IEEE International Joint Conference on Biometrics

GaitRef: Gait Recognition with Refined Sequential Skeletons

Haidong Zhu*, Wanrong Zheng*, Zhaoheng Zheng and Ram Nevatia
University of Southern California

IEEE International Joint Conference on Biometrics (IJCB) 2023

Task Definition

- Gait Recognition:
  - Identify individuals based on their walking patterns
  - Gait can be observed at distance and does not require active cooperation of the subject

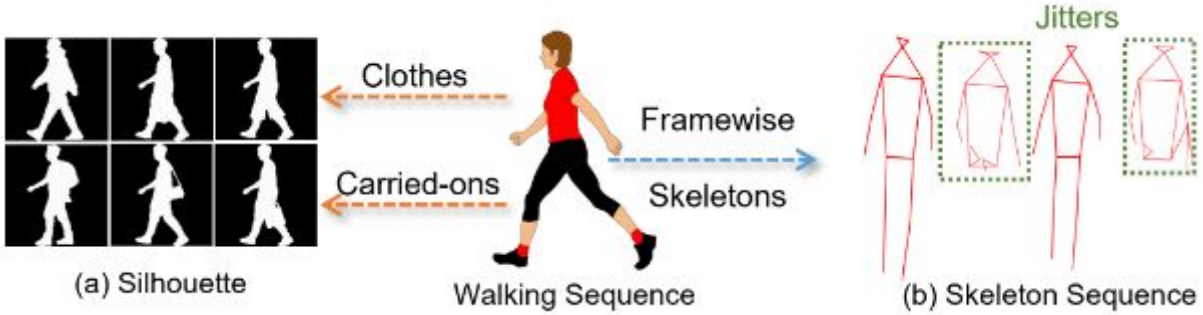
OpenGait

Background

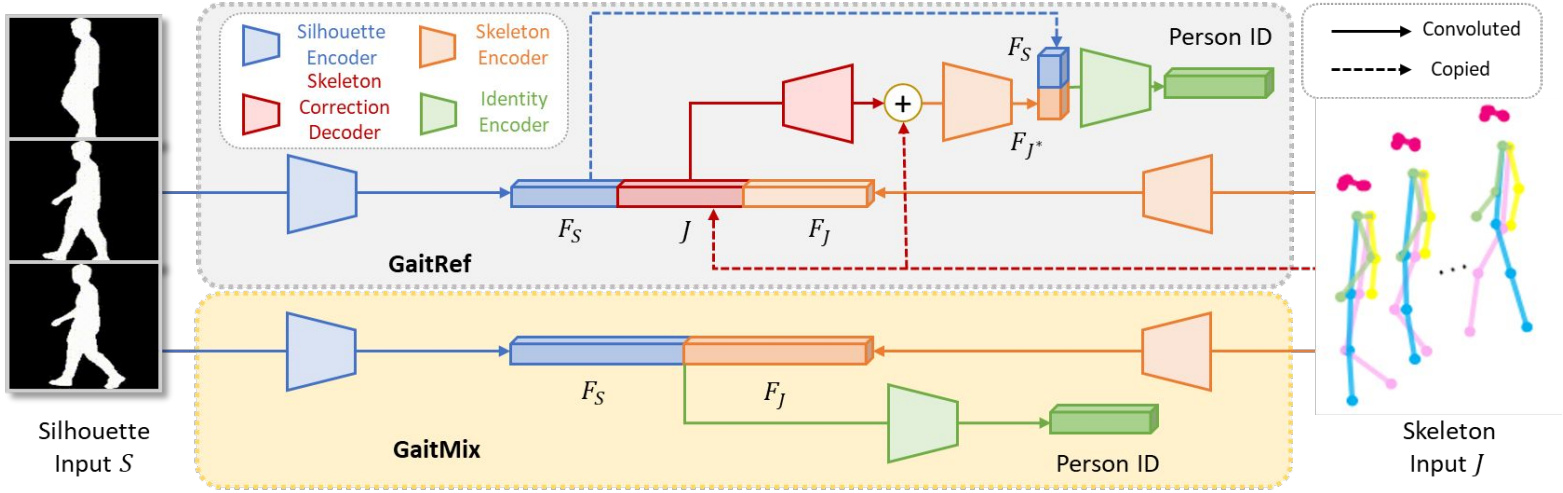
- Person ID needs to work with variations in pose, clothing, illumination and other variations
- Silhouettes and skeletons have been used in many recent methods
  - Silhouettes: Sensitive to clothing and carried objects
  - Skeletons: Rich information but have jitter with noisy frames
- RGB appearance can be also provide helpful features but many public datasets do not release RGB videos due to privacy issues
  - We do not use RGB videos for this work

