Fine-grained Unsupervised Domain Adaptation for Gait Recognition

Kang Ma[1], Ying Fu[1*], Dezhi Zheng[1*], Yunjie Peng[2], Chunshui Cao[3], Yongzhen Huang[3,4*]
[1]Beijing Institute of Technology, [2]Beihang University, [3]WATRIX.AI, [4]Beijing Normal University
kangx.ma@gmail.com, {fuying, zhengdezhi}@bit.edu.cn, YunjiePeng@buaa.edu.cn
chunshui.cao@watrix.ai, huangyongzhen@bnu.edu.cn

Abstract

Gait recognition has emerged as a promising technique for the long-range retrieval of pedestrians, providing numerous advantages such as accurate identification in challenging conditions and non-intrusiveness, making it highly desirable for improving public safety and security. However, the high cost of labeling datasets, which is a prerequisite for most existing fully supervised approaches, poses a significant obstacle to the development of gait recognition. Recently, some unsupervised methods for gait recognition have shown promising results. However, these methods mainly rely on a fine-tuning approach that does not sufficiently consider the relationship between source and target domains, leading to the catastrophic forgetting of source domain knowledge. This paper presents a novel perspective that adjacent-view sequences exhibit overlapping views, which can be leveraged by the network to gradually attain cross-view and cross-dressing capabilities without pre-training on the labeled source domain. Specifically, we propose a fine-grained Unsupervised Domain Adaptation (UDA) framework that iteratively alternates between two stages. The initial stage involves offline clustering, which transfers knowledge from the labeled source domain to the unlabeled target domain and adaptively generates pseudo-labels according to the expressiveness of each part. Subsequently, the second stage encompasses online training, which further achieves cross-dressing capabilities by continuously learning to distinguish numerous features of source and target domains. The effectiveness of the proposed method is demonstrated through extensive experiments conducted on widely-used public gait datasets.

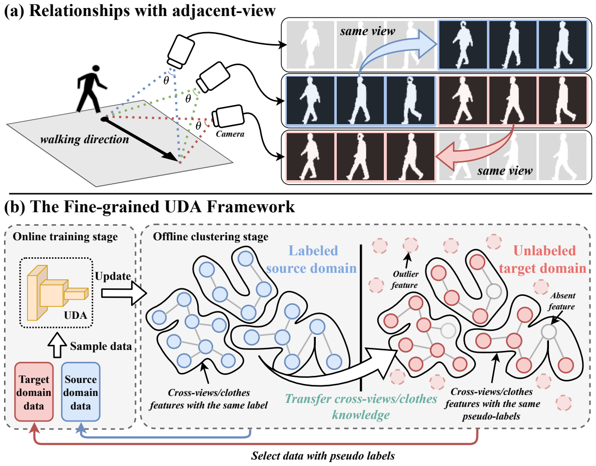
(a) Relationships with adjacent-view
same view
θ
θ
θ
Camera
walking direction
same view
(b) The Fine-grained UDA Framework
Online training stage
UDA
Update
Sample data
Target domain data
Source domain data
Offline clustering stage
Labeled source domain
Outlier feature
Unlabeled target domain
Absent feature
Cross-views/clothes features with the same label
Transfer cross-views/clothes knowledge
Cross-views/clothes features with the same pseudo-labels
Select data with pseudo labels

Figure 1. (a) The prior knowledge of the existence of angle overlap between adjacent-view. (b) The proposed Fine-grained Unsupervised Domain Adaptation (UDA) framework for gait recognition.

1. Introduction

Gait recognition, a biometric technology [37, 41, 15] that identifies individuals based on their unique walking patterns, has gained significant attention for its ability to recognize people at long distances, regardless of their attire. The applications [6, 31, 12] of gait recognition have been greatly facilitated by advances in deep learning and the availability of vast amounts of labeled data. However, accurately annotating large amounts of data remains a daunting challenge, especially under long-distance cameras, where critical information such as the human face is often blurred, making it difficult to label gait sequences of the same person in different clothing, views, and carrying conditions. As a solution, we propose a fine-grained Unsupervised Domain Adaptation (UDA) framework that transfers knowledge from a labeled source domain to an unlabeled target domain.

In the gait recognition task, an essential prior knowledge is the existence of overlapping angles in adjacent-view sequences as shown in Fig. 1 (a), which enables networks to build cross-view capabilities through a chain-like approach. *Leveraging this premise, the UDA framework for gait recognition can gradually learn cross-dressing capabilities in the target domain through knowledge transfer and clustering, without the need for pre-training the network.* We aim to reveal it by addressing the following aspects:
(i) Offline clustering stage. In the offline clustering phase,

[*]Corresponding Authors

as shown in Fig. 1, the primary goal is to assign accurate pseudo-labels to unlabeled data by leveraging cross-view chaining relationships. To achieve this, the network extracts essential insights for gait from a labeled source domain and subsequently transfers this acquired knowledge to the unlabeled target domain. Following this knowledge transfer, the precise pseudo-label is assigned to the cross-view and cross-dressing data present within the target domain. This assignment hinges on the observed global feature space distribution within the aforementioned target domain. Furthermore, previous methods [3, 10, 16, 25, 19, 28] demonstrate that part-based methods outperform other approaches. To this end, we leverage the varied expressive capabilities of different body parts and adaptively generate a global feature space, which results in a more accurate assignment of pseudo-labels to unlabeled data.

(ii) Online training stage. During the online training stage, the core objective is to continually learn the gait features in both the source and target domains. As depicted in Fig. 1 (a), the angles of the sequences vary continuously during human walking, with identical angles and actions appearing in sequences from adjacent viewpoints (indicated by the blue and red arrows). Consequently, the network can establish associations of neighboring perspectives by focusing on temporal features and gain cross-view capabilities through the association of all perspectives. Moreover, as shown in Fig. 1 (b), the network is trained by continuously selecting features from the unlabelled target domain that have been assigned the same pseudo-label as the training set. By continually adding data to the target domain in this way, the network focuses on more robust motion patterns to distinguish features from different domains. Prior methods [25, 26, 28] have demonstrated the criticality of motion pattern extraction in cross-dressing recognition. Consequently, the network gradually develops the cross-dressing capability, even in cases where it is absent from the source domain data.

Driven by this analysis, we propose a simple yet effective fine-grained UDA framework, wherein knowledge learned from the labeled source domain is transferred to obtain pseudo-labels for the unlabeled target domain. Subsequently, we utilize both labeled source and target domain data to extract fine-grained spatio-temporal motion patterns. As illustrated in Fig. 3, the offline clustering stage calculates the confidence scores for each part-based feature. These confidence scores are employed to derive the global feature space and assign pseudo-labels for unlabeled data. And the hybrid memory is initialized based on these pseudo-labels. Additionally, information scores are utilized to supervise the hybrid memory and facilitate the computation of the part-based contrastive loss. Furthermore, the online training stage is shown in Fig. 2, we propose a novel Spatio-Temporal Aggregation Network (STANet) for learning global motion patterns, which contains spatial designed Cycle Temporal Shift convolution (CTS-Conv) and Global context block. At each iteration, the hybrid memory is dynamically updated using the global motion patterns.

