Learning Instance-level Spatial-Temporal Patterns for Person Re-identification

Min Ren[1 2]*, Lingxiao He[3], Xingyu Liao[3],Wu Liu[3], Yunlong Wang[2], Tieniu Tan[2]
[1]University of Chinese Academy of Sciences
[2]CRIPAC NLPR, Institute of Automation Chinese Academy of Sciences
[3]JD AI Research
{min.ren, yunlong.wang}@cripac.ia.ac.cn, {helingxiao3, liaoxingyu5, liuwu1}@jd.com, tnt@nlpr.ia.ac.cn

Abstract

Person re-identification (Re-ID) aims to match pedestrians under dis-joint cameras. Most Re-ID methods formulate it as visual representation learning and image search, and its accuracy is consequently affected greatly by the search space. Spatial-temporal information has been proven to be efficient to filter irrelevant negative samples and significantly improve Re-ID accuracy. However, existing spatial-temporal person Re-ID methods are still rough and do not exploit spatial-temporal information sufficiently. In this paper, we propose a novel Instance-level and Spatial-Temporal Disentangled Re-ID method (InSTD), to improve Re-ID accuracy. In our proposed framework, personalized information such as moving direction is explicitly considered to further narrow down the search space. Besides, the spatial-temporal transferring probability is disentangled from joint distribution to marginal distribution, so that outliers can also be well modeled. Abundant experimental analyses are presented, which demonstrates the superiority and provides more insights into our method. The proposed method achieves mAP of 90.8% on Market-1501 and 89.1% on DukeMTMC-reID, improving from the baseline 82.2% and 72.7%, respectively. Besides, in order to provide a better benchmark for person re-identification, we release a cleaned data list of DukeMTMC-reID with this paper: https://github.com/RenMin1991/cleaned-DukeMTMC-reID/

1. Introduction

Person re-identification aims to retrieve pedestrians across non-overlapping camera views. Most existing person re-identification methods focus on the visual feature representations of pedestrian images [6, 11, 14, 27, 32, 35, 38, 36, 42], such as appearance, clothes, and textures. The auxiliary information of person images is also adopted recently, such as parsing information [5, 13, 15, 22, 25], pose of the pedestrians [3, 19, 20], or human body key points [33]. However, the performances of these methods are still far from the requirements of real-world situations. Because it is hard for visual representations to discriminate pedestrian with similar appearance and clothes.

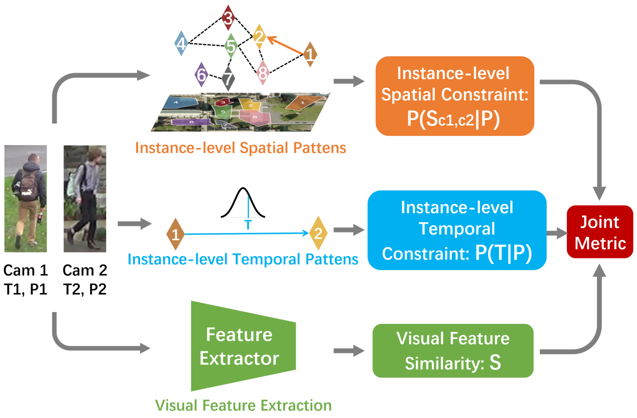
3
4
2
5
1
6
7
8
Instance-level Spatial Pattens
Instance-level Spatial Constraint: P(Sc1,c2|P)
Cam 1
T1, P1
Cam 2
T2, P2
1
T
2
Instance-level Temporal Pattens
Instance-level Temporal Constraint: P(T|P)
Joint Metric
Feature Extractor
Visual Feature Similarity: S
Visual Feature Extraction

Figure 1. For each pair of pedestrian images, instance-level spatial and temporal constraints are provided separately by the proposed framework. Then they are adaptively combined with the visual feature similarity for matching.

Recent methods model spatial-temporal patterns [8, 18, 21, 34] to filter out the irrelevant candidates and narrow down the search space. Specifically, these methods mainly formulate spatial-temporal pattern as a joint distribution $P(S_{c_i,c_j}, T)$, where S_{c_i,c_j} means moving from *camera i* to *camera j*, T means time interval. It has been proven to be efficient to significantly improve re-identification accuracy. However, there are two problems of the existing methods. Firstly, the existing spatial-temporal methods only consider camera-level but neglect instance-level information. The state information of each pedestrian is neglected while it is essential for spatial-temporal patterns of the person. Secondly, existing methods formulate spatial-temporal patterns as a joint distribution, meaning that only those candidates

*This work is done when Min Ren is an intern at JD AI Research.

matching both spatial and temporal priors can be matched. They are not robust to the outliers.

To solve these problems, we propose a novel method named Instance-level and Spatial-Temporal Disentangled Re-ID (InSTD) to model the instance-level and spatial-temporal disentangled patterns. Firstly, the traditional spatial-temporal pattern is updated to be conditional on instance-level state information. Its formulation looks like $p(S_{c_i,c_j}, T|P)$, where P is instance-level pedestrian information. The walking direction of the pedestrian, which is the key instance-level state information, is taken into consideration in this paper. The walking direction of a pedestrian is complimentary information of pedestrian detection and tracking. It is useful because it is highly correlated with spatial-temporal patterns. For example, a pedestrian, who is walking towards the west in the view of a camera, is more probable to appear in the view of the western cameras later, rather than the eastern cameras. Meanwhile, it is economical because pedestrian detection and tracking are necessary steps before person re-identification in practice.

Secondly, we disentangle the spatial-temporal pattern by constructing their marginal distribution, *i.e.* transmission probability $P(S_{c_i,c_j}|P)$ and time interval distribution $P(T|P)$. They are modeled separately and adaptively combined to handle outliers. If the temporal (spatial) pattern of a pedestrian is unusual, the person may be normal in the term of spatial (temporal) patterns. The similarity metric should focus on the spatial (temporal) pattern. For example, a runner, who is moving faster than most pedestrians, is an outlier from the view of temporal pattern. But the runner can be quite normal in terms of spatial transmission perspective. It is harmful to model this runner by joint distribution of spatial and temporal patterns. To this end, we propose a novel fusion approach to adaptively combine the spatial and temporal patterns. The spatial patterns and temporal patterns are complementary, rather than in conflict as existing methods, so that outliers can also be well modeled.

