FactorizePhys: Matrix Factorization for Multidimensional Attention in Remote Physiological Sensing

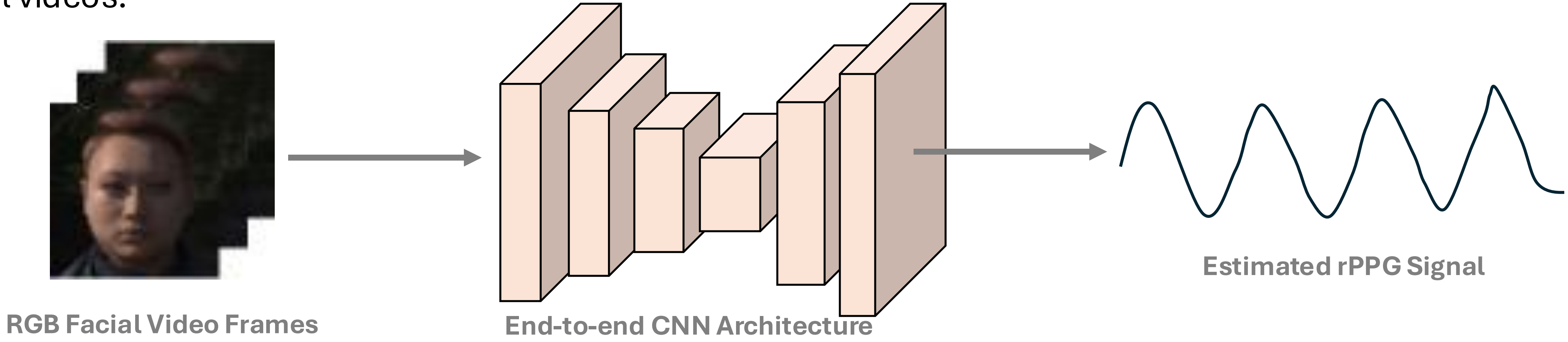
Jitesh Joshi[1], **Sos S. Agaian**[2] and **Youngjun Cho**[1]

[1] Department of Computer Science, University College London, UK

[2] Department of Computer Science, City University of New York, USA

Remote Physiological Sensing

- Remote photoplethysmography (rPPG) enables non-invasive extraction of blood volume pulse (BVP) signals from facial videos.



RGB Facial Video Frames

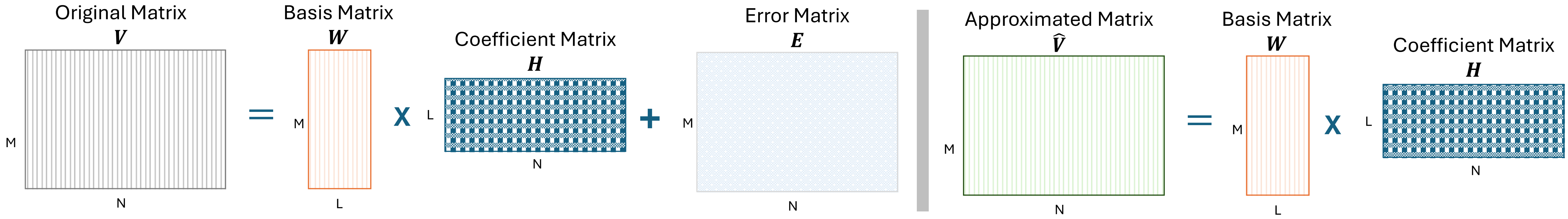
End-to-end CNN Architecture

Estimated rPPG Signal

- End-to-end estimation of BVP requires the network to selectively learn spatial-temporal features, while being invariant to head movement, lighting, and skin tones.
- Existing convolutional attention mechanisms derive attention in spatial, temporal, and channel dimensions disjointly.
- This highlights the complexity of this spatial-temporal task and presents a valuable opportunity to explore attention mechanisms within a multidimensional feature space.

