Diversity Regularized Spatiotemporal Attention
for Video-based Person Re-identification

Shuang Li[1*]   Slawomir Bak[2*]   Peter Carr[2*]   Xiaogang Wang[1]

[1]CUHK-SenseTime Joint Lab,
The Chinese University of Hong Kong
{sli,xgwang}@ee.cuhk.edu.hk

[2]Argo AI

{sbak,pcarr}@argo.ai

Abstract

Video-based person re-identification matches video clips of people across non-overlapping cameras. Most existing methods tackle this problem by encoding each video frame in its entirety and computing an aggregate representation across all frames. In practice, people are often partially occluded, which can corrupt the extracted features. Instead, we propose a new spatiotemporal attention model that automatically discovers a diverse set of distinctive body parts. This allows useful information to be extracted from all frames without succumbing to occlusions and misalignments. The network learns multiple spatial attention models and employs a diversity regularization term to ensure multiple models do not discover the same body part. Features extracted from local image regions are organized by spatial attention model and are combined using temporal attention. As a result, the network learns latent representations of the face, torso and other body parts using the best available image patches from the entire video sequence. Extensive evaluations on three datasets show that our framework outperforms the state-of-the-art approaches by large margins on multiple metrics.

1. Introduction

Person re-identification matches images of pedestrians in one camera with images of pedestrians from another, non-overlapping camera. This task has drawn increasing attention in recent years due to its importance in applications, such as surveillance [42], activity analysis [32] and tracking [47]. It remains a challenging problem because of complex variations in camera viewpoints, human poses, lighting, occlusions, and background clutter.

In this paper, we investigate the problem of video-based

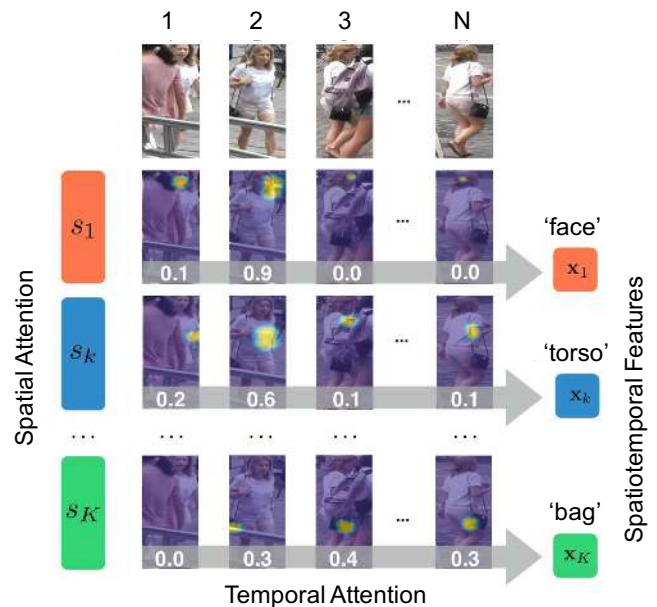*Work done while employed at Disney Research Pittsburgh.

Figure 1. **Spatiotemporal Attention.** In challenging video re-identification scenarios, a person is rarely fully visible in all frames. However, frames in which only part of the person is visible often contain useful information. For example, the face is clearly visible in the frames 1 and 2, the torso in frame 2, and the handbag in frames 2, 3 and N. Instead of averaging full frame features across time, we propose a new spatiotemporal approach which learns to detect a set of K diverse salient image regions within each frame (superimposed heatmaps). An aggregate representation of each body part is then produced by combining the extracted per-frame regions across time (weights shown as white text). Our spatiotemporal approach creates a compact encoding of the video that exploits useful partial information in each frame by leveraging multiple spatial attention models, and combining their outputs using multiple temporal attention models.

person re-identification, which is a generalization of the standard image-based re-identification task. Instead of

matching image pairs, the algorithm must match pairs of video sequences (possibly of different durations). A key challenge in this paradigm is developing a good latent feature representation of each video sequence.

Existing video-based person re-identification methods represent each frame as a feature vector and then compute an aggregate representation across time using average or maximum pooling [52, 28, 46]. Unfortunately, this approach has several drawbacks when applied to datasets where occlusions are frequent (Fig. 1). The feature representation generated for each image is often corrupted by the visual appearances of occluders. However, the remaining visible portions of the person may provide strong cues for re-identification. Assembling an effective representation of a person from these various glimpses should be possible. However, aggregating features across time is not straightforward. A person's pose will change over time, which means any aggregation method must account for spatial misalignment (in addition to occlusion) when comparing features extracted from different frames.

In this paper, we propose a new spatiotemporal attention scheme that effectively handles the difficulties of video-based person re-identification. Instead of directly encoding the whole image (or a predefined decomposition, such as a grid), we use multiple spatial attention models to localize discriminative image regions, and pool these extracted local features across time using temporal attention. Our approach has several useful properties:

- Spatial attention explicitly solves the alignment problem between images, and avoids features from being corrupted by occluded regions.

- Although many discriminative image regions correspond to body parts, accessories like sunglasses, backpacks and hats; are prevalent and useful for re-identification. Because these categories are hard to predefine, we employ an unsupervised learning approach and let the neural network automatically discover a set of discriminative object part detectors (spatial attention models).

- We employ a novel diversity regularization term based on the Hellinger distance to ensure multiple spatial attention models do not discover the same body part.

- We use temporal attention models to compute an aggregate representation of the features extracted by each spatial attention model. These aggregate representations are then concatenated into a final feature vector that represents all of the information available from the entire video.

We demonstrate the effectiveness of our approach on three challenging video re-identification datasets. Our technique out performs the state-of-the-art methods under multiple evaluation metrics.

2. Related Work

Person re-identification was first proposed for multi-camera tracking [42, 38]. Gheissari *et al.* [11] designed a spatial-temporal segmentation method to extract visual cues and employed color and salient edges for foreground detection. This work defined the image-based person re-identification as a specific computer vision task.

