Lightweight Part-wise Temporal Self-Attention for Silhouette Gait Recognition on CASIA-B

Federico Iannini 1931748
Gloria Fiammengo 2254256

Abstract

Gait recognition identifies people by their walking patterns and is commonly evaluated as a cross-view retrieval problem. In this project we present a lightweight silhouette-based architecture designed for the CASIA-B benchmark. Our pipeline extracts frame-level spatial features with a small CNN, summarizes body regions using horizontal part pooling with Generalized Mean (GeM) aggregation, and models gait dynamics with a shared Transformer encoder (self-attention) applied independently to each body part over clips of 30 frames. We follow the standard LT-74 protocol (gallery NM#1–4, probe NM#5–6/BG#1–2/CL#1–2, excluding same-view matches) and report competitive Rank-1 results, with detailed ablations on part division and temporal positional encodings.

1 Introduction

Gait recognition is a biometric identification task: given a probe walking sequence, the goal is to retrieve the correct identity from a gallery. Compared to face recognition, gait can work at distance and in low-resolution scenarios, but it is strongly affected by *view changes* (camera angle), *carrying condition* (BG), and *clothing changes* (CL). For these reasons, CASIA-B is often used to test robustness across 11 views (0°–180°) and multiple walking conditions [1].

A key design choice in gait models is how to aggregate temporal information. Some approaches collapse time early or treat sequences as order-invariant sets, which can work well but may ignore phase evolution and longer-range motion patterns. In this project we adopt a simple structural idea: **factorize space and time**. We first build per-frame part descriptors, then we apply temporal self-attention on short clips to model how each body region evolves over time.

2 Related Work

Strong silhouette-based methods often differ mainly in **(a) spatial partitioning** and **(b) temporal aggregation**. **GaitSet** treats a gait sequence as an unordered set and performs set pooling, which makes the representation less sensitive to the input frame order [2]. **GaitPart** emphasizes that different body regions contribute differently, using explicit horizontal partitions and part-wise temporal modeling [3]. **GaitGL** shows that strong global-local representations and careful pooling strategies (including GeM) are effective for cross-view gait recognition [4].

Recent work also explores attention and multi-modal cues. **MMGaitFormer** combines silhouette and skeleton information using Transformer-style attention for cross-modal fusion [5]. **GaitW** focuses on gait recognition in the wild, where background and capture conditions are less controlled [6]. Our method stays in the **silhouette-only** setting, aiming to be compact while using explicit temporal self-attention.

2.1 Diagnostic Study on Temporal Aggregation in GaitGL

Before proposing our new architecture, we conducted a diagnostic study on GaitGL [4] with the goal of understanding to what extent a CNN-based SOTA exploits temporal information to achieve its performance.
In this phase, we refrained from introducing structural modifications to the original model: preprocessing, backbone, loss functions, and evaluation protocol remained unchanged. The only intervention concerned the temporal aggregation block, which we replaced with Transformer-based alternatives, experimenting with both absolute and relative positional encoding. This approach allowed us to isolate the role of temporal aggregation without introducing confounding factors.
The results highlight that GaitGL achieves high performance despite being largely invariant to temporal ordering, suggesting that temporal max pooling represents an effective shortcut on CASIA-B [7], but at the cost of collapsing the structure of temporal dynamics.
This study provided the conceptual motivation to design an alternative architecture in which temporal modeling is explicit and central from the very beginning.

3 Proposed Method

3.1 Input and data format

We use pre-extracted silhouette frames (no raw RGB videos). Each sequence is organized as `ID/condition/view/frame.png`. During training we sample fixed-length clips of $T = 30$ frames; during evaluation we can use either a center-crop clip protocol or a full-sequence protocol via chunking and averaging.

3.2 Data augmentation and preprocessing

We apply two simple augmentations during training: **(1) Horizontal flipping** randomly mirrors silhouettes to encourage left-right robustness. **(2) Random Erasing** randomly removes a rectangular region of the input, simulating partial occlusions and forcing the model to rely on multiple body regions. Both augmentations are applied only during training, while evaluation uses the original silhouettes to keep the benchmark protocol comparable.

