



Survey of GaitSet-Based Frameworks for Low-Resolution Gait Recognition

¹Feng Jiale, ^{2,*}Kazem Chamran

¹City Graduate School, City University Malaysia, Kuala Lumpur, Malaysia;

²Department of Information Technology, City University Malaysia, Kuala Lumpur, Malaysia

Email: ¹feng.jiale864@gmail.com, ²kazem.chamran@city.edu.my

*Corresponding author: Kazem Chamran (kazem.chamran@city.edu.my)

Abstract- Gait recognition remains useful at distances at which facial detail is unavailable, but low-resolution surveillance video removes the spatial cues required by many silhouette models. GaitSet represents a sequence as an unordered set of frames and is relatively tolerant of view variation and temporal misalignment. Its fixed max or mean pooling, however, cannot distinguish reliable silhouettes from frames damaged by downsampling, compression, blur, occlusion, or segmentation error. This focused narrative survey reviews GaitSet and closely related low-resolution gait methods across three intervention points: input enhancement, multi-scale representation learning, and set-level aggregation. Later architectures are separated into a Transformer used after convolutional frame encoding, reliability-aware or multi-scale aggregation that retains a GaitSet-style backbone, and an end-to-end Transformer gait encoder. Source-reported Rank-1 results are compared only within their stated protocols. The literature provides few matched evaluations at both standard 64×44 or 64×64 resolution and explicit 32×32 or compression conditions. Results from CASIA-B, Gait3D, and GREW are also not directly interchangeable because the datasets and evaluation protocols differ substantially. Across the reviewed evidence, reliability-aware aggregation addresses the immediate failure of low-resolution sequences more directly than adding attention alone. A practical design should estimate frame and region quality before set aggregation, then add temporal refinement or a Transformer backbone only when the training data and deployment budget support the added complexity.

Keywords: Gait Recognition, Low-Resolution Surveillance, Gaitset, Set Pooling, Reliability-Aware Aggregation, Transformer, Silhouette Quality.

I. Introduction

Gait recognition identifies people from patterns in walking and can operate without close-range imaging or active cooperation. These properties make it relevant to surveillance, but the same operating conditions that make gait attractive also damage the signal: a distant subject may occupy only a small part of the frame, and the extracted silhouette may contain blur, compression artifacts, occlusion, and segmentation error. Broad reviews of deep gait recognition consistently identify data quality and transfer from controlled datasets to operational video as persistent limitations (Munusamy & Senthilkumar, 2024; Sepas-Moghaddam & Etemad, 2023; Shen et al., 2025).

Resolution loss weakens limb contours, joint configuration, and local motion cues. It also increases variation among frames from the same sequence. A recognizer must therefore solve two related problems: extract discriminative information from small silhouettes and prevent the poorest frames from dominating the sequence representation. Evaluations of silhouette-based systems show that protocol design, silhouette quality, and implementation details can alter reported performance substantially (Fan et al., 2023b; Hou et al., 2023). Results should consequently be interpreted within the dataset and protocol reported by the source.

GaitSet is central to this discussion because it models a gait sequence as a set rather than a strictly ordered clip (Chao et al., 2019, 2022). A convolutional encoder processes each silhouette independently, and permutation-invariant pooling produces a sequence representation. The design avoids explicit temporal alignment and works with variable numbers of frames. Its weakness under degraded video is equally clear: conventional max or mean pooling does not estimate whether a frame is sharp, complete, or correctly segmented.

Research that responds to this weakness has developed along several paths. Input enhancement attempts to restore or normalize small silhouettes. Multi-scale and part-based encoders preserve structure at different spatial or temporal ranges. Attention-based methods refine relations among frames, while quality-aware models assign different importance to frames or regions. Transformer models appear in more than one location in the pipeline, so describing every attention-based method as a Transformer backbone obscures meaningful architectural differences.