Image-based person re-identification mainly focuses on two categories: extracting discriminative features [13, 9, 33, 19, 43] and learning robust metrics [37, 50, 18, 36, 2]. In recent years, researchers have proposed numerous deep learning based methods [1, 24, 8, 20, 44] to jointly handle both aspects. Ahmed *et al.* [1] input a pair of cropped pedestrian images to a specifically designed CNN with a binary verification loss function for person re-identification. In [8], Ding *et al.* minimize feature distances between the same person and maximize the distances among different people by employing a triplet loss function when training deep neural networks. Xiao *et al.* [44] jointly train the pedestrian detection and person re-identification in a single CNN model. They propose an Online Instance Matching loss function which learns features more efficiently in large scale verification problems.

Video-based person re-identification. Video-based person re-identification [35, 52, 46, 41, 53, 34] is an extension of image-based approaches. Instead of pairs of images, the learning algorithm is given pairs of video sequences. In [46], You *et al.* present a top-push distance learning model accompanied by the minimization of intra-class variations to optimize the matching accuracy at the top rank for person re-identification. McLaughlin *et al.* [35] introduce an RNN model to encode temporal information. They utilize temporal pooling to select the maximum activation over each feature dimension and compute the feature similarity of two videos. Wang *et al.* [41] select reliable space-time features from noisy/incomplete image sequences while simultaneously learning a video ranking function. Ma *et al.* [34] encode multiple granularities of spatiotemporal dynamics to generate latent representations for each person. A Time Shift Dynamic Time Warping model is derived to select and match data between inaccurate and incomplete sequences.

Attention models for person re-identification. Attention models [45, 22, 21] have grown in popularity since [45]. Zhou *et al.* [52] combine spatial and temporal information by building an end-to-end deep neural network. An attention model assigns importance scores to input frames according to the hidden states of an RNN. The final feature is a temporal average pooling of the RNN's outputs. However, if trained in this way, corresponding weights at different time steps of the attention model tend to have the

same values. Liu *et al.* [30] proposed a multi-directional attention module to exploit the global and local contents for image-based person re-identification. However, jointly training multiple attentions might cause the mode collapse. The network has to be carefully trained to avoid attention models focusing on similar regions with high redundancy. In this paper, we combine spatial and temporal attentions into spatiotemporal attention models to address the challenges in video-based person re-identification. For spatial attention, we use a penalization term to regularize multiple redundant attentions. We employ temporal attention to assign weights to different salient regions on a per-frame basis to take full advantage of discriminative image regions. Our method demonstrates better empirical performance, and decomposes into an intuitive network architecture.

3. Method

We propose a new deep learning architecture (Fig. 2) to better handle video re-identification by automatically organizing the data into sets of consistent salient subregions. Given an input video sequence, we first use a restricted random sampling strategy to select a subset of video frames (Sec. 3.1). Then we send the selected frames to a multi-region spatial attention module (Sec. 3.2) to generate a diverse set of discriminative spatial gated visual features—each roughly corresponding to a specific salient region of a person (Sec. 3.3). The overall representation of each salient region across the duration of the video is generated using temporal attention (Sec. 3.4). Finally, we concatenate all temporal gated features and send them to a fully-connected layer which represents the latent spatiotemporal encoding of the original input video sequence. An OIM loss function, proposed by Xiao *et al.* [44], is built on top of the FC layer to supervise the training of the whole network in an end-to-end fashion. However, any traditional loss function (like softmax) could also be employed.

3.1. Restricted Random Sampling

Previous video-based person re-identification methods [35, 34, 52] do not model long-range temporal structure because the input video sequences are relatively short. To some degree, this paradigm is only slightly more complicated than image-based re-identification since consecutive video frames are highly correlated, and the visual features extracted from one frame do not change drastically over the course of a short sequence. However, when input video sequences are long, any re-identification methodology must be able to cope with significant visual changes over time, such as different body poses and angles relative to the camera.

Wang *et al.* [39] proposed a temporal segment network to generate video snippets for action recognition. Inspired by them, we propose a restricted random sampling strategy to generate compact representations of long video sequences that still provide good representations of the original data. Our approach enables models to utilize visual information from the entire video and avoids the redundancy between sequential frames. Given an input video $\mathbf{V}$, we divide it into N chunks $\{C_n\}_{n=1,N}$ of equal duration. From each chunk C_n, we randomly sample an image I_n. The video is then represented by the ordered set of sampled frames $\{I_n\}_{n=1,N}$.

3.2. Multiple Spatial Attention Models

We employ multiple spatial attention models to automatically discover salient image regions (body parts or accessories) useful for re-identification. Instead of pre-defining a rigid spatial decomposition of input images (*e.g.* a grid structure), our approach automatically identifies multiple disjoint salient regions in each image that consistently occur across multiple training videos. Because the network learns to identify and localize these regions (e.g. automatically discovering a set of object part detectors), our approach mitigates registration problems that arise from pose changes, variations in scale, and occlusion. Our approach is not limited to detecting human body parts. It can focus on any informative image regions, such as hats, bags and other accessories often found in re-identification datasets. Feature representations directly generated from entire images can easily miss fine-grained visual cues (Fig. 1). Multiple diverse spatial attention models, on the other hand, can simultaneously discover discriminative visual features while reducing the distraction of background contents and occlusions. Although spatial attention is not a new concept, to the best of our knowledge, this is first time that a network has been designed to automatically discover a diverse set of attentions within image frames that are consistent across multiple videos.

As shown in Fig. 2, we adopt the ResNet-50 CNN architecture [14] as our base model for extracting features from each sampled image. The CNN has a convolutional layer in front (named $conv1$), followed by four residual blocks. We exploit $conv1$ to $res5c$ as the feature extractor. As a result, each image I_n is represented by an 8×4 grid of feature vectors $\{\mathbf{f}_{n,\ell}\}_{\ell=1,L}$, where $L = 32$ is the number of grid cells, and each feature is a $D = 2048$ dimensional vector.

