Variational Discriminative Aggregation for Video Person Re-identification

Anonymous ECCV submission

Paper ID 3875

Abstract. Video-based person re-identification (ReID) aims at retrieving the video clips of the same person across non-overlapping cameras. Most existing approaches apply various attention mechanisms to aggregate the features of multiple frames to learn the representation of the image sequence for the following retrieval. However, such formulations give less consideration to the structure of the sets of videos, which is useful information to the sets matching problem like video-based ReID. Hence, we propose to utilize the variational autoencoder (VAE) to explicitly capture the structural information through the reconstruction of the feature sets. The structure information can, therefore, be encoded into the intermediate latent embedding of VAE. Given the nature of the identification problem, besides the reconstruction loss and KL-divergence used by ordinary VAE model, we further employ discriminative loss to supervise the learning of the intermediate latent embeddings. As a result, we can obtain the representation of the image sequence with high structural information and the discriminative ability for the retrieval. Extensive experiments on three benchmarks strongly demonstrate our intuition by achieving state-of-the-art results.

Keywords: Video Person Re-identification, Variational Autoencoder

1 Introduction

Recent years have witnessed the great development of deep learning, which has brought tremendous improvement in many areas in computer vision, such as person re-identification (ReID). Person re-identification aims at identifying a query individual from a large set of candidates under the non-overlapping camera views. It plays an important role in applications like security and surveillance, thus it has been studied extensively in the past few years [23, 24, 43, 51, 61]. However, the studies in this task still suffer from many negative factors like variations of viewpoints and pose, occlusion or lightning conditions because the information in a single image is limited to fully overcome these obstacles. Hence, many works turn to the extension of image-based person re-identification where they match pairs of image sequences instead of only a single image, namely video-based person re-identification.

The main difference of video-based ReID, compared to image-based one, lies in its demand for learning the temporal correlation among different frames

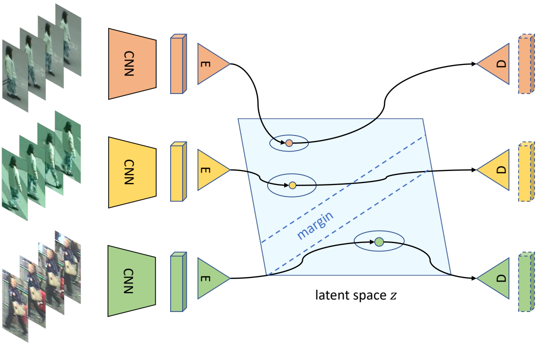

Fig. 1. The key idea of our method. We propose to utilize variational autoencoder (VAE) to explicitly capture the structural information through the reconstruction of the feature sets. The structure information can, therefore, be encoded into the intermediate latent embedding of VAE. A discriminative loss is also added to boost person re-identification.

in the video sequence. The most intuitive way of temporal modeling is just simply aggregating the learned features of each frame in the sequence by using max/mean pooling [27, 63]. Such formulation totally omits the ordering of the frames in video sequences. Therefore, many recent works [2, 4, 22] attempt to design various attention mechanism to learn how to aggregate the information across time by assigning different weights to different frames. However, in such a way we can not explicitly model the global structural information of the features extracted from all frames in the sequence. Intuitively, the features of each frame in the sequence reside on a manifold in the high-dimensional space and the intrinsic structure in this space implies how features interact with other features, which can not be captured by state-of-the-art attention-based approaches.

Given the useful information provided by the knowledge of the set structure, it is important to apply this structural information to identify whether two sequences are matching with each other. Motivated by this idea, we propose, in this paper, to explicitly model the set structure in the feature embedding space to learn the representation for the video sequence, which can be used for the following retrieval step. Specifically, we adopt variational autoencoder (VAE), a commonly used generative model, to capture the underlying structure of the sequence in the embedding space. VAE encodes the information of the structure into an intermediate latent embedding, whose prior is assumed to be the unit multivariate normal distribution, for the need for the information of the structure for the reconstruction process of the sequence. Nevertheless, noting that different from reconstruction task whose goal is to faithfully reconstruct the data distribution, ReID emphasizes more on the identification, that is to say, we expect the learned VAE can capture more of the discriminative property of the sets and reconstruct the *useful* data distribution. To this end, we also employ the classification loss as a part of the reconstruction loss to the whole generative model. Moreover, the learned intermediate latent embedding is what we will use for the following identification step, so discriminative loss is adopted to supervise the learning of intermediate latent embedding to directly boost the discriminative ability of the latent space. Specifically, we utilize metric learning techniques, namely triplet loss [35] or N-pair loss [37], as discriminative loss to supervise the intermediate latent embedding with the goal to draw the latent representations of the same person near to each other while the ones coming from different subjects far from each other. Figure 1 shows the key idea of our method.

Aside from the newly-proposed supervision signal, given that different from single image, video clip often contains the motion information of the moving subjects inside it, we also devise a new two-stream network pipeline to learn the motion-aware representation for each sequence. Our newly-proposed model takes the image sequence and optical flow as input for two independent CNN [20] to extract appearance feature and motion feature respectively. We then merge the features from two streams by using element-wise addition. The fused feature is finally fed into a high-level ConvNet to obtain the global feature map for the image sequence, which is then processed by the variational discriminative aggregation (VAD) module described above. Note that different from [2] and [52] which directly combine optical flow and image as the input, our method designs a unique optical flow stream and two streams in the middle level, which effectively utilizes the optical flow information.