Our contributions can be summarized in three aspects.

- We propose a unified UDA framework structure based on the prior knowledge of the existence of an overlap between adjacent-view sequences, which enables our network to gradually acquire cross-view and cross-dressing capabilities without pre-training.
- We present an innovative STANet that effectively captures fine-grained motion patterns while enabling the acquisition of cross-view and cross-dressing capabilities. Our approach demonstrates superior performance than current state-of-the-art methods, particularly under the cross-dressing condition.
- The effectiveness of the proposed UDA framework is demonstrated through experimental results, which show superior performance on the CASIA-B [41], OUMVLP [32], and GREW [48] datasets. Moreover, comprehensive ablation experiments performed on the CASIA-B dataset further corroborate the efficacy of the proposed method within the UDA framework.

2. Related Work

In this section, we review the most relevant research in the area of unsupervised domain adaptation, part-based approaches, and spatial-temporal modeling.

2.1. Unsupervised Domain Adaptation Approaches

The Unsupervised Domain Adaptation (UDA) methods [30, 11, 44, 42, 14, 40, 8, 13, 7] for person re-identification (re-ID) have made significant progress. Advanced UDA methods for re-ID tasks often employ clustering algorithms, such as K-means [47], DBSCAN [9, 18], and InfoMAP [5], to generate pseudo-labels for unlabeled data from global features. For instance, MMT [13] used mutual mean-teaching to softly refine the pseudo labels in the target domain, AD-Cluster [42] incorporated style-translated images using a novel discriminative clustering method, and SpCL [14] employed a self-paced approach to create more reliable clusters gradually and refine the hybrid memory and the targets. Despite significant progress in recent years, adapting these methods for gait recognition tasks remains challenging due to inherent differences between gait and re-ID data. Specifically, re-ID data typically contains informative features such as texture and color, which enable the association of individuals without explicit labels. In contrast, the silhouette sequence in gait lacks such information, which poses a significant challenge for unsupervised tasks. Recent pioneering UDA methods for gait recognition [36, 45]

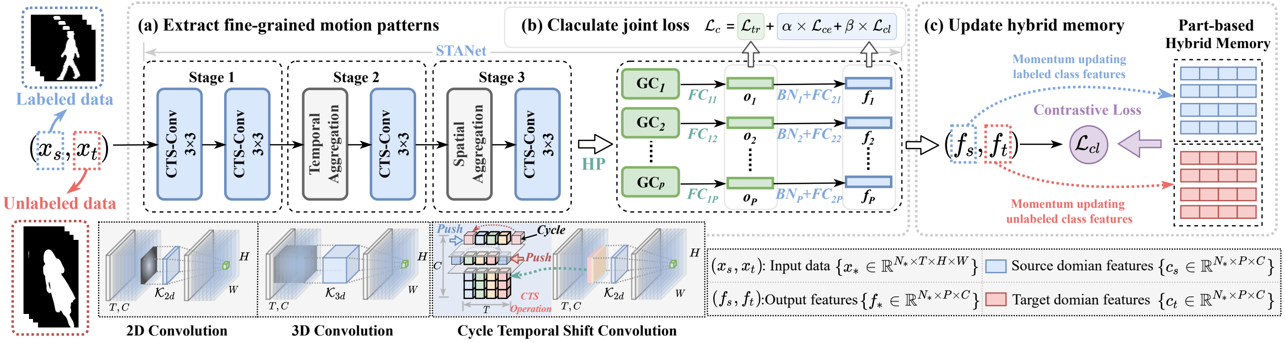
(a) Extract fine-grained motion patterns
(b) Claculate joint loss $\mathcal{L}_c = \mathcal{L}_{tr} + \alpha \times \mathcal{L}_{ce} + \beta \times \mathcal{L}_{cl}$
(c) Update hybrid memory
Part-based Hybrid Memory
Labeled data
Unlabeled data
(x_s, x_t)
STANet
Stage 1
Stage 2
Stage 3
CTS-Conv 3×3
CTS-Conv 3×3
Temporal Aggregation
CTS-Conv 3×3
Spatial Aggregation
CTS-Conv 3×3
HP
GC_1
GC_2
GC_P
FC_{11}
FC_{12}
FC_{1P}
o_1
o_2
o_P
BN_1+FC_{21}
BN_2+FC_{22}
BN_P+FC_{2P}
f_1
f_2
f_P
Momentum updating labeled class features
Contrastive Loss
(f_s, f_t)
$\mathcal{L}_{cl}$
Momentum updating unlabeled class features
T, C
$\mathcal{K}_{2d}$
H
W
$\mathcal{K}_{3d}$
Push
Cycle
Push
C
T
CTS Operation
2D Convolution
3D Convolution
Cycle Temporal Shift Convolution
(x_s, x_t): Input data $\{x_* \in \mathbb{R}^{N_* \times T \times H \times W}\}$
(f_s, f_t):Output features $\{f_* \in \mathbb{R}^{N_* \times P \times C}\}$
Source domian features $\{c_s \in \mathbb{R}^{N_* \times P \times C}\}$
Target domian features $\{c_t \in \mathbb{R}^{N_* \times P \times C}\}$

Figure 2. The overview of the online training stage in the proposed UDA framework, where “HP” represent the Horizontal Pooling, and “GC” represents the Global Context block. (a) The fine-grained motion patterns are extracted by Spatio-Temporal Aggregation Network (STANet). (b) The hybrid memory is used to supervise the training data and dynamically updated based on these features. (c) The features extracted by GC are used to calculate triplet loss, and the normalized features are used to calculate cross-entropy and contrastive losses.

have made some achievements on unlabeled datasets. However, these methods do not sufficiently consider the relationship between source and target domains, which ultimately restricts their performance and practical applications. As a solution, we propose a novel UDA framework that transfers valuable knowledge from a labeled source domain to an unlabeled target domain. By leveraging UDA, gait can be applied in more challenging real-world situations.

2.2. Part-based Approaches

The extraction of fine-grained features of body parts in gait recognition is a critical insight that plays a pivotal role to distinguish different individuals. Extensive research [3, 10, 16, 26, 19] has emphasized the significance of horizontally dividing these features into distinct parts. For example, GaitSet [3] introduced the Horizontal Pyramid Mapping (HPM) module, which partitions gait features into 62 strips along the height dimension. Since then, most gait recognition algorithms [16, 17, 25, 2] have adopted a part-based approach, training features with distinct parts separately. GaitPart [10] divided gait features horizontally into 16 strips, while GaitGL [26] separated global and local features into 32 strips, respectively. 3DLocal [19] proposed a localization module to learn the adaptive spatial and temporal scales. Therefore, gait recognition relies on the representation of multiple motion patterns with critical parts. However, the clustering algorithms typically assign pseudo-labels based on the global feature. To this end, we develop a fine-grained UDA framework for gait recognition that leverages the part-based features. Specifically, we evaluate the expressiveness of different parts to calculate confidence scores and then adaptively obtain the pseudo-labels.