Nonnegative Matrix Factorization as Global Context Block

- Nonnegative matrix factorization (NMF) [1] is a dimensionality reduction paradigm that decomposes $M \times N$ matrix $V = [v_1, v_2, \dots, v_N] \in R_{\geq \mathbb{0}}^{M \times N}$into nonnegative $M \times L$ basis matrix $W = [w1, w2, \dots, wL] \in R_{\geq \mathbb{0}}^{M \times L}$ and nonnegative $L \times N$ coefficient matrix $H = [h1, h2, \dots, hN] \in R_{\geq \mathbb{0}}^{L \times N}$.



- A recent work formulated NMF as an approach to compute low-rank embeddings as a global context block [2], which was shown to be effective in semantic segmentation and image generation tasks.

- Our work focuses on deploying NMF as a multidimensional attention to learn robust spatial-temporal features, without reducing the tensor dimensions.

[1] Lee, D. D., & Seung, H. S. (1999). Learning the parts of objects by non-negative matrix factorization. nature, 401(6755), 788-791.

[2] Geng Z. et al., (2021), Is attention better than matrix decomposition? International Conference on Learning Representations.

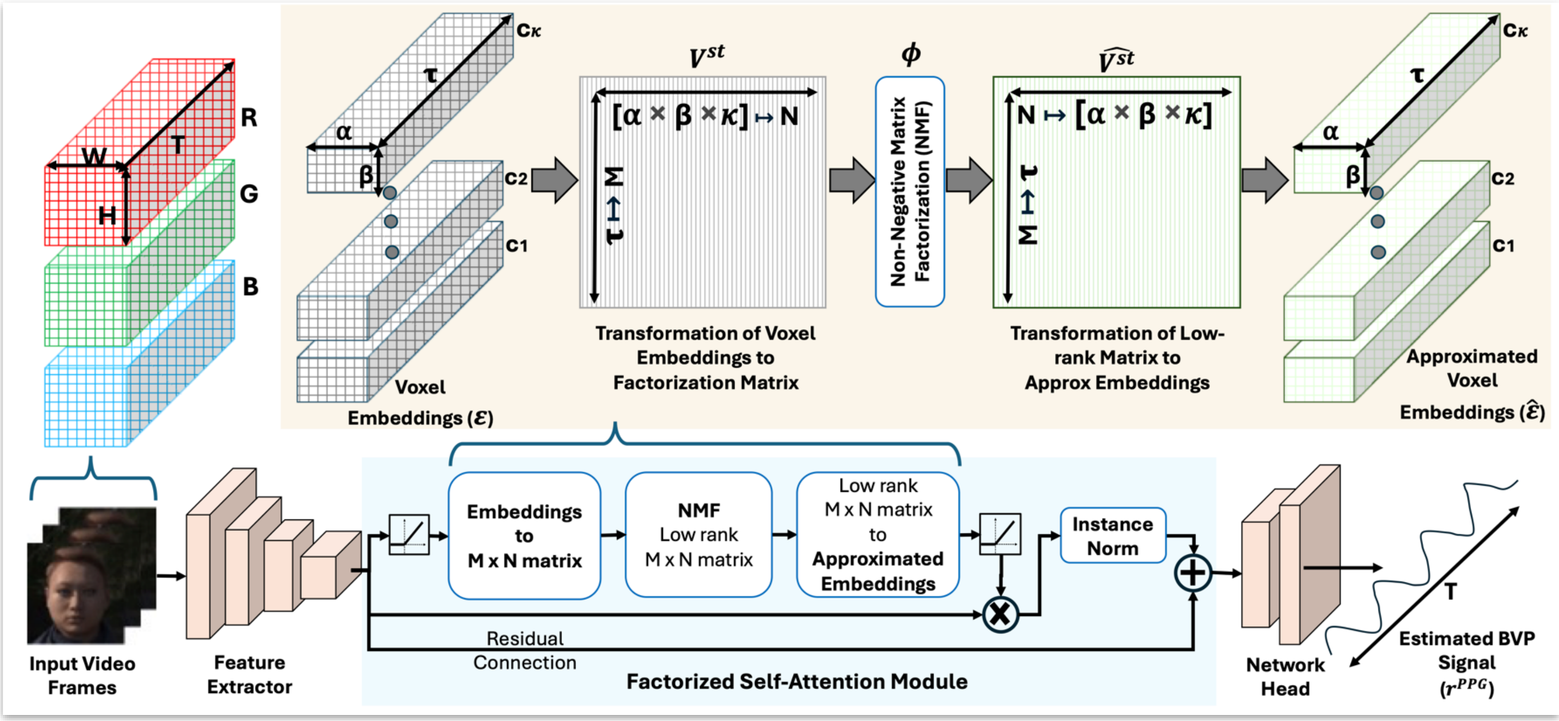
[3] Hu, J., Shen, L., & Sun, G. (2018). Squeeze-and-excitation networks. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 7132-7141).

