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

UC RIVERSIDE

UNITED IMAGING | United Imaging Intelligence

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

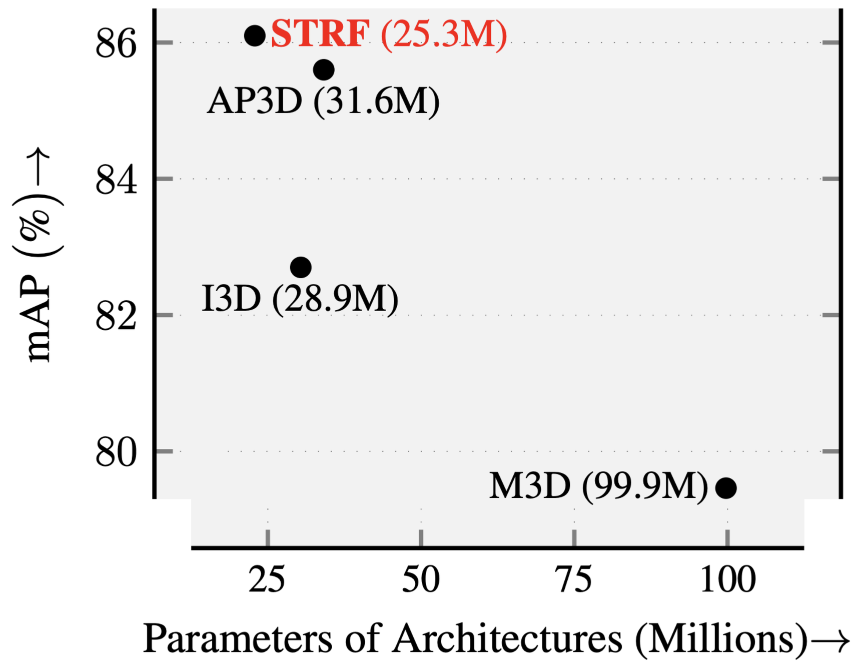
[†]United Imaging Intelligence, Cambridge MA, [‡]University of California, Riverside CA

2021 ICCV OCTOBER 11-17 VIRTUAL

Spatio-Temporal Representation Factorization (STRF)

- A flexible new computational unit that enhances most existing 3D convolutional neural network architectures for re-ID.
- Key Innovations:
  - ✓ Temporal factorization to learn static features (e.g., the color of clothes) that do not change much over time, and dynamic features (e.g., walking patterns) that change over time.
  - ✓ Spatial factorization to learn both global (coarse segments) and local (finer segments) appearance features, with the local features particularly useful in cases of occlusion or misalignment.
  - ✓ STRF shows new state-of-the-art results on three benchmarks.

