

# Causality-inspired Discriminative Feature Learning in Triple Domains for Gait Recognition

Haijun Xiong<sup>✉</sup>, Bin Feng<sup>✉</sup>, Xinggang Wang<sup>✉</sup>, and Wenyu Liu<sup>✉</sup>

Hubei Key Lab of Smart Internet, School of EIC, Huazhong University of Science and Technology, Wuhan, China

{xionghj, fengbin, xgwang, liuwu}@hust.edu.cn

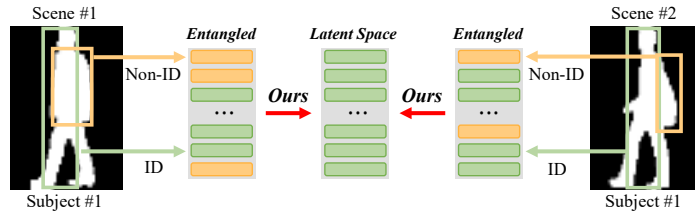
**Abstract.** Gait recognition is a biometric technology that distinguishes individuals by their walking patterns. However, previous methods face challenges when accurately extracting identity features because they often become entangled with non-identity clues. To address this challenge, we propose CLTD, a causality-inspired discriminative feature learning module designed to effectively eliminate the influence of confounders in triple domains, *i.e.*, spatial, temporal, and spectral. Specifically, we utilize the Cross Pixel-wise Attention Generator (CPAG) to generate attention distributions for factual and counterfactual features in spatial and temporal domains. Then, we introduce the Fourier Projection Head (FPH) to project spatial features into the spectral space, which preserves essential information while reducing computational costs. Additionally, we employ an optimization method with contrastive learning to enforce semantic consistency constraints across sequences from the same subject. Our approach has demonstrated significant performance improvements on challenging datasets, proving its effectiveness. Moreover, it can be seamlessly integrated into existing gait recognition methods.

**Keywords:** Gait recognition · Causality · Triple Domains · Discriminative representations

## 1 Introduction

Gait recognition is a long-distance biometric identification technology that uses unique walking patterns for individual recognition. It has garnered significant attention due to its wide applications in surveillance and healthcare [40, 41]. However, accurately identifying individuals presents a challenge due to various factors influencing performance.

There are two categories of gait recognition methods: appearance-based [3, 9, 21, 31] and model-based [11, 12, 20, 46]. Appearance-based approaches have gained more attention recently because of their superior performance compared to model-based approaches. However, the entanglement of discriminative identity (ID) features with non-identity (non-ID) clues makes it difficult to effectively and efficiently extract features. Here, non-ID clues such as noise, clothing, and carrying conditions, act as extrinsic confounders to identities. As depicted



**Fig. 1:** Motivation. Illustration of entanglement between non-ID and ID clues. With our approach, the impact of non-ID clues is systematically eliminated.

in Fig. 1, most of the current approaches focus on extracting ID-intrinsic features directly from the entangled space, which is challenging. This entanglement biases gait recognition models towards non-ID clues, ultimately compromising performance [27]. The core idea of this work is to establish a consistent latent space where non-ID clues can be effectively eliminated.

The Structural Causal Model (SCM) provides a fundamental framework for describing causal mechanisms within a system [55], with applications spanning diverse fields such as computer vision [52], natural language processing [1], and recommender systems [5]. Recently, causal analysis has been explored in gait recognition, with notable contributions from the Generative Counterfactual Intervention framework (GaitGCI) [8]. GaitGCI focuses on mitigating the influence of confounders during the feature learning process, but it has several limitations. Firstly, it only addresses confounders at the output stage of the backbone network, overlooking the intricate formation and effects of confounders across different stages. Secondly, its causal intervention is confined to spatial features, neglecting potential confounders in other domains such as temporal and spectral. Lastly, the absence of explicit semantic consistency constraints may compromise the consistency of learned identity representations for the same subject, hindering overall performance.

In contrast, we propose to leverage the relationship between spatial and spectral domain signals. We believe that if confounders exist in spatial signals, they are also likely present in the spectral domain. We can effectively describe and separate signals by exploiting spectral representation, thereby eliminating confounders. This notion has been validated in various computer vision tasks [25, 26, 33, 59], underscoring its potential efficacy in gait recognition.

Based on these analyses, we propose a novel module called CLTD, which is designed to mitigate the influence of confounders within spatial, temporal, and spectral domains. First, we perform the Cross Pixel-wise Attention Generator (CPAG) to capture contextual information for each pixel in the time, height, and width dimensions, thereby generating high-quality attention distributions. This ensures the effective processing of spatial and temporal information. Next, we present the Fourier Projection Head (FPH), which utilizes the Fast Fourier Transform (FFT) to project the spatial feature map into the Fourier spectral domain. This transformation preserves essential information while significantly

reducing computational costs, enabling efficient processing in the spectral domain. Furthermore, we introduce contrastive learning to supervise the generated factual and counterfactual features, thus ensuring consistent semantic representations of ID-intrinsic clues across sequences from the same subject. Moreover, we propose to utilize CLTD at multiple stages of the network, not just in the final stage, to comprehensively eliminate the effects of confounders. Importantly, CLTD serves as a plug-and-play training paradigm, compatible with various gait recognition methods. We validate the effectiveness and versatility of CLTD through experiments conducted on multiple datasets, including OU-MVLP [43], CASIA-B [53], GREW [60], and Gait3D [58]. The results demonstrate significant performance improvements when CLTD is integrated with various gait recognition methods, leading to state-of-the-art results on multiple datasets, with a maximum improvement of **11.1%** (from 60.2% to 71.3%, refer to Tab. 4).

The main contributions of our work can be summarized as follows:

- We solve the gait recognition problem from a new perspective, *i.e.*, causality-based contrastive learning. This way aims to establish a consistent latent space, reducing the bias introduced by non-ID clues.
- We propose CLTD, which leverages the Cross Pixel-wise Attention Generator and Fourier Projection Head to eliminate the impact of confounders in triple domains effectively. In addition, we incorporate contrastive learning to ensure the consistency of ID-intrinsic semantics information across sequences from the same subject.
- Demonstrating the effectiveness and versatility of CLTD through extensive experiments, achieving state-of-the-art results on various datasets and significantly improving performance on multiple baseline methods.

## 2 Related Work

**Gait Recognition.** Gait recognition methods can be generally categorized into two groups: model-based and appearance-based. Model-based approaches [11, 12, 20, 45, 46, 54] use a priori model representing human shape and dynamics, extracting gait features by fitting these models. For instance, methods in [45, 46] utilize ResGCN for spatial-temporal feature extraction from skeleton sequences. In contrast, GPGait [11] focuses on efficient fine-grained feature extraction. On the other hand, appearance-based methods [2, 3, 8, 10, 15, 18, 19, 21, 29, 31, 42, 49, 58] directly extract gait features from silhouette sequences. Han *et al.* [15] introduced Gait Energy Images (GEIs) obtained by averaging multiple silhouettes from a gait cycle sequence. GaitSet [3] is the first method to treat gait as a set for learning discriminative feature representations. Huang *et al.* [19, 21] presented fine-grained feature representations and multi-scale temporal motion modeling methods. SMPLGait [58] introduces a multimodal method, which uses 3D human meshes [30] to enhance silhouette information. Chai *et al.* [2] implemented a second-order motion extraction module and view embedding based on 3D convolution to learn motion information. Wang *et al.* [50] presented DyGait focusing on extracting feature representations of dynamic body parts. Additionally, some

other methods [29, 49] explored the influence of part-based multi-scale motion information.