This article reviews GaitSet-centered work with emphasis on low-resolution and otherwise degraded surveillance video. It distinguishes a late Transformer refiner, adaptive set aggregation, and an end-to-end Transformer backbone. The comparison retains only results reported by the original publications and separates laboratory CASIA-B scores from in-the-wild Gait3D and GREW results. The purpose is not to identify a universal winner, but to clarify which part of the recognition pipeline each method changes and what evidence supports its use under degraded input.

II. Literature Review

2.1 Datasets and Evaluation Practice

CASIA-B remains a common controlled benchmark for clothing, carrying, and view changes (Yu et al., 2006). OU-MVLP expands the number of subjects and viewpoints and is useful for large-scale cross-view evaluation (Takemura et al., 2018). Neither dataset reproduces the full combination of distance, compression, background clutter, and segmentation failure found in open surveillance systems.

GREW and Gait3D move evaluation into less constrained environments. GREW contains naturally collected sequences with large variation in background, appearance, and image quality (Zhu et al., 2021). Gait3D combines in-the-wild video with dense 3D body representations and multi-camera capture (Zheng et al., 2022). These benchmarks are important for practical testing, but their Rank-1 results cannot be ranked directly against CASIA-B normal, bag, and clothing scores.

Low-resolution protocols are less standardized. Some studies downsample silhouettes to a fixed size, some evaluate highly compressed video, and others use in-the-wild datasets in which resolution varies together with many other factors. Niculae et al. (2024) show that compression itself deserves explicit study rather than being treated as an incidental nuisance. A reliable comparison must report the resolution, degradation process, split, gallery protocol, and whether identical views are excluded.

2.2 Technical Taxonomy and Reported Evidence

Figure 1 organizes the literature by the point at which a method responds to low-resolution input. The three lower branches are complementary. A system may combine enhancement, multi-scale representation, and adaptive aggregation, but the contribution of each component should be measured separately.

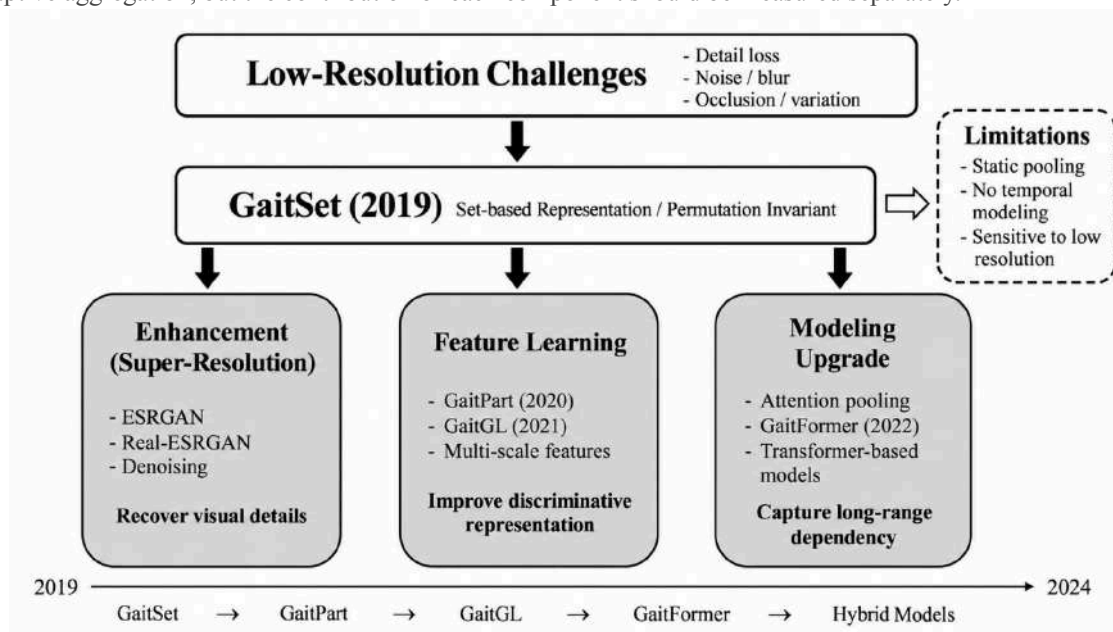


Figure 1. Technical responses to low-resolution gait input in GaitSet-centered recognition.

