Leveraging Text-to-Video Diffusion Models for Synthetic Gait Dataset Generation and Appearance-Robust Person Recognition

Leveraging Text-to-Video Diffusion Models for Synthetic Gait Dataset Generation and Appearance-Robust Person Recognition

Federico Boscolo
Dept. of Control and Computer Engineering
Politecnico di Torino
Turin, Italy
federico.boscolo@polito.it

Fabrizio Lamberti
Dept. of Control and Computer Engineering
Politecnico di Torino
Turin, Italy
fabrizio.lamberti@polito.it

***Abstract*—Gait recognition accuracy drops 10-25% when subjects wear coats or carry bags, yet collecting diverse training data is costly. This paper proposes a text-to-video diffusion pipeline to synthesize realistic gait data. Unlike prior silhouette generation, we produce full video sequences with identity-consistent faces, enabling direct application to multimodal face-gait access control systems. Using a custom three-stage pipeline, we generate 200 synthetic identities across Normal Walking (NM), Coat-wearing (CL), and Bag-carrying (BG) conditions in indoor and outdoor environments. Sim-to-real transfer learning on CASIA-B demonstrates that synthetic pre-training followed by real NM fine-tuning boosts CL accuracy from 31% to 95%. This matches fully-supervised baselines without requiring real appearance-varied data, enabling balanced 50-50 fusion of face and gait modalities that reaches 99.6% accuracy.**

***Index Terms*—face recognition, gait recognition, text-to-image, image-to-video, synthetic data, deep learning**

I. Introduction

Access control systems increasingly rely on multimodal face-gait recognition to identify subjects approaching from a distance [1], [2]. While face recognition is robust at close range, gait enables distance identification but suffers severe accuracy drops (10–25%) under clothing variations like coats or bags [3], [4], [5], [6].

Addressing this vulnerability is hindered by the scarcity of diverse training data. Existing synthetic approaches generate silhouettes [7], discarding facial information essential for multimodal identity matching. To address the lack of attention to facial and other contextual data, we present a generative AI pipeline that uses text-to-video diffusion models to synthesize full RGB gait videos with controlled appearance variations and identity-consistent faces, for use in multimodal face-gait fusion approaches. Our three-stage approach utilizes text-to-image synthesis, vision-language appearance editing, and pose-guided video generation to enable sim-to-real transfer learning.

This work is part of the project PNRR-NGEU which has received funding from the MUR – DM 117/2023.

Evaluating on frontal-view CASIA-B sequences [3], synthetic pre-training boosts coat-condition accuracy from 31% to 95%, surpassing fully-supervised baselines without real appearance-varied data.

The main contributions of this work are:

- A three-stage generative pipeline yielding appearance-varied, identity-consistent RGB gait videos for both unimodal and multimodal recognition.
- A sim-to-real transfer strategy that learns appearance-invariant gait features.
- Experimental validation matching fully-supervised performance on CASIA-B without real appearance-varied data.
- Demonstration of balanced, highly accurate multimodal face-gait fusion using synthetic-enhanced gait features.

The remainder of this paper is organized as follows: Section II reviews related work, Section III details our pipeline, Section IV presents results, and Section V provides conclusions.

II. Related Work

We review relevant literature in three areas that inform our approach: methods for handling appearance variations in gait recognition, generative models for synthesizing biometric data, and techniques for bridging the synthetic-to-real domain gap.

A. Gait Recognition Under Appearance Variations

Gait recognition methods are mainly categorized into model-free and model-based approaches. Many model-free techniques, particularly Gait Energy Image (GEI) [8], rely on silhouette-based representations, making them easier to implement but highly sensitive to variations in clothing and carrying conditions. Model-based methods construct explicit models of human body structure and estimate joint movements, offering greater robustness at the cost of computational complexity.

With the advent of deep learning, silhouette-based approaches using CNN backbones became dominant due to their

strong performance on standard benchmarks. GaitSet [9] introduced set-based processing of silhouette sequences, achieving 96.1% rank-1 accuracy on CASIA-B under normal walking (NM) conditions but experiencing significant degradation under appearance variations (coat-wearing, CL, and bag-carrying, BG). GaitPart [10] extended this approach with temporal part-based modeling, extracting spatiotemporal features from different body regions.

