Learning Visual Prompt for Gait Recognition

Kang Ma[1], Ying Fu[1]*, Chunshui Cao[2], Saihui Hou[2,3], Yongzhen Huang[2,3], Dezhi Zheng[1]*
[1]Beijing Institute of Technology, [2]WATRIX.AI, [3]Beijing Normal University
kangx.ma@gmail.com, fuying@bit.edu.cn, chunshui.cao@watrix.ai, housaihui@bnu.edu.cn,
huangyongzhen@bnu.edu.cn, zhengdezhi@bit.edu.cn

Abstract

Gait, a prevalent and complex form of human motion, plays a significant role in the field of long-range pedestrian retrieval due to the unique characteristics inherent in individual motion patterns. However, gait recognition in real-world scenarios is challenging due to the limitations of capturing comprehensive cross-viewing and cross-clothing data. Additionally, distractors such as occlusions, directional changes, and lingering movements further complicate the problem. The widespread application of deep learning techniques has led to the development of various potential gait recognition methods. However, these methods utilize convolutional networks to extract shared information across different views and attire conditions. Once trained, the parameters and non-linear function become constrained to fixed patterns, limiting their adaptability to various distractors in real-world scenarios. In this paper, we present a unified gait recognition framework to extract global motion patterns and develop a novel dynamic transformer to generate representative gait features. Specifically, we develop a trainable part-based prompt pool with numerous key-value pairs that can dynamically select prompt templates to incorporate into the gait sequence, thereby providing task-relevant shared knowledge information. Furthermore, we specifically design dynamic attention to extract robust motion patterns and address the length generalization issue. Extensive experiments on four widely recognized gait datasets, i.e., Gait3D, GREW, OUMVLP, and CASIA-B, reveal that the proposed method yields substantial improvements compared to current state-of-the-art approaches.

1. Introduction

Gait, as a distinct and intricate form of movement [4, 43], holds significant implications for long-distance recognition [36, 49, 60] owing to its remarkable uniqueness. However, gait recognition suffers from various complicating factors in real-world scenarios. Silhouette data from the same individual can exhibit significant variation depending on viewing angles and clothing conditions [44, 58], resulting in more intra-class variation than inter-class variation. In addition, potential obstructions from crowds and actions of the observed subjects [61, 63], such as turning, lingering, or stopping, may manifest within a single gait sequence. Despite the ability of human visual perception to seamlessly establish correspondences between occlusions, viewpoint changes, and object transformations, it still poses a formidable challenge for the gait recognition network.

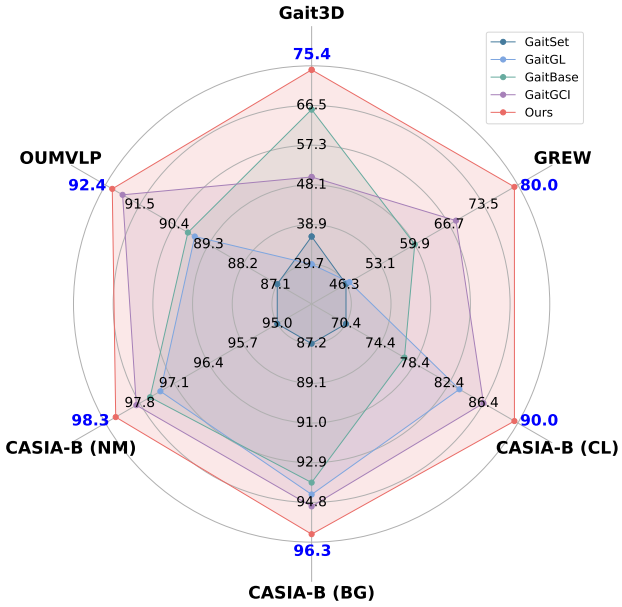

Figure 1. Performance comparison of prevailing methods and the proposed method in both real-world and laboratory scenarios.

Learning robust motion patterns from multiple variations all at once is not a trivial task. With the widespread application of deep learning, some promising methods [7, 14, 22, 33, 35] have emerged. However, *these methods preserve shared information across different viewpoints and attire conditions while discarding task-relevant knowledge*

*Corresponding Authors

relevant to particular conditions. Intuitively, the knowledge is crucial for the gait network to learn valuable invariances. Thus, we aim to reveal it from the following aspects:

(i) Prompt Templates. A widely acknowledged assumption [7, 35, 54] regarding gait is that individuals maintain consistent motion patterns regardless of external factors. Central to all of these capabilities is the ability to establish associations despite occlusions, viewpoint changes, and varying object appearances. Indeed, in controlled laboratory settings marked by abundant data and well-managed conditions, stable associations tend to form readily. Conversely, addressing scenarios in the wild, constrained by limited data and influenced by obstructions and crowded environments, significantly increases the complexity of establishing stable associations. Consequently, *a simple and effective approach is to provide task-relevant knowledge using a learnable prompt pool, which is expected to adaptively select prompt templates based on sequence attributes to establish stable associations under various disturbances.*

(ii) Length Generalization. Similar to natural language processing (NLP) tasks [47], gait also faces the problem of length generalization [17, 42, 55]. In particular, the majority of the most promising gait recognition methods involve training by randomly selecting a subset of the sequence and testing by extracting features from the entire sequence. Nevertheless, these methods frequently neglect the challenges presented by the longer sequence. To this end, we implement a dynamic context window mask for gait, enabling the network to focus its attention on the features within the specified window. Furthermore, *we propose to establish a correlation between the localized window and the prompt templates, with the aim of emphasizing the significance of distinctive gait features within the sequence*.

Based on the above analysis, we present a comprehensive Visual Prompt Network (VPNet) for gait recognition. VPNet leverages prompt templates to augment the task-relevant information associated with the gait feature. Dynamic attention is employed to establish connections between prompt templates and local window features, enabling the network to capture crucial invariant gait features. As illustrated in Fig. 2, VPNet initially extracts features from the sampled sequence using a backbone network. These features are then compared with a trainable part-based prompt pool, with prompt templates dynamically selected and integrated into the gait sequence. The merged features are further input into the dynamic transformer to enhance the learning of global motion patterns. VPNet exhibits superior performance, especially in realistic scenarios, compared to state-of-the-art (SOTA) methods on publicly available datasets as shown in Fig. 1. A brief overview of our main contributions can be summarized as follows:

- We introduce an innovative visual prompt that integrates key-value pairs with gait features, encoding factors such as viewpoint, attire, occlusion, etc. These prompt templates provide task-related information to the network, facilitating the extraction of representative gait features.
- We design an efficient dynamic transformer with a context window mask structure to correlate prompt templates with local motion patterns, addressing the length generalization problem for gait recognition.
- We present a unified framework that yields competitive results on publicly accessible datasets, namely Gait3D [61], GREW [63], OUMVLP [44], and CASIA-B [58]. In addition, a series of rigorous ablation experiments further validate the effectiveness of our approach.

