Keypoint Message Passing for Video-Based Person Re-Identification

Di Chen[†],[1,3] **Andreas Döring**[†],[2] **Shanshan Zhang**[*],[1] **Jian Yang**[*],[1] **Juergen Gall**,[2] **Bernt Schiele**[3]

[1] PCA Lab, Key Lab of Intelligent Perception and Systems for High-Dimensional Information of Ministry of Education, Jiangsu Key Lab of Image and Video Understanding for Social Security, Nanjing University of Science and Technology

[2] University of Bonn

[3] Max Planck Institute for Informatics

{dichen,shanshan.zhang,csjyang}@njust.edu.cn, {doering,gall}@iai.uni-bonn.de, {dichen,schiele}@mpi-inf.mpg.de

Abstract

Video-based person re-identification (re-ID) is an important technique in visual surveillance systems which aims to match video snippets of people captured by different cameras. Existing methods are mostly based on convolutional neural networks (CNNs), whose building blocks either process local neighbor pixels at a time, or, when 3D convolutions are used to model temporal information, suffer from the misalignment problem caused by person movement. In this paper, we propose to overcome the limitations of normal convolutions with a human-oriented graph method. Specifically, features located at person joint keypoints are extracted and connected as a spatial-temporal graph. These keypoint features are then updated by message passing from their connected nodes with a graph convolutional network (GCN). During training, the GCN can be attached to any CNN-based person re-ID model to assist representation learning on feature maps, whilst it can be dropped after training for better inference speed. Our method brings significant improvements over the CNN-based baseline model on the MARS dataset with generated person keypoints and a newly annotated dataset: PoseTrackReID. It also defines a new state-of-the-art method in terms of top-1 accuracy and mean average precision in comparison to prior works.

Introduction

Visual surveillance systems play a vital role in modern society to ensure public safety. The massive data collected by the systems raises the need for automatic visual understanding techniques such as person re-identification (re-ID). Person re-ID aims to associate the same person across cameras with non-overlapping views, which is usually achieved by calculating the similarities between the representations of images. Compared to images, video sequences provide much richer information which is beneficial to address visual ambiguities. Therefore, video-based person re-ID emerged recently as a parallel research field to image-based person re-ID.

The central problem of learning discriminative video representations is how to exploit both spatial and temporal information. Most existing solutions propose to capture spatial and temporal information separately, *i.e.* using a 2D convolutional neural network (CNN) for spatial representation learning, while handling temporal information by aggregating the high-level outputs of CNNs with pooling (Zheng et al. 2016a), recurrent neural networks (Chung, Tahboub, and Delp 2017; McLaughlin, Del Rincon, and Miller 2016; Chen et al. 2018; Xu et al. 2017; Zhou et al. 2017) or temporal attention (Fu et al. 2019; Zhou et al. 2017; Xu et al. 2017; Liu, Yan, and Ouyang 2017; Li et al. 2018). Other works (Liu et al. 2019b,a; Li, Zhang, and Huang 2019) learn concurrent spatial-temporal representations with 3D convolutions. The core operation of both types of methods is convolution, which only processes information at a small local range, especially when it is located at shallow hierarchies of a network. Meanwhile, 3D convolution layers also have the misalignment problem, *i.e.* the same position in adjacent frames may belong to different body parts due to person movement, such that the appearance representations learned by 3D convolution are polluted with the wrong body part or background. The spatially local and temporally misalignment problems are shown in Fig. 1 (a) and (b).

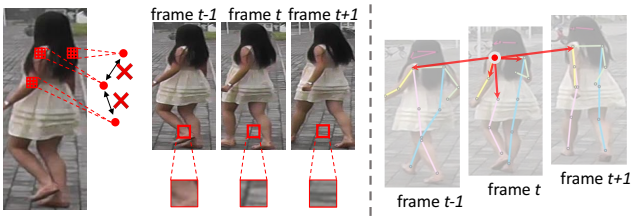


Figure 1: Problems using 2D/3D convolution: (a) spatially local and (b) temporal misalignment; We propose (c) keypoint message passing, where keypoint features on frame t (marked with white circle) are updated with information from connected nodes (red arrow) by graph convolution, which is not bounded by the fixed shape and location of normal convolutions. The complete spatial-temporal graph structure is shown in Fig. 3.

In this paper, we propose to overcome the limitations of convolutions in video-based person re-ID with the help of person pose estimation and graph convolution. Since hu-

[†]Equal contribution.

[*]Corresponding author.

man body keypoints are representative for distinct shapes of different people, we construct the graph based on keypoints. Specifically, the features located at person joint keypoints are cropped from the CNN feature maps, which are then used as the node features for constructing a spatial-temporal graph. The graph-structured data is then processed with graph convolutions, which is not bounded by small-scale, rectangular-shaped kernels of normal convolution. A brief illustration is shown in Fig. 1 (c). We can see that features located at, *e.g*. one's left shoulder, interact with features from the right shoulder, left elbow and left hip simultaneously, which is otherwise hard to reach for a single convolutional layer due to the large spatial distances. Meanwhile, the left shoulder features at video frame t also receive information from the same location at adjacent frames without misalignment problems. The same process happens to all other keypoints on the human body, enriching the local keypoint features with non-local, temporal-aligned and human-oriented information. Since the core idea is based on the message passing mechanism (Gilmer et al. 2017) of graph convolutions, we name our method *keypoint message passing*.

When we attach a graph convolutional network (GCN) to a CNN, it comes with significant computational cost and extra memory consumption, especially when the graph scale is large and the GCN is deep. To this end, we propose a flexible design which enables graph convolutions during training but does not require the graph during inference. An illustration is shown in Fig. 2. The GCN serves as a parallel branch along side the CNN, which functions as a training guide. Supervision signals with spatial-temporal information flows back from the GCN to the CNN by back-propagation. Therefore, each keypoint location on the CNN feature map receives feedback from a more diverse set of spatial-temporal locations, especially for the shallow layers which have small receptive fields. Once the model is trained, the whole GCN branch, keypoints and graphs can be dropped, leaving no extra computational burden other than the CNN branch. Moreover, the choice of the CNN is not limited, *i.e*. our design is generalizable to any CNN-based re-ID model. We name our method 'KMPNet'.