Multiple attention models are then trained to locate discriminative image regions (distinctive object parts) within the training data. For the kth model, $k \in (1, \ldots, K)$, the amount of spatial attention $s_{n,k,\ell}$ given to the feature vector in cell ℓ is based on a response $e_{n,k,\ell}$ generated by passing the feature vector through two linear transforms and a ReLU activation in between. Specifically,

$$e_{n,k,\ell} = (\mathbf{w}'_{s,k})^{\mathsf{T}} \max(\mathbf{W}_{s,k}\mathbf{f}_{n,\ell} + \mathbf{b}_{s,k}, 0) + b'_{s,k}, \quad (1)$$

where $\mathbf{w}'_{s,k} \in \mathbb{R}^d$, $\mathbf{W}_{s,k} \in \mathbb{R}^{d \times D}$, $\mathbf{b}_{s,k} \in \mathbb{R}^d$ and $b'_{s,k} \in \mathbb{R}$
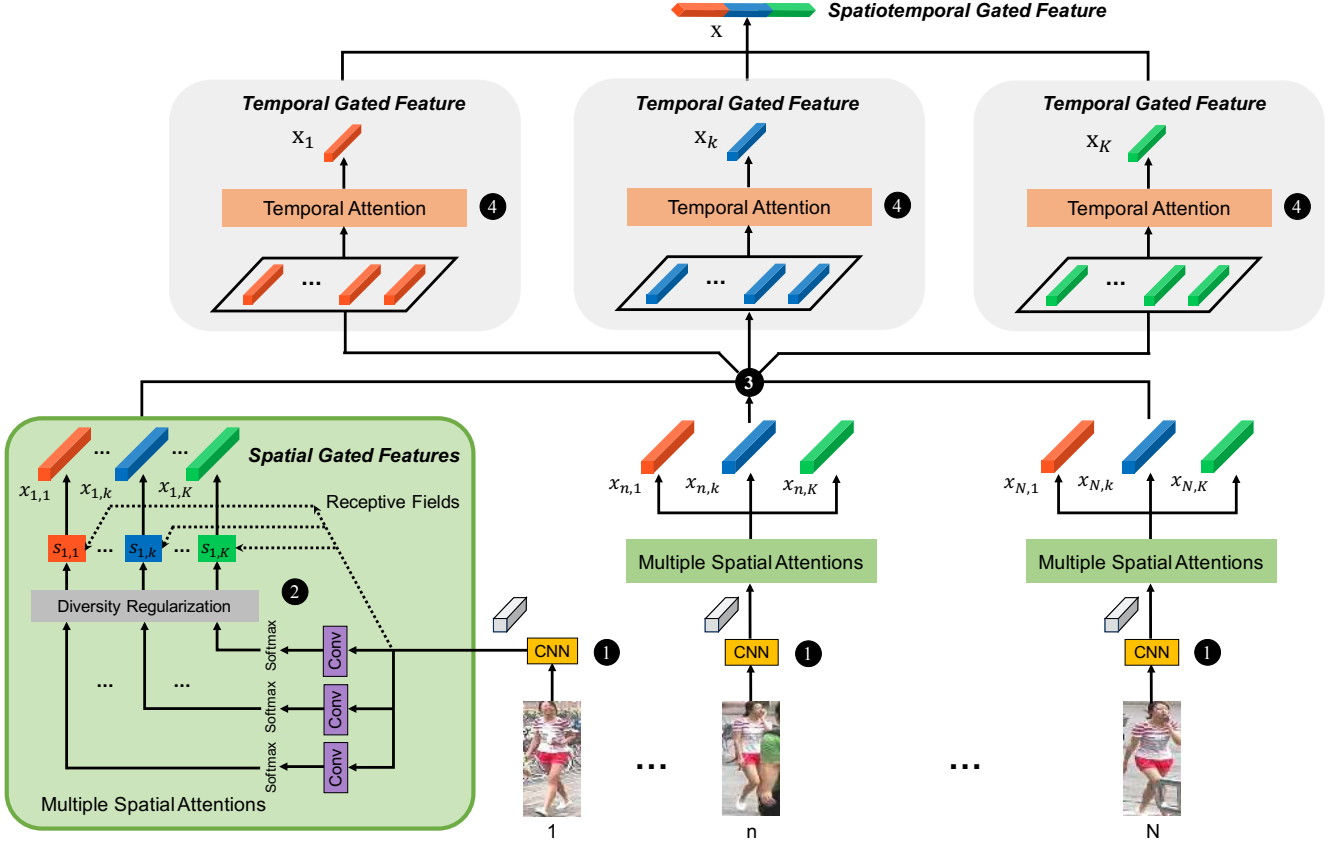
Figure 2. **Spatiotemporal Attention Network Architecture**. The input video is reduced to N frames using restricted random sampling. (1) Each image is transformed into feature maps using a CNN. (2) These feature maps are sent to a conventional network followed by a softmax function to generate multiple spatial attention models and corresponding receptive fields for each input image. A diversity regularization term encourages learning spatial attention models that do not result in overlapping receptive fields per image. Each spatial attention model discovers a specific salient image region and generates a spatial gated feature (Fig. 3). (3) Spatial gated features from all frames are grouped by spatial attention model. (4) Temporal attentions compute an aggregate representation for the set of features generated by each spatial attention model. Finally, the spatiotemporal gated features for all body parts are concatenated into a single feature which represents the information contained in the entire video sequence.

are parameters to be learned for the kth spatial attention model. The first linear transform projects the original feature to a lower $d = 256$ dimensional space, and the second transform produces a scalar value for each feature/cell. The attention for each feature/cell is then computed as the softmax of the responses

$$s_{n,k,\ell} = \frac{\exp(e_{n,k,\ell})}{\sum_{j=1}^{L} \exp(e_{n,k,j})}. \tag{2}$$

The set $\mathbf{s}_{n,k} = [s_{n,k,1}, \ldots, s_{n,k,L}]$ of weights defines the *receptive field* of the kth spatial attention model (part detector) for image I_n. By definition, each receptive field is a probability mass function since $\sum_{\ell=1}^{L} s_{n,k,\ell} = 1$.

