

# Gait Recognition Using 3-D Human Body Shape Inference

Haidong Zhu   Zhaoheng Zheng   Ram Nevatia  
University of Southern California  
{haidongz | zhaoheng.zheng | nevatia@usc.edu}

## Abstract

Gait recognition, which identifies individuals based on their walking patterns, is an important biometric technique since it can be observed from a distance and does not require the subject’s cooperation. Recognizing a person’s gait is difficult because of the appearance variants in human silhouette sequences produced by varying viewing angles, carrying objects, and clothing. Recent research has produced a number of ways for coping with these variants. In this paper, we present the usage of inferring 3-D body shapes distilled from limited images, which are, in principle, invariant to the specified variants. Inference of 3-D shape is a difficult task, especially when only silhouettes are provided in a dataset. We provide a method for learning 3-D body inference from silhouettes by transferring knowledge from 3-D shape prior from RGB photos. We use our method on multiple existing state-of-the-art gait baselines and obtain consistent improvements for gait identification on two public datasets, CASIA-B and OUMVLP, on several variants and settings, including a new setting of novel views not seen during training.

## 1. Introduction

Many biometrics, such as face ID [8, 34], have been developed for automated human identification. One such biometric is gait, which has the advantage of being able to be acquired from long distance and without the subjects’ cooperation. Gait recognition [12, 36, 40, 42] aims to find the uniqueness for a sequence of walking patterns and posture of a person in the binarized silhouette sequence describing human boundaries. However, appearance variances in 2-D images, like camera positions, carried-on objects, and clothing, introduce additional disparity in the human shape and make the task of recognition challenging. Fig. 1 (a) shows these variations in extracted silhouette for a person under three different appearance variances.

To address the issue of appearance variances, researchers have developed part-based deep-learning models that focus on local patterns. For example, GaitPart [10] splits the im-

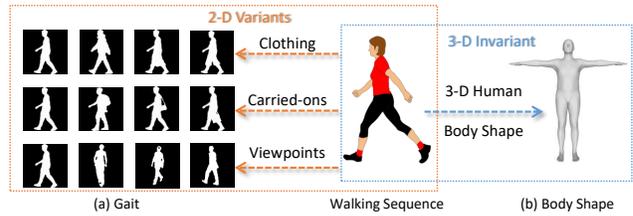


Figure 1. (a) 2-D silhouette sequences suffer from different appearance variances, such as clothing, carried-on bags and camera viewpoints. (b) However, 3-D skinned human body shape is robust and shows consistent output for the same person.

age into several patches to encode the part-based features for gait recognition, whereas GaitGL [23] utilizes local features along with the global ones for the analysis. By limiting the variances to a small portion of the feature, these strategies aim to reduce the influence of the variances. However, features encoded by these approaches are still impacted.

In this paper, we propose inferring 3-D human body shape representations directly from silhouette sequences by using knowledge distillation to learn from a small number of RGB samples. We observe that the 3-D body shape of the human, as illustrated in Fig. 1 (b) for example, is, in principle, invariant to viewpoint, carrying objects, and clothing, and might therefore be useful for human identification under such challenging scenarios. Nevertheless, inference of the underlying 3-D shape is difficult in and of itself. Recent work [31, 28, 45] has created numerous approaches to infer 3-D shape from RGB images, but no work directly infers the body shape from a silhouette sequence. Many gait datasets lack RGB photos due to confidentiality, making such inferences more difficult.

To infer 3-D body shapes from the silhouette sequence, we exploit a temporal shift between the features obtained from adjacent frames in the silhouette sequence. Considering the consistency of the body shape of the same individual in a video sequence, we extract and reconstruct a single body shape for a video sequence. We combine body shapes acquired from our approach with 2-D gait features collected from certain state-of-the-art gait recognition methods [6, 10, 23, 15] to build our module on each of them. To

supervise the generation of the body shape from the silhouette sequence, we distill and transfer the knowledge from a small set of RGB images, denoted as human body prior, and propagate it to gait. We demonstrate that adding 3-D body shape feature inferred from silhouette sequence significantly improves gait recognition accuracy on two public datasets (CASIA-B [43] and OUMVLP [37]), particularly for novel viewpoints that were not observed during training with fewer available training instances, which is a new setting in our experiment.

A recent paper, Gait3D [45], has also proposed using 3-D body shape for gait recognition. However, our work differs in the following manner: instead of inferring 3-D body shapes from all RGB frames, we infer 3-D shapes from silhouettes via distilling knowledge from a small set of images. Another is in our use of temporal information in 3-D inference. Gait3D extracts framewise body shape, while we extract video-level body shape using temporal consistency.

In summary, our contributions are summarized as follows: 1) We apply the 3-D human body inferred from gait to eliminate the effects of different clothing, carried-on objects and viewpoints for gait recognition; 2) We distill the knowledge of human body prior from limited single-frame RGB images and transfer to the silhouette sequence for body shape reconstruction directly from gait; and 3) We explore the setting for gait recognition on novel camera positions to assess the generalization of gait recognition models with fewer available data.

## 2. Related Work

**Gait Recognition.** For a silhouette sequence describing a person’s walking pose, gait recognition is to find the corresponding identity of the person in the gallery. Recently, researchers have proposed different methods [38, 36, 6, 23, 15, 10, 45, 21, 9, 46, 20] for extracting the identity information from the gait sequence for recognition. GaitSet [6] treats the whole sequence as a set of different images for set pooling and feature fusion. GaitPart [10] splits the gait image into different parts and extracts the feature from each local pattern for analysis. GLN [15] utilizes both silhouette-level and set-level features and fuses them for different gait analyses at different stages. GaitGL [23] introduces using both local and global features: local features are computed by splitting an image into several patches and encoding the feature for each patch; global features are framewise encoded features and combine them together for gait recognition. GaitNet [36] and GaitGraph [38] use the consistency between RGB images and graph recognition network for recognizing the identity of the human in the sequence. These methods focus on extracting and distinguishing information directly from the 2-D gait sequence. Other methods, such as PoseGait [22] and ModelGait [20], require the RGB images for all the training instances, which are sometime

difficult to get due to privacy issues.

Gait3D [45] uses the 3-D body shape extracted from RGB images, which has extra input compared with other methods. Since the gait sequences are binarized images, when people have different clothing or carried-on objects and are shot by the camera from different positions, predictions from the features extracted are affected.

**Knowledge Distillation.** The task of knowledge distillation is to transfer the knowledge from one model to another. Knowledge transfer has been successfully applied in tasks such as visual and speech recognition [11, 7] and between different modalities [39], etc. Researchers have proposed several methods for knowledge distillation and transfer [1, 13, 14, 17, 19, 27, 41, 44]. For these methods, their inputs for different models are mostly from the same modalities: both the source and target are data sequences or single frames. Knowledge distillation and transfer from RGB images to gait sequences require understanding both gait sequences and single-frame images. To transfer the knowledge from an image to a video, we need to distribute the information to individual frames of the video.

**3-D Body Shape Reconstruction.** A model needs lots of the knowledge [16, 18] to reconstruct the 3-D body shapes. Methods for 3-D body shape reconstruction can be divided into two mainstreams: parametric methods, such as SMPLify-X [28] and SMPLify [3], and non-parametric methods, such as PIFu [31] and PIFuHD [32].

For parametric methods, SMPLify [3] and SMPLify-X [28] reconstruct the human body shape based on the pre-defined parameterized skinned models, SMPL [25] and SMPL-X [28]. As non-parametric methods, PIFu [31] and PIFuHD [32] utilize the implicit function for representing the shape and predict whether points in the space are inside or outside the object. These non-parametric methods do not record locations of points on the object surface but understand the whole body shape correspondingly. With the introduction of NeRF [26], researchers also introduce Animatable NeRF [29] and Neural Body [29] for reconstructing the human body shape in the video sequence with SMPL priors. Compared with these methods, due to the lack of RGB images in gait datasets, we use the distilled body prior from a small set of examples and extract body shapes from the silhouette sequence instead of RGB images.