2. Relate Work

In this section, we present a review of the most relevant studies on gait recognition and prompt engineering.

2.1. Gait Recognition

Gait recognition approaches can be broadly classified into two categories based on their modeling methodology, *i.e*., model-based methods and appearance-based methods.

Model-based method. Model-based approaches [2, 48, 59] aim to capture the inherent biomechanical characteristics of pedestrians. Early approaches [3, 28, 56] focused on distinguishing identities by estimating pedestrian motion parameters. However, such methods were constrained by predefined empirical points and yielded limited recognition accuracy. With the rapid advancement of deep learning, researchers have shifted their focus to extracting training data using pose estimation or Skinned Multi-Person Linear (SMPL) models. Subsequently, a well-designed network [21, 29, 30, 45, 46] is employed to further learn the gait features based on joint-based datasets. Despite substantial advancements, the efficacy of model-based methods remains constrained by the empirical design of predefined points and the estimation of results from low-quality images.

Appearance-based method. Appearance-based methods aim to learn gait features directly from silhouette data. Previous methods [18, 37, 54] have employed an efficient approach by directly applying weighted averaging to the sequence, *i.e*., Gait Energy Image (GEI). These methods have achieved significant advancements across various viewpoints and clothing conditions. However, the weighted averaging of video sequence data inevitably leads to the loss of substantial motion details. To this end, current research [22, 23, 26, 32] emphasizes the utilization of video input to learn gait features in order to address this limitation. Some of the 2D convolution-based methods [7, 14] independently extract local spatial features from each frame and subsequently utilize set-based pooling techniques, *i.e*., MaxPooling and MeanPooling, to directly extract global temporal features from the variable-length sequence. These methods have yielded impressive experimental results, significantly

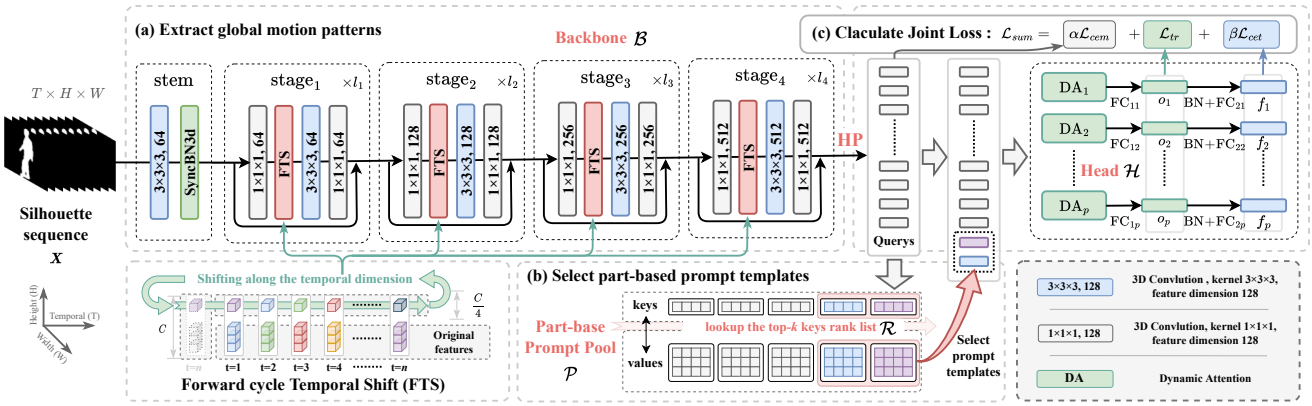

Figure 2. Overview of the proposed Visual Prompt Network (VPNet), where "HP" denotes the Horizontal Pooling. (a) VPNet incorporates the FTS module into the backbone based on ResNet, focusing on the extraction of global motion patterns without any extra parameters. (b) VPNet utilizes a trainable prompt pool comprising numerous key-value pairs, dynamically selecting prompt templates and integrating them into gait features to provide task-relevant knowledge. (c) Three different loss functions are combined for supervised network training.

advancing the field of gait recognition. However, gait represents a coordinated whole-body motion process, and processing the features of each frame individually disregards the preservation of robust local motion patterns, which are essential for accurate gait information. Consequently, some 3D convolution-based methods [6, 12, 33, 35, 36] proposed to local spatio-temporal features from adjacent frame images using a shared 3D convolution kernel, while further modeling the unfixed length sequence to obtain robust global features. The approach presented in this paper falls within the category of appearance-based methods. In contrast to previous methods, we aim to explore further improvements within the deep network structure.

2.2. Prompt Engineering

Prompt Engineering [5, 9, 40] is a methodology that involves the incorporation of additional information as a condition in the inference process using a model. It has gained considerable popularity [16, 19, 41] in the field of natural language processing (NLP) due to its ability to integrate task-specific cues for each individual task. In recent years, there has been a notable surge in prompt-based research [31, 38, 57, 62] within the field of computer vision. It integrates visual and linguistic tasks, using NLP to reference or define visual concepts. To address inherent disparities in information densities between visual and NLP tasks, researchers have developed methodologies [1, 24, 27] that involve pre-filling model inputs with adaptive parameters. These parameters, functioning as prompts, can be fine-tuned through gradient-based optimization, thereby facilitating their contribution to visual tasks. Specifically, VPT [27] and VP [1] load a set of trainable prompts into the model inputs while maintaining the architectural integrity of the core network. This strategic implementation consistently yields noteworthy performance improvements across a range of downstream tasks. In contrast, DAM-VP [24] adopts a divide-and-conquer scheme, segregating datasets associated with downstream tasks into smaller, homogeneous subsets. Each subset is endowed with its distinct cue, and subsequently, optimization is carried out independently for each subset. We extend the concept of prompt learning from prior research [52, 53], utilizing it to provide task-related information and extract robust motion patterns.

3. Methodology

In this section, we present the overall architecture of the proposed Visual Prompt Network (VPNet) for gait and provide an in-depth explanation of the motivation in Sec. 3.1. We then provide a comprehensive elucidation of the backbone architecture in Sec. 3.2. Finally, we present the specific structure of the dynamic transformer in Sec. 3.3.

