Learning A Unified Template for Gait Recognition

Panjian Huang[1], Saihui Hou[1], Junzhou Huang[2], Yongzhen Huang[1,3]*
[1] School of Artificial Intelligence, Beijing Normal University
[2] Department of Computer Science and Engineering, The University of Texas at Arlington
[3] WATRIX.AI

Abstract

"What I cannot create, I do not understand." *Human wisdom reveals that creation is one of the highest forms of learning. For example, Diffusion Models have demonstrated remarkable semantic structure and memory in image generation, understanding, and restoration, which intuitively benefits representation learning. However, current gait networks rarely embrace this perspective, relying primarily on learning by contrasting gait samples under varying complex conditions, leading to semantic inconsistency and uniformity issues. To address these issues, we propose* ***Origins*** *with generative capabilities whose underlying philosophy is that* ***different entities are generated from a unified template****, inherently regularizing gait representations within a consistent and diverse semantic space to capture accurate gait differences. Admittedly, learning this unified template is exceedingly challenging, as it requires the comprehensiveness of the template to encompass gait representations with various conditions. Inspired by Diffusion Models, Origins diffuses the unified template into* ***timestep templates*** *for gait generative learning, and meanwhile transfers the unified template for gait representation learning. Especially, gait generative and representation learning serve as a unified framework for end-to-end joint training. Extensive experiments on CASIA-B, CCPG, SUSTech1K, Gait3D, GREW and CCGR-MINI demonstrate that Origins performs unified generative and representation learning, achieving superior performance.*

1. Introduction

> *"The Tao produced One; One produced Two; Two produced Three; Three produced All things."*
>
> — *Laozi, Tao Te Ching, ch. 42*

According to the associative theory of creativity, individuals with higher creativity possess richer semantic structure and memory that support an expansive associative search, facilitating the combination of distant concepts into novel ideas [1, 27, 36]. Indeed, modern Generative AI (*e.g.*, Diffusion Models) has demonstrated remarkable visual control, memory and imagination in image generation [20], understanding [26], and restoration [61]. Intuitively, a network, capable of controllably generating diverse entities across different categories, implies a powerful representation space [5, 28, 49, 62, 67].

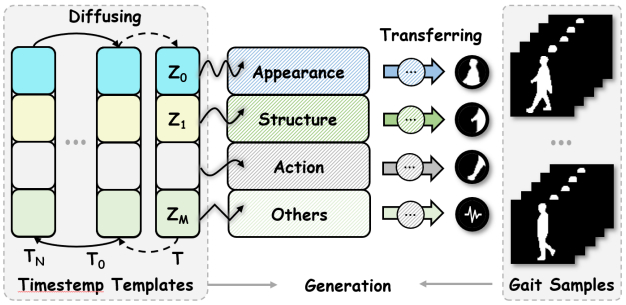


Figure 1. Origins performs unified generative and representation learning. The unified template $\mathcal{T}$ diffuses to the timestep templates, which generates gait representations. Meanwhile, the unified template $\mathcal{T}$ transfers to capture accurate gait differences.

Within the scope of representation learning, gait serves as a descriptor of walking patterns for long-distance human recognition [44, 46]. Although current gait recognition has made significant progress, the gait representation learning obtains identity information by contrasting gait samples under varying complex conditions (*e.g.*, cross-view, cross-clothing, occlusion and illumination conditions), causing two problems: **(i) Semantic Inconsistency.** Gait representations with different complex conditions may exhibit significant gaps. For example, semantic inconsistency has been observed in VPNet [35], where different view angles correspond to distinct prompts, indicating significant variations in the current gait representations. Additionally, cross-view gait networks potentially learn the 3D human body projections from various 2D view angles and the chaining relations between different view angles [34], while cross-

*Corresponding Author

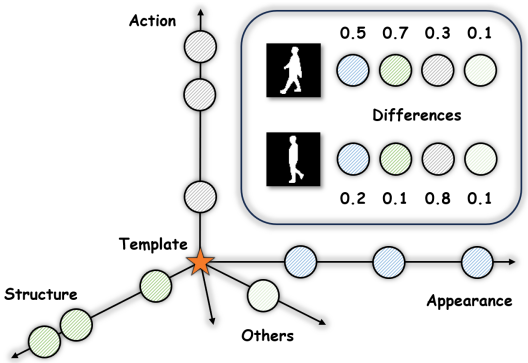


Figure 2. The unified template is the consistent and diverse space. The different semantic dimensions in this unified template linearly construct each gait representation.

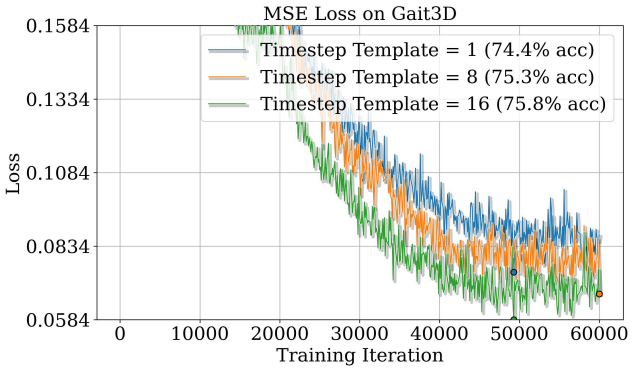


Figure 3. More timestep templates enable better convergence and more precise recognition performance.

clothing gait networks may learn pose and motion information [24]. Occluded gait networks may implicitly learn reasoning and filtering capabilities [25]. **(ii) Semantic Uniformity.** Facing the combination of multiple covariates, gait representations are constrained to a narrow representation space, maintaining usable patterns for all conditions.

To address the above issues, we present Origins where **different entities are generated from a unified template**. The philosophy is inspired by ***Tao Te Ching***, where All things are said to originate from the "Tao". Towards this goal, gait representations exhibit **Phenomena 1** (Figure 2). Gait samples with different conditions align within the semantic space. **Phenomena 2** (Figure 4). The semantic space possesses sufficient diversity to encompass all gait samples with various conditions.

Admittedly, learning this unified template is exceedingly challenging. Aligning and generating highly different gait samples with the template distribution encounters the convergence difficulty. As shown in Figure 1, Origins diffuses a unified template (*i.e.*, $\mathcal{T}$) into timestep templates (*i.e.*, $\mathcal{T}_0, ..., \mathcal{T}_N$), which generates gait representations, and meanwhile transfers the unified template information ($\mathcal{T}$) to capture accurate gait differences. Figure 3 illustrates that more timestep templates enable better convergence and more precise recognition performance. We specifically clarify relations between Origins and Diffusion Models:

Inspirations. Origins is inspired by diffusion models, lies in: (i) The step-by-step diffusion mechanism. In the vanilla diffusion training, for each sample, a random timestep is chosen, mapped through the time series and MLP, and conditioned for generation. Analogously, in the Origins training, each gait sample randomly selects a timestep template, which is transformed from the unified template through the time series and MLP, and conditioned to generate gait representations; (ii) A consistent and diverse semantic space. The generative paradigm needs to generate every gait sample independent of cross-covariate conditions, which forces the unified template to span the entire semantic manifold.

