Watch Where You Head: A View-biased Domain Gap in Gait Recognition and Unsupervised Adaptation

Gavriel Habib* Noa Barzilay* Or Shimshi Rami Ben-Ari Nir Darshan
OriginAI, Israel
{gavrielh, noab, ors, ramib, nir}@originai.co

Abstract

Gait Recognition is a computer vision task aiming to identify people by their walking patterns. Although existing methods often show high performance on specific datasets, they lack the ability to generalize to unseen scenarios. Unsupervised Domain Adaptation (UDA) tries to adapt a model, pre-trained in a supervised manner on a source domain, to an unlabelled target domain. There are only a few works on UDA for gait recognition proposing solutions to limited scenarios. In this paper, we reveal a fundamental phenomenon in adaptation of gait recognition models, caused by the bias in the target domain to viewing angle or walking direction. We then suggest a remedy to reduce this bias with a novel triplet selection strategy combined with curriculum learning. To this end, we present Gait Orientation-based method for Unsupervised Domain Adaptation (GOUDA). We provide extensive experiments on four widely-used gait datasets, CASIA-B, OU-MVLP, GREW, and Gait3D, and on three backbones, GaitSet, GaitPart, and GaitGL, justifying the view bias and showing the superiority of our proposed method over prior UDA works.

1. Introduction

Gait recognition is a biometric method that analyzes and identifies individuals based on their walking patterns. It involves the analysis of various characteristics such as stride, body shape, motion postures, and limb movements. Gait recognition has been widely used in different application areas such as sport [45], health monitoring [34], forensic analysis [31] as well as security and surveillance [33]. However, compared to different biometrics, like face and fingerprint recognition, gait is a challenging task due to complex conditions such as clothes changing, object carrying, and diverse camera views. Existing methods achieve high performance on specific datasets [4,9,29]. GaitSet [4] learns from a set of gait silhouettes, GaitPart [9] proposes body part-based model, and GaitGL [29] utilizes global and local visual information. However, the ability of these methods to generalize to new datasets is limited due to domain gaps. Domain gaps can be caused by e.g. different camera settings or clothing conditions (changed between seasons), making the gait sequences appear differently across use cases.

Unsupervised Domain Adaptation (UDA) aims to transfer a model learned from a source domain with labeled data to an unseen target domain without labels. UDA is common in different computer vision tasks, such as object classification [42,50], semantic segmentation [5,16], and person re-identification [3,21,40].

UDA for Gait Recognition addresses the case in which the source and target domains include different identities. One main concern for gait recognition task is the ability to recognize an individual over different camera views and walking directions. The interplay between camera positioning and an individual’s walking direction within the 3D realm defines what is referred to as the viewing angle in literature [8,35,44]. Gait recognition models are trained to generalize over different viewing angles, particularly tested by robustness on such scenarios [35,44].

In this paper, we consider a case where a trained model on a labeled source domain should be adapted to a target domain without any access to the target labels or the source data (that could be unavailable due to privacy issues). To this end, we first use the source model to generate an initial representation of the target domain followed by an adaptation scheme to adapt the pre-trained backbone to the new domain. Such adaptation is often carried out by applying a contrastive loss such as triplet loss [14]. Without access to the labels in the target domain, prior works used different cues to distinguish (approximately) between the samples [39,47]. In this paper, we propose GOUDA - Gait Orientation-based method for Unsupervised Domain Adaptation. Our method is driven mainly by a strong observation of the gait view distribution in the target domain (we consider yaw angle here). Fig. 1 shows an example of view distribution on OU-MVLP dataset [35]. The depicted UMAPs

