Progressive Feature Learning for Realistic Cloth-Changing Gait Recognition

Xuqian Ren[1*], Juho Kannala[2,3], and Esa Rahtu[1]

[1] Tampere University, Finland {xuqian.ren, esa.rahtu}@tuni.fi
[2] Aalto University, Finland juho.kannala@aalto.fi
[3] University of Oulu, Finland

Abstract. Existing datasets and methods for gait recognition struggle to address the challenging issue of cloth-changing scenarios effectively. In practice, gait recognition models are often trained using automatically labeled data, where sequences are constrained by specific view and clothing conditions. Specifically, cross-view datasets typically include sequences under normal walking conditions without cloth changes. In contrast, cross-cloth datasets feature clothing variations but are limited to front views. This imbalance leads to suboptimal performance under realistic conditions. To tackle this issue, we define the task as Realistic Cloth-Changing Gait Recognition (abbreviated as RCC-GR) and we construct two benchmarks: CASIA-BN-RCC and OUMVLP-RCC, to simulate real-world scenarios. Furthermore, we propose a novel framework, Progressive Feature Learning, compatible with existing backbone models, to improve recognition performance in RCC-GR. Our framework includes two key components: Progressive Mapping and Progressive Uncertainty. These techniques first focus on extracting cross-view features and then incorporate cross-cloth features, ensuring that features from the cross-view dataset dominate the feature space. This approach mitigates the uneven distribution caused by cross-cloth variations. Experiments conducted on our benchmarks demonstrate that the proposed framework significantly improves recognition accuracy, particularly under cloth-changing conditions.

Keywords: Realistic Cloth-Changing · Gait Recognition · Progressive Feature Learning.

1 Introduction

Gait recognition is an important tool in video surveillance for identity verification and behavior analysis [1,13]. However, handling cloth-changing scenarios remains a significant challenge [15,27]. Among them, a key challenge is to deal with the cloth-changing condition in real-world applications. Previous research for gait recognition is mostly conducted with the assumption that the sequences of all views and clothes for each person are available, which, however, **is nearly impossible in practice** when collecting

* Corresponding Author

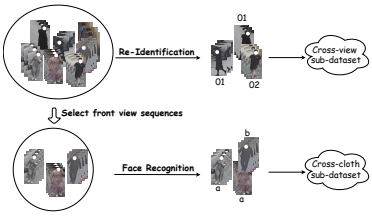


Fig. 1: The automatic label pipeline. The images come from DeepChange dataset [25].

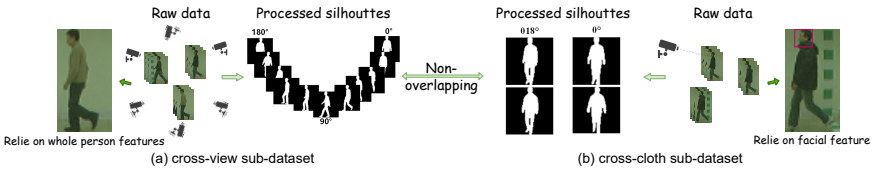


Fig. 2: The composition of a real cloth-changing benchmark. (a) Cross-view sub-dataset. Each person has view variations but only has the normal walking condition (NM). (b) Cross-cloth sub-dataset. Each identity has walking in different coat conditions (CL) but only has limited views (only front views).

large labeled datasets, especially in the wild [30]. While sequences for each person may include multiple views and clothing variations, manually associating these sequences is time-consuming and costly. Therefore, it is more practical to use automatic methods to assign labels to these sequences in order to get a labeled dataset.

As illustrated in Figure 1, when collecting the gait benchmarks in the wild automatically, it is relatively easier to obtain the cross-view sequences for each person aided by techniques such as person re-identification (ReID) [5, 29]. These methods rely primarily on clothing features, such as color and type, resulting in sub-datasets that include only view variations with consistent clothing. On the other hand, clustering cross-cloth sequences depends on facial features [7, 8], which are typically available only in front-view sequences. This limitation significantly reduces the size of the cross-cloth sub-dataset compared to the cross-view dataset. Furthermore, the identities in the cross-view and cross-cloth subsets nearly have no overlap because we use two different data collection methods to get the two sub-datasets. Even when there is some overlap, differences in assigned labels make it challenging for models to simultaneously learn cross-view and cross-cloth recognition. The datasets generated through this automatic labeling approach have the following key characteristics: (1) The cross-view sub-dataset includes view variations but no clothing changes. (2) The cross-cloth sub-dataset consists only of front-view sequences with clothing variations. (3) The cross-view sub-dataset is significantly larger than the cross-cloth sub-dataset. (4) The two sub-datasets nearly have no overlap. Figure 2 illustrates the structure of a realistic cloth-changing dataset. Unlike manually labeled datasets collected in the lab or under controlled situations, this approach closely resembles real-world challenges. To improve the performance of existing gait recognition methods on such data, developing a new framework specifically designed for these conditions is essential.

In this work, we introduce a new task, Realistic Cloth-Changing Gait Recognition (RCC-GR), to address the practical challenges of gait recognition in real-world scenarios, which is much more meaningful but has not attracted enough attention. To support our research, we first construct two benchmarks: **CASIA-BN-RCC** and **OUMVLP-RCC**, based on two famous datasets: CASIA-B [27] and OU-MVLP [23]. Each benchmark consists of two sub-datasets: **Cross-View Sub-dataset**: Contains sequences with view variations under normal walking conditions (NM). **Cross-Cloth Sub-dataset**: Includes sequences with clothing variations (normal walking and walking with coats) but limited to front and back views (0° and 180°). Particularly, when we use appearance-based methods, we take silhouettes as the input for gait recognition in this work, so the sequences around 180° can also be included for the cross-cloth subset due to the horizontal symmetry to enlarge the dataset volume. This construction ensures the benchmarks simulate real-world conditions by separating cross-view and cross-cloth features, providing a practical evaluation setting for gait recognition methods.

