You Can Run but not Hide: Improving Gait Recognition with Intrinsic Occlusion Type Awareness

Ayush Gupta, Rama Chellappa

Johns Hopkins University

3400 North Charles Street, Baltimore, MD, 21218

{agupt120, rchella4}@jhu.edu

Abstract

While gait recognition has seen many advances in recent years, the occlusion problem has largely been ignored. This problem is especially important for gait recognition from uncontrolled outdoor sequences at range - since any small obstruction can affect the recognition system. Most current methods assume the availability of complete body information while extracting the gait features. When parts of the body are occluded, these methods may hallucinate and output a corrupted gait signature as they try to look for body parts which are not present in the input at all. To address this, we exploit the learned occlusion type while extracting identity features from videos. Thus, in this work, we propose an occlusion aware gait recognition method which can be used to model intrinsic occlusion awareness into potentially any state-of-the-art gait recognition method. Our experiments on the challenging GREW and BRIAR datasets show that networks enhanced with this occlusion awareness perform better at recognition tasks than their counterparts trained on similar occlusions.

1. Introduction

Gait-based recognition and identification techniques focus on analyzing and recognizing an individual's walking pattern, and can be very helpful when other identifiers like face are not available. Every person has a characteristic gait pattern [21] consisting of various bio-mechanic features like stride length, walking speed, body posture, and arm swing [28]. These features can be explicitly measured through wearable sensors [14]; however, this is not practical in remote biometrics applications. Thus, vision-based gait recognition has become more popular [22, 23] because of its non-intrusive nature and its potential to identify subjects at a distance.

There have been many indoor datasets for vision-based gait recognition [11, 32]. However, performance on these has started to saturate [13, 22] and focus is shifting towards the outdoor, in-the-wild gait recognition task [3, 35]. Such
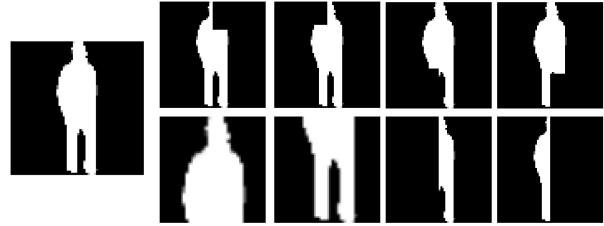
Figure 1. Examples of synthetic occlusions used in our method. The leftmost image is an original frame taken from the BRIAR [3] dataset, and the other images show the synthetic occlusions applied. We use eight types of synthetic occlusions; placing black patches on any corner of the frame (#1-#4), removing the top or bottom part of the frame (#5-#6), or blocking out the left or right half of the frame(#7-#8). The size and position of the patch is decided randomly for each video. The subject has consented to the use of these images.

uncontrolled sequences pose many challenges to gait recognition, such as severe turbulence, changes in illumination, extreme altitudes and viewpoints. Because the camera can be at a large distance from the subject, the quality of the image can not be guaranteed in such in-the-wild data. One of the important challenges in uncontrolled sequences is occlusion, where the input itself may be partially blocked due to an obstruction between the subject and the camera. Gait recognition identifies subjects by observing the correlations between the motion of different body parts, and when some of these body parts are blocked from view, it can be difficult to extract a unique gait signature. We aim to address this problem of occlusion in this work.

Occlusion has been relatively well studied in the Person Re-ID field, with large scale datasets specifically targeting occlusion [33, 34, 36]. With only one occluded gait dataset [10], that too captured in controlled conditions, occlusion has not yet gained the attention it deserves within the gait recognition community. Due to the lack of a large scale outdoor occluded dataset, many methods either work on small data collected on their own [24] or simulate occlu-

sions on holistic datasets [4, 27, 30].

Current gait recognition techniques assume the availability of the complete body to extract features and extract correlations between the motion of different body parts. When the frame is partially occluded, these correlations get corrupted as the model looks for missing body parts. To address this issue, some attempts have been made for developing algorithms for occluded gait recognition using generative models [4, 27] which complete the occluded frames. However, generative models can often hallucinate unnecessary or irrelevant details. [30] uses 3D human mesh modeling and predicts the locations of the occluded body parts by further computations on the SMPL [16] features. However, 3D mesh modeling is not reliable from low-quality data captured in uncontrolled environments, which can further corrupt the predictions.

To tackle these issues, we propose a novel method for occluded gait recognition with intrinsic occlusion type awareness. Our method can be integrated into potentially any state-of-the-art gait recognition model to enhance performance in occluded scenarios. We argue that occlusion type awareness is important to generate discriminative gait signatures from partial, occluded inputs. Without such occlusion awareness, the gait recognition model would try to look for body parts which might not even be present in the input at all - leading to the corruption of the gait signature. Thus, we propose an auxiliary occlusion detection module, which is able to inject occlusion relevant information into the network. The occlusion detector is trained for the occlusion classification task on a variety of different occlusions as shown in Figure 1. This information is later used to guide the gait recognition backbone. Our experiments show that embedding such occlusion aware encodings into the network improves the gait recognition performance on challenging outdoor datasets like BRIAR [3] and GREW [35].

In summary, our main contributions are as follows:

- We propose a novel model-agnostic, occlusion aware method for gait recognition in uncontrolled environments which can be used with potentially any state-of-the-art method to enhance performance under occlusion.

- We design an auxiliary *occlusion detection module* to generate occlusion encodings and show that they contain useful information which can facilitate learning of effective discriminative features for occluded gait recognition.

- Through extensive experimentation, we explore various approaches to effectively utilize this occlusion information to optimize model training under occlusion scenarios.

