Temporal Coherence or Temporal Motion: Which is More Critical for Video-based Person Re-identification?

Guangyi Chen[1,2,3*], Yongming Rao[1,2,3*], Jiwen Lu[1,2,3], and Jie Zhou[1,2,3,4]

[1] Department of Automation, Tsinghua University, China
[2] State Key Lab of Intelligent Technologies and Systems, China
[3] Beijing National Research Center for Information Science and Technology, China
[4] Tsinghua Shenzhen International Graduate School, Tsinghua University, China
chen-gy16@mails.tsinghua.edu.cn, raoyongming95@gmail.com,
{lujiwen,jzhou}@tsinghua.edu.cn

Abstract. Video-based person re-identification aims to match pedestrians with the consecutive video sequences. While a rich line of work focuses solely on extracting the motion features from pedestrian videos, we show in this paper that the temporal coherence plays a more critical role. To distill the temporal coherence part of video representation from frame representations, we propose a simple yet effective Adversarial Feature Augmentation (AFA) method, which highlights the temporal coherence features by introducing adversarial augmented temporal motion noise. Specifically, we disentangle the video representation into the temporal coherence and motion parts and randomly change the scale of the temporal motion features as the adversarial noise. The proposed AFA method is a general lightweight component that can be readily incorporated into various methods with negligible cost. We conduct extensive experiments on three challenging datasets including MARS, iLIDS-VID, and DukeMTMC-VideoReID, and the experimental results verify our argument and demonstrate the effectiveness of the proposed method.

Keywords: Video-based person re-identification, Temporal coherence, Feature augmentation, Adversarial learning.

1 Introduction

Person re-identification (ReID) matches pedestrians in a non-overlapping camera network, which has great potential in surveillance applications [25], such as suspect tracking and missing elderly retrieval. Conventional image-based ReID methods [2, 23, 48, 49] face many challenges, such as pose variations, illumination changes, partial occlusions and clutter background, due to the complicated intra-class variances and the limited clues in the single image. To tackle these challenges, many works [1, 3, 28, 29, 54] tend to use videos instead of a single image to identify the persons.

[*] Equal contribution. The corresponding author is Jiwen Lu.

Compared with image data, surveillance videos avoid complex pre-processing and preserve more abundant identity clues from different view angles and poses. To obtain these identity clues from pedestrian videos, many existing works attempt to extract the temporal motion features, such as the gait of person. For example, some works [1, 17, 18, 43, 45] extract the shallow motion features like HOG3D [19] or learn the deep motion features from the optical flow [7]. Other methods [26, 29, 54] further model the motion information with the recurrent model like RNN or LSTM. Besides, Some works [21, 24] learn the motion clues with 3D convolution neural network (3D-CNN). Different from them, in this paper, we show that the temporal coherence is more critical than than temporal motion, which offers a new perspective on learning better representation for video-based person ReID.

Many video ReID methods capturing the motion clues (e.g. RNN, 3D-CNN, or optical flow based two-stream networks) are inspired by the video recognition tasks. However, different from these video-based recognition tasks (e.g. video classification, action recognition), the video ReID task focuses on the object (person) itself, rather than the pure motion of object. It motivates us to capture the invariant features about the person itself, but not the variant features over time. In Fig. 1, we show a visual comparison between temporal coherence features and optical flows about temporal motion. We can observe that temporal motion features are more instable for different view angles, different actions and other moving occlusions. For example, the optical flows of Tracklet 1 focus on the other occlusion person, while the ones of Tracklet 2 focus on the motion of arm due to the person's action. As a contrast, the temporal coherence features capture the clues that are invariant over time (the cloth of the person), which is more related to the identity and more discriminative. We agree that videos contain the temporal motion clues which are beneficial to person ReID, e.g. the gaits. However, these temporal motion clues also bring intra-class noise like the change of poses, especially for the aggregation stage. As proved in [45], the temporal motion features are harder to be applied to distinguish the person videos, due to the large intra-class variance and small inter-class variance. Thus, we argue that the temporal coherence features are more appropriate for the video-based person ReID task and give a proof-of-concept in this paper.

To distill the temporal coherence feature, we propose a simple yet effective adversarial feature augmentation (AFA) method which generates the adversarial augmented features with the temporal motion noise. In this paper, we use invariant and variant features to represent temporal coherence and temporal motion respectively. Specifically, we disentangle the video representation into the expectation and variance of embeddings of different frames. In the training process, we randomly vary the magnitude of the temporal motion features as an adversarial interference to highlight the temporal coherence feature. On the one hand, the various temporal motion noises break the discriminant of the video representation. On the other hand, the video representation highlights the temporal constance information and adversarially reduce the influence of motion noises. In the testing process, we only use the temporal coherence feature for similar-

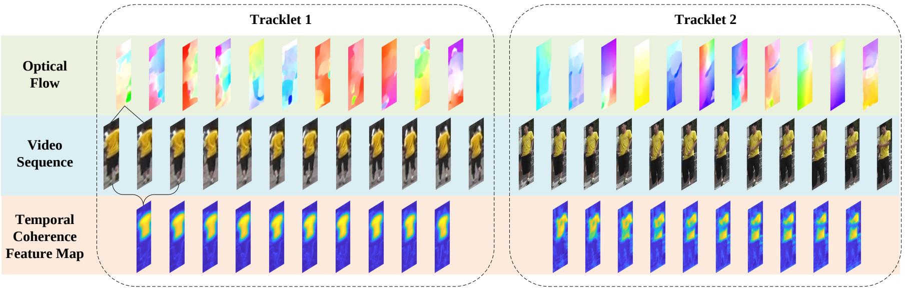


Fig. 1. The visual comparison between temporal coherence features and optical flows of temporal motion. For two tracklets from the same identity under different camera views, the temporal motion optical flows may vary dramatically due to different actions of one pedestrian, while the temporal coherence features are more discriminative since they focus on the invariance of the video.