For each image I_n, we generate K spatial gated visual features $\{\mathbf{x}_{n,k}\}_{k=1,K}$ using attention weighted averaging

$$\mathbf{x}_{n,k} = \sum_{\ell=1}^{L} s_{n,k,\ell} \mathbf{f}_{n,\ell}. \tag{3}$$

Each gated feature represents a salient part of the input image (Fig. 3). Because $\mathbf{x}_{n,k}$ is computed by pooling over the entire grid $\ell \in [1, L]$, the spatial gated feature contains no information about the image location from which it was extracted. As a result, the spatial gated features generated for a particular attention model across multiple images are all roughly aligned—*e.g.* extracted patches of the face all tend to have the eyes in roughly the same pixel location.

Similar to fine-grained object recognition [26], we pool information across frames to created an enhanced variant

$$\widehat{\mathbf{x}}_{n,k} = E(\mathbf{x}_{n,k}) \tag{4}$$
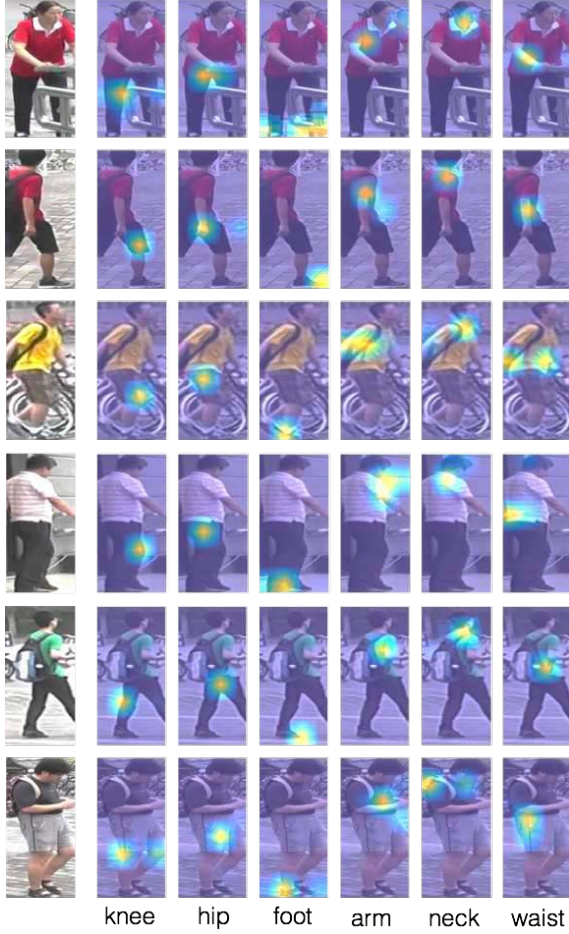
| knee | hip | foot | arm | neck | waist |

Figure 3. **Learned Spatial Attention Models**. Example images and corresponding receptive fields for our diverse spatial attention models when $K = 6$. Our methodology discovers distinctive image regions which are useful for re-identification. The attention models primarily focus on foreground regions and generally correspond to specifc body parts. Our interpretation of each is indicated at the bottom of each column.

of each spatial gated feature. The enhancement function $E()$ follows the past work on second-order pooling [5]. See the supplementary material for further details.

3.3. Diversity Regularization

The outlined approach for learning multiple spatial attention models can easily produce a degenerate solution. For a given image, there is no constraint that the receptive field generated by one attention model needs to be different from the receptive field of another model. In other words, multiple attention models could easily learn to detect the same body part. In practice, we need to ensure each of the N spatial attention models focuses on different regions of the given image.

Since each receptive field $\mathbf{s}_{n,k}$ has a probabilistic inter-

pretation, one solution is to use the Kullback-Leibler divergence to evaluate the diversity of the receptive fields for a given image. For notational convenience, we define the matrix $\mathbf{S}_n \in \mathbb{R}^{K \times L}$ as the collection of receptive fields generated for image I_n by the K spatial attention models

$$\mathbf{S}_n = [\mathbf{s}_{n,1}, \ldots, \mathbf{s}_{n,K}]. \tag{5}$$

Typically, the attention matrix has many values close to zero after the softmax() function, and these small values drop sharply when passed though the $\log()$ operation in the Kullback-Leibler divergence. In this case, the empirical evidence suggests the training process is unstable [27].

To encourage the spatial attention models to focus on different salient regions, we design a penalty term which measures the overlap between different receptive fields. Suppose $\mathbf{s}_{n,i}$ and $\mathbf{s}_{n,j}$ are two attention vectors in attention matrix $\mathbf{S}_n$. Employing the probability mass property of attention vectors, we use the Hellinger distance [4] to measure the similarity of $\mathbf{s}_{n,i}$ and $\mathbf{s}_{n,j}$. The distance is defined as

$$H(\mathbf{s}_{n,i}, \mathbf{s}_{n,j}) = \frac{1}{\sqrt{2}} \sqrt{\sum_{\ell=1}^{L} (\sqrt{s_{n,i,\ell}} - \sqrt{s_{n,j,\ell}})^2}, \tag{6}$$

$$= \frac{1}{\sqrt{2}} \|\sqrt{\mathbf{s}_{n,i}} - \sqrt{\mathbf{s}_{n,j}}\|_2. \tag{7}$$

Since $\sum_{\ell=1}^{L} s_{n,k,\ell} = 1$:

$$H^2(\mathbf{s}_{n,i}, \mathbf{s}_{n,j}) = 1 - \sum_{\ell=1}^{L} (\sqrt{s_{n,i,\ell} s_{n,j,\ell}}). \tag{8}$$

To ensure diversity of the receptive fields, we need to maximize the distance between $\mathbf{s}_{n,i}$ and $\mathbf{s}_{n,j}$, which is equivalent to minimizing $1 - H^2(\mathbf{s}_{n,i}, \mathbf{s}_{n,j})$. We introduce $\mathbf{R}_n = \sqrt{\mathbf{S}_n}$ for notation convenience, where each element in $\mathbf{R}_n$ is the square root of the corresponding element in $\mathbf{S}_n$. Thus, the regularization term to measure the redundancy between receptive fields per image is

$$Q = \|(\mathbf{R}_n \mathbf{R}_n^\top - \mathbf{I})\|_F^2, \tag{9}$$

where $\| \cdot \|_F$ denotes the Frobenius norm of a matrix and $\mathbf{I}$ is a K-dimensional identity matrix. This regularization term Q will be multiplied by a coefficient, and added to the original OIM loss.