2. Related Work

Gait Recognition has traditionally been done using wearable sensors [14] which capture various aspects of human movement like speed, stride length and body posture. With advances in computer vision techniques, focus has shifted to vision-based methods where gait recognition is performed directly using videos of the subject captured from camera sensors at a distance [6, 13, 19, 22, 23, 26]. Within vision based methods, most fall under the category of either skeleton-based [19, 26] or appearance-based [6, 13]. Skeleton-based methods extract human keypoints and use the motion of these keypoints to extract features. They rely on the performance of pose estimation techniques, which may not work as well on low-quality data captured from long range. Appearance-based methods, which generally rely on silhouette masks, employ a sequence of 2D and 3D CNNs to capture spatio-temporal information and extract gait-relevant features. Appearance-based methods usually outperform skeleton-based methods [26], and they have almost saturated performance on indoor datasets like CASIA-B [32]. More challenging outdoor datasets like GREW [35] and BRIAR [3] have been collected to further advance research in this domain. [8] shows that existing models do not achieve desirable performance when applied directly to these datasets because of challenges like atmospheric turbulence, changes in illumination and occlusions unique to these datasets.

Occluded Person Re-ID has been a relatively well-studied field, with large scale datasets and techniques [2, 7, 27, 31, 36]. Some works simulate different types of occlusions in clever ways to train their networks [2, 31], while others generate body-part based representations to handle cases when some parts are missing [25]. However, most of these methods do not utilize any information about the type of occlusion in the input. Some works like [7] identify the missing body parts using an occlusion awareness score as part of their architecture, which can work for an image but can not be easily extended for the video modality where the occlusion can change dynamically. On a related note, [15] computes a quality metric for multiple frames and uses this information to combine frame-wise features. However, the model itself does not become quality aware since this information is used only in the aggregation step.

Person Re-ID methods can work well when we have just a single image, but gait recognition relies on temporal information if it is available. Hence, we discuss Occluded Gait Recognition next.

Occluded gait recognition has not seen much progress, especially for outdoor, in-the-wild scenarios, mainly due to the lack of a large scale dataset targeting the problem. Most methods work using synthetic occlusions [1, 4, 17, 18, 27, 30]. Some attempts have been made to complete the occluded portions of the input sequence using generative models

[4, 27]. However, generative models can often hallucinate unnecessary and irrelevant details. [17] generates a training matrix of gait features for different occlusions and uses it during verification. [4] performs binary classification on the input sequence to detect any occlusions and later reconstructs the Gait Energy Image [9]. However, the type of occlusion is not considered during reconstruction. Works like [1, 18] try to generate better GEI variants for occlusion, but fail to utilize the temporal information effectively in the process of generating the GEI. Some other methods [30] identify and predict the positions of missing body parts in the input using 3D mesh modeling, but they require high quality data which is not always available in outdoor scenes; subsequent prediction of missing keypoints from a noisy mesh model may lead to more errors.

To tackle the shortcomings of existing methods, we present an occlusion aware approach for performing gait recognition, which can be used with potentially any state-of-the-art method to enhance performance in outdoor, in-the-wild conditions. We describe our method in detail in the next section.

3. Method

The main aim of this work is to show that occlusion-guided training yields better results when training a gait recognition network on occluded data. Given a (possibly occluded) video sequence $V^i = \{v_1, v_2, ...v_f\}$ of size $f \times H \times W$ for subject i, the goal is to find an identifying gait signature γ_i for the subject which may be used for downstream tasks.

Our approach can be divided into four main steps: 1) The preprocessing stage, where we extract silhouette masks and introduce synthetic occlusions in the videos; 2) The occlusion detector pretraining step, where we train the occlusion detector $\mathcal{D}$ on the classification task with synthetic occlusions; 3) The gait recognition backbone $\mathcal{F}$, which extracts gait features from the input; and 4) The Occlusion Awareness Module $\mathcal{M}$, which generates occlusion aware features and inserts them into the backbone. An overview of these components is described in the next section.

3.1. Overview

An overview of the proposed method is shown in Figure 2. To ensure that only gait is captured, we first preprocess V^i to generate silhouette masks if the original video belongs to RGB or grayscale modality. Next, we pass the masked video to the occlusion detector $\mathcal{D}$, which outputs the occlusion aware feature β containing information about the type of occlusion in the video. In a parallel branch, the gait recognition backbone $\mathcal{F}$ is also fed the preprocessed video V^i. The occlusion feature β is inserted in the intermediate layers of $\mathcal{F}$ to guide the feature generation and generate an occlusion aware gait signature γ.

3.2. Preprocessing and Synthetic Occlusions

To ensure that only gait is captured, we extract silhouette masks from the RGB videos using Detectron2 [29] for the BRIAR dataset which has RGB videos. We adopt the widely used method of centering, cropping and resizing these masks as done in [5].

We train the occlusion detector and the gait recognition backbone on various kinds of synthetic occlusions, as shown in Figure 1. To simulate occlusion, consistent black patches are placed on the input frames of a video. To better model practical scenarios, the amount of occlusion is chosen randomly between a range R. Where the frame's top or bottom half is completely blocked out, we resize the visible part of the subject to the complete frame to simulate how the occluded subject would be detected during real occlusions. Overall, we use eight broad categories of synthetic occlusions; placing a black patch on any of the four corners of the frame (#1-#4), removing a random portion of the top or bottom part of the frame (#5-#6), or removing the left or right half of the frame (#7-#8). These eight categories of occlusion along with the non-occluded category give a total of nine classes the occlusion detector $\mathcal{D}$ is trained on. While it is possible to train the occlusion detector on more occlusions classes, we feel that it is not practical to train it on every possible occlusion type. As described in Section 3.3, we discard the classification head of the occlusion detector and use only the penultimate features which can contain relevant information even for unseen occlusions.

