Video Person Re-Identification using Learned Clip Similarity Aggregation

Neeraj Matiyali
IIT Kanpur
neermat@cse.iitk.ac.in

Gaurav Sharma
NEC Labs America
grv@nec-labs.com

Abstract

We address the challenging task of video-based person re-identification. Recent works have shown that splitting the video sequences into clips and then aggregating clip-based similarity is appropriate for the task. We show that using a learned clip similarity aggregation function allows filtering out hard clip pairs, e.g. where the person is not clearly visible, is in a challenging pose, or where the poses in the two clips are too different to be informative. This allows the method to focus on clip-pairs which are more informative for the task. We also introduce the use of 3D CNNs for video-based re-identification and show their effectiveness by performing equivalent to previous works, which use optical flow in addition to RGB, while using RGB inputs only. We give quantitative results on three challenging public benchmarks and show better or competitive performance. We also validate our method qualitatively.

1. Introduction

Person re-identification is the problem of identifying and matching persons in videos captured from multiple non-overlapping cameras. It plays an important role in many intelligent video surveillance systems and is a challenging problem due to the variations in camera viewpoint, person pose and appearance, and challenging illumination along with various types and degrees of occlusions.

Visual person re-identification involves matching two images or video sequences (containing persons) to answer whether the persons in the two videos are the same or not. The general approach for it includes (a) extraction of features that are discriminative wrt. the identity of the persons while being invariant to changes in pose, viewpoint, and illumination and (b) estimating a distance metric between the features. The earlier methods for re-identification used handcrafted features in conjunction with metric learning to perform the task [7, 10, 16, 24, 43]. These works mainly leveraged intuitions for the task, while in recent years, the use of deep CNNs has become more common owing to their superior performance [1, 4, 6, 19, 39, 40].

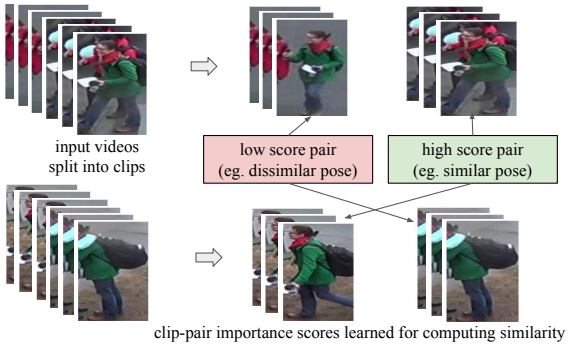

Figure 1: Illustration of the proposed method. We learn an importance scoring function for aggregating clip pairs of video sequences, for person re-identification task. The method learns to weight important clip pairs, which help in discrimination, higher than those which are not informative.

Many of the previous works on person re-identification have focused on image-based benchmarks, however, with the introduction of large-scale video re-identification benchmarks such as MARS [44], the video-based setting is becoming popular. Most existing methods on video-based re-identification extract CNN features of individual frames and aggregate them using average pooling, max pooling, temporal attention mechanisms, or RNNs [25, 41, 44, 47]. These methods, thus, represent the video sequence as a single feature vector. However, for long sequences that have a significant amount of variation in pose, illumination, etc., a single vector might not be enough to represent them.

A recent state-of-the-art video-based method by Chen et al. [3] address the problem by dividing the sequences into short clips, and embedding each clip separately using a CNN and applying a temporal attention based method. To match two given sequences, they compute similarities between all pairs of clips and compute the final similarity by aggregating a fixed percentage of top clip pair similarities. Thus, the contribution of a clip in a video sequence is dynamically determined, based on its similarities to the clips in the other sequence. Chen et al. [3] assume that the similarity between a pair of clips is indicative of the informativeness of the clip pair. We argue that this assumption is not necessarily true in practice, e.g., a pair of clips with low

similarity can be utilized as evidence for the fact that the persons in the two clips are different. Such clip-pairs get discarded while computing the final similarity, which may hurt the re-identification performance. Another shortcoming of the method is that it uses a fixed percentage of the clip-pairs for all pairs of sequences. This limits the performance of the method since for different pairs of sequences, the number of informative clip-pairs can vary.

We address the above shortcomings of Chen et al. [3], and propose an end-to-end trainable model to estimate the similarity between two video sequences. Our model takes pairs of clips as input in a sequence and predicts an importance score for each clip pair. It computes the final similarity between the two sequences by taking an average of the clip-pair similarities weighted by their corresponding importance scores. Thus, our model allows the filtering of non-informative or distracting clip-pairs while focusing only on clip-pairs relevant for estimating the similarities. While [3] aim to filter non-informative or distracting clip-pairs, like here, the measure of informativeness is different. [3] uses clip-level similarity as a proxy for the informativeness, while our method uses a learnable scoring function optimized for the task at the video level. Consider a clip-pair without any artefact, but with a low clip-similarity due to different persons being present. While [3] would reject such a pair despite it being informative, our scoring function would give it high importance to maintain a low overall similarity.

As another contribution, we show the effectiveness of 3D CNNs [36, 2] for obtaining clip features. 3D CNNs, which have been used for various video-based tasks such as action recognition in recent years, remain largely unexplored for the task of video-based person re-identification. We show their effectiveness on this task, by reporting performances equivalent to previous works which use optical flow in addition to RGB, while using RGB inputs only.

We give quantitative results on three video-based person re-identification benchmarks, MARS [44], DukeMTMC-VideoReID [28, 38] and PRID2011 [12]. We show that our trainable similarity estimation model performs better than the top clip-similarity aggregation proposed by Chen et al. [3]. To simulate more challenging situations, we also report experiments with partial frame corruption, which could happen due to motion blur or occlusions, and show that our method degrades gracefully and performs better than the competitive baseline. We also provide qualitative results that verify the intuition of the method.

2. Related Work

Image based Re-Identification. Initial works on person re-identification focused on designing and extracting discriminative features from the images [10, 7, 24, 16, 43]. These works mainly leveraged intuitions for the task and proposed hand-designed descriptors that capture the shape, appearance, texture and other visual aspects of the person. Other works proposed better metric learning methods for the task of person reidentification [10, 27, 46, 15, 20, 26]. This line of work mainly worked with standard features and innovated on the type and better applicability of metric learning algorithms for the task.

