MimicGait: A Model Agnostic approach for Occluded Gait Recognition using Correlational Knowledge Distillation

Ayush Gupta
Johns Hopkins University
3400 N. Charles St
agupt120@jh.edu

Rama Chellappa
Johns Hopkins University
3400 N. Charles St
rchella4@jhu.edu

Abstract

Gait recognition is an important biometric technique over large distances. State-of-the-art gait recognition systems perform very well in controlled environments at close range. Recently, there has been an increased interest in gait recognition in the wild prompted by the collection of outdoor, more challenging datasets containing variations in terms of illumination, pitch angles and distances. An important problem in these environments is that of occlusion, where the subject is partially blocked from camera view. While important, this problem has received little attention. Thus, we propose ***MimicGait****, a model-agnostic approach for gait recognition in the presence of occlusions. We train the network using a multi-instance correlational distillation loss to capture both inter-sequence and intra-sequence correlations in the occluded gait patterns of a subject, utilizing an auxiliary Visibility Estimation Network to guide the training of the proposed mimic network. We demonstrate the effectiveness of our approach on challenging real-world datasets like GREW, Gait3D and BRIAR. We release the code in https://github.com/Ayush-00/mimicgait.*

1. Introduction

Gait is an important biometric feature which can be used for identifying humans [28], especially when the face is not visible. There has been significant progress in the field of Gait Recognition - the problem of recognizing subjects based on their walking pattern. Gait recognition can be performed by placing wearable sensors on subjects [10, 17], however, such methods require the subjects' cooperation and are not scalable. With progress in computer vision, the popularity of gait recognition techniques using only vision-based modalities has risen significantly. As a result, gait has gained a unique importance among all biometric signatures as one of the few identifying characteristics in humans that can be captured effectively at a distance.

There has been significant progress in the field of vision-based gait recognition [32], with some methods achieving almost perfect scores on indoor controlled datasets [5, 23]. With saturation in controlled scenarios, there has been an increased focus on outdoor, in the wild scenarios [2, 44]. Such datasets pose much bigger challenges to gait recognition - due to large variations in viewpoint, altitude, clothing, background, illumination changes and occlusion.

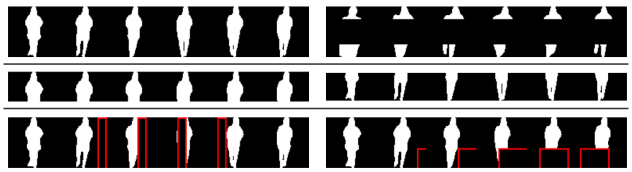

Figure 1. Visualizations of the synthetic occlusions used in our experiments, taken from the GREW dataset. The original holistic video is shown in the top left. Middle occlusions are shown in the top right. The second row shows the same video with consistent synthetic occlusions. The bottom most row shows the same video with dynamic occlusions. The boundary of the moving occlusion patch is shown in red for visualization purposes only.

A deployed gait recognition system should be able to handle occlusion scenarios. There can be many types of occlusions; arising from an obstruction between the camera and the subject, or due to improper camera placement. Occlusion can be consistent, such as an elevated sidewalk blocking a subject's feet for the entire sequence, or dynamic, such as another person or a stationary object temporarily blocking the subject of interest from view. With gait being recognized as a viable option for biometrics, it is important to address occlusion within gait recognition.

Most current work on gait recognition does not address this problem specifically; the lack of a large-scale dataset focused on occlusion has resulted in slow progress in the area. Some works that focus on this problem simulate occlusions on indoor datasets [36] or work with small datasets [33]. Moreover, most current approaches assume an ideal occlusion scenario, where the subject is close to the camera

and clearly visible. In a more practical in-the-wild scenario, the subject could be hundreds of meters away, and the camera may be situated at an altitude, and it is not easy to extend these approaches to such unconstrained data.

[34] uses a generative network to synthesize complete gait sequences in the case of occlusion. However, it will be difficult to generate such sequences when the partial input is itself of low quality. Similarly, [36] uses an SMPL-based human mesh model to construct the gait signature, but the 3D structure of the body is not easy to recover from noisy data collected at a distance of several hundred metres, that too in the presence of occlusions. [13] generates occlusion aware features and inserts them inside the gait recognition network. However, it is limited by the assumption that the network can independently learn discriminative features through an occlusion detector, neglecting the potential correlations between occluded and visible body parts.

To address these challenges, we propose *MimicGait*, a *model-agnostic* approach to generate discriminative gait features for subjects at range under occlusion using correlational knowledge distillation. We assume that temporal patterns which exist in the occluded sections of the subject are correlated with the observable motion in the gait sequence. We adopt a knowledge distillation approach to learn these correlations among the occluded and visible parts of the body, enabling the network to produce features closer to the ideal, holistic features. Building upon the work by [13], we also utilize a Visibility Estimation Network (VEN) to introduce occlusion-relevant features into our method to enhance the prediction of these missing correlations. Our approach is model-agnostic and can be used to increase the performance of any state-of-the-art gait recognition method to extend it to occlusions. Lastly, being a vision-based method, it can work with noisy data captured at a distance. We demonstrate the performance of our proposed method on the publicly available GREW [44] and Gait3D [42] datasets. Additionally, we also evaluate our approach on the BRIAR [2] dataset, which includes variations in range, altitude, clothing and walking conditions.

We also introduce some practical evaluation schemes for occluded gait recognition. We formalize the concept of *Generalizability*, introduced in [13], relating model performance on unseen occlusions. We further introduce a concept of *Adaptability*, relating to how well the model can be trained/adapted to newer occlusions. This becomes useful where a certain type of occlusion is expected to occur frequently in deployment. Lastly, we propose another metric for occluded gait recognition called *relative performance*, RP, which measures the occluded recognition performance of the model relative to performance on ideal, holistic data. We show that it is a better metric to evaluate occluded performance. Together with generalizability, adaptability and RP, we perform an extensive analysis of our method and show that it outperforms previous works.