In summary, our contribution is three-fold:

- With the help of spatial-temporal guidance provided by person joint keypoints and graph convolutions, we propose a general method to assist CNN-based re-ID model training, which overcomes the limitation of normal convolutions without extra computational burden during inference.
- We present PoseTrackReID, a new dataset for video-based re-ID featuring both person ID and keypoint annotations.
- Extensive experiments demonstrate that our model significantly improves the baseline, achieving results on par with or better than the current state-of-the-art models.

Related Work

Image-based Person Re-Identification. Image-based person re-ID models usually serve as good baselines for video-based re-ID methods. Early re-ID models mainly focus on designing discriminative hand-crafted features (Wang et al. 2007; Farenzena et al. 2010; Zhao, Ouyang, and Wang 2013; Liao et al. 2015) and distance metrics (Kostinger et al. 2012; Li et al. 2015; Zhang, Xiang, and Gong 2016). Nowadays, designing re-ID models based on CNNs has become main stream. These methods typically formulate re-ID as a classification (Xiao et al. 2016; Zheng et al. 2016b; Fan et al. 2018; Xiang et al. 2018) or ranking (Yi et al. 2014; Li et al. 2014; Ahmed, Jones, and Marks 2015; Varior et al. 2016; Liu et al. 2017; Xu et al. 2018) problem at training time, and use the optimized backbone network as a feature extractor during inference. Instead of extracting global features with global average pooling, recent methods (Sun et al. 2018; Wei et al. 2017; Zhao et al. 2017; Yao et al. 2019) propose to divide the final CNN feature maps into several parts and use average pooling separately. For example, PCB (Sun et al. 2018) partitions the feature maps into horizontal stripes and then concatenates the pooled stripe features to generate the final features, which contain richer spatial information and thus achieve better performance than the simple global features. In our ablation studies, we also choose PCB as the base CNN of the visual branch, whereas it is worth to notice that any CNN based model is a candidate.

Video-based Person Re-Identification. The most direct way for video-based re-ID is to lift image-based re-ID methods by aggregating multi-frame features via different operations, such as mean/max pooling (Zheng et al. 2016a), recurrent neural networks (RNNs) (Chung, Tahboub, and Delp 2017; Chen et al. 2018; Xu et al. 2017; Zhou et al. 2017) and temporal attention (Fu et al. 2019; Zhou et al. 2017; Xu et al. 2017; Liu, Yan, and Ouyang 2017; Li et al. 2018). Another strategy is to capture the spatial and temporal information simultaneously by 3D convolution (Liu et al. 2019b,a; Li, Zhang, and Huang 2019). Despite their favorable performance, 3D convolutions usually require more computational and memory resources, which limits their potential for real-world applications. The graph convolution in our method shares similar concept with 3D convolution on concurrent spatial-temporal information modeling. However, our method differs from 3D CNNs in that the input data has a non-local structure, whereas the input to 3D convolutions must be a rectangular-shaped local range of pixels. Besides, our method does not suffer from the temporal misalignment problem as 3D convolutions do, since the cross-frame features are always extracted at the same location on the human body. Additionally, all the graph convolution operations are only performed at training time, thus no extra computational costs are required during inference.

Pose-assisted Person Re-Identification. Benefiting from recent advances on pose estimation (Rafi et al. 2020; Cao et al. 2019; Xiao, Wu, and Wei 2018; Sun et al. 2019), person keypoints have been utilized to facilitate person re-ID models. Some works (Ge et al. 2018; Liu et al. 2018) focus on generating person images with keypoints which are later used as extra training data. Others propose to use person keypoints as a spatial cue for aligning body parts (Suh et al. 2018; Su et al. 2017; Wu et al. 2020) or highlight-

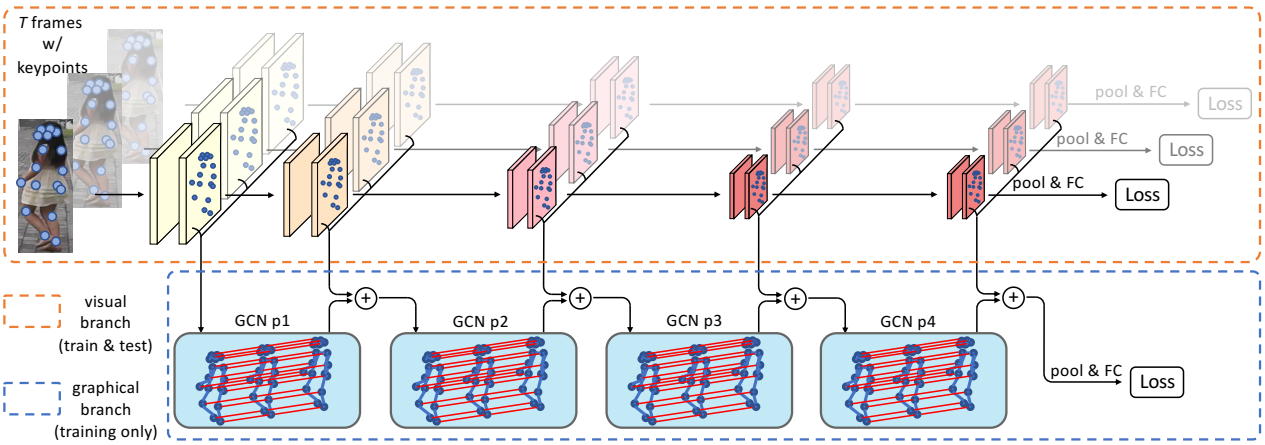


Figure 2: Overall pipeline for our method. The visual branch is a base CNN which is a typical 5-stage model and takes in a video with T frames as input. The graphical branch is a GCN divided into 4 parts. Given the keypoint locations (either annotated or generated with a pose estimation model), we extract features according to the locations on the first CNN stage and use them as the inputs to the GCN. At the end of each stage, keypoint features are fused with the intermediate representations from GCN with an element-wise sum. The fused features serve as the new inputs to the subsequent GCN layers. The detailed graph topology is shown in Fig. 3. Both the CNN and GCN are supervised with cross-entropy loss during training. Note, that the GCN branch including keypoint estimation can be dropped during inference.

ing non-occluded regions (Miao et al. 2019). For instance, Su et al. (2017) crop out person parts from the input images according to provided keypoints and re-assemble them into a pose-normalized synthetic image. Miao et al. (2019) use the keypoint heatmaps as spatial attention which is then multiplied with the feature maps element-wise before average pooling. In our method, person keypoints also play a key role. However, they are introduced with a different motivation, *i.e*. refining CNN features by capturing non-local human-oriented information within video frames as well as temporal information between frames.