3.3. Occlusion Detection Module

The occlusion detector $\mathcal{D}$ is a convolutional neural network which is trained to classify the type of occlusion in the input video V^i. It consists of three convolutional layers and two additional linear layers, with the final layer classifying the video into nine classes - for eight types of occlusions and the non-occluded case. The network is trained on this classification task using the cross entropy loss.

The occlusion detector is trained separately from the gait recognition backbone. During its training, the classifier head is used for loss calculation. However, when $\mathcal{D}$ is used to guide the training of the gait recognition backbone $\mathcal{F}$ in the next step, this classifier head is removed. It is important to note that the weights of the occlusion detector are frozen when $\mathcal{F}$ is being trained, to ensure that $\mathcal{D}$ does not lose its intrinsic occlusion awareness.

When guiding $\mathcal{F}$, the occlusion detector can operate in two modes - 1) Cumulative occlusion detection and 2) Transient occlusion detection, depending on the position of inserting of β in the backbone. In the cumulative mode, the occlusion detector outputs a single occlusion feature for the whole video. In the transient mode, the occlusion detector outputs different occlusion features for each frame in the input V. In summary, the occlusion detection module per-
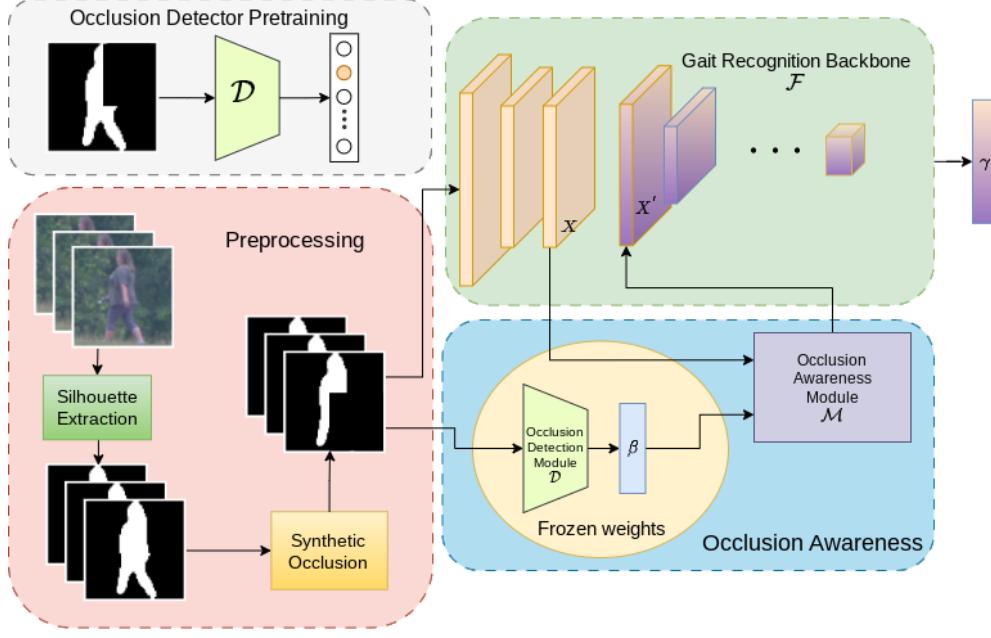
Figure 2. An overview of the proposed approach. First, silhouette masks are extracted from input video frames in the preprocessing step. Next, the video is synthetically occluded. This video is passed through the occlusion detection module to give occlusion aware embeddings β, which are subsequently used to guide the gait recognition backbone $\mathcal{F}$ in generating occlusion aware gait signatures γ. It is important to note that the occlusion detection module is frozen during training of $\mathcal{F}$ to preserve its occlusion awareness.

forms the following computation:

$$\beta = \mathcal{D}(V^i) \qquad (1)$$

where β is the occlusion aware feature containing information about the occlusion type in the video V^i. If $\mathcal{D}$ is operating in the transient mode, the features are denoted by β_t, which consists of the output of $\mathcal{D}$ for every frame. On the contrary, if $\mathcal{D}$ is operating in the cumulative mode, the features are represented by β_c and denote the output of the occlusion detector averaged across the temporal dimension. In our experiments, we observe that the position where we insert β_c into the backbone utilizes occlusion awareness more effectively than the position we insert β_t.

3.4. Gait Recognition Backbone

The Gait Recognition backbone $\mathcal{F}$ can potentially be any state-of-the-art network for gait recognition. Though state-of-the-art models can take occluded videos as input without any change, they assume the visibility of the entire body. As a result, the performance is impacted when complete gait information is not present since the network is not able to capture the correlations of movement between the visible and occluded body parts.

To address this problem, we use the occlusion detector $\mathcal{D}$ to guide the training process of $\mathcal{F}$. Occlusion aware features are inserted into the forward pass of $\mathcal{F}$ so that it can identify the missing body parts and calculate the cor-

relations between the visible and occluded portions accordingly. The occlusion aware features are provided by the occlusion awareness module $\mathcal{M}$, as described in the next section.