More recent methods have started leveraging CNN features for the task of person re-identification. These methods explore various CNN architectures and loss functions. Li et al. [19] proposed a CNN architecture specifically for the re-identification task, which was trained using a binary verification loss. Ding et al. [6] proposed a triplet loss to learn CNN features. Ahmed et al. [1] proposed a siamese CNN architecture and used binary verification loss for training. Cheng et al. [4] used a parts-based CNN model for re-identification, which was learned using a triplet loss. Xiao et al. [39] used domain guided dropout that allowed learning of CNN features from multiple domains. They used a softmax classification loss to train the model. Xiao et al. [40] jointly trained a CNN for pedestrian detection and identification. They proposed online instance matching (OIM) loss, which they showed to be more efficient than the softmax classification loss.

Another line of work [34, 45, 42, 35] leverages human pose estimators and uses parts-based representations for person re-identification. For example, Suh et al. [35] used a two-stream framework with an appearance and a pose stream, which were combined using bilinear pooling to get a part-aligned representation.

Video-based person re-identification. The methods working with videos commonly rely on CNNs to extract features from the individual frames, while using different ways for aggregating frame-wise CNN features, e.g. Yan et al. [41] used LSTM to aggregate the frame-wise features. Zheng et al. [44] aggregated the CNN features using max/average pooling and also used metric learning schemes such as KISSME [15] and XQDA [20] to improve the re-identification performance. McLaughlin et al. [25] used RNN on top of CNN features followed by temporal max/average pooling.

More recent works have also started exploring temporal and spatial attention based methods for video-based re-identification. Zhou et al. [47] used a temporal attention mechanism for the weighted aggregation of frame features. Li et al. [18] employed multiple spatial attention units for discovering latent visual concepts that are discriminative for re-identification. They combined the spatially gated frame-wise features from each spatial attention unit using temporal attention mechanisms and concatenation.

Liu et al. [21] used the two-stream framework for video re-identification, which consists of an appearance and a motion stream, to exploit the motion information in the video

sequences. Instead of using pre-computed optical flow, however, they learned the motion context from RGB images in an end-to-end manner.

3. Approach

We assume humans have been detected and tracked and we are provided with cropped videos that contain a single human. We view the videos as ordered sequences of tensors (RGB frames). We formally define the problem we address as that of learning a parameterized similarity between two ordered sequences of tensors. Denote the query and the gallery video sequence as, $\mathbf{X}_q = \{\mathbf{x}_{q,1}, \mathbf{x}_{q,2}, \ldots, \mathbf{x}_{q,n}\}$, and $\mathbf{X}_g = \{\mathbf{x}_{g,1}, \mathbf{x}_{g,2}, \ldots, \mathbf{x}_{g,m}\}$, with $\mathbf{x}_{i,k} \in \mathcal{F} = \mathbb{R}^{3\times H\times W}$ being an RGB frame. We are interested in learning a function $\psi_\Theta : \mathcal{F}^n \times \mathcal{F}^m \to \mathbb{R}$, with parameters Θ, which takes as input two sequences, $\mathbf{X}_q, \mathbf{X}_g$ and outputs a real-valued similarity between them $\psi_\Theta(\mathbf{X}_q, \mathbf{X}_g)$, where a high (low) similarity indicates that they are (not) of the same person.

3.1. Learning Clip Similarity Aggregation

The similarity function we propose is based on a learned aggregation of clip-pairs sampled from the video sequences. Fig. 2 gives a full block diagram of our method. We uniformly sample M clips of length L from both the query and the gallery sequences, denoted by $\{\mathbf{s}_q^1, \ldots, \mathbf{s}_q^M\}$ and $\{\mathbf{s}_g^1, \ldots, \mathbf{s}_g^M\}$, where, $\mathbf{s}_q^i, \mathbf{s}_g^i \in \mathbb{R}^{L\times 3\times H\times W}$. The number of clips could also be different for the two sequences being compared, but for brevity and implementation ease we keep them to be the same, allowing potential overlap of the clips if the number of frames in the sequence(s) is less than ML.

We first forward pass the clips through a state-of-the-art 3D CNN $f_\xi(\cdot)$ with parameters ξ to obtain D-dimensional features $\mathbf{x}_q \to \{\mathbf{f}_q^1, \ldots, \mathbf{f}_q^M\}$ and $\mathbf{x}_g \to \{\mathbf{f}_g^1, \ldots, \mathbf{f}_g^M\}$, where, $\mathbf{f}_q^i = f_\xi(\mathbf{s}_q^i)$, $\mathbf{f}_g^i = f_\xi(\mathbf{s}_g^i)$ and $\mathbf{f}_q^i, \mathbf{f}_g^i \in \mathbb{R}^D$. We then learn to estimate which pairs of clips are informative, considering all the M^2 combinations. This is in contrast to many sequence modeling approaches, like those based on max/average pooling [44] or attention-based temporal pooling [47], which encode the clip sequences individually with the intuition that some clips might be bad due to occlusion, difficult pose or high motion blur, etc. In our case, we argue that even if some clips have partial artefacts, due to the various nuisance factors, they might still match with a similarly (partially) corrupted clip from another video, and thus should not be discarded. Hence, in the proposed method we consider all the quadratic combinations of pairs of clips and learn to weight them according to their importance. We run the importance estimation in a sequential manner and condition on the information that we have already accumulated at any step t. We estimate the importance score of the clip pair at step t, α_t, using a small neural network $g_\theta(\cdot)$ which takes as input the difference of the aggregated representation r_t till that point and the combined representation c_t of current clip pair. The combined representation used for a pair of clips is an element-wise dot product (denoted as $\odot$) of the clip features, and the pooling process, at step $t = 2, \ldots, M^2$ is given by,

$$\mathbf{c}_t = \mathbf{f}_{q,t} \odot \mathbf{f}_{g,t}, \quad \alpha_t = g_\theta(\mathbf{r}_{t-1} - \mathbf{c}_t) \tag{1}$$

$$\mathbf{r}_t = \frac{1}{\mathcal{A}_t}\left\{\left(\sum_{i=1}^{t-1}\alpha_i\right)\mathbf{r}_{t-1} + \alpha_t\mathbf{c}_t\right\} = \frac{1}{\mathcal{A}_t}\sum_{i=1}^{t}\alpha_i\mathbf{c}_i \tag{2}$$

with, $\mathcal{A}_t = \sum_{i=1}^{t}\alpha_i, \mathbf{r}_1 = \mathbf{f}_{q,1} \odot \mathbf{f}_{g,1}$. This gives the final combined representation $\mathbf{r}_{qg} = \mathbf{r}_{M^2}$.

