JUSTC

Article

http://justc.ustc.edu.cn

Received: September 26, 2021; Accepted: April 10, 2022

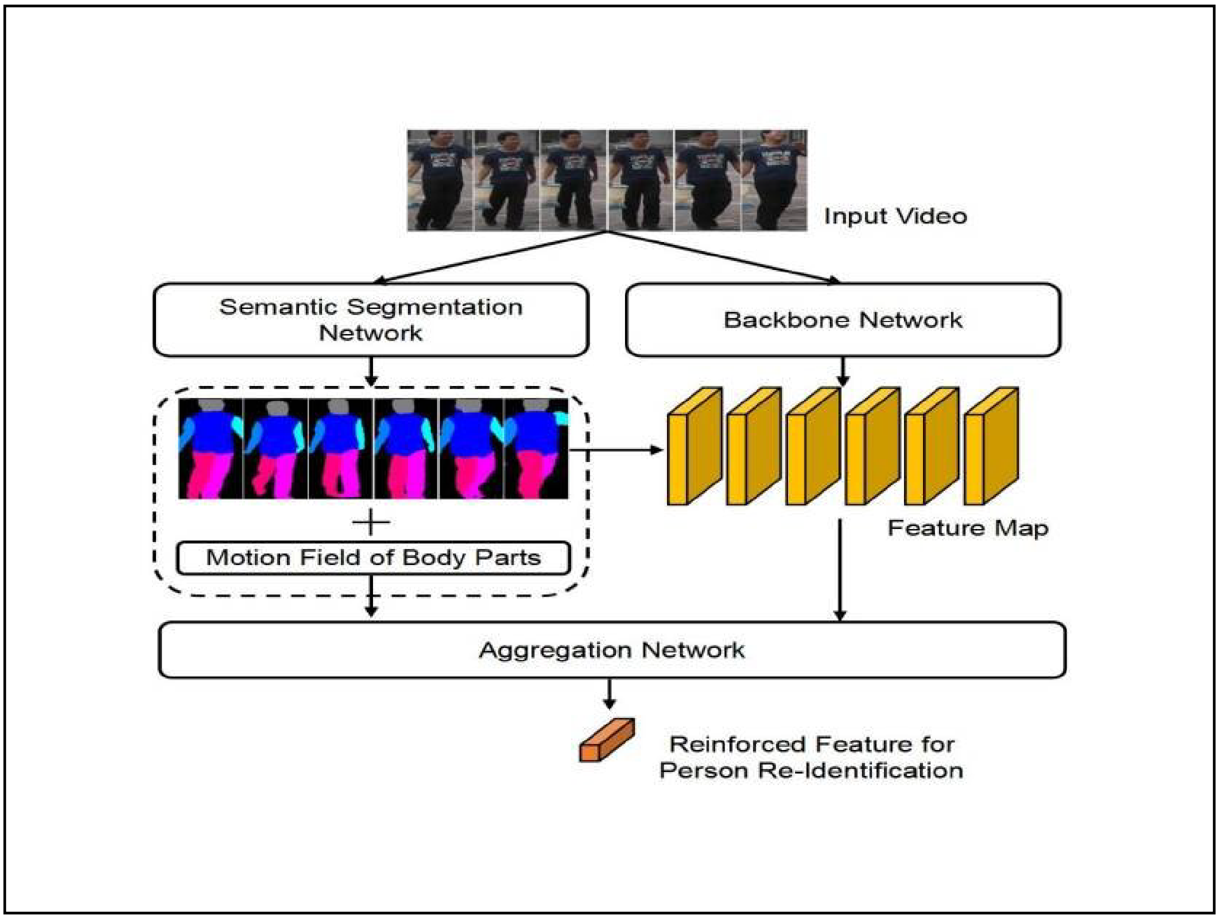
Self-supervised human semantic parsing for video-based person re-identification

Wei Wu, and Jiawei Liu ✉

School of Information Science and Technology, University of Science and Technology of China, Hefei 230027, China

✉Correspondence: Jiawei Liu, E-mail: jwliu6@ustc.edu.cn

Graphical abstract


Input Video
Semantic Segmentation Network
Backbone Network
Motion Field of Body Parts
Feature Map
Aggregation Network
Reinforced Feature for Person Re-Identification

The basic structure of self-supervised human semantic parsing approach (SS-HSP).

Public summary

- A self-supervised human semantic parsing approach is proposed for video-based person re-identification.
- We employ self-supervised learning to adaptively segment the human body by estimating the motion information of each body part between consecutive frames.
- We explore complementary temporal relations for pursuing reinforced appearance and motion representations.

Citation: Wu W, Liu J W. Self-supervised human semantic parsing for video-based person re-identification. *JUSTC*, **2022**, 52(9): 5. DOI: 10.52396/JUSTC-2021-0212

JUSTC

Article

http://justc.ustc.edu.cn

Received: September 26, 2021; Accepted: April 10, 2022

Self-supervised human semantic parsing for video-based person re-identification

Wei Wu, and Jiawei Liu ✉

School of Information Science and Technology, University of Science and Technology of China, Hefei 230027, China

✉Correspondence: Jiawei Liu, E-mail: jwliu6@ustc.edu.cn

Cite This: *JUSTC*, **2022**, 52(9): 5 (11pp)

Read Online

Abstract: Video-based person re-identification is an important research topic in computer vision that entails associating a pedestrian’s identity with non-overlapping cameras. It suffers from severe temporal appearance misalignment and visual ambiguity problems. We propose a novel self-supervised human semantic parsing approach (SS-HSP) for video-based person re-identification in this work. It employs self-supervised learning to adaptively segment the human body at pixel-level by estimating motion information of each body part between consecutive frames and explores complementary temporal relations for pursuing reinforced appearance and motion representations. Specifically, a semantic segmentation network within SS-HSP is designed, which exploits self-supervised learning by constructing a pretext task of predicting future frames. The network learns precise human semantic parsing together with the motion field of each body part between consecutive frames, which permits the reconstruction of future frames with the aid of several customized loss functions. Local aligned features of body parts are obtained according to the estimated human parsing. Moreover, an aggregation network is proposed to explore the correlation information across video frames for refining the appearance and motion representations. Extensive experiments on two video datasets have demonstrated the effectiveness of the proposed approach.

Keywords: person re-identification; self-supervised learning; semantic parsing

CLC number: TP181 **Document code:** A

1 Introduction

Person re-identification (Re-ID) is the task of associating individuals across non-overlapping camera views. It has drawn increasing attention in recent years, as it plays a significant role in various practical applications, such as intelligent surveillance, activity analysis, smart retail, etc.[1–4] The surge of deep learning techniques has been reflected in the task of person Re-ID, achieving exciting progresses on many benchmark datasets. Nevertheless, it remains challenging in real scenario, due to cluttered background, partial occlusion, heavy illumination changes, viewpoint variations, etc.[5–7]

Person Re-ID is often approached with either image or video data for representation[8]. Most existing approaches recognize pedestrians in static image setting, mainly focusing on learning image-level discriminative representations. In parallel with the impressive progress of image-based person Re-ID, video-based person Re-ID has recently attracted significant attention. Compared to an image with limited appearance information, a video sequence captures abundant visual details in a long time, presenting appearance under diverse posture and viewpoint variations. Hence, a video provides crucial knowledge to alleviate visual ambiguity. Besides, it also contains rich motion patterns of pedestrians, e.g., walking style and moving direction[9, 10], contributing to identifying pedestrians apart from appearance. The key to video-based person re-identification is effectively excavating appearance and motion information from video sequences. Fig. 1 illustrates some sample video sequences on the two datasets, i.e., MARS[11] and iLIDS-VID[12].