The contributions of this paper can be summarized as follows:

- We present a novel instance-level method to model spatial-temporal patterns for person re-identification. The proposed method provides personalized predictions by leveraging the instance-level state information of each pedestrian.
- The instance-level spatial-temporal patterns are decoupled into transmission probabilities and time interval distributions between cameras in the proposed method. The spatial and temporal patterns become complementary rather than in conflict as existing methods.
- Without bells and whistles, the proposed method surpasses the baseline model based on visual features by 16.9% on DukeMTMC-reID and 8.6% on Market-1501 in the term of mAP, and outperforms the state-of-the-art method based on spatial-temporal patterns by 4.8% on DukeMTMC-reID and 2.2% on Market-1501.

2. Related Work

2.1. Visual Features based Re-ID

Person re-identification addresses the problem of matching pedestrian images across non-overlapping camera views [29, 40]. Many studies exploit discriminative visual features [2, 26, 37].

Deep learning algorithms foster significant improvements in the field of person re-identification. Some researchers attempt to explore effective convolutional neural networks [1, 4, 6, 9, 16, 17, 10, 30, 31, 35, 36, 41, 44]. Some studies explore training strategies and loss functions for person re-identification [1, 9, 10, 31, 43]. Recently, some studies leverage the structure information of person images, such as parsing information [5, 15, 22, 25], pose of the pedestrians [3, 19, 20, 24], or human body key points [33].

However, appearance-based methods are still far from practical applications. They are not discriminative enough in complex scenarios where pedestrians may exhibit similar appearance and clothes. It is hard to further improve the performance using only appearance-based features.

2.2. Spatial-temporal Person Re-ID

There are some researchers who have paid attention to the topology of cameras since the spatial-temporal patterns implied in the topology are essential for cross-camera retrieval. The spatial-temporal constraints are utilized to filter out the irrelevant gallery images [8, 18, 21, 34]. Huang *et al.* [34] propose a method to take both visual feature representation and spatial-temporal constraints into consideration for person re-identification. However, this method makes a strong assumption that the time intervals between cameras follow Weibull distribution. This assumption is invalid for complex scenarios. Cho *et al.* [21] propose a framework to integrate camera network topology into person re-identification. However, the temporal constraints are simply realized by time thresholds, which cannot handle massive gallery images and complex cases in practice. Lv *et al.* [18] propose a method that leverage the spatial-temporal constraints for cross-dataset person re-identification. The spatial-temporal constraints improve the performance on the target dataset by enhancing the pseudo label during training. It is not proper to be directly applied to general person re-identification tasks. Wang *et al.* [8] propose a two-stream architecture to apply spatial-temporal constraints to person re-identification. However, the spatial-temporal constraints are coupled together in this method,

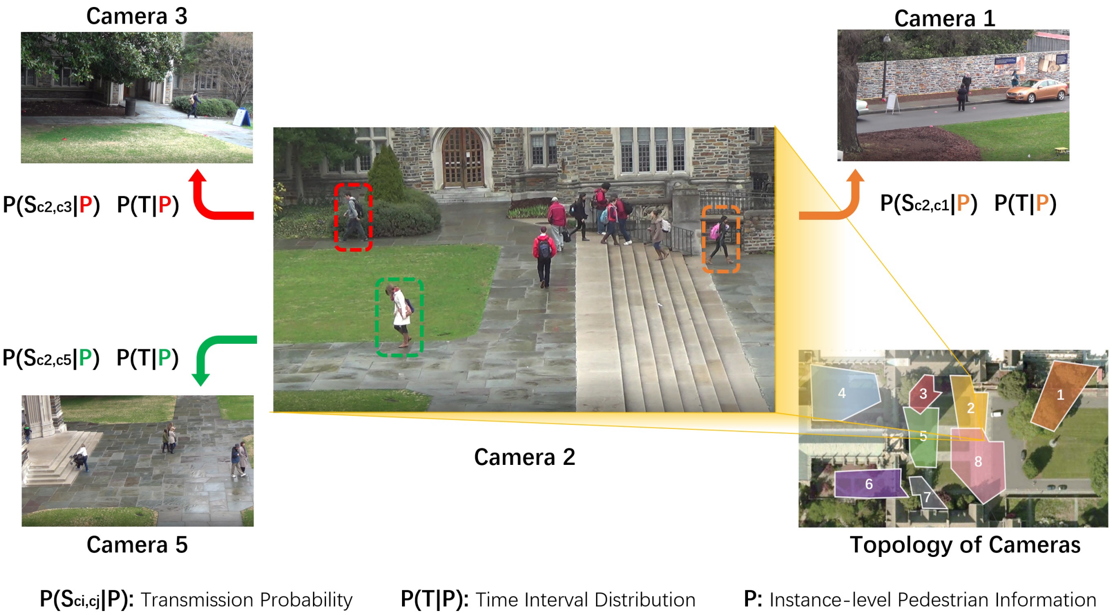

Figure 2. Spatial-temporal patterns are implied in the topology of cameras. The spatial-temporal pattern between two cameras of a pedestrian are highly correlated with his/her moving direction. Pedestrians in the view of camera 2 may appear in camera 1, camera 3, or camera 5 after a certain time lapse. However, pedestrians with different states will appear in different cameras at different times. For example, the pedestrian in the red bounding box is much more likely to appear in Camera 3 than Camera 1, because he is moving towards the field of view of Camera 3. In the proposed method, the instance-level state information is adopted rather than modeling the spatial-temporal patterns on camera-level as the existing methods.

which is harmful to recalling positive samples. All these spatial-temporal person Re-ID methods establish the patterns based on the camera-level information, which means they can not provide fine-grained constraints for each instance.

Different from existing spatial-temporal methods, our method models spatial-temporal patterns at the instance-level to filter out more irrelevant gallery images and provide personalized predictions. And the spatial-temporal patterns are decoupled into transmission probabilities and time interval distributions to make them mutually beneficial rather than in conflict.

3. Method

The instance-level spatial constraint *i.e.* transmission probabilities, and temporal constraint *i.e.* time interval distributions are detailed separately in this section. Then the adaptive combined metric is presented.