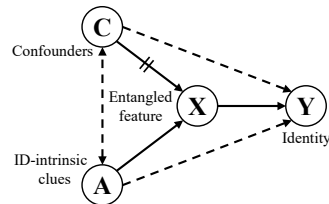
**Causality in Computer Vision.** Causality has emerged as a crucial concept in various computer vision tasks, aiding in the representation and reasoning of uncertain knowledge. Several recent works have leveraged causal models to address different challenges. Yang *et al.* [51] proposed a causal attention model to eliminate confounding effects in vision-language attention models, enhancing the interpretability and robustness of the model. Chen *et al.* [4] presented a meta-causal learning method that learns to infer the causes of domain shift between the auxiliary and source domains during training, facilitating domain adaptation and transfer learning tasks. Miao *et al.* [34] developed CauSSL a framework for semi-supervised medical image segmentation, which enhances algorithmic independence between two branches using causal mechanisms. Dou *et al.* [8] proposed GaitGCI, first introducing causality into gait recognition. GaitGCI focuses on intervening causally on the final spatial feature to mitigate confounding effects. In contrast to GaitGCI, our proposed approach eliminates the impact of confounders in triple domains, namely spatial, temporal, and spectral domains. This allows for a more comprehensive elimination of confounders and improves the discriminative ability of the gait recognition model.

**Contrastive Learning.** Contrastive learning is a powerful technique, aiming to decrease the distance between similar samples while increasing the gap between dissimilar ones. Guo *et al.* [13] proposed AimCLR, emphasizing extreme augmentations to improve the universality of learned representations. Rao and Miao [39] presented TransG to capture skeleton relationships and spatial-temporal semantic information. Quan *et al.* [38] introduced SemCL, specifically targeting object extraction, significantly improving the spatial comprehension of pre-trained models. Inspired by these, CLTD ensures semantic consistency between sequences of the same subject by incorporating contrastive learning.

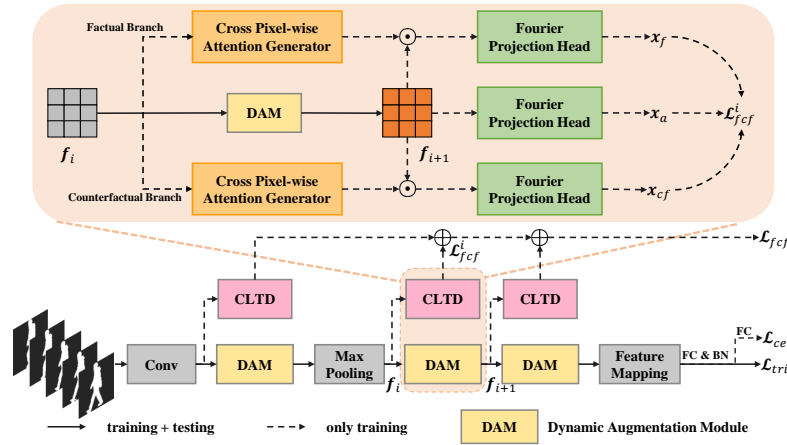
## 3 Method

### 3.1 Gait Recognition from Causal View

Causal inference [37] refers to the process of determining the cause-and-effect relationship between different factors or events. Enlightened by previous excellent works [4, 24, 28, 52], we introduce causality into gait recognition to separate various confounders and ID-intrinsic clues. The causal graph in Fig. 2 is constructed using SCM [55] to understand causality in gait recognition. In this representation,  $\mathbf{X}$  denotes features generated by the gait model,  $\mathbf{A}$  denotes ID-intrinsic clues,  $\mathbf{C}$  stands for confounders (non-ID clues) entangled with  $\mathbf{A}$ , and  $\mathbf{Y}$  represents the ground truth, *i.e.*, the identity. The solid line indicates a direct causal relationship, *e.g.*,  $\mathbf{X} \rightarrow \mathbf{Y}$  denotes  $\mathbf{Y}$  is caused by  $\mathbf{X}$ , while the dotted line indicates there exists statistical depen-



**Fig. 2:** Causal Graph illustration of our approach.



**Fig. 3: Overview of our approach.** We illustrate CLTD using DyGait [50] as a backbone. In Sec. 4, we will show its versatility to various gait recognition models. Multiple CLTDs are used along the backbone of DyGait. Each CLTD consists of two branches: the Factual Branch and the Counterfactual Branch. These branches aim to generate the factual feature  $x_f$  and the counterfactual feature  $x_{cf}$ , respectively. Notably, CLTDs are only used for training and excluded during testing.

dence between two factors. In the ideal scenario,  $\mathbf{X}$  should be learned solely from ID-intrinsic clues, denoted as  $\mathbf{A} \rightarrow \mathbf{X} \rightarrow \mathbf{Y}$ . However, in practical cases, the existence of confounders often entangles  $\mathbf{A}$  with non-ID clues  $\mathbf{C}$ , represented as  $(\mathbf{C}, \mathbf{A}) \rightarrow \mathbf{X} \rightarrow \mathbf{Y}$ , where  $(\mathbf{C}, \mathbf{A})$  denotes that the two factors are entangled. It is difficult to distill  $\mathbf{X}$  from confounders  $\mathbf{C}$  due to the severe entanglement of various influencing factors with  $\mathbf{A}$ . Performing intervention on confounders  $\mathbf{C}$ , such as *do*-operation denoted as  $do(\mathbf{C})$  for specifying the exact value of  $\mathbf{C}$  and isolating its cause, allows us to eliminate the corresponding effects. In Fig. 2,  $\parallel$  denotes the *do*-operation.

### 3.2 Architecture

Based on the analysis of causality, we introduce CLTD to enhance the learning of discriminative features by eliminating the influence of confounders, while preserving ID-intrinsic information. The overall architecture of our approach is illustrated in Fig. 3. For clarity, we adopt DyGait [50] as the backbone to construct our approach, and the versatility of CLTD will be demonstrated in Sec. 4.4. DyGait primarily consists of several fundamental DAM blocks. Given the wide adoption of multi-scale feature representation [17, 21, 49, 56] in various computer vision tasks, it is reasonable to infer that confounding information is distributed across multiple feature scales. Therefore, we propose incorporating CLTD at multiple stages of the network, not solely in the final stage, to effectively and comprehensively eliminate the effect of confounders. Specifically, we integrate a

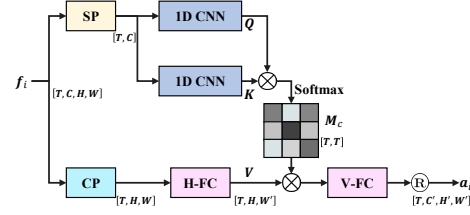
CLTD for each DAM to learn more discriminative feature representations across multiple feature grains. Each CLTD is responsible for eliminating the impact of confounders in triple domains by generating factual and counterfactual features simultaneously, supervised with a Factual and Counterfactual Loss. This ensures the gait recognition model remains free from confounders throughout the backbone. The experimental results in Sec. 4.4 will demonstrate the effectiveness of this approach. It is crucial to clarify that CLTD is only employed during training, which incurs no additional computational cost or memory consumption during inference.

