GaitGCI: Generative Counterfactual Intervention for Gait Recognition

Huanzhang Dou[1] Pengyi Zhang[1] Wei Su[1] Yunlong Yu[2] Yining Lin[3] Xi Li[1,4,5,6*]
[1]College of Computer Science & Technology, Zhejiang University
[2]College of Information Science & Electronic Engineering, Zhejiang University
[3]SupreMind [4]Shanghai Institute for Advanced Study, Zhejiang University
[5]Shanghai AI Laboratory [6]Zhejiang – Singapore Innovation and AI Joint Research Lab

Abstract

*Gait is one of the most promising biometrics that aims to identify pedestrians from their walking patterns. However, prevailing methods are susceptible to confounders, resulting in the networks hardly focusing on the regions that reflect effective walking patterns. To address this fundamental problem in gait recognition, we propose a **G**enerative **C**ounterfactual **I**ntervention framework, dubbed GaitGCI, consisting of **C**ounterfactual **I**ntervention **L**earning (CIL) and **D**iversity-**C**onstrained **D**ynamic **C**onvolution (DCDC). CIL eliminates the impacts of confounders by maximizing the likelihood difference between factual/counterfactual attention while DCDC adaptively generates sample-wise factual/counterfactual attention to efficiently perceive the sample-wise properties. With matrix decomposition and diversity constraint, DCDC guarantees the model to be efficient and effective. Extensive experiments indicate that proposed GaitGCI: 1) could effectively focus on the discriminative and interpretable regions that reflect gait pattern; 2) is model-agnostic and could be plugged into existing models to improve performance with nearly no extra cost; 3) efficiently achieves state-of-the-art performance on arbitrary scenarios (in-the-lab and in-the-wild).*

1. Introduction

Gait recognition aims to utilize walking patterns to identify pedestrians without explicit cooperation, thus drawing rising attention. Current gait recognition research focuses on in-the-lab [53, 69] and in-the-wild scenarios [73, 76] for theoretical analysis and practical application, respectively.

The key to addressing gait recognition is to fully capture the effective visual cues of the gait patterns, *i.e.*, the regions close to the body boundary [39, 60] for both in-the-lab scenarios and in-the-wild scenarios. However, the attention analysis [4, 59, 68] on prevailing methods in Fig. 1 indicates that the existing methods hardly capture the effective gait patterns and tend to collapse into the suboptimal attention regions, which would deteriorate the gait representation. We argue that this phenomenon is caused by the network's susceptibility to the *confounders* [21, 32], which may provide *shortcuts* [21, 32] for the models rather than the valid gait-related patterns. For example, the attention regions of prevailing methods are related to viewpoints [64] or walking conditions [30]. As shown in Fig. 1, the prevailing network tends to focus on the head under the front view and the head/feet under the side view. However, the majority of the gait-related information close to the boundary is neglected. Therefore, how to alleviate the impact of confounders is a fundamental problem to model discriminative and interpretable gait representation.

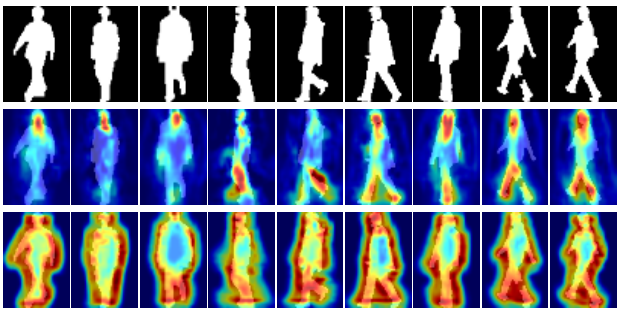


Figure 1. Network attention comparison. From top to down: silhouette, existing method, and proposed GaitGCI. The confounders make the existing model collapse into suboptimal attention regions. By contrast, GaitGCI could effectively focus on the discriminative and interpretable regions (*i.e.*, close to the boundary [39, 60]) that could represent walking patterns.

Motivated by this, we propose a generative counterfactual intervention framework, named GaitGCI, consisting of Counterfactual Intervention Learning (CIL) and Diversity-Constrained Dynamic Convolution (DCDC). The core idea of CIL is to leverage the counterfactual-based causal inference to alleviate the impact of confounders and mine the

*Corresponding author.

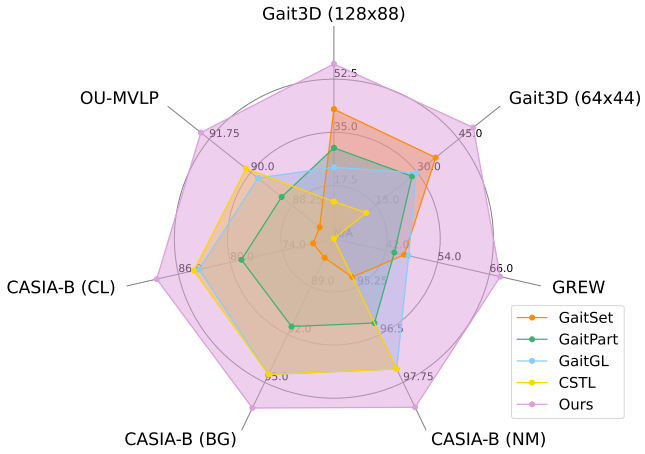


Figure 2. GaitGCI could achieve state-of-the-art performance under arbitrary scenarios, including in-the-lab scenarios [53, 69] and in-the-wild scenarios [75, 76].

direct causality link between factual attention and prediction. Specifically, we first construct a causal analysis tool (*i.e.*, Structural Causal Model [47]) to formulate the causality links among the input, attention, and prediction. Then, the training objective is modified from maximizing the original likelihood that contains confounders to maximizing the likelihood difference between the factual/counterfactual attention, which forces the network to focus on the direct causality between the factual attention and the prediction instead of collapsing into the confounders.

Further, considering that the previous network to produce factual attention is static and the mainstream counterfactual is pre-defined distribution [11, 49] (*e.g.*, random or normal distribution), which limits the ability of the network to perceive the sample-wise properties. Therefore, we propose a Diversity-Constrained Dynamic Convolution (DCDC) to efficiently produce the sample-adaptive kernel, which aims to generate factual/counterfactual attention. Specifically, we first decouple the dynamic convolution [57, 67] into the sample-agnostic convolution and sample-adaptive convolution. Then, to improve the efficiency, we apply the matrix decomposition to decompose sample-adaptive convolution into two bases and a generative affinity matrix, which transforms dense convolution integration in high-dimensional space into the aggregation of bases in low-dimensional space. Besides, to guarantee the representation power, we propose a rank-based diversity constraint on two bases of the sample-adaptive convolution.

By alleviating the impact of confounders, the proposed method: (1) could effectively focus on the discriminative and interpretable regions instead of collapsing into the confounders; (2) is model-agnostic and could boost the performance of prevailing methods; (3) could efficiently achieve state-of-the-art performance under arbitrary scenarios (in-the-lab and in-the-wild) as shown in Fig. 2.

The main contributions are summarized as follows:

- We present counterfactual intervention learning (CIL) to alleviate the impact of confounders. CIL could effectively force the model to focus on the regions that reflect gait patterns by maximizing the likelihood difference between factual/counterfactual attention.
- We present diversity-constrained dynamic convolution (DCDC) to generate factual/counterfactual attention in a sample adaptive manner. Matrix decomposition and diversity constraint guarantee efficiency and representation power, respectively.
- Extensive experiments demonstrate that the proposed framework efficiently achieves state-of-the-art performance in arbitrary scenarios. Besides, the proposed methods could serve as a plug-and-play module to boost the performance of prevailing models.