After observing the statistics of the datasets, we propose a new framework that utilizes progressive feature learning to realize gait recognition. The framework mainly consists of two methods to map features extracted from the backbone into discriminative feature spaces in a cascaded way, called Progressive Mapping and Progressive Uncertainty, respectively. Specifically, the previous method used a one-stage module to map features in the cross-view and cross-cloth feature space simultaneously. However, in our benchmarks, the cross-view sub-dataset has richer intra-class diversity and has a wider feature span in the high dimensional space, whereas the cross-cloth sub-dataset has a relatively smaller spatial span in the feature space due to its scarcity in intra-class variety. So if the feature extracted from the two sub-datasets is mapped together, the spatial space will be distorted [18]. Therefore, our Progressive Mapping applies a two-stage module, which contains a Cross-View Mapping module (CVM) and a Cross-Cloth Mapping module (CCM) to fully span the cross-view feature in the spatial space first, and then extract the cross-cloth feature on the basis. In this fashion, we relieve the uneven distribution in the feature space.

Besides, the Progressive Uncertainty maps each sequence as a Cross-View Gaussian distribution and a Cross-Cloth Gaussian distribution instead of a point. Particularly, mapping each sequence as Gaussian distributions allows the network to increase gradient weight from easy learning cross-view sequences and prevent the network from overfitting on the hard cross-cloth sequences. Furthermore, by re-sampling new embeddings from the Gaussian distributions, we can gain more knowledge from the sequence that can be used to classify each identity. Therefore, mapping sequences as Gaussian distributions can further make the feature extracted from the cross-view sub-dataset have more effects in model training.

To sum up, our contributions mainly lie in three folds:

- We define a new setting for cloth-changing gait recognition, requiring models to learn from two non-overlapping sub-datasets.
- We introduce a Progressive Feature Learning framework, leveraging Progressive Mapping and Progressive Uncertainty to improve feature representation in a cascaded way.

- Extensive experiments on CASIA-BN-RCC and OUMVLP-RCC demonstrate the effectiveness and superior performance of our framework in real-world applications.

Our framework is compatible with existing backbone models, providing a practical and impactful solution for gait recognition in complex, real-world settings. We hope this work serves as a foundation for further advancements in the field.

2 Related Work

Model-based Gait Recognition [14, 16] aims to extract human pose parameters that are more invariant to the view, clothing, and carrying variations from videos to realize gait recognition. This approach provides a physical explanation for gait recognition and works best with high-resolution inputs. In contrast, appearance-based gait recognition focuses on extracting robust and discriminative features from silhouettes. These methods perform well even at lower resolutions, making them ideal for outdoor applications.

Data uncertainty learning has been increasingly used in tasks like face recognition [20], person re-identification [22,24,28], and other computer vision applications [6, 10,11,26,31]. This approach enhances robustness by helping networks handle noisy inputs and avoid false recognition. SCFU [21] exploits spatial-wise and channel-wise uncertainty to solve the occluded person re-identification problem. Inspired by this work, we introduce uncertainty learning to the cross-cloth gait recognition problem. The variance generated by our framework can model the optimization difficulty caused by the quantity imbalance and silhouette variation diversity. In this way, the class lacking intra-class diversity and the parts with more variations in each feature will have larger variance and contribute less to the gradient, which makes the model learn a robust template from the classes that have rich feature diversity.

3 Our Approach

In this paper, we propose a new task called **Realistic Cloth-Changing Gait Recognition**, which focuses on the cloth-changing problem in practice. In this section, we will formally define the problem and then introduce a framework with Progressive Feature Learning to fully extract the cross-view and cross-cloth information in realistic datasets. Specifically, we will describe the details of the two new methods we used in our framework: Progressive Mapping and Progressive Uncertainty.

3.1 Problem Definition

The training set is represented as $(\mathcal{X}_v \cup \mathcal{X}_c)$, where $\mathcal{X}_v$ contains cross-view sequences under normal walking (NM) conditions, and $\mathcal{X}_c$ includes sequences with NM and clothing variations (CL) restricted to front views. These sub-datasets are non-overlapping, with $\mathcal{X}_v$ being significantly larger due to the ease of collecting cross-view data using ReID. During testing, sequences from all conditions and views are used for evaluation. So when considering the properties of the proposed benchmarks, we need to alleviate the imbalance in datasets and fully explore the cross-view and cross-cloth information.

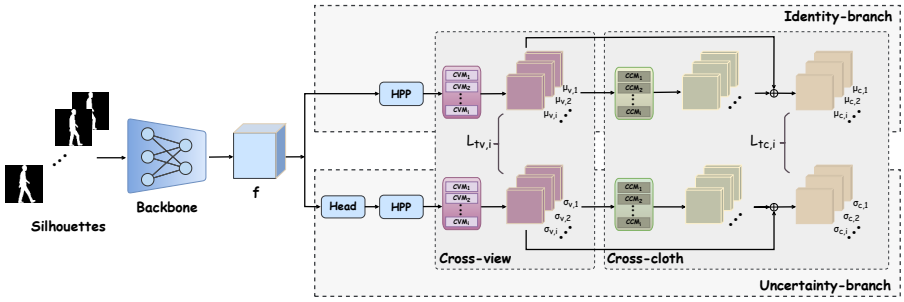


Fig. 3: The structure of our framework. Backbone is used to extract features from silhouettes. The feature f outputs from the backbone will be put into two branches, and each uses Progressive Mapping with Cross-View Mapping module (CVM) and Cross-Cloth Mapping module (CCM) to generate $\mu_{j,i}/\sigma_{j,i};\ j \in v, c;\ i \in 1, 2, ..., S$. *Head* is a light block used to isolate features of two branches. Progressive Uncertainty constructs two Gaussian distributions with learned mean and variance features to relax constraints on feature mapping.

