Re-ID-AR: Improved Person Re-identification in Video via Joint Weakly Supervised Action Recognition - Supplementary Material

Aishah Alsehaim[1,2]

Toby P. Breckon[1]

[1] Department of Computer Science Durham University, UK

[2] Department of Computer Science, Princess Nourah Bint Abdulrahman University, SA

1 Introduction

In the following we provide additional details and results for our approach.

2 Datasets

MARS [15] is a large-scale Re-ID dataset consisting of 1261 identities and 20,715 tracklet under 6 camera views of a single scene and is considered as the largest contemporary video Re-ID benchmark. LPW [10], considered a more challenging dataset in terms of relative Re-ID performance [10, 13], contains around 2,731 identities and 7,694 tracklets across three scenes, where each identity is captured by 2 or 4 cameras. We provide 1261 and 3771 supplementary action labels corresponding to the tracklist present in MARS and LPW respectively. Despite MARS [15] and LPW [10] being selected as containing higher action variations than other Re-ID datasets surveyed, the imbalance in the action distribution per dataset remains significant (Table 1). In general, there are no alternative datasets across the action recognition domain that additionally meets the desired qualities for effective Re-ID evaluation (i.e. multiple non overlapping cameras, challenging outdoor surveillance scenes comprising 1000+ subjects).

3 IDentification Network (IDN)

Our IDN branch is trained using combined person ID labeling and deep metric learning (DML) losses that are used to learn the embedding features to effectively capture semantic similarity knowledge between data points, knowing that each video C consists of set of images $(I_1, I_2, ..., I_m)$ with one ID label C_{id}. The optimisation of this IDN sub-network is guided by the use of four loss functions ID_L, RLL_L, $center_L$ and E_L to produce optimal person features representation for the Re-ID process [2]. Most recent Re-ID approaches [2, 7, 8]

use a modified version of the cross-entropy loss to prevent overfitting, Label Smoothing (LS) [11], as it encourages the model to be less confident on the training set in order to enforce generalisation to unseen examples. This loss is usually called identification (ID) loss, as it calculates the loss of predicted ID to the truth labels, defined as:

$$ID_L = \sum_{i=1}^{N} -q_i \log(pre_i). \tag{1}$$

where N is the number of person subjects and, given a video of person i, pre_i is the ID prediction likelihood of class i and q_i is constructed as follows:

$$q_i = \begin{cases} 1 - \frac{N-1}{N}\varepsilon, & if\ i = y \\ \frac{\varepsilon}{N}, & otherwise \end{cases} \tag{2}$$

where y as the ground-truth ID label and ε is used to encourage the model to be less confident on the training set.

In addition, we make use of deep metric learning (DML) that uses the embedding extracted by the model to learn semantic similarity information among data points, which can boost the model learning, such as center loss [14], triplet loss [9] or rank list loss [12]. Our model is trained using Ranked List Loss (RLL) [12] and center loss [14] as embedding loss functions.

The RLL [12] learns a hypersphere for each class and forces the distance between a positive Re-ID pairing to be smaller than a constant margin. Learning the hypersphere for each class will avoid intra-class data distribution that may occur in other loss functions, such as triplet loss [9]. Forcing a distance between negative samples to be greater than specific threshold α, whilst positive samples are pulled closer than a threshold of $\alpha - m$, where m is the margin. The RLL loss can be defined as the following:

$$RLL_{\mathrm{L}}(\mathbf{x}_i^c; f) = (1-\lambda) L_{\mathrm{P}}(\mathbf{x}_i^c; f) + \lambda L_{\mathrm{N}}(\mathbf{x}_i^c; f) \tag{3}$$

where L_{P} is used to minimize the distance between positive samples and create hyper-sphere of each class, L_{N} is used to push negative samples beyond the boundary α and λ is used to balance the optimization of positive and negative objectives. Consequently, center loss [14] is also applied to support the RLL loss by learning the centre of the deep feature representation for each class and decrease the distances between the embedding and the class they belong to. The center loss can be defined as the following:

$$center_L = \frac{1}{2} \sum_{i=1}^{B} ||f_i - c_{y_i}||_2^2 \tag{4}$$

where y_i is the label of video i in the mini-batch, while c_{y_i} is the class center of its deep features and B is the batch size. The use of center loss supports the intra-class distance minimisation by the RLL loss [12].

One of the main issues ID loss adversely effecting Re-ID is occlusions. The use of Random Erasing Augmentation (REA) [17] deals with this issue in Re-ID videos by randomly erasing a rectangular region within the training imagery during learning. This type of augmentation deals with the partial occlusion and improves the generalisation ability of Re-ID model. To enhance the impact of this augmentation strategy, a high attention score is given to the frame containing the erased region. Following the methodology proposed by [8], by

labelling the erased frames $Erase_L$ by 1 and others by 0 the Erasing-loss E_L can be calculated as the following:

$$E_L = \frac{1}{T}\sum_{t=1}^{T} Erase_L \; a_i^t \tag{5}$$

where a_i^t is the frame level score given by the temporal attention.

