ViSA-Gait: Leveraging Vision Foundation Models for Semantic Anchored Gait Recognition

Xiangru Li[1,5], **Deqiang Yin**[2,5], **Yifan Xie**[3], **Guojian Li**[5,4], **Zebang Cheng**[4,5] and **Fei Ma**[5*]
[1]Guangdong University of Technology
[2]Xi'an Jiaotong-Liverpool University
[3]Tsinghua University
[4]Shenzhen University
[5]Guangdong Laboratory of Artificial Intelligence and Digital Economy (SZ)
2112404378@mail2.gdut.edu.cn, deqiang.yin24@student.xjtlu.edu.cn, ivanxie416@gmail.com, liguojian@gml.ac.cn, zebang.cheng@gmail.com, mafei@gml.ac.cn

Abstract

Gait recognition has achieved remarkable success in constrained environments, yet its performance often degrades significantly in cross-domain and cross-vertical-view scenarios. This is primarily due to the fact that domain-specific silhouette geometry causes models to overfit to extrinsic geometric characteristics rather than learning generalizable motion patterns. To address this issue, we present **ViSA-Gait**, which utilizes Vision Foundation Models (VFMs) as a "Semantic Compass" for stable universal human body knowledge guidance, enabling a lightweight backbone to robustly capture fine-grained gait dynamics. Specifically, we introduce a token-based distillation mechanism where a Spatial Token Learner (STL) and a Temporal Token Learner (TTL) filter dense VFM features into motion-consistent descriptors. These descriptors are then adaptively injected into a 3D-CNN backbone via a Gated Cross-Attention (GCA) mechanism, functioning as "Semantic Anchors" that regularize the feature space and guide the model to focus on intrinsic body motion. Extensive experiments demonstrate that ViSA-Gait achieves SOTA performance on cross-domain and cross-vertical-view benchmarks while remaining competitive within-domain, offering a new perspective on bridging the gap between geometry-based analysis and universal semantic understanding.

1 Introduction

Gait recognition has emerged as a pivotal behavioral biometric for long-range, non-intrusive, and continuous person identification [Sepas-Moghaddam and Etemad, 2022; Shen *et al.*, 2024]. Compared with face or fingerprint recognition [Marasco and Ross, 2014; Zhu *et al.*, 2016; Cheng *et al.*, 2024; Xue *et al.*, 2025], gait can be captured without subject cooperation and remains reliable under low-resolution or low-light conditions, making it indispensable for surveillance and public security applications.

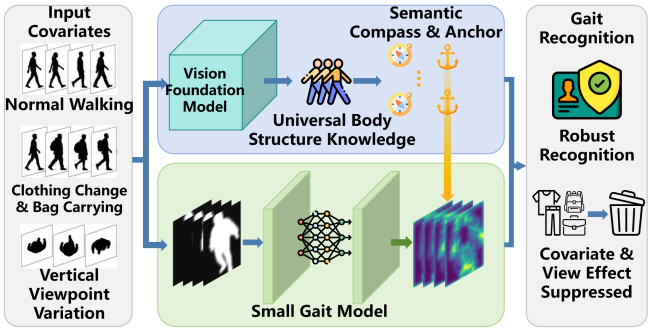

Figure 1: Motivation of ViSA-Gait. ViSA-Gait utilizes VFMs as a "Semantic Compass" for stable universal human body knowledge guidance. These universal semantic priors are distilled and adaptively injected into the backbone to function as robust "Semantic Anchors", ensuring consistent identity representation across cross-domain and cross-vertical-view shifts.

Enhancing the intrinsic robustness of gait recognition is fundamental for transitioning from laboratory success to practical real-world deployment. Such reliability hinges on the model's capability to maintain discriminative power across disparate domains and handle extreme perspective shifts. Despite substantial progress on controlled indoor datasets, achieving this robustness remains a formidable challenge, primarily due to the fragility of cross-domain and cross-vertical-view generalization [Shen *et al.*, 2024; Singloo *et al.*, 2024; Li *et al.*, 2025]. Conventional gait models, predominantly based on CNNs or Transformers, rely heavily on local spatial-temporal cues and the precise geometry of silhouettes. However, these geometric representations are highly sensitive to viewpoint shifts. As camera angles change, the 2D projection of the human body undergoes significant distortion, leading to what we define as "geometric collapse". In such scenarios, the model tends to overfit to extrinsic geometric artifacts rather than learning transferable motion patterns [Ma *et al.*, 2023; Ma *et al.*, 2024], which severely limits the adaptability of existing methods to unseen

*Corresponding author.

domains.

Meanwhile, Vision Foundation Models (VFMs) such as ViT [Dosovitskiy *et al.*, 2021] and DINOv2 [Oquab *et al.*, 2023], pretrained on billions of diverse images, have demonstrated exceptional capability in learning semantically rich and broadly invariant visual representations[Ma *et al.*, 2025; Xie *et al.*, 2025]. We argue that while the geometric representation of a silhouette often suffers from severe "geometric collapse" under extreme vertical-view changes or fails to generalize across disparate domains, the underlying anatomical semantics and human motion logic remain remarkably consistent. VFMs possess a "semantic common sense" that can recognize body structures even under extreme distortion, offering a promising direction: Leveraging VFMs as Semantic Compass to stabilize and guide the feature extraction of compact, task-specific gait models. However, directly applying VFMs to gait recognition is non-trivial due to their single-frame design, high computational cost, and the potential mismatch between universal semantics and fine-grained motion modeling.

To effectively bridge these modality and task gaps, we propose **ViSA-Gait**, a novel Hybrid Semantic Fusion framework that systematically injects high-level universal human body knowledge from VFMs into a lightweight 3D-CNN backbone. Rather than a naive feature replacement, ViSA-Gait utilizes the VFM as a "Semantic Compass" to provide view-invariant guidance for regularizing the task-specific feature space.

