extrinsic pose differences (caused by cameras) within a space, but not the intrinsic and diverse ones (caused by human itself) over time. Thus, it would be a challenging task to adapt such re-ID approaches to multi-person tracking. That is, we not only want to learn a feature space describing the observed samples like multi-view re-ID but also need to link corresponding detections that have distinct poses across time steps.

As depicted in Fig. 1, to apply person re-ID to tracking dance videos, one needs to extract pose-invariant information from targets before and after occlusion/vanishing. With the

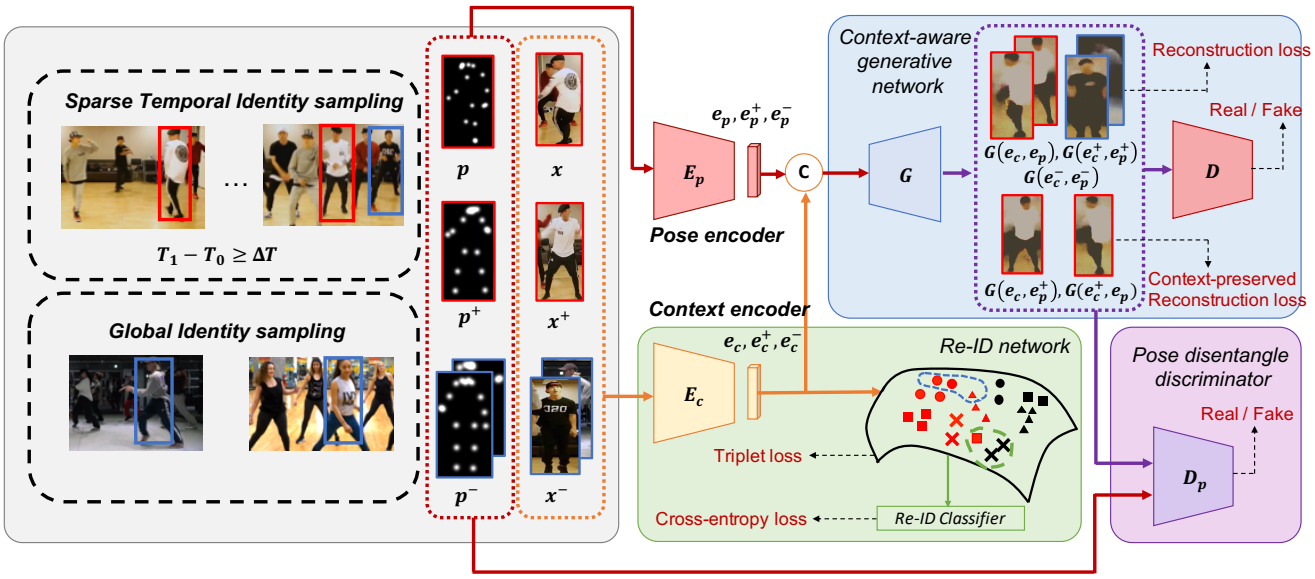

Fig. 2. **Overview of our Sparse-Temporal ReID Network.** Our Sparse-Temporal ReID Network consists of three main components including a context encoder (in green) for learning re-ID features, a context-aware generative network (in blue) for image recovery, and a discriminator (in purple) which disentangles pose information from images.

recent success of pose-guided feature representations learning [10, 11], the architectures of generative neural networks have also been exploited. However, they do not explicitly address the issue of losing tracks in videos, plus a multi-stage training procedure is needed. To mitigate the above limitations, we propose an end-to-end deep learning framework *Sparse-Temporal ReID Network*. To be detailed, our network architecture disentangles human pose information from the training images in a unique image reconstruction scheme, while jointly learns a pose-invariant feature embedding for re-ID across video sequences. To further tackle with losing of tracking in videos, an efficient *Sparse-Temporal Identity sampling* technique is performed to ensure the affinity between the features of the identical IDs across sparse video time steps.