ity measuring. AFA is a general module that can be readily incorporated into various video-based person ReID methods. It is lightweight and effective which brings significant performance improvement with negligible computing cost. We conduct experiments to verify our argument that the temporal coherence is more critical than the motion clues for video-based person re-identification. The consistent improvements on three challenging datasets including MARS, iLIDS-VID, and DukeMTMC-VideoReID demonstrate the effectiveness of the proposed AFA method. We summarize the contributions of this work as:

1) We show that the temporal coherence is more critical than the motion clues, which offers a new perspective on learning better representation for video-based person ReID.
2) Based on the observation, we propose a simple yet effective method (AFA) to distill the temporal coherence features in an adversarial manner. The proposed AFA model is a lightweight and efficient component that can be readily incorporated into other methods.
3) We conduct extensive experiments to demonstrate the superiority of our AFA method, and achieve the state-of-the-art performance on several large scale video person ReID benchmarks.

2 Related Work

2.1 Video-based Person Re-identification

Video sequences provide abundant and diverse person samples, which indicate more real sample distribution. For learning more robust representation from these video sequences, existing video-based person ReID methods mainly take great efforts to: 1) mine the motion clues in the person video; 2) aggregate the video sequence embeddings. To extract discriminative motion clues, many early works [17, 18, 39, 45] directly employ the temporal motion features like HOG3D [19] as the extra features. While some deep learning methods [1, 26, 29, 43] learn the motion features from the optical flow [7]. In addition, many methods [26, 29, 44, 54] model the motion process with the recurrent model like RNN or LSTM. Recently, 3D convolution neural network [21, 24] (3D-CNN) has been applied for video person ReID to jointly learn the appearance and motion clues. Different from these methods which are mainly inspired by video (action) recognition methods, we argue that the temporal coherence is more critical than the motion clues for video-based person ReID, since the ReID focuses on the maker of the motion but not the motion itself. Thus, we focus on learning the robust temporal consecutive features of pedestrian videos rather than the temporal motion ones.

The aggregation of embeddings of the video sequence is another popular research field for video-based person ReID, which aims to obtain a discriminative video embedding from a sequence of image embeddings. As the baseline methods, [29, 54] apply a temporal pooling layer to average all embeddings of the video. While some attention based methods [1, 3, 4, 22, 28, 43] select key frames of the video to avoid the misleading from noisy frames. Besides, some works [30, 31] sequentially discard confounding frames until the last one, which enlarges the discrimination and reduce the computing cost for video matching. Despite these recent progresses, the aggregation is still difficult due to the large intra-video variance of different frames, especially when the temporal motion features are highlighted. To solve this problem, we distill the temporal coherence feature and reduce the variance of temporal embeddings in an adversarial manner.

2.2 Data Augmentation

Data augmentation is an explicit form of regularization to learn robustness representation and prevent deep models from overfitting by generating extra data. It gains great success in various fields, such as image classification [12, 36], object detection [27, 32] and video analysis [8, 40]. The common data augmentation strategies include flipping, cropping, rotation, color jittering, and adding noises. While Zhang *et al.* [46] proposed to use the convex combinations of pairs as the augmented data. Besides, many works [10, 52] apply the GAN to generate the augmented images. Data augmentation have been also applied for person ReID to learn robust representation. For example, Zhong *et al.* [53] selected a rectangle region in an image and erases its pixels with random values. While Huang *et al.* [16] adversarially occluded samples as the data augmentation. D-ifferent from these methods which augments data, our proposed AFA method distills the temporal coherence feature from video representation by adversarially augmenting the features. Inspired by these adversarial data augmentation methods, we disentangle the video representation into temporal coherence and temporal motion parts and generate the adversarial augmented features with the variable temporal motion features.

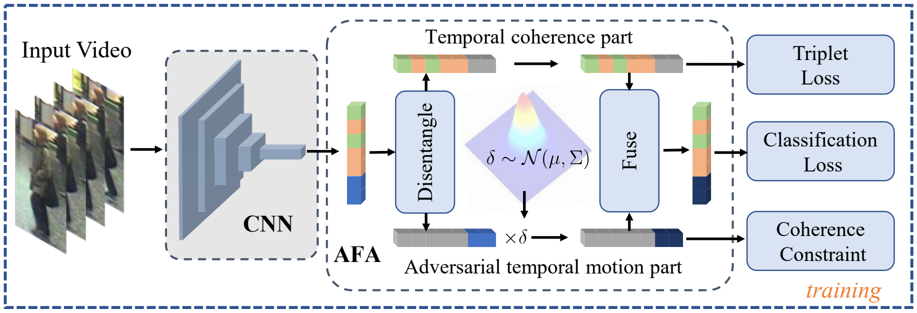


Fig. 2. Illustration of the training process of the adversarial feature augmentation method. The AFA model takes the video representation as input and disentangles it into temporal coherence and temporal motion parts. Then, the AFA model changes the scale of temporal motion features by an adversarial coefficient generated from the Gaussian distribution. The original temporal coherence features and the changed motion features are fused as the new augmented video representation. Finally, the temporal coherence features, temporal motion features and the augmented video representation are feed into the objective functions. The AFA model can be readily incorporated into any video ReID system as a general component. Best viewed in color.

3 Approach

In this section, we first present our adversarial feature augmentation (AFA) method and then employ it to distill the temporal coherence features for video ReID. Finally, we describe the optimization procedure and implementation details for the proposed AFA method.