Table 1 lists representative source-reported Rank-1 results. It is a protocol map, not a league table. Values from different datasets, splits, resolutions, and gallery rules should not be compared as though they measured the same task.

Table 1. Representative source-reported Rank-1 results and evaluation protocols.

Method	Protocol	Rank-1 (%)	Primary change	Source
GaitSet	CASIA-B, 64×44 , identical view excluded	95.0 / 87.2 / 70.4 NM/BG/CL	Set encoder with fixed pooling	Chao et al. (2019)
GaitSet	OU-MVLP, 64×44	87.1 mean	Set encoder with fixed pooling	Chao et al. (2019)
GaitPart	CASIA-B, 64×44	96.2 / 91.5 / 78.7	Part-level temporal pooling	Fan et al. (2020)
GaitGL	CASIA-B, 64×44	97.4 / 94.5 / 83.6	Global-local representation	Lin et al. (2022)
MFCF-Gait	CASIA-B, 64×64	96.49 / 91.42 / 78.24	Super-resolution plus multi-scale fusion	Song et al. (2024)
MFCF-Gait	CASIA-B, 32×32	94.23 / 87.68 / 71.57	Super-resolution plus multi-scale fusion	Song et al. (2024)
QAGait	Gait3D quality-aware evaluation	67.0	Quality assessment and quality-aware loss	Wang et al. (2024)
TransGait	CASIA-B, identical view excluded	98.1 / 94.9 / 85.8	Multimodal features plus Set Transformer	Li et al. (2023)
QAT-GaitSet	CASIA-B with LR, JPEG, and blur; non-standard split	95.8	Frame quality estimation and adaptive fusion	Jiale & Chamran (2025)
Gait Transformer	Gait3D / GREW	77.1 / 82.6	End-to-end Transformer backbone	Hou et al. (2026)
SwinGait-3D	Gait3D / GREW	75.0 / 79.3	Deep hybrid Transformer model	Fan et al. (2023a)

Note. NM = normal walking; BG = bag carrying; CL = clothing change; LR = low resolution. Results are reproduced from the cited sources and remain tied to their original protocols.

2.3 Input Enhancement

Super-resolution and denoising attempt to improve the silhouette before recognition. This strategy is intuitive when the input is smaller than the model's nominal size, but reconstructed edges are estimates rather than recovered measurements. Hallucinated contours can introduce identity cues that were not present in the captured sequence. The extra preprocessing stage also creates another model to train, validate, and deploy.

MFCF-Gait combines super-resolution preprocessing with multi-scale feature cross-fusion. Its matched CASIA-B results at 64×64 and 32×32 provide unusually direct evidence of the cost of reduced spatial resolution: the clothing-change score falls from 78.24% to 71.57% even after enhancement and fusion (Song et al., 2024). The result supports enhancement as a useful aid but not a complete solution. Compression-aware evaluation reaches a similar conclusion: image quality degradation must be represented in testing and cannot be inferred from nominal pixel dimensions alone (Niculae et al., 2024).

2.4 Multi-Scale and Part-Based Representation

Part-based and global-local models preserve information that whole-body pooling may suppress. GaitPart uses temporal aggregation within body regions and improves CASIA-B performance over early set baselines (Fan et al., 2020). GaitGL combines global and local representations, while GaitASMS adapts structured spatial representation and multi-scale temporal aggregation (Lin et al., 2022; Sun et al., 2024). GaitMGL fuses global-local information across temporal scales, and MCAT uses multi-scale temporal context with Transformer-based aggregation (Zhang et al., 2024; Zhu et al., 2022).

