Big Data and Information Analytics

https://www.aimspress.com/journal/BDIA

BDIA, 9: 328–349.
DOI:10.3934/bdia.2025015
Received: 16 November 2025
Revised: 02 December 2025
Accepted: 03 December 2025
Published: 04 December 2025

Research article

Human structure modeling for video-based person re-identification without body-part labels

Haotian Chen[1,2,*], Jianyuan Guo[3], Chao Zhang[1] and Zhouchen Lin[4,5]

[1] School of Intelligence Science and Technology, Peking University, No. 5 Yiheyuan Road, Haidian District, Beijing 100871, China

[2] North King Information Technology Co., Ltd., 7th Floor, Qingzheng Building, No. 25 Xisanhuan North Road, Haidian District, Beijing 100089, China

[3] City University of Hong Kong, 83 Tat Chee Avenue, Kowloon Tong, Hong Kong 999077, China

[4] State Key Lab of General AI, School of Intelligence Science and Technology, Peking University, No. 5 Yiheyuan Road, Haidian District, Beijing 100871, China

[5] Institute for Artificial Intelligence, Peking University, No. 5 Yiheyuan Road, Haidian District, Beijing 100871, China

* **Correspondence:** Email: archristy@pku.edu.cn.

Abstract: With the observation that most frames in video-based person re-identification (VReID) capture human figures and that continuous movements intrinsically separate the foreground from the background, this paper demonstrates that it is unnecessary to model human structures with extra key-point estimators that are elaborately pre-trained on costly body-part labels. Specifically, we propose a novel human structure modeling (HTML) module to generate discriminative part-level features from the foreground for VReID without using exclusively annotated part labels. Under the guidance of a simple humanoid topology, HTML extracts coarse body shapes (body proportions) by mapping image patches to both the topological body parts and the background. To compensate for the lack of supervision, a regularization loss is designed for the training of the HTML module to compensate for the lack of part-label supervision. Furthermore, a spatial-temporal part mixer and a spatial-temporal patch mixer are introduced to make its output more discriminative and reliable. Extensive experiments show that our approach achieves competitive performance with a favorable accuracy-efficiency trade-off across multiple benchmarks.

Keywords: unsupervised body-part alignment; human structure modeling; humanoid topology prior; video-based person re-identification; spatial-temporal feature integration; GNN; transformer-based feature fusion; deep video analytics

1. Introduction

Due to the richness in appearance and temporal information, *video-based person re-identification (VReID)*, where each person is represented by a video sequence, has drawn increasing attention over the past few years. Unfortunately, this task remains challenging for video feature representation learning across multiple frames.

A typical VReID pipeline takes *tracklets* (an equal number of frames sampled from a video sequence of a person) as the input, and extracts and integrates spatial-temporal cues to generate discriminative features.

While aggregating *global features* of the whole frames suffers from misalignment and occlusion, *local features* integrate spatiotemporal region/part-level clues in the tracklets, thus making it easier to achieve alignment [1, 2]. As a direct approach, a few works [3, 4] have utilized off-the-shelf pose estimators to locate *body parts* (key-points or joints of humans) for the purpose of extracting aligned part features. In addition to these part-based approaches, recent VReID methods have increasingly adopted large-scale pretrained backbones (e.g., ViT-B/16 [5, 6] and CLIP-based encoders [7, 8]) to further boost the performance. Although effective, such models introduce substantial computational overhead and latency, which results in a clear accuracy-efficiency trade-off that limits their applicability in practical surveillance scenarios.

Despite the promising results of methods relying on part annotations or pretrained estimators, *this paper questions the* ***necessity*** *for VReID to depend on additional semantic labels and extra pre-training*. First, we observe that annotations are not required for foreground segmentation in videos. Indeed, it is difficult for vision-based models to extract the foreground from a static image without supervision. However, for videos (as shown in Figure 1), continuous movements intrinsically separate the foreground from the background. Meanwhile, after automatic human detection and tracking, most frames in VReID tasks, except for those with misdetection, capture human figures. Therefore, human shapes are inducible just by scanning the video frames without annotations (as shown in Figures 7 and 12).

Second, we argue that, contrary to human parsing tasks, imperfect body parts are good enough for a downstream task such as VReID (as shown in Figure 1). VReID aims to retrieve a person of interest across multiple cameras, and thus does not require precise joint positions as part of its final output. Instead, the role of pre-trained estimators in the existing methods is mainly to provide approximate spatial anchors that help align local features in the vicinity of the joints and yield robust representations after concatenation. Therefore, as long as the alignment can be achieved, precise joint locations offer little additional benefit, and approximate ones are sufficiently effective for VReID. This observation indicates that utilizing pre-trained estimators is not strictly necessary for this task.

Third, we demonstrate that aligning local features only requires a small amount of prior knowledge. Specifically, human structures can be modeled by fitting the induced human shape with a simple humanoid topology given as a prior, so as to acquire the approximate positions of the joints (topological nodes, not semantic parts, as shown in Figure 7). As discussed earlier, the corresponding local features aligned in this way are already robust enough for VReID. Therefore, expensive part-level labels are unnecessary in VReID, not to mention the need for extra pre-trained estimators.

As a result, a novel human structure modeling (HTML) module (Section 3.1) is proposed for VReID to pursue discriminative and robust features while modeling body shapes (body proportions)

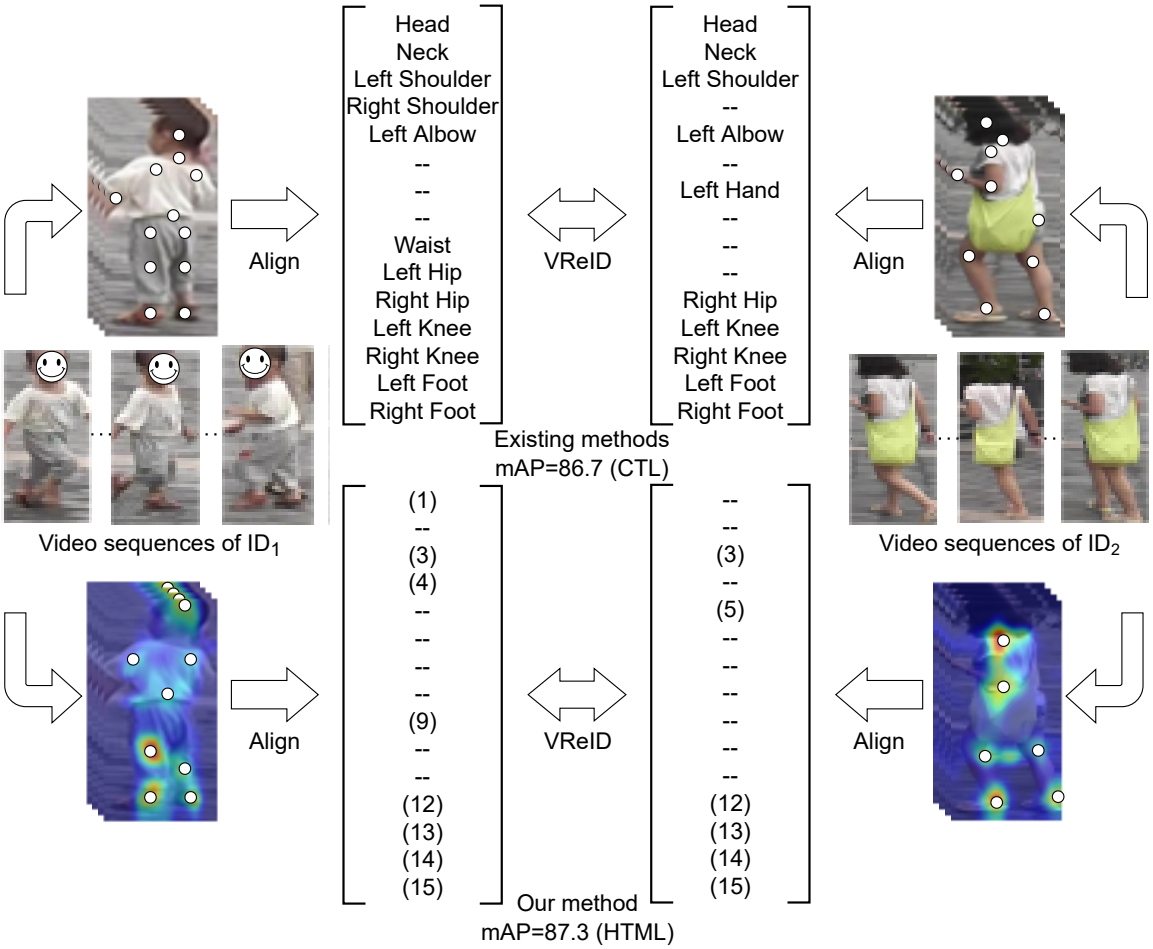

Figure 1. Comparison of existing methods and our HTML module. The heat maps are generated by Grad-CAM [9]. The hollow circles illustrate the joints and topological nodes of pre-trained estimators and the HTML module, respectively. Although the topological nodes lack precise locations, completeness, and clear semantics, the aligned features are still good enough for VReID (see full results in Table 1).

without any extra part annotations. More specifically, it maps frame patches to both the body parts within the same person and the background. Generating part-level features from the foreground, the patch-to-part mapping is also used to extract a body shape from each frame by fitting a humanoid topology provided as a prior. To compensate for the lack of supervision, a regularization loss that maximizes the similarities among the body shapes is designed to train the HTML module. Furthermore, by integrating the spatial and temporal cues, a spatial-temporal part mixer (Section 3.2) and a spatial-temporal patch mixer (Section 3.3) are introduced to make the output features of the HTML module more discriminative and reliable. Eventually, the final part features are aggregated with the global ones in a multi-branch pipeline for the VReID task.

