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

Supplementary Material

8. Introduction

In this supplementary material section, we first provide more details about the BRIAR dataset and the synthetic occlusions we use in our experiments. Next, we provide more details for implementing VEN and the Mimic networks. We also provide further clarification regarding our evaluation metrics, and the evaluation protocol we use on the GREW dataset. We also evaluate VEN on the occlusion classification and the occlusion amount regression task, along with discussing the additional overhead introduced by VEN. Further, we provide more information regarding the reproducibility of our experiments, and we also evaluate our method on multiple different occlusion settings. Next, we provide some results on indoor gait recognition datasets like CASIA-B and OUMVLP. Lastly, we discuss some failure cases of our model.

9. BRIAR Dataset

BRIAR [2] is a dataset collected for Person Re-ID in challenging outdoor conditions. It comprises of both images and video modalities, however, we utilize only the videos for this work. Some videos contain only the face information, which we discard while evaluation our gait recognition approach.

The subset of BRIAR we utilize in our experiments comprises approximately 60,000 videos for training and 10,000 videos for testing, each with a duration ranging from 1 to 2 minutes, recorded at 30 frames per second (fps). This extensive collection of videos offers a diverse array of gait sequences captured under various conditions, enabling comprehensive training and evaluation of gait recognition models.

The dataset encompasses a wide range of viewpoints and distances, including indoor controlled environments, close-range elevated viewpoints, and aerial perspectives captured by Unmanned Aerial Vehicles (UAVs). Distances from the subjects to the cameras span from 100 meters to 1000 meters, introducing varying levels of spatial resolution and turbulence challenges associated with long-distance capture. Furthermore, the use of UAVs with moving cameras adds another layer of complexity to the dataset.

Each distance category in the dataset employs different sensors, capturing gait sequences in both black-and-white and color formats. Even within the color spectrum, different cameras introduce variations in image quality, contrast, color balance, dynamic ranges, and lens distortions. Some of these variations can be seen in Fig. 3 of the main paper. In Fig. 5, we visualize some more frames captured from the BRIAR dataset, including some examples from the controlled indoor sequences. This huge diversity in the image quality necessitates robustness in gait recognition models to a large number of such variations.

The dataset encompasses a wide range of walking conditions, including random walks, structured walks, carrying a large cardboard box, wearing backpacks, using cellphones while walking, and even scenarios where subjects point at cameras while walking. These diverse conditions introduce variations in gait patterns, postures, and object interactions, enhancing the dataset's realism and applicability to real-world scenarios.

BRIAR also provides a predefined evaluation protocol along with a probe-gallery split for assessing gait recognition performance. Indoor sequences captured in controlled environments serve as the gallery set, while outdoor sequences, presenting more challenging scenarios, are designated as probe samples. Notably, standing sequences are excluded from the evaluation protocol to focus on walking-based gait recognition.

Different sets of videos of the same subject in different clothing are also collected. Further, some videos have significant occlusions present where the lower portion of the subject is not visible. With these large variations in acquisition conditions, atmospheric turbulence and changes in illumination introduced by long ranges and different camera sensors, the outdoor portion of the dataset is the more challenging part and is meant to be used as the probe set.

In the indoor setups the subjects have different clothing conditions and either a random or structured way of walking. However, there are many more viewpoint variations in these indoor environments. But the indoor data has much smaller variations in illumination, turbulence, and noise compared to the outdoor dataset, and is thus meant to be used as the gallery set.

Only subjects who have explicitly consented to appear in the videos are included in the dataset, ensuring compliance with ethical guidelines and data protection regulations. Furthermore, subjects are given the option to decide whether their images can be used for publication purposes, with only those consenting to both inclusion in the dataset and publication being featured in visualizations within the paper.

10. Synthetic Occlusions

Dynamic Occlusions: To simulate dynamic occlusions, we place moving patches of various sizes in the video.