We address the single-frame limitation of VFMs through a dual-tokenizer architecture: a Spatial Token Learner (STL) is first employed to distill dense VFM features into hierarchical structural body descriptors, ensuring structural integrity even under the severe geometric collapse typically observed in aerial perspectives. Complementing this, a Temporal Token Learner (TTL) aggregates these spatial cues to capture motion-consistent "Semantic Anchors" that remain stable across disparate domains. These anchors are then adaptively integrated into the gait backbone via a Gated Cross-Attention (GCA) mechanism, which prevents representation contamination while facilitating multi-level semantic distillation. By enforcing an autonomous gating strategy, ViSA-Gait preserves fine-grained identity nuances in well-controlled views while prioritizing stable semantic anchors in unconstrained scenarios. This synergy effectively guides the model to focus on intrinsic body motion rather than overfitting to domain-specific geometric artifacts.

Our contributions are summarized as follows:

- We propose ViSA-Gait, a novel VFM-Empowered paradigm anchoring gait representations to universal semantics from VFMs. By treating VFMs as a "Semantic Compass", our framework effectively overcomes silhouette geometric fragility in unconstrained environments.
- We design a dual-stage architecture comprising Spatial/Temporal Token Learners and Gated Cross-Attention. This mechanism adaptively distills and injects "Semantic Anchors" to stabilize the feature space across extreme viewpoints and disparate domains.
- We achieve SOTA cross-view and cross-domain performance while remaining competitive within-domain, on DroneGait, Gait3D, and GREW, providing a robust solution for altitude-invariant and "in-the-wild" gait recognition.

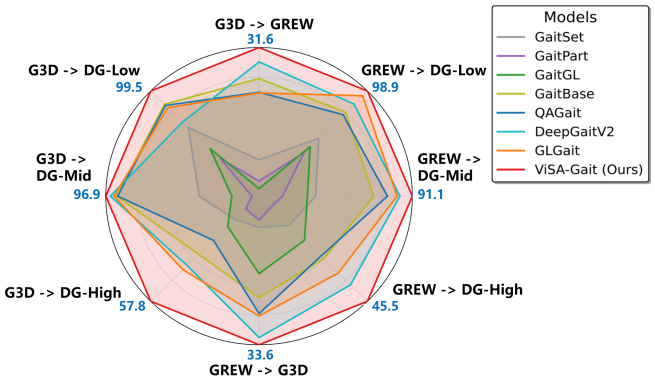

Figure 2: Performance comparison of existing methods and our ViSA-Gait in both cross-vertical-view and cross-domain condition.

2 Related Work

2.1 Gait Recognition: From Controlled to In-the-wild

Early research (GaitSet [Chao *et al.*, 2019], GaitPart [Fan *et al.*, 2020], GaitGL [Lin *et al.*, 2022]) primarily focus on controlled datasets like CASIA-B [Yu *et al.*, 2006] and OU-MVLP [Takemura *et al.*, 2018] with horizontal viewpoints. Transitioning to "in-the-wild" (GREW [Zhu *et al.*, 2021], Gait3D [Zheng *et al.*, 2022]) and UAV-based (DroneGait [Li *et al.*, 2023]) scenarios introduces complexities like background clutter and extreme vertical viewpoint shifts. In these settings, existing models often suffer from "geometric collapse" due to non-linear deformations, limiting their robustness in aerial or unconstrained outdoor environments.

2.2 Cross-View and Cross-Domain Generalization

Cross-view methods using transformation mappings [Makihara *et al.*, 2006] or disentangled learning [Wu *et al.*, 2016] often struggle with drastic altitude shifts [Li *et al.*, 2023]. Meanwhile, cross-domain strategies like UDA [Luo *et al.*, 2022] typically require target-domain access or complex adversarial training. Unlike these methods, ViSA-Gait bypasses domain-specific alignment by utilizing universal semantic anchors from VFMs. This provides a domain-invariant reference that enhances robustness across views and datasets without additional data overhead.

2.3 Vision Foundation Models and Knowledge Injection

VFMs like DINOv2 [Oquab *et al.*, 2023] offer rich semantic representations and zero-shot generalization. Strategies typically involve using them as extractors [He *et al.*, 2022] or for knowledge distillation [Hinton *et al.*, 2015]. While some gait methods use VFMs for initialization [Ye *et al.*, 2024; Ye *et al.*, 2025], using the human-structure priors of VFMs to counteract distortions caused by viewpoint and domain shifts