3.2 Framework Overview

To solve this realistic cloth-changing task, we propose a framework that utilizes Progressive Feature Learning. Our framework can be applied to both appearance-based and model-based methods. To simplify the illustration, we take silhouettes as input which is more suitable for realistic low-resolution conditions.

According to the sequence quantity proportion in $\mathcal{X}_v$ and $\mathcal{X}_c$, there is a data imbalance in benchmarks, which will make the model biased if it learns the cross-view and cross-cloth ability at the same time. If the cross-view feature and cross-cloth feature are treated equally, the uneven data distribution will make it hard to separate from each other and distort feature spaces. So, making features from $\mathcal{X}_v$ domain the feature space first can train the model more robust. To achieve this, we design Progressive Mapping and Progressive Uncertainty in our framework, and the overview is shown in Figure 3. The training pipeline is as followings. First, the feature extracted from a sequence through the backbone is named f. f will be transferred into two branches, called the identity-branch and uncertainty-branch separately. Each branch employs Cross-View Mapping (CVM) to extract cross-view features (μ_v, σ_v) from $\mathcal{X}_v$. These features are then used as a foundation for Cross-Cloth Mapping (CCM), which extracts cross-cloth features (μ_c, σ_c) from $\mathcal{X}_c$. Next, Progressive Uncertainty will map each sequence as a cross-view Gaussian distribution $\mathcal{N}\left(\mu_v, \sigma_v^2 \mathrm{I}\right)$ and a cross-cloth Gaussian distribution $\mathcal{N}\left(\mu_c, \sigma_c^2 \mathrm{I}\right)$ in feature spaces. Mapping features as Gaussian distributions can make the feature extracted from the $\mathcal{X}_v$ domain the feature space first and then span cross-cloth features from $\mathcal{X}_c$. Finally, Uncertainty-aware Triplet Loss [4] $\mathcal{L}_{tv}$ and $\mathcal{L}_{tc}$ are employed to train the whole framework.

3.3 Progressive Mapping

Progressive Mapping aims to make features from $\mathcal{X}_v$ domain the feature space first, making the model learn to separate each identity better. This method has two key components: A two-stage module and Progressive-aware Triplets.

Two-stage module: A two-stage module is designed to span the feature used for cross-view and cross-cloth in a cascaded way. Similar to approaches in GaitSet [3] and GaitPart [9], a Horizontal Pyramid Pooling (HPP) [3] first divides the feature horizontally into S parts:

$$z_i = HPP(f_i);\ i \in 1, 2, ..., S \tag{1}$$

Each feature part, z_i, is processed through the Cross-View Mapping module (CVM) to extract features from $\mathcal{X}_v$. Subsequently, the Cross-Cloth Mapping module (CCM) processes the output of CVM to extract features from $\mathcal{X}_c$. Both CVM and CCM work in parallel across the S parts of the feature. A residual connection is added between CVM and CCM to retain cross-view information. CCM can alleviate the pressure on models that are trained to extract the cross-view and the cross-cloth features simultaneously. We take the identity-branch as an example and define the module with the following formulation:

$$\mu_{v,i} = CVM_i(z_i) \tag{2}$$

$$\mu_{c,i} = \mu_{v,i} + CCM_i(\mu_{v,i}) \tag{3}$$

where $\mu_{v,i}$ and $\mu_{c,i}$ are outputs from CVM and CCM, respectively. By sequentially learning cross-view and cross-cloth features, the model can better focus on identity classification first and then address cross-cloth variations.

Progressive-aware Triplets: To maximize the effectiveness of the two-stage module, a redesigned triplet selection method is introduced. We denote a sample triplet as (a, p, n), where a denotes the anchor, p denotes a sample whose label is as same as the anchor a, and n is a sample whose labels are different from that of a. In our work, we only choose persons from $\mathcal{X}_v$ in each batch to construct $\mathcal{T}_v = (a, p, n)$ when calculating the loss behind CVM, making it only extract the cross-view feature. Then we use the whole batch to construct $\mathcal{T}_c = (a, p, n)$ when calculating loss behind CCM, so the CCM can make full use of the whole information to cross-cloth as well as maintain the cross-view ability. Also, like in previous research, we select $p \times k$ sequences in each iteration, where p represents the number of persons and k is the number of training sequences each person has in the batch. To make the triplets maximum in cross-cloth triplets as well as guarantee the equality, we choose to randomly select $(p/2, k)$ batch from $\mathcal{X}_v$, and $(p/2, k)$ batch from $\mathcal{X}_c$ with $k/2$ from the NM and $k/2$ from CL. More specifically, for CASIA-BN-RCC, the (a, p, n) are chosen from the 1-44id for $\mathcal{L}_{tv}$, and the (a, p, n) are chosen from the 1-74id for $\mathcal{L}_{tc}$; for OUMVLP-RCC, the (a, p, n) are chosen from the first 3200ids of the training dataset for $\mathcal{L}_{tv}$, and the (a, p, n) are chosen from the whole 5143id from training set for $\mathcal{L}_{tc}$. This method ensures the model prioritizes learning features from $\mathcal{X}_v$ and then extracts features from both $\mathcal{X}_v$ and $\mathcal{X}_c$.