2.3. Spatio-temporal Modeling

Gait recognition approaches can be broadly categorized into two primary domains: model-based [20, 34, 23, 24, 22] and appearance-based [39, 15, 29, 33, 34]. The appearance-based approaches have become the mainstream for gait recognition, which can be broadly classified into two categories, including silhouette-based methods [43, 3, 16, 10, 17] and video-based methods [25, 26, 2, 19, 38]. The silhouette-based methods generally employ 2D convolution (2D-Conv) neural networks to extract spatial features on either a single Gait Energy Image (GEI) or a collection of randomly arranged silhouette sets. For example, CNN-LB [39] introduced a CNN-based technique for gait feature learning from Gait Energy Images (GEI), while GaitSet [3] employed 2D convolution for feature extraction, assuming that the silhouette's appearance contained its positional information. However, these methods neglect temporal features which is the robust information in gait sequences. On the other hand, video-based techniques [10, 25, 26, 19] strive to acquire spatio-temporal features, which have been established as advantageous for enhancing gait recognition. For example, GaitGL [26] employed a 3D convolution (3D-Conv) neural network to extract spatio-temporal features. However, these methods use 3D-Conv to extract temporal features with a high amount of parameters. Different from these methods, we propose a more efficient solution named Cycle-Temporal Shit Convolution (CTS-Conv) that enables both low-level and high-level spatio-temporal features with a low computational cost. Further, we refined a global context modeling to extract robust motion patterns.

3. Method

In this section, we begin by introducing the overarching structure of fine-grained UDA framework in Sec. 3.1. Following that, we outline the offline clustering stage in Sec. 3.2, and subsequently detail the online training stage in Sec. 3.3. Finally, we present the joint loss in Sec. 3.4.

3.1. Overview

The UDA framework for gait recognition aims to transfer knowledge gained from labeled source domains to un-

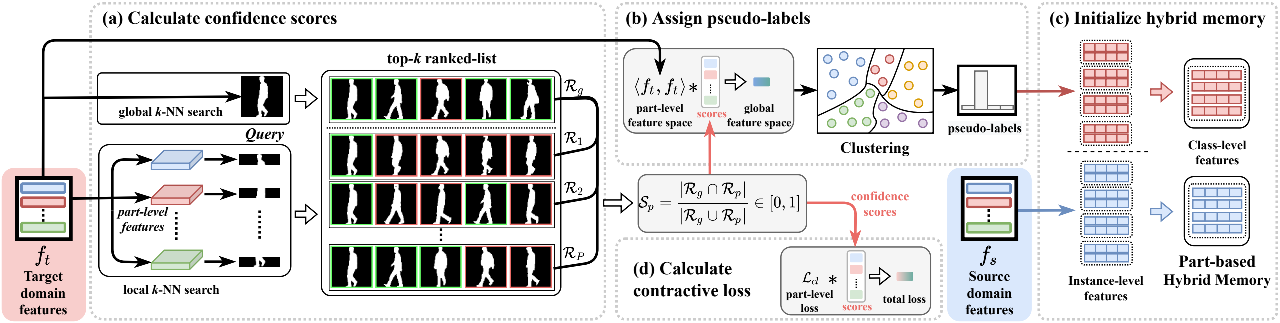

Figure 3. The offline clustering stage in the proposed UDA framework. (a) Calculating the confidence scores using a k-nearest neighbor search algorithm. (b) Generating pseudo labels using a clustering algorithm in the global space. (c) Initializing the hybrid memory using class-level features. (d) Using the confidence scores to further optimize contrastive loss.

labeled target domains, enabling accurate identification of individuals in unlabeled data. Following the previous UDA methods [50, 35, 4, 46, 21], as shown in Fig. 1 (b), we divide the training of the UDA framework into two stages, including the stage of offline clustering and online training.

In the offline clustering stage, we first assign pseudo-labels to unlabeled data. Then, we extract class-level central features from both source and target domains, which are used to initialize the hybrid memory module. Formally, Let $\mathcal{D}_s = \{(x_s, y_s)|x_s \in \mathbb{R}^{N_S \times T \times H \times W}, y_s \in \mathbb{R}^{N_S}\}$ and $\mathcal{D}_t = \{x_t|x_t \in \mathbb{R}^{N_T \times T \times H \times W}\}$ denote the labeled and unlabeled data, where N_S and N_T are the numbers of sequences. The generation of pseudo-labels y_t can be expressed as

$$y_t = \mathcal{C}\left(\langle \mathcal{F}_\theta(x_t), \mathcal{F}_\theta(x_t) \rangle\right), \tag{1}$$

where $\mathcal{C}$ represents the clustering algorithm, $\langle \cdot, \cdot \rangle$ denotes the operation to measure the similarity between gait features space, $\mathcal{F}_\theta$ is a learnable network parameterized by θ. During the online training stage, the pseudo-labeled data can be sampled with the labeled data. The gait features (f_s, f_t) extraction can be expressed as

$$(f_s, f_t) = (\mathcal{F}_\theta(x_s), \mathcal{F}_\theta(x_t)). \tag{2}$$

To perform gait recognition on unlabeled sequences, TranND [45] utilizes a fine-tuning approach, which may lead to catastrophic forgetting of the knowledge learned in the source domain when trained on the target domain. In this paper, the proposed UDA framework is an end-to-end approach that utilizes alternating “offline clustering stage” and “online training stage” to simultaneously train the data in both source and target domains.

3.2. Offline Clustering Stage

This part provides a comprehensive overview of the offline clustering stage, which generates the pseudo-labels for unlabeled data and initializes the hybrid memory.

Generating pseudo-labels. The offline clustering stage, as shown in Fig. 3, plays a crucial role in UDA tasks by generating pseudo-labels for unlabeled data. The effectiveness of the UDA algorithm relies on the distribution that features with the same label cluster densely in feature space, while those with different labels are scattered. However, *one of the main challenges in the UDA framework is generating accurate pseudo-labels by effectively utilizing the relationships between the part-level features*, given that features are comprised of multiple spatio-temporal motion patterns.