In summary, our contributions are highlighted as follows: 1) We propose a novel end-to-end trainable network which could learn pose-invariant human representations in a unique context-aware reconstruction scenario. 2) Unlike the existing multi-view re-ID methods, we explicitly resolve the issue of losing tracks by learning pose-invariant features with re-ID guarantees from the sparse time samples. 3) We newly collect a large scale benchmark dataset *DanceReID* to evaluate re-ID in multi-person tracking. Compared to the mutli-view pedestrian datasets, our dataset contains more diverse poses and captures human pose changes over time.

2. MODELS AND METHODS

Given a dataset $\mathcal{D}$ consisting of human RGB image corps and identity labels pairs $(x, y) \in \mathcal{D}$, we aim to learn a feature embedding where the distances between an unseen query and any samples with an identical label (i.e., positive samples) are always smaller than the distances with those distinct ones (i.e., negative samples). To further achieve pose disentanglement from input images, additional pose information is exploited in the training phase. We first extract 17 human pose landmarks by PyraNet [12] from the input image x and convert such landmarks into a 17-channel human pose map p.

Fig. 2 illustrates our entire pipeline: A training tuple consisting of a positive sample x^+ and a negative sample x^- is first selected via *Sparse Temporal Identity sampling* strategy. Given a training image x and its corresponding pose map p, a pair of encoders E_c and E_p then return the context and pose features e_c and e_p respectively. As guided by the pose feature, the image generator G would recover desirable outputs, while pose-invariant information for re-ID is well-preserved in the context feature. Properties of each component will be further discussed in the following subsections.

2.1. Sparse temporal identity sampling

The goal for multi-person tracking re-ID is to learn pose-invariant features from effective pairs across time steps. Most existing methods rely on either drawing samples randomly [10] or mining for batch-hard samples [9]. However, such methods could be insufficient for long video sequences due to their repetitive content. To this end, we propose an efficient *Sparse Temporal Identity sampling* method based on observations in human dance videos: 1) Positive samples only appear in the same video sequence while negative samples can be selected from arbitrary videos. Furthermore, nega-

tive samples from the same video sequence and time step are likely to share similar background, clothes, and poses with the anchor, which can be served as hard-negative samples during training. 2) Subjects within a small period ΔT might share too similar appearances and poses. Thus, sparsely drawing the positive pairs would help learning of pose-invariant features by observing a diversity of human poses.

By deploying the above strategy, one could obtain very efficient training pairs for learning pose-invariant features across time. We note that under our training scenario, a small ratio of negative samples will still be drawn from distinct videos (i.e., Global Identity sampling in Fig. 2) to make sure that features across videos are distinguishable.

2.2. Learning of pose-invariant features

As shown in Fig. 2, our proposed Sparse-Temporal ReID Network is a combination of multiple neural network components: re-ID network, context-aware generative network, and pose disentangle discriminator.

Re-ID network: The context encoder E_c based on ResNet 50 [13] architecture would extract human context features for the re-ID purpose. To introduce supervised signals in the learning process, we jointly optimize the triplet loss and the cross-entropy loss. To be specific, given a training tuple $\left(x, x^+, x^-\right)$, the re-ID loss can be calculated as:

$$\mathcal{L}_{ID} = \max\{\mathcal{M} - d_{l2}(e_c, e_c^-) + d_{l2}(e_c, e_c^+), 0\} + \mathcal{L}_{CE}, \quad (1)$$

where $d_{l2}(e, e')$ returns the L2 norm of a pair of features, and $\mathcal{M}$ denotes the margin for metric learning. The $\mathcal{L}_{CE}$ term is the cross-entropy loss for the re-ID classifier.