To leverage appearance information, some preliminary methods learned frame-level appearance features by considering the whole frames and then aggregating them through pooling operation or recurrent neural network[13]. The ubiquitous presence of temporal appearance misalignment problem caused by partial occlusions, inaccurate detection or human pose variations, etc., leads to severe performance degradation for these preliminary methods. Recent works attempted to address the misalignment issue: including fixed partition based methods which directly partition video frames into rigid horizontal stripes[14, 15], and attention based methods which discover distinctive body parts by using diverse spatial attentions and crucial frames by using temporal attentions[16, 17]. However, they are rough with much background noise in their located partial regions, thus can not accurately extract features from body parts. Instead of utilizing these self-learned styles, some other methods exploit augmented information, including object segmentation[18] or pose estimation[19, 20], to achieve part alignment at pixel level. Nevertheless, they depend on the accuracy of the pre-trained semantic segmentation or pose estimation models by additional datasets with annotations and are thus susceptible to dataset discrepancy.

Compared to appearance representation, motion patterns own strong robustness to the variations in illumination and viewpoint, which provide complementary cues for alleviating visual ambiguity and realizing precise matching. To leverage motion information from video sequences, most existing ap-

DOI: 10.52396/JUSTC-2021-0212

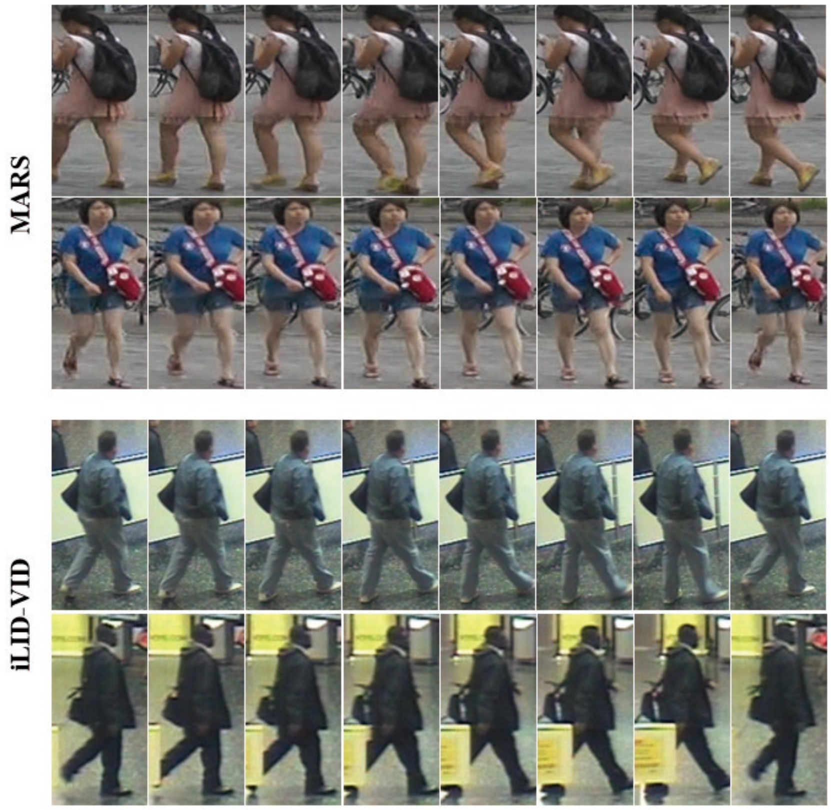


Fig. 1. Example video sequences in the MARS and iLIDS-VID person re-identification datasets.

proaches either employ 3D convolution or resort to hand-crafted optical flow in an offline way[21, 22]. The 3D convolution operation is limited by high computational complexity and the unsatisfied ability for video analysis[23]. However, pre-computed optical flow is independent of video-based person re-identification, which may not be optimal for this task.

Apart from the methods mentioned above, some gait-based [24] and true motion-based[25] person Re-ID methods are also proposed to leverage motion information from video sequences. The method in Ref. [24] introduces gait recognition as an auxiliary task to drive person Re-ID models to learn more effective representations by leveraging personal unique and cloth-independent gait information. However, the learned gait feature is heavily dependent on the pre-extracted input silhouettes and pre-trained GaitSet (a set-based gait recognition model), which is thus susceptible to dataset discrepancy, deteriorating its ability. Moreover, the method in Ref. [25] formulates a FIne moTion encoDing (FITD) model based on dynamic cues, which characters motion patterns by the trajectory-aligned descriptors in a three-level body-action pyramid. However, it can’t obtain robust motion features by only using a fixed partition strategy to capture trajectory-aligned descriptors, due to the widespread misalignment issue.

In this work, we propose a novel self-supervised human semantic parsing approach (SS-HSP) for video-based person Re-ID. It is the first work for video-based person Re-ID exploring self-supervised learning to precisely locate human body parts at pixel-level by estimating the motion of each body part between consecutive frames without any manual annotation. By constructing a pretext task of predicting future frames, SS-HSP learns segmentation maps of body parts and the corresponding motion field and explores temporal relations across video frames to generate reinforced appearance and motion representations. As illustrated in Fig. 2, SS-HSP consists of a backbone network for extracting low-level visual representation, a semantic segmentation network for predicting future frames, and an aggregation network for Re-ID. The semantic segmentation network composes a segmentation module and a prediction module. The former is in charge of extracting the segmentation maps of body parts and the optical flows corresponding to these body parts. The latter is in charge of predicting the next frame by employing the current frame and the output of the segmentation module. Several customized loss functions optimize the semantic segmentation network. After obtaining the segmentation maps and optical flows, the aggregation network extracts frame-level appearance and motion features. It explores the complementary relation information across video frames via a temporal relation block to generate discriminative video-level representations. We conduct extensive experiments to evaluate SS-HSP on two challenging datasets and report superior performance over state-of-the-art methods.

Although human semantic parsing has been explored in person Re-ID[5, 26], these works are deigned for image-based person Re-ID without considering video human semantic parsing and temporal motion information. Thus they can not be directly applied to video-based person Re-ID. Moreover, they heavily rely on the performance of the pre-trained human parsing models by auxiliary datasets. They do not efficiently handle large-scale video datasets due to iteratively cluster the pixels of all training samples’ feature maps simultaneously.

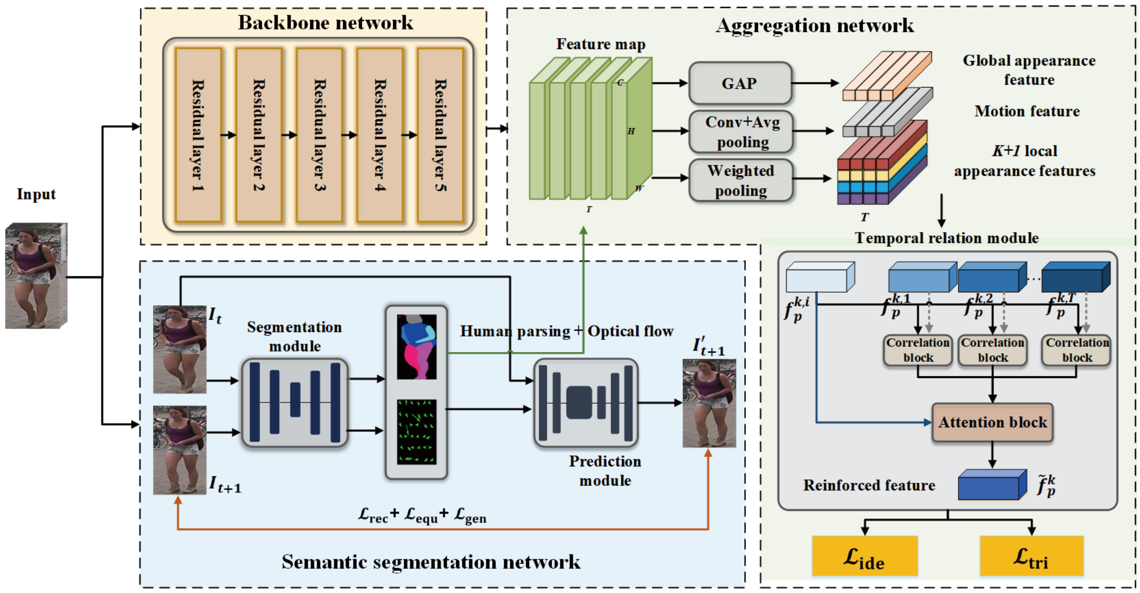


Fig. 2. The overall architecture of the proposed SS-HSP. It consists of a backbone network, a semantic segmentation network as well as an aggregation network.