To accurately distribute unlabeled data, we take into account the expressive capabilities of different body parts and calculate the confidence score for each part feature accordingly. Formally, let $f_t \in \mathbb{R}^{N_T \times P \times C}$ denote the target domain part-level features, where P is the number of horizontally sliced parts, and C is the feature channels. Specifically, we perform the k-nearest neighbor search algorithm on both the global feature space and each part feature space to generate ranked lists $\{\mathcal{R}_g, \mathcal{R}_1, \ldots, \mathcal{R}_P\}$ for each unlabeled feature. The $\mathcal{R}_g$ represents the sets of global feature indices for the top-k samples in the ranked list, while $\mathcal{R}_p$ denotes the sets of p-th part feature indices. We then compute the confidence scores $\mathcal{S}_p$ between the global feature indices $\mathcal{R}_g$ and the p-th part feature indices $\mathcal{R}_p$ as follows,

$$\mathcal{S}_p = \frac{|\mathcal{R}_g \cap \mathcal{R}_p|}{|\mathcal{R}_g \cup \mathcal{R}_p|} \in [0, 1], \tag{3}$$

where $|\cdot|$ is a non-negative real number denoting the cardinality of a set. The confidence scores are further used to refine the part-based space to global feature space. We then use a clustering algorithm $\mathcal{C}$ (*e.g.*, DBSCAN [9]) to assign pseudo-labels on the global feature space, *i.e.*,

$$y_t = \mathcal{C}\left\{\mathcal{J}\left(\frac{1}{P}\sum_{p=1}^{P}\left[\langle f_t^p, f_t^p \rangle * \frac{\mathcal{S}_p}{\sum_{p=1}^{P} \mathcal{S}_p}\right]\right)\right\}, \tag{4}$$

where $\mathcal{J}$ is the Jaccard distance [49] computed in global feature space and “$*$” denotes the Hadamard product.

Initializing Hybrid Memory. To initialize the part-based hybrid memory, we utilize the class-level features of both source and target domains. It is worth illustrating that in the target domain, we only retain samples assigned with pseudo labels, and exclude those with discrete labels from the subsequent online training stage, *i.e.*,

$$\{c_s, c_t\} = \left\{ \frac{1}{|\mathcal{N}_s|} \sum_{f_s \in \mathcal{N}_s} f_s, \frac{1}{|\mathcal{N}_t|} \sum_{f_t \in \mathcal{N}_t} f_t \right\}, \tag{5}$$

where $\{c_s, c_t\}$ denotes the source and target domain class-level features, $(\mathcal{N}_s, \mathcal{N}_t)$ is the same label set or cluster set.

3.3. Online Training Stage

This part provides a detailed explanation of the Cycle Temporal Shift Convolution (CTS-Conv) and Global Context Block. Then, we introduce the utilization of class-level momentum updating for the part-based hybrid memory.

Cycle Temporal Shift Convolution. Gait recognition is a challenging task due to the impact of various external factors, including carrying conditions, wearing coats, and varying viewing angles, leading to a much higher intra-class variance than inter-class. Previous studies [25, 26] proposed to use 3D convolution (3D-Conv) neural network to capture temporal information in gait sequences and extract robust features. However, when applied to part-based UDA tasks, 3D-Conv is limited by its high computational cost and parameter-rich structure, increasing the risk of over-fitting. To address this, we propose a novel CTS-Conv, which separates temporal feature learning from spatial feature learning based on the well-designed temporal shift convolution (TS-Conv) [27]. In CTS-Conv, the input features are denoted as $\Upsilon = \{\upsilon_1, \upsilon_2, \ldots, \upsilon_E\} \in \mathbb{R}^{C \times T \times H \times W}$, where υ_i represents the input features divided equally into E pieces, and T denotes the number of frames. As shown in Fig. 2, the CTS-Conv operation can be expressed as

$$\overrightarrow{\upsilon}_1^t = \upsilon_1^{t-1}, \overrightarrow{\upsilon}_1^0 = \upsilon_1^T, \overleftarrow{\upsilon}_2^t = \upsilon_2^{t+1}, \overleftarrow{\upsilon}_2^T = \upsilon_2^0, \tag{6}$$

where $\overrightarrow{\upsilon}_1$ and $\overleftarrow{\upsilon}_2$ denote the first two pieces which are shifted to the right and left along the temporal dimension, respectively. The remaining pieces $[\upsilon_3, \upsilon_4, \cdots, \upsilon_E]$ are unshifted. The CTS-Conv operation in STANet learns temporal features by exchanging the features of adjacent frames, without introducing any additional parameters.

Algorithm 1: The training process of UDA for Gait

```
Require: Labeled data $\mathcal{D}_s$ and Unlabeled data $\mathcal{D}_t$;
Require: Random initialize pipeline $\mathcal{F}_\theta$;
Require: $m$ for Eq. (9) and $\tau$ of Eq. (10);
for $i \leq N_e$ do
    Offline Clustering Stage, $N_e$ is the total iterations;
    Assign pseudo-label for $\mathcal{D}_t$ with Eq. (4);
    Initialize the hybrid memory with Eq. (5);
    for $j \leq T_e$ do
        Online Training Stage, $T_e$ is the train iterations;
        Sample mini-batch data $(x_s, x_t)$;
        Encoder part-based features $(f_s, f_t)$;
        Compute the joint loss with Eq. (11);
        Update part-based hybrid memory with Eq. (9);
    end
end
```

Global Context Block. After extracting the spatio-temporal features using the CTS-Conv, we further refine the global context modeling technique called global context (GC) block base on [1, 28]. The GC block employs multiple heads to map features during context modeling, thereby facilitating the comprehensive extraction of global motion patterns. To generate the output feature, a horizontal pooling method (HP) is first utilized, resulting in horizontally sliced features denoted as $H = \{h^1, h^2, \ldots, h^P\} \in \mathbb{R}^{P \times T \times C}$, where P represents the number of horizontally sliced parts. Then, the global motion patterns g^p are derived using the subsequent expression:

$$\tilde{g}^p = \frac{\exp\left(W_k W_{v1} h_t^p\right)}{\sum_{t=1}^{T} \exp\left(W_k W_{v1} h_t^p\right)} \otimes W_{v1} h^p, \tag{7}$$

$$g^p = \delta\left(W_{v2} \tilde{g}^p\right) \oplus h^p, \tag{8}$$

where $\tilde{g}^p$ represents the context modeling, W_k, W_{v1}, and W_{v2} are learnable parameters, δ denotes the activation function, $\otimes$ signifies matrix multiplication, and $\oplus$ indicates the broadcast element-wise addition.

Updating Hybrid Memory. Each mini-batch of the sampled data is passed through a pipeline that extracts part-based features (f_s, f_t). It is worth noting that our proposed UDA framework is highly versatile and can be readily applied to other network architectures [3, 26], as illustrated in Fig. 2, where STANet can be replaced by any of the existing popular networks. At each iteration, the extracted features in each mini-batch would be involved in part-based hybrid memory momentum updating, *i.e.*,

$$\left\{ \begin{array}{l} c_s \leftarrow m \cdot f_s + (1-m) \cdot \frac{1}{|\mathcal{B}_s|} \sum_{f_s \in \mathcal{B}_s} f_s \\ c_t \leftarrow m \cdot f_t + (1-m) \cdot \frac{1}{|\mathcal{B}_t|} \sum_{f_t \in \mathcal{B}_t} f_t \end{array} \right\}, \tag{9}$$

where (c_s, c_t) denotes the class-level features, $(\mathcal{B}_s, \mathcal{B}_t)$ represents the set of features belonging to the source or target domain class and $m \in [0, 1]$ is a momentum coefficient.