Motivation and Approach

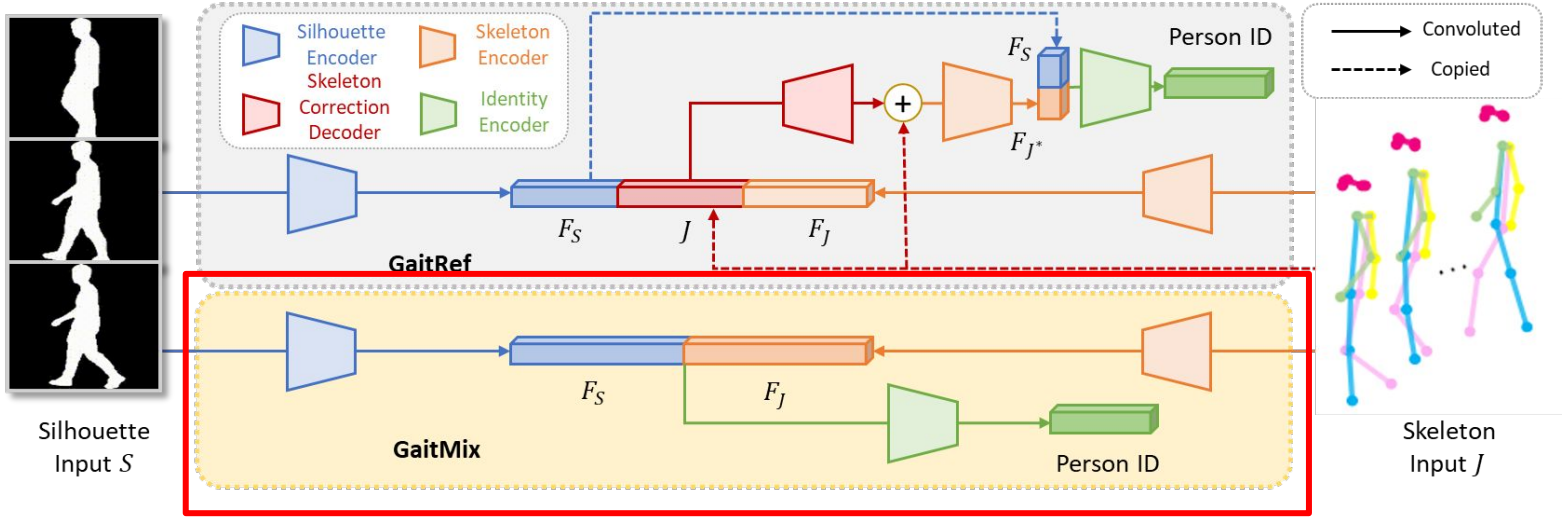
- Skeletons and silhouette describe the same sequence
- Combining the two modalities should help overcome some limitations of each (GaitMix in our design)
- We can use silhouettes to correct skeletons as silhouettes seem to have less temporal noise (GaitRef in our design)