* The authors contributed equally

6109
show the distribution of gait representations color coded by corresponding viewing angles. While in the source domain (Fig. 1a) the distribution is nearly uniform indicating that relations in the embedding space are invariant to the view, the clustered colors in the target domain (Fig. 1b) imply a strong bias toward the viewing angle. In the target domain, nearby points are likely to be from similar view and not necessarily the same identity (see Rank-1 accuracies in the figure). This view bias causes a significant drop in recognition. Fig. 1c shows that GOUDA is capable of disentangling the view bias, modifying the embedding space in the target domain to improve the recognition accuracy. We show that this fundamental observation and correction method is a key contributor to gait UDA methods as evaluated over various gait backbones and cross domain datasets. Essentially, we suggest a novel triplet selection strategy, based on this key observation. Our method also uses a self-supervised method and curriculum learning, to further boost the results. To summarize, these are our key contributions:

- We reveal a fundamental bias in cross-domain gait recognition models toward viewing angles.
- We propose GOUDA, a UDA method with a novel view-based Triplet Selection strategy, mitigating the observed viewing angle bias and leading to a significant improvement in gait UDA task.
- We conduct extensive experiments with three backbones and with 14 source and target combinations, outperforming prior state-of-the-art by a large gap.

2. Related Work

Existing gait recognition methods can be grouped into two primary categories: model-based approaches and appearance-based approaches.

Model-based approaches [24, 27, 36, 43] focus on capturing the underlying structure and kinematic features of the human body to learn the walking patterns for recognition. These methods utilize structural models, such as 2D or 3D skeleton keypoints [24, 27, 36, 49] as their input representation, aiming to achieve robustness to clothing and object carrying conditions. However, they often face challenges in estimating body parameters in low-resolution videos. Appearance-based methods extract gait representations directly from RGB images or binary silhouette sequences without explicitly modeling the human body. In this context, Silhouette-based recognition is widely used in the literature e.g., template-based methods [12, 23, 38, 38] that treat entire gait sequences as templates, sequence-based methods that consider gait sequences as videos [7, 17, 18, 26, 28, 29] or as an unordered image set [4, 15]. In our study, we use appearance-based gait recognition models, specifically employing silhouette sequences or unordered sets as input data.

2.1. Unsupervised Domain Adaptation (UDA)

In general UDA task [30, 41], domain-invariant feature learning aims to align the source and target distributions by minimizing divergence measures [20] or by utilizing adversarial training [11, 37]. Other UDA approaches employ ensemble methods [10] or adapt batch normalization statistics from source to target domain [25]. A different approach applies clustering using pseudo-labels [3, 21, 40], in the target domain. Our approach draws inspiration from aligning the source and target distributions.

UDA has been employed in the re-identification task coping with variations in view, camera settings, and environmental conditions, over different domains. In this respect, UNRN [48] is a clustering-based method with uncertainty-guided optimization driven by a teacher-student framework. SECRET [13] utilizes global and local features for a self-consistent pseudo-label refinement. A triplet stra-
We use the setting with GaitSet and OU-MVLP (source) to CASIA-B (target). TraND [47] usually selects positive samples with the exact same view as the anchors due to its strategy to bring adjacent samples closer. Contrary to that, GOUDA selects positives with different views by default. Here, the anchor view is 54°, and the positive sample selected by GOUDA is of view 108°. In both cases, the positives are correct identity-wise. By bringing closer samples of different views, our method improves the cross-view benchmarks and outperforms TraND, as shown in Table 2.

Despite the necessity of UDA in Gait Recognition, only a few studies have addressed this specific task [39, 47]. Both of these methods attempt for a fine-tuning adaptation on the target domain. From Indoor to Outdoor (FI2O) [39] conducts an adaptation step by using a pseudo-labeling strategy. Second, the Triplet Selection algorithm is applied, aiming to push away sequences of similar views and bring closer different ones in the embedding space. The colored shapes depict the selected triplets. The triplets are then used to update the backbone using triplet loss.

3. Method

Figure 3 presents an overview of our method with three main stages. First, a gait model, pre-trained on a source domain, is used as a feature extractor for an unseen target domain. In the embedding space, each sequence is described by its evaluated viewing angle, while its label (identity) is unknown. Second, our Triplet Selection approach is applied, aiming to push away sequences of similar views while bringing closer different views. In order to increase the probability that the triplets are identity-wise correct, we further adopt a Curriculum Learning framework [2] based on a certainty measure. Lastly, the selected triplets are used to fine-tune the network using standard triplet loss.