## 3. Method

Our network consists of two branches, one for gait feature extraction and the other for body shape feature extraction from RGB images, which is shown in Fig. 2. For gait inputs, we have a silhouette feature encoder and a body shape feature encoder to encode the gait and body shape feature. To supervise the generation of body shapes, we simultaneously extract knowledge from selected RGB frames using a human body reconstruction model. Then, we extract

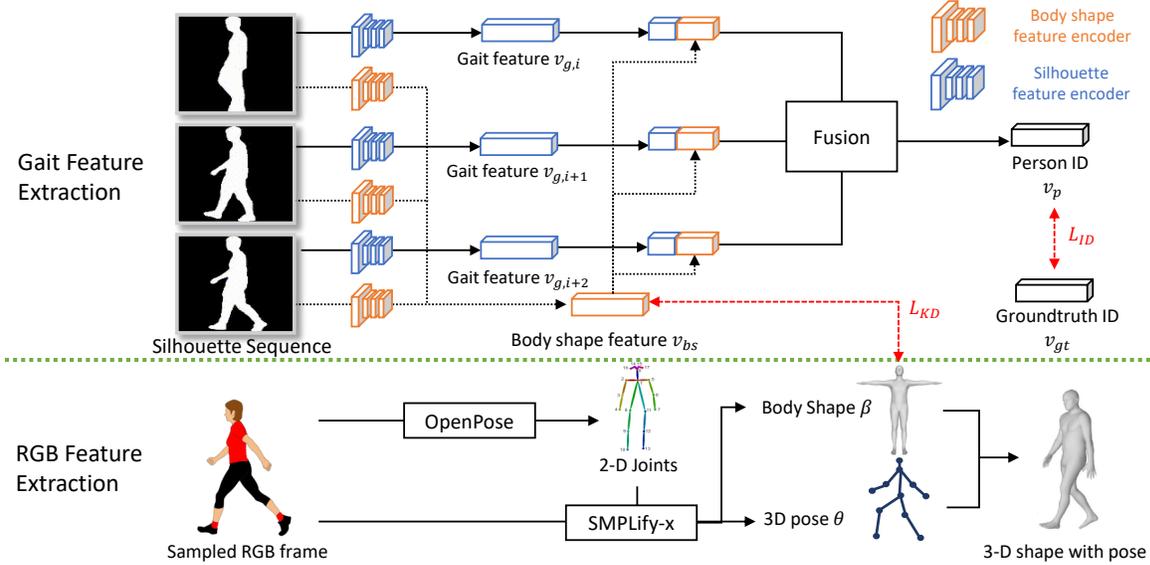


Figure 2. Our Proposed method for gait recognition with 3-D human body shape. During inference, only the features extracted from gait branch are used. Features from RGB images (below the green line) are only used for training when corresponding images are available.

and transfer these inferred body shape information from RGB frames to the gait branch’s body shape encoder.

In the remaining of this section, we will first introduce our gait pipeline in Sec. 3.1 for how gait features and body shape features are extracted for identification, and then discuss how body shape of selected RGB images are generated and used as the supervision for the gait branch in Sec. 3.2.

### 3.1. Gait Feature Extraction

We propose two feature encoders to extract features from gait images: a silhouette feature encoder to extract walking patterns from the gait sequence and a body shape feature encoder to extract the body shape.

**Silhouette Feature Encoder.** The silhouette feature encoder projects the individual frames  $\{f_i\}_{i=1,\dots,m}$  of a gait sequence  $G$  to their feature representations  $\{v_{g,i}\}_{i=1,\dots,m}$ , where  $m$  is the number of frames. To verify the generality of using the body shape features to improve the gait recognition network, we choose four state-of-the-art gait recognition methods as the gait feature encoder for comparison: GaitSet [6], GaitPart [10], GaitGL [23], and GLN [15].

**Body Shape Feature Encoder.** To extract the body shape feature from the silhouette sequence, we input the gait sequence  $G$  and project it to the feature space  $v_{bs}$  representing the body shape of the person in the video. Extracting the body shape from a single gait sequence is difficult since the single binary silhouette only provides the boundary of a human body and lacks essential information. Thus we need the neighbor frames to help complete the missing information for extracting the whole human body shape. We show our proposed body shape feature encoder in Fig. 3. The encoder consists of  $n$  blocks, where every block includes

a convolution and a temporal shifting (TS) operation. The convolution operator takes the frame-wise feature from the raw gait sequence or the previous layer as input, and operators in the same block share weight. After the convolution operation in each block, we follow [24] to exchange 12.5% of the features of part of the channels between the previous and future segments of video for temporal shifting.

We preserve the first frame of the sequence’s content, which should be exchanged with the previous frame since there is no frame before it. We also keep the feature from the future segments for the last frame. After the last layer of feature shifting, we do a temporal average pooling on the extracted feature sequences to generate the final body shape feature  $v_{bs}$ . With the features from two encoders, we concatenate the body shape feature  $v_{bs}$  to each of the gait features  $\{v_{g,i}\}_{i=1,\dots,m}$ . We maxpool the features along the temporal and horizontal dimension following the implementation of [6, 10, 23, 15] and apply two fully-connected layers, whose dimensionalities match with the backbone network we used as the silhouette feature encoder, to generate feature  $v_p$ , representing the person’s identity. We calculate the similarity between  $v_p$  and the groundtruth  $v_{gt}$  and calculate the identity loss  $L_{ID}$  following [6, 10, 15, 23].

### 3.2. Human Body Prior Distillation and Transfer

The purpose of inferring 3-D human body shapes is to separate movement patterns from variations in the appearance of 2-D silhouettes. Due to the absence of ways to directly reconstruct the 3-D human body from the gait, we first extract the shape priors from a small set of RGB frames in the gait sequence, then distill and transfer this information to the body shape feature encoder in the gait branch if

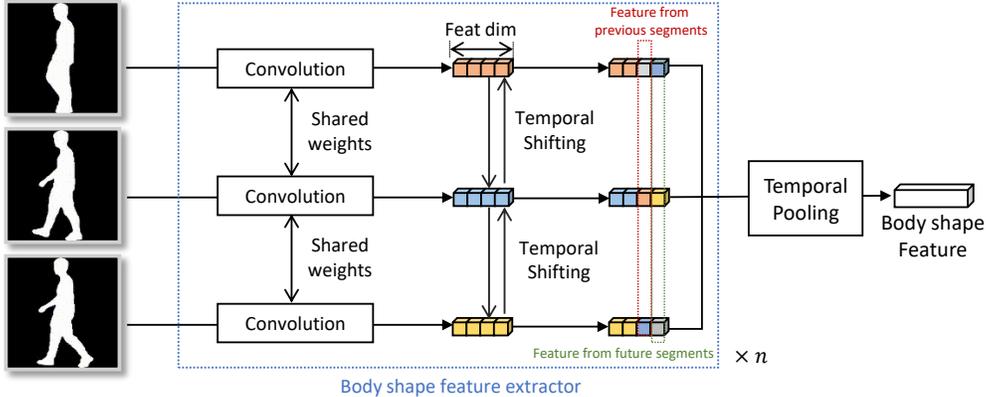


Figure 3. Our proposed body shape encoder for silhouette sequence input.  $n$  represents the repeated time for the operations in the block.

corresponding RGB images are available.

**Body Prior Inference.** To infer 3-D body prior from the RGB images, we follow [28] for using SMPL-X reconstructed from SMPLify-X as body shapes. Compared with other 3-D body reconstruction models such as PIFu [31] and PIFuHD [32], SMPLify-X models human bodies with a strong human prior for the skinned body and outputs consistent results for the same person with different appearances, such as clothing, helping gait encoders to distinguish body shape from appearance variances in 2-D silhouettes. In addition, as a parametric method, SMPLify-X provides us the body shape feature decoupled from its pose and its strong prior can help us generate the complete shape even with some mild occlusions.

