Supplementary Material
Camera Alignment and Weighted Contrastive Learning for Domain Adaptation in Video Person ReID

1. Ablation: Impact of Frames per Tracklets

Table 1 analyses the impact on performance of growing the number of frames per tracklet (fpt). A CNN backbone has been trained with the supervised contrastive loss function. Each tracklet is cut into equal size to maintain scalability across a dataset. We can observe that the temporal information remains stable from 4 frames per tracklets, and performances does not differ significantly beyond that point. In this context, setting the number of fpt to the minimal value that provides the best accuracy (4 in our case) is suitable parameter setting for real-time video-based ReID applications. It is less costly than longer tracklets in term of memory consumption.

# fpt	PRID2011		iLIDS-VID		MARS	
	mAP	Rank-1	mAP	Rank-1	mAP	Rank-1
2	87.2	89.6	72.5	78.4	81.8	75.0
4	90.1	92.5	74.0	82.0	84.9	79.8
8	89.9	92.2	74.2	82.0	84.6	78.6
16	90.1	92.9	73.7	82.2	85.1	79.3

Table 1. Accuracy for a supervised training using a different number of frames per tracklets (fpt).

2. Visual Results

Figure 1 displays examples of activation maps [2, 1] for SPCL and our proposed method obtained from three bounding box images, featuring different individuals, in the target domain (ILIDS-VID dataset). The activation maps indicates the regions of interest of the backbone CNN when extract feature representations. The figure shows that our method provides a better localization of the person. Compared to the baseline SPCL, less background information is captured, allowing the model to focus on strong identity-based features.

References

[1] Abhishek Aich, Meng Zheng, Srikrishna Karanam, Terrence Chen, Amit K Roy-Chowdhury, and Ziyan Wu. Spatio-temporal representation factorization for video-based person re-identification. In *ICCV*, 2021. 1

[2] Sergey Zagoruyko and Nikos Komodakis. Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer. *ICLR*, 2017. 1

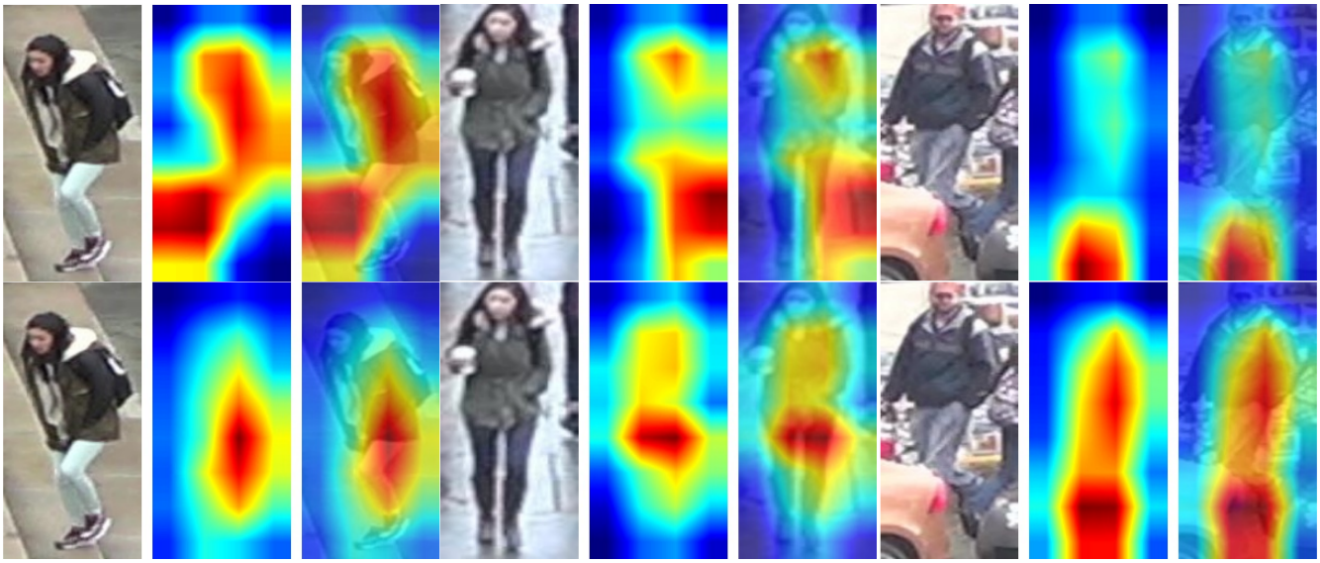

Figure 1. Visualization of images from the target dataset. Top row is from the baseline SPCL and bottom is our proposed approach. The first image represent the sample, the middle is the activation map and the last image is the superposition of the sample with the activation map.