3.1. View Extraction

In gait recognition, the view is defined as the walking direction of the person relative to the camera. To estimate the view of a gait sequence in the target domain, we use the provided 2D keypoints (representing the posture of human body parts). First, we apply MHFormer [22] to regress the 3D keypoints per frame. Then, we estimate the sequence view (yaw angle) using the median 3D keypoints of relevant body parts - hips and shoulders. For real application uses, the 2D keypoints can be estimated from RGB images.

3.2. Triplet Selection

This section provides a description of the Triplet Selection algorithm, which aims to improve performance on the target domain by disentangling the view bias. Ideally, we aim to find correct triplets such that, given an anchor, the positive sample has the same identity but a different view, and the negative sample belongs to a different identity but of similar view.

Negative Sample Selection. Based on the observation that gait recognition models strongly emphasize view-based
features in the target domain (Fig. 1), we assume that within a neighborhood of samples with a similar view, there are likely to exist samples of different identities.

**Positive Sample Selection.** Gait models demonstrate a certain level of learned general gait features. Hence, we assume that among all samples with different views, the ones with the same identities are reasonable to be the closest\(^1\).

In Fig. 1b, it is implied by the cross-view Rank-1 accuracy (25.7%), which is much higher than random (±0.02%).

In the embedding space, we prioritize pushing away similar-view samples and bringing closer together cross-view samples. We use various techniques, such as curriculum learning, to increase the identity-wise certainty.

The input to the algorithm (pseudo-code presented in Alg. 1) is an unlabeled target domain set \( S = \{ (x_r, v_r) \}_{r=1}^R \), where \( R \) is the number of gait sequences, \( x_r \in \mathbb{W} \times H \times T \) is the \( r \)-th sequence with \( T \) frames of size \( W \times H \), and \( v_r \in \mathbb{R} \) is its estimated view. Each gait sequence \( x_r \), is represented by its embedding \( e_r = f(x_r) \), where \( f(\cdot) \) is a gait model pre-trained in a supervised manner on the source domain. In addition, we define the following hyper-parameters: margin \( m \), and view thresholds \( T_s, T_c \) such that \( T_s \leq T_c \). We use \( v_j \) as an indicator (in the range of \([0, 360)\)), for separation of the data for each anchor \( a \) into two mutually exclusive sets. Considering an anchor \( a \), the sample \( j \) belongs to the set \( \chi_{a, \text{similar}}^a \), if \( |v_a - v_j| < T_s \), and belongs to \( \chi_{a, \text{cross}}^a \), if \( |v_a - v_j| > T_c \). The triplets \( \{a, p, n\} \) are selected such that \( n \in \chi_{a, \text{similar}}^a \), and \( p \in \chi_{a, \text{cross}}^a \).

Given a set of embeddings and views \( \{ (e_r, v_r) \}_{r=1}^R \), we now calculate the distances matrix \( D \in \mathbb{R}^{R \times R} \), such that \( D_{i,j} := d(e_i, e_j) \), where \( d \) is the cosine distance metric. In order to increase the certainty of triplets to be identity-wise correct, we select them as follows. The positive \( p \) is defined as the closest sample to \( a \) in \( \chi_{a, \text{cross}}^a \) (line 4 in Alg. 1). We choose \( n \in \chi_{a, \text{similar}}^a \) to be the farthest sample from \( a \) within a fixed margin \( m \) from \( p \) (line 5 in Alg. 1). Eventually, the triplet \( \{a, p, n\} \) is “valid” if exists at least one sample in \( \chi_{a, \text{similar}}^a \) that is closer to \( a \) than \( n \) (line 6 in Alg. 1). Both restrictions on \( n \) are used to increase the probability that \( a \) and \( n \) do not share the same identity, assuming that the closest similar-view sample is more likely to share the same identity. The margin \( m \) is incorporated into the negative sample selection (line 5 in Alg. 1) in order to align with the objective of triplet loss. An illustration of the Triplet Selection algorithm is shown in Fig. 4, and described in Algorithm 1. The algorithm is applied iteratively during the training process (Section 3.3).