These models do not all target low resolution explicitly, but their design is relevant because coarse global motion and local body regions degrade at different rates. Multi-scale features reduce dependence on one spatial granularity. They still require enough signal for the encoder to distinguish a useful region from segmentation noise, and they do not automatically decide which frames deserve more weight.

2.5 GaitSet Under Low Resolution

GaitSet's set formulation is well suited to irregular sampling and variable sequence length (Chao et al., 2019, 2022). The difficulty is not permutation invariance itself; it is the use of fixed aggregation after frame encoding.

Low-resolution surveillance produces sequences in which frame quality varies because blur, compression, occlusion, and segmentation failure do not occur uniformly over time.

QAGait approaches the problem from a quality perspective and reports a Gait3D Rank-1 increase from a GaitBase reference of 59.7% to 67.0% (Wang et al., 2024). QAT-GaitSet estimates degradation at frame level and applies adaptive multi-scale fusion within a GaitSet pipeline. Under its reported degraded CASIA-B protocol, the full model reaches 95.8% Rank-1 and lowers equal error rate relative to the plain baseline (Jiale & Chamran, 2025). Because that split is not the standard CASIA-B protocol, the figure should not be compared directly with the 95.0% normal-walking score reported by Chao et al. (2019).

2.6 Transformer Integration and Adaptive Aggregation

2.6.1 Mode A Transformer Refiner

Mode A retains a convolutional frame encoder and applies attention after spatial features have been extracted. TransGait uses multimodal gait features with a Set Transformer and reports strong CASIA-B performance, particularly under clothing change (Lee et al., 2019; Li et al., 2023). Recurring GaitSet similarly adds a recurring Transformer-style layer to filter ineffective features without replacing the GaitSet stem (Kuang & Li, 2023).

The benefit is broader temporal or relational context at lower cost than a full Transformer backbone. The limitation is architectural: if refined tokens are still passed through fixed set pooling, the model has not solved unequal frame reliability. Context-sensitive temporal learning offers a related non-Transformer example of refining temporal context after convolutional features (Huang et al., 2021).

2.6.2 Mode B Adaptive Set Aggregation

Mode B changes the aggregation rule. Frame quality, scale, or region reliability becomes an explicit weight in set pooling or horizontal pyramid mapping. This is the most direct response to a sequence in which some silhouettes are substantially cleaner than others. QAGait, MFCF-Gait, and QAT-GaitSet implement different versions of this idea through quality-aware learning, multi-scale fusion, or degradation-conditioned weighting (Jiale & Chamran, 2025; Song et al., 2024; Wang et al., 2024).

Adaptive weighting cannot reconstruct detail that is absent from every frame. It is most useful when quality varies within the sequence and cleaner observations still exist. Its additional cost is usually smaller than that of an end-to-end Transformer backbone, which makes it attractive for edge or multi-camera deployment.

2.6.3 Mode C Transformer Backbone

Mode C uses a Transformer as the primary gait encoder rather than a module attached after GaitSet pooling. Gait Transformer combines hybrid patch embedding, decomposed spatial-temporal mixing, and positional encoding in an end-to-end silhouette model (Hou et al., 2026). The model reports 77.1% Rank-1 on Gait3D and 82.6% on GREW. SwinGait is another deep Transformer-oriented series evaluated on practical datasets (Fan et al., 2023a). Self-supervised vision Transformer features and specialized spatial or temporal variants further demonstrate the breadth of this design space (Cosma et al., 2023; Sheng, 2025).

A Transformer backbone can model relations throughout the network, but attention increases memory and compute as sequence length grows. Data requirements also rise because the model has weaker built-in spatial inductive bias than a conventional convolutional encoder. These costs matter in low-resolution research, where realistic labeled data are already limited. Robustness studies reinforce the need to evaluate corruption types explicitly rather than assuming that strong average accuracy implies stable performance under degradation (Sayera et al., 2026).

2.7 Comparison and Open Issues

The three modes intervene at different points. Mode A adds context after convolutional encoding. Mode B changes which observations influence the set representation. Mode C replaces the principal encoder. They can be combined, but their effects should not be conflated.