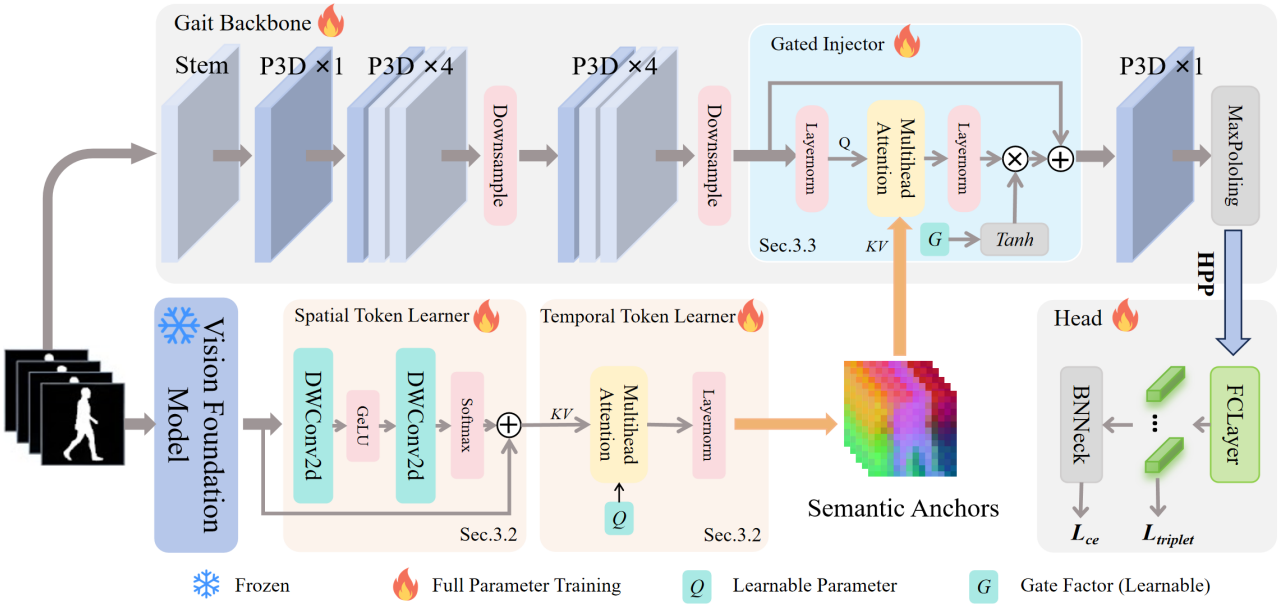

Figure 3: ViSA-Gait Pipeline. VFM serves as a Semantic Compass to provide universal human body knowledge, which is distilled through **Spatial and Temporal Token Learners** (Sec. 3.2). These distilled features are adaptively injected into the 3D-CNN backbone via **Gated Cross-Attention** (Sec. 3.3) to function as "Semantic Anchors".

remains underexplored. To bridge this gap, ViSA-Gait leverages the VFM as a "Semantic Compass" to provide universal human body knowledge for stable guidance. By distilling and injecting these priors as "Semantic Anchors", our framework effectively regularizes the feature space to mitigate geometric collapse in cross-vertical-view and cross-domain scenarios.

3 Method

3.1 Overview

The proposed **ViSA-Gait** framework is designed to overcome the inherent limitations of purely silhouette-based representations, particularly the susceptibility to *geometric collapse* under extreme viewpoint variations and domain shifts. The core philosophy of our approach is to treat the VFM as a "Semantic Compass" that provides stable universal human body knowledge guidance, while the backbone focuses on capturing fine-grained gait dynamics.

As illustrated in Figure 3, the architecture follows a dual-branch paradigm: (i) **Task-specific Tokenizer** (DeepGait [Fan *et al.*, 2025]): A 3D-CNN based backbone that extracts local spatio-temporal features directly from silhouette sequences. This branch is responsible for preserving identity-related motion nuances but is vulnerable to projection distortions. (ii) **Semantic Prior Branch** (ViT / DINOv2): A VFM branch that processes the input to provide high-level, view-invariant semantic descriptors.

The synergy between these two branches is achieved through two meticulously designed stages. First, the **Semantic Anchor Discovery** stage distills dense VFM features into a set of global semantic anchors that represent consistent anatomical structures. Second, the **Gated Semantic Injection** stage adaptively fuses these anchors into the intermediate layers of the gait backbone. This hybrid design allows the model to leverage the vast "semantic common sense" of VFMs to regularize the feature space, ensuring that the final identity representation remains robust even when the subject's geometric profile is severely distorted by camera altitudes or environmental noise.

3.2 Semantic Anchor Discovery

A critical challenge in leveraging VFMs for gait recognition is the resolution of the modality and task gap. While VFMs provide dense, high-dimensional patch features, only a subset of these features is relevant to the fine-grained intrinsic body motion required for identification. To condense the information and filter out task-irrelevant background noise, we propose a two-stage aggregation process that transforms dense features into robust "Semantic Anchors".

Given an input silhouette sequence $S \in \mathbf{R}^{B\times T\times 1\times H\times W}$, we first adapt it to the VFM's input requirements. The VFM backbone processes each frame independently to extract dense patch-level representations:

$$F_{vfm} = \text{VFM}(S) \in \mathbf{R}^{(B\cdot T)\times N_p\times D} \quad (1)$$

where N_p and D denote the number of patches and the embedding dimension, respectively.

Spatial Token Learning. Unlike static global pooling, STL treats the LVM feature map $\mathcal{F} \in \mathbf{R}^{B\times T\times D\times H_p\times W_p}$ as a reservoir of universal body structure knowledge. As illustrated in Figure 3, STL employs a lightweight bottleneck (consisting of DWConv2d and GeLU layers) to generate M soft-attention maps. This allows the model to adaptively "focus" on key body regions regardless of spatial displacement or distortion:

$$A = \text{Softmax}(\text{DWConv}(\text{GeLU}(\text{DWConv}(\mathcal{F})))) \quad (2)$$

$$T_{spatial} = A \cdot \mathcal{F}_{flat} \in \mathbf{R}^{B\times T\times M\times D} \tag{3}$$

where $Tspatial$ distills M localized semantic descriptors per frame. By projecting dense features into these spatial tokens, STL ensures that even under extreme vertical-view compression, the universal body semantics are preserved and aligned as stable references.

Temporal Token Learning. Since gait is inherently a sequential process, single-frame semantic tokens are insufficient for capturing motion dynamics. We therefore employ a Temporal Token Learner to aggregate information across the entire sequence. We define a set of K learnable queries $Q_{learn} \in \mathbf{R}^{1\times K\times D}$, which act as "probes" to distill motion-consistent information:

$$T_{global} = \text{LayerNorm}(\text{MHA}(Q, K, V)) \tag{4}$$

where $Q = Q_{learn}, K = T_{spatial}, V = T_{spatial}$, MHA denotes the Multi Head Attention mechanism. This stage enables the model to bridge the gap between frame-level spatial cues and sequence-level temporal representations. By learning to attend to the most representative temporal segments, the resulting T_{global} provides a set of "Semantic Anchors" that remain stable across disparate camera viewpoints and domains.

3.3 Gated Semantic Injection

The primary objective of the injection stage is to bridge the modality and representation gap between the VFM branch and the task-specific gait backbone. While the Semantic Anchors T_{global} carry rich view-invariant anatomical priors, they reside in a different feature space than the spatio-temporal maps of the DeepGait backbone. Direct summation of these features may lead to representation contamination, especially in simple scenarios where the silhouette's geometry is already sufficient for identification. To address this, we propose a **Gated Cross-Attention (GCA)** mechanism that adaptively regularizes the gait features.