3.3. Curriculum Learning

Inspired by [2, 47], we apply the Triplet Selection algorithm in four different stages during training, in a from-

---

\(^1\)In the supplementary materials, we present further intuition and justification, including a shared visualization of views and identities.

---

**Algorithm 1 Triplet Selection**

0: function Triplet Selection\((D, \{v_r\}_{r=1}^R)\)
1: for \( a \leftarrow 1 \) to \( R \) do
2: \( \chi_{a, \text{similar}}^a \leftarrow \{ j \mid |v_a - v_j| < T_s, a \neq j \} \)
3: \( \chi_{a, \text{cross}}^a \leftarrow \{ j \mid |v_a - v_j| > T_c \} \)
4: \( p \leftarrow \chi_{a, \text{cross}}^a \arg\min_j \{D_{a,j}\} \)
5: \( n \leftarrow \chi_{a, \text{similar}}^a \arg\max_j \{D_{a,j} \mid D_{a,j} < D_{a,p} + m\} \)
6: if \( \exists j \in \chi_{a, \text{similar}}^a \) s.t. \( D_{a,j} < D_{a,n} \) then
7: \( \mathcal{V} : \text{Valid Triplets} \leftarrow \{a, p, n\} \)
8: \( \mathcal{C} : \text{Confidence Scores} \leftarrow 1 - D_{a,p} \)
9: Selected Triplets \( \leftarrow \mathcal{V}[\arg\max \mathcal{C}] \)
10: return Selected Triplets

---

Figure 4. Triplet selection illustration (shape describes view). Given an anchor \( a \), the positive \( p \) is the closest sample with a different view. The negative \( n \) is the farthest sample (up to margin \( m \) from \( p \)) with a similar view as \( a \). A triplet \( \{a, p, n\} \) is “valid” if exists at least one similar-view sample closer to \( a \) than \( n \).

easy-to-hard manner. We define a triplet confidence score as the cosine similarity between the anchor and the positive samples \((1 - D_{a,p})\). We select the top \( q\% \) valid triplets \( \mathcal{V} \) according to their confidence scores \( \mathcal{C} \), while \( q \) gets higher in each stage. We keep improving target domain performance by starting from updating the model with a small amount but high-quality data and carefully extending the dataset size.

3.4. Stopping Criteria

In contrast to previous works [39, 47], we introduce an indicative stopping criteria (SC) using a validation set. Although training with a fixed number of iterations is simple, it lacks control over the distribution obtained in the embedding space, due to an over-fitting risk. Our SC is driven by the desire to achieve greater diversity in view within the neighborhoods. Given a validation set of size \( N_{val} \), for each sample \( a \) we compute its \( K \) nearest neighbors, \( KNN(a) \), using the cosine distance metric. Then, we count how many of the \( K \) samples are of similar view as \( a \). The final SC value is the average of all counts (Eq. 1). Eventually, we
Gait3D [46] is a 3D representation-based gait dataset captured in the supermarket. The dataset contains 4,000 subjects and over 25,000 sequences extracted from 39 cameras. The training set includes 3,000 subjects, and the test set includes the remaining 1,000 subjects.

GREW [51] is a large-scale in-the-wild gait recognition dataset. The dataset contains 26,345 subjects and 128,671 sequences extracted from 882 cameras. We adopt 2 protocols to evaluate the dataset [51], [39]. According to the first protocol [51], the training set contains 20,000 subjects and the test set includes 6,000 subjects. The second protocol [39] splits the original training set, the first 2,000 subjects are utilized for training and the last 2,000 subjects are utilized for test. We refer to the latter partition as GREW*.

4.2. Implementation details

We use three different well-known gait backbones - GaitSet [4], GaitPart [9], and GaitGL [29] to evaluate the proposed GOUDA. The input silhouettes in all datasets are resized to 64×44. Each backbone was first trained on the source domain data, using the implementations provided in OpenGait code [8]. The training set in each target domain was partitioned into separate training and validation sets. Specifically, we created the validation set by selecting the last 10% of subjects from the training set.