To demonstrate the effectiveness of our approach described above, we conduct experiments on MARS [60], DukeMTMC-VideoReID [50], and iLIDS-VID [48], which are popular benchmark for the video-based person re-identification task. Our model improved the performance on Mars, DukeMTMC-VideoReID, and iLIDS-VID compared to state-of-the-art method. The experimental results strongly advocate our intuition. An ablation study is also done thoroughly to prove the necessity of each component in our method. Code for our model will be made publicly available.

We summarize the contributions of this work as follows:

1) We propose a variational discriminative aggregation method to explicitly model the structure of the feature space of the image sequence to learn the representation of the sequence by employing variational autoencoder and discriminative loss.
2) We learn a motion-aware representation which extracts coarse-grained appearance and motion information by a two-stream network and fuses them by the following network.
3) The experimental results demonstrate the superiority of our method, which outperforms the state-of-the-art methods on several large scale datasets including Mars, DukeMTMC-VideoReID, and iLIDS-VID.

2 Related Work

Image-based Person Re-identification: Most previous image-based person ReID methods focus on designing effective deep neural networks, person image preprocessing, and capturing extra supervisory signals. To improve the effectiveness of model, part model [4, 42, 43, 46] are proposed for learning local information, and attention model [1, 3, 36] are also widely used to address feature misalignments. To reduce the intraclass variance, many works pre-process the person image by cleaning backgrounds [38] or generative adversarial networks [10,32,62]. Besides, some other methods attempt to leverage additional supervisory signals, including human parsing [14, 56, 57], pose estimation [32, 40],and attribute labels [44, 58] or monitor camera information [45].

Video-based Person Re-identification: Compared with image data, video sequences contain continuously moving person samples, which provide the obvious motion characteristics and sufficient sample diversity. Many existing works focus on learn the discriminative motion features to assist appearance representation. For instance, many previous methods [15, 16, 47, 54] employ handcraft motion features like HOG3D [18]. Some methods [2, 27, 30] apply optical flow to extract the motion information to capture dynamic gist. Inspired by the success of deep learning, the recurrent neural neural [29, 53, 63] and 3D convolution network [25] have been employed to learn the motion feature of person videos. Beside, the aggregation of video sequence embeddings is another popular research field. The most common manner of embedding aggregation is average pooling [29, 63]. Taking average pooling as the baseline, attention pooling [2,4,28,52] have been proposed to adaptively select key frames. Besides, some works formulate the frame selection as a markov decision process and use reinforcement learning to sequentially discard confounding frames [31, 33]. Different with these aggregation methods, we propose to aggregate the image set while preserving the structure information as well as the discrimiative nature of the set in the feature space by our proposed variational discriminative aggregation method.

Variational Autoencoder: Variational autoencoder is a commonly used generative model, which is first proposed by et al. [17]. It can learn complex density model for data via latent variables. Specifically, VAE employs an encoder that represents the input as a latent variable with Gaussian distribution assumption and an another decoder that reconstructs the input from the latent variable. VAE can be widely applied to many tasks to learn the distribution of the input data as its representation. VAE has been widely used in recent unsupervised learning researches as a highly expressive model. Aside from image synthesis, VAE is also adopted in many other tasks, such as zero-shot learning [19], 3D representation learning from point cloud [11] and metric learning [26]. In this work, we utilize VAE model to learn the representation for the video clips that encodes the structural information for the need of reconstruction.

3 Approach

In this section, we will describe our approach in detail. Suppose we have an image sequence now, our pipeline firstly calculates the optical flow of the sequence,

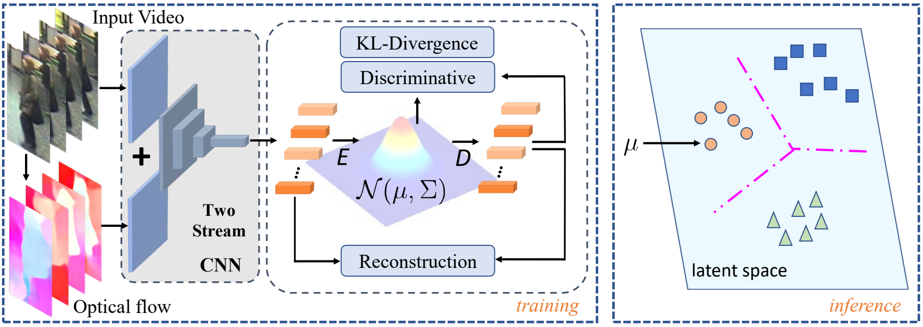

Fig. 2. Our proposed VDA framework. We extract a feature of size R^C for each frame in the input video and take them as input for the VDA module. The encoder in VDA firstly maps the input feature map of size $R^{T \times C}$ for the whole sequence to mean μ and variance Σ which represent the structure of the input image set. Then we can sample from the gaussian distribution defined by μ and Σ by applying the reparametrization trick and use the decoder to reconstruct the input feature set for the video clip. Aside from the reconstruction loss and KL-divergence as ordinary VAE models, the discriminative loss is also employed to supervise the learning of the VDA module. During inference phase, the final decision for retrieval is based on the distance between learned representation μ of different video clips in the latent embedding space.

then both of them are fed into a two-stream network before they are merged by a following high-level ConvNet to extract the global features for the image sequence. The learned features are then taken by the variational discriminative aggregation module, which is encoder-decoder architecture, as input to obtain the representation for the video clips, which is the output of the encoder. Finally, we use this representation to perform the retrieval. The overall framework of our method is illustrated in Figure 2.

