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

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

**Abstract**

We uniquely consider the task of joint person re-identification (Re-ID) and action recognition in video as a multi-task problem. In addition to the broader potential of joint Re-ID and action recognition within the context of automated multi-camera surveillance, we show that the consideration of action recognition in addition to Re-ID results in a model that learns discriminative feature representations that both improve Re-ID performance and are capable of providing viable per-view (clip-wise) action recognition. Our approach uses a single 2D Convolutional Neural Network (CNN) architecture comprising a common ResNet50-IBN backbone CNN architecture, to extract frame-level features with subsequent temporal attention for clip level feature extraction, followed by two sub-branches:- the IDentification (sub-)Network (IDN) for person Re-ID and the Action Recognition (sub-)Network for per-view action recognition. The IDN comprises a single fully connected layer while the ARN comprises multiple attention blocks on a one-to-one ratio with the number of actions to be recognised. This is subsequently trained as a joint Re-ID and action recognition task using a combination of two task-specific, multi-loss terms via weakly labelled actions obtained over two leading benchmark Re-ID datasets (MARS, LPW). Our consideration of Re-ID and action recognition as a multi-task problem results in a multi-branch 2D CNN architecture that outperforms prior work in the field (rank-1 (mAP) – MARS: 93.21%(87.23%), LPW: 79.60%) without any reliance 3D convolutions or multi-stream networks architectures as found in other contemporary work. Our work represents the first benchmark performance for such a joint Re-ID and action recognition video understanding task, hitherto unapproached in the literature, and is accompanied by a new public dataset of supplementary action labels for the seminal MARS and LPW Re-ID datasets.

## 1 Introduction

The tasks of person re-identification and action recognition are central pillars within any future fully-automated video surveillance system. Person re-identification in video refers to the task of matching a person in a query surveillance video, to the same person within other videos from multiple non-overlapping cameras whilst action recognition considers what activity a given person is doing within that video sequence. Within real-world surveillance video, both are very challenging problems due to the large variations of human pose, occlusion, differing camera viewpoints, illumination and background scene clutter. In this work we aim to leverage the significant challenge posed by action recognition in such surveillance

© 2021. The copyright of this document resides with its authors.
It may be distributed unchanged freely in print or electronic forms.

video sequences, as a conduit to achieving improved person Re-Identification (Re-ID) under the same challenging conditions.

Video based Re-ID [3, 7, 21, 28, 50, 51] aims to match video of an individual against a gallery of candidates. It benefits from richer multi-frame spatiotemporal information within video that is used to address this task as one of cross-video instance matching. The availability of both visual and temporal features within video Re-ID can be more robust to noise and occlusions, in addition to aligning more naturally with broader video-stream surveillance tasks such as tracking or action recognition. Consequently, the spatiotemporal features present for video-based person Re-ID can also be used for action recognition and one can imagine such a common spatiotemporal feature extraction pipeline being key component within a multi-faceted, multi-camera intelligent surveillance systems [34]. However, despite the obvious alignment of these tasks, contemporary state of the art work in the field generally tackles either one of person Re-ID (in video) [14, 20, 21, 22, 23, 24, 31, 38, 57] or action recognition [6, 8, 35, 42, 43, 44] in isolation. By contrast to recent prior work in the field, here we consider video person Re-ID and action recognition jointly via the use of shared temporal features across both tasks with a view to leveraging the additional spatiotemporal feature requirements for action recognition as a driver for improved Re-ID performance. The consideration of joint Re-ID and action recognition within a shared multi-task computational framework results in a model that learns a more discriminative spatiotemporal feature representation from a given video sequence, in turn improving Re-ID performance whilst being additionally capable of providing clip-wise action recognition. To the best of our knowledge, this work is the first study to consider both tasks within in a single shared architecture applied to the real-world challenges of multi-camera video surveillance under the challenging conditions imposed by existing video Re-ID benchmarks (e.g. MARS [58], LPW [36]).

As a new area of research spanning both video Re-ID and action recognition, there are no readily available benchmark datasets spanning both domains that are representative of the challenges of multi-camera video surveillance *"in the wild"*. Leading Re-ID benchmark datasets [13, 36, 45, 58] contain only person ID annotations, whilst leading action recognition datasets [2, 10, 18, 37] contain no ID annotations and are often devoid of the challenges of occlusion, differing camera viewpoints, illumination and background scene clutter that are the mainstay of challenging Re-ID benchmarks. Whilst some action recognition datasets do contain person ID and action (e.g. NTU-RGB [33]), they do not however meet the Re-ID task requirements of multiple, non-overlapped cameras. Conversely, the breath of human actions present in many leading benchmark Re-ID datasets is very limited (often only *walking*) with the fortuitous exception of MARS [58] and LPW [36]. Within these datasets, we are able to identify and annotate a set of up to eight actions to support this work. In this paper, we thus consider the task of joint person re-identification (Re-ID) and action recognition in video as a multi-task problem. Our approach uses a single 2D Convolutional Neural Network (CNN) architecture comprising a common ResNet50-IBN backbone CNN architecture, to extract frame-level features with subsequent temporal attention for clip level feature extraction, followed by two sub-branches:- the IDentification (sub-)Network (IDN) for person Re-ID and the Action Recognition (sub-)Network for per-view action recognition. Our method is jointly optimized as a multi-task problem using multiple Re-ID (IDN) and the action recognition (ARN) loss terms via weakly labelled actions obtained over two leading benchmark Re-ID datasets (MARS [58], LPW [36]). The main contributions of this paper are:

– we present the first study to consider joint person Re-ID in video and action recognition in a single deep learning (CNN-based) framework, with accompanying bench-

mark task performance and reference dataset annotations.[1]

- we propose an efficient novel joint architecture based solely on 2D convolution operations, capable of achieving state of the art Re-ID performance on MARS [58] and LPW [36] datasets outperforming the prior contemporary work of [14, 20, 21, 23, 24, 36, 38, 57] via the addition of an action recognition sub-branch to the shared CNN backbone that can both learn discriminative feature representations to improve Re-ID performance and is capable of providing secondary (clip-wise) action recognition.
- we introduce supplementary action label annotation for the seminal MARS [58] and LPW [36] Re-ID datasets ( MARS: 1261 / LPW: 3771 action labels).
- we report state of the art Re-ID performance on the MARS (93.21%) and LPW (79.60%) for rank-1 accuracy and furthermore provide an initial benchmark for multi-label action recognition across these two seminal Re-ID datasets that itself outperforms the leading contemporary action recognition approaches of [17, 25].