Highlights. Origins aims to recognize individuals with generative capabilities. (i) No noise addition. Origins integrates the generative process to learn a consistent yet diverse semantic space without sampling new gait samples from a known noise distribution (*e.g.*, Gaussian distribution). (ii) Implicit timestep template relations. Origins does not follow a conventional diffusion model where timestep relations are explicitly equivalent to the Markov chain. Instead, it implicitly learns the timestep template relations by randomly sampling to generate gait representations. (iii) No identity constraint. The generative evolution aims to learn a consistent and diverse space (*i.e.*, a unified template), independent of identity retention, where Origins only adopt the unified template and real samples for identification.

Our main contributions can be summarized as follows:

- We propose a novel framework Origins where the unified generative and representation learning regularizes gait representations within a consistent and diverse semantic space, addressing the semantic inconsistency and uniformity issues.
- We design Diffusing Timestep Templates to alleviate the convergence difficulty, and Transferring Unified Template to capture accurate gait differences.
- We evaluate Origins on six public benchmarks, CASIA-B, CCPG, SUSTech1K, Gait3D, GREW and CCGR-MINI, demonstrating the effectiveness and achieving superior performance.

2. Related Work

2.1. Diffusion Models and Representation Learning

Diffusion Models (DMs). Diffusion models gradually add noise to input data and learn to reverse this process for data generation, making them a probabilistic generative framework. (i) Architectures. DDPMs [20] leverages a U-Net architecture for denoising, which enables effective noise pre-

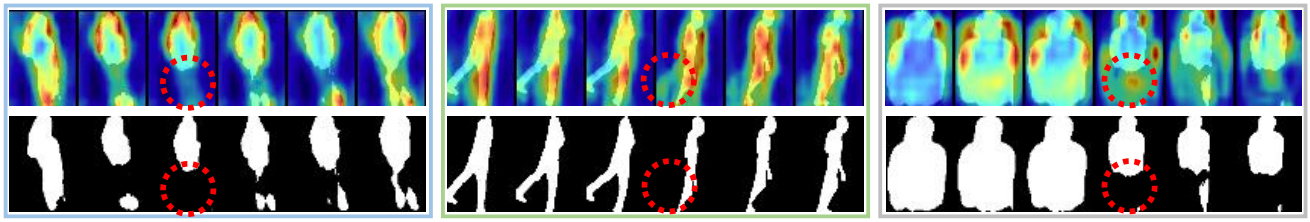


Figure 4. The heatmaps show that the unified template generates the missing human body parts for information supplementation and matching (*e.g.*, the regions in the red circle of the left and right figures), and rectifies temporal information (*e.g.*, the region in the red circle of the middle figure).

diction across varying timesteps. LDMs [42] apply diffusion processes in the latent space of a pre-trained variational autoencoder. This innovation significantly reduces computational overhead while maintaining high generation quality. DiT [38] adapts Vision Transformers (ViTs) by tokenizing inputs and attention-based mechanisms for more flexibility. (ii) Guidance Techniques. Classifier Guidance [7] uses gradients from pre-trained noise-robust classifiers to guide the diffusion process. Classifier-Free Guidance [19] avoids reliance on external classifiers by training a diffusion model with both conditional and unconditional inputs. Self-Guided Diffusion [23] introduces a self-supervised framework that generates guidance signals via clustering.

Representation Learning Based on DMs. Representation learning based on DMs can be broadly categorized into five types. (i) Leveraging Intermediate Activations. DDPM-Seg [3] extracts intermediate feature activations from decoder blocks of DDPMs for semantic segmentation. (ii) Knowledge Transfer. RepFusion [64] employs reinforcement learning to dynamically extract representations from diffusion models, which are distilled into student networks for downstream tasks. (iii) Reconstructing Diffusion Models. l-DAE [5] reconstructs DDPMs into autoencoder for self-supervised learning, highlighting the role of denoising in representation learning. (iv) Generative Augmentation. GAM [2] employs latent diffusion models to create augmented views of training data, enhancing the generalization of learned representations across diverse datasets. (v) Joint Diffusion Models. HybirdViT [65] and ADDP [49] combine generative and discriminative objectives in a single model, improving performances in both. Origins aims to construct a consistent and diverse gait representation space with generative capabilities, which falls under this scope.

2.2. Gait Recognition

Model-Based Gait Recognition. These methods focus on human structure representations. PoseGait [31] uses 3D human body pose features, including joint angles, limb lengths, and motion patterns, to enhance gait robustness. GaitGraph [47] and GaitGraph2 [48] employ Graph Convolutional Networks (GCNs) to model robust spatio-temporal information. GaitTR [68] and GaitMixer [41] incorporate self-attention mechanisms to capture long-range spatial correlations. GPGait [13] proposes a generalized pose-based gait framework, improving cross-domain generalization by transforming pose data into a unified representation. SMPLGait [69] introduces the SMPL model to integrate dense 3D mesh representations. SkeletonGait [11], HiH [56], and GaitHeat [14] introduce gaussian-approximated skeleton maps for structural analysis and shape details.

Appearance-Based Gait Recognition. These methods primarily employ human shape representations. GaitSet [4] proposes a set-based method with a flexible, permutation-invariant framework. GaitPart [9] introduces a temporal part-based framework with fine-grained body-part motion. GaitGL [32] combines global and local features with 3D CNNs to capture fine-grained temporal-spatial patterns. GaitBase [10] proposes a simple yet robust baseline for the real-world applications. DANet [33], DyGait [57], HSTL [55], VPNet [35], GLGait [39] and GaitMoE [25] focus on dynamic local-global gait representations. GaitGCI [8], GaitCSV [52], CLTD [63], GaitC^{3}I [54], QAGait [60], Free Lunch [53] and GaitAttack [22] address confounders and noises with the interpretability. In addition, there are some research with other gait modalities, such as GaitEdge [30] with RGB data; GaitParsing [58], LandmarkGait [59] and ParsingGait [70] with parsing information; LidarGait [45] with point clouds; MMGaitFormer [6] and CL-Gait [16] with multimodal data.

3. Methodology

In this section, we first introduce the gait network with Origins in Sec. 3.1, then present the overview of Origins in Sec. 3.2, and finally, describe the unified gait generative and representation learning as a unified framework for end-to-end joint training in Sec. 3.3.

3.1. Overview

As shown in Figure 5, the vanilla gait framework typically consists of a Visual Encoder ($\mathcal{E}$), a Horizontal Partition ($\mathcal{HP}$), a Recognition Head ($\mathcal{RH}$), and a Joint Loss ($\mathcal{L}$). In this work, the Visual Encoder employs a Stem and

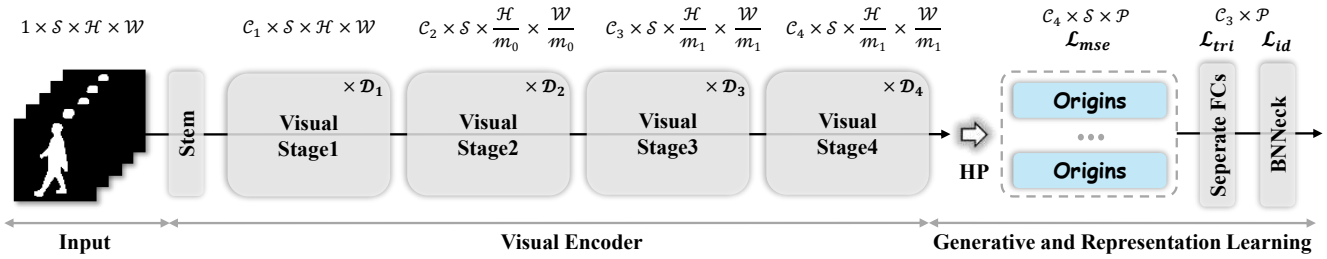


Figure 5. The gait network with Origins. HP represents Horizontal Partition, and Visual Stage consists of basic convolution blocks. After the gait silhouette sequence passes through Visual Encoder and HP, the part representations are fed into the respective Origins and Recognition Head for generative and representation learning.