3.1. Instance-level Spatial Constraint

The spatial constraint between two cameras is described by the transmission probability of the cameras, which means how tightly the two cameras correlate. Formally, we model the transmission probability by a conditional probability:

$$p_{i,j} = Pr(C_l = j | C_e = i) \tag{1}$$

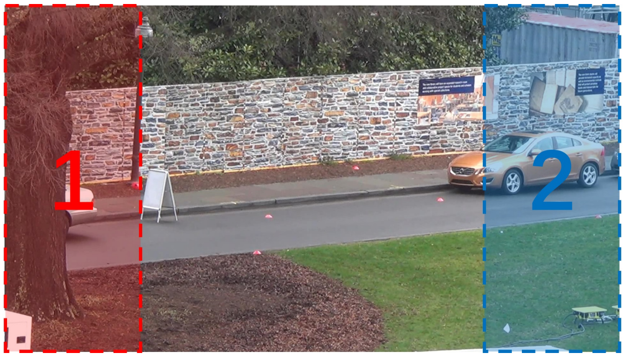

Figure 3. View of the first camera of DukeMTMC-reID. The state set of this camera contains two states: walking towards the red zone and walking towards the blue zone.

where i and j are the indexes of cameras, C_e is the camera that a person appears earlier, C_l is the camera that the same person appears later. It is the probability that a person appears in the view of camera j later on the condition that this person has appeared in the view of camera i. The conditional probability in Eq. 1 can be easily calculated:

$$p_{i,j} = \frac{Pr(C_l = j, C_e = i)}{Pr(C_e = i)} \tag{2}$$

The higher the conditional probability means the person in the view of camera i is more likely to appear in the view of camera j later. The time interval between camera i and j is not involved here. Note that $p_{i,j} \neq p_{j,i}$ in most cases.

However, the spatial patterns of persons appear in the same camera can be different, as shown in Fig. 2 . To address this problem, we introduce the instance-level state information of a person into the conditional probability:

$$p_{i,j}^s = Pr(C_l = j | C_e = i, s_e = s) \tag{3}$$

where s_e is the state of a person in the view of C_e, $s \in S_i$, S_i is the set of states:

$$S_i = \{s_1, s_2, ..., s_{n_i}\} \tag{4}$$

where n_i is the number of states of camera i.

The instance-level states are represented by walking directions of pedestrians. For example, the view of the first camera of DukeMTMC-reID [7] is shown in Fig. 3. The state set of this camera contains two states: walking towards the red zone and walking towards the blue zone. The state sets of the rest cameras are defined similarly, and the illustrations of other cameras can be found in the supplementary material.

Hence, the instance-level transmission probability can be calculated:

$$p_{i,j}^s = \frac{Pr(C_l = j, C_e = i, s_e = s)}{Pr(C_e = i, s_e = s)} \tag{5}$$

3.2. Instance-level Temporal Constraint

The temporal constraint is described by the time interval distribution, which represents the time lapse for a pedestrian to transfer between two cameras. Formally, we model the time interval distribution by a conditional probability density function:

$$f_{i,j}(\delta) = \frac{\mathrm{d}F_{i,j}(\delta)}{\mathrm{d}\delta} \tag{6}$$

where δ is the transfer time, $F_{i,j}(\delta)$ is the cumulative distribution function, which is a conditional probability:

$$F_{i,j}(\delta) = Pr(\Delta \leq \delta | C_e = i, C_l = j) \tag{7}$$

It can be harmful to recalling positive samples that fitting $f_{i,j}(\delta)$ or $F_{i,j}(\delta)$ into a closed-form probability distribution. Hence, a non-parameter estimation method is adopted in our method. Specially, we use Parzen window with Gaussian kernel to estimate $f_{i,j}(\delta)$:

$$f_{i,j}(\delta) = \frac{1}{Z_{i,j}} \sum_n H_{i,j}(\delta) K(n - \delta) \tag{8}$$

$$H_{i,j}(\delta) = \begin{cases} 1 & \delta \in \mathcal{D}_{i,j} \\ 0 & otherwise \end{cases} \tag{9}$$

where $K(\cdot)$ is the kernel function, $Z_{i,j} = \sum_\delta H_{i,j}(\delta)$ is a normalized factor, $\mathcal{D}_{i,j}$ is the time interval set between camera i and camera j of training samples.

However, the time interval distribution of persons appear in the same camera can be different, as we have mentioned before. Similar to the instance-level transmission probabilities, the instance-level state information of a person is introduced into the conditional probability to address this problem:

$$f_{i,j}^s(\delta) = \frac{\mathrm{d}F_{i,j}^s(\delta)}{\mathrm{d}\delta} \tag{10}$$

$$F_{i,j}^s(\delta) = Pr(\Delta \leq \delta | C_e = i, C_l = j, s_e = s) \tag{11}$$

where s_e is the instance-level state of a person in the view of C_e. The moving direction is also considered as the key state information. Similar with Eq. 8 , instance-level time interval distribution $f_{i,j}^s(\delta)$ can be estimated:

$$f_{i,j}^s(\delta) = \frac{1}{Z_{i,j}^s} \sum_n H_{i,j}^s(\delta) K(n - \delta) \tag{12}$$

$$H_{i,j}^s(\delta) = \begin{cases} 1 & \delta \in \mathcal{D}_{i,j}^s \\ 0 & otherwise \end{cases} \tag{13}$$

where the normalized factor $Z_{i,j}^s = \sum_\delta H_{i,j}^s(\delta)$, $\mathcal{D}_{i,j}^s$ is a subset of $\mathcal{D}_{i,j}$, it contains the samples subject to $s_e = s$.

3.3. Joint Metric

Given the transmission probability and time interval distribution affiliated to instance-level state information, the spatial-temporal probability is the fusion of them:

$$\mathcal{P} = \mathcal{F}(p_{spa}, p_{tem}) \tag{14}$$

where $p_{spa} = p_{i,j}^s$ is the instance-level transmission probability of two images in Eq. 5, and $p_{tem} = f_{i,j}^s(\delta)$ is the instance-level time interval probability in Eq. 10. And the final joint metric of two images is:

$$\mathcal{S} \cdot \mathcal{P} = \mathcal{S} \cdot \mathcal{F}(p_{spa}, p_{tem}) \tag{15}$$

where $\mathcal{S}$ is the visual feature similarity.