Two evidence gaps remain. First, few papers evaluate one model under both a standard silhouette size and a severe low-resolution or compression protocol. MFCF-Gait is a useful exception because it reports matched 64×64 and 32×32 results. Second, in-the-wild datasets mix resolution loss with pose, background, occlusion, camera, and segmentation changes. A higher or lower Rank-1 score may reflect a different task rather than a better response to low resolution.

III. Survey Methodology

This article is a focused narrative survey rather than a systematic review. The retained literature was limited to work that defines GaitSet, evaluates silhouette quality or degradation, reports a low-resolution or compression condition, introduces a closely related aggregation strategy, or provides a Transformer design that clarifies

where attention enters the gait pipeline. Conference papers, journal articles, and identifiable arXiv preprints were accepted when their bibliographic records and original manuscripts could be verified.

Quantitative values were transcribed only when the source reported the dataset, protocol, and metric. No missing score was estimated, and no laboratory result was converted into an in-the-wild equivalent. CASIA-B NM/BG/CL values are presented together, while Gait3D and GREW values remain labeled by dataset. A method was assigned to one primary architectural mode according to the component it changes most directly: late relational refinement, adaptive set aggregation, or backbone replacement.

The comparison emphasizes model structure, evidence under degradation, and deployment cost. It does not rank methods across incompatible protocols. The synthesis excludes non-visual sensing, gait-emotion classification, and publications whose reported experiments do not address low-resolution visual recognition.

IV. Results and Discussion

4.1 Principal Findings

The reviewed evidence supports a clear distinction between preserving information and deciding which information to trust. Super-resolution and multi-scale features address representation loss. Temporal attention improves relations among observations. Reliability-aware aggregation determines how much each observation contributes. Under mixed degradation, this last decision is unavoidable because two frames with the same nominal size may have very different recognition values.

Reported results also show why resolution alone is an incomplete descriptor. MFCF-Gait records a measurable decline from 64×64 to 32×32 , while QAGait and QAT-GaitSet improve performance by modeling quality rather than pixel count alone. The practical implication is that low-resolution benchmarks should include blur, compression, and segmentation error, report their severity, and retain a clean baseline for comparison.

4.2 Remaining Technical Challenges

Spatial information loss is irreversible once every frame contains only a coarse outline. Enhancement can regularize shape but may also invent edges. A useful recognizer must therefore combine conservative preprocessing with features that exploit stable body proportions and motion without treating reconstructed texture as evidence.

Quality varies across time and body region. A single scalar frame score may miss the fact that the upper body is occluded while the legs remain informative, or that compression affects narrow limbs more than the torso. Region-level uncertainty and frame-level quality should be calibrated together. This requirement aligns naturally with horizontal-part representations but remains under-evaluated.

Cross-resolution generalization is still weak because training and deployment distributions differ in camera distance, sensor processing, compression, and segmentation. Synthetic downsampling is useful for controlled experiments but does not reproduce the full capture pipeline. Compression-aware work and in-the-wild benchmarks provide a better starting point, although they make causal attribution more difficult (Niculae et al., 2024; Zheng et al., 2022; Zhu et al., 2021).

Efficiency constrains architectural choice. A full Transformer can model long-range relations but may be unnecessary when a small quality estimator and adaptive pool address the dominant failure. Conversely, quality weighting cannot recover information when every frame is poor. The appropriate design depends on sequence length, hardware, latency, and the expected distribution of degradations.

4.3 Performance Robustness and Efficiency Trade-Offs

Table 2. Trade-offs among low-resolution gait recognition strategies.