2. Related Work

2.1. Gait Recognition

Prior research focuses on the in-the-lab scenario. However, VersatileGait [73] has pioneered the more challenging in-the-wild gait recognition via synthetic datasets. This problem draws increasing attention, resulting in the emergence of real-world datasets for in-the-wild scenarios [75, 76]. And mainstream methods could be grouped as follows:

Silhouette-based Methods. This fashion [17, 31] extracts gait patterns from the silhouette sequence. GaitSet [12] deems each sequence as an unordered set, GaitPart [19] proposes part-based modeling, and GaitGL [42] extracts features from global/local representation. This paradigm is sensitive to covariates but is more popular for its efficiency.

Skeleton-based Methods. Many methods [3, 6, 7, 23, 33, 40, 61] utilize pose estimation to model gait patterns. For example, Teepe *et al.* [54] model the skeleton as a graph and utilize GCN [35]. Li *et al.* [37] propose to jointly utilize 2D/3D keypoints information to model gait representation. These methods should be more robust to the covariates but rely on accurate pose estimation.

Methods using Other Modalities. Recently, more gait modalities have emerged. Several methods [37, 39, 74] extract features from RGB video. Castro *et. al.* [9] leverage optical flow to obtain abundant motion information. The depth information [44] and 3D mesh [37, 75] are also introduced to use extra information. Further, several works [8, 27, 75] conduct multi-modal learning to achieve informative representation.

2.2. Vision Causal Inference

Causal inference [20, 48, 63] arouses widespread attention to endow networks with the ability to analyze the

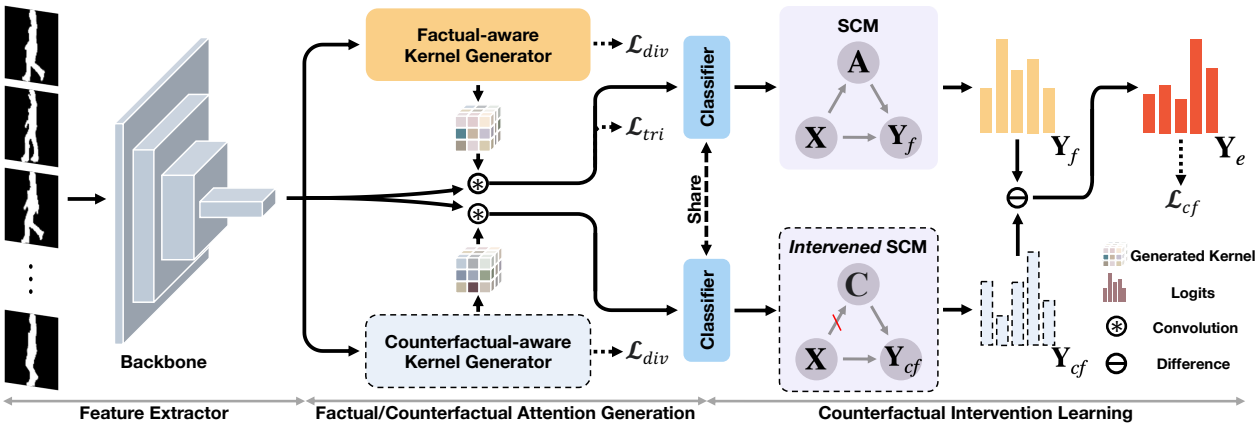


Figure 3. Overview of GaitGCI. The factual/counterfactual-aware generator is implemented by the proposed diversity-constrained dynamic convolution to efficiently generate factual/counterfactual attention based on the sample-wise properties. Then, counterfactual intervention learning is performed to maximize the likelihood difference between factual/counterfactual attention. The optimization objective is the combination of triplet loss, counterfactual loss, and diversity constraint.

causal effect. The causal inference has been successfully used in various areas, including visual explanation [22, 25], semantic segmentation [16], and few-shot/zero-shot learning [70, 71]. Previous vision causal inference methods with counterfactuals [1, 34] focus on the analysis of the outcome intervened by sorts of pre-defined counterfactuals. By contrast, we leverage dynamic convolution to adaptively perceive the sample-wise factual/counterfactual attention.

2.3. Dynamic Deep Neural Networks

Dynamic network [24] aims to boost the network capacity and generalizability via adapting its parameters or structures based on the input during inference. Dynamic convolution [57, 67] aggregates multiple candidate convolutions via the SE-style attention mechanism [29]. DRConv [13] proposes grouped dynamic convolution to adaptively select channels from groups. Besides, weight adjustment could be performed by soft attention over the spatial dimension of the convolutional weights [2, 52, 72]. In this paper, we propose to leverage matrix decomposition [38] and diversity constraint to guarantee the efficiency and representation power of dynamic convolution, respectively.

3. Method

3.1. Overview

As shown in Fig. 3, the silhouette is first fed to the backbone with low-rank 3D CNN. Then, factual/counterfactual attention is generated by the corresponding kernel generator (diversity-constrained dynamic convolution). Finally, GaitGCI is optimized with counterfactual loss, triplet loss, and diversity constraint. The feature aggregation (temporal pooling/separate FC [12]) is omitted for simplicity.

3.2. Counterfactual Intervention Learning

We propose Counterfactual Intervention Learning (CIL) to alleviate the impact of confounders. First, we formulate the learning process with the causality analysis tool, *i.e.*, the Structural Causal Model (SCM) [45, 47]. Then, the counterfactual intervention is introduced to analyze the direct causality link between factual attention and prediction.

Structural Causal Model Formulation. To represent the causality links among input $\mathbf{X}$, attention $\mathbf{A}$, and prediction $\mathbf{Y}$, we formulate them with the SCM $\mathcal{G} = \{\mathbf{N}, \mathbf{E}\}$, where $\mathbf{N}$ and $\mathbf{E}$ represent the variable nodes and causality links, respectively. The causality links denotes: *cause*$\rightarrow$ *effect*. Therefore, the causality could be formulated as $\mathbf{X} \rightarrow \mathbf{Y}$: the conventional model. $\mathbf{X} \rightarrow \mathbf{A}$: the model produces the corresponding attention. $\mathbf{X} \rightarrow \mathbf{Y} \leftarrow \mathbf{A}$: the final prediction $\mathbf{Y}$ is determined by $(\mathbf{X}, \mathbf{A})$ jointly. With SCM, The causality links between the variables can be directly analyzed via variable intervention, which means manipulating the value of specific variables and then observing the effect.

Counterfactual Intervention. Ideally, $\mathbf{A}$ decides to predict $\mathbf{Y}$ entirely by sensing the effective properties of $\mathbf{X}$. However, there are confounders in $\mathbf{X}$, which confuses the network's learning process and makes the network collapse into the suboptimal attention regions. Therefore, we propose to leverage the counterfactual intervention $Do(\cdot)$, which could cut off the causality link between the confounders and the factual attention.