We then predict the similarity score between $\mathbf{x}_q$ and $\mathbf{x}_g$ by taking an average of all clip-pair cosine similarities weighted according to the importance scores,

$$s(\mathbf{x}_q, \mathbf{x}_g) = \frac{1}{\sum_{t=1}^{M^2}\alpha_t}\sum_{t=1}^{M^2}\alpha_t\frac{\mathbf{f}_{q,t}\cdot\mathbf{f}_{g,t}}{\|\mathbf{f}_{q,t}\|_2\,\|\mathbf{f}_{g,t}\|_2} \tag{3}$$

If the clip features $\mathbf{f}_{q,t}$ and $\mathbf{f}_{g,t}$ are ℓ_2-normalized, then the final similarity (3) can be directly computed using the final combined representation $\mathbf{r}_{qg}$ as

$$s(\mathbf{x}_q, \mathbf{x}_g) = \sum_{l=1}^{D} r_{qg}^l, \tag{4}$$

where $\mathbf{r}_{qg} = \left[r_{qg}^1, \ldots, r_{qg}^D\right]$. The expression (4) can be obtained from (3), with $\mathbf{c}_t = \left[c_t^1, \ldots, c_t^D\right]$, as follows,

$$s(\mathbf{x}_q, \mathbf{x}_g) = \frac{1}{\sum_{t=1}^{T}\alpha_t}\sum_{t=1}^{T}\alpha_t(\mathbf{f}_{q,t}\cdot\mathbf{f}_{g,t}) = \frac{1}{\sum_{t=1}^{T}\alpha_t}\sum_{t=1}^{T}\left(\alpha_t\sum_{d=1}^{D}c_t^d\right) \tag{5}$$

$$= \sum_{d=1}^{D}\left(\frac{1}{\sum_{t=1}^{T}\alpha_t}\sum_{t=1}^{T}\alpha_t c_t^d\right) = \sum_{l=1}^{D} r_{qg}^l. \tag{6}$$

3.2. Learning

Our method allows us to learn all the parameters, $\Theta = (\xi, \theta)$ end-to-end and jointly for the task using the standard backpropagation algorithm for neural networks. However, due to computational constraints, we design the training as a two-step process. First, we learn the parameters of 3D CNNs, then we fix the 3D CNNs and learn the clip-similarity aggregation module parameters. We now describe each of these steps.

3D CNN. In each training iteration, following [11], we randomly sample a batch of PK sequences belonging to P person identities with K sequences from each identity. Then, we randomly sample one clip of length L frames from each sampled sequence to form the mini-batch. We use a combination of the hard mining triplet loss [11] and

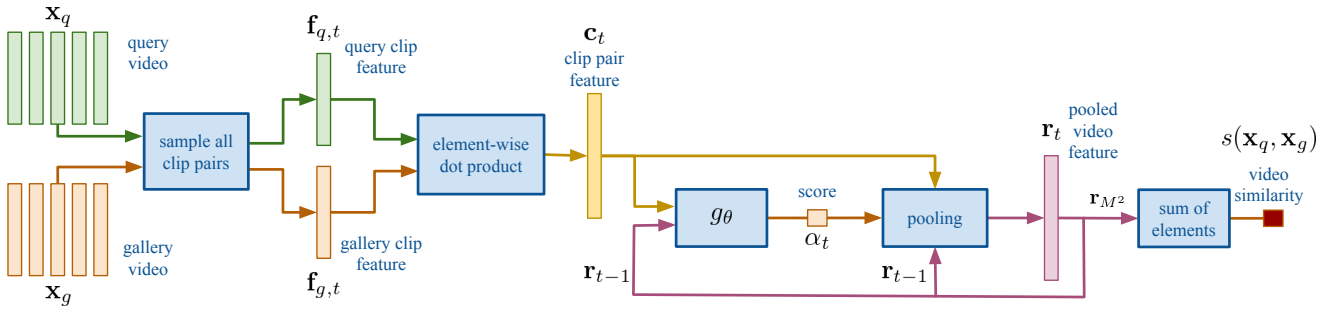

Figure 2: The block diagram of the proposed video similarity estimation method. The video sequences are first split into clips, which are combined to give a clip features. The combined clip features are then pooled with an importance score as a weight. The final pooled representation vector is then used to compute the similarity.

the cross-entropy loss as our objective, $\mathcal{L}(\xi) = \mathcal{L}_{\text{triplet}}(\xi) + \mathcal{L}_{\text{softmax}}(\xi)$.

The hard mining triplet loss is given as, $\mathcal{L}_{\text{triplet}}(\xi) =$

$$\sum_{i=1}^{P}\sum_{a=1}^{K}\left[m + \max_{p=1,\dots,K} d(\mathbf{x}_{a,i}, \mathbf{x}_{p,i}) - \min_{\substack{j=1,\dots,P \\ n=1,\dots,K \\ j\neq i}} d(\mathbf{x}_{a,i}, \mathbf{x}_{n,j})\right]_+, \quad (7)$$

where, $d(\mathbf{x}_1, \mathbf{x}_2) = \|\mathbf{x}_1 - \mathbf{x}_2\|_2$, $\mathbf{x}_{k,i}$ is the 3D CNN feature vector of the k-th clip of the i-th person in the batch, and m is the margin, and $[\cdot]_+ = \max(0, \cdot)$.

We add a classification layer on top of our 3D-CNN network with C classes, where C is the total number of identities in the training set. Let $\{\mathbf{w}_1, \dots, \mathbf{w}_C\}$ be the weights of the classification layer. The softmax cross-entropy loss is given by,

$$\mathcal{L}_{\text{softmax}}(\xi) = -\sum_{i=1}^{P}\sum_{k=1}^{K}\left[\log \frac{\exp(\mathbf{w}_{y^i} \cdot \mathbf{f}_{k,i})}{\sum_{c=1}^{C}\exp(\mathbf{w}_c \cdot \mathbf{f}_{k,i})}\right], \quad (8)$$

where y^i is the person index of the i-th person in the batch. Note that, while learning 3D CNN parameters, ξ, we do not use our clip-similarity aggregation module.

Clip similarity aggregation module. For learning θ, we use the same batch sampling process as described for the learning of 3D CNN parameters ξ, except now we uniformly sample M clips of length L instead of a single clip from each sampled sequence. We extract features $\mathbf{x}_i$ of the clips, with the above learned 3D CNNs, and normalize them. Then, we compute the similarity scores between all pairs of sequences in the batch using (4). We use the hard mining triplet loss similar to (7) as the objective, with the euclidean distances replaced by negative clip similarities as defined above in (3)–(6).