Diversity regularization was recently employed for text embedding using recurrent networks [27]. In this case, the authors employed a variant

$$Q' = \|(\mathbf{S}_n \mathbf{S}_n^\top - \mathbf{I})\|_F^2 \tag{10}$$

of our proposed regularization. Although Q and Q' have similar formulations, the regularization effects are very different. Q is based on probability mass distributions with

the constraint $\sum_{\ell=1}^{L} \mathbf{s}_{n,k,\ell} = 1$ while Q' can be formulated on any matrix. Q' encourages $\mathbf{S}_n$ to be sparse – preferring only non-zero elements along the diagonal of $\mathbf{S}_n$. Although Q' forces the receptive fields not to overlap, it also encourages them to be concentrated to a single cell. Q, on the other hand, allows large salient regions like "upperbody" while discouraging receptive fields from overlapping. We compare the performances of the two regularization terms Q and Q' in Section 4.3.

3.4. Temporal Attention

Recall that each frame I_n is represented by a set $\{\widehat{\mathbf{x}}_{n,1}, \ldots, \widehat{\mathbf{x}}_{n,K}\}$ of K enhanced spatial gated features, each generated by one of the K spatial attention models. We now consider how best to combine these features extracted from individual frames to produce a compact representation of the entire input video.

All parts of an object are seldom visible in every video frame—either because of self-occlusion or from an explicit foreground occluder (Fig. 1). Therefore, pooling features across time using a per-frame weight t_n is not sufficiently robust, since some frames could contain valuable partial information about an individual (*e.g.* face, presence of a bag or other accessory, etc.).

Instead of applying the same temporal attention weight t_n to all features extracted from frame I_n, we apply multiple temporal attention weights $\{t_{n,1}, \ldots, t_{n,K}\}$ to each frame—one for each spatial component. With this approach, our temporal attention model is able to assess the importance of a frame based on the merits of the different salient regions. Temporal attention models which only operate on whole frame features could easily lose fine-grained cues in frames with moderate occlusion.

Similarly, basic temporal aggregation techniques (compared to temporal attention models) like average pooling or max pooling generally weaken or over emphasize the contribution of discriminative features (regardless of whether the pooling is applied per-frame, or per-region). In our experiments, we compare our proposed per-region-per-frame temporal attention model to average and maximum pooling applied on a per-region basis, and indeed find that maximum performance is achieved with our temporal attention model.

Similar to spatial attention, we define the temporal attention $t_{n,k}$ for the spatial component k in frame n to be the softmax of a linear response function

$$e_{n,k} = (\mathbf{w}_{t,k})^{\mathsf{T}} \widehat{\mathbf{x}}_{n,k} + b_{t,k}, \tag{11}$$

where $\widehat{\mathbf{x}}_{n,k} \in \mathbb{R}^D$ is the enhanced feature of the kth spatial component in the nth frame, and $\mathbf{w}_{t,k} \in \mathbb{R}^D$ and $b_{t,k}$ are parameters to be learned.

$$t_{n,k} = \frac{e_{n,k}}{\sum_{j=1}^{N} e_{j,k}}. \tag{12}$$

The temporal attentions are then used to gate the enhanced spatial features on a per component basis by weighted averaging

$$\mathbf{x}_k = \sum_{n=1}^{N} t_{n,k} \widehat{\mathbf{x}}_{n,k}. \tag{13}$$

Combining (3), (4) and (13) summarizes how we apply attention on a spatial then temporal basis to extract and align portions of each raw feature $\mathbf{f}_{n,\ell}$ and then aggregate across time to produce a latent representation of each distinctive object region/part

$$\mathbf{x}_k = \sum_{n=1}^{N} t_{n,k} E \left(\sum_{\ell=1}^{L} s_{n,k,\ell} \mathbf{f}_{n,\ell} \right). \tag{14}$$

Finally, the entire input video is represented by a feature vector $\mathbf{x} \in \mathbb{R}^{K \times D}$ generated by concatenating the temporally gated features of each spatial component

$$\mathbf{x} = [\mathbf{x}_1, \ldots, \mathbf{x}_K]. \tag{15}$$

3.5. Re-Identification Loss

In this paper, we adopt the Online Instance Matching loss function (OIM) [44] to train the whole network. Typically, re-identification uses a multi-class softmax layer as the objective loss. Often, the number of mini-batch samples is much smaller than the number of identities in the training dataset, and network parameter updates can be biased. Instead, the OIM loss function uses a lookup table to store features of all identities appearing in the training set. In each forward iteration, a mini-batch sample is compared against all the identities when computing classification probabilities. This loss function has shown to be more effective than softmax when training re-identification networks.

4. Experiments

4.1. Datasets

We evaluate the proposed algorithm on three commonly used video-based person re-identification datasets: PRID2011 [15], iLIDS-VID [40], and MARS [48]. PRID2011 consists of person videos from two camera views, containing 385 and 749 identities, respectively. Only the first 200 people appear in both cameras. The length of each image sequence varies from 5 to 675 frames. iLIDS-VID consists of 600 image sequences of 300 subjects. For each person we have two videos with the sequence length ranging from 23 to 192 frames with an average duration of 73 frames. The MARS dataset is the largest video-based person re-identification benchmark with 1,261 identities and around 20,000 video sequences generated by DPM

detector [10] and GMMCP tracker [7]. Each identity is captured by at least 2 cameras and has 13.2 sequences on average. There are 3,248 distractor sequences in the dataset.

For PRID2011 and iLIDS-VID datasets, we follow the evaluation protocol from [40]. Datasets are randomly split into probe/gallery identities. This procedure is repeated 10 times for computing averaged accuracies. For the MARS dataset, we follow the original splits provided by [48] which use the predefined 631 identities for training and the remaining identities for testing.