Graph Convolutional Networks. The growing need for processing non-Euclidean data has motivated research on graph convolutional networks (Kipf and Welling 2016; Chen, Ma, and Xiao 2018; Hamilton, Ying, and Leskovec 2017; Huang et al. 2018; Li et al. 2020). Some computer vision researchers also take advantage of GCNs on tasks such as action recognition (Yan, Xiong, and Lin 2018), video classification (Wang and Gupta 2018) and gait recognition (Li, Zhao, and Ma 2020). Meanwhile, some person re-ID works (Wu et al. 2020; Shen et al. 2018; Yang et al. 2020; Yan et al. 2020) also exploit GCNs for unstructured relationship learning. Shen et al. (2018) builds a graph based on probe-gallery image pairs and utilize a GCN for better similarity estimation. Shen et al. (2018),Yang et al. (2020) and Yan et al. (2020) propose to model the relationship of intra-frame spatial parts and inter-frame temporal cues for a video. Our method differs from these works in the following three aspects. Firstly, instead of using horizontal parts as the graph node, we use person keypoints which provide better localization on the human body, and thus avoid the misalignment problem naturally. Secondly, the graph convolution for spatial-temporal refinement is applied to all levels of CNN features, whereas Shen et al. (2018),Yang et al. (2020) and Yan et al. (2020) only use a GCN for high-level features. Thirdly, the graphical branch with graph convolutions is only used to assist CNN feature training. While during inference, it can be dropped entirely to save computation and memory resources.

Methodology

The Framework

A brief overview of our model is shown in Fig. 2. Our model is mainly composed of two branches, *i.e*. the *visual* branch and the *graphical* branch. The visual branch is a base CNN model, as a canonical choice for person re-identification. The input to the visual branch is a video with T frames, while each frame is processed as an individual image. The graphical branch is a GCN model, which is manually partitioned into 4 stages to match the hierarchies of the base CNN. We extract the person keypoint features with an RoIAlign (He et al. 2017) operation from the last layer of each CNN stage, which are further constructed into a spatial-temporal graph and serve as input to the corresponding GCN stages for long-range interaction. The keypoint locations are manually annotated or generated with an off-the-shelf pose estimation model. Specifically, keypoint features from CNN stage 1 are used as the initial input to GCN part 1. After passing through several graph convolution layers, the outputs of GCN part 1 are then fused with keypoint features from CNN stage 2 with an element-wise sum. The "convolution-fusion" paradigm is repeated until the end of both branches. During training, both of the branches are supervised with cross-entropy losses. Gradients from the graphical branch flow back to the visual branch, which enhances the features located at person joints with long-range information. During inference, we only use the enhanced visual features pro-

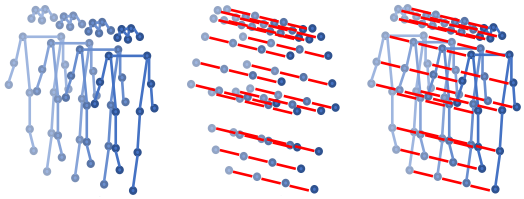


Figure 3: Spatial-temporal graph for a sequence of person joints. Figures from left to right denote spatial edges E_S (blue), temporal edges E_T (red) and all edges E respectively. We use E in our KMP method unless specified otherwise.

duced by the CNN for person matching. At this time, the whole GCN branch can be dropped to reduce the computational burden.

Graph Formulation

Following (Yan, Xiong, and Lin 2018) and (Li, Zhao, and Ma 2020), we represent a sequence of person joints with a spatial-temporal graph $G = (V, E)$. The node set $V = \{v_{ti}|t = 1, \ldots, T, i = 1, \ldots, N\}$ includes all the joints in this sequence, where T is the number of frames and N is the number of keypoints in each frame. The edge set E is composed of two subsets, namely *spatial* set E_S and *temporal* set E_T. The spatial edges, denoted as $E_S = \{(v_{ti}, v_{tj})|(i, j) \in H, t = 1, \ldots, T\}$, is a direct representation of natural human topology H in each frame, as is shown by the blue lines in Fig. 3. The temporal set $E_T = \{(v_{ti}, v_{(t+1)i}), |t = 1, \ldots, T, i = 1, \ldots, N\}$ consists of connections of the same joints between frame t and $t + 1$. It is illustrated as the red lines in Fig. 3.

There are different options for features of each graph node. In (Yan, Xiong, and Lin 2018) and (Li, Zhao, and Ma 2020), the keypoint coordinates are used as node features in order to capture the action or gait information. Differently, we use visual features cropped from CNN feature maps as graph node feature, since our motivation is to capture the non-local, temporal-aligned relationships between different locations, rather than modeling movement information. Specifically, the feature for node v at layer l is represented as $\mathbf{h}_v^{(l)}$. It has three variants depending on which layer it is processed. For the first layer of GCN p1, $\mathbf{h}_v^{(l)}$ is the plain keypoint feature cropped from the feature maps of CNN stage 1. For the hidden layers inside each GCN part, $\mathbf{h}_v^{(l)}$ is the latent outputs from layer $l - 1$. The input features for the first layers of GCN p2 to GCN p4 are linear combinations of the latent outputs and CNN features:

$$\mathbf{h}_v^{(l)} \leftarrow \mathbf{W}_{down}^T(\mathbf{W}_{up}^T\mathbf{h}_v^{(l)} + f(i, j)), \tag{1}$$

where f denotes the CNN feature map and i, j are the spatial coordinates of node v; $\mathbf{W}_{up}$ and $\mathbf{W}_{down}$ are weights of fully-connected layers to match the dimensions of $\mathbf{h}_v^{(l)}$ and $f(i, j)$.

Keypoint Message Passing

Once the graph topology and node features are defined, graph convolutions could be applied to update the node features. We adopt the improved version of graph convolution block from (Li et al. 2020), which takes advantage of generalized message aggregation, modified skip connections and a novel normalization method. The block consists of a series of operations, including normalization, non-linear activation, dropout, graph convolution and residual addition. For simplicity, we only introduce the graph convolution operation since the message passing between different nodes only happens at this step.