## 2 Related work

We briefly review relevant prior work in video Re-ID where we find reliable feature representations in contemporary work are generally extracted by tailor-made architectures [39, 40, 48] or generic convolutional neural network (CNN) architectures [1, 20, 31, 57]. Such tailor-made architectures are designed to consider the structure of the human body to reduce the effect of occlusion and to alleviate false detection. More recent research uses generic CNN architectures as a feature extraction network such as ResNet-50 [1, 20, 31] and ResNet-18 [57]. In addition to spatial features, temporal information is a significant component of contemporary video based Re-ID with varying temporal feature aggregation strategies spanning optical flow [3, 7, 28, 51], recurrent neural networks (RNN) [58], temporal pooling [4, 9, 26], spatiotemporal attention [24, 40, 52] or spatiotemporal 3D CNN [21].

The use of optical flow as a temporal aggregation strategy [3, 7, 28, 51] is computationally demanding, requiring significant off-line sample pre-processing, making it impractical for real-time Re-ID in addition to limiting overall robustness to occlusion Re-ID events. RNN are similarly commonplace for temporal feature aggregation in many video analysis tasks [28, 50, 58, 61] but commonly fail to effectively aggregate low-level temporal features effectively in favour of high-level temporal feature connections. In the temporal pooling strategy of [3, 7, 28, 51], all frames are treated equally with clip-features as the average or maximum pooling of all the video frame features. By contrast, many attention-based methods weight each frame and subsequently aggregate frames features are dependant on that weight [1, 9, 22, 24, 31]. More recently 3D convolution has been adopted for spatiotemporal feature learning in video person Re-ID, as it directly extracts spatial-temporal features [21]. However, such 3D CNN approaches are both computationally expensive and require an increased memory footprint whilst recent state of the art approaches show comparable accuracy without such an overhead [1]. In addition, the use of graph neural networks for video Re-ID is introduced in [54], where two separate graph networks for spatial and temporal features are created and jointly optimised to extract video spatial-temporal features.

By contrast, we build directly upon the effectiveness of the leading state-of-the-art 2D convolutional pipeline of [1] (MARS: 89.62%, PRID2011: 97.75%, iLIDS-VID: 97.33% rank-1 accuracy - Table 2, S2DN) and extend this to our joint multi-task Re-ID (IDN sub-network) and action recognition (ARN sub-network) architecture (Figure 1).

[1] Datasets Actions annotation https://github.com/AishahAADU/Re-ID-AR.

# 3 Method

We present an overview of our multi-task approach (Section 3.1) followed by a detailed description of the two sub-networks (branches): the IDentification Network (IDN, Section 3.2) and the Action Recognition Network (ARN, Section 3.3).

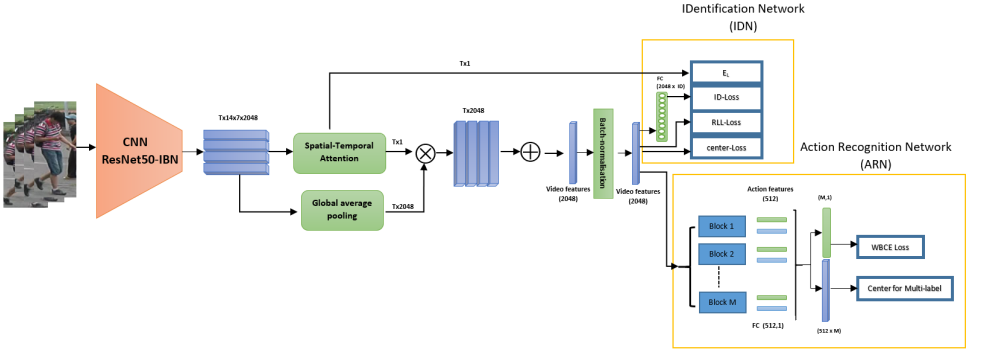


Figure 1: The proposed architecture with shared backbone (ResNet-50-IBN-a [30] and spatial-temporal attention [9]) followed by the IDN and ARN sub-network branches. ($\otimes$ indicates pairwise multiplication; $\oplus$ indicates summation).

## 3.1 Multi-task Re-ID and Action Recognition

Whilst there is a wide range of multi-task person Re-ID and attribute recognition research with the shared target to learn pedestrian descriptions (attributes) [32, 53, 60], no such multi-task formulation with action has been made in the literature to date.

By contrast, we propose multi-task person Re-ID and action recognition as two co-joined video understanding tasks within automated surveillance. Our proposed hybrid architecture uses single shared backbone stream that predicts both ID and action within a shared computational cost. In such video based methods there are two ways to process each video clip: (1) via 3D convolution to extract spatiotemporal features from a temporal block of frames; or (2) via 2D convolution with subsequent temporal aggregation. In general, 3D convolution requires a significantly larger number of parameters to be optimised resulting in both additional computational complexity and an increased memory footprint for both training and inference. By contrast the use of 2D convolution followed by temporal aggregation show comparable state-of-the-art results in recent work [1, 22, 23, 31, 54].