The main contribution of this work is three-fold: (i) We propose a novel self-supervised human semantic parsing approach (SS-HSP) for video-based person Re-ID. (ii) We design a semantic segmentation network for precisely locating human body parts and estimating each body part's motion information between consecutive frames. (iii) We develop an aggregation network to explore the complementary relation information among video frames for learning reinforced appearance and motion representations.

2 Related work

Existing person re-identification approaches can be summarized into image-based person Re-ID and video-based person Re-ID. We briefly review the two categories of related works in this section.

2.1 Image-based person Re-ID

Conventional approaches for image-based person Re-ID mainly focus on designing hand-crafted descriptors[27, 28] or learning appropriate distance metric[27, 29]. Recently, deep learning based methods have been widely proposed for learning distinctive features[17, 30]. For example, Zhou et al.[30] proposed a local-refining based deep neural network for person Re-ID, which contained a main branch network and a pose branch network to fuse pose and attribute information in a consistent way. Zhang et al.[31] proposed a Relation-Aware Global Attention (RGA) module, which captured the global structural information for better attention learning. Jin et al.[32] designed a semantics aligning network (SAN) for learning semantics-aligned feature representations from images under the joint supervision of re-identification and semantics-aligned texture generation.

2.2 Video-based person Re-ID

Early approaches for video-based person Re-ID concentrated on hand-crafted video-level descriptors or distance metric learning[12, 33]. Recent works mostly utilized deep learning techniques to extract discriminative representations from videos. Some methods[13] were proposed to formulate video-based Re-ID as an extension of image-based Re-ID simply. They extracted appearance features from each frame by various deep learning models, and aggregated frame-level features across time by pooling operation or RNN. For example, McLaughlin et al.[13] proposed a Siamese network, which captured pedestrian features and then employed a recurrent layer and a temporal pooling layer to abstract video-level features. For learning effective appearance features against the misalignment issue, rigid stripe partition[14] and attention mechanism[1] have been widely applied to plenty of person Re-ID methods. For example, Li et al.[16] proposed a spatio-temporal attention model automatically discovering a diverse set of distinctive body parts and extracting useful information from all frames against occlusions and misalignments. Moreover, a few works[18, 19] utilized the augmented information to enhance feature representation. For example, Jones et al.[19] proposed a pose-guided alignment network which mimicked the top-down attention of the human visual cortex. On the other hand, for learning motion representation, some existing methods introduce 3D convolution[34] or pre-computed optical flow[35]. For example, Liu et al.[21] proposed a Dense 3D-Convolutional Network (D3DNet) to jointly learn spatio-temporal and appearance representation from videos by 3D convolution.

3 Method

In this section, we first present the overall architecture of SS-HSP and then introduce each component of SS-HSP in the following subsections.

3.1 Architecture overview

For video person Re-ID, we aim at learning effective and discriminative appearance and motion representations from videos. The overall architecture of SS-HSP is illustrated in-

Fig. 2. It consists of a backbone network, a semantic segmentation network and an aggregation network. Supposing a video sequence is denoted as $\{\boldsymbol{I}_t\}_{t=1}^{T}$, where T is the sequence length. The backbone network takes each frame as an input to extract the initial feature maps $\{\boldsymbol{X}_t|\boldsymbol{X}_t \in \mathbb{R}^{C\times H\times W}\}_{t=1}^{T}$, where C, H, and W denote the channel, height, and width of the feature maps, respectively. The backbone network is built on the ResNet-50 model[36], which contains five residual layers, each of them is composed of several convolution layers, batch normalization (BN) layers, rectified linear units (ReLU) layers, and max-pooling layers. The semantic segmentation network consists of a segmentation module and a prediction module. The network learns the segmentation maps of human body parts and the corresponding optical flows among consecutive frames by constructing a pretext task of predicting future frames. The estimated segmentation maps and motion information together with the feature map $\boldsymbol{X}_t$ are fed into the aggregation network to generate reinforced clip-level appearance and motion features. These features are finally taken into a classifier and optimized by two Re-ID loss functions.

3.2 Semantic segmentation network

The misalignment and visual ambiguity issues are ubiquitous in person Re-ID, which deteriorate the ability of the extracted representation and compromise the performance. Considering that the annotation of human body parts is unavailable, we introduce a self-supervised learning strategy inspired by the work[37] to design a semantic segmentation network for adaptively locating body parts of pedestrians and extracting the corresponding motion information across consecutive frames. As shown in Fig. 3, the network consists of a segmentation module and a prediction module. The semantic segmentation network employs the current frame with the motion information of body parts between consecutive frames to predict the next frame.

The segmentation module takes a pair of frames $\boldsymbol{I}_t$ and $\boldsymbol{I}_{t+1}$ sampled from a video as input and generates the segmentation maps and motion field of body parts between the two consecutive frames. Concretely, the module is based on U-Net architecture[38], which consists of four 3×3 convolution layers followed by BN layer, ReLU layer, and average pooling layer, and four 3×3 up-sampling convolution layers followed by BN layer and ReLU layer. The resolution of the input frames is $H'\times W'$, and the outputs of the modules are two $(6K+1)$-channel tensors $\boldsymbol{S}_t, \boldsymbol{S}_{t+1} \in \mathbb{R}^{(6K+1)\times H'\times W'}$. The tensor $\boldsymbol{S}_t$ composes one $(K+1)$-channel tensor, one K-channel tensor, and one $4K$-channel tensor. The K-channel and $4K$-channel tensors are used to calculate the motion field between the two consecutive frames. The remaining $(K+1)$-channel tensors are applied with a channel-wise softmax operation to generate the segmentation maps $\boldsymbol{M}_t$, $\boldsymbol{M}_{t+1} \in [0,1]^{(K+1)\times H'\times W'}$ of body parts ($K+1$ denotes K body parts of a pedestrian with additional background).

Moreover, we employ optical flows to represent the motion field between the two consecutive frames, which maps each position of pixels in $\boldsymbol{I}_{t+1}$ to its corresponding position in $\boldsymbol{I}_t$. Note that the module does not use external optical flow estimators to calculate the motion between the two frames. Instead, it models the temporal motion of the pixels within each body part by an affine transformation. Therefore, the backward optical flow $\mathcal{G}$ between the two consecutive frames can be approximately by combining the affine transformation of each body part. Let $\boldsymbol{Z}_{t+1}^k = \{z\,|\,\boldsymbol{M}_{t+1}^k[z]=1\}$ denotes the locations in the segmentation map associated to kth body part for frame $\boldsymbol{I}_{t+1}$. Following the previous work[37], the optical flow for kth body part is computed as follows:

$$\mathcal{G}^k(z) = \boldsymbol{e}_t^k + \boldsymbol{A}_t^k \boldsymbol{A}_{t+1}^{k^{-1}}(z - \boldsymbol{e}_{t+1}^k), \tag{1}$$

where $\boldsymbol{e}_t^k$, $\boldsymbol{e}_{t+1}^k \in \mathbb{R}^2$ denote the shift parameters of kth body part for the two frames, which are estimated by performing the soft-argmax operation on the K-channel tensors in $\boldsymbol{S}_t$, $\boldsymbol{S}_{t+1}$. $\boldsymbol{A}_t^k$, $\boldsymbol{A}_{t+1}^k \in \mathbb{R}^{2\times 2}$ denote the affine parameters, which are estimated by performing the spatial weighted average operation on the $4K$-channel tensors in $\boldsymbol{S}_t$, $\boldsymbol{S}_{t+1}$. After that, the partial optical flow fields $\{\boldsymbol{G}^k \in \mathbb{R}^{2\times H'\times W'}\}_{k=1}^{K}$ for K body parts is obtained by repeating $\mathcal{G}^k(z)$ for $H'\times W'$ times. Supposing that the background is static, an addition optical flow field for the background is $\boldsymbol{G}^{K+1} = z$. Consequently, the final overall optical flow field $\hat{\boldsymbol{G}}$ is modeled as follows:

$$\hat{\boldsymbol{G}} = \sum_{k=1}^{K+1} \boldsymbol{M}_{t+1}^k \odot \boldsymbol{G}^k, \tag{2}$$