3.1 Motion-Aware Representation Learning

Different from the image-based person ReID, the video data contains motion characteristics like gait or speed, which can be used to facilitate to learn a representation of higher quality. To extract the motion-aware person representation, we propose a two-stream representation network which respectively captures the motion and appearance features and fuses them into the final embeddings. We employ two CNN models to learn the coarse-grained motion and appearance features, whose inputs are optical flows and image sequences respectively. The input channels of two streams are 2 (vertical and horizontal channels) and 3(RGB channels) for optical flow and image data while the output see channels of two streams are the same for efficient information fusion. In each batch, we feed the video clips and the corresponding optical flows into our two-steam network and fuse the outputs by direct element-wise addition. The optical flows between adjacent frames are calculated by Flownet [7]. Note that we repeat the last optical flow map in each video clips to avoid the misalignment of the video length, due to that the length of optical flows is always one less than original video frames for the frame-by-frame difference. After extracting the motion feature and appearance feature of a video clip, we then fuse them by element-wise addition and fed into the following high-level ConvNet to obtain the global feature map of the image sequence. Then for each frame, we perform simple mean pooling strategy to aggregate information across the spatial domain to get the final feature map for the whole image sequence. Given a video clip of a pedestrian $X \in R^{3\times T\times H\times W}$, its corresponding optical flows F should be of the size $R^{2\times T-1\times H\times W}$, the final feature map $g \in R^{C\times T}$ can be expressed as:

$$g = \text{mean}_{H,W}(\mathcal{F}_g(\mathcal{F}_{RGB}(X) \oplus \mathcal{F}_{flow}(F))) \tag{1}$$

where $\mathcal{F}_{RGB}$ and $\mathcal{F}_{flow}$ represent appearance stream network and motion stream network respectively, while $\mathcal{F}_g$ represents the high-level ConvNet The final obtained feature map g is then processed by our proposed variational discriminative aggregation module to extract the representation for the whole sequence.

3.2 Variational Discriminative Aggregation

We aim at aggregating features of different frames to a discriminative video representation while keeping the structures information of the video for recognition. To achieve this goal, we propose a variational discriminative aggregation (VDA) network to model the structural information of a video and perform temporal feature aggregation. Before introducing our method, we start by reviewing the variational autoencoder (VAE) [17], on which our method is built. Variational inference aims at capturing the true conditional probability distribution over the latent variables $p_\phi(z|x)$, where z the latent variables and x is the input observation. Since perfectly capturing this distribution is quite intractable, this distribution can be approximated by finding its closet proxy posterior $q_\theta(z|x)$ by minimizing the distance between them using a variational lower bound limit. Therefore, the objective function of a VAE can be represented as the variational lower bound on the marginal likelihood of a given observation x, which can be written as:

$$\mathcal{L}_{\text{VAE}} = \mathbb{E}_{q_\phi(z|x)}[\log p_\theta(x|z)] - D_{KL}(q_\phi(z|x)||p_\theta(z)), \tag{2}$$

where the first term is the reconstruction error and the second term is the Kullback-Leibler divergence between the inference model $q(z|x)$ and $p(z)$. Following common practice [17], the prior $p(z)$ is a multivariate standard Gaussian distribution. To capture the latent distribution, a neural network encoder can be trained to predict μ and $\varSigma$ of this Gaussian distribution such that $q_\phi(z|x) = \mathcal{N}(\mu, \varSigma)$. Based on this distribution, latent vectors z can generate using the reparametrization trick [17].

Different from the original VAE model that aims to capture data distribution, our goal is to learn discriminative embeddings to improve recognition performance. Therefore, we combine the idea of deep metric learning with the VAE framework to produce embeddings that can simultaneously capture the video structural information and improve the discriminative power of the original features. Specifically, given a set of video features $g_1, ..., g_T$, an encoder model E is proposed to aggregate features from different frames to the mean and variance of the distribution:

$$\mu, \Sigma = E(g_1, g_2, ..., g_T). \tag{3}$$

Based on this distribution, we can further sample M features and employ a decoder model D to obtain the reconstructed video frame features:

$$g'_i = D(z_i), z \sim \mathcal{N}(\mu, \Sigma), i = 1, 2, ..., M. \tag{4}$$

Since our goal is to capture the most discriminative information from input frames, we do not use the original VAE reconstrcution error that enforces the autoencoder to faithfully reconstruct the input data. We propose to reconstrcut the useful part of the input data, which can be formulated as:

$$\mathcal{L}_{\text{recon}} = ||\frac{1}{M}\sum_{i=1}^{M} g'_i - \frac{1}{M}\sum_{i=1}^{M} g_i||_2 + \frac{1}{M}\sum_{i=1}^{M}\mathcal{L}_{\text{cls}}(g'_i), \tag{5}$$

where $\mathcal{L}_{\text{cls}}$ is the classification loss. The first term of the reconstruction loss aims to align the first-order moment of the input feature distribution and the reconstructed feature distribution, which serves as the basic supervision signal for reconstruction. The second term encourages the network to focus on features that are more discriminative and beneficial for classification task.