3.1 Adversarial Feature Augmentation

Many existing works focus on extracting temporal motion features for video-based person ReID, such as optical flow [1,43], RNN [29,44,54], and 3D-CNN [21, 24]. However, in this paper, we argue that the temporal coherence is more critical than the motion clues for the video-based person ReID task, since person ReID focuses on the maker of the motion but not the motion itself. To highlight the temporal coherence of video representation, we propose an adversarial feature augmentation (AFA) method which disentangles the video representation into temporal coherence and motion parts and randomly changes the temporal motion part as an adversarial noise for feature augmentation.

We first describe the feature disentangling process. As shown in Fig. 2, given the video representation which is learned by a CNN model $\mathcal{X} = \mathcal{F}_\theta(\mathcal{V})$, we disentangle the video representation into the temporal coherence and temporal motion parts as:

$$\mathcal{X} = \mathcal{X}_C + \mathcal{X}_D, \tag{1}$$

where $\mathcal{X} = \{X^t \in R^d\}_{t=1:T}$ denotes the video representation, and $\mathcal{X}_C$ and $\mathcal{X}_D$ respectively denote the temporal coherence and motion features. The T is the number of video frames and X^t is the visual embedding of the tth frame, and θ denotes the parameters of the CNN. The temporal coherence feature $\mathcal{X}_C$ represents the invariance in the video sequence. In the pedestrian video, the temporal coherence features mainly focus on the identity information which is invariant for different poses and views in different frames. While the temporal motion features $\mathcal{X}_D$ represent the variety and motion in the video. These temporal motion features not only contain motion clues like gaits but also contain many noises such as pose changing. As quantitatively proved in Fig. 2 of [45], temporal motion features always have more ambiguities than temporal coherence.

Inspired by the prototypical network [33], we assume all embeddings of different frames $\{X^t\}_{t=1:T}$ in the video lie in a manifold, and calculate the video prototype as the center of class:

$$\mathcal{X}_C = \frac{1}{T} \sum_{X^t \in \mathcal{X}} X^t. \tag{2}$$

This prototype $\mathcal{X}_C \in R^d$ denotes the temporal coherence part of the video representation, e.g. the identity information. Then we disentangle the temporal motion features from the video representation as:

$$\mathcal{X}_D = \{X_D^t \in R^d | X_D^t = X^t - \mathcal{X}_C\}. \tag{3}$$

In above definition, we classify all other clues as the temporal motion features except for the identity-related temporal coherence. These temporal motion features may include the motions, the varying backgrounds and other noises. In this paper, we regard temporal motion features as the adversarial noises and utilize them to distill the temporal coherence features.

Given the disentangled temporal coherence and temporal motion features, we design an adversarial coefficient δ to generate the adversarial features as:

$$\mathcal{X}' = \{X'^t | X'^t = \mathcal{X}_C + \delta X_D^t\}, \tag{4}$$

where $\mathcal{X}'$ is the augmented new feature with the various motion noise as shown in Fig. 2. The adversarial coefficient δ randomly varies in the training stage following a Gaussian distribution $\mathcal{N}$:

$$\delta \sim \mathcal{N}(\mu = 1, \varSigma), \tag{5}$$

where $\mu = 1$ indicates the expectation of adversarial coefficient is 1, and the standard deviation $\varSigma$ is a hyperparameter to control the amplitude of the noise. The larger standard deviation indicates to increase the noises. In the experiments, we set the standard deviation as $\varSigma = 0.025$.

In the training process, we sample δ to generate the adversarial augmented features with the temporal motion noise. These variable temporal motion noises break the discriminant, while the video representation will highlight the temporal constance information and adversarially reduce the influence from motion noise. By the adversarial training for these motion noise, the learned video representation can be robust for the large intra-class variance, including different poses, occlusions, and cluttered background. Note that, compared with the baseline model, our AFA model only introduces a pool layer with negligible computing cost. While in the testing, we fixed $\delta = 0$, which is equal to remove the adversarial feature augmentation and only use the temporal coherence features $\mathcal{X}_C$ for evaluation. It requires **no** extra cost in the inference process.

3.2 Optimization

Given the new video feature $\mathcal{X}'$ augmented by our adversarial feature augmentation method, we optimize it to reduce the influence from motion noise. Instead of calculating the objective function with the single video representation which is aggregated from the embeddings of the video sequence, we separately optimize each augmented feature X'^t to constrain the temporal coherence of the video representation.

The objective function of our method is formulated as follows:

$$\mathcal{L}(\mathcal{X}', \mathcal{X}_C, \mathcal{X}_D) = \mathcal{L}_{cls}(\mathcal{X}') + \mathcal{L}_{tri}(\mathcal{X}_C) + \lambda \mathcal{L}_{coh}(\mathcal{X}_D), \tag{6}$$

which contains three parts: classification loss, triplet loss, and coherence constraint. The λ is a rate to balance different loss functions.

1) Classification Loss: We apply the cross entropy loss function as the classification loss to learn the identify-specific video representation. For each augmented feature $X'^t \in \mathcal{X}'$, we first apply a batch normalization layer before the classifier to normalize the scales, since the classification loss is sensitive to the scale of features. Then we calculate the predicted probabilities of each frame with a linear classifier:

$$p^k(X'^t) = \frac{exp(W_k X'^t)}{\sum_j exp(W_j X'^t)}, \tag{7}$$

where W_k indicate the kth column of the linear classifier, and $p^k(X'^t)$ is the predicted probability of the frame X'^t for kth class. Then we aggregate the classification results of the frames as the video-based classification result:

$$p^k(\mathcal{X}') = \frac{1}{T} \sum_{t=1}^{T} p^k(X'^t). \tag{8}$$

Classification results of all frames are concentrative to the same identity, which constraints the temporal coherence of the video representation. Finally, we apply the cross-entropy loss to supervise the classifier and representation model:

$$\mathcal{L}_{cls}(\mathcal{X}') = \frac{1}{|\Omega|} \sum_{\mathcal{X}' \in \Omega} \sum_{k=1}^{K} y^k(\mathcal{X}') \log(p^k(\mathcal{X}')), \tag{9}$$

where $y^k(\mathcal{X}') = 1$ denotes the ground truth identity of the video clip $\mathcal{X}'$ is k, and 0 denotes not. We use Ω to denote the whole training set.