For GCN layer l, the graph convolution is mainly composed of three actions, namely *message construction* $\boldsymbol{\rho}$, *message aggregation* $\boldsymbol{\zeta}$ and *vertex update* $\boldsymbol{\phi}$. For message construction, we focus on node features only and omit the edge features. Therefore, the message construction function $\boldsymbol{\rho}^{(l)}(\cdot)$ is simply defined as:

$$\mathbf{m}_{vu}^{(l)} = \boldsymbol{\rho}^{(l)}(\mathbf{h}_v^{(l)}, \mathbf{h}_u^{(l)}) = \text{ReLU}(\mathbf{h}_u^{(l)}) + \epsilon, u \in \mathcal{N}(v) \tag{2}$$

where $\mathbf{m}_{vu}^{(l)}$ indicates the message passed from node u to v; $\mathcal{N}(v)$ denotes the neighbour nodes of vertex v; ϵ is a small constant introduced for numerical stability. Eqn. 2 means the rectified node features are directly used as neighbour messages. For the message aggregation function $\boldsymbol{\zeta}^{(l)}(\cdot)$, we construct the aggregated message $\mathbf{m}_v^{(l)}$ for node v with the form of softmax with a learnable temperature τ:

$$\mathbf{m}_v^{(l)} = \sum_{u \in \mathcal{N}(v)} \frac{e^{\frac{1}{\tau}\mathbf{m}_{vu}^{(l)}}}{\sum_{i \in \mathcal{N}(v)} e^{\frac{1}{\tau}\mathbf{m}_{vi}^{(l)}}} \cdot \mathbf{m}_{vu}^{(l)} \tag{3}$$

which can be regarded as a weighted summation of all the neighbour messages. The aggregated message $\mathbf{m}_v^{(l)}$ is then used to update the node feature of v with a function $\boldsymbol{\phi}^{(l)}(\cdot)$:

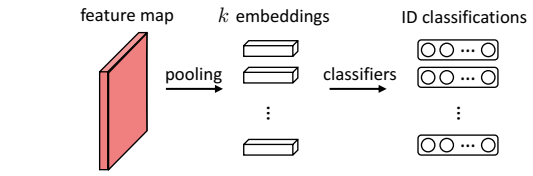
$$\mathbf{h}_v^{(l+1)} = \boldsymbol{\phi}^{(l)}(\mathbf{h}_v^{(l)}, \mathbf{m}_v^{(l)}) = \text{MLP}(\mathbf{h}_v^{(l)} + \mathbf{m}_v^{(l)}) \tag{4}$$

where $\text{MLP}(\cdot)$ is a multi-layer perceptron with 2 fully-connected layers.

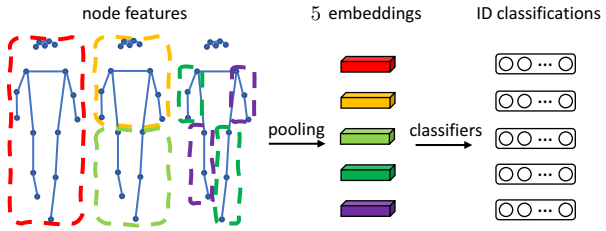
At the end of the GCN, the node features are pooled and mapped into discriminative embeddings. An illustration is shown in Fig. 4(b). We adopt the pooling operation in (Li, Zhao, and Ma 2020), which combines multiple partition patterns and uses average pooling within each part. Specifically, there are three types of discriminative embeddings, namely 1) average feature of the whole body, 2) the upper and lower body separately averaged features and 3) averaged pair features of (left arm, right leg) and (right arm, left leg). All the features are then mapped with a fully connected layer without weight sharing. In total, we obtain 5 discriminative embeddings, denoted as $\{\mathbf{x}_g^i|i = 1, \ldots, 5\}_t$ on frame t, which are further supervised by training objectives.

Training & Inference

For a video sequence with T frames, a forward pass of our model generates two sets of embeddings from the visual branch and the graphical branch respectively. The CNN



(a) Visual branch



(b) Graphical branch

Figure 4: Pooling and classification for visual and graphical branches. (a) Spatial pooling on CNN feature maps. (b) Graph pooling (Li, Zhao, and Ma 2020) on node features. Nodes wrapped with the same color are average pooled into a single embedding.

embeddings from the visual branch is denoted as $\{\mathbf{x}_c^i | i = 1, \ldots, k\}_t, t = 1, \ldots, T$, where k is a hyper-parameter depending on the CNN design. For example, $k = 1$ for an IDE model (Zheng, Zheng, and Yang 2017) since global average pooling is applied on the final feature map, producing one single embedding for a frame. For a PCB model (Sun et al. 2018), k equals the number of horizontal stripes which is typically set to 4 or 6. Similarly, the output of the graphical branch is the GCN embeddings $\{\mathbf{x}_g^i | i = 1, \ldots, 5\}_t, t = 1, \ldots, T$. During training, all embeddings are input into $k+5$ classifiers respectively. Each classifier is composed of a fully connected layer and a softmax activation, which is supervised by a cross-entropy loss.

During inference, the embeddings on each frame are *concatenated* and then *averaged* over the temporal dimension, thus the CNN and GCN embeddings for the whole sequence is denoted as $\mathbf{x}_c$ and $\mathbf{x}_g$ respectively. We find through experiments that $\mathbf{x}_c$ shows better performance than $\mathbf{x}_g$. Therefore, we report the re-ID results using only $\mathbf{x}_c$ unless mentioned otherwise.

Discussion: Why discard GCN during inference? Similar to dropping the discriminator in GANs and classifier layer in re-ID models, we also drop some part of our model during inference. In our case, the entire graphical branch is dropped since we only use the CNN embedding $\mathbf{x}_c$. The reason why we discard the graphical branch is twofold: 1) The forward pass of CNN does not rely on any information provided by the GCN. 2) The CNN embedding $\mathbf{x}_c$ performs better than GCN embedding $\mathbf{x}_g$. It is worth to notice that although GCN is not used, it benefits CNN training via the back-propagated gradients. In this way, $\mathbf{x}_c$ is enriched with long-range semantic dependencies and aligned temporal information, which is the fundamental advantage brought by keypoint message passing.