The ID_L and RLL_L play different roles to guide the model to produce a robust feature representation for person Re-ID. The ID_L supports the model to learn more discriminative features, while the RLL_L is used to make similar samples closer in the embedding space and make dissimilar samples have greater separation using a predefined distance measurement. The E_L guides the model to deal with occlusions in the video. Consequently, these loss functions appear with equal weight and different roles in IDN total loss function. On the other hand, the $center_L$ is used as support for RLL_L to find the center of each class, following the suggestion of [7] we set the weight of the center loss to β.

4 Ablation Studies

We compare our secondary action recognition task performance against leading contemporary action recognition techniques, that have been similarly adapted with an IDN Re-ID branch output and multi label action recognition (Table 1) - our proposed multitask technique is shown to offer superior action recognition performance. We also compare against alternative strategies for dealing with dataset imbalance (Table 2), where our approach is shown to offer comparable performance to the use of CB loss [3] for Re-ID alone but lesser performance for the secondary task of action recognition across the board.

Methods	Re-ID	Action		
	rank-1 (mAP)	mAP	recall	f1
VTN[4] (WBCE)	0.1(0.4)	22.52	33.33	26.88
TEA [5] (WBCE)	65.22(53.00)	**46.88**	39.84	41.33
Ours (WBCE)	**92.72 (86.29)**	42.40	**47.22**	**49.20**

Table 1: Ablation comparison for action recognition against state-of-the-art for 3 actions recognition on MARS [16].

Methods	Re-ID	Action		
	rank-1 (mAP)	mAP	recall	f1
Ours + DiscCentroidsLoss [6]+BCE	91.63(84.82)	27.44	15.05	16.71
Ours + CB loss [3]	**93.15(87.06)**	**28.66**	4.2	7.07
Ours + imbalance sampler[1]+BCE	86.36(76.12)	24.21	15.60	16.85
Ours + WBCE+ L_{center}	92.17 (86.21)	27.55	**16.51**	**17.82**

Table 2: Ablation comparison on dealing with dataset imbalance for 8 Actions recognition problem: MARS [16].

References

[1] Imbalanced Dataset Sampler. https://github.com/ufoym/imbalanced-dataset-sampler. Accessed: 2020-10-12.

[2] Toby P. Breckon and Aishah Alsehaim. Not 3d re-id: Simple single stream 2d convolution for robust video re-identification. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 5190–5197. IEEE, 2021.

[3] Yin Cui, Menglin Jia, Tsung-Yi Lin, Yang Song, and Serge Belongie. Class-balanced loss based on effective number of samples. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9268–9277, 2019.

[4] Alexander Kozlov, Vadim Andronov, and Yana Gritsenko. Lightweight network architecture for real-time action recognition. In *Proceedings of the 35th Annual ACM Symposium on Applied Computing*, pages 2074–2080, 2020.

[5] Yan Li, Bin Ji, Xintian Shi, Jianguo Zhang, Bin Kang, and Limin Wang. Tea: Temporal excitation and aggregation for action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 909–918, 2020.

[6] Ziwei Liu, Zhongqi Miao, Xiaohang Zhan, Jiayun Wang, Boqing Gong, and Stella X Yu. Large-scale long-tailed recognition in an open world. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2537–2546, 2019.

[7] Hao Luo, Wei Jiang, Youzhi Gu, Fuxu Liu, Xingyu Liao, Shenqi Lai, and Jianyang Gu. A strong baseline and batch normalization neck for deep person re-identification. *IEEE Transactions on Multimedia*, pages 2597–2609, 2019.

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

[9] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 815–823, 2015.

[10] Guanglu Song, Biao Leng, Yu Liu, Congrui Hetang, and Shaofan Cai. Region-based quality estimation network for large-scale person re-identification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, page 7347–7354, 2018.

[11] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2818–2826, 2016.

[12] Xinshao Wang, Yang Hua, Elyor Kodirov, Guosheng Hu, Romain Garnier, and Neil M Robertson. Ranked list loss for deep metric learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5207–5216, 2019.

[13] Xinshao Wang, Elyor Kodirov, Yang Hua, and Neil M Robertson. Id-aware quality for set-based person re-identification. *arXiv preprint arXiv:1911.09143*, 2019.

[14] Yandong Wen, Kaipeng Zhang, Zhifeng Li, and Yu Qiao. A discriminative feature learning approach for deep face recognition. In *European Conference on Computer Vision*, pages 499–515. Springer, 2016.

[15] Liang Zheng, Zhi Bie, Yifan Sun, Jingdong Wang, Chi Su, Shengjin Wang, and Qi Tian. Mars: A video benchmark for large-scale person re-identification. In *European Conference on Computer Vision*, pages 868–884. Springer, 2016.

[16] Liang Zheng, Zhi Bie, Yifan Sun, Jingdong Wang, Chi Su, Shengjin Wang, and Qi Tian. Mars: A video benchmark for large-scale person re-identification. In *European Conference on Computer Vision*, pages 868–884. Springer, 2016.

[17] Zhun Zhong, Liang Zheng, Guoliang Kang, Shaozi Li, and Yi Yang. Random erasing data augmentation. *arXiv preprint arXiv:1708.04896*, 2017.