Exploiting Temporal Coherence for Self-Supervised One-shot Video Re-identification (Supplementary materials)

Dripta S. Raychaudhuri and Amit K. Roy-Chowdhury

University of California, Riverside, CA 92521, USA
{draychaudhuri,amitrc}@ece.ucr.edu

1 Evaluation on few-example setting

Our method can be extended to the few example setting very easily, by acquiring more labeled samples before training. Thus, the labeled dataset may contain more than one tracklet per identity. We report the performance of our framework with varying ratios of labeled data in Table 1.

Table 1. Comparison to the state-of-the-art supervised methods on MARS. We report performance in the semi-supervised (few-example) setting. The number in the bracket indicates the percentage of used labeled training data.

Type	Method	R-1	R-5	R-20	mAP
Supervised	ResNet50-3D [2]	82.9	93.7	96.8	76.2
	IDTriplet [4]	79.8	91.4	-	67.7
	Baseline (100%)	80.8	92.1	96.1	67.4
Semi-supervised	Ours (10%)	72.0	85.3	91.4	56.5
	Ours (20%)	78.2	89.9	94.4	64.4

On the MARS dataset, using only 20% of the training data as the labeled set, our method achieves 78.2% Rank-1 accuracy and 64.4% mAP, which is very close to the fully supervised methods which utilize the entire training data with labels. Although this setting requires more annotations than the one-shot task, it can easily achieve competitive results compared to the supervised methods.

2 Initial selection of tracklets

We choose the labeled tracklets in a manner identical to the previous works [10, 6]. More importantly, our method is designed to be robust to the selection of the labeled set - this is an advantage of our consistency losses, which promote discriminative feature learning regardless of labels. This robust behavior is demonstrated in Table 2

Table 2. Results on Duke for $p = 0.2$ across two random selections of the labeled set

Split	R-1	mAP
1	74.7	65.5
2	74.4	65.2

3 Analysis on range parameter r

The range parameter r plays an important role in the context of the inter-sequence consistency criterion. Choosing r too high will lead to sampling of easy negatives, which will not contribute too much to learning (zero gradients). On the flip side, choosing r too low can lead to positives being interpreted as negatives, which can hamper learning. We demonstrate this behavior in Table 3.

Table 3. Performance on MARS for $\mathcal{L}_{\text{inter}}$ and $p = 0.20$ as the range parameter r is varied.

r	R-1	mAP
1	53.1	30.2
2	53.6	30.6
3	50.8	27.9

4 Additional results

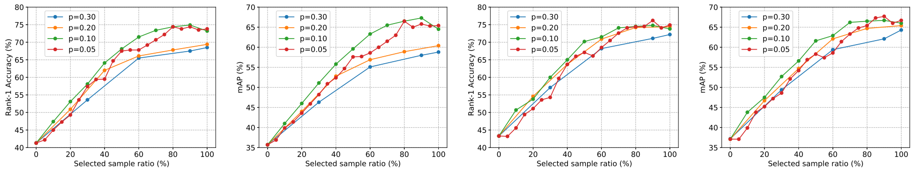
In Table 4, we present the case when the enlarging factor $p = 0.30$. This indicates a very aggressive incorporation of pseudo-labels and increases the chance for erroneous label estimation. However, even in this case TCPL is able to perform better than competing methods. Especially on the DukeMTMC-VideoReID dataset, TCPL outperforms [9] in mAP by **8**%. In Figure 1, we present the learning curves on DukeMTMC-VideoReID.

5 Dataset overview

The **MARS** dataset [11] is the largest video person re-identification dataset for the person and was collected in a university campus. The dataset contains 17503 tracklets for 1261 identities and 3248 distractor tracklets, which are captured by six cameras. The dataset is split into 625 identities for training and 636 identities for testing. Every identity in the training set has approximately 13 video tracklets on average and 800 frames on average. The bounding boxes are detected and tracked using the Deformable Part Model (DPM) and GMMCP tracker.

Table 4. One-Shot Performance for the enlarging parameter $p = 0.30$. The best and second best results are in red/blue respectively.