In summary, our main contributions are

- We propose **MimicGait**, a novel model-agnostic approach to generate robust gait features under occlusions in various conditions, viewpoints, and distances.
- We demonstrate the utility of a *multi-instance correlational knowledge distillation* approach to learn the correlations between occluded and visible motion patterns across multiple gait instances of a subject.
- We improve upon the auxiliary occlusion detector proposed in previous works and propose a *Visibility Estimation Network* to enhance gait recognition performance under occlusions.
- We introduce the concepts of *generalizability, adaptability,* and a new metric *RP* for evaluating occluded gait recognition performance. Across these benchmarks, our approach outperforms other works on GREW, Gait3D, and the challenging real-world BRIAR dataset.

2. Related Work

2.1. Gait Recognition

Traditionally, gait recognition was performed using wearable motion sensors [10, 24]. With advances in computer vision techniques, gait recognition has become an attractive method for human identification at a distance [8]. These techniques can be classified into two categories based on the data modality; 1) Skeleton-based and 2) Vision-based. Skeleton-based gait recognition systems [6,9,14,21,22,40] first use pose estimation techniques [3] to extract the joints/keypoints from the input image or video. This introduces a bottleneck in the form of the pose estimation method. Vision-based gait recognition systems usually operate on silhouettes [4, 5, 7, 23, 44]. Some progress has been made to utilize the RGB modality [41] for gait recognition. However, RGB videos have a lot of irrelevant information like background, texture, and color. To account for this, works like [19,20] adopt an end-to-end approach while learning silhouettes.

Performance of gait recognition methods has almost saturated [5, 23] on indoor controlled datasets like CASIA-B [39]. To further advance research in this field, more challenging real-world datasets like GREW [44], Gait3D [42] and BRIAR [2] have been collected with large variations in viewpoints, altitude, and other conditions. To overcome the new challenges posed by such outdoor datasets, works like [12] use multiple modalities for the recognition task.

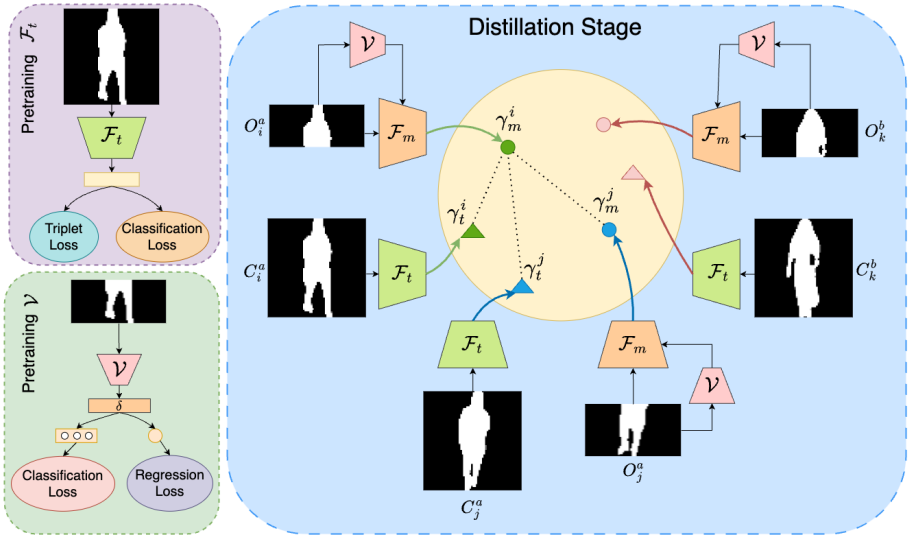

Figure 2. Overview of the proposed approach. The training procedure consists of two stages. In the pretraining stage, the Visibility Estimation Network $\mathcal{V}$ (also called VEN) and the teacher network $\mathcal{F}_t$ are trained. In the distillation stage, a new mimic network $\mathcal{F}_m$ is trained by $\mathcal{F}_t$ using a multi-instance correlational KD loss. $\mathcal{V}$ is used to guide $\mathcal{F}_m$ by injecting occlusion-relevant features. The i^{th} video of subject a may be occluded (denoted by O_i^a) or holistic (denoted by C_i^a). Their representations in the latent space are denoted by γ_m and γ_t respectively. The three types of anchor-positive pairs sampled by the proposed loss seen in the figure are described in Sec. 3.3.

2.2. Occluded Person Re-ID and Gait Recognition

Occlusions can severely hamper gait recognition and Person Re-ID systems [27, 38]. Occlusion has already been recognized as an important problem in Person Re-ID [30], with datasets targetting the occlusion problem specifically [25, 45]. Current works deal with simulated occlusions [1] as well as real occlusions [26]. [1] categorizes occlusions into different broad categories and uses this knowledge to better identify the subject, while [26] uses pose estimation techniques to identify visible body parts, generating different representations for each part.

Unlike the Person Re-ID problem, occlusion has not received much attention in gait recognition due to lack of large-scale datasets. Some available works focus on reconstruction [33, 34] of the silhouette sequence using generative networks or estimate an SMPL-based 3D mesh model [36] to infer the missing body parts. However, these approaches are not easily extendable to long ranges or noisy data, as would be the case in a real-world scenario.

Some works simulate occlusions in existing datasets and utilize auxiliary networks to gain additional information about occlusions [13, 37]. [37] performs silhouette registration to enhance the input, but only works with top occlusions on indoor datasets. [13] works on outdoor data and a diverse set of occlusions, but neglects the potential correlations between occluded and visible body parts due to exclusive exposure to occluded data during training.

2.3. Knowledge Distillation

Broadly, Knowledge Distillation (KD) involves a student-teacher learning framework where one model passes on its 'knowledge' to another model [11]. Initially, KD was used as a technique for model compression and acceleration [16], where a large model was used to train a smaller network to reduce memory and computing requirements. However, the utility of KD was shown in other areas as well. [31] used the idea of data distillation to train a student network using a single teacher by applying multiple transforms on the input. [15] used distillation to teach a student network to dehaze an image by applying consistency loss in intermediate features between the student and teacher. [29] proposes correlational congruence in knowledge distillation, utilizing correlation across multiple instances to teach their student network. [43] uses KD in the context of gait recognition, transferring knowledge from RGB to silhouette encoders to infer 3D body features.

To work with occluded data captured from a distance, we utilize KD to capture correlations among the occluded and visible body parts to learn a gait signature for the subject.