The main contributions of this paper are as follows:

- By proposing a novel HTML module, we provide an alternative to costly body-part annotations and pre-trained estimators for VReID. Under the guidance of a simple humanoid topological prior,

the HTML maps patch features to part features, thus generating aligned representations without the need for expensive annotated part labels.
- Observing that videos in VReID provide enough clues for human structure modeling, HTML can be regarded as a novel way of utilizing spatial-temporal information. We further demonstrate that HTML is compatible with other spatiotemporal integration approaches (i.e., the part and patch mixers) to improve the representations.
- Extensive experiments on multiple VReID datasets demonstrate that our approach achieves strong and competitive results, even when built on a lightweight backbone.

2. Related works

2.1. Using global features

Since deep neural networks are thriving in image classification, the primary choice for VReID is to use global features. With a rich appearance and temporal information provided by videos, many works have focused on *temporal integration* methods, such as temporal fusion [10] and temporal disentanglement [11], to generate more consistent global representations than those of single-frame approaches. For example, Jiang et al. [10] proposed a general temporal fusion framework for both semantic and temporal aspects. Additionally, some works have also attempted to combine temporal fusion and disentanglement to strengthen video-level representations. For example, Liu et al. [12] separated the high- and low-correlated features under the guidance of the fused global features and designed an recurrent-neural-network(RNN)-based module to sequentially accumulate the disentangled information. However, these channel-wise feature aggregation methods suffer from misalignment. Meanwhile, a few works have concentrated on *enhancing* features in different frames. For instance, Hou et al. [13, 14] extracted complementary features from consecutive frames and developed building blocks that can be inserted into a network at any stage. Nonetheless, they still fused the features channel-wise after the enhancement. On the other hand, with global feature extraction as one of the branches, our pipeline keeps its discriminative power while addressing misalignment by combining it with local features.

2.2. Using local features

The body parts used for local features in VReID are either generated from coarse *horizontal division* or extracted by automatic *human parsing*.

To address partial occlusion and inaccurate detection, attention mechanisms [15], transformer-based methods [16], and graph-convolutional-network(GCN)-based methods [17] have been employed to align region features, thereby aiming to pair different regions that contain the same body parts. For example, Yang et al. [17] proposed an spatial-temporal (ST)-GCN whose spatial branch extracts structural information from horizontally partitioned patches. He et al. [16] combined a multi-grained convolutional neural network (CNN) encoder and a Transformer decoder to yield fine-grained spatial-temporal features. However, features extracted from coarsely partitioned regions are sensitive to background clutter, which degrades the performance.

Human parsing aims to segment a human image into different semantic body parts, such as the head, torso, arms, and legs. With automatic human parsing, one solution for VReID is to utilize semantic

attributes [18], and another is to extract human skeletons from each frame [3,4]. In a similar way to our work, Liu et al. [3] took advantage of the correlation and topology of the human body and constructed a ST graph, however, their graph was built on multiple granularities, whereas ours is built on a single granularity. Although the above methods intend to attain the aligned local part features, they depend on extra key-point estimators pre-trained on costly part labels. In contrast, HTML does not require additional data or pre-training, but is jointly trained in an unsupervised manner.

2.3. Unsupervised human parsing in ReID

A few recent works have extended unsupervised human parsing and body-part segmentation to ReID. Specifically, instead of skeleton extraction, Subramaniam et al. [19] used frame-level co-segmentation to generate body parts. Jiang et al. [4] applied attention mechanisms to discover the same parts in different images. However, their module was better trained with a semi-supervised learning strategy, whereas our approach is entirely free from part labels. Moreover, our work leverages a simple human topology as a prior to obtain discriminative and robust local representations for VReID.

3. Our approach

In this section, we will first elaborate on HTML that models human structures without pre-trained key-point estimators or body-part labels. Then, we will describe the two mixers that integrate ST information into reliable patch and part features. Finally, we will give an overview of our pipeline, which is a multi-branch architecture composed of HTML and the aforementioned two mixers.

3.1. HTML

HTML is proposed to realize body-shape modeling without part labels. As illustrated in Figure 2, it does not locate every semantic body part in each frame. Instead, it calculates an attention map from frame patches to both the body parts of the same person and the background. Focusing on the foreground, the patch-to-part mapping is used to extract a body shape from each frame by fitting a humanoid topology provided as a prior. With the intuition that the body shapes of different people are similar to each other, a regularization loss is introduced for training the HTML module, thereby encouraging it to maximize the similarities among the body shapes.

Formally, we denote the input tracklet as $\{\boldsymbol{I}_t | t = 1, \cdots, T\}$, where T is the number of frames. A pre-trained backbone network (ResNet-50 [20] in our work) is used to initialize the frame-level features. Each stage of the network generates a feature map $\boldsymbol{F}^t \in \mathbb{R}^{C \times H \times W}$ for the t^{th} frame, where C, H, and W denote the channel size, height, and width of the feature maps, respectively, omitting the stage indicator s for simplicity.

Let the patch features be the pixels on $\boldsymbol{F}^t$, denoted by $\boldsymbol{F}^t_{\text{patch}} = \{\boldsymbol{f}^{t,p}_{\text{patch}} | \boldsymbol{f}^{t,p}_{\text{patch}} \in \mathbb{R}^C,\ p = 1, \cdots, H \cdot W\}$, where the number of patches equals to $H \cdot W$. The part features are denoted as $\boldsymbol{F}^t_{\text{part}} = \left\{ \boldsymbol{f}^{t,p}_{\text{part}} \middle| \boldsymbol{f}^{t,p}_{\text{part}} \in \mathbb{R}^C, p = 1, \cdots, P \right\}$, where P corresponds to the number of body parts defined by the humanoid topology introduced later in this section.

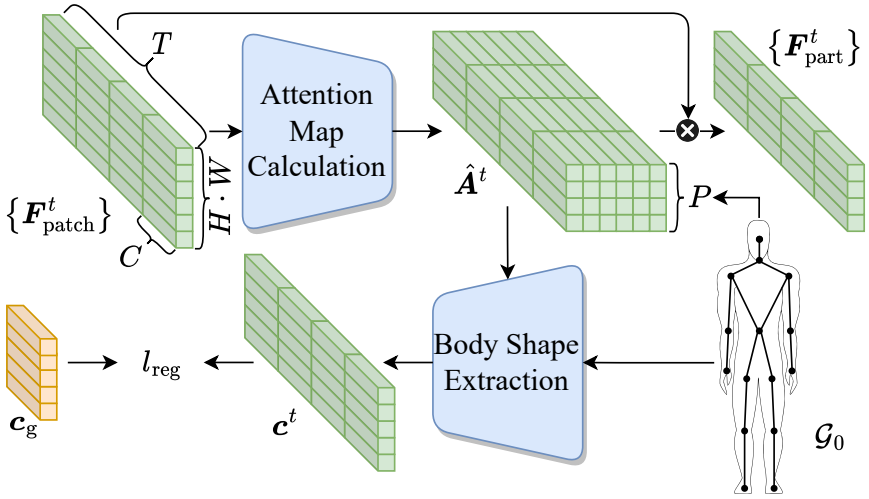

Figure 2. The detailed structure of our HTML module.

3.1.1. Attention map calculation

Patches in each frame contribute differently to different body parts: some contain a single body part, some cover several parts, and the rest correspond to background regions that contain features irrelevant to VReID. With the assumption that continuous body movements can separate the foreground from the background, we introduce an attention classifier to calculate the weights of patches for every body part: $\boldsymbol{A}^t = \left(a^t_{i,j}\right) = \text{FC}\left(\boldsymbol{F}^t_{\text{patch}}\right) \in \mathbb{R}^{(H\cdot W)\times(P+1)}$. Here, FC$(\cdot)$ is a fully connected (FC) neural network that maps from the C-dimensional feature space to $P+1$ classes, in which the additional class is used as a placeholder for the background.