4.2. Implementation details and evaluation metrics

We divide each input video sequence into $N = 6$ chunks of equal duration. We first pretrain the ResNet-50 model on image-based person re-identification datasets, including CUHK01 [23], CUHK03 [24], 3DPeS [3], VIPeR [12], DukeMTMC-reID [51] and CUHK-SYSU [44]. and then fine-tune it on PRID2011, iLIDS-VID and MARS training sets. Once finished, we fix the CNN model and train the set of multiple spatial attention models with average temporal pooling and OIM loss function. Finally, the whole network, except the CNN model, is trained jointly. The input image is resized to 256×128. The network is updated using batched Stochastic Gradient Descent with an initial learning rate set to 0.1 and then dropped to 0.01. The aggregated feature vector after the last FC layer is embeded into 128-dimensions and L2-normalized to represent each video sequence. During the training stage, we utilize the Restricted Random Sampling to select training samples. For each video, we extract its L2-normalized feature and sent it to the OIM loss function to supervise the training process. During testing, we use the first image from each of N segments as a testing sample and its L2-normalized features are utilized to compute the similarity of the spatiotemporal gated features generated for the pair of videos being assessed.

Re-identification performance is reported using the rank-1 accuracy. On the MARS dataset we also evaluate the mean average precision (mAP) [48]. Since mAP takes recall into consideration, it is more suitable for the MARS dataset which has multiple videos per identity.

4.3. Component Analysis of the Proposed Model

We investigate the effect of each component of our model by conducting several analytic experiments. In Tab. 1, we list the results of each component in the proposed network. **Baseline** corresponds to ResNet-50 trained with OIM loss on image-based person re-id datasets and then jointly fine-tuned on video datasets: PRID2011, iLIDS-VID, and MARS. **SpaAtn** consists of the subnetwork of ResNet-50 (from $res2x$ to $res5x$) and multiple spatial attention models. All spatial gated features generated by the same attention model are grouped together and averaged over all frames. For each video sequence, there will be K

METHOD	PRID2011	iLIDS-VID	MARS
Baseline	82.7	61.2	73.4 (58.1)
SpaAtn	84.2	64.9	74.5 (59.3)
SpaAtn+Q'	86.5	64.5	74.0 (58.2)
SpaAtn+Q	86.7	68.6	77.0 (60.9)
SpaAtn+Q+MaxPool	86.9	68.2	76.8 (60.5)
SpaAtn+Q+TemAtn	88.4	69.7	77.1 (61.2)
SpaAtn+Q+TemAtn+Ind	**93.2**	**80.2**	**82.3 (65.8)**

Table 1. Component analysis of the proposed method: rank-1 accuracies are reported. For MARS we provide mAP in brackets. **SpaAtn** is the multi-region spatial attention, $\mathbf{Q'}$ and $\mathbf{Q}$ are two regularization terms, **MaxPool** and **TemAtn** are max temporal pooling and the proposed temporal attention respectively. **Ind** represents fine-tuning the whole network to each dataset independently.

averaged feature vectors. We concatenate the K features and then send them to the last FC layer and OIM loss function to train the neural network. Compared with **Baseline**, **SpaAtn** improves the rank-1 accuracy by 1.5%, 3.7%, and 1.1% on PRID2011, iLIDS-VID and MARS, respectively. This shows that multiple spatial attention models are effective at finding persistent discriminative image regions which are useful for boosting re-identification performance.

SpaAtn+Q' has the same network architecture as **SpaAtn** but with the text embedding diversity regularization term Q' [27]. **SpaAtn+Q** uses our proposed diversity regularization term Q based on Hellinger distance. From the results, we can see that our proposed Hellinger regularization improves accuracy. We believe the improvement comes from being able to learn multiple attention models with sufficiently large (but minimally overlapping) receptive fields (see Fig.3 for sample receptive fields generated for the learned attention models using **SpaAtn+Q**). **SpaAtn+Q** and **SpaAtn+Q+MaxPool** are strategies for average temporal pooling and maximum temporal pooling, respectively. **SpaAtn+Q+TemAtn** applies multiple temporal attentions to each frame—one for each diverse spatial attention model. The assigned temporal attention weights reflect the pertinence of each spatially attended region (*e.g.* is the part fully visible and easy to detect?). We finally fine-tune the whole network, including the CNN model, to each video dataset independently. **SpaAtn+Q+TemAtn+Ind** is the final result of our proposed framework.

Different number of spatial attention models: We also carry out experiments to investigate the effect of varying the number K of spatial attention models (Tab. 2). When $K = 1$, the framework is limited to a single spatial attention model, which tends to cover the whole body. As K is increased, the network is able to discover a larger set of body parts, and since the receptive fields are regularized to have minimal overlap, the reception fields tend to shrink as

K	PRID2011	iLIDS-VID	MARS
1	86.2	64.7	76.0
2	83.4	64.6	75.7
4	86.9	64.6	**77.2**
6	**88.4**	**69.7**	77.1
8	88.0	66.9	76.7

Table 2. The rank-1 accuracy using different number K of diverse spatial attention models.

METHOD	PRID2011	iLIDS-VID	MARS
STA [29]	64.1	44.3	-
DVDL [16]	40.6	25.9	-
TDL [46]	56.7	56.3	-
SI2DL [53]	76.7	48.7	-
mvRMLLC+Alignment [6]	66.8	69.1	-
AMOC+EpicFlow [28]	82.0	65.5	-
RNN [35]	70.0	58.0	-
IDE [49] + XQDA [25]	-	-	65.3 (47.6)
GEI+Kissme [48]	19.0	10.3	1.2 (0.4)
end AMOC+EpicFlow [28]	83.7	68.7	68.3 (52.9)
Mars [48]	77.3	53.0	68.3 (49.3)
SeeForest [52]	79.4	55.2	**70.6** (**50.7**)
QAN [31]	90.3	68.0	-
PAM-LOMO+KISSME [17]	**92.5**	**79.5**	-
Ours	**93.2**	**80.2**	**82.3** (**65.8**)

Table 3. Comparisons of our proposed approach to the state-of-the-art on PRID2011, iLIDS-VID, and MARS datasets. The rank-1 accuracies are reported and for MARS we provide mAP in brackets. The best and second best results are marked by red and blue colors, respectively.

K gets bigger. Interestingly, there is a general drop in perform when K is increased from 1 to 2. This implies treating a person as a single region instead of two distinct body parts is better. However, when a sufficiently large $K = 6$ number of spatial models is used, the network achieves maximum performance.