The counterfactual intervention $Do(\cdot)$ could remove the impact of specific variables. Note that counterfactual [36, 58] means "counter to the facts," and the intervention is impossible to occur in the real world. Thus the process of $Do(\cdot)$ is called *counterfactual intervention*, which is

achieved by an imaginary intervention to replace the variables' state. For example, the value of the counterfactual intervention $Do(\mathbf{A} = C)$ means that the counterfactual C is assigned to $\mathbf{A}$ and breaks the causality link between $\mathbf{A}$ and its all parent nodes, which forces the variable to no longer be affected by the confounders. Therefore, the direct causality link between the factual attention $\mathbf{A}$ and the prediction $\mathbf{Y}$ could be analyzed. Specifically, the value A and C of factual attention $\mathbf{A}$ and counterfactual attention $\mathbf{C}$ is produced by the process $\mathcal{A}(\cdot)$ and $\mathcal{C}(\cdot)$, respectively.

$$A = \mathcal{A}(\mathbf{X}) = \{\mathbf{A}_0, \mathbf{A}_1, ..., \mathbf{A}_{M-1}\}, \tag{1}$$

$$C = \mathcal{C}(\mathbf{X}) = \{\mathbf{C}_0, \mathbf{C}_1, ..., \mathbf{C}_{M-1}\}, \tag{2}$$

where M is the channel number of A and C to control the capacity to perceive the sample-wise properties. In prevailing implementations, $\mathcal{A}(\cdot)$ is a static network, and $\mathcal{C}(\cdot)$ is a manually pre-defined distribution (*e.g.*, random or normal distribution). Then, the likelihood of counterfactual intervention $\mathbf{P}(\mathbf{Y}|Do(\mathbf{A} = C))$ could be leveraged to analyze the direct causality link between $\mathbf{A}$ and $\mathbf{Y}$ excluding the confounders. The likelihood of factual attention $\mathbf{Y}_f$ and counterfactual intervention $\mathbf{Y}_{cf}$ could be formulated as:

$$\mathbf{Y}_f = \mathbf{P}(\mathbf{Y}|\mathbf{A} = A) = \mathbb{E}_{A\sim\mathcal{A}(\mathbf{X})}(\mathbf{X} * A), \tag{3}$$

$$\mathbf{Y}_{cf} = \mathbf{P}(\mathbf{Y}|Do(\mathbf{A} = C)) = \mathbb{E}_{C\sim\mathcal{C}(\mathbf{X})}(\mathbf{X} * C). \tag{4}$$

The former $\mathbf{Y}_f$ is the key to model discriminative and interpretable gait representation with gait-related properties, and the latter $\mathbf{Y}_{cf}$ denotes the context-specific confounders, which is expected to be removed from the likelihood prediction. Then, we calculate the likelihood difference [46] between the factual attention and the counterfactual attention to obtain the direct causality effect $\mathbf{Y}_e$ between the factual attention $\mathbf{A}$ and the corresponding prediction $\mathbf{Y}$:

$$\mathbf{Y}_e = \mathbf{Y}_f - \mathbf{Y}_{cf}. \tag{5}$$

Maximizing the likelihood difference $\mathbf{Y}_e$ could force the network to focus on factual attention learning instead of collapsing into the confounders represented by the counterfactuals. Thus, counterfactuals can be regarded as additional supervision to alleviate the impact of confounders.

Note that CIL is model-agnostic and could be a plug-and-play module. Besides, the impact of confounders is a fundamental problem, thus CIL could theoretically be applied to arbitrary scenarios. Further, CIL is only used during training and is discarded at the inference stage.

3.3. Diversity-Constrained Dynamic Convolution

We propose Diversity-Constrained Dynamic Convolution (DCDC) to adaptively generate factual/counterfactual attention based on the following observations. First, the existing attention module is static, which hinders models from perceiving the sample-wise properties of the sparse silhouette. Second, previous counterfactuals are from predefined distribution, which cannot adaptively represent the confounders of specific samples.

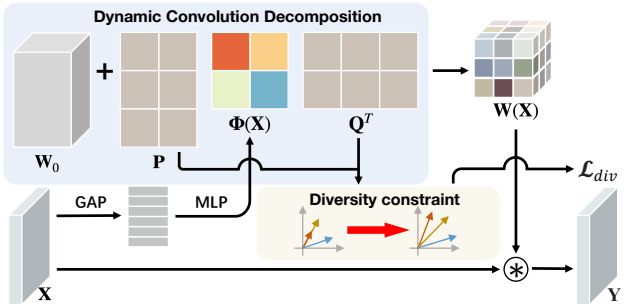


Figure 4. Illustration of Diversity-Constrained Dynamic Convolution. DCDC is formulated as a sample-agnostic convolution $\mathbf{W}_0$ and sample-adaptive one, which could be decomposed into two bases $\mathbf{P}$/$\mathbf{Q}$ and an affinity matrix $\mathbf{\Phi}(\mathbf{X})$. Rank-based diversity constraint on two bases aims to guarantee the representation power.

Vanilla Dynamic Convolution. The main idea of dynamic convolution [57, 67] $\mathbf{W}(\mathbf{X})$ is to linearly combine S static candidate convolutions $\{\mathbf{W}_s\}$ through the score $\{\pi_s(\mathbf{X})\}$ adaptively produced by the SE-style attention [29] as:

$$\mathbf{W}(\mathbf{X}) = \sum_{s=1}^{S} \pi_s(\mathbf{X})\mathbf{W}_s \text{ s.t. } 0 \le \pi_s(\mathbf{X}) \le 1, \sum_{s=1}^{S} \pi_s(\mathbf{X}) = 1. \tag{6}$$

Reformulation with Matrix Decomposition. To avoid the high costs from the high-dimensional computation [15], we reformulate the dynamic convolution with matrix decomposition. First, each candidate convolution $\mathbf{W}_s$ could be re-defined as the combination of a sample-agnostic kernel $\mathbf{W}_0$ and the corresponding offset kernel $\Delta\mathbf{W}_s$, *i.e.*, $\mathbf{W}_s = \mathbf{W}_0 + \Delta\mathbf{W}_s$, where $\mathbf{W}_0 = \frac{1}{S}\sum_{s=1}^{S}\mathbf{W}_s$. Thus, the dynamic convolution $\mathbf{W}(\mathbf{X})$ could be reformulated as:

$$\begin{aligned}\mathbf{W}(\mathbf{X}) &= \sum_{s=1}^{S} \pi_s(\mathbf{X})\mathbf{W}_0 + \sum_{s=1}^{S} \pi_s(\mathbf{X})\Delta\mathbf{W}_s \\ &= \mathbf{W}_0 + \sum_{s=1}^{S} \pi_s(\mathbf{X})\Delta\mathbf{W}_s.\end{aligned} \tag{7}$$

Specifically, $\mathbf{W}_0$ and $\{\pi_s(\mathbf{X})\Delta\mathbf{W}_s\}$ could be regarded as the kernel to extract sample-agnostic features and sample-adaptive features, respectively. Further, we propose to leverage low-rank decomposition on the sample-adaptive kernel to improve the efficiency as follows:

$$\begin{aligned}\mathbf{W}(\mathbf{X}) &= \mathbf{W}_0 + \sum_{i=1}^{L}\sum_{j=1}^{L} \mathbf{p}_i\phi_{i,j}(\mathbf{X})\mathbf{q}_j^T \\ &= \mathbf{W}_0 + \mathbf{P}\mathbf{\Phi}(\mathbf{X})\mathbf{Q}^T,\end{aligned} \tag{8}$$

where $\mathbf{P} \in \mathbb{R}^{C_{out} \times L}$ and $\mathbf{Q} \in \mathbb{R}^{k \times C_{in} \times L}$ are bases to interact the input in low-dimensional latent space $\mathbb{R}^L$. k is the kernel size. $\mathbf{\Phi}(\cdot) \in \mathbb{R}^{L \times L}$ denotes affinity matrix to adaptively interact $\mathbf{P}$ and $\mathbf{Q}$. Therefore, the adaptiveness of dynamic convolution is transformed from the attention-based linear combination to the generative aggregation of two bases. And $\mathbf{\Phi}(\mathbf{X})$ could be generated by an MLP:

$$\mathbf{\Phi}(\mathbf{X}) = \mathbf{W}_{fc2} \times \delta(\mathbf{W}_{fc1} \times (GAP(\mathbf{X}))), \quad (9)$$

where $\mathbf{W}_{fc1} \in \mathbb{R}^{C/r \times C}$ and $\mathbf{W}_{fc2} \in \mathbb{R}^{L^2 \times C/r}$. $\delta(\cdot)$ denotes the Sigmoid. In this way, the decomposition-base dynamic convolution could efficiently reduce the dimension of the latent space from SC to L ($SC \gg L$).