3. Proposed Method

Given an input video sequence $S^i = \{v_1, v_2, ...v_n\}$ for subject i, the goal is to find a discriminative gait signature γ for the subject. We assume that the input sequence consists of binary silhouette masks. In the occlusion scenario, this mask may not be completely visible in some or all of the

frames which makes the task more challenging.

The overview of the proposed approach is shown in Fig. 2. The entire training procedure consists of two stages - the pretraining stage, and the distillation stage. In the pretraining stage, a state-of-the-art gait recognition model $\mathcal{F}_t$ is trained on the original unoccluded video frames. Separate from $\mathcal{F}_t$, the VEN $\mathcal{V}$ is also trained to identify the amount and type of synthetic occlusions present in the video. These two networks are used in the next stage with frozen weights.

In the distillation stage, a new network $\mathcal{F}_m$, the mimic network, is initialized with the same architecture as $\mathcal{F}_t$. $\mathcal{F}_m$ is trained to output a gait signature γ_m by taking the occluded video O^i as input. At the same time, $\mathcal{F}_t$ outputs the gait signature γ_t by taking the corresponding full body video C^i as input. A multi-instance correlational distillation loss is used to train $\mathcal{F}_m$ to bring the distributions of γ_m and γ_t closer in the latent space. VEN is used to regulate $\mathcal{F}_m$ by injecting visibility information into the backbone. The method is described in detail in the following sections.

3.1. Visibility Estimation Network

VEN is a CNN that predicts the type of occlusion, if any, present in the input video and also a measure of the amount of this occlusion, which we call visibility estimation. This module is inspired by [13], where an auxiliary network was used to identify the occlusion class. VEN builds on top of [13] by performing the visibility estimation task while simultaneously predicting the occlusion type.

In the pretraining stage, VEN is trained on synthetic occlusions. Some examples of the occlusions are shown in Figure 1. It consists of a sequence of convolutional and linear layers, with two parallel heads - one classification head for the occlusion classification task, and one regression head for the visibility estimation task. Accordingly, cross entropy and L2 regression losses are used on the two heads to train the network, making it learn occlusion-relevant features.

When VEN is used in the distillation stage, the two heads are removed and the penultimate feature vector δ is used to guide $\mathcal{F}_m$. It is important to note that the weights of VEN are frozen during this stage, to ensure that it retains the visibility awareness learned during training.

3.2. Mimic Network

The main idea behind the mimic network is that the features corresponding to the missing/occluded body parts are correlated with the observable motion of the subject. Every moving body part is correlated to a global pattern and also has its own distinctive local motion, which is why pyramid structures operating on multi-scale input are popular in gait recognition [7,23]. It is these patterns of motion that constitute gait, and when some of these patterns are missing, the mimic network uses its learned correlations and the available input to fill in the gaps in the gait signature.

During the pretraining stage, a state-of-the-art gait recognition backbone $\mathcal{F}_t$ is trained on the original, unoccluded videos to generate discriminative gait signatures γ_t. During the distillation stage, $\mathcal{F}_m$ is trained to output a discriminative gait signature γ_m using the occluded videos O^i as input. Since occlusions can be of many different types, and the internal architecture of $\mathcal{F}_m$ does not target any occlusion, we use the occlusion-relevant features from VEN to guide $\mathcal{F}_m$, similar to [13]. Specifically,

$$\gamma_m = \mathcal{F}^{'}{}_m(O^i) = T(\mathcal{F}_m(O^i) \oplus \mathcal{V}(O^i)) \tag{1}$$

where $\mathcal{V}$ refers to VEN, $\oplus$ is the concatenation operation, and T is a linear transformation to make the feature size compatible after concatenation. By introducing visibility features from VEN into $\mathcal{F}_m$ in such a manner, the network gains information about the visibility of the subject in the input, which in turn helps it to generate features closer to the optimal holistic features.

In the distillation stage, $\mathcal{F}_t$ acts as a teacher network and $\mathcal{F}^{'}_m$ tries to mimic the holistic features generated by $\mathcal{F}_t$. However, the difference is that $\mathcal{F}^{'}_m$ is only allowed to see the occluded videos during training. In the process of trying to mimic the holistic features from the occluded inputs, $\mathcal{F}_m$ is able to learn better gait representations for occlusion scenarios. This is achieved by using a multi-instance correlational distillation loss as described in the next section.

3.3. Multi-instance Correlational Distillation Loss

Inspired by [29], we use a multi-instance correlational KD loss modified for the occluded gait recognition task. We recognize that while there is correlation within the motion of different body parts of the same gait instance, there is also correlation among the motion of the body parts across multiple gait instances of the same subject. The distillation loss, modeled as a Triplet Loss [35], is able to capture both these correlations. As shown in Fig. 2, the outputs of the mimic network, γ_m, and those of the teacher network, γ_t, are used to sample anchor-positive pairs of the following three types: 1) γ^i_m - γ^i_t, representing student-teacher learning within the same gait instance, 2) γ^i_m - γ^j_t, representing student-teacher learning across instances, and 3) γ^i_m - γ^j_m, representing mimic network correlation across instances.

These considerations lead to a triplet margin loss:

$$L = \sum_i [D^i_{a,p} - D^i_{a,n} + m]_+ \tag{2}$$

where the summation is over all anchor-positive-negative triplets and $D_{a,p}$, $D_{a,n}$ refer to the Euclidean distance between the anchor-positive (AP) or anchor-negative (AN) pairs. The AP pairs are sampled from the previously mentioned three types, while the AN pairs are sampled from other identities, and m is the margin.

Figure 3. Some samples images taken from the BRIAR dataset for two subjects. From left to right, the range of capture increases from close range to 1000m for each subject. A large variation in the quality of the videos and the collection conditions can be seen. Subjects have consented to the use of these images in publication.

Inference: During inference, when only the occluded video is available, the mimic network $\mathcal{F}_m$ guided by the VEN is used to generate the gait features.