Modality Alignment. We first align the Semantic Anchors to the backbone's hidden dimension C_l at layer l. A linear projection layer, followed by Layer Normalization, is employed to transform T_{global} into the aligned anchors $\hat{T}$:

$$\hat{T} = \text{LayerNorm}(\text{Linear}(T_{global})) \in \mathbf{R}^{B\times K\times C_l} \tag{5}$$

where $\hat{T}$ preserves the K key semantic descriptors while matching the channel capacity of the gait backbone.

Gated Cross-Attention Mechanism. We treat the localized spatial features from the gait backbone as the query, and the aligned semantic anchors as the key-value pairs. This allows each spatial location in the gait feature map to selectively "query" the most relevant anatomical information from the anchors. Let $X_l \in \mathbf{R}^{B\times C_l\times T\times H_l\times W_l}$ be the intermediate feature map. We flatten the spat-temporal dimensions and compute the cross-attention:

$$Y = \text{MHA}(Q = X_l^{flat}, K = \hat{T}, V = \hat{T}) \tag{6}$$

The cross-attention output Y represents the "semantic correction" signal that compensates for the backbone's geometric deficiencies.

Adaptive Gating Control. To ensure training stability and prevent the VFM priors from dominating the backbone's established discriminative power, we introduce a gated fusion strategy. The updated feature map X_{l+1} is formulated as:

$$X_{l+1} = X_l + \text{Dropout}(\text{LN}(Y)) \cdot (\tanh(g) \cdot s) \tag{7}$$

where g and s are learnable gate and scale parameters, both initialized to zero. This initialization enforces a "Gait-first" policy at the start of training: the model initially functions as a standard gait network. As training progresses, the network autonomously optimizes g to adaptively modulate the semantic injection intensity. Although the gating factor $\tanh(g)$ is learned implicitly, its functional role as a "semantic compass" is manifested through the model's ability to maintain identity consistency when low-level geometric cues vanish.

3.4 Learning Objective

The final embeddings are extracted via Horizontal Pooling Pyramid (HPP) [Fu *et al.*, 2019] and Separate FCs. The model is trained end-to-end using a combination of Triplet Loss ($\mathcal{L}_{triplet}$), Label Smoothing Cross-Entropy Loss ($\mathcal{L}_{ce}$):

$$\mathcal{L} = \mathcal{L}_{triplet} + \mathcal{L}_{ce} \tag{8}$$

4 Experiment

We evaluate on Gait3D [Zheng *et al.*, 2022], GREW [Zhu *et al.*, 2021], and Dronegait [Li *et al.*, 2023] for within-domain and unconstrained performance, validating "Semantic Compass" cross-domain stability and "Semantic Anchors" resilience to extreme vertical views.

4.1 Datasets and Implementation Details

The dataset information and implementation details in our experiments are as follows.

Gait3D [Zheng *et al.*, 2022] is a large-scale comprehensive dataset for gait recognition. It comprises 4,000 subjects, 25,309 sequences, and 39 distinct camera viewpoints, divided into two parts with 3,000 and 1,000 subjects as training set and test set. Collected in supermarket environments, the dataset contains abundant human silhouettes from elevated perspectives, presenting enhanced challenges for gait recognition tasks.

GREW [Zhu *et al.*, 2021] is one of the largest gait datasets in the wild, including Silhouettes, GEIs, and 2D/3D human poses data types. The raw videos are collected from 882 cameras in large public areas. 7,533 video clips are used, containing nearly 3,500 hours of 1,920 × 1,080 streams. It has 26,345 subjects and 128,671 sequences, divided into two parts with 20,000 and 6,000 subjects as training set and test set.

DroneGait [Li *et al.*, 2023] is a drone-based gait dataset focusing on high vertical view recognition, with 96 subjects and 22,000 sequences with vertical angles from 0° to 80°. It categorizes data into three view groups: low (0°), medium (30° – 60°), and high (60° – 80°), with frame-level alignment for cross-view analysis. Following the standard protocol, we split the data into 70 training and 26 testing subjects, evaluating recognition performance through probe-gallery matching where gallery sequences cover multi-view conditions.

Train	Method	Venue	VFM	Test Dataset								
				Gait3D			GREW			DroneGait		
				R-1	R-5	mAP	R-1	R-5	R-10	Low	Mid	High
Gait3D	GaitSet	AAAI19	–	36.7	58.3	30	15.3	27.7	33.5	96.2	86.6	32.3
	GaitPart	CVPR20	–	28.2	47.6	21.6	12.2	22.2	28.6	93.4	80.8	29.0
	GaitGL	ICCV21	–	63.8	80.5	55.9	11.1	20.4	25.7	94.3	83.0	34.5
	GaitBase	CVPR23	–	64.6	73.9	45.5	27.1	41.7	48.2	98.3	95.7	43.2
	QAGait	AAAI24	–	67.0	81.5	56.5	25.1	39.4	45.7	98.2	95.6	38.8
	GLGait	MM24	–	77.7	88.9	70.6	27.4	40.1	47.7	98.4	96.9	46.9
	DeepGaitV2	TPAMI25	–	74.4	88	65.8	29.5	44.1	50.5	96.7	96.4	46.7
	ViSA-Gait	Ours	VIT-S	74.8	88.1	66.8	30.9	44.6	50.7	99.3	96.9	57.4
	ViSA-Gait	Ours	DINOV2-S	75	88.5	**68.4**	31.1	45	51.2	**99.6**	96.6	**57.8**
	ViSA-Gait	Ours	VIT-B	**76.6**	**89.0**	68.0	**31.6**	**45.2**	**51.8**	99.5	**96.9**	**57.8**
GREW	GaitSet	AAAI19	–	18.9	32.6	13.9	46.3	63.6	70.3	94.9	84.1	30.7
	GaitPart	CVPR20	–	18	30.1	12.9	44	60.7	67.3	93.7	81.7	27.2
	GaitGL	ICCV21	–	24.7	36.6	17.5	68	80.7	85	94.2	83.5	33.6
	GaitBase	CVPR23	–	27.7	42.9	19.3	64.6	–	–	97.1	88.3	37.2
	QAGait	AAAI24	–	29.7	46.0	21.1	59.1	74.0	79.2	96.9	89.3	36.7
	GLGait	MM24	–	30.0	46.8	22.3	82.8	91.1	93.5	96.4	88.7	41.4
	DeepGaitV2	TPAMI25	–	32.7	49.5	23.9	77.7	87.9	90.6	97.8	90.2	42.3
	ViSA-Gait	Ours	VIT-S	33.1	49.3	24.0	77.7	88.9	92.0	97.4	90.3	44.1
	ViSA-Gait	Ours	DINOV2-S	33.3	49.2	24.1	78.0	**89.8**	**92.9**	**98.9**	91.0	45.2
	ViSA-Gait	Ours	VIT-B	**33.6**	**49.6**	**24.3**	**78.1**	89.5	92.7	**98.9**	**91.1**	**45.5**