Rank-based Diversity Constraint. To guarantee the representation power, we propose to diversify two bases $\mathbf{P}$ and $\mathbf{Q}$. The diversity of the weight matrix $\mathbf{W} \in \mathbb{R}^{m \times n}$ could be represented by the rank function as:

$$Rank(\mathbf{W}) = \sigma_1^0 + \sigma_2^0 + \cdots + \sigma_r^0 = \lim_{p \to 0} ||\mathbf{W}||_{S_p}^p, \quad (10)$$

where σ_i is the i^{th} singular value of the weight matrix $\mathbf{W}$ and $r = \min\{m, n\}$. The rank function $Rank(\cdot)$ has similar form with Schatten p-norm $||\cdot||_{S_p}^p$ ($p \to 0$) [55, 62], which could be defined as:

$$||\mathbf{W}||_{S_p} = (\sigma_1^p + \sigma_2^p + \cdots + \sigma_r^p)^{1/p}. \quad (11)$$

However, optimizing rank is NP-hard and the Schatten p-norm ($p \neq 1$) is non-convex [66]. Further, Schatten 1-norm (nuclear norm) $||\mathbf{W}||_{S_1}$ has been verified to be a convex approximation [43] to $Rank(\mathbf{W})$ and is differentiable as:

$$\frac{\partial ||\mathbf{W}||_{S_1}}{\partial \mathbf{W}} = \frac{tr(\partial \mathbf{\Sigma})}{\partial \mathbf{W}} = \frac{tr(\mathbf{U}^T \partial(\mathbf{W}) \mathbf{V})}{\partial \mathbf{W}} = \mathbf{U}\mathbf{V}^T, \quad (12)$$

where $\mathbf{W}$ is decomposed into $\mathbf{U}\mathbf{\Sigma}\mathbf{V}^T$ by singular value decomposition (SVD), which introduces nearly no extra computation since the representation is low-dimensional. Thus, we propose to leverage Schatten 1-norm as the diversity constraint to maximize $Rank(\mathbf{W})$:

$$\mathcal{L}_{div} = -\sum_{i \in \{\mathbf{A}, \mathbf{C}\}} ||\mathbf{P}_i||_{S_1} - \sum_{j \in \{\mathbf{A}, \mathbf{C}\}} ||\mathbf{Q}_j||_{S_1}. \quad (13)$$

3.4. Optimization

To effectively optimize GaitGCI, the objective is composed of counterfactual loss $\mathcal{L}_{cf}$, triplet loss $\mathcal{L}_{tri}$ [26], and diversity constraint $\mathcal{L}_{div}$. Specifically, $\mathcal{L}_{cf}$ can be easily implemented with cross-entropy loss by replacing the original prediction $\mathbf{Y}$ with causality effect $\mathbf{Y}_e$.

$$\mathcal{L}_{total} = \underbrace{\mathcal{L}_{ce}(\mathbf{Y}_e, y)}_{Counterfactual\ Loss} + \mathcal{L}_{tri} + \lambda \mathcal{L}_{div}, \quad (14)$$

where y is the ground truth and λ is the weight of diversity constraint, respectively.

4. Experiments

4.1. Dataset

OU-MVLP [53]. It is one of the largest gait datasets, which includes 10307 subjects and each subject contains two sequences. The viewpoints are uniformly distributed between [0°,90°] and [180°,270°]. Following the mainstream protocol [12], the first sequence of each ID is deemed as the gallery, and the rest are the probe during the evaluation.

CASIA-B [69]. CASIA-B contains 124 subjects, and the viewpoints are distributed in [0°, 180°]. Besides, 10 groups of three conditions are included in each subject, *i.e.*, 6 normal (NM), 2 with a bag (BG), and 2 with a coat (CL). For evaluation, we adopt the mainstream protocol [12], which selects the first 74 subjects as the training set and the rest as the test set. During the evaluation, the sequences (NM01-NM04) are the gallery, and the rest are the probe.

GREW [76]. GREW is one of the largest in-the-wild datasets, including 26345 subjects and 128671 sequences. It contains 4 modalities: silhouettes, optical flow, 2D/3D pose. GREW is divided into training set, validation set, and test set, containing 20000, 345, and 6000 subjects, respectively. During the evaluation, each subject contains 2 sequences as the probe and another 2 sequences as the gallery.

Gait3D [75]. Gait3D is the latest in-the-wild dataset containing 4000 subjects and 25309 sequences, which are collected in a large supermarket from 39 cameras. Following the protocol [75], 3000 subjects are selected as the training set, and the rest are the test set. For evaluation, one sequence of each subject is regarded as the query, and the other sequences become the gallery. Further, Gait3D provides 3D annotations to study model-based applications.

4.2. Implementation Details

For common settings, the backbone is composed of 4 3D low-rank convolution layers. In the training stage, the frame number of each sequence is set to 30. The optimizer is Adam (lr=1e-4). The loss weight λ is 0.1. The latent dimension L and reduction ratio r are set to 8 and 4, respectively. During the evaluation, all frames are fed into the framework. More details are in the supplementary material.

For CASIA-B, the channel C of the backbone is set to (32, 64, 128, 128). We train the model for 80k iterations with batch size of (8,8). M is set to 2. For other datasets, the network capacity should be increased [28,42]. We add extra

Table 1. Rank-1 (%) performance comparison on OU-MVLP, excluding the identical-view cases.

Method	Venue	Probe View														Mean
		0°	15°	30°	45°	60°	75°	90°	180°	195°	210°	225°	240°	255°	270°	
GaitSet [12]	AAAI19	79.5	87.9	89.9	90.2	88.1	88.7	87.8	81.7	86.7	89.0	89.3	87.2	87.8	86.2	87.1
GaitPart [19]	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 [28]	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
CSTL [30]	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 [31]	ICCV21	86.1	91.2	92.6	**92.9**	92.2	91.3	91.1	86.9	90.8	**92.2**	92.3	91.3	91.1	90.2	90.9
GaitGL [42]	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
GaitMPL [18]	TIP22	83.9	90.1	91.3	91.5	91.2	90.6	90.1	85.3	89.3	90.7	90.7	90.7	89.8	88.9	89.6
Lagrange [10]	CVPR22	85.9	90.6	91.3	91.5	91.2	91.0	90.6	88.9	89.2	90.5	90.6	89.9	89.8	89.2	90.0
GaitGCI	–	**91.2**	**92.3**	**92.6**	92.7	**93.0**	**92.3**	**92.1**	**92.0**	**91.8**	91.9	**92.6**	**92.3**	**91.4**	**91.6**	**92.1**

Table 2. Rank-1 (%), parameters (M), and computation cost (G MACs) comparison at the inference stage on CASIA-B.

Method	Venue	NM	BG	CL	Mean	Param.	MACs
GaitSet [12]	AAAI19	95.0	87.2	70.4	84.2	2.59	3.27
GaitPart [19]	CVPR20	96.2	91.5	78.7	88.8	1.20	56.96
GLN [28]	ECCV20	96.9	94.0	77.5	89.5	14.70	22.14
MT3D [41]	MM20	96.7	93.0	81.5	90.4	3.20	36.59
CSTL [30]	ICCV21	97.8	93.6	84.2	91.9	9.09	6.43
3DLocal [31]	ICCV21	97.5	94.3	83.7	91.8	4.26	11.20
GaitGL [42]	ICCV21	97.4	94.5	83.6	91.8	2.49	12.62
GaitMPL [18]	TIP22	95.5	92.9	87.9	92.1	–	–
Lagrange [10]	CVPR22	96.9	93.5	86.5	92.3	–	–
GaitGCI-T	–	97.9	95.0	86.4	93.1	1.09	5.41
GaitGCI-M	–	98.2	96.1	87.6	94.0	2.45	12.13
GaitGCI-L	–	**98.4**	**96.6**	**88.5**	**94.5**	4.35	21.54

Table 3. Rank-1 (%), Rank-5 (%), Rank-10 (%), and Rank-20 (%) performance comparison on GREW.