To further improve the discriminative ability of the aggregated feature, we add a discriminative loss to the mean of the learned distribution like triplet loss [35] or N-pair loss [37]:

$$\mathcal{L}_{\text{triplet}} = \max(\alpha + ||\mu_a - \mu_p||_2 - ||\mu_a - \mu_n||, 0), \tag{6}$$

or

$$\mathcal{L}_{\text{N}-\text{pair}} = \log(1 + \sum_i(\mu_a^\top \mu_p - \mu_a^\top \mu_i) \tag{7}$$

where α is the margin value, μ_a, μ_p, μ_n denote the anchor, the postive, the negative samples and μ_i is the i-th negative sample in N-pair loss. For the sake of simplicity, we use the triplet loss as an example in the following descriptions.

We can add the classification loss to form the discriminative loss:

$$\mathcal{L}_{\text{dis}} = \mathcal{L}_{\text{triplet}} + \mathcal{L}_{\text{cls}}(\mu). \tag{8}$$

By combining VAE loss and the discrimniative loss, we arrive at the final objective of VDA:

$$\mathcal{L}_{\text{VDA}} = \mathcal{L}_{\text{dis}} + \mathcal{L}_{\text{recon}} - D_{KL}(q_\phi(z|x)||p_\theta(z)). \tag{9}$$

During inference, we directly use the μ as the aggregated representation of the input video. Note that although the distribution information is not explicitly used during inference, the reconstruction loss in VAE framework ensures the representation ability of arbitrary feature inside the distribution. As the center of the distribution, μ can be viewed as the most representative one. Besides, since features are randomly sampled from the distribution, μ is much more robust compared to models without using VAE, where VAE loss can be regarded as a regularizer for learning more robust features.

3.3 VDA Architecture

To learn the representation of the image set that encodes the structural information, a good architecture of the encoder is critical. To this end, we propose 3 different encoders, namely LSTM Encoder, Attention Encoder, and Dynamic Filter Encoder. We will explain them in detail.

LSTM Encoder: We apply LSTM [12] to aggregate the features of different frames to get the representation of the video clips. Specifically, the LSTM takes the ordered frame features as input and we can take the average of the output of LSTM as the representation of the distribution of the image set, which can be expressed as:

$$r = \text{mean}_T(LSTM(g_1, \cdots, g_T)) \tag{10}$$

where g denotes the input feature map of the image set to LSTM, T denotes the length of input sequence and z denotes the output embedding of size R^C for the image sequence.

Attention Encoder: Another design is to first perform 1×1 convolution on the input feature map g and use Softmax function to compute the normalized scores for the features of each frame. Then we use the normalized scores as weights to do the weighted average over the input features of each frame as the representation of the video clips, which can be expressed as:

$$r = \text{mean}_T(\text{Softmax}(\text{conv}_{1\times1}(g)) \otimes g) \tag{11}$$

where $\text{conv}_{1\times1}$ only has 1 output channel since we only use it to calculate a scalar weight for the weighted average. The $\otimes$ denotes multiply the feature of each frame by its corresponding calculated normalized weight.

Dynamic Filter Encoder: For a given set, we first perform mean pooling to calculate the mean μ of the set and concatenate μ with each feature in the set. Then we perform 1×1 convolution on the concatenated feature map to get a weight matrix of the same size as the input feature map g. We also perform 1×1 convolution on the input feature map, with whose result we element-wisely multiply by the learned weight matrix. Finally, we perform the max-pooling to get the representation of the video clips. The whole procedure can be expressed as:

$$r = \max_T(\text{conv}_{1\times1}(\text{concat}(\text{mean}(g), g)) \otimes \text{conv}_{1\times1}(g)) \tag{12}$$

where both 1×1 convolution have the same output channels as input g. And $\otimes$ indicate element-wise multiplication.

Algorithm 1 : Training Procedure for VDA

```
Input: Training set: $\mathcal{X} = \{\boldsymbol{X}_i\}_{i=1}^{n}$, ID Label:$\boldsymbol{D} = \{\boldsymbol{D}_i\}_{i=1}^{n}$, K, P
Output: Decoder model $D$ and encoder model $E$
Initialize model parameters;
for $t = 1, 2, \ldots, M$ do
    Randomly sample K videos separately for P identities from $\mathcal{X}$;
    Feed mini-batch into backbone CNN model to get feature sets;
    Generate the structure parameters $\mu$ and $\Sigma$ of each feature setusing Eq. 3;
    Reconstruct the input feature sets using (4);
    Calculate the reconstruction loss using (5);
    Calculate the discriminative loss using (5);
    if ($t \geq 1$ and $|L_{VDA}(t) - L_{VDA}(t-1)| < \varepsilon$) go to return;
    Update all parameters by minimizing (9);
end for
return model $f(x; \theta)$;
```

Among the above 3 encoders, the experimental results show that dynamic filter encoder works the best compared to the other 2 architectures.

Nevertheless, given the need for VAE model, what we should learn as the intermediate embedding should be the estimated mean and variance of the posterior $p(z|g)$, so we add another two following independent branches to each of the encoder architecture described above to learn the mean μ and variance Σ that encodes the structural information for the reconstruction respectively. Each branch is composed of a multi-layer perceptron (MLP) with two layers.

$$\mu = \mathrm{MLP}_1(r) \tag{13}$$

and

$$\Sigma = \mathrm{MLP}_2(r) \tag{14}$$

Then we adopt reparametrization trick [17] to sample from the multivariate Gaussian distribution $\mathcal{N}(\mu, \Sigma)$ and feed the sample points $z_1, \cdots, z_T$ to the decoder D, which is a multi-layer fully connected neural network, to reconstruct the input feature set as $g'_1, \cdots, g'_T$.

Algorithm 1 summarizes the detailed procedure of our proposed VDA method.