STRF w.r.t. Related 3D-CNN Works

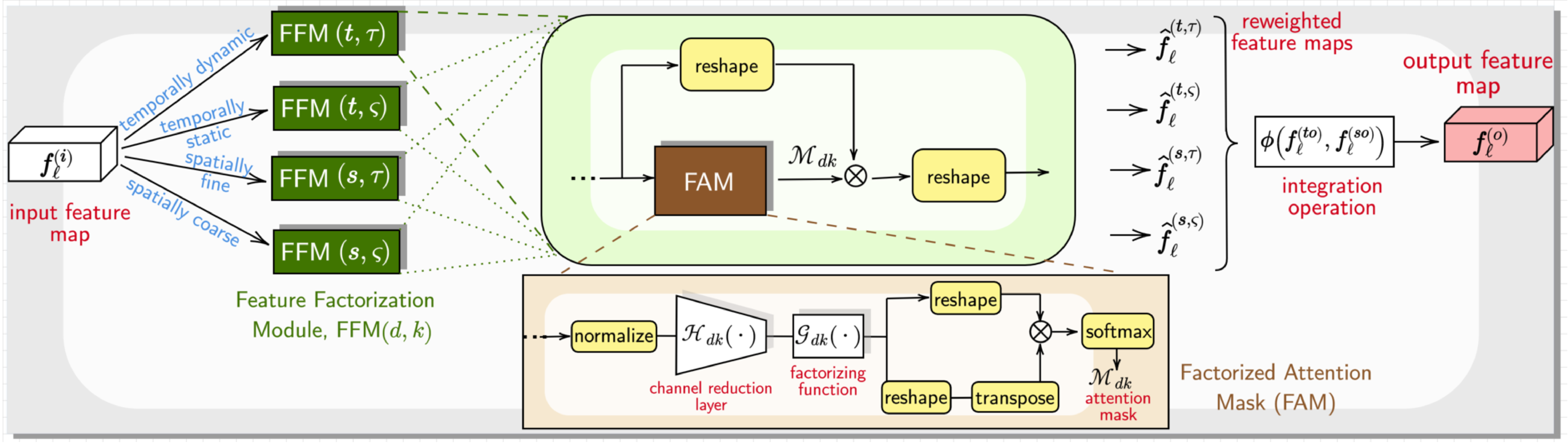


Baseline Improvements

MODEL	P(M)	DATASETS			
		MARS		DukeMTMC	
		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**

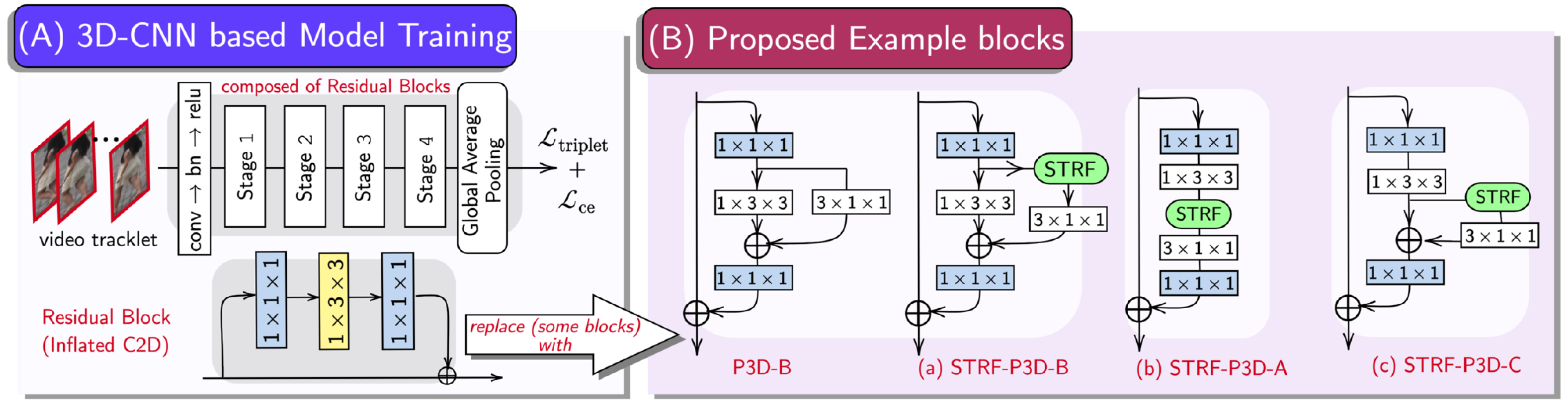
Spatio-Temporal Representation Factorization Unit

- Each STRF module has four factorization units applied on input feature $\boldsymbol{f}^{(i)}$ at ℓth layer.
- STRF module extracts static/coarse and dynamic/fine information and generates richer feature representation from $\boldsymbol{f}^{(i)}$, while adding only 0.5 million parameters (w.r.t. best baseline).
- Each factorization unit is made of a Feature Factorization Module (FFM) aided by our proposed Factorized Attention Mask (FAM) block.
- Outputs of all units are integrated to create the final feature $\boldsymbol{f}^{(o)}$ to be passed to $\ell + 1$th layer.



How to employ STRF?

- With inflated (time dimension of kernel set to 1) C2D residual network as backbone, STRF enhances its feature representation by replacing some blocks at different stages (see below left).
- **e.g.** We replace some C2D blocks with STRF-P3D blocks (see below right).