4. Experiments and Results

4.1. Datasets

MARS. The MARS dataset [44] is a large scale video-based person re-identification benchmark. It contains 20,478 pedestrian sequences belonging to 1261 identities. The sequences are automatically extracted using the DPM pedestrian detector [8] and GMMCP tracker [5]. The lengths of the sequences range from 2 to 920 frames. The videos are captured from six cameras and each identity is captured from at least two cameras. The training set consists of 8,298 sequences from 625 identities while the remaining 12,180 sequences from 636 identities make up the test set which consists of a query and a gallery set.

DukeMTMC-VideoReID. The DukeMTMC-VideoReID [28, 38] is another large benchmark of video-based person re-identification. It consists of 702 identities for training, 702 identities for testing. The gallery set contains additional 408 identities as distractors. There is a total of 2,196 sequences for training and 2,636 sequences for testing and distraction. Each sequence has 168 frames on average.

PRID2011. PRID2011 dataset [12] contains 400 sequences of 200 person identities captured from two cameras. Each image sequence has a length of 5 to 675 frames. Following the evaluation protocol from [37, 44], we discard sequences shorter than 21 frames and use 178 sequences from the remaining for training and rest 178 sequences for testing.

4.2. Implementation Details

3D CNN Architecture. We use the PyTorch implementation[1] of the Inception-V1 I3D network [2] pre-trained on the Kinetics action recognition dataset. We remove the final classification layer from the I3D network and replace the original average pooling layer of kernel $2 \times 7 \times 7$ with a global average pooling layer. The resulting I3D network takes an input clip of size $L \times 3 \times 256 \times 128$ and outputs a 1024-dimensional feature vector ($D = 1024$).

Clip similarity aggregation module architecture. The clip-pair similarity aggregation module takes as input a pair of tensors $(M \times D, M \times D)$ representing I3D features of M clips sampled from the two sequences to be matched. In

[1] https://github.com/piergiaj/pytorch-i3d

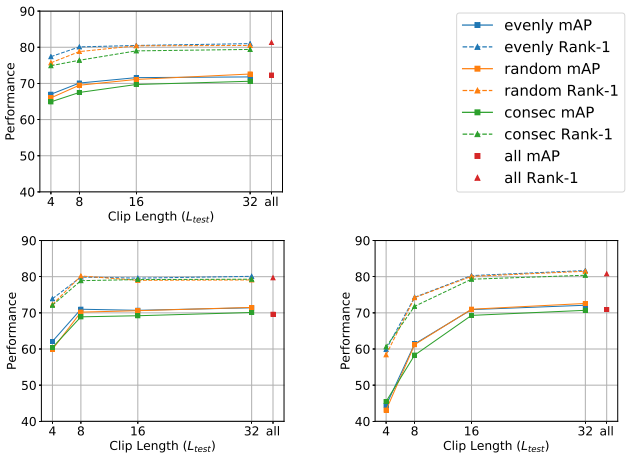

Figure 3: Effect of frame sampling methods and test clip length on MARS. $L_{\text{train}} = 4, 8, 16$ counterclockwise.

our experiments, we set the number of clips M to 8 and the clip length L to 4 frames, this setting was faster than higher L and smaller M while giving similar performance (kindly see the supplementary document for complete ablation experiment). The importance scoring function $g_\theta(\cdot)$ consists of two hidden layers with 1024 units in both layers. The output layer has a single unit that represents the estimated importance score. The hidden layers have the ReLU activation function while the output layer has the softplus activation function, $\sigma(x) = \log(1 + \exp(x))$. The softplus function, a smooth approximation of ReLU function, constraints the importance score to always be positive. We also use a dropout layer [33] with dropout probability 0.5 and a batch normalization layer [14] after both hidden layers.

Training details. Due to lack of space, we include the complete training details of the 3D CNN and the Clip Similarity Aggregation module in the supplementary document.

Evaluation protocol and evaluation metrics. We follow the experimental setup of [37], [44] and [38] for PRID2011, MARS and DukeMTMC-VideoReID respectively. For MARS and DukeMTMC-VideoReID, we use the train/test split provided by [44] and [38], respectively. For PRID2011, we average the re-identification performance over 10 random train/test splits. We report the re-identification performance using CMC (cumulative matching characteristics) at selected ranks and mAP (mean average precision).

4.3. Analysis of I3D Features for Re-Identification

Frame sampling method and clip length. In the scenario, where we use a single clip to represent a sequence, it becomes important how we sample the frames from the sequence to form a clip. In this experiment, we explore multiple frame sampling methods given in Tab. 1 and their effect

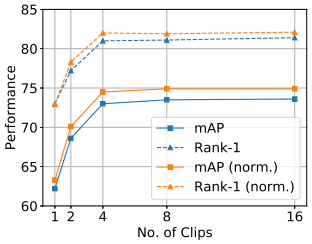

Figure 4: Test performance (MARS) with averaging of I3D features of multiple clips, with and without ℓ_2-normalization.

`consec`	Randomly sample a clip of L consecutive frames
`random`	Randomly sample L frames (arrange in order)
`evenly`	Sample L frames uniformly
`all`	Take all frames

Table 1: Frame sampling methods for clip construction

on re-identification performance.

Note that, all sampling methods in Tab. 1 result in a clip of length L except the `all` sampling method. We train three I3D models with different clip lengths, $L_{\text{train}} \in \{4, 8, 16\}$. The frames are sampled consecutively (`consec`) to form a clip during training. During the evaluation, each test sequence is represented by I3D features of a single clip sampled in one of the ways described above. Given a query, the gallery sequences are ranked based on the distances of their I3D features. We evaluate the three models with different frame sampling methods and test clip lengths $L_{\text{test}} \in \{4, 8, 16, 32\}$.

Figure 3 shows the plots of re-identification performance as a function of L_{test} with different frame sampling methods. We observe that the performance improves as we increase the clip-length during testing, although with diminishing returns. We also observe that when tested on longer clips (e.g. $L_{\text{test}} = 16, 32$), models trained on different clip-lengths ($L_{\text{train}} = 4, 8, 16$) show similar performance to each other. However, when tested on shorter clips (e.g. $L_{\text{test}} = 4$), a model trained on shorter clips performs better than the model trained on longer clips.

The sampling methods `random` and `evenly` perform better than the `consec` sampling method, especially for smaller clip lengths. This can be explained by the fact that `random` and `evenly` have larger temporal extents than `consec` and do not rely on frames only from a narrow temporal region which could be non-informative because of a difficult pose, occlusion, etc.

Averaging features of multiple clips. Since sequences in the MARS dataset can be up to 920 frames long, using single short clips to represent these sequences is not optimal. In this experiment, we take the average of the I3D features of multiple clips evenly sampled from the original sequence to represent these sequences. We vary the number of clips in $\{1, 2, 4, 8, 16\}$ on the MARS dataset. We use the model trained with $L_{\text{train}} = 4$ and we keep the same clip length $L_{\text{test}} = 4$ during the evaluation. We also evaluate with and without the ℓ_2-normalization of clip-features. Figure 4

Figure 5: Example of an uncorrupted and a corrupted clip.