4 Experiment

We evaluate our method on three widely used datasets including MARS [60], Duke-MTMC-VideoReID [50], and iLIDS-VID [48]. The following describes the details of our experiments and results.

4.1 Datasets

MARS: MARS is the largest video-based person re-identification dataset with 1261 subjects and around 20000 video sequences. Each identity in the sequences in Mars is captured by 6 cameras at most and 2 cameras at least and has 13.2 sequences on average. The bounding-box annotation of Mars is generated by DPM detector [8] and GMMCP tracker [6] instead of hand annotating.

DukeMTMC-VideoReID: DukeMTMC-VideoReID is a subset of the DukeMTMC tracking dataset [34] for video-based person re-identification. The dataset consists of 702 identities for training, 702 identities for testing, and 408 identities as distractors. In total there are 2,196 videos for training and 2,636 videos for testing. Each video clip contains person images sampled every 12 frames.

iLIDS-VID: The iLIDS-VID dataset includes 600 videos of 300 randomly selected persons who were observed in two non-overlapping cameras from the i-LIDS Multiple-Camera Tracking Scenario (MCTS), where each person has a pair of videos captured from two camera views in an airport arrival hall. The videos have variable lengths ranging from 23 to 193 image frames, with an average of 73. This dataset is challenging for person re-identification as there are large variations of lighting, viewpoint, cluttered background and occlusions in this dataset. Moreover, clothing similarity among different persons is also high.

Experiment Setup: For the Mars dataset, we follow the evaluation protocol proposed in [60] that uses 625 subjects for training and others for testing. For DukeMTMC-VideoReID, a video for each ID is used as the query and the remaining videos are placed in the gallery during testing as in [50]. For iLIDS-VID, we repeat experiments 10 times and calculated the average accuracy by splitting the dataset into equal-sized training and testing sets. To avoid the noise from dataset splitting and make sure a fair evaluation, we selected the identical 10 splits in [48], instead of random splits. The cumulative matching characteristic (CMC) curve is adopted as the evaluation metric. For the Mars dataset, mean Average Precision (mAP) is also utilized. CMC curves record the true matching within the top n ranks, while mAP considers precision and recall to evaluate the overall performance of the method. Given the Mars has multiple video clips for each identity, it is quite reasonable to use mAP to evaluate the performance on this dataset.

Implementation Details: We select Pytorch as the basic toolbox to implement our experiments. The backbone network architecture of the feature representation model is ResNet50. All CNN models in this work are pre-trained on ImageNet and fine-tuned with video person re-identification datasets. For iLIDS-VID dataset, we fuse the RGB stream and optical flow stream after a residual block. While for Mars and DukeMTMC-VideoReID, the fusing stage starts at the first pooling layer. In the training stage, e.g., the iLIDS-VID dataset we train our model on a single GTX 1080 Ti GPU machine for 500 epochs by Adam optimizer. The initial learning rate is 0.0002 and reduces by half with every 60 epochs. We randomly select 16 video clips from 4 persons in a batch, where each clip consists of 6 frames (for MARS, we train the model on 2 GPUs for 800 epochs with 48 clips from 12 persons in a batch). For the inputs of the original RGB-based video data and the optical-flow data, we apply randomly mirror and erase as data augmentation and resize them to 256 × 128.

Table 1. Comparison with the state-of-the-art person re-identification methods on the iLIDS-VID and MARS datasets.

Method	**iLIDS-VID**				**MARS**		
	R=1	**R=5**	**R=10**	**R=20**	**R=1**	**R=5**	**mAP**
CNN+RNN [29]	58.0	84.0	91.0	96.0	56.0	69.0	-
CNN+XQDA [60]	54.1	80.7	88.0	95.4	65.3	82.0	47.6
AMOC+ EpicFlow [27]	68.7	94.3	98.3	99.3	68.3	81.4	52.9
TRL [5]	57.7	81.7	-	94.1	80.5	91.8	69.1
TAM+SRM [63]	55.2	86.5	-	97.0	70.6	90.0	50.7
QAN [28]	68.0	86.6	95.4	97.4	73.7	84.9	51.7
ASTPN [52]	62.0	86.0	94.0	98.0	44.0	70.0	-
DSAN [49]	61.2	80.7	90.3	97.3	69.7	83.4	-
DRSTA [22]	80.2	-	-	-	82.3	-	65.9
CSSA+CASE [2]	85.4	96.7	98.8	99.5	86.3	94.7	76.1
RQEN [39]	76.1	92.9	97.5	99.3	73.7	84.9	51.7
STAL [4]	82.8	95.3	97.7	98.8	82.2	92.8	73.5
SDM [55]	60.2	84.7	91.7	95.2	71.2	85.7	-
SPL [31]	70.5	91.4	96.8	99.1	74.8	86.7	-
COSAM [41]	79.6	95.3	-	-	84.9	95.5	79.9
GLTR [21]	63.1	77.2	83.8	88.4	87.0	95.8	78.5
VRSTC [13]	83.4	95.5	97.7	99.5	**88.5**	**96.5**	82.3
ADFDTA [59]	86.3	97.4	-	99.7	87.0	95.4	78.2
VDA	**91.3**	**98.5**	**99.5**	**99.7**	87.6	95.6	**82.6**

Table 2. Comparison with the state-of-the-art person re-identification methods on the DukeMTMC-VideoReID datasets.