Then, $\boldsymbol{A}^t$ is normalized into a patch-to-part attention map $\widehat{\boldsymbol{A}}^t = \left(\widehat{a}^t_{i,j}\right) \in \mathbb{R}^{(H\cdot W)\times P}$ by a softmax function: $\widehat{a}^t_{i,j} = \frac{\exp(a^t_{i,j})}{\sum_{k=1}^{P}\exp(a^t_{i,k})}$. It is important to emphasize that the last class is dropped, since it does not correspond to any body part.

After acquiring $\widehat{\boldsymbol{A}}^t$, we use $\boldsymbol{F}^t_{\text{part}} = \widehat{\boldsymbol{A}}^{t\intercal}\boldsymbol{F}^t_{\text{patch}} \in \mathbb{R}^{P\times C}$, to calculate the part features, where $\intercal$ indicates the transpose operation of a matrix.

3.1.2. Body shape extraction

To extract body shapes, we need to transform the spatial structures of patches into body parts and fit them to a humanoid topology.

First, we define a distance matrix $\boldsymbol{D} = \left(d_{i,j}\right) \in \mathbb{R}^{(H\cdot W)\times(H\cdot W)}$ in the image, where $d_{i,j} = \text{d}\left(\text{patch}_i, \text{patch}_j\right)$. Here, d$(\cdot,\cdot)$ calculates the Euclidean distance between the two elements, and patch_i is the coordinate of the i^{th} patch in an image: $\text{patch}_i = (h, w)$, $1 \leqslant h \leqslant H \in \mathbb{Z}$, and $1 \leqslant w \leqslant W \in \mathbb{Z}$. Then, we define a modified distance matrix $\widetilde{\boldsymbol{D}}$ by adding an identity matrix $\boldsymbol{I}$ to $\boldsymbol{D}$, so that the diagonal elements of $\widetilde{\boldsymbol{D}}$ equal the area of each patch, and the off-diagonal elements represent the distance between two patches.

The spatial structures provided by $\widetilde{\boldsymbol{D}}$ can be transformed from the patch space to the part space via $\widehat{\boldsymbol{A}}^t$, and the body shape matrix of frame t can be calculated by $\boldsymbol{S}^t = \widehat{\boldsymbol{A}}^{t\intercal}\widetilde{\boldsymbol{D}}\widehat{\boldsymbol{A}}^t \in \mathbb{R}^{P\times P}$. $\boldsymbol{S}^t$ contains three types of measures of body shape: the sizes of body parts, the structural relationships between

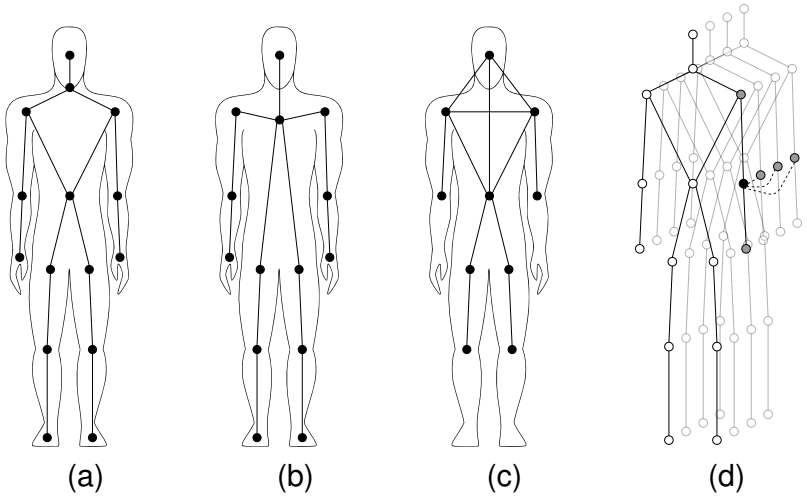

Figure 3. Humanoid topological graphs. The nodes represent the body parts or key-points. The neighboring nodes in each frame are linked with solid lines. (a) Defined by OpenNI [21]. (b) Extracted by OpenPose [22] from COCO [23]. (c) Simplified version of skeleton. (d) Spatiotemporal graph designed for the ST-GCN. For a given node (black), its neighbors are colored in gray, while the rest are hollow. The dashed curves link the same body part in different frames.

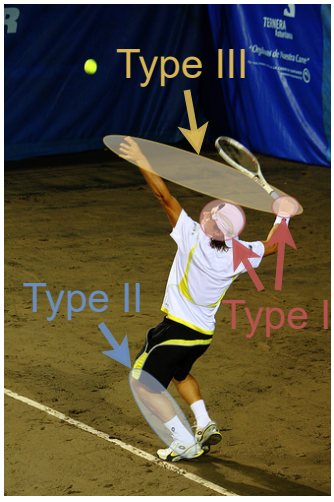

Figure 4. Three types of measures of body shape. In this example, the size of the head (Type I) is proportional to the size of the right hand (Type I) and to the length between the left knee and left foot (Type II), even when the person is playing tennis. However, this proportionality does not hold for the distance between the two hands (Type III).

physically connected body parts, and the relative distances between non-adjacent parts.

To capture a stable body shape, the third type, which varies across different frames, should be filtered out (as illustrated in Figure 4), which can be achieved by providing a small amount of prior knowledge of humanoid topology. Let $\mathcal{G}_0 = (\mathcal{V}_0, \mathcal{E}_0)$ be a predefined topological graph of a human

body (defined by OpenNI [21] in our work, as shown in Figure 3(a)), where $\boldsymbol{v}_i \in \mathcal{V}_0, i = 1, \cdots, P$ refers to a body part, and $(\boldsymbol{v}_i, \boldsymbol{v}_j) \in \mathcal{E}_0$ represents the structural link between the connected parts. As mentioned before, $\mathcal{G}_0$ also determines the number of body parts P. Given $\mathcal{G}_0$, a constant mask $\boldsymbol{M} = \left(m_{i,j}\right) \in \mathbb{R}^{P \times P}$, where

$$m_{i,j} = \begin{cases} 1, & \text{if } (\boldsymbol{v}_i, \boldsymbol{v}_j) \in \mathcal{E}_0 \text{ or } i = j, \\ 0, & \text{otherwise} \end{cases}, \tag{3.1}$$

is used for body shape extraction: $\widetilde{\boldsymbol{S}}^t = \boldsymbol{S}^t \odot \boldsymbol{M}$, where $\odot$ is the Hadamard product operation. Finally, $\widetilde{\boldsymbol{S}}^t$ is reshaped into a body-shape vector $\boldsymbol{c}^t$ by discarding the zeroed elements.

3.1.3. Regularization loss

Since the body shapes on each frame (i.e., $\boldsymbol{c}^t$) are extracted without supervision, we introduce a regularization loss l_{reg} for the HTML module's training.

It is commonly acknowledged that the body shapes of different people are similar to each other. This is especially true when the inputs of VReID are the results of automatic human detection, tracking, and resizing (e.g., the boy and the woman shown in Figure 1). Thus, l_{reg} is defined as follows: $l_{\text{reg}} = \frac{1}{T}\sum_{t=1}^{T} \text{MSE}\left(\boldsymbol{c}^t, \boldsymbol{c}_{\text{g}}\right)$, where MSE stands for the mean squared error. $\boldsymbol{c}_{\text{g}}$ is a globally shared cluster center of $\{\boldsymbol{c}^t\}$ and can be viewed as an average body shape corresponding to a "canonical" body, which is akin to the Vitruvian Man. By minimizing l_{reg} during training, the HTML module is encouraged to generate body shapes with high mutual similarity.

3.2. Spatial-temporal part mixer

We develop a ST part mixer to make the output of the HTML module more discriminative for VReID, as spatial and temporal cues provide both ST dependencies and structural information of the body parts.

Following [3], we employ a small (2-layer) ST-GCN to further propagate part-related messages according to a ST graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ (as illustrated in Figure 3(d)) defined on a tracklet.