3.5. Occlusion Awareness Module

The occlusion awareness module $\mathcal{M}$ is used to insert occlusion information into the gait recognition backbone $\mathcal{F}$. It takes the output of the occlusion detector β, and combines it with intermediate features in $\mathcal{F}$; this is used to guide the gait recognition backbone. Both $\mathcal{F}$ and $\mathcal{M}$ are trained together end to end, while weights of the occlusion detector $\mathcal{D}$ are frozen. The generation of occlusion aware features is described by the following equation:

$$X^{'} = \mathcal{M}(X, \beta) \qquad (2)$$

where X is the intermediate feature computed by $\mathcal{F}$, β is the occlusion awareness feature from $\mathcal{D}$, and $X^{'}$ is the new occlusion aware intermediate feature which is replaced by X in the backbone. Thus, occlusion information is injected into the forward pass of $\mathcal{F}$ by the occlusion awareness module $\mathcal{M}$ according to Equations 1 and 2.

We experiment with different functions for $\mathcal{M}$ as described in Section 5.3, namely adding β to X, concatenating X and β and passing them through a learnable transformation, and choosing different positions of X from the backbone. We find that introducing a learnable layer in the inter-

Figure 3. Some sample frames taken from the GREW [35] dataset. Only the silhouette masks have been publicly released due to privacy concerns.
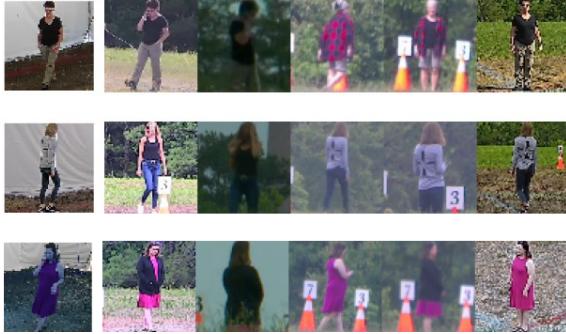
Figure 4. Examples of original video frames taken from the BRIAR [3] dataset. The leftmost column shows images taken from the indoor sequences used in the gallery set. Next columns have images taken from 100m, 200m, 400m, 500m, and elevated angle respectively for the same subject. Illumination changes and turbulence make the task challenging. These subjects have consented to use of these images in publication.

mediate layer of $\mathcal{F}$ to insert X', which we call the Deferred Concat method, yields the best performance.

4. Experiments

4.1. Datasets

BRIAR [3] is a recently collected outdoor dataset containing many challenging variations in clothing, illumination conditions, viewpoints, distances and altitudes. Videos are captured at multiple distances from different camera sensors, including some aerial videos. Some videos are captured in controlled environments from high quality cameras as well, and these are used as gallery sequences during evaluation. The outdoor videos captured from long range and UAVs are of low quality because of atmospheric turbulence and are used as the probe set. Furthermore, the clothing of the subjects is different in the gallery and probe. Our subset of the BRIAR dataset contains a total of 212 training subjects and 90 test subjects, both sets being mutually exclusive. Some examples of the original frames taken from BRIAR are presented in Figure 4. Additional details have been included in the supplementary material.

GREW [35] 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. Only the preprocessed silhouettes are publicly available due to privacy concerns. Some example frames from the GREW dataset are shown in Figure 3.

4.2. Data preprocessing

To ensure that only gait is used and appearance features are filtered out, we calculate segmentation masks using Detectron 2 [29] on all the input frames for the BRIAR dataset. We crop, centre and resize all these silhouette masks to a standard $H \times W = 64 \times 64$ size.

For the GREW dataset, we follow the widely adopted preprocessing technique as in [5], where we crop the images to include only the bounding box around the subject and resize it to 64×64. Since the GREW dataset is already in the form of silhouette masks, we do not run any segmentation model on it.

4.3. Backbones and Baselines

We experiment with three different state-of-the-art models for our backbone $\mathcal{F}$, namely GaitGL [13], GaitPart [6] and GaitBase [5]. All of these are CNN-based networks. We use the implementations of these models from the Open-Gait [5] repository. All backbones are trained using a partwise triplet loss as done in [5], with GaitGL and GaitBase having an additional cross-entropy loss.

We perform experiments with two different types of baselines as mentioned in Tables 1 and 2. Baseline-1 is when we train the backbone $\mathcal{F}$ on complete videos of the subjects without any occlusion, and perform zero-shot evaluation on occluded sequences. Baseline-2 has these same architectures which are trained and evaluated on similar kind of synthetically occluded data.

4.4. Implementation Details

The gait recognition backbone $\mathcal{F}$ accepts a video of silhouette masks. We set the size of the input frames to 64×64. During training, we also fix the number of input frames f to 30 during training. The occlusion detection module outputs an occlusion aware embedding β for the video. Depending on the mode of operation of the occlusion detection module, β can be a single 1-D vector of size 64 (in the cumulative mode), or a 2-D vector of size $f \times 64$, where f is the number of frames in the input (in the transient mode). β is later fed into $\mathcal{F}$ at different positions as described in Section 5.3.

We use the PyTorch [20] framework to conduct our experiments. We use the backbone models provided in the OpenGait [5] repository. We train our baselines and occlusion aware models with the Adam optimizer [12], a learning rate of 0.0001 and a batch size of $(8, 8)$ for training GaitGL and GaitPart and $(32, 4)$ for GaitBase.

4.5. Evaluation

BRIAR [3] dataset provides the evaluation protocol in terms of the gallery and probe set split. The gallery set consists of high quality controlled videos captured indoors, while the probe set contains more challenging outdoor conditions. More details about the evaluation protocol have been included in the supplementary material. We compute the top-K rank retrieval accuracy at different distances for evaluation on BRIAR.

For the **GREW** [35] dataset, we follow the standard protocol mentioned in Section 4.1, and follow [5] for computing top-K rank retrieval accuracy.

Multiple evaluations: Since our evaluation results are based on randomly chosen occlusions, we repeat the evaluation ten times and report the mean and the standard deviation across these multiple runs to check the effect of randomness in our main experiments.

