Appearance and Motion Enhancement for Video-Based Person Re-Identification

Shuzhao Li,[1] **Huimin Yu,**[1,2*] **Haoji Hu**[1]
[1]College of Information Science and Electronic Engineering, Zhejiang University, Hangzhou, China
[2]The State Key Laboratory of CAD and CG, Zhejiang University, Hangzhou, China
{leeshuz, yhm2005, haoji_hu}@zju.edu.cn

Abstract

In this paper, we propose an Appearance and Motion Enhancement Model (AMEM) for video-based person re-identification to enrich the two kinds of information contained in the backbone network in a more interpretable way. Concretely, human attribute recognition under the supervision of pseudo labels is exploited in an Appearance Enhancement Module (AEM) to help enrich the appearance and semantic information. A Motion Enhancement Module (MEM) is designed to capture the identity-discriminative walking patterns through predicting future frames. Despite a complex model with several auxiliary modules during training, only the backbone model plus two small branches are kept for similarity evaluation which constitute a simple but effective final model. Extensive experiments conducted on three popular video-based person ReID benchmarks demonstrate the effectiveness of our proposed model and the state-of-the-art performance compared with existing methods.

Introduction

Person Re-identification (ReID) aims at matching images of a person in different non-overlapping cameras, which has attracted increasing attention during recent years due to its wide application in practical scenarios, such as criminal retrieval, video surveillance and so on. Many researches have been conducted on this topic while it is still full of challenges including human pose variation, occlusions and different camera viewpoints.

Currently, the prominent progress of person ReID is mainly gathered in the static image setting, which only exploits a static image and its spatial information to perform the matching process. Despite easier to implement and less complexity, image-based methods present many disadvantages compared with video-based methods. On one hand, the single image is sensitive to pose variation and occlusions while the multiple images contained in one sequence provide more samples against these problems. On the other hand, using only one image cannot capture the motion pattern of people walking, which is another important clue for identifying someone apart from the appearance. Besides, video format data is more accessible in practical scenarios. To this end, more and more researchers have shifted their attention to the video-based setting, which is also the topic we mainly focus on in this paper.

To take full advantage of the temporal cues provided by video data, multiple common methods for video analysis have been integrated into the ReID system, including RNN, optical flow learning, and 3D CNN. Although these are powerful tools for temporal feature extraction and action recognition, they may be not well suited for person ReID task when directly applied since all data contains only one action category – walking, which is similar among different samples, and the subtle difference contained in walking styles is hard to capture without specific design. Attention models also play important roles in fusing multiple image features into a sequence-level one, whereas neglecting most of the temporal cues. Apart from this, considering the special property of human walking pattern from which people can easily identify someone, gait recognition has aroused wide attention over the years. However, most approaches take Gait Energy Image (GEI) as input which makes strong assumptions on the initial walking tracklets, such as the uncluttered background, aligned sequence, silhouette extraction and so on. These are difficult to acquire especially from the surveillance data in complex scenarios. To this end, we seek to design a novel model to simultaneously capture the temporal cues and take into account the specialty of walking patterns from different people.

Attribute learning for Person ReID task has been studied in recent years and proven to be of great help when treated as one kind of mid-level semantic feature. While in most existing works, attributes are only exploited in the image-based setting, few works consider to incorporate it into video analysis. The reasons may lie in two aspects. On one hand, there are no existing video datasets containing attribute labels and it takes heavy manual labor to make explicit annotations. On the other hand, some attributes may occur only in certain frames in one sequence due to the pose variation or occlusions, it will be ambiguous to decide the sequence-level labels. To deal with these, we propose to take advantage of the existing image-based attribute dataset to assist the learn-

*Corresponding author

ing of video-based attributes. As a result, the appearance and semantic information in the human walking tracklets can be largely strengthened.

Based on above analysis, in this paper we propose a video-based person ReID model called **A**ppearance and **M**otion **E**nhancement **M**odel (AMEM) which enriches the two types of information in the backbone networks to better leverage the abundant information provided by video data. To enrich the appearance information, an Appearance Enhancement Module (AEM) is proposed where human attribute learning is exploited on video data to pay attention to different aspects of human appearance. Compared with single-image based attribute learning, multiple images in a sequence can cope with the pose variation and occlusions which cause the absence of certain attributes in some frames, leading to more robust attribute learning and more accurate appearance information for the target identity. For the motion information enhancement, a generative model named Motion Enhancement Module (MEM) is designed to capture the walking style of a certain identity through predicting successive frames. The motivation is based on the hypothesis that if continuous frames can be correctly predicted, the model can successfully capture the latent specific walking style of one identity which is discriminative from others'. In this way, the motion information can be explicitly enriched and simultaneously, the properties of human walking activity can be fully studied in our proposed model. Despite a large-scale model the AMEM is in the training stage, only the backbone network plus two simple branches are kept for the final feature extraction during the inference, with both the appearance and motion feature enhanced while bringing about limited increment on computational complexity. We evaluate our AMEM on three public video-based datasets to verify the effectiveness of our ideas and demonstrate that it can achieve state-of-the-art performance compared with other video-based ReID methods.