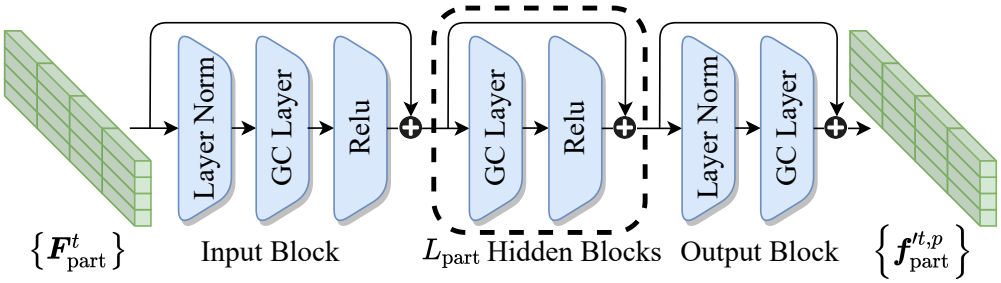

Figure 5. Our ST-GCN implemented with GC layers [24], residual connections [20], and layer normalization [25]. The blocks inside the dashed line are repeated L_{part} times.

Next, we build the GCN-based ST part mixer with graph convolutional (GC) layers [24], residual connections [20], and layer normalization [25]. As shown in Figure 5, our version of the ST-GCN consists of an input block, L_{part} hidden blocks, and an output block. After the mixing, we process the final part features $\left\{\boldsymbol{f}'^{t,p}_{\text{part}} \middle| \boldsymbol{f}'^{t,p}_{\text{part}} \in \mathbb{R}^C, p = 1, \cdots, P, t = 1, \cdots, T\right\}$ through spatial concatenation and

temporal average pooling (TAP) to generate the final local features: $\boldsymbol{f}_{\mathrm{l}} = \left[\mathrm{TAP}\left(\boldsymbol{f}_{\mathrm{part}}^{\prime t,1}\right); \mathrm{TAP}\left(\boldsymbol{f}_{\mathrm{part}}^{\prime t,2}\right); \cdots; \mathrm{TAP}\left(\boldsymbol{f}_{\mathrm{part}}^{\prime t,P}\right)\right]$.

3.3. Spatial-temporal patch mixer

HTML requires patch features as the input. Since the backbone network is built for single frames and does not blend temporal cues with spatial ones, we construct a ST patch mixer to generate more reliable patch features.

The mixer is designed with a spatial embedding followed by L_{patch} layers of Transformer encoders [26]. By fusing ST information, it also changes the feature dimension from C to C', where C' is the size of the hidden layers. Experiments show that a shallow (2-layer) transformer with low computational overhead achieves a superior performance.

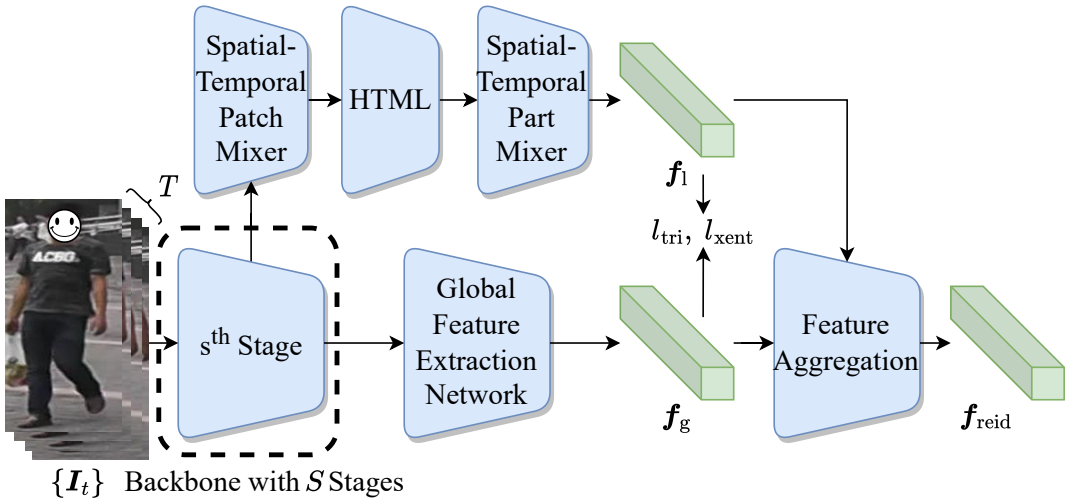

Figure 6. The overview of our pipeline for VReID. The module inside the dashed line repeats S times. Only the output of certain stages is used for multi-branch feature extraction.

3.4. Pipeline for VReID

Multi-branch architectures have achieved great success in vision tasks [14]. Therefore, to further enhance the capacity of VReID representations, we design a pipeline (as shown in Figure 6) that contains multiple branches for global and local feature extraction at different stages. The local features $\boldsymbol{f}_{\mathrm{l}}$ are extracted using the proposed network (HTML and the two mixers); then, they are aggregated with the global features $\boldsymbol{f}_{\mathrm{g}}$ to yield the final representations $\boldsymbol{f}_{\mathrm{reid}}$ for VReID.

During training, the frame sampling strategy plays an important role in efficient learning [27]. Therefore, we modify the popular PK sampler (a mini-batch sampling method that includes P distinct person identities, each represented by K samples) [28], thus making it more informative and more efficient for VReID. Specifically, we arbitrarily select Q persons and then randomly choose K tracklets per person, thus ensuring that most of them come from different cameras.

To train our pipeline, we calculate a batch triplet loss l_{tri} and a cross entropy loss l_{xent} following [13] for both $\boldsymbol{f}_{\mathrm{g}}$ and $\boldsymbol{f}_{\mathrm{l}}$ (as shown in Figure 6). To encourage local feature diversity, we also adopt the diversity loss l_{div} for $\boldsymbol{F}_{\mathrm{part}}^{t}$ proposed in [3]. In the end, the total loss l for model optimization is the

combination of the four losses: $l = \lambda_1 \cdot l_{\text{tri}} + \lambda_2 \cdot l_{\text{xent}} + \lambda_3 \cdot l_{\text{div}} + \lambda_4 \cdot l_{\text{reg}}$, where $\lambda_{1\sim4}$ are the balance weights.

Finally, $\boldsymbol{f}_{\text{g}}$ and $\boldsymbol{f}_{\text{l}}$ obtained from different stages of the backbone network are aggregated via concatenation to form a single representation vector $\boldsymbol{f}_{\text{reid}}$ for the VReID task.

Table 1. Comparison with state-of-the-art methods without large-scale pretrained backbones across MARS [29], LS-VID [30], iLIDS-VID [31], and PRID-2011 [32]. The methods are grouped into temporal integration (TI), temporal enhancement (TE), extra data (ED), horizontal division (HD), and human parsing (HP), and are sorted by mAP on MARS within each group. The best results are shown in **bold**, and the second-best are underlined. * indicates results that are not reported in the original papers but are reproduced using their public code.

Methods		MARS		LS-VID		iLIDS-VID		PRID-2011	
Groups	Names	mAP(%)↑	R-1(%)	mAP(%)	R-1(%)	R-1(%)	R-5(%)	R-1(%)	R-5(%)
TI	GRL [12]	84.8	91.0	62.1*	74.5*	90.4	98.3	96.2	**99.7**
	Jiang et al. [10]	85.2	87.1	–	–	87.7	–	95.8	–
	PSTA [33]	85.8	91.5	–	–	91.5	98.1	95.6	98.9
	STRF [11]	86.1	90.3	–	–	89.3	–	–	–
TE	STMN [34]	84.5	90.5	69.2	82.1	–	–	–	–
	TCLNet [13]	85.1	89.8	74.4*	83.5*	86.6	–	–	–
	BiCnet-TKS [14]	86.0	90.2	75.1	84.6	–	–	–	–
	SINet [35]	86.2	91.0	79.6	87.4	**92.5**	–	96.5	–
ED	DL+CF-AAN [36]	86.5	91.3	–	–	–	–	–	–
HD	STGCN [17]	83.7	90.0	–	–	–	–	–	–
	SGWCNN [38]	85.7	90.0	–	–	87.8	96.0	–	–
	KMPNet (MGH) [37]	86.6	**92.0**	–	–	–	–	–	–
	DenseIL [16]	87.0	90.8	–	–	92.0	98.0	–	–
HP	AMEM [18]	79.3	86.7	–	–	87.2	97.7	93.3	98.7
	Hu et al. [39]	79.6	89.9	–	–	87.9	93.6	95.9	99.6
	SSN3D [4]	86.2	90.1	–	–	88.9	–	–	–
	CTL [3]	86.7	91.4	–	–	89.7	97.0	–	–
HP	HTML (Ours)	**87.3**	91.9	**80.1**	**87.6**	92.0	**99.3**	**96.6**	98.9

4. Experiments and discussion

4.1. Experimental settings

4.1.1. Datasets

Following [35], the proposed network is evaluated on two large-scale VReID datasets*(i.e., MARS [29] and LS-VID [30]) and on two small-scale datasets (i.e., iLIDS-VID [31] and PRID-2011 [32]).