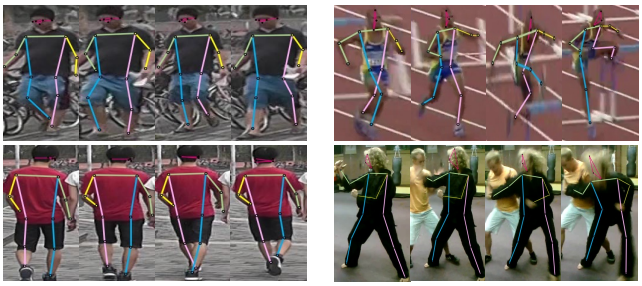


Figure 5: Sample tracklets in MARS (left) and PoseTrackReID (right). Person keypoints on MARS are generated by an off-the-shelf pose estimator (Rafi et al. 2020), whilst on PoseTrackReID the keypoints are manually annotated.

Experiments

In this section, we first introduce the datasets, evaluation protocols and implementation details. Then we compare our method to state-of-the-art methods, followed by extensive ablation studies. More analytical experiments are included in our supplementary material.

Datasets and Evaluation Protocol

MARS (Zheng et al. 2016a) is a large-scale benchmark dataset for video-based person re-ID. All the videos are collected with 6 stationary cameras on a university campus, from which a DPM detector (Felzenszwalb et al. 2009) and GMMCP tracker (Dehghan, Modiri Assari, and Shah 2015) are used to crop out the person regions. The training set consists of $8,298$ tracklets of 625 identities, while the testing set includes $1,980$ tracklets of 626 identities for query and $6,082$ tracklets of 620 identities for gallery. The person keypoints are generated with a top-down pose estimation model proposed in (Rafi et al. 2020) on each frame. Some frames with visualized poses are shown in Fig. 5.

PoseTrackReID is a new dataset proposed in this work to facilitate more comprehensive experiments for video-based person re-ID. It is a cropped subset of the PoseTrack 2018 dataset (Andriluka et al. 2018) which is originally proposed for multi-person pose estimation and articulated tracking. The videos are captured in various scenes with large amounts of pose, appearance and scale variation. They also contain challenging scenes with severe body part occlusion and truncation. The person keypoints are manually annotated at an interval of 4 frames at the beginning and at the end of a sequence, whereas the center of a sequence contains 30 consecutive manually annotated keypoints. Based on PoseTrack 2018, we construct PoseTrackReID by adding additional annotations of person bounding box and global person ID. The person regions are then cropped out for video-based person re-ID. The training set of PoseTrackReID is gathered from the training set of PoseTrack 2018, including $7,725$ tracklets of $5,350$ identities. The query set consists of 847 tracklets of 830 identities, while the gallery set includes $1,965$ tracklets of $1,696$ identities. Both the query and gallery sets are collected from the validation set of PoseTrack 2018. Some example frames are shown in Fig. 5.

Method	top-1	mAP
CNN + XQDA (Zheng et al. 2016a)	65.3	47.6
SeeForest (Zhou et al. 2017)	70.6	50.7
DuATM (Si et al. 2018)	81.2	67.7
Snippet (Chen et al. 2018)	86.3	76.1
ADFD (Zhao et al. 2019)	87.0	78.2
COSAM (Subramaniam et al. 2019)	84.9	79.9
GLTR (Li et al. 2019)	87.0	78.5
TCLNet (Hou et al. 2020)	**88.8**	83.0
KMPNet (ResNet-50)	86.7	**84.4**
ASTPN (Xu et al. 2017)	44.0	-
VRSTC (Hou et al. 2019)	88.5	82.3
MG-RAFA (Zhang et al. 2020)	88.8	85.9
STGCN (Yang et al. 2020)	90.0	83.7
MGH (Yan et al. 2020)	90.0	85.8
STRF (Aich et al. 2021)	90.3	86.1
KMPNet (PCB)	89.7	86.5
KMPNet (MGH)	**92.0**	**86.6**

Table 1: Performance comparison with state-of-the-art video-based person re-ID methods on MARS. ResNet-50, PCB and MGH in the brackets denote the base CNN of our visual branch. Methods in the upper block use global average pooling while the ones in the lower block use part-based pooling. Best performances in each block are marked bold.

Evaluation Protocols. We use Cumulative Matching Characteristics (CMC) and mean Average Precision (mAP) as the evaluation metrics, which is the standard on the MARS benchmark. On PoseTrackReID, we follow the rules of MARS by using CMC and mAP as well.

Implementation

We choose ResNet-50 (He et al. 2016) as the base CNN for the visual branch. We adopt the 28-layer GCN model in (Li et al. 2020) and remove the first graph convolution layer to match the visual branch. We then partition the remaining 27 layers *in proportion to* the design of ResNet-50. It is also straightforward to replace the ResNet-50 with other backbones. The dimension for the latent node features of GCN is set to 64. Please refer to the supplementary material for more details. The new dataset and code will be released at https://github.com/DeanChan/KeypointMessagePassing.

Comparison to the State-of-the-Art

In this section, we compare our KMPNet to state-of-the-art methods on both MARS and PoseTrackReID. We choose three representative base CNNs as the visual branch of our KMPNet, namely ResNet-50 (He et al. 2016), PCB (Sun et al. 2018) and MGH (Yan et al. 2020). The results on MARS are summarized in Tab. 1. For clear comparison, we group the methods in the table according to the pooling method used at the top convolutional layer. Methods in the upper block adopt global average pooling, while the lower block features part-based pooling.

Method	top-1	mAP
ResNet-50 (He et al. 2016)	75.1	79.4
PCB (Sun et al. 2018)	77.9	81.5
MGH (Yan et al. 2020)	82.7	84.2
KMPNet (ResNet-50)	78.7	82.7
KMPNet (PCB)	79.2	82.7
KMPNet (MGH)	**83.3**	**84.9**

Table 2: Performance comparison on PoseTrackReID.