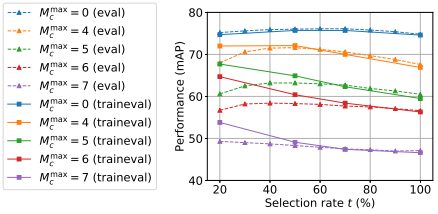

Figure 6: MARS test mAP vs. selection rate (t) for top-$t\%$ clip-similarity aggregation.

shows the test re-identification performance for different number of clips with and without ℓ_2-normalization of clip-features. We observe that averaging features from multiple clips significantly improves the re-id performance. The performance improves up to around 8 clips beyond which there is little improvement. We also find that ℓ_2-normalization of clip features leads to consistent improvement in performance.

4.4. Evaluation of Learned Clip Similarity Aggregation on MARS

In this section, we present the re-identification performance results of our learned clip similarity aggregation method on the MARS test set. We also investigate the robustness of our method by evaluating it with varying degrees of input corruption. We randomly corrupt clips during training and evaluation as follows. For every training or test sequence $\mathbf{x}$, we first randomly pick a number $M_c(\mathbf{x})$ with $0 \leq M_c(\mathbf{x}) \leq M_c^{\max}$. Here, $M_c^{\max}$ denotes the maximum number of corrupt clips in a sequence with $M_c^{\max} \leq M$. Next, we apply a corruption transformation function to randomly selected $M_c(\mathbf{x})$ of the M clips sampled from the sequence $\mathbf{x}$. The corruption transformation function consists of first scaling of every frame in the clip down by a factor of 5, JPEG compression of resulting scaled-down frames, and finally rescaling of the frames up to the original size. Figure 5 shows examples of uncorrupted and corrupted clips.

Let $\{\mathbf{f}_q^1, \ldots, \mathbf{f}_q^M\}$ and $\{\mathbf{f}_g^1, \ldots, \mathbf{f}_g^M\}$ be the ℓ_2-normalized I3D features of M clips sampled from a query sequence $\mathbf{x}_q$ and a gallery sequence $\mathbf{x}_q$ respectively. As described in Section 3, the similarity between $\mathbf{x}_q$ and $\mathbf{x}_g$, as estimated by our method, is given by (3) or (4). We train and evaluate our clip-similarity aggregation module for different rates of input corruption. The rate of input corruption is changed via the parameter $M_c^{\max}$. We use the I3D network trained only on uncorrupted clips and keep it fixed throughout the experiment.

We compare our method with the **top-$t\%$ clip-similarity aggregation** (`top-t%`) baseline, which is based on [3]. It takes $t\%$ of the clip-pairs with the highest similarity and averages their similarities to estimate the overall similarity between the two sequences. By taking only top $t\%$ and not all clip-pairs into account, the resulting similarity becomes more robust and improves re-identification performance [3]. In our implementation, we learn a linear layer that projects the D-dimensional I3D features to a new D-dimensional space. We define the similarity between two given clips as the cosine similarity between their projected I3D features. Let $\{\mathbf{f'}_q^1, \ldots, \mathbf{f'}_q^M\}$ and $\{\mathbf{f'}_g^1, \ldots, \mathbf{f'}_g^M\}$ be the projected clip features and let $\widehat{P}_t(\mathbf{x}_q, \mathbf{x}_g)$ be the set of $t\%$ clip-pairs with the highest similarity. Then, the `top-t%` similarity between the two sequences is given by,

$$s_{\texttt{top-t\%}}(\mathbf{x}_q, \mathbf{x}_g) = \frac{1}{\left|\widehat{P}_t(\mathbf{x}_q, \mathbf{x}_g)\right|} \sum_{\substack{(i,j) \in \\ \widehat{P}_t(\mathbf{x}_q, \mathbf{x}_g)}} \frac{\mathbf{f'}_q^i \cdot \mathbf{f'}_g^j}{\left\|\mathbf{f'}_q^i\right\|_2 \left\|\mathbf{f'}_g^i\right\|_2}. \quad (9)$$

We implement two variants of this method. In the first variant `top-t%-eval`, we perform the top-$t\%$ similarity aggregation only during the evaluation. In the second variant, `top-t%-traineval`, we perform the top-$t\%$ similarity aggregation during the evaluation as well as during the training. This means that the loss gradients are back-propagated only for the clips that are included in the top $t\%$ of the clip-pairs.

Figure 6 shows the test re-identification performance vs t plots for `top-t%-eval` and `top-t%-eval` respectively with different values of $M_c^{\max}$. As expected, the re-identification performance deteriorates as the value of $M_c^{\max}$ is increased. We also observe that top-$t\%$ aggregation during training significantly improves the re-identification performance, especially with the smaller selection rates.

Table 2 shows the re-identification performance of our method and the baselines on the MARS test set. Our method has comparable performance to the top-$t\%$ clip-similarity aggregation when the corruption rate is low i.e. $M_c^{\max}$ is small. However, it significantly outperforms the top-$t\%$ clip-similarity aggregation baseline for higher rates of input corruption, *e.g.* for $M_c^{\max} = 7$, the maximum mAP for the baseline `topt-e` is 49.3 (for $t = 20\%$), while our method degrades more gracefully to give 69.6 mAP. This highlights the advantage of the proposed method for learning of clip similarity aggregation.

4.5. Comparison with the state-of-the-art

In Table 3, we compare our method with the state-of-the-art techniques on MARS dataset. Our method achieves 75.9% mAP and 82.7% Rank-1 accuracy. In terms of mAP, our method is on par with all the methods, except for the recently published visual distributional representation based method of Hu and Hauptmann [13], who achieve an mAP of