Context-aware generative network: While the context encoder E_c performs feature extraction to the input image, we adopt a unique image reconstruction scheme to ensure only pose-invariant information would be captured. The image generator G would try to recover the input image from the context feature conditioned on additional pose information. For instance, the input image x should be recovered by e_c and e_p. As such, the image reconstruction loss for a training tuple can be computed as:

$$\begin{aligned}\mathcal{L}_R(G) = &\left\|G(e_c, e_p) - x\right\|_1 + \left\|G(e_c^+, e_p^+) - x^+\right\|_1 + \\ &\left\|G(e_c^-, e_p^-) - x^-\right\|_1.\end{aligned} \quad (2)$$

Additionally, we also expect the context feature from the same ID could also recover other images given different pose features. Thus we apply another context-aware reconstruction loss by swapping the context features from the positive pair:

$$\mathcal{L}_{CR}(G) = \left\|G(e_c^+, e_p) - x\right\|_1 + \left\|G(e_c, e_p^+) - x^+\right\|_1, \quad (3)$$

Finally, the image discriminator D is deployed to ensure the generator produce perceptually realistic images. That is, the image generator tries to fool the discriminator by producing realistic recoveries from the features while the discriminator aims to distinguish the real/recovered ones. The adversarial loss of the discriminator D is thus written as

$$\begin{aligned}\mathcal{L}_D(G, D) = &\mathbb{E}_{x\sim\mathcal{D}}[\log D(x)] + \\ &\mathbb{E}_{x,p\sim\mathcal{D}}[\log(1 - D(G(e_c, e_p)))].\end{aligned} \quad (4)$$

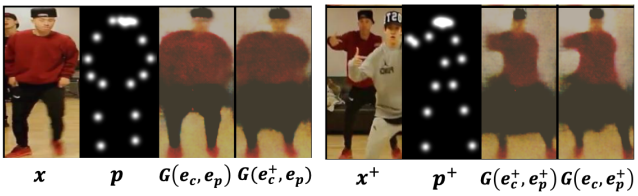

Fig. 3. Visualization of image recovery results. It can be seen that our method preservers only context information for the re-ID purpose and thus the influence of occlusion can be removed.

Pose disentangle discriminator: Despite the use of the above context-aware generator for recovering positive image pairs, there is no guarantee that the pose encoder Ep would disentangle pose information from the corresponding images. To achieve this goal, we further employ a pose discriminator D_p based on PatchGAN structure [14] as suggested in [10]. D_p is designed to discriminate real/recovered images conditioned on pose maps, and its loss function is formulated as follow:

$$\begin{aligned}\mathcal{L}_P(G, D_p) = &\mathbb{E}_{x,p\sim\mathcal{D}}[\log D_p(p, x)] \\ &+ \mathbb{E}_{x,p\sim\mathcal{D}}[\log(1 - D_p(p, G(e_c, e_p)))] \\ &+ \mathbb{E}_{x,p\sim\mathcal{D}}[\log(1 - D_p(p^+, x)))].\end{aligned} \quad (5)$$

It is worth noting that our network allows end-to-end training by updating overall loss with weighting factors of 1.0, 10.0, and 0.1 for $\mathcal{L}_{ID}$, $\left(\mathcal{L}_R + \mathcal{L}_{CR}\right)$ and $\left(\mathcal{L}_D + \mathcal{L}_P\right)$, respectively. Besides, only RGB images are utilized in the inference stage since our model has learned to capture pose-invariant information with pose-guided training.

3. DATASETS

In this section, we describe the datasets (including our proposed one) we consider for experiments. *Market-1501* [15] and *DukeMTMC-reID* [16] are two large scale multi-view pedistrain datasets that have been commonly used in most re-ID works. Market-1501 consists of 1501 IDs captured from 6 camera views in total 32668 images, while DukeMTMC-reID contains 36411 images of 1404 IDs from 8 cameras.

Nevertheless, these multi-view pedestrian datasets are still far from ideal for our purpose (i.e., to re-identify moving subjects when losing tracks in multi-person tracking) since they only recorded identities in different viewpoints but not intrinsic pose changes across time steps. To this end, we collect a