Skeleton-based approaches instead directly model body joint positions. Gait-TR [11] combines spatial transformers with temporal CNNs, GaitGraph [12] applies graph convolutional networks to 2D pose estimates, while PoseGait [13] uses 3D poses with specially designed spatial-temporal features. Despite these advances, collecting sufficient training data across diverse appearance conditions remains highly expensive, motivating synthetic data approaches. Notably, this appearance sensitivity affects not only standalone gait recognition but also multimodal systems where gait serves as a complementary modality to face recognition [1]; improving gait robustness is therefore critical for enhancing overall person identification performance in such systems.

B. Generative Models for Synthetic Biometric Data

Diffusion models have revolutionized generation of synthetic data. Recent work addresses motion synthesis with body shape conditioning [14] and temporal consistency [15], [16].

For gait-specific synthesis, Sawicki [17] compared GAN, VAE, and LSTM-MDN models for synthesizing gait data from 100 VR-captured subjects. LSTM-MDN synthetic samples improved F1-score from 0.928 to 0.966 for biometric identification, while GAN-generated samples performed poorly with attention-based classifiers.

GAN-based approaches have shown promise for gait domain translation. Alsaggaf et al. [7] use CycleGANs to translate gait energy images affected by bags or coats to normal GEIs in an unsupervised manner. Gao et al. [18] demonstrate GAN-based perspective transformation achieving 98.32% on CASIA-B through view normalization. These works focus on transforming existing real gait data rather than generating synthetic data from scratch. Moreover, all existing approaches operate on silhouette or skeleton representations, discarding facial appearance information. This limitation prevents their application to multimodal systems requiring identity-consistent faces across appearance variations. Our work addresses this gap by generating complete video sequences that preserve both gait patterns and facial identity.

C. Synthetic Data for Computer Vision

The effectiveness of synthetic data for computer vision tasks is well-established across multiple domains. The NOVA framework [19] demonstrates Unity-based photorealistic rendering with procedurally generated humans, featuring 36 major body types and diverse clothing. Domain adaptation techniques bridge the synthetic-to-real gap. Liao et al. [20] demonstrate end-to-end Sim2Real transfer using CycleGAN for image translation combined with hybrid transfer learning. Domain randomization [21] successfully transferred deep networks trained only on simulated RGB images to real-world robotic control. Recent meta-analyses show that, for urban scene understanding, models trained on 50% real + 50% synthetic data performed best [22], while other works achieved 96% recall using purely synthetic training [23]. These findings demonstrate that synthetic pre-training generally allows learning appearance-invariant features without collecting costly real-world data.

TABLE I: Performance degradation of state-of-the-art gait recognition methods on CASIA-B under cross-condition scenarios.

Method	Type	NM	BG	CL	Drop†
GaitSet [9]	Silhouette	95.0	87.2	70.4	24.6
GaitPart [10]	Silhouette	96.2	91.5	78.7	17.5
GaitBase [5]	Silhouette	97.6	94.0	77.4	20.2
CSTL [6]	Silhouette	98.7	96.2	88.7	10.0
PoseGait [13]	Skeleton	68.7	66.3	49.5	19.2
GaitGraph [12]	Skeleton	87.7	74.8	66.3	21.4
Gait-TR [11]	Skeleton	96.0	91.3	90.0	6.0

†Drop = NM − CL accuracy difference (percentage points).

III. Methodology

This section presents our approach to generating synthetic gait training data with controlled appearance variations. We first describe our three-stage generation pipeline, then detail the sim-to-real transfer learning strategy and integration with a multimodal framework.

A. Overview and Motivation

Table I summarizes the performance degradation of state-of-the-art gait recognition methods under appearance variations on CASIA-B. Even the best-performing methods exhibit accuracy drops of 10–25 percentage points between normal walking (NM) and coat-wearing (CL) conditions, confirming that appearance sensitivity remains a critical challenge.

