Spatio-Temporal Representation Factorization for Video-based Person Re-Identification

Abhishek Aich, Meng Zheng, Srikrishna Karanam, Terrence Chen, Amit K. Roy-Chowdhury, Ziyan Wu

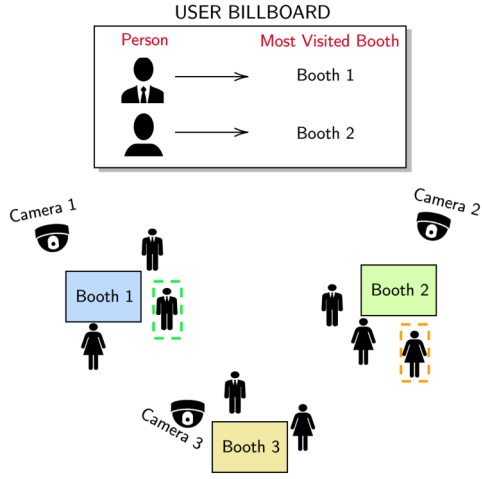
UC RIVERSIDE | UNITED IMAGING | United Imaging Intelligence

2021 ICCV OCTOBER 11-17 VIRTUAL

Introduction

Problem Scenario

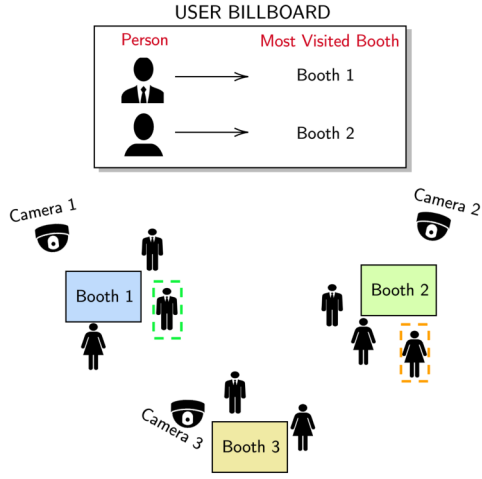
Example Scenario: Suppose we want to estimate which booth does a particular person visits the most in a business conference.



Problem Scenario

Example Scenario: Suppose we want to estimate which booth does a particular person visits the most in a business conference.

What tools do we need?

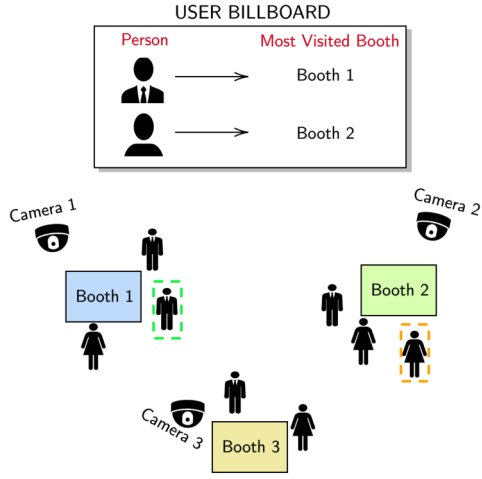


Problem Scenario

Example Scenario: Suppose we want to estimate which booth does a particular person visits the most in a business conference.

What tools do we need?

- Camera system
- Tracking system
- **Person Re-Identification system**

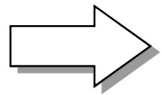


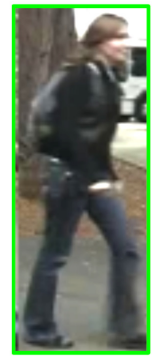
Our Objective:
Person Re-Identification

What is Person Re-Identification?

What is Person Re-Identification?

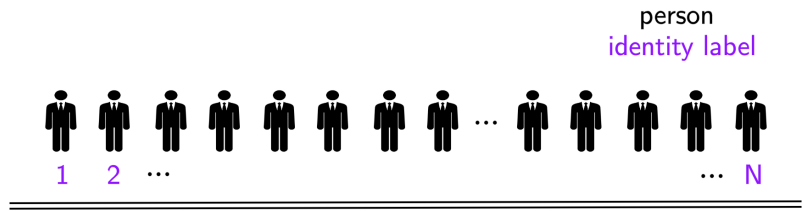
What is Person Re-Identification?