*The DukeMTMC-VideoReID dataset has been retracted.

4.1.2. Evaluation metrics

We adopt standard evaluation metrics (i.e., Rank-1 (R-1) and Rank-5 (R-5) from the cumulative matching characteristics (CMC) [40]), and the mean average precision (mAP) [41] to assess the performance of different methods. R-1 measures the single top-1 accuracy, while the mAP evaluates the overall retrieval ranking.

4.1.3. Implementation details

Following [13], a tracklet consists of $T = 4$ frames sampled from a video sequence. We apply temporal data augmentation by randomly cropping with a stride of eight. Every mini-batch contains $Q = 8$ persons, each with $K = 4$ tracklets. The input frames are resized to 256×128, and horizontal flipping and random erasing [42] are adopted for spatial data augmentation. An Adam optimizer [43] with a weight decay of 5×10^{-4} is used for 150 epochs of training. We employ a warmup strategy [44], thereby linearly increasing the learning rate from 3.5×10^{-5} to 3.5×10^{-4} during the first 10 epochs. Then, the learning rate decays by a factor of 0.1 at the 40^{th}, 80^{th}, and 120^{th} epochs.

For global feature extraction, we adopt the temporal saliency erasing (TSE) module introduced in [13]. The backbone is branched at stages three and four for both the TSE and the proposed network. Both L_{patch} and L_{part} are set to two. P is 15, which is the number of nodes in Figure 3(a). C' is set to 768, following the standard configuration of ViT [26]. $\lambda_{1\sim3}$ are all set to one following [3], and $\lambda_4 = 1$ is selected through a grid search (as shown in Table 4). Our method is implemented in PyTorch 1.7.1 and trained on a 2-GPU (Titan XP 12 GB) machine running Ubuntu 18.04.

Following [13], each testing video sequence is split into 4-frame tracklets, whose features are extracted with the proposed pipeline and averaged into a single representation vector. Finally, cosine distances between the features are computed for VReID retrieval.

4.2. Comparison with existing methods

To isolate the contribution of our approach from differences in the backbone capacity or pretraining scale, we first evaluate it against state-of-the-art methods without large-scale pretrained backbones across multiple datasets, as shown in Table 1. The methods grouped under temporal integration (TI), temporal enhancement (TE), and extra data (ED) are those without local features, while the methods that involve local features are categorized into horizontal division (HD) and human parsing (HP). In addition, the methods within each group are sorted by their mAP performance on MARS.

Several conclusions can be drawn from the results on MARS:

- Instead of exhibiting a trade-off between mAP and R-1, our method simultaneously maintains high performances on both metrics simultaneously.
- DenseIL [16] achieves the second-best mAP of 87.0%. It is constructed with CNN encoders and deep Transformer decoders and is trained on a 4-GPU setup. In contrast, our method only uses a shallow Transformer (two layers), requires half the number of GPUs, and surpasses DenseIL by 1.1% in R-1 accuracy.
- Compared to our approach, KMPNet (MGH) [37] trades a 0.1% gain in R-1 for a 0.6% drop in mAP. It employs a 27-layer GCN model along with an off-the-shelf pose estimator. In contrast, our method is jointly trained without additional body-part labels, and our ST-GCN module is much more lightweight (only two layers).

- CTL [3] also leverages human topology by constructing a spatial-temporal graph. However, it relies heavily on a pre-trained key-point estimator, and its graph is defined across multiple granularities. Instead, our method operates on a single granularity and surpasses CTL by 0.6% in mAP and 0.5% in R-1.
- SSN3D [4] employs an attention mechanism similar to ours. However, rather than using a simple topological prior under an unsupervised scheme, SSN3D relies on a complex semi-supervised strategy. Consequently, our method significantly outperforms SSN3D on all metrics.

Then, we extend the analysis to methods that incorporate large-scale pretrained backbones by additionally reporting their model complexity and performance together with ours (Table 2). The large-scale pretrained backbones (e.g., ViT-B/16 [5, 6] and CLIP-ViT-B/16 [7, 8]) contain substantially larger capacities and considerably higher computational costs compared to our ResNet-50 backbone [45, 46]. Despite these differences in backbone capacity, our method achieves a comparable performance to these large-scale pretrained models on both the MARS and LS-VID benchmarks—outperforming some methods on certain metrics while producing slightly lower results on others. This demonstrates that the HTML module provides a strong accuracy-efficiency trade-off, thereby enabling competitive video-based ReID performance with a significantly more lightweight backbone.

Table 2. Comparison of model complexity and performance with state-of-the-art methods using large-scale pretrained backbones.

Methods	Backbone			MARS		LS-VID	
	Models	Params	FLOPs	mAP(%)	R-1(%)	mAP(%)	R-1(%)
CAViT [5]	ViT-B/16	86.6 M	17.5G	87.2	90.8	79.2	89.2
TCViT [6]	ViT-B/16	86.6 M	17.5 G	87.6	91.7	83.1	90.1
TF-CLIP [7]	CLIP-ViT-B/16	86.6 M	17.5 G	89.4	93.0	83.8	90.4
HAMoBE [8]	CLIP-ViT-B/16	86.6 M	17.5 G	91.1	94.6	85.2	92.1
HTML (Ours)	ResNet-50	25.6 M	4.1 G	87.3	91.9	80.1	87.6

Table 3. Ablation study on the key components.

Methods	Components				MARS	
	HTML	Part mixer	Patch mixer	Regularization loss	mAP(%)	R-1(%)
Baseline	–	–	–	–	85.9	90.6
Ours	✓	–	–	–	86.6	91.4
Ours	✓	✓	–	–	86.8	91.4
Ours	✓	✓	✓	–	87.0	91.2
Ours	✓	✓	✓	✓	**87.3**	**91.9**

Table 4. Ablation study on the design choices of different modules. "Norm.", "attn.", "spa.", "temp." and "emb." stand for normalization, attention, spatial, temporal, and embedding, respectively.

Modules	Methods	mAP(%)	R-1(%)
	Baseline	85.9	90.6
HTML	Ours (w/ simplified prior)	86.7	90.9
	Ours (w/ OpenPose [22])	87.0	91.3
	Ours (w/ shape norm.)	85.7	90.4
Part mixer	Ours ($L_{\text{part}} = 0$)	87.0	91.4
	Ours ($L_{\text{part}} = 1$)	86.9	91.4
	Ours ($L_{\text{part}} = 3$)	**87.3**	91.8
	Ours (w/ batch norm.)	**87.3**	91.4
	Ours (w/ dropout)	85.8	90.6
	Ours (w/ multiplicative attn.)	86.2	90.6
	Ours (w/ additive attn.)	79.0	86.1
Patch mixer	Ours (w/o spa. emb.)	87.0	91.2
	Ours (w/ temp. emb.)	87.0	91.2
	Ours ($L_{\text{patch}} = 1$)	86.8	91.3
	Ours ($L_{\text{patch}} = 3$)	80.6	87.4
	Ours ($H = 6$)	87.2	91.6
	Ours ($H = 24$)	82.9	88.5
Further studies	Ours (branched@stage 3)	86.2	90.6
	Ours (branched@stage 4)	85.5	90.6
	Ours (w/ PK sampler)	86.8	90.8
	Ours ($\lambda_4 = 0.5$)	86.7	91.5
	Ours ($\lambda_4 = 0.75$)	86.6	91.4
	Ours ($\lambda_4 = 1.25$)	87.1	90.9
	Ours ($\lambda_4 = 1.5$)	86.9	91.1
	Ours	**87.3**	**91.9**

4.3. Ablation study

4.3.1. Key components

Table 3 reports the ablation results of the key components. **i)** The baseline is built upon our multi-branch pipeline by replacing the three proposed modules with FC layers. **ii)** The performance is significantly improved by applying the HTML module, which indicates that it is beneficial to model body shapes from videos and to generate more discriminative features for VReID. **iii)** The mixers further improve the performance, thereby demonstrating that the HTML module is highly compatible with other spatiotemporal integration methods. Meanwhile, since we implement only a shallow Transformer and a small ST-GCN (both with two layers), the additional computational cost is negligible. **iv)** Applying the patch mixer lowers R-1 by 0.2%, but boosts the mAP by 0.2% at the same time. One possible explanation is that, by blending features across time and space, the patch

mixer may slightly reduce the discriminability of the top-1 clip representations while improving those from other ranks. **v)** With the regularization loss, both the mAP and R-1 are further improved, thus showing the effectiveness of encouraging a greater similarity among the extracted body shapes.