Considering the time consumption for inferring body prior with SMPLify-X [28], it is not feasible to extract the body prior for all frames in the video or image sequence. Since the identity of the person within the same video is consistent, the inferred 3-D body prior without the pose should be identical across all frames in which the person is discernible. Consequently, we infer the prior form based on one of the frames taken from the RGB sequence or video in conjunction with the gait sequence  $G$ .

To select this frame, we first extract the skeletons  $\{s_i\}_{i=1,\dots,m}$  of the whole sequence using OpenPose [4], followed by finding the longest sequence in the segments of  $\{s_i\}$  with skeletons detectable and use the middle frame of this segment, which we annotate as  $s_r$ , to represent the body prior of the whole video. In this way, we can guarantee the quality of image  $s_r$  used to infer the 3-D body prior since the longest segments with skeletons detectable can ensure stable and consistent performance for pose detection and estimation, making it easier for body prior extraction.

We then reconstruct the whole human body prior for  $s_r$  by generating the shape feature  $\beta$  and 3-D pose  $\theta$  following [28].  $\beta$  is a 1-D vector with a size of 10 describing the appearance of the reconstructed body prior, and  $\theta$  only includes 3-D joint positions. By combining the  $\beta$  and  $\theta$ , we

can reconstruct a full 3-D body model for a specific pose. In our experiments, we only use the  $\beta$  as the body prior feature  $v_{br}$  to guide the body shape based on silhouette  $v_{bs}$ . No skeleton information is used for gait recognition to avoid the different accuracies of the prediction of skeleton.

**Knowledge Distillation and Transfer.** With the body prior features  $v_{br}$  from the selected RGB frame and  $v_{bs}$  from the silhouette sequence, we utilize  $v_{br}$  to guide the training of feature  $v_{bs}$  from the body shape feature encoder in the gait branch. We use CRD (Contrastive Representation Distillation) [39] for distilling knowledge between features from  $v_{br}$  and  $v_{bs}$  following

$$L_{KD} = \mathbb{E}_{q(v_{br}, v_{bs}|C=1)}[\log h(v_{br}, v_{bs})] + \mathbb{E}_{q(v_{br}, v_{bs}|C=0)}[\log(1 - h(v_{br}, v_{bs}))] \quad (1)$$

$$h(s, t) = \frac{\exp(f_1(v_{bs})^T \cdot f_2(v_{br}))}{\exp(f_1(v_{bs})^T \cdot f_2(v_{br})) + \frac{N}{M}}$$

where  $f_1$  and  $f_2$  are two linear projection layer with  $L_2$  norm for projecting  $v_{br}$  and  $v_{bs}$ .  $N$  is the batch size and  $M$  is the cardinality of the dataset.  $C$  is 1 while  $v_{br}$  and  $v_{bs}$  are from the same identity and 0 if not. We will compare CRD with some other knowledge distillation methods in ablation studies. During training, we have two different loss functions,  $L_{ID}$  for gait recognition loss and  $L_{KD}$  for knowledge distillation between the features of inferred 3-D body prior from the selected RGB frame and the gait sequence,  $v_{br}$  and  $v_{bs}$ . We use a hyperparameter  $\lambda$  for balancing two losses. The final objective is shown as

$$L = \lambda_1 L_{ID} + \lambda_2 L_{KD} \quad (2)$$

We set  $\lambda_1$  to 1 empirically. Following ablations in the supplementary material, we set  $\lambda_2$  to 1 for knowledge transfer for the examples with RGB images and 0 for others.

## 4. Experiments

In this section, we show the details of our implementation for the experiment and the results. We first discuss our

setups for the experiments in Sec. 4.1, followed by our analysis based on the experiment results in Sec. 4.2.

#### 4.1. Experimental Setup

For experimental setup, we discuss datasets with the baseline methods and criteria. We also introduce the new setting of gait recognition where camera positions for training and test are mutually exclusive.

**Datasets.** We conduct our experiments on two public datasets, CASIA-B [43] and OU-MVLP [37].

CASIA-B [43] is a gait recognition dataset with 124 objects with 10 different walking variants for each subject, where 6 are for normal walking (NM), 2 for walking while carrying bags (BG) and 2 for different clothing (CL). Each variant is recorded from 11 different camera viewpoints between  $0^\circ$  and  $180^\circ$  with 18 as the gap, making 110 videos for each subject. We follow [6, 10, 15, 23] to use silhouette sequences of the first 74 subjects for training. During inference, we use the first four walking variances in normal walking conditions (NM) as the gallery set, which is the identity library for the test set. The remaining 2 variants in NM, along with the sequences in BG and CL for the remaining 50 subjects, are used as probes for evaluation to find the correct identity in the gallery set.

In addition to using all the camera positions for supervised gait recognition, we introduce a new zero-shot setting where camera viewpoints used for training and testing are mutually exclusive. For all the camera viewpoints in the dataset, we only sample partial angles between  $0^\circ$  and  $90^\circ$  for training and use the viewpoints between  $108^\circ$  and  $180^\circ$  for inference to assess the model’s performance when encountering novel viewpoints. We will further discuss about this dataset in the supplementary material.

OUMVLP [37] is a large gait recognition dataset with 10,307 subjects. Each subject has 2 different sequences for normal walking (NM) recorded from 14 different camera positions, resulting in 28 gait sequences for each subject. The camera viewpoints are evenly distributed from  $0^\circ$  to  $90^\circ$  and  $180^\circ$  and  $270^\circ$ , with a 15-degree gap. Following [6, 10, 15, 23], we use the 5,153 subjects with an odd index between the 1-*st* and 10,305-*th* as training examples and the remaining 5,154 for inference, where the first sequence for each subject is the gallery set and the second as the probe.

**Implementation Details.** To extract the silhouette features, we follow the original settings of baseline methods [6, 10, 23, 15] for setting the hyperparameters for the model. For GaitPart [10], GaitSet [6] and GaitGL [23], we resize input gait sequence  $g$  to the size of  $64 \times 44$ . We use Adam optimizer with  $1e-4$  as the learning rate and 0.9 as the momentum. We set the margin in separate triplet loss as 0.2. Batch size is set to (8, 16) for CASIA-B, and (32, 16) for OUMVLP. We set the maximum iteration and weight decay following [6, 10, 23, 15]. For GLN [15], the initial gait se-

quence is sampled to  $128 \times 88$ . We use SGD with 0.1 as the initial learning rate and reduce it to  $\frac{1}{10}$  three times during training. The learning rate is reduced every 10,000 steps for CASIA-B and every 50,000 steps for OUMVLP.

For the body shape encoder at the gait branch, we apply the temporal shifting modules to MobileNet-v2 [33] and set  $n$  to 6 following [24] with the same learning rates and hyperparameters as the silhouette feature extraction model. We use CRD as our knowledge distillation method, where the ablation studies for other methods can be found in the ablation studies in the supplementary material.

To fuse the inferred body shape feature with the gait features from the silhouette feature encoder, we append the features before the last fully-connected layers for each backbone, since the features before the temporal or set pooling are the high-level feature representing the frame-level identity, and the inferred 3-D body shape representation can give additional guidance for identity encoding. For GaitPart, we append the 3-D body shape feature to all the part features to help each feature for a specific part understand the global body shape along with its local patterns. The input of fully-connected layers is set to the original size of the identity feature plus 10 (size of  $v_{bs}$ ) for each model [6, 10, 23, 15] after feature concatenation.

**RGB Data for Knowledge Distillation.** For 3-D human body prior extraction, we use the latent feature  $\beta$  in SMPL-X [25] model and normalize features in training set to  $(0, 0.1)$  gaussian distribution. To supervise the generation of body shape feature in the gait branch, we select 20% of sequences in the CASIA-B sequence for the data distillation and transfer. Since OUMVLP does not provide the RGB video sequences, we apply the body shape feature encoder for the gait branch pretrained on the CASIA-B subset and keep it frozen during training for feature extraction for all the examples in the OUMVLP dataset.