3.3 Spatial backbone (frame-wise CNN)

Given an input clip

$$\mathbf{X} \in \mathbb{R}^{B\times 1\times T\times H\times W}, \quad T = 30,$$

we apply a lightweight CNN to each frame to obtain feature maps:

$$\mathbf{F}_t = \mathrm{CNN}(\mathbf{X}_t), \quad \mathbf{F}_t \in \mathbb{R}^{C\times H'\times W'}.$$

In implementation we reshape $B \times T$ into $B{\cdot}T$ so the CNN runs efficiently on GPU while sharing the same weights across frames.

3.4 Part-based spatial aggregation (stripes + GeM)

Following common gait practice, we split each feature map into P horizontal stripes (parts). This introduces an anatomical bias: upper body and lower body are summarized separately, and legs can be emphasized without requiring keypoints. For each stripe we apply GeM aggregation [8]:

$$\mathrm{GeM}(\{x_i\}) = \left(\frac{1}{N}\sum_{i=1}^{N} x_i^p\right)^{\frac{1}{p}}.$$

GeM generalizes average pooling ($p \approx 1$) and max pooling ($p \to \infty$), allowing the model to retain strong activations while remaining stable. After pooling, each frame is represented as P part vectors $\mathbf{z}_{t,p} \in \mathbb{R}^D$.

3.5 Temporal modeling (Transformer encoder, per part)

For each part p, we build a temporal sequence:

$$\mathbf{Z}_p = [\mathbf{z}_{1,p}, \ldots, \mathbf{z}_{T,p}] \in \mathbb{R}^{T\times D}.$$

We process $\mathbf{Z}_p$ with a shared **Transformer encoder** (encoder-only, no decoder) [9]. Self-attention lets each time step attend to all others, capturing gait dynamics beyond local differences. We prepend a dedicated `[CLS]` token and use its output as a compact part embedding. We experiment with multiple positional encoding strategies (learned absolute, sinusoidal, cycle, conditional/convolutional, and relative encodings).

3.6 Embedding, BNNeck, and training losses

We combine the P part embeddings into the final gait descriptor used for retrieval. Training uses a hybrid objective: (i) cross-entropy classification over training identities, and (ii) triplet loss to shape the embedding space for retrieval. We also use a BNNeck, which stabilizes feature distributions and often improves retrieval performance [10].

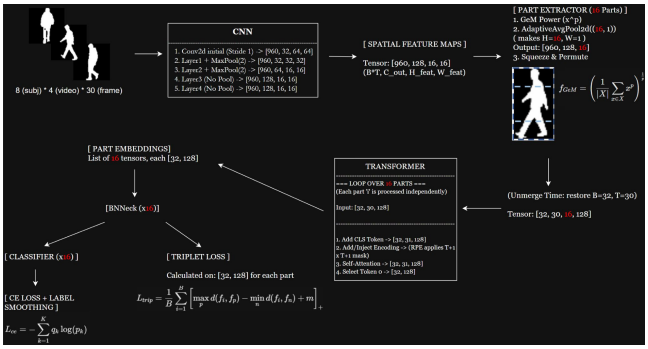

Figure 1: Overview of our model. A clip of $T = 30$ silhouettes is processed by a lightweight CNN, split into P horizontal parts, aggregated with GeM, and modeled over time using a shared Transformer encoder per part. Part embeddings are combined into a final descriptor and trained with cross-entropy and triplet loss (BNNeck).

4 Dataset and Benchmark

4.1 CASIA-B and LT-74 protocol

CASIA-B contains 124 subjects (ID 001–124), three walking conditions (NM, BG, CL), and 11 camera views (0°–180°) [1]. We follow the standard LT-74 split: 74 subjects for training (ID 001–074) and 50 for testing (ID 075–124). The protocol does not define an official validation set, so we keep fixed hyperparameters and no early stopping.