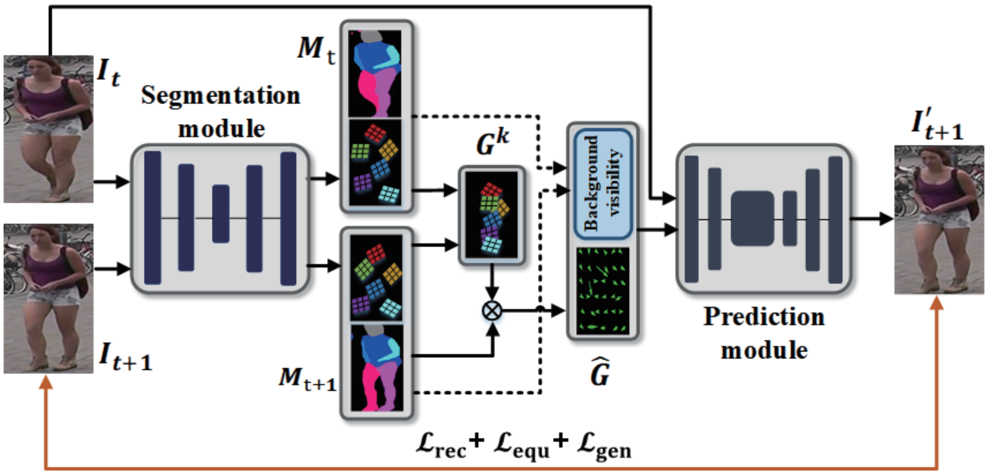


Fig. 3. Detailed structure of the semantic segmentation network.

where the segmentation map $\boldsymbol{M}^k_{t+1}$ allocates the partial optical flow field $\boldsymbol{G}^k$ of human body parts to each location.

The prediction module takes the current frame $\boldsymbol{I}_t$ and the estimated optical flow field as input. It then warps the feature of $\boldsymbol{I}_t$ according to the estimated optical flow field between the two frames for predicting the next frame $\boldsymbol{I}'_{t+1}$. The module is based on the encoder-decoder architecture[39], which contains two down-sampling layers, a deformation layer[40], five residual layers and two up-sampling layers. The deformation layer is employed to warp the feature of $\boldsymbol{I}_t$ with the optical flow field, which is defined as follows:

$$\begin{aligned} \boldsymbol{v}'_s &= \boldsymbol{O} \odot f_w(\boldsymbol{v}_s, \hat{\boldsymbol{G}}), \\ \boldsymbol{O} &= 1 - (\boldsymbol{M}^{K+1}_{t+1} \odot \sum_{k=1}^{K} \boldsymbol{M}^k_t), \end{aligned} \tag{3}$$

where $\boldsymbol{v}_s$ denotes the extracted feature map of $\boldsymbol{I}_t$ after two down-sampling layers, $f_w(\cdot,\cdot)$ denotes the back-warping function. $\boldsymbol{O}$ refers to the background visibility map, which indicates the pixels of the background in $\boldsymbol{I}_{t+1}$ are occluded by the foreground body parts in $\boldsymbol{I}_t$. The background visibility map suppresses the occluded regions' information and provides an important regularization to enforce superior foreground/background region segmentation. Finally, the transformed feature map $\boldsymbol{v}'_s$ is fed to subsequent layers of this module for rendering the next frame $\boldsymbol{I}'_{t+1}$. The whole semantic segmentation network is trained by several losses, including reconstruction loss, equivariance loss, and geometric concentration loss.

3.3 Aggregation network

The aggregation network is designed with a temporal relation block to employ temporal relation among video frames and generate reinforced clip-level appearance and motion features. The network receives the initial feature map $\boldsymbol{X}$ from the backbone network, the estimated segmentation map, and the optical flow field from the semantic segmentation network. The initial feature map $\boldsymbol{X}$ is fed to a global average pooling (GAP) layer and a fully connected (FC) layer to produce the global appearance features $\{\boldsymbol{f}_g\}_{t=1}^T$. Moreover, the feature map $\boldsymbol{X}$ and the K segmentation maps $\{\boldsymbol{M}^1, \boldsymbol{M}^2, ..., \boldsymbol{M}^K\}$ of body parts are applied with a weighed pooling layer and a FC layer to generate the $(K+1)$ local appearance features $\{\boldsymbol{f}_p^1, \boldsymbol{f}_p^2, ..., \boldsymbol{f}_p^{K+1}\}_{t=1}^T$. The last local appearance feature $\boldsymbol{f}_p^{K+1}$ is associated with the overall foreground region $\sum_{k=1}^K \boldsymbol{M}^k$ of video frames. The motion information of $\hat{\boldsymbol{G}}$ are fed into a 1×1 convolution layer, an average pooling layer, and an FC layer to obtain the motion features $\{\boldsymbol{f}_m\}_{t=1}^T$.

In order to effectively explore the complementary relation information across video frames and enhance the frame-level appearance and motion features, we develop a temporal relation block to refine the frame-level features by their relation with features of the other frames and aggregate them into robust clip-level features. Specifically, the frame-level features are firstly fed into a correlation block, which produces the informative and compact relation features. The formulation of this block is defined as follows:

$$\begin{aligned} \boldsymbol{r}_{t,m} &= h_1([\phi(\boldsymbol{f}_t), \varphi(\boldsymbol{f}_t - \boldsymbol{f}_m)]), \\ \tilde{\boldsymbol{r}}_t &= \mathrm{Concat}([\boldsymbol{r}_{t,1}, \boldsymbol{r}_{t,2}, ..., \boldsymbol{r}_{t,T}]), \\ \hat{\boldsymbol{f}}_t &= h_2([\boldsymbol{f}_t, \tilde{\boldsymbol{r}}_t]), \end{aligned} \tag{4}$$

where $\boldsymbol{f}_t$ denotes the local appearance feature, the global appearance feature, or the motion feature of tth video frame. ϕ, φ, h_1, h_2 are the embedding functions implemented by a full connected layer with a BN layer and a ReLU layer. The relation feature $\hat{\boldsymbol{f}}_t$ aggregates the global relation information of all other frame-level features. Afterwards, the generated relation features go through an attention block to infer the temporal attention score for each video frame and form the reinforced clip-level appearance and motion features by weighted sum operation. It is formulated as follows:

$$\begin{aligned} \boldsymbol{a}_t &= \mathrm{Sigmoid}(\mathrm{BN}(\boldsymbol{W}_a[\boldsymbol{f}_t, \tilde{\boldsymbol{r}}_t])), \\ \tilde{\boldsymbol{f}} &= \frac{\sum_{t=1}^{T} \boldsymbol{a}_t \cdot \hat{\boldsymbol{f}}_t}{\sum_{t=1}^{T} \boldsymbol{a}_t}, \end{aligned} \tag{5}$$

where $\boldsymbol{a}_t$ is a temporal attention value, $\boldsymbol{W}$ is parameter matrix and $\tilde{\boldsymbol{f}}$ denotes the reinforced appearance or motion clip-level features. After the temporal relation block, the generated reinforced global and local appearance features $\tilde{\boldsymbol{f}}_g$, $\{\tilde{\boldsymbol{f}}_p^k\}_{k=1}^{K+1}$ are supervised by identification loss. Meanwhile, the reinforced appearance and motion features $\tilde{\boldsymbol{f}}_g$, $\{\tilde{\boldsymbol{f}}_p^k\}_{k=1}^{K+1}$, and $\tilde{\boldsymbol{f}}_m$ are concatenated, and then supervised by triplet loss. In the testing stage, the final video representation of pedestrians is formed by concatenating these features.