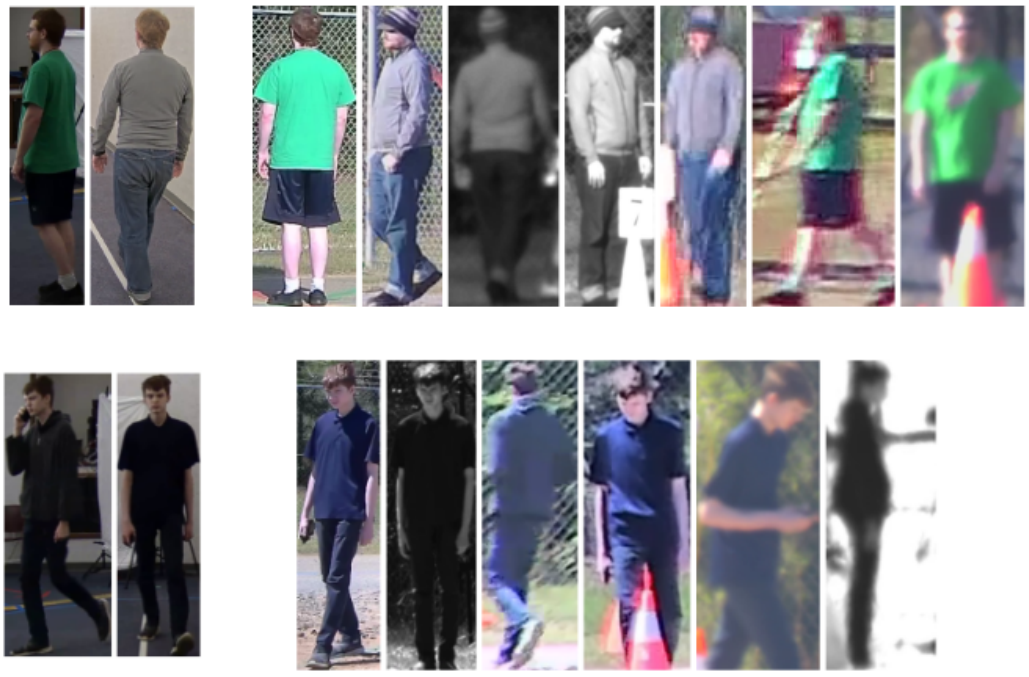

Figure 5. Some more sample frames taken from videos present in the BRIAR dataset. Subjects have consented to use of these images. Each row consists of images of one subject. The two leftmost images in both rows show examples of the indoor, controlled gallery sequences. The remaining images are captured outdoors, and they make up the probe set. As the distance increases from left to right, the quality of the frames drops significantly.

Some examples have been shown in Fig. 1 of the main paper. These occlusions try to simulate small stationary/moving obstacles which might obstruct the subject from camera view as the subject moves. Stationary objects like tall grass, trees, traffic signs and poles can block the subject, but as the camera follows the subject, the occlusion pattern from these objects appears to be moving from the frame of reference of the subject.

We consider two types of patches - small rectangular patches which can not cover the entire height of the frame, or tall rectangular patches which cover the entire height of the frame. The occlusion type to be applied on the video is randomly chosen to be either a small rectangular patch, or a tall rectangular patch as shown in Fig. 6. If it is a small rectangular patch, the height and width of the patch are randomly chosen from the range R, which is set to (0.4, 0.6). If a tall rectangular patch is to be applied, the width of the patch is chosen within a different range R_t. R_t is a smaller range than R because we assume that tall objects like poles are thin and may not cover the entire width of the frame. We set R_t to be $(0.2, 0.4)$, meaning the width of the tall patch may be between 20%-40% of the width of the frame.

Since these patches move acorss the frame, the direction and speed of movement needs to be decided. The direction of motion is randomly chosen to be from left to right or right to left. For the speed of movement, we visualize patches with different speeds and empirically decide which range of patch speeds look the most realistic for dynamic occlusion. We set the range of speeds of these moving patches $R_s = (0.5, 1.0)$ pixels/frame. Thus, for each video, the speed of the patch is also selected randomly within this range.

Some more examples of the synthetic dynamic occlusions we use are shown in Fig. 6 of this supplementary material section.

Consistent Occlusions: We use three types of consistent occlusions, namely 1) top occlusion, where the torso and head of the person may be occluded; 2) bottom occlusion, where the legs and lower body may be occluded, and 3) middle occlusions, where the middle part of the body is occluded. The portion of the frame to be cropped out is chosen randomly from the fixed range R.