The main contributions of our work are summarized as follows: (1) We propose a novel end-to-end trainable framework for video-based Person ReID called Appearance and Motion Enhancement Model (AMEM), which enriches both the appearance and motion information in the final feature representation. (2) We propose an Appearance Enhancement Module (AEM) which exploits video-based human attribute learning to improve the appearance learning for the backbone network. (3) We propose a generative model for the motion feature enhancement. By means of predicting consecutive frames, the model can capture the specific walking style of each identity. To our best knowledge, it is the first time that human walking styles are explicitly studied to provide id-discriminative information in person ReID field. (4) The performance of the final model achieves a large improvement without integrating complicated modules.

Related Works

Video-based Person ReID. Current video-based ReID approaches can be categorized into two classes, video-based methods and multi-image based methods. For video-based methods, the model takes the input tracklet as video data and tries to capture the temporal information in it. Li *et al.* (Li, Zhang, and Huang 2019) propose 3D convolutional network with multi-scale temporal convolutions and residual attention layers. Chung *et al.* (Chung, Tahboub, and Delp 2017) design a two stream siamese network which integrates the optical flow learning for timing information learning. However, the process of optical flow extraction is too slow and complicated that can hardly be put into practical use. All the methods above have the ability to capture the temporal cues in the video data, while none of them takes into consideration the special properties of human walking patterns.

For the other category, video data is usually taken as multiple independent images, where attention models are adopted for fusing multiple image features. Fu *et al.* (Fu et al. 2019) design the STA model to generate attention scores both intra and inter different frames, which are finally combined according to their scores. Zhang *et al.* (Zhang et al. 2019) propose the SCAN model including two types of attention modules, namely self-attention and collaborative attention network. These methods are with less computational complexity while they do not make full use of the abundant temporal information provided by video data.

Attribute Learning. Treated as one kind of middle level semantic feature, attributes have been widely studied in the image-based person ReID. Su *et al.* (Su et al. 2016) train an attribute learning model treating the deep attribute predictions as final representation. Deng *et al.* (Deng et al. 2014) have released a large-scale pedestrian attribute dataset PETA. Recently, a video-based attribute learning model is proposed by Zhao *et al.* (Zhao et al. 2019) which disentangles the frame feature into several sub-features for different groups of attributes. All the sub-features are aggregated in the temporal dimension to produce final predictions. Our work is similar to this while we put our main energy into the learning for human walking patterns.

Video Prediction. Visual forecasting is a vital part of computer vision. It has been exploited as a self-supervised method for video representation learning in many works recently. Srivastava *et al.* (Srivastava, Mansimov, and Salakhudinov 2015) try to learn the video representation in an unsupervised manner by predicting future frames apart from reconstructing previous frames. Walker *et al.* (Vaswani et al. 2017) generate future frames through predicting future human poses. In this paper, we focus on a special case among various activities – human walking, and seek to enhance the motion feature through prediction as well.

Approaches

In this section, an end-to-end trainable framework is formulated to enhance both the appearance and motion information in the backbone network. The overall architecture is illustrated in Fig. 1, where given an input walking sequence, it firstly goes through the backbone network to perform an initial feature extraction, then the two auxiliary modules named Appearance Enhancement Module (AEM) and Motion Enhancement Module (MEM) work independently to enrich the appearance and motion information contained in the final feature. Finally, only the backbone model and two small branches are kept for similarity estimation.

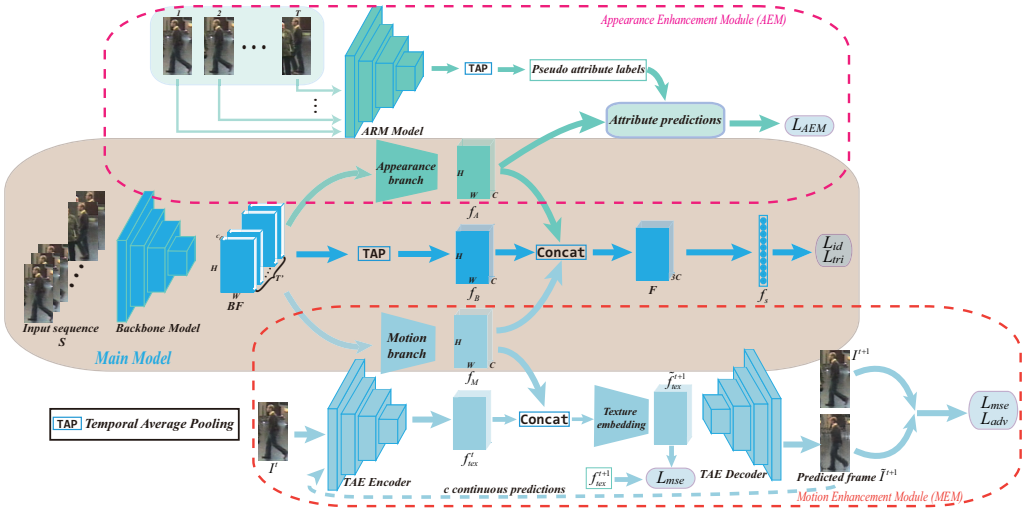

Figure 1: The overall architecture of our proposed AMEM. Only the *Main Model* is exploited during the inference. (Best viewed in color.)

Table 1: Some examples of the partitioned attribute groups in our proposed method. The attribute names are consistent with the names defined in PETA.