4.3.2. HTML

Different topological priors affect both the extracted body shape and the structure of the spatial-temporal graph for ST-GCN. We compare three types of humanoid topologies: OpenNI [21], OpenPose [22], and a simplified skeleton (as shown in Figure 3(a)–(c), respectively. The "HTML" part in Table 4 shows that even the simplest prior achieves a better mAP and R-5 than the baseline. Moreover, the network equipped with the OpenNI-defined topological prior [21] achieves the best performance, which is likely due to its richer structural complexity.

In addition, we attempt to normalize the body-shape vector, since the body shapes of different people should be proportional rather than identical. However, the result in the last row of the "HTML" part suggests the opposite trend. This is because the input videos have already been preprocessed via human detection, tracking, and resizing. Therefore, the body shapes do not require redundant normalization. Furthermore, normalization becomes challenging due to the incompleteness of human bodies in certain frames caused by misdetection and occlusion.

4.3.3. Spatial-temporal part mixer

Different design choices are explored for the ST part mixer. The "part mixer" part in Table 4 reports the results for various network depths. The performance improves as L_{part} increases but peaks at $L_{\text{part}} = 2$. In addition, deeper GCNs may suffer from over-smoothing, which explains the performance saturation.

Additionally, we examine different regularization strategies within the GCN layers. We observe that the normalization methods outperform the dropout [47], and that layer normalization yields better results than batch normalization [48].

Attention mechanisms are a common design option for GCN modules. However, the last two rows in the "part mixer" section indicate that two popular attention mechanisms actually degrade the performance in our setting.

4.3.4. Spatial-temporal patch mixer

The first two rows of the "patch mixer" section in Table 4 investigate the effects of applying spatial and temporal embeddings before the Transformer layers in the patch mixer. Notably, while both the mAP and R-1 are boosted by the spatial embedding, adding the temporal embedding has a negative impact. We conjecture that one possible reason is that the tracklet length ($T = 4$) is too short for the temporal embedding to be effective.

To further study the Transformer layers, we train and test the network with different numbers of layers and attention heads. As shown in the "patch mixer" section of Table 4, increasing L_{patch} and H initially improves the performance but eventually worsens it due to overfitting. In addition, deeper networks require substantially more computational resources (4 GPUs). Therefore, we keep the Transformer shallow (2 layers).

4.3.5. Further studies

Additional experiments are conducted to explore different design choices for the remaining modules. The first two rows of the "further study" section in Table 4 report the results when branching the pipeline at different stages of the backbone, thus showing that the performance degrades in both under-branching settings. Conversely, over-branching requires unaffordable computational resources (more than 4 GPUs). Moreover, the last row of the "further study" section indicates that training with our modified sampler improves both the mAP and R-1.

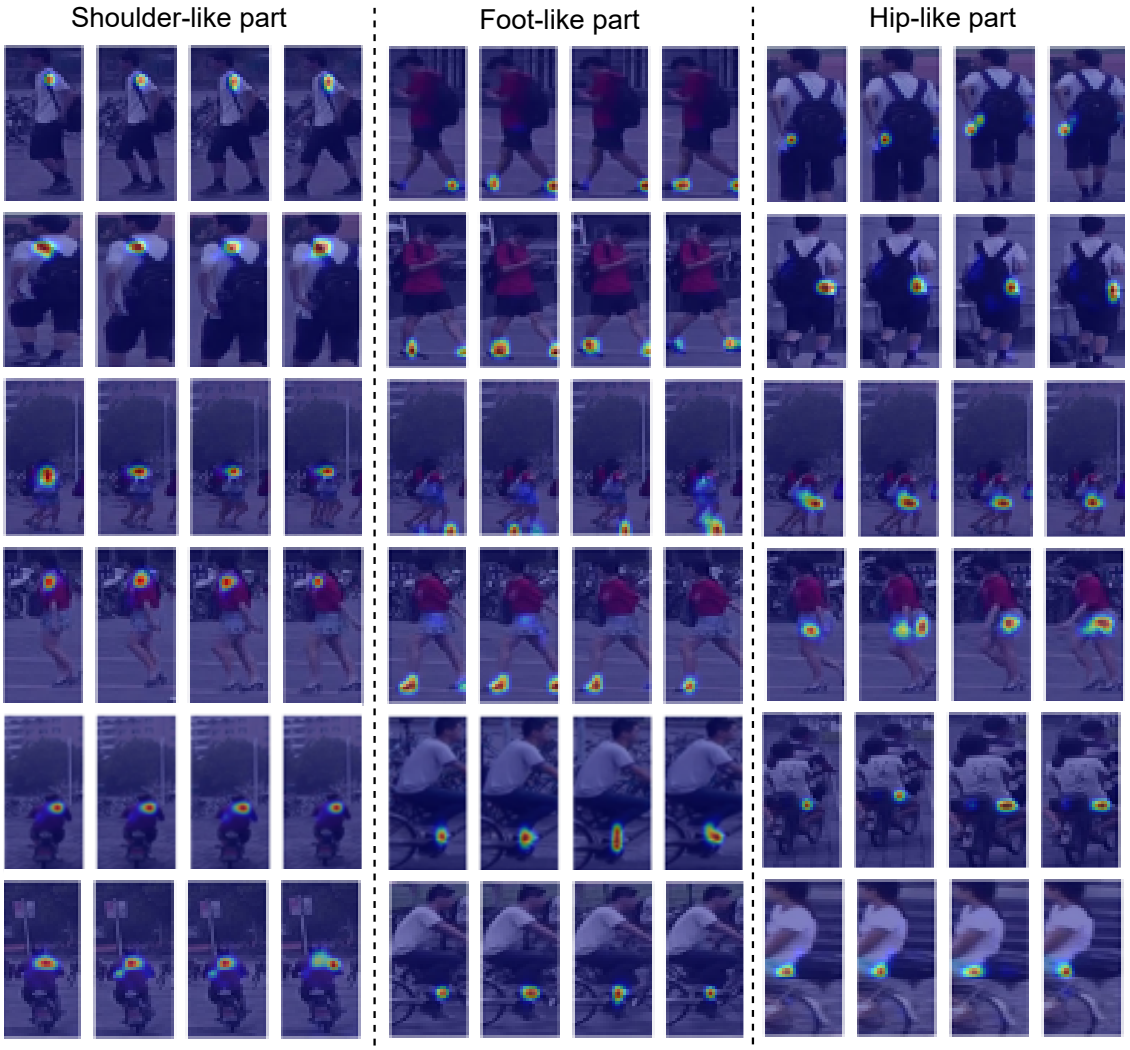

Figure 7. Visualization of the corresponding regions of shoulder-like, foot-like, and hip-like parts, which are well suited for alignment. Best viewed in color.

4.3.6. Qualitative results

We visualize the corresponding regions of the parts in $\widehat{\boldsymbol{A}}^t$. Figure 7 shows that the HTML module indeed locates aligned parts even without semantic labels.

We offer qualitative justifications for why the HTML module outperforms pre-trained key-point

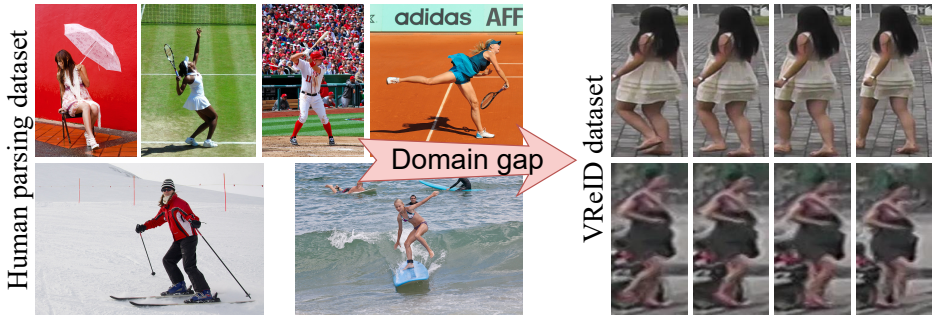

Figure 8. Samples from COCO [23] and MARS [29], resized to the same height.

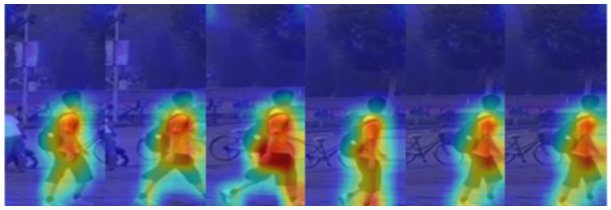

Figure 9. Features learned by CTL [3]. Best viewed in color.