3.4 Progressive Uncertainty

Progressive Uncertainty represents each sequence as a Gaussian distribution in cross-view and cross-cloth feature spaces, relaxing the constraints on the model and improv-

ing robustness. The identity branch learns the mean feature, $\mu_{j,i}$, which represents the sequence's characteristics and is used during inference. The uncertainty branch, discarded during inference, generates the variance feature, $\sigma_{j,i}$, which measures the uncertainty in feature representation. In this branch, the original feature f outputs from the backbone will first pass through a Head module, which is a light block with several convolution layers, to make the feature isolated from the identity branch. Then, like the identity-branch, a two-stage module is employed in the Uncertainty-branch to generate two variance features as follows:

$$\sigma_{v,i} = Relu\left(CVM_i(HPP\left(Head(f)\right))\right) \tag{4}$$

$$\sigma_{c,i} = Relu\left(\sigma_{v,i} + CCM_i\left(\sigma_{v,i}\right)\right) \tag{5}$$

where $\sigma_{v,i}$ is the output from each module in CVM and $\sigma_{c,i}$ is the result that added the outputs from CVM and CCM. To ensure the variances are positive, we apply the Rectified linear unit. With mean feature $\mu_{v,i}$ and $\mu_{c,i}$ and variance feature $\sigma_{v,i}$ and $\sigma_{c,i}$, we construct Cross-View Gaussian distributions $\mathcal{N}\left(\mu_{v,i}, \sigma_{v,i}^2\mathrm{I}\right)$ and Cross-Cloth Gaussian distributions $\mathcal{N}\left(\mu_{c,i}, \sigma_{c,i}^2\mathrm{I}\right)$. In this way, the representation of each sequence is not point embeddings but stochastic embeddings sampled from $\mathcal{N}\left(\mu_{j,i}, \sigma_{j,i}^2\mathrm{I}\right), j \in v, c$. $\mu_{j,i}$ used during inference can be regarded as sampled from the center point of the Gaussian distributions. The variances measure the uncertainty of the classification result of each stage for the Progressive Mapping, which indirectly benefits the supervision information of the label and optimizes the training process [21].

The re-parameterization trick: To enrich the information that can be used in classification, we can also re-sample a new embedding from each Gaussian distribution. However, the sampling operation mentioned above is not differentiable and prevents the backpropagation of the gradients flow during training [2]. To solve the problem, we employ a re-parameterization trick [12] to let the model backpropagate normally like in [2, 21, 28]. A random noise ϵ, which is independent of the model parameters, is sampled from a normal distribution and then used to generate the embedding $e_{j,i}$, making them be regarded as re-sampled representations from Gaussian distributions mapped by a sequence.

$$e_{j,i} = \mu_{j,i} + \varepsilon\sigma_{j,i};\ \epsilon \sim \mathcal{N}(0, \mathrm{I});\ j \in v, c;\ s \in 1, 2, ..., S \tag{6}$$

where $e_{j,i}$ is the embedding re-sampled with the help of $\mu_{j,i}$ and $\sigma_{j,i}$.

Uncertainty-aware Triplet Loss: The mean feature $\mu_{j,i}$, variance feature $\sigma_{j,i}$, and the sampled embedding $e_{j,i}$ are used to control the triplet loss. We employ an uncertainty-aware triplet loss [4] after CVM and CCM separately:

$$\begin{aligned}
\mathcal{L}_{tv,i} =& \frac{1}{|\mathcal{T}_v|} \sum_{(a,p,n)\in\mathcal{T}_v} \left[\frac{\left\|\mu_{v,i}^a - \mu_{v,i}^p\right\|^2}{\sigma_{v,i}^{a^2} + \sigma_{v,i}^{p^2}} - \frac{\left\|\mu_{v,i}^a - \mu_{v,i}^n\right\|^2}{\sigma_{v,i}^{a^2} + \sigma_{v,i}^{n^2}} + m \right]_+ \\
\mathcal{L}_{tc,i} =& \frac{1}{|\mathcal{T}_c|} \sum_{(a,p,n)\in\mathcal{T}_c} \left[\frac{\left\|\mu_{c,i}^a - \mu_{c,i}^p\right\|^2}{\sigma_{c,i}^{a^2} + \sigma_{c,i}^{p^2}} - \frac{\left\|\mu_{c,i}^a - \mu_{c,i}^n\right\|^2}{\sigma_{c,i}^{a^2} + \sigma_{c,i}^{n^2}} + m \right]_+
\end{aligned} \tag{7}$$

where $|\mathcal{T}_i|, i = v, c$ is the number of triplets. We can also calculate $\mathcal{L}_{tv,i}^e$ and $\mathcal{L}_{tv,i}^e$ by replacing $\mu_{v,i}$ and $\mu_{c,i}$ with $e_{v,i}$ and $e_{c,i}$.

In total, the final loss in our framework can be represented as:

$$\mathcal{L} = \mathbb{E}(\sum_{i=0}^{S} \mathcal{L}_{tv,i} + \mathcal{L}_{tc,i} + \mathcal{L}^{e}_{tv,i} + \mathcal{L}^{e}_{tc,i}) \tag{8}$$

$\mathbb{E}(\cdot)$ means we sum each triplet loss and average them to gain the final loss.

4 Experiment

In this section, we evaluate the effectiveness of our framework using two representative backbones: GaitSet [3] and GaitGL [17]. GaitSet operates on image set inputs, while GaitGL processes video inputs. It should be noted that our framework can be adopted with both model-based or appearance-based backbones to improve their accuracy in RCC-GR task. All models are implemented with PyTorch [19], and we train with CASIA-BN-RCC on TITAN-XP and trained with OUMVLP-RCC on V100 GPUs.

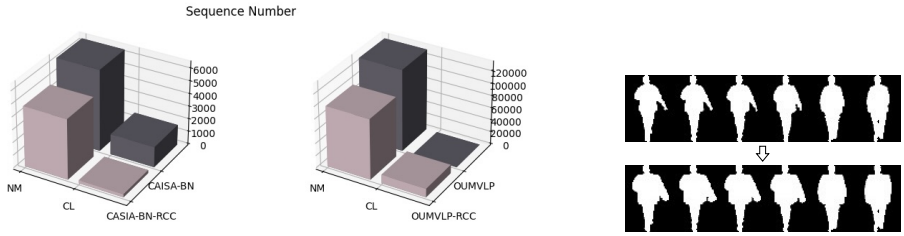
4.1 Datasets Construction

Datasets To support the proposed task, we constructed two benchmarks, CASIA-BN-RCC and OUMVLP-RCC, based on the popular CASIA-B [27] and OU-MVLP [23] datasets. The construction follows three principles: 1) The training set collected in real scenes has a sub-dataset that only has NM conditions with view variations, and a sub-dataset has NM/CL conditions only in front views. 2) The two cross-view and cross-cloth sub-datasets have no overlap. 3) The number of sequences in the cross-view sub-dataset is larger than in the cross-cloth sub-dataset.