### 3.3 CLTD

The proposed CLTD is a plug-and-play module designed to eliminate the impact of confounders in the spatial, temporal, and spectral domains. It consists of two primary components: the Cross Pixel-wise Attention Generator and the Fourier Projection Head.

**Cross Pixel-wise Attention Generator.** The Cross Pixel-wise Attention Generator, inspired by CCNet [23], is introduced to establish high-quality factual and counterfactual attention. This is achieved by employing a separation operation in time, height, and width dimensions. This operation calculates correlations only between pixels that cross a specific pixel, effectively reducing computational complexity in time and space.

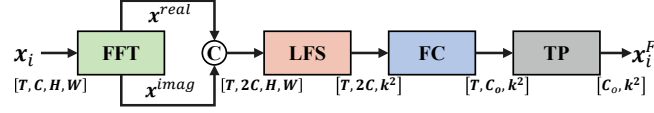
The detailed structure of CPAG depicted in Fig. 4, involves several key steps. Given the input feature map  $\mathbf{f}_i \in \mathbb{R}^{T \times C \times H \times W}$ , CPAG starts with a Spatial Pooling operation and two 1D CNNs applied to  $\mathbf{f}_i$ , generating distinct feature maps  $\mathbf{Q}$  and  $\mathbf{K}$ , where  $\mathbf{Q}, \mathbf{K} \in \mathbb{R}^{T \times C}$ . Subsequently, the temporal correlation matrix  $\mathbf{M}_c \in \mathbb{R}^{T \times T}$  is generated through a matrix multiplication operation and a softmax layer. The process is formulated as:



**Fig. 4:** The detailed structure of CPAG. The process is formulated as:

$$\begin{aligned}
 \mathbf{Q} &= \text{Conv1D}(\text{SP}(\mathbf{f}_i)), \\
 \mathbf{K} &= \text{Conv1D}(\text{SP}(\mathbf{f}_i)), \\
 \mathbf{M}_c &= \text{Softmax}\left(\frac{\mathbf{Q}\mathbf{K}^\top}{\gamma}\right),
 \end{aligned} \tag{1}$$

where  $\text{SP}(\cdot)$  denotes the Spatial Pooling operation, and  $\gamma$  is a learnable scaling parameter to balance the learning magnitude of the key-query dot product. Simultaneously, CPAG performs an average pooling operation on  $\mathbf{f}_i$  along the channel dimension. It is then horizontally mapped into  $\mathbf{V} \in \mathbb{R}^{T \times H \times W'}$  via the Horizontal separate FC layer (H-FC)  $\mathbf{W}_H \in \mathbb{R}^{H \times W \times W'}$ , where  $W'$  denotes the width of the  $i$ -th DAM's output  $\mathbf{f}_{i+1}$ . After obtaining the temporal correlation matrix  $\mathbf{M}_c$  and feature map  $\mathbf{V}$ , CPAG utilizes matrix multiplication, Vertical



**Fig. 5:** Details of FPH. The symbol  $\odot$  stands for the concatenating operation.

separate FC layer (V-FC), and channel repeat operations to produce the output  $\mathbf{a}_i$ , denoted as:

$$\mathbf{a}_i = \text{Repeat} \left( (\mathbf{M}_c \mathbf{V})^\top \mathbf{W}_V \right), \quad (2)$$

where  $\mathbf{a}_i \in \mathbb{R}^{T \times C' \times H' \times W'}$ , and  $C'$  and  $H'$  refer to the channel and height of  $\mathbf{f}_{i+1}$ , respectively.  $\mathbf{W}_V \in \mathbb{R}^{W' \times H \times H'}$  represents the V-FC operation like H-FC, and  $\text{Repeat}(\cdot)$  represents the channel repeat operation.

In naive self-attention [6, 47], the computational complexity of the matrix multiplication is quadratic to the spatial-temporal resolution of the input, i.e.,  $\mathcal{O}(2t^2h^2w^2c)$  for a feature map with  $t \times c \times h \times w$  pixels. In contrast, our approach, which involves the separation operations of time, height, and width, significantly reduces the computational complexity to  $\mathcal{O}(t^2(c + hw))$ . Therefore, CPAG dramatically lowers the computational cost compared to Self-Attention. Furthermore, horizontal and vertical decoupling enables CPAG to aggregate pixels at different locations along the horizontal and vertical directions, facilitating enhanced spatial interaction at the pixel level. This, in turn, generates high-quality distributions of factual and counterfactual attention.

**Fourier Projection Head.** The distribution of information in the spatial domain is often scattered, posing challenges for effective extraction. In contrast, the Fourier spectrum typically concentrates most of the energy in the low-frequency region, making it more amenable to information extraction. Motivated by this observation, we introduce the Fourier Projection Head to transform the spatial feature map into the Fourier spectral domain through FFT. The subsequent step involves a Low-frequency Selector (LFS) to reduce dimensionality while preserving essential information.

As depicted in Fig. 5, FPH takes  $\mathbf{x}_i = \mathbf{a}_i \odot \mathbf{f}_{i+1} \in \mathbb{R}^{T \times C \times H \times W}$  as input and first employs FFT to convert  $\mathbf{x}_i$  into the Fourier spectrum. Subsequently, the LFS is applied to the Fourier spectrum with a window of  $k \times k$  to reduce dimensionality. The information is then projected into a vector  $\mathbf{x}_i^F$  through an FC layer  $\mathbf{W}_P \in \mathbb{R}^{2C \times 2C_o}$  and a Temporal Pooling (TP) operation, where  $C_o$  is the output dimension. The overall FPH process can be formulated as:

$$\begin{aligned} \mathbf{x}^{real}, \mathbf{x}^{imag} &= \text{FFT}(\mathbf{x}_i), \\ \mathbf{x}^k &= \text{LFS}_{k \times k}(\mathbf{x}^{real} \odot \mathbf{x}^{imag}), \\ \mathbf{x}_i^F &= \text{TP}(\mathbf{x}^k \mathbf{W}_P), \end{aligned} \quad (3)$$

where  $\mathbf{x}_i^F \in \mathbb{R}^{C_o \times k^2}$ ,  $\odot$  denotes concatenating the real and imaginary components, and  $\text{LFS}_{k \times k}$  represents the low-frequency selector with a window of  $k \times k$ ,

defined as:

$$\text{LFS}_{k \times k}(x) = \text{Sigmoid}(x[\dots, \frac{h-k}{2} : \frac{h-k}{2} + k, \frac{w-k}{2} : \frac{w-k}{2} + k]). \quad (4)$$