4. Experimental Setup

4.1. Datasets

BRIAR: We use the BRIAR [2] dataset to conduct our experiments. This is a recently collected dataset that contains many variations in the walking conditions of subjects. The subset of BRIAR data we use comprises of 776 training subjects and 856 test subjects. The dataset contains videos of subjects walking in indoor, controlled environments as well as outdoor field environments. The outdoor data is captured from many different camera sensors, ranges, altitudes and viewpoints. Some example images are shown in Fig. 3.

Videos of walking subjects in outdoor environments are captured systematically at distances ranging from 100m to 1000m. Additionally, videos are also captured using an UAV and at close range with extreme viewpoint.

Each subject either walks randomly inside a fixed boundary (random), or along well-defined straight lines (structured), or both, in different videos. The subject may be carrying a large object like a big cardboard box in some videos, while some may be using their cellphones while walking.

GREW: GREW [44] is a large scale publicly available dataset for gait recognition, comprising of 20,000 subjects in the training split and 6,000 in the testing split. The dataset is captured from in-the-wild videos in a variety of conditions from multiple different cameras with varying viewpoints. We utilize the provided 2D silhouettes in this work.

Gait3D: Gait3D [42] is a publicly available in-the-wild gait recognition dataset, comprising of 3,000 subjects in the training set and 1,000 subjects in the testing set. We utilize the provided 2D silhouettes of the dataset in this work.

4.2. Synthetic Occlusions

Broadly, occlusions can be classified into two categories - 1) consistent, where the occlusion stays roughly the same for the entire length of the video, and 2) dynamic, where they can change with time.

In uncontrolled environments, consistent occlusions occur when there are obstacles like an elevated sidewalk or tall grass between the subject and the camera, or they might arise due to a bad camera angle. Dynamic occlusions occur when objects, or other people, temporarily block the subject of interest from view. We try to simulate both consistent and dynamic occlusions by placing stationary or moving black patches on the input frames, as shown in Fig. 1. For consistent occlusions, we remove either the top, bottom or middle part of the frame. Other works on occluded gait recognition [13] simulate more type of occlusions, however many of these occlusions are not likely to occur in a practical scenario. We focus on the the top and bottom occlusions in our main results and perform generalizability and adaptability evaluations using middle and dynamic occlusions in Sec. 5.

Occlusions are introduced randomly in the videos during training and evaluation. We randomly occlude a portion of the frame as shown in Fig. 1. The amount of occlusion is selected randomly within a fixed range R, which we set to (0.4, 0.6) for all our experiments. This means that the portion of the synthetic occlusion is chosen randomly between 40%-60% of the spatial dimensions of the input video. More details about the synthetic occlusions have been included in the supplementary material.

4.3. Baselines

To showcase the model-agnosticity of our method, we experiment with existing architectures like GaitBase [5], GaitGL [23] and deeper networks like DeepGaitV2 [4].

Following [13], we train these networks on holistic videos and evaluate them directly on synthetic occlusions in a zero-shot setting. This is called Baseline-1. Since these architectures do not address the occlusion problem specifically, we retrain them on synthetically occluded data as Baseline-2. We also compare our results with [13].

4.4. Implementation Details

Visibility Estimation Network: VEN consists of three convolutional layers and two linear layers. VEN consists of two heads - a classification head for classifying occlusions and a regression head for visibility estimation. In our experiments, VEN is trained to classify the input into three classes, namely, no occlusion, top occlusion and bottom occlusion. This set is updated as more occlusion types are introduced for training the mimic network. More details about VEN are included in the supplementary material.

Mimic Network: We localize the subjects in an $H \times W = 64 \times 64$ bounding box during preprocessing and crop these bounding boxes from the video. The mimic network architecture is the same as the gait recognition backbone $\mathcal{F}_t$. Training occurs in two stages. We train $\mathcal{F}_t$ on holistic videos in the pretraining stage. Next, we train $\mathcal{F}_m$ using features obtained by $\mathcal{F}_t$ in the distillation stage. We choose the same optimizer as used by $\mathcal{F}_t$ during pretraining, namely Adam [18] for GaitGL and SGD for GaitBase and Deep-

Backbone	Method	Gait3D		GREW		BRIAR		
		Rank-1	Rank-5	Rank-1	Rank-5	Rank-1	Rank-20	TAR@0.01 FAR
GaitBase	Baseline-1	7.6 (0.11)	15.71 (0.19)	14.85 (0.27)	25.55 (0.35)	1.34 (0.04)	12.05 (0.16)	2.46 (0.04)
	Baseline-2	17.12 (0.24)	31.43 (0.37)	16.42 (0.30)	30.38 (0.42)	6.13 (0.16)	27.52 (0.36)	9.81 (0.17)
	Occlusion Aware [13]	18.22 (0.26)	34.94 (0.41)	22.55 (0.41)	37.95 (0.53)	8.70 (0.23)	35.38 (0.46)	12.83 (0.22)
	Mimic Network (ours)	**22.72 (0.32)**	**40.84 (0.48)**	**28.38 (0.51)**	**45.43 (0.63)**	**10.93 (0.28)**	**42.25 (0.55)**	**12.92 (0.22)**
GaitGL	Baseline-1	3.00 (0.11)	6.70 (0.15)	3.60 (0.10)	6.90 (0.13)	0.69 (0.05)	6.70 (0.18)	1.90 (0.09)
	Baseline-2	4.4 (0.16)	9.71 (0.22)	6.3 (0.17)	11.92 (0.22)	1.94 (0.15)	13.17 (0.35)	3.42 (0.16)
	Occlusion Aware [13]	4.8 (0.18)	12.7 (0.29)	7.05 (0.19)	12.72 (0.24)	4.02 (0.30)	22.85 (0.61)	4.19 (0.19)
	Mimic Network (ours)	**5.6 (0.21)**	**13.5 (0.31)**	**7.53 (0.21)**	**13.97 (0.26)**	**5.4 (0.40)**	**24.75 (0.66)**	**5.14 (0.23)**
DeepGaitV2	Baseline-1	2.20 (0.03)	6.21 (0.07)	3.72 (0.05)	6.85 (0.08)	2.38 (0.05)	11.40 (0.13)	1.80 (0.03)
	Baseline-2	6.71 (0.09)	14.21 (0.16)	9.65 (0.13)	16.20 (0.19)	6.52 (0.13)	27.47 (0.32)	7.40 (0.11)
	Occlusion Aware [13]	14.01 (0.18)	27.83 (0.32)	13.38 (0.18)	22.87 (0.27)	7.39 (0.15)	32.74 (0.38)	7.12 (0.10)
	Mimic Network (ours)	**16.82 (0.22)**	**33.03 (0.38)**	**14.38 (0.20)**	**24.93 (0.29)**	**11.49 (0.24)**	**44.71 (0.52)**	**20.73 (0.30)**

Table 1. Our results on the Gait3D [42], GREW [44] and BRIAR [2] datasets, for different gait recognition backbones. Baseline-1 refers to zero shot evaluation on occluded data. Baseline-2 refers to training the network on occlusions. The values in (.) denote the relative performance (RP) values with respect to the ideal, no occlusion scenario. We can see that the mimic network outperforms other methods on occluded data, achieving at least 20-30% RP across datasets and backbones.