5. Results and Discussion

5.1. Occlusion Detector

The occlusion detector is trained separately from the backbone $\mathcal{F}$ for classifying the occlusion type. It is able to identify the correct occlusion type with 98% accuracy on the BRIAR test set. We only train it on the relatively smaller BRIAR dataset and use it directly for our experiments with GREW to demonstrate its robustness across domains for the task of occlusion detection. Additional cross-domain evaluation results for the occlusion detector are included in the supplementary material.

5.2. Occlusion Aware Gait Recognition

Our results are summarized in Tables 1 and 2. Baseline-1, which refers to zero-shot evaluation on occluded sequences, generally performs the worst since the model has never seen occluded sequences during training. Baseline-2 performs slightly better but is still not optimal, since it is difficult for a single network to work on different occlusion types on its own. Our results show that using an auxiliary occlusion detector to introduce occlusion awareness helps the model to generate better identity features under occlusions.

Our results also show that even though occlusions are chosen randomly during evaluation, the results are mostly consistent across multiple evaluations, with the exception of the aerial condition on BRIAR which deviates significantly from the mean. This is because only thirty subjects have aerial videos out of the ninety test subjects, and the small dataset size causes randomness to become a larger factor in evaluation.

We also observe that Baseline-1 performs better than Baseline-2 for the GaitGL backbone. Since these are data-driven methods, we do not have a proper explanation for this

behavior; this may be because it is difficult to optimize the slightly larger GaitGL model on the turbulent BRIAR data with synthetic occlusions. However, the occlusion aware network still performs better than both the two baselines.

5.3. Analysis and Ablations

We perform various experiments to analyze the performance of the model with different occlusion awareness methods and on different occlusions. All experiments in this section are done using the GaitGL backbone on the GREW dataset, unless stated otherwise.

5.3.1 Optimal location for occlusion awareness

In this section, we experiment with variants of the occlusion awareness module $\mathcal{M}$ to identify the optimal method for introducing occlusion awareness in the backbone and producing the features X'. The results are presented in Tables 3 and 4, and are discussed in this section.

Guided Add: The first variant we tried adds the outputs of the occlusion detector directly to the second convolution layer of the CNN-based backbone, as described by:

$$X' = X + \beta_t \tag{3}$$

where β_t means that the occlusion detector operates in the transient mode, generating a β feature for each frame. However, we observe that this method does not perform well. Even though β_t contains occlusion information, it may not be directly compatible to be used with the backbone features X as they come from different networks. Hence, we introduce a learnable transformation over these features in the next experiment to make the features compatible.

Learnable 3DConv: In this variant, instead of adding X and β_t, they are concatenated across the channel dimension and passed through a learnable 3D convolutional layer. It is important to note that the occlusion detector is still not trained, but an additional transformation layer is introduced so that both X and β_t become compatible, resulting in better performance. This method is described by:

$$X' = \mathcal{M}(X \oplus \beta_t) \tag{4}$$

where $\oplus$ denotes the concatenation operation across the channel dimension and the occlusion awareness module $\mathcal{M}$ is a learnable 3D convolutional layer.

Deferred Concat: Here, we experiment with a different position to introduce occlusion awareness in the backbone. Specifically, we introduce the occlusion aware features at the *Head0* layer of GaitGL and GaitPart, and FCs layer of GaitBase, as defined in the OpenGait repository [5]. These layers are one of the deeper layers of the network, which is why we call this method Deferred Concat. We follow a similar procedure as Learnable 3D Conv, concatenating the output across the channel dimension as shown by:

$$X' = \mathcal{M}(X \oplus \beta_c) \tag{5}$$

Backbone	Method	100m		400m		500m		Elevated Angle		Aerial	
		Rank-1	Rank-20	Rank-1	Rank-20	Rank-1	Rank-20	Rank-1	Rank-20	Rank-1	Rank-20
GaitBase [5]	Baseline-1	16.34 ± 1.934	50.34 ± 2.538	13.72 ± 1.263	48.98 ± 2.940	10.96 ± 1.422	46.33 ± 3.582	14.06 ± 1.394	48.17 ± 3.329	14.76 ± 7.680	49.84 ± 15.409
	Baseline-2	24.50 ± 1.738	83.59 ± 1.661	20.00 ± 2.070	78.67 ± 1.829	16.62 ± 1.925	75.39 ± 2.128	24.60 ± 1.277	79.30 ± 1.242	23.81 ± 6.809	71.59 ± 7.185
	Occlusion Aware	**31.99 ± 2.593**	**86.63 ± 1.591**	**24.79 ± 1.875**	**80.89 ± 1.665**	**23.38 ± 1.765**	**78.60 ± 1.915**	**29.58 ± 0.865**	**82.32 ± 0.837**	**32.38 ± 5.226**	**78.57 ± 5.418**
GaitGL [13]	Baseline-1	12.76 ± 1.434	63.24 ± 2.094	6.43 ± 0.984	53.39 ± 3.338	6.62 ± 1.439	51.74 ± 2.130	9.12 ± 0.843	56.94 ± 1.091	9.37 ± 2.879	59.84 ± 7.206
	Baseline-2	11.58 ± 1.186	65.60 ± 2.586	5.19 ± 1.357	50.21 ± 2.281	4.31 ± 0.828	50.17 ± 2.219	7.52 ± 0.640	59.11 ± 2.024	4.60 ± 3.363	65.08 ± 10.836
	Occlusion Aware	**34.66 ± 1.549**	**81.81 ± 1.485**	**20.14 ± 1.366**	**70.00 ± 1.729**	**16.90 ± 1.546**	**70.91 ± 2.204**	**26.72 ± 1.506**	**75.29 ± 1.232**	**26.83 ± 8.413**	**81.27 ± 3.237**
GaitPart [6]	Baseline-1	4.65 ± 0.717	45.41 ± 1.186	2.32 ± 0.414	34.49 ± 1.472	2.21 ± 0.673	35.03 ± 1.519	3.85 ± 0.661	43.01 ± 1.201	4.60 ± 3.131	50.63 ± 5.958
	Baseline-2	12.55 ± 1.424	73.12 ± 1.871	5.59 ± 0.940	51.18 ± 2.515	5.14 ± 1.230	48.81 ± 1.970	9.14 ± 0.617	64.49 ± 2.397	10.16 ± 5.079	77.14 ± 5.776
	Occlusion Aware	**21.17 ± 1.913**	**78.80 ± 1.956**	**11.58 ± 1.146**	**66.91 ± 1.420**	**11.57 ± 1.303**	**65.35 ± 2.302**	**15.59 ± 0.977**	**72.72 ± 1.400**	**11.59 ± 5.769**	**75.87 ± 6.098**