Method	Venue	Rank-1	Rank-5	Rank-10	Rank-20
PoseGait [40]	PR20	0.2	1.1	2.2	4.8
GaitGraph [54]	ICIP21	1.3	3.5	5.1	7.5
GEINet [51]	ICB16	6.8	13.4	17.0	21.0
TS-CNN [65]	TPAMI16	13.6	24.6	30.2	37.0
GaitSet [12]	AAAI19	46.3	63.6	70.3	76.8
GaitPart [19]	CVPR20	44.0	60.7	67.3	73.5
GaitGL [42]	ICCV21	47.3	63.6	69.3	74.2
GaitGCI	–	**68.5**	**80.8**	**84.9**	**87.7**

2 layers with 128 channels. The batch size and M are set to (32,8) and 8, respectively. The iterations are 200k, 200k, and 150k for OU-MVLP, GREW, and Gait3D, respectively.

4.3. Results under in-the-lab Scenario

OU-MVLP. The comparison of Tab. 1 indicates that GaitGCI outperforms previous methods by a considerable margin, which reveals the effectiveness and generalizability of GaitGCI. In detail, GaitGCI achieves the best performance at almost all viewpoints. Specifically, performance at 0°/180° with less information is significantly improved, which may be attributed to reducing the impact of confounders so that the gait pattern is relatively salient.

CASIA-B. The comparison of Tab. 2 demonstrates that GaitGCI could efficiently outperform previous methods. Considering that GaitGCI is lightweight and increasing the number of channels could improve the network's capacity, we design three variants of GaitGCI, *i.e.*, GaitGCI-T, GaitGCI-M, and GaitGCI-L with the channel C, $1.5C$, and $2C$, respectively. Specifically, GaitGCI-T could efficiently achieve 93.1% rank-1 accuracy only with 1.09 M parameters and 5.41 G MACs. Further, GaitGCI-L could achieve 94.5% rank-1 accuracy with acceptable costs. As a trade-off, GaitGCI-M could outperform GaitGL by 2.2% with similar parameters and computation costs. Moreover, GaitGCI greatly improves the performance on BG/CL conditions, which suggests that confounders may hinder the development of existing methods on challenging conditions. The results of each view are in the supplementary material.

4.4. Results under in-the-wild Scenario

GREW. The performance comparison of skeleton-based, GEI-based, and silhouette-based methods on GREW is shown in Tab. 3. Several conclusions could be drawn. First, the performance of the previous methods dramatically deteriorates when migrated to the in-the-wild scenario. Second, silhouette-based methods dominate the single-modality in-the-wild scenarios compared to skeleton/GEI-based methods. Third, GaitGCI significantly outperforms previous methods by over 20% and achieves 3^{rd} in the GREW competition [76] only using silhouette sequences. The results of GREW competition are in the supplementary material.

Gait3D. The comparison on the latest in-the-wild dataset Gait3D is conducted in Tab. 4, including skeleton-based, silhouette-based, and multi-modal methods. GaitGCI outperforms prevailing silhouette-based methods by 14.6% and 13.6% in terms of rank-1 accuracy at the resolution of 128×88 and 64×44, respectively. Besides, the improvement of mAP and mINP fully illustrates the superior retrieval performance of GaitGCI. Further, silhouette-based GaitGCI exceeds SMPLGait [75], which introduces extra 3D SMPL to perform multi-modal learning.

Table 4. Rank-1 (%), Rank-5 (%), mAP (%), and mINP (%) comparison on Gait3D at the resolution of 128×88 and 64×44. As skeleton-based methods are unrelated to the resolution, we only report one group of results. "*" denotes the method with extra 3D modality.

Input Size (H×W)		128×88				64×44			
Methods	Venue	Rank-1	Rank-5	mAP	mINP	Rank-1	Rank-5	mAP	mINP
PoseGait [40]	PR20	0.2	1.1	0.5	0.3	-	-	-	-
GaitGraph [54]	ICIP21	6.3	16.2	5.2	2.4	-	-	-	-
GaitSet [12]	AAAI19	42.6	63.1	33.7	19.7	36.7	58.3	30.0	17.3
GaitPart [19]	CVPR20	29.9	50.6	23.3	13.2	28.2	47.6	21.6	12.4
GLN [28]	ECCV20	42.2	64.5	33.1	19.6	31.4	52.9	24.7	13.6
GaitGL [42]	ICCV21	23.5	38.5	16.4	9.2	29.7	48.5	22.3	13.3
CSTL [30]	ICCV21	12.2	21.7	6.4	3.3	11.7	19.2	5.6	2.6
SMPLGait* [75]	CVPR22	53.2	71.0	42.4	26.0	46.3	64.5	37.2	22.2
GaitGCI	–	**57.2**	**74.5**	**45.0**	**27.6**	**50.3**	**68.5**	**39.5**	**24.3**

Table 5. Ablation on counterfactual intervention learning (CIL) and diversity-constrained dynamic convolution (DCDC), which includes generative factual attention (GFA) and generative counterfactual attention (GCA).

CIL	GFA	GCA	NM	BG	CL	Mean
			96.5	92.9	80.9	90.1
✓			97.1	93.8	84.2	91.7
✓	✓		97.8	94.8	85.2	92.6
✓		✓	97.7	94.5	85.3	92.5
✓	✓	✓	**97.9**	**95.0**	**86.4**	**93.1**

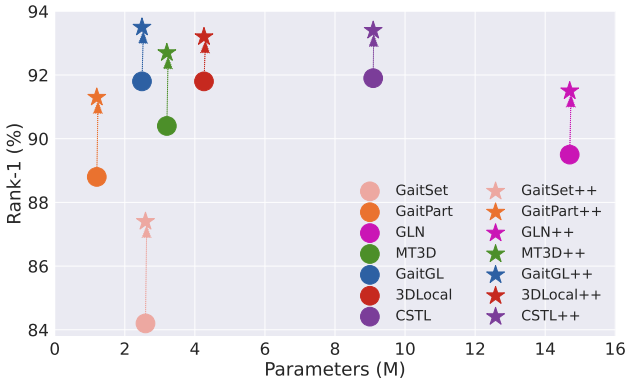


Figure 5. Performance comparison of prevailing methods and those equipped with CIL and DCDC (denoted with suffix '++').

Summary. First, prevailing methods experience a dramatic performance decrease under in-the-wild scenarios, which indicates that the confounders under in-the-wild scenarios are more complex than those under in-the-lab scenarios. Second, the superior performance of GaitGCI under in-the-wild scenarios demonstrates the necessity for alleviating the impact of confounders. Third, although multi-modal methods dominate in-the-wild scenarios, silhouette-based methods have considerable performance improvement potential.

Table 6. Analysis on DCDC. MD and DC denote matrix decomposition and diversity constraint, respectively.

