ShARc: Shape and Appearance Recognition for Person Identification In-the-wild

Haidong Zhu, Wanrong Zheng, Zhaoheng Zheng and Ram Nevatia

University of Southern California

JAN 4-8 WACV 2024 WAIKOLOA HAWAII

USC Viterbi School of Engineering, Department of Computer Science

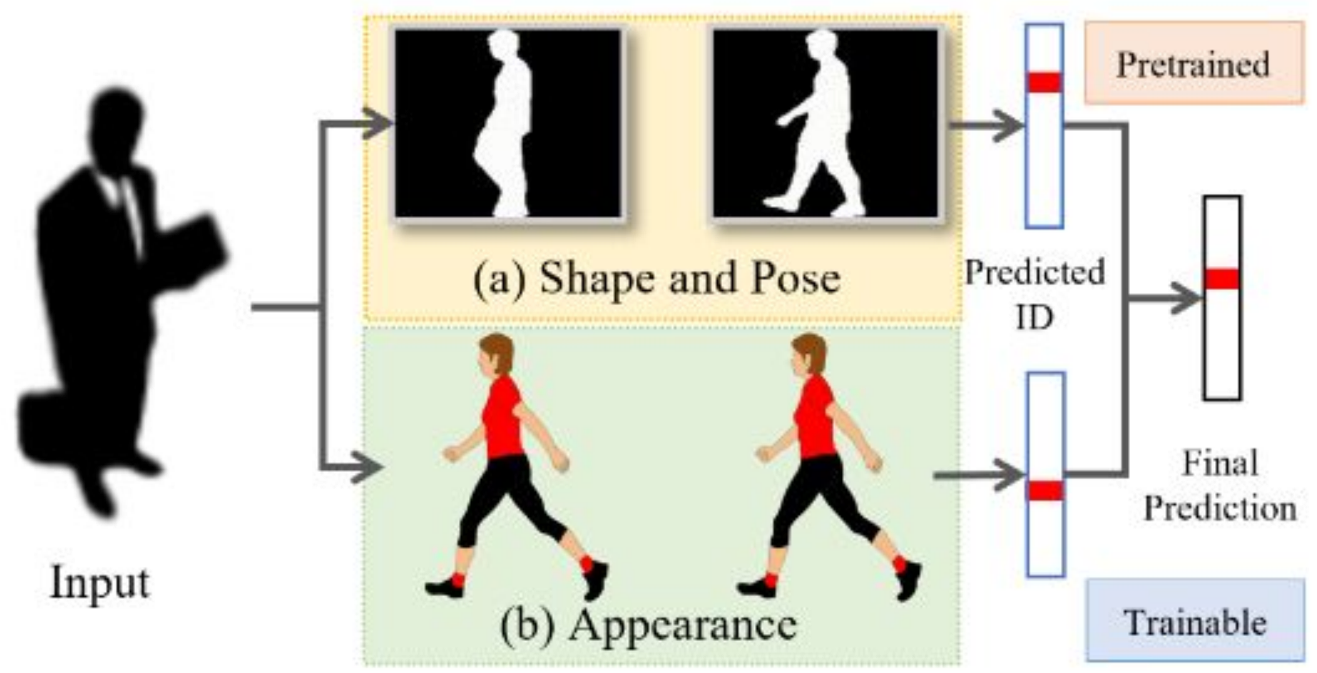
Paper

INTRODUCTION

Gallery Frame	Standing Videos	Different Clothing	Turbulence & Occlusion
Gait		✓	✗
Body shape	✗	✗	✓
Appearance	✓	✗	✗
Ours	✓	✓	✓

- Different modalities have their own pros and cons for recognizing the person's identity; some of them have the limitation of specific actions and conditions to work with.
- We combine and investigate the performance of different modalities for person identification using shape and appearance, named as ShARc, **Sh**ape and **A**ppearance **Rec**ognition.