3.4. Joint Loss Function

In this work, three types of part-based loss functions are employed, namely, the triplet loss $\mathcal{L}_{tp}$, the cross-entropy loss $\mathcal{L}_{ce}$, and the contrastive loss $\mathcal{L}_{cl}$. These loss functions are used to constrain the features of each part individually. As shown in Fig. 2, the output features (o_s, o_t) are used to calculate $\mathcal{L}_{tp}$, while the normalized output features (f_s, f_t) are used to calculate $\mathcal{L}_{ce}$ and $\mathcal{L}_{cl}$. Additionally, we consider the varying expressiveness of different body parts and

Table 1. The Rank-1 (%) performance comparison of fully supervised and unsupervised domain adaptation methods on CASIA-B. Where "OUMVLP→CASIA-B" denotes the use of OUMVLP as the source domain and CASIA-B as the target domain.

	Method	Probe View											Mean
		0°	18°	36°	54°	72°	90°	108°	126°	134°	162°	180°	
Fully supervised methods with all labels on CASIA-B													
NM	CNN-LB [39]	82.6	90.3	96.1	94.3	90.1	87.4	89.9	94.0	94.7	91.3	78.5	89.9
	GaitSet [3]	90.8	97.9	**99.4**	96.9	93.6	91.7	95.0	97.8	98.9	96.8	85.8	95.0
	GaitGL [26]	96.0	98.3	99.0	97.9	96.9	95.4	97.0	98.9	99.3	98.8	94.0	97.4
	STANet(*ours*)	**96.4**	**99.4**	99.3	**98.9**	**97.0**	**95.8**	**98.2**	**99.2**	**99.6**	**99.2**	**96.0**	**98.1**
BG	CNN-LB [39]	64.2	80.6	82.7	76.9	64.8	63.1	68.0	76.9	82.2	75.4	61.3	72.4
	GaitSet [3]	83.8	91.2	91.8	88.8	83.3	81.0	84.1	90.0	92.2	94.4	79.0	87.2
	GaitGL [26]	92.6	96.6	96.8	95.5	93.5	89.3	92.2	96.5	98.2	96.9	91.5	94.5
	STANet(*ours*)	**94.4**	**98.2**	**98.9**	**97.5**	**94.1**	**91.2**	**93.9**	**97.4**	**98.5**	**97.8**	**94.0**	**96.0**
CL	CNN-LB [39]	37.7	57.2	66.6	61.1	55.2	54.6	55.2	59.1	58.9	48.8	39.4	54.0
	GaitSet [3]	61.4	75.4	80.7	77.3	72.1	70.1	71.5	73.5	73.5	68.4	50.0	70.4
	GaitGL [26]	76.6	90.0	90.3	87.1	84.5	79.0	84.1	87.0	87.3	84.4	69.5	83.6
	STANet(*ours*)	**83.4**	**94.6**	**96.8**	**93.9**	**89.6**	**86.0**	**88.2**	**92.1**	**93.9**	**90.2**	**78.1**	**89.7**
Unsupervised Domain Adaptation methods without any labels on CASIA-B (OUMVLP→CASIA-B)													
NM	TranND [45]	-	-	-	-	-	-	-	-	-	-	-	63.4
	UDA(GaitSet)	**84.7**	**95.8**	**97.2**	**95.8**	**95.1**	**92.9**	**94.8**	**96.1**	**97.3**	**96.6**	79.2	**93.2**
	UDA(GaitGL)	76.7	88.3	88.8	87.1	88.4	82.9	85.9	90.6	93.1	93.9	73.0	86.2
	UDA(STANet)	84.3	95.2	95.5	91.1	90.7	85.9	89.4	93.3	94.9	93.7	**82.4**	90.6
BG	TranND [45]	-	-	-	-	-	-	-	-	-	-	-	42.7
	UDA(GaitSet)	**77.5**	87.5	**90.0**	**89.2**	**86.2**	**81.2**	**87.1**	**89.9**	**94.0**	**91.1**	70.9	**85.9**
	UDA(GaitGL)	69.1	85.3	84.7	83.6	82.9	76.1	79.9	86.8	89.1	88.0	63.6	80.8
	UDA(STANet)	75.4	**88.7**	86.6	83.6	82.9	76.2	82.6	86.9	89.3	87.5	**73.7**	83.0
CL	TranND [45]	-	-	-	-	-	-	-	-	-	-	-	15.8
	UDA(GaitSet)	24.2	38.9	42.3	39.6	41.1	38.1	37.9	43.3	43.4	29.6	25.4	36.7
	UDA(GaitGL)	43.9	65.3	66.5	66.9	68.8	60.0	66.4	71.2	69.3	61.8	42.2	62.0
	UDA(STANet)	**54.0**	**73.2**	**75.7**	**72.0**	**71.4**	**64.4**	**70.1**	**74.6**	**73.7**	**69.2**	**51.4**	**68.2**

use the confidence scores to accurately constrain the features by constraining the contrastive loss for each part. The contrastive loss $\mathcal{L}_{cl}$ can be formulated as follows:

$$\begin{aligned}\mathcal{L}_{cl} = &-\sum_{p=1}^{P} \frac{\mathcal{S}_p}{N_{cs}} \left\{ -\log \frac{\exp(f^p \cdot c_s^{p,+}/\tau)}{\sum_{k=1}^{|\mathcal{N}_s|} \exp(f^p \cdot c_s^{p,k}/\tau)} \right\} \\ &-\sum_{p=1}^{P} \frac{\mathcal{S}_p}{N_{ct}} \left\{ -\log \frac{\exp(f^p \cdot c_t^{p,+}/\tau)}{\sum_{k=1}^{|\mathcal{N}_t|} \exp(f^p \cdot c_t^{p,k}/\tau)} \right\},\end{aligned} \tag{10}$$

where N_{cs}, N_{ct} is a positive integer obtained by multiplying the mini-batch with parts, τ is a hyper-parameter, c_s, c_t denote the class-level features of source and target domains, and $c_s^{p,+}, c_t^{p,+}$ indicate the positive class prototype corresponding to f^p. The combined loss $\mathcal{L}_c$ can be expressed as

$$\mathcal{L}_c = \mathcal{L}_{tp} + \alpha\mathcal{L}_{ce} + \beta\mathcal{L}_{cl}, \tag{11}$$

where α and β are the hyper-parameters to balance the combined loss and are empirically set to 0.3.

4. Experiments

In this section, we evaluate the proposed method on three large-scale gait recognition benchmarks, including two widely adopted datasets (*i.e.*, CASIA-B and OUMVLP) and one real-world dataset (*i.e.*, GREW). Then, we compare our method to state-of-the-art methods. Finally, comprehensive ablation studies are conducted to substantiate the effectiveness of our proposed method.