### 3.4 Loss Function

We introduce the Factual and Counterfactual Loss as a supervision mechanism for each CLTD in our approach. For the  $i$ -th CLTD, we obtain the factual feature  $\mathbf{x}_f$ , and counterfactual feature  $\mathbf{x}_{cf}$ . They are used along with the anchor feature  $\mathbf{x}_a$ , the output of FPH taking  $\mathbf{f}_{i+1}$  as input, to formulate contrastive learning. Here we have omitted the symbol  $i$  for simplicity. Specifically, we employ InfoNCE [14, 35] to minimize the gap between entangled and ID-intrinsic features while increasing the gap between entangled and confounder features, ensuring consistent semantic representations of ID-intrinsic clues across sequences from the same subject. This is formulated as:

$$\mathcal{L}_{NCE}^{fcf} = - \sum_{\mathbf{x}^+ \in \mathcal{S}_f} \log \frac{\text{sim}(\mathbf{x}_a, \mathbf{x}^+)}{\text{sim}(\mathbf{x}_a, \mathbf{x}^+) + \sum_{\mathbf{x} \in \mathcal{S}_{cf}} \text{sim}(\mathbf{x}_a, \mathbf{x})}, \quad (5)$$

where  $\text{sim}(\mathbf{a}, \mathbf{b}) = \frac{\mathbf{a} \cdot \mathbf{b}}{\|\mathbf{a}\| \|\mathbf{b}\|}$  represents the similarity between two feature vectors  $\mathbf{a}$  and  $\mathbf{b}$ .  $\mathcal{S}_f$  and  $\mathcal{S}_{cf}$  denote the sets of factual features and counterfactual features, respectively, from the subject that  $\mathbf{x}_a$  belongs to. Additionally, inspired by the work of Tang *et al.* [44] in alleviating context bias through the total direct effect (TDE) in causal inference for unbiased scene graph generation, we eliminate the impact of confounders by maximizing TDE and the factual probability. This is defined as:

$$\begin{aligned} \mathbf{Y}_f &= \mathbf{P}(\mathbf{Y} | \mathbf{X} = \mathbf{x}_f) = \mathbf{W}_c \mathbf{x}_f, \\ \mathbf{Y}_{cf} &= \mathbf{P}(\mathbf{Y} | do(\mathbf{X} = \mathbf{x}_{cf})) = \mathbf{W}_c \mathbf{x}_{cf}, \\ \mathbf{TDE} &= \mathbf{Y}_f - \mathbf{Y}_{cf}, \\ \mathcal{L}_{ce}^{fcf} &= \mathcal{L}_{ce}(\mathbf{TDE}, y) + \mathcal{L}_{ce}(\mathbf{Y}_f, y), \end{aligned} \quad (6)$$

where  $\mathbf{W}_c$  and  $y$  denote the ID classifier and the ground truth, respectively. The minimization of  $\mathcal{L}_{ce}(\mathbf{Y}_f, y)$  and  $\mathcal{L}_{ce}(\mathbf{TDE}, y)$  enforces the factual feature  $\mathbf{x}_f$  comprising all ID-intrinsic information, and the counterfactual feature  $\mathbf{x}_{cf}$  only comprising non-ID clues, respectively. Thus, the influence of confounders can be eliminated by the TDE operation. The Factual and Counterfactual Loss of the  $i$ -th CLTD is expressed as:

$$\mathcal{L}_{fcf}^i = \mathcal{L}_{NCE}^{fcf} + \mathcal{L}_{ce}^{fcf}. \quad (7)$$

The overall loss function is the combination of the triplet loss  $\mathcal{L}_{tri}$  [16], cross-entropy loss  $\mathcal{L}_{ce}$ , and multi-stage Factual and Counterfactual Loss  $\mathcal{L}_{fcf}$ :

$$\mathcal{L} = \mathcal{L}_{tri} + \mathcal{L}_{ce} + \sum_{i=1}^3 \lambda_i \cdot \mathcal{L}_{fcf}^i, \quad (8)$$

where  $\lambda_i$  is a hyper-parameter to balance the differences among features at various stages.

## 4 Experiments

We conduct extensive experiments on four popular datasets, *i.e.*, OU-MVLP [43], CASIA-B [53], GREW [60], and Gait3D [58], to demonstrate the effectiveness of the proposed approach.

### 4.1 Datasets

OU-MVLP [43] encompasses a significant number of indoor gait samples, totaling 10307 subjects. Each subject is represented by 14 view angles, evenly distributed between  $0^\circ$  to  $90^\circ$  and  $180^\circ$  to  $270^\circ$ , with each angle containing two sequences (Seq#00-01). According to the protocol outlined in [29, 43], a total of 5153 subjects are employed as the training set, while the remaining 5154 are allocated for testing. CASIA-B [53] comprises 124 subjects, each captured from 11 distinct view angles evenly spread across a range from  $0^\circ$  to  $180^\circ$ . Within each view angle, there are three different walking conditions: normal walking (NM#1-6), walking with bags (BG#1-2), and walking with coats (CL#1-2). Following the established protocol detailed in [3], the initial 74 sequences are used for training. GREW [60] is a large-scale wild dataset containing a vast array of 26345 subjects and 128671 sequences. GREW is divided into three distinct parts: the training set (20000 subjects), the validation set (345 subjects), and the test set (6000 subjects). Gait3D [58] is a wild dataset comprising 4000 subjects and 25309 sequences. Following its protocol, 3000 subjects are designated for the training set, while the remaining subjects are allocated for the test set.

### 4.2 Implementation Details

Our approach is implemented using the PyTorch framework [36], and all experiments are conducted on NVIDIA GeForce RTX3090 GPUs. For all experiments, the default input silhouette size is  $64 \times 44$ .

**Baseline Models.** To demonstrate the effectiveness of CLTD, we employ DyGait [50] as a backbone for comparison with some advanced methods. Additionally, to showcase the versatility of CLTD, we also use GaitSet [3], GaitPart [10], GaitGL [29], GaitGCI [8], and GaitBase [9] as baselines, which illustrates the improvements CLTD achieves across various frameworks on diverse datasets.

**Training Strategies.** The experimental settings align with the established training strategies of baseline models, including the choice of the optimizer, scheduler, iteration count, and batch size. This way ensures a fair and consistent evaluation framework across all models. More training strategies are provided in the supplementary material.

**Hyper Parameters.** To adapt our approach to the unique characteristics of each dataset, we tailor certain parameters accordingly. Specifically, for CASIA-B, we set the output dimension of FPH, denoted as  $C_o$  to 128, and the weights  $(\lambda_1, \lambda_2, \lambda_3)$  in Eq. (8) to (0.05, 0.10, 0.15), respectively. For OU-MVLP, GREW, and Gait3D, we increase  $C_o$  and the weights  $(\lambda_1, \lambda_2, \lambda_3)$  to 256 and (0.1, 0.2, 0.3), respectively. Additionally, for all datasets, we set the number of CLTDs to 3, and the window size  $k \times k$  of LFS in FPH to  $7 \times 7$ .

**Table 1:** Comparison results of Rank-1 (%) on OU-MVLP, excluding identical-view cases. The best result is indicated in **bold**. The results of the last 6 lines are obtained by removing invalid probe sequences.