3.4 Loss function and optimization

We adopt reconstruction loss, equivariance loss and geometric concentration loss to optimize the semantic segmentation network. The reconstruction loss is based on the perceptual loss[39], which assesses the reconstruction quality between the predicted next frame and the ground-true next frame. The formulation of this loss is defined as follows:

$$\mathcal{L}_{\mathrm{rec}} = \sum_{i=1}^{N} \|\phi_i(\boldsymbol{I}'_{t+1}) - \phi_i(\boldsymbol{I}_{t+1})\|_2^2, \tag{6}$$

where $\phi_i(\cdot)$ denotes the i-th channel feature extracted from a pre-trained and fixed VGG-19 model[37]. This loss calculates on three resolutions of 256×128, 128×64, and 64×32 for $\boldsymbol{I}_{t+1}$ and $\boldsymbol{I}'_{t+1}$. The equivariance loss encourages the learned segmentation maps and the optical flow field to be robust against the appearance variation and consistent with geometric transformation. It is formulated as follows:

$$\mathcal{L}_{\mathrm{equ}} = D_{\mathrm{KL}}(\hat{\boldsymbol{M}}_t \| T_s(\boldsymbol{M}_t)) + \|\hat{\boldsymbol{A}}_t - T_s(\boldsymbol{A}_t)\|_2^2 + \|\hat{\boldsymbol{e}}_t - T_s(\boldsymbol{e}_t)\|_2^2, \tag{7}$$

where D_{KL} denotes the Kullback-Leibler divergence distance, $\hat{\boldsymbol{M}}_t$, $\hat{\boldsymbol{A}}_t$ and $\hat{\boldsymbol{e}}_t$ are the estimated segmentation map, the affine and shift parameters from the transformed frame $\hat{\boldsymbol{I}}_t = T_s(T_c(\boldsymbol{I}_t))$. T_s refers to spatial transformation by thin plate splines, and T_c refers to appearance perturbation by color transforms. The geometric concentration loss enforces all the pixels that belong to a body part and are spatially close to the center of this body part[41]. The formulation of this loss is defined as follows:

$$
\begin{aligned}
&\mathcal{L}_{\mathrm{geo}} = \sum_{k}\sum_{h,w}\|(h,w)-(c_h^k,c_w^k)\|_2^2\cdot\frac{\boldsymbol{M}_t(k,h,w)}{z^k},\\
&c_h^k = \sum_{h,w} h\cdot\frac{\boldsymbol{M}_t(k,h,w)}{z^k},\ z^k = \sum_{h,w}\boldsymbol{M}_t(k,h,w),
\end{aligned}
\tag{8}
$$

where c_h^k denotes the center of the k-th body part along dimension h, and z^k represents transforming the segmentation map of k-th body part into a spatial probability distribution function. The total loss for the semantic segmentation network is sum of the three losses ($\lambda_1\mathcal{L}_{\mathrm{rec}}+\lambda_2\mathcal{L}_{\mathrm{equ}}+\lambda_3\mathcal{L}_{\mathrm{geo}}$).

Identification loss and triplet loss are the widely-used losses for person Re-ID[42]. Thus, we adopt triplet loss with hard mining strategy $\mathcal{L}_{\mathrm{tri}}$[43] and identification loss with label smoothing regularization $\mathcal{L}_{\mathrm{ide}}$[44] to optimize the aggregation network. Thus, the total loss for this network is sum of the two losses ($\lambda_4\mathcal{L}_{\mathrm{ide}}+\lambda_5\mathcal{L}_{\mathrm{tri}}$). The whole training process of SS-HSP contains two stages. In the first stage, the semantic segmentation network is trained until convergence for the task of predicting future frames. In the second stage, the backbone network is followed by the aggregation network, and the pre-trained segmentation module is optimized until convergence for person re-identification.

Although our semantic segmentation network is inspired by the method mentioned in Ref. [37] (hereinafter, the previous method), there are significant differences between SS-HSP and the previous method. (i) The previous method was proposed for image animation tasks that use a representation consisting of a set of learned key points along with their local affine transformations to encode the motion information. Thus, It cannot directly locate body parts of pedestrians and learn the corresponding motion information, which are the main purpose of video person Re-ID. (ii) The previous method requires at least 2 frames to predict key-point neighborhoods, even during the inference, which makes its predictions highly dependent on the other frame in a pair. In contrast, SS-HSP encodes more semantically meaningful body parts by making independent frame-based predictions, thus can adaptively locate body parts of pedestrians for a single image during the inference. (iii) Different from the previous method using reconstruction loss and equivariance loss, SS-HSP employs reconstruction loss, reinforced equivariance loss, and geometric concentration loss to impel the semantic segmentation network and to estimate more accurate segmentation map of body parts and learn more effective motion feature.

4 Experiments

In this section, we conduct several experiments on two widely-used video datasets to evaluate the effectiveness of SS-HSP. These experiments consist of comparative analysis with state-of-the-art methods and ablation studies.

4.1 Experimental settings

Datasets. MARS dataset[11] is one of the largest video-based person Re-ID dataset, consisting of 1261 identities and a total of 20715 video sequences. Each identity contains 13.2 video sequences on average, and the length of each video sequence varies from 2 to 920 frames, with an average number of 59.5. Following the work[11], we fixedly divide this dataset into 625 identities for training and remain 636 identities for testing. iLIDS-VID dataset [12] is another video person Re-ID dataset. It contains 300 identities from 600 video sequences. Each identity captured from two cameras has a pair of video sequences. Each video sequence contains variable lengths ranging from 23 to 192 video frames, with an average number of 73. Following the work[12], this dataset is randomly divided into 150 identities for training and 150 identities for testing.

Evaluation metrics. Cumulative matching characteristic (CMC) is widely used to quantitatively evaluate the performance of person Re-ID algorithms. The rank-k recognition rate in the CMC curve indicates the probability that an approach retrieves the ground-truth identity in the top-k position. Another evaluation metric is the mean average precision (MAP), which evaluates the algorithms in a multi-shot setting.

Implementation details. The implementation of SS-HSP is based on the PyTorch framework with four Titan RTX GPUs. We randomly select $T=8$ frames from a variable-length video sequence as the input clip. Each min-batch contains 16 pedestrians and 4 input clips for each pedestrian. All video frames are resized to the dimension of $3\times256\times128$, which are then normalized with 1.0/256. The input frames are enlarged by data augmentation including random horizontal flipping and random erasing probability of 0.3. The parameters of $H'\times W'$ are set to 64×32, and K is set to 6. The dimensions of $\tilde{\boldsymbol{f}}_g$, $\tilde{\boldsymbol{f}}_p^k$ and $\tilde{\boldsymbol{f}}_m$ are 256. The hyper-parameters of λ_1, λ_2, ..., λ_5 are set to 1. We adopt the Adam optimizer with the initial learning rate (lr) of 3e^{-4}, the weight decay of 5e^{-4}, and the Nesterov momentum of 0.9. In the first stage, the semantic segmentation network is trained for 120 epochs, which takes about 10 h. lr is decreased by 0.1 after every 40 epochs. In the second stage, the whole model is optimized for 400 epochs, which takes about 12 h. lr is decreased by 0.1 after every 150 epochs.

4.2 Comparison to state-of-the-arts

MARS: In Table 1, 14 state-of-the-art methods of person Re-ID are compared with SS-HSP. The first two approaches belong to image-based person Re-ID, and the remaining approaches belong to video-based person Re-ID. From the results, SS-HSP achieves superior performance in terms of both Rank-1 accuracy and mAP over most of the state-of-the-art methods, especially for image-based Re-ID algorithms. The Rank-1 accuracy and mAP of SS-HSP reach 91.0% and 85.9%, respectively. Compared with the 2nd best method DenseIL, SS-HSP improves Rank-1 accuracy by 0.2%. Considering the high performance, the improvement of SS-HSP is appreciable. The comparison indicates the effectiveness of SS-HSP for learning reinforced appearance and motion representations from video sequences against temporal appearance misalignment and visual ambiguity problems.

iLIDS-VID: In Table 2, 12 state-of-the-art methods of person Re-ID are compared with the proposed SS-HSP. SS-HSP surpasses all the existing methods from Rank-1 to Rank-20 by a large margin, except for the method DenseIL. Especially on Rank-1 accuracy, it boosts the compared method AP3D by 1.6%. Moreover, compared with the 2nd best method DenseIL, SS-HSP improves Rank-5 accuracy by 0.4%. The

Table 1. Performance comparison to the state-of-the-art methods on MARS dataset.