In our multi-task method, we adopt a 2D CNN architecture (ResNet50-IBN-a [30]), to generate frames level features, followed by spatiotemporal attention method to aggregate temporal features through the video frames and to produce video level features. The method learns to produce video features, at the training stage by choosing random frames, $T$, from the tracklet. At inference time, all of the images in the video are used to produce the video level feature by dividing the tracklet into several clips as $V_1$ $(C_1, C_2, ..., C_m)$, each clip has $T$ frames, where $T$ is the number of tracklet frames the model was originally trained with. Our 2D CNN architecture thus extracts features from each frame in the video, and these features are then aggregated using spatiotemporal attention layers to represent video level features (Figure 1). At inference time, our approach extracts clip-level features that are then fused by taking the average of all the clip-level features to represent the Re-ID gallery and query videos. The overall architecture of our proposed method is shown in Figure 1.

The results of a previous comparison study [9], show that temporal attention is the most efficient way to capture temporal information among the sequence of frames in the video

as compared to average/max pooling and Recurrent Neural Network (RNN) aggregation. Temporal attention is preformed to obtain an attention score $a_i^t$ for each frame $f_i^t$ in clip $C_i$ where $t \in [1,T]$. The frame feature $f_i^t$ of a clip $C_i$ are weighted and averaged to represent clip level features. The spatial-temporal attention is preformed using 2D convolution with an input dimensionality of 2048, from the 2D CNN feature extractor (ResNet50-IBN-a [30]), with a 256 dimensional output following [9]. This spatial attention followed by temporal 1D convolution on the frame-level features generates temporal attentions $s_i^t$. The final frame attention score $a_i^t$ is calculated using $softmax()$ [62] with the resulting video features used across the two subsequent sub-network branches (IDN and ARN).

## 3.2 IDentification Network (IDN)

To maximise simplicity and efficacy, the IDN comprises of a single fully connected layer following the shared multi-task spatiotemporal video feature extractor. Following the experiment settings of [1] this branch is trained using four loss functions Label Smoothing ($ID_L$) [41], Ranked List Loss ($RLL_L$) [46], center loss ($center_L$) [49] and Erasing-loss ($E_L$). From these four contributory losses (see supplementary material for details), the overall loss function for IDN sub-network can be formulated as:

$$IDN_{loss} = ID_L + RLL_L + \beta\ center_L + E_L \tag{1}$$

## 3.3 Action Recognition Network (ARN)

This sub-network branch is used to predict the action performed by the subject in a given video. Conventionally in action recognition, each video $C$ has one action label $C_a$ but in our task, with videos originating from real-world Re-ID surveillance datasets, significantly more scene noise, partial occlusion and action transition is present. Consequently, the ARN branch is trained using multi-label action, such that we convert the one action label to multi-label by simply extending one-hot to multi-hot labelling. This is required due to the high probability of action transition that occurs within the Re-ID datasets (MARS, LPW) we are using for training and the weak labelling methology used to obtain the action label ground truth (Section 4). As a result, the ARN is trained to learn multi-label action recognition in the form of the independent likelihood of each action, where $a_i \in [0,1]$ is the likelihood of action $i$ in a given video and hence outputs the likelihood of all the defined actions in the dataset, $\{a_i\}$, for each video.

Following the common 2D CNN architecture to extract video features, that is shared with the IDN, the ARN comprises several separate attention blocks equal to the number of dataset action labels, $M$, to encourage the model to learn discriminative features for each action (Figure 1). Each action attention block consists of one linear layer along with batch-normalisation, ReLU, and dropout layer to generate an attention map. Subsequently, an attention map is generated for each available action label which is then passed to a fully connected layer followed by a sigmoid activation output layer of dimension, 1.

Our reasoning behind the use of an attention block for each action is two-fold: (1) common pedestrian actions can have close appearance features with only subtle differences, as illustrated in Figure 2 and there are highly likely to be action transitions in the pedestrian tracklet video; (2) the highly imbalanced action samples (Table 1) needs to be addressed via weighting. Hence, the main roles for these blocks are to generate attention maps for each action and then assign appropriate weight for each action map through a weighted loss function. As such, the weighted attention map can address the issue of the highly imbalanced action samples in a dataset as the attention maps can be weighted according to the number of

samples in the training dataset by using the weighted binary cross entropy loss [19]. Furthermore, adding attention blocks equal to number of actions helps the model to deal with action transitions and to more appropriately generalize across highly imbalanced dataset such as those considered here.

The optimisation of this branch is performed using the widely used Binary Cross Entropy (BCE) loss, or weighted Binary Cross Entropy (WBCE) [19], on a per multi-label basis. As such BCE loss can be defined for our action recognition task as follows:

$$L_{BCE} = -\frac{1}{M}\sum_{i=1}^{M} y_i log(p_i) + (1-y_i) log(1-p_i) \tag{2}$$

where $M$ is the number of actions, $p_i$ is prediction probability of the action $i$ and $y_i$ is the ground truth label. In contrast, the weighted BCE loss defined in [19] works to alleviate the imbalance distribution of actions in the training datasets and is defined as follows:

$$L_{WBCE} = \frac{1}{V}\sum_{i=1}^{V}\sum_{j=1}^{M} \omega_j (y_i \log(\sigma(p_{ij})) + (1 - y_{ij}) \log(1 - \sigma(p_{ij}))) \tag{3}$$

where $\omega_j$ is the weight assigned to each action based on its distribution in the training dataset, $p_{ij}$ is the output of action $j$ classifier, and $\sigma(z) = 1/(1+e^{-z})$ such that: $V$ is number of videos in the dataset and $M$ is number of actions.

$$\omega_j = \begin{cases} e^{1-r_j}, & y_j = 1 \\ e^{r_j}, & y_j = 0 \end{cases} \tag{4}$$

where $r_j$ is the ratio of the number of action $j$ samples in training dataset.