3.1. Preliminary and Motivation

A vanilla vision gait recognition framework typically comprises two essential components: the backbone $\mathcal{B}$ and the head $\mathcal{H}$. Formally, given a gait silhouette sequence $\boldsymbol{X} \in \mathbb{R}^{T\times H\times W}$, where H, W, and T represent the height, width, and number of frames, respectively. Firstly, the gait sequence is fed into the backbone to extract the local motion patterns, followed by its passage through the head to derive the global motion patterns. The extraction of gait features can be briefly represented as

$$\boldsymbol{F} = \mathcal{H}(\mathcal{B}(\boldsymbol{X})), \tag{1}$$

where $\boldsymbol{F} \in \mathbb{R}^{P\times C}$ denotes the output features, P is the number of horizontal slices, and C is the feature dimension. Most of the existing research [8, 15, 33] places a significant

focus on improving the efficiency of the backbone component for extracting complex spatio-temporal features. Despite their progress, these approaches struggle to address the challenges of gait recognition in real-world scenarios. To this end, as shown in Fig. 2, we intend to introduce a comprehensive framework for gait recognition. In particular, it involves utilizing a common backbone to directly extract global motion patterns, designing a trainable prompt pool $\mathcal{P}$ that adaptively selects task-relevant templates to concatenate with the original gait features, and employing dynamic attention to learn representative gait features, *i.e.*,

$$\boldsymbol{F} = \mathcal{H}(\text{concat}[\overbrace{\mathcal{P}(\mathcal{B}(\boldsymbol{X}))}^{\text{prompt templates}}, \mathcal{B}(\boldsymbol{X})]), \tag{2}$$

where concat represents the concatenation along the temporal dimension. *The prompt templates, functioning as trainable embeddings, essentially encode the gait task. These templates include various aspects such as viewpoint, attire, carrying conditions, occlusion, and so on.* They provide instructions to the model, specifying the task it executes.

3.2. Backbone Architecture

We utilize the bottleneck structure in ResNet [20] to establish the backbone for gait recognition and introduce a plug-and-play gait temporal-shift operation [34, 36], coupled with a 3D convolutional network, as the foundational motion pattern extraction element within the bottleneck. By modifying the depth of the backbone network, we sequentially construct three models with distinct parameters in Sec. 4.1, namely VPNet-T, VPNet-M, and VPNet-L.

Temporal Shift Bottleneck. 3D convolution is commonly used in gait recognition tasks due to its proficiency in local spatio-temporal aggregation. However, its ability to be used in real-world scenarios with numerous long silhouette sequences may pose certain challenges. Inspired by temporal shift operation [34, 36], as shown in Fig. 2, we design a novel Forward cycle Temporal Shift (FTS) module for the gait recognition network. *Considering that forward walking is the predominant daily practice, whereas backward walking is less frequent, we employ a forward cyclic shifting approach along the temporal dimension instead of the zero padding technique* used in TSConv [34]. Specifically, the FTS module divides the features $f \in \mathbb{R}^{C\times T\times H\times W}$ of each pixel into four segments. It only cyclically shifts the first segment $f_1 \in \mathbb{R}^{\frac{C}{4}\times T\times H\times W}$ along the temporal dimension. The extraction of motion patterns with a global receptive field can be achieved by incorporating the FTS module before 3D convolution, without the additional parameters.

3.3. Dynamic Transformer

We present a dynamic transformer as an integral head layer $\mathcal{H}$ in the network architecture. As shown in Fig. 2, it employs prompt engineering to provide task-specific shared

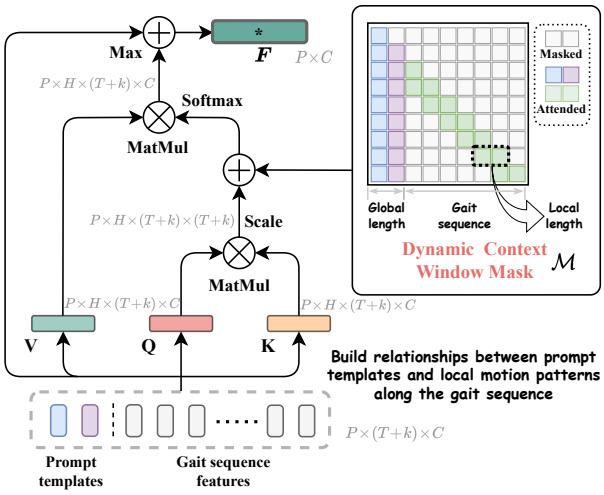

Figure 3. The dynamic attention architecture. We utilize both prompt templates and gait sequence as inputs, employing the Context Window Mask to overcome the length generalization problem.

knowledge information for gait sequence and introduces a dynamic context window mask to effectively alleviate the problem of length generalization in real-world scenarios.

Part-based Prompt Pool. Prompt learning has emerged as a methodology in NLP, and its application in gait recognition is motivated by three main considerations. Firstly, the adaptable visual prompt templates facilitate the integration of shared insights concerning the gait sequence. Secondly, the shared knowledge seamlessly incorporates into the gait sequence as a global feature through a flexible plug-and-play process. Finally, specific visual prompts are associated with salient features within the gait sequence. To address this, we introduce a part-based prompt pool with shared parameters, delivering task-specific information. The j-th prompt pool can be defined as

$$\mathcal{P}_j = \{\boldsymbol{P}_{j,1}, \boldsymbol{P}_{j,2}, \ldots, \boldsymbol{P}_{j,N} | \boldsymbol{P}_{j,i} = (k_{j,i}, v_{j,i})\}, \tag{3}$$

where $\{\mathcal{P}_j | j = 1, ..., P\}$ is the j-th prompt pool, P denotes the number of horizontal slices, $\boldsymbol{P}_{j,i}$ is the i-th prompt token, N denotes the number of trainable tokens in the prompt pool, $k_{j,i} \in \mathbb{R}^C$ and $v_{j,i} \in \mathbb{R}^C$ represent the key and value, respectively. We utilize key-value pairs as prompt tokens, matching the key with the input gait sequence. Following this alignment, the top-k values are selected as prompt templates based on their similarity, and these chosen templates are smoothly integrated with the original sequence along the temporal dimension. Specifically, given a gait sequence feature $\boldsymbol{S} \in \mathbb{R}^{P\times T\times C}$, for the j-th part sequence feature $S_j \in \mathbb{R}^{T\times C}$, we first compute the average feature along the temporal dimension $\bar{S}_j \in \mathbb{R}^C$ and then perform similarity matching with key embeddings in the prompt pool to select the top-k values with the highest similarity. Denote $\mathcal{R}$ as the function to lookup the top-k keys in the ranked list, the

selection process for the key can be represented as

$$\mathcal{K}_j = \mathcal{R}\left(\frac{\bar{S}_j \otimes k_{j,i}}{\|\bar{S}_j\| \otimes \|k_{j,i}\|}\right), \tag{4}$$

where $\mathcal{K}_j$ denotes a list of sets, generating a top-k index that is linked to the query sequence, and $\otimes$ represents matrix multiplication. The index value is employed to retrieve k values, denoted as $V_j = \{v_{j,\kappa} | \kappa \in \mathcal{K}_s\}$. Subsequently, we combine the indexed value V_j with the feature S_j, *i.e.*,

$$S_j^* = \text{concat}\left[V_j,\ S_j\right], \tag{5}$$

where concat represents the concatenation of prompt templates and gait vectors along the temporal dimension. The resulting gait sequence, denoted as $S_j^* \in \mathbb{R}^{(T+k)\times C}$, integrates the prompt information, $(T + k)$ is the length of the gait sequence, and k is empirically set to 6. Similar to the codebook collapse issue faced in VQ-VAE [10, 39], there exists a risk that all index $\mathcal{K}_j$ map to a limited set of prompt templates. To address this, we implement a randomized restart strategy: the key vector in the prompt pool is randomly reset to one of the encoder outputs if its average utilization falls below a predefined threshold ($\tau = 0.001$).