METHOD



- We separate the pipeline to two different branches, one for shape and one for appearance, and process each modality separately.
  - Shape-based recognition with PSE (Pose and Shape Encoder).
  - Appearance-based recognition with AAE (Aggregated Appearance Encoder)
- We train two networks separately
  - For PSE, we build the loss following

$$\mathcal{L}_{shape} = 0.1\,\mathcal{L}_{triplet} + \mathcal{L}_{CE}$$

  - For AAE, we build the loss following

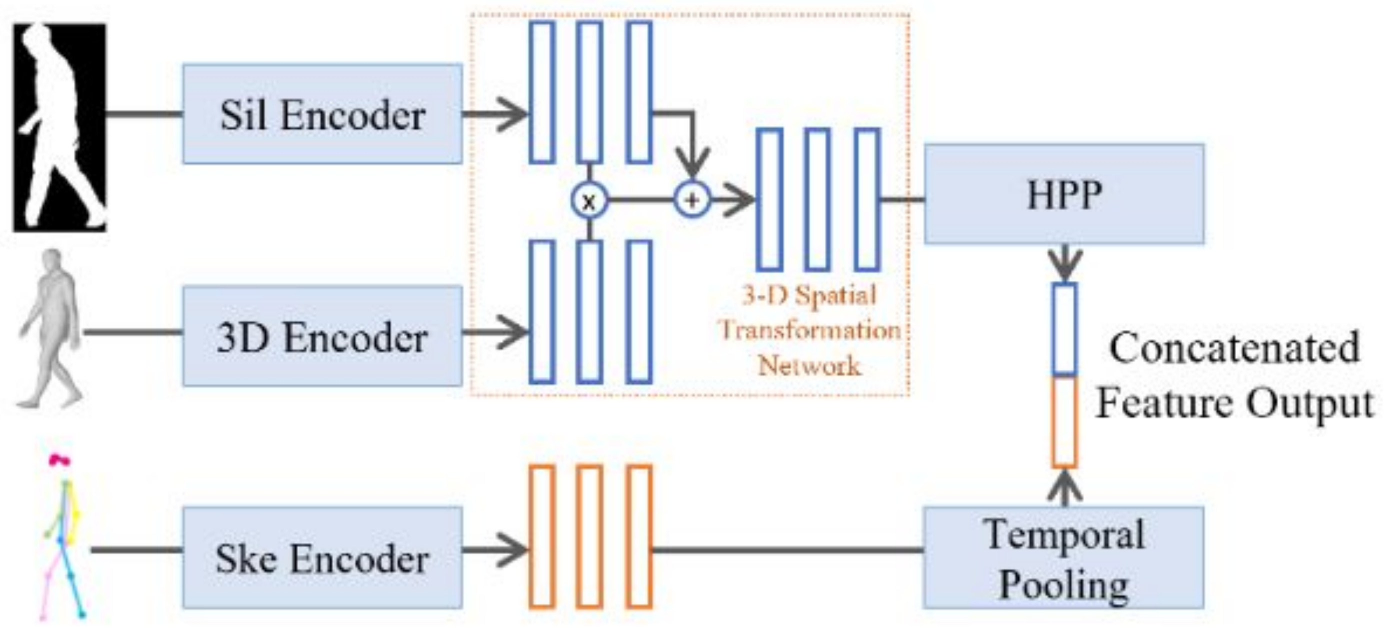
$$\mathcal{L}_{app} = \mathcal{L}_{triplet} + \mathcal{L}_{CE} + \mathcal{L}_{cen} + 5e^{-4}\,\mathcal{L}_{CTL}$$

- During inference, we add two cosine similarity scores using features generated by two branches as the final prediction