Methods Overview

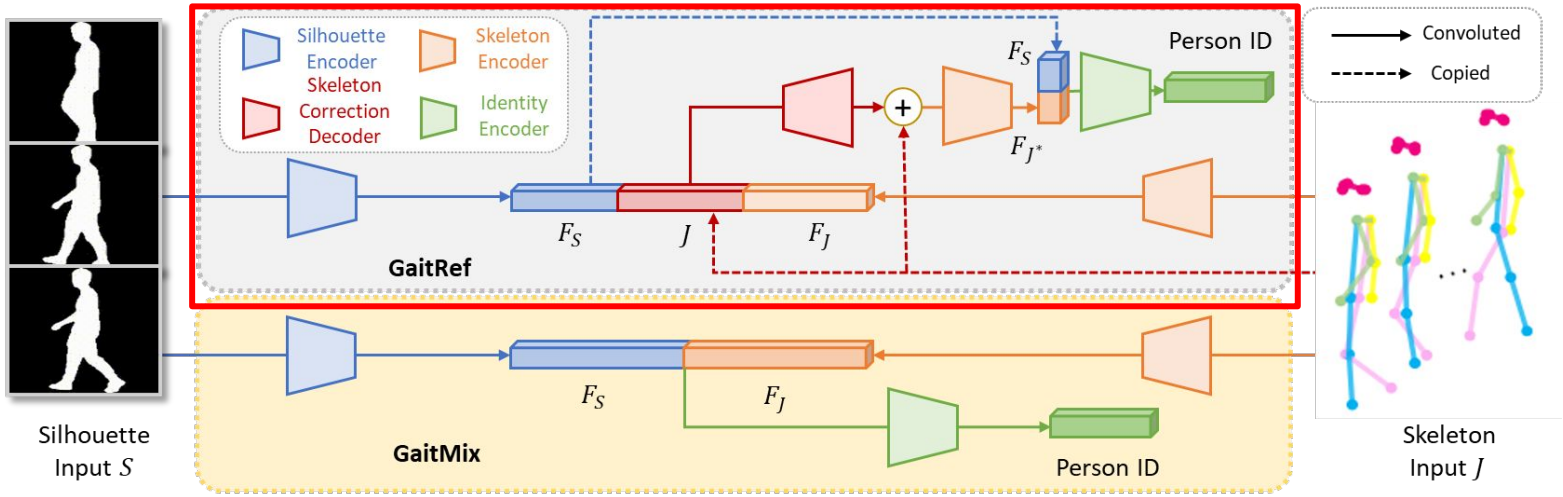


Methods Overview



- GaitMix
  - Fuse two modalities as a baseline approach

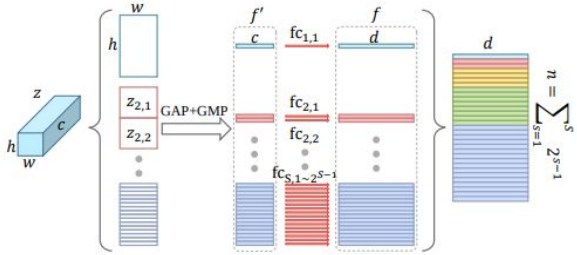
Methods Overview



- GaitRef
  - Refine the skeletons with temporal consistency from silhouettes

Methods

- GaitMix
  - Combine silhouette and skeleton features for identification
  - Three encoders for extracting and aggregating features
    - Skeleton encoder: ST-GCN [1]
    - Silhouette encoder: GaitGL[2] and SMPLGait[3] (for Gait3D only)
      - Apply SOTA methods as encoders fo
    - Identity encoder: Part-based FC[4] layers



[1] Yan, Sijie, *et al*. "Spatial temporal graph convolutional networks for skeleton-based action recognition." *AAAI* 2018.
[2] Lin, Beibei, *et al*. "Gait recognition via effective global-local feature representation and local temporal aggregation." *ICCV* 2021.
[3] Zheng, Jinkai, *et al*. "Gait recognition in the wild with dense 3d representations and a benchmark." *CVPR* 2022.
[4] Chao, Hanqing, *et al*. "Gaitset: Regarding gait as a set for cross-view gait recognition." *AAAI* 2019.