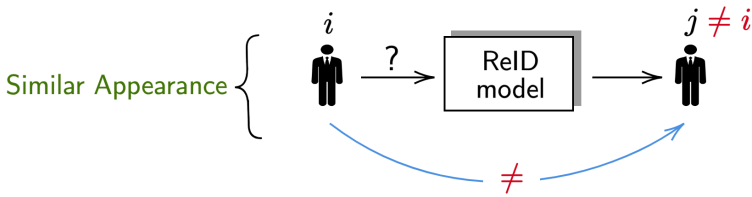


Challenges

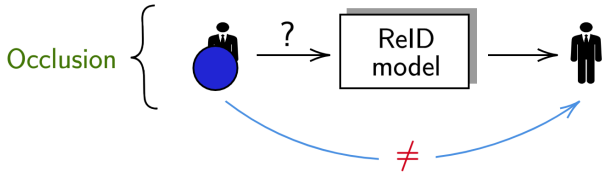
What Are Our Challenges?



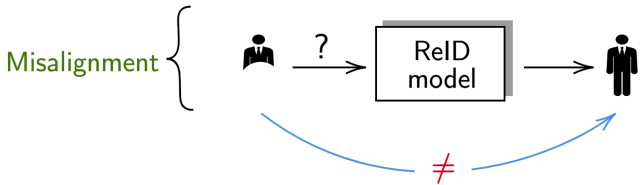
What Are Our Challenges?



What Are Our Challenges?

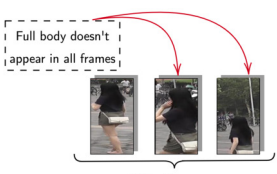


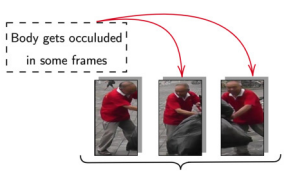
What Are Our Challenges?

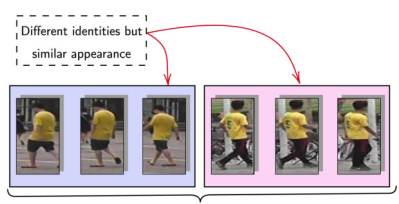


What Are Our Challenges?

Occlusion, misalignment, and similar appearance between different identities are inherent issue.



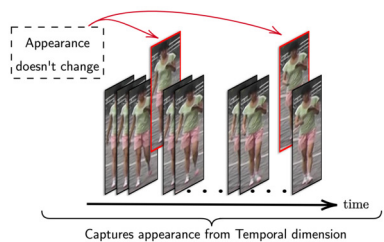




Proposed Formulation

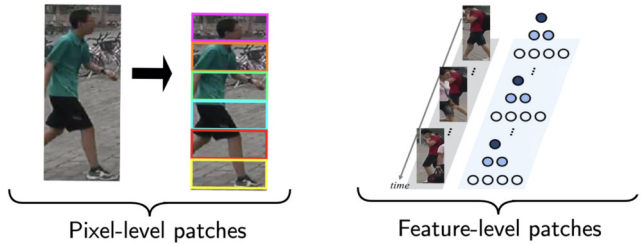
Design Motivation

M.1 Current methods do not explicitly address re-ID challenges in **both Temporal** and **Spatial** dimension.



Design Motivation

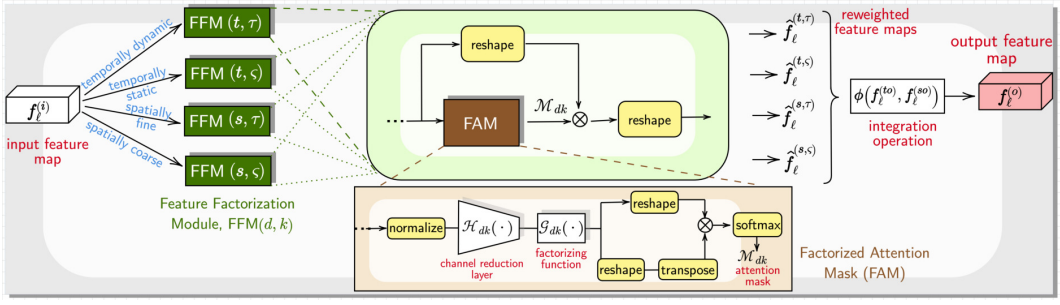
M.2 Multi-granularity[1]/patch division[2] helps in localizing individual specific features.



[1] Yichao Yan et al. "Learning Multi-Granular Hypergraphs for Video-based Person Re-Identification". *CVPR*. 2020.