**Details for Identity Loss  $L_{ID}$ .** For the selection of identity loss function  $L_{ID}$ , we follow the implementation of each baseline method [6, 10, 15, 23]. For GaitSet-HBS, GaitPart-HBS and GLN-HBS, We use the triplet loss with its margin set to 0.2 as  $L_{ID}$ . For GaitGL-HBS, in addition to the triplet loss with the same margin, we use a cross-entropy loss for predicting the identity, which is represented as a one-hot vectors; weights for both losses are set to 1.

**Baseline Methods.** Since our method is an additional to the existing gait recognition methods, we compare with four state-of-the-art deep-learning gait recognition methods: GaitSet [6], GaitPart [10], GaitGL [23] and GLN [15]. We compare the baseline methods with and without inferred 3-D human body shape on both datasets. For ablation studies, we conduct our experiments on GaitGL [23] and GLN [15], since these are the two state-of-the-art methods for gait recognition. We exclude GaitView [5] and Gait3D [45] as they have extra supervision or additional in-

Probe	Method	Camera Positions											Mean
		0°	18°	36°	54°	72°	90°	108°	126°	144°	162°	180°	
NM #5-6	GaitSet [6]	91.1	98.0	<b>99.6</b>	97.8	95.4	93.8	95.7	97.5	98.1	97.0	88.2	95.6
	GaitPart [10]	94.0	98.7	99.3	98.8	94.8	92.6	96.4	98.3	99.0	97.4	91.2	96.4
	GLN [15]	93.8	98.5	99.2	98.0	95.2	92.9	95.4	98.5	99.0	99.2	91.9	96.5
	GaitGL [23]	95.3	97.9	99.0	97.8	96.1	95.3	97.2	<b>98.9</b>	99.4	98.8	<b>94.5</b>	97.3
	GaitSet-HBS	92.2	98.7	99.2	97.9	95.1	93.4	95.7	98.4	98.2	97.9	89.0	96.0
	GaitPart-HBS	93.2	<b>98.9</b>	99.4	<b>98.9</b>	95.1	91.9	96.5	98.8	<b>99.5</b>	98.4	91.7	96.6
	GLN-HBS	93.8	98.1	99.1	98.2	95.2	94.2	95.4	98.4	99.2	<b>99.4</b>	93.2	96.8
	GaitGL-HBS	<b>96.0</b>	98.3	99.2	97.8	<b>96.4</b>	<b>95.9</b>	<b>97.4</b>	98.7	99.2	98.7	<b>94.5</b>	<b>97.5</b>
BG #1-2	GaitSet [6]	87.0	93.8	94.6	92.9	88.2	83.0	86.6	92.6	95.7	92.9	83.4	90.1
	GaitPart [10]	89.5	94.5	95.3	93.5	88.5	83.9	89.0	93.6	96.0	94.1	85.3	91.2
	GLN [15]	92.2	95.6	96.7	94.3	91.8	87.8	91.4	95.1	96.3	95.7	87.2	93.1
	GaitGL [23]	<b>93.0</b>	95.7	97.0	<b>95.9</b>	93.3	<b>90.0</b>	91.9	96.8	97.5	96.9	90.7	94.4
	GaitSet-HBS	89.7	93.8	95.9	93.3	87.1	83.1	87.4	91.9	94.1	93.7	85.1	90.5
	GaitPart-HBS	90.1	93.6	95.7	94.4	89.9	85.8	89.9	94.0	96.0	92.7	86.4	91.7
	GLN-HBS	91.7	<b>96.6</b>	96.6	95.2	90.9	88.1	91.5	95.4	96.6	96.8	89.8	93.6
	GaitGL-HBS	<b>93.0</b>	96.0	<b>97.3</b>	<b>95.9</b>	<b>93.7</b>	89.5	<b>92.9</b>	<b>97.0</b>	<b>98.3</b>	<b>97.4</b>	<b>92.2</b>	<b>94.8</b>
CL #1-2	GaitSet [6]	71.0	82.6	84.0	80.0	71.7	69.1	72.1	76.7	78.5	77.2	63.4	75.1
	GaitPart [10]	72.5	82.8	86.0	82.2	79.5	71.0	77.7	80.8	82.9	81.4	67.7	78.6
	GLN [15]	78.5	90.4	90.3	85.1	80.2	75.8	78.1	81.8	80.9	83.2	<b>72.6</b>	81.5
	GaitGL [23]	71.7	<b>90.5</b>	<b>92.4</b>	89.4	<b>84.9</b>	<b>78.1</b>	83.1	<b>87.5</b>	89.1	83.9	67.4	83.5
	GaitSet-HBS	72.9	84.1	83.7	79.6	73.0	70.5	73.1	76.6	79.8	78.3	64.6	76.0
	GaitPart-HBS	75.9	84.8	86.5	84.6	77.4	74.4	78.6	82.4	83.5	80.5	67.6	79.7
	GLN-HBS	77.7	89.4	91.9	87.0	84.1	<b>78.1</b>	81.6	83.8	85.2	83.8	<b>72.6</b>	83.2
	GaitGL-HBS	<b>75.8</b>	<b>90.5</b>	92.3	<b>90.0</b>	84.0	77.9	<b>83.3</b>	87.3	<b>89.3</b>	<b>85.1</b>	69.8	<b>84.1</b>

Table 1. Gait recognition results on CASIA-B dataset, excluding identical-view cases.

Probe	Stats	HBS	Method				Avg. Change
			GaitSet [6]	GaitPart [10]	GLN [15]	GaitGL [23]	
NM #5-6	Mean (↑)	<b>X</b>	95.6	96.4	96.5	97.3	<b>+0.3</b>
		✓	96.0	96.6	96.8	97.5	
		Δ	+0.3	+0.2	+0.3	+0.2	
	STD. (↓)	<b>X</b>	3.4	2.8	2.8	1.7	<b>-0.1</b>
		✓	3.2	3.0	2.4	1.6	
		Δ	-0.2	+0.2	-0.4	-0.1	
BG #1-2	Mean (↑)	<b>X</b>	90.1	91.2	93.1	94.4	<b>+0.4</b>
		✓	90.5	91.7	93.6	94.8	
		Δ	+0.4	+0.5	+0.5	+0.4	
	STD. (↓)	<b>X</b>	4.6	4.2	3.3	2.7	<b>-0.3</b>
		✓	4.2	3.5	3.2	2.7	
		Δ	-0.4	-0.7	-0.1	0.0	
CL #1-2	Mean (↑)	<b>X</b>	75.1	78.6	81.5	83.5	<b>+1.1</b>
		✓	76.0	79.7	83.2	84.1	
		Δ	+0.9	+1.1	+1.7	+0.6	
	STD. (↓)	<b>X</b>	6.2	5.8	5.5	8.0	<b>-0.4</b>
		✓	5.9	5.6	5.5	7.0	
		Δ	-0.3	-0.2	0.0	-1.0	

Table 2. Statistics analysis for supervised results on CASIA-B dataset, excluding identical-view cases. (↑) indicates that larger values show better performance, while (↓) indicates that lower values are better. Δ indicates the change between the method with and without HBS.

put modality (framewise skeletons and body meshes) from RGB images. We also exclude earlier methods, such as [40, 35, 36], which not show state-of-the-art performance.

**Inference and Metrics.** We assess  $L_2$  similarity between features extracted from examples from gallery and probe sets, excluding the identical-view cases. We calculate the top-1 accuracies for finding the response with the

smallest  $L_2$  distance among the examples in the gallery to each example in the probe set.