We find that the use of either of the binary cross entropy losses improved the Re-ID task over the baselines (Section 5). However, we find that using the weighted binary cross entropy loss alleviates the class imbalance problem by weighting the action based on their distribution in the training dataset with a slight reduction in Re-ID performance. We also integrate the center loss for multi-label [29] to cluster the actions by learning the central features of each action then penalize the distance between extracted features and their class center. The use of this improves the accuracy of action recognition as in Table 4 and Table 5. Subsequently, our ARN sub-network is trained by jointly optimising weighted BCE loss and the center loss for multi-label actions as follows:

$$ARN_{loss} = L_{WBCE} + \beta\, L_{center} \tag{5}$$

### 3.4 Multi-Task Network Loss

Our overall multi-task architecture, with our IDN and ARN sub-networks as detailed (Sections 3.2 / 3.3), is constructed using a common attention-enabled 2D CNN backbone (Section 3.1) that is then optimised jointly using combined IDN and ARN losses as follows:

$$L_{total} = \lambda\, IDN_{loss} + (1-\lambda)\, ARN_{loss} \tag{6}$$

## 4 Weakly Labelled Action Annotation

We produce supplementary action labels for the MARS [58] and LPW [36] based on manual annotation of the Re-ID target person in each video sequence as one of the following set of

actions: *{walking, riding, holding item one hand, holding item both hands, holding hands, holding phone to ear, holding phone to face, pulling/pushing trolley}* (Figure 2).

Our labelling is weak in the sense that we follow labelling strategies that produce an imprecise or inexact action labelling approximation. For MARS [58], as the dataset is collected from one scene (via multiple cameras) with very few action transitions within the sequence, we follow a weak identity level labeling strategy by manually assigning each person a single action label for all the videos that he/she appears, based on their primary action characteristic, in even if there are secondary action transitions present. In contrast for LPW [36], as the dataset was collected across three separate scenes (via multiple cameras) with a very high prevalence of action transitions, we instead use a per-video sequence labelling strategy (i.e. per subject, per camera view) but again manually assign each video a single action label, based on the primary action characteristic from that view, in even if there are secondary action transitions present. Due to the imbalance of action labels present (Table 1), we additional consider broader action definitions by grouping certain subsets of actions to form even more weakly defined super-labels. For example, grouping all actions relating to a hand movement/gesture results in three action labels: - *{walking, riding, using hand with object}* . Similarly, with MARS we also consider a set of five action labels by merging the three actions labels with much lower occurrences (i.e. *{holding hands, holding phone to ear, pulling/pushing trolley}* ) into the *{holding item one hand}* action label. In our subsequent evaluation this gives us consideration of three such multi-task Re-ID/Action Recognition problems:- 8 actions, 5 actions (MARS only) and 3 actions (Tables 4 / 5).

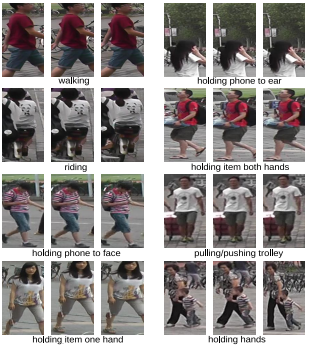


Figure 2: Action examples within MARS dataset [58].

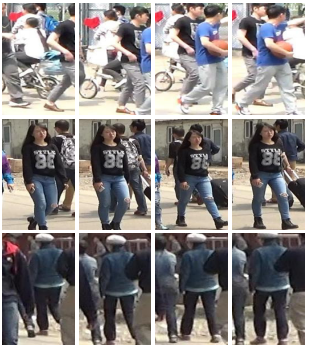

Figure 3: Illustrative examples of challenging issues in LPW [36].

| Action | MARS [58] | | LPW [36] | |
|---|---|---|---|---|
| | Train Set | Test Set | Train Set | Test Set |
| Walking | 396 (63.36%) | 439(69.24%) | 946(35.02%) | 488(45.61%) |
| Riding | 20 (3.2%) | 10(1.58%) | 9 (0.33%) | 92(8.60%) |
| Holding item in one hand | 151(23.82%) | 116 (18.30%) | 368 (13.62%) | 97(9.07%) |
| Holding item in both hands | 33 (5.28%) | 28(4.42%) | 939 (34.76%) | 284(26.54%) |
| Holding hands | 3 (0.48%) | 4 (0.63%) | 36 (1.33%) | 36(3.36%) |
| Holding phone to the ear | 5 (0.8%) | 8(1.26%) | 86 (3.18%) | 14 (1.31%) |
| Holding phone to the face | 14 (2.24%) | 27(4.26%) | 189 (6.10%) | 53(4.95%) |
| Pulling/pushing a trolley | 3 (0.48%) | 2 (0.32%) | 128 (4.74%) | 6(0.56%) |

Table 1: Distribution of action labels in MARS/LPW datasets.

# 5 Evaluation

We evaluate our approach on the MARS [58] and LPW [36] datasets, as described in Section 4. For Re-ID, via the IDN sub-network, we report the Cumulative Match Characteristic (CMC) and mean Average Precision (mAP) metrics. CMC measures the prevalence of the ground truth identity within the top-k ranked matches whereby we report rank-1 accuracy

(for MARS and LPW). The mAP metric measures model performance in multi-shot Re-ID datasets such as MARS [58] and is reported to follow the common practice (for MARS only). For action recognition, via the ARN sub-network, on our imbalanced multi-label recognition task we report mean Average Precision (mAP), recall and $F1$ in average score.