We address this challenge through text-to-video diffusion models. Two key design choices distinguish our approach from prior work:

Full video generation instead of silhouettes. Prior silhouette-based synthesis suffices for unimodal recognition but discards facial appearance. Our pipeline generates complete RGB video sequences with identity-consistent faces, enabling direct application to multimodal face-gait systems.

Frontal viewing angle (0°)**.** We focus exclusively on frontal-view sequences because it represents the dominant acquisition scenario in access control, and restricting to a single viewpoint reduces domain complexity during sim-to-real transfer.

As illustrated in Figure 1, our hierarchical approach decouples identity, appearance, and motion into separately controllable factors.

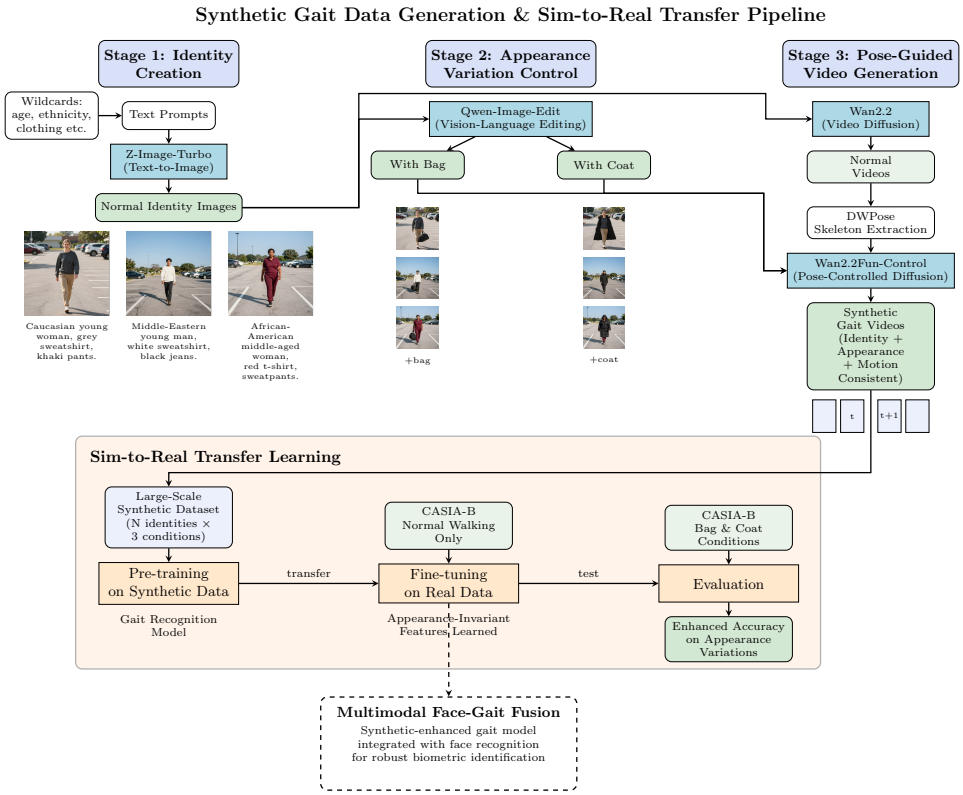

Fig. 1: Overview of the proposed synthetic gait data generation pipeline. Stage 1 creates diverse anchor identities via text-to-image synthesis. Stage 2 generates appearance variations (coat, bag) through vision-language editing while preserving facial identity. Stage 3 produces walking videos using pose-guided synthesis, ensuring motion consistency across appearance conditions. The synthetic dataset enables sim-to-real transfer learning for appearance-robust gait recognition.

B. Synthetic Data Generation Pipeline

We generate synthetic data through a three-stage process, separating the generation of identity, clothing attributes, and walking motion.