4.2 Evaluation protocol (cross-view retrieval)

Gallery sequences are NM#1–4, probes are NM#5–6, BG#1–2, CL#1–2. We exclude same-view pairs and report Rank-1 accuracy averaged over the 11 views, as commonly done in CASIA-B reports [2].

For full-sequence testing, we extract embeddings on multiple 30-frame chunks and average them into a single descriptor per sequence.

5 Results

5.1 Training setup

We use identity-balanced batches to support metric learning. Specifically, we use a P×K sampler with $P = 8$ identities and $K = 4$ clips per identity (batch size 32). We optimize $L = L_{\text{triplet}} + L_{\text{CE}}$ using AdamW (weight

decay 10^{-4}), with a step schedule that reduces the learning rate by a factor of 10 at epochs 30 and 50 for a total of 60 epochs. We train with mixed precision (FP16) to reduce memory usage and improve throughput.

5.2 Experimenting number of parts and positional encoding

We study how horizontal partitioning affects performance by varying P (e.g., 1, 2, 4, 8, 16, and a coarse 3-region split head/torso/legs inspired by MMGaitFormer's region mapping) [5]. The results show that part-based representations matter: too few parts can mix upper and lower body cues, while a moderate-to-high number of parts better isolates discriminative motion patterns.

Condition	1 Part *(LPE)*	Corpse *(LPE)*	8 Parts *(LPE)*	16 Parts *(LPE)*
NM (Normal)	92.18%	95.66%	96.92%	**97.81%**
BG (Bag)	81.43%	93.59%	95.12%	**96.23%**
CL (Clothing)	40.91%	74.82%	76.91%	**80.27%**
Average	*71.51%*	*88.02%*	*89.65%*	***91.44%***

Figure 2: Rank-1 comparison across different numbers of horizontal parts.

We also compare multiple temporal positional encodings (sinusoidal, cycle, learned absolute, CPE, RPE). With $T = 30$, we observe similar performance across encodings, consistent with short temporal windows and already-strong part features.

Encoding	NM *(Normal)*	BG *(Bag)*	CL *(Clothing)*	Average *(Mean)*
LPE (Absolute)	97.81%	95.79%	**80.27%**	91.29%
CPE (Conditional)	97.76%	96.15%	**81.27%**	91.73%
Cycle (Periodic)	97.32%	95.32%	**81.64%**	91.43%
RPE (Relative)	98.06%	96.57%	**82.00%**	92.21%

Figure 3: Comparison of positional encodings.

5.3 Comparison to SOTA

Our model adopts horizontal body partitioning, but uses it in a direct way (fixed P stripes + GeM per stripe, then per-stripe temporal attention).

GaitSet uses horizontal partitions inside its *Horizontal Pyramid Mapping (HPM)* module: the set-level feature map is split into multiple horizontal strips at multiple pyramid scales; each strip is pooled (max + average) and mapped by independent fully connected layers, then concatenated into the final representation [2]. **GaitPart** applies *Horizontal Pooling (HP)* by splitting each frame-level feature map horizontally into n parts; each part is reduced by global average pooling plus global max pooling (summed), forming a part-by-time representation that is temporally aggregated per part using MCM [3]. **MMGaitFormer** uses a coarse 3-region split along the height (top quarter / middle half / bottom quarter) to build predefined attention masks that restrict cross-attention to corresponding silhouette regions and skeleton joint groups during fusion [5]. Compared to these works, we keep the stripe-based bias (part-specific descriptors), while performing temporal modeling with a shared Transformer encoder per part in a silhouette-only and compact design.