Group Name	Attribute names
Personal	Less15, ... , Larger60, Female
Hair	Black, ..., Short
UpperBody	Black, Casual, ..., Jacket
LowerBody	Formal, Shorts, ..., White
Footwear	Boots, Sneakers, ..., Blue
Accessory	Backpack, Suitcase, ..., Umbrella

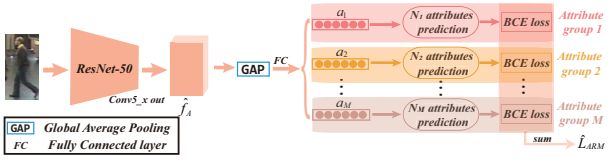

Figure 2: The overall architecture of the Attribute Recognition Model (ARM) (Best viewed in color).

Some necessary notations are firstly introduced. Let $S = \{I_1, I_2, ..., I_T\}$ and y denote the input video sequence with T continuous frames and its corresponding identity label in the dataset. Given an input sequence S, it is sent into a backbone network to perform feature extraction, whose process is denoted as a function $\boldsymbol{BF} = \phi(S, \theta_B), \boldsymbol{BF} \in \mathbb{R}^{C \times T' \times H \times W}$, where θ_B represents the network parameters.

Appearance Enhancement Module (AEM)

Human attributes are basic representations for human appearance and have been studied extensively as mid-level semantic features in recent years. Motivated by these, we propose AEM to take advantage of attributes to enrich the appearance and semantic information during model learning.

Pseudo Attribute Labels Generation Since there are no available video-based datasets containing attribute labels and it requires expensive labor to manually annotate, we seek to take advantage of existing large-scale attribute datasets to obtain a robust attribute recognition model (ARM) firstly. Concretely, we train a ResNet-50 model on the PETA dataset (Deng et al. 2014), which is a commonly used benchmark for person attribute recognition. As presented in Fig. 2, we use the output of Conv5_x block in ResNet-50 as the final feature map $\hat{f}_A$ for attribute recognition. Since the number of pre-defined attributes N is large ($N = 105$) which may cause ambiguous learning for the classifiers, we manually divide all attributes into M groups according to different locations or types like in (Zhao et al. 2019). Each group contains an attribute feature a_m and several types of attributes. The details of attribute groups can be referred to in Tab. 1. For each attribute feature a_m, it is acquired from $\hat{f}_A$ by applying a global average pooling layer and a fully connected layer, and is responsible for predicting corresponding attributes in the same group. Suppose there are N_m attributes in each group, the probability $\hat{p}_i$ of each attribute occurrence is predicted by applying a fully-connected layer and Sigmoid layer on a_m, then the loss in the m-th group is calculated by Binary Cross-Entropy (BCE) loss:

$$\hat{\mathcal{L}}_m = -\sum_{i}^{N_m} l_i^m \log \hat{p}_i + (1 - l_i^m) \log(1 - \hat{p}_i) \quad (1)$$

where l_i^m is the binary-value label of the i-th attribute in the m-th attribute group. The spatial size of $\hat{f}_A$ is set to $2048 \times 16 \times 8$ and the dimension of a_m is set to 256. The total loss for training ARM is defined by summing all the group ones:

$$\hat{\mathcal{L}}_{ARM} = \sum_{m=1}^{M} \hat{\mathcal{L}}_m \quad (2)$$

After obtaining a powerful attribute recognition model ARM, we take advantage of it to generate pseudo labels

Table 2: The structure of appearance and motion branch. ($out\_channel \times kernel\_size$).

Input (c=1024)			
branch 0	branch 1	branch 2	branch 3
384×1	192×1	48×1	MaxPooling
	384×3	128×3	128×1
Concat, Temporal Average Pooling, Output			

for each sequence in the video ReID dataset. Firstly, each frame in one sequence is sent into ARM to obtain image-level probabilities for all types of attributes. Secondly, for each attribute, we perform the temporal average pooling on the predicted probabilities of T frames to obtain the averaged prediction. Finally, if the averaged predicted probability of a certain attribute is beyond 0.5, we set its pseudo label $\hat{l}_i = 1$, else $\hat{l}_i = 0$. We denote the generated pseudo attribute labels as $\hat{l}_1, \hat{l}_2, \ldots, \hat{l}_N$.

Appearance Enhancement Next, the pseudo attribute labels are taken as the supervision for our AEM training. After acquiring ***BF***, we send it into an appearance branch to generate an attribute feature map $f_A \in \mathbb{R}^{C \times H \times W}$. The structure is listed in Tab. 2 which adopts the inception block in I3D (Carreira and Zisserman 2017). Each Conv-layer is followed by a batch normalization layer and a ReLU layer. Afterwards, M attribute features and N attribute predictions are obtained in the same way as in ARM. Finally, the attribute loss for the input sequence supervised by pseudo attribute labels is computed by:

$$\mathcal{L}_{AEM} = -\sum_{i=1}^{N} \hat{l}_i \log(p_i) + (1 - \hat{l}_i) \log(1 - p_i) \qquad (3)$$

where p_i is the predicted probability for the i-th attribute.