Method	t (%)	mAP								Rank-1							
		$M_c^{\max}$								$M_c^{\max}$							
		0	1	2	3	4	5	6	7	0	1	2	3	4	5	6	7
`topt-e`	10	73.9	72.6	71.3	68.5	63.9	57.0	53.5	49.2	78.9	78.3	78.0	76.9	73.0	67.5	64.7	62.4
	20	75.2	74.7	73.7	72.3	68.0	60.6	56.7	49.3	80.9	80.7	80.9	80.0	76.8	70.7	67.6	62.1
	30	75.6	75.1	74.6	73.7	70.6	62.5	58.2	49.0	81.2	81.1	81.6	81.1	79.7	73.7	69.3	61.8
	40	75.9	75.3	75.0	74.1	71.5	63.2	58.4	48.7	81.8	81.5	82.6	81.6	80.2	73.9	70.1	61.5
	50	76.0	**75.4**	75.1	74.3	71.6	63.2	58.3	48.3	82.5	81.5	82.6	82.2	**80.3**	73.9	70.0	61.3
	60	**76.1**	**75.4**	74.8	74.1	71.2	63.0	58.1	47.9	**83.2**	**82.3**	82.2	82.2	79.7	72.9	70.1	60.7
	70	**76.1**	75.2	74.5	73.4	70.6	62.7	57.7	47.5	**83.2**	82.2	82.4	81.4	79.2	72.8	69.6	60.4
	80	75.8	75.0	73.9	72.7	69.7	61.9	57.5	47.3	82.9	82.1	82.2	81.0	78.6	72.2	69.4	60.2
	90	75.4	74.5	73.3	71.9	68.8	61.3	57.1	47.0	82.6	82.0	82.0	80.4	77.6	71.9	68.5	60.4
	100	74.8	73.6	72.5	70.9	67.6	60.5	56.6	47.1	82.2	81.2	81.1	79.4	76.8	71.5	67.9	59.9
`topt-te`	20	74.7	74.7	74.4	74.1	72.0	67.7	64.7	53.8	80.5	80.7	80.5	80.9	79.1	77.1	74.9	67.1
	50	75.7	**75.4**	**75.2**	74.7	72.1	64.9	60.4	49.1	82.2	81.6	82.6	**82.4**	79.9	75.0	72.0	63.0
	70	75.7	75.3	74.4	73.3	70.0	62.3	58.4	47.4	82.9	82.3	**82.7**	81.8	78.5	73.5	70.6	60.7
	100	74.6	73.4	72.3	70.5	66.9	59.5	56.3	46.6	82.5	81.0	81.7	79.0	76.7	70.7	67.9	59.9
Ours	n/a	75.9	**75.4**	**75.2**	**75.2**	**74.4**	**73.7**	**73.1**	**69.9**	82.7	81.4	81.4	81.5	79.8	**80.0**	**80.3**	**78.6**

Table 2: The MARS test performance (mAP and rank-1 accuracy) of our method learned clip similarity aggregation and of the top-t% aggregation baseline. The blocks of rows labeled `topt-e` and `topt-te` show the results of the `top-t%-eval` and `top-t%-traineval` variants of the baseline top-t% clip-similarity aggregation, respectively.

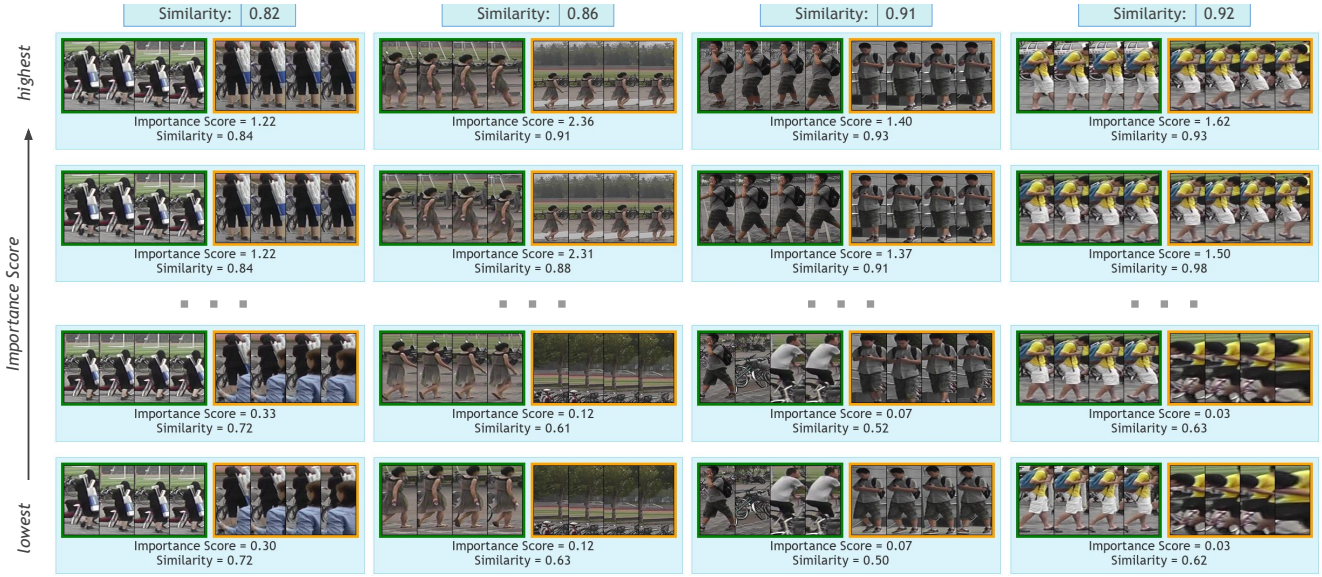

Figure 7: Each column gives an example of different clips from query and gallery videos, where in each row, the left clip is a query clip (green outline) and the right one is a gallery clip (orange outline). Notice how the method predicts low importance score when the query clip is very different from the gallery clip, and thus effectively ignores the pair even if the similarity is predicted to be non trivially high by the feature matching.

81.8%, which is significantly higher than ours (we discuss below). In terms of mAP performance, our method is very close to the part-aligned bilinear representations (PABR) [35] and CSSA-CSE + Flow [3]. However, the performance of [3] is much lower than ours when optical flow is not used (see CSSA-CSE in Table 3). Among methods that use 3D CNNs as their backbone (marked * in Table 3), our method achieves the best mAP performance.

Table 4 shows the comparison of our method with the state-of-the-art on the DukeMTMC-VideoReID dataset. There are only a few works with results on this dataset.We achieve 88.5% mAP and 89.3% Rank-1 accuracy, which is significantly better than the baseline presented in [38]. However, the performance of Hu and Hauptmann [13] and [9] is better than our method.