## 4.2. Results and Analysis

In this subsection, we present the results and analysis on CASIA-B [43] and OUMVLP [37]. We further conduct ablations on CASIA-B for the selection of 3-D body shape

Probe	Method	Training Viewpoints						Test Viewpoints					Mean	Avg. Diff.	
		0°	18°	36°	54°	72°	90°	108°	126°	144°	162°	180°			
NM #5-6	GLN	(All Camera Positions)						95.4	98.5	99.0	99.2	91.9	96.8	+0.3	
	GLN-HBS	(All Camera Positions)						95.4	98.4	99.2	99.4	93.1	97.1		
	GLN	✓	✓	✓	✓	✓	✓	82.3	91.8	95.8	89.3	78.0	87.4	+1.3	
	GLN	✓		✓		✓		79.7	88.2	94.7	87.5	78.0	85.7		
	GLN	✓			✓			74.0	89.3	93.5	83.8	76.3	83.3		
	GLN-HBS	✓	✓	✓	✓	✓	✓	85.5	93.5	97.3	89.0	82.3	89.5		
	GLN-HBS	✓		✓		✓		82.0	88.5	93.5	91.8	77.5	86.7		
	GLN-HBS	✓			✓			77.2	89.5	94.5	83.2	76.5	84.2		
	GaitGL	(All Camera Positions)						97.2	98.9	99.4	98.8	94.5	97.8	-0.1	
	GaitGL-HBS	(All Camera Positions)						97.4	98.7	99.2	98.7	94.5	97.7		
	GaitGL	✓	✓	✓	✓	✓	✓	84.5	93.3	95.8	92.3	75.0	88.2	+1.4	
	GaitGL	✓		✓		✓		81.5	90.0	92.8	89.5	69.5	84.7		
	GaitGL	✓			✓			76.3	91.0	91.3	86.2	69.7	82.9		
	GaitGL-HBS	✓	✓	✓	✓	✓	✓	88.0	93.8	95.5	92.5	78.8	89.7		
	GaitGL-HBS	✓		✓		✓		82.8	90.5	93.0	89.7	71.7	85.6		
	GaitGL-HBS	✓			✓			79.0	92.0	91.7	88.5	71.5	84.6		
	BG #1-2	GLN	(All Camera Positions)						91.4	95.1	96.3	95.7	87.2	93.1	+0.9
		GLN-HBS	(All Camera Positions)						91.5	95.4	96.6	96.8	89.8	94.0	
GLN		✓	✓	✓	✓	✓	✓	72.0	83.0	87.3	80.1	75.0	79.5	+1.3	
GLN		✓		✓		✓		70.7	79.2	88.5	80.0	73.5	78.4		
GLN		✓			✓			65.0	81.5	86.5	79.6	65.5	75.6		
GLN-HBS		✓	✓	✓	✓	✓	✓	74.5	85.0	88.8	82.1	74.0	80.9		
GLN-HBS		✓		✓		✓		73.2	82.0	88.3	86.4	72.7	80.5		
GLN-HBS		✓			✓			69.5	81.5	86.5	77.3	65.2	76.0		
GaitGL		(All Camera Positions)						91.9	96.8	97.5	96.9	90.7	94.8	+0.8	
GaitGL-HBS		(All Camera Positions)						92.9	97.0	98.3	97.4	92.2	95.6		
GaitGL		✓	✓	✓	✓	✓	✓	74.3	83.8	90.0	88.1	69.3	81.1	+1.6	
GaitGL		✓		✓		✓		72.2	81.0	85.7	84.1	61.5	76.9		
GaitGL		✓			✓			64.7	82.7	86.5	78.5	66.8	75.9		
GaitGL-HBS		✓	✓	✓	✓	✓	✓	75.5	87.8	91.8	87.4	70.5	82.6		
GaitGL-HBS		✓		✓		✓		70.2	81.2	89.3	84.6	66.3	78.3		
GaitGL-HBS		✓			✓			70.8	83.2	87.2	80.8	67.0	77.8		
CL #1-2		GLN	(All Camera Positions)						78.1	81.8	80.9	83.2	72.6	79.3	+2.1
		GLN-HBS	(All Camera Positions)						81.6	83.8	85.2	83.8	72.6	81.4	
	GLN	✓	✓	✓	✓	✓	✓	57.3	60.0	67.0	56.0	46.3	57.3	+2.3	
	GLN	✓		✓		✓		50.0	62.5	67.5	58.5	44.8	56.5		
	GLN	✓			✓			45.0	54.7	59.3	52.0	44.5	51.1		
	GLN-HBS	✓	✓	✓	✓	✓	✓	57.8	62.5	68.3	61.5	46.8	59.4		
	GLN-HBS	✓		✓		✓		54.8	62.5	66.5	62.7	44.3	58.2		
	GLN-HBS	✓			✓			47.5	58.0	64.0	55.3	45.5	54.1		
	GaitGL	(All Camera Positions)						83.1	87.5	89.1	83.9	67.4	82.2	+0.8	
	GaitGL-HBS	(All Camera Positions)						83.3	87.3	89.3	85.1	69.8	83.0		
	GaitGL	✓	✓	✓	✓	✓	✓	58.8	68.5	73.3	66.8	44.0	62.3	+1.8	
	GaitGL	✓		✓		✓		53.2	63.7	71.2	63.5	41.0	58.5		
	GaitGL	✓			✓			48.5	62.0	64.2	51.5	43.3	53.9		
	GaitGL-HBS	✓	✓	✓	✓	✓	✓	61.5	70.8	76.0	72.3	47.3	65.6		
	GaitGL-HBS	✓		✓		✓		59.8	63.5	71.8	60.5	44.0	59.9		
	GaitGL-HBS	✓			✓			49.0	62.3	69.8	51.0	41.3	54.7		

Table 3. Gait recognition results for novel camera viewpoints on CASIA-B dataset. Viewpoints used for the training and inference stages are mutually exclusive. Supervised results, where all viewpoints are available for training, are shown at the top of each set.

features along with knowledge distillation and transfer.

**Results on CASIA-B.** We show the results for CASIA-B in Table 1. Methods ending with ‘HBS’, which is the abbreviation of **H**uman **B**ody **S**hape, are the ones with inferred 3-D human body features compared with baseline methods. In addition, we summarize the statistics for the performance on CASIA-B in Table 2, where we compare the models with and without features for the inferred human body. Mean and STD values in Table 2 refer to the average and standard de-

viation values of performance for different viewpoints for the same model. We have the following observations:

- Better performance.** Table 2 shows that the models with inferred human body shapes outperform the original ones on all four baselines for all three splits. For most of the viewpoints shown, the best performances among all models also appear in the model with the inferred 3-D body shape. With the knowledge of the boundary of the skinned human body model, gait

Method	Camera Positions														Mean
	0°	15°	30°	45°	60°	75°	90°	180°	195°	210°	225°	240°	255°	270°	
GEINet [35]	23.2	38.1	48.0	51.8	47.5	48.1	43.8	27.3	37.9	46.8	49.9	45.9	45.7	41.0	42.5
GaitSet [6]	79.2	87.7	89.9	90.1	87.9	88.6	87.7	81.7	86.4	89.0	89.2	87.2	87.7	86.2	87.0
GaitPart [10]	82.8	89.2	90.9	91.0	89.7	89.9	89.3	85.1	87.7	90.0	90.1	89.0	89.0	88.1	88.7
GaitGL [23]	84.2	89.8	91.3	<b>91.7</b>	90.8	<b>91.0</b>	90.4	88.1	88.2	<b>90.5</b>	<b>90.5</b>	89.5	<b>89.7</b>	88.8	89.6
GaitSet-HBS	79.0	87.9	90.4	90.6	88.4	89.2	88.4	82.3	87.1	89.6	89.6	87.7	88.4	86.9	87.5
GaitPart-HBS	82.4	89.1	91.1	91.3	89.8	90.2	89.7	84.8	88.0	90.3	90.3	89.2	89.4	88.4	88.9
GaitGL-HBS	<b>84.7</b>	<b>90.2</b>	<b>91.4</b>	<b>91.7</b>	<b>90.9</b>	<b>91.0</b>	<b>90.5</b>	<b>88.4</b>	<b>88.7</b>	<b>90.5</b>	<b>90.5</b>	<b>89.6</b>	89.6	<b>88.9</b>	<b>89.8</b>

Table 4. Gait recognition results on OUMVLP dataset, excluding identical-view cases.

recognition models are capable of focusing on the motions instead of the appearances in 2-D silhouettes.