Dynamic Attention. As shown in Fig. 3, we propose to extract global motion patterns for each body part based on prompt templates using the multi-head attention [11, 47]. Formally, the j-th part sequence feature S_j^* is mapped as query $\boldsymbol{Q}$, key $\boldsymbol{K}$ and value $\boldsymbol{V}$. The mapping process is as

$$\boldsymbol{Q} = S_j^* W^Q, \boldsymbol{K} = S_j^* W^K, \boldsymbol{V} = S_j^* W^V, \tag{6}$$

where $W^Q \in \mathbb{R}^{C\times h\cdot C'}$, $W^K \in \mathbb{R}^{C\times h\cdot C'}$ and $W^V \in \mathbb{R}^{C\times h\cdot C'}$ are learnable parameters, h denotes the number of heads, and C' represents the feature dimension of each head. Then, we employ $\boldsymbol{Q}$ and $\boldsymbol{K}$ to establish similarity relationships between each frame. It is notable that prompt templates can be associated with the features of each frame, thereby giving rise to the establishment of correlations, *i.e.*,

$$\text{sim}(\boldsymbol{Q}, \boldsymbol{K}) = \boldsymbol{Q}\boldsymbol{K}^T, \tag{7}$$

where $\text{sim}(\boldsymbol{Q}, \boldsymbol{K})$ denotes the similarity matrix. Similar to natural language processing, gait recognition also faces the challenge of length generalization [17, 42, 55]. During training, a limited number of frames are randomly selected, but during testing, the longer sequence can distract the attention of the network. To effectively utilize the shared information provided by prompt templates in discerning distinctive global motion patterns within varying length sequences, we introduce a context window mask structure. This design dynamically establishes connections between prompt templates and gait sequence. As depicted in Fig. 3, the architecture comprises a global branch on the left and a local branch on the right. The global branch allows each feature in the sequence to focus on establishing relationships with prompt templates, while the local branch fosters associations between each feature and other features within a defined local length range. Features outside of these two branches are excluded from the attention process, *i.e.*,

$$\text{Att}(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}) = \text{softmax}(\frac{\text{sim}(\boldsymbol{Q}, \boldsymbol{K})}{\sqrt{C'}} \oplus \mathcal{M})\boldsymbol{V}, \tag{8}$$

where $\sqrt{C'}$ is the scaling factor, $\mathcal{M}$ is the dynamic context window mask, and $\oplus$ denotes element-wise sum operation.

3.4. Learning Details

In this work, we introduce three loss functions during training: triplet loss $\mathcal{L}_{tp}$, cross-entropy loss $\mathcal{L}_{cet}$, and cosine embedding loss $\mathcal{L}_{cem}$ in Fig. 2. The $\mathcal{L}_{tp}$ and $\mathcal{L}_{cet}$ serve to supervise part-based features individually, while the $\mathcal{L}_{cem}$ is applied to supervise the key embedding $(k_{j,i})$ within the prompt pool. $\mathcal{L}_{cem}$ requires that the selected key embedding be closely related to the query sequence $\bar{S}_i$, *i.e.*,

$$\mathcal{L}_{cem} = \underset{\kappa\in\mathcal{K}_j}{\arg\min} \sum_{i=1}^{P} \frac{-1}{N_{cem}} \left(\frac{\bar{S}_i \otimes k_{j,\kappa}}{\|\bar{S}_i\| \otimes \|k_{j,\kappa}\|}\right), \tag{9}$$

where N_{cem} is a positive integer obtained by multiplying the batch size with the number of parts. In each training step, we adhere to the previously described strategy to select $\mathcal{K}_s$. The selected prompt templates are then merged with gait sequence vectors and fed into the dynamic transformer and BNNeck [15] to extract representative gait features. In general, the training process involves the use of a combined loss following the end-to-end method, *i.e.*,

$$\mathcal{L}_{sum} = \mathcal{L}_{tp} + \alpha\mathcal{L}_{cem} + \beta\mathcal{L}_{cet}, \tag{10}$$

where α and β are the hyper-parameters to balance the combined loss, α is empirically set to 0.01 and β is set to 0.3.

4. Experiments

4.1. Datasets and Implementation Details

The gait datasets are divided into two subsets, *i.e.*, in-the-wild and in-the-lab, distinguished by their respective collection environments in Tab. 1. A standardized preprocessing procedure [7] is uniformly applied to these datasets, and the data sizes used in this paper are all resized to 64×64.

In-the-wild datasets. In-the-wild datasets introduce numerous unforeseen challenges, such as occlusions, clothing changes, viewpoint variations, and carrying conditions. Pedestrians may also pause or alter their walking patterns. These datasets are essential benchmarks for evaluating gait recognition methods under various external disturbances.

Table 1. Comparison of the specific values of the adopted datasets and corresponding parameters. Where "#Ide., #Seq., #Cam." refer to numbers of identities, sequences, and cameras respectively. Additionally, "lr, dr, wd" denote learning rate, decay rate, and weight decay.

Environment	Datasets	Train		Test		# Cam.	Batch Size	Optimizer	Scheduler	
		# Ide.	# Seq.	# Ide.	# Seq.				step size	epochs
In-the-wild	Gait3D [61]	3,000	18,940	1,000	6,369	Diverse	(32, 4)	SGD	20	80
	GREW [63]	20,000	102,887	6,000	24,000	Diverse	(64, 4)	$lr = 0.1$	50	200
In-the-lab	OUMVLP [44]	5,153	144,284	5,154	144,312	14	(32, 8)	$dr = 0.1$	50	200
	CASIA-B [58]	74	8,140	50	5,500	11	(8, 16)	$wd = 0.0005$	10	40

Table 2. The detailed structure of the backbone in VPNet-L, where FTSConv denotes the combined of FTS module and convolution.

Layer			Output size
conv_1	$3 \times 3 \times 3$, 64, stride $(1, 1, 1)$		$T \times H \times W$
stage_1	$\begin{bmatrix} 1 \times 1 \times 1, & 64, & (0,0,0) \\ \text{FTSConv}, & 64, & (1,1,1) \\ 1 \times 1 \times 1, & 64, & (0,0,0) \end{bmatrix}$	$\times 3$	$T \times H \times W$
stage_2	$\begin{bmatrix} 1 \times 1 \times 1, & 128, & (0,0,0) \\ \text{FTSConv}, & 128, & (1,2,2) \\ 1 \times 1 \times 1, & 128, & (0,0,0) \end{bmatrix}$	$\times 4$	$T \times \frac{H}{2} \times \frac{W}{2}$
stage_3	$\begin{bmatrix} 1 \times 1 \times 1, & 256, & (0,0,0) \\ \text{FTSConv}, & 256, & (1,2,2) \\ 1 \times 1 \times 1, & 256, & (0,0,0) \end{bmatrix}$	$\times 6$	$T \times \frac{H}{4} \times \frac{W}{4}$
stage_4	$\begin{bmatrix} 1 \times 1 \times 1, & 512, & (0,0,0) \\ \text{FTSConv}, & 512, & (1,1,1) \\ 1 \times 1 \times 1, & 512, & (0,0,0) \end{bmatrix}$	$\times 3$	$T \times \frac{H}{4} \times \frac{W}{4}$

We evaluate our method on the widely adopted Gait3D [61] and GREW [63] datasets in Tab. 1. *To enhance the robustness of our model, we employ three data augmentation techniques: rotation, flipping, and perspective.* We conduct training using VPNet-L on both Gait3D and GREW datasets and report rank-1 (%) and rank-5 (%) accuracy.