In dynamic and middle occlusions, the occlusion patch zeros out the pixel values of the occluded region. However, in the case of top and bottom occlusions, the occluded portion of the frame is cropped out completely. The remaining part of the frame is then resized to the fixed $H \times W = 64 \times 64$ size of the original frame. Since we work with binary silhouette masks, resizing using linear interpolation causes the output image to become 8-bit non-binary

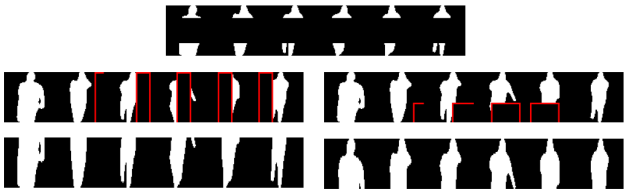

Figure 6. More visualizations of the synthetic occlusions on a video sequence taken from the GREW dataset. The top row shows middle occlusions. The second tow shows dynamic occlusions, and the bottom row shows top and bottom occlusions. The occlusion patch is shown with a red boundary in dynamic occlusions for visualization purposes only. In the synthetic dynamic occlusions, the width of the patch is more when the patch does not cover the entire height of the video. If it covers the whole height of the frame, the patch is relatively thinner. In top and bottom occlusions, the occluded portion is cropped out and the remaining frame is resized to the original height and width.

integer type. Thus, we re-binarize the resized image using a threshold of 128. This simulates how an object detector would detect a subject in the case of real consistent occlusions. Some more examples of the consistent occlusions we use are shown in Fig. 6.

11. Visibility Estimation Network

VEN, $\mathcal{V}$, is a three layer convolutional neural network with one hidden linear layer and two parallel linear heads for the classification and regression tasks. The complete architecture of $\mathcal{V}$ is presented in Tab. 5.

Layer Name	Input shape	Output Shape
Conv1	$64 \times 64 \times 1$	$64 \times 64 \times 32$
ReLU, MaxPool1	$64 \times 64 \times 32$	$32 \times 32 \times 32$
Conv2	$32 \times 32 \times 32$	$32 \times 32 \times 64$
ReLU, MaxPool2	$32 \times 32 \times 64$	$16 \times 16 \times 64$
Conv3	$16 \times 16 \times 64$	$16 \times 16 \times 128$
ReLU, MaxPool3	$16 \times 16 \times 128$	$8 \times 8 \times 128$
AdaptiveAvgPool	$8 \times 8 \times 128$	128
FC1	128	64
Classification Head	64	3
Regression Head	64	1

Table 5. The architecture of VEN. It is a three layer convolutional neural network, with one hidden linear layer, and two parallel linear heads which can predict the type and amount of occlusion in the input.

Based on how many occlusion types the mimic network is supposed to be trained on, the classification head classifies the input into the occlusion classes or the no occlusion category. The cross entropy loss is used to train VEN through the classification head, so the network gains occlusion type awareness.

The above mentioned classes are simply broad categories of occlusion types, and the amount of synthetic occlusion within one category can also vary within a range R. In our experiments, we set the range of occlusions R to be $(0.4, 0.6)$, so the output of the regression head is trained to be close to 0 when there are no occlusions in the input, and x when the amount of occlusion is x. x is sampled uniformly within the range R for each video. As seen from Tab. 4 of the main paper, the regression task helps the network gain occlusion amount awareness as well.

11.1. Additional overhead of VEN

VEN introduces additional parameters to the backbone during the inference stage - in terms of an extra convolutional network to generate occlusion relevant features. Specifically, the architecture of VEN we use introduces 0.1M additional parameters at inference time. This is relatively small compared to the gait recognition backbone - for example, GaitBase has roughly 7M parameters.

The number of additional parameters introduced is exactly the same as the occlusion detector in [13], since the only difference between VEN and [13] is the occlusion regression head which is discarded during inference. It should be noted that we are counting only the parameters used during the inference stage and not in the training stage. This means we are excluding the BNNeck layer in GaitBase, and the classification and regression heads in VEN in the numbers reported above. These layers are used only during training time.

12. Implementation Details

In this section, we elaborate on the implementation details of our method. For ease of reproducibility, we will release our source code upon acceptance.