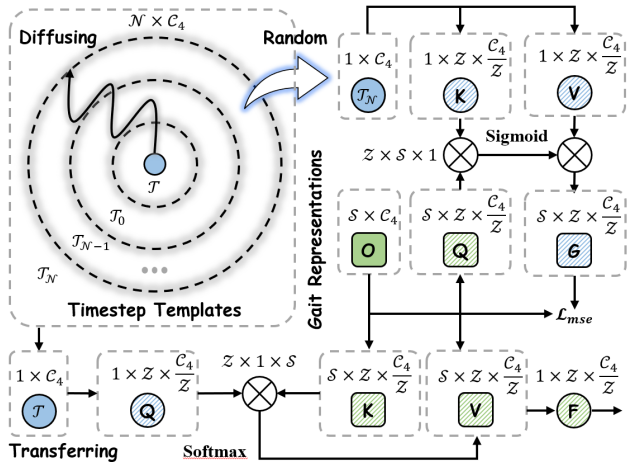


Figure 6. The overview of Origins, best viewed in colors and shapes. The $\mathcal{N}$, $\mathcal{C}_4$ and $\mathcal{Z}$ denote the number, dimension and multi-heads of timestep templates. The unified template $\mathcal{T}$ diffuses timestep templates by time serialization for generation and transfers information to gait representations for recognition.

several Visual Stages (as in ResNet [17]). Notably, we introduce the general backbone rather than sophisticated designs for better validating the effectiveness of generative learning. The $\mathcal{HP}$ horizontally splits the human body ($\mathcal{P}$), extracting and matching finer-grained identity information. Finally, the $\mathcal{RH}$ performs feature mapping for optimization with Joint Loss (*i.e.*, Triplet Loss [18] and Cross-Entropy Loss [10]). The proposed Origins is embedded after the $\mathcal{HP}$ and individually on each human part, regularizing the part representations into a consistent and diverse semantic space. ***Here, we omit the part index for simplicity.*** Formally, given the gait silhouette sequence $\mathcal{X} \in \mathbb{R}^{1\times\mathcal{S}\times\mathcal{H}\times\mathcal{W}}$, where $1, \mathcal{S}, \mathcal{H}, \mathcal{W}$ represent channel, consecutive $\mathcal{S}$ frames, height and width dimensions, the process of Visual Encoder and Horizontal Partition is formulated as follows:

$$\mathcal{O} = \mathcal{HP}(\mathcal{E}(\mathcal{X})) \tag{1}$$

where $\mathcal{O} \in \mathbb{R}^{\mathcal{C}_4\times\mathcal{S}\times\mathcal{P}}$ is the part representations. Similar to [42], generative learning in the latent space is generally more efficient and flexible than in the pixel space.

3.2. Origins

Existing gait paradigms primarily rely on learning the invariant gait representations by contrasting gait samples with different conditions. However, the complex real-world environment (*e.g.*, cross-view and cross-clothing conditions) inevitably causes semantic inconsistency and uniformity (*e.g.*, the view angle chaining relations vs. the motion information). To this end, we propose Origins with the generative capability to regularize gait representations within a consistent and diverse semantic space to capture accurate gait differences. As shown in Figure. 6, Origins presents the semantic consistency as prior, applying a unified template to generate gait representations for the entire gait database, implying that each gait sequence can be constructed within the space spanned by this unified template.

Diffusing Timestep Templates. Admittedly, learning this unified template is extremely challenging, as it requires the comprehensiveness of the template to encompass gait representations with various conditions, facing the convergence difficulty as many generative models (*e.g.*, GANs [15] and VQVAEs [50]). Inspired by the step-by-step mechanism in Diffusion Models [7, 37, 42], Origins diffuses this unified template into timestep templates, where each is derived from the unified template through time series modeling and MLP. Formally, Given the learnable unified template $\mathcal{T} \in \mathbb{R}^{1\times\mathcal{C}_4}$, the diffusion process for timestep templates is as follows:

$$\mathcal{T}_\mathcal{N} = \mathcal{T} + \mathrm{MLP}(t_\mathcal{N}) \tag{2}$$

$$t_\mathcal{N} = \big[\cos(\omega_1\mathcal{N}), \ldots, \cos(\omega_{\frac{\mathcal{C}_4}{2}}\mathcal{N}), \sin(\omega_1\mathcal{N}), \ldots, \sin(\omega_{\frac{\mathcal{C}_4}{2}}\mathcal{N})\big] \tag{3}$$

$$\omega_k = 10000^{-\frac{2(k-1)}{\mathcal{C}_4}}, \qquad k = 1, \ldots, \tfrac{\mathcal{C}_4}{2} \tag{4}$$

where $\mathcal{T}_\mathcal{N} \in \mathbb{R}^{1\times\mathcal{C}_4}$, $\mathcal{N}$ is the scalar timestep, ω_k defines a frequency schedule. Similar to the training stage

Table 1. DataBase and Architectures. Id. and Seq. denote the number of identities and sequences. CV, BG and CL refer to cross-view and carrying bags and cross-clothing conditions. $\mathcal{D}$ and $\mathcal{C}$ denote the number of conv blocks and the channels in each visual stage.

Environment	Dataset	Train		Test		Condition	Stage	Channels	Strides
		Id.	Seq.	Id.	Seq.		$[\mathcal{D}_1, \mathcal{D}_2, \mathcal{D}_3, \mathcal{D}_4]$	$[\mathcal{C}_1, \mathcal{C}_2, \mathcal{C}_3, \mathcal{C}_4]$	
Constrained	CASIA-B [66]	74	8,140	50	5,500	CV, BG, CL	[1, 1, 1. -]	[64, 128, 256, -]	[1, 2, 1, -]
	CCPG [29]	100	8,187	100	8,095	CV, BG, CL	[1, 1, 1, 1]	[64, 128, 256, 512]	[1, 2, 2, 1]
In-the-wild	SUSTech1K [45]	200	5988	850	19,228	Real-world	[1, 1, 1, 1]	[64, 128, 256, 512]	[1, 2, 2, 1]
	CCGR-MINI [72]	571	27507	399	20377	Real-world	[1, 4, 4, 1]	[64, 128, 256, 512]	[1, 2, 2, 1]
	Gait3D [69]	3,000	18,940	1,000	6,369	Real-world	[1, 4, 4, 1]	[64, 128, 256, 512]	[1, 2, 2, 1]
	GREW [71]	20,000	102,887	6,000	24,000	Real-world	[3, 4, 6, 3]	[64, 128, 256, 512]	[1, 2, 2, 1]

in Diffusion Models, Origins does not require generation through all sequential timestep templates. Instead, it implicitly learns the timestep template relations by randomly sampling to generate gait representations. Given a gait representation sequence $\mathcal{O} \in \mathbb{R}^{\mathcal{S} \times \mathcal{C}_4}$ and a ***Random*** sampled timestep template $\mathcal{T}_{\mathcal{N}} \in \mathbb{R}^{1 \times \mathcal{C}_4}$, the generation process is as follows:

$$\mathcal{Q} = \mathcal{W}^{\mathcal{Q}}\mathcal{O}, \quad \mathcal{K} = \mathcal{W}^{\mathcal{K}}\mathcal{T}_{\mathcal{N}}, \quad \mathcal{V} = \mathcal{W}^{\mathcal{V}}\mathcal{T}_{\mathcal{N}} \tag{5}$$

$$\mathcal{G} = \text{Sigmoid}(\mathcal{Q} \otimes \mathcal{K}^T) \otimes \mathcal{V} \tag{6}$$

where $\mathcal{W}^{\mathcal{Q}}$, $\mathcal{W}^{\mathcal{K}}$ and $\mathcal{W}^{\mathcal{V}} \in \mathbb{R}^{\mathcal{C}_4 \times \mathcal{Z} \times \frac{\mathcal{C}_4}{\mathcal{Z}}}$ are mapping functions and $\mathcal{Z}$ denotes the number of multi-heads in the attention mechanism, similar to Transformers [51]. Each head denotes one type of semantic information. This generation process is optimized by MSE Loss:

$$\mathcal{L}_{mse} = \frac{1}{S}(\mathcal{G} - \text{detach}(\mathcal{O}))^2 \tag{7}$$

The generated gait representation $\mathcal{G}$ is composed of $\mathcal{Z}$ multiple heads of the timestep template ($\mathcal{Z}$ is empirically set to 16). Diffusing Timestep Templates regularizes all gait representations into a consistent and diverse space.

Transferring Unified Template. To further exploit template information for capturing accurate gait differences, Origins transfers the unified template $\mathcal{T}$ to compress a gait sequence into one token. Specifically, as Diffusing Timestep Template progresses, the unified template gradually accumulates rich gait knowledge from the entire gait database. Therefore, the unified template enables to compresses each gait sequence as completely and effectively as possible, which is inspired by the idea of "Compression as Intelligence" [43]. The process is as follows:

$$\mathcal{Q} = \mathcal{W}^{\mathcal{Q}}\mathcal{T}, \quad \mathcal{K} = \mathcal{W}^{\mathcal{K}}\mathcal{O}, \quad \mathcal{V} = \mathcal{W}^{\mathcal{V}}\mathcal{O} \tag{8}$$

$$\mathcal{F} = \text{Softmax}(\mathcal{Q} \otimes \mathcal{K}^T) \otimes \mathcal{V} \tag{9}$$

where $\mathcal{W}^{\mathcal{Q}}$, $\mathcal{W}^{\mathcal{K}}$ and $\mathcal{W}^{\mathcal{V}} \in \mathbb{R}^{\mathcal{C}_4 \times \mathcal{Z} \times \frac{\mathcal{C}_4}{\mathcal{Z}}}$ are mapping functions and $\mathcal{Z}$ denotes the number of multi-heads in the attention mechanism. Finally, this one token $\mathcal{F}$ is fed into the following Recognition Head.

Summarize. Origins includes that (i) Only the unified template serves as ($\mathcal{Q}$) transfers to extract discriminative representations from real gait samples ($\mathcal{K}$, $\mathcal{V}$); (ii) Meanwhile, the unified template diffuses into timestep templates; (iii) Only timestep templates serve as bases ($\mathcal{K}$, $\mathcal{V}$) conditioned by real gait samples ($\mathcal{Q}$) for generative learning, where each gait sample randomly selects a timestep template.

3.3. Training Details

Origins unifies gait generative and representation learning for end-to-end joint training. The joint loss is as follows:

$$\mathcal{L} = \mathcal{L}_{mse} + \mathcal{L}_{tp} + \mathcal{L}_{ce} \tag{10}$$

where $\mathcal{L}_{mse}$, $\mathcal{L}_{tp}$ and $\mathcal{L}_{ce}$ denote MSE Loss, Triplet Loss [18] and Cross Entropy Loss.

4. Experiments

4.1. Datasets

Gait datasets are generally divided into two subsets: constrained and in-the-wild, based on their collection environments. As shown in Table. 1, constrained datasets (*i.e.*, CASIA-B [66] and CCPG [29]) typically provide quantified conditional benchmarks but include a relatively small number of identities, whereas in-the-wild datasets (*i.e.*, SUSTech1K [45], Gait3D [69], GREW [71]) and CCGR-MINI [72] involve more complex environments and a large number of identities. Origins is thoroughly evaluated the effectiveness across these widely-used gait benchmarks.

CASIA-B contains 124 subjects with 11 camera views and 3 scenarios: normal walking (NM), carrying a bag (BG) and cloth-changing condition (CL).

CCPG provides the challenge of rich cross-clothing conditions and contains 200 subjects with over 16,000 sequences.

SUSTech1K consists of 1,050 subjects under various real-world conditions such as Clothing, Occlusion and Night.

Gait3D collects gait data in a supermarket, addressing practical gait recognition. it includes 3,000 subjects and 25,309 sequences, divided into a training set of 2,000 subjects and a testing set of 1,000 subjects.

GREW is a large-scale gait dataset in the wild, including 26,345 subjects and 128,671 sequences recorded by 882

Table 2. The performance comparisons on CCPG under various conditions, including full-body cloth changes (CL), upper-body cloth changes (UP), lower-body cloth changes (DN), and backpacks changes (BG), reported with Rank-1 accuracy (%).

Paradigm	Method	Venue	Gait Evaluation Protocol					ReID Evaluation Protocol				
			CL	UP	DN	BG	Mean	CL	UP	DN	BG	Mean
Model	GaitGraph2 [47]	CVPRW22	5.0	5.3	5.8	6.2	5.1	5.0	5.7	7.3	8.8	6.7
	Gait-TR [68]	ES23	15.7	18.3	18.5	17.5	17.5	24.3	28.7	31.1	28.1	28.1
	MSGG [40]	MTA23	29.0	34.5	37.1	33.3	33.5	43.1	52.9	57.4	49.9	50.8
	SkeletonGait [11]	AAAI24	40.4	48.5	53.0	61.7	50.9	52.4	65.4	72.8	80.9	67.9
Appearance	GaitSet [4]	AAAI19	60.2	65.2	65.1	68.5	64.8	77.5	85.0	82.9	87.5	83.2
	GaitPart [9]	CVPR20	64.3	67.8	68.6	71.7	68.1	79.2	85.3	86.5	88.0	84.8
	OGBase [29]	CVPR23	52.1	57.3	60.1	63.3	58.2	70.2	76.9	80.4	83.4	77.7
	GaitBase [10]	CVPR23	71.6	75.0	76.8	78.6	75.5	88.5	92.7	93.4	93.2	92.0
	DeepGaitV2 [12]	TPAMI25	78.6	84.8	80.7	89.2	83.3	90.5	96.3	91.4	96.7	93.7
	Origins-S (**ours**)	-	**84.3**	**90.2**	**86.4**	**93.6**	**88.6**	**93.4**	**97.6**	**94.6**	**97.6**	**95.8**

Table 3. The performance comparisons on SUSTech1K are reported with Rank-1 and Rank-5 accuracy (%).