In-the-lab datasets. In-the-lab datasets encompass various controllable conditions, including camera angles, predetermined walking routes, and dress variations. These datasets play a pivotal role in evaluating the ability of the gait network to handle cross-viewing and cross-dressing scenarios. We select the well-established OUMVLP [44] and CASIA-B [58] datasets to evaluate our approach in Tab. 1. Specifically, the OUMVLP provides an abundance of cross-viewing data, while the CASIA-B offers valuable cross-dressing data. *Limited datasets, particularly those obtained from laboratory scenarios, may result in overfitting if the network parameters are excessively large.* Therefore, we employ VPNet-M for training on OUMVLP and VPNet-T for training on CASIA-B. We present the rank-1 (%) accuracy for all perspectives, excluding the identical-view cases.

Implementation Details. We establish the network parameters based on the dataset scale and complexity, as outlined in Tab. 1. The batch size is represented as $(\mathcal{S}, \mathcal{T})$, indicating that each mini-batch involves $\mathcal{S}$ subjects. For each subject, $\mathcal{T}$ sequences are sampled, and within each sequence, a random selection of 20 to 40 frames is chosen following DANet [35]. The optimization strategy employs Stochastic Gradient Descent (SGD) with an initial learning rate of 0.1, and each epoch consists of 1000 iterations. The learning rate undergoes periodic reduction based on the *step size* in Tab. 1, with a *decay rate* of 0.1. The entire model is built by sequentially stacking the temporal shift bottleneck in Sec. 3.2. We construct three models with distinct parameters, namely VPNet-T, VPNet-M, and VPNet-L, by modifying the depth of the backbone network. Specifically, VPNet-T is composed of three stages, each comprising a specific number of layers (1, 1, 1) and channels $(64, 128, 256)$. In contrast, both VPNet-M and VPNet-L consist of four stages with channel configurations (64, 128, 256, 512) in Tab. 2. VPNet-M employs a layer stack of (2, 2, 2, 2), while VPNet-L utilizes a stack of layers (3, 4, 6, 3).

Table 3. The performance comparisons on Gait3D are reported with rank-1 and rank-5 accuracy (%).

Methods	Publication	rank-1 (%)	rank-5 (%)
GaitSet [7]	AAAI 2019	36.7	58.3
GaitPart [14]	CVPR 2020	28.2	47.6
GLN [22]	ECCV 2020	31.4	52.9
GaitGL [33]	ICCV 2021	29.7	48.5
CSTL [25]	ICCV 2021	11.7	19.2
SMPLGait [61]	CVPR 2022	46.3	64.5
GaitBase [15]	CVPR 2023	65.6	-
GaitGCI [13]	CVPR 2023	50.3	68.5
DANet [35]	CVPR 2023	48.0	69.7
HSTL [50]	ICCV 2023	61.3	76.3
DyGait [51]	ICCV 2023	66.3	80.8
VPNet-L	-	**75.4**	**87.1**

Table 4. The performance comparisons on GREW are reported with rank-1 and rank-5 accuracy (%).

Methods	Publication	rank-1 (%)	rank-5 (%)
GaitSet[7]	AAAI 2019	46.3	63.6
GaitPart [14]	CVPR 2020	44.0	60.7
GaitGL [33]	ICCV 2021	47.3	63.6
GaitGraph [45]	ICIP 2021	1.3	3.5
GaitBase [15]	CVPR 2023	60.1	-
GaitGCI [13]	CVPR 2023	68.5	80.8
HSTL [50]	ICCV 2023	62.7	76.6
DyGait [51]	ICCV 2023	71.4	83.2
VPNet-L	-	**80.0**	**89.4**

Table 5. Rank-1 accuracy (%) on OUMVLP under all view angles, excluding the identical-views cases.

Method	Probe View														Mean
	0°	15°	30°	45°	60°	75°	90°	180°	195°	210°	225°	240°	255°	270°	
GaitSet [7]	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 [14]	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 [22]	83.8	90.0	91.0	91.2	90.3	90.0	89.4	85.3	89.1	90.5	90.6	89.6	89.3	88.5	89.2
GaitGL [33]	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
GaitBase [15]	-	-	-	-	-	-	-	-	-	-	-	-	-	-	90.0
GaitGCI [13]	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
DANet [35]	87.7	91.3	91.6	91.8	91.7	91.4	91.1	90.4	90.3	90.7	90.9	90.5	90.3	89.9	90.7
HSTL [50]	91.4	92.9	**92.7**	**93.0**	92.9	**92.5**	**92.5**	92.7	92.3	**92.1**	92.3	92.2	91.8	91.8	**92.4**
VPNet-M	**91.9**	**93.0**	92.4	92.7	**93.2**	**92.5**	92.3	**92.9**	**92.4**	91.9	92.1	**92.5**	**91.9**	**91.9**	**92.4**

Table 6. The performance comparisons on CASIA-B are reported with rank-1 accuracy (%), excluding the identical-view cases.

Methods	Publication	NM	BG	CL	Mean
GaitSet[7]	AAAI 2019	95.0	87.2	70.4	84.2
GaitPart[14]	CVPR 2020	96.2	91.5	78.7	88.8
GLN [22]	ECCV 2020	96.9	94.0	77.5	89.5
MT3D [32]	MM 2020	96.7	93.0	81.5	90.4
GaitGL [33]	ICCV 2021	97.4	94.5	83.6	91.8
GaitBase [15]	CVPR 2023	97.6	94.0	77.4	89.8
GaitGCI-T [13]	CVPR 2023	97.9	95.0	86.4	93.1
DANet [35]	CVPR 2023	98.0	95.9	89.9	94.6
HSTL [50]	ICCV 2023	98.1	95.9	88.9	94.3
DyGait [51]	ICCV 2023	**98.4**	96.2	87.8	94.1
VPNet-T	-	98.3	**96.3**	**90.0**	**94.9**

4.2. Results under in-the-wild Scenario

Gait3D. VPNet-L demonstrates superior performance in terms of rank-1 and rank-5 accuracy in Tab. 3, surpassing state-of-the-art methods. Notably, it exhibits a **9.1**% improvement over DyGait [51], a silhouette-based approach, and a remarkable **29.1**% improvement compared to SMPLGait [61], a multimodal method. These experimental results underscore the efficacy of our method in extracting robust features in the presence of diverse external disturbances.