Preprocessing: If the input video S^i is of RGB modality, we first extract binary masks from the video. For this, we use Detectron2 [36], to obtain the masks around the subject for each frame. This is done to filter out any covariates like background, color and texture hampering gait recognition.

On obtaining the silhouette masks, we center the subject and resize all the frames to a uniform size of $H \times W$ similar to [5]. In the frames where no subject is detected, we leave an empty black frame in the output video to keep the number of frames consistent.

VEN: We train VEN using the Adam optimizer [18] with a learning rate of 1e-4. During training, the classification loss L_{ce} and the regression loss L_r are multiplied by loss weights λ_{ce} and λ_r to calculate the final loss L for the backward pass as shown below

$$L = \lambda_{ce} L_{ce} + \lambda_r L_r \tag{4}$$

Empirically, we find that setting $\lambda_{ce} = 1.0$ and $\lambda_r = 10.0$ yields the best performance in the proxy tasks of occlusion classification and regression.

Mimic Network: For pretraining of $\mathcal{F}_t$, Triplet and Cross Entropy losses are used as done by [5]. In the distillation stage, the multi-instance correlational distillation loss is modelled as a TripletMarginLoss [35] with a margin $m = 0.05$.

For training GaitGL models, we randomly sample $n = 30$ contiguous frames from the full video. For GaitBase and DeepGaitV2, n is chosen uniformly between (20, 40) for each video. A batch size of (32, 4) is used for training, meaning that each batch of training data has 32 identities and 4 sequences per identity.

Apart from the randomly generated synthetic occlusions, we also use data augmentation techniques like Random Horizontal Flipping, Random Cropping and Random Perspective to train $\mathcal{F}_t$ and $\mathcal{F}_m$. We use the same data augmentation settings as used by [5].

For introducing occlusion-relevant features into the gait recognition backbone, we adopt the approach used by [13]. Specifically, we utilize the fully connected layers in the later parts of the gait recognition backbones as positions for inserting the occlusion features provided by VEN.

13. Evaluation Details

We use the Top-K rank retrieval accuracy to evaluate gait recognition performance for all datasets. In addition, we perform verification as well on the BRIAR [2] dataset, computing the True Acceptance Rate(TAR) at a False Accept Rate(FAR) of 0.01.

The respective datasets provide their protocols which tell us which video sequences should be used as probes and which ones should be used as a part of the gallery set. As described in Sec. 14, we use a local evaluation protocol for GREW introduced by [5] so that we can evaluate our models locally.

To evaluate our approach on synthetic occlusions, we use the same gallery-probe split from the dataset protocol but introduce synthetic occlusions in each video during the data loading stage. It should be noted that the occlusion type and amount is chosen for each video independently - thus, there is no correlation between the occlusions in the probe and gallery sequences of a particular subject. The probe and gallery sequences may have different types and amounts of occlusions, which makes our task formulation more challenging and better suited for practical application.

We compute the gait signatures for each element in the gallery set, and compare each probe signature with each gallery to find the Top-K subject matches. We use Euclidean distance to compare probe and gallery elements. If the true identity of the probe is present in the Top-K matches, the probe is considered to be recognized correctly. The percentage of probes recognized correctly is reported as the rank retrieval accuracy for Rank-K. This process is repeated for different values of K for more comprehensive evaluation.

14. GREW evaluation protocol

The GREW dataset [45] does not provide identity labels for their probe set. Hence, local evaluation of a model is not possible. According to the official protocol, the matching scores for each gallery and probe video have to be uploaded on the GREW competition website, which computes the accuracy of the model. This is limiting for our experiments, especially since we can not compare our results to other papers directly and have to re-train previous methods on our synthetically occluded data.

Thus, as mentioned in the main paper, we use a slightly different evaluation protocol for GREW which enables local evaluation. This protocol was introduced by [5] and has also been used in some existing works [13]. We explain this modified protocol below.

Each of the 6,000 subjects in the test set of GREW have two gait sequences, giving a total of 12,000 videos in the test set. Instead, one video of each subject is chosen as gallery and the other video is chosen as the probe, giving each subject one sequence in the gallery set. The rank-retrieval task is performed on this probe-gallery split and corresponding Rank-K metrics are computed. For the purposes of this protocol, the unlabelled videos in the 'probe' directory of GREW are ignored.