1) Stage 1: Text-to-Image Identity Synthesis: We employ a latent diffusion model (Stable Diffusion 3 Z-Image-Turbo [24]) to create a diverse population of anchor identities. To ensure demographic diversity representative of real-world populations, we implement a procedural prompt generation strategy using wildcard-based sampling. Wildcards sample uniformly from predefined sets covering 10 gender/age combinations, 9 ethnicities, 640 clothing combinations (10 colors $\times$ 8 tops $\times$ 8 bottoms), and 4 weather conditions (for the outdoor environment).

To ensure cross-environment robustness, we generate $N = 200$ unique identities at 1024×1024 resolution under the Normal Walking (NM) condition, split equally across two distinct backgrounds: an outdoor parking lot and an indoor underground garage. These anchor images serve as the identity reference for subsequent appearance modification and video generation stages.

2) Stage 2: Appearance Variation via Vision-Language Editing: Direct text-to-image generation of subjects wearing coats or carrying bags can produce inconsistent identities due to the randomness introduced by the diffusion step.

Instead, we use Qwen-Image-Edit [25] for instruction-guided image modification. Given an anchor image I_{NM}, we prompt the model to "*Make the person wear a heavy black winter coat...*" (coat-wearing, CL) or "*Make the person carry a large black duffel bag...*" (bag-carrying, BG).

This approach produces three aligned start frames $\{I_{\text{NM}}, I_{\text{CL}}, I_{\text{BG}}\}$ for each generated identity, with consistent facial features and body structure across conditions. The resulting identity-condition pairs form the input for video generation.

3) Stage 3: Pose-Guided Video Generation: The final stage synthesizes walking videos from the static appearance frames, ensuring motion consistency across appearance conditions.

Substep 3a: Normal Walking Video Generation. We generate an NM video from I_{NM} using Wan2.2 5B TI2V [26] with a natural gait prompt. To mitigate identity drift during synthesis, we incorporate a FaceDetailer module [27] that enhances facial regions across frames.

Substep 3b: Appearance Variant Video Generation. To ensure CL and BG videos exhibit identical gait dynamics to their NM counterparts, we employ Wan2.2 Fun-Control, which accepts video-based motion conditioning. The generated

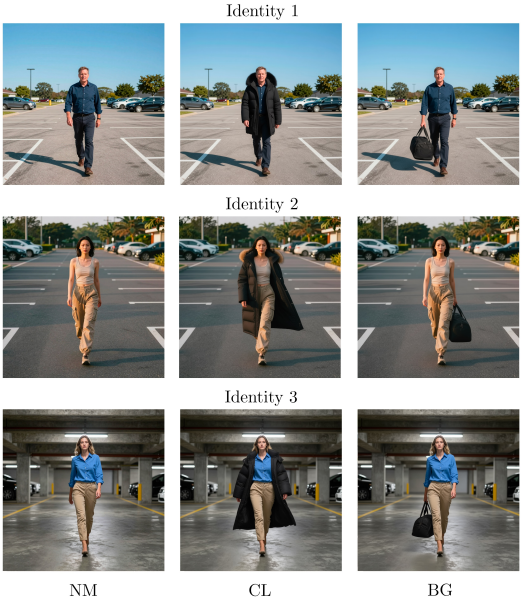

Fig. 2: Synthetic walking sequences: three identities (rows) across NM, CL, and BG conditions (columns), showing identity consistency with controlled appearance variation. Identity 3 is from the indoor environment.

NM video serves as the motion reference, while I_{CL} or I_{BG} defines the clothing. The model synthesizes new pixel content following the exact motion trajectory of the reference video.

The resulting dataset comprises 800 videos (two NM sequences, one CL, one BG per subject, each 121 frames at 24 fps) of subjects that preserve identical temporal gait dynamics, differing only in clothing appearance (Figure 2).

C. Training Strategy

Our approach leverages the synthetic dataset to precondition the model before exposure to real data.

1) Pre-training on Synthetic Data: We train GaitBase [5] on the 200 synthetic identities (800 sequences total) using triplet loss. Pre-training proceeds for 10,000 iterations with Adam optimizer and learning rate 10^{-4}.