Methods	DanceReID		
	mAP	top-1	top-5
Baseline (ResNet-50)	43.2	43.0	74.1
ResNet-50 + Softmax [7]	74.4	73.1	94.7
ResNet-50 + Siamese [8]	77.5	75.9	96.9
ResNet-50 + Triplet [9]	78.4	77.2	97.6
FD-GAN [10]	80.9	79.2	97.5
Triplet + Softmax	78.7	77.5	97.6
Ours w/o $\mathcal{L}_{CR}$	83.6	82.1	98.0
Ours w/ random sample	84.1	82.9	98.3
Ours w/o $\mathcal{L}_{D}$	84.4	83.1	98.5
Ours w/o $\mathcal{L}_{P}$	84.5	83.3	98.4
Ours	**86.1**	**84.9**	**98.7**

Table 1. Performance comparisons and ablation studies on our proposed DanceReID dataset. The top-1 and the top-5 scores are calculated following the CMC-CUHK03 protocol.

new multi-person tracking re-ID dataset *DanceReID* containing various categories of dance videos from YouTube Creative Commons. We first perform human detection via the YOLO detector [17] and a tracking procedure based on PoseFlow [18]. The resulting bounding-boxes are then assigned with pseudo-labels by training one classifier per video. After manually filtering out label misalignment, our dataset contains 86238 frames and 346K detections in a total of 100 IDs. In our experiments, the dataset is downsampled and split into the training/test set with 41242/21356 images of 71/29 IDs.

4. EXPERIMENT

4.1. Experimental setup

We follow the standard evaluation protocol of re-ID to calculate the cumulative matching curve (CMC) and mean average precision (mAP) based on the original dataset setting. As for the evaluation on our DanceReID, we adopt the "single gallery shot" protocol (CMC-CUHK03 [19]) due to the large number of query images in videos. We note that only the ResNet features will be utilized in the inference stage for fair comparisons since in [10] (denoted as FD-GAN‡) a two-stage evaluation is performed, which requires additional embedding features and a re-rank [20] technique. Additional model implementation details can be found in our source code repository.

4.2. Comparison and discussion

Table 1 summarizes the quantitative results of our proposed method against baselines and state-of-the-art methods on DanceReID. The results show that pose-guided training methods (the use of pose and context encoder) basically outperformed single model methods (ResNet-50 + re-ID losses).

	Methods	Market-1501		DukeMTMC	
		mAP	top-1	mAP	top-1
w/ Re-rank	Siamese‡	72.5	88.2	61.3	78.2
	FD-GAN‡	77.7	90.5	64.5	80.0
w/o Re-rank	Softmax [7]	59.8	81.4	40.7	62.5
	Siamese [8]	69.7	86.8	58.9	75.9
	Triplet [9]	67.9	85.1	54.6	73.1
	FD-GAN [10]	76.3	90.4	62.9	**79.4**
	Ours	**78.2**	**91.9**	**64.6**	79.1

Table 2. Comparisons on the multi-view pedestrian benchmarks. We note that these datasets do not capture temporal pose changes as ours.

While the state-of-the-art FD-GAN [10] could achieve better performance with the guidance of pose map, their method is not favorable to multi-person tracking plus a 3-stage training procedure is needed. Our full model not only outperformed their method by a large margin, but it was also well trained in an *end-to-end* manner. We also show comparisons on the multi-view re-ID datasets in Table 2 (although they are not originally created for the multi-person tracking purpose). We note that since both datasets do not capture intrinsic pose changes over time as ours, and also the occlusion issue is not as severe as in the dance videos. However, it can be seen that our method still produced comparable results even though Sparse Temporal Identity sampling is inapplicable under such dataset settings.

We also perform an ablation study to analyze each design choice of our model under the end-to-end training scheme in Table 1. The use of Sparse Temporal Identity sampling and pose-guided image recovery both provided a significant performance gain for our model. A further improvement could be yielded by enforcing the discriminators for adversarial learning, which also produced more realistic recovered images as shown in Fig. 3. Thus, our model design and integration of the above components are desirable for multi-person tracking re-ID.

5. CONCLUSION

In this work, we propose *Sparse-Temporal ReID Network* to tackle re-ID of subjects for multi-person tracking. Our network uniquely learns pose-invariant features in an end-to-end image recovery fashion. Moreover, we adopt a sampling strategy which allows the mining of time-sparse hard samples more efficiently. In addition to the evaluation on the multi-view re-ID benchmarks, we additionally collect a new pose-rich dance video dataset. Experimental results confirmed the effectiveness and robustness of our proposed framework.