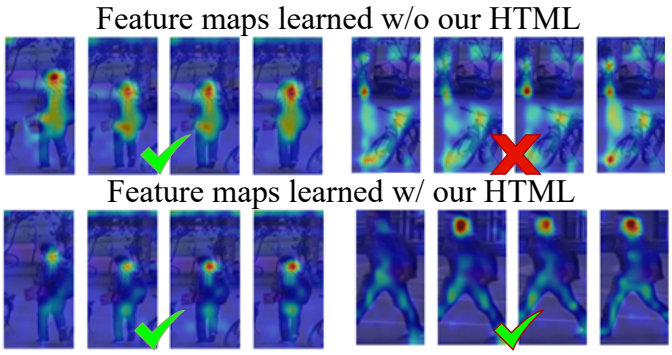

Figure 10. Features learned with and without HTML. Best viewed in color.

estimators. **i)** Methods with pre-trained key-point estimators may suffer from a substantial *domain gap* due to varying real-world conditions (e.g., differences in camera/image resolutions (Human Parsing: 640 × 480 in COCO [23] *vs*. VReID: 256 × 128 / 128 × 64 in MARS [29] / PRID [32])) and scene backgrounds (COCO contains richer background diversity, while VReID datasets often involve road or campus scenes), as illustrated in Figure 8. **ii)** Methods such as CTL [3] highlight the entire human figure in each frame (see Figure 9). However, considering all parts for VReID may degrade performance when different subjects wear similar clothing. In contrast, the HTML module encourages the network to focus on body parts that are more discriminative for VReID (see Figure 10). **iii)** The HTML module's performance is further improved by steering the attention maps toward human-related patches (as shown in Figure 11).

Additionally, we provide both successful and failure cases in Figure 12. **i)** Although the part estimation is not perfect due to the absence of supervision, imperfect body shapes are sufficient for alignment in downstream tasks such as VReID (as shown in Table 1). **ii)** The HTML module adapts

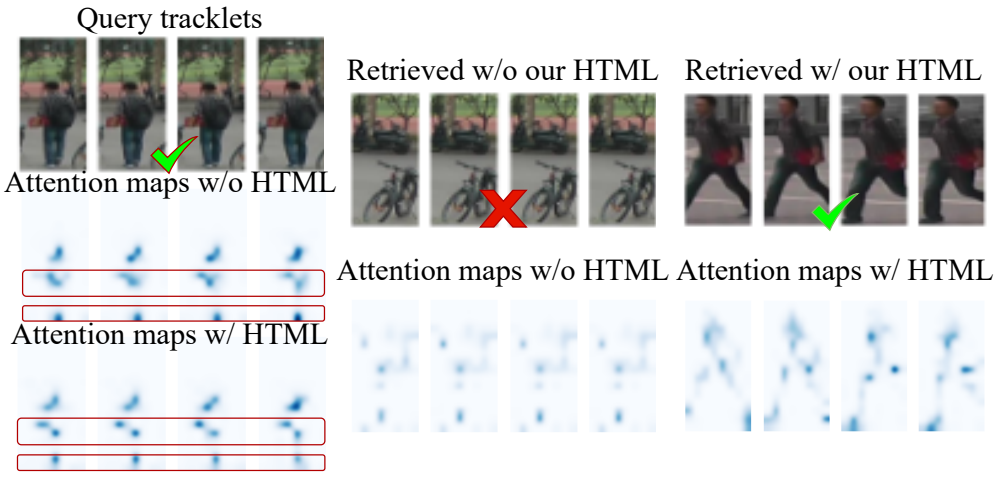

Figure 11. Retrieval results and attentions with and without HTML. Without HTML, some attention responses are scattered or incorrectly focused on the ground. Best viewed in color.

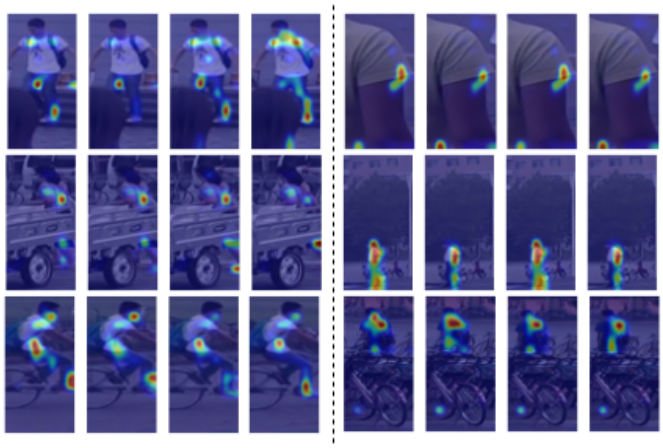

Figure 12. Success (left) and failure (right) cases of the HTML module using the same topological prior. Best viewed in color.

well under the same topological prior, even in the presence of occlusions or when the subject is riding a bike. **iii)** The HTML module may fail when improper crops occur due to the lack of supervision, or when the subject is too small for reliable parsing. In addition, it may mistakenly identify a person as riding a bike when the person is simply standing behind it.

5. Conclusions

In this paper, we proposed a novel HTML module for VReID, and demonstrated that extra pre-trained key-point estimators and additional body-part labels are not necessary to model human structures. By leveraging a minimal humanoid topological prior, it models body shapes from video as a side task, thus resulting in aligned and discriminative body-part features. To compensate for the lack of supervision, a regularization loss was designed to encourage HTML to model body shapes with

maximal similarity across frames. Moreover, by introducing two ST mixers, we show that HTML is compatible with other ST integration approaches, the further enhancing the reliability and discriminative power of the learned representations for VReID. The experiments demonstrated that our method delivers competitive performances across multiple datasets while retaining a lightweight backbone design.

Use of AI tools declaration

The authors declare they have not used Artificial Intelligence (AI) tools in the creation of this article.

Acknowledgments

This work was supported by the NSF China (No. 62276004).

Conflict of interest

The authors declare that there is no conflict of interest.

References

1. Ma H, Zhang C, Zhang Y, Li Z, Wang Z, Wei C, (2024) A review on video person re-identification based on deep learning. *Neurocomputing* 609: 128479. https://doi.org/10.1016/j.neucom.2024.128479
2. Saad RSM, Moussa MM, Abdel-Kader NS, Farouk H, Mashaly S, (2024) Deep video-based person re-identification (deep vid-reid): Comprehensive survey. *EURASIP J Adv Signal Process* 2024: 63. https://doi.org/10.1186/s13634-024-01139-x
3. Liu J, Zha Z, Wu W, Zheng K, Sun Q, (2021) Spatial-temporal correlation and topology learning for person re-identification in videos, In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021: 4370–4379. https://doi.org/10.1109/CVPR46437.2021.00435
4. Jiang X, Qiao Y, Yan J, Li Q, Zheng W, Chen D, (2021) SSN3D: Self-separated network to align parts for 3d convolution in video person re-identification, In: *Proceedings of the AAAI Conference on Artificial Intelligence*, 35: 1691–1699. https://doi.org/10.1609/aaai.v35i2.16262
5. Wu J, He L, Liu W, Yang Y, Lei Z, Mei T, et al. (2022) CAViT: Contextual alignment vision transformer for video object re-identification, In: *European Conference on Computer Vision*, 2022: 549–566. https://doi.org/10.1007/978-3-031-19781-9_32
6. Wu P, Wang L, Zhou S, Hua G, Sun C, (2024) Temporal correlation vision transformer for video person re-identification, In: *Proceedings of the AAAI Conference on Artificial Intelligence*, 38: 6083–6091. https://doi.org/10.1609/aaai.v38i6.28424
7. Yu C, Liu X, Wang Y, Zhang P, Lu H, (2024) Tf-clip: Learning text-free clip for video-based person re-identification, In: *Proceedings of the AAAI Conference on Artificial Intelligence*, 38: 6764–6772. https://doi.org/10.1609/aaai.v38i7.28500