Paradigm	Method	Venue	Probe Sequence (Rank-1)								Overall	
			Normal	Bag	Clothing	Carrying	Umbrella	Uniform	Occlusion	Night	Rank-1	Rank-5
Model	GaitGraph2 [47]	CVPRW22	22.2	18.2	6.8	18.6	13.4	19.2	27.3	16.4	18.6	40.2
	Gait-TR [68]	ES23	33.3	31.5	21.0	30.4	22.7	34.6	44.9	23.5	30.8	56.0
	MSGG [40]	MTA23	67.11	66.16	35.92	63.31	61.58	58.07	66.59	17.88	33.8	-
	SkeletonGait [11]	AAAI24	67.9	63.5	36.5	61.6	58.1	67.2	79.1	50.1	63.0	83.5
Appearance	GaitSet [4]	AAAI19	69.1	68.2	37.4	65.0	63.1	61.0	67.2	23.0	65.0	84.8
	GaitPart [9]	CVPR20	62.2	62.8	33.1	59.5	57.2	54.8	57.2	21.7	59.2	80.8
	GaitGL [32]	ICCV21	67.1	66.2	35.9	63.3	61.6	58.1	66.6	17.9	63.1	82.8
	GaitBase [10]	CVPR23	81.5	77.5	49.6	75.8	75.5	76.7	81.4	25.9	76.1	89.4
	DeepGaitV2 [12]	TPAMI25	87.4	84.1	53.4	81.3	86.1	84.8	88.5	28.8	82.3	92.5
	Origins-S (**ours**)	-	**91.4**	**88.1**	**64.8**	**86.0**	**89.8**	**88.9**	**92.8**	**29.6**	**86.9**	**94.2**

cameras. The benchmark consists of 20,000 subjects for training and 6,000 for testing.

CCGR-MINI is provided by the CCGR [72] team, which serves as an alternative to CCGR, offers fast training, and maintains equivalent covariates.

4.2. Implementation Details

The details of the training process are as follows. **Inputs.** Each input gait sequence consists of 30 frames, and all silhouettes are resized to 64 × 44. The batch size $]\mathcal{I}$, $\mathcal{J}]$ is consistent with [10], where $\mathcal{I}$ represents the number of subjects sampled per mini-batch, and $\mathcal{J}$ represents the number of sequences sampled per subject. **Networks.** As shown in Table. 1, we provide Origins-T, Origins-S, Origins-M, and Origins-L based on network depths. Origins-T consists of 3 Bottleneck3D Stages with block numbers [1, 1, 1] and channels [64, 128, 256]. Origins-S comprises 4 Bottleneck3D Stages with block numbers [1, 1, 1, 1] and channels [64, 128, 256, 512]. Origins-M consists of 1 Basic2D Stage and 3 Pseudo3D Stages, with block numbers [1, 4, 4, 1] and channels [64, 128, 256, 512]. Origins-L is composed of 4 Bottleneck3D Stages, with block numbers [3, 4, 6, 3] and channels [64, 128, 256, 512]. **Optimization.** We use the optimizer of SGD with an initial learning rate of 0.1, which is decreasing by a factor of 0.1 per $[20K, 40K, 50K]$, $[20K, 40K, 50K]$, $[20K, 30K, 40K]$, $[20K, 40K, 50K]$, $[80K, 120K, 150K]$, $[30K, 55K, 65K]$ for CASIA-B (Total $60K$), CCPG (Total $60K$), SUSTech1K (Total $50K$), Gait3D (Total $60K$), GREW (Total $180K$) and CCGR-MINI (Total $80K$). All the models are trained on NVIDIA 8×3090 GPUs.

4.3. Results on Constrained Scenario

We first validate the effectiveness of Origins across various complex scenarios on CASIA-B and CCPG.

CASIA-B. As shown in Table. 4, Origins-T achieves state-of-the-art (SoTA) performance on all scenarios, with an average accuracy of 95.7%, demonstrating that the unified template possesses a consistent and diverse semantic space to improve the individual distinctiveness and discriminability of gait representations under different conditions.

CCPG. As shown in Table. 2, Origins-S achieves outstanding performance (*e.g.*, 88.6% in Gait Evaluation Protocol and 95.8% in ReID Evaluation Protocol) in more challenging clothing-change scenarios (*e.g.*, full-body, upper-body, lower-body, and backpacks changes), which indicates the ability to capture accurate gait differences, significantly alleviating one of the biggest bottlenecks in current gait

Table 4. The performance comparisons on CASIA-B are reported with Rank-1 accuracy (%).

Method	Venue	NM	BG	CL	Mean
GaitSet [4]	AAAI19	95.0	87.2	70.4	84.2
GaitPart [9]	CVPR20	96.2	91.5	78.7	88.8
GLN [21]	ECCV20	96.9	94.0	77.5	89.5
GaitGL [32]	ICCV21	97.4	94.5	83.6	91.8
QAGait [60]	AAAI24	97.9	94.6	78.2	90.2
GaitBase [10]	CVPR23	97.6	94.0	77.4	89.8
DANet [33]	CVPR23	98.0	95.9	89.9	94.6
GaitGCI [8]	CVPR23	97.9	95.0	86.4	93.1
DyGait [57]	ICCV23	98.4	96.2	87.8	94.1
HSTL [55]	ICCV23	98.1	95.9	88.9	94.3
VPNet [35]	CVPR24	98.3	96.3	90.0	94.9
DeepGaitV2 [12]	TPAMI25	-	-	-	89.6
CLTD [63]	ECCV24	98.6	96.4	89.3	94.8
Free Lunch [53]	ECCV24	98.1	94.1	77.9	90.0
Origins-T (**ours**)	-	**99.3**	**97.4**	**90.3**	**95.7**

Table 5. The performance comparisons on Gait3D are reported with Rank-1, Rank-5 accuracy and mAP (%).

Method	Venue	Rank-1	Rank-5	mAP
GaitSet [4]	AAAI19	36.7	58.3	30.0
GaitPart [9]	CVPR20	28.2	47.6	47.6
GaitGL [32]	ICCV21	29.7	48.5	22.3
SMPLGait [69]	CVPR22	46.3	64.5	37.2
MTSGait [69]	MM22	48.7	67.1	37.6
QAGait [60]	AAAI24	67.0	81.5	56.5
GaitBase [10]	CVPR23	64.6	-	-
DANet [33]	CVPR23	48.0	69.7	-
GaitGCI [8]	CVPR23	50.3	68.5	39.5
DyGait [57]	ICCV23	66.3	80.8	56.4
HSTL [55]	ICCV23	61.3	76.3	55.5
VPNet [35]	CVPR24	75.4	87.1	-
DeepGaitV2 [12]	TPAMI25	74.4	88.0	65.8
CLTD [63]	ECCV24	69.7	85.2	-
GaitMoE [25]	ECCV24	73.7	-	66.2
Free Lunch [53]	ECCV24	70.1	-	61.9
Origins-M (**ours**)	-	**75.8**	**86.8**	**67.0**

recognition: the clothing-change problem.

4.4. Results on in-the-wild Scenario

We then validate the robustness of Origins against more complex and unknown covariates on in-the-wild datasets, SUSTech1K, Gait3D, GREW and CCGR-MINI.

SUSTech1K. As shown in Table. 3, Origins-S demonstrates superior performance in more diverse and complex environments, surpassing state-of-the-art methods DeepGaitV2 [12] by a significant margin 4.6% in Overall Rank-1 accuracy. Notably, even facing the combination of various covariates, Origins achieves remarkable performance (*e.g.*, 64.8% Rank-1 accuracy in the Clothing benchmark), demonstrating its ability to preserve diverse semantic space during gait representation learning.

Table 6. The performance comparisons on GREW are reported with Rank-1, Rank-5 and Rank-10 accuracy (%).