Table 1: Performance comparisons of Rank-1, 5 and 10 accuracies, mean Average Precision (%) of the proposed ViSA-Gait against existing methods on Gait3D, GREW, and DroneGait datasets. The evaluations encompass within-domain, cross-domain, and cross-vertical-view scenarios, where DroneGait altitudes (Low, Mid, High) represent discrete samplings of vertical perspective variations. The "VFM" column specifies the foundation model used as the semantic prior.

Implementation Details. Our network implementation and reproduction experiments of SOTA models are both deployed through the OpenGait codebase [Fan *et al.*, 2025], and are conducted on a system with four NVIDIA RTX A6000 GPUs. Each input gait sequence consists of 30 frames, and all silhouettes are resized to 64×44. The batch size $[I, J]$ is consistent with [Fan *et al.*, 2025], where I represents the number of subjects sampled per mini-batch, and J represents the number of sequences sampled per subject.

The optimization setup for ViSA-Gait employs AdamW with a base learning rate of 1×10^{-4} and weight decay 0.05, coupled with a OneCycleLR scheduler whose maximum learning rate is 6×10^{-4}. The total training steps are $80{,}000$ for DroneGait and Gait3D, and $180{,}000$ for GREW, with a warm-up phase that covers 6% of the cycle.

4.2 Comparison with Other SOTA Methods

To verify the effectiveness of our method, we introduce several latest gait recognition methods for comparison, including GaitSet [Chao *et al.*, 2019], GaitPart [Fan *et al.*, 2020], GaitGL [Lin *et al.*, 2022], GaitBase [Fan *et al.*, 2025], QAGait [Wang *et al.*, 2024], GLGait [Peng *et al.*, 2024] and DeepGaitV2 [Fan *et al.*, 2023].

Within-domain Evaluation. Within-domain evaluation (Gait3D $\rightarrow$ Gait3D and GREW $\rightarrow$ GREW) reflects the basic discriminative capability when training and testing distributions are aligned. As summarized in Table 1:

- On the Gait3D dataset, ViSA-Gait-VIT-B achieves a Rank-1 accuracy of 76.6%, outperforming the advanced baseline DeepGaitV2 (74.4%) and showing competitive performance against the latest GLGait.
- On the GREW dataset, our VIT-B variant reaches 78.1% Rank-1 accuracy, yielding a consistent improvement over geometric-only methods.

These results indicate that the injection of high-level semantic priors does not interfere with the model's ability to capture domain-specific motion patterns, but instead provides a more robust feature representation even in standard "in-the-wild" horizontal views.

Cross-domain Evaluation. Cross-domain evaluation (Gait3D $\leftrightarrow$ GREW) assesses the model's generalization ability against environmental shifts (e.g., changes in lighting, background, and sensor types).

- In the Gait3D $\rightarrow$ GREW task, ViSA-Gait-VIT-B improves the Rank-1 accuracy to 31.6%, a 2.1% increase over DeepGaitV2.
- For the GREW $\rightarrow$ Gait3D transfer, our model also maintains a leading position with 33.6% Rank-1 accuracy.

The performance gains across different data distributions suggest that semantic anchors from the VFM provide universal structural regularizations. These priors are inherently more stable than raw geometric silhouettes, which are often sensitive to the domain-specific noise present in outdoor datasets.

Cross-vertical-view Evaluation. The evaluation on the DroneGait dataset is of high practical significance as it closely simulates real-world surveillance deployment. In actual deployment scenarios, the perspective angle between a fixed-high camera and a moving subject varies continuously. The three altitudes—Low (parallel view), Mid (oblique view), and High (near-top-down view)—serve as discrete samplings of this angular variation.

Scenario Adaptation: At the Low altitude representing parallel views, most methods achieve near-saturation performance, with ViSA-Gait reaching 99.5%. However, as the

Strategy	Gait3D		GREW	
	Rank-1	mAP	Rank-1	Rank-5
Baseline	74.4	65.8	77.7	87.9
Full Fine-tuning	71.8	63.2	71.3	80.1
LoRA Fine-tuning	72.3	64.5	73.5	83.2
Unfreeze Last-3	72.0	63.7	71.7	81.5
ViSA-Gait	**76.6**	**68.0**	**78.1**	**89.5**

Table 2: Within-domain performance comparison of different fine-tuning strategies on Gait3D and GREW datasets. All methods share the same ViT-B backbone; ViSA-Gait keeps it frozen while Full, LoRA (r=8, α=16) and Last-3 respectively unfreeze all, adapt attention, or unlock the final three blocks.