GREW. VPNet-L achieves the advanced accuracy of rank-1 and rank-5 as shown in Tab. 4, outperforming GaitGCI [13] by **11.5**% and GaitGraph [45] by **79.7**%. These experimental results demonstrate that our method effectively handles various unexpected external factors in realistic scenes, yielding effective global motion patterns in gait sequences.

Summary. We specifically design VPNet-L to tackle the intricacies of outdoor scenes, leading to state-of-the-art performance on real-world datasets. **(1)** Our method introduces a unified backbone architecture designed for the extraction of global motion patterns in real-world scenarios. **(2)** Prompt templates convey task-related knowledge, and a dynamic transformer is subsequently employed to effectively alleviate the problem of length generalization. **(3)** Experimental results validate the effectiveness of the proposed framework in the wild datasets.

4.3. Results under in-the-lab Scenario

OUMVLP. The experimental results of VPNet-M in Tab. 5 exhibit superior performance compared to the majority of methods, underscoring its robustness and effectiveness. Notably, VPNet-M attains average accuracies on par with HSTL [50]. It is noteworthy that the accuracy of VPNet-M is further improved after excluding invalid probe sequences.

CASIA-B. The experimental outcomes in Tab. 6 underscore the competitive performance of VPNet-T, surpassing other gait recognition methods. It is worth noting that, excluding the cases of identical views, VPNet-T achieved an average accuracy of over **90**% in cross-viewing and cross-dressing.

Summary. We customize VPNet-T and VPNet-M based on data size and complexity, achieving state-of-the-art performance on in-the-lab datasets. **(1)** Previously, different depth networks have been designed on unified architectures are common, such as ResNet-18 [20] and ResNet-50 [20]. Our proposed VPNet architecture fills a gap in the design of network architectures for gait recognition across diverse datasets. **(2)** In indoor scenarios, performance has reached its peak due to the utilization of continuously improving methods. Therefore, it is expected that gait recognition will increasingly concentrate on outdoor scenarios in the future.

4.4. Ablation Study

Effectiveness of core designs. We perform ablation experiments of the proposed module in Gait3D [61] and CASIA-B [58] datasets. **(1)** The experimental results reveal that the core design module, comprising the FTS module, prompt pool, and dynamic attention module, notably influenced cross-dressing and carrying conditions within the CASIA-B dataset. In contrast, the introduced module exhibits a more pronounced impact on Gait3D. **(2)** The results show that replacing the dynamic attention with the max+mean maintains competitive accuracy in CASIA-B. However, the results show a significant decrease in Gait3D, highlighting the

Table 7. The ablation study on Gait3D and CASIA-B are reported with rank-1 accuracy(%), excluding the identical-views cases.

Method	Gait3D	CASIA-B		
	R-1 (%)	NM (%)	BG (%)	CL (%)
VPNet	**75.4**	**98.3**	96.3	**90.0**
Analysis of each component in VPNet				
- *w/o* FTS	73.2	98.0	96.1	89.2
- *w/o* prompt	72.8	98.0	96.0	89.1
- *w/o* mask	73.0	98.3	96.3	90.0
w max+mean	72.1	98.0	95.7	88.5
Analysis of temporal shift bottleneck				
2D Conv	69.7	98.0	95.7	86.8
3D Conv	73.2	98.0	96.1	89.2
2D Conv + FTS	74.1	98.1	96.2	88.2
3D Conv + FTS	75.4	98.3	96.3	90.0
Analysis of the local length in dynamic attention				
l=8	74.9	98.0	96.3	89.8
l=16	75.4	98.1	**96.4**	89.8
l=32	73.2	98.3	96.3	90.0

effectiveness of our module in real-world scenarios.

Analysis of temporal shift bottleneck. We perform an empirical analysis comparing convolution outcomes within the temporal shift bottleneck in Tab. 7. **(1)** The network utilizing 2D convolution experienced a noteworthy decline in performance, representing a substantial improvement over using the FTS module in conjunction with 2D convolution. These results indicate that the FTS module plays a role in acquiring global motion patterns. **(2)** The network utilizing 3D convolution effectively captures local motion patterns, whereas the incorporation of the FTS module into 3D convolution facilitates the extraction of global motion patterns.

Analysis of the local length. Different length affects the receptive fields in dynamic attention. **(1)** The experimental results demonstrate that an optimal range of receptive fields (l=16) is essential for real data, as an excessively long receptive field (l=32) may introduce noise. **(2)** In laboratory scenarios with relatively short sequence lengths, the impact of local receptive fields is minimal. **(3)** The issue of training and testing inconsistency can be alleviated by incorporating local receptive fields in the real world.

Visualization of the prompt selection. We quantify the frequency of prompt template selection across various viewpoints and dress conditions on CASIA-B, as detailed in Fig. 4. The visualization results reveal variability in prompt template selection influenced by external factors, illustrating the encoding of task-relevant knowledge. **(1)** The statistical results show that different dressings and viewpoints have obvious differences in the frequency of prompt template selection. **(2)** The visualization results reveal variability in prompt template selection influenced by external factors, illustrating the encoding of task-relevant knowledge.

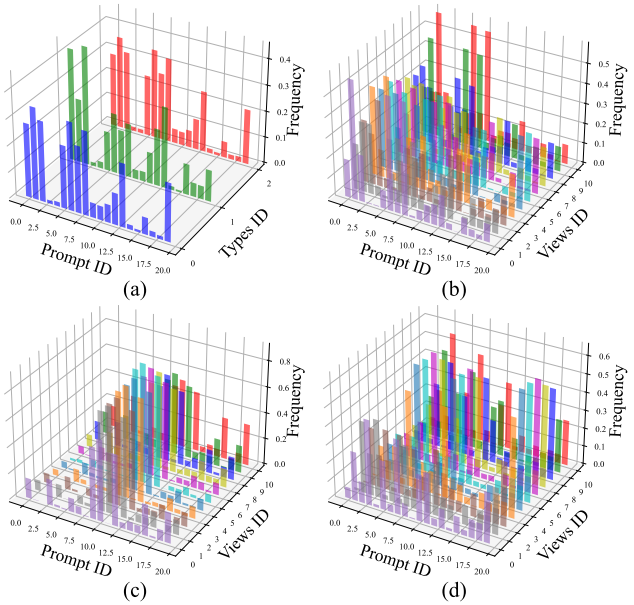

Figure 4. Histogram of prompt template selection on CASIA-B, where "Prompt ID", "View ID", and "Type ID" denote the index of prompt, view angle, and dressing, respectively. (a) Comparison of the selection frequency of prompt templates under cross-dressing and carrying conditions. (b)(c)(d) Comparison of prompt template selection frequency across various views and body parts.

