Re-identifying People in Video via Learned Temporal Attention and Multi-modal Foundation Models –Supplementary Material–

Cole Hill[1,2], Florence Yellin[1], Krishna Regmi[1], Dawei Du[1], and Scott McCloskey[1]
[1]Kitware Inc., USA {firstname.lastname}@kitware.com
[2]University of South Florida, USA coleh@usf.edu

Table 1. Comparison with prior work on the MEVID dataset for the Change-of-Location Challenge. The best scores are shown in **bold** whereas the second best scores are underlined.

Method	Same Location					Different Location				
	mAP	Rank				mAP	Rank			
		1	5	10	20		1	5	10	20
BiCnet-TKS [5]	5.1	14.5	25.3	30.6	37.5	4.7	9.4	19.5	24.6	34.3
PiT [10]	12.1	25.0	45.4	53.6	60.5	10.1	19.2	34.0	39.1	49.2
STMN [2]	10.1	16.8	31.0	36.1	43.1	10.0	22.2	41.6	52.1	58.4
AP3D [4]	14.5	31.0	45.1	51.8	63.3	12.0	24.1	37.6	43.4	50.9
TCLNet [6]	20.7	38.8	52.6	60.9	68.8	18.8	33.0	42.4	48.5	55.9
PSTA [8]	20.0	36.8	54.6	63.5	71.4	16.5	28.6	41.4	49.5	57.6
AGRL [9]	18.1	27.6	42.8	48.8	57.6	**22.5**	**41.1**	**57.6**	**64.8**	**70.1**
Attn-CL [7]	16.1	35.1	48.2	54.6	64.9	14.2	26.4	40.7	47.8	55.6
Attn-CL+rerank [7]	23.9	41.5	53.4	58.1	63.9	19.6	33.9	44.7	50.2	55.9
CAL [3]	**24.7**	42.1	56.6	63.2	72.0	22.2	35.0	49.8	55.2	62.0
VCLIP(Ours)	**24.7**	**50.5**	**65.2**	**70.6**	**75.7**	20.9	39.7	52.5	58.6	67.1

1. VCLIP Performance on MEVID Location Difference Challenge

The MEVID [1] dataset provides several challenges to evaluate Re-ID algorithms. In Tab. 1 we compare the performance of our method, VCLIP, against prior work. The results show that our method is on par with prior work for this challenge, with VCLIP achieving the highest performance for the same location condition for all of our metrics and performing second best for all metrics, except mAP for the different location condition.

References

[1] Daniel Davila, Dawei Du, Bryon Lewis, Christopher Funk, Joseph Van Pelt, Roderic Collins, Kellie Corona, Matt Brown, Scott McCloskey, Anthony Hoogs, and Brian Clipp. Mevid: Multi-view extended videos with identities for video person re-identification. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 1634–1643, January 2023.

[2] Chanho Eom, Geon Lee, Junghyup Lee, and Bumsub Ham. Video-based person re-identification with spatial and temporal memory networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12036–12045, 2021.

[3] Xinqian Gu, Hong Chang, Bingpeng Ma, Shutao Bai, Shiguang Shan, and Xilin Chen. Clothes-changing person re-identification with rgb modality only. In *CVPR*, 2022.

[4] Xinqian Gu, Hong Chang, Bingpeng Ma, Hongkai Zhang, and Xilin Chen. Appearance-preserving 3d convolution for video-based person re-identification. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *Computer Vision – ECCV 2020*, pages 228–243, Cham, 2020. Springer International Publishing.

[5] Ruibing Hou, Hong Chang, Bingpeng Ma, Rui Huang, and Shiguang Shan. Bicnet-tks: Learning efficient spatial-temporal representation for video person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer*

Vision and Pattern Recognition (CVPR), pages 2014–2023, June 2021.

[6] Ruibing Hou, Hong Chang, Bingpeng Ma, Shiguang Shan, and Xilin Chen. Temporal complementary learning for video person re-identification. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *Computer Vision – ECCV 2020*, pages 388–405, Cham, 2020. Springer International Publishing.

[7] Priyank Pathak, Amir Erfan Eshratifar, and Michael Gormish. Video person re-id: Fantastic techniques and where to find them (student abstract). *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(10):13893–13894, Apr. 2020.

[8] Yingquan Wang, Pingping Zhang, Shang Gao, Xia Geng, Hu Lu, and Dong Wang. Pyramid spatial-temporal aggregation for video-based person re-identification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 12026–12035, October 2021.

[9] Yiming Wu, Omar El Farouk Bourahla, Xi Li, Fei Wu, Qi Tian, and Xue Zhou. Adaptive graph representation learning for video person re-identification. *IEEE Transactions on Image Processing*, 29:8821–8830, 2020.

[10] Xianghao Zang, Ge Li, and Wei Gao. Multidirection and multiscale pyramid in transformer for video-based pedestrian retrieval. *IEEE Transactions on Industrial Informatics*, 18(12):8776–8785, Dec. 2022.