In the model updating stage, all backbones were trained using the loss function as described in Eq. 4, with $m = 0.2$. View thresholds $T_\mathcal{s}$ and $T_\mathcal{c}$ are set to 10° and 20° respectively. We adopt Adam optimizer [19] with a learning rate of 1e-5 and weight decay of 5e-4. In every training iteration, a batch size of 32×44 was employed, where the number 4 corresponds to the inclusion of two different augmentations applied to the anchor sample $(a)\,\tilde{a} = \tilde{a}$, a positive sample $(p)$, and a negative sample $(n)$. The final weights of the model were selected according to the optimal results of the stopping criteria (cf. Sec. 3.4) on the validation set. $K$ value of the SC was set to 5. The curriculum learning strategy includes four stages, where each stage involves selecting a higher percentage of top $q\%$ valid triplets: 10, 25, 50, and 100%. The amount of training iterations after each stage is proportional to the number of selected triplets. Specifically, we move to the next stage after each selected triplet appears in training 10 times on average.

5. Evaluation

5.1. Results

Comparison to Direct Testing. Our method does not assume special architectural conditions and therefore can be integrated with different kinds of deep backbones. We evaluate the performance of GOUDA when combined with various backbones, including GaitSet [4], GaitPart [9], and GaitGL [29], as shown in Table 1. These results are then compared to Direct Testing performances, i.e. models that are trained solely on the source domain in a supervised manner.

\begin{equation}
SC = \frac{1}{N_{val}} \sum_{a=1}^{N_{val}} \sum_{i \in KNN(a)} \mathbb{I}[i \in \chi_{similar}] \rightarrow \min \tag{1}
\end{equation}

3Ablation study on $q$ values is shown in the Supplementary Materials.

\begin{equation}
L_{GOUDA} = \sum_{i=1}^{N} [D_{a_i,p_i} - D_{a_i,n_i} + m]_+ \tag{2}
\end{equation}

where $N$ represents the number of triplets, $m$ is the triplet loss margin and $[x]_+$ is max($x, 0$). Note that $a$ and $n$ share a similar-view, different from $p$. Our self-supervision loss (SSL) is defined by:

\begin{equation}
L_{SSL} = \sum_{i=1}^{N} [D_{a_i,\tilde{a}_i} - D_{a_i,n_i} + m]_+ \tag{3}
\end{equation}

In the self-supervision loss we replace $p$ with $\tilde{a}$, an augmented version of $a$. The augmentations are two random sub-sequences of the same gait sequence. Eventually, the losses are linearly combined to our domain adaptation loss:

\begin{equation}
L = L_{GOUDA} + L_{SSL} \tag{4}
\end{equation}

The selected triplets are randomly extracted for training.

4. Experiments

4.1.Datasets

In this paper we evaluate the proposed GOUDA on four well-known gait datasets, CASIA-B [44], OU-MVLP [35], Gait3D [46], and GREW [51].

CASIA-B [44] is a popular indoor gait dataset. It contains 13,640 gait sequences, including 124 subjects with 11 views. Each subjects’ videos include 3 different settings, normal walking (NM), walking with a bag (BG), and walking with a coat (CL). The first 74 subjects are used for the training set and the rest 50 subjects are used for the test set.

OU-MVLP [1, 35] is a large indoor gait dataset. It contains 10,307 subjects with 14 views. According to the provided protocol [4], 5,153 subjects are used for the training set, and the remaining 5,154 subjects are used for testing.

Gait3D [46] is a 3D representation-based gait dataset captured in the supermarket. The dataset contains 4,000 subjects and over 25,000 sequences extracted from 39 cameras. The training set includes 3,000 subjects, and the test set includes the remaining 1,000 subjects.