As a sanity check to confirm whether we reproduce the existing methods correctly, we take the teacher model $\mathcal{F}_t$, which is trained on complete videos, and evaluate it directly on the original GREW data without any occlusions. We evaluate this model both on the official GREW protocol and the modified protocol. The official protocol results are directly comparable to the results in the original papers, indicating we have correctly reproduced these works. The results are summarized in Tab. 6. We observe that the local evaluation protocol consistently has lower rank retrieval accuracy than when we use the official protocol. We think this is because in our local evaluation protocol, there is only one gallery sequence per subject. However, in the official protocol, there are two gallery sequences per subject - making it a bit easier for the model to match the probe to the appropriate gallery.

15. VEN Results

We train VEN on the proxy tasks of occlusion type classification and occlusion amount regression. This helps VEN to learn occlusion-relevant features, which are useful for occluded gait recognition as seen in Tab. 4 - when we remove VEN, the 'vanilla' mimic network performs worse.

However, we also perform a sanity check evaluation of VEN on these same proxy tasks themselves, similar to [13]. More specifically, we compute the classification accuracy

Rank-1/Rank-5	Local Evaluation	Official protocol
GaitBase [5]	55.3/72.1	59.9/74.7
DeepGaitV2 [4]	73.1/85.3	78.4/88.6

Table 6. Comparing the evaluation results on the official protocol (using the submission website) with our local evaluation protocol. The models are trained and evaluated on the original GREW dataset, making them identical to the teacher model $\mathcal{F}_t$ in our method. The numbers are the Rank-1/Rank-5 accuracies on the GREW dataset. The numbers in the 'official protocol' column are directly comparable to the results in the corresponding papers. We perform this experiment as a sanity check to see whether we reproduce the original methods correctly.

Accuracy/MSE		Train	
		BRIAR	GREW
Test	BRIAR	99.9/0.00060	99.7/0.00060
	GREW	99.0/0.00161	99.9/0.00031

Table 7. Cross dataset evaluation of VEN, on BRIAR and GREW datasets on top, bottom and no occlusion classes. The proxy tasks of occlusion type classification and occlusion amount regression are used to perform this evaluation. The first value of each cell represents classification accuracy of the occlusion type, the second value represents the MSE in occlusion amount prediction. VEN performs reasonably well on both these tasks, indicating that it has learned occlusion-relevant information. We can also see that VEN is robust to domain shifts, since switching to a different dataset for evaluation does not decrease the performance as much.

of the occlusion type, and the mean squared error in the regression amount for frames taken from the test set, of the same or a different dataset than it was trained on. This helps us get an idea of the robustness of VEN to domain shifts.

The results of this evaluation of VEN are shown in Tab. 7. We observe that VEN is able to perform well on these proxy tasks and is also able to generalize to other domains/datasets.

16. Additional Experiments

16.1. Reproducibility of results

Since our evaluations are based on random occlusions, we also perform multiple evaluations of our network to check whether the results are reproducible across evaluation runs. Hence, we perform 10 repeat evaluation runs on the mimic network on the GREW dataset using the GaitBase backbone. We observe a standard deviation of 0.35% in the Rank-1 accuracy, indicating there is very little change in overall performance due to the introduction of randomness in the evaluation process.

Changing occlusion types	Rank-1	Rank-5
Baseline-2	9.73	19.52
Occlusion Aware [13]	14.63	27.65
Mimic Network	16.05	29.75

Table 8. Effect of flipping the occlusion type in the middle of a video, between top and bottom occlusion cases. The mimic network is able to deal with changing occlusion types better than other methods even though it too has not seen such data during training. This further demonstrates the generalizability of our method.

Restricting Occlusion Types	Rank-1	Rank-5
Top occlusion only	30.17	47.27
Middle occlusion only	29.62	48.67
Bottom occlusion only	14.17	26.6

Table 9. Restricting the occlusion types during evaluation of the mimic network. This shows the relative difficulty of different occlusion types. We can see that bottom occlusions are the most difficult for the mimic network, since the gait is much more difficult to observe when legs are not visible.