## 5.1 Implementation Details

For initial video frame feature extraction we use a variant of ResNet, ResNet50-IBN-a [30] pre-trained on ImageNet [12], because of its ability to maintain effective discriminative features and eliminate appearance variance, which is the most significant challenge within Re-ID. In our architecture, the last spatial down-sampling stride of ResNet50-IBN-a is changed to 1 as suggested by [20], to bring higher spatial resolution without additional parameters and with a low computational cost. Video frames are resized to $244 \times 112$ and the resized image frame is zero-padded by 10 pixels. It is then randomly cropped into $244 \times 112$ rectangular image samples and each sample is flipped horizontally with 0.5 probability. The RGB channels are normalised by subtracting $(0.485, 0.456, 0.406)$ and then dividing by $(0.229, 0.224, 0.225)$, following ImageNet [12]. Our model is trained using four frames for each video, $T = 4$, following the suggestion of [1, 9, 31] and using the combined loss across the IDN and ARN sub-network branches (Eqn. 6).

The IDN losses are ID loss [41], center loss [49], Ranked List Loss (RLL) [46] and Erasing-loss [31]. In our experiments, $\varepsilon = 0.1$ in ID loss [41]. RLL is used to force a distance between negative ID samples to be greater than specific threshold $\alpha$, in our experiments $\alpha = 2.0$. In addition, the positive ID samples are pulled to be closer than a threshold $\alpha - m$, where $m$ is 1.3 in our experiments. For center loss [49], the center of each ID class is learned using the SGD optimizer with 0.5 learning rate. To balance its weight we follow the suggestion of [27] and multiply the center loss by factor $\beta = 0.0005$.

The ARN losses having one corresponding fully connected layer output that is optimised using weighted Binary Cross Entropy loss to enable multi-action labelling and hence help alleviate the dataset imbalance problem within model learning process. We also apply center loss for the multi-label action loss [29] with factor $\beta = 0.0005$. The center of each action is learned using the SGD optimizer as its dedicated optimizer, with 0.1 learning rate.

Overall our multi-task model is trained, using the IDN and ARN losses functions as formalised in Equation 6 with $\lambda = 0.5$. The model is trained for 120 epochs and is validated every 10 epochs. Adam [16] is used as the optimiser for our model with base learning rate of 0.00035 and an adaptive learning rate warm-up strategy following [27].

## 5.2 Comparison with the State-of-the-art Methods

Our experiments show that the use of our single stream multi-task approach improves Re-ID performance with a 3%+ margin on the MARS [58] dataset and a 8%+ margin on the LPW [36] dataset when compared to prior work in the field (Table 2, including very recent single-task work [1, 8, 11, 15, 22, 23, 23, 52, 56] on MARS ). Our action recognition accuracy, as a secondary task on the basis of multi-label classification output is shown in Table 4 (MARS [58]) and Table 5 (LPW [36]) where we can see moderate performance on these otherwise challenging, imbalanced and weakly labelled datasets. Furthermore, in comparison to leading contemporary techniques [17, 25], our approach outperforms the current state of the art on the more balanced three action problem (see supplementary material for details).

## 5.3 Ablation Studies

Comparing our approach, with and without action recognition as a multi-task problem (i.e. with/without ARN branch) we similarly see a a 3%+ margin of Re-ID improvement on

| Methods | Publication | MARS[58] rank-1 (mAP) | LPW[36] rank-1 |
|---|---|---|---|
| RQEN [36] | AAAI 2018 | 73.7 (51.7) | 57.1 |
| SAN [24] | CVPR 2018 | 82.3 (65.8) | - |
| Att-Driven [31] | CVPR 2019 | 87.0 (78.2) | - |
| VRSTC [14] | CVPR 2019 | 88.5 (82.3) | - |
| Co-Segment [33] | ECCV 2019 | 84.9 (79.9) | - |
| GLTR [20] | ICCV 2019 | 87.02 (78.47) | - |
| M3D [21] | IEEE-T IP 2020 | 88.63 (79.46) | - |
| ID-aware [10] | arXiv 2019 | 83.3 (71.7) | 70.9 |
| VPRFT [31] | AAAI 2020 | 88.6 (82.9) | - |
| TACAN [22] | WACV 2020 | 89.1 (84.0) | - |
| STGCN [34] | CVPR 2020 | 89.95 (83.70) | - |
| S2DN [1] | ICPR 2020 | 89.62 (84.61) | - |
| MG-TCN [23] | IEEE-T CS-VT 2021 | 87.1 (77.7) | - |
| AP3D[11] | ECCV 2020 | 90.1(85.1) | - |
| MG-RAFA [26] | CVPR 2020 | 88.8(85.9) | - |
| AFA [8] | ECCV 2020 | 90.2(82.9) | - |
| TCLNet [15] | ECCV 2020 | 88.8(83.0) | - |
| MGH [32] | CVPR 2020 | 90(85.8) | - |
| Re-ID-AR (Ours) | – | **93.21 (87.23)** | **79.60** |

Table 2: Re-ID State-of-the-art Comparison: MARS / LPW.