In the adaptation stage, we optimize the backbone $f$ with two triplet loss objectives: GOUDA and self-supervision loss. Suppose $\{a_i, p_i, n_i\}$ is a selected triplet, GOUDA loss is given by:

\begin{equation}
L_{GOUDA} = \sum_{i=1}^{N} [D_{a_i,p_i} - D_{a_i,n_i} + m]_+ \tag{2}
\end{equation}

where $N$ represents the number of triplets, $m$ is the triplet loss margin and $[x]_+$ is max($x, 0$). Note that $a$ and $n$ share a similar-view, different from $p$. Our self-supervision loss (SSL) is defined by:

\begin{equation}
L_{SSL} = \sum_{i=1}^{N} [D_{a_i,\tilde{a}_i} - D_{a_i,n_i} + m]_+ \tag{3}
\end{equation}

In the self-supervision loss we replace $p$ with $\tilde{a}$, an augmented version of $a$. The augmentations are two random sub-sequences of the same gait sequence. Eventually, the losses are linearly combined to our domain adaptation loss:

\begin{equation}
L = L_{GOUDA} + L_{SSL} \tag{4}
\end{equation}

The selected triplets are randomly extracted for training.
ner. It is evident that the direct testing results exhibit a certain level of competence in learning general gait features, which contributes to achieving reasonably decent results in the target domains. Nevertheless, these results are still unsatisfactory. Referring to Table 1, it becomes apparent that our method yields substantial improvements in Rank-1 accuracy across almost all settings. Notably, GOUDA demonstrates improvements across multiple backbones, underscoring the algorithm’s robustness independent of any specific architecture. This finding reinforces the observation presented in Section 1. GOUDA demonstrates superior performance compared to Direct Testing results in 49 out of 54 cases. In some cases, our method improves the Rank-1 accuracy by a significant margin, for example from 23.9% to 43.6%, when using GaitPart trained on GREW (source) to OU-MVLP (target). In the supplementary materials, we present even higher improvements when using ground truth views provided as meta-data in CASIA-B and OU-MVLP datasets, avoiding the noise produced by the View Extraction module. It may suggest that the potential of GOUDA is higher, if using an improved 3D keypoints estimator.

It is apparent that the Rank-1 improvements achieved by GOUDA are considerably higher when employing CASIA-B or OU-MVLP datasets as the target domain, compared to GREW or Gait3D. This disparity can be attributed to the nature of the datasets themselves. CASIA-B and OU-MVLP are indoor datasets captured in controlled environments, whereas GREW and Gait3D were collected in-the-wild. Consequently, the latter datasets present more challenging conditions for gait recognition, potentially leading to lower gains in performance compared to the former datasets. Furthermore, our method relies on estimating a single view through the entire sequence. However, such an approach does not accurately reflect the conditions encountered in datasets captured in the wild. In practice, one can decompose the gait sequence into snippets with piece-wise constant views and analyze them separately.

**Comparison with State-of-The-Art.** We compare GOUDA to state-of-the-art UDA gait recognition methods [39, 47] and present them in Table 2. To ensure a fair comparison, we employ the same backbone in each comparison, GaitGL when comparing with [39] and GaitSet when comparing with [47]. In order to compare to TraND [47], we trained their model using the implementations provided by the authors. As a reference, we further report the results of UDA methods for person re-identification, UNRN [48] and SECRET [13], applied by [39]. In these cases, the entire gait sequence was compressed to a single template (GEI [12]) as these models get a single image as an input.

As shown in Table 2, GOUDA significantly outperforms prior works in 9 out of 10 settings. The difference in results between GOUDA and TraND [47] can be attributed to the fundamental principle on which the TraND approach relies. This approach is centered around an entropy-first neighbor selection strategy focusing solely on the nearest neighbor sample that is likely to share the same view. We believe that due to this reason, TraND is less robust to different views.