Motion Enhancement Module (MEM)

Albeit only single kind of human activity is contained in the human walking sequences, people belonging to different identities still possess different walking styles which can be easily distinguished by human beings. However, this subtle difference can not be easily captured by common video learning structures like RNN or 3D-Conv without specific design. To this end, we propose the MEM which employs a generative method for motion feature learning through predicting future frames. If frames can be successfully predicted, the model is considered to be able to capture the id-discriminative walking patterns.

Texture AutoEncoder The AutoEncoder structure (Hinton and Salakhutdinov 2006) is exploited to encode and decode the input human walking images, which is called Texture AutoEncoder (TAE) here. Like a common AutoEncoder, taking as the input of TAE a pedestrian image, the encoder will embed the image into a texture feature map $f_{tex} \in \mathbb{R}^{C \times H \times W}$, then the decoder will reconstruct the original image depending on it. The overall architecture of TAE is presented in Fig. 3. Concretely, we exploit ResNet-18 model (He et al. 2016) as the encoder, and the decoder is composed of four de-convolution blocks. Each block comprises one de-convolution layer with 3×3 kernel size followed by a batch normalization layer. ReLU layer is added except in the last block. Finally, a Sigmoid layer is applied to normalize the output image.

To obtain a self-contained texture of the pedestrian image, we pretrain the TAE on a large-scale pedestrian image dataset: Market-1501 (Zheng et al. 2015). The Mean Squared Error (MSE) loss is adopted for reconstruction task. In addition, as the images reconstructed by AutoEncoder often suffer from the blurry problem, for better visualization results, we add a discriminator D_{TAE} to judge whether the generated image is real or fake. The structure of D_{TAE} is the same as in DCGAN (Yu et al. 2017). In summary, the loss function for TAE training is defined by:

$$\begin{aligned} \hat{\mathcal{L}}_{TAE} = & \overbrace{\|\tilde{I} - \hat{I}\|_2}^{\hat{\mathcal{L}}_{mse}} \\ & + \underbrace{E_{\hat{I} \sim p_{\hat{I}}}[\log D(\hat{I})] + E_{f \sim p_f}[\log(1 - D(G(f_{tex})))]}_{\hat{\mathcal{L}}_{adv}} \end{aligned} \qquad (4)$$

where $\hat{I}$ is the input image to TAE, and $\tilde{I}$ is the reconstructed result of $\hat{I}$. G, D, f is short for the decoder of TAE, D_{TAE} and f_{tex}. $p_{\hat{I}}$ and p_f denote the sample distributions in the image and texture feature space. The D_{TAE} is optimized to maximize the $\hat{\mathcal{L}}_{adv}$ while the TAE is to minimize.

Motion Enhancement through Prediction After pretraining the TAE, it is used to help the motion feature extraction. We randomly select one frame I^t in the input sequence ($0 < t < T - c$) and predict the next frame I^{t+1} according to I^t. To simplify the task, we consider the I^{t+1} to be generated from the texture feature f_{tex}^{t+1} through the decoder of TAE. Furthermore, the texture feature f_{tex}^{t+1} can be disentangled into two components: the texture feature of current frame f_{tex}^t, and the motion feature f_M representing the movement between two continuous frames. In particular, the texture feature f_{tex}^t is extracted by the encoder of TAE, and the motion feature $f_M \in \mathbb{R}^{C \times H \times W}$ is provided by the backbone network $\phi(S, \theta_B)$. Same structure while different parameters compared with the appearance branch, a small motion branch is applied on ***BF*** to generate f_M. In this way, the backbone model can concentrate on exploring the motion properties while not being distracted by human appearance or cluttered background during the predicting process.

Then the texture feature f_{tex}^t is concatenated with the motion feature f_M and sent into a small texture embedding module to obtain f_{tex}^{t+1} for the next frame. The embedding module comprises two convolutional layers with kernel sizes of 3×3 and 1×1. A batch normalization layer is applied on each convolutional layer and a ReLU layer only on the first one. After that, f_{tex}^{t+1} is sent into TAE decoder for $\tilde{I}^{t+1}$.

However, predicting only one frame cannot capture the complete walking pattern of one person since the movement between two adjacent frames is often small and it usually

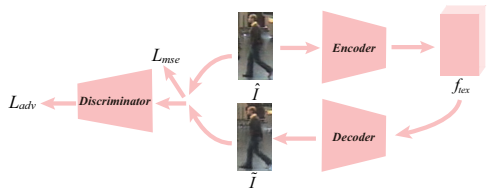

Figure 3: The overall architecture of the TAE.

takes several continuous frames for people to get the walking style and identify someone. In this case, we resolve to predict c continuous frames for better motion feature learning. In particular, after first round of prediction which generates the $\tilde{I}^{t+1}$, we take $\tilde{I}^{t+1}$ as the new "current frame" and predict $\tilde{I}^{t+2}$. The process is repeated for c times until the $\tilde{I}^{t+c}$. With the intention that the motion feature should capture the complete walking style in one sequence, we keep the predicting process at different time stamps sharing the same motion feature f_M. In this way, the model can be aware of a continuous motion pattern and successfully capture the global walking style of someone at the same time.