Method	Venue	Rank-1	Rank-5	Rank-10
GaitSet [4]	AAAI19	46.3	63.6	70.3
GaitPart [9]	CVPR20	44.0	60.7	67.3
GaitGL [32]	ICCV21	47.3	63.6	-
MTSGait [69]	MM22	55.3	71.3	76.9
QAGait [60]	AAAI24	59.1	74.0	79.2
GaitBase [10]	CVPR23	60.1	-	-
GaitGCI [8]	CVPR23	68.5	80.8	84.9
DyGait [57]	ICCV23	71.4	83.2	86.8
HSTL [55]	ICCV23	62.7	76.6	81.3
VPNet [35]	CVPR24	80.0	89.4	-
DeepGaitV2 [12]	TPAMI25	77.7	88.9	91.8
CLTD [63]	ECCV24	78.0	87.8	-
GaitMoE [25]	ECCV24	79.6	89.1	-
Free Lunch [53]	ECCV24	65.5	78.7	83.3
Origins-L (**ours**)	-	**80.8**	**89.6**	**92.1**

Table 7. The performance comparisons on CCGR-MINI are reported with Rank-1 accuracy, mAP and mINP (%).

Method	Venue	Rank-1	mAP	mINP
GaitSet [4]	AAAI19	13.77	15.39	5.75
GaitPart [9]	CVPR20	8.02	10.12	3.52
GaitGL [32]	ICCV21	17.51	18.12	6.85
GaitBase [10]	CVPR23	26.99	24.89	9.72
DeepGaitV2 [12]	TPAMI25	39.37	36.01	16.77
Origins-M (**ours**)	-	**41.45**	**38.31**	**24.71**

Gait3D. Origins-M achieves the SoTAs shown in Table. 5, surpassing the latest appearance-based method CLTD [63] by 6.1% in Rank-1 accuracy. Compared to the parameter-heavy GaitMoE [25], it still maintain the improvements, indicating that Origins-M learns meaningful semantic space rather than noises.

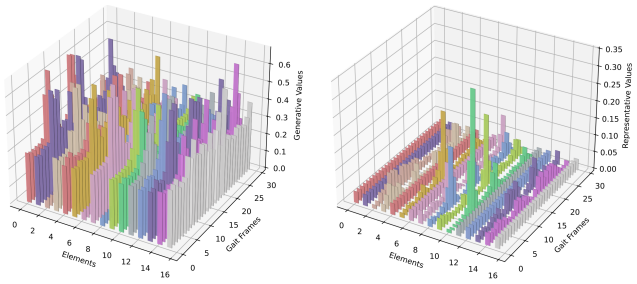
GREW. Origins-L achieves competitive performance shown in Table. 6, surpassing VPNet [35] by 0.8% Rank-1 accuracy. Origins-L adopts Free Lunch [53] (*i.e.*, logits as gait representations) to achieve more stable results without introducing any additional computational complexity.

CCGR-MINI. As shown in Table. 7, Origins-M outperforms the previous SoTA DeepGaitV2 by 2.1% Rank-1 accuracy, which further reveals the robustness.

4.5. Ablation Study

We first validate the core modules of Origins on Gait3D. Then, we analyze the number of timestep templates on Gait3D and CASIA-B. Finally, we perform visualization to better understand the mechanisms of Origins.

The core modules of Origins. As shown in Table. 8, Origins achieved significant improvments compared to the Baseline with 2.4% higher Rank-1 accuracy on Gait3D.



(a) ID 77 with the timestep template (b) ID 77 with the unified template

(c) ID 75 with the timestep template (d) ID 75 with the unified template

Figure 7. The "elements" axe is $\mathcal{Z}$ that is the number of multi-heads in the template. The visualization shows that different subjects exhibit differences in the template.

Specifically, the combination of Transferring Unified Template and Diffusing Timestep Templates enables further improving performance, proving the effectiveness of unified gait generative and representation learning. Additionally, we find that the number of multi-heads $\mathcal{Z}$ in attention mechanism is set to 16 for better performance shown in Table. 8.

The number of timestep templates. As shown in Table. 9, the number of timestep templates represents a trade-off. For a large gait database with complex covariates, having fewer timestep templates may result in a narrow semantic space, potentially affecting the expression of gait representations. Conversely, more timestep templates could lead to an overwhelming semantic space that introduces noise and confusion into the representations. For the smaller datasets, the number of timestep templates has less impact, which is make sense as gait representation learning already enable to capture sufficient gait patterns.

The visualization of the unified Template. As shown in (a, c) of Fig. 7, we perform the generative differences to better understand the mechanism of Diffusing Timestep Templates. Specifically, we select the 8th timestep template to generate gait representations for different IDs. It can be observed that different IDs impact the signal from the timestep template. However, they can be linearly combined by these elements (*i.e.*, the multi-heads $\mathcal{Z}$), which means they are regularized to a consistent space. As shown in (b, d) of Fig. 7, we perform the discriminative differences to better understand the mechanism. Transferring Unified Template aims to compresses each gait sequence into one token as completely and effectively as possible. It can be observed that not all information within a gait sequence effectively contribute to recognition. Instead, the unified template captures accurate gait differences in the multi-heads $\mathcal{Z}$ attention mechanism for recognition. We also observe the gait periodicity in some heads, highlighting the impact of fine-grained body motion.

Table 8. The core module analysis on Gait3D.

Method	Gait3D		
	Rank-1	Rank-5	mAP
Origins-M	75.8	86.8	67.0
The analysis on the core modules			
-w/o Origins	73.6	86.9	65.1
- w/o Transferring Unified Template	74.4	87.9	67.1
- w/o Diffusing Timestep Templates	75.3	86.4	66.3
The analysis on the number of multi-heads			
$\mathcal{Z} = 8$	74.6	86.6	66.7
$\mathcal{Z} = 16$	75.8	86.8	67.0
$\mathcal{Z} = 32$	75.2	86.4	66.1

Table 9. The timestep template analysis on Gait3D and CASIA-B.

Method	Gait3D			CASIA-B			
	Rank-1	Rank-5	mAP	NM	BG	CL	Mean
Origins-M	75.8	86.8	67.0	99.3	97.4	90.3	95.7
Baseline	73.6	86.9	65.1	98.2	95.8	84.4	92.8
The analysis on the number of timestep templates							
$\mathcal{N} = 8$	75.3	87.0	66.5	99.1	97.3	90.4	95.6
$\mathcal{N} = 16$	75.8	86.8	67.0	99.3	97.4	90.3	95.7
$\mathcal{N} = 32$	73.7	86.3	65.4	99.2	97.3	90.0	95.5

5. Conclusion and Limitations

In this work, we propose Origins with generative capabilities, regularizing gait representations within a consistent and diverse semantic space and addressing semantic inconsistency and uniformity in complex scenarios. Origins learns a unified template through diffusing timestep templates for gait generative learning, addressing the convergence difficulty, and meanwhile transfers the unified template for gait representation learning, capturing accurate gait differences. Extensive experiments demonstrate that Origins performs unified generative and representation learning, achieving superior performance.

Limitations. While diffusion models have been widely adopted to generate new samples at the pixel level, enhancing the usability, Origins performs generative learning in the representation space. In the future, we will explore generating new gait samples to further enhance general representations, such as in pretraining paradigms.

Acknowledgement

This work is jointly supported by National Natural Science Foundation of China (62276025, 62206022, 62476027) and the Fundamental Research Funds for the Central Universities (2253200026).