Factorized Self-Attention Module (FSAM)

- For rPPG estimation task, the spatial-temporal input data is expressed as $\mathcal{I} \in R^{T\times C\times H\times W}$ where:
  - T = Total frames in a video
  - C = Channels in a frame (e.g. for RGB)
  - $H, and\ W$ = Height and width
- Feature extractor generates voxel embeddings $\boldsymbol{\varepsilon} \in R^{\tau\times\kappa\times\alpha\times\beta}$, with temporal (τ), channel (κ) and spatial (α, β) dimensions.
- $\boldsymbol{\varepsilon}$ is transformed to $V^{st} \in R^{M\times N}$, such that:

1 $$V^{st} \in R^{M\times N} = \Gamma^{\tau\kappa\alpha\beta\mapsto MN}\left(\xi_{pre}\left(\varepsilon \in R^{\tau\times\kappa\times\alpha\times\beta}\right)\right)$$
$$\ni \kappa\times\alpha\times\beta \mapsto N; \tau \mapsto M$$

FSAM illustrated for a 3D-CNN architecture for rPPG estimation.



R
G
B
W
T
H
τ
α
β
Cκ
C2
C1
Voxel Embeddings (ℰ)
V^{st}
[α × β × κ] ↦ N
τ ↦ M
Transformation of Voxel Embeddings to Factorization Matrix
ϕ
Non-Negative Matrix Factorization (NMF)
$\widehat{V^{st}}$
N ↦ [α × β × κ]
M ↦ τ
Transformation of Low-rank Matrix to Approx Embeddings
Approximated Voxel Embeddings (ℰ̂)
Input Video Frames
Feature Extractor
Embeddings to M x N matrix
NMF Low rank M x N matrix
Low rank M x N matrix to Approximated Embeddings
Instance Norm
Residual Connection
Factorized Self-Attention Module
Network Head
T
Estimated BVP Signal (r^{PPG})

- V^{st} is factorized, and remapped to embeddings dimension:

2 $$\widehat{V^{st}} = \phi\left(V^{st}\right)$$

3 $$\hat{\varepsilon} = \Gamma^{MN\mapsto\tau\kappa\alpha\beta}\left(\widehat{V^{st}} \in R^{M\times N}\right)$$

- Low rank $\hat{\boldsymbol{\varepsilon}}$ serves as an attention for rPPG estimation.

4 $$r^{ppg} = \omega\left(\varepsilon + \mathcal{IN}\left(\varepsilon \odot \xi_{post}(\hat{\varepsilon})\right)\right)$$