Method	Venue	Probe View														Mean
		0°	15°	30°	45°	60°	75°	90°	180°	195°	210°	225°	240°	255°	270°	
GaitSet [3]	AAAI19	70.3	87.9	90.0	90.1	88.0	88.7	87.7	81.8	86.5	89.0	89.2	87.2	87.6	86.2	87.1
GaitPart [10]	CVPR20	82.6	88.9	90.8	91.0	89.7	89.9	89.5	85.2	88.1	90.0	90.1	89.0	89.1	88.2	88.7
GLN [18]	ECCV20	83.8	90.0	91.0	91.2	90.3	90.0	89.4	85.3	89.1	90.5	90.6	89.6	89.3	88.5	89.2
GaitGL [29]	ICCV21	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
CSTL [21]	ICCV21	87.1	91.0	91.5	91.8	90.6	90.8	90.6	89.4	90.2	90.5	90.7	89.8	90.0	89.4	90.2
3DLocal [22]	ICCV21	86.1	91.2	92.6	92.9	92.2	91.3	91.1	86.9	90.8	<b>92.2</b>	92.3	91.3	91.1	90.2	90.9
MetaGait [7]	ECCV22	88.2	92.3	<b>93.0</b>	<b>93.5</b>	93.1	<b>92.7</b>	<b>92.6</b>	89.3	91.2	92.0	<b>92.6</b>	92.3	<b>91.9</b>	91.1	91.9
DANet [31]	CVPR23	87.7	91.3	91.6	91.8	91.7	91.4	91.1	90.4	90.3	90.7	90.9	90.5	90.3	89.9	90.7
GaitBase [9]	CVPR23	-	-	-	-	-	-	-	-	-	-	-	-	-	-	90.8
GaitGCI [8]	CVPR23	91.2	92.3	92.6	92.7	93.0	92.3	92.1	92.0	<b>91.8</b>	91.9	<b>92.6</b>	92.3	91.4	91.6	92.1
<b>Ours</b>	-	<b>91.6</b>	<b>92.5</b>	92.7	92.6	<b>93.2</b>	92.4	92.4	<b>92.5</b>	<b>91.8</b>	<b>92.2</b>	91.9	<b>92.5</b>	<b>91.9</b>	<b>91.8</b>	<b>92.3</b>
GaitSet [3]	AAAI19	84.5	93.3	96.7	96.6	93.5	95.3	94.2	87.0	92.5	96.0	96.0	93.0	94.3	92.7	93.3
GaitPart [10]	CVPR20	88.0	94.7	97.7	97.6	95.5	96.6	96.2	90.6	94.2	97.2	97.1	95.1	96.0	95.0	95.1
GLN [18]	ECCV20	89.3	95.8	97.9	97.8	96.0	96.7	96.1	90.7	95.3	97.7	97.5	95.7	96.2	95.3	95.6
GaitGL [29]	ICCV21	90.5	96.1	98.0	98.1	97.0	97.6	97.1	94.2	94.9	97.4	97.4	95.7	96.5	95.7	96.2
DyGait [50]	ICCV23	96.2	98.2	99.1	99.0	98.6	99.0	98.8	97.9	97.6	98.8	98.6	98.1	98.3	98.2	98.3
<b>Ours</b>	-	<b>97.0</b>	<b>98.3</b>	<b>99.5</b>	<b>99.2</b>	<b>98.9</b>	<b>99.2</b>	<b>99.1</b>	<b>98.1</b>	<b>98.0</b>	<b>99.1</b>	<b>98.8</b>	<b>98.5</b>	<b>98.7</b>	<b>98.3</b>	<b>98.6</b>

### 4.3 Comparison with State-of-the-art Methods

In this section, we compare our approach with several methods [2, 3, 7–10, 18, 21, 22, 29, 31, 48–50, 57, 58]. More experimental results are provided in the supplementary material.

**Evaluation on OU-MVLP.** As indicated in Tab. 1, we compare CLTD with previous methods on OU-MVLP. Our proposed approach outperforms the existing methods across most view angles (9 out of 14), which demonstrates its effectiveness. Specifically, it surpasses GaitGCI [8] and DyGait [50] by 0.2% and 0.3%, respectively, establishing SOTA performance and affirming its effectiveness.

**Evaluation on CASIA-B.** As illustrated in Tab. 2, our approach undergoes a comparative analysis with published methods on three walking conditions under identical evaluation settings.

Observing the table, our approach achieves an average accuracy of 94.8%, which is 0.2% and 0.3% higher than DANet [31] and GaitGCI [8], respectively, establishing itself as the SOTA. Furthermore, compared to DyGait [50], the integration of CLTD yields remarkable performance improvement, with an average improvement of 0.7%, particularly evident under the CL condition, where it

Method	Venue	NM	BG	CL	Mean
GaitSet [3]	AAAI19	95.0	87.2	70.4	84.2
GaitPart [10]	CVPR20	96.2	91.5	78.7	88.8
GLN [18]	ECCV20	96.9	94.0	77.5	89.5
GaitGL [29]	ICCV21	97.4	94.5	83.6	91.8
3DLocal [22]	ICCV21	97.5	94.4	83.7	91.8
CSTL [21]	ICCV21	97.8	93.6	84.2	91.9
LagrangeGait [2]	CVPR22	96.9	93.5	86.5	92.3
MetaGait [7]	ECCV22	98.1	95.2	86.9	93.4
DANet [31]	CVPR23	98.0	95.9	<b>89.9</b>	94.6
GaitBase [9]	CVPR23	97.6	94.0	77.4	89.6
GaitGCI [8]	CVPR23	98.4	<b>96.6</b>	88.5	94.5
HSTL [49]	ICCV23	98.1	95.9	88.9	94.3
DyGait [50]	ICCV23	98.4	96.2	87.8	94.1
<b>Ours</b>	-	<b>98.6</b>	96.4	89.3	<b>94.8</b>

**Table 2:** Comparison results of Rank-1 (%) with SOTA methods on CASIA-B, excluding identical-view case.

**Table 3:** Comparison of Rank-1 and Rank-5 (%) on GREW and Giat3D.

Method	Venue	GREW		Gait3D	
		Rank-1	Rank-5	Rank-1	Rank-5
GaiSet [3]	AAAI19	46.3	63.6	36.7	58.3
GaitPart [10]	CVPR20	44.0	60.7	28.2	47.6
GaitGL [29]	ICCV21	47.3	63.6	29.7	48.5
SMPLGait [58]	CVPR22	-	-	46.3	64.5
MTSGait [57]	ACM MM22	55.3	71.3	48.7	67.1
DANet [31]	CVPR23	-	-	48.0	69.7
GaitBase [9]	CVPR23	60.1	-	64.6	-
GaitGCI [8]	CVPR23	68.5	80.8	50.3	68.5
HSTL [49]	ICCV23	62.7	76.6	61.3	76.3
DyGait [50]	ICCV23	71.4	83.2	66.3	80.8
GaitCSV [48]	ACM MM23	64.9	-	69.1	-
<b>Ours</b>	-	<b>78.0</b>	<b>87.8</b>	<b>69.7</b>	<b>85.2</b>

increases by 1.5%. This underscores the strong discriminative feature learning capability of our approach, attributed to the effective elimination of confounders.

It is important to note that the performance improvements on OU-MVLP and CASIA-B may not appear particularly significant, with Rank-1 accuracy improvement of 0.3% and 0.7% respectively. This is likely due to the relative simplicity of these datasets, where performance tends to reach saturation. However, the benefits of CLTD will become more pronounced when applied to more challenging datasets, as demonstrated in the following experiments.

