

# ShARc: Shape and Appearance Recognition for Person Identification In-the-wild

Haidong Zhu    Wanrong Zheng    Zhaoheng Zheng    Ram Nevatia  
University of Southern California

{haidongz, wanrongz, zhaoheng.zheng, nevatia}@usc.edu


## Abstract

Identifying individuals in unconstrained video settings is a valuable yet challenging task in biometric analysis due to variations in appearances, environments, degradations, and occlusions. In this paper, we present ShARc, a multimodal approach for video-based person identification in uncontrolled environments that emphasizes 3-D body shape, pose, and appearance. We introduce two encoders: a Pose and Shape Encoder (PSE) and an Aggregated Appearance Encoder (AAE). PSE encodes the body shape via binarized silhouettes, skeleton motions, and 3-D body shape, while AAE provides two levels of temporal appearance feature aggregation: attention-based feature aggregation and averaging aggregation. For attention-based feature aggregation, we employ spatial and temporal attention to focus on key areas for person distinction. For averaging aggregation, we introduce a novel flattening layer after averaging to extract more distinguishable information and reduce overfitting of attention. We utilize centroid feature averaging for gallery registration. We demonstrate significant improvements over existing state-of-the-art methods on public datasets, including CCVID, MEVID, and BRIAR.

## 1. Introduction

Recognizing individuals in-the-wild [43] is a challenging yet valuable task for determining a person’s identity from images or videos, playing a crucial role in many applications. Since face images may be unreliable or unavailable for individuals at a distance or from specific viewpoints, recognizing individuals via body images or videos becomes increasingly important. In this paper, we focus on video-level appearance and body shapes to develop a robust

This research is based upon work supported in part by the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA), via [2022-21102100007]. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of ODNI, IARPA, or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for governmental purposes notwithstanding any copyright annotation therein.



| Gallery Frame | Standing Videos | Different Clothing | Turbulence & Occlusion |
|---------------|-----------------|--------------------|------------------------|
| Gait          |                 | ✓                  | ✗                      |
| Body shape    | ✓               | ✗                  | ✓                      |
| Appearance    | ✓               | ✗                  | ✗                      |
| Ours          | ✓               | ✓                  | ✓                      |

Figure 1. To identify a person, gait is unreliable in stationary videos, and appearance alters when subjects wear different clothing. The imprecise reconstruction of 3-D body shapes results in unstable predictions while the human prior assists in occlusions.

identification system suitable for various distances and camera viewpoints, utilizing multiple videos as gallery samples. We specifically address different clothing and activities in generalized scenarios by comparing and combining shape and appearance-based methods for identification.

To identify individuals from their body, research primarily focuses on appearance [58, 60] and gait [2, 12, 35, 66, 72, 73]. Unlike facial features, which are relatively constant [8, 9, 28], body appearance can vary significantly due to changes in clothing, environment, and occlusions [7], as depicted in Figure 1. Gait analysis captures an individual’s walking pattern and is less affected by environmental changes or clothing. However, it requires a walking sequence that may not always be available. Additionally, varying environmental conditions pose challenges in feature registration and matching, making the prediction of human identity more sensitive to noisy samples in gallery videos.

We introduce ShARc, a method based on SHape and Appearance ReCognition. Specifically, we employ a Pose and Shape Encoder (PSE) and an Aggregated Appearance Encoder (AAE) to project the input video into their cor-

responding embedding spaces. Leveraging body shapes with shape and motion representations [73], ShARc enables identification in diverse scenarios; a robust body prior [40] offers guidance under occlusion or variations in clothing. Alongside this, we introduce multi-level appearance features for both video-level and frame-level analysis. Importantly, these techniques show commendable performance even before combining with body shapes.

To extract the shape of a person in a sequence, we disentangle motion and poses by extracting skeletons, 3-D body shapes, and silhouettes from tracklets with our Pose and Shape Encoder (PSE). We utilize silhouettes and 3-D body shapes to represent individual frame shape patterns in 2-D and 3-D space, while employing sequential skeletons to represent motions. For the two different shape modalities, we first extract their frame-wise features and then combine them frame-by-frame using an attention mechanism for body shape feature extraction. Subsequently, we concatenate the pooled features with pose features encoded from skeletons for the final shape representation.

Parallel to body shape extraction, we also use an Aggregated Appearance Encoder (AAE) to extract features from appearances, preserving identification information from raw images. We obtain both frame-wise and video-level features and integrate them for dual-level understanding. For frame-level extraction, we introduce a novel flattening layer after averaging to extract more distinguishable information and reduce overfitting. At the video level, we employ spatial and temporal attention, as per [58], to focus on key areas for person distinction. This allows the model to concentrate on unique patterns in both frame and sequence.

After obtaining both shape and appearance features, we employ centroid feature averaging for gallery registration, using the mean features of the same ID rather than comparing to each gallery separately. This helps to mitigate variances in gallery examples with different clothing. We validate our approach on public datasets like CCVID [16], MEVID [7], and the recently-released BRIAR [5], showing state-of-the-art performance on all of them.

In summary, our contributions are as follows: 1) We introduce ShARc, a multimodal method for person identification in-the-wild using video samples, focusing on both shape and appearance; 2) We unveil a novel Pose and Shape Encoder (PSE) that captures dynamic motion and body shape features for more robust shape-based identification; 3) We deploy an Aggregated Appearance Encoder (AAE) that incorporates both frame-level and video-level features.

## 2. Related Work

**Person Identification Based on Body Appearance** is a critical task in computer vision that focuses on identifying and matching individuals across different camera views or separate instances [31, 68]. Unlike face recog-

nition [8, 9, 28], body appearance-based re-identification [21, 27, 60, 69] requires less subject cooperation and is achievable in diverse environments. With deep learning advancements, researchers focus on various ways to extract maximally useful information from single-frame inputs. Approaches include part-based methods [4, 20, 33, 54, 65, 74] and attention [15, 26] to address occlusions and others.

Besides single-frame person identification, recent research has explored video-level re-identification methods [16, 23, 58] by introducing more frames and reducing poor-quality frame impact for enhanced temporal robustness. Since the model only needs to output one person ID prediction for multiple frames, researchers either use temporal pooling [14, 30, 38, 67] or recurrent networks [6, 41, 70] for fusing frames across timestamps. Recently, attention mechanisms [13, 23, 29, 37, 49, 51, 58, 62] have been utilized for aggregating useful information from temporal and spatial dimensions for identification. However, most methods focus on videos with consistent clothing, limiting model generalizability. Researchers are now emphasizing videos with different outfits and environments [7, 16], making identification tasks more applicable for real-life scenarios.

**Gait Recognition** focuses on identifying a person based on their walking patterns. Compared with appearance-based recognition methods, gait patterns, usually captured via binarized silhouettes [55, 63] describing body shape contours, reduce the negative impact of clothing changes for identification but introduce different appearance variations with body contours. Due to the lack of RGB patterns, it is challenging to infer body information directly from silhouettes. To address this, some researchers [12, 35] focus on part-based recognition, while others [2, 11, 24] extract framewise consistencies for identification.