Method	NM	BG	CL	Mean
Static Conv	97.4	94.0	84.8	92.1
DyConv	97.6	94.4	85.8	92.6
+MD	97.7	94.7	85.7	92.7
+MD+DC	**97.9**	**95.0**	**86.4**	**93.1**

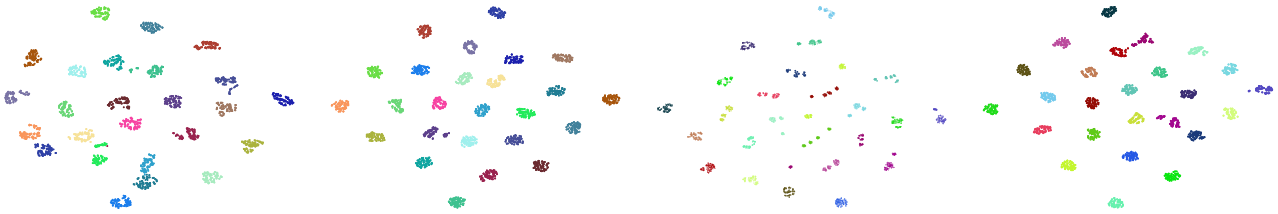
4.5. Ablation Study

In this section, we conduct a series of quantitative and qualitative ablation studies to analyze the effectiveness of GaitGCI and its components. The baseline refers to the backbone with temporal pooling and separate FC [12].

Individual Effectiveness of CIL and DCDC. The individual effects of CIL and DCDC are shown in Tab. 5, where the factual/counterfactual attention of methods without GFA/GCA is set to static convolution and pre-defined normal distribution [11, 49], respectively. CIL effectively improves 1.6% rank-1 accuracy than baseline. Further, generative factual attention and generative counterfactual attention achieve 0.9% and 0.8% performance gain, respectively. And they deliver 1.4% performance improvement in total, indicating the effectiveness and necessity of generative factual/counterfactual attention.

Generalizability of GaitGCI. As a model-agnostic module, CIL and DCDC could be plugged into prevailing methods. As shown in Fig. 5, they could effectively boost the existing methods with nearly no extra costs, which indicates the generalizability and efficiency of CIL and DCDC. Further, this study demonstrates that the confounders may limit the performance of previous silhouette-based methods.

Analysis on DCDC. To evaluate the effectiveness of diversity-constrained dynamic convolution on factual/counterfactual generation, the ablation is conducted in Tab. 6. First, DyConv [15] outperforms static convolu-



(a) Baseline on CASIA-B (b) GaitGCI on CASIA-B (c) Baseline on Gait3D (d) GaitGCI on Gait3D

Figure 6. Comparison of the feature space under in-the-lab and in-the-wild scenarios using t-SNE [56].

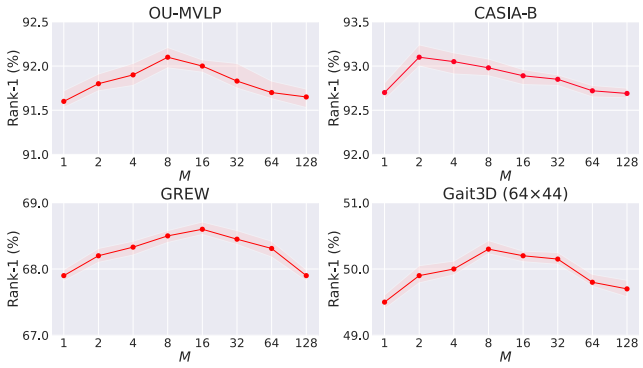


Figure 7. Analysis of attention channel number M.

tion, which indicates the necessity of adaptiveness. Second, matrix decomposition could effectively reduce the computation and parameters while maintaining comparable performance, which demonstrates the redundancy of the high-dimensional computation of dynamic convolution. Third, the rank-based diversity constraint could efficiently improve the representation power.

Analysis on M. The channel number M controls the capacity to perceive sample-wise factual/counterfactual attention. From the results in Fig. 7, we can conclude that: first, the in-the-wild dataset requires larger M, which may be due to the complexity of confounders and the dataset scale; second, the performance rises first and then falls with increasing M, which indicates that larger M brings stronger capacity while superfluous M may lead to overfitting.

Visualization of Network Attention. The visualization with Grad-CAM [50] is shown in Fig. 1. Prevailing methods tend to collapse into confounders while neglecting most regions of the body boundary that could represent gait patterns. By alleviating the impact of confounders, GaitGCI could effectively focus on the discriminative and interpretable regions for gait pattern representation.

Visualization of Feature Space. To qualitatively evaluate the retrieval performance, we visualize the feature space by t-SNE [56] in Fig. 6. First, GaitGCI could improve intra-class compactness and inter-class dispersibility under both scenarios. Second, the feature space of baseline under the in-the-lab scenario tends to have several sub-cluster in each cluster, and this phenomenon is more evident under the in-the-wild scenario, which may indicate the confounders of in-the-wild scenario are more complex. Meanwhile, it may also be why the previous model has acceptable performance under the in-the-lab scenario while the performance drops sharply under the in-the-wild scenario.

5. Conclusion and Limitations

This paper proposes a generative counterfactual intervention learning framework, which could force the network to focus on discriminative and interpretable regions. Counterfactual intervention learning leverages causal inference to analyze the direct causality link between factual attention and prediction. Further, diversity-constrained dynamic convolution, which could adaptively generate factual/counterfactual attention, utilizes matrix decomposition/diversity constraint to guarantee efficiency/representation power, respectively. Extensive experiments prove that GaitGCI could efficiently achieve state-of-the-art performance in arbitrary scenarios and could be used as a plug-and-play module.

For limitations, GaitGCI utilizes SVD, whose costs could only be ignored with low-dimensional feature representation. Besides, channel M is a hyperparameter that depends on the dataset. In future work with high-dimensional representation and multi-dataset scenarios, we could alleviate these issues with numerical iteration methods [5, 14] and attention-based channel selection, respectively.

Acknowledgements

This work is supported in part by National Natural Science Foundation of China under Grant U20A20222, National Science Foundation for Distinguished Young Scholars under Grant 62225605, National Key Research and Development Program of China under Grant 2020AAA0107400, Zhejiang – Singapore Innovation and AI Joint Research Lab, Ant Group through CCF-Ant Research Fund, and sponsored by CCF-AFSG Research Fund, CAAI-HUAWEI MindSpore Open Fund as well as CCF-Zhipu AI Large Model Fund(CCF-Zhipu202302).

References

[1] Ehsan Abbasnejad, Damien Teney, Amin Parvaneh, Javen Shi, and Anton van den Hengel. Counterfactual vision and language learning. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 10044–10054, 2020. 3

[2] Iasonas Kokkinos Adam W Harley, Konstantinos G. Derpanis. Segmentation-aware convolutional networks using local attention masks. In *Int. Conf. Comput. Vis.*, 2017. 3

[3] G. Ariyanto and M. S. Nixon. Model-based 3d gait biometrics. In *Int. Joint Conf. Bio.*, pages 1–7, 2011. 2

[4] Shutao Bai, Bingpeng Ma, Hong Chang, Rui Huang, and Xilin Chen. Salient-to-broad transition for video person re-identification. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 7339–7348, June 2022. 1

[5] Nitin Bansal, Xiaohan Chen, and Zhangyang Wang. Can we gain more from orthogonality regularizations in training deep networks? *Adv. Neural Inform. Process. Syst.*, 31, 2018. 8

[6] Robert Bodor, Andrew Drenner, Duc Fehr, Osama Masoud, and Nikolaos Papanikolopoulos. View-independent human motion classification using image-based reconstruction. *Int. Video Conf.*, pages 1194–1206, 2009. 2

[7] N. V. Boulgouris and Z. X. Chi. Gait recognition based on human body components. In *IEEE Int. Conf. Image Process.*, pages 353–356, 2007. 2

[8] Francisco M Castro, Manuel J Marin-Jimenez, Nicolás Guil, and Nicolás Pérez de la Blanca. Multimodal feature fusion for cnn-based gait recognition: an empirical comparison. *Neural Comput. Appl.*, 32(17):14173–14193, 2020. 2