Method	Rank-1 (%)	Rank-5 (%)	Rank-20 (%)	mAP (%)
CNN+XQDA [11]	68.3	82.6	89.4	49.3
QAN [45]	73.5	84.9	91.6	51.7
STAN [16]	82.3	–	–	65.8
M3D [23]	84.4	93.8	97.7	74.0
COSAM [46]	84.9	95.5	97.9	79.9
Snippet [47]	86.3	94.7	98.2	76.1
GLTR [48]	87.0	95.8	98.2	78.5
RGSAT [1]	89.4	96.9	98.3	84.0
AGRL [15]	89.8	96.1	97.6	81.1
TCLNet [42]	89.8	–	–	85.1
STGCN [14]	90.0	96.4	98.3	83.7
AP3D [34]	90.1	–	–	85.1
STRF [49]	90.3	–	–	86.1
DenseIL [50]	90.8	97.1	98.8	87.0
SS-HSP	91.0	96.9	98.6	85.9

Table 2. Performance comparison to the state-of-the-art methods on iLIDS-VID dataset.

Method	Rank-1 (%)	Rank-5 (%)	Rank-20 (%)
CNN+XQDA [11]	53.0	81.4	95.1
QAN [45]	68.0	86.8	97.4
M3D [23]	74.0	94.33	–
COSAM [46]	79.6	95.3	–
STAN [26]	80.2	–	–
AGRL [15]	83.7	95.4	99.5
Snippet [47]	85.4	96.7	99.5
RGSAT [1]	86.0	98.0	99.4
GLTR [48]	86.0	98.0	–
TCLNet [42]	86.6	–	–
AP3D [34]	86.7	–	–
DenseIL [50]	92.0	98.0	–
SS-HSP	88.3	98.4	99.9

comparison shows the effectiveness of the proposed SS-HSP on relatively small video datasets. Note that Snippet and AP3D utilize the optical flow or 3D convolution kernels to learn motion features, VRSTC, AGRL, and RGSAT attempt to learn appearance feature for handling temporal appearance misalignment problem. These methods are all inferior to SS-HSP, which indicates SS-HSP is able to learn reinforced and discriminative appearance and motion features for person Re-ID.

4.3 Ablation studies

Effectiveness of components. Table 3 summarizes the experimental results of the ablation studies for SS-HSP on MARS dataset. Basel, Basel+Part, Basel+Part+Motion, Basel+Part+Motion+TRB denote using SS-HSP to extract the global appearance feature with temporal averaging pooling (TAP), the global and local appearance features with TAP, the appearance and motion features with TAP, the reinforced appearance and motion features with the temporal relation block, respectively. Compared with Basel, Basel+Part boosts Rank-1 accuracy and mAP by 2.2% and 3.7%, respectively. The comparison shows that the semantic segmentation network can precisely learn the segmentation maps of body parts and guide SS-HSP to learn aligned part representations. By utilizing the motion representation, Basel+Part+Motion achieves obvious performance improvement over Basel+Part. The improvement indicates the semantic segmentation network can effectively extract optical flows, which contain abundant complementary information for appearance features. Moreover, by adding the temporal relation block, the best per-

Table 3. Evaluation of the effectiveness of each component within SS-HSP on MARS dataset.

Model	Rank-1 (%)	Rank-5 (%)	Rank-20 (%)	mAP (%)
Basel	86.3	94.6	97.2	79.1
Basel+Part	88.5	95.7	98.0	82.8
Basel+Part+Motion	89.9	96.4	98.4	84.9
Basel+Part+Motion+TRB	91.0	96.9	98.6	85.9

formance is obtained. The boosting demonstrates that this block effectively explores the complementary correlation information among video frames to refine the representations.

Different components of the loss function. The results in Table 4 show the influence of different components of the loss function. SS-HSP w/o $\mathcal{L}_{\text{tri}}$, SS-HSP w/o $\mathcal{L}_{\text{ide}}$, SS-HSP w/o $\mathcal{L}_{\text{equ}}$, and SS-HSP w/o $\mathcal{L}_{\text{geo}}$ denote SS-HSP is trained without triplet loss, identification loss, equivariance loss, and geometric concentration loss, respectively. By comparing SS-HSP w/o $\mathcal{L}_{\text{tri}}$ with SS-HSP w/o $\mathcal{L}_{\text{ide}}$, we can observe that triplet loss can enforce the model to learn more effective representation. Moreover, both of SS-HSP w/o $\mathcal{L}_{\text{tri}}$ and SS-HSP w/o $\mathcal{L}_{\text{ide}}$ are inferior to SS-HSP, indicating that jointly employing triplet loss and identification loss contributes to superior feature representation. Besides, the comparison results of SS-HSP w/o $\mathcal{L}_{\text{equ}}$, SS-HSP w/o $\mathcal{L}_{\text{geo}}$, and SS-HSP show that equivariance loss and geometric concentration loss can impel the semantic segmentation network to learn more accurate human semantic parsing and effective motion representation towards better feature alignment and representation.

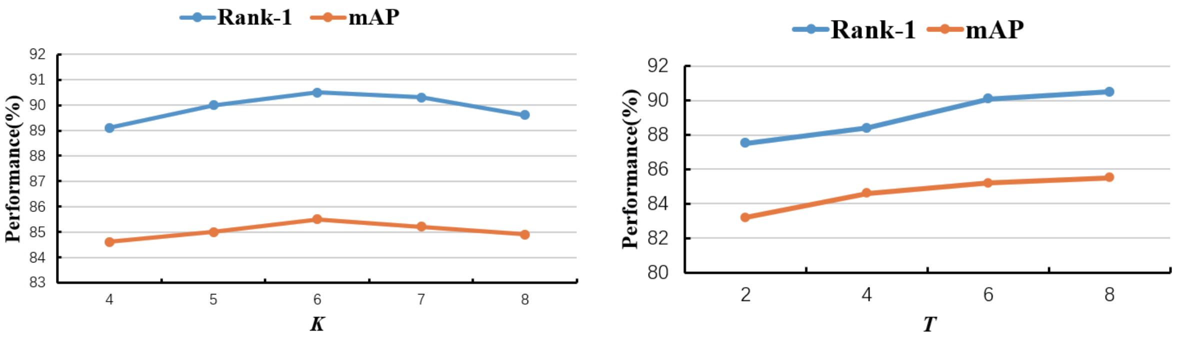
Number of body parts. In Fig. 4a, we investigate the influence of different numbers of body parts on SS-HSP and find the most suitable K. From the results, we can see that the performance of SS-HSP is robust to different values of K. As the number of body parts increases, the performance of SS-HSP improves. SS-HSP obtains the best results with the setting of 6 body parts, and the performance drops when K increases from 6 to 8. We further visualize two examples of the estimated segmentation maps of 6 body parts in Fig. 5, which validates that SS-HSP can precisely locate human body parts and extract aligned local appearance features.

Sequence with different lengths. In Fig. 4b, we investigate the influence of sequence length. We select T frames from a video sequence as the input clip. From the results, we can see that SS-HSP is robust to the variations in T. When the sequence length T increases, the model captures wider range of temporal complementary information and obtains better re-identification performance. The longer sequences bring more computation complexity. Considering the limited computation resources, We set $T = 8$ for SS-HSP in the experiments.