Due to the limited information in silhouettes, recent research [1, 18, 48, 56, 73] focuses on external modalities to assist silhouettes for identification. GaitGraph [56], GaitMix [72] and GaitRef [72] apply or refine HRNet [57] for joint detection and uses the generated pose sequence for identification. Gait3D [66], GaitHBS [73], and ModelGait [32] focus on extracting or using body shapes alongside silhouettes for gait recognition, intending to provide more information for part separation. LiDARGait [48] employs point clouds instead of silhouettes for body shape description. Some researchers [18, 34] also integrate RGB images with silhouettes for gait understanding. Since these methods still focus on gait representation, they can only apply to walking sequences for identification. Our proposed PSE combines pose with 3-D body shape for identification, inherently removing the requirement for walking sequences.

## 3. Methodology

Given a video with sequential frames  $V = \{f_i\}_n$  containing  $n$  frames of the person, ShARc decomposes it into

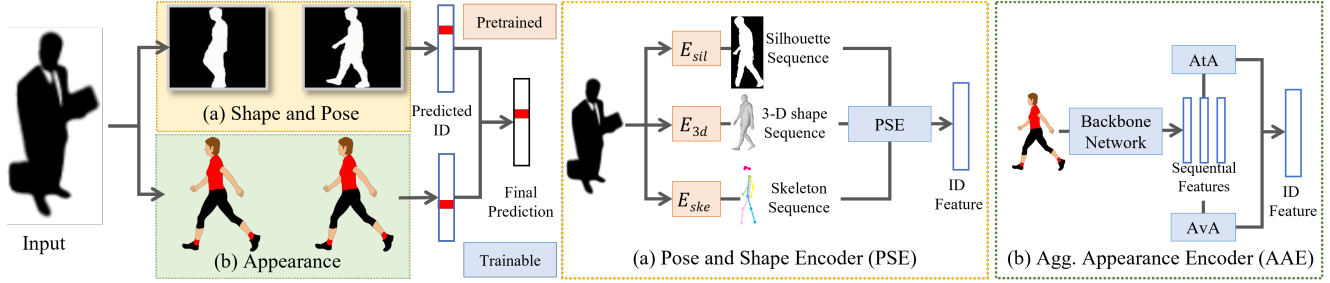


Figure 2. Our proposed method includes two sub modules: (a) a shape-based recognition system, PSE, which extracts the silhouette, 3-D body shape and skeletons sequences and fuses them for person recognition, and (b) an appearance-based recognition system, AAE, which takes both outputs from attention-based aggregation (AgA) and averaging aggregation (AvA) as input for identification.

two branches: the body shape  $\{b_i\}$  and the RGB appearance of the frames  $\{a_i\}$  that exhibit the most distinguishable patterns, as illustrated in Figure 2. By estimating their independent similarities  $S_{shape}(V)$  and  $S_{app}(V)$  compared with gallery candidates, ShARc combines the two scores together using weighted average for the final similarity  $S(V)$ .

### 3.1. Shape-based Person Recognition

For shape-based person recognition, we mitigate the influence of appearance by focusing on alternative representations, such as 3-D human body shape and silhouettes, to emphasize the individual’s body shape, as well as skeletons to capture motion in pose. Although gait recognition is useful when walking segments are available, it offers limited distinguishable information in stationary videos when the person is not walking. Unlike existing gait recognition methods [2, 35, 56, 73], our shape-based approach compensates for the absence of gait by leveraging extra body shape priors. We first extract the corresponding modalities utilized in our model, which include 3-D body shapes, skeletons, and silhouettes, and then fuse them as the final representation.

**Shape and motion extraction.** For shape-based person recognition, we focus on two crucial representations for distinguishing individuals: body shape  $P_i$  and motion  $M_i$ . Body shape encompasses specific actions or shapes a person may exhibit, while motion refers to the temporal information, representing a more specific case. If both shape and motion exist in all sequences, the task can be regarded as gait recognition. For body shape extraction, we focus on two distinct modalities: silhouettes and 3-D body shapes. Silhouettes represent the 2-D human boundary in each frame, while 3-D body shape reconstruction remains invariant to viewpoints by reconstructing the person’s 3-D shape. The combination of silhouettes and body shapes allows for the preservation of both general shape and frame-wise detailed reconstruction of the individual.

In addition to body shape, we incorporate skeletons to understand motions, as motions represent the specific movement patterns of a person. Unlike gait recognition

tasks [2], which use binarized silhouettes as input, skeletons can provide temporal understanding without the biases of body shape. Furthermore, by separating body shapes from motion analysis, the network for pose extraction can better focus on the general shape, aiding temporal understanding and helping the model to maximize the utilization of potential information in the sequence.

For the three modalities described above, we employ three extractors,  $E_{sil}(\cdot)$ ,  $E_{3d}(\cdot)$  and  $E_{ske}(\cdot)$ , to encode the corresponding representations of these three modalities for each frame  $i$  following

$$P_i = E_{sil}(f_i) + E_{3d}(f_i); \quad M_i = E_{ske}(f_i) \quad (1)$$

and extract the corresponding body shape  $P_i$  and motion  $M_i$  inputs for further processing. For silhouette input, we concatenate the silhouette and the cropped RGB images using silhouette as masks, as our input, since this can provide more separation of the human part in the body shape. Since these modals requires heavy training to ensure a stable performance, we use pretrained networks to extract these representations, which we discuss in Section 4.1.

**Multimodal Fusion.** With these three modalities, we introduce PSE for combining framewise body shape features  $P_i$  along with motion pattern  $M_i$ , as illustrated in Figure 3. For feature representation of silhouettes  $Feat_{sil}$  and 3-D body shapes  $Feat_{3d}$ , we use corresponding encoders  $F_{pose}$  to project  $E_{sil}(f_i)$  and  $E_{3d}(f_i)$  into their embedding space. We then apply the 3-D spatial transformation network [66] with skip connection and implement horizontal pyramid pooling  $HPP$  [2] with  $B$  bins after the encoder output for each frame following

$$\begin{aligned} I_{sil}, I_{3d} &= F_{pose}(E_{sil}(f_i), E_{3d}(f_i)) \\ I_{pose} &= (I_{sil} \cdot I_{3d}) + I_{sil} \\ I_{pose} &= HPP(I_{pose}) \end{aligned} \quad (2)$$

where  $I_k$  represents the feature for the modality  $k$ . For motion representation, we utilize a motion encoder  $F_{motion}$  to extract multi-level spatial and temporal skeleton information and use average pooling along the temporal dimension

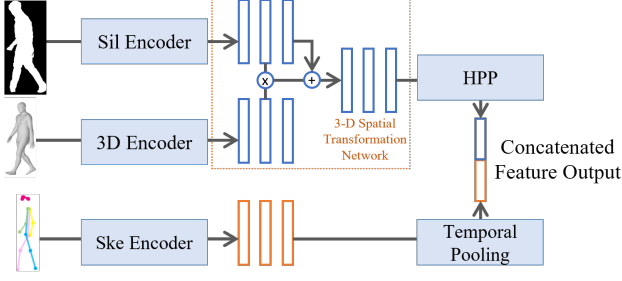


Figure 3. Architecture of PSE for combining body shape and motion information for shape-based identification.

for the generated feature of the last layer. Then, we concatenate the skeleton feature, after temporal pooling, along with the pose representation as an additional new bin in the matching process, making the concatenated  $(B + 1) \times C$  feature map our final output for shape representation:

$$\begin{aligned} I_{motion} &= AvgPooling(F_{motion}(E_{ske}(f_i))) \\ I_{shape} &= [I_{pose}, I_{motion}] \end{aligned} \quad (3)$$

where  $[\cdot, \cdot]$  represents feature concatenation.