MARS [58] dataset and 2%+ margin on LPW [36] (Table 3). Whilst this verifies the effectiveness of our multi-task method, the reason behind the slighter improvement in LPW is attributable to the challenging nature of the LPW dataset videos with multiple views across multiple scenes and a high number of action transitions per sequence as shown in Figure 3, where we can see multiple persons in each frame performing different actions. Adding our ARN branch to contemporary Re-ID work, we can see that our ARN branch improves the Re-ID performance on VPRFT [31] by a 3%+ margin on MARS [58], however for TCLNet [15] and AP3D[11] adding our ARN branch did not improve the Re-ID accuracy potentially due to its existing multi-task architecture and 3D convolution with high number of parameters, and hence by adding an additional task there is potentially a need for an additional balancing strategy.

| Methods | Re-ID only | | Multi-task Re-ID with Action | |
|---|---|---|---|---|
| | MARS[58] | LPW[36] | MARS[58] | LPW [36] |
| VPRFT [31] | 88.6 (82.9) | - | 92.2 (83.0) | - |
| TCLNet [15] | 88.8 (83.0) | - | 85.71 (78.41) | - |
| AP3D[11] | 90.1 (85.1) | - | 87.6 (81.4) | - |
| **Ours** | **89.62 (84.61)** | **77.35** | **93.21 (87.23)** | **79.60** |

Table 3: The effect of adding ARN to different Re-ID methods.

Considering the three action recognition problems set out in Section 4 (8 actions, 5 actions and 3 actions) we consider performance on 8/5/3 actions for MARS [58] and 8/3 actions for LPW [36] as it is more balanced (Table 1). As recommended by [55], the most suitable metric to evaluate our ARN accuracy as a multi-label task on an imbalanced dataset is $f1$ score. By considering varying granularity in our weak action labelling allows us to mitigate the effects of dataset imbalance somewhat and additionally study the effect of action recognition task complexity on Re-ID task performance. We also examine the use of BCE loss and WBCE in both tasks and study the effect of adding center for multi-label to ARN losses. The results in Table 4 show that using BCE gives higher performance for Re-ID in all three for action recognition problems on MARS [59]. We can also observe that the 8 action problem has the lowest action recognition accuracy attributable to extreme dataset imbalance. By contrast, we see improved action recognition performance in the 5/3 action problems at

the marginal expense of Re-ID performance. Consequently, the use of joint losses WBCE and center for multi-label actions result in peak action recognition with a slight decrease in Re-ID performance (Table 4, f1), except in the 8 actions set with some actions unrecognised. The results in Table 5 examine the effect of action recognition on Re-ID via the 8/3 action problems on LPW [36]. For the 8 action problem, we report the highest Re-ID accuracy but with the lowest action recognition accuracy due to dataset imbalance. For the 3 action problem, the action recognition accuracy improves with the same effect as the WBCE and center losses for multi-label actions as observed in MARS [58].

| Methods | Re-ID | Action | | |
|---|---|---|---|---|
| | rank-1 (mAP) | mAP | recall | f1 |
| 8 Actions+BCE | 93.10 (87.10) | 27.42 | 14.90 | 16.19 |
| 8 Action+WBCE | 92.28 (86.25) | 26.36 | **16.75** | **18.11** |
| 8 Actions+BCE+ $L_{center}$ | **93.21 (87.23)** | 25.94 | 14.80 | 16.10 |
| 8 Actions+WBCE+ $L_{center}$ | 92.17 (86.21) | **27.55** | 16.51 | 17.82 |
| 5 Actions+BCE | **92.83 (86.96)** | **48.41** | 24.98 | 28.01 |
| 5 Action+WBCE | 92.61 (86.32) | 46.88 | 27.96 | 30.85 |
| 5 Actions+BCE+ $L_{center}$ | 92.55 (86.79) | 45.6 | 23.9 | 27.02 |
| 5 Actions+WBCE+ $L_{center}$ | 92.75 (86.39) | 48.09 | **28.47** | **31.07** |
| 3 Actions+BCE | **92.88 (86.63)** | 44.42 | 42.42 | 46.09 |
| 3 Action+WBCE | 92.72 (86.29) | 42.40 | 47.22 | 49.20 |
| 3 Actions+BCE+ $L_{center}$ | 92.66 (86.50) | **45.23** | 44.56 | 48.99 |
| 3 Actions+WBCE+ $L_{center}$ | 92.55 (86.30) | 43.74 | **47.50** | **49.85** |

Table 4: Ablation comparison for action recognition: MARS [59].

| Methods | Re-ID | Action | | |
|---|---|---|---|---|
| | rank-1 | mAP | recall | f1 |
| 8 Actions+BCE | 77.75 | 26.77 | 10.62 | 13.22 |
| 8 Action+WBCE | 75.76 | **26.64** | 15.82 | **18.34** |
| 8 Actions+BCE+ $L_{center}$ | **78.15** | 26.42 | 10.76 | 13.16 |
| 8 Actions+WBCE+$L_{center}$ | 74.83 | 36.72 | **14.52** | 18.05 |
| 3 Actions+BCE | **77.09** | **72.96** | 40.53 | 36.21 |
| 3 Action+WBCE | 75.23 | 39.04 | 43.36 | 38.28 |
| 3 Actions+BCE+ $L_{center}$ | 76.69 | 72.72 | 41.46 | 39.58 |
| 3 Actions+WBCE+ $L_{center}$ | 75.76 | 39.04 | **44.45** | **39.93** |

Table 5: Ablation comparison for action recognition: LPW [36].

In terms of the effect of $\lambda$ in the learning process, our experiments shown in Table 6/7 illustrate a small to marginal effect on performance which is attributable to our single-stream network design such that the IDN and ARN branches share a common backbone.

| Method | $\lambda$ | rank1 | mAP |
|---|---|---|---|
| 8 Actions+WBCE+ $L_{center}$ | 0.8 | 93.37 | 87.28 |
| 8 Actions+WBCE+ $L_{center}$ | 0.5 | 93.21 | 87.23 |
| 8 Actions+WBCE+ $L_{center}$ | 0.4 | 92.50 | 87.03 |
| 8 Actions+WBCE+ $L_{center}$ | 0.2 | 92.83 | 87.03 |

Table 6: Effect of $\lambda$ on learning process for MARS[58] dataset.

| Method | $\lambda$ | rank1 |
|---|---|---|
| 8 Actions+WBCE+ $L_{center}$ | 0.8 | 79.60 |
| 8 Actions+WBCE+ $L_{center}$ | 0.5 | 78.15 |
| 8 Actions+WBCE+ $L_{center}$ | 0.4 | 75.50 |
| 8 Actions+WBCE+ $L_{center}$ | 0.2 | 75.50 |

Table 7: Effect of $\lambda$ on learning process for LPW [36] dataset.