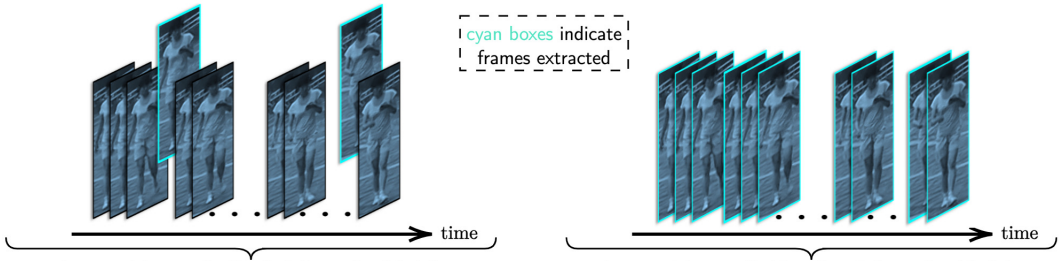
[2] Yifan Sun et al. "Beyond Part Models: Person Retrieval with Refined Part Pooling (and A Strong Convolutional Baseline)". *ECCV*. 2018.

Spatio-Temporal Representation Factorization (STRF)



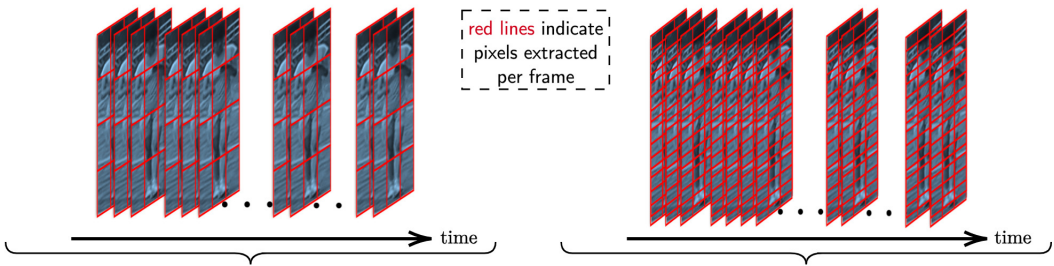
Concept of Temporal Branch

- To handle **similar appearance** and possibly **occlusion**.

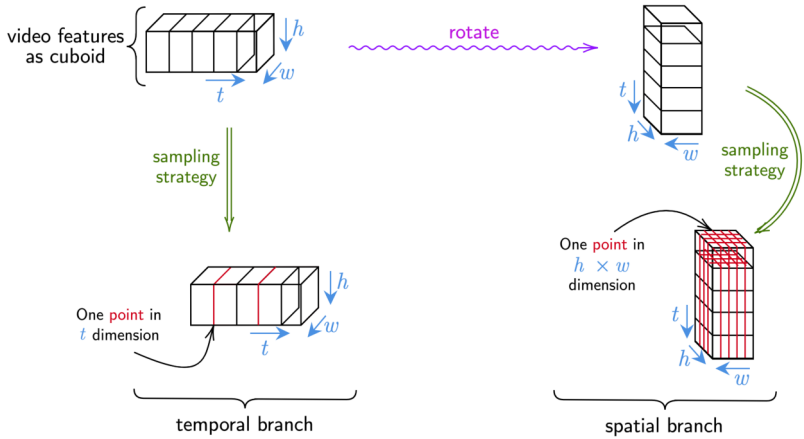


Concept of Spatial Branch

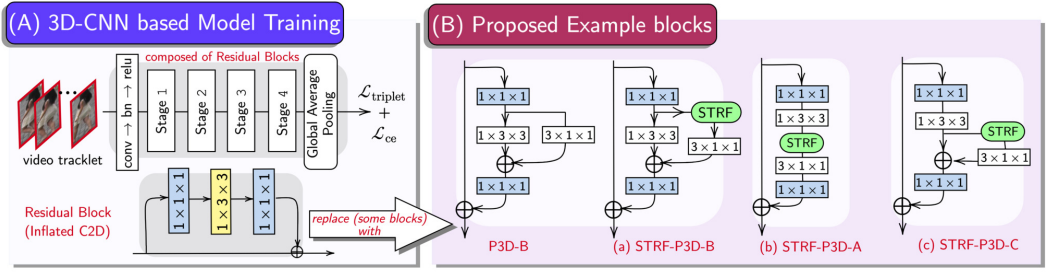
- Majority pixels of a given frame belong to the person.
- Sampling in **spatial** domain ($H \times W$) should thus alleviate **occlusions** and **misalignment** which only occur in few frame.



Overall Concept for Complete Model



How to use STRF units?



Results

Improvements over Baselines

Table 1: STRF consistently improves the performance of baseline models. **P**(M) is model size in millions.