- Stability at different viewpoints.** In addition to the average performance for all camera viewpoints, we observe the standard deviations for the accuracies at different viewpoints reduce after using inferred human body shapes. Even for those models with no improvement on the mean value, e.g., GaitPart-HBS compared with GaitPart on the NM split, the standard deviation still reduces. With the inferred 3-D body shape, consistent for all camera positions, models can show additional robustness to the camera viewpoints and have more stable performances.
- Different appearance variances.** BG and CL sets have higher average accuracy than NM, whose gait appearances are similar. In BG and CL sets, the silhouette sequence individual is carrying different bags or wearing different outfits, affecting the binarized silhouette. Focusing on appearance differences hurts the gait recognition model. Since inferred 3-D human body shapes are skinned models, they are stable and resilient to these fluctuations. Gait recognition models may detect the consistent body shapes and reduce non-human body content, exhibiting benefits.

**Zero-shot Results for Novel Viewpoints.** In addition to the results on existing viewpoints, we assess the model on the novel viewpoints on CASIA-B dataset in Table 3 with GaitGL and GLN, the two of the best performing baselines. Instead of using silhouette sequence from all the viewpoints for both training and inference, we only use part of the viewpoints for training, and viewpoints used for training and inference are mutually exclusive. We notice that when gait recognition models encounter novel viewpoints not seen before, using the inferred human body shape gives a consistent improvement compared with the baseline methods. Although these novel camera positions are unavailable during training, the consistency of the 3-D human body shape helps gait recognition models extract motion information from a new camera position for identification.

We further reduce the number of available viewpoints during training to assess the robustness of gait recognition

models learning from fewer examples. With fewer viewpoints available in the training set, performances for all the methods are decreasing. However, GaitGL [23] and GLN [15] with inferred 3-D human body shape still show a consistent improvement compared to the model without body shapes, showing the 3-D body shape can give consistent guidance at different amounts of data.

**Results on OUMVLP.** We show the results for the OUMVLP dataset in Table 4. Since the OUMVLP dataset does not provide the original RGB frames, we apply the knowledge distillation model pretrained on the training set of CASIA-B to infer human body shape directly from the silhouette sequences. Compared to baseline methods, inferring 3-D body shape for gait recognition consistently outperforms original methods, showing good generalization ability and robustness of body shape feature encoders across different datasets. Examples in OUMVLP are all normal walking with fewer variations, which explains the limited improvement as NM sets for CASIA-B.

## 5. Conclusion

In this paper, we propose the exploitation of inferring 3-D body shape from gait sequence to disentangle gait motion from appearance variances of 2-D images. In addition to the gait pattern analysis, we distill the 3-D body shape features from selected RGB frames and transfer them to gait sequences via feature exchanging between neighbor frames. We assess our method with four state-of-the-art gait recognition methods and show better results on two public datasets at both seen and novel camera viewpoints.

## Acknowledgement

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.

## References

- [1] Sungsoo Ahn, Shell Xu Hu, Andreas Damianou, Neil D Lawrence, and Zhenwen Dai. Variational information distillation for knowledge transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9163–9171, 2019.
- [2] Jimmy Ba and Rich Caruana. Do deep nets really need to be deep? *Advances in neural information processing systems*, 27, 2014.
- [3] Federica Bogo, Angjoo Kanazawa, Christoph Lassner, Peter Gehler, Javier Romero, and Michael J Black. Keep it smpl: Automatic estimation of 3d human pose and shape from a single image. In *European conference on computer vision*, pages 561–578. Springer, 2016.
- [4] Zhe Cao, Gines Hidalgo, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Openpose: realtime multi-person 2d pose estimation using part affinity fields. *IEEE transactions on pattern analysis and machine intelligence*, 43(1):172–186, 2019.
- [5] Tianrui Chai, Xinyu Mei, Annan Li, and Yunhong Wang. Silhouette-based view-embeddings for gait recognition under multiple views. In *2021 IEEE International Conference on Image Processing (ICIP)*, pages 2319–2323. IEEE, 2021.
- [6] Hanqing Chao, Yiwei He, Junping Zhang, and Jianfeng Feng. Gaitset: Regarding gait as a set for cross-view gait recognition. In *Proceedings of the AAAI conference on artificial intelligence*, pages 8126–8133, 2019.
- [7] Yevgen Chebotar and Austin Waters. Distilling knowledge from ensembles of neural networks for speech recognition. In *Interspeech*, pages 3439–3443, 2016.
- [8] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4690–4699, 2019.
- [9] Chao Fan, Saihui Hou, Jilong Wang, Yongzhen Huang, and Shiqi Yu. Learning gait representation from massive unlabelled walking videos: A benchmark. *arXiv preprint arXiv:2206.13964*, 2022.
- [10] 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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14225–14233, 2020.
- [11] Tong He, Zhi Zhang, Hang Zhang, Zhongyue Zhang, Junyuan Xie, and Mu Li. Bag of tricks for image classification with convolutional neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 558–567, 2019.
- [12] Yiwei He, Junping Zhang, Hongming Shan, and Liang Wang. Multi-task gans for view-specific feature learning in gait recognition. *IEEE Transactions on Information Forensics and Security*, 14(1):102–113, 2018.
- [13] Byeongho Heo, Minsik Lee, Sangdoo Yun, and Jin Young Choi. Knowledge transfer via distillation of activation boundaries formed by hidden neurons. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 3779–3787, 2019.
- [14] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.
- [15] Saihui Hou, Chunshui Cao, Xu Liu, and Yongzhen Huang. Gait lateral network: Learning discriminative and compact representations for gait recognition. In *European Conference on Computer Vision*, pages 382–398. Springer, 2020.
- [16] Pengpeng Hu, Edmond Shu-Lim Ho, and Adrian Munteanu. 3dbodynet: fast reconstruction of 3d animatable human body shape from a single commodity depth camera. *IEEE Transactions on Multimedia*, 24:2139–2149, 2021.
- [17] Zehao Huang and Naiyan Wang. Like what you like: Knowledge distill via neuron selectivity transfer. *arXiv preprint arXiv:1707.01219*, 2017.
- [18] Lei Jin, Xiaojuan Wang, Xuecheng Nie, Luoqi Liu, Yandong Guo, and Jian Zhao. Grouping by center: Predicting centripetal offsets for the bottom-up human pose estimation. *IEEE Transactions on Multimedia*, 2022.
- [19] Jangho Kim, SeongUk Park, and Nojun Kwak. Paraphrasing complex network: Network compression via factor transfer. *arXiv preprint arXiv:1802.04977*, 2018.
- [20] Xiang Li, Yasushi Makihara, Chi Xu, Yasushi Yagi, Shiqi Yu, and Mingwu Ren. End-to-end model-based gait recognition. In *Proceedings of the Asian conference on computer vision*, 2020.
- [21] 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.
- [22] Rijun Liao, Shiqi Yu, Weizhi An, and Yongzhen Huang. A model-based gait recognition method with body pose and human prior knowledge. *PR*, 2020.
- [23] Beibei Lin, Shunli Zhang, and Xin Yu. Gait recognition via effective global-local feature representation and local temporal aggregation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14648–14656, 2021.
- [24] Ji Lin, Chuang Gan, and Song Han. Tsm: Temporal shift module for efficient video understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7083–7093, 2019.
- [25] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. Smpl: A skinned multi-person linear model. *ACM transactions on graphics (TOG)*, 34(6):1–16, 2015.
- [26] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. In *European conference on computer vision*, pages 405–421. Springer, 2020.
- [27] Wonpyo Park, Dongju Kim, Yan Lu, and Minsu Cho. Relational knowledge distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3967–3976, 2019.