[9] Francisco M. Castro, Manuel J. Marin-Jimenez, Nicolas Guil, Santiago Lopez-Tapia, and Nicolas Perez de la Blanca. Evaluation of cnn architectures for gait recognition based on optical flow maps. In *Int. Conf. of the Bio. Special Interest Group*, pages 1–5, 2017. 2

[10] Tianrui Chai, Annan Li, Shaoxiong Zhang, Zilong Li, and Yunhong Wang. Lagrange motion analysis and view embeddings for improved gait recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 20249–20258, June 2022. 6

[11] Chun-Hao Chang, George Alexandru Adam, and Anna Goldenberg. Towards robust classification model by counterfactual and invariant data generation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 15212–15221, 2021. 2, 7

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

[13] Jin Chen, Xijun Wang, Zichao Guo, X. Zhang, and Jian Sun. Dynamic region-aware convolution. *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 8060–8069, 2021. 3

[14] Tianlong Chen, Shaojin Ding, Jingyi Xie, Ye Yuan, Wuyang Chen, Yang Yang, Zhou Ren, and Zhangyang Wang. Abd-net: Attentive but diverse person re-identification. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 8351–8361, 2019. 8

[15] Yinpeng Chen, Xiyang Dai, Mengchen Liu, Dongdong Chen, Lu Yuan, and Zicheng Liu. Dynamic convolution: Attention over convolution kernels. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 11030–11039, 2020. 4, 7

[16] Zhang Dong, Zhang Hanwang, Tang Jinhui, Hua Xiansheng, and Sun Qianru. Causal intervention for weakly supervised semantic segmentation. In *Adv. Neural Inform. Process. Syst.*, 2020. 3

[17] Huanzhang Dou, Pengyi Zhang, Wei Su, Yunlong Yu, and Xi Li. Metagait: Learning to learn an omni sample adaptive representation for gait recognition. In *Eur. Conf. Comput. Vis.*, pages 357–374, 2022. 2

[18] Huanzhang Dou, Pengyi Zhang, Yuhan Zhao, Lin Dong, Zequn Qin, and Xi Li. Gaitmpl: Gait recognition with memory-augmented progressive learning. *IEEE Trans. Image Process.*, 2022. 6

[19] Chao Fan, Yunjie Peng, Chunshui Cao, Xu Liu, Saihui Hou, Jiannan Chi, Yongzhen Huang, Qing Li, and Zhiqiang He. Gaitpart: Temporal part-based model for gait recognition. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2020. 2, 6, 7

[20] Amy Fire and Song-Chun Zhu. Inferring hidden statuses and actions in video by causal reasoning. In *IEEE Conf. Comput. Vis. Pattern Recog. Worksh.*, July 2017. 2

[21] Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, 2020. 1

[22] Yash Goyal, Ziyan Wu, Jan Ernst, Dhruv Batra, Devi Parikh, and Stefan Lee. Counterfactual visual explanations. In *Int. Conf. Mach. Learn.*, volume 97, pages 2376–2384, 2019. 3

[23] Guoying Zhao, Guoyi Liu, Hua Li, and M. Pietikainen. 3d gait recognition using multiple cameras. In *Int. Conf. Autom. Face Gesture Recog.*, pages 529–534, 2006. 2

[24] Yizeng Han, Gao Huang, Shiji Song, Le Yang, Honghui Wang, and Yulin Wang. Dynamic neural networks: A survey. *IEEE Trans. Pattern Anal. Mach. Intell.*, 44:7436–7456, 2022. 3

[25] Lisa Anne Hendricks, Zeynep Akata, Marcus Rohrbach, Jeff Donahue, Bernt Schiele, and Trevor Darrell. Generating visual explanations. *Eur. Conf. Comput. Vis.*, 2016. 3

[26] Alexander Hermans, Lucas Beyer, and Bastian Leibe. In defense of the triplet loss for person re-identification. *arXiv preprint arXiv:1703.07737*, 2017. 5

[27] Martin Hofmann, Jürgen Geiger, Sebastian Bachmann, Björn Schuller, and Gerhard Rigoll. The tum gait from audio, image and depth (gaid) database: Multimodal recognition of subjects and traits. *J. Vis. Commun. Image. Represent.*, 25(1):195–206, 2014. 2

[28] Saihui Hou, Chunshui Cao, Xu Liu, and Yongzhen Huang. Gait lateral network: Learning discriminative and compact representations for gait recognition. In *Eur. Conf. Comput. Vis.*, pages 382–398, 2020. 5, 6, 7

[29] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 7132–7141, 2018. 3, 4

[30] Xiaohu Huang, Duowang Zhu, Hao Wang, Xinggang Wang, Bo Yang, Botao He, Wenyu Liu, and Bin Feng. Context-sensitive temporal feature learning for gait recognition. In *Int. Conf. Comput. Vis.*, pages 12909–12918, October 2021. 1, 6, 7

[31] Zhen Huang, Dixiu Xue, Xu Shen, Xinmei Tian, Houqiang Li, Jianqiang Huang, and Xian-Sheng Hua. 3d local convolutional neural networks for gait recognition. In *Int. Conf. Comput. Vis.*, pages 14920–14929, October 2021. 2, 6

[32] Pavel Izmailov, Polina Kirichenko, Nate Gruver, and Andrew Gordon Wilson. On feature learning in the presence of spurious correlations. *Adv. Neural Inform. Process. Syst.*, 2022. 1

[33] Dimitris Kastaniotis, Ilias Theodorakopoulos, and Spiros Fotopoulos. Pose-based gait recognition with local gradient descriptors and hierarchically aggregated residuals. *J. Electron. Imaging*, 25(6):063019, 2016. 2

[34] Saeed Khorram and Li Fuxin. Cycle-consistent counterfactuals by latent transformations. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 10203–10212, 2022. 3

[35] Thomas N. Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. In *Int. Conf. Learn. Represent.*, 2017. 2

[36] Matt J Kusner, Joshua Loftus, Chris Russell, and Ricardo Silva. Counterfactual fairness. *Adv. Neural Inform. Process. Syst.*, 30, 2017. 3

[37] Xiang Li, Yasushi Makihara, Chi Xu, Yasushi Yagi, Shiqi Yu, and Mingwu Ren. End-to-end model-based gait recognition. In *ACCV*, 2020. 2

[38] Yunsheng Li, Yinpeng Chen, Xiyang Dai, mengchen liu, Dongdong Chen, Ye Yu, Lu Yuan, Zicheng Liu, Mei Chen, and Nuno Vasconcelos. Revisiting dynamic convolution via matrix decomposition. In *Int. Conf. Learn. Represent.*, 2021. 3

[39] Junhao Liang, Chao Fan, Saihui Hou, Chuanfu Shen, Yongzhen Huang, and Shiqi Yu. Gaitedge: Beyond plain end-to-end gait recognition for better practicality. *Eur. Conf. Comput. Vis.*, 2022. 1, 2

[40] Rijun Liao, Shiqi Yu, Weizhi An, and Yongzhen Huang. A model-based gait recognition method with body pose and human prior knowledge. *Pattern Recog.*, 98:107069, 2020. 2, 6, 7

[41] Beibei Lin, Shunli Zhang, and Feng Bao. Gait recognition with multiple-temporal-scale 3d convolutional neural network. In *ACM Int. Conf. Multimedia*, pages 3054–3062, 2020. 6

[42] Beibei Lin, Shunli Zhang, and Xin Yu. Gait recognition via effective global-local feature representation and local temporal aggregation. In *Int. Conf. Comput. Vis.*, pages 14648–14656, October 2021. 2, 5, 6, 7

[43] Canyi Lu, Jinhui Tang, Shuicheng Yan, and Zhouchen Lin. Nonconvex nonsmooth low rank minimization via iteratively reweighted nuclear norm. *IEEE Trans. Image Process.*, 25(2):829–839, 2015. 5