CASIA-BN-RCC The original CASIA-B dataset contains 124 identities, with each individual recorded under three walking conditions: normal walking (NM-1 to NM-6), carrying a bag (BG-1, BG-2), and walking with different coats (CL-1, CL-2). Each condition is captured across 11 views ranging from 0° to 180°. Based on previous research [3, 9], the first 74 identities are used as the training dataset, and the remaining 50 identities are used for evaluation. For testing, NM-1 to NM-4 are designated as the gallery, while NM-5, NM-6, BG-1, BG-2, and CL-1, CL-2 form probe subsets based on their conditions. To improve the segmentation quality of CASIA-B, we collected additional pedestrian images and trained a new segmentation model, resulting in the refined CASIA-BN dataset. In order to construct two sub-datasets for CASIA-BN, we first extract the NM, BG* sequences with all views in 1-44 identities to construct $\mathcal{X}_v$, then extract sequences in 45-74 identities, which walking conditions are NM, BG, CL but views are limited in $0^\circ/18^\circ/162^\circ/180^\circ$ to construct $\mathcal{X}_c$. The sequence ratio between $\mathcal{X}_v$ and $\mathcal{X}_c$ is approximately 3:1. The evaluation set remains consistent with previous work [3, 9].

OUMVLP-RCC The OU-MVLP dataset contains 10,307 identities, split into 5,153 for training and 5,154 for testing. Each identity is recorded across 14 views, ranging from

* Since only the middle part of silhouettes of BG is different with NM in appearance, we regard BG as NM and also extract BG sequences to enlarge the dataset



(a) (b)

Fig. 4: (a) The comparison of sequence numbers for each clothing condition across different datasets. (b) Examples of generated cloth-changing images.

Table 1: The optimization method for each backbone used in each dataset. IB means that *only train with the identity-branch*, UB means *load the pre-trained model and train with the uncertainty-branch*. Total means *the total iteration*. LR means *the learning rate at the beginning*.

Backbone	Dataset	Backbone	Branch	MileStone	Total	LR	Weight decay
GaitSet	CASIA-BN-RCC	(16,32,64)	Upper	(1k,2k,3k)	35k	1e-1	5e-4
			IB	(5k,10k,15k)	20k		
			UB	(4k,8k,12k)	15k		
	OUMVLP-RCC	(64,128,256)	Upper	(50k,100k,125k)	150k	1e-1	1e-4
			IB	(50k,100k,125k)	150k		
			UB	(50k,80k)	120k	1e-2	
GaitGL	CASIA-BN-RCC	(16,32,64)	Upper	70k	80k	1e-1	5e-4
			IB	(10k,20k,30k)	35k		
			UB	(3k,5k,8k)	10k		
	OUMVLP-RCC	(32,64,128,256)	Upper	(150k,200k)	210k	1e-1	1e-4
			IB	(150k,200k)	210k		
			UB	(50k,100k)	150k	1e-2	

0° to 90° and 180° to 270°, under a single walking condition: normal walking (NM-00 and NM-01). For testing, NM-01 is used as the gallery, and NM-00 serves as the probe. To construct $\mathcal{X}_v$ in OUMVLP-RCC, we extract all sequences of the first 3200 identities in the training dataset. $\mathcal{X}_c$ includes sequences from the last 1,954 identities, restricted to views at 0°, 15°, 180°, and 195°. Identities missing these views were excluded, resulting in 1,944 identities. To simulate cloth-changing conditions, we artificially modified NM-00 and NM-01 sequences. Specifically, the upper body was dilated using a 5×5 kernel with random shapes (rectangular, cross, or elliptical), and the dilation area was randomly selected within predefined bounds. This method ensures that the data distribution is similar to CASIA-BN. The ratio number between $\mathcal{X}_v$ and $\mathcal{X}_c$ also constraints to 3:1. The ratio of sequences in $\mathcal{X}_v$ to $\mathcal{X}_c$ is also maintained at 3:1. During evaluation, NM-01 serves as the gallery, while NM-00 and its dilated version (CL-00) are used as probes under different conditions across 14 views.

Since our benchmarks extract only a portion of sequences from the original datasets, their volume is smaller. Figure 4a compares the sequence counts for each cloth condition in the original datasets and our benchmarks. Although our benchmarks are smaller, accuracy can improve with additional data collection.

We also show some examples of the generated cloth-changing image in OUMVLP-RCC in Figure 4b. Since we selected the upper part of body and dilation methods randomly, each image has higher diversity after dilation.

Table 2: The rank-1 accuracy (%) on CASIA-BN-RCC for different probe views excluding the identical-view cases. For evaluation, the sequences of NM-01,02,03,04 for each subject are taken as the gallery. The probe sequences are divided into three subsets according to the walking conditions (*i.e.* NM, BG, CL).