- [28] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed A. Osman, Dimitrios Tzionas, and Michael J. Black. Expressive body capture: 3d hands, face, and body from a single image. In *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2019.
- [29] Sida Peng, Yuanqing Zhang, Yinghao Xu, Qianqian Wang, Qing Shuai, Hujun Bao, and Xiaowei Zhou. Neural body: Implicit neural representations with structured latent codes for novel view synthesis of dynamic humans. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9054–9063, 2021.
- [30] Adriana Romero, Nicolas Ballas, Samira Ebrahimi Kahou, Antoine Chassang, Carlo Gatta, and Yoshua Bengio. Fitnets: Hints for thin deep nets. *arXiv preprint arXiv:1412.6550*, 2014.
- [31] Shunsuke Saito, Zeng Huang, Ryota Natsume, Shigeo Morishima, Angjoo Kanazawa, and Hao Li. Pifu: Pixel-aligned implicit function for high-resolution clothed human digitization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2304–2314, 2019.
- [32] Shunsuke Saito, Tomas Simon, Jason Saragih, and Hanbyul Joo. Pifuhd: Multi-level pixel-aligned implicit function for high-resolution 3d human digitization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 84–93, 2020.
- [33] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4510–4520, 2018.
- [34] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 815–823, 2015.
- [35] Kohei Shiraga, Yasushi Makihara, Daigo Muramatsu, Tomio Echigo, and Yasushi Yagi. Geinet: View-invariant gait recognition using a convolutional neural network. In *2016 international conference on biometrics (ICB)*, pages 1–8. IEEE, 2016.
- [36] Chunfeng Song, Yongzhen Huang, Yan Huang, Ning Jia, and Liang Wang. Gaitnet: An end-to-end network for gait based human identification. *Pattern recognition*, 96:106988, 2019.
- [37] 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. *IPSN Transactions on Computer Vision and Applications*, 10(1):1–14, 2018.
- [38] Torben Teepe, Ali Khan, Johannes Gilg, Fabian Herzog, Stefan Hörmann, and Gerhard Rigoll. Gaitgraph: Graph convolutional network for skeleton-based gait recognition. *arXiv preprint arXiv:2101.11228*, 2021.
- [39] Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive representation distillation. *arXiv preprint arXiv:1910.10699*, 2019.
- [40] Zifeng Wu, Yongzhen Huang, Liang Wang, Xiaogang Wang, and Tieniu Tan. A comprehensive study on cross-view gait based human identification with deep cnns. *IEEE transactions on pattern analysis and machine intelligence*, 39(2):209–226, 2016.
- [41] Junho Yim, Donggyu Joo, Jihoon Bae, and Junmo Kim. A gift from knowledge distillation: Fast optimization, network minimization and transfer learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4133–4141, 2017.
- [42] Shiqi Yu, Haifeng Chen, Qing Wang, Linlin Shen, and Yongzhen Huang. Invariant feature extraction for gait recognition using only one uniform model. *Neurocomputing*, 239:81–93, 2017.
- [43] Shiqi Yu, Daoliang Tan, and Tieniu Tan. A framework for evaluating the effect of view angle, clothing and carrying condition on gait recognition. In *18th International Conference on Pattern Recognition (ICPR'06)*, volume 4, pages 441–444. IEEE, 2006.
- [44] Sergey Zagoruyko and Nikos Komodakis. Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer. *arXiv preprint arXiv:1612.03928*, 2016.
- [45] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20228–20237, 2022.
- [46] Zheng Zhu, Xianda Guo, Tian Yang, Junjie Huang, Jiankang Deng, Guan Huang, Dalong Du, Jiwen Lu, and Jie Zhou. Gait recognition in the wild: A benchmark. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 14789–14799, 2021.

# Gait Recognition Using 3-D Human Body Shape Inference

## Supplementary Material

In this supplementary document, we present some further experimental details and results that could not fit in the main paper. We discuss the motivation and details for the new setting for the CASIA-B dataset with novel camera viewpoints as further experiment details, followed by some experimental details and additional ablation studies for hyperparameters we choose in the main paper; these include the balancing term  $\lambda$  in the final loss function and the ratio of feature exchange in the temporal shift operation. We then show some visualization results for the inferred body shapes directly from silhouette compared with the reconstruction results by SMPLify-X [28] for selected RGB frames.

### A. Experiment Details

**Discussion for the novel view settings.** In addition to the original CASIA-B setting in which the training and test set share the same viewpoints, the new setting of CASIA-B only includes 2 to 6 viewpoints in the training set, while we evaluate the model on the test viewpoints of the remaining five camera viewpoints,  $108^\circ$ ,  $126^\circ$ ,  $144^\circ$ ,  $162^\circ$ , and  $180^\circ$ . In a real-world instance of silhouettes taken by a camera, the camera’s perspective can come from any direction, which is the primary purpose of introducing this new setting. Compared to the original setting, our setting is more suitable for evaluating the generalization capacity of the gait recognition model when meeting novel camera viewpoints.

**Variations for Silhouette Feature Encoder.** In the experiment, we choose four methods as our silhouette feature encoder: GaitSet, GaitPart, GaitGL and GLN. *GaitSet* [6] uses the frame sequence in the gait video as a *set* of independent frames. By using set processing methods, such as set pooling, GaitSet can extract set-level features for preserving spatial and temporal information. *GaitPart* [10] introduces split the gait image into four different parts and assess the motion pattern for each part separately to focus on more local movements. *GLN* [15] learns both discriminative and compact representations from the silhouettes. It extracts both silhouette-level and set-level features from different stages for gait recognition. *GaitGL* [23] applies the features from both global and local patterns by using both global visual information and local region details.

### B. Ablation studies

In this subsection, we discuss five different ablation studies for the composition of our model, including the choice

of balancing term  $\lambda_2$ , the model we use for human body shape reconstruction from selected RGB images, knowledge distillation function  $L_{KD}$  for knowledge transfer between two modalities, fusion method for backpropagating body shape feature from single image frames to silhouette sequence, and the ablation for feature exchange between neighbor frames.

**Ablations for the Balancing Term  $\lambda_2$ .** To balance the identity loss  $L_{ID}$  and knowledge distillation loss  $L_{KD}$ , we set the balancing term follow the ablations on CASIA-B [43] for all three splits, NM, CL and BG, with GLN-HBS and GaitGL-HBS for some other variations of  $\lambda_2$ . We show the results in Table 7, where top-1 accuracy is reported excluding identical-view cases. We note that when we have the balancing term  $\lambda_2$  set to 1, GLN-HBS and GaitGL-HBS both show the best performance. With  $\lambda_2$  as 1, our model can find a balancing point between distinguishing different identities from silhouette sequences and transferring knowledge from inferred 3-D body shape from selected RGB frames by SMPLify-X [28].

**Body Prior Reconstruction.** Since we need a strong human body prior to help disentangle the skinned body shape from appearance variances, to reconstruct human body prior from RGB frames, we compare the methods two skinned models, SMPL [25] from SMPLify [3] and SMPL-X [28] from SMPLify-X [28], for 3-D body reconstruction. Compared with SMPL-X, SMPL does not require the output for human skeletons extracted by OpenPose [4]. We assess both methods on the CASIA-B dataset for three settings with GLN. For SMPLify, the average accuracies are 96.7, 93.4 and 82.6 for NM, BG and CL, respectively, while for SMPLify-X, the average accuracies are 96.7, 93.6 and 83.2. Although SMPL shows some improvement compared with GLN without 3-D human body shape, the inaccurate reconstructions from SMPLify make the network unable to distinguish between body shapes and appearance variances, making it unable to beat SMPLify-X reconstructions.

**Knowledge Distillation.** We show the results for different knowledge distillation methods [27, 2, 30, 39, 17], in addition to the experiment directly using the feature output from the teacher network, in Table 5. Since GLN and GaitGL are the two state-of-the-art methods with the best performance in Table 1, we compare several knowledge distillation methods on all three variations of the CASIA-B dataset for GLN and GaitGL with SMPLify-X as the 3-D human body shape reconstruction model for RGB images.