4.1. Datasets and Implementation Details

CASIA-B [41] is a well-established dataset widely used in the field of gait recognition research. It comprises 124 subjects and encompasses 3 distinct walking conditions: normal walking (NM), walking while wearing a coat (CL), and walking with bags (BG). The dataset is captured by 11 cameras positioned uniformly between 0° and 180°, providing a diverse range of viewpoints. Each subject has a total of 110 sequences. The first 74 subjects are used for training, while the remaining 50 subjects are reserved for testing. During

Table 2. Rank-1 accuracy of fully supervised and unsupervised domain adaptation methods on OUMVLP under all view angles. Where “CASIA-B→OUMVLP” represents the use of CASIA-B as the labeled source domain and OUMVLP as the unlabeled target domain.

Method	Probe View														Mean
	0°	15°	30°	45°	60°	75°	90°	180°	195°	210°	225°	240°	255°	270°	
Fully supervised methods with all labels on OUMVLP															
GaitSet [3]	79.5	87.9	89.9	90.2	88.1	88.7	87.8	81.7	86.7	89.0	89.3	87.2	87.8	86.2	87.1
GaitGL [26]	84.9	90.2	91.1	91.5	91.1	90.8	90.3	88.5	88.6	90.3	90.4	89.6	89.5	88.8	89.7
STANet(*ours*)	**87.7**	**91.4**	**91.6**	**91.9**	**91.6**	**91.4**	**91.2**	**90.4**	**90.3**	**90.8**	**91.0**	**90.5**	**90.3**	**90.1**	**90.7**
Unsupervised Domain Adaptation methods without any labels on OUMVLP (CASIA-B→OUMVLP)															
TranND [45]	-	-	-	-	-	-	-	-	-	-	-	-	-	-	75.6
UDA(GaitSet)	45.6	60.1	75.3	80.7	75.5	73.9	73.9	49.9	59.9	75.5	81.7	74.9	73.2	72.9	69.5
UDA(GaitGL)	23.6	50.5	79.5	78.5	75.3	74.6	74.5	28.7	50.9	77.5	77.5	74.3	73.3	73.0	65.1
UDA(STANet)	**42.8**	**85.1**	**92.1**	**91.2**	**88.9**	**89.2**	**89.5**	**50.4**	**86.6**	**90.8**	**90.1**	**88.0**	**87.9**	**88.0**	**82.9**

testing, the first four “NM” conditions are utilized as the gallery, while the remaining conditions serve as the probe.

OUMVLP [32] is a prominent public gait dataset that comprises 10,307 subjects, two distinct walking sequences (labeled as “00” and “01”), and 14 cameras positioned uniformly between $(0° \rightarrow 90°)$ and $(180° \rightarrow 270°)$. To ensure fair data volume in comparison to CASIA-B, we select the first 1,000 subjects from OUMVLP for our UDA experiments without any specific instructions. In the UDA experiment, we followed the official data split and used 500 individuals for training and the remaining 500 for testing. During the test phase, the “00” sequences serve as the probe, while the remaining “01” sequences serve as the gallery.

GREW [48] is a comprehensive and extensive dataset designed for gait recognition, encompassing silhouettes and 2D/3D human body pose. In contrast to CASIA-B and OUMVLP, GREW was collected in diverse and complex scenes in the wild, making it a more challenging task for gait recognition. The GREW dataset comprises 26345 subjects, 128671 sequences, and 882 camera viewpoints. To ensure fairness in data volume with respect to CASIA-B, we use the first 4000 subjects from GREW for UDA experiments without any special instructions. During testing, we make full use of the official query and gallery set.

Implementation Details. The training process of UDA is shown in Algorithm 1, and all datasets are uniformly resized to 64 × 44 [3]. The model is trained end-to-end using the SGD optimizer, starting with an initial learning rate of 0.1, which is then decreased by a factor of 10 iteratively until convergence. The training mini-batch size for CASIA-B and OUMVLP is set to (8, 16), while for GREW dataset, it is set to (4, 32). Data augmentation techniques are utilized to preserve both the original and augmented data in real-world datasets. The convolutional channels in the STANet are set to (64, 128, 256). The network is trained for a total of N_e=100K iterations, and offline clustering is performed every T_e=5K iteration to assign pseudo-labels to unlabeled data and reinitialize the hybrid memory dictionary.

4.2. Comparison with State-of-the-art Methods

To evaluate the effectiveness of the proposed approach, we compared it with both existing fully supervised methods [39, 3, 26] and Unsupervised Domain Adaptation (UDA) method [45] for gait recognition.

Evaluation on CASIA-B. The performance of both fully supervised [39, 3, 26] and UDA methods [45] on CASIA-B is presented in Tab. 1. **(1)** The proposed STANet algorithm outperforms the fully supervised algorithm [39, 3, 26], especially in the cross-dressing condition. Experimental results show that our method has the ability to learn more robust motion patterns compared to others. **(2)** The proposed UDA framework outperforms the existing UDA algorithm [45] under all conditions. The experimental results demonstrate that our UDA framework can effectively utilize fine-grained part-based features to generate pseudo-labels that supervise network learning, leading to a significant improvement in recognition accuracy. **(3)** We conducted experiments by substituting STANet with various state-of-the-art algorithms, such as GaitSet [3] and GaitGL [26]. In the UDA framework, the GaitSet method outperforms other methods in “NM” and “BG” but has lower performance in “CL”. The experimental results demonstrate that GaitSet emphasizes learning cross-view information from the OUMVLP but struggles to capture robust motion patterns.

Evaluation on OUMVLP. To provide additional evidence of the effectiveness of our proposed method, we evaluated the proposed approach with some state-of-the-art fully supervised [3, 26] and UDA [45] methods on OUMVLP. **(1)** The proposed STANet outperforms other methods, including GaitSet [3] and GaitGL [26], under both the whole dataset and fully supervised conditions. Our experimental results demonstrate the critical importance of effectively utilizing fine-grained spatio-temporal motion patterns for achieving high performance on large-scale gait datasets. **(2)** Our proposed UDA framework outperforms the existing UDA algorithm [45]. The experimental results demonstrate that our framework can effectively transfer the norm walking information from CASIA-B to OUMVLP. **(3)** We conducted experiments to compare STANet with other algorithms, including GaitSet [3] and GaitGL [26]. The results showed that the proposed STANet in the UDA framework effectively utilized fine-grained part-based features, even in the presence of cross-dressing influence in CASIA-B.

Evaluation on GREW. We evaluated the proposed method on a wild dataset, *i.e.*, GREW, to provide additional evidence for the effectiveness of our method. **(1)** The experimental results presented in Tab. 3 demonstrate the effectiveness of the fine-grained motion patterns extracted by our method on real datasets. **(2)** The UDA experimental results confirm that knowledge gained from accurately labeled data in a laboratory setting can be transferred to real-world scenarios. This validation also confirm the efficacy of our proposed approach in extracting fine-grained motion patterns.

4.3. Ablation Study

In this section, an ablation study is performed on the proposed UDA framework to improve the understanding of the effects of different configurations.