Algorithm 1 : Adversarial Feature Augmentation

```
Input: Training video sequences: $\{\mathcal{V}\}$, maximal iterative number $I$, The standard
    deviation coefficient $\mathit{\Sigma}$.
Output: The parameters of video representation network $\theta$.
Initialize $\theta$;
for $i = 1, 2, \ldots, I$ do
    Randomly select a batch of video sequences from $\{\mathcal{V}\}$;
    Obtain the original adversarial coefficient $\mathcal{X}$;
    Disentangle the video representation into temporal coherence and temporal
    motion parts as (1),(2),(3);
    Generate an adversarial coefficient $\lambda$ from a Gaussian distribution as (5);
    Obtain the augmentation video representation $\mathcal{X}'$ as (4);
    Update $\theta \leftarrow -\frac{\partial}{\partial \theta}\mathcal{L}(\mathcal{X}', \mathcal{X}_C, \mathcal{X}_D)$ as (6);
end for
return $\theta$
```

2) Triplet Loss: We employ the triplet loss function [13] to preserve the rank relationship among a triplet of samples with a large margin, which increases the inter-class distance and reduces the intra-class one. The triplet loss is directly applied on the temporal coherence features to increase the discriminative ability:

$$\mathcal{L}_{tri}(\mathcal{X}_C) = \sum_{\mathcal{X}_C \in \mathit{\Omega}} \left[||\mathcal{X}_C - \mathcal{X}_C^+||_2^2 - ||\mathcal{X}_C - \mathcal{X}_C^-||_2^2 + m \right]_+, \tag{10}$$

where $[\cdot]_+$ indicates the max function $\max(0, \cdot)$, and $\mathcal{X}_C, \mathcal{X}_C^+, \mathcal{X}_C^-$ respectively denote as the temporal coherence features of the anchor, positive and negative sample in a triplet. m is a margin to enhance the discriminative ability of learned features. In the experiments, we apply the adaptive soft margin and hard negative mining strategies as [13] and measure the distance in the Euclidean space.

3) Coherence Constraint: To further distill the temporal coherence, we develop a coherence constraint loss, which reduces the influence of temporal motion parts. It is formulated as:

$$\mathcal{L}_{coh}(\mathcal{X}_D) = \sum_{\mathcal{X}_D \in \mathit{\Omega}} ||\mathcal{X}_D||_2, \tag{11}$$

where $||\cdot||_2$ denotes the L2 norm. By this loss function, we aim that the scales of the temporal motion parts are limited. In other view, it is equal to apply a Mean Squared Error (MSE) loss to reduce the intra-class variance of video sequence. To explain the optimization more clearly, we provide Algorithm 1 to detail the learning process of our AFA method.

3.3 Implementation Details

We employed the ResNet-50 [37] as the basic backbone network for our AFA method in the experiments, and initialized it with the ImageNet pre-trained

Table 1. The basic statistics of all datasets in the experiments.

Datasets	Identities	Sequences	Frames	Cameras	Splits	Repetitions
iLIDS-VID [38]	300	600	73	2	150/150	10 times
MARS [51]	1,261	20,715	58	6	625/636	1 time
DukeV [42]	1,404	4,832	168	8	702/702	1 time

parameters. In order to preserve the resolution of the image, we applied a convolution layer with $stride = 1$, instead of original $stride = 2$ convolution layer in the last block of ResNet-50. During training, we apply two data augmentation methods including the horizontal flipping and the designed video-based random erasing. This video random erasing data augmentation method erases the same region for all frames in the same clip, to overcame the partial occlusions. In each mini-batch, we randomly selected 8 individuals and sampled 4 video clips for each individual. Each video clips consists of 7 images for MARS and iLIDS-VID datasets, and 9 images for DukeMTMC-VideoReID dataset, since the video sequences in the DukeMTMC-VideoReID dataset are longer than others. Besides, we only use the optical flows for iLIDS-VID since the better manual alignment. Each input image is resized as 256×128. The standard deviation coefficient Σ in adversarial coefficient distribution and the balance rate λ of loss functions are respectively set as 0.025 and 0.1 in the experiments. We trained our model for 200 epochs in total by the Adam optimizer. The initial learning rate was 0.0001 and was divided by 10 every 50 epochs. The weight decay factor for L2 regularization was set to 0.00001. During evaluation, we removed the feature augmentation part and used the temporal coherence features for evaluation. We employed the Euclidean distance as the metric to measure the similarity of two features. All experiments were implemented with PyTorch 1.3.1 on 2 Nvidia GTX 1080Ti GPUs. Taking MARS dataset as the example, the whole training process took about 2.4 hours with data-parallel acceleration.

4 Experiments

In the experiments, we evaluated our method on three public video-based person ReID benchmarks. We compared the proposed method with other state-of-the-art approaches and conducted ablation studies and parameter analysis to analyze our AFA model. In addition, we conducted the transfer testing on the cross-dataset to investigate the generalization ability.