GaitV2. The learning rate for GaitGL experiments is 1e-4, while for GaitBase and DeepGaitV2 it is 1e-1. Additional details about the training procedure of the mimic network have been included in the supplementary material.

4.5. Evaluation Metrics

Rank Retrieval: Rank retrieval accuracy is a standard metric to evaluate recognition performance. We follow the gallery-probe splits of [5] for GREW and Gait3D. For GREW, the probe set labels are not publicly released. To enable local evaluation for the GREW dataset, we follow the method proposed in [5], the details of which are provided in the supplementary material. The BRIAR dataset provides its own protocol [2]. We additionally compute the verification performance for BRIAR.

Relative Performance (RP): In the model-agnostic scenario being discussed in this work, we need to measure the effectiveness of an *occlusion-mitigating method* across backbones. Considering only the performance of a model on occluded data - the *occluded performance* OP - gives an incomplete picture. This is because a low OP might be caused by other factors not related to the strength of the occlusion-mitigating method - such as the backbone being suboptimal or the dataset being too difficult. To filter out these other factors and focus on just the strength of the occlusion-mitigating method, we define a new RP metric,

$$RP = \frac{OP}{HP} \tag{3}$$

where OP is the occluded performance and HP is the holistic performance on unoccluded data for a given backbone. If the backbone is suboptimal or the dataset is too difficult, both OP and HP are low, therefore the RP is not affected very much. However, if the occlusion-mitigating method is suboptimal, only the OP is low - reducing the RP. This makes RP relatively more suited to model-agnostic evaluation of occlusion-mitigating methods, like the one introduced by [13] and our proposed mimic network.

Another way of looking at RP is that it normalizes the OP by its true upper bound HP, so changes in OP are measured with respect to HP. A small improvement Δy in OP may seem insignificant, but becomes important if HP itself is small - for example, an improvement of 1% in Rank-1 accuracy in OP is significant if the upper bound HP is itself just 10%!

Fig. 4 provides a geometric interpretation of this scenario. B1/B2 are two hypothetical backbones and M1/M2 are two occlusion-mitigating methods. RP is the slope of the line joining origin to M1/M2. For the suboptimal backbone B1, a small Δy_1 by using M2 over M1 can cause a large change in slope/RP. On the other hand, a larger increase of Δy_2 in OP is needed to cause similar improvements in slope/RP. Thus, RP gives more insight into the improvement brought about by M2 in this backbone-agnostic scenario by normalizing OP by the performance of the backbone.

Adaptability: This test aims to evaluate how well a model adapts to new occlusion types when it is further trained on them along with the original occlusions. This test provides a new perspective - it analyses the scenario when we want to extend the model's capability to a new occlusion type by additional training. The results of the adaptability evaluation are presented in Tab. 2.

Generalizability: While having prior knowledge about occlusion types is an advantage, it is not practical to train on all possible occlusion types. For a method to be deployable, it should be able to generalize to occlusions not seen during training. This idea of generalizability to different occlusions was introduced in [13]. In the generalizability test, we take models trained on top and bottom occlusions and evaluate them directly on the newer occlusion types, in a zero-shot setting. The results are presented in Tab. 2.

5. Results and Discussion

In general, we observe that the mimic network outperforms other methods for occluded gait recognition. This

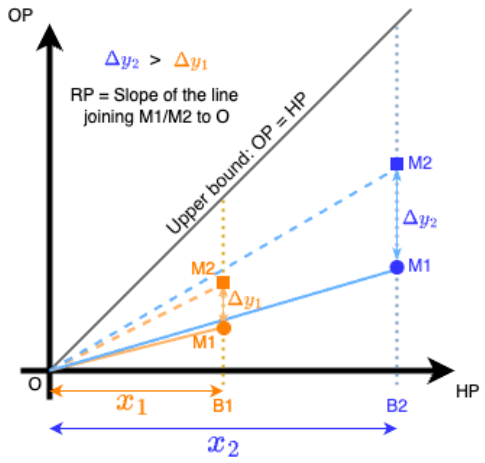

Figure 4. Comparing two hypothetical occlusion-mitigating methods M1/M2 between two backbones B1/B2 on occluded performance (OP) and holistic performance (HP). A small change Δy_1 in OP can cause a large change in the slope/RP for B1, but a larger Δy_2 is needed to cause a similar change in slope for B2. By considering the slope rather than just the OP, RP is able to better isolate the effect of M1/M2 across backbones.

statement holds across datasets and backbones, demonstrating the effectiveness and model-agnosticity of our method (Tab. 1). The generalizability and adaptability results presented in Tab. 2 confirm that the mimic network also outperforms other approaches here. Thus, we conclude that capturing correlations among occluded and visible body parts using our proposed mimic network does indeed help in occluded scenarios. However, apart from this general trend, we make several interesting observations.

Insights from the RP metric: In some cases, performance improvements are not visible just from Rank-K accuracy, such as with GaitBase on BRIAR or GaitGL on Gait3D, due to low absolute performance. This is attributed to the challenging nature of the datasets (e.g., BRIAR with extreme range and turbulence) or lower capability of backbones (e.g., GaitGL). The RP metric can filter out these factors to some extent. It is not perfect at filtering these factors, and RP might still change across backbones, but improvements are easier to see in RP even in the case of suboptimal backbones or difficult datasets.