Acknowledgments. We would like to specially acknowledge NAVER Smart Machine Learning (NSML) [21] platform for the support of the computational platform.

6. REFERENCES

[1] Laura Leal-Taixé, Anton Milan, Konrad Schindler, Daniel Cremers, Ian Reid, and Stefan Roth, "Tracking the trackers: an analysis of the state of the art in multiple object tracking," *CoRR*, vol. abs/1704.02781, 2017.

[2] Philipp Bergmann, Tim Meinhardt, and Laura Leal-Taixe, "Tracking without bells and whistles," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2019.

[3] Anton Milan, Laura Leal-Taixé, Ian Reid, Stefan Roth, and Konrad Schindler, "Mot16: A benchmark for multi-object tracking," *CoRR*, vol. abs/1603.00831, 2016.

[4] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," in *Advances in Neural Information Processing Systems (NIPS)*, 2015.

[5] Wenhan Luo, Junliang Xing, Anton Milan, Xiaoqin Zhang, Wei Liu, Xiaowei Zhao, and Tae-Kyun Kim, "Multiple object tracking: A literature review," *CoRR*, vol. abs/1409.7618, 2014.

[6] Liang Zheng, Yi Yang, and Alexander G Hauptmann, "Person re-identification: Past, present and future," *CoRR*, vol. abs/1610.02984, 2016.

[7] Tong Xiao, Hongsheng Li, Wanli Ouyang, and Xiaogang Wang, "Learning deep feature representations with domain guided dropout for person re-identification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[8] Dahjung Chung, Khalid Tahboub, and Edward J Delp, "A two stream siamese convolutional neural network for person re-identification," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017.

[9] Alexander Hermans, Lucas Beyer, and Bastian Leibe, "In defense of the triplet loss for person re-identification," *CoRR*, vol. abs/1703.07737, 2017.

[10] Yixiao Ge, Zhuowan Li, Haiyu Zhao, Guojun Yin, Shuai Yi, Xiaogang Wang, et al., "Fd-gan: Pose-guided feature distilling gan for robust person re-identification," in *Advances in Neural Information Processing Systems (NIPS)*, 2018.

[11] Luan Tran, Xi Yin, and Xiaoming Liu, "Disentangled representation learning gan for pose-invariant face recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[12] Wei Yang, Shuang Li, Wanli Ouyang, Hongsheng Li, and Xiaogang Wang, "Learning feature pyramids for human pose estimation," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017.

[13] 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 (CVPR)*, 2016.

[14] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros, "Image-to-image translation with conditional adversarial networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[15] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian, "Scalable person re-identification: A benchmark," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[16] Ergys Ristani, Francesco Solera, Roger Zou, Rita Cucchiara, and Carlo Tomasi, "Performance measures and a data set for multi-target, multi-camera tracking," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2016.

[17] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi, "You only look once: Unified, real-time object detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[18] Yuliang Xiu, Jiefeng Li, Haoyu Wang, Yinghong Fang, and Cewu Lu, "Pose Flow: Efficient online pose tracking," in *Proceedings of the British Machine Vision Conference (BMVC)*, 2018.

[19] Wei Li, Rui Zhao, Tong Xiao, and Xiaogang Wang, "Deepreid: Deep filter pairing neural network for person re-identification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014.

[20] Zhun Zhong, Liang Zheng, Donglin Cao, and Shaozi Li, "Re-ranking person re-identification with k-reciprocal encoding," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[21] Hanjoo Kim, Minkyu Kim, Dongjoo Seo, Jinwoong Kim, Heungseok Park, Soeun Park, Hyunwoo Jo, KyungHyun Kim, Youngil Yang, Youngkwan Kim, et al., "Nsml: Meet the mlaas platform with a real-world case study," *CoRR*, vol. abs/1810.09957, 2018.