Strategy	Pipeline change	Evidence under LR	Compute	Principal limitation
Input enhancement	Pixels before recognition	Useful for size normalization; direct compression and 32×32 tests exist	Medium-high	May create artificial contours; adds a separate stage
Multi-scale representation	Spatial or temporal features	Matched 64×64 and 32×32 results show partial resilience	Low-medium	Cannot replace detail missing from all scales
GaitSet baseline	Unordered set with fixed pooling	Strong standard benchmarks; limited matched LR tests	Low	Poor frames can influence the pooled representation
Mode A refiner	Attention after frame encoding	Strong standard results; dedicated LR evidence remains limited	Medium	Fixed pooling may remain after refinement



Strategy	Pipeline change	Evidence under LR	Compute	Principal limitation
Mode B aggregation	Reliability or scale weights in pooling	Direct quality-aware and degraded-protocol evidence	Medium	Weighting cannot recover universally bad input
Mode C backbone	Transformer-based encoder	Strong Gait3D and GREW results	High	Data, memory, and latency requirements

Table 2 suggests a practical ordering. If frame quality varies, aggregation should account for reliability before a heavier temporal refiner or backbone is introduced. When all frames are severely degraded, representation learning and additional sensing become more important than weighting.

4.4 Deployment Implications

A deployable system needs evaluation at the capture conditions it will encounter. Reporting only resized silhouettes can hide the effect of codec settings, exposure, motion, and segmentation. Validation should include camera-distance bands, compression levels, blur severity, occlusion, and failure cases, together with throughput and memory use.

The comparison also favors calibrated abstention. If a quality estimator finds that an entire sequence is unreliable, forcing an identity prediction can be worse than requesting more frames or returning an uncertain result. Current Rank-1 protocols rarely measure this behavior. Open-set identification, calibration, and failure detection deserve more attention alongside closed-set accuracy.

Privacy and governance are outside the quantitative scope of the reviewed benchmarks but remain part of real deployment. Long-range biometric identification affects people who may not know that they are being analyzed. Technical evaluation should therefore be paired with lawful purpose, access controls, retention limits, and independent review.

4.5 Research Priorities

Future studies should publish matched resolution ladders and mixed-degradation protocols, including the exact preprocessing and gallery rules. A single model evaluated at 64×44 , 32×32 , and realistic compression levels would be more informative than isolated best results on unrelated datasets.

Reliability should be estimated at both frame and region levels and evaluated for calibration, not only for final Rank-1. Models should show whether predicted quality corresponds to actual recognition error and whether weighting remains stable when degradation types change.

Hybrid systems should be compared under equal resource budgets. A small Mode B module, a Mode A refiner, and a Mode C backbone may have similar accuracy but very different latency and memory. Reporting parameter count, operations, sequence length, and device throughput would make architectural choices easier to evaluate.

Cross-dataset testing is necessary to separate benchmark adaptation from genuine low-resolution robustness. Training on one capture pipeline and testing on another should become routine, with synthetic-to-real transfer reported separately from within-dataset performance.

V. Conclusion

GaitSet remains a useful foundation for low-resolution gait recognition because it handles variable-length, irregularly sampled sequences without requiring precise temporal alignment. Its fixed pooling is also the point at which degraded frames can contaminate the sequence representation.

The literature separates into three complementary responses. A late Transformer refiner models relations after convolutional encoding. Adaptive aggregation estimates which frames, scales, or regions should count. A Transformer backbone models spatial-temporal relations throughout the encoder. Source-reported results do not identify one universal architecture because the datasets and protocols differ. The most consistent design principle is to model reliability explicitly before increasing backbone complexity.

Progress will depend on matched low-resolution evaluations, realistic mixed degradations, calibrated quality estimates, and resource reporting. These practices would make it possible to distinguish genuine robustness from gains that depend on a specific benchmark or preprocessing pipeline.

Ethical Approval

This article is a literature review and reports no new study involving human participants or animals; therefore, institutional ethical approval was not required.

Data Availability

No new dataset was generated or analyzed for this review. The quantitative results discussed in the article are available in the cited publications and their associated repositories where provided by the original authors.