Similar to the TAE training, we apply the MSE loss and adversarial loss on the MEM training. While differently, to preserve the semantic texture for a better prediction, a texture MSE loss is added which measures the similarity between the predicted and real texture feature ($\tilde{f}_{tex}^{t+1}$ and f_{tex}^{t+1}). In summary, the total loss for MEM training is computed by:

$$\mathcal{L}_{MEM} = \sum_{0<t<T-c}^{t+c} \overbrace{\|\tilde{I}^{t+1} - I^{t+1}\|_2 + \|\tilde{f}_{tex}^{t+1} - f_{tex}^{t+1}\|_2}^{\mathcal{L}_{mse}} + \underbrace{E_{I\sim p_I}[\log D(I^{t+1})] + E_{f\sim p_f}[\log(1 - D(G(\tilde{f}_{tex}^{t+1})))]}_{\mathcal{L}_{adv}} \tag{5}$$

where G, D, f is short for the TAE decoder, Discriminator and f_{tex}, respectively. The parameter c is experimentally set to 3 in this paper. Moreover, during training, the parameters of the TAE encoder are frozen for better texture extraction, and the decoder is fine-tuned to make better predictions.

Intuitively, the MEM proposed in our model can be interpreted from two other aspects. On one hand, the prediction task adopted in MEM can be taken as a special kind of reconstruction, since it is conducted inside the original sequence while not predicting for unseen frames. While different from existing works (Srivastava, Mansimov, and Salakhudinov 2015), we conduct the "reconstruction" in the forward order and start from a randomly selected frame, which seems to be a combination of reconstruction and prediction. On the other hand, capturing the movement between two successive frames seems similar to the optical flow. However, we mainly focus on the movement of human walking while the optical flow tries to precisely capture all possible differences between frames and may easily be distracted by various changes in cluttered background or other objects. Moreover, the accurate estimation of optical flow itself is still a challenging task with high computation complexity.

The walking patterns also share many similarities with human gaits. Nevertheless, due to the strong assumptions on input data and specific scenario settings used in the gait recognition field, none of current gait recognition methods can be extended to common person ReID tasks. In this way, we propose the MEM for learning human gaits on common pedestrian data. To our best knowledge, there are no existing works leveraging frame prediction in ReID and for walking patterns learning, also the prediction results can serve as an evaluation for the quality of motion feature learning.

Optimization

To integrate the enhanced appearance and motion information into our backbone network to improve the performance of person re-identification, we concatenate the f_A and f_M with f_B along the channel dimension to constitute the final feature map $\boldsymbol{F}$. The $f_B \in \mathbb{R}^{C\times H\times W}$ is the backbone feature acquired from $\boldsymbol{BF}$ by adding a temporal average pooling layer. Followed by a global average pooling layer and a fully-connected layer, the final feature representation f_s for the whole sequence is acquired. The whole model containing the backbone network, AEM and MEM is trained in end-to-end manner. Apart from the $\mathcal{L}_{AEM}$ and $\mathcal{L}_{MEM}$, the identity softmax loss $\mathcal{L}_{id}$ and triplet loss $\mathcal{L}_{tri}$ are applied to f_s for ReID task. The total loss for the whole model is defined by:

$$\mathcal{L}_{total} = \mathcal{L}_{id} + \mathcal{L}_{tri} + \lambda_A\mathcal{L}_{AEM} + \lambda_M\mathcal{L}_{MEM} \tag{6}$$

$$\begin{aligned} \mathcal{L}_{id} &= -\frac{1}{L}\sum_{i=1}^{L} y_i \log(q_i) \\ \mathcal{L}_{tri} &= \frac{1}{K}\sum_{i=1}^{K}[d_i^p - d_i^n + k]_+ \end{aligned} \tag{7}$$

where L and K are the numbers of samples and triplets in one batch. y_i and q_i are the identity label and predicted probability for the target identity. $[*]_+ = \max(*, 0)$ is the hinge loss, d_i^p, d_i^n denote the feature distance in positive and negative pairs, k is the margin to separate them. λ_A and λ_M are determined by cross-validation to balance different losses.

During the inference stage, only the backbone network plus the two small branches for appearance and motion feature extraction are preserved, as shown in the *Main Model* block in Fig. 1. f_s is used for the similarity evaluation after the L2 normalization. The model with appearance and motion information enriched largely improves the performance of re-identification while the increment on computational complexity is limited.

Experiments

In this section, we report the experimental results on several standard datasets and a detailed ablation study is conducted over different modules of the AMEM. Extensive experiments on three widely used benchmarks demonstrate that our approach can achieve state-of-the-art performances.

Implementation Details

We implement our proposed algorithm based on PyTorch framework on two GTX 1080Ti GPUs with 11GB memory. We adopt the I3D model (Carreira and Zisserman 2017)

Table 3: Ablation study on each module of our proposed method on MARS.

Models	MARS			
	rank-1	rank-5	rank-20	mAP
Baseline (I3D)	83.3	93.3	96.4	75.1
Baseline + AEM	86.0	93.6	96.7	78.4
Baseline + MEM (c=1)	85.4	94.1	97.0	78.3
Baseline + MEM (c=2)	86.0	94.0	96.9	78.7
Baseline + MEM (c=3)	86.2	**94.1**	**97.2**	79.0
Baseline + MEM (c=4)	85.6	94.1	97.1	78.6
Baseline (After Enhancement)	85.0	93.7	96.7	77.3
Baseline + AEM + MEM (AMEM)	**86.7**	94.0	97.1	**79.3**

Table 4: Experimental results conducted on different backbone networks. 'B' stands for Baseline and 'A', 'M' for AEM and MEM respectively.