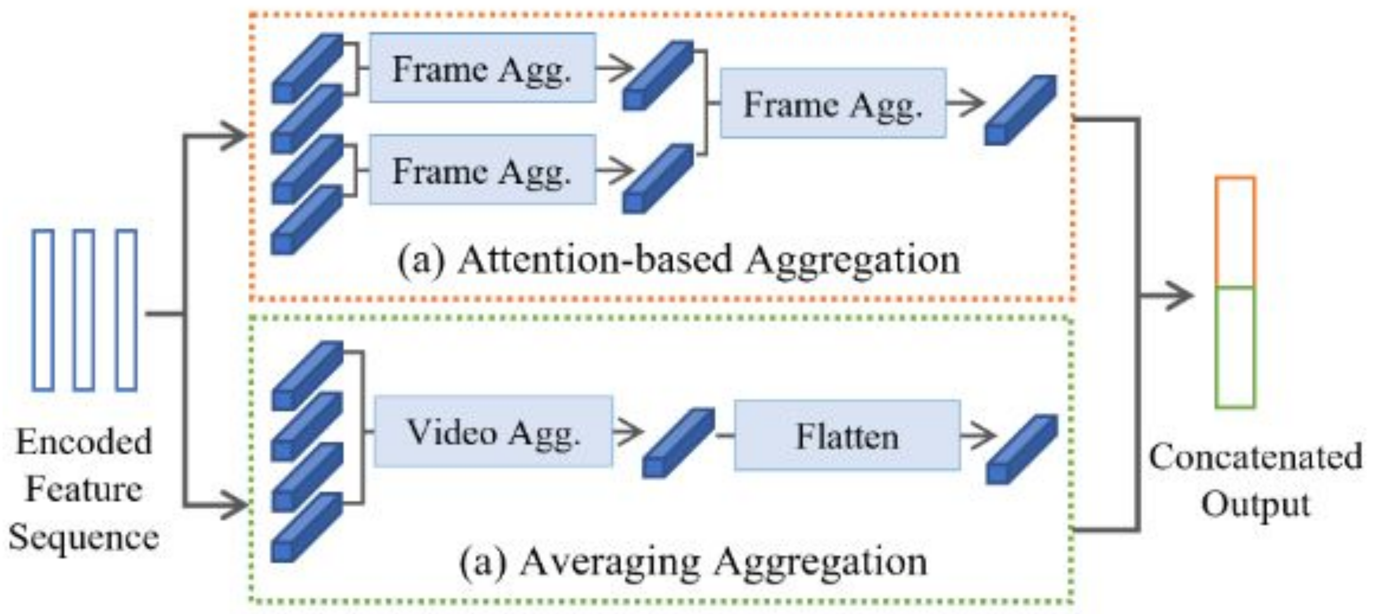
$$S(V) = \alpha S_{shape}(V) + (1-\alpha) S_{app}(V)$$

NETWORK DETAILS

- Extraction of shape-related patterns
  - Gait - DeepLab-v3 for silhouette extraction
  - 3-D body shape - ROMP for SMPL extraction
  - Skeleton - HRNet for skeleton extraction



- Shape and pose encoder (PSE)
  - Silhouette encoder for gait pattern extraction
  - 3-D body shape encoder for framewise body shape encoding
    - Two features are aggregated framewise with 3-D spatial transformation network



- Aggregated appearance encoder (AAE)
  - Attention-based aggregation (AtA)
    - Aggregate 2 consecutive frames at a time
    - Pyramid-like aggregation till last layer
  - Averaging Aggregation (AvA)
    - Average features from all input frames
    - Append a flatten layer for averaged feature

$$A_{avg} = sgn(A_{avg}) \cdot ||A_{avg}||^{\gamma}$$

DATASETS

- We use three datasets for our evaluation which include clothes change cases
  - CCVID
    - Include same-clothes cases
    - 75 IDs for training, 151 for inference
  - MEVID
    - Include same-clothes cases
    - 104 IDs for training, 54 for inference
  - BRIAR
    - 407 IDs for training, 642 for inference

RESULTS

- CCVID

	General		Clothes Changes	
Method	Rank 1	mAP	Rank 1	mAP
GaitNet	62.6	56.5	57.7	49.0
GaitSet	81.9	79.2	71.0	62.1
CAL	82.6	81.3	81.7	79.6
ShARc	**89.8**	**90.2**	**84.7**	**85.2**