### 3.2. Appearance-based Person Recognition

Compared to shape-based methods, which depend on the accuracy of body shape and contours, appearance provides richer and lossless RGB information for distinguishing individuals. We implement both attention-based and averaging appearance aggregation for identification. As people may wear different clothing and be in varying environmental conditions, we incorporate temporal and spatial information with attention-based appearance aggregation to focus on the relevant parts for differentiation between nearby frames. Moreover, to avoid overfitting on specific body parts or frames, we also employ video-level averaging aggregation to equally utilize spatial and temporal features.

**Attention-based Aggregation.** For attention-based aggregation, we follow Figure 4 (a) for building spatial and temporal attention (STA) for the features extracted from the backbone network, encoding each frame  $F_i$  to their corresponding features  $A_i$ . We follow [58] to combine the features of two frames using a 3-level pyramid following

$$A_t^{l+1} = SA(A_t^l) + SA(A_{t+1}^l) + TA(A_t^l, A_{t+1}^l) \quad (4)$$

where  $l$  is the current layer in the pyramid, and  $t$  is the temporal stamp for the current frame.  $TA$  and  $SA$  are two attention generation layers following [58]. For each layer of the pyramid, we reduce the number of available appearance features to half the size of its previous layer, until we get the output feature representation in the last layer. This means the network, as an example, can handle at most 8 frames for the final feature  $A_{attn}(V)$  with a three layers of pyramid.

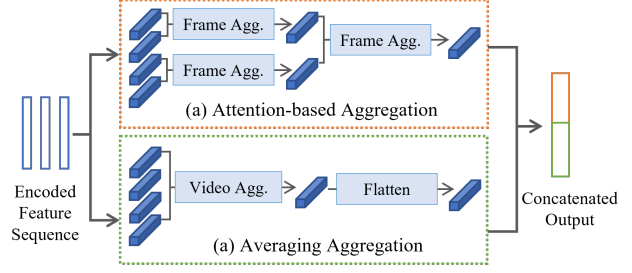


Figure 4. Architecture of the AAE with an example of sequence length  $n = 4$ . AAE aggregate the video frames in two ways: 1) attention-based aggregation, which mines the connection between nearby frames with attention, and 2) averaging aggregation, which takes all the frames together equally.

It is important to note that if attention-based aggregation is not combined with averaging aggregation and its backbone feature encoder not shared, it is degraded to the existing method PSTA [58] encoder.

**Averaging Aggregation.** As attention mechanism may create overfitting when there is shift between training and testing domain, we add averaging aggregation, as illustrated in Figure 4 (b), for global representation extraction. Video-level appearance focuses on finding the corresponding features of each frame and treating all the frames equally. After extracting the framewise appearance feature  $A_i$ , we average the features of all the frames in the same video following

$$A_{avg} = \frac{1}{n} \sum_{i=1}^n A_i \quad (5)$$

We then use Gamma Correction  $\gamma$  in the range of  $[0, 1]$  to flatten the features as a feature flatten layer following

$$A_{avg} = sgn(A_{avg}) \cdot ||A_{avg}||^\gamma \quad (6)$$

where  $sgn(\cdot)$  is the sign function operated on channel-wise elements. Since the videos include multiple frames that may capture the person from different aspects, some of the specific representative features of this person may not be captured in all the frames. With  $\gamma < 1$ , the new feature are different from the old one in cases. When the feature value is close to the zero point (0), flattening layer makes the original value more distinguishable by increasing its absolute value. In addition, the flattening layer can also reduce the maximum value and avoid overfitting with feature values far from 0, making the network focus on more patterns instead of on just a few of them for making predictions.

### 3.3. Registration and Fusion

For person identification in the wild, it is essential to handle videos of individuals with varying clothing conditions, as gallery videos also exhibit differences in clothing, leading to variances in appearance. To address this issue, we



follow [60] and construct a centroid representation for registering gallery examples. Assuming we have  $k \times c$  features with a same ID, we average the  $k$  features and use the  $1 \times c$  feature for representing this ID in the gallery. The averaging operation helps to mitigate the biases arising from different clothing, as clothing across videos are assumed to be randomly distributed, while the appearance remains consistent.

Since shape and appearance are distinct modalities, we compute the features independently for each and match them with their corresponding modalities in the gallery candidates to obtain two matching scores,  $S_{shape}(V)$  and  $S_{app}(V)$ . We then use a weighted average function to combine these two scores following

$$S(V) = \alpha S_{shape}(V) + (1 - \alpha) S_{app}(V), \quad (7)$$

where  $S(V)$  is the final similarity score,  $S_{shape}(V)$  and  $S_{app}(V)$  are the shape-based and appearance-based similarity scores, respectively, and  $\alpha$  is a weight parameter that balances the contributions of the two modalities. By adjusting  $\alpha$ , we can find the optimal combination that leads to the best overall identification performance. Based on our ablation results in Section 4.2, we set  $\alpha$  to 0.1 in our experiment.

### 3.4. Objectives

Considering that ShARc is a model for video-based identification, we train shape and appearance models separately, using end-to-end training for each. For the shape-based recognition model, PSE, we follow [66] and combine triplet loss  $\mathcal{L}_{triplet}$  [46] with a margin of 0.2, along with cross-entropy loss  $\mathcal{L}_{CE}$  as follows:

$$\mathcal{L}_{shape} = 0.1 \mathcal{L}_{triplet} + \mathcal{L}_{CE} \quad (8)$$

For the appearance model, we apply four losses following [60], which combines a Triplet loss  $\mathcal{L}_{triplet}$  [46] with 0.3 as margin, a Center Loss  $\mathcal{L}_{cen}$  [59], a Cross Entropy loss  $\mathcal{L}_{CE}$ , and a Centroid Triplet Loss  $\mathcal{L}_{CTL}$  [60], as follows:

$$\mathcal{L}_{app} = \mathcal{L}_{triplet} + \mathcal{L}_{CE} + \mathcal{L}_{cen} + 5e^{-4} \mathcal{L}_{CTL} \quad (9)$$

## 4. Experiments and Results

### 4.1. Experimental Details

**Datasets.** In our experiment, we primarily compare our method with other state-of-the-art methods on three challenging, public, video-based datasets: CCVID [16], MEVID [7], and BRIAR [5]. We include the statistics for these three datasets in Table 1. CCVID [16] and MEVID [7] are recent datasets featuring the same and different clothes and include more than one outfit for each identity, with 226 and 158 identities, respectively. Unlike CCVID, which has only one viewpoint from the same location, MEVID includes 33 viewpoints and multiple scales of images from

| Dataset | Split   | #frames   | #identities | #tracklets |
|---------|---------|-----------|-------------|------------|
| BRIAR   | train   | 4,366,198 | 407         | 37,466     |
|         | query   | 189,819   | 192         | 886        |
|         | gallery | 2,326,111 | 544         | 4,379      |
| CCVID   | train   | 116,799   | 75          | 948        |
|         | query   | 118,613   | 151         | 834        |
|         | gallery | 112,421   | 151         | 1,074      |
| MEVID   | train   | 3,609,156 | 104         | 6,338      |
|         | query   | 205,044   | 52          | 316        |
|         | gallery | 981,207   | 54          | 1,438      |

Table 1. Statistics for the three datasets in our experiment.

33 different settings. BRIAR is a large, in-the-wild person identification dataset with varying distances, conditions, activities, and outfits for identification.