Example learned spatial attention models and corresponding receptive fields are shown in Fig. 3. The receptive fields generally correspond to specific body parts and have varying sizes dependent on the discovered concept. In constrast, the receptive fields generated by [30] tend to include background clutter and exhibit substantial overlap between different attention models. Our receptive fields, on the other hand, have minimal overlap and focus primarily on the foreground regions.

4.4. Comparison with the State-of-the-art Methods

Table 3 reports the performance of our approach with other state-of-the-art techniques. On each dataset, our method attains the highest performance. We achieve maximum improvement on MARS dataset, where we improve the state-of-the-art by 11.7%. The previous best reported results are from PAM-LOMO+KISSME [17] (which learns

signature representation to cater for high variance in a person's appearance) and from SeeForest [52] (which combines six spatial RNNs and temporal attention followed by a temporal RNN to encode the input video). In contrast, our network architecture is intuitive and straightforward to train. MARS is the most challenging data (it contains distractor sequences and has a substantially larger gallery set) and our methodology achieves a significant increase in mAP accuracy. This result suggests our spatiotemporal model is very effective for video-based person re-identification in challenging scenarios.

5. Summary

A key challenge for successful video-based person re-identification is developing a latent feature representation of each video as a basis for making comparisons. In this work, we propose a new spatiotemporal attention mechanism to achieve better video representations. Instead of extracting a single feature vector per frame, we employ a diverse set of spatial attention models to consistently extract similar local patches across multiple images (Fig. 3). This approach automatically solves two common problems in video re-identification: aligning corresponding image patches across frames (because of changes in body pose, orientation relative to the camera, etc.) and determining whether a particular part of the body is occluded or not.

To avoid learning redundant spatial attention models, we employ a diversity regularization term based on Hellinger distance. This encourages the network to discover a set of spatial attention models that have minimal overlap between receptive fields generated for each image. Although diversity regularization is not a new topic, we are the first to learn a diverse set of spatial attention models for video sequences, and illustrate the importance of Hellinger distance for this task (our experiments illustrate how a diversity regularization term used in text embedding is less effective for images).

Finally, temporal attention is used to aggregate features across frames on a per-spatial attention model basis—*e.g.* all features from the facial region are combined. This allows the network to represent each discovered body part based on the most pertinent image regions within the video. We evaluated our proposed approach on three datasets and performed a series of experiments to analyze the effect of each component. Our method outperforms the state-of-the-art approaches by large margins which demonstrates its effectiveness in video-based person re-identification.

References

[1] E. Ahmed, M. Jones, and T. K. Marks. An improved deep learning architecture for person re-identification. In *Computer Vision and Pattern Recognition*, pages 3908–3916, 2015. 2

[2] S. Bak and P. Carr. One-shot metric learning for person re-identification. In *Computer Vision and Pattern Recognition*, 2017. 2

[3] D. Baltieri, R. Vezzani, and R. Cucchiara. 3dpes: 3d people dataset for surveillance and forensics. In *Proceedings of the 2011 joint ACM workshop on Human gesture and behavior understanding*, pages 59–64. ACM, 2011. 7

[4] R. Beran. Minimum hellinger distance estimates for parametric models. *The Annals of Statistics*, pages 445–463, 1977. 5

[5] J. Carreira, R. Caseiro, J. Batista, and C. Sminchisescu. Semantic segmentation with second-order pooling. In *European Conference on Computer Vision*, 2012. 5

[6] J. Chen, Y. Wang, and Y. Y. Tang. Person re-identification by exploiting spatio-temporal cues and multi-view metric learning. *IEEE Signal Processing Letters*, 23(7):998–1002, 2016. 8

[7] A. Dehghan, S. Modiri Assari, and M. Shah. Gmmcp tracker: Globally optimal generalized maximum multi clique problem for multiple object tracking. In *Computer Vision and Pattern Recognition*, pages 4091–4099, 2015. 7

[8] 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. 2

[9] M. Farenzena, L. Bazzani, A. Perina, V. Murino, and M. Cristani. Person re-identification by symmetry-driven accumulation of local features. In *Computer Vision and Pattern Recognition*, pages 2360–2367. IEEE, 2010. 2

[10] 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. 7

[11] N. Gheissari, T. B. Sebastian, and R. Hartley. Person reidentification using spatiotemporal appearance. In *Computer Vision and Pattern Recognition*, volume 2, pages 1528–1535. IEEE, 2006. 2

[12] D. Gray, S. Brennan, and H. Tao. Evaluating appearance models for recognition, reacquisition, and tracking. In *Proc. IEEE International Workshop on Performance Evaluation for Tracking and Surveillance*, pages 1–7, 2007. 7

[13] 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. 2

[14] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Computer Vision and Pattern Recognition*, pages 770–778, 2016. 3

[15] M. Hirzer, C. Beleznai, P. M. Roth, and H. Bischof. Person re-identification by descriptive and discriminative classification. In *Scandinavian conference on Image analysis*, pages 91–102. Springer, 2011. 6

[16] S. Karanam, Y. Li, and R. J. Radke. Person re-identification with discriminatively trained viewpoint invariant dictionaries. In *International Conference on Computer Vision*, pages 4516–4524, 2015. 8

[17] F. M. Khan and F. Bremond. Multi-shot person re-identification using part appearance mixture. In *Winter Conference on Applications of Computer Vision*, pages 605–614, 2017. 8

[18] M. Koestinger, M. Hirzer, P. Wohlhart, P. M. Roth, and H. Bischof. Large scale metric learning from equivalence constraints. In *Computer Vision and Pattern Recognition*, pages 2288–2295. IEEE, 2012. 2

[19] 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. 2

[20] D. Li, X. Chen, Z. Zhang, and K. Huang. Learning deep context-aware features over body and latent parts for person re-identification. In *Computer Vision and Pattern Recognition*, 2017. 2

[21] S. Li, T. Xiao, H. Li, W. Yang, and X. Wang. Identity-aware textual-visual matching with latent co-attention. In *International Conference on Computer Vision*, 2017. 2

