Unbiased Spatio-Temporal Representation with Uncertainty Control for Person Re-identification

Xiu Zhang and Bir Bhanu, *Life Fellow, IEEE*

***Abstract*—For person re-identification, most current research aims to encode the spatial and temporal information by using convolutional neural networks (CNNs) to extract spatial features and recurrent neural networks (RNNs) or their variations to discover the time dependencies. However, it ignores the effect of the complex background, which leads to a biased spatial representation. Further, it often uses the backpropagation through time (BPTT) to train RNNs. Unfortunately, it is hard to learn the long-term dependency via BPTT due to the gradient vanishing or exploding. The significance of a frame should not be biased by its position in a given sequence. In this paper, a new method is proposed to learn an unbiased semantic representation for video-based person re-identification. To handle the background clutter and occlusion, a two-branch CNN model is used to obtain the enriched representation from both the foreground person and original pedestrian images. Then, an unbiased bidirectional convolutional neural network architecture is developed to learn the unbiased spatial and temporal representation. Experimental results on three public datasets demonstrate the effectiveness of the proposed method.**

***Index Terms*—person re-identification (re-id), unbiased representation, sparse attentive backtracking, pedestrian detection, uncertainty control, bidirectional recurrent neural networks**

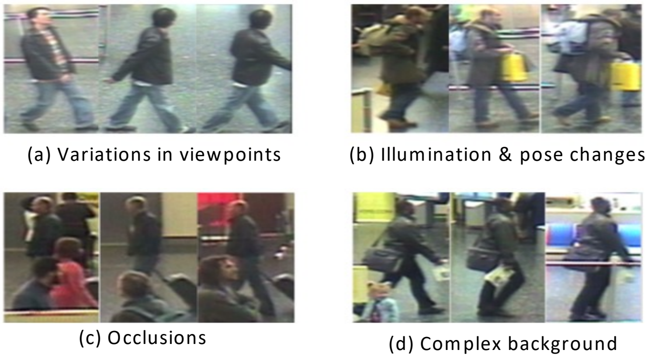
(a) Variations in viewpoints
(b) Illumination & pose changes
(c) Occlusions
(d) Complex background

Fig. 1. Challenges in re-id task. In (a), the first image is taken from the front view and the other two images are taken from the side view. In (b), the first image is taken from the first camera, while the other two frames are captured by the second camera. Due to the illumination changes, the yellow color of the backpack differs a lot in different frames. Also, the person changes the location of bag from the back to hands to hold the bag. In (c), in the subsequent frames, people are blocked by other persons. In (d), the background keeps changing as the person moves.

I. Introduction

The amount of video data has been rapidly increasing due to the prevalence of both the Internet and recording devices e.g., cell-phones with quality cameras, compact video cameras and video surveillance systems. Video analytics has become a hot topic with increasing demands for developing automatic processing tools. In this paper, we focus on the person re-identification (re-id) task, which aims to identify pedestrians across non-overlapping cameras. It has attracted much attention from both the research community and industry, and plays an important role in various surveillance applications [1], [2]. For example, when a girl gets lost in a theme park, the re-id techniques can be used to search the child in the video camera network. Re-id can be used for continuous tracking in video games using a Kinect sensor [3]. Still another example, we could use re-id to track a specific soccer player and obtain video analytics for each player for the duration of the game [4].

In spite of significant research in recent years, person re-id remains very challenging. As the common surveillance setting moves into the unconstrained environment, it becomes very difficult to re-identify pedestrians due to the variations in view angle of the camera, pose of pedestrian, illumination conditions, background clutter and occlusions as shown in Fig. 1.

In this paper, we address re-id problem in the video context. Compared to the still images, video-based re-id provides various samples to learn a more discriminative and robust appearance representation, especially, when frames include occlusions or complex backgrounds. Another benefit for using the videos instead of images is that the useful temporal information, such as gait, pose, movement captured in video may help to distinguish people in difficult scenarios.

Existing video-based person re-id methods extract frame-level features by using convolutional neural networks (CNNs) and aggregate the representation with recurrent neural networks (RNNs) across time [5]–[7]. However, these methods have several drawbacks. First, most methods learn the person representation either from the full-body or integrate different body parts from the estimated regions of interest (ROIs). Usually, the ROIs are in the form of rectangular bounding boxes, which may not capture the silhouettes well and may even include the complex backgrounds and occlusions. As shown in Fig. 2 (a), the background differs with the frames. In frame 1, grass and flowers are detected which disappear in the frame 2 to 4. Then, a patch of green grass appears in frame 5 and 6. Also, the other two persons with different shirts (striped shirt and white shirt) walk in front of the target person from the camera view. Thus, learning a good

X. Zhang is with the Department of Computer Science, University of California at Riverside, Riverside, CA, 92521 USA (e-mail: xzhan060@ucr.edu).

B. Bhanu is with the Department of Electrical and Computer Engineering, University of California at Riverside, Riverside, CA, 92521 USA (e-mail: bhanu@ece.ucr.edu).

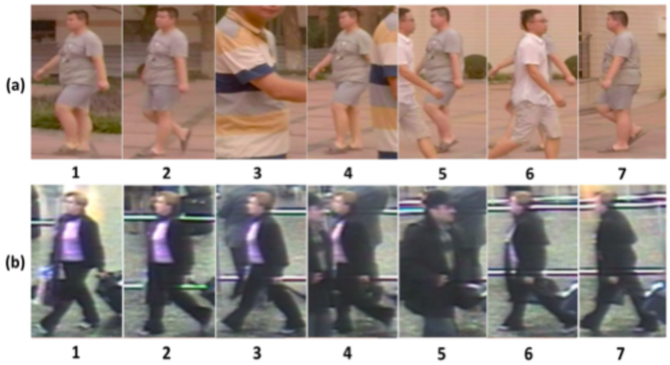

Fig. 2. Examples of walking pedestrians in two sequences. For (a) and (b), 7 frames are selected from two videos. In (a), as the person walks, the background changes. At the same time, the person is blocked by two different people from frame 3 to 6. In (b), the walking person is captured from different angles. The early images represent the front angle, and we could see the purple T-shirt, scarf, and the shopping bag in her right hand, which disappear gradually as shown in the following images. Occlusions also exist with varying degrees from frame 3 to 5.

spatial representation, without bias, for the background is essential, which may help filter the changing background (grass) and other pedestrians (frames 4 and 5) with less overlapping with frame 1 in the example. More recent work [8], [9] has exploited segmentation techniques to emphasize the foreground information to avoid the background clutter. However, such methods have the following drawbacks: segmentation methods are usually noisy and do not capture the perfect silhouette information, especially when there are more than one pedestrians. Further, the useful information is lost when the background is removed since there are connections between the person and background, e.g., backpacks, the carried briefcase, and other belongings. Thus, a hard cut-off of the background information is detrimental to the performance of the re-identification techniques.

Second, it is very intuitive to obtain the global representation of a given sequence with RNNs. For our task, the information that we want to capture is the identity of the person in the given frame, which is consistent along the sequence. Thus, the importance of each frame should not be dependent on its position. But, backpropagation through time (BPTT), which is now commonly used to train RNNs, is not able to capture the long-term dependencies due to the well-known gradient vanishing or exploding problem [10]. Although Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU) are proposed to alleviate this problem, it is still doubtful how much a fixed-length vector can memorize over a long sequence. All these limitations result in the difficulty in assigning enough or at least fair credit to the earlier time steps in a long sequence, while looking at the entire sequence would be considerably better than relying on the last few frames only. For example, as shown in Fig. 2, the first two frames capture the front side of the person and the other frames represent the side view. Both angles are important to achieve a comprehensive representation for this specific target. The first two frames should be given at least the equal emphasis as the latter frames. Frame 3 and 6 in Fig.2 (a) and frame 5 in Fig.2 (b) should be assigned with lower weights as they are not helpful in recognizing the target persons as they are completely blocked by the other people.

In light of the above discussions, this paper proposes to learn an unbiased spatio-temporal semantic representation for person re-identification. Specifically, our pipeline first uses a pedestrian detection method to obtain pixel-level mask for the body, and then replaces the complex background with a unified representation of the background to only capture the foreground pedestrian. Unlike other research [11]–[13] that learn features from either the whole image or foreground image, we use a two-branch network which includes both the masked foreground pedestrian and the original image. In order to learn the contribution of each branch (masked foreground branch and original image branch), we introduce homoscedastic uncertainty to combine the loss functions without manual tuning. Further, instead of BPTT, sparse attentive backtracking mechanism [14] is used to train RNNs in both forward and backward directions to get the unbiased temporal representation for pedestrians.

The paper is organized as follows. Section 2 summarizes the related work and contributions of this paper. Section 3 presents the framework of the proposed method and describes each component in detail. Experimental results on three video-based person re-id public datasets are shown and discussed in Section 4. Finally, the paper is concluded in Section 5.

II. Related work and Contributions

A. Related Work

The re-id task has been extensively explored in the last few years. The current approaches generally fall into two categories: (a) developing robust features for the given image or video and (b) designing discriminative metric learning methods that push the same person to be close by increasing similarity and pull different persons to be away by decreasing similarity. Recently, deep learning methods have been successfully applied to learn feature representation and similarity distance metric jointly.

For feature-based methods, different cues are used to learn discriminative and robust feature representation. A large part of person re-id methods subdivide the whole body into parts and then integrate different local and global low-level features [15]. One example SDALF [16] separates the body into parts and extracts three sets of entities, which are color histogram, maximally stable color regions, and recurrent high-structured patches. It also applies symmetric information to obtain good view invariance. Ma et al. [17] combine the Gabor filters and covariance descriptor to get the BiCov descriptor and dense color histogram [18]. Some approaches extract features from the whole body: treat the body as a whole and represent it using various kinds of features: haar-like features [19]; SIFT-like interest points [20]; texture (Schmid and Gabor filters) and color (histograms in different color spaces) [21]; color-position histogram [22]; 4-D multicolor height histogram; transform-normalized RGB (illumination invariant) features [23] and patch-based saliency features [24]. Some methods learn feature representations from multi-scales [25], [26]. Some other methods adopt semantic attributes and co-occurrence properties to model the consistent features across different views [27], [28].