References

[1] Creativity: Discipline for imagination. *Nature*, 226:1196–1197, 1970. 1

[2] Sana Ayromlou, Arash Afkanpour, Vahid Reza Khazaie, and Fereshteh Forghani. Can generative models improve self-supervised representation learning? *arXiv preprint arXiv:2403.05966*, 2024. 3

[3] Dmitry Baranchuk, Ivan Rubachev, Andrey Voynov, Valentin Khrulkov, and Artem Babenko. Label-efficient semantic segmentation with diffusion models. *arXiv preprint arXiv:2112.03126*, 2021. 3

[4] Hanqing Chao, Yiwei He, Junping Zhang, and Jianfeng Feng. Gaitset: Regarding gait as a set for cross-view gait recognition. In *Proceedings of the AAAI conference on artificial intelligence*, pages 8126–8133, 2019. 3, 6, 7

[5] Xinlei Chen, Zhuang Liu, Saining Xie, and Kaiming He. Deconstructing denoising diffusion models for self-supervised learning. *arXiv preprint arXiv:2401.14404*, 2024. 1, 3

[6] Yufeng Cui and Yimei Kang. Multi-modal gait recognition via effective spatial-temporal feature fusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17949–17957, 2023. 3

[7] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021. 3, 4

[8] Huanzhang Dou, Pengyi Zhang, Wei Su, Yunlong Yu, Yining Lin, and Xi Li. Gaitgci: Generative counterfactual intervention for gait recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5578–5588, 2023. 3, 7

[9] 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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14225–14233, 2020. 3, 6, 7

[10] Chao Fan, Junhao Liang, Chuanfu Shen, Saihui Hou, Yongzhen Huang, and Shiqi Yu. Opengait: Revisiting gait recognition towards better practicality. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9707–9716, 2023. 3, 4, 6, 7

[11] Chao Fan, Jingzhe Ma, Dongyang Jin, Chuanfu Shen, and Shiqi Yu. Skeletongait: Gait recognition using skeleton maps. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 1662–1669, 2024. 3, 6

[12] Chao Fan, Saihui Hou, Junhao Liang, Chuanfu Shen, Jingzhe Ma, Dongyang Jin, Yongzhen Huang, and Shiqi Yu. Opengait: A comprehensive benchmark study for gait recognition towards better practicality. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–18, 2025. 6, 7

[13] Yang Fu, Shibei Meng, Saihui Hou, Xuecai Hu, and Yongzhen Huang. Gpgait: Generalized pose-based gait recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 19595–19604, 2023. 3

[14] Yang Fu, Saihui Hou, Shibei Meng, Xuecai Hu, Chunshui Cao, Xu Liu, and Yongzhen Huang. Cut out the middleman: Revisiting pose-based gait recognition. In *European Conference on Computer Vision*, pages 112–128. Springer, 2024. 3

[15] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020. 4

[16] Wenxuan Guo, Yingping Liang, Zhiyu Pan, Ziheng Xi, Jianjiang Feng, and Jie Zhou. Camera-lidar cross-modality gait recognition. In *European Conference on Computer Vision*, pages 439–455. Springer, 2024. 3

[17] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 4

[18] Alexander Hermans, Lucas Beyer, and Bastian Leibe. In defense of the triplet loss for person re-identification. *arXiv preprint arXiv:1703.07737*, 2017. 4, 5

[19] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022. 3

[20] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 1, 2

[21] Saihui Hou, Chunshui Cao, Xu Liu, and Yongzhen Huang. Gait lateral network: Learning discriminative and compact representations for gait recognition. In *European conference on computer vision*, pages 382–398. Springer, 2020. 7

[22] Saihui Hou, Zengbin Wang, Man Zhang, Chunshui Cao, Xu Liu, and Yongzhen Huang. Edge-oriented adversarial attack for deep gait recognition. *International Journal of Computer Vision*, 133(4):1549–1563, 2025. 3

[23] Vincent Tao Hu, David W Zhang, Yuki M Asano, Gertjan J Burghouts, and Cees GM Snoek. Self-guided diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18413–18422, 2023. 3

[24] Panjian Huang, Saihui Hou, Chunshui Cao, Xu Liu, Xuecai Hu, and Yongzhen Huang. Integral pose learning via appearance transfer for gait recognition. *IEEE Transactions on Information Forensics and Security*, 2024. 2

[25] Panjian Huang, Yunjie Peng, Saihui Hou, Chunshui Cao, Xu Liu, Zhiqiang He, and Yongzhen Huang. Occluded gait recognition with mixture of experts: an action detection perspective. In *European Conference on Computer Vision*, pages 380–397. Springer, 2024. 2, 3, 7

[26] Drew A Hudson, Daniel Zoran, Mateusz Malinowski, Andrew K Lampinen, Andrew Jaegle, James L McClelland, Loic Matthey, Felix Hill, and Alexander Lerchner. Soda: Bottleneck diffusion models for representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23115–23127, 2024. 1

[27] Yoed N Kenett. The role of knowledge in creative thinking. *Creativity Research Journal*, pages 1–8, 2024. 1

[28] Tianhong Li, Dina Katabi, and Kaiming He. Self-conditioned image generation via generating representations. *arXiv preprint arXiv:2312.03701*, 2023. 1

[29] Weijia Li, Saihui Hou, Chunjie Zhang, Chunshui Cao, Xu Liu, Yongzhen Huang, and Yao Zhao. An in-depth exploration of person re-identification and gait recognition in cloth-changing conditions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13824–13833, 2023. 5, 6

[30] Junhao Liang, Chao Fan, Saihui Hou, Chuanfu Shen, Yongzhen Huang, and Shiqi Yu. Gaitedge: Beyond plain end-to-end gait recognition for better practicality. In *European conference on computer vision*, pages 375–390. Springer, 2022. 3

[31] Rijun Liao, Shiqi Yu, Weizhi An, and Yongzhen Huang. A model-based gait recognition method with body pose and human prior knowledge. *Pattern Recognition*, 98:107069, 2020. 3

[32] Beibei Lin, Shunli Zhang, and Xin Yu. Gait recognition via effective global-local feature representation and local temporal aggregation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14648–14656, 2021. 3, 6, 7

[33] Kang Ma, Ying Fu, Dezhi Zheng, Chunshui Cao, Xuecai Hu, and Yongzhen Huang. Dynamic aggregated network for gait recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22076–22085, 2023. 3, 7

[34] Kang Ma, Ying Fu, Dezhi Zheng, Yunjie Peng, Chunshui Cao, and Yongzhen Huang. Fine-grained unsupervised domain adaptation for gait recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11313–11322, 2023. 1

[35] Kang Ma, Ying Fu, Chunshui Cao, Saihui Hou, Yongzhen Huang, and Dezhi Zheng. Learning visual prompt for gait recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 593–603, 2024. 1, 3, 7

[36] Sarnoff Mednick. The associative basis of the creative process. *Psychological review*, 69(3):220, 1962. 1

[37] Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*, 2021. 4

[38] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4195–4205, 2023. 3

[39] Guozhen Peng, Yunhong Wang, Yuwei Zhao, Shaoxiong Zhang, and Annan Li. Glgait: A global-local temporal receptive field network for gait recognition in the wild. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 826–835, 2024. 3

[40] Yunjie Peng, Kang Ma, Yang Zhang, and Zhiqiang He. Learning rich features for gait recognition by integrating skeletons and silhouettes. *Multimedia Tools and Applications*, 83(3):7273–7294, 2024. 6