**Evaluation on GREW and Gait3D.** Validation of the effectiveness of our approach on two wild datasets, namely GREW and Gait3D, is presented in Tab. 3. Compared with DyGait [50] and GaitCSV [48], our approach outperforms by 6.6% and 13.1% on GREW, and by 3.4% and 0.6% on Gait3D, respectively. This highlights the capability of our approach to eliminate the influence of confounders effectively, capture more discriminative features, and contribute to better performance. When considering the results collectively from Tab. 1, Tab. 2, and Tab. 3, it is evident that other methods experience a significant performance drop on wild datasets due to increased confounders. However, our approach still achieves 78.0% Rank-1 accuracy on GREW, establishing itself as the SOTA, and demonstrating strong robustness across diverse scenarios.

#### 4.4 Ablation Study

In this section, we conduct ablation experiments on GREW and Gait3D to verify the design of CLTD.

**Versatility of CLTD.** In Tab. 4, we investigate the versatility of CLTD with different gait recognition models, including GaitSet [3], GaitPart [10], GaitGL [29], GaitBase [9], GaitGCI [8], and DyGait [50]. From Tab. 4, we can summarize the following valuable findings: (1) When integrated with our approach, all methods exhibit performance improvements on four datasets, indicating the efficacy of

**Table 4:** Performance improvements (Rank-1 (%)) when using CLTD with different baselines. Identical-view cases are excluded on CASIA-B and OU-MVLP. The symbol  $\dagger$  represents excluding invalid probe sequences on OUMVLP. \* indicates GaitGCI without its CIL block.

Backbone	Venue	Testing Dataset			
		CASIA-B	OU-MVLP	GREW	Gait3D
GaiSet [3]	AAAI19	84.2	87.1	46.3	36.7
GaiSet w/CLTD	-	<b>88.1</b> <sup>†2.2</sup>	<b>89.3</b> <sup>†2.2</sup>	<b>51.8</b> <sup>†5.5</sup>	<b>40.8</b> <sup>†4.1</sup>
GaitPart [10]	CVPR20	88.8	88.7	44.0	28.2
GaitPart w/CLTD	-	<b>91.5</b> <sup>†2.7</sup>	<b>90.1</b> <sup>†1.4</sup>	<b>51.1</b> <sup>†7.1</sup>	<b>31.4</b> <sup>†3.2</sup>
GaitGL [29]	ICCV21	91.8	89.7	47.3	29.7
GaitGL w/CLTD	-	<b>93.5</b> <sup>†1.7</sup>	<b>90.8</b> <sup>†1.1</sup>	<b>57.4</b> <sup>†10.1</sup>	<b>33.2</b> <sup>†3.5</sup>
GaitBase [9]	CVPR23	89.6	90.8	60.1	64.6
GaitBase w/CLTD	-	<b>92.4</b> <sup>†2.8</sup>	<b>91.3</b> <sup>†0.5</sup>	<b>70.4</b> <sup>†10.3</sup>	<b>71.9</b> <sup>†7.3</sup>
GaitGCI [8]	CVPR23	93.1	92.1	68.5	50.3
GaitGCI*	-	90.1	90.2	60.2	45.8
GaitGCI* w/CLTD	-	<b>93.5</b> <sup>†3.4</sup>	<b>92.3</b> <sup>†2.1</sup>	<b>71.3</b> <sup>†11.1</sup>	<b>52.2</b> <sup>†6.4</sup>
DyGait <sup>†</sup> [50]	ICCV23	94.1	98.3	71.4	66.3
DyGait <sup>†</sup> w/CLTD	-	<b>94.8</b> <sup>†0.7</sup>	<b>98.6</b> <sup>†0.3</sup>	<b>78.0</b> <sup>†6.6</sup>	<b>69.7</b> <sup>†3.4</sup>

CLTD. (2) Particularly noteworthy is the significant performance improvement observed on GREW, with a maximum improvement of **11.1%** (from 60.2% to 71.3%) when using GaitGCI without its causal module CIL as the backbone.

**Effectiveness of proposed components.** To assess the effectiveness of the proposed components, we conduct ablation experiments on GREW and Gait3D. As shown in Tab. 5, it can be observed that the integration of proposed modules leads to consistent performance improvements. The CPAG contributes to the improvement of recognition performance, with an average improvement of 1.9% on GREW compared to the baseline. Additionally, employing FPH and Factual and Counterfactual Loss together yields 5.7% performance improvement on GREW, underscoring the effectiveness of these components.

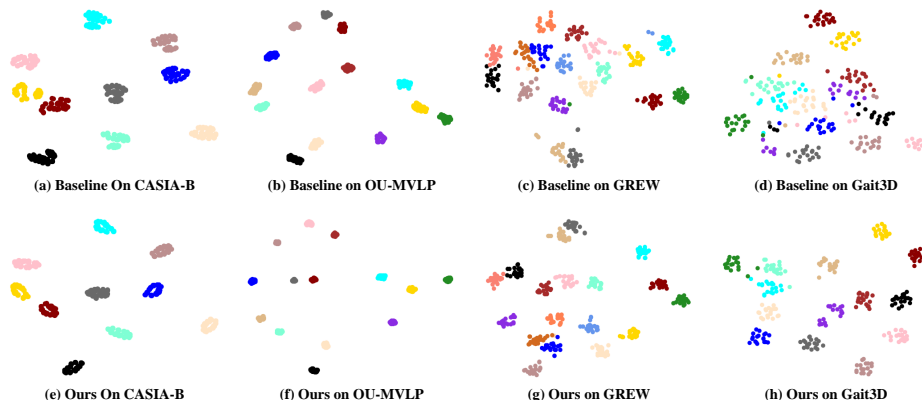
**Impact of multiple stages.** In Tab. 6, we analyze the impact of the number and position of CLTDs in terms of accuracy. From the results, we note that: (1) The impact of using CLTD at different positions on performance varies. The deeper layer it is used at, the more significant performance improvement is achieved. This phenomenon indicates the importance of eliminating confounders closer to the output layer. (2) The best

Method	GREW	Gait3D
Baseline	71.4	66.3
Baseline + CPAG	73.3	67.1
Baseline + CPAG + FPH	76.0	68.6
<b>Ours</b>	<b>78.0</b>	<b>69.7</b>

**Table 5:** Study the effectiveness of proposed modules in CLTD on GREW and Gait3D, including CPAG, FPH and Factual and Counterfactual Loss.

Method	1-st	2-nd	3-rd	GREW	Gait3D
Baseline				71.4	66.3
a	✓			73.6	67.0
b		✓		74.2	67.5
c			✓	75.1	68.4
d	✓	✓		75.4	68.6
e	✓		✓	76.5	69.1
f		✓	✓	77.1	69.3
g	✓	✓	✓	<b>78.0</b>	<b>69.7</b>

**Table 6:** Impact of multiple stages.



**Fig. 6:** t-SNE visualization examples of feature distributions between the baseline and our approach on CASIA-B, OU-MVLP, GREW, and Gait3D, respectively. Different colors denote distinct identities. Best viewed by color and zooming in.

performance is achieved when CLTDs are used simultaneously at multiple positions, resulting in a 6.6% improvement compared to the baseline on GREW. This highlights the complementary nature of CLTD when integrated at different positions within the network.

**Effectiveness of factual cross-entropy loss and InfoNCE loss.** As reported in Tab. 7, we conduct experiments to investigate the effectiveness of factual cross-entropy loss and InfoNCE loss. We find that both losses contribute to improved recognition performance, and using them together results in even higher performance, demonstrating their complementary properties.