Retrieval results. Fig. 6 shows the retrieval results of three pedestrians by SS-HSP on the MARS dataset. We can observe that Rank-1 retrieval results by SS-HSP are all matching. This indicates CTL effectively alleviates the problem of misalignment and occlusion, viewpoint variation, etc. and realizes precise re-identification.

Table 4. Evaluation of the effectiveness of each component of the loss function on MARS dataset.

Model	Rank-1 (%)	Rank-5 (%)	Rank-20 (%)	mAP (%)
SS-HSP w/o $\mathcal{L}_{\text{tri}}$	87.2	94.7	97.8	81.8
SS-HSP w/o $\mathcal{L}_{\text{ide}}$	88.3	95.5	98.1	83.0
SS-HSP w/o $\mathcal{L}_{\text{equ}}$	89.0	96.0	98.4	84.1
SS-HSP w/o $\mathcal{L}_{\text{geo}}$	90.3	96.5	98.5	85.2
SS-HSP	91.0	96.9	98.6	85.9



(a) (b)

Fig. 4. Parameter analysis of (a) the number of body parts K and (b) the sequence length T on the MARS dataset.

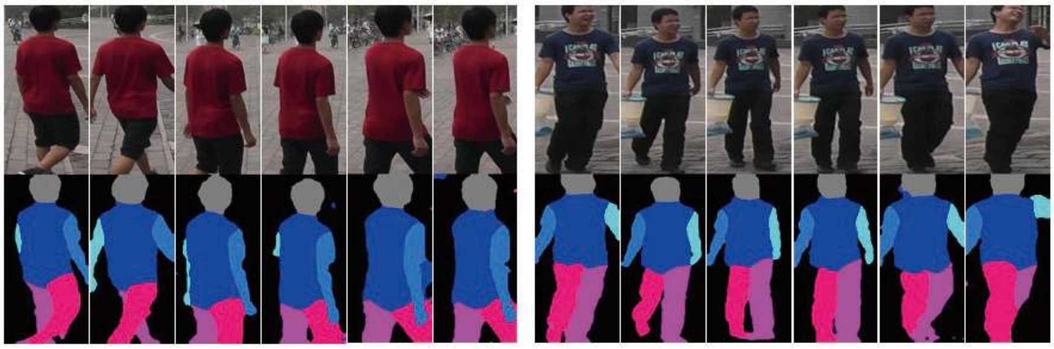


Fig. 5. Visualization results of the estimated segmentation maps of two video sequences.

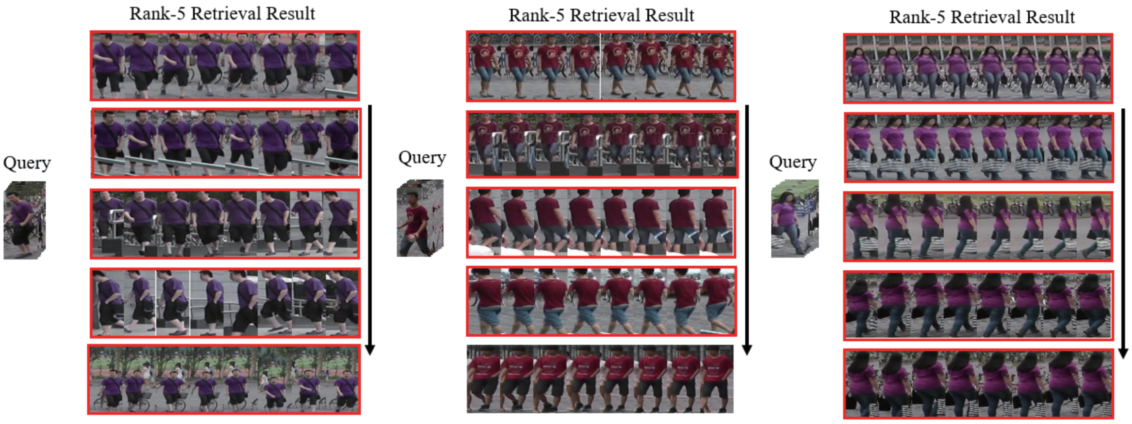


Fig. 6. Example of retrieval results by SS-HSP on MARS dataset. Correct matches are highlighted red.

5 Conclusions

In this work, we propose a novel self-supervised human semantic parsing approach (SS-HSP) for video-based person re-identification. It explores self-supervised learning to precisely locate body parts of pedestrians at pixel-level by estimating the corresponding optical flows between consecutive frames and utilizes the temporal relation information across video frames to learn reinforced appearance and motion representations. The semantic segmentation network builds a pretext task of predicting future frames in a self-supervised learning manner and learns the segmentation maps of body parts and the optical flow field from video sequences. The aggregation network refines the frame-level features by their relation to features of the other frames for accurate matching. Extensive experiments on the two challenging benchmarks have shown that the proposed SS-HSP achieves superior performance over a wide range of state-of-the-art methods.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (62106245), the Fundamental Research Funds for the Central Universities (WK2100000021), and China Postdoctoral Science Foundation Funded Project (2020M671898).

Conflict of interest

The authors declare that they have no conflict of interest.

Biographies

Wei Wu received her B.E. degree in Electronic Information Engineering from the University of Science and Technology of China (USTC) in 2020, and is pursing a Ph.D. degree in the School of Cyber Science and Technology at USTC. Her research interests mainly include computer vision and multimedia.

Jiawei Liu received his B.E. degree from Hefei University of Technology in 2013 and received his Ph.D. degree from the University of Science and Technology of China (USTC) in 2019. He is currently an associate research fellow in the School of Information Science and Technology at USTC. His research interests mainly include computer vision and multimedia.

References

[1] Li X, Zhou W, Zhou Y, et al. Relation-guided spatial attention and temporal refinement for video-based person re-identification. *Proceedings of the AAAI Conference on Artificial Intelligence,* **2020**, *34* (7): 11434–11441.

[2] Cheng Z, Dong Q, Gong S, et al. Inter-task association critic for cross-resolution person re-identification. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, USA: IEEE, **2020**: 2602–2612.

[3] Huang Y, Zha Z J, Fu X, et al. Real-world person re-identification via degradation invariance learning. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, USA: IEEE, **2020**: 14072–14082.

[4] Ding Y, Fan H, Xu M, et al. Adaptive exploration for unsupervised person re-identification. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM),* **2020**, *16* (1): 1–19.

[5] Kalayeh M M, Basaran E, Gökmen M, et al. Human semantic parsing for person re-identification. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE, **2018**: 1062–1071.

[6] Liang X, Gong K, Shen X, et al. Look into person: Joint body parsing & pose estimation network and a new benchmark. *IEEE Transactions on Pattern Analysis and Machine Intelligence,* **2019**, *41* (4): 871–885.

[7] Song C, Huang Y, Ouyang W, et al. Mask-guided contrastive attention model for person re-identification. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE, **2018**: 1179–1188.

[8] Ye M, Yuen P C. PurifyNet: A robust person re-identification model with noisy labels. *IEEE Transactions on Information Forensics and Security,* **2020**, *15*: 2655–2666.

[9] Liu H, Jie Z, Jayashree K, et al. Video-based person re-identification with accumulative motion context. *IEEE Transactions on Circuits and Systems for Video Technology,* **2018**, *28* (10): 2788–2802.

[10] Wang Z, Luo S, Sun H, et al. An efficient non-local attention network for video-based person re-identification. In: ICIT 2019: Proceedings of the 2019 7th International Conference on Information Technology: IoT and Smart City. Shanghai, China: Association for Computing Machinery, **2019**: 212–217.

[11] Zheng L, Bie Z, Sun Y, et al. MARS: A video benchmark for large-scale person re-identification. In: Leibe B, Matas J, Sebe N, et al. editors. Computer Vision – ECCV 2016. Cham, Switzerland: Springer, **2016**: 868–884.

[12] Wang T, Gong S, Zhu X, et al. Person re-identification by video ranking. In: Fleet D, PajdlaT, Schiele B, et al. editors. Computer Vision – ECCV 2014. Cham, Switzerland: Springer, **2014**: 688–703.

