Video-based Gait Analysis with Temporal Transformers

Data Source & Protocol

Dataset origin

CASIA-B Gait Dataset

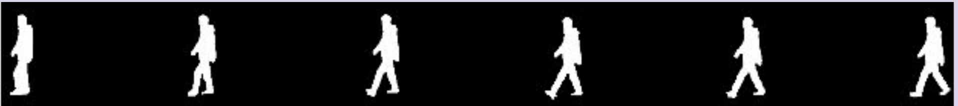
provided by: *Institute of Automation, Chinese Academy of Sciences (CASIA)*
public benchmark for gait recognition and re-identification

Data content

1. **124 subjects** (ID 001–124)
2. **silhouette image sequences** (frame-based, grayscale)
3. **3 walking conditions**:
    a. *nm* — normal walking
    b. *bg* — walking with bag
    c. *cl* — walking with coat
4. **11 camera views**: 0° to 180°

Data format used in this project

- pre-extracted silhouette frames (no raw videos)
- directory structure:
  `ID / condition / view / frame.png`
- each sequence treated as a **temporal signal of silhouettes**



Train / Test Split (Standard Protocol)

- **training subjects**: ID **001–074** (74 subjects)
- **test subjects**: ID **075–124** (50 subjects)
- **no validation set**
    - fixed hyperparameters
    - no early stopping

Dataset Size (Order of Magnitude)

- **~110 sequences per subject** (10 sequences × 11 views)
- **training set**: **≈ 8,000 gait sequences**
  (randomly sampled into 30-frame clips during training)
- **test set**: **≈ 5,500 gait sequences**
  (evaluated with center-crop or full-sequence protocol)

Evaluation Protocol

cross-view gait identification (retrieval task)
gallery: `nm-01` to `nm-04`
probe: `nm-05/06`, `bg-01/02`, `cl-01/02`
matching performed **excluding same-view pairs**

Why GaitGL? A Structural Question on Temporal Modeling

1. **Strong CNN-based SOTA with minimal architectural noise**
2. Temporal information is **present but implicitly handled** via **spatio-temporal convolutions**

Key Structural Bias in Temporal Aggregation

Temporal aggregation in GaitGL:

$$F_{gait}(x,y,c)=\max_{t} F(t,x,y,c)$$

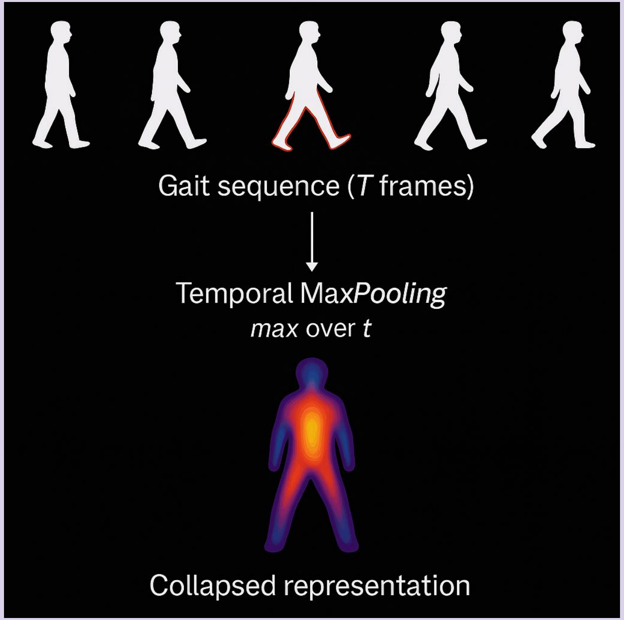
Time is collapsed via temporal max pooling.

Consequence

Order-invariant aggregation induces a strong structural bias toward static extrema.

Implication

sequences with identical per-frame extrema → **same final representation**
temporal ordering and phase evolution are **not required to achieve high accuracy**
long-range gait dynamics are **discarded**, not modeled



Pivot: when “temporalizing GaitGL” wasn’t enough

What we tried

add temporal modeling on top of GaitGL’s strong spatial backbone (ABS / RPE / conv head)

What we observed

performance stayed ~flat on NM/BG and inconsistent on CL.