Strategy	Params	FLOPs	Gait3D		GREW	
	(M)	(G)	Rank-1	mAP	Rank-1	Rank-5
Baseline	11.1	2.9	74.4	65.8	77.7	87.9
ViSA-Gait-VIT-S	26.6	7.1	74.8	88.1	77.7	88.9
ViSA-Gait-DINOV2-S	26.6	32.3	75.0	88.4	78.0	89.8
ViSA-Gait-VIT-B	28.9	19.8	**76.6**	**89.0**	**78.1**	**89.5**

Table 3: Within-domain Parameter and GFLOPs comparison.

subject moves into Mid and High perspective zones, the silhouettes undergo severe vertical compression.

Resilience to Geometric Collapse: The most critical challenge occurs at the High altitude, where the viewpoint approaches a bird's-eye perspective. This leads to the "geometric collapse" of silhouettes, rendering traditional geometric features uninformative. Under this extreme yet realistic shift, ViSA-Gait-VIT-B (trained on Gait3D) achieves 57.8% Rank-1 accuracy, outperforming DeepGaitV2 (46.7%) by a substantial 11.1%. Even when trained on the larger GREW dataset, ViSA-Gait maintains a significant lead at 45.5%. These results confirm our core hypothesis: While geometric information dissipates under steep viewing angles, the VFM branch successfully preserves view-invariant anatomical semantics.

The stability across the entire angular range proves that ViSA-Gait offers a robust solution for real-world surveillance where camera-to-subject perspectives are in constant flux.

4.3 Ablations

Effectiveness of Fine-tuning Strategies

We investigate the optimal utilization of VFM priors by comparing our strategy against three adaptive methods: Full Fine-tuning, LoRA, and Unfreeze Last-3. As summarized in Table 2, all strategies involving VFM adaptation result in a performance drop relative to the Baseline and our Frozen approach. Notably, *Full Fine-tuning* and *Unfreeze Last-3* cause significant degradation, with GREW Rank-1 accuracy dropping by over 6%.

We conclude that fine-tuning on abstract gait silhouettes leads to the **catastrophic forgetting** of universal anatomical priors. By keeping the VFM frozen, ViSA-Gait preserves the model's "semantic common sense", allowing it to function as a stable "semantic compass" that regularizes the backbone. This strategy prevents over-specialization to domain-specific noise and ensures that the injected features

Strategy	DroneGait			GREW	
	Low	Mid	High	Rank-1	Rank-5
GMP	99.5	96.7	55.8	29.8	43.9
GAP	99.4	96.7	56.7	30.6	44.3
STL + TTL (Ours)	**99.5**	**96.9**	**57.8**	**31.6**	**45.2**

Table 4: Comparison of different feature extraction strategies for VFM branch. GMP: Global Mean Pooling, GAP: Global Max Pooling, STL: Spatial Token Learner, TTL: Temporal Tokene Learner.

Mechanism	Interaction Type	DroneGait			GREW	
		Low	Mid	High	Rank-1	Rank-5
Addition	Element-wise Add	98.6	96.8	57.1	24.7	39.1
Concat + MLP	Channel Fusion	98.9	96.5	57.5	25.1	39.9
Film	Linear Modulation	99.3	96.7	57.1	29.7	43.6
GCA(Ours)	**Attention-based**	**99.5**	**96.9**	**57.8**	**31.6**	**45.2**

Table 5: Impact of different injection mechanisms.

remain task-agnostic and view-invariant, ultimately providing the best balance between semantic preservation and motion modeling.

Analysis of Model Complexity and Efficiency

Table 3 evaluates the complexity-performance trade-off across ViSA-Gait variants. Notably, ViSA-Gait-DINOv2-S, which targets the same parameter scale as ViT-S, incurs a disproportionate computational cost (32.3 GFLOPs) for marginal gains. We attribute this to DINOv2's smaller 14×14 patch size, which quadratically increases token-level attention and introduces significant redundancy for fine-grained gait modeling. Conversely, ViSA-Gait-ViT-B achieves the superior Rank-1 (76.6%) and mAP (89.0%) results with a much lower footprint (19.8 GFLOPs) than DINOv2-S. This proves that higher-capacity universal human body knowledge is a more effective "Semantic Compass" than high-density but computationally expensive features. Given the diminishing returns of DINOv2-S and our commitment to a lightweight design, ViSA-Gait-ViT-B represents the most optimal scientific balance for robust recognition.

Effectiveness of Feature Extraction Strategy

We compare STL + TTL against standard global pooling (GMP, GAP) in Table 4. While performance is similar at lower altitudes, STL + TTL significantly excels in challenging scenarios, achieving 57.8% accuracy (+1.1%) in DroneGait and a 1.0% Rank-1 gain on GREW. Unlike static pooling that treats patches equally, STL adaptively identifies anatomical anchors while TTL aggregates temporal dependencies, capturing more robust "semantic common sense" for gait representation.

Effectiveness of Injection Mechanism

We evaluate the proposed Gated Cross-Attention (GCA) against standard fusion strategies in Table 5. While element-wise addition and channel fusion provide basic integration, they lack the selectivity required to handle complex perspective distortions. FiLM [Perez *et al.*, 2018] offers more flexible modulation but remains suboptimal on the large-scale GREW dataset. In contrast, GCA consistently achieves superior performance, particularly improving the Rank-1 ac-

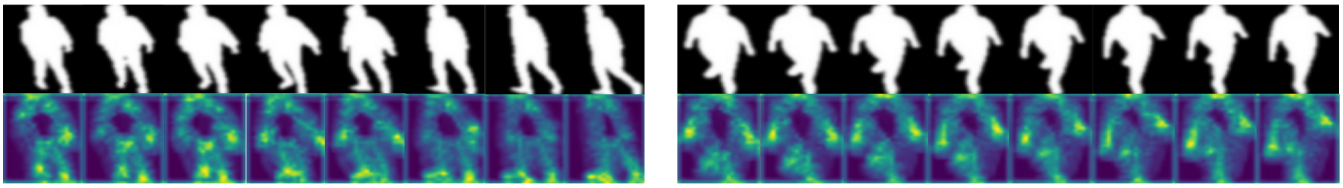

Figure 4: Comparison of feature activation maps for the same individual across eye-level (left) and high-angle (right) perspectives. ViSA-Gait demonstrates consistent anatomical tracking across diverse viewing conditions.

curacy on GREW by 1.9% over the FiLM baseline. These results demonstrate that the attention-based mechanism can more adaptively inject VFM-derived "semantic anchors" into the backbone, ensuring precise feature compensation and enhanced generalization across diverse environmental conditions.