References

1. Chao, H., He, Y., Zhang, J., & Feng, J. (2019). GaitSet: Regarding gait as a set for cross-view gait recognition. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(1), 8126-8133. <https://doi.org/10.1609/aaai.v33i01.33018126>
2. Chao, H., Wang, K., He, Y., Zhang, J., & Feng, J. (2022). GaitSet: Cross-view gait recognition through utilizing gait as a deep set. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(7), 3467-3478. <https://doi.org/10.1109/TPAMI.2021.3057879>
3. Cosma, A., Catruna, A., & Radoi, E. (2023). Exploring self-supervised vision transformers for gait recognition in the wild. *Sensors*, 23(5), 2680. <https://doi.org/10.3390/s23052680>
4. Fan, C., Hou, S., Huang, Y., & Yu, S. (2023a). Exploring deep models for practical gait recognition. *arXiv*. <https://doi.org/10.48550/arXiv.2303.03301>
5. Fan, C., Liang, J., Shen, C., Hou, S., Huang, Y., & Yu, S. (2023b). OpenGait: Revisiting gait recognition toward better practicality. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 9707-9716). <https://doi.org/10.1109/CVPR52729.2023.00936>
6. Fan, C., Peng, Y., Cao, C., Liu, X., Hou, S., Chi, J., Huang, Y., Li, Q., & He, Z. (2020). GaitPart: Temporal part-based model for gait recognition. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 14213-14221). <https://doi.org/10.1109/CVPR42600.2020.01423>
7. Hou, S., Fan, C., Cao, C., Liu, X., & Huang, Y. (2023). A comprehensive study on the evaluation of silhouette-based gait recognition. *IEEE Transactions on Biometrics, Behavior, and Identity Science*, 5(2), 196-208. <https://doi.org/10.1109/TBIOM.2022.3216857>
8. Hou, S., Lang, W., Wang, J., Huang, Y., Wang, L., & Huang, Y. (2026). Gait Transformer: End-to-end Transformer backbone for gait recognition. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(6), 4771-4779. <https://doi.org/10.1609/aaai.v40i6.42479>
9. Huang, X., Zhu, D., Wang, H., Wang, X., Yang, B., He, B., Liu, W., & Feng, B. (2021). Context-sensitive temporal feature learning for gait recognition. In *2021 IEEE/CVF International Conference on Computer Vision* (pp. 12889-12898). <https://doi.org/10.1109/ICCV48922.2021.01267>
10. Jiale, F., & Chamran, M. K. (2025). Robust gait recognition in complex degraded surveillance videos using a quality-guided, temporally-aware GaitSet model. In *2025 International Conference on Computers, Information Processing and Advanced Education* (pp. 249-253). <https://doi.org/10.1109/CIPAE66821.2025.00049>
11. Kuang, R., & Li, P. (2023). Recurring GaitSet: A gait recognition method based on deep learning and recurring layer Transformer. In *2023 5th International Academic Exchange Conference on Science and Technology Innovation* (pp. 576-579). <https://doi.org/10.1109/IAECST60924.2023.10503545>
12. Lee, J., Lee, Y., Kim, J., Kosiorek, A., Choi, S., & Teh, Y. W. (2019). Set Transformer: A framework for attention-based permutation-invariant neural networks. *Proceedings of Machine Learning Research*, 97, 3744-3753. <https://proceedings.mlr.press/v97/lee19d.html>
13. Li, G., Guo, L., Zhang, R., Qian, J., & Gao, S. (2023). TransGait: Multimodal-based gait recognition with set Transformer. *Applied Intelligence*, 53(2), 1535-1547. <https://doi.org/10.1007/s10489-022-03543-y>
14. Lin, B., Zhang, S., Wang, M., Li, L., & Yu, X. (2022). GaitGL: Learning discriminative global-local feature representations for gait recognition. *arXiv*. <https://doi.org/10.48550/arXiv.2208.01380>
15. Munusamy, V., & Senthilkumar, S. (2024). Emerging trends in gait recognition based on deep learning: A survey. *PeerJ Computer Science*, 10, e2158. <https://doi.org/10.7717/peerj-cs.2158>
16. Niculae, A., Catruna, A., Cosma, A., Rosner, D., & Radoi, E. (2024). Gait recognition from highly compressed videos. In *2024 IEEE 18th International Conference on Automatic Face and Gesture Recognition* (pp. 1-7). <https://doi.org/10.1109/FG59268.2024.10581961>
17. Sayera, R., Kumar, A., Mitra, S., Kamtam, P., & Rawat, Y. S. (2026). RobustGait: Robustness analysis for appearance-based gait recognition. In *2026 IEEE/CVF Winter Conference on Applications of Computer Vision*. <https://doi.org/10.1109/WACV61042.2026.00250>
18. Sepas-Moghaddam, A., & Etemad, A. (2023). Deep gait recognition: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(1), 264-284. <https://doi.org/10.1109/TPAMI.2022.3151865>
19. Shen, C., Yu, S., Wang, J., Huang, G. Q., & Wang, L. (2025). A comprehensive survey on deep gait recognition: Algorithms, datasets, and challenges. *IEEE Transactions on Biometrics, Behavior, and Identity Science*, 7(2), 270-292. <https://doi.org/10.1109/TBIOM.2024.3486345>
20. Sheng, H. (2025). GaitTriViT and GaitVViT: Transformer-based methods emphasizing spatial or temporal aspects in gait recognition. *PeerJ Computer Science*, 11, e3061. <https://doi.org/10.7717/peerj-cs.3061>
21. Song, C., Yun, L., & Li, R. (2024). MFCF-Gait: Small silhouette-sensitive gait recognition algorithm based on multi-scale feature cross-fusion. *Sensors*, 24(17), 5500. <https://doi.org/10.3390/s24175500>
22. Sun, Y., Long, H., Feng, X., & Nixon, M. (2024). GaitASMS: Gait recognition by adaptive structured spatial representation and multi-scale temporal aggregation. *Neural Computing and Applications*, 36(13), 7057-7069. <https://doi.org/10.1007/s00521-024-09445-z>
23. Takemura, N., Makiyama, Y., Muramatsu, D., Echigo, T., & Yagi, Y. (2018). Multi-view large population gait dataset and its performance evaluation for cross-view gait recognition. *IPSJ Transactions on Computer Vision and Applications*, 10, 4. <https://doi.org/10.1186/s41074-018-0039-6>