5.2. Ablation Study

To demonstrate the efficacy of the proposed method and assess the influence of various key components, we carried out an ablation study using GaitGL with OU-MVLP (source) to CASIA-B (target) setting. The experimental results are reported in Table 3.
Analysis of Various Key Components. We trained various models that emphasize the impact of key components in GOUDA (rows 2-8 in Table 3). Row 2 highlights the impact of incorporating $m$ in the Triplet Selection process. In this experiment, we used $m = 0$ when selecting the negative samples (line 5 in Alg. 1). Interestingly, the results are even lower than those achieved with Direct Testing. This outcome can be attributed to the fact that excluding the margin in the Triplet Selection algorithm leads to a significantly reduced number of valid triplets available for training. Thus, it may potentially result in an overfitted model, as it lacks the necessary diversity and complexity in the training data. In row 3, we demonstrate the effectiveness of incorporating a curriculum learning strategy. In this training process, the model undergoes four stages of triplet selection as well. However, unlike GOUDA, all valid triplets are selected during training ($q = 100\%$ in all four stages).

Loss Components. The models in rows 4-5 were trained without $L_{SSL}$ and $L_{GOU DA}$, respectively, i.e., all triplets related to each one of these loss terms were excluded from $L$. We observe the significant impact of these loss terms on the results, especially of $L_{GOU DA}$, which is the key component of the loss.

Hyperparameters. In row 6, we explore a different Stopping Criteria (SC) for GOUDA. Instead of using 5 neighbors ($K = 5$), we examine the SC with a single nearest neighbor ($K = 1$). In this case, the results are comparable to GOUDA. However, selecting an environment of 5 neighbors might be less prone to noise compared to solely considering the nearest neighbor. Rows 7-8 present experiments with different values of $T_c$ and $T_s$ (defined in lines 2, 3 of Alg. 1), where in each study only one threshold was changed. These results are inferior or comparable to those achieved with the original values.

TABLE 2. Comparison with the state-of-the-art UDA gait recognition methods, FI2O [39] and TraND [47], and with UDA methods for person re-identification, UNRN [48] and SECRET [13]. Results are reported by Rank-1 (R1), Rank-5 (R5) and Rank-10 (R10). GOUDA is compared to each method in two different settings. GREW* uses the protocol presented in [39]. Direct testing and domain adaptation results of the UDA gait methods are shown with relevant backbones (as used in the corresponding papers).

Analysis

## Table 2. Comparison with the state-of-the-art UDA gait recognition methods, FI2O [39] and TraND [47], and with UDA methods for person re-identification, UNRN [48] and SECRET [13]. Results are reported by Rank-1 (R1), Rank-5 (R5) and Rank-10 (R10). GOUDA is compared to each method in two different settings. GREW* uses the protocol presented in [39]. Direct testing and domain adaptation results of the UDA gait methods are shown with relevant backbones (as used in the corresponding papers).

<table>
<thead>
<tr>
<th>Method</th>
<th>CASIA-B → GREW*</th>
<th>OU-MVLP → GREW*</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>R1</td>
<td>R5</td>
</tr>
<tr>
<td>UNRN</td>
<td>20.2</td>
<td>32.1</td>
</tr>
<tr>
<td>SECRET</td>
<td>12.3</td>
<td>19.6</td>
</tr>
<tr>
<td>GaitGL</td>
<td>30.0</td>
<td>40.5</td>
</tr>
<tr>
<td>F12O</td>
<td>37.4</td>
<td>49.2</td>
</tr>
<tr>
<td>GOU DA</td>
<td>35.6</td>
<td>49.9</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Method</th>
<th>CASIA-B → OU-MVLP</th>
<th>OU-MVLP → CASIA-B</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>NM</td>
<td>BG</td>
</tr>
<tr>
<td>GaitSet</td>
<td>9.6</td>
<td>74.0</td>
</tr>
<tr>
<td>TraND</td>
<td>15.8</td>
<td>74.7</td>
</tr>
<tr>
<td>GOU DA</td>
<td>27.9</td>
<td>87.0</td>
</tr>
</tbody>
</table>

**Table 3. Ablation experiments with domain adaptation from OU-MVLP (source) to CASIA-B (target). Average Rank-1 accuracies across all 11 views are shown, excluding identical-view cases.**