Backbones	Ranks	MARS			
		B	B+A	B+M	AMEM
R3D	rank-1	56.1	61.7	61.8	62.8
	mAP	41.5	46.6	48.5	49.1
P3D	rank-1	68.8	73.5	74.0	74.5
	mAP	54.3	61.2	62.6	63.1
I3D	rank-1	83.3	86.0	86.2	**86.7**
	mAP	75.1	78.4	79.0	**79.3**

which is pretrained on Kinetics (Kay et al. 2017) as our backbone network. The whole network is optimized using Adam optimizer in an end-to-end manner. The initial learning rate is set to 1e-3, and decreased by 0.2 every 60 epochs. The weight decay is set to 5e-4. The length of the input sequence T is empirically set to 8. The input frames are resized to 256×128. The sizes of the feature maps in our model are set to $H = 16, W = 8, C = 1024, T' = 3$, and the dimension of the final feature f_s is set to 512. The hyper-parameters k, λ_A, λ_M are set to $0.2, 0.1, 10$ respectively.

Datasets and Evaluation Metric

MARS (Zheng et al. 2016) is currently one of the largest video-based person re-identification datasets, which consists of 1261 identities and around 20000 human walking sequences. Among them, 625 identities are used for training and 8298 tracklets of the rest 636 identities are for testing.

iLIDS-VID (Wang et al. 2014) is composed of 600 sequences belonging to 300 different pedestrians from two non-overlapping cameras. The sequence length is varied from 23 to 192 frames, which has 73 frames on average.

PRID-2011 (Hirzer et al. 2011) consists of 749 different identities from one camera, and 385 identities from the other, with only the first 200 people appear in both cameras. Each sequence has a length between 5 and 675 frames.

Following the evaluation metrics widely adopted, we adopt cumulative matching characteristics (CMC) for evaluating the three datasets. Besides, for MARS which have multiple ground truths in the gallery, we also report mAP scores. We follow the original splits provided by MARS, and for iLIDS-VID and PRID-2011, we follow the evaluation protocol from previous works (Wang et al. 2014) where the dataset is randomly split into the train/test set for 10 times, then the averaged accuracies are reported.

Ablation study

Baseline Comparison In Tab. 3, we list the performance of the baseline model and the model after the enhancement. From the results we can find that, either adding the AEM or MEM can boost the performance of the baseline model, which demonstrates the effectiveness of our proposed modules for enriching the appearance and motion information in the backbone network. Furthermore, by integrating both modules, the best performance is achieved.

We also find that compared with the AEM, the MEM can bring more improvement on the performance, especially on rank5, 20 and mAP. The reason may be that the multiple images contained in a sequence have provided enough information to cope with different appearance variations. In this way, the promotion brought by attribute learning may be limited. While for motion information extraction, the baseline model cannot capture the special patterns contained in human walking styles without MEM. Therefore, the MEM can lead to full promotion for the final performance.

Moreover, to intuitively demonstrate the impact of AEM and MEM on the backbone feature f_B, we further made an experiment on not concatenating the f_A and f_M with f_B, which means only the f_B is exploited both in ReID task training and testing stage. In this way, the inference model can discard the two branches and has the same structure as the baseline model, whose results are shown in the second last row of Tab. 3. The improved results compared to the initial baseline also verified that the backbone feature f_B has benefited from our designed AEM and MEM.

Scalability To empirically demonstrate the scalability of our designed AEM and MEM for the appearance and motion enhancement. We conducted an experiment on exploiting different backbone networks. Since our model involves both spatial and temporal feature extraction, we choose some variants of 3D-Conv models apart from the adopted I3D, including 3D-ResNet (R3D) (Hara, Kataoka, and Satoh 2018) and Pseudo 3D (P3D) (Qiu, Yao, and Mei 2017) networks. The experimental results are listed in Tab. 4, where the proposed AEM and MEM improve the performances for all three different backbones. The results reveal that our designed AMEM can be easily integrated into different backbone networks to mine the appearance and motion information contained in video data, which verified its scalability. The reason that R3D and P3D show inferior performance to the I3D model may lie in the huge amount of parameters in R3D and P3D which cause over-fitting for the training set.

Model Size We also compare the model size between the baseline model and our final model. The model size of the backbone network is 50.2MB, and 67.7MB for the final model. The increment of model size mainly comes from the appearance and motion branches. Thanks to the special design of 3D Inception block in I3D, it can still maintain a relatively small size after adding two small branches. We also did an experiment on designing the two branches with other structures. Using one simple 3D-Conv layer brings about 90MB increment for one branch, and 32.6MB when using two 2D-Conv layers. If more complicated modules like at-

Table 5: Comparisons with state-of-the-arts methods on several datasets.