# 6 Conclusions

In this paper, we propose a single stream 2D CNN architecture as the first approach to jointly consider person re-identification (Re-ID) and action recognition in video as a multi-task problem. Our work shows that consideration of action recognition in addition to Re-ID results in an improved discriminative feature representation that both improves Re-ID performance against prior contemporary work in the field [1, 14, 20, 21, 22, 23, 24, 31, 38, 57] including recent multi-task work [8, 11, 15, 52, 56] and is additionally capable of providing viable per-view (clip-wise) action recognition beyond that of leading action recognition approaches in the field [17, 25] for the challenging datasets considered. Our use of weakly labelled actions, over two leading benchmark Re-ID datasets (MARS [58], LPW [36]), for training as a joint Re-ID and action recognition task using a combination of two task-specific multi-loss terms notably outperforms prior work in the field (rank-1 (mAP) – MARS: 93.21%(87.23%), LPW: 79.60%) without any reliance 3D convolutions or multi-stream networks architectures as found in other contemporary work [4, 21, 28, 51, 54]. This represents the first benchmark performance for such a joint Re-ID and action recognition video understanding task based on our generation and use of supplementary action labels for the seminal MARS and LPW Re-ID datasets. Future work will continue to expand the use of multi-task optimisation for Re-ID and broader aspects of automated visual surveillance.

# References


[1] 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.

[2] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6299–6308, 2017.

[3] Dapeng Chen, Hongsheng Li, Tong Xiao, Shuai Yi, and Xiaogang Wang. Video person re-identification with competitive snippet-similarity aggregation and co-attentive snippet embedding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1169–1178, 2018.

[4] Dapeng Chen, Hongsheng Li, Tong Xiao, Shuai Yi, and Xiaogang Wang. Video person re-identification with competitive snippet-similarity aggregation and co-attentive snippet embedding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1169–1178, 2018.

[5] Guangyi Chen, Yongming Rao, Jiwen Lu, and Jie Zhou. Temporal coherence or temporal motion: Which is more critical for video-based person re-identification? In *European Conference on Computer Vision*, pages 660–676. Springer, 2020.

[6] R Christoph and Feichtenhofer Axel Pinz. Spatiotemporal residual networks for video action recognition. *Advances in neural information processing systems*, pages 3468–3476, 2016.

[7] Dahjung Chung, Khalid Tahboub, and Edward J Delp. A two stream siamese convolutional neural network for person re-identification. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1983–1991, 2017.

[8] Christoph Feichtenhofer, Axel Pinz, and Richard P Wildes. Spatiotemporal multiplier networks for video action recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4768–4777, 2017.

[9] Jiyang Gao and Ram Nevatia. Revisiting temporal modeling for video-based person reid. *arXiv preprint arXiv:1805.02104*, 2018.

[10] Raghav Goyal, Samira Ebrahimi Kahou, Vincent Michalski, Joanna Materzynska, Susanne Westphal, Heuna Kim, Valentin Haenel, Ingo Fruend, Peter Yianilos, Moritz Mueller-Freitag, et al. The" something something" video database for learning and evaluating visual common sense. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5842–5850, 2017.

[11] Xinqian Gu, Hong Chang, Bingpeng Ma, Hongkai Zhang, and Xilin Chen. Appearance-preserving 3d convolution for video-based person re-identification. In *European Conference on Computer Vision*, pages 228–243. Springer, 2020.

[12] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1026–1034, 2015.

[13] Martin Hirzer, Csaba Beleznai, Peter M Roth, and Horst Bischof. Person re-identification by descriptive and discriminative classification. In *Scandinavian Conference on Image Analysis*, pages 91–102. Springer, 2011.

[14] Ruibing Hou, Bingpeng Ma, Hong Chang, Xinqian Gu, Shiguang Shan, and Xilin Chen. Vrstc: Occlusion-free video person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7183–7192, 2019.

[15] Ruibing Hou, Hong Chang, Bingpeng Ma, Shiguang Shan, and Xilin Chen. Temporal complementary learning for video person re-identification. In *European conference on computer vision*, pages 388–405. Springer, 2020.

[16] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *International Conference on Learning Representations*, 2015.

[17] 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.

[18] Hildegard Kuehne, Hueihan Jhuang, Estíbaliz Garrote, Tomaso Poggio, and Thomas Serre. Hmdb: a large video database for human motion recognition. In *International Conference on Computer Vision*, pages 2556–2563. IEEE, 2011.

[19] Dangwei Li, Xiaotang Chen, Zhang Zhang, and Kaiqi Huang. Pose guided deep model for pedestrian attribute recognition in surveillance scenarios. In *IEEE International Conference on Multimedia and Expo*, pages 1–6. IEEE, 2018.

[20] Jianing Li, Jingdong Wang, Qi Tian, Wen Gao, and Shiliang Zhang. Global-local temporal representations for video person re-identification. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3958–3967, 2019.

[21] Jianing Li, Shiliang Zhang, and Tiejun Huang. Multi-scale temporal cues learning for video person re-identification. *IEEE Transactions on Image Processing*, pages 4461–4473, 2020.

[22] Mengliu Li, Han Xu, Jinjun Wang, Wenpeng Li, and Yongli Sun. Temporal aggregation with clip-level attention for video-based person re-identification. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3376–3384, 2020.