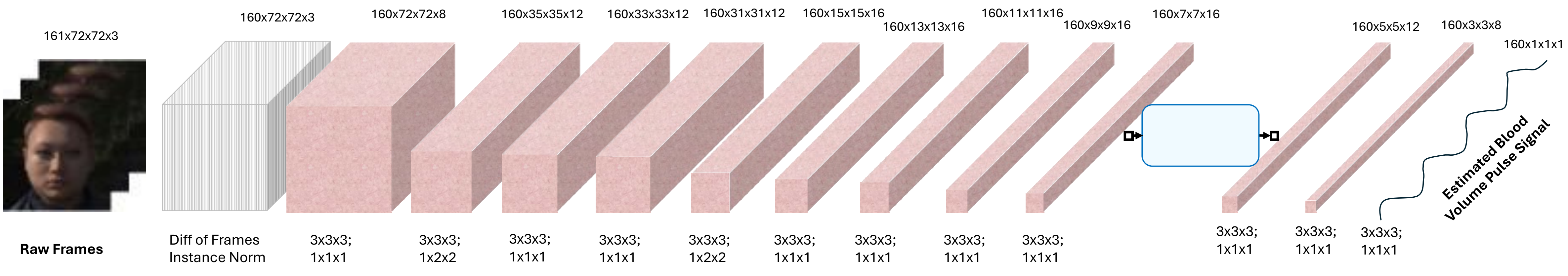
$\xi_{pre}, \xi_{post} = Conv$ layers

ω = Network Head

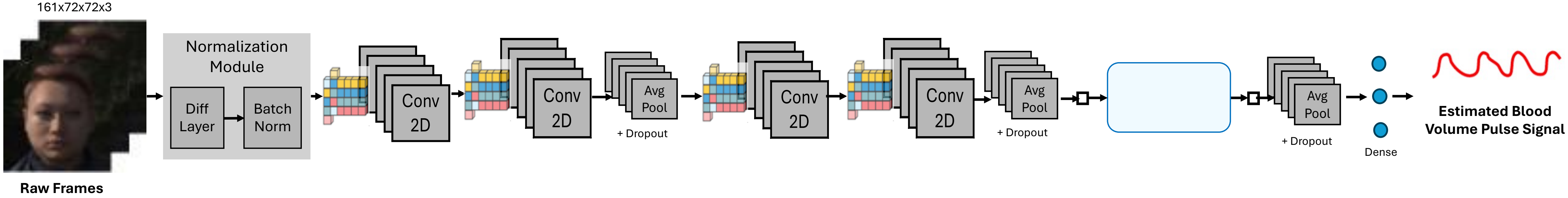
$\mathcal{IN} = Instance\ Normalization$

FSAM Implemented for 3D-CNN and 2D-CNN Architectures

FactorizePhys, Proposed 3D-CNN Architecture, that Implements FSAM



Adaptation of FSAM for EfficientPhys [4], the SOTA 2D-CNN rPPG method



[4] Liu, et al., (2023). Efficientphys: Enabling simple, fast and accurate camera-based cardiac measurement. WACV (pp. 5008-5017).

Main Results

Cross-Dataset Evaluation on PURE Dataset

Training Dataset	Model	Attention Module	MAE (HR) ↓	RMSE (HR) ↓	MAPE (HR)↓	Corr (HR) ↑	SNR (BVP) ↑
iBVP	PhysNet	-	7.78	19.12	8.94	0.59	9.90
	PhysFormer	TD-MHSA	6.58	16.55	6.93	0.76	9.75
	EfficientPhys	SASN	0.56	1.40	0.87	0.998	11.96
	EfficientPhys	FSAM	**0.44**	**1.19**	**0.64**	**0.999**	12.64
	FactorizePhys	FSAM	0.60	1.70	0.87	0.997	**15.19**
SCAMPS	PhysNet	-	26.74	36.19	46.73	0.45	-2.21
	PhysFormer	TD-MHSA	16.64	28.13	30.58	0.51	0.84
	EfficientPhys	SASN	6.21	18.45	12.16	0.74	4.39
	EfficientPhys	FSAM	8.03	19.09	15.12	0.73	3.81
	FactorizePhys	FSAM	**5.43**	**15.80**	**11.1**	**0.80**	**11.40**
UBFC-rPPG	PhysNet	-	10.38	21.14	20.91	0.66	11.01
	PhysFormer	TD-MHSA	8.90	18.77	17.68	0.71	8.73
	EfficientPhys	SASN	4.71	14.52	7.63	0.80	8.77
	EfficientPhys	FSAM	3.69	13.27	5.85	0.83	9.65
	FactorizePhys	FSAM	**0.48**	**1.39**	**0.72**	**0.998**	**14.16**