Table 1. Comparison of the baselines and occlusion aware networks on the BRIAR [3] dataset on synthetic occlusions. Rank-K retrieval accuracies averaged across ten evaluation runs are reported for different ranges along with the standard deviation. Baseline-1 is zero shot evaluation on occluded data, Baseline-2 is the network which has seen synthetic occlusions during training. We can see the training on synthetic occlusions with occlusion awareness performs the best for occluded data.

Backbone	Method	Rank-1	Rank-5	Rank-10	Rank-20
GaitBase [5]	Baseline-1	15.15 ± 0.263	25.34 ± 0.437	30.67 ± 0.465	36.41 ± 0.385
	Baseline-2	27.26 ± 0.554	42.66 ± 0.478	49.81 ± 0.425	57.08 ± 0.476
	Occlusion Aware	**34.57 ± 0.522**	**50.73 ± 0.486**	**57.93 ± 0.488**	**64.82 ± 0.396**
GaitGL [13]	Baseline-1	5.95 ± 0.327	10.40 ± 0.313	12.97 ± 0.316	15.78 ± 0.370
	Baseline-2	10.59 ± 0.177	20.73 ± 0.319	26.96 ± 0.435	34.40 ± 0.523
	Occlusion Aware	**13.77 ± 0.280**	**25.91 ± 0.383**	**32.70 ± 0.506**	**40.60 ± 0.594**
GaitPart [6]	Baseline-1	5.18 ± 0.237	9.99 ± 0.375	12.76 ± 0.359	16.00 ± 0.396
	Baseline-2	8.71 ± 0.323	18.11 ± 0.423	23.79 ± 0.442	30.56 ± 0.399
	Occlusion Aware	**13.77 ± 0.441**	**25.84 ± 0.329**	**32.42 ± 0.484**	**39.90 ± 0.627**

Table 2. Comparison of the baselines and the occlusion aware networks on the GREW [35] dataset after introducing synthetic occlusions. Rank retrieval accuracies averaged across ten evaluation runs are reported along with the standard deviation. We can see that occlusion awareness helps across different backbones.

where X are the features taken from the deeper *Head0* or *FCs* layer and X' are the obtained occlusion aware features. Since these are flat layers, $\mathcal{M}$ is also a learnable fully connected layer. Since this layer does not contain any temporal information, the occlusion detector operates in the Cumulative mode in this method, generating a single occlusion feature for the entire video denoted by β_c. We observe that this performs better than Learnable 3D Conv, and we think it is because occlusion awareness may be more useful in the later layers when the network looks at the global scale of the image, instead of the earlier low level layers.

Deferred Concat + 3D Conv: In this experiment, we inject occlusion information at multiple points in the backbone $\mathcal{F}$. To do this, we combine both the above two methods, namely Learnable 3D Conv and Deferred Concat. We observe that performance actually deteriorates compared to using just the Deferred Concat method, as the model can get distracted with too much occlusion information.

Complex Deferred Concat: In this experiment, we add an additional linear layer to the deferred concat awareness module $\mathcal{M}$ to increase the complexity of the occlusion aware features X'. For the GREW dataset, we observe that this improves performance, but not for the BRIAR dataset. We think this is because the larger GREW dataset requires

more complex representations of occlusion awareness because of its larger size, but the complex model may overfit on the relatively smaller BRIAR dataset.

5.3.2 Optimal location across backbones

All experiments in the previous section were performed using the GaitGL [13] backbone. Here, we experiment with another backbone to see if the observations about the optimal location for occlusion awareness hold true across different gait recognition models. Thus, we perform another experiment with GaitPart [6], the results of which are shown in Table 5. We observe that the Deferred Concat method still works better. We conjecture that for arbitrary CNN backbones, inserting occlusion aware features would work better in the later layers since the model can extract local level features in the earlier layers and consolidate these with occlusion awareness in the later layers.