[13] McLaughlin N, del Rincon J M, Miller P. Recurrent convolutional network for video-based person re-identification. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Las Vegas, USA: IEEE, **2016**: 1325–1334.

[14] Yang J, Zheng W S, Yang Q, et al. Spatial-temporal graph convolutional network for video-based person re-identification. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, USA: IEEE, **2020**: 3286-3296.

[15] Wu Y, Bourahla O E F, Li X, et al. Adaptive graph representation learning for video person re-identification. *IEEE Transactions on Image Processing,* **2020**, *29*: 8821–8830.

[16] Li S, Bak S, Carr P, et al. Diversity regularized spatiotemporal attention for video-based person re-identification. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE, **2018**: 369–378.

[17] Zhou Z, Huang Y, Wang W, et al. See the forest for the trees: Joint spatial and temporal recurrent neural networks for video-based person re-identification. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Honolulu, USA: IEEE, **2017**: 4747-4756.

[18] Li X, Loy C C. Video object segmentation with joint re-identification and attention-aware mask propagation. In: Ferrari, V, Hebert M, Sminchisescu C, et al. editors. Computer Vision – ECCV 2018. Cham, Switzerland: Springer, **2018**: 93–110.

[19] Jones M J, Rambhatla S. Body part alignment and temporal attention for video-based person re-identification. In: Sidorov K, Hicks Y, editors. Proceedings of the British Machine Vision Conference (BMVC). London: BMVA Press, **2019**, 115: 1−12.

[20] Gao C, Chen Y, Yu J G, et al. Pose-guided spatiotemporal alignment for video-based person re-identification. *Information Sciences,* **2020**, *527*: 176–190.

[21] Liu J, Zha Z J, Chen X, et al. Dense 3D-convolutional neural network for person re-identification in videos. *ACM Transactions on Multimedia Computing, Communications, and Applications,* **2019**, *15* (1s): 1–19.

[22] Chung D, Tahboub K, Delp E J. A two stream siamese convolutional neural network for person re-identification. In: 2017 IEEE International Conference on Computer Vision (ICCV). Venice, Italy: IEEE, **2017**: 1992-2000.

[23] Li J, Zhang S, Huang T. Multi-scale 3D convolution network for video based person re-identification. In: AAAI'19: AAAI Conference on Artificial Intelligence. Honolulu, USA: AAAI Press, **2019**: 1057.

[24] Jin X, He T, Zheng K, et al. Cloth-changing person re-identification from a single image with gait prediction and regularization. [2021-09-01]. https://arxiv.org/abs/2103.15537.

[25] Zhang P, Wu Q, Xu J, et al. Long-term person re-identification using true motion from videos. In: 2018 IEEE Winter Conference on Applications of Computer Vision (WACV). Lake Tahoe, USA: IEEE, **2018**: 494–502.

[26] Zhu K, Guo H, Liu Z, et al. Identity-guided human semantic parsing for person re-identification. In: Vedaldi A, Bischof H, Brox T, et al. editors. Computer Vision – ECCV 2020. Cham, Switzerland: Springer, **2020**: 346-363.

[27] Liao S C, Hu Y, Zhu X Y, et al. Person re-identification by local maximal occurrence representation and metric learning. In: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Boston, USA: IEEE, **2015**, 2197–2206.

[28] Bazzani L, Cristani M, Murino V. Symmetry-driven accumulation of local features for human characterization and re-identification. *Computer Vision and Image Understanding,* **2013**, *117* (2): 130–144.

[29] Zhang L, Xiang T, Gong S. Learning a discriminative null space for person re-identification. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Las Vegas, USA: IEEE, **2016**: 1239-1248.

[30] Zhou Q, Zhong B, Lan X, et al. LRDNN: Local-refining based deep neural network for person re-identification with attribute discerning. In: Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence. Macao: International Joint Conferences on Artificial Intelligence Organization, **2019**: 1041−1047.

[31] Zhang Z, Lan C, Zeng W, et al. Relation-aware global attention for person re-identification. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, USA: IEEE, **2020**: 3183-3192.

[32] Jin X, Lan C, Zeng W, et al. Semantics-aligned representation learning for person re-identification. *Proceedings of the AAAI Conference on Artificial Intelligence,* **2020**, *34* (7): 11173–11180.

[33] You J, Wu A, Li X, et al. Top-push video-based person re-identification. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Las Vegas, USA: IEEE, **2016**: 1345–1353.

[34] Gu X, Chang H, Ma B, et al. Appearance-preserving 3D convolution for video-based person re-identification. In: Vedaldi A, Bischof H, Brox T, et al. editors. Computer Vision – ECCV 2020. Cham, Switzerland: Springer, **2020**: 228–243.

[35] Li S, Yu H, Hu H. Appearance and motion enhancement for video-based person re-identification. *Proceedings of the AAAI Conference on Artificial Intelligence,* **2020**, *34* (7): 11394–11401.

[36] He K, Zhang X, Ren S, et al. Deep residual learning for image recognition. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Las Vegas, USA: IEEE, **2016**: 770–778.

[37] Siarohin A, Lathuilière A, Tulyakov S, et al. First order motion model for image animation. In: Wallach H, Larochelle H, Beygelzimer A et al. editors. Advances in Neural Information Processing Systems. Red Hook, NY: Curran Associates, Inc, **2019**: 3854.

[38] Ronneberger O, Fischer P, Brox T. U-net: Convolutional networks for biomedical image segmentation. In: Navab N, Hornegger J, Wells W, et al. editors. Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015. Cham, Switzerland: Springer,

DOI: 10.52396/JUSTC-2021-0212

2015: 234–241.

[39] Johnson J, Alahi A, Li F F. Perceptual losses for real-time style transfer and super-resolution. In: Leibe B, Matas J, Sebe N, et al. editors. Computer Vision – ECCV 2016. Cham, Switzerland: Springer, **2016**: 694-711.

[40] Siarohin A, Sangineto E, Lathuiliere S, et al. Deformable GANs for pose-based human image generation. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE, **2018**: 3408−3416.

[41] Hung W C, Jampani V, Liu S F, et al. SCOPS: Self-supervised co-part segmentation. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, USA: IEEE, **2019**: 869–878.

[42] Hou R, Chang H, Ma B, et al. Temporal complementary learning for video person re-identification. [2021-09-01]. https://arxiv.org/abs/2007.09357.

[43] Hermans A, Beyer L, Leibe B. In defense of the triplet loss for person re-identification. [2021-09-01]. https://arxiv.org/abs/1703.07737.

[44] Liu J, Zha Z J, Chen D, et al. Adaptive transfer network for cross-domain person re-identification. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach, USA: IEEE, **2019**: 7195–7204.

[45] Liu Y, Yan J, Ouyang W. Quality aware network for set to set recognition. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Honolulu, USA: IEEE, **2017**: 4694–4703.

[46] Subramaniam A, Nambiar A, Mittal A, et al. Co-segmentation inspired attention networks for video-based person re-identification. In: 2019 IEEE/CVF International Conference on Computer Vision (ICCV). Seoul, Korea (South): IEEE, **2019**: 562–572.

[47] Chen D, Li H, Xiao T, et al. Video person re-identification with competitive snippet-similarity aggregation and co-attentive snippet embedding. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE, **2018**: 1169–1178.

[48] Li J, Zhang S, Wang J, et al. Global-local temporal representations for video person re-identification. In: 2019 IEEE/CVF International Conference on Computer Vision (ICCV). Seoul, Korea(South): IEEE, **2019**: 3957–3966.

[49] Aich A, Zheng M, Karanam S, et al. Spatio-temporal representation factorization for video-based person re-identification. In: 2021 IEEE/CVF International Conference on Computer Vision (ICCV). Montreal, Canada: IEEE, **2021**: 152–162.

[50] He T Y, Jin X, Shen X, et al. Dense interaction learning for video-based person re-identification. In: 2021 IEEE/CVF International Conference on Computer Vision (ICCV). Montreal, Canada: IEEE, **2021**: 1470–1481.