The effectiveness of STANet. To comprehensively evaluate the impact of STANet in the UDA framework, we conducted a thorough comparative experiment in Tab. 4. Specifically, **(1)** The "w/o CTS-Conv" represents replacing the CTS-Conv with 2D convolution. The experiment demonstrates a decrease in the performance of the "CL", thereby confirming the ability of CTS-Conv to acquire resilient temporal features in the UDA framework. **(2)** Furthermore, we conducted additional experiments to remove the GC block, denoted as "w/o GC block", in which the GC block was replaced with Max Pooling along the temporal dimension. The experiment results showed that STANet exhibited performance comparable to that of GaitGL [26].

The effectiveness of confidence scores. To explore the impact of confidence scores on the offline clustering stage, we completely remove the process of generating confidence scores and generate the global feature space directly averaged over the part-based feature space. The experimental results "w/o Confidence scores" from the Tab. 4 show that removing the confidence score decreased the performance.

Analysis of the number of spatial slices parts. We conducted an in-depth analysis to assess the influence of different spatial bin numbers of the UDA framework in Tab. 4. **(1)** The experimental results indicate that the performance improves as the number of N_{part} increases. **(2)** In the field of gait recognition, many popular approaches [10, 26] have confirmed that the number of parts has a significant impact on the recognition results. Specifically, increasing the number of parts requires the network to focus more on fine-grained parts. Experimental results show that N_{part} also has a significant impact on unlabeled feature learning.

Table 3. Rank-1 (%) performance comparison of fully supervised and unsupervised domain adaptation methods on GREW.

Method	Source	Rank-1 (%)
Fully supervised methods on GREW		
CNN-LB [39]	TPAMI 2017	12.2
GaitSet [3]	AAAI 2019	33.5
GaitGL [26]	ICCV 2021	32.7
STANet(*ours*)	-	**41.3**
UDA methods without any labels (OUMVLP→GREW)		
UDA(Gaitset)	-	15.3
UDA(GaitGL)	-	13.7
UDA(STANet)	-	**26.2**

Table 4. The ablation study on CASIA-B with rank-1 accuracy.

Method	OUMVLP→CASIA-B			
	NM	BG	CL	Mean
Analysis of the UDA framework				
Baseline	**90.6**	**83.0**	**68.2**	**80.6**
- *w/o* CTS-Conv	91.7	84.7	62.5	79.6
- *w/o* GC block	86.3	79.9	62.3	76.1
- *w/o* Confidence scores	88.6	81.3	66.8	78.9
Analysis of the number spatial bins N_{part}				
N_{part} = 2	85.2	72.3	25.7	61.1
N_{part} = 4	83.5	76.2	51.3	70.3
N_{part} = 8	84.7	76.4	54.8	72.0
N_{part} = 16	84.2	77.1	59.4	73.6
N_{part} = 32	**90.6**	**83.0**	**68.2**	**80.6**

5. Conclusion

In this paper, we present a novel part-based Unsupervised Domain Adaptation (UDA) framework for gait recognition. Our framework consists of an offline clustering stage and an online training stage, which facilitate the knowledge transformation from a labeled source domain to an unlabeled target domain. Our method enhances the reliability of spatio-temporal features and extracts robust gait features for further knowledge transfer. We evaluate the proposed method on three popular gait datasets: CASIA-B, OUMVLP, and GREW, and our experimental results demonstrate the effectiveness of the UDA framework and its potential for practical applications. In future work, we plan to conduct more extensive investigations of UDA methods for gait recognition to accelerate network training and improve the accuracy of pseudo-labels.

6. Acknowledgement

This work was supported by the National Key R&D Program of China (2022YFC3300700), the National Natural Science Foundation of China (62171038, 62088101, 62276031 and 62276025), and the Fundamental Research Funds for the Central Universities.

References

[1] Yue Cao, Jiarui Xu, Stephen Lin, Fangyun Wei, and Han Hu. Gcnet: Non-local networks meet squeeze-excitation networks and beyond. In *ICCV*, 2019. 5

[2] Tianrui Chai, Annan Li, Shaoxiong Zhang, Zilong Li, and Yunhong Wang. Lagrange motion analysis and view embeddings for improved gait recognition. In *CVPR*, 2022. 3

[3] Hanqing Chao, Yiwei He, Junping Zhang, and Jianfeng Feng. Gaitset: Regarding gait as a set for cross-view gait recognition. In *AAAI*, 2019. 2, 3, 5, 6, 7, 8

[4] Yanbei Chen, Xiatian Zhu, and Shaogang Gong. Instance-guided context rendering for cross-domain person re-identification. In *ICCV*, 2019. 4

[5] Yoonki Cho, Woo Jae Kim, Seunghoon Hong, and Sung-Eui Yoon. Part-based pseudo label refinement for unsupervised person re-identification. In *CVPR*, 2022. 2

[6] Omar Costilla-Reyes, Ruben Vera-Rodriguez, Abdullah S Alharthi, Syed U Yunas, and Krikor B Ozanyan. Deep learning in gait analysis for security and healthcare. *Deep learning: algorithms and applications*, 2020. 1

[7] Yongxing Dai, Jun Liu, Yifan Sun, Zekun Tong, Chi Zhang, and Ling-Yu Duan. Idm: An intermediate domain module for domain adaptive person re-id. In *ICCV*, 2021. 2

[8] Weijian Deng, Liang Zheng, Qixiang Ye, Guoliang Kang, Yi Yang, and Jianbin Jiao. Image-image domain adaptation with preserved self-similarity and domain-dissimilarity for person re-identification. In *CVPR*, 2018. 2

[9] Martin Ester, Hans-Peter Kriegel, Jörg Sander, Xiaowei Xu, et al. A density-based algorithm for discovering clusters in large spatial databases with noise. In *kdd*, 1996. 2, 4

[10] 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 *CVPR*, 2020. 2, 3, 8

[11] Yang Fu, Yunchao Wei, Guanshuo Wang, Yuqian Zhou, Honghui Shi, and Thomas S Huang. Self-similarity grouping: A simple unsupervised cross domain adaptation approach for person re-identification. In *ICCV*, 2019. 2

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

[13] Yixiao Ge, Dapeng Chen, and Hongsheng Li. Mutual mean-teaching: Pseudo label refinery for unsupervised domain adaptation on person re-identification. In *ICLR*, 2020. 2

[14] Yixiao Ge, Feng Zhu, Dapeng Chen, Rui Zhao, et al. Self-paced contrastive learning with hybrid memory for domain adaptive object re-id. *NIPS*, 2020. 2

[15] Jinguang Han and Bir Bhanu. Individual recognition using gait energy image. *PAMI*, 2005. 1, 3

[16] Saihui Hou, Chunshui Cao, Xu Liu, and Yongzhen Huang. Gait lateral network: Learning discriminative and compact representations for gait recognition. In *ECCV*, 2020. 2, 3

[17] Xiaohu Huang, Duowang Zhu, Hao Wang, Xinggang Wang, Bo Yang, Botao He, Wenyu Liu, and Bin Feng. Context-sensitive temporal feature learning for gait recognition. In *ICCV*, 2021. 3

[18] Yangru Huang, Peixi Peng, Yi Jin, Yidong Li, and Junliang Xing. Domain adaptive attention learning for unsupervised person re-identification. In *AAAI*, 2020. 2