Compared to CCVID and MEVID, BRIAR [5] encompasses more variations of distances, viewpoints, and candidate IDs, which models the person identification problem in the wild. In addition, BRIAR has more distractor IDs in the gallery for the open-set problem evaluation, as well as featuring more images from elevated cameras and UAVs, introducing greater difficulty for final template matching. Since the BRIAR dataset is continuously expanding, we use the version including both BGC1 and 2 following [5], which is an extended version compared to [18].

**Implementation Details.** We first discuss the detailed architecture used for shape and appearance-based networks, followed by the training and inference details.

*Shape-based Modalities Extraction.* For shape-based recognition, our model requires three different inputs: silhouettes, 3-D body shapes, and skeletons. For silhouette extraction  $E_{sil}(\cdot)$ , we use DeepLab-v3 [3] with ResNet-101 [19] pretrained on the Pascal VOC dataset as the backbone to identify the pixels predicted as the ‘person’ category for silhouettes. For the 3-D human body shape extraction  $E_{3d}(\cdot)$ , we use ROMP [53] pretrained on Human3.6M [25] and MPI-INF-3DHP [42] to extract three vectors: a 3-D camera parameter, a 10-D vector body shape, and a 72-D vector representing the rotation of the joints. These three vectors form an 85-D SMPL [40,71] representation for each frame. Since there is only one person in each frame sequence, we use the first SMPL body shape predicted by ROMP as our body shape representation. For skeletons  $E_{ske}(\cdot)$ , we follow [56] and use HRNet [52] with architecture ‘pose\_hrnet\_w32’ and  $384 \times 288$  as input size, which is pretrained on the MS COCO dataset [36] for 2-D pose estimation as the skeleton representation.

With different input modalities available for shape-based modal extraction, we use ResNet-9 [11] as the gait encoder, a 4-layer MLP [66] for 3-D body shape encoding, and MS-G3D [39] for skeleton encoding. All these three models are trained together with PSE end-to-end with the shape-based recognition model.

| Method              | All Activities |             | Walking Sequences |             | Stationary Sequences |             |
|---------------------|----------------|-------------|-------------------|-------------|----------------------|-------------|
|                     | Rank 1         | Rank 20     | Rank 1            | Rank 20     | Rank 1               | Rank 20     |
| GaitSet [2]         | 15.3           | 40.5        | 27.7              | 64.5        | 7.3                  | 24.9        |
| GaitPart [12]       | 14.1           | 41.7        | 25.7              | 67.8        | 6.6                  | 24.8        |
| GaitGL [35]         | 15.6           | 45.1        | 28.0              | 67.2        | 7.5                  | 30.8        |
| GaitMix [72]        | 15.9           | 46.5        | 27.6              | 65.3        | 8.1                  | 33.9        |
| GaitRef [72]        | 17.7           | 50.2        | 29.9              | 69.4        | 9.5                  | 37.2        |
| SMPLGait [66]       | 18.8           | 51.9        | 25.2              | 63.4        | 14.6                 | 44.3        |
| PSE (Ours)          | 21.2           | 65.3        | 23.2              | 68.6        | 19.9                 | 63.2        |
| DME [18]            | 25.0           | 63.8        | 30.4              | 68.8        | 21.5                 | 60.5        |
| PSTA [58]           | 33.6           | 67.3        | 32.1              | 66.0        | 34.5                 | 68.1        |
| CAL [16]            | 34.9           | 71.4        | 34.7              | 71.0        | 35.0                 | 71.7        |
| TCL Net [23]        | 31.3           | 65.6        | 31.0              | 65.1        | 31.5                 | 65.9        |
| Attn-CL+rerank [44] | 27.6           | 61.8        | 26.9              | 60.5        | 28.1                 | 62.6        |
| AAE (Ours)          | 38.3           | 81.8        | 37.6              | 79.0        | 39.5                 | 83.7        |
| ShARc               | <b>41.1</b>    | <b>83.0</b> | <b>39.4</b>       | <b>80.7</b> | <b>42.2</b>          | <b>84.5</b> |

Table 2. Identification results on BRIAR dataset.

*Appearance-based Recognition Model.* For input frames  $f_i$ , we first employ a ResNet-50 [19] network which is pre-trained on ImageNet [45] dataset for feature encoding to get their  $H \times W \times C$  feature maps  $A_i$  before spatial pooling. For AAE, we follow [58] and use the patch level encoding for building a three-layer pyramid architecture with two different levels of attentions: temporal attention (TA) between two consecutive frames, and spatial attention (SA) of each frame. TA and SA of the same layer of the pyramid share weight, while those from different layers do not. The output attention is the same size as the input feature  $A_i$ , so we apply point-wise production for each input attention-feature pair and sum them up as the output, which is the input for the next level of the pyramid. For averaging aggregation, we set  $\gamma$  as 0, which degrades the function to a binarized representation, following our results for ablation studies in Sec. 4.1. After having the two features from AAE, we concatenate them to represent the appearance of the person.

*Training and Inference.* Due to the network’s complexity, we do not combine shape and appearance during training but train them individually end-to-end with their own inputs. For the shape-based network, we use the Adam optimizer for 180,000 iterations and set the initial learning rate as  $1e^{-3}$ . The learning rate is decayed to  $\frac{1}{10}$  three times at iterations 30,000, 90,000, and 150,000. For the appearance-based method, we follow [58] and train the network for 500 epochs, using the Adam optimizer with an initial learning rate of  $3.5e^{-4}$ . We decay the learning rate by 0.3 at steps 70, 140, 210, 310, and 410 during training.

During inference, we follow [60] by using centroid representation when registering the features of gallery examples via averaging all the features with the same ID. If there are multiple single frames, as gallery examples in BRIAR, we first combine the frames for the same ID as a ‘pseudo video’ before sending it into the network for feature extraction. When querying an example with the gallery, we use the cosine distance to find the highest score in the gallery for shape score  $S_{shape}(V)$  and Euclidean distance for appearance score  $S_{app}(V)$  following existing gait recognition works [2]. If videos are shorter than 8 frames, we resample

| Methods             | mAP         | Rank-1      | Rank-5      | Rank-10     | Rank-20     |
|---------------------|-------------|-------------|-------------|-------------|-------------|
| BiCnet-TKS [22]     | 6.3         | 19.0        | 35.1        | 40.5        | 52.9        |
| PiT [64]            | 13.6        | 34.2        | 55.4        | 63.3        | 70.6        |
| STMN [10]           | 11.3        | 31.0        | 54.4        | 65.5        | 72.5        |
| AP3D [17]           | 15.9        | 39.6        | 56.0        | 63.3        | 76.3        |
| TCLNet [23]         | 23.0        | 48.1        | 60.1        | 69.0        | 76.3        |
| PSTA [58]           | 21.2        | 46.2        | 60.8        | 69.6        | 77.8        |
| AGRL [61]           | 19.1        | 48.4        | 62.7        | 70.6        | 77.9        |
| Attn-CL [44]        | 18.6        | 42.1        | 56.0        | 63.6        | 73.1        |
| Attn-CL+rerank [44] | 25.9        | 46.5        | 59.8        | 64.6        | 71.8        |
| CAL [16]            | 27.1        | 52.5        | 66.5        | 73.7        | 80.7        |
| PSE                 | 10.6        | 25.9        | 39.9        | 48.7        | 62.7        |
| AAE                 | <b>29.6</b> | <b>59.2</b> | <b>70.3</b> | <b>77.2</b> | <b>83.2</b> |
| ShARc               | <b>29.6</b> | <b>59.5</b> | <b>70.3</b> | <b>77.2</b> | 82.9        |

Table 3. Rank accuracy and mAP on MEVID dataset. Results for existing methods are from official MEVID [7] implementation.

the frames until we have 8 frames for appearance feature extraction, and if the video is longer than 8 frames, we separate the video into several groups of 8 frames and average the results after extracting the features from all the groups.