1. RPE: NM 80.08 / BG 72.72 / CL 46.26
2. ABS: NM 80.51 / BG 71.75 / CL 43.51
3. Conv: NM 80.14 / BG 73.97 / CL 48.83

Interpretation: temporal modeling injected **too late** after a pipeline that already compresses/filters temporal cues

The insight

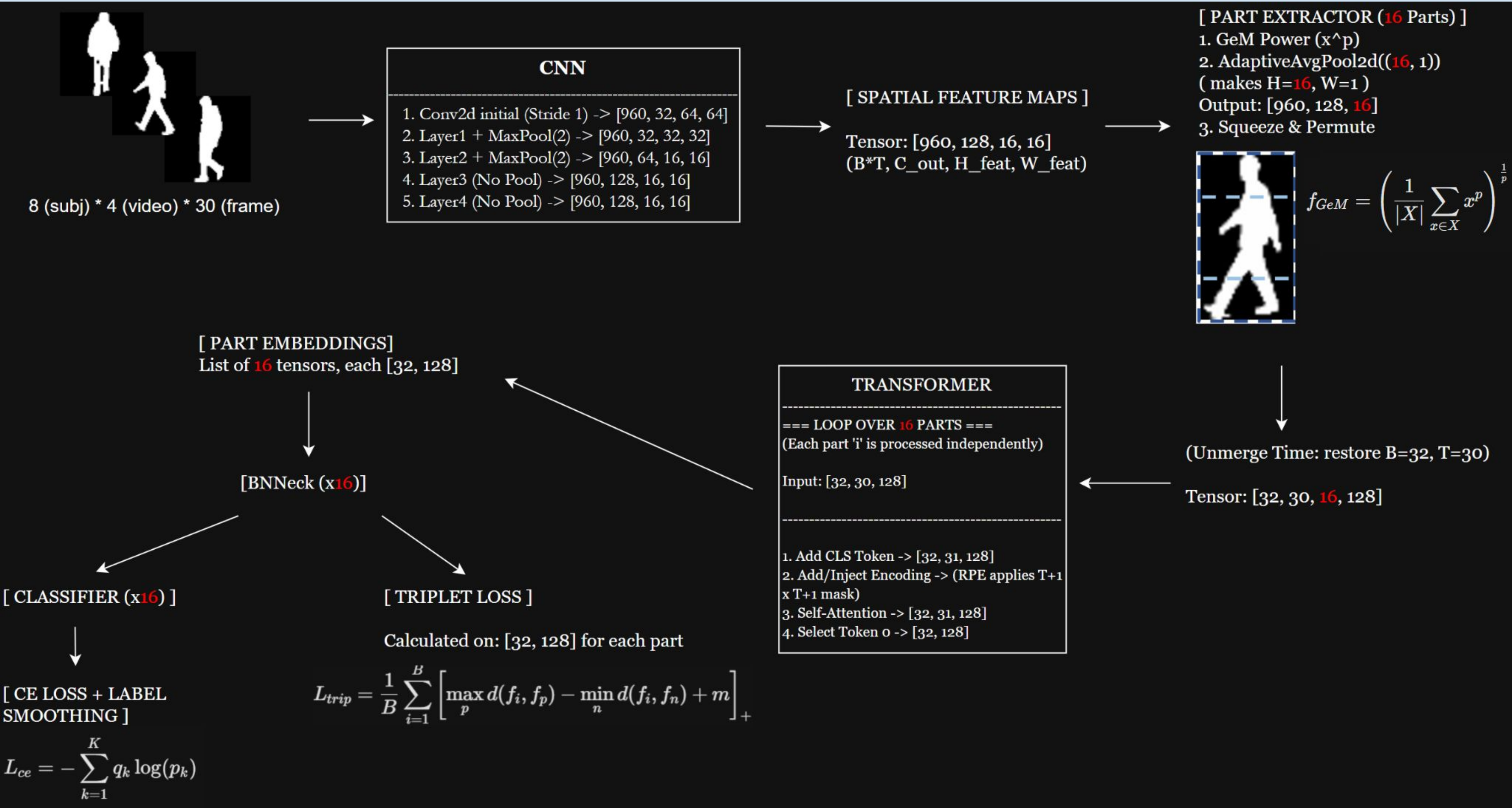
temporal modeling must be **structural**, not a bolt-on

The pivot

- move from “GaitGL + patch”
- to **explicit space–time factorization** *(frame-wise CNN + shared temporal Transformer)*