5.3.3 Generalization to unseen occlusions

We train and evaluate our models on a fixed set of broad occlusion types, since we feel that training a model on all possible occlusion types is not practical. We try to verify whether this approach generalizes to different occlusions by evaluating these trained models directly on unseen occlusion types. Specifically, we evaluate our approach on

Occlusion Awareness Method	100m		400m		500m		Extreme Angle		Aerial	
	Rank1	Rank20	Rank1	Rank20	Rank1	Rank20	Rank1	Rank20	Rank1	Rank20
Guided Add	21.31	82.76	12.36	70.47	13.38	71.97	16.65	73.55	15.87	60.32
Learnable 3DConv	27.3	81.05	14.15	72.39	13.38	74.31	21.37	73.55	19.05	82.54
3D Conv + Deferred Concat	15.52	73.23	7.28	58.79	5.1	52.65	9.8	65.54	3.17	74.6
Complex Deferred Concat	32.44	79.87	16.48	67.03	14.86	66.45	25.91	75.6	31.75	87.3
Deferred Concat	34.58	82.12	21.15	70.19	18.47	70.91	25.73	78.27	28.57	82.54

Table 3. A comparison of different methods of injecting occlusion awareness into the gait recognition backbone, on the BRIAR [3] dataset. We observe that the Deferred Concat method performs best and adding complexity to the occlusion awareness model may help in some cases. Also, we see that too much occlusion information can distract the network as observed from the 3D Conv + Deferred Concat row.

Occlusion Awareness Method	Rank-1	Rank-5	Rank-10	Rank-20
Guided Add	7.73	16.27	21.57	28.23
Learnable 3DConv	8.13	17.47	22.92	29.77
3D Conv + Deferred Concat	11.73	22.05	28.78	35.88
Deferred Concat	13.9	26.2	32.68	40.68
Complex Deferred Concat	14.25	26.48	33.08	40.3

Table 4. A comparison of different occlusion awareness methods on the synthetically occluded GREW dataset. Too much occlusion information distracts the network as seen from the third row. The Complex Deferred Concat method works better here because of the larger size of the GREW dataset.

	Rank-1	Rank-5	Rank-10	Rank-20
Learnable 3D Conv	12.45	25.07	31.60	39.40
Deferred Concat	13.93	26.00	32.73	40.33

Table 5. Inserting occlusion awareness in different positions in the GaitPart backbone, evaluated on GREW. We observe that even in GaitPart, occlusion awareness is more useful in the deeper layers using the Deferred Concat method.

Method	Synthetic Dynamic Occlusions				Real Jagged Occlusions	
	Rank-1	Rank-5	Rank-10	Rank-20	Rank-1	Rank-20
Baseline - 2	20.75	34.00	39.87	45.85	4.94	51.33
Occlusion aware	25.18	40.2	46.37	52.32	23.86	85.78

Table 6. Generalizability to unseen occlusions with the GaitGL backbone with and without occlusion awareness. Synthetic dynamic occlusions are applied on GREW and real jagged occlusions are used from BRIAR.

synthetic dynamic occlusions and also on real jagged occlusions present in the BRIAR dataset. Descriptions about these new occlusion types have been included in the supplementary material. Our results, presented in Table 6, show that our approach can generalize to unseen occlusions as well. This is because even though $\mathcal{D}$ is trained on nine occlusion classes, its penultimate layer can still contain occlusion relevant information about which body parts are present in the input. This occlusion awareness can be utilized by the backbone to generate better features.

6. Limitation and Future Work

We demonstrate that occlusion type awareness helps in generating more discriminative features for occluded gait recognition. Our current method identifies the occlusion type and passes along the information to guide the gait recognition backbone. We believe that introducing changes within the backbone architecture itself will help further since their current design assumes the availability of the full body. Additionally, we mostly deal with synthetic occlusions in this work, and it is still not clear how well these synthetic occlusions can simulate real world occlusions. As such, datasets specifically addressing the occlusion problem in gait recognition are needed to further advance research in this area. Lastly, we only experiment with gait recognition methods on silhouettes. We believe that exploiting all modalities and combining Person Re-ID and face recognition methods with gait recognition can improve recognition performance and we leave this to future work.

7. Conclusion

Existing gait recognition methods are not effective when occlusions are present in the input. Thus, in this work, we propose a model-agnostic approach to introduce *intrinsic occlusion awareness* into existing models to enhance their performance in occlusion scenarios. We perform experiments on the BRIAR and GREW datasets, both of which contain uncontrolled outdoor sequences. We train a domain-robust auxiliary occlusion detector and use it to inject occlusion aware features into three different state-of-the-art gait recognition backbones. We find that occlusion awareness is best used in the intermediate layer of the backbones. We compare our method with two different baselines and conclude that intrinsic occlusion awareness can help in increasing performance on occluded gait recognition.

8. Acknowledgement

References

[1] Frame difference energy image for gait recognition with incomplete silhouettes. *Pattern Recognition Letters*, 30(11):977–984, 2009. 2, 3

[2] 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. 2

[3] 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, 7, 8

[4] Dhritimaan Das, Ayush Agarwal, and Pratik Chattopadhyay. Gait Recognition from Occluded Sequences in Surveillance Sites. In Leonid Karlinsky, Tomer Michaeli, and Ko Nishino, editors, *Computer Vision – ECCV 2022 Workshops*, Lecture Notes in Computer Science, pages 703–719, Cham, 2023. Springer Nature Switzerland. 2, 3

[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. 3, 5, 6, 7

[6] 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. pages 14225–14233, 2020. 2, 5, 7

[7] Guangyu Gao, Qianxiang Wang, Jing Ge, and Yan Zhang. AONet: Attentional Occlusion-Aware Network for Occluded Person Re-identification. In Lei Wang, Juergen Gall, Tat-Jun Chin, Imari Sato, and Rama Chellappa, editors, *Computer Vision – ACCV 2022*, volume 13845, pages 21–36. Springer Nature Switzerland, Cham, 2023. Series Title: Lecture Notes in Computer Science. 2

[8] 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

[9] J. Han and Bir Bhanu. Individual recognition using gait energy image. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(2):316–322, 2006. 3