A straightforward way to fuse both components is multiplying p_{spa} and p_{tem} together. However, the constraint realized by directly multiplying is too strict for person re-identification. If a spatial/temporal pattern of a pedestrian is unusual, the person may be normal in terms of temporal/spatial patterns. This kind of samples should not be removed recklessly. Hence, the spatial-temporal probability $\mathcal{F}(p_{spa}, p_{tem})$ should be fairly high when only one of them is high. Fusion by multiplying directly is not proper here obviously.

The spatial-temporal probability in our method is defined as:

$$\mathcal{P} = \frac{1}{1 + e^{-(\alpha p_{spa} + \beta p_{tem})}} \tag{16}$$

where α and β are scaling parameters of similarity fusion. The spatial and temporal constraints are adjusted by α and β separately.

The spatial-temporal factor is scaled into $[0.5, 1)$. The constraint is relaxed properly when the spatial-temporal probability is low. And the value of $\mathcal{P}$ stays stable when p_{spa} or p_{tem} is low.

3.4. Implementation Details

More details are presented in this subsection. The moving direction of a pedestrian is complimentary information of pedestrian detection and tracking, which is a necessary step before person re-identification in practice. There is no need to predict the moving direction by an extra model or manual annotations. In our experiments, the moving directions of samples in the field of one camera are confirmed by tracking them in the original video of this camera. Actually, the moving direction of a pedestrian can be confirmed within five consecutive frames in most cases, which can be easily derived from existing tracking methods.

A pretrained ResNet-50 is adopted as baseline for feature extraction. We set the standard deviation of the Gaussian kernel for distribution estimation to 100. As for the scaling parameters, α, β in Eq. 16 are set to 0.15 and 1 respectively.

4. Experiments

In this section, we evaluate our method on two large scale person re-identification benchmark datasets, *i.e.* Market-1501 [23] and DukeMTMC-reID [7]. Then, more experimental analysis is presented.

4.1. Datasets and Evaluation Protocol

Market-1501 dataset [23] is collected at a university campus. A total of six cameras are used, including 5 high-resolution cameras, and one low-resolution camera. It contains 32,668 annotated bounding boxes of 1,501 identifies, plus a distractor set of over 500K images. The pedestrians are detected by Deformable Part Model (DPM). Among them, the training set consists of 12,936 images from 751 identities, the gallery set contains 19,732 images from other 750 identities and all the distractors. 3,368 hand-drawn bounding boxes from 750 identities are used as the query images. In this dataset, each image contains its camera index and time stamp.

DukeMTMC-reID [7] is a subset of DukeMTMC dataset for image-based person re-identification. There are eight cameras in total. 1,404 identities appear in more than one camera and 408 identities (distractor) appear in only one camera. 702 identities are used for training, and the other 702 identities plus distractors are used for testing. One image for each identity in each camera is picked as a query, and the other images are put into the gallery. Each image contains its camera index and time stamp.

We use two performance indexes as in most person re-identification literature. The first is mean average precision (mAP). The average precision (AP) of a query is the area under the Precision-Recall curve, which means both precision and recall rate is taken into consideration. Hence, the mean average precision among all query images is a comprehensive performance index for person re-identification. The second is the cumulative matching characteristic (CMC) *i.e.* the top-k accuracy. Hence, the cumulative matching characteristic emphasizes precision rather than recall rate.

Methods	**mAP**	**Rank-1**	**Rank-5**	**Rank-10**
PCB [38]	77.4%	92.3%	97.2%	98.2%
VPM [39]	80.8%	93.0%	97.8%	98.8%
BOT [12]	85.9%	94.5%	-	-
SPReID [25]	81.3%	92.5%	97.15%	98.1%
MGCAM [5]	74.3%	83.8%	-	-
MaskReID [22]	75.4%	90.4%	-	-
FPR [15]	86.6%	95.4%	-	-
PDC [3]	63.4%	84.1%	-	-
Pose-transfer [20]	68.9%	87.7%	-	-
PSE [28]	69.0%	87.7%	94.5%	96.8%
PGFA [19]	76.8%	91.2%	-	-
HOReID [33]	84.9%	94.2%	-	-
Baseline	82.2%	93.6%	98.4%	99.0%
Baseline+st-ReID [8]	88.6%	96.9%	99.2%	99.5%
Baseline+InSTD	**90.8%**	**97.6%**	**99.5%**	**99.7%**

Table 1. Comparison with the state-of-the-arts methods on Market-1501. Group 1: vanilla deep learning based methods. Group 2: human-parsing information based methods. Group 3: pose or key points based methods. Group 4: spatial-temporal methods.

4.2. Comparisons to the State-of-the-Art

The proposed method is compared with fourteen existing state-of-the-art methods, which can be categorized into four groups. The first group of methods extract visual features directly from the person images, including PCB [38], VPM [39], and BOT [12]. These methods explore various aspects of visual feature extraction, including the structure of convolutional neural networks, training strategy, data augmentation, and loss function. The second group of methods adopt human parsing information for person re-identification, including SPReID [25], MGCAM [5], MaskReID [22] and FPR [15]. The third group of methods leverage the pose or key points of person images, including PDC [3], Pose-transfer [20], PSE [28], PGFA [19] and HOReID [33]. The fourth group of methods utilize the spatial-temporal information to enhance the person re-identification, including TFusion-sup [18] and st-ReID [8]. These methods use hard or soft constraints to narrow the number of gallery images.

The experiment results on Market-1501 are shown in Tab. 1, and the results on DukeMTMC-reID are shown in Tab. 2. Our method outperforms all of the existing methods on both datasets. Comparing to the baseline model,

Methods	mAP	Rank-1	Rank-5	Rank-10
PCB [38]	66.1%	81.7%	89.7%	91.9%
VPM [39]	72.6%	83.6%	91.7%	94.2%
BOT [12]	76.4%	86.4%	-	-
SPReID [25]	70.9%	84.4%	91.8%	93.7%
MGCAM [5]	46.0%	46.7%	-	-
MaskReID [22]	61.89%	78.86%	-	-
FPR [15]	78.4%	88.6%	-	-
Pose-transfer [20]	56.9%	78.5%	-	-
PSE [28]	62.0%	79.8%	89.7%	92.2%
PGFA [19]	65.5%	82.6%	-	-
HOReID [33]	75.6%	86.9%	-	-
Baseline	72.7%	85.7%	90.9%	93.5%
Baseline+st-ReID [8]	84.3%	94.1%	96.3%	97.2%
Baseline+InSTD	**89.1%**	**95.7%**	**97.2%**	**98.0%**

Table 2. Comparison with state-of-the-arts for person re-identification on DukeMTMC-reID [7]. Group 1: vanilla deep learning based methods. Group 2: human-parsing information based methods. Group 3: pose or key points based methods. Group 4: spatial-temporal methods.

which is a ResNet-50, our method improves the mAP by 8.6% on Market-1501 and 16.5% on DukeMTMC-reID, improve the Rank-1 accuracy by 4% on Market-1501 and 10% on DukeMTMC-reID. Our method achieves significant improvements especially in terms of mAP.