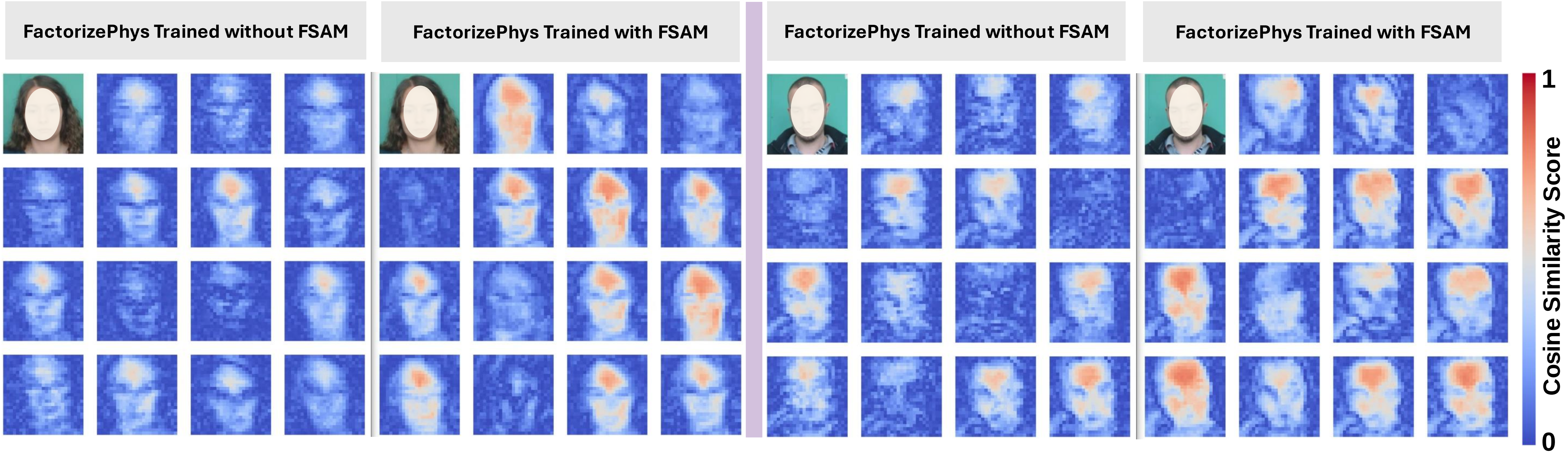
TD-MHSA: Temporal Difference Multi-Head Self-Attention [5]; SASN: Self-Attention Shifted Network [4]

❖ For complete set of main and supplementary results on multiple datasets, please visit our paper.

[4] Liu, et al., (2023). Efficientphys: Enabling simple, fast and accurate camera-based cardiac measurement. WACV (pp. 5008-5017).