Increasing RP with Rank: An interesting observation is the effect of rank. As rank increases, both absolute accuracy and RP increase across all experiments. The increase in RP is non-trivial because RP is the ratio of occluded to holistic accuracy, and the rising RP indicates that occlusion accuracy grows faster than holistic accuracy as rank increases! This is possibly because the model is better at leveraging partial information to correctly identify occluded instances when given more opportunities (higher rank) to match. Holistic data already benefits from complete information, leading to a relatively slower increase in accuracy.

Deeper networks: DeepGaitV2 is considered to be better than GaitBase as a backbone [4]. Interestingly, we observe that it performs worse under occlusions in many of our experiments. We think that the larger number of parameters and the larger depth of DeepGaitV2 make it harder to optimize under occluded conditions where the input is more sparse. Regardless, the mimic network still performs better than the occlusion aware method for this backbone as well.

Relative difficulty of different occlusions: Comparing Tab. 1 and the adaptability section Tab. 2 suggests that middle and dynamic occlusions are easier than top and bottom occlusions, since the model is able to perform better on the former set. We investigate this further in the supplementary material by evaluating on these occlusions individually.

5.1. Ablation Studies

In this section, we perform various ablations on the proposed network, removing or modifying different parts of the model to see their effects on performance. For all the experiments in this section, we use the GREW [44] dataset and the GaitBase [5] backbone unless stated otherwise.

Effect of Multi-instance Correlational KD loss: In this section, we analyse the effect of the proposed multi-instance correlational distillation (MiCKD) loss on the network. To isolate the effects of the MiCKD loss, we remove the VEN in the experiments in this section, and deal with a 'Vanilla Mimic' network. If we completely remove the distillation loss from this Vanilla Mimic network, the method becomes the same as Baseline-2, which does not capture any occluded-visible body part correlations. Next, we try a simpler approach for the distillation stage by considering correlations among γ_m and γ_t only within the same instance, minimizing the L2 distance between them (L2 KD). Comparing this to MiCKD in Tab. 3 isolates the effect of utilizing multiple instances for feature learning.

We observe that the latter performs better, possibly because local gait patterns remain consistent across walking instances. The model is able to leverage this consistency to learn correlations among body parts which are occluded in one instance but visible in another.

Adding Cross Entropy Loss: Based on the training techniques of $\mathcal{F}_t$, we hypothesize that adding cross entropy loss using a BNNeck layer as done in [5] would further help the model along with MiCKD loss. However, as shown in Tab. 3, our hypothesis is negated and we observe that this approach actually reduces the model performance. We are unsure why this occurs, and hypothesize that the losses could conflict with each other. Based on this, we choose to exclude cross-entropy loss in the final model.

Effect of guidance by VEN: In this section, we analyze the role of VEN in the mimic network. We compare the vanilla mimic network with the 'Mimic + VEN' row of

Additional Occlusion Type	Method	Generalizability		Adaptability	
		Rank1	Rank5	Rank1	Rank5
Middle	Baseline-1	13.12 (0.24)	24.62 (0.34)	-	-
	Baseline-2	13.87 (0.25)	25.72 (0.36)	21.25 (0.38)	36.5 (0.51)
	Occlusion Aware [13]	17.93 (0.32)	32.15 (0.45)	26.7 (0.48)	43.82 (0.61)
	Mimic Network (ours)	**21.73 (0.39)**	**37.37 (0.52)**	**34.78 (0.63)**	**52.75 (0.73)**
Dynamic	Baseline-1	17.27 (0.31)	28.05 (0.39)	-	-
	Baseline-2	17.48 (0.32)	31.53 (0.44)	32.1 (0.58)	49.68 (0.69)
	Occlusion Aware [13]	21.27 (0.38)	36.5 (0.51)	34.87 (0.63)	52.07 (0.72)
	Mimic Network (ours)	**26.77 (0.48)**	**42.9 (0.60)**	**36.65 (0.66)**	**53.15 (0.74)**

Table 2. The Adaptability (additional training) and Generalizability (zero-shot) evaluations using the GaitBase backbone on GREW dataset, using additional occlusion types. RP values are shown in (.). Note that adaptability is not applicable for Baseline-1, since Baseline-1 is not trained on any occlusions. The mimic network outperforms other approaches in these scenarios.

Method	Rank-1	Rank-5
No KD (Baseline-2)	16.42	30.38
L2 KD	20.43	35.95
MiCKD (Vanilla Mimic)	**25.75**	**42.2**
MiCKD + XE	20.22	34.93

Table 3. Different distillation strategies for the mimic network, using GaitBase on the GREW dataset. We see that the proposed multi-instance correlational knowledge distillation loss (MiCKD) indeed helps in learning better occlusion features.

Method	Mimic	Proxy Tasks		Accuracy	
		Classif.	Reg.	Rank-1	Rank-5
Baseline-2				16.42	30.38
Occlusion aware [13]		✓		22.55	37.95
VEN		✓	✓	23.52	39.68
Vanilla Mimic	✓			25.75	42.2
Mimic + Occlusion Aware	✓	✓		27.52	44.15
Mimic + VEN	✓	✓	✓	**28.38**	**45.43**

Table 4. Effect of the mimic training strategy and the proxy tasks involved in the pretraining of the occlusion detector. Classif. and Reg. refer to the classification and regression tasks respectively. In general, training the auxiliary network on more tasks and using the mimic training strategy improves performance under occlusions.

Tab. 4. The mimic network benefits from VEN guidance; without VEN, the network must independently determine which body parts are visible, complicating the extraction of gait patterns. In contrast, external guidance from VEN simplifies this task, allowing the network to focus on gait pattern extraction rather than occlusion identification.

Different proxy tasks for training VEN: In the VEN pretraining stage, we train it to jointly output the occlusion type and the occlusion amount, with two different loss functions for each task. In this section, we investigate how useful these two individual tasks are for learning useful occlusion aware features. As such, we train a network without the occlusion amount regression head, making it similar to the occlusion detector in [13] and compare it with VEN.