2) Fine-tuning on Real Data: The pre-trained model is fine-tuned on CASIA-B subjects 1–74 using *only* normal walking sequences ($lr = 10^{-5}$). Since the model never observes real CL or BG conditions during training, any improvement on these variations is direct evidence of learned appearance invariance.

D. Integration with Multimodal Recognition

Generating RGB video enables application to multimodal biometric systems where both modalities must be extracted from the same input sequence. We evaluate our features within a score-level fusion framework [1], [2]. We extract face features using GhostFaceNets [28] and gait features using our synthetic-enhanced GaitBase model. Identity scores are combined through weighted fusion:

$$s_{\text{fusion}} = \alpha \cdot s_{\text{face}} + (1 - \alpha) \cdot s_{\text{gait}} \tag{1}$$

where α is determined empirically. By improving gait robustness under appearance variations, our synthetic pre-training reduces reliance on the face modality when gait would otherwise fail, improving overall system reliability.

IV. Experiments

We evaluate our synthetic data generation approach on the CASIA-B benchmark, focusing on recognition performance under appearance variations never observed during real-data training.

A. Experimental Setup

Datasets & Protocol. To ensure cross-environment robustness, our expanded synthetic dataset comprises 200 identities generated across two diverse backgrounds (100 in an outdoor parking lot, 100 in an indoor underground garage). Each identity is rendered under NM, CL, and BG conditions at a frontal viewing angle (0°). For each subject, one NM sequence is used to guide generation of the CL, BG and second NM sequences. Silhouettes are extracted using HumanSegV2 [29] and resized to 64×44. We evaluate on the CASIA-B dataset [3]; following standard protocol, subjects 1–74 are used for training and subjects 75–124 for testing. Since our synthetic data generation targets frontal approaches (0° viewing angle), we focus our evaluation on this angle to ensure alignment between synthetic pre-training and real-data fine-tuning conditions. For the test set (subjects 75–124), NM01 serves as the gallery for each viewing angle, while all remaining sequences (NM02–06, BG01–02, CL01–02) constitute the probe set.

Implementation Details. We employ GaitBase [5] via the OpenGait framework. We compare four configurations: *Real (NM, BG, CL)* (trained on all real CASIA-B conditions), *Real, NM-only* (trained solely on real NM sequences), *Synthetic (NM, BG, CL)* (trained on the 200 synthetic subjects), and *Ours* (pre-trained on synthetic data, fine-tuned on real NM).

For the baseline experiments (Real, All Conditions and NM-Only), we train for 60,000 iterations with initial learning rate 0.1, reduced by factor 0.1 at iterations 20K, 40K, and 50K. The Synthetic-Only experiment follows the same schedule.

For our pre-training + fine-tune approach, pre-training on synthetic data proceeds for 40,000 iterations with learning rate 0.1 and milestones at 15K, 30K, and 35K iterations. Fine-tuning on CASIA-B uses a reduced learning rate of 0.01 for 20,000 iterations, with learning rate decay at 10K and 15K iterations.

Extracted synthetic silhouettes preserve core gait patterns (stride length, arm swing) with clean boundaries, showing comparable quality to real CASIA-B silhouettes without requiring explicit segmentation labeling. A comparison between samples of our generated silhouettes and CASIA-B silhouettes is shown in Figure 3.

B. Gait Recognition Results

Table II compares recognition accuracy across configurations. Synthetic pre-training dramatically improves robustness

(a) Synthetic silhouettes

(b) CASIA-B silhouettes

Fig. 3: Comparison between synthetic (top) and CASIA-B (bottom) silhouettes. The synthetic data achieves comparable visual quality and preserves gait-relevant features.

TABLE II: Rank-1 accuracy (%) on CASIA-B test set (subjects 75–124) at 0° viewing angle.

Method	NM	BG	CL	Overall
Real (NM, BG, CL)	**100.0**	**99.0**	91.0	97.8
Real (NM-only)	**100.0**	94.0	31.0	83.3
Synthetic (NM, BG, CL)	92.8	83.0	52.0	81.6
Ours (Pre-train + FT)	**100.0**	98.0	**95.0**	**98.4**

to appearance variations, most notably on the challenging CL (coat/clothing) condition.