Backbone	Condition	Method	Probe View											Average
			0°	18°	36°	54°	72°	90°	108°	126°	134°	162°	180°	
GaitSet	NM-5,6	RCC-Base	84.1	91.3	98.4	94.5	86.3	83.1	86.9	93.0	97.9	92.4	78.5	89.7
		RCC-Ours	84.3	93.1	97.3	94.3	87.6	82.7	87.6	93.4	97.2	93.3	81.7	**90.2**
	BG-1,2	RCC-Base	79.6	86.6	91.7	86.2	77.2	71.2	76.7	81.5	89.4	87.7	74.0	82.0
		RCC-Ours	78.5	86.0	91.8	87.5	78.9	75.5	79.8	83.2	90.6	87.6	74.6	**83.1**
	CL-1,2	RCC-Base	43.5	53.2	54.8	47.4	45.1	36.3	36.1	43.3	50.6	49.3	37.5	45.2
		RCC-Ours	**50.7**	**60.9**	**61.9**	**55.0**	**47.9**	**40.3**	**45.5**	**53.1**	**55.2**	**56.4**	**38.4**	**51.4**
GaitGL	NM-5,6	RCC-Base	88.0	94.1	96.9	94.3	86.8	82.8	89.2	94.5	97.4	93.3	79.4	90.6
		RCC-Ours	87.4	93.5	96.6	94.4	87.7	82.6	89.6	94.8	97.3	92.0	80.3	**90.6**
	BG-1,2	RCC-Base	81.0	89.1	94.2	88.7	79.6	73.1	78.6	86.8	93.4	89.4	68.5	83.9
		RCC-Ours	81.7	89.3	93.7	89.3	80.9	76.4	79.2	88.5	92.9	90.8	70.5	**84.8**
	CL-1,2	RCC-Base	57.5	66.3	64.6	59.5	59.5	52.7	53.3	57.2	63.6	59.8	39.7	57.6
		RCC-Ours	**58.2**	**71.3**	**73.5**	**64.6**	**60.9**	**53.7**	**54.5**	**60.4**	**68.1**	**65.6**	38.0	**60.8**

Table 3: The rank-1 accuracy (%) on OUMVLP-RCC for different probe views excluding the identical-view cases. For evaluation, the sequences of NM-01 for each subject are taken as the gallery and the probe sequences of the two walking conditions, NM-00 and CL-00, are respectively evaluated.

Backbone	Condition	Method	Probe View													Average
			0°	15°	30°	45°	60°	75°	90°	180°	195°	210°	225°	240°	255°	270°	
GaitSet	NM-00	RCC-Base	76.4	85.4	88.7	88.8	84.2	86.3	84.4	78.9	84.1	87.8	87.9	83.8	85.2	82.8	84.6
		RCC-Ours	76.8	85.9	89.1	89.3	84.8	86.7	84.8	79.1	84.4	88.0	88.2	84.3	85.8	83.2	**85.0**
	CL-00	RCC-Base	66.9	78.1	79.6	65.0	30.1	31.1	22.3	70.2	77.4	80.4	70.2	36.1	32.5	21.8	54.4
		RCC-Ours	**67.2**	**78.3**	**80.9**	**68.6**	**35.5**	**35.7**	**26.4**	**70.3**	**77.7**	**81.5**	**73.2**	**41.8**	**38.3**	**27.0**	**57.3**
GaitGL	NM-00	RCC-Base	86.5	90.4	91.1	91.3	90.6	90.5	90.0	89.5	89.0	90.2	90.2	89.3	89.2	88.6	89.7
		RCC-Ours	86.6	90.4	91.9	91.4	90.8	90.7	90.2	89.7	88.8	90.2	90.4	89.5	89.4	88.8	**89.8**
	CL-00	RCC-Base	83.1	88.4	88.9	85.5	67.7	67.5	58.6	86.9	86.9	88.6	86.1	69.5	67.0	55.8	77.2
		RCC-Ours	**83.3**	**88.2**	**89.1**	**86.6**	**71.6**	**72.8**	**65.8**	**87.0**	86.4	**88.7**	**86.8**	**73.0**	**69.1**	**59.5**	**79.1**

4.2 Performance Comparison

This section evaluates the performance of our framework on the CASIA-BN-RCC and OUMVLP-RCC benchmarks.

Network Structure The Network structure we employed for GaitSet [3] and GaitGL [17] backbone is shown in Table 1.

Baselines To assess the effectiveness of our framework, we compare it against the following baselines:

(1) *RCC-Base*. Results using the baseline model on CASIA-BN-RCC and OUMVLP-RCC with the original framework.

(2) *RCC-Ours*. Results using our Progressive Feature Learning framework optimized on the same benchmarks.

CASIA-BN-RCC The performance comparison on CASIA-BN-RCC is shown in Table 2. The probe has been divided into three subsets according to the walking conditions and has been evaluated separately. It should be noted that since our framework

Table 4: (a) The performance of Progressive Mapping. The results are reported in the rank-1 accuracy. T-S: *two-stage module*. P-T: *Progressive-aware Triplets*. Without P-T, triples are selected from both $\mathcal{X}_v$ and $\mathcal{X}_c$. (b) Progressive Uncertainty Performance in our framework. The results are reported in the rank-1 accuracy. *e* represents *re-sampling an additional embedding from Gaussian distributions*. Underlined numbers indicate *the comprehensive best result.* ‡ means the residual connection is abandoned.

(a)

Settings				CASIA-BN-RCC			OUMVLP-RCC	
T-S	$L_{nv,i}$	$L_{nc,i}$	P-T	NM ↑	BG ↑	CL ↑	NM ↑	CL ↑
		✓		89.7	82.0	45.2	84.6	54.4
✓		✓		88.8	81.6	45.9	84.0	54.4
✓	✓	✓		89.0	82.0	46.0	84.1	54.8
✓	✓	✓	✓	**89.7**	**82.6**	**48.5**	**84.7**	**55.6**

(b)

Dataset	CASIA-BN-RCC			OUMVLP-RCC	
Settings	NM ↑	BG ↑	CL ↑	NM ↑	CL ↑
$\sigma_{v,i}$ + $\sigma_{c,i}$	90.1	82.8	**51.6**	85.0	57.3
$\sigma_{v,i}$ + $\sigma_{c,i}$ + *e*	<u>90.2</u>	<u>83.1</u>	<u>51.4</u>	<u>85.0</u>	**57.3**
$\sigma_{c,i}$ (CCM) + *e*	**90.3**	**83.2**	50.7	85.2	56.6
$\sigma_{c,i}$ (CVM + CCM)‡ + *e*	90.2	83.0	50.3	**85.2**	56.6

aims at improving the precision in realistic cloth-changing condition, we take **the accuracy for CL as the main criteria**, since it is the hardest target in our task. That is, it is necessary to improve the accuracy in CL condition when we could only extract cross-view and cross-cloth features from two non-overlapping datasets. The characteristics of datasets will weaken the ability of baseline model to combine information. However, our framework can bring large improvement for the baseline in both backbones for the CL conditions (+6.2% for GaitSet, +3.2% for GaitGL), indicating we can handle the more challenging cloth-changing problem. The accuracy of NM and BG also show improvements compared with baseline, and the results with our framework in these two conditions are also reasonable.