Comparing to the methods in the first three groups, the advantage of our method is obvious. Besides, the compared methods in the second and third groups need expensive annotations, such as key points, pixel-wise parsing maps, and masks, to match the query and gallery images. Our method adopts economical information *i.e.* camera ID, timestamp, and state information.

The disadvantages of the methods in the fourth group have been interpreted in Sec. 2. And the interpretations have been demonstrated by the results of experiments in this subsection. Given the same baseline model, our method outperforms the st-ReID by a remarkable margin, especially in terms of mAP (2.2% on Market-1501 and 4.8% on DukeMTMC-reID). These results indicate that our method has evident advantages over existing spatial-temporal methods.

To show the effect of spatial-temporal constraint, an example from DukeMTMC-reID is presented in Fig. 4. The appearance of the pedestrian in the red bounding box, who is mistakenly ranked first, is similar to the query image. The visual representation cannot distinguish it from the correct identifies as shown in Fig. 4 (a). The incorrect pedestrian, which is difficult to discriminate for the visual representation, is filtered out by the spatial-temporal constraint as shown in Fig. 4 (b).

The effect of instance-level information are shown in Fig. 5. The spatial-temporal constrains may be misguided in complex scenarios, as shown in Fig. 5 (a). The instance-level information can make the spatial-temporal constrains more reliable for person re-identification. The incorrect pedestrians are filtered out by the instance-level state as shown in Fig. 5 (b).

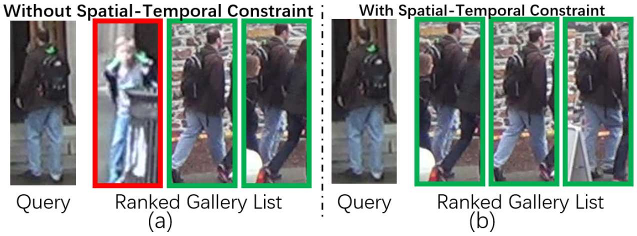
Without Spatial-Temporal Constraint
With Spatial-Temporal Constraint
Query
Ranked Gallery List
(a)
Query
Ranked Gallery List
(b)

Figure 4. **(a)**: The appearance of the pedestrian in the red bounding box, who is mistakenly ranked first, is similar to the query image. **(b)**: The incorrect pedestrian is filtered out by the spatial-temporal constraint.

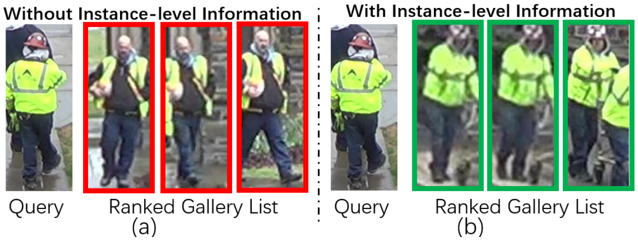
Without Instance-level Information
With Instance-level Information
Query
Ranked Gallery List
(a)
Query
Ranked Gallery List
(b)

Figure 5. **(a)**: The top three of the ranked list are wrong samples because the spatial-temporal constraints are misguided without instance-level information. **(b)**: The spatial-temporal constraints are more reliable because of the instance-level state.

4.3. Experiments on Different Feature Extractors

The proposed method can be applied to different feature extractors. To verify its effectiveness, we evaluate the proposed method based on other two feature extractors: PCB [38], and VPM [39].

The results are shown in Tab. 3. Our method consistently improves the performance of all feature extractors. Our method gains significant 20%/11% improvement in mAP/rank-1 accuracy for PCB [38], and 16%+/10%+ improvement for the other two feature extractors. Comparing to st-ReID [8], which is also based on spatial-temporal constraints, our method achieves consistent improvements too. Our method outperforms st-ReID [8] by 4%+/0.6%+ improvement in mAP/rank-1 accuracy for all of the feature extractors.

The results show that our method can be generalized to different feature extractors. Moreover, the results demonstrate the advantages of our method comparing to the existing spatial-temporal based method.

4.4. Analysis of Scaling Parameters

To investigate the impact of two scaling parameters, α and β in Eq. 16, we conduct two sensitivity analysis experiments on α and β. The results are shown in Fig. 6. When

Methods	mAP	Rank-1	Rank-5	Rank-10
PCB [38]	66.1%	81.7%	89.7%	91.9%
PCB [38]+st-ReID [8]	80.9%	92.1%	95.4%	96.6%
PCB [38]+InSTD	86.1%	92.7%	96.5%	97.6%
VPM [39]	72.6%	83.6%	91.7%	94.2%
VPM [39]+st-ReID [8]	84.9%	94.2%	96.1%	96.9%
VPM [39]+InSTD	89.3%	95.1%	97.0%	97.9%

Table 3. Effects on different feature extractors. The experiments are conducted on DukeMTMC-reID [7]

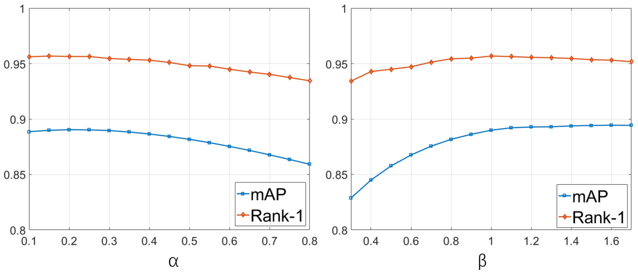

Figure 6. Result of sensitivity analysis experiments on α and β in Eq. 16. When analyzing one of them, the other one is fixed as its optimal value. The experimental results show that our method is is insensitive to fusion parameters.

analyzing one of them, the other one is fixed as its optimal value: $\alpha = 0.15$, $\beta = 1$. As we can observe, our method nearly keeps the best performance when α is in the range of 0.1 to 0.3 or β is in the range of 1 to 1.7. The results show that our method is insensitive to fusion parameters.