Compared to the most recent methods, including MG-RAFA (Zhang et al. 2020), STGCN (Yang et al. 2020), MGH (Yan et al. 2020) and TCLNet (Hou et al. 2020), our KMPNet with a simple PCB (Sun et al. 2018) as the visual branch achieves higher mAP and comparable top-1 accuracy. Apart from the backbone CNN, all these methods require some extra computations such as spatial-temporal attention (Zhang et al. 2020), recursive feature erasing (Hou et al. 2020) and graph convolution (Yang et al. 2020; Yan et al. 2020). In contrast, our method only requires graph convolution at the training stage. During inference, no other computations are needed other than the backbone CNN, which makes our method computationally efficient.

Moreover, our method could also be used to boost the performance of other models by replacing the PCB baseline in the visual branch. For example, applying our keypoint message passing method to the MGH (Yan et al. 2020) model[1] improves the top-1 accuracy and mAP by 2.0 and 0.8 pp. respectively, setting a new state-of-the-art performance. This improvement also indicates that our method has good generalization ability w.r.t. different visual baselines.

We also show the re-ID results on PoseTrackReID in Tab. 2. The re-implemented CNNs are grouped in the upper block, while the corresponding GCN enhancements are listed in the lower block. We can see from Tab. 2 that for all the three visual baselines, KMPNet consistently improves their top-1 accuracies and mAPs. Compared to MARS, PoseTrackReID is more diverse with various background and human poses. The fact that KMPNet also excels on PoseTrackReID suggests that our method is also generalizable to different scenes.

Ablation Study

In this subsection, we conduct analytical experiments on both MARS and PoseTrackReID. For simplicity, we focus on one specific design which uses PCB (Sun et al. 2018) as the visual branch of our KMPNet. The corresponding results and conclusions also apply to other base CNNs.

Analysis on the spatial temporal relationships. Compared to the PCB baseline, our model mainly benefits from two aspects, *i.e*. the non-local spatial dependencies and the cross-frame temporal information. In order to better understand the contribution of the two components, we conduct ablation studies on the spatial temporal structures of the graph.

[1]During training, the visual branch is initialized with the published model weight of MGH.

Model Variants	MARS		PoseTrackReID	
	top-1	mAP	top-1	mAP
PCB baseline	85.3	84.6	77.9	81.5
+ fine-tune	84.7	84.6	77.2	81.4
+ spatial	88.6	85.5	78.9	82.1
+ temporal	88.7	85.3	77.9	81.8
+ both	89.7	86.5	79.2	82.7

Table 3: Ablation study on spatial and temporal connections.

Embedding	MARS		PoseTrackReID	
	top-1	mAP	top-1	mAP
$\mathbf{x}_c$	89.7	86.5	79.2	82.7
$\mathbf{x}_g$	61.7	57.3	48.3	51.6
$\text{cat}(\mathbf{x}_c, \mathbf{x}_g)$	62.4	57.7	49.0	52.2

Table 4: Performance for visual/graphical embeddings. 'cat' stands for concatenation.

Starting from the PCB baseline, which is basically the visual branch of our KMPNet, we add the graphical branch with different graph structures, namely *spatial-only graph*, *temporal-only graph* and *spatial-temporal graph*. The three variants of graph structure are demonstrated in Fig. 3.

The comparison is shown in the lower block of Tab. 3. We can see that adding spatial information with our method improves the performance of the PCB baseline by 3.3 and 0.9 pp. w.r.t. to top-1 and mAP on MARS, which means that adding non-local information during training is beneficial for re-ID feature learning. Meanwhile, the efficacy of temporal information is also clear: increasing top-1 and mAP by 3.4 and 0.7 pp. Similarly, the experiment results on PoseTrackReID also reveal the same conclusion. Finally, our final KMPNet featuring both spatial and temporal graph achieves the best performance over either of them alone.

On the other hand, we also show in the upper block of Tab. 3 a control experiment where the model is trained longer with the same learning rate and epochs as our KMPNet but without the assistance of the graphical branch. The re-ID accuracy of this model was not increased, which suggests that the performance gain is not due to the extra fine-tuning stage but the message passing via graphs.

Based on the above results, we could draw the conclusion that the strategy of guiding CNN training with spatial-temporal information and graph convolution is effective.

Embedding choices. The two branches of our KMPNet produce two sets of embeddings respectively, *i.e*. the CNN and GCN embeddings $\mathbf{x}_c$ and $\mathbf{x}_g$. In practice, we only use $\mathbf{x}_c$ for calculating the similarity between probe-gallery pairs. What if we also take $\mathbf{x}_g$ into consideration? How would it affect the re-ID performance? The answer to this question lies in Tab. 4, from which we can see that neither the top-1 accuracy nor the mAP of $\mathbf{x}_g$ are comparable to that of $\mathbf{x}_c$. Meanwhile, concatenating $\mathbf{x}_g$ to $\mathbf{x}_c$ would also lower the performance of $\mathbf{x}_c$. The performance degeneration of $\mathbf{x}_g$ suggests that keypoint features processed by graph convolutions are not as expressive as CNN features, since they are just a sampled subset of CNN feature maps. However, it does not obliterate the contribution of the GCN since it significantly boosts the performance of the CNN embedding $\mathbf{x}_c$, as is shown in Tab. 3. Based on the analysis above, we decide to use only $\mathbf{x}_c$ for matching the query and gallery persons, which makes it possible to remove the whole graphical branch during inference. Therefore, the computation and memory resources needed are drastically reduced.

Figure 6: Feature map visualizations for our KMPNet and the PCB baseline. Warmer color denotes stronger activation.

Feature map visualizations. We visualize the feature maps of some representative samples in Fig. 6. The activation maps are obtained from the channel-wise max of features in 'conv1'. We can see from Fig. 6 that our KMPNet has stronger activation on the human body region than the baseline PCB model, despite that they have the same architecture during inference. This is because the graphical branch in our KMPNet helps to cast stronger feedback signals onto the keypoint locations on CNN feature maps during training. As a result, the learned visual branch tends to focus more on the human body region. Therefore, more discriminative information could be discovered on human body.

Conclusion

In this paper, we present KMPNet, a spatial-temporal enhanced model for video-based person re-identification. A graphical branch featuring a graph convolutional network is attached alongside a visual branch, which can be initialized with any CNN-based person re-ID model. In the training stage, the graphical branch assists the CNN training by passing spatial and temporal messages on the feature maps, where spatial messages are passed among joint keypoints on the human body and temporal messages are passed between the same keypoints of adjacent video frames. During inference, the entire graphical branch can be dropped for efficiency, while the visual branch alone shows superior performance over the initial CNN model. Extensive experiments on the MARS and PoseTrackReID dataset demonstrate the effectiveness of our method.