OUMVLP-RCC The performance comparison on OUMVLP-RCC is shown in Table 3. Notice that we also divide the probe into two subsets according to the walking conditions: NM and CL, and they have been evaluated separately, and we **take the accuracy for CL as the main criteria**. Similar to CASIA-BN-RCC, the baseline model struggles to extract cross-view and cross-cloth features in the CL condition. Our framework improves CL accuracy by +2.9% for GaitSet and +1.9% for GaitGL, confirming its robustness in handling the most difficult task. For the NM subset, the baseline model's accuracy is already near the upper bound, leaving limited room for improvement. Nonetheless, our framework maintains competitive performance in NM while excelling in CL, demonstrating its overall effectiveness.

4.3 Ablation Study

To evaluate the effectiveness of our framework, we conducted several ablation studies with various settings on CASIA-BN-RCC and OUMVLP-RCC systematically using GaitSet as the backbone for simplicity

Impact of Progressive Mapping We compare the performance of our Progressive Mapping with another three settings and show the results in Tabel 4a, to show the advance of our design. In the first row, we show the performance of the one-stage module, which is also the baseline of our task. All the other experiments employ the identity-branch and are equipped with the two-stage module, the difference lies in the triplet

loss. In the other three settings, we choose to use the normal triplet loss $\mathcal{L}_{n,j},\ j \in v, c$ behind CVM/CCM, and normal triplets are selected from $\mathcal{X}_v$ and $\mathcal{X}_c$.

$$\begin{aligned}\mathcal{L}_{nv,i} =& \frac{1}{|\mathcal{T}_v|} \sum_{(a,p,n)\in\mathcal{T}_v} \left[\left\| \mu^a_{v,i} - \mu^p_{v,i} \right\|^2 - \left\| \mu^a_{v,i} - \mu^n_{v,i} \right\|^2 + m \right]_+ \\ \mathcal{L}_{nc,i} =& \frac{1}{|\mathcal{T}_c|} \sum_{(a,p,n)\in\mathcal{T}_c} \left[\left\| \mu^a_{c,i} - \mu^p_{c,i} \right\|^2 - \left\| \mu^a_{c,i} - \mu^n_{c,i} \right\|^2 + m \right]_+ \end{aligned} \tag{9}$$

In the experiment of the second line, we only add $\mathcal{L}_{nc,i}$ behind CCM, and the result shows that only constraining the two-stage module at the end of the final stage is not enough. Compared with our methods, the results in the third row are also slightly poorer, for the negative repel effect of $\mathcal{L}_{nv,i}$ enlarges the intra-class distance of samples in $\mathcal{X}_c$, which is larger than the positive aggregate effect of $\mathcal{L}_{nc,i}$ since the diversity is richer in $\mathcal{X}_v$, so the adversarial effect will add difficulty to the optimization of CCM. With the four experiments, we can demonstrate that our improvement does not lie in the increased module parameter but in the two-stage module and the design of the triplets, making the CL condition improvement (GaitSet: +3.3%, GaitGL: + 1.2%). It's important to recognize that our approach is not solely focused on achieving a small increase in accuracy; rather, it can be scalable with additional data and computing resources, allowing for the potential to achieve even greater levels of accuracy.

Impact of Progressive Uncertainty We evaluate the effect of Progressive Uncertainty in Table 4b. The first row shows results using the uncertainty branch with the Uncertainty-aware Triplet Loss $\mathcal{L}_{tv,i}$ and $\mathcal{L}_{tc,i}$. In the second row, embeddings are resampled from Gaussian distributions $\mathcal{N}(\mu_{j,i}, \sigma^2_{j,i}\mathrm{I})$. These results demonstrate that variance optimization alone improves performance in the CL condition, with most of the improvement attributed to the Uncertainty-aware Triplet Loss. In the third and fourth rows, we calculate $\sigma_{c,i}$ and $\mathcal{L}_{tc,i}$ but map sequences as points after CVM. The difference is that in the third row, only CCM generates $\sigma_{c,i}$, while in the fourth row, both CVM and CCM generate $\sigma_{c,i}$ without residual connections. The results indicate that mapping sequences as a single Gaussian distribution is insufficient for optimal performance. Our results demonstrate that only mapping each sequence as one Gaussian distribution cannot have the best performance.

5 Conclusion

In this work, we tackle the cloth-changing gait recognition problem in a realistic application with automatic data collection. We propose a new framework that can be applied with off-the-shelf gait recognition backbones to boost performance in the RCC-GR task. The key methods of our Progressive Feature Learning framework are that we design Progressive Mapping and Progressive Uncertantity to span the cross-view and cross-cloth features in a cascaded way. Extensive experiments were conducted to validate the effectiveness of our framework. It shows that our framework could improve the performance of state-of-the-art methods in walking with different coats (CL) conditions in RCC-GR.