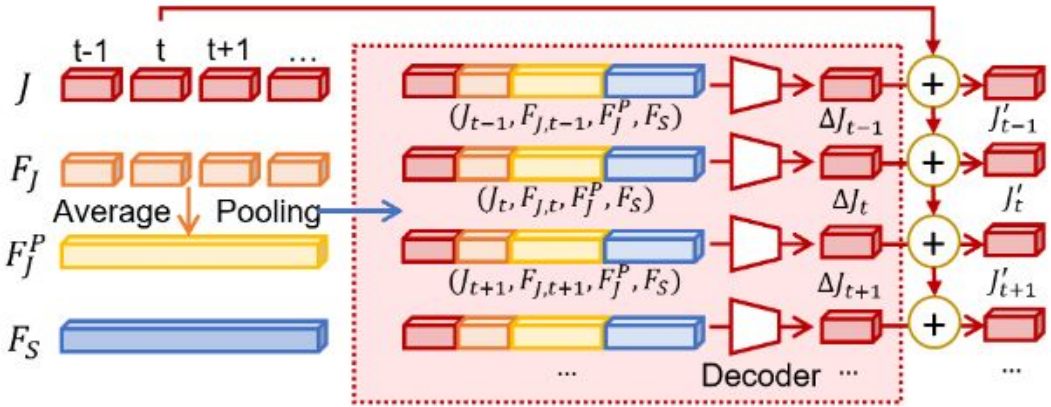
Methods

- GaitRef
  - Refine skeleton representation with temporal consistency from silhouettes, and fuse the refined skeleton for identification
  - An additional “decoder” added to GaitMix
    - Skeleton encoder: ST-GCN [1]
    - Silhouette encoder: SMPLGait (for Gait3D only) or GaitGL
    - Identity encoder: Part-based FC layers
    - Skeleton correction decoder: Reversed ST-GCN
      - Reversed stgcn is of the same structure st-gcn but with channel size reducing

[1] Yan, Sijie, *et al*. "Spatial temporal graph convolutional networks for skeleton-based action recognition." *AAAI* 2018.

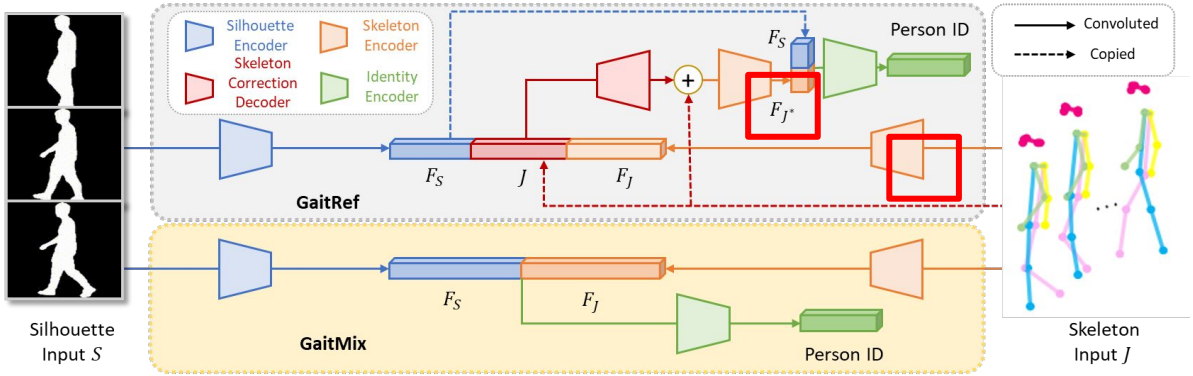
Methods

- Skeleton correction network
  - Refine skeletons with temporal information from silhouettes
  - Inputs: Per-frame input: skeletons and encoded skeleton features
    Video-level input: silhouette and averaged skeleton features
  -



Methods

- Skeleton correction network
  - Output: Per-frame joint correction, will be added on original skeleton
  - Alignment between input and output skeletons
    - Re-use the same skeleton encoder to force them in the same distribution as the encoder input



Experiment Datasets (1 and 2)

- CASIA-B:
  - 124 subjects, 74 for training and 50 for inference
  - 3 different cases - 6 segments for normal walking (NM), 2 for bag carrying (BG) and 2 for different clothing (CL)
  - Use first 4 NM videos as galleries and remaining as probes
- OUMVLP:
  - 10307 subjects with 28 videos for each subject
  - 5153 subjects for training and 5154 for inference
  - All sequences are NM

Experiment Datasets (3 and 4)

- Gait3D*
  - 4,000 identities with 25,309 sequences
  - 3,000 identities for training and 1,000 for inference
- GREW*
  - 26,345 identities with 128,671 sequences.
  - 20,000 identities for training, 345 identities for validation, and the remaining 6,000 for inference

* Dataset collected in the wild.