- MEVID

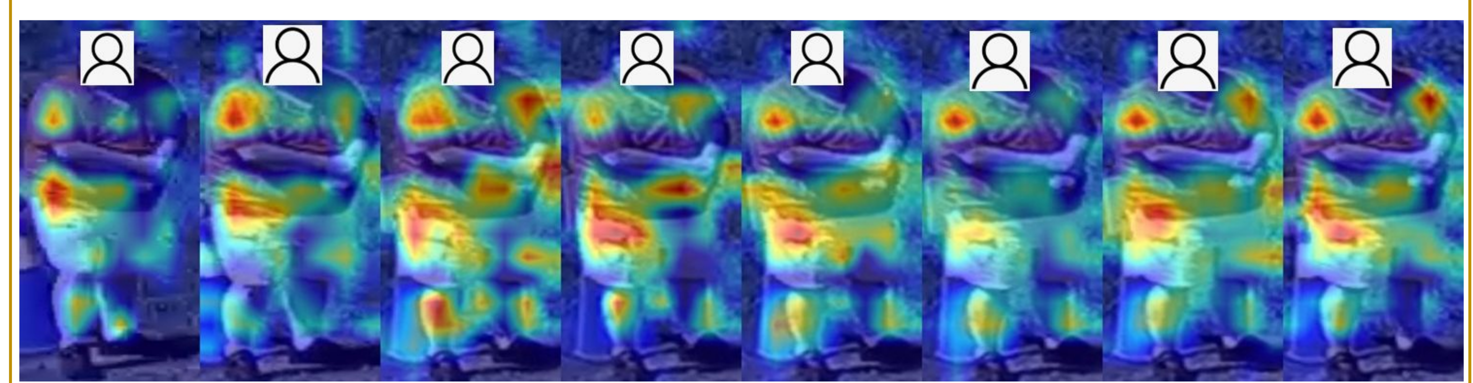
Method	Rank 1	Rank 5	Rank 10	Rank 20
PSTA	46.2	60.8	69.6	77.8
ARGL	48.4	62.7	70.6	77.9
Attn-CL	42.1	56.1	63.6	73.1
Attn-CL+RR	46.5	59.8	64.6	71.8
CAL	52.5	66.5	73.7	80.7
ShARc	**59.5**	**70.3**	**77.2**	**82.9**

- BRIAR

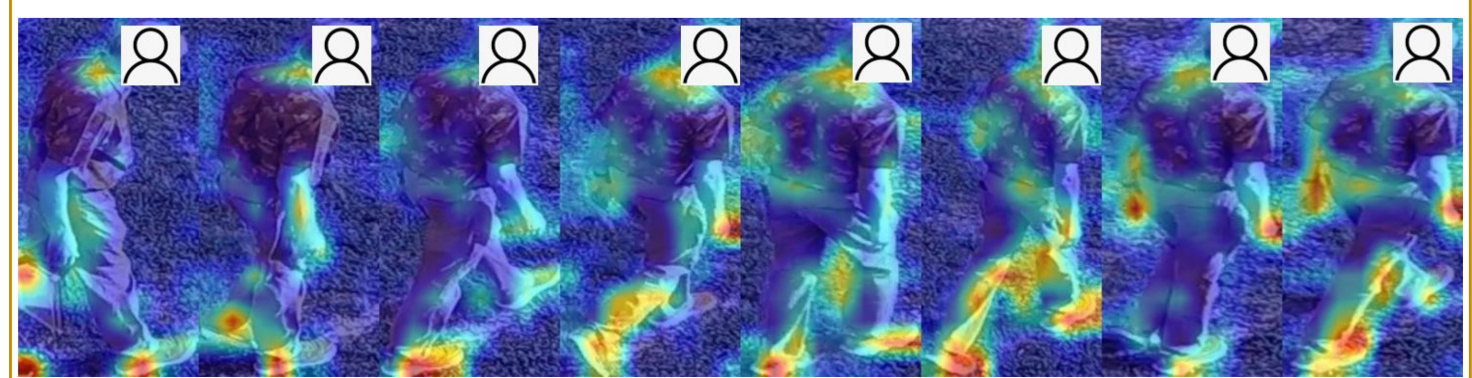
Method	Rank 1	Rank 20
GaitGL	15.6	45.6
GaitRef	17.7	50.2
PSTA	33.6	67.3
Attn-CL+RR	27.6	61.8
CAL	34.9	71.4
ShARc	**41.1**	**83.0**

ATTENTION MAP VISUALIZATION

- We include visualization map for sequences with different attention maps using GradCam.



  - Standing videos
    - Model focuses more on body shape and visible skins for making decisions.



  - Walking videos
    - Model focuses on the end of the legs and the visible skins on body for decision making.