Dataset	Setting	Methods	R-1	R-5	R-20	mAP
DukeMTMC	$p = 0.30$	EUG [10]	63.8	78.6	87.0	54.6
		One-Shot Progressive [9]	66.1	79.8	88.3	56.3
		TCPL -$\mathcal{L}_{\text{intra}}$	72.2	83.2	90.3	64.3
		TCPL -$\mathcal{L}_{\text{inter}}$	68.5	80.8	88.6	58.8
MARS	$p = 0.30$	EUG [10]	42.8	56.5	67.2	21.1
		One-Shot Progressive [9]	44.5	58.7	70.6	22.1
		TCPL -$\mathcal{L}_{\text{intra}}$	45.3	57.6	66.7	23.8
		TCPL -$\mathcal{L}_{\text{inter}}$	45.7	59.6	69.3	23.9



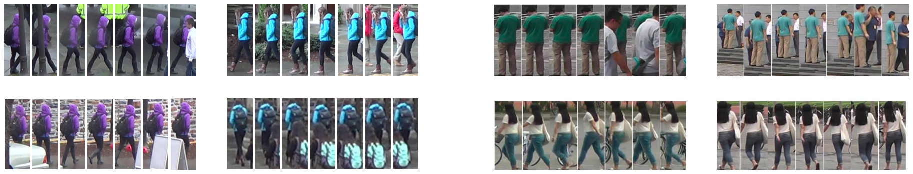
(a) (b) (c) (d)

Fig. 1. Comparison with different values of enlarging factor on DukeMTMC-VideoReID. Figures (a) and (b) represent the Rank-1 accuracy and mAP while using $\mathcal{L}_{\text{inter}}$. Figures (c) and (d) represent the Rank-1 accuracy and mAP while using $\mathcal{L}_{\text{intra}}$.

The **DukeMTMC** dataset [8] was released with the aim of developing multi-camera tracking algorithms. The dataset was captured in outdoor scenes with noisy background and suffers from illumination, pose, and viewpoint change and occlusions. The **DukeMTMC-VideoReID** [10] is a subset of the DukeMTMC dataset created for video re-identification. The dataset is manually annotated and each identity has a singular tracklet under a camera. The dataset contains 702 identities for training, 702 identities for testing, and 408 identities as the distractors. In total there are $369,656$ frames of $2,196$ tracklets for training, and $445,764$ frames of $2,636$ tracklets for testing and distractors.

6 Implementation details

We use PyTorch [7] for all experiments. For our model, we use a ResNet-50 [3] pre-trained on ImageNet [1] - the last classification layer removed and a fully-connected layer with batch normalization [5] and a classification layer are added at the end of the model. We adopt stochastic gradient descent (SGD) with momentum 0.5 and weight decay 0.0005 to optimize the parameters for 70



(a) DukeMTMC-VideoReID (b) MARS

Fig. 2. A total of 8 sample tracklets from the two datasets used in our experiments. Each column represents a distinct individual, with the rows denoting two different views of the same person from two different cameras. We can see that across cameras, the tracklets of the same person vary significantly due to changes in illumination, occlusion etc. Even within a tracklet, the background varies significantly.

epochs, with batch size 16 in each iteration. We set $\lambda = 1$ in for the DukeMTMC-VideoReID dataset and $\lambda = 0.8$ for the MARS dataset (due to the huge disparity in the number of labeled and unlabeled tracklets as a result of fragmentation in MARS). The learning rate is initialized to 0.1. In the last 15 epochs, to stabilize the model training and prevent overfitting, we change the learning rate to 0.01 and set $\lambda = 0$.

References

1. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 248–255 (2009)
2. Gao, J., Nevatia, R.: Revisiting temporal modeling for video-based person reid. arXiv preprint arXiv:1805.02104 (2018)
3. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 770–778 (2016)
4. Hermans, A., Beyer, L., Leibe, B.: In defense of the triplet loss for person re-identification. arXiv preprint arXiv:1703.07737 (2017)
5. Ioffe, S., Szegedy, C.: Batch normalization: Accelerating deep network training by reducing internal covariate shift. arXiv preprint arXiv:1502.03167 (2015)
6. Liu, Z., Wang, D., Lu, H.: Stepwise metric promotion for unsupervised video person re-identification. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 2429–2438 (2017)
7. Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al.: Pytorch: An imperative style, high-performance deep learning library. In: Advances in neural information processing systems. pp. 8026–8037 (2019)
8. Ristani, E., Solera, F., Zou, R., Cucchiara, R., Tomasi, C.: Performance measures and a data set for multi-target, multi-camera tracking. In: Proceedings of the European Conference on Computer Vision. pp. 17–35. Springer (2016)

9. Wu, Y., Lin, Y., Dong, X., Yan, Y., Bian, W., Yang, Y.: Progressive learning for person re-identification with one example. IEEE Transactions on Image Processing **28**(6), 2872–2881 (2019)
10. Wu, Y., Lin, Y., Dong, X., Yan, Y., Ouyang, W., Yang, Y.: Exploit the unknown gradually: One-shot video-based person re-identification by stepwise learning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 5177–5186 (2018)
11. Zheng, L., Bie, Z., Sun, Y., Wang, J., Su, C., Wang, S., Tian, Q.: Mars: A video benchmark for large-scale person re-identification. In: Proceedings of the European Conference on Computer Vision. pp. 868–884. Springer (2016)