4.4 Visualization

Cross-View Robustness and Generalization

To intuitively evaluate the efficacy of ViSA-Gait, we first visualize the feature activation maps across disparate viewing conditions in Figure 4. While the left side captures an eye-level laboratory perspective with standard geometric integrity, the right side presents a challenging outdoor scenario where the subject undergoes severe geometric collapse. Remarkably, despite the drastic silhouette compression, the heatmaps reveal that our model consistently maintains a stable focus on core key body regions, such as the head and moving limbs, mirroring the response patterns observed in constrained environments. This cross-view semantic alignment confirms that the framework has successfully transitioned from a dependency on fragile silhouette geometry to a robust reliance on intrinsic body motion. Such generalization ensures that the captured gait patterns remain discriminative even under extreme viewpoint-induced variations.

Semantic Consistency and Noise Suppression

We visualize the raw feature maps extracted from the VFM using PCA color mapping. The PCA operation projects high-dimensional semantic vectors into a 3-D RGB space, where pixels possessing semantic similarity are represented by similar color clusters. As illustrated in the Figure 5, the VFM exhibits a strong capability for holistic semantic perception, mapping the entire pedestrian body to a consistent and contiguous color manifold while effectively suppressing background noise. This stark contrast between the foreground subject and the environmental noise demonstrates that the VFM serves as a robust "semantic anchor".

Global Semantic Anchoring and Feature Generalization

Building on these spatial priors, our STL + TTL mechanism distills these spatial priors into a compact set of global tokens. As shown in Figure 6, the resulting "semantic anchor" exhibit highly structured activation patterns across feature dimensions. The consistent vertical striping confirms that the model captures stable, identity-related attributes while filtering residual background noise. By distilling universal VFM priors into these compact signatures, ViSA-Gait fosters a highly generalized feature space, enabling the capture of stable identity patterns that transcend specific environmental and viewpoint-induced variations.

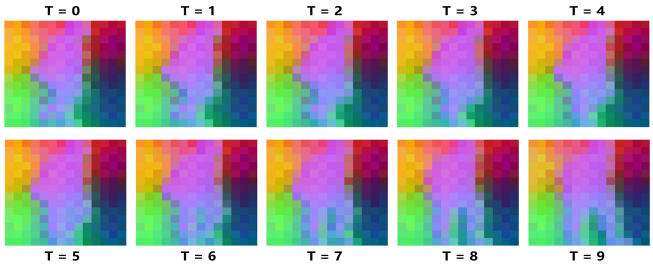

Figure 5: Visualization of VFM semantic anchors via PCA color mapping.

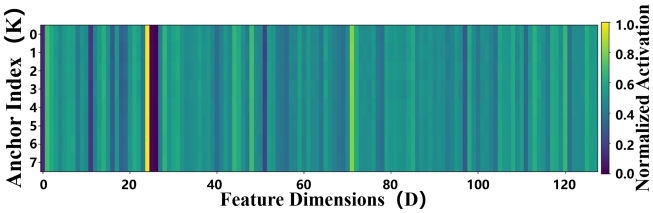

Figure 6: Visualization of global token signatures extracted by the STL + TTL mechanism, illustrating the consistent and structured activation patterns across feature dimensions.

5 Conclusion

In this paper, we present ViSA-Gait, a framework that transcends the inherent limitations of silhouette-based gait recognition by incorporating high-level semantic priors from VFMs. By utilizing the VFM as a "Semantic Anchor", our STL + TTL and GCA mechanisms effectively distill and inject view-invariant universal human body knowledge, guiding the model to focus on intrinsic intrinsic body motion rather than overfitting to domain-specific geometric artifacts. Extensive experiments confirm that preserving the VFM's "semantic common sense" through a frozen strategy is essential for cross-domain robustness. Ultimately, ViSA-Gait achieves SOTA performance on challenging benchmarks, successfully bridging the gap between conventional geometry-based gait analysis and universal semantic understanding in diverse real-world scenarios.

Acknowledgments

This work was supported by the Guangdong Basic and Applied Basic Research Foundation (2026A1515010184).

Contribution Statement

Xiangru Li and Deqiang Yin contributed equally to this work. Xiangru Li: conceptualization, methodology, software, experiments, data analysis, writing – original draft, writing – review & editing. Deqiang Yin: experiments, data analysis, writing – review & editing. Yifan Xie: writing – review & editing. Guojian Li: writing – review & editing. Zebang Cheng: writing – review & editing. Fei Ma: supervision, writing – review & editing, funding, corresponding author

References

[Chao *et al.*, 2019] 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*, volume 33, pages 8126–8133, 2019.

[Cheng *et al.*, 2024] Zebang Cheng, Zhi-Qi Cheng, Jun-Yan He, Kai Wang, Yuxiang Lin, Zheng Lian, Xiaojiang Peng, and Alexander Hauptmann. Emotion-llama: Multimodal emotion recognition and reasoning with instruction tuning. *Advances in Neural Information Processing Systems*, 37:110805–110853, 2024.

[Dosovitskiy *et al.*, 2021] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *ICLR*, 2021.

[Fan *et al.*, 2020] 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.

[Fan *et al.*, 2023] Chao Fan, Saihui Hou, Yongzhen Huang, and Shiqi Yu. Exploring deep models for practical gait recognition. *arXiv preprint arXiv:2303.03301*, 2023.

[Fan *et al.*, 2025] 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*, 2025.