5. Conclusion and Limitations

In this paper, we present a Visual Prompt Network (VPNet) for gait recognition. VPNet employs temporal shift bottleneck as the foundational module of a backbone network to extract global motion patterns and utilizes prompt engineering to adaptively select templates with corresponding prompt information tailored for different gait sequences. To enhance the robustness of global motion patterns, the relationships between these templates and global features of gait sequence are further established using dynamic transformer structures. VPNet demonstrates superior performance in both in-the-wild and in-the-lab scenarios, highlighting its significant potential in real-world applications.

Limitations. Prompt learning has gained prominence in natural language processing for retraining large-scale pre-trained language models, whereas the proposed network does not incorporate pre-training. In future work, we will investigate the applications of prompt engineering for large-scale gait recognition pre-training in real-world scenarios.

6. Acknowledgement

This work was supported by the National Key R&D Program of China (2022YFC3300700), the National Natural Science Foundation of China (62331006, 62171038, 62088101, 62276031, 62276025, and 62276025), and the Fundamental Research Funds for the Central Universities.

References

[1] Hyojin Bahng, Ali Jahanian, Swami Sankaranarayanan, and Phillip Isola. Exploring visual prompts for adapting large-scale models. *arXiv preprint arXiv:2203.17274*, 2022. 3

[2] A.F. Bobick and A.Y. Johnson. Gait recognition using static, activity-specific parameters. In *CVPR*, pages I–I, 2001. 2

[3] Robert Bodor, Andrew Drenner, Duc Fehr, Osama Masoud, and Nikolaos Papanikolopoulos. View-independent human motion classification using image-based reconstruction. *Image Vision Comput.*, 27(8):1194–1206, 2009. 2

[4] Dennis M Bramble and David R Carrier. Running and breathing in mammals. *Science*, 219(4582):251–256, 1983. 1

[5] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *NeurIPS*, 33:1877–1901, 2020. 3

[6] Tianrui Chai, Annan Li, Shaoxiong Zhang, Zilong Li, and Yunhong Wang. Lagrange motion analysis and view embeddings for improved gait recognition. In *CVPR*, pages 20249–20258, 2022. 3

[7] Hanqing Chao, Yiwei He, Junping Zhang, and Jianfeng Feng. Gaitset: Regarding gait as a set for cross-view gait recognition. In *AAAI*, pages 8126–8133, 2019. 1, 2, 5, 6, 7

[8] Hanqing Chao, Kun Wang, Yiwei He, Junping Zhang, and Jianfeng Feng. Gaitset: Cross-view gait recognition through utilizing gait as a deep set. *IEEE TPAMI*, 44(7):3467–3478, 2021. 3

[9] Yulong Chen, Yang Liu, Li Dong, Shuohang Wang, Chenguang Zhu, Michael Zeng, and Yue Zhang. Adaprompt: Adaptive model training for prompt-based nlp. *arXiv preprint arXiv:2202.04824*, 2022. 3

[10] Prafulla Dhariwal, Heewoo Jun, Christine Payne, Jong Wook Kim, Alec Radford, and Ilya Sutskever. Jukebox: A generative model for music. *arXiv preprint arXiv:2005.00341*, 2020. 5

[11] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 5

[12] Huanzhang Dou, Pengyi Zhang, Wei Su, Yunlong Yu, and Xi Li. Metagait: Learning to learn an omni sample adaptive representation for gait recognition. In *ECCV*, pages 357–374, 2022. 3

[13] Huanzhang Dou, Pengyi Zhang, Wei Su, Yunlong Yu, Yining Lin, and Xi Li. Gaitgci: Generative counterfactual intervention for gait recognition. In *CVPR*, pages 5578–5588, 2023. 6, 7

[14] Chao Fan, Yunjie Peng, Chunshui Cao, Xu Liu, Saihui Hou, Jiannan Chi, Yongzhen Huang, Qing Li, and Zhiqiang He. Gaitpart: Temporal part-based model for gait recognition. In *CVPR*, pages 14225–14233, 2020. 1, 2, 6, 7

[15] Chao Fan, Junhao Liang, Chuanfu Shen, Saihui Hou, Yongzhen Huang, and Shiqi Yu. Opengait: Revisiting gait recognition towards better practicality. In *CVPR*, pages 9707–9716, 2023. 3, 5, 6, 7

[16] Tianyu Gao, Adam Fisch, and Danqi Chen. Making pre-trained language models better few-shot learners. *arXiv preprint arXiv:2012.15723*, 2020. 3

[17] Chi Han, Qifan Wang, Wenhan Xiong, Yu Chen, Heng Ji, and Sinong Wang. Lm-infinite: Simple on-the-fly length generalization for large language models. *arXiv preprint arXiv:2308.16137*, 2023. 2, 5

[18] Jinguang Han and Bir Bhanu. Individual recognition using gait energy image. *IEEE TPAMI*, 28(2):316–322, 2005. 2

[19] Xu Han, Weilin Zhao, Ning Ding, Zhiyuan Liu, and Maosong Sun. Ptr: Prompt tuning with rules for text classification. *AI Open*, 3:182–192, 2022. 3

[20] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016. 4, 7

[21] Nadia Hosni and Boulbaba Ben Amor. A geometric convnet on 3d shape manifold for gait recognition. In *CVPRW*, 2020. 2

[22] Saihui Hou, Chunshui Cao, Xu Liu, and Yongzhen Huang. Gait lateral network: Learning discriminative and compact representations for gait recognition. In *ECCV*, pages 382–398, 2020. 1, 2, 6, 7

[23] Hung-Min Hsu, Yizhou Wang, Cheng-Yen Yang, Jenq-Neng Hwang, Hoang Le Uyen Thuc, and Kwang-Ju Kim. Gaittake: Gait recognition by temporal attention and keypoint-guided embedding. In *ICIP*, pages 2546–2550, 2022. 2

[24] Qidong Huang, Xiaoyi Dong, Dongdong Chen, Weiming Zhang, Feifei Wang, Gang Hua, and Nenghai Yu. Diversity-aware meta visual prompting. In *CVPR*, pages 10878–10887, 2023. 3

[25] Xiaohu Huang, Duowang Zhu, Hao Wang, Xinggang Wang, Bo Yang, Botao He, Wenyu Liu, and Bin Feng. Context-sensitive temporal feature learning for gait recognition. In *ICCV*, pages 12909–12918, 2021. 6

[26] Xiaohu Huang, Xinggang Wang, Botao He, Shan He, Wenyu Liu, and Bin Feng. Star: Spatio-temporal augmented relation network for gait recognition. *IEEE Transactions on Biometrics, Behavior, and Identity Science*, pages 115–125, 2022. 2

[27] Menglin Jia, Luming Tang, Bor-Chun Chen, Claire Cardie, Serge Belongie, Bharath Hariharan, and Ser-Nam Lim. Visual prompt tuning. In *ECCV*, pages 709–727, 2022. 3