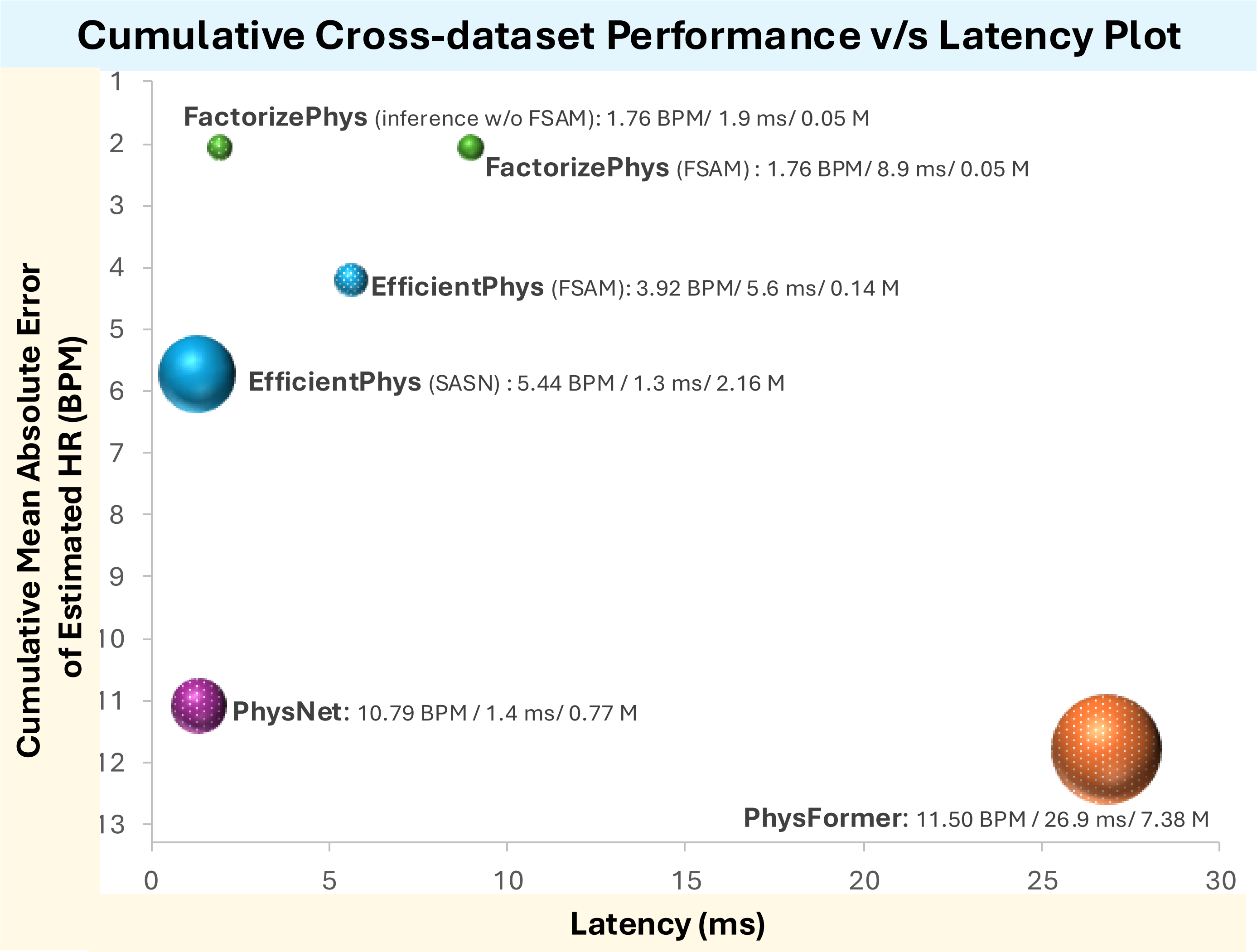
[5] Yu et al., (2022). Physformer: Facial video-based physiological measurement with temporal difference transformer. CVPR(pp. 4186-4196).

Learned Spatial Temporal Features



Each tile represents cosine similarity (CSIM) map computed between the ground-truth PPG signal and the temporal dimension of each channel of the embedding layer. The magnitude and spatial spread of CSIM scores, as depicted with the heatmaps, highlight the effectiveness of the FSAM.