**Baseline Methods and Metrics.** In our experiment, we compare our method with some state-of-the-art person-reID methods on different datasets. For MEVID [7], we compared with CAL [16], AGRL [61], BiCnet-TKS [22], TCLNet [23], PSTA [58], PiT [64], STMN [10], Attn-CL [44], Attn-CL+rerank [44], and AP3D [17] following the official results in MEVID [7]. For CCVID [16], we compared with CAL [16] following their original paper setting. For the comparison on BRIAR, we select some re-ID methods [16, 23, 44, 58] based on their performance on MEVID, as well as including some gait-based recognition methods [2, 12, 18, 35] for comparison. For evaluation metrics, we use rank accuracies and mAP (mean average precision) for evaluation on these datasets.

## 4.2. Results and Analysis

To compare with existing methods, we present the results for different baseline methods on the BRIAR, MEVID, and CCVID datasets in Tables 2, 3, and 4, respectively. In addition, we conduct some further ablation studies along with visualizations of the attention generated by the appearance branch for analysis of why the appearance model still works for clothes changing cases.

**Results for person identification.** As our main experiment, we have compared with all the three datasets with state-of-the-art methods in Table 2, 3 and 4 respectively. Note that all these three datasets are describing the clothes change settings in the re-ID task, which is more complex than the existing person re-ID tasks with same outfit. We have the following observations.

(i) *Identification Performance.* Our proposed method, ShARc, demonstrates significant performance improvements on all three datasets when compared to other state-of-the-art methods. For instance, on the BRIAR dataset,

| Method            | General     |             | CC          |             |
|-------------------|-------------|-------------|-------------|-------------|
|                   | Rank-1      | mAP         | Rank-1      | mAP         |
| GaitNet [50]      | 62.6        | 56.5        | 57.7        | 49.0        |
| GaitSet [2]       | 81.9        | 73.2        | 71.0        | 62.1        |
| PSE (Ours)        | 83.9        | 86.5        | 77.1        | 85.0        |
| CAL-baseline [16] | 78.3        | 75.4        | 77.3        | 73.9        |
| CAL Triplet [16]  | 81.5        | 78.1        | 81.1        | 77.0        |
| CAL [16]          | 82.6        | 81.3        | 81.7        | 79.6        |
| AAE (Ours)        | 89.7        | 89.9        | 84.6        | 84.8        |
| ShARc             | <b>89.8</b> | <b>90.2</b> | <b>84.7</b> | <b>85.2</b> |

Table 4. Rank-1 accuracy and mAP on CCVID dataset. CC includes the videos specifically for clothes changing, while general include both same and different clothing.

SHARc, after combining shape and appearance, achieves a 6.2% and 11.6% improvement in rank-1 and rank-20 accuracy, substantially outperforming other state-of-the-art methods. Moreover, on the other clothes-changing datasets, our method attains a 2.5% and 7.5% improvement in mAP and Rank-1 accuracy on MEVID [7], as well as a 4.6% and 8.0% improvement on CCVID [16]. Note that we follow [7] not using centroid averaging for gallery on MEVID. In addition, unlike BRIAR and CCVID, activities in MEVID do not include specific walking patterns, which results in a limited contribution from the PSE when combined with the appearance-based method, AAE.

Apart from the overall dataset results, we note that gait-based methods [2, 12, 18, 35] and appearance-based methods [16, 23, 44, 58] display different performance differences for the two types of activities, standing and walking. On the BRIAR dataset, gait-based methods [2, 12, 35] struggle with stationary sequences. Although DME [18]<sup>1</sup> demonstrates reasonable performance by incorporating masked RGB images into the gait branch, it still faces challenges when gait information is not available. In contrast, appearance-based methods exhibit slightly better performance with stationary videos compared to walking sequences, as stationary videos have less blurred boundaries due to reduced motion.

*(ii) Shape and Appearance Analysis.* Apart from comparing our method with existing methods, we also separate the shape and appearance models, PSE and AAE, to evaluate their individual contributions in ShARc. We present the results in Table 2, 3, and 4. Our appearance-based approach, AAE, demonstrates a substantial improvement over other appearance-based methods and achieves the best performance. This suggests that the averaging aggregation is indeed effective in providing supplementary information not captured by attention-based methods, thus helping to alleviate the overfitting problem. Furthermore, our shape-based model, PSE, not only outperforms other gait-based methods but also shows relatively robust performance on stationary

<sup>1</sup>The BRIAR dataset has included more subjects compared to the version used in DME, making it considerably more challenging.

| Distances | 200m | 400m | 500m | 1000m | UAV  |
|-----------|------|------|------|-------|------|
| PSE       | 38.5 | 38.2 | 35.7 | 5.3   | 25.9 |
| AAE       | 60.6 | 56.3 | 51.2 | 10.5  | 30.7 |
| ShARc     | 64.3 | 60.4 | 56.0 | 10.5  | 36.4 |

Table 5. Rank-1 accuracy for different distances in BRIAR.

| Distances          | Rank 1 | Rank 5 | Rank 20 |
|--------------------|--------|--------|---------|
| PSE                | 21.2   | 44.9   | 65.3    |
| w/o binarized sil. | 8.7    | 20.7   | 40.1    |
| w/o skeletons      | 19.7   | 35.6   | 63.4    |
| w/o 3-D shape      | 8.7    | 20.1   | 37.6    |
| AAE                | 38.3   | 63.7   | 81.8    |
| w/o att.           | 29.1   | 51.3   | 68.9    |
| w/o avg.           | 33.0   | 57.2   | 77.5    |
| w/o centroid [60]  | 30.9   | 56.1   | 75.4    |

Table 6. Ablation results for different components in ShARc. ‘att’ and ‘avg’ are attention-based and averaging aggregations.

videos, indicating that the integration of body shape features allows the model to better understand and distinguish between individuals, particularly when gait is unavailable.

It is worth noting that on datasets involving clothes-changing scenarios, such as BRIAR where the outfits between gallery and query videos are strictly different, appearance-based methods consistently outperform shape-based methods, even when both gait and body shape information are available. As shown in Table 2, appearance-based methods continue to surpass gait and body shape-based methods under different clothing conditions. One possible explanation for this observation is that the process of generating body shape (SMPL) and gait (silhouettes) features directly from RGB frames introduces noise or increases information loss during the preprocessing stage. This results in a degradation of the extracted features’ quality and their effectiveness in the re-identification task.

On the other hand, appearance-based methods can effectively leverage the rich information provided by RGB images to focus on relevant areas, even when the patterns of outfits differ between gallery and probe videos. This finding highlights the potential limitations of human-designed features, such as gait patterns or 3-D body shape, which despite being specifically and carefully designed for certain tasks, may still lead to information loss and underperform when compared to machine-designed features. In the final part of this section, we will present visualizations that further illustrate the effectiveness of our appearance-based method in handling clothes-changing scenarios.