Models	MARS				iLIDS-VID			PRID-2011		
	rank-1	rank-5	rank-20	mAP	rank-1	rank-5	rank-20	rank-1	rank-5	rank-20
CNN+XQDA (Zheng et al. 2016)	68.3	82.6	89.4	49.3	53.0	81.4	95.1	77.3	93.5	99.3
RQEN (Song et al. 2018)	77.8	88.8	94.3	71.1	76.1	92.9	99.3	92.4	98.8	100.0
CSA-CSE (Chen et al. 2018)	81.2	92.1	-	69.4	79.8	91.8	-	88.6	**99.1**	-
DR-STAN (Li et al. 2018)	82.3	-	-	65.8	80.2	-	-	93.2	-	-
3D-NLA (Liao et al. 2018)	84.3	94.6	-	77.0	81.3	-	-	91.2	-	-
3D PersonVLAD (Wu et al. 2019)	82.8	**94.8**	**99.0**	64.7	70.7	88.2	99.2	88.0	96.2	99.7
STMP (Liu et al. 2019)	84.4	93.2	96.3	72.7	84.3	96.8	99.5	92.7	98.8	99.8
M3D (Li, Zhang, and Huang 2019)	84.4	93.9	97.8	74.1	74.0	94.3	-	91.0	-	-
SCAN (Zhang et al. 2019)	86.6	**94.8**	97.1	76.7	81.3	93.3	**100.0**	92.0	98.0	100.0
Baseline	83.3	93.3	96.4	75.1	82.6	96.3	99.0	88.8	98.1	99.8
Ours (AMEM)	**86.7**	94.0	97.1	**79.3**	**87.2**	**97.7**	99.5	**93.3**	98.7	**100.0**

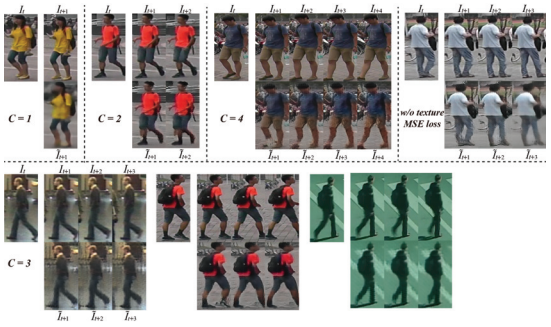

Figure 4: The randomly selected samples of frame prediction from different datasets and settings. The first row of each sample presents the original images and the second row presents the generated ones (Best viewed in color).

tention models are designed and integrated, the model size will be further increased. In summary, our proposed algorithm can achieve enhancement on both appearance and motion fields while maintaining a light-weight model.

The Number of Frames Prediction We empirically study the number of frames prediction in our MEM. The visualization results and accuracies are shown in Fig. 4 and Tab. 3 respectively. The simplest case is to predict only one future frame ($c = 1$), where the MEM can make correct but blurry predictions and the accuracy is relatively inferior. The reason we analyze is that predicting only one frame is not enough for MEM to extract the walking pattern of someone since it usually takes several frames for human beings to identify. In this way, we consider to predict more continuous frames and started from $c = 2$, both the generation results and the test performance are further improved, which verified our analysis. However, when $c = 4$, the rank1 accuracy falls down again, which may be due to the vanishing gradient during the too long range of prediction. Finally, to balance the performance and efficiency, we adopted $c = 3$ in our model.

Different Components in the Loss Function We empirically analyze the different components in our loss function. The $\mathcal{L}_{id}$ and $\mathcal{L}_{tri}$ are commonly used losses for person ReID tasks. For $\mathcal{L}_{AEM}$, it is the attribute recognition loss which helps enrich the appearance information in the final model. In $\mathcal{L}_{MEM}$ there are three elements, of which the most important is the MSE loss for prediction. It provides the supervision for our model to capture the walking patterns contained in the input sequences. The MSE loss for texture and the adversarial loss mainly aim at achieving sound generation results. Take texture-MSE loss as an example, we did an experiment on removing it and the visualization results are shown in Fig. 4. From which we can find that the model can still make predictions while the texture information decreases over time, leading to the blurry and noisy generation results. Since it can capture the walking patterns, the final performance is similar to our final model thus is not listed here. In conclusion, removing $\mathcal{L}_{adv}$ or $\mathcal{L}_{mse}$ for texture will not bring too much impact on the ReID accuracy, whereas the quality of generated images drops a lot.

Comparison with the State-of-the-arts

In this section, we present the comparison with several state-of-the-art algorithms which are listed in Tab. 5. From the results we can observe that, our method can achieve the best performances on all three datasets, with nearly the minimum complexity increment on the backbone network. Similar 3D-Conv based works like M3D and 3D-NLA all aimed at modifying the Conv operations to enhance the spatial and temporal feature learning, while were hard to discover more latent information like the human attributes and walking patterns. Our method outperforms M3D by 2.3%@rank1 on PRID-2011 and despite a spatial feature stream is further added in M3D, AMEM can still outperform it by 2.3%@rank1, 13.2%@rank1 on MARS and iLIDS-VID respectively. Compared with multi-images based models CSA-CSE and SCAN, which designed complicated attention modules to aggregate frame features, our model still perform better since the temporal cues contained in the sequences are explicitly studied by the designed MEM.