Training GaitBase on NM sequences alone yields only 31.0% on CL probes. Our two-stage approach achieves 95% on CL, surpassing the fully-supervised baseline trained on real CL data. This confirms that the text-guided appearance variations capture sufficient diversity to generalize to real-world clothing changes.

Direct sim-to-real transfer (Synthetic-only) achieves 52.0% on CL—substantially better than NM-only—but its 92.8% NM accuracy reveals a domain gap. Fine-tuning bridges this gap while preserving the appearance-robust representations learned during pre-training.

Ablation on Synthetic Scale and Environment. To address the impact of dataset size and environmental diversity, we perform an ablation study comparing pre-training on 100 identities (outdoor only) versus 200 identities (outdoor + indoor garage). As shown in Table III, expanding the synthetic population and introducing cross-environment variations improves CASIA-B Rank-1 accuracy on CL conditions from 91% to 95%. This confirms that scaling our generative pipeline yields direct performance gains in sim-to-real transfer, forcing the model to learn truly background-invariant gait features.

C. Integration with Multimodal Recognition

A key motivation for our work is improving gait robustness within multimodal biometric systems, where face and gait recognition are combined for enhanced person identification. As established in the introduction, such systems are particularly relevant for access control scenarios where subjects approach from frontal angles. We evaluate whether synthetic pre-training improves not only unimodal gait recognition but also the overall multimodal pipeline.

TABLE III: Ablation study on dataset scale and environmental diversity (Rank-1 accuracy in % on CASIA-B at 0°).

Pre-training Data	NM	BG	CL	Overall
Synthetic, 100 IDs	91.2	82.0	50.0	80.0
Synthetic, 200 IDs	92.8	83.0	52.0	81.6
Ours, 100 IDs	100.0	98.0	91.0	97.6
Ours, 200 IDs	**100.0**	**98.0**	**95.0**	**98.4**

TABLE IV: Multimodal face-gait fusion on CASIA-B at 0° viewing angle.

Configuration	Rank-1 (%)	α
Face only	96.9	—
Gait only (NM-only)	83.6	—
Gait only (Ours)	98.7	—
Fusion (NM-only Gait)	98.2	0.8
Fusion (Ours)	**99.6**	0.5

Following established fusion methodologies [1], we employ weighted mean score-level fusion combining normalized face and gait similarity scores (Eq. 1) where α is optimized on a validation set. Face features are extracted using GhostFaceNets [28] with temporal averaging across frames.

Table IV reveals a critical finding: when gait recognition is trained only on normal walking sequences (NM-only), it becomes the weak link in the multimodal pipeline. The NM-only gait model achieves just 83.6% accuracy, forcing the fusion system to compensate by heavily weighting face features ($\alpha = 0.8$). Even with this compensation, fusion accuracy reaches only 98.2%.

In contrast, our synthetic pre-training restores gait to 98.7% accuracy, enabling balanced fusion ($\alpha = 0.5$) where both modalities contribute equally. This results in 99.6% fusion accuracy, demonstrating that robust gait features are essential for effective multimodal recognition. The shift in optimal α from 0.8 to 0.5 quantifies how synthetic pre-training transforms gait from a liability into a genuine asset for multimodal systems.

D. Discussion

Our experiments demonstrate that synthetic data generated through text-to-video diffusion models can effectively address appearance variation challenges in gait recognition.

The 64 percentage point improvement on coat-wearing conditions (Table II: 31% to 95%) achieved without any real appearance-varied training data indicates that our synthetic pipeline generates sufficiently realistic appearance variations for the model to learn invariant features.

Table IV reveals that weak gait features compromise the entire multimodal pipeline. When gait accuracy drops to 83.6% (NM-only), fusion must compensate through face-heavy weighting ($\alpha = 0.8$), limiting system reliability when

face recognition is degraded. Our synthetic pre-training restores balanced fusion.