**Ablation results.** Since the BRIAR dataset provides valuable information, such as the exact distance at which images are captured and the impact of different types of activities in the sequences, we conduct ablation experiments on a sampled validation set derived from the training sequences to analyze the selection of weights when fusing the

| Gamma  | 1    | 0.2  | 0.1  | 0    |
|--------|------|------|------|------|
| Rank 1 | 35.1 | 36.6 | 37.5 | 38.3 |

Table 7. Rank-1 accuracy for feature flattening for AvA.

|        |      |      |      |      |      |
|--------|------|------|------|------|------|
| App.   | 0.95 | 0.9  | 0.8  | 0.7  | 0.6  |
| Shape  | 0.05 | 0.1  | 0.2  | 0.3  | 0.4  |
| Rank 1 | 91.1 | 91.4 | 91.0 | 90.2 | 88.4 |

Table 8. Rank-1 accuracy for the selection of  $\alpha$ .

scores from the shape and appearance models.

*Distances.* We present the performance of our method across various distances in Table 5. We select five distance variations from the BRIAR dataset: 200 meters, 400 meters, 500 meters, 1000 meters, and video captured from UAV cameras. Generally, performance is better at shorter distances. However, we see a significant performance drop at 1000 meters, where the bodies in images are nearly indistinguishable. The results for UAV-captured images aren't as strong as those at 200 meters. This is due to the incomplete body images, as the UAV images are taken with the head occluding the whole body. The performance decline of PSE is less compared to AAE, showing its relative robustness in identification when occlusion is present.

*Model Components Ablations.* Our pipeline consists of multiple sub-modules, and we analyze the individual contribution of each component in both branches. For gait representation, we have two components: masked RGB and binarized silhouettes. We investigate the contributions of binarized silhouette masks and masked RGB images independently. It is important to note that the masked RGB images in this case are resized to a smaller scale, similar to binarized silhouettes, to provide information about the separation of body parts rather than directly using appearance for training. To remove each component in the network, we zero out the corresponding input for analysis.

We show the results in Table 6. For the shape-based branch, masked RGB contributes the most, while 3-D body shape and binarized silhouettes contribute almost equally. Compared to other modalities, 3-D masked RGB images precisely provide more internal content for the gait branch, enabling the network to understand the boundary of different body parts and the movement of each part. For the appearance branch, we find that both aggregation contribute similarly to the final performance, and the combination of both yields the best results. Furthermore, using centroid averaging [60] when registering gallery examples also has a significant contribution to the final performance.

*Feature Flattening.* For the flattening layer in averaging aggregation, we analyze the different Gamma and their corresponding results in Table 7. When Gamma is 1, we have simple averaging across all the features. We observe that

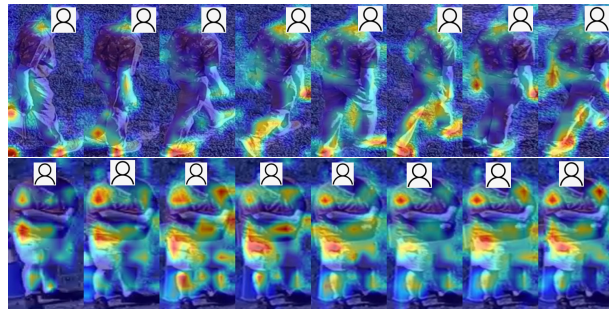


Figure 5. Attention generated from appearance model for (a) a walking sequence and (b) a stationary video for two examples taken from 100 meters distance category.

with higher gamma values, our performance improves, indicating that the results exhibit more discriminative patterns. When gamma is infinity, the final feature representation becomes binarized and yields the best performance.

*Choice of  $\alpha$ .* To combine the two modalities, we construct a small validation set from the training data to analyze the weights between appearance and shape models, and present the results in Table 8. We find that when the weight is 0.9 for appearance and 0.1 for shape, the model achieves the best performance. For shape-based methods, we use Euclidean distance instead of cosine distance; thus, 0.1 does not imply that it contributes minimally, but rather serves as a scaling factor for  $S_{shape}$  during combination. For generalizability, we use this  $\gamma$  and  $\alpha$  for all datasets.

**Visualization for Appearance Branch.** In the BRIAR dataset, where query and gallery images feature distinct outfits, we use GradCam [47] to visualize network focus. Figure 5 presents two examples taken from 100-meter-distance cameras, one during walking and another while stationary. For walking videos, the network focuses mainly on the lower body and arms, suggesting implicit pose pattern extraction. In stationary scenarios, attention is directed towards the waist and shoulders, important areas for discerning body shape. We observe this trend across multiple examples, although quantification has not been performed.

## 5. Conclusion

In this paper, we introduce ShARc, a shape and appearance-based method for identification in-the-wild. Our approach explicitly explores the contribution of body shape and appearance to the model with two encoders, pose and shape encoder for body shape and motion, and aggregated appearance encoder for human appearance. ShARc is able to handle most of the challenges for identification in the wild, such as occlusion, non-walking sequences, change of clothes, and image degradations. We have compared our method on three public datasets, including BRIAR, CCVID, and MEVID, and show state-of-the-art performance.