[Fu *et al.*, 2019] Yang Fu, Yunchao Wei, Yuqian Zhou, Honghui Shi, Gao Huang, Xinchao Wang, Zhiqiang Yao, and Thomas Huang. Horizontal pyramid matching for person re-identification. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 8295–8302, 2019.

[He *et al.*, 2022] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16000–16009, 2022.

[Hinton *et al.*, 2015] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

[Li *et al.*, 2023] Aoqi Li, Saihui Hou, Qingyuan Cai, Yang Fu, and Yongzhen Huang. Gait recognition with drones: A benchmark. *IEEE Transactions on Multimedia*, 26:3530–3540, 2023.

[Li *et al.*, 2025] Xiangru Li, Wei Song, Yingda Huang, Wei Meng, Le Chang, and Hongyang Li. CVVNet: A cross-vertical-view network for gait recognition. *arXiv preprint arXiv:2505.01837*, 2025.

[Lin *et al.*, 2022] Beibei Lin, Shunli Zhang, Ming Wang, Lincheng Li, and Xin Yu. Gaitgl: Learning discriminative global-local feature representations for gait recognition. *arXiv preprint arXiv:2208.01380*, 2022.

[Luo *et al.*, 2022] Huoling Luo, Congcong Wang, Xingguang Duan, Hao Liu, Ping Wang, Qingmao Hu, and Fucang Jia. Unsupervised learning of depth estimation from imperfect rectified stereo laparoscopic images. *Computers in biology and medicine*, 140:105109, 2022.

[Ma *et al.*, 2023] 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.

[Ma *et al.*, 2024] 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.

[Ma *et al.*, 2025] Fei Ma, Yifan Xie, Yukan Li, Ying He, Yi Zhang, Hongwei Ren, Zhou Liu, Wei Yao, Fuji Ren, and Fei Richard Yu. A review of human emotion synthesis based on generative technology. *IEEE Transactions on Affective Computing*, 2025.

[Makihara *et al.*, 2006] Yasushi Makihara, Ryusuke Sagawa, Yasuhiro Mukaigawa, Tomio Echigo, and Yasushi Yagi. Gait recognition using a view transformation model in the frequency domain. In *European conference on computer vision*, pages 151–163. Springer, 2006.

[Marasco and Ross, 2014] Emanuela Marasco and Arun Ross. A survey on antispoofing schemes for fingerprint recognition systems. *ACM Computing Surveys*, 47(2):1–36, 2014.

[Oquab *et al.*, 2023] Maxime Oquab, Timothée Darcet, Theo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Russell Howes, Po-Yao Huang, Hu Xu, Vasu Sharma, Shang-Wen Li, Wojciech Galuba, Mike Rabbat, Mido Assran, Nicolas Ballas, Gabriel Synnaeve, Ishan Misra, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision, 2023.

[Peng *et al.*, 2024] 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.

[Perez *et al.*, 2018] Ethan Perez, Florian Strub, Harm De Vries, Vincent Dumoulin, and Aaron Courville. Film: Visual reasoning with a general conditioning layer. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.

[Sepas-Moghaddam and Etemad, 2022] Alireza Sepas-Moghaddam and Ali Etemad. Deep gait recognition: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(1):264–284, 2022.

[Shen *et al.*, 2024] Chuanfu Shen, Shiqi Yu, Jilong Wang, George Q. Huang, and Liang Wang. A comprehensive survey on deep gait recognition: Algorithms, datasets, and challenges. *IEEE Transactions on Biometrics, Behavior, and Identity Science*, 2024.

[Singloo *et al.*, 2024] Hemal Singloo, Manoj Kumar, Simran Namdev, Kinsukh Suni Sonbawne, and Nishant Indoria. Gait recognition: A cross-dataset evaluation of deep learning models. In *Proceedings of the 2024 IEEE International Students' Conference on Electrical, Electronics and Computer Science (SCEECS)*, pages 1–10. IEEE, 2024.

[Takemura *et al.*, 2018] 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 transactions on Computer Vision and Applications*, 10(1):4, 2018.

[Wang *et al.*, 2024] 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*, volume 38, pages 5785–5793, 2024.

[Wu *et al.*, 2016] Jiajun Wu, Chengkai Zhang, Tianfan Xue, Bill Freeman, and Josh Tenenbaum. Learning a probabilistic latent space of object shapes via 3d generative-adversarial modeling. *Advances in neural information processing systems*, 29, 2016.

[Xie *et al.*, 2025] Yifan Xie, Mingyang Li, Shoujie Li, Xingting Li, Guangyu Chen, Fei Ma, Fei Richard Yu, and Wenbo Ding. Universal visuo-tactile video understanding for embodied interaction. *arXiv preprint arXiv:2505.22566*, 2025.

[Xue *et al.*, 2025] Haiwei Xue, Xiangyang Luo, Zhanghao Hu, Xin Zhang, Xunzhi Xiang, Yuqin Dai, Jianzhuang Liu, Zhensong Zhang, Minglei Li, and Jian Yang. Human motion video generation: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.

[Ye *et al.*, 2024] Dingqiang Ye, Chao Fan, Jingzhe Ma, Xiaoming Liu, and Shiqi Yu. Biggait: Learning gait representation you want by large vision models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 200–210, 2024.

[Ye *et al.*, 2025] Dingqiang Ye, Chao Fan, Zhanbo Huang, Chengwen Luo, Jianqiang Li, Shiqi Yu, and Xiaoming Liu. Biggergait: Unlocking gait recognition with layer-wise representations from large vision models. *arXiv preprint arXiv:2505.18132*, 2025.

[Yu *et al.*, 2006] 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)*, volume 4, pages 441–444. IEEE, 2006.

[Zheng *et al.*, 2022] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 20228–20237, 2022.

[Zhu *et al.*, 2016] Xiangyu Zhu, Zhen Lei, Xiaoming Liu, Hailin Shi, and Stan Z. Li. Face alignment across large poses: A 3d solution. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 146–155, 2016.

[Zhu *et al.*, 2021] 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 (ICCV)*, pages 14789–14799, 2021.