[44] João Ferreira Nunes, Pedro Miguel Moreira, and João Manuel RS Tavares. Benchmark rgb-d gait datasets: A systematic review. In *ECCOMAS Thematic Conf. on Comput. Vis. and Med. Image. Process.*, pages 366–372. Springer, 2019. 2

[45] Nick Pawlowski, Daniel Coelho de Castro, and Ben Glocker. Deep structural causal models for tractable counterfactual inference. *Adv. Neural Inform. Process. Syst.*, 33:857–869, 2020. 3

[46] Judea Pearl. Direct and indirect effects. In *Conference on Uncertainty in Artificial Intelligence*, pages 373–392, 2001. 4

[47] Judea Pearl et al. Models, reasoning and inference. *Cambridge, UK: CambridgeUniversityPress*, 19(2), 2000. 2, 3

[48] Vasili Ramanishka, Yi-Ting Chen, Teruhisa Misu, and Kate Saenko. Toward driving scene understanding: A dataset for learning driver behavior and causal reasoning. In *IEEE Conf. Comput. Vis. Pattern Recog.*, June 2018. 2

[49] Yongming Rao, Guangyi Chen, Jiwen Lu, and Jie Zhou. Counterfactual attention learning for fine-grained visual categorization and re-identification. In *Int. Conf. Comput. Vis.*, pages 1005–1014, 2021. 2, 7

[50] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Int. Conf. Comput. Vis.*, pages 618–626, 2017. 8

[51] Kohei Shiraga, Yasushi Makihara, Daigo Muramatsu, Tomio Echigo, and Yasushi Yagi. Geinet: View-invariant gait recognition using a convolutional neural network. In *2016 International Conference on Biometrics (ICB)*, pages 1–8, 2016. 6

[52] Hang Su, Varun Jampani, Deqing Sun, Orazio Gallo, Erik Learned-Miller, and Jan Kautz. Pixel-adaptive convolutional neural networks. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 11166–11175, 2019. 3

[53] Noriko Takemura, Yasushi Makihara, Daigo Muramatsu, Tomio Echigo, and Yasushi Yagi. Multi-view large population gait dataset and its performance evaluation for cross-view gait recognition. *IPSJ Trans. Comput. Vis. Appl.*, 10(1):4, 2018. 1, 2, 5

[54] Torben Teepe, Ali Khan, Johannes Gilg, Fabian Herzog, Stefan Hörmann, and Gerhard Rigoll. Gaitgraph: Graph convolutional network for skeleton-based gait recognition. In *IEEE Int. Conf. Image Process.*, pages 2314–2318, 2021. 2, 6, 7

[55] Ryota Tomioka and Taiji Suzuki. Convex tensor decomposition via structured schatten norm regularization. *Adv. Neural Inform. Process. Syst.*, 26, 2013. 5

[56] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *J. Mach. Learn. Res.*, 9(11), 2008. 8

[57] Thomas Verelst and Tinne Tuytelaars. Dynamic convolutions: Exploiting spatial sparsity for faster inference. In *IEEE Conf. Comput. Vis. Pattern Recog.*, June 2020. 2, 3, 4

[58] Sahil Verma, John Dickerson, and Keegan Hines. Counterfactual explanations for machine learning: A review. *arXiv preprint arXiv:2010.10596*, 2020. 3

[59] Chen Wang, Junping Zhang, Jian Pu, Xiaoru Yuan, and Liang Wang. Chrono-gait image: A novel temporal template for gait recognition. In *Eur. Conf. Comput. Vis.*, pages 257–270, 2010. 1

[60] Liang Wang, Huazhong Ning, Weiming Hu, and Tieniu Tan. Gait recognition based on procrustes shape analysis. In *IEEE Int. Conf. Image Process.*, volume 3, pages III–III, 2002. 1

[61] Liang Wang, Huazhong Ning, Tieniu Tan, and Weiming Hu. Fusion of static and dynamic body biometrics for gait recognition. *IEEE TCSVT*, 14(2):149–158, 2004. 2

[62] Qianqian Wang, Fang Chen, Quanxue Gao, Xinbo Gao, and Feiping Nie. On the schatten norm for matrix based subspace learning and classification. *Neurocomputing*, 216:192–199, 2016. 5

[63] Tan Wang, Jianqiang Huang, Hanwang Zhang, and Qianru Sun. Visual commonsense representation learning via causal inference. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 378–379, 2020. 2

[64] Haoqian Wu, Jian Tian, Yongjian Fu, Bin Li, and Xi Li. Condition-aware comparison scheme for gait recognition. *IEEE Trans. Image Process.*, 30:2734–2744, 2021. 1

[65] Zifeng Wu, Yongzhen Huang, Liang Wang, Xiaogang Wang, and Tieniu Tan. A comprehensive study on cross-view gait based human identification with deep cnns. *IEEE Trans. Pattern Anal. Mach. Intell.*, 39(2):209–226, 2016. 6

[66] Chen Xu, Zhouchen Lin, and Hongbin Zha. A unified convex surrogate for the schatten-p norm. In *AAAI*, 2017. 5

[67] Brandon Yang, Gabriel Bender, Quoc V Le, and Jiquan Ngiam. Condconv: Conditionally parameterized convolutions for efficient inference. In *Adv. Neural Inform. Process. Syst.*, volume 32, 2019. 2, 3, 4

[68] Wenjie Yang, Houjing Huang, Zhang Zhang, Xiaotang Chen, Kaiqi Huang, and Shu Zhang. Towards rich feature discovery with class activation maps augmentation for person re-identification. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 1389–1398, 2019. 1

[69] Shiqi Yu, Daoliang Tan, and Tieniu Tan. A framework for evaluating the effect of view angle, clothing and carrying condition on gait recognition. In *Int. Conf. Pattern Recog.*, pages 441–444, 2006. 1, 2, 5

[70] Zhongqi Yue, Tan Wang, Hanwang Zhang, Qianru Sun, and Xiansheng Hua. Counterfactual zero-shot and open-set visual recognition. *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 15399–15409, 2021. 3

[71] Zhongqi Yue, Hanwang Zhang, Qianru Sun, and Xian-Sheng Hua. Interventional few-shot learning. In *Adv. Neural Inform. Process. Syst.*, 2020. 3

[72] Pengyi Zhang, Huanzhang Dou, Yunlong Yu, and Xi Li. Adaptive cross-domain learning for generalizable person re-identification. In *Eur. Conf. Comput. Vis.*, pages 215–232, 2022. 3

[73] Pengyi Zhang, Huanzhang Dou, Wenhu Zhang, Yuhan Zhao, Zequn Qin, Dongping Hu, Yi Fang, and Xi Li. A large-scale synthetic gait dataset towards in-the-wild simulation and comparison study. *ACM Trans. Multimedia Comput. Commun. Appl.*, 19(1):1–23, 2023. 1, 2

[74] Ziyuan Zhang, Luan Tran, Feng Liu, and Xiaoming Liu. On learning disentangled representations for gait recognition. *IEEE Trans. Pattern Anal. Mach. Intell.*, 44:345–360, 2022. 2

[75] Jinkai Zheng, Xinchen Liu, Wu Liu, Lingxiao He, Chenggang Yan, and Tao Mei. Gait recognition in the wild with dense 3d representations and a benchmark. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 20228–20237, June 2022. 2, 5, 6, 7

[76] Zheng Zhu, Xianda Guo, Tian Yang, Junjie Huang, Jiankang Deng, Guan Huang, Dalong Du, Jiwen Lu, and Jie Zhou. Gait recognition in the wild: A benchmark. In *Int. Conf. Comput. Vis.*, pages 14789–14799, 2021. 1, 2, 5, 6