<table>
<thead>
<tr>
<th>#</th>
<th>Method</th>
<th>CASIA-B → GREW*</th>
<th>OU-MVLP → GREW*</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>GaitGL Direct Testing</td>
<td>81.7</td>
<td>71.5</td>
</tr>
<tr>
<td>2</td>
<td>GOU DA w/o m</td>
<td>68.2</td>
<td>43.6</td>
</tr>
<tr>
<td>3</td>
<td>GOU DA w/o curriculum learning</td>
<td>88.2</td>
<td>72.0</td>
</tr>
<tr>
<td>4</td>
<td>GOU DA w/o $L_{SSL}$</td>
<td>90.8</td>
<td>76.7</td>
</tr>
<tr>
<td>5</td>
<td>GOU DA w/o $L_{GOU DA}$</td>
<td>87.5</td>
<td>72.9</td>
</tr>
<tr>
<td>6</td>
<td>GOU DA w/ SC $K = 1$</td>
<td>92.1</td>
<td>79.6</td>
</tr>
<tr>
<td>7</td>
<td>GOU DA w/ $T_c = 10^5$</td>
<td>87.8</td>
<td>70.2</td>
</tr>
<tr>
<td>8</td>
<td>GOU DA w/ $T_s = 20^5$</td>
<td>92.1</td>
<td>80.2</td>
</tr>
<tr>
<td>9</td>
<td>Triplet Selection Oracle</td>
<td>96.3</td>
<td>90.3</td>
</tr>
<tr>
<td>10</td>
<td>Supervised fine-tuning</td>
<td>98.3</td>
<td>94.7</td>
</tr>
<tr>
<td>11</td>
<td>GOU DA</td>
<td>92.8</td>
<td>80.9</td>
</tr>
</tbody>
</table>
Figure 5. Curriculum Learning Analysis. Triplet Selection is applied 4 times during training (x-axis). Y-axis shows how many of the selected triplets are correct considering person identities. The colors represent the top confident triplets, and in each stage the chosen percentage is marked by a rectangle (see Section 3.3).

Figure 6. Curriculum Learning Analysis (as described in Fig. 5), using the ground-truth views instead of estimated views. The trend remains, but the percentage of correct triplets is higher.

Each stage of Triplet Selection, selecting a smaller number of triplets with the highest confidence scores is better in terms of triplets’ correctness. Moreover, the percentage of correct triplets increases with stages, as the model performance on the target domain improves. Both insights emphasize the importance of learning in an easy-to-hard manner. Although the percentage of correct triplets is approximately 15%-20% (Fig. 5), GOUDA significantly improves direct testing results, due to its ability to reduce the target domain view bias. Higher percentage (up to 90%) is achieved when applying GOUDA using ground-truth views (see Fig. 6), highlighting the importance of quality view estimations.

View Analysis. Here we show the recognition accuracy as a function of view. Let us consider a triplet \( \{a, p, n\} \) from our triplet selection algorithm. As described in Section 3.2, \( p \) is the closest sample to \( a \) out of all cross view samples. This results in a limited diversity among the selected triplets, as each view is closely associated with a specific set of views based on its gait features (see Fig. 7). Despite that, we achieve significant improvements on the target domain test set. Fig. 8 presents the accuracy distribution on the test set according to view, before and after applying GOUDA. It is evident that the results improved for most of the cross view combinations, due to the indirect effect of changing the target domain view distribution (brighter colors in the right figure, implying higher accuracies). Furthermore, our method maintains the initial direct testing results on the exact same view combinations (diagonal), avoiding any degradation effects.

7. Summary

In this paper, we introduce a core problem of Gait Recognition models, often exhibiting a strong bias toward viewing angles when transferred to a new target domain. To reduce the presented gap causing a significant performance drop, we propose GOUDA - an unsupervised domain adaptation method, based on a new triplet selection strategy. We evaluate our method on multiple gait backbones and datasets, showing significant improvements and demonstrating superiority over prior works. We view our work as a first step toward exploring approaches to adapt gait models to target domains by looking at their view distributions.
References