## References

- [1] Weizhi An, Shiqi Yu, Yasushi Makihara, Xinhui Wu, Chi Xu, Yang Yu, Rijun Liao, and Yasushi Yagi. Performance evaluation of model-based gait on multi-view very large population database with pose sequences. *TBIOM*, 2(4):421–430, 2020. [2](#)
- [2] Hanqing Chao, Yiwei He, Junping Zhang, and Jianfeng Feng. Gaitset: Regarding gait as a set for cross-view gait recognition. In *AAAI*, pages 8126–8133, 2019. [1](#), [2](#), [3](#), [6](#), [7](#)
- [3] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587*, 2017. [5](#)
- [4] Yoonki Cho, Woo Jae Kim, Seunghoon Hong, and Sung-Eui Yoon. Part-based pseudo label refinement for unsupervised person re-identification. In *CVPR*, pages 7308–7318, 2022. [2](#)
- [5] David Cornett, Joel Brogan, Nell Barber, Deniz Aykac, Seth Baird, Nicholas Burchfield, Carl Dukes, Andrew Duncan, Regina Ferrell, Jim Goddard, et al. Expanding accurate person recognition to new altitudes and ranges: The briar dataset. In *WACV*, pages 593–602, 2023. [2](#), [5](#)
- [6] Ju Dai, Pingping Zhang, Dong Wang, Huchuan Lu, and Hongyu Wang. Video person re-identification by temporal residual learning. *TIP*, 28(3):1366–1377, 2018. [2](#)
- [7] Daniel Davila, Dawei Du, Bryon Lewis, Christopher Funk, Joseph Van Pelt, Roderic Collins, Kellie Corona, Matt Brown, Scott McCloskey, Anthony Hoogs, et al. Mevid: Multi-view extended videos with identities for video person re-identification. In *WACV*, pages 1634–1643, 2023. [1](#), [2](#), [5](#), [6](#), [7](#)
- [8] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *CVPR*, pages 4690–4699, 2019. [1](#), [2](#)
- [9] Yueqi Duan, Jiwen Lu, and Jie Zhou. Uniformface: Learning deep equidistributed representation for face recognition. In *CVPR*, pages 3415–3424, 2019. [1](#), [2](#)
- [10] Chanho Eom, Geon Lee, Junghyup Lee, and Bumsub Ham. Video-based person re-identification with spatial and temporal memory networks. In *ICCV*, pages 12036–12045, 2021. [6](#)
- [11] Chao Fan, Junhao Liang, Chuanfu Shen, Saihui Hou, Yongzhen Huang, and Shiqi Yu. Opengait: Revisiting gait recognition toward better practicality. *arXiv preprint arXiv:2211.06597*, 2022. [2](#), [5](#)
- [12] Chao Fan, Yunjie Peng, Chunshui Cao, Xu Liu, Saihui Hou, Jiannan Chi, Yongzhen Huang, Qing Li, and Zhiqiang He. Gaitpart: Temporal part-based model for gait recognition. In *CVPR*, pages 14225–14233, 2020. [1](#), [2](#), [6](#), [7](#)
- [13] Yang Fu, Xiaoyang Wang, Yunchao Wei, and Thomas Huang. Sta: Spatial-temporal attention for large-scale video-based person re-identification. In *AAAI*, volume 33, pages 8287–8294, 2019. [2](#)
- [14] Jiyang Gao and Ram Nevatia. Revisiting temporal modeling for video-based person reid. *arXiv preprint arXiv:1805.02104*, 2018. [2](#)
- [15] Zan Gao, Hongwei Wei, Weili Guan, Jie Nie, Meng Wang, and Shenyong Chen. A semantic-aware attention and visual shielding network for cloth-changing person re-identification. *arXiv preprint arXiv:2207.08387*, 2022. [2](#)
- [16] Xinqian Gu, Hong Chang, Bingpeng Ma, Shutao Bai, Shiguang Shan, and Xilin Chen. Clothes-changing person re-identification with rgb modality only. In *CVPR*, pages 1060–1069, 2022. [2](#), [5](#), [6](#), [7](#)
- [17] Xinqian Gu, Hong Chang, Bingpeng Ma, Hongkai Zhang, and Xilin Chen. Appearance-preserving 3d convolution for video-based person re-identification. In *ECCV*, pages 228–243. Springer, 2020. [6](#)
- [18] Yuxiang Guo, Cheng Peng, Chun Pong Lau, and Rama Chelappa. Multi-modal human authentication using silhouettes, gait and rgb. *arXiv preprint arXiv:2210.04050*, 2022. [2](#), [5](#), [6](#), [7](#)
- [19] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016. [5](#), [6](#)
- [20] Tianyu He, Xu Shen, Jianqiang Huang, Zhibo Chen, and Xian-Sheng Hua. Partial person re-identification with part-part correspondence learning. In *CVPR*, pages 9105–9115, 2021. [2](#)
- [21] Peixian Hong, Tao Wu, Ancong Wu, Xintong Han, and Wei-Shi Zheng. Fine-grained shape-appearance mutual learning for cloth-changing person re-identification. In *CVPR*, pages 10513–10522, 2021. [2](#)
- [22] Ruibing Hou, Hong Chang, Bingpeng Ma, Rui Huang, and Shiguang Shan. Bicnet-tks: Learning efficient spatial-temporal representation for video person re-identification. In *CVPR*, pages 2014–2023, 2021. [6](#)
- [23] Ruibing Hou, Hong Chang, Bingpeng Ma, Shiguang Shan, and Xilin Chen. Temporal complementary learning for video person re-identification. In *ECCV*, pages 388–405. Springer, 2020. [2](#), [6](#), [7](#)
- [24] Saihui Hou, Chunshui Cao, Xu Liu, and Yongzhen Huang. Gait lateral network: Learning discriminative and compact representations for gait recognition. In *ECCV*, pages 382–398, 2020. [2](#)
- [25] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. Human3.6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *TPAMI*, 36(7):1325–1339, 2013. [5](#)
- [26] Zilong Ji, Xiaolong Zou, Xiaohan Lin, Xiao Liu, Tiejun Huang, and Si Wu. An attention-driven two-stage clustering method for unsupervised person re-identification. In *ECCV*, pages 20–36, 2020. [2](#)
- [27] Xin Jin, Tianyu He, Kecheng Zheng, Zhiheng Yin, Xu Shen, Zhen Huang, Ruoyu Feng, Jianqiang Huang, Zhibo Chen, and Xian-Sheng Hua. Cloth-changing person re-identification from a single image with gait prediction and regularization. In *CVPR*, pages 14278–14287, 2022. [2](#)
- [28] Minchul Kim, Anil K Jain, and Xiaoming Liu. Adaface: Quality adaptive margin for face recognition. In *CVPR*, pages 18750–18759, 2022. [1](#), [2](#)
- [29] Jianing Li, Jingdong Wang, Qi Tian, Wen Gao, and Shiliang Zhang. Global-local temporal representations for video person re-identification. In *ICCV*, pages 3958–3967, 2019. [2](#)