Metric Learning methods aim to learn the distance between the given images or videos across different camera views by finding the mapping functions, which make the distance between the matched persons smaller but larger for different people in the learned space. For instance, KISSME [29] has the assumption that the distance follows a Gaussian distribution and the metric is formulated as a log-likelihood ratio test. Liao et al. [30] propose XQDA to learn a discriminative subspace by linear discriminant analysis (LDA). Mignon et al. [31] utilize a sparse set of pairwise similarity constraints to learn the distance metric. An et al. [32] propose a modified cosine similarity to measure the matching scores between probe and gallery. Chen et al. [33] learn the similarity from an explicit polynomial kernel feature map. Zheng et al. [34] formulate the problem as a relative distance comparison problem and present a probabilistic solution. An et al. [35], [36] learn a subspace in which the correlations of the reference data from different cameras are maximized using regularized canonical correlation analysis (RCCA).

Recent advances in deep learning provide a joint solution to integrate the feature representation and distance metric in a supervised manner. Li et al. [37] propose a filter pairing neural network (FPNN) to handle the body part displacements. Ahmed et al. [38] include a new layer which encodes the cross-input neighborhood differences and a subsequent layer that summarizes these differences in a siamese architecture. Ding et al. [39] develop an effective triplet generation scheme and use triplet loss to train the model. Zhao et al. [40] propose a center-triplet model which jointly learns the robust feature representation and optimizes the metric loss function.

Beyond the image-based person re-id, researchers have exploited temporal information across frames for video-based re-id. Early work used gait [41]–[43] or HOG3D descriptors [44]. More recently, McLaughlin et al. [35] introduced RNNs to explore the temporal information and added an additional pooling layer to summarize a video. Instead of using regular RNNs, Varior et al. [6] and Zhang et al. [7] adopted the LSTMs and bi-directional RNNs to select and encode more information. Liu et al. [45] constructed motion net to encode the motion information in their framework. Xu et al. [46] used a similar architecture but added one spatial pooling layer to select regions from each frame, and added another attentive temporal pooling layer to select informative frames. Zhou et al. [47] used a similar attention temporal pooling mechanism, but they employed spatial RNNs to integrate the neighborhood similarities within and across the frames. All of the above methods used BPTT to train the networks and thus, they introduced bias along time. These networks inevitably put more emphasis on the last few frames even when the earlier frames may contain more useful information.

As the quality of frames along a video differs significantly, attention schemes are widely employed to associate weight and select the informative frames. Li et al. [48] learned multiple spatial attention models with a diversity regularization term to localize body parts and combine features using temporal attention. Wu et al. [49] proposed a Siamese attention architecture that jointly optimized spatio-temporal video representations and their similarity metrics. Subramaniam et al. [50] activated a common set of salient features across multiple frames of a video with mutual consensus. To deal with the varying lengths of videos, Chen et al. [51] divided the long video sequences into multiple short snippets and aggregated the top-ranked snippets to estimate the sequence-level similarity. Similarly, Fu et al. [52] computed clip-level feature representation by aggregating frame-level representations. Gu et al. [53] proposed Appearance Preserving 3D Convolution (AP3D) model to learn better appearance representation for the video data. Yang et al. [54] applied the dynamic pyramid strategy to exploit multi-scale features under attention mechanism to maximally capture discriminative features.

Bayesian models mainly include two types of uncertainty: epistemic uncertainty and aleatoric uncertainty. Aleatoric uncertainty is inherent in data observations and cannot be reduced even if more data is collected. Kendall and Gal [55] divided it into homoscedastic uncertainty and heteroscedastic uncertainty. Homoscedastic uncertainty is captured independent of the input data and varies between different tasks. Heteroscedastic uncertainty is data dependent and varies across different data inputs. Epistemic uncertainty refers to the uncertainty in the model and can be reduced with enough training data. Modeling uncertainty can help to improve both the robustness and the interoperability of various algorithms.

B. Contributions of this paper

In this paper, we propose a new scheme to learn an unbiased semantic (spatial and temporal) representation to handle these difficulties of video-based person re-identification. The contributions of this paper are:

(1) A novel framework is proposed to effectively model the temporal correlations among frames by a sparse attentive backtracking mechanism [14], [56], which emphasizes the importance of learning long-term dependencies in re-id. We enable all possible interframe relations among any RNN units instead of restricting the information flow only within the adjacent RNN units from BPTT. Then we use a temporal attention to select the important routes to perform backtracking.

(2) A two-branch CNN model is used to enrich the unbiased spatial representation. It enables the model to get rid of the background clutter and occlusions.

(3) Homoscedastic uncertainty is used to balance the original branch and the masked branch instead of a naive manually tuned approach for estimating weights for the identification terms.

(4) Three public datasets are used for evaluation and comparison with other state-of-the-art methods. The importance of each component of our framework is validated experimentally.

This paper is an extension of our previous work [57]. We make the following major advancements compared with [57].

(a) We improve our elementary framework by incorporating a two-branch CNN to learn the unbiased spatial representation along with a principled way to combine the two identification loss terms. We conduct more experiments on more data sets and achieve state-of-the-art results, and provide visualization to explain the dynamic progress of using attentive weights for backtracking.

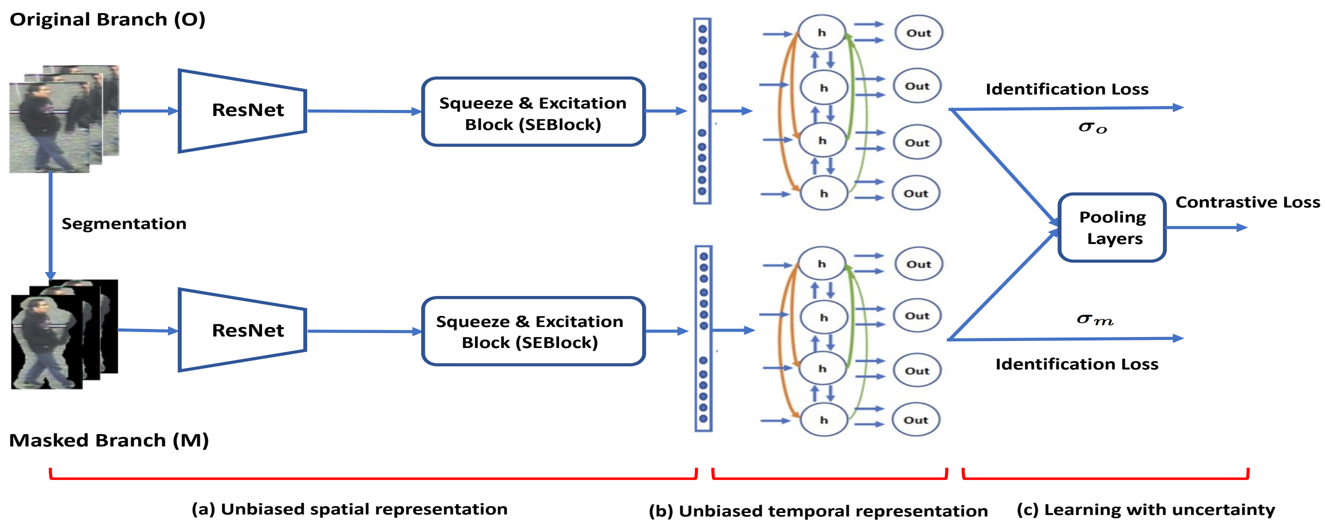

Fig. 3. Framework of the proposed method. There are three stages: (a) unbiased spatial representation, (2) unbiased temporal representation and (c) multitask learning with uncertainty. The first stage (a) includes a two branch convolutional neural network (ResNet). The first branch (O branch) is for the original images, and the second branch (M branch) is for the foreground images where the pedestrians are detected and their mask is used to replace the background to a uniform color. Both branches are passed through the SEBblock (Squeeze and Excitation Block). In the second stage (b), bi-directional RNNs encode temporal information, and use the sparse attentive backtracking method to train RNNs. The connections among the hidden units are shown in the figure, where the orange lines represent the forward connections among hidden units while the green lines illustrate the backward links. In the last stage (c), the features from both branches are fused by pooling layers with global average pooling layer (GAP) and one fully connected layer (FC). The whole network is trained end-to-end with both the contrastive loss and identification loss, where the identification loss is computed by aggregating with homoscedastic uncertainty σ_o and σ_m (best viewed in color).

(b) We exploit the temporal information in a bi-directional way to get a more complete representation of a video, as the identity of the person remains consistent along both the forward and backward directions.

(c) We explore the related work more extensively, and more datasets and ablation studies are included for a better understanding.

III. Our approach

A. Unbiased spatial representation

As the person re-id task often involves an unconstrained environment, the regions-of-interest (ROIs) usually contain occlusions and complex background information, which is usually considered as noise. It is very important to achieve a good spatial representation by extracting features which could resist the interference by noise.

Invariant background generation: Most of the early work takes the entire image as the input to the CNNs to extract spatial features. However, the presence of occlusions and variations in the background make it difficult to get the discriminative representation. Therefore, we are motivated to use a pedestrian detection method to get a segmented person, and then replace the complex background with a uniform color. Taking into account the scenario of re-id, we aim to have a method that is not sensitive to the background clutter, complex poses, and occlusions. Besides, we assume that there is only one target person in a frame. We used the pre-trained Deep Decompositional Network (DDN) [58] to get the estimated mask for the human body. DDN jointly estimates occluded regions and segments body parts by stacking occlusion estimation layers, completion layers, and decomposition layers.

When we apply DDN to the images, the masks usually contain sharp boundaries that are neither appropriate to describe the human silhouette nor good for further feature extraction. Gaussian smoothing method (kernel 3 x 3) [59] is used to smooth the undesired boundaries. As shown in Fig. 4, The first row shows the original images and the second row displays the corresponding masked images. After applying the smoothing method, the final outputs are illustrated in the third row, which excludes the occlusions and other background clutter. Then, we replace the backgrounds of all the frames to a unified background, which is black background in our case.

Two branch architecture: We design a two branch architecture to balance both the foreground pedestrian and background information [60] . One pair of images for each person (both the original image and the masked image) are fed into the CNN. Both branches share the same architecture, but their network parameters are not shared. In order to model the inter-dependencies between channels, the SEBlock (Squeeze and Excitation Block) [61] is added to re-calibrate the feature responses of the residual block to enhance spatial structure information. As shown in Fig. 5, the features X are first passed through a squeeze operation, which aggregates the feature maps across spatial dimensions H x W to produce a channel descriptor for C Channels. Then an extraction operation (Fully connected layer, ReLU, Fully Connected Layer, Sigmoid, Scale) is added to fully capture channel-wise dependencies. The reduction factor r can also help to adjust the cost to improve model efficiency. Unlike the channel attention, spatial attention concentrates on processing information into specific locations in space. Inspired by [62], [63], we add one self-attention module after the SEBlock to generate spatial

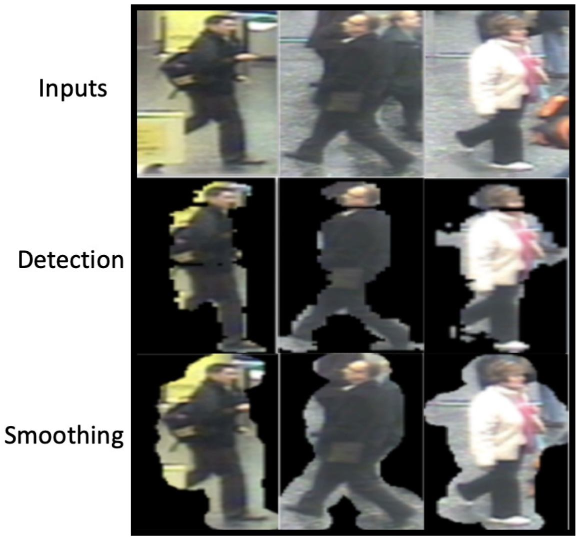

Fig. 4. Examples of invariant background generation. The three rows show the original frames, the results of pedestrian detection and Guassian smoothing, respectively (best viewed in color).

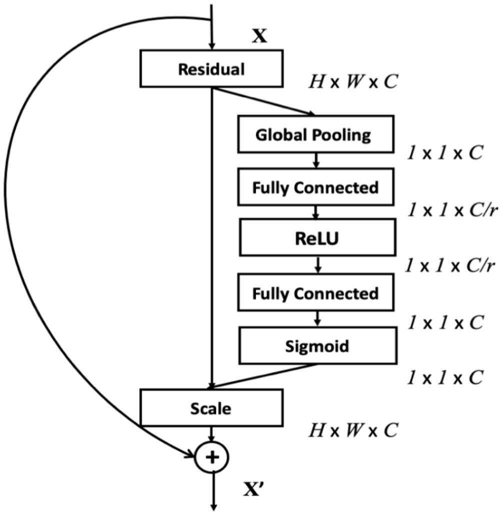

Fig. 5. The schema for SEBlock (Squeeze and Excitation Block), which aggregates the feature maps across spatial dimensions H x W for C channels with the reduction factor r).

attention map to align the features.

B. Unbiased temporal representation

It is intuitive to use RNNs to capture the time-series dynamics for it tracks the information of previous frames to predict the states of the current node. However, RNNs, which are trained using BPTT, suffer from the well-known exploding- or vanishing-gradient problems and they tend to forget the early inputs in case of long term sequences. Existing methods try to solve this by adding a pooling layer to summarize all the outputs from all the hidden units including the early ones. This could somehow include the information from the early frames, but it is still biased because the interference made from the early frames do not include the information from the latter frames. We expect to make the decision after seeing what has happened across all the previous timesteps, i.e., if we use the one-directional RNNs, the final decision is made based on $h_t(t \in 1....N)$ when we use $h_1, \cdots, h_{t-1}$ as the sequence of the hidden units. To address this problem, we use the sparse attentive backtracking mechanism [14], which is capable of learning long term dependencies but not lean towards the last few frames. Unlike the previous work [57] which uses the sparse attentive backtracking in one direction, we apply it to train bi-directional RNNs. The hidden state h_t at time t is a concatenation of the hidden state h_t^f in the forward direction and h_t^b in the backward direction. To compute h_t^f, we split the input into two sources: 1) the hidden unit from last timestep h_{t-1}; 2) all the hidden units prior to t. Likewise, when computing h_t^b, we need h_{t+1}^b and all the hidden units after t.

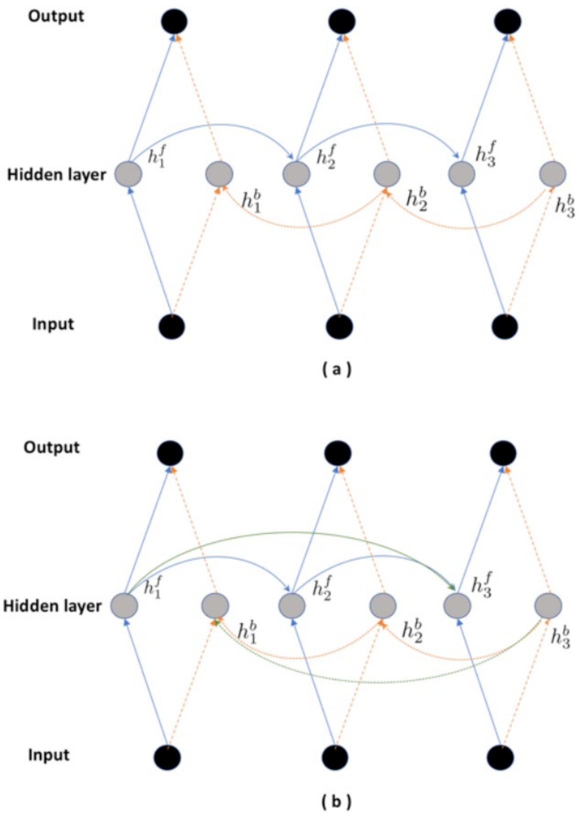

Fig. 6. Illustration for the forward pass with BPTT and sparse attentive backtracking to compute the hidden unit h_3 (which is $[h_3^f; h_3^b]$) in our bidirectional RNNs. We take the forward direction as an example. In (a), the only way that h_3^f gets the information from h_1^f is through h_2 . In (b), h_3^f could selectively choose any previous hidden units (h_1^f and h_2^f) for direct interaction (best viewed in color).

Additionally, the attention mechanism is adopted to assign credits for each former or latter states to compute h_t^f or h_t^b. We follow the attention process in [64] to compute the weights. The sparse attentive backtracking process is formulated as follows:

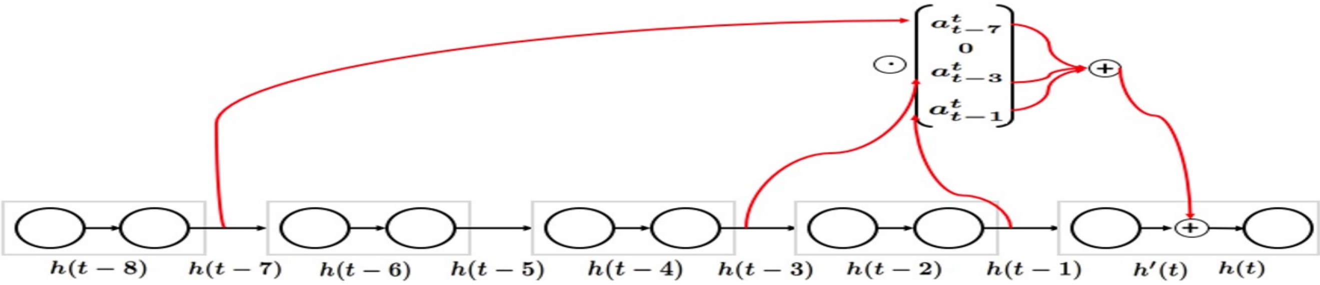

Fig. 7. This figure illustrates the forward pass in sparse attentive backtracking when the clip length is $T = 2$, and only the top $M = 3$ clips will be backpropagated. Black arrows describe how attention weights $h(t)$ are calculated by current provisional hidden state $h'(t)$ against the set of all memories H. The system selects and normalizes only the top k attention weights, while the others are zeroed out. Red arrows show the selected non-zero sparsified attention weights (best viewed in color).

$$\begin{aligned}
h_t^f &= tanh(W_v^f v_t + W_h^f h_{t-1}^f + W_{\tilde{h}}^f \sum_{i=1}^{t-1} \alpha_i^f h_i^f) \\
\alpha_i^f &= w_{\tilde{h}}^f [h_i^f ; h_t^f] \\
h_t^b &= tanh(W_v^b v_t + W_h^b h_{t+1}^b + W_{\tilde{h}}^b \sum_{i=(t+1)}^{N} \alpha_i^b h_i^b) \\
\alpha_i^b &= w_{\tilde{h}}^b [h_i^b ; h_t^b] \qquad (1)
\end{aligned}$$

where $W_{\tilde{h}} \in \mathbb{R}^{n \times n}$ and $w_{\tilde{h}} \in \mathbb{R}^{2n}$.

The final output of bi-directional RNNs is represented as:

$$\begin{aligned}
y_t &= W_y h_t \\
h_t &= [h_t^f ; h_t^b] \qquad (2)
\end{aligned}$$

where y_t is the output of the RNNs at time t, and it is used as a sequence-level feature representation. Its computation is dependent not only on the previous frames but also the upcoming frames. W_y represents the parameters of projections from the hidden layer to the output y_t based on the combined hidden states of h_t^f and h_t^b. In this work, we apply one RNN layer for both the forward pass and the backward pass.

As described above, this sparse attentive backtracking strategy explores all possibilities for correlations among the hidden states, and use attention mechanism to select the key routes to do the backtracking. In order to deal with the varying lengths of videos and speedup the backtracking, we first divide the given video into multiple small clips, where each clip includes T consecutive frames. We only select the clips with the top M highest attention weights to backpropagate as shown in Fig. 7. The use of bi-directional RNNs adds a stronger constraint for the consistent identity among the frames of any given video. We employ both techniques (sparse attentive backtraking and bi-directonal RNNs) to discover the potential long-term dependency patterns and learn an unbiased temporal representation.

C. Learning with uncertainty

We employ both contrastive loss and identification loss over the training samples and the two loss terms are denoted as $L_{contrastive}$ and L_{id}, respectively.

The contrastive loss aims to push the feature representation of the same person close and pull the features of different persons away. We employ the triplet loss with hard mining for contrastive loss [65]. To form a batch, we randomly sample P identities and randomly sample K clips for each identity (each clip contains T frames). The total number of clips in a batch is PK. For each sample a in the batch, the hardest positive, and the hardest negative samples within the batch are selected. The $L_{triplet}$ is defined as:

$$\begin{aligned}
L_{contrastive} = \overbrace{\sum_{i=1}^{P} \sum_{a=1}^{K}}^{all\ anchors} [H + \overbrace{\underset{p=1 \cdots K}{max} D(f(x_a^i), f(x_p^i))}^{hardest\ positive} \\
\underbrace{- \underset{\substack{j=1 \cdots P \\ n=1 \cdots K \\ j \neq n}}{min} D(f(x_a^i), f(x_n^i))]}_{hardest\ negative} \qquad (3)
\end{aligned}$$

where x_a^i is the anchor, x_p^i is the positive sample which has the same identity as x_a^i, x_n^i is the negative sample with different identity from x_a^i. $D()$ means the euclidean distance and H is the hyperparameter margin in hard-batch triplet loss. Hard-batch triplet loss makes sure that given an anchor x_a^i, x_p^i is closer to x_a^i than x_n^i.

The second term is an identity related loss. We use the cross-entropy loss function which is presented as follows:

$$L_{id} = \lambda_o L_{id}^o + \lambda_m L_{id}^m \qquad (4)$$

where L_{id}^o and L_{id}^m are the losses of the original branch O and masked branch M respectively.

To optimize λ_o and λ_m, one common option is to use a heuristic approach to weight the losses with grid search [28]. Another solution is to use network learning with validation loss to determine the weights for the task losses. Both methods require additional validation data. Model performance is extremely sensitive to weight selection and remains a challenging problem for the community.

To solve this problem, we formulate this problem as a multi-task learning (MTL) process. The uncertainty of MTL is homoscedastic in nature, which is task dependent. Thus, we infer the weights for the task loss from the observable homoscedastic uncertainty noise [66], [67].

We derive L_{id} based on maximizing the Gaussian likelihood with homoscedastic uncertainty. Let f^W be the output of a neural network with weights W on input x. The classification likelihood of a Bayesian probabilistic model is defined as:

$$p(y|f^W(x,\sigma)) = Softmax(\frac{1}{\sigma^2}f^W(x)) \tag{5}$$

where y refers to the model output and σ is the observation noise. The log likelihood for the output is:

$$\begin{aligned} logp(y=c|f^W(x),\sigma) &= \frac{1}{\sigma^2}f_c^W(x) \\ &-log\sum_{c'} exp\frac{1}{\sigma^2}f_{c'}^W(x) \end{aligned} \tag{6}$$

Then, the identification loss for a given class c can be formulated as $-logp(y = c|f^W(x),\sigma)$, c' refers to any possible class of the classification task. The identification loss could be defined as $L(x,W) = -logSoftmax(y, f^W(x))$, which could be simplified to

$$L(x,W,\sigma) \approx \frac{1}{\sigma^2}L(x,W) + log\sigma \tag{7}$$

Next, the joint identification loss of two branches is given as:

$$\begin{aligned} L_{id} = L(x,W,\sigma_o,\sigma_m) \approx \frac{1}{{\sigma_o}^2}L_o(x,W) + \frac{1}{\sigma_m^2}L_m(x,W) \\ +log\sigma_o\sigma_m \end{aligned} \tag{8}$$

The final training objective is the combination of contrastive loss and identification loss as:

$$L = L_{contrastive} + L_{id} \tag{9}$$

According to the above equation, the contrastive loss and the identification loss terms are assigned the same weight. During training, we use the given identity labels of the videos as the outputs of the network. We alternatively feed the positive (same person) and negative (different people) pairs of sequences as the inputs. The sparse attentive backtracking is used to train the bidirectional RNNs over the time steps in the network. While in the test phase, we discard the softmax layer and use the network as a feature extractor. Then we compute the distance of the extracted features against the gallery set. Similar pedestrians are closer in the Euclidean distance space.

IV. Experiments

In this section, we evaluate the proposed approach on four of the most popular public video datasets: iLIDs-VID [44], PRID 2011 [80], MARS [42], and DukeMTMC-VideoReID [81] and compare our method with other state-of-the-art methods.

A. Datasets

The iLIDs-VID dataset [44] contains 300 persons, which are recorded at an airport arrival hall using a CCTV network. Each person has 2 acquisition of videos with the sequence length varying from 23 to 192 frames. This dataset is very challenging due to the clothing similarities among people, changing illumination conditions and viewpoints, cluttered background, and the presence of occlusions.

The PRID 2011 dataset [80] consists of 749 persons captured by two adjacent camera views. Only the first 200 pairs of videos are taken from both cameras. The length of the image sequences ranges from 5 to 675, with an average of 100 frames. As compared to the iLIDs-VID dataset, this dataset is less challenging because it is taken under the uncrowded outdoor scenes, and it has relatively simple background and rare occlusions. We use the first 200 persons for evaluation as the other compared works [5], [43], [46].

The MARS dataset [43] includes 20,478 tracklets of 1,261 pedestrians which are captured at a university campus from 6 non-overlapping camera views. The dataset is divided into a training set with 625 pedestrians and a testing set with 626 pedestrians. There are 8,298 tracklets for the training set and 12,180 tracklets for the testing set. MARS dataset is one of the largest publicly available video-based person re-identification datasets.

The DukeMTMC-VideoReID is a subset of a large-scale re-id dataset DukeMTMC [81], which is recorded in an outdoor environment. This dataset is very challenging due to the changing illumination and viewpoint conditions, varying poses, noisy background and presence of occlusions. It includes 2,196 tracklets of 702 identities for training and 2,636 tracklets of another 702 identities for testing. Each identity has only one tracklet from one camera.

B. Experimental Setup

We use cumulative matching characteristic (CMC) curve and mean average precision (mAP) as the evaluation metrics to evaluate the performance. Cumulative Match Characteristics (CMC) curve which shows the identification rate vs. the rank for a closed set consisting of persons to be re-identified. To be specific, during evaluation, our USTRU model is used as a feature extractor for both the gallery sequences and the target sequence. After training for any test sequence, we sort all the gallery sequences by their nearest distance to the test sequence arranged in an ascending order. The recognition score at rank R means target persons are identified within the top R ranks. We report rank-1, rank-5 and rank-20 scores to display the CMC curve.

The MARS and DukeMTMC-VideoReID datasets have provided the splits for the training set and the testing set, which means the testing identities are fixed. Thus, we report the mAP for these two datasets. However, the PRID2011 and iLIDS-VID datasets do not provide the splits for the training and testing sets. We follow the common strategy to randomly divide the datasets into training sets and testing sets 10 times as in other papers [5], [74], [75] and report the average CMC curve for these two datasets.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCDS.2022.3159557, IEEE Transactions on Cognitive and Developmental Systems

TABLE I
ILIDS-VID DATASET: COMPARISONS OF THE RECOGNITION RATES AT DIFFERENT RANKS (%). THE TOP THREE SCORES ARE INDICATED IN **RED**, **ORANGE** AND **GREEN**, RESPECTIVELY. IF TWO SCORES ARE IDENTICAL, WE HAVE LABELED ALL THOSE SCORES WITH THE SAME COLOR.

Methods	Reference	Backbone	iLIDs-VID			
			rank r=1	rank r=5	rank r=10	rank r=20
RQEN [68]	AAAI18	GoogleNet	77.1	93.2	7.7	99.4
STAN [48]	CVPR18	ResNet50	80.2	-	-	-
ADFD [69]	CVPR19	ResNet50	86.3	97.4	-	99.7
VRSTC [70]	CVPR19	ResNet50	83.4	95.5	97.7	99.5
COSAM [50]	ICCV19	ResNet50	79.6	95.3	-	-
GLTR [71]	ICCV19	ResNet50	86.0	98.0	-	-
SCAN [72] w/o optical	TIP19	ResNet50	81.3	93.3	96.0	98.0
SCAN [72] w optical	TIP19	ResNet50	88.0	96.7	98.0	100
MGH [73]	CVPR20	ResNet50	85.6	97.1	-	99.5
AP3D [53]	ECCV20	AP3D	86.7	-	-	-
DCGN [74]	Multimed Tools Appl21	ResNet	78.5	94.5	-	98.5
PS-GNN [75]	IEEE Signal Process. Lett.21	ResNet50	89.3	98.0	-	99.3
STRF [76]	ICCV21	3D CNN	89.3	-	-	-
USTR (Ours w/o uncertainty)	-	ResNet50	87.2	97.5	98.1	99.7
USTRU (Ours w/ uncertainty)	-	ResNet50	89.7	97.8	98.3	100

TABLE II
PRID 2011 DATASET: COMPARISONS OF THE RECOGNITION RATES AT DIFFERENT RANKS (%). THE TOP THREE SCORES ARE INDICATED IN **RED**, **ORANGE** AND **GREEN**, RESPECTIVELY. IF TWO SCORES ARE IDENTICAL, WE HAVE LABELED ALL THOSE SCORES WITH THE SAME COLOR.

Methods	Reference	Backbone	PRID 2011			
			rank r=1	rank r=5	rank r=10	rank r=20
RQEN [68]	AAAI18	GoogleNet	91.8	98.4	**99.3**	**99.8**
STAN [48]	CVPR18	ResNet50	93.2	-	-	-
ADFD [69]	CVPR19	ResNet50	93.9	**99.5**	-	**100**
GLTR [71]	ICCV19	ResNet50	**95.5**	**100**	-	-
SCAN [72] w/o optical	TIP19	ResNet50	92.0	98.0	**100.0**	**100.0**
SCAN [72] w optical	TIP19	ResNet50	**95.3**	99.0	**100.0**	**100.0**
MGH [73]	CVPR20	ResNet50	**94.8**	**99.3**	-	**100**
DCGN [74]	Multimed Tools Appl21	ResNet50	90.8	96.3	-	**98.9**
PS-GNN [75]	IEEE Signal Process. Lett.21	ResNet50	**95.5**	**100**	**100**	**100**
USTR (Ours w/o uncertainty)	-	ResNet50	94.5	98.7	**100.0**	**100.0**
USTRU (Ours w/ uncertainty)	-	ResNet50	**95.3**	99.2	**100.0**	**100.0**

The hinge margin value in our experiment is set to 4 as we use bi-directional RNN and the dimension of the final representation is 512, which is doubled as compared to [5]. We randomly crop and flip each image in each dataset to augment the data. To train the network, we set the initial learning rate to be $1e-2$ for the first 100 epochs and then change to $1e-3$ for the remaining epochs, the momentum of 0.9, dropout rate of 0.7, and the number of epochs to be 600. The frame features are first extracted by ResNet50, and then the average temporal pooling is used to obtain the sequence feature. Input images are resized to 256×128. The batch size is set to 32. We analyze the use of the snippet representation and also competitive similarity aggregation. Our standard version sets the clip length $L = 4$, each clip includes $T = 4$ images.

C. Experimental Results and Discussion

Comparisons with state-of-the-art methods: We compare our method with the related state-of-the-art results on iLIDs-VID dataset, PRID 2011 dataset, MARS dataset and DukeMTMC-VideoReID dataset, shown in Table I, Table II, Table III, and Table IV, respecitivly. These tables also show the results using RQEN [68], STAN [48], ADFD [69], VRSTC [70], COSAM [50], GLTR [71], SCAN [72] MGH [73], AP3D [53], DCGN [74], PS-GNN [75], DPRM [54], STRF [76], and Bicnet-TKS [79]. We list the quantitative recognition results at different ranks by our method and the above approaches.

We achieve rank 1 recognition rates of 89.7%, 95.3%, 90.1% , and 96.7% for iLIDs-VID dataset, PRID 2011 dataset, MARS dataset, and DukeMTMC-VideoReID dataset, which are reported in Tables I, II, III, and IV, respectively. USTR (**Ours without w/o uncertainty**) refers to the case where we take the same weight for λ_o and λ_m of 0.5, while USTRU (**Ours with w/ uncertainty**) refers the full model using homoscedastic uncertainty to weight the two identification loss terms. The proposed USTRU outperforms USTR at all ranks on all the three datasets.

We highlight the top three identification rates at each rank and mean average precision (mAP) in Table I, II, III, IV. Our model outperforms other compared methods with rank 1 accuracy on iLIDs-VID dataset (Table I) and ranks into the top three places for all other datasets (Table II, III, IV). For the MARS dataset (Table III), our rank 1 recognition rate 90.1% is 0.2% lower than the best 90.3% from STRF [76] with a different backbone of 3D CNN. The 0.2% difference means that our model approximately makes only one more mistake when recognizing around 600 pedestrians compared to the best

TABLE III
MARS dataset: Comparisons of the recognition rates at different ranks (%). The top three scores are indicated in **red**, **orange** and **green**, respectively. If two scores are identical, we have labeled all those scores with the same color.

Methods	Reference	Backbone	MARS			
			rank r=1	rank r=5	rank r=20	mAP
RQEN [68]	AAAI18	GoogleNet	77.8	88.8	94.3	71.1
DuATM [77]	CVPR18	DenseNet121	78.7	90.9	95.8	62.3
STAN [48]	CVPR18	ResNet50	82.3	-	-	65.8
Part-Aligned [78]	ECCV18	InceptionV1	84.7	94.4	97.5	75.9
STA [52]	Arxiv19	ResNet50	86.3	95.7	98.1	80.8
ADFD [69]	CVPR19	ResNet50	87.0	95.4	**98.7**	78.2
VRSTC [70]	CVPR19	ResNet50	88.5	96.5	97.4	82.3
COSAM [50]	ICCV19	ResNet50	84.0	95.5	97.9	79.9
GLTR [71]	ICCV19	ResNet50	87.0	95.8	98.2	78.5
SCAN [72] w/o optical	TIP19	ResNet50	86.6	94.8	97.1	76.7
SCAN [72] w optical	TIP19	ResNet50	87.2	95.2	98.1	77.2
MGH [73]	CVPR20	ResNet50	90.0	**96.7**	**98.5**	**85.8**
AP3D [53]	ECCV20	AP3D	**90.1**	-	-	85.1
DCGN [74]	Multimed Tools Appl21	ResNet50	89.6	96.5	98.3	81.8
PS-GNN [75]	IEEE Signal Process. Lett.21	ResNet50	87.1	94.9	97.1	76.0
DPRM [54]	TIP21	DPRM	89.0	**96.6**	98.3	83.0
Bicnet-TKS [79]	CVPR21	Bicnet	**90.2**	-	-	**86.0**
STRF [76]	ICCV21	3D CNN	90.3	-	-	86.1
USTR (Ours w/o uncertainty)	-	ResNet50	89.2	96.3	98.2	82.4
USTRU (Ours w/ uncertainty)	-	ResNet50	**90.1**	**96.8**	**98.6**	**85.8**

TABLE IV
DukeMTMC-VideoReID dataset: Comparisons of the recognition rates at different ranks (%). The top three scores are indicated in **red**, **orange** and **green**, respectively. If two scores are identical, we have labeled all those scores with the same color.

Methods	Reference	Backbone	DukeMTMC-VideoReID			
			rank r=1	rank r=5	rank r=20	mAP
DuATM [77]	CVPR18	DenseNet121	81.8	90.2	95.3	64.6
Part-Aligned [78]	ECCV18	InceptionV1	84.4	92.2	95	69.3
STA [52]	Arxiv19	ResNet50	96.2	**99.3**	99.6	94.9
VRSTC [70]	CVPR19	ResNet50	95	99.1	99.4	93.5
COSAM [50]	ICCV19	ResNet50	95.4	**99.3**	-	94.1
GLTR [71]	ICCV19	ResNet50	96.3	**99.3**	-	93.7
AP3D [53]	ECCV20	AP3D	96.3	-	-	95.6
DCGN [74]	Multimed Tools Appl21	ResNet50	95.4	98.7	99.4	93.1
PS-GNN [75]	IEEE Signal Process. Lett.21	ResNet50	95.9	**99.2**	**99.7**	93.1
DPRM [54]	TIP21	DPRM	**97.1**	**99.4**	**100**	95.6
Bicnet-TKS [79]	CVPR21	Bicnet	96.3	-	-	**96.1**
STRF [76]	ICCV21	3D CNN	97.4	-	-	96.4
USTR (Ours w/o uncertainty)	-	ResNet50	95.9	99.1	**99.9**	93.5
USTRU (Ours w/ uncertainty)	-	ResNet50	**96.7**	**99.3**	**99.9**	**95.8**

TABLE V
Comparison results of methods tested on all datasets. The top two scores are indicated in **red** and **orange**, respectively. Not all the methods listed in previous Table I, II, III, IV are shown in this table since not all the methods are tested on all datasets. If two scores are identical, we have labeled all those scores with the same color.

Datasets Methods	MARs		DukeMTMC-VideoReID		iLIDs-VID	PRID 2011
	rank r=1	mAP	rank r=1	mAP	rank r=1	rank r=1
GLTR [71]	87.0	78.5	**96.3**	**93.7**	86.0	**95.5**
DCGN [74]	**89.6**	**81.8**	95.4	93.1	78.5	90.8
PS-GNN [75]	87.1	76	95.9	93.1	**89.3**	**95.5**
USTR (Ours w/o uncertainty)	89.2	82.4	95.9	93.5	87.2	94.5
USTRU (Ours w/ uncertainty)	**90.1**	**85.8**	**96.7**	**95.8**	**89.7**	**95.3**

TABLE VI
Spatial component analysis on iLIDs-VID, PRID, MARS and DukeMTMC-VideoReID (Duke) datasets.

Inputs	Rank 1			
	iLIDs-VID	PRID 2011	MARS	Duke
RGB branch	81.2	89.1	82.6	92.6
Mask-RGB branch	83.4	90.7	83.7	93.0
USTRU	89.7	95.3	90.1	96.7

TABLE VII
Temporal Component Analysis on TiLIDs-VID, PRID, MARS and DukeMTMC-VideoReID (Duke) datasets.

Inputs	Rank 1			
	iLIDs-VID	PRID 2011	MARS	Duke
average pooling	82.7	93.7	86.1	95.5
max pooling	81.9	92.7	85.1	94.8
RNN	80.5	92.5	84.3	94.2
BRNN	81.4	93.1	85.2	94.7
BLSTM	84.9	93.8	86.6	95.1
USTRU	89.7	95.3	90.1	96.7

model. For the PRID 2011 dataset, we get the second highest rank 1 score of 95.3% compared to the best 95.5% (Table II). For DukeMTMC-VideoReID dataset, our USTRU rank 1 score is in the third place (Table IV). Our model is capable in handling the more challenging datasets, e.g, iLIDs-VID and MARS datasets.

Note that not all the methods have been tested on all the four datasets. Table V shows the comparison results for those algorithms in Table I-IV with results across the four datasets. We highlight the top two scores in red and orange colors, respectively. There is no model that could get the best results in all datasets. Our USTRU approach is relatively more stable and could consistently demonstrate relatively good results on different datasets (see Table V) for it achieves the best results on all datasets except PRID 2011 (2nd), while the other methods have bias on some specific datasets.

Evaluating the contribution of the spatial component of the approach: The contributions of the masked foreground image and original image information to the re-ID system. The results of comparison shown in Table VI, from which we make the following observations. The background information contributes to re-ID. Rank-1 results drop by 6.3%, 4.6%, 6.4%, and 3.7% when only the masked foreground (second row) is used without background information for iLIDs-VID, PRID 2011, MARS, and DukeMTMC-VideoReID, respectively. Modeling foreground and original image (that has background) in two branches improves the results significantly. The two-branch model reaches rank 1 accuracy of 89.7%, surpassing the one branch model RGB by 8.5% for the most challenging iLIDs-VID dataset.

TABLE VIII
Performance of different sequence lengths on iLIDs-VID, PRID, MARS and DukeMTMC-VideoReID (Duke) datasets.

Clips	Rank 1			
	iLIDs-VID	PRID 2011	MARS	Duke
N = 2	84.6	94.1	87.4	95.3
N = 4	89.7	95.3	90.1	96.7
N = 8	88.3	94.9	88.8	96.2

Evaluating the contribution of the temporal component of the approach: Compared to image-based person re-identification, most video-based re-id methods encode the temporal information either by applying the pooling layers to summarize the features from the frames or by using RNNs and their variants to embed the temporal information. We evaluate the results while applying different temporal embedding components including average pooling, max pooling, RNN, bi-directional RNN (BRNN) and bi-directional LSTM (BLSTM), respectively. The results on MARS dataset are shown in Table VII. Our USTRU with a bi-directional sparse attentive backtracking model achieved the best results. We also find that average pooling and max pooling perform better than RNN and BRNN. The pooling operation captures and summarizes the long-term information along the sequence, while the RNN and BRNN are not good at learning long-term dependencies. Further, both BLSTM and our model embed the self-attention mechanism, while BLSTM is still inferior to our model due to its inherent BPTT (backpropagation through time) backbone.

Additionally, to show the efficiency of our method, we provide an analysis of the runtime of our method. We implement our model with PyTorch and train it end-to-end. For iLIDs-VID dataset, It takes about 8.5 hours, 6 hours and 13.5 hours to train USTRU (Our model), BRNN model and BLSTM models, respectively, using the Nvidia GTX-1080 GPU. Our model runs much faster than the BLSTM model due to the complicated calculation of gaits in BLSTM. On the other hand, our model runs a little slower compared to the BRNN model since we add the attention mechanism. However, we achieved 89.7%. for rank 1 accuracy, which is 8.3% higher than the compared BRNN model (81.4%) for that our model is capable of selecting important frames for backtraking.

Evaluation for different sequence lengths of the approach: In order to capture both the long-term and short-term information, the video sequences are first divided into clips, where each clip includes the adjacent $T = 4$ image frames. Then our model selects the top N clips according to the attentive weights for the backpropagation. We investigate the effect of N on the performance. Table VIII shows the comparison results for using different numbers of clips on three dataset. When $N = 4$, our USTRU model achieves the best ranking scores when 16 frames are selected.

Cross Dataset Generalization: Due to the various conditions in the process of data collection, the data distributions of different datasets may have a great bias. The performance of the model trained on one dataset may drop a lot on another one. To evaluate the generalization ability of the proposed model, we conduct cross-dataset validation with the following setting [82]: we use iLIDs-VID, MARS and DukeMTMC-VideoReID datasets as the training sets, respectively and use the PRID 2011 as the testing set. Table IX reports the results for rank 1, 5, 20. Our USTRU model achieves approximately best recognition rates among all the listed methods.

Visualization of attention weights for backpropagation: We investigate the effects of the spatial and temporal cues in our method. As shown in Fig. 8-11 the color of the bar under

TABLE IX
CROSS DATASET MATCHING RESULTS ON PRID 2011 DATASET. THE FIRST ROW INDICATES THE TRAINING DATASET.

Training dataset	iLIDs-VID			MARS			DukeMTMC-VideoReID		
	rank r=1	rank r=5	rank r=20	rank r=1	rank r=5	rank r=20	rank r=1	rank r=5	rank r=20
CNN-RNN [5]	28.0	57.0	81.0	-	-	-	-	-	-
ASTPN [46]	30.0	58.0	85.0	-	-	-	-	-	-
TPL [82]	29.5	59.4	82.2	35.2	69.6	89.3	-	-	-
SCAN [72]	42.8	71.6	88.9	46.0	69.0	91.0	-	-	-
USTRU (Ours)	43.1	69.7	91.3	48.6	71.7	94.3	45.2	68.5	92.1

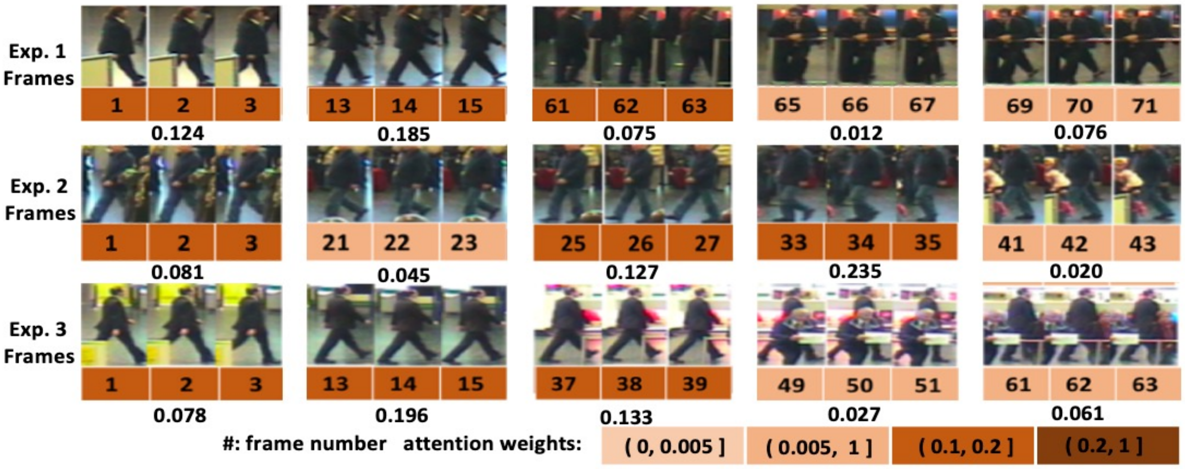

Fig. 8. iLIDs-VID dataset: Visualization of attention weights for backpropagation. Three examples are chosen (Exp. 1, Exp. 2. and Exp. 3) In each row, selected frames for the given video are listed. The number under each image represents the frame number from the original video and the color of the rectangular box represents the attention weights for the sparse attentive backtracking. The darker the color is, the higher the weight is. The reference ruler for the weight is shown on the bottom.

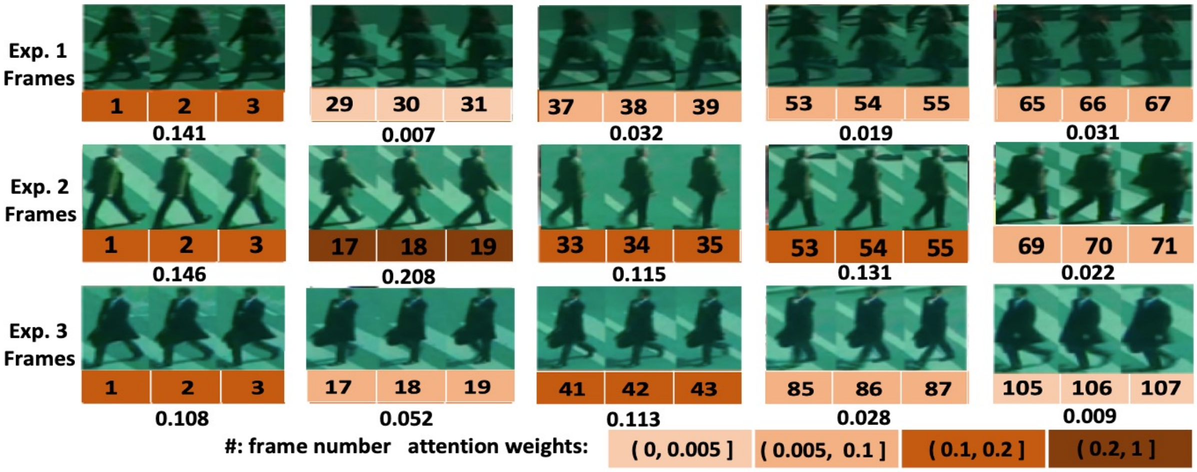

Fig. 9. PRID 2011 dataset: Visualization of attention weights for backpropagation. Three examples are chosen (Exp. 1, Exp. 2. and Exp. 3). In each row, selected frames for the given video are listed. The number under each image represents the frame number from the original video and the color of the rectangular box represents the attention weights for the sparse attentive backtracking. The darker the color is, the higher the weight is. The reference ruler for the weight is shown on the bottom.

each image is an indicator of the attention weights for the backtracking. As the color gets darker, the weight is higher.

a) iLIDs-VID Dataset: Fig. 8 shows three successful examples for our method. The three target people are walking in the lobby where they get occluded by different people and other objects from time to time. Our method is able to assign higher weights to the frames where there are fewer occlusions, for example, frame 13, 14, 15 for the first person, frame 33, 34, 35 for the second person and frame 13, 14, 15 for the third person. Also, the attention weights become lower when the overlapping is more. For example in example 3, at the very beginning, the first person is overlapped by a yellow board, the given attention weight (frame 1, 2, 3) is just a little lower than for frame 13, 14, 15. When the person is gradually blocked

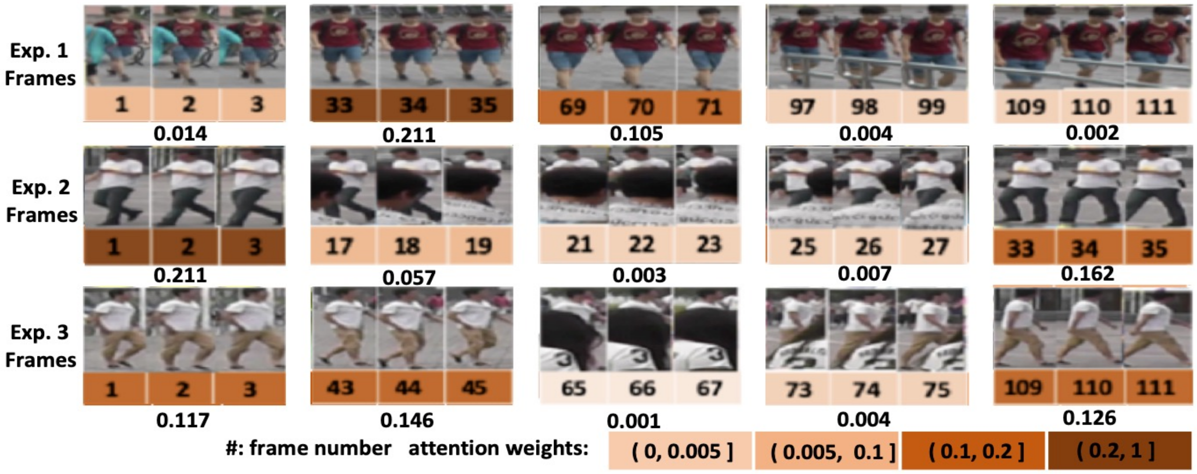

Fig. 10. MARS dataset: Visualization of attention weights for backpropagation. Three examples are chosen (Exp. 1, Exp. 2. and Exp. 3). In each row, selected frames for the given video are listed. The number under each image represents the frame number from the original video and the color of the rectangular box represents the attention weights for the sparse attentive backtracking. The darker the color is, the higher the weight is. The reference ruler for the weight is shown on the bottom.

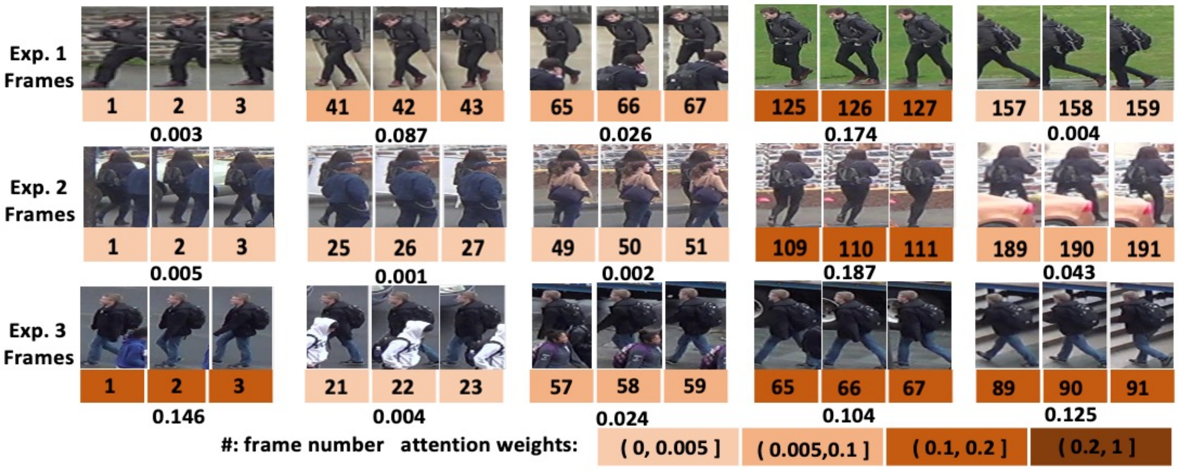

Fig. 11. DukeMTMC-VideoReID dataset: Visualization of attention weights for backpropagation. Three examples are chosen (Exp. 1, Exp. 2. and Exp. 3). In each row, selected frames for the given video are listed. The number under each image represents the frame number from the original video and the color of the rectangular box represents the attention weights for the sparse attentive backtracking. The darker the color is, the higher the weight is. The reference ruler for the weight is shown on the bottom.

by the other people, the attention weights drop significantly for frame 49, 50, 51 and frame 61, 62, 63.

b) PRID 2011 Dataset: Fig. 9 illustrates another three correct instances. In general, the colors for this dataset has a less diffused distribution than the other two datasets shown in Fig. 8 and Fig. 10. The reason is that PRID 2011 dataset is relatively simple and includes less complex background and less instances of occlusions. The learned attention weights are consistent with the degree of blurriness for the given frames. For example, the leg of the second person is not very clear in frames 69, 70, 71. Similarly, we could hardly see one leg of the first person in frames 29, 30, 31.

c) MARS Dataset: Fig. 10 presents the other three detected examples at their first attempt. The target person is obscured by other interfering pedestrians as in the iLIDs-VID dataset. If we check frames 17, 18, 19 and 21, 22, 23 for the second example, the learned weights decrease as the target person is gradually covered by the other guy. For the following frames 25, 26, 27, the weights increase when the other pedestrian passes by.

d) DukeMTMC-VideoReID Dataset: Fig. 11 displays another three right cases. The target pedestrians suffer a lot from occlusions as in the iLIDs-VID and Mars dataset. Besides, there are more frames in each tracklet than the other three datasets. This indicates that the background changes a lot when the target person is walking. For the first example, the person walks on the flat ground in frame 1, 2, 3, and then goes up the stairs in frame 41, 42, 43, and then walks on the grassland in frame 125, 126, 127.

In summary, the proposed approach is capable to learn the unbiased representation by focusing on the main parts of a person (spatial unbiased representation) and finding the useful frames (temporal unbiased representation) regardless of the position where the frame is in a sequence.

V. Conclusions

This paper proposed an unbiased spatio-temporal learning framework to address video-based person re-id. The proposed framework explicitly removed the background clutter by learning a two branch CNN network. Homoscedastic uncertainty is used to balance the original and masked foreground branches. In addition, the long term dependency issue is handled with sparse attentive backtracking. Extensive experiments are conducted on three person re-ID benchmark datasets, where the proposed framework achieved favorable performance compared with the recent state-of-the-art methods.

Acknowledgment

This work was supported in part by NSF grants 1552454, 1330110 and Bourns Endowment funds. The contents of the information do not reflect the position or policy of US Government.

References

[1] W. Hu, T. Tan, L. Wang, and S. Maybank, "A survey on visual surveillance of object motion and behaviors," *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, vol. 34, no. 3, pp. 334–352, 2004.

[2] S. Zhang, J. Wang, Z. Wang, Y. Gong, and Y. Liu, "Multi-target tracking by learning local-to-global trajectory models," *Pattern Recognition*, vol. 48, no. 2, pp. 580–590, 2015.

[3] Z. Imani and H. Soltanizadeh, "Person re-identification using local pattern descriptors and anthropometric measures from videos of kinect sensor," *IEEE Sensors Journal*, vol. 16, no. 16, pp. 6227–6238, 2016.

[4] R. Theagarajan, F. Pala, X. Zhang, and B. Bhanu, "Soccer: Who has the ball? generating visual analytics and player statistics," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2018, pp. 1749–1757.

[5] N. McLaughlin, J. M. del Rincon, and P. Miller, "Recurrent convolutional network for video-based person re-identification," in *CVPR*, 2016.

[6] R. R. Varior, B. Shuai, J. Lu, D. Xu, and G. Wang, "A siamese long short-term memory architecture for human re-identification," in *ECCV*, 2016.

[7] W. Zhang, X. Yu, and X. He, "Learning bidirectional temporal cues for video-based person re-identification," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 28, no. 10, pp. 2768–2776, 2017.

[8] M. Tian, S. Yi, H. Li, S. Li, X. Zhang, J. Shi, J. Yan, and X. Wang, "Eliminating background-bias for robust person re-identification," in *CVPR*, 2018.

[9] C. Song, Y. Huang, W. Ouyang, and L. Wang, "Mask-guided contrastive attention model for person re-identification," in *CVPR*, 2018.

[10] Y. Bengio, P. Simard, and P. Frasconi, "Learning long-term dependencies with gradient descent is difficult," *IEEE Transactions on Neural Networks*, vol. 5, no. 2, pp. 157–166, 1994.

[11] T. Xiao, H. Li, W. Ouyang, and X. Wang, "Learning deep feature representations with domain guided dropout for person re-identification," in *CVPR*, 2016.

[12] F. Liu and L. Zhang, "View confusion feature learning for person re-identification," in *ICCV*, 2019.

[13] A. Khatun, S. Denman, S. Sridharan, and C. Fookes, "Semantic consistency and identity mapping multi-component generative adversarial network for person re-identification," in *WACV*, 2020.

[14] N. R. Ke, A. Goyal, O. Bilaniuk, J. Binas, L. Charlin, C. Pal, and Y. Bengio, "Sparse attentive backtracking: Long-range credit assignment in recurrent networks," *arXiv preprint arXiv:1711.02326*, 2017.

[15] R. Satta, "Appearance descriptors for person re-identification: a comprehensive review," *arXiv preprint arXiv:1307.5748*, 2013.

[16] M. Farenzena, L. Bazzani, A. Perina, V. Murino, and M. Cristani, "Person re-identification by symmetry-driven accumulation of local features," in *CVPR*, 2010.

[17] B. Ma, Y. Su, and F. Jurie, "Bicov: a novel image representation for person re-identification and face verification," in *BMVC*, 2012.

[18] R. Zhao, W. Ouyang, and X. Wang, "Unsupervised salience learning for person re-identification," in *CVPR*, 2013.

[19] E. Corvee, F. Bremond, M. Thonnat *et al.*, "Person re-identification using Haar-based and DCD-based signature," in *AVSS*, 2010.

[20] R. Satta, G. Fumera, F. Roli, M. Cristani, and V. Murino, "A multiple component matching framework for person re-identification," in *ICIAP*, 2011.

[21] D. Gray and H. Tao, "Viewpoint invariant pedestrian recognition with an ensemble of localized features," in *ECCV*, 2008.

[22] D.-N. T. Cong, L. Khoudour, C. Achard, C. Meurie, and O. Lezoray, "People re-identification by spectral classification of silhouettes," *Signal Processing*, vol. 90, no. 8, pp. 2362–2374, 2010.

[23] F. Pala, R. Satta, G. Fumera, and F. Roli, "Multimodal person re-identification using rgb-d cameras," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 26, no. 4, pp. 788–799, 2016.

[24] Y. Liu, Y. Zhang, S. Coleman, B. Bhanu, and S. Liu, "A new patch selection method based on parsing and saliency detection for person re-identification," *Neurocomputing*, vol. 374, pp. 86–99, 2020.

[25] Y. Wu, K. Zhang, D. Wu, C. Wang, C.-A. Yuan, X. Qin, T. Zhu, Y.-C. Du, H.-L. Wang, and D.-S. Huang, "Person re-identification by multi-scale feature representation learning with random batch feature mask," *IEEE Transactions on Cognitive and Developmental Systems*, 2020.

[26] X. Qian, Y. Fu, T. Xiang, Y.-G. Jiang, and X. Xue, "Leader-based multi-scale attention deep architecture for person re-identification," *IEEE transactions on pattern analysis and machine intelligence*, vol. 42, no. 2, pp. 371–385, 2019.

[27] X. Zhang, F. Pala, and B. Bhanu, "Attributes co-occurrence pattern mining for video-based person re-identification," in *AVSS*, 2017.

[28] C. Su, F. Yang, S. Zhang, Q. Tian, L. S. Davis, and W. Gao, "Multi-task learning with low rank attribute embedding for multi-camera person re-identification," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 5, pp. 1167–1181, 2018.

[29] M. Koestinger, M. Hirzer, P. Wohlhart, P. M. Roth, and H. Bischof, "Large scale metric learning from equivalence constraints," in *CVPR*, 2012.

[30] S. Liao, Y. Hu, X. Zhu, and S. Z. Li, "Person re-identification by local maximal occurrence representation and metric learning," in *CVPR*, 2015.

[31] A. Mignon and F. Jurie, "Pcca: A new approach for distance learning from sparse pairwise constraints," in *CVPR*, 2012.

