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 λ 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°, 126°, 144°, 162°, and 180°. 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 sihouette 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 λ_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 λ_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 λ_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 λ_2 set to 1, GLN-HBS and GaitGL-HBS both show the best performance. With λ_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	NM #5-6		BG #1-2		CL #1-2	
Function L_{KD}	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 temporally 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.

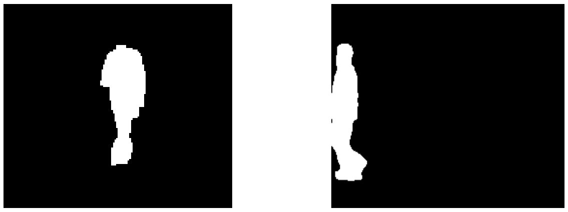
We visualize some reconstructions of human body shapes to assess the quality of inferred body shape v_{bs} from silhouette sequences. We convert v_{bs} to the form of the body shape feature β used by the skinned human body reconstruction model SMPL-X [28] in the reverse way that we normalize it. Since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

Balancing	NM #5-6		BG #1-2		CL #1-2	
Term λ_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 λ_2 used for balancing L_{KD} and L_{ID}.

Exchange	NM #5-6		BG #1-2		CL #1-2	
Ratio	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°, 36°, 72° and 108°.

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.

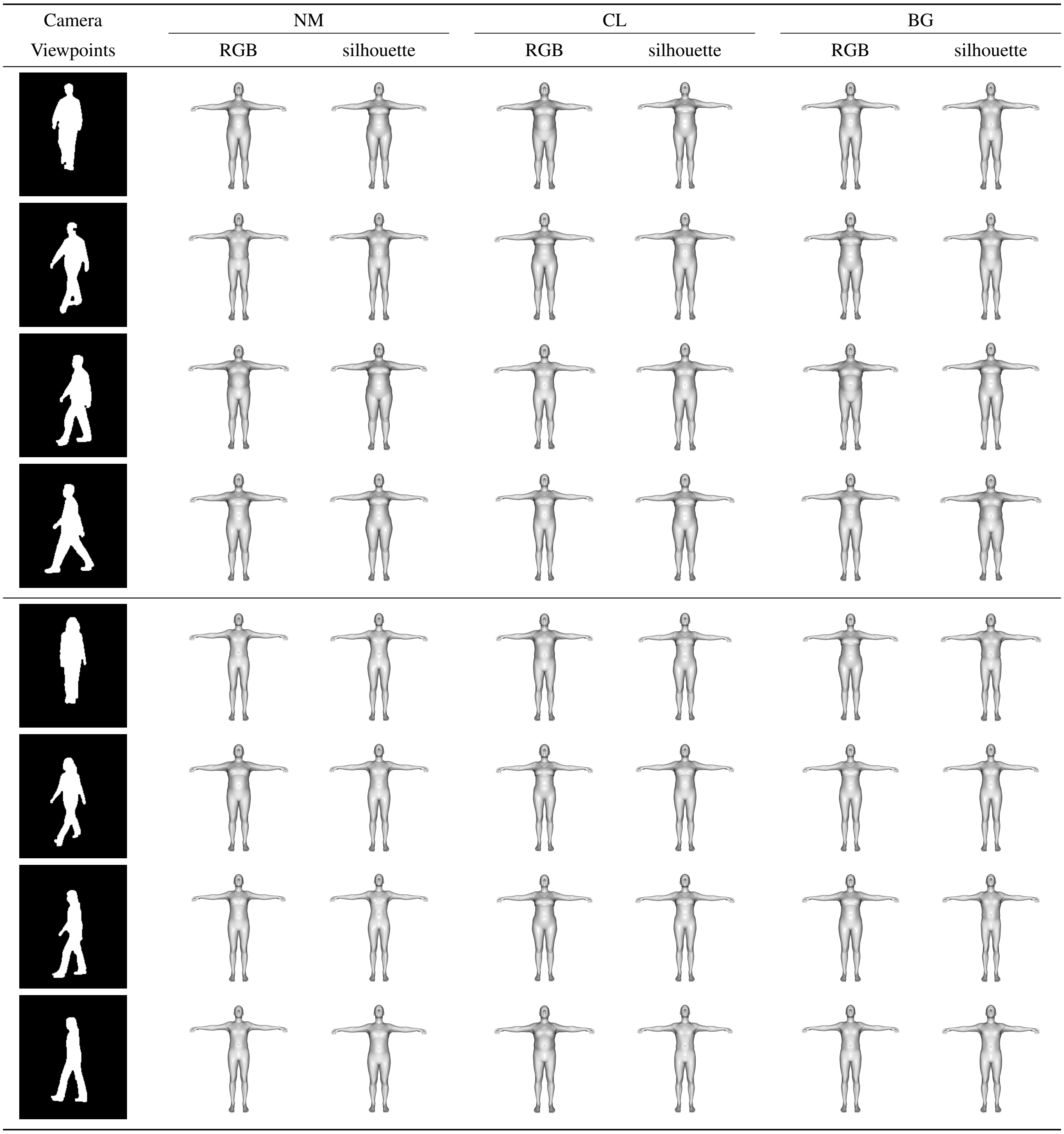


Figure 5. Visualizations for reconstructed human body shapes of two identities from selected RGB frames and silhouettes in the CASIA-B test set. For each example, the camera position from top to down is 0°, 36°, 72° and 108° respectively. We align the camera position to the front view for all variations and plot T-pose shapes for each person with the β 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.