8. Su Y, Shi Y, Liu F, Liu X, (2025) Hamobe: Hierarchical and adaptive mixture of biometric experts for video-based person reid, In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2025: 11525–11536.
9. Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D, (2017) Grad-cam: Visual explanations from deep networks via gradient-based localization, In: *Proceedings of the IEEE International Conference on Computer Vision*, 2017: 618–626. https://doi.org/10.1109/ICCV.2017.74
10. Jiang X, Gong Y, Guo X, Yang Q, Huang F, Zheng W, et al., (2020) Rethinking temporal fusion for video-based person re-identification on semantic and time aspect, In: *Proceedings of the AAAI Conference on Artificial Intelligence*, 34: 11133–11140. https://doi.org/10.1609/aaai.v34i07.6770
11. Aich A, Zheng M, Karanam S, Chen T, Roy-Chowdhury AK, Wu Z, (2021) Spatio-temporal representation factorization for video-based person re-identification, In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021: 152–162. https://doi.org/10.1109/ICCV48922.2021.00022
12. Liu X, Zhang P, Yu C, Lu H, Yang X, (2021) Watching you: Global-guided reciprocal learning for video-based person re-identification, In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021: 13334–13343. https://doi.org/10.1109/CVPR46437.2021.01306
13. Hou R, Chang H, Ma B, Shan S, Chen X, (2020) Temporal complementary learning for video person re-identification, In: *European Conference on Computer Vision*, 2020: 388–405. https://doi.org/10.1007/978-3-030-58545-7_23
14. Hou R, Chang H, Ma B, Huang R, Shan S, (2021) Bicnet-tks: Learning efficient spatial-temporal representation for video person re-identification, In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021: 2014–2023. https://doi.org/10.1109/CVPR46437.2021.00204
15. Shu X, Li G, Wei L, Zhong J, Zang X, Zhang S, et al., (2021) Diverse part attentive network for video-based person re-identification, *Pattern Recognition Letters*, 149: 1–8. https://doi.org/10.1016/j.patrec.2021.07.015
16. He T, Jin X, Shen X, Huang J, Chen Z, Hua X, (2021) Dense interaction learning for video-based person re-identification, In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021: 1490–1501. https://doi.org/10.1109/ICCV48922.2021.00153
17. Yang J, Zheng W, Yang Q, Chen Y, Tian Q, (2020) Spatial-temporal graph convolutional network for video-based person re-identification, In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020: 3289–3299. https://doi.org/10.1109/CVPR42600.2020.00332
18. Li S, Yu H, Hu H, (2020) Appearance and motion enhancement for video-based person re-identification, In: *Proceedings of the AAAI Conference on Artificial Intelligence*, 34: 11394–11401. https://doi.org/10.1609/aaai.v34i07.6890
19. Subramaniam A, Nambiar A, Mittal A, (2019) Co-segmentation inspired attention networks for video-based person re-identification, In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019: 562–572. https://doi.org/10.1109/ICCV.2019.00061

20. He K, Zhang X, Ren S, Sun J, (2016) Deep residual learning for image recognition, In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2016: 770–778. https://doi.org/10.1109/CVPR.2016.90

21. Han F, Reily B, Hoff W, Zhang H, (2017) Space-time representation of people based on 3d skeletal data: A review, *Comput Vision Image Understanding*, 158: 85–105. https://doi.org/10.1016/j.cviu.2016.12.003

22. Cao Z, Hidalgo G, Simon T, Wei S, Sheikh Y, (2019) Openpose: Realtime multi-person 2d pose estimation using part affinity fields, *IEEE Trans Pattern Anal Mach Intell*, 43: 172–186. https://doi.org/10.1109/TPAMI.2019.2929257

23. Lin T, Maire M, Belongie S, Hays J, Perona P, Ramanan D, et al. (2014) Microsoft coco: common objects in context, In: *European Conference on Computer Vision*, 2014: 740–755. https://doi.org/10.1007/978-3-319-10602-1_48

24. Kipf TN, Welling M, (2017) Semi-supervised classification with graph convolutional networks, In: *Proceedings of the International Conference on Learning Representations*, 2017: 1–14.

25. Ba JL, Kiros JR, Hinton GE, (2016) Layer normalization, In: *Proceedings of the Neural Information Processing Systems*, 2016: 1–14.

26. Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, et al. (2021) An image is worth 16x16 words: Transformers for image recognition at scale, In: *Proceedings of the International Conference on Learning Representations*, 2021: 1–22.

27. Ye M, Shen J, Lin G, Xiang T, Shao L, Hoi SC, (2021) Deep learning for person re-identification: a survey and outlook, *IEEE Trans Pattern Anal Mach Intell*, 44: 2872–2893. https://doi.org/10.1109/TPAMI.2020.3039709

28. Hermans A, Beyer L, Leibe B, (2019) In defense of the triplet loss for person re-identification, In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2019: 1–10. https://doi.org/10.1109/CVPRW.2017.137

29. Zheng L, Bie Z, Sun Y, Wang J, Su C, Wang S, et al. (2016) Mars: A video benchmark for large-scale person re-identification, In: *European Conference on Computer Vision*, 2016: 868–884. https://doi.org/10.1007/978-3-319-46475-6_53

30. Li J, Wang J, Tian Q, Gao W, Zhang S, (2019) Global-local temporal representations for video person re-identification, In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019: 3958–3967. https://doi.org/10.1109/ICCV.2019.00404

31. Wang T, Gong S, Zhu X, Wang S, (2014) Person re-identification by video ranking, In: *European Conference on Computer Vision*, 2014: 688–703. https://doi.org/10.1007/978-3-319-10599-4_44

32. Hirzer M, Beleznai C, Roth PM, Bischof H, (2011) Person re-identification by descriptive and discriminative classification, In: *Proceedings of the Scandinavian Conference on Image Analysis*, 2011: 91–102. https://doi.org/10.1007/978-3-642-21227-7_9

33. Wang Y, Zhang P, Gao S, Geng X, Lu H, Wang D, (2021) Pyramid spatial-temporal aggregation for video-based person re-identification, In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021: 12026–12035. https://doi.org/10.1109/ICCV48922.2021.01183

34. Eom C, Lee G, Lee J, Ham B, (2021) Video-based person re-identification with spatial and temporal memory networks, In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021: 12036–12045. https://doi.org/10.1109/ICCV48922.2021.01184

35. Bai S, Ma B, Chang H, Huang R, Chen X, (2022) Salient-to-broad transition for video person re-identification, In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022: 7339–7348. https://doi.org/10.1109/CVPR52688.2022.00722

36. Liu C, Chen J, Chen C, Chien S, (2021) Video-based person re-identification without bells and whistles, In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2021: 1491–1500. https://doi.org/10.1109/CVPRW53098.2021.00156

37. Chen D, Zhang Y, Yuan J, Gao S, Bai X, (2022) Keypoint message passing for video-based person re-identification, In: *Proceedings of the AAAI Conference on Artificial Intelligence*, 36: 239–247. https://doi.org/10.1609/aaai.v36i1.19771

38. Yao Y, Jiang X, Fujita H, Fang Z, (2022) A sparse graph wavelet convolution neural network for video-based person re-identification, *Pattern Recognit*, 129: 108708. https://doi.org/10.1016/j.patcog.2022.108708

39. Hu X, Wei D, Wang Z, Shen J, Ren H, (2021) Hypergraph video pedestrian re-identification based on posture structure relationship and action constraints, *Pattern Recognit*, 111: 107688. https://doi.org/10.1016/j.patcog.2020.107688

40. Zheng L, Shen L, Tian L, Wang S, Wang J, Tian Q, (2015) Scalable person re-identification: A benchmark, In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2015: 1116–1124. https://doi.org/10.1109/ICCV.2015.129

41. Bolle RM, Connell JH, Pankanti S, Ratha NK, Senior AW, (2005) The relation between the roc curve and the cmc, In: *Fourth IEEE Workshop on Automatic Identification Advanced Technologies (AutoID'05)*, 2005: 15–20. https://doi.org/10.1109/AUTOID.2005.1529697

42. Zhong Z, Zheng L, Kang G, Li S, Yang Y, (2020) Random erasing data augmentation, In: *Proceedings of the AAAI Conference on Artificial Intelligence*, 34: 13001–13008. https://doi.org/10.1609/aaai.v34i07.6999

43. Kingma DP, Ba J, (2015) Adam: A method for stochastic optimization, In: *Proceedings of the International Conference on Learning Representations*, 2015: 1–15.

44. Luo H, Gu Y, Liao X, Lai S, Jiang W, (2019) Bag of tricks and a strong baseline for deep person re-identification, In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2019: 4321–4329. https://doi.org/10.1109/CVPRW.2019.00463

45. Chen X, Xie S, He K, (2021) An empirical study of training self-supervised vision transformers, In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021: 9640–9649. https://doi.org/10.1109/ICCV48922.2021.00957

46. Bai J, Yuan L, Xia S, Yan S, Li Z, Liu W, (2022) Improving vision transformers by revisiting high-frequency components, In: *European Conference on Computer Vision*, 2022: 1–18. https://doi.org/10.1007/978-3-031-19800-7_1

47. Srivastava N, Hinton G, Krizhevsky A, Sutskever I, Salakhutdinov R, (2014) Dropout: A simple way to prevent neural networks from overfitting, *J Mach Learn Res*, 15: 1929–1958. https://doi.org/10.1145/2627435.2670313

48. Ioffe S, Szegedy C, (2015) Batch normalization: Accelerating deep network training by reducing internal covariate shift, In: *Proceedings of the International Conference on Machine Learning*, 37: 448–456.

AIMS Press