[41] Ekkasit Pinyoanuntapong, Ayman Ali, Pu Wang, Minwoo Lee, and Chen Chen. Gaitmixer: skeleton-based gait representation learning via wide-spectrum multi-axial mixer. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023. 3

[42] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 3, 4

[43] Jürgen Schmidhuber. Driven by compression progress: A simple principle explains essential aspects of subjective beauty, novelty, surprise, interestingness, attention, curiosity, creativity, art, science, music, jokes. In *Workshop on anticipatory behavior in adaptive learning systems*, pages 48–76. Springer, 2008. 5

[44] Alireza Sepas-Moghaddam and Ali Etemad. Deep gait recognition: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 45(1):264–284, 2022. 1

[45] Chuanfu Shen, Chao Fan, Wei Wu, Rui Wang, George Q Huang, and Shiqi Yu. Lidargait: Benchmarking 3d gait recognition with point clouds. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1054–1063, 2023. 3, 5

[46] Chunfeng Song, Yongzhen Huang, Weining Wang, and Liang Wang. Casia-e: a large comprehensive dataset for gait recognition. *IEEE transactions on pattern analysis and machine intelligence*, 45(3):2801–2815, 2022. 1

[47] Torben Teepe, Ali Khan, Johannes Gilg, Fabian Herzog, Stefan Hörmann, and Gerhard Rigoll. Gaitgraph: Graph convolutional network for skeleton-based gait recognition. In *2021 IEEE international conference on image processing (ICIP)*, pages 2314–2318. IEEE, 2021. 3, 6

[48] Torben Teepe, Johannes Gilg, Fabian Herzog, Stefan Hörmann, and Gerhard Rigoll. Towards a deeper understanding of skeleton-based gait recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1569–1577, 2022. 3

[49] Changyao Tian, Chenxin Tao, Jifeng Dai, Hao Li, Ziheng Li, Lewei Lu, Xiaogang Wang, Hongsheng Li, Gao Huang, and Xizhou Zhu. Addp: Learning general representations for image recognition and generation with alternating denoising diffusion process. *arXiv preprint arXiv:2306.05423*, 2023. 1, 3

[50] Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. *Advances in neural information processing systems*, 30, 2017. 4

[51] A Vaswani. Attention is all you need. *Advances in Neural Information Processing Systems*, 2017. 5

[52] Jilong Wang, Saihui Hou, Yan Huang, Chunshui Cao, Xu Liu, Yongzhen Huang, and Liang Wang. Causal intervention for sparse-view gait recognition. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 77–85, 2023. 3

[53] Jilong Wang, Saihui Hou, Yan Huang, Chunshui Cao, Xu Liu, Yongzhen Huang, Tianzhu Zhang, and Liang Wang. Free lunch for gait recognition: A novel relation descriptor. In *European Conference on Computer Vision*, pages 39–56. Springer, 2024. 3, 7

[54] Jilong Wang, Saihui Hou, Xianda Guo, Yan Huang, Yongzhen Huang, Tianzhu Zhang, and Liang Wang. Gaitc 3 i: Robust cross-covariate gait recognition via causal intervention. *IEEE Transactions on Circuits and Systems for Video Technology*, 2025. 3

[55] Lei Wang, Bo Liu, Fangfang Liang, and Bincheng Wang. Hierarchical spatio-temporal representation learning for gait recognition. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 19582–19592. IEEE, 2023. 3, 7

[56] Lei Wang, Yinchi Ma, Peng Luan, Wei Yao, Congcong Li, and Bo Liu. Hih: A multi-modal hierarchy in hierarchy network for unconstrained gait recognition. *arXiv preprint arXiv:2311.11210*, 2023. 3

[57] Ming Wang, Xianda Guo, Beibei Lin, Tian Yang, Zheng Zhu, Lincheng Li, Shunli Zhang, and Xin Yu. Dygait: Exploiting dynamic representations for high-performance gait recognition. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 13424–13433, 2023. 3, 7

[58] Zengbin Wang, Saihui Hou, Man Zhang, Xu Liu, Chunshui Cao, and Yongzhen Huang. Gaitparsing: Human semantic parsing for gait recognition. *IEEE Transactions on Multimedia*, 2023. 3

[59] Zengbin Wang, Saihui Hou, Man Zhang, Xu Liu, Chunshui Cao, Yongzhen Huang, and Shibiao Xu. Landmarkgait: intrinsic human parsing for gait recognition. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 2305–2314, 2023. 3

[60] Zengbin Wang, Saihui Hou, Man Zhang, Xu Liu, Chunshui Cao, Yongzhen Huang, Peipei Li, and Shibiao Xu. Qagait: Revisit gait recognition from a quality perspective. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 5785–5793, 2024. 3, 7

[61] Bin Xia, Yulun Zhang, Shiyin Wang, Yitong Wang, Xinglong Wu, Yapeng Tian, Wenming Yang, and Luc Van Gool. Diffir: Efficient diffusion model for image restoration. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 13095–13105, 2023. 1

[62] Weilai Xiang, Hongyu Yang, Di Huang, and Yunhong Wang. Denoising diffusion autoencoders are unified self-supervised learners. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15802–15812, 2023. 1

[63] Haijun Xiong, Bin Feng, Xinggang Wang, and Wenyu Liu. Causality-inspired discriminative feature learning in triple domains for gait recognition. In *European Conference on Computer Vision*, pages 251–270. Springer, 2024. 3, 7

[64] Xingyi Yang and Xinchao Wang. Diffusion model as representation learner. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 18938–18949, 2023. 3

[65] Xiulong Yang, Sheng-Min Shih, Yinlin Fu, Xiaoting Zhao, and Shihao Ji. Your vit is secretly a hybrid discriminative-generative diffusion model. *arXiv preprint arXiv:2208.07791*, 2022. 3

[66] Shiqi Yu, Daoliang Tan, and Tieniu Tan. 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)*, pages 441–444. IEEE, 2006. 5

[67] Sihyun Yu, Sangkyung Kwak, Huiwon Jang, Jongheon Jeong, Jonathan Huang, Jinwoo Shin, and Saining Xie. Representation alignment for generation: Training diffusion transformers is easier than you think. *arXiv preprint arXiv:2410.06940*, 2024. 1

[68] Cun Zhang, Xing-Peng Chen, Guo-Qiang Han, and Xiang-Jie Liu. Spatial transformer network on skeleton-based gait recognition. *Expert Systems*, 40(6):e13244, 2023. 3, 6

[69] Jinkai Zheng, Xinchen Liu, Xiaoyan Gu, Yaoqi Sun, Chuang Gan, Jiyong Zhang, Wu Liu, and Chenggang Yan. Gait recognition in the wild with multi-hop temporal switch. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 6136–6145, 2022. 3, 5, 7

[70] Jinkai Zheng, Xinchen Liu, Shuai Wang, Lihao Wang, Chenggang Yan, and Wu Liu. Parsing is all you need for accurate gait recognition in the wild. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 116–124, 2023. 3

[71] 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 *Proceedings of the IEEE/CVF international conference on computer vision*, pages 14789–14799, 2021. 5

[72] Shinan Zou, Chao Fan, Jianbo Xiong, Chuanfu Shen, Shiqi Yu, and Jin Tang. Cross-covariate gait recognition: A benchmark. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 7855–7863, 2024. 5, 6