16.2. Changing occlusions within a video

In all our previous experiments, we have assumed that the occlusion type remains the same across the video. Here, we conduct an experiment where the occlusion type can change among the frames in the video. More specifically, we flip the occlusion type from top to bottom and vice versa in the middle of a video to see whether this hampers the performance of the model. Indeed, when comparing it to Tab. 1 of the main paper, there is a drop in performance for all methods. However, we observe that the mimic network still outperforms other methods in this changing occlusion scenario as shown in Tab. 8. This further demonstrates the generalizability of our method.

16.3. Difficulty of different occlusion types

In this section, we compare the relative difficulty of different occlusion types for the mimic network. For this, we take a model trained on top, bottom and middle occlusions. However, during evaluation, we restrict the occlusions to one type at a time, for top, bottom and middle occlusions. The results are presented in Tab. 9. As one might expect, bottom occlusions are the most difficult for the network to work on, since the legs - the body part where the most obvious gait patterns appear - are not visible in these occlusions. Interestingly, top and middle occlusions are roughly equally difficult for the model.

16.4. Training on all occlusion types

So far, our experiments have focused on top and bottom occlusions, with some analysis being done on middle

All occlusion types	Rank-1	Rank-5
Baseline-2	30.3	46.2
Occlusion Aware	35.5	52.4
Mimic Network	43.0	60.2

Table 10. Performance of different methods on the mixed occlusion set, comprising of top, middle, bottom, dynamic small and dynamic tall patches. These results are on the GREW dataset using the GaitBase backbone.

and dynamic occlusions. Since it is not practical to train on all possible occlusion types that may occur, a model needs to be able to generalize to newer occlusion types. Hence, we do a generalizability evaluation in the 'Zero-shot evaluation' columns of Tab. 2 in the main paper, taking a model trained on top and bottom occlusions and applying it on middle or dynamic occlusions.

However, if a particular type of occlusion is anticipated in the task setup - maybe due to camera placement constraints in the final use case - it is possible to prepare the model for the specific occlusion type. Generally, the trend is that training a model on a specific occlusion type improves performance on that occlusion set. This is demonstrated in the Training columns of Tab. 2 of the main paper. In these training columns, we add an additional occlusion type (middle or dynamic) in the training set, in addition to top and bottom occlusions.

While training on newer occlusion types helps, we reiterate that it is not practical to train (and evaluate) a gait recognition model for every possible occlusion type which might occur. Hence, in our main paper, we focus on only a limited set of occlusion types (top and bottom occlusions) for training the model in our main results and then try to see whether this model generalizes well to other occlusion types - rather than training the model on all the occlusion types at once.

In this supplementary material section, for completeness, we also present the results from training new networks on the combined set of all occlusion types we have used in our experiments - Top occlusions, Middle occlusions, Bottom occlusions, dynamic small patch occlusions and dynamic tall patch occlusions. The results are summarized in Tab. 10. We observe that the mimic network is still able to outperform other approaches on this combined occlusion set.

16.5. Different occlusion ranges

In our work, we pick an occlusion range R, which we set to (0.4, 0.6) for most of our experiments in consistent occlusions. This means the amount of occlusion is randomly sampled to be 40-60% of the frame dimension. In this section, we explore how the performance changes on different occlusion ranges. We present the results in Tab. 11. As expected, the performance improves as we reduce the amount of occlusion. This is reflected in both the rank retrieval accuracy and the RP values.

Range	Rank-1	Rank-5
40-60%	28.38 (0.51)	45.43 (0.63)
30-50%	33.62 (0.61)	51.33 (0.71)
20-40%	40.25 (0.73)	57.68 (0.80)
10-30%	45.35 (0.82)	61.85 (0.86)

Table 11. Performance of the mimic network on multiple ranges of top and bottom occlusion, on the GREW dataset using GaitBase backbone. The RP values are shown in (.). The first row denotes the experiments with the standard occlusion range we use throughout the paper. Both the rank retrieval accuracy and the RP increase as we reduce the occlusion range, as expected.

Our results on different occlusion types or different occlusion settings are some basic analyses which we performed to investigate how different occlusions impact the gait recognition problem. This is in no way a comprehensive analysis of which occlusion setting is more practical, or more likely to occur in real scenarios. We maintain that the focus of this work is to propose an approach which works better than other approaches for some given occlusion settings. The analysis of different occlusion types is out of the scope of this work and we leave it up to future research.