[10] Martin Hofmann, Jürgen Geiger, Sebastian Bachmann, Björn Schuller, and Gerhard Rigoll. The tum gait from audio, image and depth (gaid) database: Multimodal recognition of subjects and traits. *Journal of Visual Communication and Image Representation*, 25(1):195–206, 2014. 1

[11] H. Iwama, M. Okumura, Y. Makihara, and Y. Yagi. The ouisir gait database comprising the large population dataset and performance evaluation of gait recognition. *IEEE Trans. on Information Forensics and Security*, 7, Issue 5:1511–1521, Oct. 2012. 1

[12] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980, 2014. 5

[13] 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, 5, 7

[14] Xin Liu, Chen Zhao, Bin Zheng, Qinwei Guo, Xiaoqin Duan, Aziguli Wulamu, and Dezheng Zhang. Wearable Devices for Gait Analysis in Intelligent Healthcare. *Frontiers in Computer Science*, 3, 2021. 1, 2

[15] Yu Liu, Junjie Yan, and Wanli Ouyang. Quality aware network for set to set recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5790–5799, 2017. 2

[16] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. *SMPL: A Skinned Multi-Person Linear Model*. Association for Computing Machinery, New York, NY, USA, 1 edition, 2023. 2

[17] Daigo Muramatsu, Yasushi Makihara, and Yasushi Yagi. Gait regeneration for recognition. In *2015 International Conference on Biometrics (ICB)*, pages 169–176. IEEE, 2015. 2, 3

[18] Javier Ortells, Ramón A Mollineda, Boris Mederos, and Raúl Martín-Félez. Gait recognition from corrupted silhouettes: a robust statistical approach. *Machine Vision and Applications*, 28:15–33, 2017. 2, 3

[19] Honghu Pan, Yongyong Chen, Tingyang Xu, Yunqi He, and Zhenyu He. Toward complete-view and high-level pose-based gait recognition. *IEEE Transactions on Information Forensics and Security*, 18:2104–2118, 2023. 2

[20] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. In *NIPS-W*, 2017. 5

[21] Todd C. Pataky, Tingting Mu, Kerstin Bosch, Dieter Rosenbaum, and John Y. Goulermas. Gait recognition: highly unique dynamic plantar pressure patterns among 104 individuals. *Journal of the Royal Society Interface*, 9(69):790–800, Apr. 2012. 1

[22] Alireza Sepas-Moghaddam and Ali Etemad. Deep Gait Recognition: A Survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(01):264–284, Jan. 2023. Publisher: IEEE Computer Society. 1, 2

[23] Chuanfu Shen, Shiqi Yu, Jilong Wang, George Q. Huang, and Liang Wang. A Comprehensive Survey on Deep Gait Recognition: Algorithms, Datasets and Challenges, June 2022. arXiv:2206.13732 [cs]. 1, 2

[24] Jasvinder Pal Singh, Sakshi Arora, Sanjeev Jain, and Uday Pratap Singh SoM. A multi-gait dataset for human recognition under occlusion scenario. In *2019 International Conference on Issues and Challenges in Intelligent Computing Techniques (ICICT)*, volume 1, pages 1–6, 2019. 1

[25] Vladimir Somers, Christophe De Vleeschouwer, and Alexandre Alahi. Body part-based representation learning for occluded person re-identification. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1613–1623, 2023. 2

[26] Torben Teepe, Johannes Gilg, Fabian Herzog, Stefan Hormann, and Gerhard Rigoll. Towards a Deeper Understanding of Skeleton-based Gait Recognition. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1568–1576, New Orleans, LA, USA, June 2022. IEEE. 2

[27] Md. Zasim Uddin, Daigo Muramatsu, Noriko Takemura, Md. Atiqur Rahman Ahad, and Yasushi Yagi. Spatio-temporal silhouette sequence reconstruction for gait recognition against occlusion. *IPSJ Transactions on Computer Vision and Applications*, 11(1):9, Nov. 2019. 2, 3

[28] David M Wert, Jennifer Brach, Subashan Perera, and Jessie M VanSwearingen. Gait biomechanics, spatial and temporal characteristics, and the energy cost of walking in older adults with impaired mobility. *Physical therapy*, 90(7):977–985, 2010. 1

[29] Yuxin Wu, Alexander Kirillov, Francisco Massa, Wan-Yen Lo, and Ross Girshick. Detectron2. https://github.com/facebookresearch/detectron2, 2019. 3, 5

[30] 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. 2, 3

[31] Cheng Yan, Guansong Pang, Jile Jiao, Xiao Bai, Xuetao Feng, and Chunhua Shen. Occluded Person Re-Identification with Single-scale Global Representations. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 11855–11864, Montreal, QC, Canada, Oct. 2021. IEEE. 2

[32] 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. 1, 2

[33] Zhedong Zheng, Zhi Bie, Yifan Sun, Jingdong Wang, Chi Su, and Shengjin Wang. Partial person re-identification, 2018. 1

[34] Zhedong Zheng, Shaogang Gong, and Tao Xiang. Person re-identification by probabilistic relative distance comparison. In *2011 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 649–656. IEEE, 2011. 1

[35] Tian Yang Junjie Huang Jiankang Deng Guan Huang Dalong Du Jiwen Lu Jie Zhou Zheng Zhu, Xianda Guo. Gait recognition in the wild: A benchmark. In *IEEE International Conference on Computer Vision (ICCV)*, 2021. 1, 2, 5, 6, 7

[36] Jiaxuan Zhuo, Zeyu Chen, Jianhuang Lai, and Guangcong Wang. Occluded person re-identification. *arXiv preprint arXiv:1804.02792*, 2018. 1, 2