Acknowledgements

The authors would like to thank the AC and the anonymous reviewers for their critical and constructive comments and suggestions. This work has been funded by the National Science Fund of China (NSFC) (Grant No. U1713208 and 62172225), Funds for International Cooperation and Exchange of NSFC (Grant No. 61861136011), the Fundamental Research Funds for the Central Universities (No. 30920032201), National Key R&D Program of China (2017YFC0820601, 2021YFA1001100) and the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) - GA 1927/8-1.

References

Ahmed, E.; Jones, M.; and Marks, T. K. 2015. An improved deep learning architecture for person re-identification. In *CVPR*.

Aich, A.; Zheng, M.; Karanam, S.; Chen, T.; Roy-Chowdhury, A. K.; and Wu, Z. 2021. Spatio-temporal representation factorization for video-based person re-identification. In *ICCV*.

Andriluka, M.; Iqbal, U.; Insafutdinov, E.; Pishchulin, L.; Milan, A.; Gall, J.; and Schiele, B. 2018. Posetrack: A benchmark for human pose estimation and tracking. In *CVPR*.

Cao, Z.; Hidalgo Martinez, G.; Simon, T.; Wei, S.; and Sheikh, Y. A. 2019. OpenPose: Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields. *TPAMI*.

Chen, D.; Li, H.; Xiao, T.; Yi, S.; and Wang, X. 2018. Video person re-identification with competitive snippet-similarity aggregation and co-attentive snippet embedding. In *CVPR*.

Chen, J.; Ma, T.; and Xiao, C. 2018. Fastgcn: fast learning with graph convolutional networks via importance sampling. *arXiv preprint arXiv:1801.10247*.

Chung, D.; Tahboub, K.; and Delp, E. J. 2017. A two stream siamese convolutional neural network for person re-identification. In *ICCV*.

Dehghan, A.; Modiri Assari, S.; and Shah, M. 2015. Gmmcp tracker: Globally optimal generalized maximum multi clique problem for multiple object tracking. In *CVPR*, 4091–4099.

Fan, X.; Jiang, W.; Luo, H.; and Fei, M. 2018. SphereReID: Deep Hypersphere Manifold Embedding for Person Re-Identification. *arXiv preprint arXiv:1807.00537*.

Farenzena, M.; Bazzani, L.; Perina, A.; Murino, V.; and Cristani, M. 2010. Person re-identification by symmetry-driven accumulation of local features. In *CVPR*.

Felzenszwalb, P. F.; Girshick, R. B.; Mcallester, D.; and Ramanan, D. 2009. Object detection with discriminatively trained part based models. *TPAMI*, 32(9): 1627–1645.

Fu, Y.; Wang, X.; Wei, Y.; and Huang, T. 2019. Sta: Spatial-temporal attention for large-scale video-based person re-identification. In *AAAI*.

Ge, Y.; Li, Z.; Zhao, H.; Yin, G.; Yi, S.; Wang, X.; et al. 2018. Fd-gan: Pose-guided feature distilling gan for robust person re-identification. In *NeurIPS*.

Gilmer, J.; Schoenholz, S. S.; Riley, P. F.; Vinyals, O.; and Dahl, G. E. 2017. Neural Message Passing for Quantum Chemistry. In *ICML*.

Hamilton, W.; Ying, Z.; and Leskovec, J. 2017. Inductive representation learning on large graphs. In *NeurIPS*.

He, K.; Gkioxari, G.; Dollár, P.; and Girshick, R. 2017. Mask r-cnn. In *ICCV*.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *CVPR*.

Hou, R.; Chang, H.; Ma, B.; Shan, S.; and Chen, X. 2020. Temporal Complementary Learning for Video Person Re-Identification. In *ECCV*.

Hou, R.; Ma, B.; Chang, H.; Gu, X.; Shan, S.; and Chen, X. 2019. Vrstc: Occlusion-free video person re-identification. In *CVPR*.

Huang, W.; Zhang, T.; Rong, Y.; and Huang, J. 2018. Adaptive sampling towards fast graph representation learning. In *NeurIPS*.

Kipf, T. N.; and Welling, M. 2016. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*.

Kostinger, M.; Hirzer, M.; Wohlhart, P.; Roth, P. M.; and Bischof, H. 2012. Large scale metric learning from equivalence constraints. In *CVPR*.

Li, G.; Xiong, C.; Thabet, A.; and Ghanem, B. 2020. Deepergcn: All you need to train deeper gcns. *arXiv preprint arXiv:2006.07739*.

Li, J.; Wang, J.; Tian, Q.; Gao, W.; and Zhang, S. 2019. Global-local temporal representations for video person re-identification. In *ICCV*.

Li, J.; Zhang, S.; and Huang, T. 2019. Multi-scale 3d convolution network for video based person re-identification. In *AAAI*.

Li, N.; Zhao, X.; and Ma, C. 2020. A model-based Gait Recognition Method based on Gait Graph Convolutional Networks and Joints Relationship Pyramid Mapping. *arXiv preprint arXiv:2005.08625*.

Li, S.; Bak, S.; Carr, P.; and Wang, X. 2018. Diversity regularized spatiotemporal attention for video-based person re-identification. In *CVPR*.

Li, W.; Zhao, R.; Xiao, T.; and Wang, X. 2014. DeepReID: Deep filter pairing neural network for person re-identification. In *CVPR*.

Li, X.; Zheng, W. S.; Wang, X.; Xiang, T.; and Gong, S. 2015. Multi-scale learning for low-resolution person re-identification. In *ICCV*.

Liao, S.; Hu, Y.; Zhu, X.; and Li, S. Z. 2015. Person Re-Identification by Local Maximal Occurrence Representation and Metric Learning. In *CVPR*.

Liu, C.-T.; Wu, C.-W.; Wang, Y.-C. F.; and Chien, S.-Y. 2019a. Spatially and temporally efficient non-local attention network for video-based person re-identification. *arXiv preprint arXiv:1908.01683*.

Liu, H.; Feng, J.; Qi, M.; Jiang, J.; and Yan, S. 2017. End-to-end comparative attention networks for person re-identification. *TIP*, 26(7): 3492–3506.