**Impact of the window size  $k \times k$  in FPH.**

To study the impact of the value  $k$ , we conduct five experiments detailed in Tab. 8 with  $k$  of 3, 5, 7, 9, and 11, respectively. The average accuracy initially increases as  $k$  rises, reaching a peak, and then degrades with larger  $k$ . This trend indicates that very small  $k$  may not adequately explore discriminative features, while very large  $k$  may distract discriminative features with less discriminative features. Based on the results in Tab. 8,  $k = 7$  is chosen for the best performance.

Method	Cross-Entropy	InfoNCE	GREW	Gait3D
Baseline			71.4	66.3
a	✓		76.0	68.6
b		✓	77.3	69.2
c	✓	✓	<b>78.0</b>	<b>69.7</b>

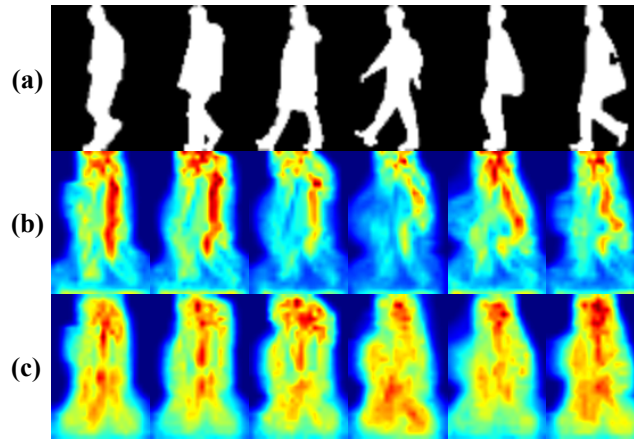
**Table 7:** Impact of loss function.

$k \times k$	GREW		Gait3D	
	Rank-1	Rank-5	Rank-1	Rank-5
$3 \times 3$	76.4	85.2	66.7	81.6
$5 \times 5$	77.0	86.4	68.3	83.9
$7 \times 7$	<b>78.0</b>	<b>87.8</b>	<b>69.7</b>	<b>85.2</b>
$9 \times 9$	77.2	86.6	68.9	84.7
$11 \times 11$	76.3	84.9	67.4	82.5

**Table 8:** Impact of the window size  $k \times k$  in FPH.

#### 4.5 Qualitative results

**Visualization of feature distributions.** We employ t-SNE [32] to visualize the feature distributions between the baseline and our approach on CASIA-B,



**Fig. 7:** Visualization of heatmaps representing counterfactual and factual features after CLTD. Best viewed in color. (a) Original input. (b) Counterfactual features. (c) Factual features.

OU-MVLP, GREW, and Gait3D. As shown in Fig. 6, we observe that the feature distributions produced by our approach are more compact for the same subject compared to the baseline. Consequently, identities become easier to distinguish. These visualizations confirm that our approach effectively eliminates the interference of confounders and extracts the discriminative features that are genuinely relevant for identification. This provides additional support for the effectiveness of our approach.

**Visualization of heatmaps.** As illustrated in Fig. 7, we present visualizations of the feature heatmaps processed after CLTD. The objective is to illustrate the superiority of our approach qualitatively. A notable observation is that counterfactual attention tends to focus more on confounders such as clothing and bagging, which are unrelated to identity. In contrast, factual attention concentrates on ID-intrinsic information. This observation demonstrates that CLTD can effectively eliminate the influence of confounders.

## 5 Conclusion

In this work, we present a discriminative feature learning module CLTD based on causality. CLTD effectively eliminates the impact of confounders in spatial, temporal, and spectral domains. Thorough quantitative and qualitative experimental analyses demonstrate the effectiveness and versatility of our approach. Future work will explore the application of causality in other computer vision tasks, such as action recognition, person re-identification, and image restoration.

**Limitation.** FPH can improve performance while reducing time costs, but feature selection is fixed. Next, we consider using the attention mechanism for adaptive selection.

## Acknowledgements

This work was supported by the National Key R&D Program of China under project 2023YFF0905401.

## References

1. Abraham, E.D., D'Oosterlinck, K., Feder, A., Gat, Y., Geiger, A., Potts, C., Reichart, R., Wu, Z.: Cebab: Estimating the causal effects of real-world concepts on nlp model behavior. *Advances in Neural Information Processing Systems* **35**, 17582–17596 (2022)
2. Chai, T., Li, A., Zhang, S., Li, Z., Wang, Y.: Lagrange motion analysis and view embeddings for improved gait recognition. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 20249–20258 (2022)
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*. pp. 8126–8133 (2019)
4. Chen, J., Gao, Z., Wu, X., Luo, J.: Meta-causal learning for single domain generalization. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 7683–7692 (2023)
5. Ding, S., Feng, F., He, X., Liao, Y., Shi, J., Zhang, Y.: Causal incremental graph convolution for recommender system retraining. *IEEE Transactions on Neural Networks and Learning Systems* (2022)
6. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929* (2020)
7. Dou, H., Zhang, P., Su, W., Yu, Y., Li, X.: Metagait: Learning to learn an omni sample adaptive representation for gait recognition. In: *European Conference on Computer Vision*. pp. 357–374. Springer (2022)
8. Dou, H., Zhang, P., Su, W., Yu, Y., Lin, Y., Li, X.: Gaitgci: Generative counterfactual intervention for gait recognition. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 5578–5588 (2023)
9. Fan, C., Liang, J., Shen, C., Hou, S., Huang, Y., Yu, S.: Opengait: Revisiting gait recognition towards better practicality. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 9707–9716 (2023)
10. 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)
11. Fu, Y., Meng, S., Hou, S., Hu, X., Huang, Y.: Gpgait: Generalized pose-based gait recognition. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. pp. 19595–19604 (October 2023)
12. Guo, H., Ji, Q.: Physics-augmented autoencoder for 3d skeleton-based gait recognition. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. pp. 19627–19638 (2023)
13. Guo, T., Liu, H., Chen, Z., Liu, M., Wang, T., Ding, R.: Contrastive learning from extremely augmented skeleton sequences for self-supervised action recognition. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. pp. 762–770 (2022)