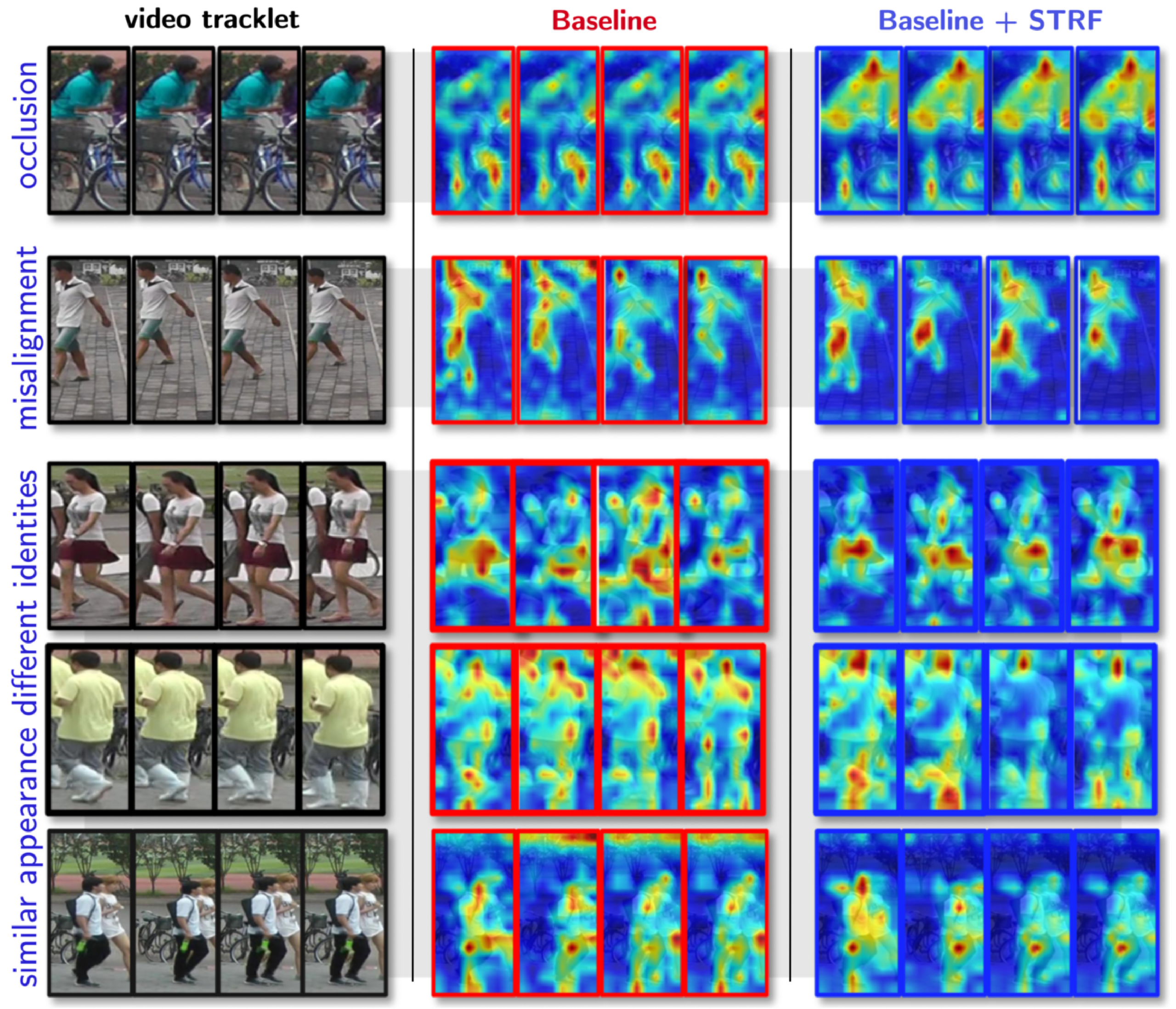
Learning Objective

Any STRF-aided network can be trained in an end-to-end manner with following objective $\mathcal{L}$:

$$\mathcal{L} = \mathcal{L}_{\text{ce}} + \mathcal{L}_{\text{triplet}}$$

where $\mathcal{L}_{\text{ce}}$ is the standard cross-entropy loss, and $\mathcal{L}_{\text{triplet}}$ is the cosine distance based triplet loss with batch-hard mining.

Attention Map Visualization



Comparison to SOTA

METHODS	VENUE	DATASETS				
		MARS		DukeMTMC		iLiDS-VID
		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

(best results in red, second best in blue, and third best results in green.)

Conclusions

- We proposed a novel computational unit that learns complementary spatio-temporal feature representations to deal with real-world re-ID challenges.
- Extensive evaluations with various architectures on benchmark re-ID datasets show STRF's efficacy and generality.