Liu, J.; Ni, B.; Yan, Y.; Zhou, P.; Cheng, S.; and Hu, J. 2018. Pose transferrable person re-identification. In *CVPR*.

Liu, Y.; Yan, J.; and Ouyang, W. 2017. Quality aware network for set to set recognition. In *CVPR*.

Liu, Y.; Yuan, Z.; Zhou, W.; and Li, H. 2019b. Spatial and temporal mutual promotion for video-based person re-identification. In *AAAI*.

McLaughlin, N.; Del Rincon, J. M.; and Miller, P. 2016. Recurrent convolutional network for video-based person re-identification. In *CVPR*.

Miao, J.; Wu, Y.; Liu, P.; Ding, Y.; and Yang, Y. 2019. Pose-guided feature alignment for occluded person re-identification. In *ICCV*.

Rafi, U.; Doering, A.; Leibe, B.; and Gall, J. 2020. Self-supervised Keypoint Correspondences for Multi-Person Pose Estimation and Tracking in Videos. In *ECCV*.

Shen, Y.; Li, H.; Yi, S.; Chen, D.; and Wang, X. 2018. Person re-identification with deep similarity-guided graph neural network. In *ECCV*.

Si, J.; Zhang, H.; Li, C.-G.; Kuen, J.; Kong, X.; Kot, A. C.; and Wang, G. 2018. Dual attention matching network for context-aware feature sequence based person re-identification. In *CVPR*.

Su, C.; Li, J.; Zhang, S.; Xing, J.; Gao, W.; and Tian, Q. 2017. Pose-driven deep convolutional model for person re-identification. In *ICCV*.

Suh, Y.; Wang, J.; Tang, S.; Mei, T.; and Mu Lee, K. 2018. Part-aligned bilinear representations for person re-identification. In *ECCV*.

Sun, K.; Xiao, B.; Liu, D.; and Wang, J. 2019. Deep High-Resolution Representation Learning for Human Pose Estimation. In *CVPR*.

Sun, Y.; Zheng, L.; Yang, Y.; Tian, Q.; and Wang, S. 2018. Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline). In *ECCV*.

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

Wang, X.; Doretto, G.; Sebastian, T.; Rittscher, J.; and Tu, P. 2007. Shape and appearance context modeling. In *ICCV*.

Wang, X.; and Gupta, A. 2018. Videos as space-time region graphs. In *ECCV*.

Wei, L.; Zhang, S.; Yao, H.; Gao, W.; and Tian, Q. 2017. Glad: Global-local-alignment descriptor for pedestrian retrieval. In *ACM'MM*.

Wu, Y.; Bourahla, O. E. F.; Li, X.; Wu, F.; Tian, Q.; and Zhou, X. 2020. Adaptive graph representation learning for video person re-identification. *TIP*, 29: 8821–8830.

Xiang, W.; Huang, J.; Qi, X.; Hua, X.-S.; and Zhang, L. 2018. Homocentric Hypersphere Feature Embedding for Person Re-identification. *arXiv preprint arXiv:1804.08866*.

Xiao, B.; Wu, H.; and Wei, Y. 2018. Simple baselines for human pose estimation and tracking. In *ECCV*.

Xiao, T.; Li, H.; Ouyang, W.; and Wang, X. 2016. Learning Deep Feature Representations with Domain Guided Dropout for Person Re-identification. In *CVPR*.

Xu, J.; Zhao, R.; Zhu, F.; Wang, H.; and Ouyang, W. 2018. Attention-Aware Compositional Network for Person Re-identification. In *CVPR*.

Xu, S.; Cheng, Y.; Gu, K.; Yang, Y.; Chang, S.; and Zhou, P. 2017. Jointly attentive spatial-temporal pooling networks for video-based person re-identification. In *ICCV*.

Yan, S.; Xiong, Y.; and Lin, D. 2018. Spatial Temporal Graph Convolutional Networks for Skeleton-Based Action Recognition. In *AAAI*.

Yan, Y.; Qin, J.; Chen, J.; Liu, L.; Zhu, F.; Tai, Y.; and Shao, L. 2020. Learning Multi-Granular Hypergraphs for Video-Based Person Re-Identification. In *CVPR*.

Yang, J.; Zheng, W.-S.; Yang, Q.; Chen, Y.-C.; and Tian, Q. 2020. Spatial-Temporal Graph Convolutional Network for Video-Based Person Re-Identification. In *CVPR*.

Yao, H.; Zhang, S.; Hong, R.; Zhang, Y.; Xu, C.; and Tian, Q. 2019. Deep representation learning with part loss for person re-identification. *TIP*, 28(6): 2860–2871.

Yi, D.; Lei, Z.; Liao, S.; and Li, S. Z. 2014. Deep metric learning for person re-identification. In *ICPR*.

Zhang, L.; Xiang, T.; and Gong, S. 2016. Learning a Discriminative Null Space for Person Re-Identification. In *CVPR*.

Zhang, Z.; Lan, C.; Zeng, W.; and Chen, Z. 2020. Multi-Granularity Reference-Aided Attentive Feature Aggregation for Video-based Person Re-identification. In *CVPR*.

Zhao, L.; Li, X.; Zhuang, Y.; and Wang, J. 2017. Deeply-learned part-aligned representations for person re-identification. In *ICCV*.

Zhao, R.; Ouyang, W.; and Wang, X. 2013. Unsupervised salience learning for person re-identification. In *CVPR*.

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

Zheng, L.; Bie, Z.; Sun, Y.; Wang, J.; Su, C.; Wang, S.; and Tian, Q. 2016a. Mars: A video benchmark for large-scale person re-identification. In *ECCV*.

Zheng, L.; Bie, Z.; Sun, Y.; Wang, J.; Su, C.; Wang, S.; and Tian, Q. 2016b. MARS: A Video Benchmark for Large-Scale Person Re-Identification. In *ECCV*.

Zheng, Z.; Zheng, L.; and Yang, Y. 2017. A discriminatively learned cnn embedding for person reidentification. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 14(1): 1–20.

Zhou, Z.; Huang, Y.; Wang, W.; Wang, L.; and Tan, T. 2017. See the forest for the trees: Joint spatial and temporal recurrent neural networks for video-based person re-identification. In *CVPR*.