Conclusion

In this paper, an Appearance and Motion Enhancement Model (AMEM) for video-based person ReID is proposed aiming to simultaneously enrich the two types of information contained in the final feature. The appearance information is enhanced by the human attribute information and the motion information is enriched through future frames prediction. We demonstrate that, apart from the appearance information can be improved by attribute recognition which has been verified in image-based person ReID, the proposed MEM can capture the id-discriminative walking pat-

terns which were hardly studied in previous works. The final model used for inference has similar complexity to the backbone network while with both appearance and motion information enhanced. Experiments on three popular video-based ReID datasets verified the effectiveness of our model. In future work, we will dig more into understanding human recognition mechanism for re-identifying people, including but not limited to human attributes or walking patterns.

Acknowledgments

This work is sponsored by Zhejiang Lab (No.2019KD0AB01), Sichuan Science and Technology Project (NO.2019YJ0680), Tencent IEG VASD, Artificial Intelligence Research Foundation of Baidu Inc, the National Natural Science Foundation of China (NO.61471321).

References

Carreira, J., and Zisserman, A. 2017. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 6299–6308.

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 *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1169–1178.

Chung, D.; Tahboub, K.; and Delp, E. J. 2017. A two stream siamese convolutional neural network for person re-identification. In *Proceedings of the IEEE International Conference on Computer Vision*, 1983–1991.

Deng, Y.; Luo, P.; Loy, C. C.; and Tang, X. 2014. Pedestrian attribute recognition at far distance. In *Proceedings of the 22nd ACM international conference on Multimedia*, 789–792. ACM.

Fu, Y.; Wang, X.; Wei, Y.; and Huang, T. 2019. Sta: Spatial-temporal attention for large-scale video-based person re-identification. In *Proceedings of the Association for the Advancement of Artificial Intelligence*.

Hara, K.; Kataoka, H.; and Satoh, Y. 2018. Can spatiotemporal 3d cnns retrace the history of 2d cnns and imagenet? In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 6546–6555.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 770–778.

Hinton, G. E., and Salakhutdinov, R. R. 2006. Reducing the dimensionality of data with neural networks. *science* 313(5786):504–507.

Hirzer, M.; Beleznai, C.; Roth, P. M.; and Bischof, H. 2011. Person re-identification by descriptive and discriminative classification. In *Scandinavian conference on Image analysis*, 91–102. Springer.

Kay, W.; Carreira, J.; Simonyan, K.; Zhang, B.; Hillier, C.; Vijayanarasimhan, S.; Viola, F.; Green, T.; Back, T.; Natsev, P.; et al. 2017. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*.

Li, S.; Bak, S.; Carr, P.; and Wang, X. 2018. Diversity regularized spatiotemporal attention for video-based person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 369–378.

Li, J.; Zhang, S.; and Huang, T. 2019. Multi-scale 3d convolution network for video based person re-identification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 8618–8625.

Liao, X.; He, L.; Yang, Z.; and Zhang, C. 2018. Video-based person re-identification via 3d convolutional networks and non-local attention. In *Asian Conference on Computer Vision*, 620–634. Springer.

Liu, Y.; Yuan, Z.; Zhou, W.; and Li, H. 2019. Spatial and temporal mutual promotion for video-based person re-identification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 8786–8793.

Qiu, Z.; Yao, T.; and Mei, T. 2017. Learning spatio-temporal representation with pseudo-3d residual networks. In *proceedings of the IEEE International Conference on Computer Vision*, 5533–5541.

Song, G.; Leng, B.; Liu, Y.; Hetang, C.; and Cai, S. 2018. Region-based quality estimation network for large-scale person re-identification. In *Proceedings of AAAI Conference on Artificial Intelligence*.

Srivastava, N.; Mansimov, E.; and Salakhudinov, R. 2015. Unsupervised learning of video representations using lstms. In *International conference on machine learning*, 843–852.

Su, C.; Zhang, S.; Xing, J.; Gao, W.; and Tian, Q. 2016. Deep attributes driven multi-camera person re-identification. In *European Conference on Computer Vision*, 475–491.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, 5998–6008.

Wang, T.; Gong, S.; Zhu, X.; and Wang, S. 2014. Person re-identification by video ranking. In *European conference on computer vision*, 688–703. Springer.

Wu, L.; Wang, Y.; Shao, L.; and Wang, M. 2019. 3-d personvlad: Learning deep global representations for video-based person reidentification. *IEEE transactions on neural networks and learning systems*.

Yu, Y.; Gong, Z.; Zhong, P.; and Shan, J. 2017. Unsupervised representation learning with deep convolutional neural network for remote sensing images. In *International Conference on Image and Graphics*, 97–108. Springer.

Zhang, R.; Li, J.; Sun, H.; Ge, Y.; Luo, P.; Wang, X.; and Lin, L. 2019. Scan: Self-and-collaborative attention network for video person re-identification. *IEEE Transactions on Image Processing*.

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 *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 4913–4922.

Zheng, L.; Shen, L.; Tian, L.; Wang, S.; Wang, J.; and Tian, Q. 2015. Scalable person re-identification: A benchmark. In *Proceedings of the IEEE International Conference on Computer Vision*, 1116–1124.

Zheng, L.; Bie, Z.; Sun, Y.; Wang, J.; Su, C.; Wang, S.; and Tian, Q. 2016. Mars: A video benchmark for large-scale person re-identification. In *European Conference on Computer Vision*, 868–884. Springer.