Not 3D Re-ID: Simple Single Stream 2D Convolution for Robust Video Re-identification

Aishah Alsehaim and Toby P. Breckon

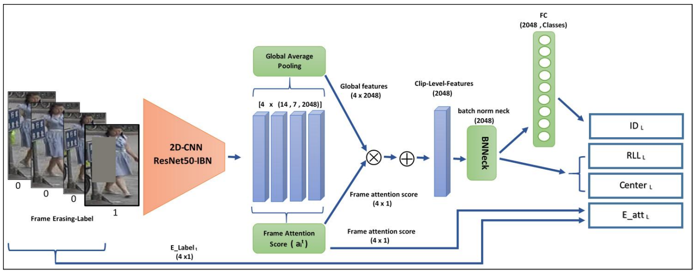
Department of Computer Science
Durham University

Issue: Building robust video-based person Re-ID under varying conditions.

Method :

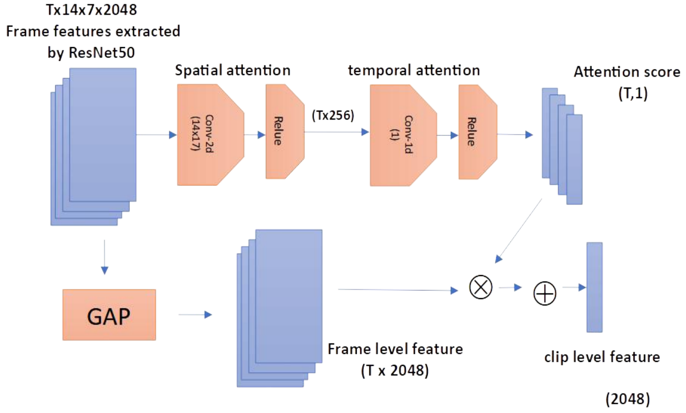
Simple video Re-ID using *{Resnet50+Two-layer spatial-temporal attention}* produce an efficient video features:

- **ResNet50-IBN-a** as a frame features extractor.
- **spatial-temporal attention** following feature extraction to produce video level features (2D convolution + 1D convolution).



Temporal features aggregation:

The use of **2D-Resnet50** as frame feature extractor is followed by a **temporal aggregation** method to produce video level features from T frames.



Spatial-temporal attention

Training:

The use of **multiple loss functions** with differing roles succeeds in guiding the **learning process** of the model without additional complexity.

$$RLL_{\mathrm{L}}\left(\mathbf{x}_i^c; f\right) = L_{\mathrm{P}}\left(\mathbf{x}_i^c; f\right) + \lambda L_{\mathrm{N}}\left(\mathbf{x}_i^c; f\right)$$

Pulls similar samples closer in the embedding space and pushes dissimilar samples apart using a predefined distance measurement.

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

Supports the model in learning more discriminative features.

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

Supports RLL loss to learn sample centric features.

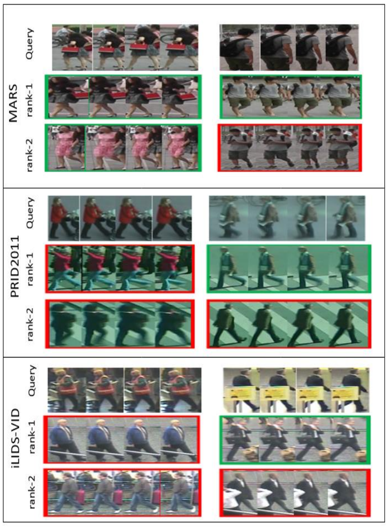
$$E\_att_L = \frac{1}{T}\sum_{t=1}^{T} E\_Label_t \ \ a_i^t$$

Guides the model to overcome partial occlusions.

Experimental Results:

Methods	Publication	MARS [38] rank-1 (mAP)	PRID2011 [11] rank-1	iLIDS-VID [32] rank-1	Input	Memory Usage (MB) Fore/Backward Pass	Params	Total Size
SAN [17]	CVPR 2018	82.3 (65.8)	93.2	80.2				
Att-Driven [37]	CVPR 2019	87.0 (78.2)	93.9	86.3				
VRSTC [12]	CVPR 2019	88.5 (82.3)	–	86.3				
Co-Segment [27]	ECCV 2019	84.9 (79.9)	–	–				
GLTR [16]	ICCV 2019	87.02 (78.47)	95.50	86.00	9.00	214.11	94.47	317.59
M3D [15]	IEEE-T IP 2020	88.63 (79.46)	96.60	86.67	9.00	1213.83	104.58	1327.41
VPRFT [22]	AAAI 2020	88.6(82.9)	93.3	–	9.19	153.92	290.58	453.69
Ours	–	**89.62 (84.61)**	96.6	89.33	9.19	171.92	290.58	471.69
VPRFT [22] (pre-trained on MARS)	AAAI 2020	–	96.6	–				
Ours (pre-trained MARS)	–	–	96.63	**97.33**				
Ours (pre-trained MARS and iLIDS-VID)	–	88.21(83.10)	**97.75**	95.33				

Statistical comparison against state-of-the-art methods.



rank-1 and **rank-2 Re-ID results** to given a query samples over 3 leading benchmark datasets

Green: true match
Red = false match

Conclusion:

- **Single stream robust video Re-ID** approach using **only 2D convolution** for video-based Re-ID.
- Using **robust training strategies** without additional complexity **exceeds state of the art accuracy.**
- Our simple 2D method **exceeds performance of prior 3D convolution and complex multi-stream based approaches.**

Full paper: