What-Where-When Attention Network for Video-based Person Re-identification

Chenrui Zhang[a,b], Ping Chen[a,*], Tao Lei[c], Yangxu Wu[a], and Hongying Meng[d]

[a] *The State Key Laboratory for Electronic Testing Technology, North University of China, Taiyuan 030051, China*
[b] *The Department of Physics, Luliang University, Luliang 033000, China*
[c] *The School of Electronic Information and Artificial Intelligence, Shaanxi University of Science and Technology, Xi'an 710021, China*
[d] *The Department of Electronic and Electrical Engineering, Brunel University London, U.K*

Abstract

Video-based person re-identification plays a critical role in intelligent video surveillance by learning temporal correlations from consecutive video frames. Most existing methods aim to solve the challenging variations of pose, occlusion, backgrounds and so on by using attention mechanism. They almost all draw attention to the occlusion and learn occlusion-invariant video representations by abandoning the occluded area or frames, while the other areas in these frames contain sufficient spatial information and temporal cues. To overcome these drawbacks, this paper proposes a comprehensive attention mechanism covering *what*, *where*, and *when* to pay attention in the discriminative spatial-temporal feature learning, namely *What-Where-When* Attention Network (W3AN). Concretely, W3AN designs a spatial attention module to focus on pedestrian identity and obvious attributes by the importance estimating layer (*What* and *Where*), and a temporal attention module to calculate the frame-level importance (*when*), which is embedded into a graph attention network to exploit temporal attention features rather than computing weighted average feature for video frames like existing methods. Moreover, the experiments on

[☆] Fully documented templates are available in the elsarticle package on CTAN.
[*] Corresponding author
Email address: support@elsevier.com (Ping Chen)
URL: www.elsevier.com (Chenrui Zhang), www.elsevier.com (Tao Lei), www.elsevier.com (Yangxu Wu), www.elsevier.com (and Hongying Meng)

three widely-recognized datasets demonstrate the effectiveness of our proposed W3AN model and the discussion of major modules elaborates the contributions of this paper.

Keywords: Person re-identification, *What-where-when* attention,
Spatial-temporal feature, Graph attention network, Attribute, Identity.

1. Introduction

Person re-identification (re-id) is a popular research on the video surveillance, due to its extensive applications in forensic search, target human tracking, and analysis of pedestrian trajectory. The goal of re-id is to find the correct person from the non-overlapped camera videos, given a target image or video sequence. To achieve satisfactory performance, person re-id must overcome the challenging variations of pose, illumination, occlusion, and background, similar to general object recognition tasks. In general, studies on re-id are mainly divided into two categories: image based and video based matching tasks. The image based person re-id attempts to retrieve the target pedestrian between single still images, while the video-based re-id is of matching between video sequences. Though image-based re-id has performed effective improvements [1, 2, 3], they are more susceptible than video analysis, when the target data are exposed to the challenging variations, especially for occlusion. Compared to still image, pedestrian video involves not only more appearance characteristics, but also temporal information to represent the object comprehensively [4, 5]. That permits the video-based person re-id task potentially relaxed from various constraints, such as occlusion, and pose variance.

The majority of existing person re-id techniques pay attention to diverse distance metric learning. Image-based person re-id has been further developed by convolutional neural networks (CNN) [2, 3], while most video based methods extract temporal cues by integrating the learned CNN features of each frame [6, 4]. Specifically, the video-based person re-identification models extract appearance spatial information from each video frame, and then aggregate these image-

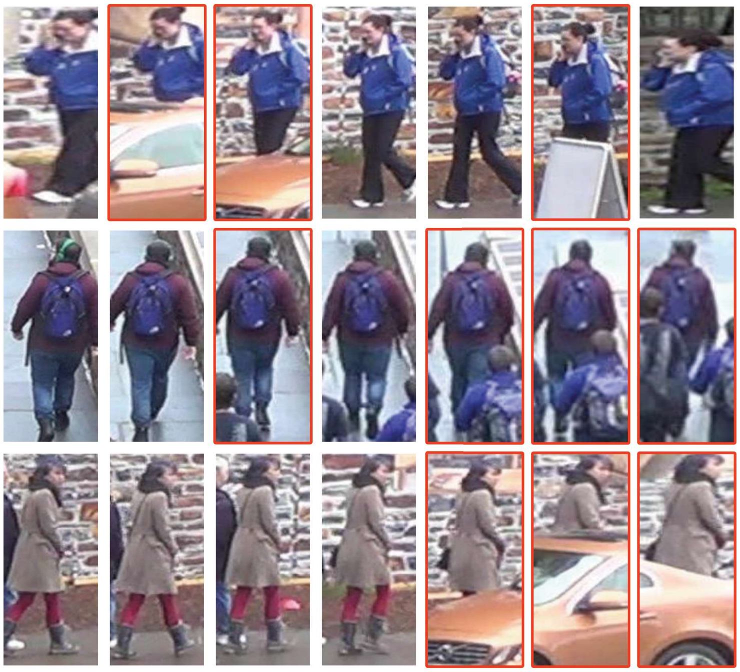


Figure 1: Examples of pedestrian video sequences. There are frequently-occurred occlusion (red rectangle), and pose variations. Besides, the pedestrian attributes devote large contribution to the identity recognition, such as wearing backpack.

level features into the final spatial-temporal feature representations, which are constrained by the distance metric learning. Besides, several researches [7, 8] introduce optical flow into motion feature learning. However, an important left problem of them is neglecting the frequently-occurred partial occlusion in video frames. Recently, several attention models have made great efforts on the problem of partial occlusion [5, 9, 10]. They employ attention module to learn occlusion-invariant video representations by discarding the occluded video frames. However, the most areas of occluded frames contain spatial information and temporal cues in the video sequences, so directly discarding these frames is inexpedient (as shown in Figure 1).

As mentioned in research [11], a faultless attention mechanism should learn *what* (objects and their local motion patterns), *where* (spatial), and *when* (temporal) to focus on. Existing attention-based person re-id models only utilize one of them, such as Rahman et al [7] exploited temporal information by aggregating the attention scores on each video frame, according to that not all frames in a video are equally informative. In this paper, we focus on *What-Where-When* (W3) attention problem in video-based person re-identification task, to integrate *what* (looking at person appearance), where (focusing on obvious pedestrian attributes), and when (emphasizing important video frames) into a unified attention model for re-id. Firstly, the model should look at the human body to learn global identity appearance information from all video frames (*what*). Then, the attention model needs to focus on where contains obvious pedestrian attributes (such as red hat, green trouser) (*where*). Finally, it should evaluate the importances of each frame because the occluded frame contains less spatial information than normal frames, due to the noise from frequently-occurred occlusion (*when*). In summary, the terms of *what* and *where* concern overcoming the occlusion in exploiting spatial feature, while the term of *when* pays attention to alleviating the impact of occluded region in temporal pattern.

In this paper, we propose a *What-Where-When* Attention Network (W3AN) to particularly cope with partial occlusion both in spatial and temporal patterns, by looking at the identity appearance, focusing on obvious attributes, and emphasizing important frames in video person re-identification task. More concretely, this paper integrates a spatial attention module and a temporal attention module to focus on pedestrian identity appearance and obvious attributes guiding by the class activation map (CAM), which can solve the *what* and *where* to pay attention. Besides, a graph attention network is proposed for the important frames in order to tackle the problem that draws attention to *when* in video-based person re-identification. Finally, the *What-Where-When* attention features are learned by the graph attention network with importance adjacent matrix of the video frames.

The contributions of our W3AN model are summarized below.

(1) We firstly study the overall attention model for video-based person re-identification, that is *What*, *Where* and *When* to pay attention, and proposes a *What-Where-When* Attention Network (W3AN) method to learn spatial-

temporal information for pedestrian video sequences, and designs a comprehensively unified attention model to explore identical information from occluded frames in re-id. Compared to traditional attention model, this paper especially focuses on exploring spatial attention features both from global appearance (*What*) and attributes (*Where*), and learning useful information from occluded frames (*When*), which are always abandoned in traditional attention models.

(2) We design a novel spatial attention module drawing attention on the pedestrian identity appearance and obvious attributes by sharing attention parameters with class activation map, which can preserve the crucial information in each video frame. Compared to recent popular transformer works, the spatial attention mechanism proposed in this paper can explore both of identity and attribute information with comprehensive attentive ability on pedestrian appearance.

(3) We propose a graph attention network to learn the frame-level importance by a temporal attention mechanism. Then the learned attention parameters are integrated into the adjacent matrix of spatially attentive frame features. The Graph Convolutional Network (GCN) are executed at the attentive matrix to learn the final video representation by global-average pooling.

(4) We validate the W3AN model by a series of compared experiments on popular datasets, and discuss the different contributions of major modules. The results elaborate its performable application in video-based person re-identification.

2. Related work

The correlated researches of this paper are divided as two groups, including image-based person re-identification, and video-based person re-identification. This section will discuss the correlations and comparisons between our research and these methods in related fields.

2.1. Image-based Person Re-identification

In recent years, person re-identification becomes a popular research topic with great challenges, which is caused by the frequently-occurred variations of viewpoints, poses, illuminations, and occlusions in realistic application. Foregoing researches mainly solve this problem by two ideologies of robust feature extraction [12, 13], and distance metric learning [14, 15]. With the development of Convolutional Neural Network (CNN), the performance of newly proposed CNN based person re-id models [1, 2, 3] has been achieved clear improvements on large scale pedestrian image data. For example, Bai et al [2] applied long short-term memory in an end-to-end way to model the pedestrian, seen as a sequence of body parts from head to foot, and it performs better results than global features; Besides, Wang et al [3] proposed a convolutional deformable part model by decoupling the complex part alignment procedure to extract robust pedestrian features representations.

Though these CNN methods have shown well-behaved performance, they only exploit spatial information from still images including global and local regions. As for local feature extraction, several attention models [1, 16] are designed besides simple part division strategy [3, 17]. Specifically, Xu et al [1] introduced an attention-aware compositional network for person re-id, consisting of pose-guided part attention, and attention-ware feature composition, which achieved state-of-the-art results on several public datasets. Tay et al [16] leveraged body parts and integrated the key attribute information in a unified framework, containing global person re-id task, part detection task, and crucial attribute detection task to draw attention on the pedestrian attributes. Although these methods have achieved improvements, they neglect the temporal information compared to video-based person re-identification, and they only conduct *where* to pay attention with susceptible sensitivity, rather than the unified *what-where-when* attention.

2.2. Video-based Person Re-identification

In real practice, image-based person re-id is easily developed into multi-shot re-id by given pedestrian video sequences. Thus, many recent works gradually exploit video-based person re-id problem. Specially, the works [7, 8] employ optical flow mechanism to learn motion information from pedestrian adjacent frames, which provides temporal correlations in video sequence. In detail, Rahman et al [7] proposed a temporal attention approach for aggregating frame-level features by adding the optical flow into the network; Gong et al [8] et al designed a flow-guided feature enhancement network that leverages flow information to enhance low-level features to achieve superior performance and outperform most of existing methods. Besides, the works [18, 19] conducted long short-term memory (LSTM) or recurrent neural network (RNN) to learn the temporal features from sequential CNN features, and they usually utilized average pooling strategy to investigate the temporal features. More concretely, Xu et al [18] presented a joint spatial and temporal attention pooling network enabling the feature extractor to be aware of the current input video sequences, which are fed into a sequence of RNN units to learn temporal information; Chen et al [19] employed competitive snippet-similarity aggregation and co-attentive snippet embedding to divide long pedestrian sequences into multiple short video snippets and aggregated the top-ranked snippet similarities for sequence-similarity estimation. As the popular technologies, several researches have leveraged spatial-temporal attention and graph convolutional network into video-based person re-id [20, 21, 22, 23, 24]. Specifically, Chen et al [20] proposed a spatial-temporal attention-aware learning model to attend the salient parts of persons in videos jointly in both spatial and temporal domains; Li et al [21] designed a relation-guided spatial attention to explore the discriminative regions globally and proposed a relation-guided temporal refinement module to further refine the feature representations across frames; Fu et al [22] integrated a spatial-temporal attention approach, adopting a more effective way for producing robust clip-level feature representation; Yang et al [24] developed a spatial-temporal graph convolutional network to identify the visually similar negative samples and utilize structural information from pedestrian video frames. Liu et al [23] proposed a non-local video attention network to incorporate video characteristics into the representation at multiple feature levels, and further introduced a spatially and temporally efficient non-local video attention network to reduce the computation complexity, but this approach only exploited multi-level features, without further considering the attribute and identity information from the spatial-temporal feature representations for video-based person re-id task.

Though these mentioned methods focus on learning spatial-temporal feature representation, or attend region of interest for person videos, they neglect the informative attribute information (*where* to draw attention) and treat the occluded frames equally or directly discard a few less important frames, which has much spatial-temporal information in the un-occluded regions. In contrary, our proposed W3AN model utilizes all frames weighted by the learned importances without neglecting any frame. At the meanwhile, our attention mechanism also focuses on *what* and *where* to draw attention. Detailed illustration can be seen in Section III.

3. Proposed W3AN Model

3.1. Overview

Our *What-Where-When* Attention Network (W3AN) can pay more attention on global identity appearance (*what*), obvious pedestrian attributes (*where*), and important un-occluded frames (*when*) by a multi-task network to learn spatial-temporal *What-Where-When* (W3) attention features. This paper designs a Spatial Attention Network (SAN) to address *what* and *where*, and a Temporal Attention Network (TAN) to tackle *when*. In detail, SAN generates Identity Activation Map (IAM) and Attribute Activation Map (AAM) by the guiding of Class Activation Map (CAM) [26] from two individual tasks, which includes identity and attribute classification tasks. In contrary, TAN achieves frame-level importance estimation by a fully convolutional attention module. Next,

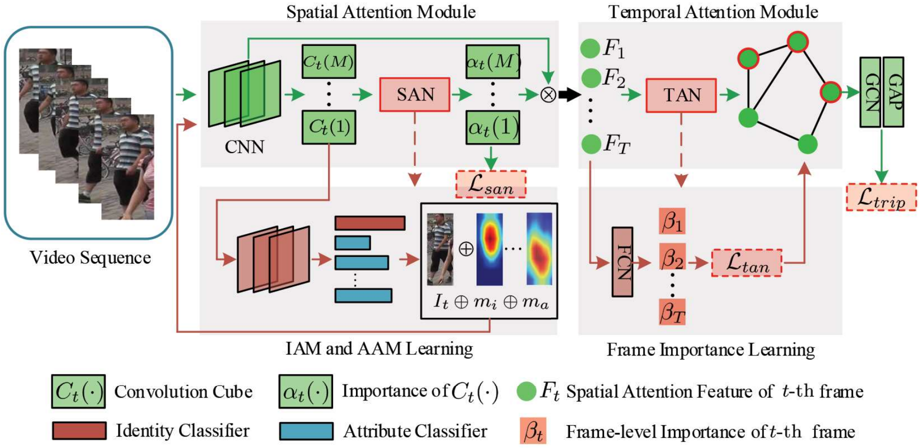


Figure 2: Overview of the proposed W3AN. The model consists of spatial and temporal attention modules, following a graph convolutional network. Firstly, pedestrian video frames are fed into a CNN backbone network to extract convolutional cube, with its element-wise importance by the supervision of identity and attribute activation map learning. Then, the spatial attention features extracted from CNN constitute a KNN-based graph convolutional neural network with their attention weights, learned by temporal attention module, to output final feature representations for pedestrian videos. Finally, a triple loss function optimizes the trainable parameters to guarantee the discriminant capability of pedestrian attention features.

the learned *what-where* attention features and the frame-level importance scores are fed into a Graph Attention Network (GAN) to calculate the *What-Where-When* attention features for each pedestrian video. Finally, we utilize Global-Average Pooling (GAP) strategy to generate the video representation, following a triplet loss to learn distance metrics among different video sequences. The overall illustration of our W3AN model is shown in Figure 2.

In our W3AN network, we employ ResNet [27] framework and remove its last fully connected layer as the CNN backbone network of the spatial attention module (Figure 2), due to its widely-acknowledged effectiveness in pedestrian feature extraction. The SAN subnet consists of a pedestrian identity classifier, and several attribute classifiers to generate Class Activation Map (CAM), and it provides a fully convolutional layer to estimate the importances of each convolutional cube. In addition, temporal attention module introduces Graph

Convolutional Network (GCN) to learn the temporal attention features based on the learned frame-level spatial features, and the frame-attention scores.

3.2. Spatial Attention Module

We denote a pedestrian video as $V = \{I_1, \cdots, I_t, \cdots, I_T\}$, where I_t is t-th frame in video V. The convolutional cube of I_t through CNN is $C_t \in \mathbb{R}^{H\times W\times M}$, where M denotes that it is constituted by M feature maps with $H \times W$ shape. We define the elements in the convolution cube by

$$C_t = \left\{ \begin{array}{ccc} C_t(1) & \cdots & C_t(W) \\ \vdots & \cdots & \vdots \\ C_t(W \times (H-1)+1) & \cdots & C_t(W \times H) \end{array} \right\}, \tag{1}$$

For each element in C_t, $C_t(k) \in \mathbb{R}^M$ is the feature vector at the k-th position, and $k \in \mathbb{R}^+$. This convolution cube is the base of generating spatial attention score in Spatial Attention Network (SAN).

In each pedestrian video frame, the most discriminant information is contained in the whole human body with global identity appearance (*what*), and the pedestrian attributes with different weights (*where*). Inspired by this point, we propose a spatial attention mechanism to estimate the attention score of *what* and *where*, providing the importances of each location in convolutional cube C_t, which is achieved by a spatial attention layer of fully convolution with trainable parameters

$$\hat{\alpha}_t(k) = v_t \tanh(W_s C_t(k) + b_s), \tag{2}$$

where $\hat{\alpha}_t(k) \in \mathbb{R}^M$ represents the unnormalized spatial attention score for each elements of C_t, and the attention parameter $v_t \in \mathbb{R}^{M\times M}$, $W_s \in \mathbb{R}^{M\times M}$ (weight matrix) and $b_s \in \mathbb{R}^M$ (bias) denote the attention parameters to conduct the convolution on $C_t(k)$, where W_s and b_s are shared for all frames, and v_t is distinct for each frame. The operation between W_s and $C_t(k)$ is linear transformation. The shared parameters W_s and b_s ensure that activation map of attention score $\hat{\alpha}_t$ can maintain the discriminative importance both on global identity appearance and the pedestrian attributes, and distinct v_t preserves the individual information for each specific video frame.

Then, the spatial attention score is normalized to the final score $\alpha_t(k)$ to ensure each element in [0,1], according to

$$\alpha_t(k) = \frac{\exp\{\hat{\alpha_t}(k)\}}{\sum_{j=1}^{W\times H}\exp\{\hat{\alpha_t}(j)\}}, \tag{3}$$

where $\alpha_t(k) \in \mathbb{R}^M$ and $\hat{\alpha_t}(k) \in \mathbb{R}^M$ denote the k-th feature vector of α_t and $\hat{\alpha_t}$, respectively. Thus, the spatial attention score α_t can represent the importances of each spatial feature vector from C_t, and we can calculate the spatial-attention feature for t-th pedestrian frame by

$$F_t = \sum_k \alpha_t(k) \odot C_t(k), \tag{4}$$

where $\odot$ means element-wise product and $F_t \in \mathbb{R}^M$.

This combination of spatial attention score and the convolution cube emphasizes the important regions estimated by the spatial attention layer. Here, the most important issue in SAN module is how to train the spatial attention layer to output a reasonable attention score α_t. Essentially, α_t is the coefficients that can reflect the discriminative information of identity and attributes in the pedestrian video. As we all know, the class activation map [28] expresses a weighted linear sum of the presence of visual patterns at different spatial locations, which can help us identify the image regions most relevant to the particular category, such as identity or attributes in pedestrian video. Therefore, to guarantee the effectiveness of this attention layer, this paper designs a spatial attention network loss function jointly trained with a multi-task learning strategy according to identity and attribute classification. So that it can generate the identity and attribute activation maps (IAM m_i and AAM m_a) by Class Activation Map (CAM)[28].

In our network, for a given identity or attribute class c, if we use S_c denotes the input of the softmax layer, then the output of the softmax can be computed by $\frac{\exp(S_c)}{\sum \exp(S_c)}$. The formula of S_c is shown in Eq.5.

$$S_c = \sum_k w_k^c \sum_{x,y} f_k(x, y) = \sum_{x,y} \sum_k w_k^c f_k(x, y), \tag{5}$$

where $f_k(x, y)$ represents the activation values of unit k in the last convolutional layer at spatial location (x, y), w_k^c indicates the importance of f_k for class c. If we use back-propagation to get the value of w when network training is done, the CAM of identity/attribute class c is denoted by IAM(m_i)/AAM(m_a), which is computed by Eq. 6.

$$m_i/m_a = \sum_k w_k^c f_k(x, y) \tag{6}$$

The purpose of using CAM is that it has excellent capability of explaining meaningful regions in original input images, which make our model can easily pay attention to important pixels with identity and attribute information. In detail, the IAM and AAM are adding together to the original frame

$$I_t' = I_t \oplus m_i \oplus m_a, \tag{7}$$

where I_t' is the baseline image with importances of IAM (m_i) and AAM (m_a), obtained according to [26]. Specifically, AAM m_a is calculated by multi-label task and this paper fuses CAMs from each attribute task by directly adding operator $\oplus$. Here, we can obtain spatial attention score α_t' when we feed I_t' into the same spatial attention network. Therefore, we utilize this baseline attention score to train SAN module by

$$\mathcal{L}_{san} = \sum_{t=1}^{T} \sum_{k=1}^{W \times H} \|\alpha_t(k) - \alpha_t'(k)\|^2. \tag{8}$$

This loss function can achieve that the spatial attention score is similar to the IAM and AAM, so that makes our network focuses on global identity appearance and attributes. Note that, the class activation map has been adopted into convolutional neural network for feature representation [29, 30, 31] similar to Eq.8, because the pixel-level supervision from IAM and AAM expresses abundant appearance and attribute information with attentive importance. In

Table 1: The employed pedestrian attribute groups in our model.

Group	Attributes
Gender & Age	Female, Ageless16, $\cdots$, Age31-45
Head-Shoulder	Hat, Glasses, $\cdots$, Black Hair
Up-Body	Shirt, Suit Up, $\cdots$, up-Blue
Low-Body	Dress, Skirt, $\cdots$, low-Black
Shoes	Sport, Leather, $\cdots$, shoes-White
Attach	Backpack, Hand Bag, $\cdots$, Plastic Bag

addition, we also introduce the multi-task training strategy for identity and attribute classification losses following the form of

$$\mathcal{L}_{cam} = \sum_{i=1}^{N} -\lambda_i \cdot y_i \log(\mathrm{softmax}(FC_i(C'_t))), \tag{9}$$

where $C'_t \in \mathbb{R}^{(H \cdot W \cdot M)}$ is the feature vector reshaped from convolutional cube $C_t \in \mathbb{R}^{H \times W \times M}$, i denotes i-th classification task, N represents the number of pedestrian identity and attribute classification tasks, y_i is the label of i-th task, and FC_i denotes the i-th fully connected layer to transform C'_t into the dimension of class number in i-th task. Here, each fully connected layer FC_i is specifically designed for each attribute and identity classification task, which can exploit more information than utilizing a shared FC layer. To demonstrate the employed attributes, we summarize the employed attribute groups in Table 1, which is totally 68 specific attributes following [6], with $N = 69$ in Eq 9.

3.3. Temporal Attention Module

In spatial attention module, the *what* and *where* to pay attention has been solved by the proposed spatial attention network. Nevertheless, the occlusion in pedestrian video is frequently-occurred, which appears in random video frames, and contains considerable discriminant identity information except for the occluded regions. Hence, directly removing these frames is ridiculous, and this paper designs a Temporal Attention Network (TAN) to solve *when* to pay attention in video-based person re-id task. TAN can estimate the frame's importance adding into the video feature learning to eliminate the influence of occluded frames rather than directly discarding them.

As for the given pedestrian video $V = \{I_1, \cdots, I_t, \cdots, I_T\}$, we have obtained its frame-level spatial attention feature collection $F = \{F_1, \cdots, F_t, \cdots, F_T\}$ by the proposed SAN module. As shown in Figure 2, the temporal attention module illustrates that we employ a fully connected network to roughly estimate the frame score for each frame feature (frame importance learning in Figure 2), and propose a TAN loss to train the performable parameters by modeling the frame relations.

Mathematically, the calculation of t-th frame importance can be represented by

$$\beta_t = \text{sigmoid}(W_f F_t + b_f), \tag{10}$$

where $\beta_t \in [0, 1]$ is the estimated temporal attention score for t-th frame, W_f and b_f are the trainable parameters in the fully connected layer, and the sigmoid function is the adopted activation function after fully connected layer.

To make TAN estimate the temporal attention score β_t more accurate, we propose a TAN loss to enforce that the minimum attention score β_{min} should be smaller than the overall one $\beta_{overall}$, which is calculated by feeding the average spatial attention feature into Eq.10 ($\beta_{overall} = \text{sigmoid}(W_f \frac{1}{T} \sum_{t=1}^{T} F_t)$). Formally, it can be represented by

$$\mathcal{L}_{tan} = \max\{0, m_1 - (\beta_{overall} - \beta_{min})\}, \tag{11}$$

where m_1 denotes the margin parameter. This TAN loss constrains the learned parameters $\beta = \{\beta_1, \cdots, \beta_t, \cdots, \beta_T\}$ more realistic, and provides accurate self-attention score for the spatial-temporal feature learning in next subsection.

3.4. Graph Attention Network

In spatial and temporal attention modules, we can learn the frame-level spatial attention features $F = \{F_1, \cdots, F_t, \cdots, F_T\}$ and its related frame attention score $\beta = \{\beta_1, \cdots, \beta_t, \cdots, \beta_T\}$ for the given pedestrian video sequence V. The features and frame attention scores contain the most informative characteristics of spatial and temporal representations, whereas they are expressing sufficient knowledge from individual aspects. As concluded, these two intermediate modules mine *what*, *where*, and *when* to individually pay attention in person re-id without further combination. The remain major step in W3AN approach is to integrate them into a spatial-temporal attention feature. Inspired by the outstanding capability of automatically correlation modeling of Graph Convolutional Network (GCN)[32], which has been widely employed to mine spatio-temporal correlations in various computer vision tasks [33, 34, 35, 24], we propose to develop GCN with temporal attention mechanism on modeling relations among different pedestrian video frames. For each identity, the consecutive temporal information is expressed among video frames, and there often contains occlusion and pose variations. Thus, constructing graph for frame-level features to achieve GCN can effectively model the spatial-temporal information for pedestrian video. As for the graph construction, we treat each spatial attention feature as nodes, and connect each other according to K nearest neighbor algorithm, as illustrated in Figure 3, added by the calculated temporal attention scores to balance the different weights of each frame in GCN. Finally, W3AN utilizes a Global-Average Pooling (GAP) layer on the learned graph attention features to compute the final spatial-temporal attention feature vector for the given pedestrian video.

Firstly, we utilize KNN algorithm to compute the connected nodes for each frame-level spatial attention feature F_t, and generate the adjacent matrix $A \in \mathbb{R}^{T \times T}$ for the learned graph , where A_{ij} denotes relationship between i-th and j-th nodes. Specifically, $A_{ij} = 1$ denotes nodes i and j are connected when they belong to the K-nearest neighbors for each other, or $A_{ij} = 0$ when they belong to other relations. Secondly, we transform the temporal attention scores as the

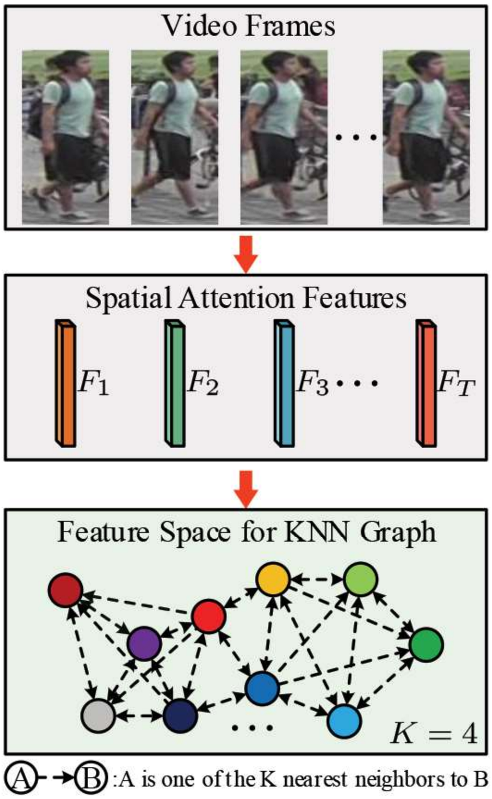

Video Frames
Spatial Attention Features
F_1
F_2
F_3 · · ·
F_T
Feature Space for KNN Graph
$K = 4$
A → B :A is one of the K nearest neighbors to B

Figure 3: Illustration of KNN graph building for spatial attention features from each pedestrian video.

temporal attention matrix A_t by

$$A_t = \left\{ \begin{array}{cccc} \beta_1 & 0 & \cdots & 0 \\ 0 & \beta_2 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & \cdots & \cdots & \beta_T \end{array} \right\}, \tag{12}$$

which is integrated into the overall adjacent matrix $\tilde{A}$ by

$$\tilde{A} = A + A_t, \tag{13}$$

Thus, W3AN constructs a graph representation $G(F, \tilde{A})$, with our graph convolutional network containing one input layer and a number of hidden layers.

Given the input $F^{(0)} = F$ and the temporal attention graph $\tilde{A}$, GCN conducts the graph convolution in hidden layers by

$$F^{(k+1)} = \sigma(D^{-1/2}\tilde{A}D^{-1/2}F^{(k)}W_g^{(k+1)}), \tag{14}$$

where $F^{(k+1)}$ is the output graph representation set of $k+1$ GCN layer, $k = 0, 1, \cdots, K-1$, D $(D_{ii} = \sum_{j=1}^{T} A_{ij})$ is the diagonal degree matrix of $\tilde{A}$, σ represents the nonlinear activation function. At the last GCN layer, the graph representation for each frame is denoted by $F^g = \{F_1^g, \cdots, F_t^g, \cdots, F_T^g\}$. Different with existing temporal attention mechanism in video-based person reidentification [7] that they fused the temporal attention score into the weighted average pooling on the learned frame-level features, our W3AN integrates the scores into GCN to further improve the involvement of the temporal attention.

Finally, we compute the spatial-temporal attention feature for each video sequence by

$$F^v = \frac{1}{T}\sum_{t=1}^{T} F_t^g, \tag{15}$$

where F^v is the spatial-temporal attention feature representation as well as *What-Where-When* (W3) attention feature for pedestrian video V.

Moreover, we utilize the batch-hard triplet loss [36] to further boost the person re-id performance on the learned W3 attention features. Formally, there are N videos in each training batch, containing P identities, and the loss is described by

$$\begin{aligned}\mathcal{L}_{triplet} = \frac{1}{P}\sum_{i=1}^{P}[m_2 + \max \|F_a^v(i) - F_p^v(i)\|_2 \\ - \min_{j \neq i} \|F_a^v(i) - F_n^v(j)\|_2]_+,\end{aligned} \tag{16}$$

where $F_a^v(i)$ and $F_p^v(i)$ are positive spatial-temporal attention video representation pair of the i-th identity, $F_n^v(j)$ is of j-th pedestrian identity which is negative to $F_a^v(i)$, and m_2 denotes the margin to constrain the distance across positive and negative pairs.

By integrating the losses proposed in this paper together, the overall end-to-end loss function can be defined as

$$\mathcal{L} = \gamma_1 \mathcal{L}_{san} + \gamma_2 \mathcal{L}_{cam} + \gamma_3 \mathcal{L}_{tan} + \gamma_4 \mathcal{L}_{triplet}, \tag{17}$$

where γ_1, γ_2, γ_3 and γ_4 are balance parameters to leverage the trainable terms of SAN, CAM, TAN and triplet identification, and the detail optimization is summarized in Algorithm 1.

Algorithm 1 *What-Where-When* Attention Network (W3AN)

```
Input: Pedestrian Video $V$ and its identity/attribute labels $y_0/y_i|_{i=1}^{68}$, each frame is resized as $224 \times 224$, and parameters $m_1$, $m_2$, and $\lambda_i|_{i=0}^{N}$; Each video contains consecutive frames and the size of mini-batch is 16; The backbone network parameters are initialized by a pre-trained model from ImageNet.
Output: The optimized network parameters.
for $n = 1$ to $N_t$ (number of iterations) do:
    for $t = 1$ to $T$ (frames of video) do:
        Learn convolutional cube $C_t$ for each frame;
        Estimate spatial attention score $\alpha_t$ for each frame;
        Calculate the frame-level feature $F_t$ by Eq. 4.
    end for
    Estimate the importance for each frame;
    Establish temporal-attention graph correlations $\tilde{A}$ for each video;
    Conduct graph convolution on the graph according to Eq.14;
    Compute the spatial-temporal attention feature $F^v$ by Eq.15;
    Optimizing the network parameters by minimizing Eq 17.
end for
```

4. Experimental Results

In this section, we elaborate the extensive person re-identification experiments to show the superiority of our proposed W3AN. To train W3AN model more efficiently, the experiments adopt MARS [37], iLIDS-VID [38], and PRID2011 [39] datasets (Figure 4), which have been widely evaluated in video-based person re-identification, to provide training and testing pedestrian videos. Moreover, we discuss the major attention modules to validate the combined influence of overall *what-where-when* attention mechanism.

4.1. Datasets

The **MARS** dataset is a recently large scale person re-id dataset, containing 1261 pedestrian identities with 20715 video sequences from six cameras. The bounding boxes are produced by DPM detector [40] and GMMCP tracker [41]. This dataset meets significant challenges because of the poor quality from the failure of detection or tracking, while this dataset is close to realistic person re-id task.

The **iLIDS-VID** dataset contains 600 image sequences from 300 pedestrian identities captured from two cameras, which the number of video frames is ranged from 23 to 192 with average of 73. Different from Mars, the bounding boxes are manually annotated and occlusion appears frequently.

The **PRID2011** dataset consists of shared 200 pedestrian identities from two cameras and another 734 identities only appears in one camera. We utilize these shared identities with 5 to 675 video frames. Their bounding boxes are also annotated by human power.

It should be noted that, these three datasets do not contain attribute labels while our W3AN network requires attribute labels to generate class activation map with supporting spatial attention module. To solve this problem, we employ the transfer learning strategy for attribute recognition following [6]. It employs maximum mean discrepancy (MMD) [42] to measure their distance to an attribute-labeled dataset RAP [43], and then transfers the same attribute knowledge into the datasets employed in this paper, by a weighted binary cross-entropy loss. Detail transfer learning can be find in [6].

The **RAP** dataset is a large-scale pedestrian attribute dataset and it provides 91 fine-grained binary attributes for each image. This paper also selects 68 specific attributes (e.g. black hair, T-shirt) and neglects others (e.g. talking, face right) following [6].

(a) MARS (b) iLIDS-VID (c) PRID2011

Figure 4: Pedestrian video sequences from MARS, iLID-VID, and PRID2011 datasets.

4.2. Settings and Protocols

Settings. In our *what-where-when* attention network, we adopt ResNet-18 [27] pre-trained on the ImageNet [44], and remove the fully connected layer in W3AN. Each video frames are reshaped to 144×288, randomly cropped to 128×256 and resize them into 224×224 before feeding into the network. For the network training, this paper employs SGD optimizer to train the parameters and set the video batch size to $N = 16$ from $P = 8$ identities. The learning rate is set by 2e-2 which multiply 0.1 in every 50 epochs and will be decayed to 0 in the last 10 epochs. Besides, the maximum epoch is 300 and the length of video sequence is $T = 10$, randomly sampled from image sequences. Specifically, we implement W3AN by 4 GPUs of NVIDIA Geforce 2080Ti on Ubuntu 16.04 system, and firstly predict the attribute labels for the experimental datasets to support the overall W3AN model. As for the balance parameters in loss function 17, we directly set $\lambda_i = 1|_{i=0}^{69}$ of Eq.9 and the margin parameters $m_1 = 0.08$, $m_2 = 0.55$. Besides, for the balance parameters in Eq.17, the final values of $[\gamma_j|_{j=1}^{4}]$ in our experiments are [0.3,0.15,0.2,0.35]. As for the parameter settings of GCN, we set $K = 4$ in KNN algorithm to connect 4 neighbors for each pedestrian video, and the number of GCN layers in W3AN is 2.

Protocols. This paper utilizes the standard measurements to evaluate the performance of our W3AN model on video-based person re-identification, including Cumulative Matching Characteristic (CMC) accuracies (Rank-m) and mean Average Precision (mAP). Each datasets are divided as training and testing data equally, and the experiments are repeated 10 times with randomly training/testing division to calculate average results for each evaluation metric.

4.3. Compared with recent works

In our experiments, we report the rank-m accuracies of W3AN compared with several recent works, and the results are summarized in Table 2. CNN+XQDA [37] combines convolutional neural network and XQDA features to extract space-time descriptors for pedestrian video. SeeForest [45] automatically chooses the most discriminative frames from pedestrian videos by a temporal attention model, and it integrates the surrounding information at each location by a spatial recurrent model when measuring the similarity with another pedestrian video. Snippet [46] divides long pedestrian video sequences into multiple short video snippets and aggregates the top-ranked snippet similarities for sequence-similarity estimation, which is achieved by temporal co-attention for snippet embedding. RQEN [47] is a region based quality estimation network in which the ingenious training mechanism enables the effective learning to extract the complementary region-based information between different frames. Attribute [48] is an attribute-driven method for feature disentangling and frame re-weighting to enhance the most informative regions of each frame and contribute to a more discriminative sequence representation. STMP [49] proposes a refining recurrent unit that recovers the missing parts and suppresses noisy parts of the current frame's features by referring historical frames, and a spatial-temporal clues integration module to mine the spatial-temporal information from those upgraded features. STPN [50] utilizes a weighted triple-sequence loss to optimize the video-based feature and reduce the impact of outliers, and it also designs a spatial transformed partial network to jointly learn image-level and video-level features to generate more robust representation. MSTA [51] designs a multi-scale spatial-temporal attention model to measure the regions of each frame in different scales from the perspective of whole video sequence, which focuses on exploiting the importance of local regions to the whole video representation in both spatial and temporal domains. AMEM [52] proposes an appearance and motion enhancement model to enrich these two kinds of information contained in the backbone network in a more interpretable way. FGRA [53] integrates a frame-guided region-aligned model for discriminative representation learning in video-based person re-identification in an end-to-end manner. MGH [54] proposes a multi-granular hypergraph framework to pursue better representational abilities for pedestrian videos by modeling spatio-temporal dependencies in terms of multiple granularities.

Table 2: Comparative results on MARS, iLIDS-VID and PRID2011 datasets. Measured by rank-m accuracies and mAP (%).

Dataset	MARS				iLIDS-VID			PRID2011		
Models	mAP	r-1	r-5	r-20	r-1	r-5	r-20	r-1	r-5	r-20
CNN+XQDA (ECCV16) [37]	49.3	68.3	82.6	89.4	53.0	81.4	95.1	77.3	93.5	99.3
SeeForest (CVPR17) [45]	50.7	70.6	90.0	97.6	55.2	86.5	97.0	79.4	94.4	99.0
Snippet (CVPR18)[46]	76.1	86.3	94.7	98.2	85.4	96.7	99.5	93.0	99.3	**100**
RQEN (AAAI18) [47]	51.7	73.7	84.9	91.6	76.1	92.9	99.3	92.4	98.8	**100**
Attribute (CVPR19)[48]	78.2	87.0	95.4	98.7	86.3	97.4	**99.7**	93.9	99.5	**100**
STMP (AAAI19)[49]	72.7	84.4	93.2	96.3	84.3	96.8	99.5	92.7	98.8	99.8
STPN (NeuroC20) [50]	77.9	85.9	94.6	97.3	82.2	94.5	99.0	95.2	99.1	**100**
MSTA (TIP20) [51]	79.7	84.1	93.5	98.0	70.1	88.7	97.6	91.2	98.7	99.7
AMEM (AAAI20) [52]	79.3	86.7	94.0	97.1	87.2	97.7	99.5	93.3	98.7	**100**
FGRA (AAAI20) [53]	81.2	87.3	96.0	98.1	88.0	96.7	99.3	95.5	**100**	100
MGH (CVPR20) [54]	85.8	90.0	96.7	98.5	85.6	97.1	99.5	94.8	99.3	100
W3AN (Ours)	**86.3**	**91.1**	**96.8**	**98.9**	**89.2**	**98.1**	99.6	**95.8**	99.5	**100**

From Table 2, the comparison elaborates the superior performance of our W3AN, compared to the recent works. The best results in rank-m are in bold

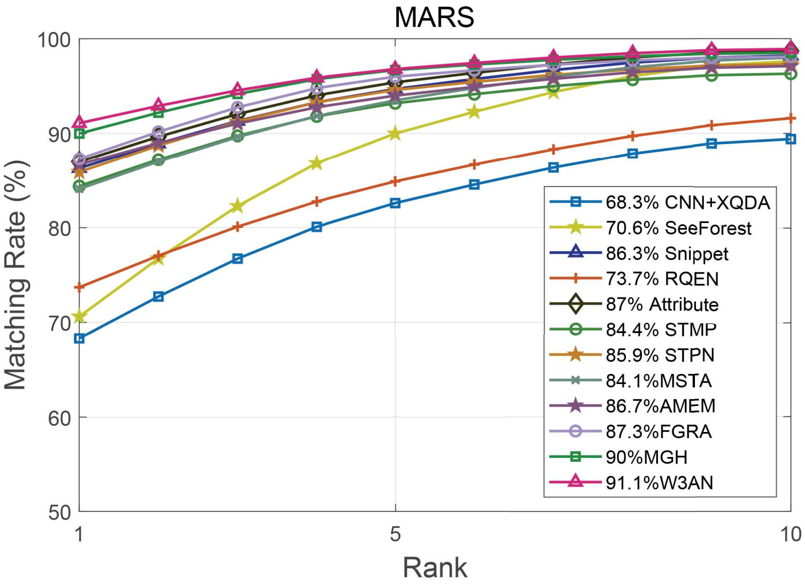


Figure 5: CMC curves on MARS dataset with comparison to baselines.

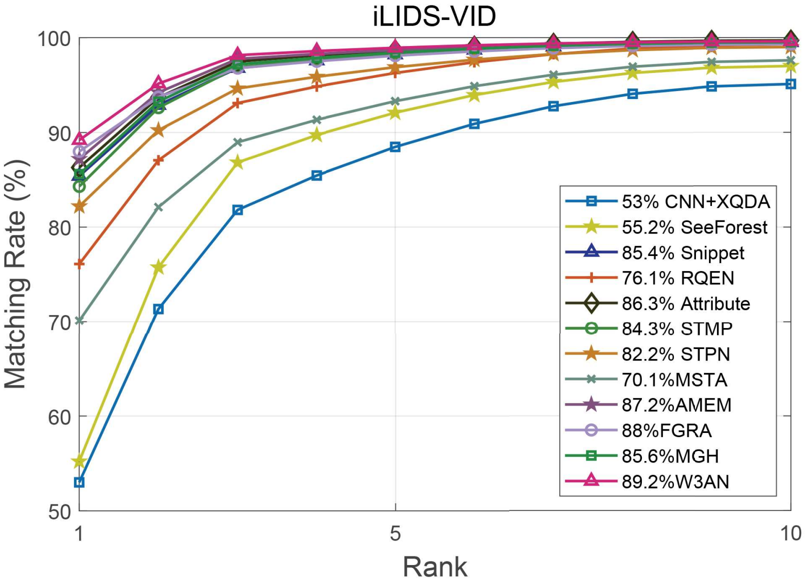


Figure 6: CMC curves on iLIDS-VID dataset with comparison to baselines.

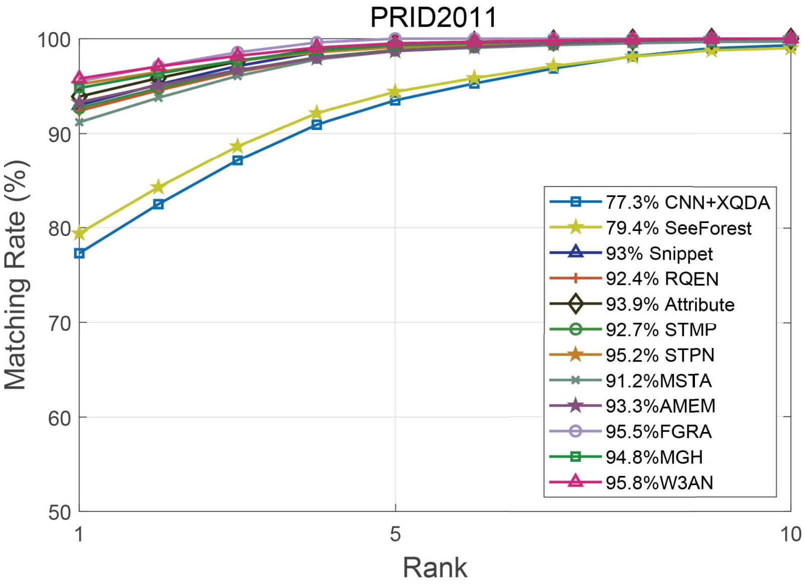


Figure 7: CMC curves on PRID2011 dataset with comparison to baselines.

and the second is in underline. For MARS dataset, W3AN obtains the best results in each rank-m accuracy and mAP (mAP 86.3% , rank-1 91.1%, rank-5 96.8%, and rank-20 98.9%). Moreover, our W3AN achieves similar performance on iLIDS-VID dataset, which realizes best rank-1, rank-5 accuracies of 89.2%, 98.1%, and second best rank-20 accuracy of 99.6%. As for PRID2011 dataset, the best rank-1 and rank-20 accuracies are 95.8% and 100% (achieved by our model), while rank-5 also performs well (second best of 99.5%). It can be easily observed that the average performance of our W3AN approach is superior to the chosen state-of-the-art methods. In overall, Our proposed W3AN achieves improvements ranged from 0.3% to 1.2% of rank-1 accuracy on three datasets, and 0.5% of mAP on MARS dataset. Besides, our model also achieves better performance on other evaluation metrics. To visualize the comparative performance, Figures 5, 6, and 7 draw the CMC curves on three datasets with comparison to baselines, which also proves the advantages of our method. The major reason of the W3AN's effectiveness is that the *what-where-when* attention mechanism exploits more discriminative information from pedestrian video sequences than other spatial-temporal feature representation models. Particularly, our W3AN model contains three principle components: spatial attention module to achieve *what-where* to attend, temporal attention module to pay attention on *when*, and a graph attention network to learn final spatial-temporal attention feature representations by the integrated graph convolutional network. The validations of these components are discussed in the Ablation Study.

4.4. Compared with spatial-temporal attention and GCN works

Moreover, we also compare W3AN with several spatial-temporal attention and GCN works mentioned in Section II (2.2), including STAL [20], RGSA [21], STA [22], NVAN [23], and STGCN [24]. These methods are proposed to exploit spatial-temporal information, combined with attention and GCN. The compared results are summarized in Table 3, where the best performance is achieved by our W3AN model. Therefore, this comparison demonstrates the excellent effectiveness and superiority of W3AN, compared to existing spatial-temporal attention and GCN based person re-identification methods. The main reason is that our W3AN model can address the problems of attending *where* (attributes), and learning useful information from occluded frames.

Table 3: Comparative results of spatial-temporal and GCN methods. Measured by rank-m accuracies and mAP (%).

Dataset	MARS				iLIDS-VID			PRID2011		
Models	mAP	r-1	r-5	r-20	r-1	r-5	r-20	r-1	r-5	r-20
STAL (TIP19) [20]	73.5	82.2	92.8	98.0	82.8	95.3	98.8	92.7	98.8	100
RGSA (AAAI20)[21])	84.0	89.4	96.9	98.3	86.0	98.0	99.4	93.7	99.0	100
STA (AAAI19) [22]	80.8	86.3	95.7	98.1	-	-	-	-	-	-
NVAN (BMVC19)[23]	82.8	90.0	-	-	-	-	-	-	-	-
STGCN (CVPR20)[24]	83.7	90.0	96.4	98.3	-	-	-	-	-	-
W3AN (Ours)	**86.3**	**91.1**	**96.8**	**98.9**	**89.2**	**98.1**	**99.6**	**95.8**	**99.5**	**100**

4.5. Compared with attribute-based re-id works

To fairly reveal the superiority of our proposed W3AN model, we also compare it with three advanced attribute-based video re-id models, including FDTA [55], TALNet [56], and AITL[57], on MARS dataset. Specifically, FDTA [55] proposed an attribute-driven method for feature disentangling and frame reweighting in video-based person re-id task; TALNet [56] designed a temporal attribute-appearance learning network for video-based person re-id, which can simultaneously exploits human attributes and appearance to learn comprehensive and effective pedestrian representations from videos; AITL [57] introduced a metric learning method for video-based person re-id, which is the attribute-aware identity-hard triplet loss to reduces the intra-class variations among positive samples via calculating attribute distance. The comparative results are reported in Table 4, and W3AN makes the best mAP and rank-m accuracies, while the second one is the AITL (with distances of 1.9% mAP and 2.9% rank-1 to W3AN). It is obvious that W3AN outperforms other attribute-based video re-id methods for pedestrian videos. The dominant reason is W3AN proposes novel modules of spatial attention and temporal attention to focus on *What-Where-When* to extract video representations, including attribute, identity and temporal discriminative clues in video-based person re-identification.

Table 4: Comparative results of advanced attribute-based methods on MARS dataset. Measured by rank-m accuracies and mAP (%).

Dataset	MARS			
Models	mAP	r-1	r-5	r-20
FDTA (CVPR19) [55]	78.2	87.0	95.4	98.7
TALNet (Arxiv)[56])	82.3	89.1	96.1	98.5
AITL (Arxiv)[57]	84.4	88.2	96.5	98.4
W3AN (Ours)	**86.3**	**91.1**	**96.8**	**98.9**

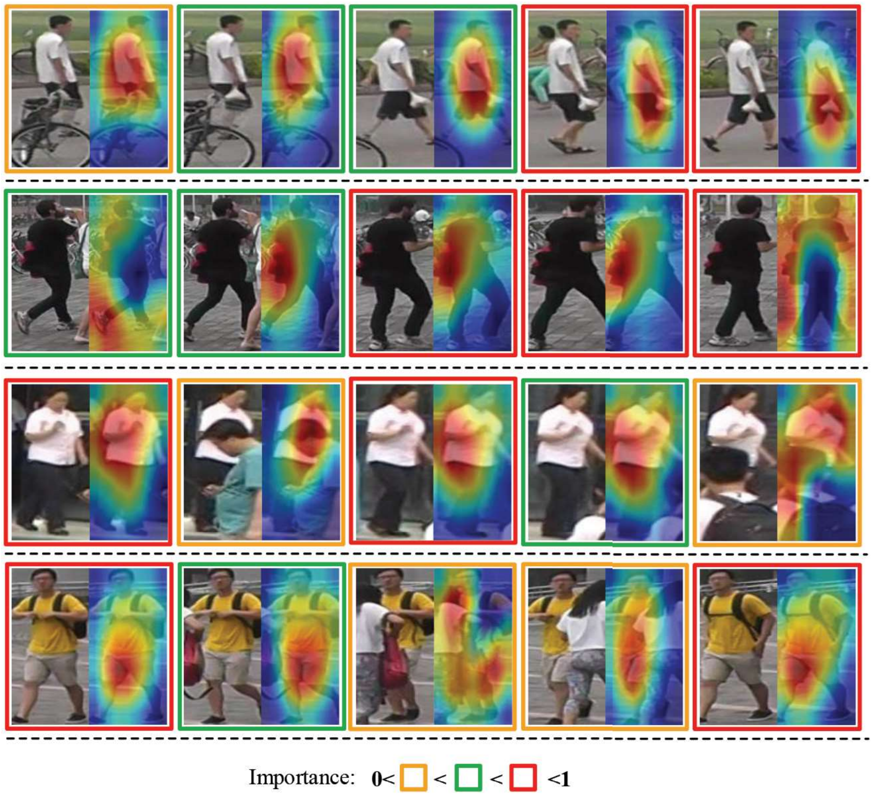


Figure 8: Attention Visualization for SAN and TAN modules. The class activation maps from identity classification are visualized in pedestrian video frames, and the frame-level importances are marked by lines around each frame. Note that, the yellow, green and red borders divide the importance as three average ranges.

4.6. Attention Visualization

To exhibit the visualization of *What-Where-When* attention, the CAM and frame-level importance of pedestrian video sequences are integrated for validating the attention mechanism. In Figure 8, the class activation maps for specific identity labels for each video frame are set in original pedestrian images, and frame-level importance from TAN is marked by border colors in three levels.

Four pedestrian video frames are employed, and each video contains five images for visualization. As shown in Figure 8, the CAM from SAN covers major pedestrian body and neglects occlusion, backgrounds or other unrelated regions, and our model marks the occluded frame with a lower temporal importance according to the occluded severity. Furthermore, the yellow border denotes lower spatial attention weight to our model by larger occluded body regions, green borders denotes a spot of occlusion and red border is non-occlusion. The visualization results in Figure 8 explain the efficiency of our designed *What-Where-When* attention mechanism more specifically on video-based person re-identification task.

Table 5: Ablation study on the major components of our W3AN on MARS, iLIDS-VID, and PRID2011 datasets (%).

Dataset	MARS				iLIDS-VID			PRID2011		
Models	mAP	r-1	r-5	r-20	r-1	r-5	r-20	r-1	r-5	r-20
WAN (non-spatial)	77.0	81.3	90.2	96.2	81.8	90.8	95.0	91.7	96.4	98.3
WAN (non-IAM&AAM)	75.4	79.8	87.3	91.0	78.8	86.2	90.2	89.8	93.2	96.4
W2AN (non-temporal)	79.8	85.6	88.9	93	84.7	94.2	96.7	92.3	96.0	97.3
W3WAN (non-GCN)	83.6	87.6	92.6	95.8	86.9	95.7	98.0	93.5	97.1	99.0
W3WAN (non-occluded frame)	84.8	89.5	94.3	96.6	88.0	96.5	98.8	94.9	98.4	99.2
AP-W3AN	85.5	90.3	93.2	95.0	87.6	94.8	96.2	93.0	95.2	97.0
AL-W3AN	84.8	88.6	93.9	95.5	86.5	93.9	95.7	91.8	93.7	96.0
W3AN(Complete)	**86.3**	**91.1**	**96.8**	**98.9**	**89.2**	**98.1**	**99.6**	**95.8**	**99.5**	**100**

4.7. Ablation Study

We evaluate the major modules of spatial attention, temporal attention, and graph attention network to demonstrate the contributions of our proposed

W3AN. Moreover, the parameter analysis is also conducted to prove the influence of the major modules.

Effectiveness of the Spatial Attention Module. We remove the spatial attention layer and SAN loss, which directly feed the CNN features into the TAN and GAN modules (named *When* Attention Network, WAN) to evaluate the spatial attention model. Table 5 shows the WAN achieves mAP with 77.0% and rank-1 accuracy of 81.3% on MARS dataset. Combining the performance on other datasets, the rank-1 accuracies of WAN are less than W3AN about 4.1% to 9.8%. In order to further discuss the impact of using IAM and AAM in Eq.7 as the guidance in SAN, we evaluate them by making I_t' close to I_t (achieved by adding a random sparse matrix ranged in [0,0.001] on I_t), named WAN (non-IAM&AAM) in Table 5. Compared to directly removing spatial attention layer and SAN loss, adding a random sparse matrix on I_t performs worse results on re-id datasets, since random matrix mislead s the training of spatial attention layer. This evaluation experiment demonstrates that adopting IAM and AMM is a reasonable solution and they provide much more helpful attentive information for spatial attention. Hence, the spatial attention module improves the effectiveness in a substantial distance because of solving the *what* and *where* to attention, that demonstrates the spatial attention module devotes a major contribution to the video-based person re-identification.

Effectiveness of the Temporal Attention Module. Similarly, we abandon the temporal attention module and directly employ the adjacent matrix A to conduct GCN without adding the temporal attention matrix A_t (*What-Where* Attention Network, W2AN). From Table 5, W2AN achieves rank-1 accuracies of 85.6%, 84.7%, and 92.3% on MARS, iLIDS-VID, and PRID2011 datasets separately, with a distance from 3.5% to 5.5% to W3AN. This comparison also proves the effectiveness of temporal attention module. Compared to the performance of WAN, it can be seen that the spatial attention module contributes more discriminative information, that illustrates the spatial appearance provides the major effectiveness and the temporal cues further develop the performance with a relative-less improvement.

Effectiveness of Graph Attention Network (GAN). Different from other frame-level attention models that calculate the weighted average frame feature with the frame attention score, our W3AN integrates the frame attention score into the adjacent matrix and implement GCN on them. To prove this novel feature representation method, we replace the graph attention network by computing the weighted average frame feature (What-Where-When Weighted Attention Network, W3WAN). The W3WAN only achieves 83.6% mAP and 87.6% rank-1 accuracy on MARS dataset as reported in Table 5, and the performance on other datasets also verifies the effectiveness of our graph attention network, because directly calculating weighted average features is an inflexible method without considering the correlations among consecutive video frames. Moreover, the primary issue addressed by graph attention network is to exploit useful information from occluded frames, which are always abandoned in previous methods. To prove the abandoned frames containing helpful features, we discard the half of video frames, which have less temporal attention scores than other frames, and then conduct graph attention network on the frames with higher scores. This experiment is defined by 'W3AN(non-occluded frames)' in Table 5, and it can be observed that abandoning occluded frames can reduce the rank-m accuracy and mAP results on all datasets. That further demonstrates the necessary of graph attention network on pedestrian video matching, and the occluded frames can not be discarded due to its discriminative information conveyed by un-occluded regions.

Impacts of Employed Attribute Groups. As summarized in Table 1, our W3AN model adopts six attribute groups to implement *Where* to draw attention. As we all know, different pedestrian attributes preserve various identity information and their impacts on video-based person re-identification can convey their contributions to our attention model. Hereby, we evaluate the influence of each attribute group by removing target group one by one, and report their rank-1 accuracy results of MARS dataset, as show in Figure 9. From the results, it is obvious that our W3AN achieves limited reduction when removing gender & age and shoes groups, and obtains prominent improvements by up-body, head-shoulder, and attach groups. The most contributed attribute group is up-body, because its appearance provides major identical information for person re-identification as attention from human beings. From this analysis, the contribution of attribute groups and their detailed influence for our attention mechanism are elaborated, demonstrating the reasonability and scalability of the proposed *What-Where-When* attention model.

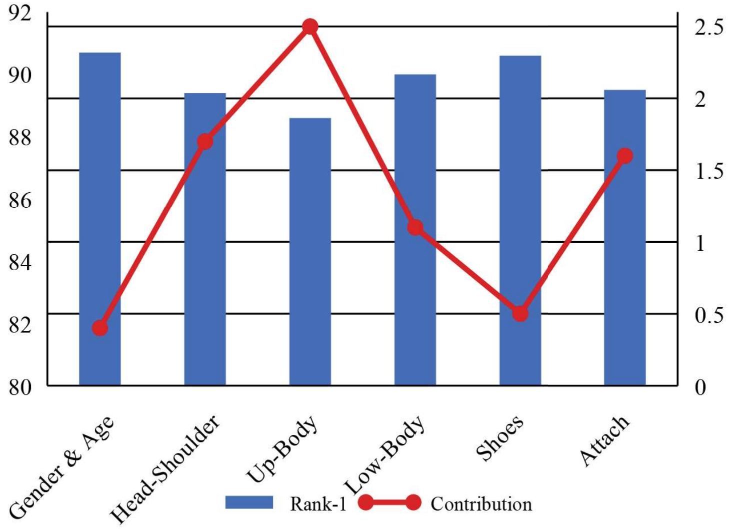
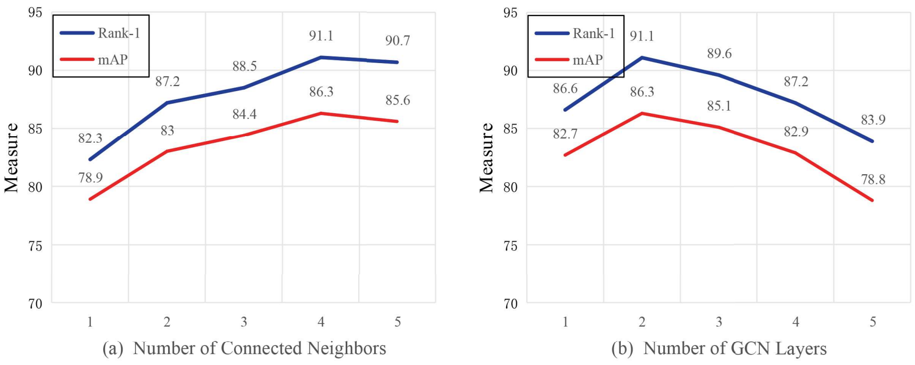


Figure 9: Evaluation for attribute groups on MARS dataset. Blue histograms is the rank-1 accuracy (%) when we removing each attribute group and the red points denotes contributed rank-1 improvements (%) from the removed groups, which is the reduction, compared to W3AN involving six attribute groups.

Evaluation for Alternative Spatial Attention models. As demonstrated above, the major contribution of this paper is to propose a novel framework about three points (*What-Where-When*) to pay attention for video-based person re-identification, which is achieved by modules of spatial attention, temporal attention, and graph attention network, involving identity, attribute and temporal clues. Here, the effectiveness of spatial attention is alternatively implemented by two another models (AP-CNN [58], and AL-network [59]) to comprehensively reveal the superiority of our proposed *What-Where-When* attention network. AP-CNN [58] proposes an attention pyramid convolutional neural network to integrate low-level information (e.g., color, edge junctions, texture patterns) into distilling high-level features to enhance the feature representation and accurately locate discriminative regions, which can improve the fine-grained image classification task. AL-network [59] is an attention long short-term memory network for fine-grained classification task, which can extract local features of category-sensitive regions by the long short-term memory unit. We re-implement the W3AN experiments by replacing spatial attention module by these two modules, keeping the same settings with original papers [59, 59], and report the results in Table 5 (AP+W3AN, and AL+W3AN). Compared with W3AN model, AP/AL-W3AN models perform relative lower mAPs and rank-m accuracies on MARS, iLIDS-VID, and PRID2011 datasets. Though these two methods can sufficiently locate discriminative regions by their proposed attention mechanism, the main reason of W3AN's superiority is that the attribute and identity information contributes precise discriminative clues compared to the self-learned local regions from AP-CNN [58], and AL-network[59].

Influence of different K values in GCN. As described in Section 3.4, K is a key factor in graph attention network to construct graph structure, which denotes how many neighbors are connected to each feature. To evaluate the influence of different K values, we conduct re-id experiments on MARS dataset by changing K from 1 to 5, and report the results in Figure 10 (a). It is obviously to observe that W3AN achieves the best performance when $K = 4$, and the re-id performance is improved with the increasing of K when it is lower than 4. The reason is that, the optimal neighbors can produce positive impact on GCN, and too many neighbors may bring much noisy information to cause negative influence.

Impacts of GCN layers. In the experiments of W3AN, we involve 2 GCN layers in graph attention network. The GCN layers is in charge of learning graph feature representations, which is a major cues for re-id. Here, we evaluate the impact of different number of GCN layers on MARS dataset, and the evaluation for different numbers of GCN layers (1,2,3,4,5) are visualized in Figure 10 (b), which reveals that two GCN layers have more excellent feature learning capability than other numbers in our proposed W3AN re-id model.



(a) Number of Connected Neighbors (b) Number of GCN Layers

Figure 10: Parameter analysis for the numbers of connected neighbors and GCN layers on MARS dataset.

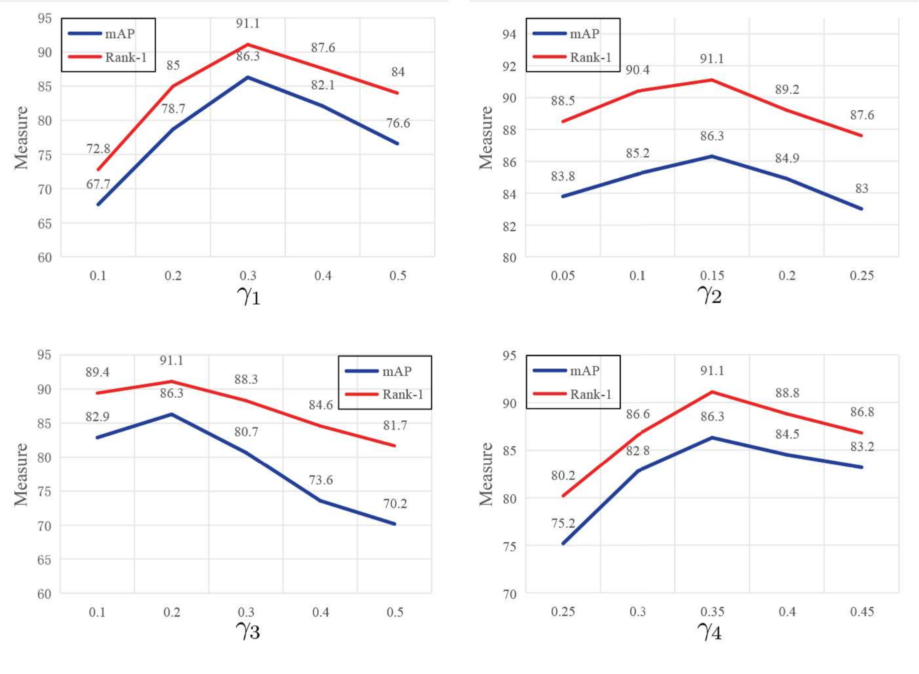


Figure 11: The balance of SAN and TAN modules on MARS dataset.

Parameter Analysis for Losses of SAN, CAM, TAN, and Triplet. To further evaluate the contributions of our proposed spatial and temporal attention modules, we implement the balance parameter analysis for them. We change the parameters γ_1 and γ_3 from 0.1 to 0.5, γ_2 in [0.05:0.05:0.25] and γ_4 in [0.25:0.05:0.45], and report the results in Figure 11. When we set a smaller parameter for both SAN and TAN, the performance reduction of SAN is more obvious than TAN, and it quickly reduces to the rank-1 accuracy and mAP when increasing the value of γ_3. This tendency demonstrates the importance of those two modules of SAN and TAN, showing the SAN contributes more information than TAN. On the other hand, SAN and TAN are both important for our W3AN since it generates a major improvement when we alleviate the balance parameter for them. Besides, CAM loss expresses a stable variation compared to triplet loss, that reveals the influences from adjusting CAM and triplet loss in different weights.

5. Conclusion

In this work, we propose a comprehensive attention framework for video-based person re-identification, namely *What-Where-When* Attention Network (W3AN), which can focus on the pedestrian identity appearance, obvious attributes and the important frames in pedestrian video representation learning. Specifically, the Spatial Attention Network (SAN) employs class activation map of identity and attributes to guide the attention layer estimating the spatial importance; Then a Temporal Attention Network (TAN) learns the frame-level importances, which is integrated into a Graph Attention Network to extract final *what-where-when* attention feature representations. Relatively speaking, the proposed W3AN approach covers the overall attention aspects in video-based person re-identification. Furthermore, the extensive experiments on three recognized datasets demonstrate the superiority of our approach, and the ablation study discusses different contributions of SAN, TAN, and GAN.

Acknowledgements

This work was supported in part by the National Natural Science Foundation of China (61801437, 61871351, 61971381); Natural Science Foundation of

Shanxi Province (201801D221206, 201801D221207); Scientific and Technological Innovation Programs of Higher Education Institutions in Shanxi (2020L0683); The National Natural Science Foundation of China under Grant 61461025, Grant 61871259, Grant 61811530325 (IECnNSFCn170396, Royal Society, U.K.), and Grant 61861024; The Key research and development plan of Luliang City (2020GXZDYF21).

References

[1] J. Xu, R. Zhao, F. Zhu, H. Wang, W. Ouyang, Attention-aware compositional network for person re-identification, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 2119–2128.

[2] X. Bai, M. Yang, T. Huang, Z. Dou, R. Yu, Y. Xu, Deep-person: Learning discriminative deep features for person re-identification, Pattern Recognit. 98. doi:10.1016/j.patcog.2019.107036.

[3] K. Wang, C. Ding, S. J. Maybank, D. Tao, CDPM: convolutional deformable part models for semantically aligned person re-identification, IEEE Trans. Image Process. 29 (2020) 3416–3428. doi:10.1109/TIP.2019.2959923.

[4] J. Li, S. Zhang, T. Huang, Multi-scale temporal cues learning for video person re-identification, IEEE Trans. Image Process. 29 (2020) 4461–4473. doi:10.1109/TIP.2020.2972108.

[5] R. Hou, B. Ma, H. Chang, X. Gu, S. Shan, X. Chen, VRSTC: occlusion-free video person re-identification, in: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019, 2019, pp. 7183–7192. doi:10.1109/CVPR.2019.00735.

[6] Y. Zhao, X. Shen, Z. Jin, H. Lu, X. Hua, Attribute-driven feature disentangling and temporal aggregation for video person re-identification, in: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019, 2019, pp. 4913–4922. doi:10.1109/CVPR.2019.00505.

[7] T. Rahman, M. Rochan, Y. Wang, Convolutional temporal attention model for video-based person re-identification, in: IEEE International Conference on Multimedia and Expo, ICME 2019, Shanghai, China, July 8-12, 2019, 2019, pp. 1102–1107. doi:10.1109/ICME.2019.00193.

[8] W. Gong, B. Yan, C. Lin, Flow-guided feature enhancement network for video-based person re-identification, Neurocomputing 383 (2020) 295–302. doi:10.1016/j.neucom.2019.11.050.

[9] W. Yang, Y. Yan, S. Chen, Adaptive deep metric embeddings for person re-identification under occlusions, Neurocomputing 340 (2019) 125–132. doi:10.1016/j.neucom.2019.02.042.

[10] Z. Zhou, Y. Huang, W. Wang, L. Wang, T. Tan, See the forest for the trees: Joint spatial and temporal recurrent neural networks for video-based person re-identification, in: 2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017, 2017, pp. 6776–6785. doi:10.1109/CVPR.2017.717.

[11] J. Perez-Rua, B. Martinez, X. Zhu, A. Toisoul, V. Escorcia, T. Xiang, Knowing what, where and when to look: Efficient video action modeling with attention, CoRR abs/2004.01278. arXiv:2004.01278.

[12] I. Kviatkovsky, A. Adam, E. Rivlin, Color invariants for person reidentification, IEEE Trans. Pattern Anal. Mach. Intell. 35 (7) (2013) 1622–1634. doi:10.1109/TPAMI.2012.246.

[13] R. Zhao, W. Ouyang, X. Wang, Learning mid-level filters for person re-identification, in: 2014 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2014, Columbus, OH, USA, June 23-28, 2014, 2014, pp. 144–151. doi:10.1109/CVPR.2014.26.

[14] S. Liao, S. Z. Li, Efficient PSD constrained asymmetric metric learning for person re-identification, in: 2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015, 2015, pp. 3685–3693. doi:10.1109/ICCV.2015.420.

[15] S. Paisitkriangkrai, C. Shen, A. van den Hengel, Learning to rank in person re-identification with metric ensembles, in: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015, 2015, pp. 1846–1855. doi:10.1109/CVPR.2015.7298794.

[16] C. Tay, S. Roy, K. Yap, Aanet: Attribute attention network for person re-identifications, in: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019, 2019, pp. 7134–7143. doi:10.1109/CVPR.2019.00730.

[17] Y. Sun, L. Zheng, Y. Yang, Q. Tian, S. Wang, Beyond part models: Person retrieval with refined part pooling (and A strong convolutional baseline), in: Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part IV, 2018, pp. 501–518. doi:10.1007/978-3-030-01225-0_30.

[18] S. Xu, Y. Cheng, K. Gu, Y. Yang, S. Chang, P. Zhou, Jointly attentive spatial-temporal pooling networks for video-based person re-identification, in: IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017, 2017, pp. 4743–4752. doi:10.1109/ICCV.2017.507.

[19] D. Chen, H. Li, T. Xiao, S. Yi, X. Wang, Video person re-identification with competitive snippet-similarity aggregation and co-attentive snippet embedding, in: 2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018, 2018, pp. 1169–1178. doi:10.1109/CVPR.2018.00128.

[20] G. Chen, J. Lu, M. Yang, J. Zhou, Spatial-temporal attention-aware learning for video-based person re-identification, IEEE Transactions on Image Processing 28 (9) (2019) 4192–4205.

[21] X. Li, W. Zhou, Y. Zhou, H. Li, Relation-guided spatial attention and temporal refinement for video-based person re-identification, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 34, 2020, pp. 11434–11441.

[22] Y. Fu, X. Wang, Y. Wei, T. Huang, Sta: Spatial-temporal attention for large-scale video-based person re-identification, in: Proceedings of the AAAI conference on artificial intelligence, Vol. 33, 2019, pp. 8287–8294.

[23] C. Liu, C. Wu, Y. F. Wang, S. Chien, Spatially and temporally efficient non-local attention network for video-based person re-identification, in: 30th British Machine Vision Conference 2019, BMVC 2019, Cardiff, UK, September 9-12, 2019, BMVA Press, 2019, p. 243.
URL `https://bmvc2019.org/wp-content/uploads/papers/0398-paper.pdf`

[24] J. Yang, W.-S. Zheng, Q. Yang, Y.-C. Chen, Q. Tian, Spatial-temporal graph convolutional network for video-based person re-identification, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 3289–3299.

[25] Z. Wang, S. Luo, H. Sun, H. Pan, J. Yin, An efficient non-local attention network for video-based person re-identification, in: Proceedings of the 2019 7th International Conference on Information Technology: IoT and Smart City, 2019, pp. 212–217.

[26] B. Zhou, A. Khosla, À. Lapedriza, A. Oliva, A. Torralba, Learning deep features for discriminative localization, in: 2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016, 2016, pp. 2921–2929. `doi:10.1109/CVPR.2016.319`.

[27] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 770–778.

[28] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, A. Torralba, Learning deep features for discriminative localization, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 2921–2929.

[29] T. Pfister, J. Charles, A. Zisserman, Flowing convnets for human pose estimation in videos, in: Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 1913–1921.

[30] J. Tompson, A. Jain, Y. LeCun, C. Bregler, Joint training of a convolutional network and a graphical model for human pose estimation, in: Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, K. Q. Weinberger (Eds.), Advances in Neural Information Processing Systems 27: Annual Conference on Neural

Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada, 2014, pp. 1799–1807.
URL https://proceedings.neurips.cc/paper/2014/hash/e744f91c29ec99f0e662c9177946c627-Abstract.html

[31] W. Du, Y. Wang, Y. Qiao, Rpan: An end-to-end recurrent pose-attention network for action recognition in videos, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 3725–3734.

[32] T. N. Kipf, M. Welling, Semi-supervised classification with graph convolutional networks, in: 5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings, 2017.

[33] Z.-M. Chen, X.-S. Wei, P. Wang, Y. Guo, Multi-label image recognition with graph convolutional networks, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 5177–5186.

[34] X. Wang, A. Gupta, Videos as space-time region graphs, in: Proceedings of the European conference on computer vision (ECCV), 2018, pp. 399–417.

[35] S. Yan, Y. Xiong, D. Lin, Spatial temporal graph convolutional networks for skeleton-based action recognition, in: Proceedings of the AAAI conference on artificial intelligence, Vol. 32, 2018.

[36] A. Hermans, L. Beyer, B. Leibe, In defense of the triplet loss for person re-identification, CoRR abs/1703.07737. arXiv:1703.07737.

[37] L. Zheng, Z. Bie, Y. Sun, J. Wang, C. Su, S. Wang, Q. Tian, MARS: A video benchmark for large-scale person re-identification, in: Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part VI, 2016, pp. 868–884. doi:10.1007/978-3-319-46466-4_52.

[38] T. Wang, S. Gong, X. Zhu, S. Wang, Person re-identification by video ranking, in: Computer Vision - ECCV 2014 - 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part IV, 2014, pp. 688–703. doi:10.1007/978-3-319-10593-2_45.

[39] M. Hirzer, C. Beleznai, P. M. Roth, H. Bischof, Person re-identification by descriptive and discriminative classification, in: Image Analysis - 17th Scandinavian Conference, SCIA 2011, Ystad, Sweden, May 2011. Proceedings, 2011, pp. 91–102. doi:10.1007/978-3-642-21227-7_9.

[40] P. F. Felzenszwalb, R. B. Girshick, D. A. McAllester, D. Ramanan, Object detection with discriminatively trained part-based models, IEEE Trans. Pattern Anal. Mach. Intell. 32 (9) (2010) 1627–1645. doi:10.1109/TPAMI.2009.167.

[41] A. Dehghan, S. M. Assari, M. Shah, GMMCP tracker: Globally optimal generalized maximum multi clique problem for multiple object tracking, in: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015, 2015, pp. 4091–4099. doi:10.1109/CVPR.2015.7299036.

[42] M. Long, H. Zhu, J. Wang, M. I. Jordan, Deep transfer learning with joint adaptation networks, in: Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017, 2017, pp. 2208–2217.

[43] D. Li, Z. Zhang, X. Chen, H. Ling, K. Huang, A richly annotated dataset for pedestrian attribute recognition, CoRR abs/1603.07054. arXiv:1603.07054.

[44] A. Krizhevsky, I. Sutskever, G. E. Hinton, Imagenet classification with deep convolutional neural networks, in: Advances in neural information processing systems, 2012, pp. 1097–1105.

[45] Z. Zhou, Y. Huang, W. Wang, L. Wang, T. Tan, See the forest for the trees: Joint spatial and temporal recurrent neural networks for video-based person re-identification, in: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 6776–6785.

[46] D. Chen, H. Li, T. Xiao, S. Yi, X. Wang, Video person re-identification with competitive snippet-similarity aggregation and co-attentive snippet embedding, in: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018.

[47] G. Song, B. Leng, Y. Liu, C. Hetang, S. Cai, Region-based quality estimation network for large-scale person re-identification, in: Proceedings of the Thirty-Second

AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018, 2018, pp. 7347–7354.

[48] Y. Zhao, X. Shen, Z. Jin, H. Lu, X.-s. Hua, Attribute-driven feature disentangling and temporal aggregation for video person re-identification, in: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2019.

[49] Y. Liu, Z. Yuan, W. Zhou, H. Li, Spatial and temporal mutual promotion for video-based person re-identification, in: The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019, 2019, pp. 8786–8793. `doi:10.1609/aaai.v33i01.33018786`.

[50] M. Jiang, B. Leng, G. Song, Z. Meng, Weighted triple-sequence loss for video-based person re-identification, Neurocomputing 381 (2020) 314–321. `doi:10.1016/j.neucom.2019.11.088`.

[51] W. Zhang, X. He, X. Yu, W. Lu, Z. Zha, Q. Tian, A multi-scale spatial-temporal attention model for person re-identification in videos, IEEE Transactions on Image Processing 29 (2020) 3365–3373.

[52] S. Li, H. Yu, H. Hu, Appearance and motion enhancement for video-based person re-identification., in: AAAI, 2020, pp. 11394–11401.

[53] Z. Chen, Z. Zhou, J. Huang, P. Zhang, B. Li, Frame-guided region-aligned representation for video person re-identification., in: AAAI, 2020, pp. 10591–10598.

[54] Y. Yan, J. Qin, J. Chen, L. Liu, F. Zhu, Y. Tai, L. Shao, Learning multi-granular hypergraphs for video-based person re-identification, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 2899–2908.

[55] Y. Zhao, X. Shen, Z. Jin, H. Lu, X.-s. Hua, Attribute-driven feature disentangling and temporal aggregation for video person re-identification, in: Proceedings of

the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 4913–4922.

[56] J. Liu, X. Zhu, Z.-J. Zha, Temporal attribute-appearance learning network for video-based person re-identification, arXiv preprint arXiv:2009.04181.

[57] Z. Chen, A. Li, S. Jiang, Y. Wang, Attribute-aware identity-hard triplet loss for video-based person re-identification, arXiv preprint arXiv:2006.07597.

[58] Y. Ding, Z. Ma, S. Wen, J. Xie, D. Chang, Z. Si, M. Wu, H. Ling, Ap-cnn: weakly supervised attention pyramid convolutional neural network for fine-grained visual classification, IEEE Transactions on Image Processing 30 (2021) 2826–2836.

[59] K. Zhang, N. Liu, X. Yuan, X. Guo, C. Gao, Z. Zhao, Z. Ma, Fine-grained age estimation in the wild with attention lstm networks, IEEE Transactions on Circuits and Systems for Video Technology 30 (9) (2019) 3140–3152.