24. Wang, Z., Hou, S., Zhang, M., Liu, X., Cao, C., Huang, Y., Li, P., & Xu, S. (2024). QAGait: Revisit gait recognition from a quality perspective. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(6), 5785-5793. <https://doi.org/10.1609/aaai.v38i6.28391>
25. Yu, S., Tan, D., & Tan, T. (2006). A framework for evaluating the effect of view angle, clothing and carrying condition on gait recognition. In *18th International Conference on Pattern Recognition* (pp. 441-444). <https://doi.org/10.1109/ICPR.2006.67>
26. Zhang, Z., Wei, S., Xi, L., & Wang, C. (2024). GaitMGL: Multi-scale temporal dimension and global-local feature fusion for gait recognition. *Electronics*, 13(2), 257. <https://doi.org/10.3390/electronics13020257>
27. Zheng, J., Liu, X., Liu, W., He, L., Yan, C., & Mei, T. (2022). Gait recognition in the wild with dense 3D representations and a benchmark. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 20228-20237).
28. Zhu, D., Huang, X., Wang, X., Yang, B., He, B., Liu, W., & Feng, B. (2022). Multi-scale context-aware network with Transformer for gait recognition. *arXiv*. <https://doi.org/10.48550/arXiv.2204.03270>
29. Zhu, Z., Guo, X., Yang, T., Huang, J., Deng, J., Huang, G., Du, D., Lu, J., & Zhou, J. (2021). Gait recognition in the wild: A benchmark. In *2021 IEEE/CVF International Conference on Computer Vision* (pp. 14769-14779). <https://doi.org/10.1109/ICCV48922.2021.01452>