Experiments and Results

- Rank-1 accuracies on CASIA-B (left) and OU-MVLP (right)

Method	NM	BG	CL
GaitGL	97.3	94.4	83.5
CSTL	97.8	93.6	84.2
ModelGait	97.9	93.1	77.6
GaitMix	97.7	95.2	85.8
GaitRef	**98.1**	**95.9**	**88.0**

Method	Accuracy
GLN	89.2
GaitGL	89.6
CSTL	**90.2**
GaitMix	89.9
GaitRef	**90.2**

Experiments and Results

- Results on Gait3D (left) and GREW (right)

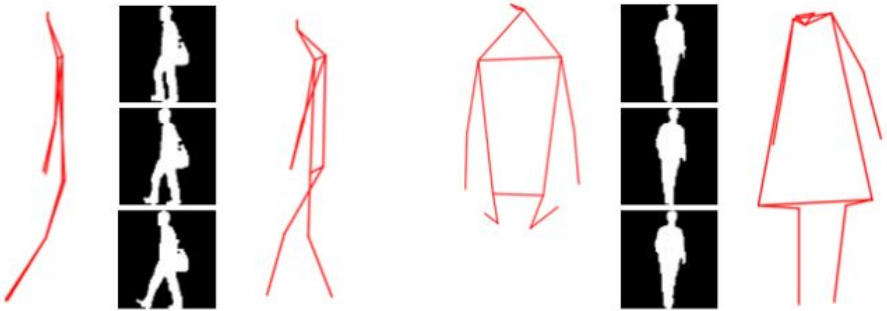
Method	Rank 1	Rank 5	mAP	mINP
GaitGL	29.70	48.50	22.29	13.26
OpenGait	42.90	63.90	35.19	20.83
CSTL	11.70	19.20	5.59	2.59
SMPLGait*	46.30	64.50	37.16	22.23
GaitMix	45.80	65.60	36.74	22.09
GaitRef	**49.00**	**69.30**	**40.69**	**25.26**

Method	Rank 1	Rank 5	Rank 10	Rank 20
GaitSet	46.3	63.6	70.3	76.8
GaitPart	44.0	60.7	67.4	73.5
CSTL	50.6	65.9	71.9	76.9
GaitGL	51.4	67.5	72.8	77.3
GaitMix	52.4	67.4	72.9	77.2
GaitRef	**53.0**	**67.9**	**73.0**	**77.5**

* SMPLGait has 3-D body shape as input.

Experiments and Results

- Visualization of corrected skeletons compared with its input frames
  - Two examples with the order of input skeleton - nearby silhouettes - corrected skeletons
  - Even the corrected skeletons are not perfect, their prediction is still correct while the original skeleton leads to wrong prediction



Experiments and Results

- Ablation results on different encoder combination
  - Other skeleton networks, such as MS-G3D, can show further improvements, but they come with heavier time consumption.
  - The refinement network works better than smoothing networks.

Methods	Encoder	Decoder	NM	CL	BG
GaitMix	ST-GCN	N/A	97.7	95.2	85.8
GaitMix	MS-G3D	N/A	98.0	95.5	86.4
GaitRef	ST-GCN	ST-GCN	**98.1**	**95.9**	88.0
GaitRef	ST-GCN	MS-G3D	**98.1**	95.7	**88.5**
GaitRef	MS-G3D	ST-GCN	**98.1**	**95.9**	88.3
GaitMix	Average Smoothing		97.6	95.0	85.6
GaitMix	Gaussian Smoothing		97.7	95.2	85.9
GaitMix	SmoothNet [39]		97.4	94.4	83.8

Conclusions and Future Directions

- Demonstrated the effectiveness of combining silhouettes and skeletons
- Demonstrated use of silhouette sequence to improve skeletons
- Future Directions
  - Combine with other modalities (such as 3D shape)
  - Use of appearance features as in ReID methods
  - Above require access to RGB videos which are available for real applications but not in many public datasets
  - Face ID datasets do provide images, gait may need to do the same

Thank you!