Knowledge Distillation Function $L_{KD}$	NM #5-6		BG #1-2		CL #1-2	
	GLN [15]	GaitGL [23]	GLN [15]	GaitGL [23]	GLN [15]	GaitGL [23]
Origin Method	96.5	97.3	93.1	94.4	81.5	83.5
+ RGB Body Prior	96.7	97.5	93.5	95.0	83.3	84.4
+ RKD [27]	96.1	97.0	92.9	94.0	82.2	83.6
+ Hint [30]	96.8	97.4	93.3	94.4	83.1	84.0
+ $L_2$ [2]	96.7	96.9	93.2	94.1	82.9	84.0
+ NST [17]	96.8	97.2	93.3	94.4	82.8	84.1
+ CRD [39]	96.8	97.5	93.6	94.9	83.3	84.3

Table 5. Ablation results for different knowledge distillation methods. Results are reported in mean accuracies on CASIA-B. ‘RGB body prior’ indicates features used are directly encoded from the teacher model, SMPLify-X [28] for selected RGB frames.

Fusion Methods	NM #5-6		BG #1-2		CL #1-2	
	GLN [15]	GaitGL [23]	GLN [15]	GaitGL [23]	GLN [15]	GaitGL [23]
Origin Method	96.5	97.3	93.1	94.4	81.5	83.5
+ MaxPool	95.0	95.9	92.2	92.6	79.3	81.0
+ AvgPool	96.4	97.2	93.0	94.4	82.6	83.7
+ RNN	96.5	97.2	93.0	94.3	82.1	83.6
+ LSTM	96.4	97.3	93.4	94.6	82.9	84.0
+ GRU	96.7	97.5	93.3	94.6	83.0	83.9
+ TS	96.8	97.7	93.6	94.8	83.2	84.1

Table 6. Ablation results for different feature fusion methods for propagating inferred human body shape feature from RGB images to gait sequence on CASIA-B. TS represents temporal shifting. MaxPool and AvgPool are max pooling and average pooling respectively. Results are reported in mean accuracies.

Among all the knowledge distillation methods, CRD shows the best performance, and we choose to use CRD as our  $L_{KD}$  for features of 3-D body shape transfer from RGB frame  $s_r$  to gait sequence  $g$ . In addition, we also note from the table that using the distilled feature from CRD is comparable to the body prior directly extracted from selected RGB frames by the teacher network, SMPLify-X [28], and even better at some splits. With knowledge distillation, body shape from gait sequence can be more stable than using a single RGB image for reconstruction.

**Fusion.** In addition to the method selection for knowledge distillation, we further show different methods for propagating the single frame RGB features to gait sequences in Table 6. We assess different fusion methods on CASIA-B using CRD for knowledge distillation and transfer. In addition to the temporal shifting, annotated as TS in the table, we assess two pooling and three RNN variations. We note that the max-pooling results are worse than the original methods, indicating that the model starts overfitting on a few frames. Compared with average pooling and three RNN variations, temporal shifting introduces the most significant improvement. The ability to propagate single frame information back to all frames and exchange the features between nearby frames introduce more stability and consistency for knowledge transfer.

**Ablation for the Ratio of feature exchange.** To tempo-

rally shift the features extracted from the body shape feature encoder in the gait feature extraction branch, we follow [24] to set the ratio of feature exchange to 12.5%. This number indicates that we use 75% of features from the current frame, 12.5% from future frames, and 12.5% from the previous frame for the next step’s convolution operation. We further research several different ratios of feature exchange in Table 8. We note that when we exchange 12.5%, following [24], as what we did in the main paper, our models show the best performance. When we increase the exchange ratio to 33.3%, the feature from the current frame is the same amount as the feature from the previous and next frames. At this ratio, the model cannot extract enough information from the current frame to identify the person in the sequence. When we set the exchange ratio as 0%, the model degenerates to the average pooling case, where no features are exchanged for temporal fusion before the average pooling layer.

### C. Visualizations for Inferred Body Shapes.

We visualize some reconstructions of human body shapes to assess the quality of inferred body shape  $v_{bs}$  from silhouette sequences. We convert  $v_{bs}$  to the form of the body shape feature  $\beta$  used by the skinned human body reconstruction model SMPL-X [28] in the reverse way that we normalize it. Since we do not predict human poses  $\theta$  from

Balancing Term $\lambda_2$	NM #5-6		BG #1-2		CL #1-2	
	GLN-HBS	GaitGL-HBS	GLN-HBS	GaitGL-HBS	GLN-HBS	GaitGL-HBS
0.5	96.6	97.5	93.4	94.6	82.8	84.0
1	96.8	97.7	93.6	94.8	83.2	84.1
2	96.6	97.4	93.5	94.8	82.6	83.9
5	96.2	97.2	92.9	94.4	81.6	83.2

Table 7. Ablation results for different  $\lambda_2$  used for balancing  $L_{KD}$  and  $L_{ID}$ .

Exchange Ratio	NM #5-6		BG #1-2		CL #1-2	
	GLN-HBS	GaitGL-HBS	GLN-HBS	GaitGL-HBS	GLN-HBS	GaitGL-HBS
0%	96.4	97.2	93.0	94.4	82.6	83.7
10%	96.7	97.7	93.5	94.8	83.2	83.9
12.5%	96.8	97.7	93.6	94.8	83.2	84.1
25%	96.5	97.2	93.0	94.4	81.9	83.1
33.3%	95.7	96.8	92.6	93.5	81.2	82.9

Table 8. Ablations for ratio used for feature exchange in the body shape feature encoder.



(a) Incomplete cases

(b) Boundary cases

Figure 4. Sampled silhouette visualization for error prediction.

silhouette with our model, we plot body shapes as T-poses for all reconstructions. We choose two examples in the test set of CASIA-B [43] with all three variants. To assess the stability among different camera positions, we select four camera positions for each subject:  $0^\circ$ ,  $36^\circ$ ,  $72^\circ$  and  $108^\circ$ .

We show the visualizations of inferred body shapes in Fig. 5, along with one of the silhouettes sampled at each camera viewpoint. We note that reconstructions from both methods, SMPLify-X [28] and our body shape feature encoder, are pretty accurate for reconstructing human body shapes in the selected frames or silhouettes. For example, the first person is broader than the second, which can be reflected in most reconstructed meshes. In addition, both reconstructed shapes show good robustness again different appearance variations and different viewpoints, while shapes reconstructed from silhouette sequences by our body shape feature encoder are more consistent for the same person. Compared with a single frame of selected RGB images, a sequence input gives more information for reconstructing the human body shape and is more precise in describing the shape using information from neighbor frames.

## D. Limitation and Error Analysis

To distill and transfer knowledge from limited RGB images to the body shape feature encoder of the gait branch, we use SMPLify-X [28] as our body prior extraction model for providing body shapes. The quality of the generated body prior from SMPLify-X is important. Although the distillation network is able to correct some mistakes generated from SMPLify-X as Figure 5, if there are too many mistakes from SMPLify-X, the distillation model will be unable to generate any useful body shapes for the training of body shape encoder in the gait branch.

During inference, our model has only one input, silhouette sequences. We note that the incomplete gait images, either from bad segmentation results or the person walking to the boundary of the image, as shown in Figure 4, increase the probability of error prediction. When these incomplete silhouette images take a relatively large part of the video, the model is more likely to give wrong predictions since the silhouette is the only modality we have during inference.

Camera Viewpoints	NM		CL		BG	
	RGB	silhouette	RGB	silhouette	RGB	silhouette

Figure 5. Visualizations for reconstructed human body shapes of two identities from selected RGB frames and silhouettes in the CASIA-B test set. For each example, the camera position from top to down is  $0^\circ$ ,  $36^\circ$ ,  $72^\circ$  and  $108^\circ$  respectively. We align the camera position to the front view for all variations and plot T-pose shapes for each person with the  $\beta$  we inferred from the human body shape encoder. ‘RGB’ and ‘silhouette’ represent the reconstruction is from the branch with selected RGB images (SMPLify-X [28]) or the gait feature extraction branch (Body Shape Feature Encoder). Silhouettes shown in the first column only indicate the IDs of the people and camera viewpoints, which are not the sequences used for body shape reconstruction.