Limitations & Generative Hallucinations. Recent studies warn that training deep learning models purely on AI-generated data can lead to model collapse and hallucinations [30]. Our approach mitigates this risk: synthetic data is utilized strictly for pre-training to learn appearance-invariant gait features. The final fine-tuning phase relies only on real-world NM data, preventing synthetic artifacts from corrupting the final feature space. Future work will explore expanding this pipeline to multi-view setups.

V. Conclusion

This paper presented a text-to-video generative pipeline to synthesize realistic, appearance-varied gait videos with identity-consistent faces. Our approach addresses gait recognition's vulnerability to clothing changes, restoring its viability as a robust modality in face-gait biometric systems. Experiments on CASIA-B prove the efficacy of our sim-to-real transfer strategy: synthetic pre-training boosts coat-condition accuracy from 31% to 95% without requiring real appearance-varied data. Furthermore, this enhanced robustness enables balanced multimodal fusion ($\alpha = 0.5$), achieving 99.6% overall identification accuracy. These results highlight the potential of generative AI to overcome data scarcity in biometrics. Future work will expand to multi-view generation and validation on larger-scale datasets across diverse environments.

VI. Acknowledgments

Federico Boscolo's work is part of the project PNRR-NGEU which has received funding from the MUR – DM 117/2023.

References

[1] A. Kale, A. Roychowdhury, and R. Chellappa, "Fusion of gait and face for human identification," in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 5. IEEE, 2004, pp. V–901.

[2] X. Zhou and B. Bhanu, "Feature fusion of face and gait for human recognition at a distance in video," in *18th International Conference on Pattern Recognition (ICPR'06)*, vol. 4. IEEE, 2006, pp. 529–532.

[3] S. Yu, D. Tan, and T. 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)*, vol. 4. IEEE, 2006, pp. 441–444.

[4] F. M. Castro, M. J. Marín-Jiménez, N. Guil, and N. Pérez de la Blanca, "Human gait recognition: A survey," *Pattern Recognition*, vol. 43, no. 11, pp. 3901–3914, 2010.

[5] C. Fan, J. Liang, C. Shen, S. Hou, Y. Huang, and S. Yu, "Opengait: Revisiting gait recognition towards better practicality," *arXiv preprint arXiv:2211.06597*, 2023.

[6] X. Huang, D. Zhu, H. Wang, X. Wang, B. Yang, B. He, W. Liu, and B. Feng, "Context-sensitive temporal feature learning for gait recognition," *IEEE Transactions on Image Processing*, vol. 31, pp. 1663–1678, 2022.

[7] W. A. Alsaggaf, I. Mehmood, E. F. Khairullah, S. Alhuraiji, M. F. Sabir, A. S. Alghamdi, A. A. Abd El-Latif, and F. Gomez-Donoso, "A smart surveillance system for uncooperative gait recognition using cycle consistent generative adversarial networks (ccgans)," *Computational Intelligence and Neuroscience*, vol. 2021, p. 3110416, 2021.

[8] J. Han and B. Bhanu, "Individual recognition using gait energy image," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 28, no. 2, pp. 316–322, 2006.

[9] H. Chao, Y. He, J. Zhang, and J. Feng, "Gaitset: Regarding gait as a set for cross-view gait recognition," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, no. 01, pp. 8126–8133, 2019.

[10] C. Fan, Y. Peng, C. Cao, X. Liu, S. Hou, J. Chi, Y. Huang, Q. Li, and Z. He, "Gaitpart: Temporal part-based model for gait recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 14 213–14 221.

[11] C. Zhang, X. Chen, G. Han, and X. Liu, "Spatial transformer network on skeleton-based gait recognition," *Expert Systems*, vol. 40, no. 6, 2023.

[12] T. Teepe, J. Gilg, F. Herzog, S. Hörmann, and G. Rigoll, "Gaitgraph: Graph convolutional network for skeleton-based gait recognition," *arXiv preprint arXiv:2101.11228*, 2021.

[13] R. Liao, C. Cao, E. B. Garcia, S. Yu, and Y. Huang, "Posegait: A model-based gait recognition method with pose estimation," in *Chinese Conference on Biometric Recognition*. Springer, 2020, pp. 12–21.