14. Gutmann, M., Hyvärinen, A.: Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In: Proceedings of the thirteenth international conference on artificial intelligence and statistics. pp. 297–304. JMLR Workshop and Conference Proceedings (2010)
15. Han, J., Bhanu, B.: Individual recognition using gait energy image. *IEEE transactions on pattern analysis and machine intelligence* **28**(2), 316–322 (2005)
16. Hermans, A., Beyer, L., Leibe, B.: In defense of the triplet loss for person re-identification. *arXiv preprint arXiv:1703.07737* (2017)
17. Hou, R., Chang, H., Ma, B., Huang, R., Shan, S.: Bicnet-tks: Learning efficient spatial-temporal representation for video person re-identification. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 2014–2023 (2021)
18. Hou, S., Cao, C., Liu, X., Huang, Y.: Gait lateral network: Learning discriminative and compact representations for gait recognition. In: European conference on computer vision. pp. 382–398. Springer (2020)
19. Huang, X., Wang, X., He, B., He, S., Liu, W., Feng, B.: Star: Spatio-temporal augmented relation network for gait recognition. *IEEE Transactions on Biometrics, Behavior, and Identity Science* **5**(1), 115–125 (2022)
20. Huang, X., Wang, X., Jin, Z., Yang, B., He, B., Feng, B., Liu, W.: Condition-adaptive graph convolution learning for skeleton-based gait recognition. *IEEE Transactions on Image Processing* (2023)
21. Huang, X., Zhu, D., Wang, H., Wang, X., Yang, B., He, B., Liu, W., Feng, B.: Context-sensitive temporal feature learning for gait recognition. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 12909–12918 (2021)
22. Huang, Z., Xue, D., Shen, X., Tian, X., Li, H., Huang, J., Hua, X.S.: 3d local convolutional neural networks for gait recognition. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 14920–14929 (2021)
23. Huang, Z., Wang, X., Huang, L., Huang, C., Wei, Y., Liu, W.: Ccnet: Criss-cross attention for semantic segmentation. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 603–612 (2019)
24. Lee, B.K., Kim, J., Ro, Y.M.: Mitigating adversarial vulnerability through causal parameter estimation by adversarial double machine learning. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 4499–4509 (2023)
25. Lee, S., Bae, J., Kim, H.Y.: Decompose, adjust, compose: Effective normalization by playing with frequency for domain generalization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11776–11785 (2023)
26. Li, C., Guo, C.L., Zhou, M., Liang, Z., Zhou, S., Feng, R., Loy, C.C.: Embedding-fourier for ultra-high-definition low-light image enhancement. In: ICLR (2023)
27. Li, X., Makihara, Y., Xu, C., Yagi, Y., Ren, M.: Gait recognition via semi-supervised disentangled representation learning to identity and covariate features. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 13309–13319 (2020)
28. Li, X., Qian, X., Liang, L., Kong, L., Dong, Q., Chen, J., Liu, D., Yao, X., Fu, Y.: Causally-aware intraoperative imputation for overall survival time prediction. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 15681–15690 (2023)

29. 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)
30. Loper, M., Mahmood, N., Romero, J., Pons-Moll, G., Black, M.J.: Smpl: A skinned multi-person linear model. *ACM Transactions on Graphics* **34**(6) (2015)
31. Ma, K., Fu, Y., Zheng, D., Cao, C., Hu, X., Huang, Y.: Dynamic aggregated network for gait recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 22076–22085 (2023)
32. Van der Maaten, L., Hinton, G.: Visualizing data using t-sne. *Journal of machine learning research* **9**(11) (2008)
33. Mao, X., Liu, Y., Liu, F., Li, Q., Shen, W., Wang, Y.: Intriguing findings of frequency selection for image deblurring. In: Proceedings of the AAAI Conference on Artificial Intelligence. pp. 1905–1913 (2023)
34. Miao, J., Chen, C., Liu, F., Wei, H., Heng, P.A.: Caussl: Causality-inspired semi-supervised learning for medical image segmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 21426–21437 (2023)
35. Oord, A.v.d., Li, Y., Vinyals, O.: Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748* (2018)
36. 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)
37. Pearl, J., Glymour, M., Jewell, N.P.: Causal inference in statistics: A primer. John Wiley & Sons (2016)
38. Quan, S., Hirano, M., Yamakawa, Y.: Semantic information in contrastive learning. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 5686–5696 (2023)
39. Rao, H., Miao, C.: Transg: Transformer-based skeleton graph prototype contrastive learning with structure-trajectory prompted reconstruction for person re-identification. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 22118–22128 (2023)
40. Sarkar, S., Phillips, P.J., Liu, Z., Vega, I.R., Grother, P., Bowyer, K.W.: The humanoid gait challenge problem: Data sets, performance, and analysis. *IEEE transactions on pattern analysis and machine intelligence* **27**(2), 162–177 (2005)
41. Shen, C., Yu, S., Wang, J., Huang, G.Q., Wang, L.: A comprehensive survey on deep gait recognition: algorithms, datasets and challenges. *arXiv preprint arXiv:2206.13732* (2022)
42. Shiraga, K., Makihara, Y., Muramatsu, D., Echigo, T., Yagi, Y.: Geinet: View-invariant gait recognition using a convolutional neural network. In: 2016 international conference on biometrics (ICB). pp. 1–8. IEEE (2016)
43. 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. *IPSN transactions on Computer Vision and Applications* **10**, 1–14 (2018)
44. Tang, K., Niu, Y., Huang, J., Shi, J., Zhang, H.: Unbiased scene graph generation from biased training. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 3716–3725 (2020)
45. Teepe, T., Gilg, J., Herzog, F., Hörmann, S., Rigoll, G.: Towards a deeper understanding of skeleton-based gait recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1569–1577 (2022)

46. Teepe, T., Khan, A., Gilg, J., Herzog, F., Hörmann, S., Rigoll, G.: Gaitgraph: Graph convolutional network for skeleton-based gait recognition. In: 2021 IEEE International Conference on Image Processing (ICIP). pp. 2314–2318. IEEE (2021)
47. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. *Advances in neural information processing systems* **30** (2017)
48. Wang, J., Hou, S., Huang, Y., Cao, C., Liu, X., Huang, Y., Wang, L.: Causal intervention for sparse-view gait recognition. In: Proceedings of the 31st ACM International Conference on Multimedia. pp. 77–85 (2023)
49. Wang, L., Liu, B., Liang, F., Wang, B.: Hierarchical spatio-temporal representation learning for gait recognition. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 19639–19649 (2023)
50. Wang, M., Guo, X., Lin, B., Yang, T., Zhu, Z., Li, L., Zhang, S., Yu, X.: Dygait: Exploiting dynamic representations for high-performance gait recognition. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 13424–13433 (October 2023)
51. Yang, X., Zhang, H., Qi, G., Cai, J.: Causal attention for vision-language tasks. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 9847–9857 (2021)
52. Yang, Z., Lin, M., Zhong, X., Wu, Y., Wang, Z.: Good is bad: Causality inspired cloth-debiasing for cloth-changing person re-identification. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1472–1481 (2023)
53. 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)
54. Zhang, C., Chen, X.P., Han, G.Q., Liu, X.J.: Spatial transformer network on skeleton-based gait recognition. *Expert Systems* p. e13244 (2023)
55. Zhang, Q.: Probabilistic reasoning based on dynamic causality trees/diagrams. *Reliability Engineering & System Safety* **46**(3), 209–220 (1994)
56. Zhao, Z., Wang, D., Zhao, X.: Movement enhancement toward multi-scale video feature representation for temporal action detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 13555–13564 (2023)
57. Zheng, J., Liu, X., Gu, X., Sun, Y., Gan, C., Zhang, J., Liu, W., Yan, C.: Gait recognition in the wild with multi-hop temporal switch. In: Proceedings of the 30th ACM International Conference on Multimedia. pp. 6136–6145 (2022)
58. Zheng, J., Liu, X., Liu, W., He, L., Yan, C., Mei, T.: Gait recognition in the wild with dense 3d representations and a benchmark. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 20228–20237 (2022)
59. Zhou, M., Huang, J., Guo, C.L., Li, C.: Fourmer: An efficient global modeling paradigm for image restoration. In: International Conference on Machine Learning. pp. 42589–42601. PMLR (2023)
60. 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)