Comparing our method to very recent works such as that of Hu and Hauptmann [13], we note that their method is significantly more costly than ours in terms of gallery storage requirements, and uses a network that is deeper than ours. While we use a 3D CNN with 22 layers, they use an image-based DenseNet CNN with 121 layers. They compare the test video with the gallery videos by estimating the Wasserstein distance between the densities estimated using KDE. This requires them to use (and save) all the frames to make the inference. While in our case, we use a limited number of clip features (~ 8) per video. While such an ac-

Model	mAP	R1	R5	R20
RQEN+XQDA+Reranking (2018 [32])	71.1	77.8	88.8	94.3
TriNet + Reranking (2017 [11])	77.4	81.2	90.8	-
DuATM (2018 [30])	67.7	81.1	92.5	-
MGCAN-Siamese (2018 [31])	71.2	77.2	-	-
PSE (2018 [29])	56.9	72.1	-	-
PSE + ECN (2018 [29])	71.8	76.7	-	-
RRU + STIM (2018 [23]) *	72.7	84.4	93.2	96.3
Two-Stream M3D (2018 [17]) *	74.1	84.4	93.8	97.7
PABR (2018 [35])	75.9	84.7	94.4	97.5
PABR + Reranking (2018 [35])	83.9	85.1	94.2	97.4
CSSA-CSE + Flow (2018 [3])	76.1	86.3	94.7	98.2
STA (2019 [9])	80.8	86.3	95.7	98.1
STA + Reranking (2019 [9])	87.7	87.2	96.2	98.6
D + GE + D_G (2019 [13])	81.8	87.3	96.0	98.1
Ours *	75.9	82.7	94.0	97.2
Ours + Reranking *	83.3	83.4	93.4	97.4

Table 3: Comparison of our model with the state-of-the-art re-identification methods on MARS dataset. Entries in grey represent the models that use re-ranking. Models marked * use 3D CNNs as their backbone.

Model	mAP	R1	R5	R20
ETAP-Net [Supervised] (2018 [38])	78.3	83.6	94.6	97.6
STA (2019 [9])	94.9	96.2	99.3	99.6
R + GE + D_G (2019 [13])	94.9	95.6	99.3	99.9
Ours	88.5	89.3	98.3	99.4

Table 4: Comparison of our model with the state-of-the-art methods on DukeMTMC-VideoReID dataset.

Model	R1	R5	R20
CNN + XQDA (2016 [44])	77.3	93.5	99.3
E2E AMOC+EpicFlow (2017 [21])	83.7	98.3	100.0
QAN (2017 [22])	90.3	98.2	100.0
M3D+RAL (2018 [17])	91.0	-	-
CSSA-CSE (2018 [3])	88.6	99.1	-
Ours	82.9	95.8	99.1
Two-Stream M3D (2018)	94.4	100.0	-
CSSA-CSE + Flow (2018)	93.0	99.3	100.0

Table 5: Comparison with state-of-the-art re-id methods on PRID2011 dataset.

curate method achieves higher performance, it comes at a significant cost.

STA [9] is another recent method with state-of-the-art performance. While STA focuses on aggregating features effectively from a small set of input frames (4-8 frames), our method is more focused on predicting the overall similarities between two long sequences while relying on I3D for clip-level features (a clip is typically 4-16 frames long). Since the video benchmarks contain much longer sequences, our method can be used in conjunction with [9] to further boost the performance as it is complementary to it.

In Table 5, we show results on the PRID2011 dataset. Unfortunately being a video-based end-to-end method, our method seems to overfit severely on the dataset. PRID2011 dataset has only 178 videos from training cf. 8,298 in MARS. We see that we are still comparable with initial CNN based methods (eg. CNN+XQDA [44]). The more recent methods seem to utilize optical flow as input, which could be leading to some regularization by removing the appearance from the videos.

4.6. Qualitative Results

Figure 7 shows four examples of pairs of query-gallery sequences and the similarity between them as predicted by our method. For each example, we also show two clip pairs (4 frames each) with the highest importance scores and two with the lowest importance score. One of the clips in the bottom clip-pair has, from left to right, (i) a significant amount of occlusion, (ii) no person in the frame, (iii) different persons in different frames due to a tracking error, and (iv) an improperly cropped person due to poor bounding box estimation. Our method learns to correctly identify the clip pairs that are unreliable for estimating the overall similarity between the two video sequences and gives them very low importance scores (bottom two rows). Our method gives an overall high similarity (the heading of each column) to all the examples shown in Figure 7 by minimizing the effect of bad clip-pairs. Although the MARS dataset considers the gallery sequences in column 2 and 4 as distractors, the high similarity estimated by our method is reasonable since they contain the same person as in the query in many of their frames, and are annotation edge cases.

These qualitative results highlight the ability of the proposed method to identify and match reliable clip pairs, and filter out unreliable ones despite non-trivial appearance similarities estimated by the base network.

5. Conclusion

We addressed the video-based person re-identification task and, to the best of our knowledge, showed that 3D CNNs can be used competitively for the task. We demonstrated better performance with 3D CNN on RGB images only, cf., existing methods that use optical flow in addition to RGB frames. This is indicative of the fact that 3D CNNs are capable of capturing necessary motion cues relevant to the task of video-based person re-identification.

Further, we proposed a novel clip similarity learning method that identifies clip pairs which are informative for correlating the two clips. While previous methods used ad-hoc approaches to obtain such pairs, we showed that our method is capable of learning to do so. We showed with the simulated partial corruption of the input clips, that the proposed method is robust to nuisances which might occur as a result of motion blur or partial occlusions. We also verified the intuition used to develop the method qualitatively.

The proposed method can be seen as an approximate discriminative mode-matching method. There have been recent works using deeper CNN models (121 layers cf. 22 here) and more accurate distribution matching that obtain better results than the proposed method, however, they come at a computational and storage cost. A future work would systematically find the balance between the two approaches to obtain the best performance for a given budget.

References

[1] E. Ahmed, M. Jones, and T. K. Marks. An improved deep learning architecture for person re-identification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3908–3916, 2015.

[2] J. Carreira and A. 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] D. Chen, H. Li, T. Xiao, S. Yi, and X. 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] D. Cheng, Y. Gong, S. Zhou, J. Wang, and N. Zheng. Person re-identification by multi-channel parts-based cnn with improved triplet loss function. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1335–1344, 2016.

[5] A. Dehghan, S. Modiri Assari, and M. Shah. Gmmcp tracker: Globally optimal generalized maximum multi clique problem for multiple object tracking. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4091–4099, 2015.

[6] S. Ding, L. Lin, G. Wang, and H. Chao. Deep feature learning with relative distance comparison for person re-identification. *Pattern Recognition*, 48(10):2993–3003, 2015.

[7] M. Farenzena, L. Bazzani, A. Perina, V. Murino, and M. Cristani. Person re-identification by symmetry-driven accumulation of local features. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 2360–2367. IEEE, 2010.

[8] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part-based models. *IEEE transactions on pattern analysis and machine intelligence*, 32(9):1627–1645, 2010.