4.1 Datasets and Settings

We conducted experiments on three challenging datasets including iLIDS-VID [38], MARS [51], and DukeMTMC-VideoReID [42]. The detailed statistics and evaluation protocols of all datasets are summarized in Table 1. The iLIDS-VID dataset contains 600 sequences of 300 pedestrians under two camera views. MARS is one of the largest public video ReID dataset, including 1261 persons and around 20000 video sequences captured by 6 cameras. Different from other datasets, the video sequences of the MARS dataset are detected with DPM detector [9], and tracked by the GMMCP tracker [6], instead of hand-drawn bounding boxes. These bounding boxes are always misaligned which causes the large intra-class variances. DukeMTMC-VideoReID [42] is another large-scale video-based benchmark, which comprises around 4,832 videos from 1,404 identities. In the following description, we use the abbreviation "DukeV" to represent the DukeMTMC-VideoReID dataset for convenience. The video sequences in the DukeV dataset are longer than videos in other datasets, which contain 168 frames on average.

In the experiments, we adopt the protocol of [38] for iLIDS-VID datasets, which repeated experiments 10 times and calculated the average accuracy. In each repeat, the dataset was randomly split into equal-sized training and testing sets, where the videos from the first camera view are regarded as the query set and the other as the gallery set. For a fair comparison, we selected the identical 10 splits as [38], instead of random splits, to avoid the experimental bias from dataset splitting. For MARS and DukeV datasets, we followed the settings as [15, 20, 35]. Note that, all the experiments are **NOT** applied the re-ranking tricks in the evaluation. We resort to both cumulative matching characteristic (CMC) curves and mean Average Precision (mAP) as evaluation metrics.

4.2 Comparison with the State-of-the-Art Methods

As shown in the Table 2 and Table 3, we respectively compared our method with other SOTA methods on the iLIDS-VID, MARS, and DukeV datasets. We can observe that the proposed AFA method achieves superior performance over other comparing methods by a large margin on all three benchmarks, which confirms the importance of the temporal coherence in the video-based person ReID task.

For iLIDS-VID and MARS datasets, we compared our AFA methods against 10 aggregation-based methods and other 9 methods with temporal feature learning. As shown in Table 2, we summarized the aggregation based methods in the top group and temporal feature learning methods in the bottom group. For both iLIDS-VID and MARS datasets, we achieved consistent improvement on Rank-1 and mAP performance.

DUKEV is a recently proposed large scale video ReID dataset, where only a limited number of works have been evaluated and reported. Table 3 shows the performance of our AFA method and other SOTA video ReID works including STA [11], VRSTC [15], COSAM [35], and GLTR [20]. Our AFA method outperformed all other methods by a large margin, which indicates that our AFA model is also appropriate for the long term videos.

4.3 Assumption Evaluation

In this paper, we argue the temporal coherence is more critical than the temporal motion for the video-based person ReID, and propose an AFA method to distill

Table 2. Comparison with the state-of-the-art video-based person ReID methods on the iLIDS-VID and MARS datasets.

Method	**Source**	**iLIDS-VID**			**MARS**		
		R1	**R5**	**R20**	**R1**	**R5**	**mAP**
CNN+XQDA [51]	ECCV 2016	54.1	80.7	95.4	65.3	82.0	47.6
QAN [28]	CVPR 2017	68.0	86.6	97.4	73.7	84.9	51.7
ASTPN [43]	ICCV 2017	62.0	86.0	98.0	44	70	-
RQEN [34]	AAAI 2018	76.1	92.9	99.3	73.7	84.9	51.7
DRSTA [22]	CVPR 2018	80.2	-	-	82.3	-	65.9
CSSA+CASE [1]	CVPR 2018	85.4	96.7	99.5	86.3	94.7	76.1
SDM [47]	CVPR 2018	60.2	84.7	95.2	71.2	85.7	-
STAL [3]	TIP 2019	82.8	95.3	98.8	80.3	90.9	64.5
STA [11]	AAAI 2019	-	-	-	86.3	95.7	80.8
ADFDTA [50]	CVPR 2019	86.3	97.4	**99.7**	87.0	95.4	78.2
DVR [39]	TPAMI 2016	41.3	63.5	83.1	-	-	-
CNN+RNN [29]	CVPR 2016	58.0	84.0	96.0	56	69	-
AMOC+ EpicFlow [26]	TCSVT 2017	68.7	94.3	99.3	68.3	81.4	52.9
TAM+SRM [54]	CVPR 2017	55.2	86.5	97.0	70.6	90.0	50.7
DSAN [41]	TMM 2018	61.2	80.7	97.3	69.7	83.4	-
TRL [5]	TIP 2018	57.7	81.7	94.1	80.5	91.8	69.1
VRSTC [15]	CVPR 2019	83.4	95.5	99.5	88.5	96.5	82.3
COSAM [35]	ICCV 2019	79.6	95.3	-	84.9	95.5	79.9
GLTR [20]	ICCV 2019	86.0	**98.0**	-	87.0	95.8	78.5
AFA	ours	**88.5**	96.8	**99.7**	**90.2**	**96.6**	**82.9**

the main feature. To quantitatively evaluate which is better between temporal coherence and motion, we respectively supposed temporal coherence feature or motion feature is more important and applied AFA method to highlight them. As shown in the part (a) of Fig. 3, we compare the performance under these two assumptions on the MARS dataset. The red and blue curves respectively denote that we apply the AFA method to distill the temporal coherence and motion features. We can observe that using AFA method to distill the temporal coherence features obtains the dramatic improvement than the temporal motion based one. Furthermore, the performance steadily declines when we increase the standard deviation coefficient Σ, (larger Σ indicates larger augmentation). It is because the motion features may contain many noises from the occlusions, pose changing and cluttered background. Compared with the intra-class noise, the effect from beneficial clues of the temporal motion (like gaits) is limited.

4.4 Ablation Studies

In this subsection, we evaluated the generality of our AFA method for different baseline models, and investigated the contributions of different components. We summarized the comparison results on the MARS dataset in different settings in Table 4 and separately analyzed each component as follows:

Table 3. Comparison with the state-of-the-art video-based person ReID methods on the DukeMTMC-VideoReID dataset.