MODEL	P(M)	DATASETS			
		MARS[3]		DukeMTMC[4]	
		mAP (%)	R@1 (%)	mAP (%)	R@1 (%)
I3D	28.92	82.70	88.50	95.20	95.40
+ **STRF**	28.97	**83.10**	**88.70**	**95.20**	**95.90**
P3DA	25.48	83.20	88.90	95.00	95.00
+ **STRF**	25.53	**85.40**	**89.80**	**95.60**	**96.00**
P3DB	25.48	83.00	88.80	95.40	95.30
+ **STRF**	25.53	**85.60**	**90.30**	**96.40**	**97.40**
P3DC	25.48	83.10	88.50	95.30	95.30
+ **STRF**	25.53	**86.10**	**90.30**	**96.20**	**97.20**

[3] Liang Zheng et al. "MARS: A Video Benchmark for Large-Scale Person Re-Identification". *ECCV*. 2016.

[4] Yu Wu et al. "Exploit the Unknown Gradually: One-Shot Video-based Person Re-Identification by Stepwise Learning". *CVPR*. 2018.

SOTA on Video-based Person Re-ID

Table 2: STRF gives state-of-the-art performance on all datasets (best results in **red**, second best in **blue**, and third best results in **green**.)

METHODS	VENUE	DATASETS				
		MARS[3]		DukeMTMC[4]		iLiDS-VID[5]
		mAP (%)	R@1 (%)	mAP (%)	R@1 (%)	R@1 (%)
MGH	CVPR 2020	**85.80**	90.00	–	–	85.60
STGCN	CVPR 2020	83.70	89.95	95.70	**97.29**	–
MG-RAFA	CVPR 2020	**85.90**	88.80	–	–	88.60
TACAN	WACV 2020	84.00	89.10	95.40	96.20	**88.90**
M3D	TPAMI 2020	79.46	88.63	93.67	95.49	86.67
AFA	ECCV 2020	82.90	**90.20**	95.40	**97.20**	88.50
AP3D	ECCV 2020	85.60	**90.70**	**96.10**	**97.20**	**88.70**
TCLNet	ECCV 2020	85.10	89.80	**96.20**	96.90	86.60
STRF	**Ours**	**86.10**	**90.30**	**96.40**	**97.40**	**89.30**

[3] Liang Zheng et al. "MARS: A Video Benchmark for Large-Scale Person Re-Identification". *ECCV*. 2016.

[4] Yu Wu et al. "Exploit the Unknown Gradually: One-Shot Video-based Person Re-Identification by Stepwise Learning". *CVPR*. 2018.

[5] Taiqing Wang et al. "Person Re-Identification by Video Ranking". *ECCV*. 2014.

Performance w.r.t. other 3D-CNN based works

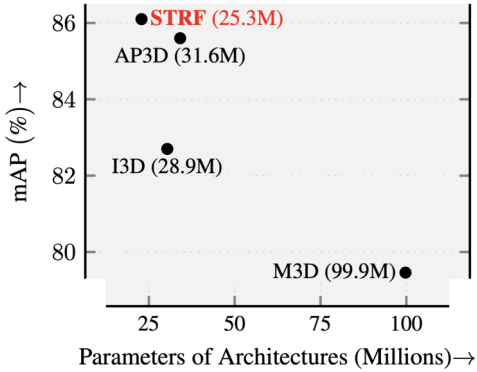
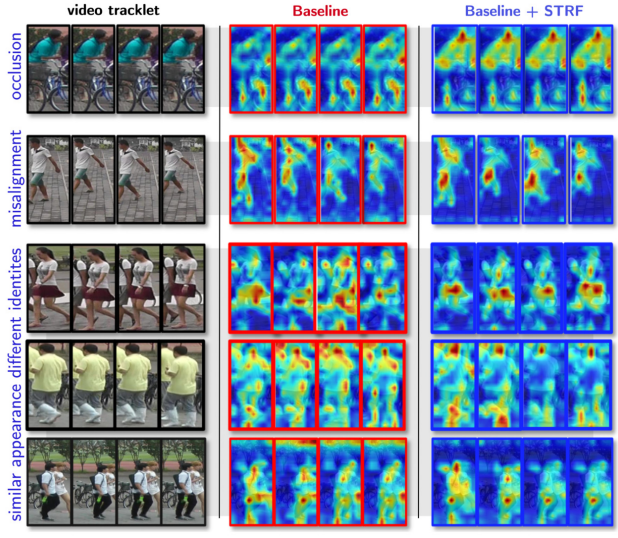


Figure 1: STRF gives state-of-the-art performance w.r.t. other 3D-CNN based methods with fewer model parameters.

Attention Map Visualization



Thank You!

- **Paper ID: 1629** → Spatio-Temporal Representation Factorization for Video-based Person Re-Identification
- **Paper Session:**
  - Session 1A → October 12, 10:00 AM – 11:00 AM (EDT)
  - Session 1B → October 14, 05:00 PM – 06:00 PM (EDT)
- **Paper available at**: `https://arxiv.org/pdf/2107.11878.pdf`