To get an estimate of the quality of occlusion awareness within these two networks, we compare the performance of gait recognition backbones trained using VEN and the occlusion detector in Tab. 4. We observe that VEN has inherently better occlusion relevant features, regardless of whether the mimic training strategy is applied or not - and so we conclude that the classification and regression proxy tasks individually contribute to performance.

6. Limitations and Future Work

Although our proposed method can perform better on synthetic occlusions, it is not perfect. We proposed a general approach without altering the backbone, and future works can explore incorporating specific architectural changes to address occlusions better. Further, we could not test our approach on real occlusions due to the absence of an occlusion category in the outdoor datasets we used. To properly evaluate our method, a large-scale dataset specifically focused on occlusions is essential to advance research in this area. Lastly, we were unable to explore why adding cross entropy loss hurts MimicGait. Future work can explore this further to achieve more gains in performance.

7. Conclusion

In this work, we proposed *MimicGait*, a novel model-agnostic approach for occluded gait recognition. We proposed a multi-instance correlational KD loss to train the mimic network in a student-teacher setting, utilizing an auxiliary Visibility Estimation Network to introduce occlusion-relevant features. We introduced generalizability and adaptability tests along with a new metric RP to evaluate occluded performance. We evaluated our approach on GREW, Gait3D and BRIAR datasets, and showed that the proposed mimic network outperforms existing approaches on occlusions on real-world data collected from large distances.

Acknowledgements: 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 [2022-21102100005]. 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 US. Government is authorized to reproduce and distribute reprints for governmental purposes notwithstanding any copyright annotation therein.

References

[1] Peixian Chen, Wenfeng Liu, Pingyang Dai, Jianzhuang Liu, Qixiang Ye, Mingliang Xu, Qi'an Chen, and Rongrong Ji. Occlude Them All: Occlusion-Aware Attention Network for Occluded Person Re-ID. pages 11833–11842, 2021. 3

[2] David Cornett, Joel Brogan, Nell Barber, Deniz Aykac, Seth Baird, Nicholas Burchfield, Carl Dukes, Andrew Duncan, Regina Ferrell, Jim Goddard, Gavin Jager, Matthew Larson, Bart Murphy, Christi Johnson, Ian Shelley, Nisha Srinivas, Brandon Stockwell, Leanne Thompson, Matthew Yohe, Robert Zhang, Scott Dolvin, Hector J. Santos-Villalobos, and David S. Bolme. Expanding accurate person recognition to new altitudes and ranges: The briar dataset. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV) Workshops*, pages 593–602, January 2023. 1, 2, 5, 6

[3] Qi Dang, Jianqin Yin, Bin Wang, and Wenqing Zheng. Deep learning based 2d human pose estimation: A survey. *Tsinghua Science and Technology*, 24(6):663–676, 2019. 2

[4] Chao Fan, Saihui Hou, Yongzhen Huang, and Shiqi Yu. Exploring deep models for practical gait recognition. *arXiv preprint arXiv:2303.03301*, 2023. 2, 5, 7

[5] Chao Fan, Junhao Liang, Chuanfu Shen, Saihui Hou, Yongzhen Huang, and Shiqi Yu. Opengait: Revisiting gait recognition towards better practicality. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9707–9716, June 2023. 1, 2, 5, 6, 7

[6] Chao Fan, Jingzhe Ma, Dongyang Jin, Chuanfu Shen, and Shiqi Yu. Skeletongait: Gait recognition using skeleton maps. *arXiv preprint arXiv:2311.13444*, 2023. 2

[7] 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 (CVPR)*, June 2020. 2, 4

[8] Claudio Filipi Gonçalves dos Santos, Diego de Souza Oliveira, Leandro A. Passos, Rafael Gonçalves Pires, Daniel Felipe Silva Santos, Lucas Pascotti Valem, Thierry P. Moreira, Marcos Cleison S. Santana, Mateus Roder, Jo Paulo Papa, and Danilo Colombo. Gait recognition based on deep learning: A survey. *ACM Comput. Surv.*, 55(2), jan 2022. 2

[9] Yang Fu, Shibei Meng, Saihui Hou, Xuecai Hu, and Yongzhen Huang. Gpgait: Generalized pose-based gait recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 19595–19604, October 2023. 2

[10] Davrondzhon Gafurov and Einar Snekkenes. Gait recognition using wearable motion recording sensors. *EURASIP Journal on Advances in Signal Processing*, 2009:1–16, 2009. 1, 2

[11] Jianping Gou, Baosheng Yu, Stephen J Maybank, and Dacheng Tao. Knowledge distillation: A survey. *International Journal of Computer Vision*, 129:1789–1819, 2021. 3

[12] Yuxiang Guo, Cheng Peng, Chun Pong Lau, and Rama Chellappa. Multi-modal human authentication using silhouettes, gait and rgb. In *2023 IEEE 17th International Conference on Automatic Face and Gesture Recognition (FG)*, pages 1–7, 2023. 2

[13] Ayush Gupta and Rama Chellappa. You can run but not hide: Improving gait recognition with intrinsic occlusion type awareness. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 5893–5902, January 2024. 2, 3, 4, 5, 6, 8

[14] Md Mahedi Hasan and Hossen Asiful Mustafa. Multi-level feature fusion for robust pose-based gait recognition using rnn. *Int. J. Comput. Sci. Inf. Secur.(IJCSIS)*, 18(1), 2020. 2

[15] Ming Hong, Yuan Xie, Cuihua Li, and Yanyun Qu. Distilling image dehazing with heterogeneous task imitation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3462–3471, 2020. 3

[16] Yu-Wei Hong, Jenq-Shiou Leu, Muhamad Faisal, and Setya Widyawan Prakosa. Analysis of model compression using knowledge distillation. *IEEE Access*, 10:85095–85105, 2022. 3

[17] Ramneet Kaur, Kaustubh Sridhar, Sangdon Park, Yahan Yang, Susmit Jha, Anirban Roy, Oleg Sokolsky, and Insup Lee. Codit: Conformal out-of-distribution detection in time-series data for cyber-physical systems. In *Proceedings of the ACM/IEEE 14th International Conference on Cyber-Physical Systems (with CPS-IoT Week 2023)*, ICCPS '23, page 120–131, New York, NY, USA, 2023. Association for Computing Machinery. 1