Method	**Source**	**DukeMTMC-VideoReID**				
		R1	**R5**	**R10**	**R20**	**mAP**
STA [11]	AAAI 2019	96.2	99.3	99.6	-	94.9
VRSTC [15]	CVPR 2019	95.0	99.1	99.4	-	93.5
COSAM [35]	ICCV 2019	95.4	99.3	-	99.8	94.1
GLTR [20]	ICCV 2019	96.3	99.3	-	99.7	93.7
AFA	ours	**97.2**	**99.4**	**99.7**	**99.9**	**95.4**

The generality for different baselines: We compared our AFA methods with two baselines, including original ResNet-50 [12] and QAN [28]. We implemented these two baselines with the same parameters and then added our AFA component. In the QAN* + AFA setting, we apply the quality attention to obtain the temporal coherence features. As shown in the top part of Table 4, our AFA module can obviously improve the baseline network by distilling the temporal coherence features.

Loss functions: The loss functions of our method including three parts: triplet loss, cross-entropy loss, and MES-based coherence constraint loss. We compared and analyzed the effectiveness of different loss functions. We employed the triplet loss as the basic objective functions in our method and use the original ResNet-50 as the baseline model. As shown the bottom part of Table 4, we achieved a superior performance when we additionally employed the cross-entropy loss to supervise the classification results of all frames are concentrative to the same identity. While the MES-based coherence constraint loss further promotes the performance. Note that both the ResNet-50 + AFA and L_{tri} + L_{cls} + L_{coh} settings denote the full AFA method. We display it twice in the both top and bottom parts of Table 4 for more clear comparison.

4.5 Parameters Analysis

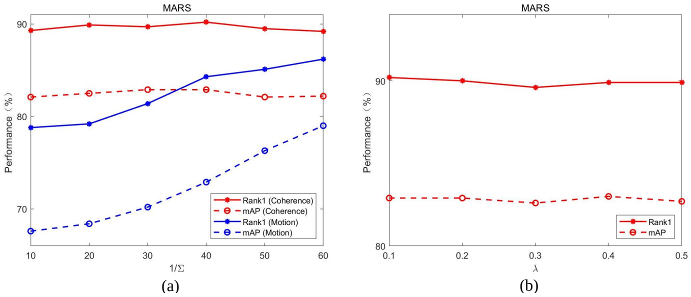
We conducted parameters analysis about the standard deviation coefficient Σ in adversarial coefficient distribution and the balance rate λ of loss functions.

Standard deviation coefficient Σ: In the AFA method, we randomly sample the adversarial coefficients δ from a Gaussian distribution $\mathcal{N}$ to generate the adversarial augmented features. The standard deviation coefficient Σ in the Gaussian distribution indicates the scale of the noisy temporal motion feature. In the part (a) of Fig. 3, the abscissa is the reciprocal of Σ and the ordinate denotes the performance on the MARS dataset. We observe that the performance is slightly lower when the standard deviation coefficient Σ is too large to interfere the training process.

Balance rate λ: We apply a trade-off parameter λ to balance different loss functions. Following the [35], we also set and fixed the rate between the triplet loss and cross-entropy loss as 1. In this subsection, we mainly discuss the balance

Table 4. Ablation studies on the MARS and DUKEV datasets, including the evaluations of AFA model, different baselines, and loss functions. The * indicates that the method is reproduced by ourself with the same backbone and hyperparameters of AFA.

Method	**MARS**				**DukeV**			
	R1	**R5**	**R10**	**mAP**	**R1**	**R5**	**R10**	**mAP**
ResNet-50	88.1	95.6	96.8	80.1	95.2	99.2	99.7	94.3
QAN*	88.6	95.2	96.9	80.9	95.5	98.8	99.6	94.5
ResNet-50 + AFA	**90.2**	96.6	**97.6**	**82.9**	**97.2**	99.4	**99.7**	95.4
QAN* + AFA	89.7	**96.8**	97.4	82.2	97.2	**99.5**	99.7	**95.5**
L_{tri} only	85.3	93.2	95.4	78.7	94.3	98.9	99.3	93.2
L_{tri} + L_{cls}	89.8	96.2	97.2	82.6	96.6	99.3	**99.7**	95.0
L_{tri} + L_{cls} + L_{coh}	**90.2**	96.6	**97.6**	**82.9**	**97.2**	99.4	**99.7**	95.4



(a)

(b)

Fig. 3. Parameters analysis on the MARS dataset about (a) the standard deviation coefficient Σ and (b) the balance rate λ.

rate on the coherence constraint loss. As shown in the part (b) in Fig. 3, the performance of different balance rates on the MARS dataset is stable, which indicates the robustness of the AFA method for the trade-off parameter of the coherence constraint loss.

4.6 Cross-Dataset Evaluation

In real surveillance systems, it requires intensive human labor to label overwhelming amount of data for training model. Thus, the cross-dataset evaluation is an important evaluation metric for person ReID systems, which measures the generalization ability of the ReID model for unseen persons and scenes. Many existing works [3, 28, 29] have conducted this cross-dataset evaluation, which train the model on the iLIDS-VID [38] dataset and test it on PRID-2011 [14].

However, this experimental setting has two main problems. First, the performance is unstable for different splits of the iLIDS-VID and PRID-2011 datasets. Thus, the comparisons from different works may be unfair. Second, the scales of the iLIDS-VID and PRID-2011 datasets are limited, which are not enough to represent the real surveillance system environment. For above reasons, we pro-

Table 5. Cross dataset evaluations between the MARS and DukeV datasets. The * indicates that the method is reproduced by ourself with the same backbone and hyperparameters of our AFA method.