16.6. Speed of adaptability

In the adaptability scenario, we extend the capability of training a model on a new type of occlusion in addition to the old occlusions it was trained on. We discuss how effectively the networks are able to adapt to different occlusion types in Tab. 2. Here, we discuss the speed of this adaptation - how much re-training is required to adapt the model to a new occlusion type. We discuss the GaitBase backbone on the GREW dataset.

GaitBase is trained on 180,000 iterations on the GREW dataset, according to the settings used by [5]. We also train the model on top and bottom occlusions for the same number of iterations. However, when adapting the model to a new occlusion type, say middle occlusions, we train the model further until the training loss converges with the new occlusion set. We observe that the training loss converges at around 20,000 iterations. Thus, the network is able to adapt to a new occlusion set with additional training of 20,000 iterations, or roughly 11% of additional training time.

16.7. Results on indoor datasets

Our main focus is on the outdoor, in-the-wild scenarios where occlusion is more likely to occur. Hence, we focus our experiments on in-the-wild datasets like GREW,

Method	CASIA-B			OUMVLP
	NM	BG	CL	NM
Baseline-1	22.76 (0.23)	20.45 (0.22)	10.03 (0.13)	2.09 (0.02)
Baseline-2	31.35 (0.32)	24.56 (0.26)	13.98 (0.18)	14.47 (0.16)
Occlusion Aware	52.80 (0.54)	47.10 (0.50)	33.46 (0.43)	17.65 (0.19)
Mimic Network (ours)	**69.42 (0.71)**	**57.12 (0.61)**	**37.35 (0.48)**	**25.42 (0.28)**

Table 12. Rank-1 accuracy for different conditions on CASIA-B and OUMVLP datasets, on top and bottom occlusions. The corresponding RP values are reported in (.). The mimic network is able to outperform other approaches on the indoor datasets as well.

Gait3D and BRIAR. However, for completeness, we evaluate our approach on indoor datasets using the simulated top and bottom occlusions as well. We report the Rank-1 accuracy on Normal Walking (NM), Baggage (BG) and Cloth Changing (CL) for the CASIA-B dataset according to the standard protocol [5]. We also report the Rank-1 accuracy on NM for OUMVLP according to the standard protocol used in [5]. Additionally, we report the RP values for all these metrics in (.). The results are presented in Tab. 12. They show a similar trend, where the mimic network is able to outperform other methods on occluded gait recognition.

17. Failure case Analysis

Though our model is able to improve performance on occluded gait recognition, it is not perfect. We have discussed the relative difficulty of different occlusion types in Sec. 16.3, concluding that the model is more likely to fail in bottom occlusions. From Tab. 11, we also empirically confirm that the model is more likely to fail when the occlusion is more severe.

In this section, we discuss some of the specific failure cases of the model. We have shown some examples of probes taken from the GREW dataset which are misidentified by our model in Fig. 7. In some of these probes, it is not possible to make out which body part is present in the input. This makes it difficult to extract its correlations with the missing body parts. In some probes, only the head is visible and there is barely any motion present in the input. It becomes difficult to recognize the subject from a generic circular silhouette without any characteristic motion.

From this observation, we can conclude that the model can not perform on any input - some basic body parts have to be recognizable for the model to extract correlations with the other body parts. Further, temporal information is necessary for the model to recognize the gait of the subject. Without any motion in the input, the model will likely not be able to identify the subject correctly. These are recognized as potential situations where there is scope for improvement and we leave this to future work.

This error analysis has been performed on the GREW dataset with the mimic network using the GaitBase backbone.

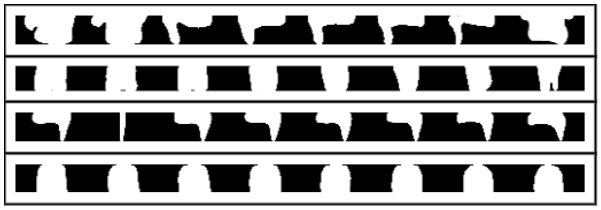

Figure 7. Visualization of some of the failure cases of the model - the misclassified probes on the GREW dataset. We see that in a lot of these examples, it is not even clear which body part is being shown; it is extremely difficult to identify these subjects clearly because of the lack of discriminative information in the input.