4.5. Ablation Study

The ablation study on the instance-level state information of pedestrians and the decoupling of spatial-temporal patterns is presented in this part.

Four protocols are taken into consideration. The first one is the proposed method itself. In the second protocol, p_{spa} and p_{tem} in Eq. 14 are replaced by $p_{i,j}$ (Eq. 2) and $f_{i,j}(\delta)$ (Eq. 8). The instance-level state information is excluded in this protocol.

In the third protocol, the spatial pattern and temporal pattern are coupled together. The normalized factor in Eq. 12 is replaced by $\hat{Z}$:

$$\hat{f}^s_{i,j}(\delta) = \frac{1}{\hat{Z}^s} \sum_n H^s_{i,j}(\delta) K(n-\delta) \tag{17}$$

$$\hat{Z}^s = \max_{i,j} Z^s_{i,j} \tag{18}$$

which means all time interval distributions share an identical denominator. The numerical relations of the area under curves indicate the transmission probabilities between cameras. And p_{spa} and p_{spa} are replaced by p_{st} Eq. 16:

$$\mathcal{P} = \frac{1}{1 + e^{\beta p_{st}}}, \tag{19}$$

Protocol	Instance Info.	ST Decouple	mAP	Rank-1
1	✓	✓	**89.1%**	**95.7%**
2	×	✓	87.1%	94.3%
3	✓	×	86.9%	95.0%
4	×	×	83.4%	93.8%

Table 4. Ablation Study results on DukeMTMC-reID [7].

$$p_{st} = \hat{f}^s_{i,j}(\delta) \tag{20}$$

In the fourth protocol, the instance-level state information is excluded based on the third protocol:

$$\hat{f}_{i,j}(\delta) = \frac{1}{\hat{Z}} \sum_n H_{i,j}(\delta) K(n-\delta) \tag{21}$$

$$\hat{Z} = \max_{i,j} Z_{i,j} \tag{22}$$

The results of these four protocols are shown in Tab. 4. The results show that the instance-level state information of pedestrians and the decoupling of spatial and temporal are both useful to improve the performance.

Besides, the instance-level state information of pedestrians is more helpful to Rank-1 accuracy, and the decoupling of spatial and temporal patterns is more contributive to mAP. These results indicate that the decoupling of spatial and temporal patterns is more helpful to improve mAP by recalling more hard positive samples. The instance-level state information of pedestrians is more helpful to improve precision by narrow the number of gallery images. The combination of these two strategies achieves the best performance.

To demonstrate the effect of introducing instance-level state information, the time interval distributions between *camera 1* and *camera 2* of DukeMTMC-reID are shown in Fig. 7. The instance-level states are not taken into consideration in Fig. 7 (a). On the other hand, distributions of two states are shown separately in Fig. 7 (b). The distribution in Fig. 7 (a) is split into two distributions, which means more irrelevant gallery images can be filtered out according to the instance-level state information.

To show the difference between spatial-temporal coupled constraint and spatial-temporal decoupled constraint, time interval distributions of *camera 1* to *camera 2* and *camera 1* to *camera 5* are shown in Fig. 8. In the coupled case, the spatial pattern is conveyed by the areas under distribution curves as shown in Fig. 8 (b). In the decoupled case, the areas under distribution curves are the same as shown in Fig. 8 (c), and the spatial pattern is decoupled from the time interval distributions as transmission probabilities.

4.6. Failure Analysis

In this part, we analyze the failure cases of the proposed method on DukeMTMC-reID. We find that the failures can be categorized as four cases:

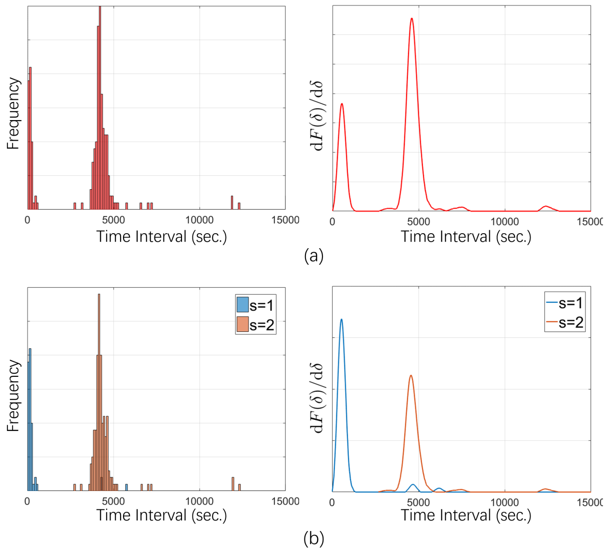

Figure 7. The time interval distributions between *camera 1* and *camera 2* of DukeMTMC-reID. **(a)**: Time interval distribution without state information. **(b)**: Time interval distributions with instance-level state information. The distribution in **(a)** is split into two distributions in **(b)**, which means more irrelevant gallery images can be filtered out according to the instance-level state information.

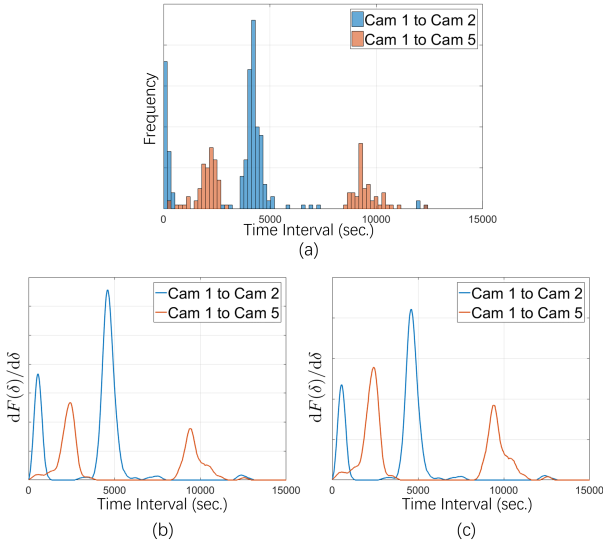

Figure 8. The time interval distributions of *camera 1* to *camera 2* and *camera 1* to *camera 5* of DukeMTMC-reID. **(a)**: Time interval frequencies. **(b)**: Time interval distributions without spatial-temporal decouple. The spatial pattern is conveyed by the areas under distribution curves. **(c)**: Time interval distributions with spatial-temporal decouple. The areas under distribution curves are the same. The spatial pattern is decoupled from the time interval distributions as transmission probabilities. (Instance-level state information is not shown for simplicity.)