Method	**R=1**	**R=5**	**R=10**	**mAP**
ETAP-Net [50]	83.6	94.6	97.6	78.3
VRST [13]	95.0	99.1	99.4	93.5
COSAM [41]	93.7	99.0	-	93.5
GLTR [21]	96.3	99.3	-	93.7
STA [9]	96.2	99.3	-	94.9
VDA	**97.0**	**99.2**	**99.7**	**96.0**

4.2 Comparison with state-of-the-art methods

We compared our method with current video-based person re-identification methods The porposed VDA improve the Rank-1 by 5.0% on the iLIDS-VID dataset (91.3% vs 86.3%) as shown in Table 1. As mentioned above, iLIDS-VID is very difficult dataset due to large variations of lighting, viewpoint, cluttered background, occlusions and high clothing similarity, so current methods usually fail to recognize the difficult sample. Our method can obtain more discriminative features than thosed methods through variational discriminative aggregation. Our VDA achieved comparable results on the MARS dataset compared with VRSTC [13] (87.6% vs 88.5% on the Rank-1 matching rate, 82.6% vs 82.3% on the mAP) on the MARS dataset, and the detailed comparison was povided in

Table 3. Rank CMC accuracy of cross-dataset evaluation.

Datasets	**iLIDS-VID/PRID-2011**			
Rank@R	**R=1**	**R=5**	**R=10**	**R=20**
QAN [28]	34.0	61.3	74.0	83.1
STAL [4]	63.7	84.0	**92.8**	98.1
VDA	**66.8**	**84.9**	**92.8**	**98.5**

the Table 1. For the DukeMTMC-VideoReID dataset, as shown in the Table 2, the poposed VDA outperform state-of-the-art methods about 1% on both the Rank-1 matching rate (97.0% vs 96.3%) and mAP (96.0% vs 94.9%).

4.3 Cross Domain Evaluation

For a fair comparison, we follow the setting of conduct cross-domain testing in QAN [28], where the model is trained by a randomly split training set of the iLIDS-VID dataset and tested on the testing set of the PRID-2011 dataset. As shown in Table 3, VDA outperforms other SOTA methods by a large margin.

4.4 Ablation Study

To evaluate the proposed methods more specifically, we conducted several ablation experiments with different loss types, feature embeddings, and aggregation manners.

Effects of different losses: We investigate the effects of different loss functions in Table 4. "Baseline" is the performance of backbone model trained using softmax loss and triplet loss, where we simply use average pooling to aggregate features of different frames. We provide the results of various versions of our model by adding different losses to the attention aggregation model. We see the baseline model cannot be consistently improved when the VAE loss is not used, which clearly shows the effectiveness of the proposed loss. We think it is because the model usually overfit to the training set when we train a temporal encoder without using VAE loss. The VAE loss is critical in training temporal aggregation models and can serve as a good regularize to avoid overfitting. We also observed that triplet loss is better than n-pair loss, which shows hard example mining is very useful in ReID problems.

Effects of network architecture: As shown in Table 5, we compare our methods with several baseline methods with different network configurations. First, we compare our method with two simple methods: Attention and LSTM, where the proposed variational framework is not applied. We see these two methods can only achieve similar results with the basic mean pooling method. On the MARS dataset, these two methods obtain even worse results than the mean pooling. We conjecture that these two methods may suffer from overfitting on the training set. We then investigate several variants of our method, where LTSM, bidirectional LTSM and attention module are used as the encoder network. Clearly,

Table 4. Ablation study on loss functions.

Method	iLIDS-VID		MARS		DukeMTM	
	R1	mAP	R1	mAP	R1	mAP
Baseline	89.7	93.9	86.9	81.5	96.4	95.9
Encoder+triplet	89.0	93.0	86.0	80.7	96.0	95.1
Encoder+triplet+cls.	89.9	94.0	86.6	81.0	96.1	95.0
Encoder+triplet+cls.+VAE	**91.3**	**94.6**	**87.6**	**82.6**	**97.0**	**96.0**
Encoder+N-pair+cls.+VAE	90.7	94.2	87.5	82.7	96.7	95.6

Table 5. Ablation study on different model designs.

Method	iLIDS-VID		MARS		DukeMTM	
	R1	mAP	R1	mAP	R1	mAP
Baseline	89.7	93.9	86.9	81.5	96.4	95.9
Attention	90.7	94.4	85.5	80.7	96.7	96.0
LSTM	90.4	94.1	86.1	80.4	96.7	95.3
VAE+LSTM	89.3	93.4	87.5	82.0	96.2	95.4
VAE+Bi-LSTM	88.9	92.9	87.2	81.9	96.4	95.4
VAE+Attention	90.8	94.5	87.5	82.0	96.2	95.4
VDA	91.3	94.6	87.6	82.6	97.0	96.0

Table 6. Ablation study on the motion-aware embeddings.

Method	iLIDS-VID		MARS		DukeMTM	
	R1	mAP	R1	mAP	R1	mAP
CNN	84.8	88.7	85.5	80.4	95.2	94.3
CNN+flow$_{high}$	88.6	93.1	86.9	81.5	96.4	95.9
CNN+flow$_{low}$	89.7	93.9	86.3	81.2	96.3	95.5
CNN+flow+VDA	91.3	94.6	87.7	82.6	97.0	96.0

our dynamic filter encoder can outperform all other baseline methods, which demonstrates the effectiveness of our design.