[9] Y. Fu, X. Wang, Y. Wei, and T. Huang. Sta: Spatial-temporal attention for large-scale video-based person re-identification. In *Proceedings of the Association for the Advancement of Artificial Intelligence*. 2019.

[10] D. Gray and H. Tao. Viewpoint invariant pedestrian recognition with an ensemble of localized features. In *European conference on computer vision*, pages 262–275. Springer, 2008.

[11] A. Hermans, L. Beyer, and B. Leibe. In defense of the triplet loss for person re-identification. *arXiv preprint arXiv:1703.07737*, 2017.

[12] M. Hirzer, C. Beleznai, P. M. Roth, and H. Bischof. Person Re-Identification by Descriptive and Discriminative Classification. In *Proc. Scandinavian Conference on Image Analysis (SCIA)*, 2011.

[13] T.-Y. Hu and A. G. Hauptmann. Multi-shot person re-identification through set distance with visual distributional representation. In *Proceedings of the 2019 on International Conference on Multimedia Retrieval*, pages 262–270. ACM, 2019.

[14] S. Ioffe and C. Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167*, 2015.

[15] M. Koestinger, M. Hirzer, P. Wohlhart, P. M. Roth, and H. Bischof. Large scale metric learning from equivalence constraints. In *2012 IEEE conference on computer vision and pattern recognition*, pages 2288–2295. IEEE, 2012.

[16] I. Kviatkovsky, A. Adam, and E. Rivlin. Color invariants for person reidentification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(7):1622–1634, 2013.

[17] J. Li, S. Zhang, and T. Huang. Multi-scale 3d convolution network for video based person re-identification. *arXiv preprint arXiv:1811.07468*, 2018.

[18] S. Li, S. Bak, P. Carr, and X. 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.

[19] W. Li, R. Zhao, T. Xiao, and X. Wang. Deepreid: Deep filter pairing neural network for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 152–159, 2014.

[20] S. Liao and S. Z. Li. Efficient psd constrained asymmetric metric learning for person re-identification. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3685–3693, 2015.

[21] H. Liu, Z. Jie, K. Jayashree, M. Qi, J. Jiang, S. Yan, and J. Feng. Video-based person re-identification with accumulative motion context. *IEEE transactions on circuits and systems for video technology*, 28(10):2788–2802, 2017.

[22] Y. Liu, J. Yan, and W. 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.

[23] Y. Liu, Z. Yuan, W. Zhou, and H. Li. Spatial and temporal mutual promotion for video-based person re-identification. *arXiv preprint arXiv:1812.10305*, 2018.

[24] B. Ma, Y. Su, and F. Jurie. Local descriptors encoded by fisher vectors for person re-identification. In *European Conference on Computer Vision*, pages 413–422. Springer, 2012.

[25] N. McLaughlin, J. Martinez del Rincon, and P. 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.

[26] S. Paisitkriangkrai, C. Shen, and A. Van Den Hengel. Learning to rank in person re-identification with metric ensembles. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1846–1855, 2015.

[27] B. J. Prosser, W.-S. Zheng, S. Gong, T. Xiang, and Q. Mary. Person re-identification by support vector ranking. In *BMVC*, volume 2, page 6, 2010.

[28] E. Ristani, F. Solera, R. Zou, R. Cucchiara, and C. Tomasi. Performance measures and a data set for multi-target, multi-camera tracking. In *European Conference on Computer Vision workshop on Benchmarking Multi-Target Tracking*, 2016.

[29] M. Saquib Sarfraz, A. Schumann, A. Eberle, and R. Stiefelhagen. A pose-sensitive embedding for person re-identification with expanded cross neighborhood re-ranking. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 420–429, 2018.

[30] J. Si, H. Zhang, C.-G. Li, J. Kuen, X. Kong, A. C. Kot, and G. Wang. Dual attention matching network for context-aware feature sequence based person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5363–5372, 2018.

[31] C. Song, Y. Huang, W. Ouyang, and L. Wang. Mask-guided contrastive attention model for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1179–1188, 2018.

[32] G. Song, B. Leng, Y. Liu, C. Hetang, and S. Cai. Region-based quality estimation network for large-scale person re-identification. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[33] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research*, 15(1):1929–1958, 2014.

[34] C. Su, J. Li, S. Zhang, J. Xing, W. Gao, and Q. Tian. Pose-driven deep convolutional model for person re-identification. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3960–3969, 2017.

[35] Y. Suh, J. Wang, S. Tang, T. Mei, and K. Mu Lee. Part-aligned bilinear representations for person re-identification. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 402–419, 2018.

[36] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri. Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 4489–4497, 2015.

[37] T. Wang, S. Gong, X. Zhu, and S. Wang. Person re-identification by video ranking. In *European Conference on Computer Vision*, pages 688–703. Springer, 2014.

[38] Y. Wu, Y. Lin, X. Dong, Y. Yan, W. Ouyang, and Y. Yang. Exploit the unknown gradually: One-shot video-based person re-identification by stepwise learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5177–5186, 2018.

[39] T. Xiao, H. Li, W. Ouyang, and X. Wang. Learning deep feature representations with domain guided dropout for person re-identification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1249–1258, 2016.

[40] T. Xiao, S. Li, B. Wang, L. Lin, and X. Wang. Joint detection and identification feature learning for person search. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3415–3424, 2017.

[41] Y. Yan, B. Ni, Z. Song, C. Ma, Y. Yan, and X. Yang. Person re-identification via recurrent feature aggregation. In *European Conference on Computer Vision*, pages 701–716. Springer, 2016.

[42] H. Zhao, M. Tian, S. Sun, J. Shao, J. Yan, S. Yi, X. Wang, and X. Tang. Spindle net: Person re-identification with human body region guided feature decomposition and fusion. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1077–1085, 2017.

[43] R. Zhao, W. Ouyang, and X. Wang. Unsupervised salience learning for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3586–3593, 2013.

[44] L. Zheng, Z. Bie, Y. Sun, J. Wang, C. Su, S. Wang, and Q. Tian. Mars: A video benchmark for large-scale person re-identification. In *European Conference on Computer Vision*, pages 868–884. Springer, 2016.

[45] L. Zheng, Y. Huang, H. Lu, and Y. Yang. Pose invariant embedding for deep person re-identification. *arXiv preprint arXiv:1701.07732*, 2017.

[46] W.-S. Zheng, S. Gong, and T. Xiang. Person re-identification by probabilistic relative distance comparison. In *CVPR 2011*, pages 649–656. IEEE, 2011.

[47] Z. Zhou, Y. Huang, W. Wang, L. Wang, and T. 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.