[22] S. Li, T. Xiao, H. Li, B. Zhou, D. Yue, and X. Wang. Person search with natural language description. In *Computer Vision and Pattern Recognition*, 2017. 2

[23] W. Li and X. Wang. Locally aligned feature transforms across views. In *Computer Vision and Pattern Recognition*, pages 3594–3601, 2013. 7

[24] W. Li, R. Zhao, T. Xiao, and X. Wang. Deepreid: Deep filter pairing neural network for person re-identification. In *Computer Vision and Pattern Recognition*, pages 152–159, 2014. 2, 7

[25] S. Liao, Y. Hu, X. Zhu, and S. Z. Li. Person re-identification by local maximal occurrence representation and metric learning. In *Computer Vision and Pattern Recognition*, pages 2197–2206, 2015. 8

[26] T.-Y. Lin, A. RoyChowdhury, and S. Maji. Bilinear cnns for fine-grained visual recognition. In *Transactions of Pattern Analysis and Machine Intelligence (PAMI)*, 2017. 4

[27] Z. Lin, M. Feng, C. N. d. Santos, M. Yu, B. Xiang, B. Zhou, and Y. Bengio. A structured self-attentive sentence embedding. *arXiv preprint arXiv:1703.03130*, 2017. 5, 7

[28] H. Liu, Z. Jie, K. Jayashree, M. Qi, J. Jiang, S. Yan, and J. Feng. Video-based person re-identification with accumulative motion context. *arXiv preprint arXiv:1701.00193*, 2017. 2, 8

[29] K. Liu, B. Ma, W. Zhang, and R. Huang. A spatio-temporal appearance representation for video-based pedestrian re-identification. In *International Conference on Computer Vision*, pages 3810–3818, 2015. 8

[30] X. Liu, H. Zhao, M. Tian, L. Sheng, J. Shao, S. Yi, J. Yan, and X. Wang. Hydraplus-net: Attentive deep features for pedestrian analysis. *arXiv preprint arXiv:1709.09930*, 2017. 3, 8

[31] Y. Liu, J. Yan, and W. Ouyang. Quality aware network for set to set recognition. In *Computer Vision and Pattern Recognition*, pages 5790–5799, 2017. 8

[32] C. C. Loy, T. Xiang, and S. Gong. Multi-camera activity correlation analysis. In *Computer Vision and Pattern Recognition*, pages 1988–1995. IEEE, 2009. 1

[33] 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. 2

[34] X. Ma, X. Zhu, S. Gong, X. Xie, J. Hu, K.-M. Lam, and Y. Zhong. Person re-identification by unsupervised video matching. *Pattern Recognition*, 65:197–210, 2017. 2, 3

[35] N. McLaughlin, J. Martinez del Rincon, and P. Miller. Recurrent convolutional network for video-based person re-identification. In *Computer Vision and Pattern Recognition*, pages 1325–1334, 2016. 2, 3, 8

[36] S. Pedagadi, J. Orwell, S. Velastin, and B. Boghossian. Local fisher discriminant analysis for pedestrian re-identification. In *Computer Vision and Pattern Recognition*, pages 3318–3325, 2013. 2

[37] B. J. Prosser, W.-S. Zheng, S. Gong, T. Xiang, and Q. Mary. Person re-identification by support vector ranking. In *British Machine Vision Conference*, 2010. 2

[38] Y. Shen and Z. Miao. Multihuman tracking based on a spatial–temporal appearance match. *IEEE transactions on Circuits and systems for video technology*, 24(3):361–373, 2014. 2

[39] L. Wang, Y. Xiong, Z. Wang, Y. Qiao, D. Lin, X. Tang, and L. Van Gool. Temporal segment networks: Towards good practices for deep action recognition. In *European Conference on Computer Vision*, pages 20–36. Springer, 2016. 3

[40] 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. 6, 7

[41] T. Wang, S. Gong, X. Zhu, and S. Wang. Person re-identification by discriminative selection in video ranking. *IEEE transactions on pattern analysis and machine intelligence*, 38(12):2501–2514, 2016. 2

[42] X. Wang. Intelligent multi-camera video surveillance: A review. *Pattern recognition letters*, 34(1):3–19, 2013. 1, 2

[43] T. Xiao, S. Li, B. Wang, L. Lin, and X. Wang. End-to-end deep learning for person search. *arXiv preprint*, 2017. 2

[44] T. Xiao, S. Li, B. Wang, L. Lin, and X. Wang. Joint detection and identification feature learning for person search. In *Computer Vision and Pattern Recognition*, 2017. 2, 3, 6, 7

[45] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, and Y. Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *International Conference on Machine Learning*, pages 2048–2057, 2015. 2

[46] J. You, A. Wu, X. Li, and W.-S. Zheng. Top-push video-based person re-identification. In *Computer Vision and Pattern Recognition*, pages 1345–1353, 2016. 2, 8

[47] S.-I. Yu, Y. Yang, and A. Hauptmann. Harry potter's marauder's map: Localizing and tracking multiple persons-of-interest by nonnegative discretization. In *Computer Vision and Pattern Recognition*, pages 3714–3720, 2013. 1

[48] 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. 6, 7, 8

[49] L. Zheng, H. Zhang, S. Sun, M. Chandraker, and Q. Tian. Person re-identification in the wild. *arXiv preprint arXiv:1604.02531*, 2016. 8

[50] W.-S. Zheng, S. Gong, and T. Xiang. Person re-identification by probabilistic relative distance comparison. In *Computer Vision and Pattern Recognition*, pages 649–656. IEEE, 2011. 2

[51] Z. Zheng, L. Zheng, and Y. Yang. Unlabeled samples generated by gan improve the person re-identification baseline in vitro. *arXiv preprint arXiv:1701.07717*, 2017. 7

[52] 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 *Computer Vision and Pattern Recognition*, 2017. 2, 3, 8

[53] X. Zhu, X.-Y. Jing, F. Wu, and H. Feng. Video-based person re-identification by simultaneously learning intra-video and inter-video distance metrics. In *International Joint Conference on Artificial Intelligence*, pages 3552–3559, 2016. 2, 8