References

1. Bouchrika, I., Goffredo, M., Carter, J., Nixon, M.: On using gait in forensic biometrics. Journal of forensic sciences **56**(4), 882–889 (2011)
2. Chang, J., Lan, Z., Cheng, C., Wei, Y.: Data uncertainty learning in face recognition. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 5710–5719 (2020)
3. Chao, H., He, Y., Zhang, J., Feng, J.: Gaitset: Regarding gait as a set for cross-view gait recognition. In: Proceedings of the AAAI conference on artificial intelligence. vol. 33, pp. 8126–8133 (2019)
4. Chen, K., Lv, Q., Yi, T., Yi, Z.: Reliable probabilistic face embeddings in the wild. arXiv preprint arXiv:2102.04075 (2021)
5. Chen, Y.C., Zhu, X., Zheng, W.S., Lai, J.H.: Person re-identification by camera correlation aware feature augmentation. IEEE transactions on pattern analysis and machine intelligence **40**(2), 392–408 (2017)
6. Cui, Y., Deng, W., Xu, X., Liu, Z., Liu, Z., Pietikäinen, M., Liu, L.: Uncertainty-guided semi-supervised few-shot class-incremental learning with knowledge distillation. IEEE Transactions on Multimedia pp. 1–14 (2022). https://doi.org/10.1109/TMM.2022.3208743
7. Deng, J., Guo, J., Ververas, E., Kotsia, I., Zafeiriou, S.: Retinaface: Single-shot multi-level face localisation in the wild. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 5203–5212 (2020)
8. Deng, J., Guo, J., Xue, N., Zafeiriou, S.: Arcface: Additive angular margin loss for deep face recognition. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 4690–4699 (2019)
9. Fan, C., Peng, Y., Cao, C., Liu, X., Hou, S., Chi, J., Huang, Y., Li, Q., He, Z.: Gaitpart: Temporal part-based model for gait recognition. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 14225–14233 (2020)
10. Guan, D., Huang, J., Xiao, A., Lu, S., Cao, Y.: Uncertainty-aware unsupervised domain adaptation in object detection. IEEE Transactions on Multimedia **24**, 2502–2514 (2022). https://doi.org/10.1109/TMM.2021.3082687
11. Hong, J., Zhang, W., Feng, Z., Zhang, W.: Dual cross-attention for video object segmentation via uncertainty refinement. IEEE Transactions on Multimedia pp. 1–16 (2022). https://doi.org/10.1109/TMM.2022.3225720
12. Kingma, D.P., Welling, M.: Auto-encoding variational bayes. arXiv preprint arXiv:1312.6114 (2013)
13. Larsen, P.K., Simonsen, E.B., Lynnerup, N.: Gait analysis in forensic medicine. Journal of forensic sciences **53**(5), 1149–1153 (2008)
14. Li, X., Makihara, Y., Xu, C., Yagi, Y., Yu, S., Ren, M.: End-to-end model-based gait recognition. In: Proceedings of the Asian conference on computer vision (2020)
15. Liao, R., Cao, C., Garcia, E.B., Yu, S., Huang, Y.: Pose-based temporal-spatial network (ptsn) for gait recognition with carrying and clothing variations. In: Chinese Conference on biometric recognition. pp. 474–483. Springer (2017)
16. Liao, R., Yu, S., An, W., Huang, Y.: A model-based gait recognition method with body pose and human prior knowledge. Pattern Recognition **98**, 107069 (2020)
17. Lin, B., Zhang, S., Yu, X.: Gait recognition via effective global-local feature representation and local temporal aggregation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 14648–14656 (2021)
18. Liu, J., Sun, Y., Han, C., Dou, Z., Li, W.: Deep representation learning on long-tailed data: A learnable embedding augmentation perspective. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 2970–2979 (2020)

19. Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al.: Pytorch: An imperative style, high-performance deep learning library. Advances in neural information processing systems **32** (2019)
20. Shi, Y., Jain, A.K.: Probabilistic face embeddings. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 6902–6911 (2019)
21. Shi, Y., Tian, W., Ling, H., Li, Z., Li, P.: Spatial-wise and channel-wise feature uncertainty for occluded person re-identification. Neurocomputing **486**, 237–249 (2022)
22. Song, X., Jin, Z.: Robust label rectifying with consistent contrastive-learning for domain adaptive person re-identification. IEEE Transactions on Multimedia **24**, 3229–3239 (2022). https://doi.org/10.1109/TMM.2021.3096014
23. Takemura, N., Makihara, Y., Muramatsu, D., Echigo, T., Yagi, Y.: Multi-view large population gait dataset and its performance evaluation for cross-view gait recognition. IPSJ Transactions on Computer Vision and Applications **10**(1), 1–14 (2018)
24. Wang, P., Ding, C., Tan, W., Gong, M., Jia, K., Tao, D.: Uncertainty-aware clustering for unsupervised domain adaptive object re-identification. IEEE Transactions on Multimedia pp. 1–1 (2022). https://doi.org/10.1109/TMM.2022.3149629
25. Xu, P., Zhu, X.: Deepchange: A large long-term person re-identification benchmark with clothes change. arXiv preprint arXiv:2105.14685 (2021)
26. Yang, X., Gao, Y., Luo, H., Liao, C., Cheng, K.T.: Bayesian denet: Monocular depth prediction and frame-wise fusion with synchronized uncertainty. IEEE Transactions on Multimedia **21**(11), 2701–2713 (2019). https://doi.org/10.1109/TMM.2019.2912121
27. Yu, S., Tan, D., Tan, T.: 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). vol. 4, pp. 441–444. IEEE (2006)
28. Yu, T., Li, D., Yang, Y., Hospedales, T.M., Xiang, T.: Robust person re-identification by modelling feature uncertainty. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 552–561 (2019)
29. Zheng, L., Yang, Y., Hauptmann, A.G.: Person re-identification: Past, present and future. arXiv preprint arXiv:1610.02984 (2016)
30. Zhu, Z., Guo, X., Yang, T., Huang, J., Deng, J., Huang, G., Du, D., Lu, J., Zhou, J.: Gait recognition in the wild: A benchmark. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 14789–14799 (2021)
31. Zhuo, J., Wang, S., Huang, Q.: Uncertainty modeling for robust domain adaptation under noisy environments. IEEE Transactions on Multimedia pp. 1–14 (2022). https://doi.org/10.1109/TMM.2022.3205457