[28] Worapan Kusakunniran, Qiang Wu, Hongdong Li, and Jian Zhang. Multiple views gait recognition using view transformation model based on optimized gait energy image. In *ICCV*, pages 1058–1064, 2009. 2

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

[30] Xiang Li, Yasushi Makihara, Chi Xu, and Yasushi Yagi. End-to-end model-based gait recognition using synchronized multi-view pose constraint. In *ICCV*, pages 4106–4115, 2021. 2

[31] Yaowei Li, Ruijie Quan, Linchao Zhu, and Yi Yang. Efficient multimodal fusion via interactive prompting. In *CVPR*, pages 2604–2613, 2023. 3

[32] Beibei Lin, Shunli Zhang, and Feng Bao. Gait recognition with multiple-temporal-scale 3d convolutional neural network. In *ACM MM*, pages 3054–3062, 2020. 2, 7

[33] Beibei Lin, Shunli Zhang, and Xin Yu. Gait recognition via effective global-local feature representation and local temporal aggregation. In *ICCV*, pages 14648–14656, 2021. 1, 3, 6, 7

[34] Ji Lin, Chuang Gan, and Song Han. Tsm: Temporal shift module for efficient video understanding. In *ICCV*, pages 7083–7093, 2019. 4

[35] Kang Ma, Ying Fu, Dezhi Zheng, Chunshui Cao, Xuecai Hu, and Yongzhen Huang. Dynamic aggregated network for gait recognition. In *CVPR*, pages 22076–22085, 2023. 1, 2, 3, 6, 7

[36] Kang Ma, Ying Fu, Dezhi Zheng, Yunjie Peng, Chunshui Cao, and Yongzhen Huang. Fine-grained unsupervised domain adaptation for gait recognition. In *ICCV*, pages 11313–11322, 2023. 1, 3, 4

[37] Yasushi Makihara, Atsuyuki Suzuki, Daigo Muramatsu, Xiang Li, and Yasushi Yagi. Joint intensity and spatial metric learning for robust gait recognition. In *CVPR*, pages 5705–5715, 2017. 2

[38] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *Proc. Int. Conf. Mach. Learn.*, pages 8748–8763, 2021. 3

[39] Ali Razavi, Aaron Van den Oord, and Oriol Vinyals. Generating diverse high-fidelity images with vq-vae-2. *NeurIPS*, 2019. 5

[40] Timo Schick and Hinrich Schütze. Exploiting cloze questions for few shot text classification and natural language inference. *arXiv preprint arXiv:2001.07676*, 2020. 3

[41] Taylor Shin, Yasaman Razeghi, Robert L Logan IV, Eric Wallace, and Sameer Singh. Autoprompt: Eliciting knowledge from language models with automatically generated prompts. *arXiv preprint arXiv:2010.15980*, 2020. 3

[42] Yutao Sun, Li Dong, Barun Patra, Shuming Ma, Shaohan Huang, Alon Benhaim, Vishrav Chaudhary, Xia Song, and Furu Wei. A length-extrapolatable transformer. *arXiv preprint arXiv:2212.10554*, 2022. 2, 5

[43] D Sutherland. The development of mature gait. *Gait & posture*, 6(2):163–170, 1997. 1

[44] 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: 1–14, 2018. 1, 2, 6

[45] Torben Teepe, Ali Khan, Johannes Gilg, Fabian Herzog, Stefan Hörmann, and Gerhard Rigoll. Gaitgraph: graph convolutional network for skeleton-based gait recognition. In *ICIP*, 2021. 2, 6, 7

[46] Torben Teepe, Johannes Gilg, Fabian Herzog, Stefan Hörmann, and Gerhard Rigoll. Towards a deeper understanding of skeleton-based gait recognition. In *CVPRW*, pages 1569–1577, 2022. 2

[47] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *NeurIPS*, 30, 2017. 2, 5

[48] David Kenneth Wagg and Mark S Nixon. On automated model-based extraction and analysis of gait. In *IEEE Int. Conf. Aut. Fac. Ges. Recog.*, pages 11–16, 2004. 2

[49] Liang Wang, Tieniu Tan, Huazhong Ning, and Weiming Hu. Silhouette analysis-based gait recognition for human identification. *IEEE TPAMI*, 25(12):1505–1518, 2003. 1

[50] Lei Wang, Bo Liu, Fangfang Liang, and Bincheng Wang. Hierarchical spatio-temporal representation learning for gait recognition. In *ICCV*, pages 19639–19649, 2023. 6, 7

[51] 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 *ICCV*, pages 13424–13433, 2023. 6, 7

[52] Zifeng Wang, Zizhao Zhang, Sayna Ebrahimi, Ruoxi Sun, Han Zhang, Chen-Yu Lee, Xiaoqi Ren, Guolong Su, Vincent Perot, Jennifer Dy, et al. Dualprompt: Complementary prompting for rehearsal-free continual learning. In *ECCV*, pages 631–648. Springer, 2022. 3

[53] Zifeng Wang, Zizhao Zhang, Chen-Yu Lee, Han Zhang, Ruoxi Sun, Xiaoqi Ren, Guolong Su, Vincent Perot, Jennifer Dy, and Tomas Pfister. Learning to prompt for continual learning. In *CVPR*, pages 139–149, 2022. 3

[54] 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 TPAMI*, 39 (2):209–226, 2016. 2

[55] Guangxuan Xiao, Yuandong Tian, Beidi Chen, Song Han, and Mike Lewis. Efficient streaming language models with attention sinks. *arXiv preprint arXiv:2309.17453*, 2023. 2, 5

[56] ChewYean Yam, Mark S Nixon, and John N Carter. Automated person recognition by walking and running via model-based approaches. *PR*, 37(5):1057–1072, 2004. 2

[57] Hantao Yao, Rui Zhang, and Changsheng Xu. Visual-language prompt tuning with knowledge-guided context optimization. In *CVPR*, pages 6757–6767, 2023. 3

[58] Shiqi Yu, Daoliang Tan, and Tieniu Tan. A framework for evaluating the effect of view angle, clothing and carrying condition on gait recognition. In *ICPR*, pages 441–444, 2006. 1, 2, 6, 7

[59] Cun Zhang, Xing-Peng Chen, Guo-Qiang Han, and Xiang-Jie Liu. Spatial transformer network on skeleton-based gait recognition. *arXiv preprint arXiv:2204.03873*, 2022. 2

[60] Yuqi Zhang, Yongzhen Huang, Liang Wang, and Shiqi Yu. A comprehensive study on gait biometrics using a joint cnn-based method. *PR*, 93:228–236, 2019. 1

[61] 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 *CVPR*, pages 20228–20237, 2022. 1, 2, 6, 7

[62] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *IJCV*, 130(9):2337–2348, 2022. 3

[63] 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 *ICCV*, pages 14789–14799, 2021. 1, 2, 6