Firstly, there are incorrect labels in DukeMTMC-reID. The proportion of failure cases caused by incorrect labels is 16.2%. It is harmful to keep the incorrect labels in the database. Hence, we release a cleaned data list of DukeMTMC-reID with this paper: `https://github.com/RenMin1991/cleaned-DukeMTMC-reID/`.

Secondly, the feature extractor is fooled because of serious occlusions. For example, the upper part of two individuals is quite similar while the lower part is occluded. The proportion of this case in all failures is 56.9%.

In the third case, the visual feature is not discriminative enough to distinguish the hard negative samples. The proportion of this case in all failures is 23.5%.

In the last case, the proposed method outputs high probabilities due to spatial-temporal patterns. However, it improperly pushes up the final joint metric. The proportion of this case in all failures is 3.4%.

The failure analysis shows that serious occlusion is the main cause of mismatching (56.9%). The second important reason is that the feature extracted by the recognition model is not discriminative enough (23.5%). Incorrect labels also degrade the performance (16.2%). The proportion of failure samples caused by the improper spatial-temporal probability in all failures is quite small (3.4%).

5. Conclusion

In this paper, we propose a method to exploit spatial-temporal patterns for person re-identification. Different from the existing spatial-temporal person re-identification methods, the proposed method adopts the walking direction of each pedestrian, as key instance-level state information, to provide personalized predictions. In addition, the spatial-temporal patterns are decoupled into transmission probabilities and time interval distributions between cameras. The spatial-temporal patterns become mutually beneficial rather than in conflict with each other as current methods. A novel joint metric is proposed to fuse the instance-level spatial constraint, temporal constraint, and visual feature similarity. The superiority of our method is demonstrated by extensive contrast experiments. And adequate experimental analyses provide more insights into our method.

Acknowledgments

The authors would like to thank Guanan Wang for his valuable help. This work is funded by National Natural Science Foundation of China (Grant No.62006225, 61806197, 62071468, U1836217) and supported by the Strategic Priority Research Program of Chinese Academy of Sciences (Grant No. XDA27040700), Tianjin Science and Technology Plan Project (17ZXRGGX00110).

References

[1] Hermans Alexander, Beyer Lucas, and Leibe Bastian. In defense of the triplet loss for person re-identification. *ArXiv 1703.07737*, 2017. 2

[2] Bedagkar-Gala Apurva and Shah Shishir K. Multiple person re-identification using part based spatio-temporal color appearance model. In *IEEE International Conference on Computer Vision Workshops (ICCV Workshops)*, 2011. 2

[3] Su C., Li J., Zhang S., Xing J., Gao W., and Q. Tian. Pose-driven deep convolutional model for person reidentification. *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2017. 1, 2, 5

[4] Weihua Chen, Xiaotang Chen, Jianguo Zhang, and Kaiqi Huang. A multi-task deep network for person re-identification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31, 2017. 2

[5] Song Chunfeng, Huang Yan, Ouyang Wanli, and Wang Liang. Mask-guided contrastive attention model for person re-identification. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 1, 2, 5, 6

[6] Ahmed Ejaz, Jones Michael, and Tim K. Marks. An improved deep learning architecture for person re-identification. In *IEEE Conference on Computer Vision & Pattern Recognition (CVPR)*, pages 3908–3916, 2015. 1, 2

[7] Ristani Ergys, Solera Francesco, Zou Roger S, Cucchiara Rita, and Tomasi Carlo. Performance measures and a data set for multi-target, multi-camera tracking. In *European Conference on Computer Vision (ECCV)*, 2016. 4, 5, 6, 7

[8] Wang Guangcong, Lai Jianhuang, Huang Peigen, and Xie Xiaohua. Spatial-temporal person re-identification. *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 8933–8940, 2019. 1, 2, 5, 6, 7

[9] Wang Guangcong, Lai Jianhuang, and Xie Xiaohua. P2snet: Can an image match a video for person re-identification in an end-to-end way? *IEEE Transactions on Circuits and Systems for Video Technology*, 2017. 2

[10] Wang Guangrun, Lin Liang, Ding Shengyong, Li Ya, and Wang Qing. Dari: Distance metric and representation integration for person verification. *Proceedings of the AAAI Conference on Artificial Intelligence*, 2016. 2

[11] Wang Guanshuo, Yuan Yufeng, Chen Xiong, Li Jiwei, and Zhou Xi. Learning discriminative features with multiple granularities for person re-identification. *2018 ACM Multimedia Conference*, 2018. 1

[12] Luo Hao, Gu Youzhi, Liao Xingyu, Lai Shenqi, and Jiang Wei. Bag of tricks and a strong baseline for deep person re-identification. In *The IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPR Workshops)*, June 2019. 5, 6

[13] Lingxiao He, Jian Liang, Haiqing Li, and Zhenan Sun. Deep spatial feature reconstruction for partial person re-identification: Alignment-free approach. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018. 1

[14] Lingxiao He, Xingyu Liao, Wu Liu, Xinchen Liu, Peng Cheng, and Tao Mei. Fastreid: A pytorch toolbox for general instance re-identification. *arXiv preprint arXiv:2006.02631*, 2020. 1

[15] Lingxiao He, Wang Yinggang, Liu Wu, Zhao He, Sun Zhenan, and Feng Jiashi. Foreground-aware pyramid reconstruction for alignment-free occluded person re-identification. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019. 1, 2, 5, 6

[16] Shuting He, Hao Luo, Pichao Wang, Fan Wang, Hao Li, and Wei Jiang. Transreid: Transformer-based object re-identification. *arXiv preprint arXiv:2102.04378*, 2021. 2