[19] Zhen Huang, Dixiu Xue, Xu Shen, Xinmei Tian, Houqiang Li, Jianqiang Huang, and Xian-Sheng Hua. 3d local convolutional neural networks for gait recognition. In *ICCV*, 2021. 2, 3

[20] Worapan Kusakunniran, Qiang Wu, Hongdong Li, and Jian Zhang. Multiple views gait recognition using view transformation model based on optimized gait energy image. In *ICCV*, 2009. 3

[21] Jianing Li and Shiliang Zhang. Joint visual and temporal consistency for unsupervised domain adaptive person re-identification. In *ECCV*, 2020. 4

[22] Xiang Li, Yasushi Makihara, Chi Xu, and Yasushi Yagi. End-to-end model-based gait recognition using synchronized multi-view pose constraint. In *ICCV*, 2021. 3

[23] Xiang Li, Yasushi Makihara, Chi Xu, Yasushi Yagi, Shiqi Yu, and Mingwu Ren. End-to-end model-based gait recognition. In *ACCV*, 2020. 3

[24] Rijun Liao, Shiqi Yu, Weizhi An, and Yongzhen Huang. A model-based gait recognition method with body pose and human prior knowledge. *PR*, 2020. 3

[25] Beibei Lin, Shunli Zhang, and Feng Bao. Gait recognition with multiple-temporal-scale 3d convolutional neural network. In *ACMMM*, 2020. 2, 3, 5

[26] Beibei Lin, Shunli Zhang, and Xin Yu. Gait recognition via effective global-local feature representation and local temporal aggregation. In *ICCV*, 2021. 2, 3, 5, 6, 7, 8

[27] Ji Lin, Chuang Gan, and Song Han. Tsm: Temporal shift module for efficient video understanding. In *ICCV*, 2019. 5

[28] Kang Ma, Ying Fu, Dezhi Zheng, Chunshui Cao, Xuecai Hu, and Yongzhen Huang. Dynamic aggregated network for gait recognition. In *CVPR*, pages 22076–22085, 2023. 2, 5

[29] Yasushi Makihara, Atsuyuki Suzuki, Daigo Muramatsu, Xiang Li, and Yasushi Yagi. Joint intensity and spatial metric learning for robust gait recognition. In *CVPR*, 2017. 3

[30] Liangchen Song, Cheng Wang, Lefei Zhang, Bo Du, Qian Zhang, Chang Huang, and Xinggang Wang. Unsupervised domain adaptive re-identification: Theory and practice. *Pattern Recognition*, 2020. 2

[31] Yingnan Sun and Benny Lo. An artificial neural network framework for gait-based biometrics. *Journal of biomedical and health informatics*, 2018. 1

[32] Noriko Takemura, Yasushi Makihara, Daigo Muramatsu, Tomio Echigo, and Yasushi Yagi. Multi-view large population gait dataset and its performance evaluation for cross-view gait recognition. *IPSJ Transactions on Computer Vision and Applications*, 2018. 2, 7

[33] Dacheng Tao, Xuelong Li, Xindong Wu, and Stephen J Maybank. General tensor discriminant analysis and gabor features for gait recognition. *PAMI*, 29, 2007. 3

[34] Chen Wang, Junping Zhang, Liang Wang, Jian Pu, and Xiaoru Yuan. Human identification using temporal information preserving gait template. *PAMI*, 2012. 3

[35] Dongkai Wang and Shiliang Zhang. Unsupervised person re-identification via multi-label classification. In *CVPR*, 2020. 4

[36] Likai Wang, Ruize Han, Wei Feng, and Song Wang. From indoor to outdoor: Unsupervised domain adaptive gait recognition. *arXiv:2211.11155*, 2022. 2

[37] Liang Wang, Tieniu Tan, Huazhong Ning, and Weiming Hu. Silhouette analysis-based gait recognition for human identification. *PAMI*, 2003. 1

[38] Thomas Wolf, Mohammadreza Babaee, and Gerhard Rigoll. Multi-view gait recognition using 3d convolutional neural networks. In *ICIP*, 2016. 3

[39] Zifeng Wu, Yongzhen Huang, Liang Wang, Xiaogang Wang, and Tieniu Tan. A comprehensive study on cross-view gait based human identification with deep cnns. *PAMI*, 2016. 3, 6, 7, 8

[40] Hong-Xing Yu, Wei-Shi Zheng, Ancong Wu, Xiaowei Guo, Shaogang Gong, and Jian-Huang Lai. Unsupervised person re-identification by soft multilabel learning. In *CVPR*, 2019. 2

[41] Shiqi Yu, Daoliang Tan, and Tieniu Tan. A framework for evaluating the effect of view angle, clothing and carrying condition on gait recognition. In *ICPR*, 2006. 1, 2, 6

[42] Yunpeng Zhai, Shijian Lu, Qixiang Ye, Xuebo Shan, Jie Chen, Rongrong Ji, and Yonghong Tian. Ad-cluster: Augmented discriminative clustering for domain adaptive person re-identification. In *CVPR*, 2020. 2

[43] Kaihao Zhang, Wenhan Luo, Lin Ma, Wei Liu, and Hongdong Li. Learning joint gait representation via quintuplet loss minimization. In *CVPR*, 2019. 3

[44] Xinyu Zhang, Jiewei Cao, Chunhua Shen, and Mingyu You. Self-training with progressive augmentation for unsupervised cross-domain person re-identification. In *ICCV*, 2019. 2

[45] Jinkai Zheng, Xinchen Liu, Chenggang Yan, Jiyong Zhang, Wu Liu, Xiaoping Zhang, and Tao Mei. Trand: Transferable neighborhood discovery for unsupervised cross-domain gait recognition. In *International Symposium on Circuits and Systems*, 2021. 2, 4, 6, 7

[46] Kecheng Zheng, Cuiling Lan, Wenjun Zeng, Zhizheng Zhang, and Zheng-Jun Zha. Exploiting sample uncertainty for domain adaptive person re-identification. In *AAAI*, 2021. 4

[47] Kecheng Zheng, Wu Liu, Lingxiao He, Tao Mei, Jiebo Luo, and Zheng-Jun Zha. Group-aware label transfer for domain adaptive person re-identification. In *CVPR*, 2021. 2

[48] Zhu Zheng, Guo Xianda, Yang Tian, Huang Junjie, Deng Jiankang, Huang Guan, Du Dalong, Lu Jiwen, and Zhou Jie. Gait recognition in the wild: A benchmark. In *ICCV*, 2021. 2, 7

[49] Zhun Zhong, Liang Zheng, Donglin Cao, and Shaozi Li. Re-ranking person re-identification with k-reciprocal encoding. In *CVPR*, 2017. 4

[50] Zhun Zhong, Liang Zheng, Zhiming Luo, Shaozi Li, and Yi Yang. Invariance matters: Exemplar memory for domain adaptive person re-identification. In *CVPR*, 2019. 4