Effects of motion-aware embeddings: We analyzed the influence of different embedding models including the original CNN backbone network, the motion-aware embedding model with low-resolution and high-resolution fusing manner, and our final model. As shown in 6, the motion clues in the optical flow improve the embedding performance by a large margin, especially for the iLIDS-VID dataset with +4.9%/5.2% on Rank-1/mAP , which is a challenging dataset with human hand-labeled. Besides, we also compare the influences of different motion fusing strategy. Specifically, "$flow_{low}$" denotes the low-resolution fusing strategy which fuses the optical flow stream and RGB stream with the 1/4 resolution stage (after a residual block). While '$flow_{high}$" denotes the high-resolution fusion with the 1/2 resolution stage (after first pooling layer). We can observe that the "$flow_{low}$" is more appropriate to iLIDS-VID dataset and "$flow_{high}$" is appropriate to others. Our VDA model can further improve the performance of the strong motion-aware embeddings which indicates its effectiveness

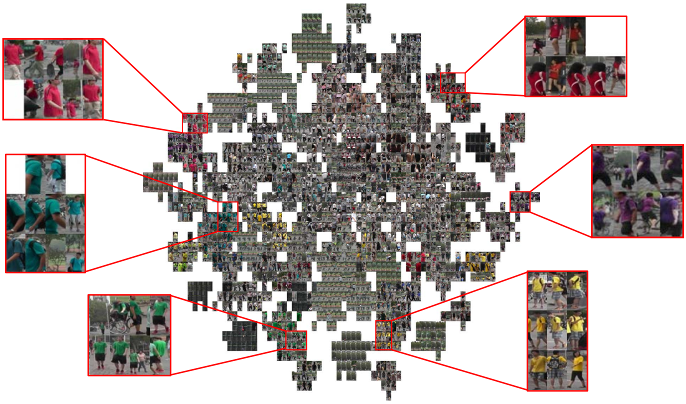

Fig. 3. The feature distribution map obtained by the proposed VDA method on the MARS dataset.

4.5 Visualization

We plotted the distributions of samples of the testing set on the MARS by using the t-SNE method to visualize the high dimensional features into the 2-D space, as shown in Figure 3. Almost of the image embeddings are trivially separable according to their identifications while the inner-class variances are large caused to the viewpoints, occlusions, and noising detections. For example, we selected four persons with blue, red, green and yellow T-shirts, and magnify them in the four corners of the image. The visual results also demonstrated that our methods can generate the representation of the image sequence with high structural information and the discriminative ability for the retrieval.

5 Conclusion

In this work, we propose a variational discriminative aggregation (VDA) approach for video-based person re-identification, which aggregates the information across frames instead of using attention or a pooling strategy for video representation. By doing so, we explicitly capture the structure information in the input feature space by encoding the structural knowledge of each video in the learned representation space, which can be used for following retrieval step. To further boost the discriminative ability of the learned representation for the retrieval, classification loss, as well as the triplet loss, are employed as supervision signals for the learning procedure of the VDA module. Moreover, considering the motion characteristics of the video-based person re-identification, we also propose a pipeline to extract the feature for each frame where we employ a two-stream network taking raw image sequence and its corresponding computed optical flow as input respectively. The experimental results and ablation studies demonstrate the importance of learning the structure of the features of the image sequence and the motion characteristics of the video clips by achieving very competitive performance compared to the state-of-the-art methods.