[17] Takashi Isobe, Dong Li, Lu Tian, Weihua Chen, Yi Shan, and Shengjin Wang. Towards discriminative representation learning for unsupervised person re-identification. *arXiv preprint arXiv:2108.03439*, 2021. 2

[18] Lv Jianming, Chen Weihang, Li Qing, and Yang Can. Unsupervised cross-dataset person re-identification by transfer learning of spatial-temporal patterns. *IEEE Conference on Computer Vision & Pattern Recognition (CVPR)*, 2018. 1, 2, 5

[19] Miao Jiaxu, Wu Yu, Liu Ping, Ding Yuhang, and Yang Yi. Pose-guided feature alignment for occluded person re-identification. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019. 1, 2, 5, 6

[20] Liu Jinxian, Ni Bingbing, Yan Yichao, Zhou Peng, and Hu Jianguo. Pose transferrable person re-identification. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 1, 2, 5, 6

[21] Cho Yeong Jun, Kim Su A, Park Jae Han, Lee Kyuewang, and Yoon Kuk Jin. Joint person re-identification and camera network topology inference in multiple cameras. *Computer Vision and Image Understanding*, 2017. 1, 2

[22] Qi Lei, Huo Jing, Wang Lei, Shi Yinghuan, and Gao Yang. Maskreid: A mask based deep ranking neural network for person re-identification. *ArXiv 1804.03864*, 2018. 1, 2, 5, 6

[23] Zheng Liang, Shen Liyue, Tian Lu, Wang Shengjin, and Tian Qi. Scalable person re-identification: A benchmark. In *IEEE International Conference on Computer Vision (ICCV)*, 2015. 5

[24] Wu Liu, Qian Bao, Yu Sun, and Tao Mei. Recent advances in monocular 2d and 3d human pose estimation: A deep learning perspective. In *CoRR abs/2104.11536*, 2021. 2

[25] Kalayeh Mahdi M., Basaran Emrah, Gokmen Muhittin, Kamasak Mustafa E., and Shah Mubarak. Human semantic parsing for person re-identification. *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 1, 2, 5, 6

[26] Bingpeng Ma, Yu Su, and Frederic Jurie. Covariance descriptor based on bio-inspired features for person reidentification and face verification. *Image and Vision Computing*, 2014. 2

[27] Sakrapee Paisitkriangkrai, Chunhua Shen, and Anton Van Den Hengel. Learning to rank in person re-identification with metric ensembles. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015. 1

[28] Sarfraz M. Saquib, Schumann Arne, Eberle Andreas, and Stiefelhagen Rainer. A pose-sensitive embedding for person re-identification with expanded cross neighborhood re-ranking. *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 5, 6

[29] Gong Shaogang, Cristani Marco, Yan Shuicheng, and Loy Chen Change. Person re-identification. *Advances*

in Computer Vision & Pattern Recognition, 42(7):301–313, 2014. 2

[30] Y. Shen, Hongsheng Li, Shuai Yi, Dapeng Chen, and X. Wang. Person re-identification with deep similarity-guided graph neural network. *ArXiv*, abs/1807.09975, 2018. 2

[31] Ding Shengyong, Lin Liang, Wang Guangrun, and Chao Hongyang. Deep feature learning with relative distance comparison for person re-identification. *Pattern Recognition*, 48(10):2993–3003, 2015. 2

[32] Dapeng Tao, Yanan Guo, Mingli Song, Yaotang Li, Zhengtao Yu, and Yuan Yan Tang. Person re-identification by dual-regularized kiss metric learning. *IEEE Transactions on Image Processing*, 25(6):2726–2738, 2016. 1

[33] Guan'an Wang, Shuo Yang, Huanyu Liu, Zhicheng Wang, Yang Yang, Shuliang Wang, Gang Yu, Erjin Zhou, and Jian Sun. High-order information matters: Learning relation and topology for occluded person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 1, 2, 5, 6

[34] Huang Wenxin, Hu Ruimin, Liang Chao, Yu Yi, and Zhang Chunjie. Camera network based person re-identification by leveraging spatial-temporal constraint and multiple cameras relations. In *International Conference on Multimedia Modeling*, 2016. 1, 2

[35] Lin Wu, Chunhua Shen, Anton Van, and Den Hengel. Deep linear discriminant analysis on fisher networks: A hybrid architecture for person re-identification. *Pattern Recognition*, 2017. 1, 2

[36] Chen Xuesong, Fu Canmiao, Zhao Yong, Zheng Feng, Song Jingkuan, Ji Rongrong, and Yang Yi. Salience-guided cascaded suppression network for person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 1, 2

[37] Yang Yang, Yang Jimei, Yan Junjie, Liao Shengcai, and Li Stan Z. Salient color names for person re-identification. *Europeon Conference on Computer Vision (ECCV)*, 2014. 2

[38] Sun Yifan, Zheng Liang, Yang Yi, Tian Qi, and Wang Shengjin. Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline). *Europeon Conference on Computer Vision (ECCV)*, 2018. 1, 5, 6, 7

[39] Sun Yifan, Xu Qin, Li Yali, Zhang Chi, Li Yikang, Wang Shengjin, and Sun Jian. Perceive where to focus: Learning visibility-aware part-level features for partial person re-identification. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 5, 6, 7

[40] Yuqi Zhang, Qian Qi, Chong Liu, Weihua Chen, Fan Wang, Hao Li, and Rong Jin. Graph convolution for re-ranking in person re-identification. *arXiv preprint arXiv:2107.02220*, 2021. 2

[41] Zhizheng Zhang, Cuiling Lan, Wenjun Zeng, Xin Jin, and Zhibo Chen. Relation-aware global attention for person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 2

[42] Chen Shi Zhe, Guo Chun Chao, and Lai Jianhuang. Deep ranking for person re-identification via joint representation learning. *IEEE Transactions on Image Processing*, pages 1–1, 2016. 1

[43] Kecheng Zheng, Wu Liu, Lingxiao He, Tao Mei, Jiebo Luo, and Zheng-Jun Zha. Group-aware label transfer for domain adaptive person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2021. 2

[44] Liang Zheng, Yi Yang, and Alexander G Hauptmann. Person re-identification: Past, present and future. *ArXiv*, 1610.02984, 2016. 2