Method	**MARS → DukeV**				**DukeV → MARS**			
	R1	**R5**	**R10**	**mAP**	**R1**	**R5**	**R10**	**mAP**
Baseline	37.6	56.3	63.9	31.8	43.3	58. 6	64.5	23.8
QAN*	38.9	**59.3**	64.5	33.0	43.3	**58.8**	**64.6**	24.1
AFA	**41.5**	59.1	**67.0**	**34.6**	**44.2**	**58.8**	**64.6**	**24.5**

pose to use the MARS and DUKEV datasets for cross-dataset evaluation, which both are large-scale benchmarks with the single fixed split.

As shown in Table 5, we trained the model with the data in the MARS dataset and tested it with the samples in the DukeV dataset, and vice versa. We compared the generalization abilities of the ResNet-50 baseline, QAN [28], and our AFA method. All the methods are using the similar backbone network and hyperparameters. For both evaluation settings, our AFA method obtained the superior performance than the baseline method and QAN method.

5 Conclusion

In this work, we have argued that the temporal coherence is more critical than motion clues for the video based person ReID task. To distill these temporal coherence clues, we have proposed an adversarial feature augmentation (AFA) method, which disentangles the video representation into the temporal coherence and temporal motion parts and highlights the temporal coherence features by generating the adversarial augmented features with the variable temporal motion noise. The proposed AFA model can be incorporated into other video ReID methods with negligible cost, as a general lightweight component. Extensive experimental results demonstrate the importance the temporal coherence and validate the effectiveness of our AFA approach.

Acknowledge

This work was supported in part by the National Key Research and Development Program of China under Grant 2017YFA0700802, in part by the National Natural Science Foundation of China under Grant 61822603, Grant U1813218, Grant U1713214, and Grant 61672306, in part by Beijing Natural Science Foundation under Grant No. L172051, in part by Beijing Academy of Artificial Intelligence (BAAI), in part by a grant from the Institute for Guo Qiang, Tsinghua University, in part by the Shenzhen Fundamental Research Fund (Subject Arrangement) under Grant JCYJ20170412170602564, and in part by Tsinghua University Initiative Scientific Research Program.

References

1. Chen, D., Li, H., Xiao, T., Yi, S., Wang, X.: Video person re-identification with competitive snippet-similarity aggregation and co-attentive snippet embedding. In: CVPR. pp. 1169–1178 (2018)
2. Chen, G., Lin, C., Ren, L., Lu, J., Zhou, J.: Self-critical attention learning for person re-identification. In: ICCV. pp. 9637–9646 (2019)
3. Chen, G., Lu, J., Yang, M., Zhou, J.: Spatial-temporal attention-aware learning for video-based person re-identification. TIP **28**(9), 4192–4205 (2019)
4. Chen, G., Lu, J., Yang, M., Zhou, J.: Learning recurrent 3d attention for video-based person re-identification. TIP **29**, 6963–6976 (2020)
5. Dai, J., Zhang, P., Wang, D., Lu, H., Wang, H.: Video person re-identification by temporal residual learning. TIP (2018)
6. Dehghan, A., Modiri Assari, S., Shah, M.: Gmmcp tracker: Globally optimal generalized maximum multi clique problem for multiple object tracking. In: CVPR. pp. 4091–4099 (2015)
7. Dosovitskiy, A., Fischer, P., Ilg, E., Hausser, P., Hazirbas, C., Golkov, V., Van Der Smagt, P., Cremers, D., Brox, T.: Flownet: Learning optical flow with convolutional networks. In: ICCV. pp. 2758–2766 (2015)
8. Feichtenhofer, C., Fan, H., Malik, J., He, K.: Slowfast networks for video recognition. In: ICCV. pp. 6202–6211 (2019)
9. Felzenszwalb, P.F., Girshick, R.B., McAllester, D., Ramanan, D.: Object detection with discriminatively trained part-based models. TPAMI **32**(9), 1627–1645 (2010)
10. Frid-Adar, M., Klang, E., Amitai, M., Goldberger, J., Greenspan, H.: Synthetic data augmentation using gan for improved liver lesion classification. In: ISBI. pp. 289–293. IEEE (2018)
11. Fu, Y., Wang, X., Wei, Y., Huang, T.: Sta: Spatial-temporal attention for large-scale video-based person re-identification. In: AAAI (2019)
12. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR. pp. 770–778 (2016)
13. Hermans, A., Beyer, L., Leibe, B.: In defense of the triplet loss for person re-identification. arXiv (2017)
14. Hirzer, M., Beleznai, C., Roth, P.M., Bischof, H.: Person re-identification by descriptive and discriminative classification. In: SCIA. pp. 91–102 (2011)
15. Hou, R., Ma, B., Chang, H., Gu, X., Shan, S., Chen, X.: Vrstc: Occlusion-free video person re-identification. In: CVPR (June 2019)
16. Huang, H., Li, D., Zhang, Z., Chen, X., Huang, K.: Adversarially occluded samples for person re-identification. In: CVPR. pp. 5098–5107 (2018)
17. Karanam, S., Li, Y., Radke, R.J.: Person re-identification with discriminatively trained viewpoint invariant dictionaries. In: ICCV. pp. 4516–4524 (2015)
18. Karanam, S., Li, Y., Radke, R.J.: Sparse re-id: Block sparsity for person re-identification. In: CVPR Workshops. pp. 33–40 (2015)
19. Klaser, A., Marszałek, M., Schmid, C.: A spatio-temporal descriptor based on 3d-gradients. In: BMVC. pp. 1–10 (2008)
20. Li, J., Wang, J., Tian, Q., Gao, W., Zhang, S.: Global-local temporal representations for video person re-identification. In: ICCV (October 2019)
21. Li, J., Zhang, S., Huang, T.: Multi-scale 3d convolution network for video based person re-identification. In: AAAI. vol. 33, pp. 8618–8625 (2019)
22. Li, S., Bak, S., Carr, P., Wang, X.: Diversity regularized spatiotemporal attention for video-based person re-identification. In: CVPR. pp. 369–378 (2018)