Summarized Performance on Multiple rPPG Datasets



Sphere Size ∝ No. of parameters

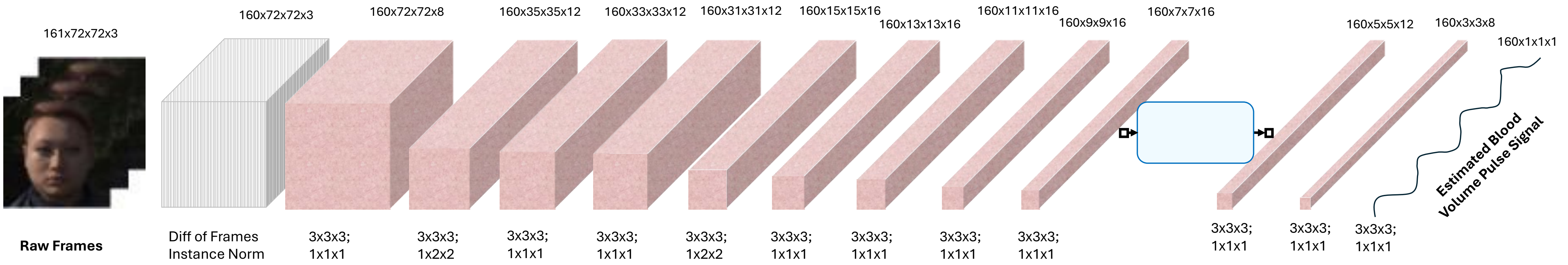
Model: MAE/ Latency/ No. of parameters

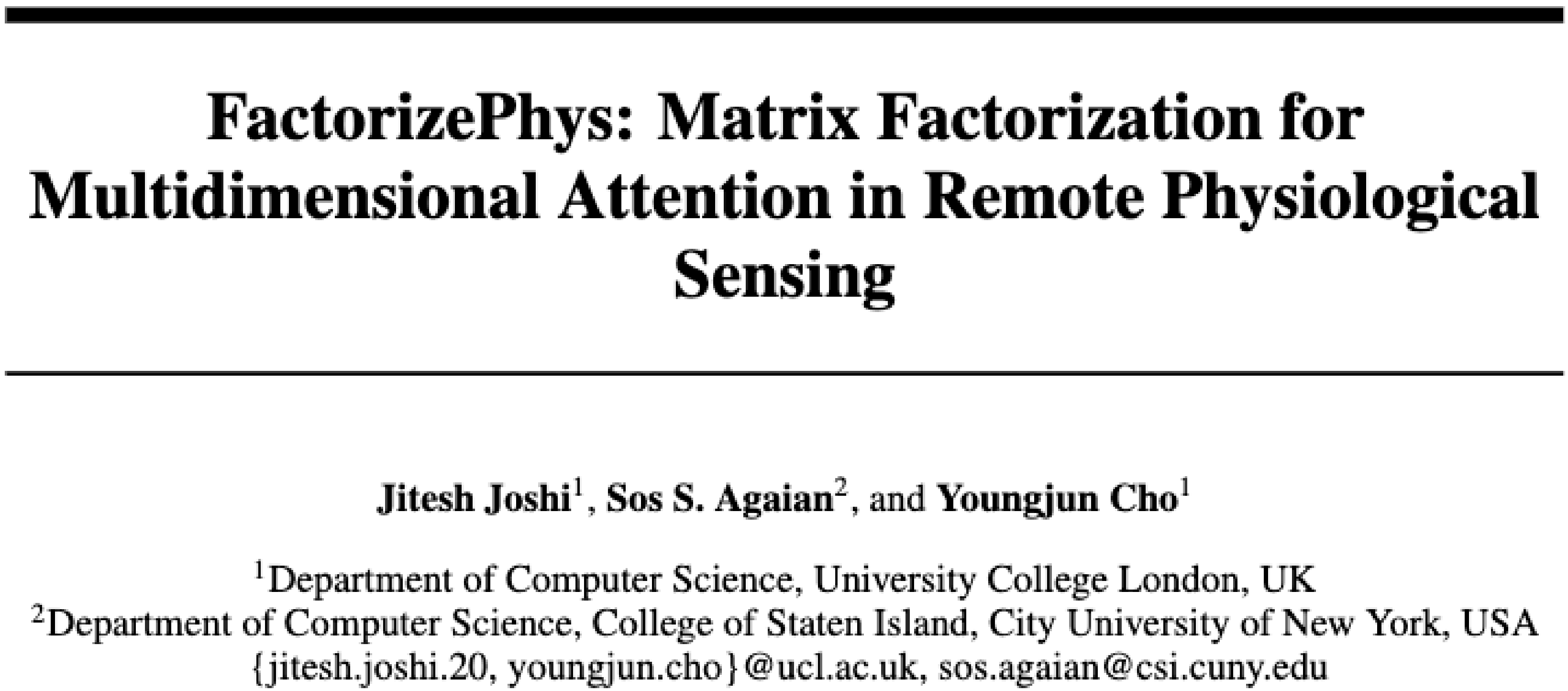
Conclusion

- FactorizePhys, equipped with FSAM, provides robust cross-dataset generalization for the extraction of rPPG signals.

- FSAM serves as an efficient multi-dimensional attention mechanism for remote physiological sensing, with potential applicability to other spatial-temporal downstream tasks.

FactorizePhys with FSAM





FactorizePhys: Matrix Factorization for Multidimensional Attention in Remote Physiological Sensing

Jitesh Joshi[1], **Sos S. Agaian**[2], and **Youngjun Cho**[1]

[1]Department of Computer Science, University College London, UK
[2]Department of Computer Science, College of Staten Island, City University of New York, USA
{jitesh.joshi.20, youngjun.cho}@ucl.ac.uk, sos.agaian@csi.cuny.edu

- Code: https://github.com/PhysiologicAILab/FactorizePhys

38th Conference on Neural Information Processing Systems (NeurIPS 2024)