References

1. Chen, B., Deng, W., Hu, J.: Mixed high-order attention network for person re-identification. In: ICCV (October 2019)
2. 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)
3. Chen, G., Lin, C., Ren, L., Lu, J., Jie, Z.: Self-critical attention learning for person re-identification. In: ICCV (2019)
4. Chen, G., Lu, J., Yang, M., Zhou, J.: Spatial-temporal attention-aware learning for video-based person re-identification. TIP (2019)
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. Felzenszwalb, P.F., Girshick, R.B., McAllester, D., Ramanan, D.: Object detection with discriminatively trained part-based models. TPAMI **32**(9), 1627–1645 (2010)
9. Fu, Y., Wang, X., Wei, Y., Huang, T.: Sta: Spatial-temporal attention for large-scale video-based person re-identification. In: AAAI (2019)
10. Ge, Y., Li, Z., Zhao, H., Yin, G., Yi, S., Wang, X., et al.: Fd-gan: Pose-guided feature distilling gan for robust person re-identification. In: NIPS. pp. 1230–1241 (2018)
11. Han, Z., Wang, X., Liu, Y.S., Zwicker, M.: Multi-angle point cloud-vae: unsupervised feature learning for 3d point clouds from multiple angles by joint self-reconstruction and half-to-half prediction. arXiv preprint arXiv:1907.12704 (2019)
12. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural computation **9**(8), 1735–1780 (1997)
13. Hou, R., Ma, B., Chang, H., Gu, X., Shan, S., Chen, X.: Vrstc: Occlusion-free video person re-identification. In: CVPR (June 2019)
14. Kalayeh, M.M., Basaran, E., Gökmen, M., Kamasak, M.E., Shah, M.: Human semantic parsing for person re-identification. In: CVPR. pp. 1062–1071 (2018)
15. Karanam, S., Li, Y., Radke, R.J.: Person re-identification with discriminatively trained viewpoint invariant dictionaries. In: ICCV. pp. 4516–4524 (2015)
16. Karanam, S., Li, Y., Radke, R.J.: Sparse re-id: Block sparsity for person re-identification. In: CVPR Workshops. pp. 33–40 (2015)
17. Kingma, D.P., Welling, M.: Auto-encoding variational bayes. arXiv preprint arXiv:1312.6114 (2013)
18. Klaser, A., Marszałek, M., Schmid, C.: A spatio-temporal descriptor based on 3d-gradients. In: BMVC. pp. 1–10 (2008)
19. Kumar Verma, V., Arora, G., Mishra, A., Rai, P.: Generalized zero-shot learning via synthesized examples. In: CVPR. pp. 4281–4289 (2018)
20. LeCun, Y., Bengio, Y., et al.: Convolutional networks for images, speech, and time series. The handbook of brain theory and neural networks **3361**(10), 1995 (1995)
21. Li, J., Wang, J., Tian, Q., Gao, W., Zhang, S.: Global-local temporal representations for video person re-identification. In: ICCV (October 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, S., Hu, Y., Zhu, X., Li, S.Z.: Person re-identification by local maximal occurrence representation and metric learning. In: CVPR. pp. 2197–2206 (2015)
25. 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)
26. Lin, X., Duan, Y., Dong, Q., Lu, J., Zhou, J.: Deep variational metric learning. In: ECCV. pp. 689–704 (2018)
27. 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)
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. McLaughlin, N., del Rincon, J.M., Miller, P.: Video person re-identification for wide area tracking based on recurrent neural networks. TCSVT (2017)
31. 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)
32. Qian, X., Fu, Y., Xiang, T., Wang, W., Qiu, J., Wu, Y., Jiang, Y.G., Xue, X.: Pose-normalized image generation for person re-identification. In: ECCV. pp. 650–667 (2018)
33. 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)
34. Ristani, E., Solera, F., Zou, R., Cucchiara, R., Tomasi, C.: Performance measures and a data set for multi-target, multi-camera tracking. In: ECCV. pp. 17–35 (2016)
35. Schroff, F., Kalenichenko, D., Philbin, J.: Facenet: A unified embedding for face recognition and clustering. In: CVPR. pp. 815–823 (2015)
36. Si, J., Zhang, H., Li, C.G., Kuen, J., Kong, X., Kot, A.C., Wang, G.: Dual attention matching network for context-aware feature sequence based person re-identification. In: CVPR. pp. 5363–5372 (2018)
37. Sohn, K.: Improved deep metric learning with multi-class n-pair loss objective. In: Advances in Neural Information Processing Systems. pp. 1857–1865 (2016)
38. Song, C., Huang, Y., Ouyang, W., Wang, L.: Mask-guided contrastive attention model for person re-identification. In: CVPR. pp. 1179–1188 (2018)
39. Song, G., Leng, B., Liu, Y., Hetang, C., Cai, S.: Region-based quality estimation network for large-scale person re-identification. In: AAAI (2018)
40. Su, C., Li, J., Zhang, S., Xing, J., Gao, W., Tian, Q.: Pose-driven deep convolutional model for person re-identification. In: ICCV (2017)
41. Subramaniam, A., Nambiar, A., Mittal, A.: Co-segmentation inspired attention networks for video-based person re-identification. In: ICCV (October 2019)
42. Sun, Y., Xu, Q., Li, Y., Zhang, C., Li, Y., Wang, S., Sun, J.: Perceive where to focus: Learning visibility-aware part-level features for partial person re-identification. In: CVPR (June 2019)
43. Sun, Y., Zheng, L., Yang, Y., Tian, Q., Wang, S.: Beyond part models: Person retrieval with refined part pooling. In: ECCV (2018)

44. Tay, C.P., Roy, S., Yap, K.H.: Aanet: Attribute attention network for person re-identifications. In: CVPR. pp. 7134–7143 (2019)
45. Wang, G., Lai, J., Huang, P., Xie, X.: Spatial-temporal person re-identification. In: AAAI. vol. 33, pp. 8933–8940 (2019)
46. Wang, G., Yuan, Y., Chen, X., Li, J., Zhou, X.: Learning discriminative features with multiple granularities for person re-identification. In: ACMMM. pp. 274–282 (2018)
47. Wang, T., Gong, S., Zhu, X., Wang, S.: Person re-identification by discriminative selection in video ranking. TPAMI **38**(12), 2501–2514 (2016)
48. Wang, X., Zhao, R.: Person re-identification: System design and evaluation overview. In: Person Re-Identification, pp. 351–370. Springer (2014)
49. 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)
50. 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)
51. Xiao, T., Li, H., Ouyang, W., Wang, X.: Learning deep feature representations with domain guided dropout for person re-identification. In: CVPR. pp. 1249–1258 (2016)
52. 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)
53. 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)
54. You, J., Wu, A., Li, X., Zheng, W.S.: Top-push video-based person re-identification. In: CVPR. pp. 1345–1353 (June 2016)
55. Zhang, J., Wang, N., Zhang, L.: Multi-shot pedestrian re-identification via sequential decision making. In: CVPR (2018)
56. Zhang, Z., Lan, C., Zeng, W., Chen, Z.: Densely semantically aligned person re-identification. In: CVPR. pp. 667–676 (2019)
57. 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)
58. 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. pp. 4913–4922 (2019)
59. 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)
60. 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. Springer (2016)
61. Zheng, L., Zhang, H., Sun, S., Chandraker, M., Yang, Y., Tian, Q., et al.: Person re-identification in the wild. In: CVPR. vol. 1, p. 2 (2017)
62. Zhong, Z., Zheng, L., Zheng, Z., Li, S., Yang, Y.: Camera style adaptation for person re-identification. In: CVPR. pp. 5157–5166 (2018)
63. 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)