23. Li, W., Zhu, X., Gong, S.: Harmonious attention network for person re-identification. In: CVPR. p. 2 (2018)
24. Liao, X., He, L., Yang, Z., Zhang, C.: Video-based person re-identification via 3d convolutional networks and non-local attention. In: ACCV. pp. 620–634. Springer (2018)
25. Lin, J., Ren, L., Lu, J., Feng, J., Zhou, J.: Consistent-aware deep learning for person re-identification in a camera network. In: CVPR (2017)
26. Liu, H., Jie, Z., Jayashree, K., Qi, M., Jiang, J., Yan, S., Feng, J.: Video-based person re-identification with accumulative motion context. TCSVT **28**(10), 2788–2802 (2017)
27. Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.Y., Berg, A.C.: Ssd: Single shot multibox detector. In: ECCV. pp. 21–37 (2016)
28. Liu, Y., Yan, J., Ouyang, W.: Quality aware network for set to set recognition. In: CVPR (2017)
29. McLaughlin, N., Martinez del Rincon, J., Miller, P.: Recurrent convolutional network for video-based person re-identification. In: CVPR. pp. 1325–1334 (June 2016)
30. Ouyang, D., Shao, J., Zhang, Y., Yang, Y., Shen, H.T.: Video-based person re-identification via self-paced learning and deep reinforcement learning framework. In: ACM MM. pp. 1562–1570 (2018)
31. Rao, Y., Lu, J., Zhou, J.: Learning discriminative aggregation network for video-based face recognition and person re-identification. IJCV **127**(6-7), 701–718 (2019)
32. Ren, S., He, K., Girshick, R., Sun, J.: Faster r-cnn: Towards real-time object detection with region proposal networks. In: NIPS. pp. 91–99 (2015)
33. Snell, J., Swersky, K., Zemel, R.: Prototypical networks for few-shot learning. In: NeurIPS. pp. 4077–4087 (2017)
34. Song, G., Leng, B., Liu, Y., Hetang, C., Cai, S.: Region-based quality estimation network for large-scale person re-identification. In: AAAI (2018)
35. Subramaniam, A., Nambiar, A., Mittal, A.: Co-segmentation inspired attention networks for video-based person re-identification. In: ICCV (October 2019)
36. Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A.: Going deeper with convolutions. In: CVPR. pp. 1–9 (2015)
37. Wang, F., Jiang, M., Qian, C., Yang, S., Li, C., Zhang, H., Wang, X., Tang, X.: Residual attention network for image classification. In: CVPR (2017)
38. Wang, T., Gong, S., Zhu, X., Wang, S.: Person re-identification by video ranking. In: ECCV. pp. 688–703 (2014)
39. Wang, T., Gong, S., Zhu, X., Wang, S.: Person re-identification by discriminative selection in video ranking. TPAMI **38**(12), 2501–2514 (2016)
40. Wang, X., Girshick, R., Gupta, A., He, K.: Non-local neural networks. In: CVPR. pp. 7794–7803 (2018)
41. Wu, L., Wang, Y., Gao, J., Li, X.: Where-and-when to look: Deep siamese attention networks for video-based person re-identification. TMM (2018)
42. Wu, Y., Lin, Y., Dong, X., Yan, Y., Ouyang, W., Yang, Y.: Exploit the unknown gradually: One-shot video-based person re-identification by stepwise learning. In: CVPR. pp. 5177–5186 (2018)
43. Xu, S., Cheng, Y., Gu, K., Yang, Y., Chang, S., Zhou, P.: Jointly attentive spatial-temporal pooling networks for video-based person re-identification. In: ICCV (2017)
44. Yan, Y., Ni, B., Song, Z., Ma, C., Yan, Y., Yang, X.: Person re-identification via recurrent feature aggregation. In: ECCV. pp. 701–716 (2016)

45. You, J., Wu, A., Li, X., Zheng, W.S.: Top-push video-based person re-identification. In: CVPR. pp. 1345–1353 (June 2016)
46. Zhang, H., Cisse, M., Dauphin, Y.N., Lopez-Paz, D.: mixup: Beyond empirical risk minimization. arXiv preprint arXiv:1710.09412 (2017)
47. Zhang, J., Wang, N., Zhang, L.: Multi-shot pedestrian re-identification via sequential decision making. In: CVPR (2018)
48. Zhang, L., Xiang, T., Gong, S.: Learning a discriminative null space for person re-identification. In: CVPR. pp. 1239–1248 (2016)
49. Zhao, H., Tian, M., Sun, S., Shao, J., Yan, J., Yi, S., Wang, X., Tang, X.: Spindle net: Person re-identification with human body region guided feature decomposition and fusion. In: CVPR (2017)
50. Zhao, Y., Shen, X., Jin, Z., Lu, H., Hua, X.s.: Attribute-driven feature disentangling and temporal aggregation for video person re-identification. In: CVPR (June 2019)
51. Zheng, L., Bie, Z., Sun, Y., Wang, J., Su, C., Wang, S., Tian, Q.: Mars: A video benchmark for large-scale person re-identification. In: ECCV. pp. 868–884 (2016)
52. Zheng, Z., Zheng, L., Yang, Y.: Unlabeled samples generated by gan improve the person re-identification baseline in vitro. In: ICCV. pp. 3754–3762 (2017)
53. Zhong, Z., Zheng, L., Kang, G., Li, S., Yang, Y.: Random erasing data augmentation. arXiv preprint arXiv:1708.04896 (2017)
54. Zhou, Z., Huang, Y., Wang, W., Wang, L., Tan, T.: See the forest for the trees: Joint spatial and temporal recurrent neural networks for video-based person re-identification. In: CVPR (July 2017)