Method	Venue	Rank-1 (CL)
GaitNet	CVPR '19	62.3%
GaitSet	AAAI '19	70.4%
GaitBase	CVPR '23	77.4%
GLN	ECCV '20	77.5%
GaitPart	CVPR '20	78.7%
SRN+CB	TBBIS '21	81.8%
Ours (16 Parts, RPE)		**82.0%**
3DLocal	ICCV '21	83.7%
CSTL	ICCV '21	84.2%
LangGait	CVPR '22	85.1%
MetaGait	ECCV '22	86.9%
GaitGL	ArXiv '21	87.3%
GaitRef	IJCB '23	88.0%
DANet	CVPR '23	89.9%
MSAFF	IJCB '23	93.3%
MSGR	TMM '23	94.0%
MMGaitFormer	CVPR '23	94.8%
GaitW		94.9%

Figure 4: Rank-1 CL condition comparison against SOTA. Highlighted methods use attention mechanisms. This trend suggests that attention-based designs are becoming increasingly common in gait recognition and may offer stronger robustness under challenging conditions such as clothing changes

6 Conclusion and Future Work

We presented a lightweight silhouette-based gait recognition pipeline that factorizes spatial and temporal modeling. The model combines a small CNN backbone, horizontal part pooling with GeM aggregation, and a shared Transformer encoder for temporal dynamics. While we achieved competitive results on CASIA-B under the LT-74 protocol, our evaluation is currently limited to this single benchmark.

A key limitation is that we could not run large-scale cross-dataset experiments on *OU-MVLP*, which is commonly reported by recent SOTA methods, due to constrained hardware (RTX 4060 GPU and 16 GB RAM) and the substantially larger scale of OU-MVLP (10,307 subjects). Future work includes cross-dataset evaluation, longer temporal windows or denser chunking at test time, and targeted strategies to improve robustness under clothing occlusions.

7 Contribution

- **Gloria Fiammengo (2254256)** was responsible for the implementation of the diagnostic study conducted on the GaitGL baseline. Furthermore, she developed the data handling pipeline, including the identity-balanced Triplet Sampler, and implemented the evaluation protocol used to benchmark

the model.

- **Federico Iannini (1931748)** focused on the design and implementation of the proposed model architecture, also managed the training workflow and the optimization pipeline.

References

[1] S. Yu, D. Tan, and T. Tan, "A framework for evaluating the effect of view angle, clothing and carrying condition on gait recognition," in *Proceedings of the 18th International Conference on Pattern Recognition (ICPR)*, pp. 441–444, 2006.

[2] H. Chao, K. Wang, Y. He, J. Zhang, and J. Feng, "Gaitset: Regarding gait as a set for cross-view gait recognition," *CoRR*, vol. abs/1811.06186, 2018.

[3] C. Fan, Y. Peng, C. Cao, S. Hou, J. Chi, Y. Huang, Y. Fu, and Q. Li, "Gaitpart: Temporal part-based model for gait recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[4] C. Fan, S. Hou, Y. Yu, and Y. Huang, "Gaitgl: Learning discriminative gait representations via global-local feature learning," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021.

[5] J. Cui, H. Wang, X. Zhang, J. Li, W. Wang, M. Wang, Y. Chen, and C. Shen, "Multi-modal gait recognition via effective spatial-temporal feature fusion," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10024–10033, June 2023.

[6] D. Thapar, A. Nigam, Q. Guan, and M. Shah, "Gaitw: Enhancing gait recognition in the wild using dynamic information," in *Proceedings of the Asian Conference on Computer Vision (ACCV)*, pp. 268–285, December 2024.

[7] L. Wang, Y. Xiong, Z. Wang, Y. Qiao, D. Lin, X. Tang, and L. Van Gool, "Temporal segment networks: Towards good practices for deep action recognition," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2016.

[8] B. Lin, S. Zhang, and F. Bao, "Gait recognition via effective global-local feature representation and local temporal aggregation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 14648–14656, October 2021.

[9] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," *CoRR*, vol. abs/1706.03762, 2017.

[10] H. Luo, Y. Gu, X. Liao, S. Lai, and W. Jiang, "Bag of tricks and a strong baseline for deep person re-identification," *CoRR*, vol. abs/1903.07071, 2019.