18. Cross-Entropy loss

In our experiments in Tab. 3, we observe that adding cross-entropy loss to our Multi-instance Correlational KD (MiCKD) loss for training the mimic network ends up reducing performance. Here, we try to analyse this issue.

We follow the cross-entropy formulation used in [5], where we use a BNNeck layer on the embeddings before applying cross entropy loss. This is shown to stabilize training and yield better performance. However, this cross-entropy (XE) formulation hurts the performance of the model when used along with the MiCKD loss. To investigate this further, we plot the t-SNE features of some samples from the GREW dataset in Fig. 8.

We observe an interesting pattern in the MiCKD + XE embeddings, where a lot of the embeddings are clustered in the center while some are scattered further away. The proximity of most of the embeddings in the center makes it difficult to match probe embeddings to the proper gallery. On the other hand, the embeddings in the MiCKD figure are clustered better than MiCKD + XE. Based on this observation, we conclude that when MiCKD and XE are used together, the embeddings are not able to cluster well - in other words, MiCKD and XE loss do not go well together.

Based on the results of [5], we know that Triplet Loss and cross-entropy loss go well together. Thus, one might try to replace MiCKD with the Triplet Loss used in [5], and use it along with cross entropy loss. However, this formulation when applied to occluded data is the same as Baseline-2 in Tab. 3; it performs worse than even MiCKD + XE. It should be noted that VEN has been removed in these set of experiments, and we are dealing with a vanilla variant of the mimic network, one without the VEN.

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

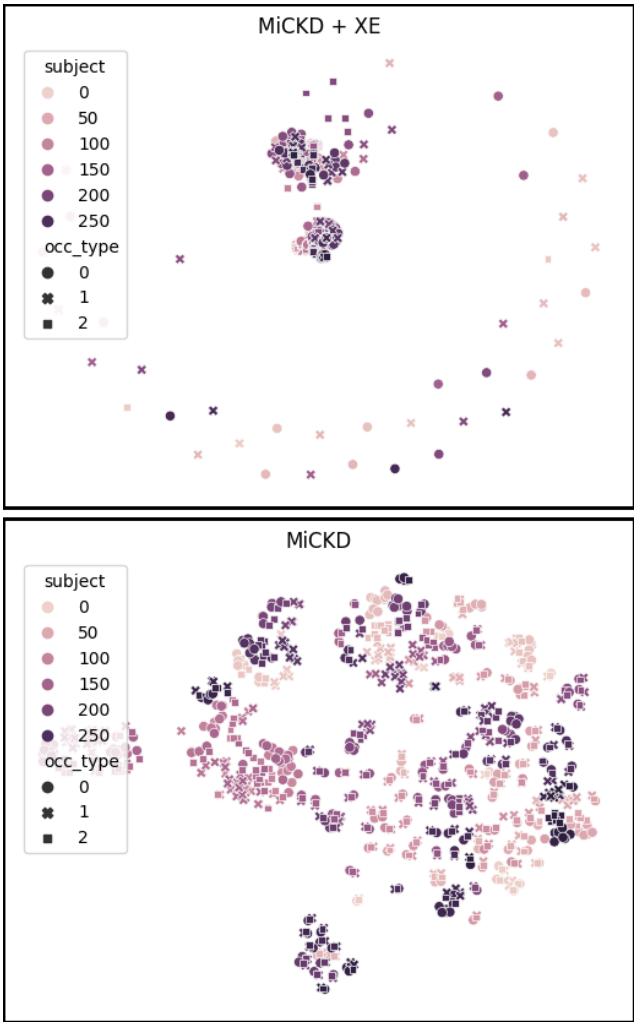

Figure 8. Visualization of t-SNE features with and without using the cross entropy (XE) loss along with the proposed MiCKD loss. Different colors denote different subjects and different shapes denote different occlusion types. Features are relatively better clustered when cross entropy loss is not used.

[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, 4

[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, 3, 4

[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, 3

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

[37] 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

[38] 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

[39] 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

[40] 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

[41] 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

[42] 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

[43] 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

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

[45] 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, 4

[46] 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