[23] P. Li, P. Pan, P. Liu, M. Xu, and Y. Yang. Hierarchical temporal modeling with mutual distance matching for video based person re-identification. *IEEE Transactions on Circuits and Systems for Video Technology*, pages 503–511, 2021.

[24] Shuang Li, Slawomir Bak, Peter Carr, and Xiaogang Wang. Diversity regularized spatiotemporal attention for video-based person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 369–378, 2018.

[25] 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.

[26] Yu Liu, Junjie Yan, and Wanli Ouyang. Quality aware network for set to set recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5790–5799, 2017.

[27] 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.

[28] Niall McLaughlin, Jesus Martinez Del Rincon, and Paul Miller. Recurrent convolutional network for video-based person re-identification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1325–1334, 2016.

[29] Sanath Narayan, Hisham Cholakkal, Fahad Shahbaz Khan, and Ling Shao. 3c-net: Category count and center loss for weakly-supervised action localization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8679–8687, 2019.

[30] Xingang Pan, Ping Luo, Jianping Shi, and Xiaoou Tang. Two at once: Enhancing learning and generalization capacities via ibn-net. In *Proceedings of the European Conference on Computer Vision*, pages 464–479. 2018.

[31] 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.

[32] Nikolaos Sarafianos, Xiang Xu, and Ioannis A Kakadiaris. Deep imbalanced attribute classification using visual attention aggregation. In *Proceedings of the European Conference on Computer Vision*, pages 680–697, 2018.

[33] Amir Shahroudy, Jun Liu, Tian-Tsong Ng, and Gang Wang. Ntu rgb+d: A large scale dataset for 3d human activity analysis. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1010–1019, 2016.

[34] Guruh Fajar Shidik, Edi Noersasongko, Adhitya Nugraha, Pulung Nurtantio Andono, Jumanto Jumanto, and Edi Jaya Kusuma. A systematic review of intelligence video surveillance: Trends, techniques, frameworks, and datasets. *IEEE Access*, pages 170457–170473, 2019.

[35] Karen Simonyan and Andrew Zisserman. Two-stream convolutional networks for action recognition in videos. In *Advances in Neural Information Processing Systems*, pages 568–576, 2014.

[36] 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.

[37] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012.

[38] Arulkumar Subramaniam, Athira Nambiar, and Anurag Mittal. Co-segmentation inspired attention networks for video-based person re-identification. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 562–572, 2019.

[39] Yumin Suh, Jingdong Wang, Siyu Tang, Tao Mei, and Kyoung Mu Lee. Part-aligned bilinear representations for person re-identification. In *Proceedings of the European Conference on Computer Vision*, pages 402–419, 2018.

[40] Yifan Sun, Liang Zheng, Yi Yang, Qi Tian, and Shengjin Wang. Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline). In *Proceedings of the European Conference on Computer Vision*, pages 480–496. 2018.

[41] 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.

[42] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4489–4497, 2015.

[43] Gül Varol, Ivan Laptev, and Cordelia Schmid. Long-term temporal convolutions for action recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1510–1517, 2017.

[44] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 5998–6008, 2017.

[45] Xiaogang Wang and Rui Zhao. Person re-identification: System design and evaluation overview. In *Person Re-Identification*, pages 351–370. Springer, 2014.

[46] 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.

[47] 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.

[48] Yicheng Wang, Zhenzhong Chen, Feng Wu, and Gang Wang. Person re-identification with cascaded pairwise convolutions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1470–1478, 2018.

[49] 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.

[50] Lin Wu, Chunhua Shen, and Anton van den Hengel. Deep recurrent convolutional networks for video-based person re-identification: An end-to-end approach. *arXiv preprint arXiv:1606.01609*, 2016.

[51] Shuangjie Xu, Yu Cheng, Kang Gu, Yang Yang, Shiyu Chang, and Pan Zhou. Jointly attentive spatial-temporal pooling networks for video-based person re-identification. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4733–4742, 2017.

[52] Yichao Yan, Jie Qin, Jiaxin Chen, Li Liu, Fan Zhu, Ying Tai, and Ling Shao. Learning multi-granular hypergraphs for video-based person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2899–2908, 2020.

[53] Jie Yang, Jiarou Fan, Yiru Wang, Yige Wang, Weihao Gan, Lin Liu, and Wei Wu. Hierarchical feature embedding for attribute recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13055–13064, 2020.

[54] Jinrui Yang, Wei-Shi Zheng, Qize Yang, Ying-Cong Chen, and Qi Tian. Spatial-temporal graph convolutional network for video-based person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3289–3299, 2020.

[55] M. Zhang and Z. Zhou. A review on multi-label learning algorithms. *IEEE Transactions on Knowledge and Data Engineering*, pages 1819–1837, 2014.

[56] Zhizheng Zhang, Cuiling Lan, Wenjun Zeng, and Zhibo Chen. Multi-granularity reference-aided attentive feature aggregation for video-based person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10407–10416, 2020.

[57] Yiru Zhao, Xu Shen, Zhongming Jin, Hongtao Lu, and Xian-sheng Hua. Attribute-driven feature disentangling and temporal aggregation for video person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4913–4922, 2019.

[58] 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.

[59] 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.

[60] Kaiyang Zhou, Yongxin Yang, Andrea Cavallaro, and Tao Xiang. Omni-scale feature learning for person re-identification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3702–3712, 2019.

[61] Zhen Zhou, Yan Huang, Wei Wang, Liang Wang, and Tieniu Tan. See the forest for the trees: Joint spatial and temporal recurrent neural networks for video-based person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4747–4756, 2017.

[62] Zhen Zhou, Yan Huang, Wei Wang, Liang Wang, and Tieniu Tan. See the forest for the trees: Joint spatial and temporal recurrent neural networks for video-based person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4747–4756, 2017.