Goal: clean, lightweight, interpretable temporal modeling

Our Model



8 (subj) * 4 (video) * 30 (frame)
CNN
1. Conv2d initial (Stride 1) -> [960, 32, 64, 64]
2. Layer1 + MaxPool(2) -> [960, 32, 32, 32]
3. Layer2 + MaxPool(2) -> [960, 64, 16, 16]
4. Layer3 (No Pool) -> [960, 128, 16, 16]
5. Layer4 (No Pool) -> [960, 128, 16, 16]
[SPATIAL FEATURE MAPS]
Tensor: [960, 128, 16, 16]
(B*T, C_out, H_feat, W_feat)
[PART EXTRACTOR (16 Parts)]
1. GeM Power (x^p)
2. AdaptiveAvgPool2d((16, 1))
(makes H=16, W=1)
Output: [960, 128, 16]
3. Squeeze & Permute
$f_{GeM} = \left(\frac{1}{|X|} \sum_{x \in X} x^p \right)^{\frac{1}{p}}$
(Unmerge Time: restore B=32, T=30)
Tensor: [32, 30, 16, 128]
TRANSFORMER
=== LOOP OVER 16 PARTS ===
(Each part 'i' is processed independently)
Input: [32, 30, 128]
1. Add CLS Token -> [32, 31, 128]
2. Add/Inject Encoding -> (RPE applies T+1 x T+1 mask)
3. Self-Attention -> [32, 31, 128]
4. Select Token 0 -> [32, 128]
[PART EMBEDDINGS]
List of 16 tensors, each [32, 128]
[BNNeck (x16)]
[CLASSIFIER (x16)]
[CE LOSS + LABEL SMOOTHING]
$L_{ce} = -\sum_{k=1}^{K} q_k \log(p_k)$
[TRIPLET LOSS]
Calculated on: [32, 128] for each part
$L_{trip} = \frac{1}{B} \sum_{i=1}^{B} \left[\max_p d(f_i, f_p) - \min_n d(f_i, f_n) + m \right]_+$

Training Strategy & Implementation Details

- **Data Protocol & Input**
  - **Dataset:** CASIA-B Large-sample Training (LT) split: **74 subjects** for training, **50** for testing.
  - **Augmentation:** We explicitly enabled **Horizontal Flipping** and **Random Erasing** to increase model robustness and prevent overfitting on specific body parts or walking directions.

- **Balanced Sampling Strategy**
  - To ensure effective Metric Learning, we used a **Triplet Sampler**.
  - **Batch Structure P x K:** We selected **P=8** different subjects and **K=4** video clips per subject.
  - **Total Batch Size:** 32 samples per step. This structure guarantees valid positive and negative pairs for the loss function.

- **Hybrid Loss Function**
  - We optimized the model using a combined objective: **L = Ltrp+ Lce**.
  - **Batch Hard Triplet Loss:** Minimizes intra-class distance and maximizes inter-class distance (Margin = 0.2).
  - **Cross Entropy Loss:** Used with **Label Smoothing (0.1)** and a **BNNeck** (Batch Normalization Neck) to stabilize convergence.

- **Optimization Details**
  - **Optimizer:** AdamW (Weight decay: 1e^-4).
  - **Schedule:** 60 Epochs total. We used a scheduler, reducing the Learning Rate by a factor of 10 at epochs **30** and **50**.
  - **Efficiency:** Training performed using **Automatic Mixed Precision (FP16)** to reduce memory usage.

- **Experiments**
  - Different horizontal part division (1,2,4,8,16,”corpse”)
  - 5 different attention mechanisms (Sinusoidal, Cycle,LPE, CPE, RPE) to identify the optimal configuration for capturing gait dynamics.

Results & Considerations

Dividing matters

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%***

Changing temporal positional encoding

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%

Results & Considerations

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%

That's all! Thanks for the attention

GaitSet
https://arxiv.org/pdf/1811.06186

GaitPart
https://openaccess.thecvf.com/content_CVPR_2020/papers/Fan_GaitPart_Temporal_Part-Based_Model_for_Gait_Recognition_CVPR_2020_paper.pdf

GaitGL
https://arxiv.org/pdf/2208.01380

MMGaitFormer
https://openaccess.thecvf.com/content/CVPR2023/papers/Cui_Multi-Modal_Gait_Recognition_via_Effective_Spatial-Temporal_Feature_Fusion_CVPR_2023_paper.pdf

GaitW
https://openaccess.thecvf.com/content/ACCV2024/papers/Thapar_GaitW_Enhancing_Gait_Recognition_in_the_Wild_using_Dynamic_Information_ACCV_2024_paper.pdf