- [30] Shuang Li, Slawomir Bak, Peter Carr, and Xiaogang Wang. Diversity regularized spatiotemporal attention for video-based person re-identification. In *CVPR*, pages 369–378, 2018. [2](#)
- [31] Wei Li, Rui Zhao, Tong Xiao, and Xiaogang Wang. Deepreid: Deep filter pairing neural network for person re-identification. In *CVPR*, pages 152–159, 2014. [2](#)
- [32] Xiang Li, Yasushi Makihara, Chi Xu, Yasushi Yagi, Shiqi Yu, and Mingwu Ren. End-to-end model-based gait recognition. In *ACCV*, 2020. [2](#)
- [33] Yulin Li, Jianfeng He, Tianzhu Zhang, Xiang Liu, Yongdong Zhang, and Feng Wu. Diverse part discovery: Occluded person re-identification with part-aware transformer. In *CVPR*, pages 2898–2907, 2021. [2](#)
- [34] Junhao Liang, Chao Fan, Saihui Hou, Chuanfu Shen, Yongzhen Huang, and Shiqi Yu. Gaitedge: Beyond plain end-to-end gait recognition for better practicality. *arXiv preprint arXiv:2203.03972*, 2022. [2](#)
- [35] Beibei Lin, Shunli Zhang, and Xin Yu. Gait recognition via effective global-local feature representation and local temporal aggregation. In *ICCV*, pages 14648–14656, 2021. [1](#), [2](#), [3](#), [6](#), [7](#)
- [36] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, pages 740–755, 2014. [5](#)
- [37] Chih-Ting Liu, Chih-Wei Wu, Yu-Chiang Frank Wang, and Shao-Yi Chien. Spatially and temporally efficient non-local attention network for video-based person re-identification. *arXiv preprint arXiv:1908.01683*, 2019. [2](#)
- [38] Hao Liu, Zequn Jie, Karlekar Jayashree, Meibin Qi, Jianguo Jiang, Shuicheng Yan, and Jiashi Feng. Video-based person re-identification with accumulative motion context. *TCSVT*, 28(10):2788–2802, 2017. [2](#)
- [39] Ziyu Liu, Hongwen Zhang, Zhenghao Chen, Zhiyong Wang, and Wanli Ouyang. Disentangling and unifying graph convolutions for skeleton-based action recognition. In *CVPR*, pages 143–152, 2020. [5](#)
- [40] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. Smpl: A skinned multi-person linear model. *TOG*, 34(6):1–16, 2015. [2](#), [5](#)
- [41] Niall McLaughlin, Jesus Martinez Del Rincon, and Paul Miller. Recurrent convolutional network for video-based person re-identification. In *CVPR*, pages 1325–1334, 2016. [2](#)
- [42] Dushyant Mehta, Helge Rhodin, Dan Casas, Pascal Fua, Oleksandr Sotnychenko, Weipeng Xu, and Christian Theobalt. Monocular 3d human pose estimation in the wild using improved cnn supervision. In *3DV*, pages 506–516, 2017. [5](#)
- [43] Christopher B Nalty, Neehar Peri, Joshua Gleason, Carlos D Castillo, Shuowen Hu, Thirimachos Bourlai, and Rama Chellappa. A brief survey on person recognition at a distance. *arXiv preprint arXiv:2212.08969*, 2022. [1](#)
- [44] Priyank Pathak, Amir Erfan Eshratifar, and Michael Gormish. Video person re-id: Fantastic techniques and where to find them (student abstract). In *AAAI*, volume 34, pages 13893–13894, 2020. [6](#), [7](#)
- [45] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115:211–252, 2015. [6](#)
- [46] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *CVPR*, pages 815–823, 2015. [5](#)
- [47] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *ICCV*, pages 618–626, 2017. [8](#)
- [48] Chuanfu Shen, Chao Fan, Wei Wu, Rui Wang, George Q Huang, and Shiqi Yu. Lidar gait: Benchmarking 3d gait recognition with point clouds. *arXiv preprint arXiv:2211.10598*, 2022. [2](#)
- [49] Jianlou Si, Honggang Zhang, Chun-Guang Li, Jason Kuen, Xiangfei Kong, Alex C Kot, and Gang Wang. Dual attention matching network for context-aware feature sequence based person re-identification. In *CVPR*, pages 5363–5372, 2018. [2](#)
- [50] Chunfeng Song, Yongzhen Huang, Yan Huang, Ning Jia, and Liang Wang. Gaitnet: An end-to-end network for gait based human identification. *PR*, 96:106988, 2019. [7](#)
- [51] Arulkumar Subramaniam, Athira Nambiar, and Anurag Mittal. Co-segmentation inspired attention networks for video-based person re-identification. In *ICCV*, pages 562–572, 2019. [2](#)
- [52] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. Deep high-resolution representation learning for human pose estimation. In *CVPR*, pages 5693–5703, 2019. [5](#)
- [53] Yu Sun, Qian Bao, Wu Liu, Yili Fu, Michael J Black, and Tao Mei. Monocular, one-stage, regression of multiple 3d people. In *ICCV*, pages 11179–11188, 2021. [5](#)
- [54] Yifan Sun, Liang Zheng, Yi Yang, Qi Tian, and Shengjin Wang. Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline). In *ECCV*, pages 480–496, 2018. [2](#)
- [55] Noriko Takemura, Yasushi Makihara, Daigo Muramatsu, Tomio Echigo, and Yasushi Yagi. Multi-view large population gait dataset and its performance evaluation for cross-view gait recognition. *TCVA*, 10(1):1–14, 2018. [2](#)
- [56] Torben Teepe, Ali Khan, Johannes Gilg, Fabian Herzog, Stefan Hörmann, and Gerhard Rigoll. Gaitgraph: Graph convolutional network for skeleton-based gait recognition. In *ICIP*, pages 2314–2318, 2021. [2](#), [3](#), [5](#)
- [57] Jingdong Wang, Ke Sun, Tianheng Cheng, Borui Jiang, Chaorui Deng, Yang Zhao, Dong Liu, Yadong Mu, Mingkui Tan, Xinggang Wang, et al. Deep high-resolution representation learning for visual recognition. *TPAMI*, 43(10):3349–3364, 2020. [2](#)
- [58] Yingquan Wang, Pingping Zhang, Shang Gao, Xia Geng, Hu Lu, and Dong Wang. Pyramid spatial-temporal aggregation for video-based person re-identification. In *ICCV*, pages 12026–12035, 2021. [1](#), [2](#), [4](#), [6](#), [7](#)

- [59] Yandong Wen, Kaipeng Zhang, Zhifeng Li, and Yu Qiao. A discriminative feature learning approach for deep face recognition. In *ECCV*, pages 499–515, 2016. [5](#)
- [60] Mikołaj Wiczcerek, Barbara Rychalska, and Jacek Dąbrowski. On the unreasonable effectiveness of centroids in image retrieval. In *ICONIP*, pages 212–223, 2021. [1](#), [2](#), [5](#), [6](#), [7](#), [8](#)
- [61] Yiming Wu, Omar El Farouk Bourahla, Xi Li, Fei Wu, Qi Tian, and Xue Zhou. Adaptive graph representation learning for video person re-identification. *TIP*, 29:8821–8830, 2020. [6](#)
- [62] Shuangjie Xu, Yu Cheng, Kang Gu, Yang Yang, Shiyu Chang, and Pan Zhou. Jointly attentive spatial-temporal pooling networks for video-based person re-identification. In *ICCV*, pages 4733–4742, 2017. [2](#)
- [63] Shiqi Yu, Daoliang Tan, and Tieniu Tan. A framework for evaluating the effect of view angle, clothing and carrying condition on gait recognition. In *ICPR*, volume 4, pages 441–444, 2006. [2](#)
- [64] Xianghao Zang, Ge Li, and Wei Gao. Multidirection and multiscale pyramid in transformer for video-based pedestrian retrieval. *TH*, 18(12):8776–8785, 2022. [6](#)
- [65] Liming Zhao, Xi Li, Yueting Zhuang, and Jingdong Wang. Deeply-learned part-aligned representations for person re-identification. In *ICCV*, pages 3219–3228, 2017. [2](#)
- [66] Jinkai Zheng, Xinchun Liu, Wu Liu, Lingxiao He, Chenggang Yan, and Tao Mei. Gait recognition in the wild with dense 3d representations and a benchmark. In *CVPR*, pages 20228–20237, 2022. [1](#), [2](#), [3](#), [5](#), [6](#)
- [67] Liang Zheng, Zhi Bie, Yifan Sun, Jingdong Wang, Chi Su, Shengjin Wang, and Qi Tian. Mars: A video benchmark for large-scale person re-identification. In *ECCV*, pages 868–884, 2016. [2](#)
- [68] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. Scalable person re-identification: A benchmark. In *ICCV*, pages 1116–1124, 2015. [2](#)
- [69] Liang Zheng, Hengheng Zhang, Shaoyan Sun, Manmohan Chandraker, Yi Yang, and Qi Tian. Person re-identification in the wild. In *CVPR*, pages 1367–1376, 2017. [2](#)
- [70] Zhen Zhou, Yan Huang, Wei Wang, Liang Wang, and Tieniu Tan. See the forest for the trees: Joint spatial and temporal recurrent neural networks for video-based person re-identification. In *CVPR*, pages 4747–4756, 2017. [2](#)
- [71] Haidong Zhu, Ye Yuan, Yiheng Zhu, Xiao Yang, and Ram Nevatia. Open: Order-preserving pointcloud encoder decoder network for body shape refinement. In *ICPR*, pages 521–527, 2022. [5](#)
- [72] Haidong Zhu, Wanrong Zheng, Zhaoheng Zheng, and Ram Nevatia. Gaitref: Gait recognition with refined sequential skeletons. *arXiv preprint arXiv:2304.07916*, 2023. [1](#), [2](#), [6](#)
- [73] Haidong Zhu, Zhaoheng Zheng, and Ram Nevatia. Gait recognition using 3-d human body shape inference. In *WACV*, pages 909–918, 2023. [1](#), [2](#), [3](#)
- [74] Kuan Zhu, Haiyun Guo, Tianyi Yan, Yousong Zhu, Jinqiao Wang, and Ming Tang. Pass: Part-aware self-supervised pre-training for person re-identification. In *ECCV*, pages 198–214, 2022. [2](#)