[32] L. An, X. Chen, S. Yang, and B. Bhanu, "Sparse representation matching for person re-identification," *Information Sciences*, vol. 355, pp. 74–89, 2016.

[33] D. Chen, Z. Yuan, G. Hua, N. Zheng, and J. Wang, "Similarity learning on an explicit polynomial kernel feature map for person re-identification," in *CVPR*, 2015.

[34] W.-S. Zheng, S. Gong, and T. Xiang, "Re-identification by relative distance comparison," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 3, pp. 653–668, 2013.

[35] L. An, M. Kafai, S. Yang, and B. Bhanu, "Person re-identification with reference descriptor," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 26, no. 4, pp. 776–787, 2015.

[36] L. An, S. Yang, and B. Bhanu, "Person re-identification by robust canonical correlation analysis," *IEEE signal processing letters*, vol. 22, no. 8, pp. 1103–1107, 2015.

[37] W. Li, R. Zhao, T. Xiao, and X. Wang, "Deepreid: Deep filter pairing neural network for person re-identification," in *CVPR*, 2014.

[38] E. Ahmed, M. Jones, and T. K. Marks, "An improved deep learning architecture for person re-identification," in *CVPR*, 2015.

[39] S. Ding, L. Lin, G. Wang, and H. Chao, "Deep feature learning with relative distance comparison for person re-identification," *Pattern Recognition*, vol. 48, no. 10, pp. 2993–3003, 2015.

[40] C. Zhao, X. Lv, Z. Zhang, W. Zuo, J. Wu, and D. Miao, "Deep fusion feature representation learning with hard mining center-triplet loss for person re-identification," *IEEE Transactions on Multimedia*, 2020.

[41] J. Han and B. Bhanu, "Individual recognition using gait energy image," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 28, no. 2, pp. 316–322, 2006.

[42] K. Liu, B. Ma, W. Zhang, and R. Huang, "A spatio-temporal appearance representation for video-based pedestrian re-identification," in *ICCV*, 2015.

[43] 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 *ECCV*, 2016.

[44] T. Wang, S. Gong, X. Zhu, and S. Wang, "Person re-identification by video ranking," in *ECCV*, 2014.

[45] 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*, 2017.

[46] S. Xu, Y. Cheng, K. Gu, Y. Yang, S. Chang, and P. Zhou, "Jointly attentive spatial-temporal pooling networks for video-based person re-identification," *arXiv preprint arXiv:1708.02286*, 2017.

[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 *CVPR*, 2017.

[48] S. Li, S. Bak, P. Carr, and X. Wang, "Diversity regularized spatiotemporal attention for video-based person re-identification," in *CVPR*, 2018.

[49] L. Wu, Y. Wang, J. Gao, and X. Li, "Where-and-when to look: Deep siamese attention networks for video-based person re-identification," *IEEE Transactions on Multimedia*, vol. 21, no. 6, pp. 1412–1424, 2018.

[50] A. Subramaniam, A. Nambiar, and A. Mittal, "Co-segmentation inspired attention networks for video-based person re-identification," in *ICCV*, 2019.

[51] 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 *CVPR*, 2018.

[52] Y. Fu, X. Wang, Y. Wei, and T. Huang, "Sta: Spatial-temporal attention for large-scale video-based person re-identification," in *AAAI*, 2019.

[53] X. Gu, H. Chang, B. Ma, H. Zhang, and X. Chen, "Appearance-preserving 3d convolution for video-based person re-identification," in *ECCV*, 2020.

[54] X. Yang, L. Liu, N. Wang, and X. Gao, "A two-stream dynamic pyramid representation model for video-based person re-identification," *IEEE Transactions on Image Processing*, vol. 30, pp. 6266–6276, 2021.

[55] A. Kendall and Y. Gal, "What uncertainties do we need in bayesian deep learning for computer vision?" in *NeurIPS*, 2017.

[56] 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.

[57] X. Zhang and B. Bhanu, "An unbiased temporal representation for video-based person re-identificatio," in *ICIP*, 2018.

[58] P. Luo, X. Wang, and X. Tang, "Pedestrian parsing via deep decompositional network," in *ICCV*, 2013.

[59] N. Ansari and E. J. Delp, "On detecting dominant points," *Pattern Recognition*, vol. 24, no. 5, pp. 441–451, 1991.

[60] D. Chen, S. Zhang, W. Ouyang, J. Yang, and Y. Tai, "Person search via a mask-guided two-stream cnn model," in *ECCV*, 2018.

[61] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *CVPR*, 2018.

[62] W. Li, X. Zhu, and S. Gong, "Harmonious attention network for person re-identification," in *CVPR*, 2018.

[63] Z. Zhang, C. Lan, W. Zeng, X. Jin, and Z. Chen, "Relation-aware global attention for person re-identification," in *CVPR*, 2020.

[64] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," *arXiv preprint arXiv:1409.0473*, 2014.

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

[66] A. Kendall, Y. Gal, and R. Cipolla, "Multi-task learning using uncertainty to weigh losses for scene geometry and semantics," in *CVPR*, 2018.

[67] C.-P. Tay, S. Roy, and K.-H. Yap, "Aanet: Attribute attention network for person re-identifications," in *CVPR*, 2019.

[68] G. Song, B. Leng, Y. Liu, C. Hetang, and S. Cai, "Region-based quality estimation network for large-scale person re-identification," in *AAAI*, 2018.

[69] Y. Zhao, X. Shen, Z. Jin, H. Lu, and X.-s. Hua, "Attribute-driven feature disentangling and temporal aggregation for video person re-identification," in *CVPR*, 2019.

[70] R. Hou, B. Ma, H. Chang, X. Gu, S. Shan, and X. Chen, "Vrstc: Occlusion-free video person re-identification," in *CVPR*, 2019.

[71] J. Li, J. Wang, Q. Tian, W. Gao, and S. Zhang, "Global-local temporal representations for video person re-identification," in *ICCV*, 2019.

[72] R. Zhang, J. Li, H. Sun, Y. Ge, P. Luo, X. Wang, and L. Lin, "Scan: Self-and-collaborative attention network for video person re-identification," *IEEE Transactions on Image Processing*, vol. 28, no. 10, pp. 4870–4882, 2019.

[73] Y. Yan, J. Qin, J. Chen, L. Liu, F. Zhu, Y. Tai, and L. Shao, "Learning multi-granular hypergraphs for video-based person re-identification," in *CVPR*, 2020.

[74] C. Chen, M. Qi, G. Huang, J. Wu, J. Jiang, and X. Li, "Learning discriminative features with a dual-constrained guided network for video-based person re-identification," *Multimedia Tools and Applications*, pp. 1–24, 2021.

[75] Y. Li, Z. Guo, H. Zhang, M. Li, and G. Ji, "Decoupled pose and similarity based graph neural network for video person re-identification," *IEEE Signal Processing Letters*, 2021.

[76] A. Aich, M. Zheng, S. Karanam, T. Chen, A. Roy-Chowdhury, and Z. Wu, "Spatio-temporal representation factorization for video-based person re-identification," in *ICCV*, 2021.

[77] 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 *CVPR*, 2018.

[78] Y. Suh, J. Wang, S. Tang, T. Mei, and K. Mu Lee, "Part-aligned bilinear representations for person re-identification," in *ECCV*, 2018.

[79] R. Hou, H. Chang, B. Ma, R. Huang, and S. Shan, "Bicnet-tks: Learning efficient spatial-temporal representation for video person re-identification," in *CVPR*, 2021.

[80] M. Hirzer, C. Beleznai, P. M. Roth, and H. Bischof, "Person re-identification by descriptive and discriminative classification," in *Scandinavian conference on Image analysis*, 2011.

[81] E. Ristani, F. Solera, R. Zou, R. Cucchiara, and C. Tomasi, "Performance measures and a data set for multi-target, multi-camera tracking," in *ECCV*, 2016.

[82] J. Dai, P. Zhang, D. Wang, H. Lu, and H. Wang, "Video person re-identification by temporal residual learning," *IEEE Transactions on Image Processing*, vol. 28, no. 3, pp. 1366–1377, 2018.

Xiu Zhang received the B.Eng. degree in computer science and technology from Xidian University, Xi'an, China, in 2012, the M.Sc degree in computer in applied technology from Fudan Unviersity, Shanghai, China, in 2015. She is currently pursuing the Ph.D. degree in computer science at the University of California at Riverside, Riverside, CA, USA. Her research interests are in computer vision and music information retrieval. Her recent research has been concerned with video-based person re-identification.

Bir Bhanu (F'95-LF'17) received the S.M. and E.E. degrees in electrical engineering and computer science from Massachusetts Institute of Technology, Cambridge, MA, USA; the Ph.D. degree in electrical engineering from the Image Processing Institute, University of Southern California, Los Angeles, CA, USA; and the M.B.A. degree from University of California, Irvine, CA, USA. He also received M.E. (with Distinction) from BITS-Pilani and B.S. (with Honors) from IIT-BHU. He is Bourns Endowed University of California Presidential Chair in Engineering and Distinguished Professor of Electrical and Computer Engineering, Cooperative Professor of Computer Science and Engineering, Mechanical Engineering, and Bioengineering with University of California, Riverside, CA, USA. He is the Director of Visualization and Intelligent Systems Laboratory (VISLab) at UCR since 1991. He has served as the Founding Director of the Center for Research in Intelligent Systems (1998-2019), the Director of NSF IGERT Program on Video Bioinformatics (2009-2016) and the Founding Chair of Electrical Engineering (1991-1994) at UCR. Prior to joining UCR in 1991, he was Senior Honeywell Fellow at Honeywell Inc. He has authored over 560 reviewed technical publications, including over 170 journal papers, 378 conference papers and 59 book chapters. He has published seven authored and five edited books. He also holds 18 patents. His research interests include computer vision, pattern recognition and data mining, machine learning, artificial intelligence, image processing, image and video database, graphics and visualization, robotics, human–computer interactions, and biological, medical, military, and intelligence applications. Dr. Bhanu is a fellow of IEEE, AAAS, AIMBE, IAPR, SPIE and NAI. He has been the Principal Investigator of various programs for NSF, DARPA, IARPA, NASA, AFOSR, ONR, ARO, and other agencies and industries in the areas of video networks, video understanding, video bioinformatics, learning and vision, image understanding, pattern recognition, target recognition, biometrics, autonomous navigation, image databases, and machine-vision applications.