[14] K. Xue, H. Seo, C. Bobenrieth, and G. Luo, "Shape-conditioned human motion diffusion model with mesh representation," *Computer Graphics Forum*, vol. 44, no. 2, p. e70065, 2025.

[15] J. Xie, W. Yu, X. Cao, Q. Tian *et al.*, "Synchronized multi-frame diffusion for video generation," *Computer Graphics Forum*, vol. 44, no. 6, p. e70095, 2025.

[16] Z. Jiang, W. Li, J. Zhang *et al.*, "Tempdiff: Temporal consistency for video generation with diffusion models," *Computer Graphics Forum*, vol. 43, no. 7, p. e15211, 2024.

[17] A. Sawicki, "Behavioral gait biometrics in vr: Is the use of synthetic samples able to increase person identification metrics?" *Computer Animation and Virtual Worlds*, vol. 36, no. 2, p. e70016, 2025.

[18] J. Gao, S. Zhang, X. Guan, X. Meng, and D. Wu, "Multiview gait recognition based on slack allocation generation adversarial network," *Wireless Communications and Mobile Computing*, vol. 2022, p. 1648138, 2022.

[19] A. Kerim, C. Aslan, U. Celikcan, E. Erdem, and A. Erdem, "Nova: Rendering virtual worlds with humans for computer vision tasks," *Computer Graphics Forum*, vol. 40, no. 6, pp. 258–272, 2021.

[20] R. Liao, Y. Zhang, H. Wang, L. Lu, Z. Chen, X. Wang, and W. Zuo, "An effective ship detection approach combining lightweight networks with supervised simulation-to-reality domain adaptation," *Computer-Aided Civil and Infrastructure Engineering*, vol. 40, no. 27, pp. 4732–4757, 2025.

[21] J. Tobin, R. Fong, A. Ray, J. Schneider, W. Zaremba, and P. Abbeel, "Domain randomization for transferring deep neural networks from simulation to the real world," in *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2017, pp. 23–30.

[22] A. Grigorev, V. Kovalenko, O. Voynov, M. Guo, E. Trnavskaya, S. Han, V. Kalogeiton, and V. Lempitsky, "Synthetic data for deep learning in computer vision and medical imaging: A means to reduce data bias," *ACM Computing Surveys*, vol. 57, no. 1, pp. 1–38, 2024.

[23] Y. Toda, F. Okura, J. Ito, S. Okada, T. Kinoshita, H. Tsuji, and D. Saisho, "Training instance segmentation neural network with synthetic datasets for crop seed phenotyping," *Communications Biology*, vol. 3, no. 1, p. 173, 2020.

[24] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 10 684–10 695.

[25] C. Wu, J. Li, J. Zhou, J. Lin, K. Gao, K. Yan *et al.*, "Qwen-image technical report," 2025. [Online]. Available: https://arxiv.org/abs/2508.02324

[26] T. Wang *et al.*, "Wan: Open and advanced large-scale video generative models," *arXiv preprint arXiv:2503.20314*, 2025.

[27] ltdrdata, "Comfyui impact pack," https://github.com/ltdrdata/ComfyUI-Impact-Pack, 2024, accessed: 2025.

[28] M. Alansari, O. Abdul Hay, S. Javed, A. Shoufan, Y. Zweiri, and N. Werghi, "Ghostfacenets: Lightweight face recognition model from cheap operations," *IEEE Access*, vol. 11, pp. 35 429–35 446, 2023.

[29] Y. Liu, L. Chu, G. Chen, Z. Wu, Z. Chen, B. Lai, and Y. Hao, "PaddleSeg: A high-efficient development toolkit for image segmentation," 2021.

[30] S. Alemohammad, J. Casco-Rodriguez, L. Luzi, A. I. Humayun, H. Babaei, D. LeJeune, A. Siahkoohi, and R. Baraniuk, "Self-consuming generative models go mad," in *International Conference on Learning Representations*, vol. 2024, 2024, pp. 53 581–53 608.