[18] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In Yoshua Bengio and Yann LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015. 5

[19] 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 (ACCV)*, November 2020. 2

[20] Junhao Liang, Chao Fan, Saihui Hou, Chuanfu Shen, Yongzhen Huang, and Shiqi Yu. Gaitedge: Beyond plain end-to-end gait recognition for better practicality. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part V*, pages 375–390. Springer, 2022. 2

[21] Rijun Liao, Chunshui Cao, Edel B Garcia, Shiqi Yu, and Yongzhen Huang. Pose-based temporal-spatial network (ptsn) for gait recognition with carrying and clothing variations. In *Biometric Recognition: 12th Chinese Conference, CCBR 2017, Shenzhen, China, October 28-29, 2017, Proceedings 12*, pages 474–483. Springer, 2017. 2

[22] Vítor C de Lima, Victor HC Melo, and William R Schwartz. Simple and efficient pose-based gait recognition method for challenging environments. *Pattern Analysis and Applications*, 24:497–507, 2021. 2

[23] Beibei Lin, Shunli Zhang, and Xin Yu. Gait Recognition via Effective Global-Local Feature Representation and Local Temporal Aggregation. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 14628–14636, Montreal, QC, Canada, Oct. 2021. IEEE. 1, 2, 4, 5

[24] Maria De Marsico and Alessio Mecca. A survey on gait recognition via wearable sensors. 52(4), aug 2019. 2

[25] Jiaxu Miao, Yu Wu, Ping Liu, Yuhang Ding, and Yi Yang. Pose-guided feature alignment for occluded person re-identification. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 542–551, 2019. 3

[26] Jiaxu Miao, Yu Wu, and Yi Yang. Identifying Visible Parts via Pose Estimation for Occluded Person Re-Identification. *IEEE Transactions on Neural Networks and Learning Systems*, 33(9):4624–4634, Sept. 2022. 3

[27] Vuong D Nguyen, Pranav Mantini, and Shishir K Shah. Temporal 3d shape modeling for video-based cloth-changing person re-identification. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 173–182, 2024. 3

[28] Gunwoo Park, Kyoung Min Lee, and Seungbum Koo. Uniqueness of gait kinematics in a cohort study. *Scientific Reports*, 11(1):15248, 2021. 1

[29] Baoyun Peng, Xiao Jin, Jiaheng Liu, Dongsheng Li, Yichao Wu, Yu Liu, Shunfeng Zhou, and Zhaoning Zhang. Correlation congruence for knowledge distillation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5007–5016, 2019. 3, 4

[30] Yunjie Peng, Saihui Hou, Chunshui Cao, Xu Liu, Yongzhen Huang, and Zhiqiang He. Deep learning-based occluded person re-identification: A survey, 2022. 3

[31] Ilija Radosavovic, Piotr Dollár, Ross Girshick, Georgia Gkioxari, and Kaiming He. Data distillation: Towards omni-supervised learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4119–4128, 2018. 3

[32] Chuanfu Shen, Shiqi Yu, Jilong Wang, George Q Huang, and Liang Wang. A comprehensive survey on deep gait recognition: algorithms, datasets and challenges. *arXiv preprint arXiv:2206.13732*, 2022. 1

[33] Jasvinder Pal Singh, Sanjeev Jain, Uday Pratap Singh, and Sakshi Arora. Hybrid neural network model for reconstruction of occluded regions in multi-gait scenario. *Multimedia Tools and Applications*, 81(7):9607–9629, 2022. 1, 3

[34] Md Uddin, Daigo Muramatsu, Noriko Takemura, Md Ahad, Atiqur Rahman, Yasushi Yagi, et al. Spatio-temporal silhouette sequence reconstruction for gait recognition against occlusion. *IPSJ Transactions on Computer Vision and Applications*, 11(1):1–18, 2019. 2, 3

[35] Daniel Ponsa Vassileios Balntas, Edgar Riba and Krystian Mikolajczyk. Learning local feature descriptors with triplets and shallow convolutional neural networks. In Edwin R. Hancock Richard C. Wilson and William A. P. Smith, editors, *Proceedings of the British Machine Vision Conference (BMVC)*, pages 119.1–119.11. BMVA Press, September 2016. 4

[36] Chi Xu, Yasushi Makihara, Xiang Li, and Yasushi Yagi. Occlusion-Aware Human Mesh Model-Based Gait Recognition. *IEEE Transactions on Information Forensics and Security*, 18:1309–1321, 2023. Conference Name: IEEE Transactions on Information Forensics and Security. 1, 2, 3

[37] Chi Xu, Shogo Tsuji, Yasushi Makihara, Xiang Li, and Yasushi Yagi. Occluded gait recognition via silhouette registration guided by automated occlusion degree estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, pages 3199–3209, October 2023. 3

[38] Mang Ye, Jianbing Shen, Gaojie Lin, Tao Xiang, Ling Shao, and Steven CH Hoi. Deep learning for person re-identification: A survey and outlook. *IEEE transactions on pattern analysis and machine intelligence*, 44(6):2872–2893, 2021. 3

[39] 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. 2

[40] Cun Zhang, Xing-Peng Chen, Guo-Qiang Han, and Xiang-Jie Liu. Spatial transformer network on skeleton-based gait recognition. *Expert Systems*, 40(6):e13244, 2023. 2

[41] Ziyuan Zhang, Luan Tran, Xi Yin, Yousef Atoum, Xiaoming Liu, Jian Wan, and Nanxin Wang. Gait recognition via disentangled representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4710–4719, 2019. 2

[42] Jinkai Zheng, Xinchen Liu, Wu Liu, Lingxiao He, Chenggang Yan, and Tao Mei. Gait recognition in the wild with dense 3d representations and a benchmark. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 2, 5, 6

[43] Haidong Zhu, Zhaoheng Zheng, and Ram Nevatia. Gait recognition using 3-d human body shape inference. pages 909–918, 01 2023. 3

[44] 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. 1, 2, 5, 6, 7

[45] Jiaxuan Zhuo, Zeyu Chen, Jianhuang Lai, and Guangcong Wang. Occluded person re-identification. In *2018 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6. IEEE, 2018. 3