RIT Rochester Institute of Technology

Multimodal Aerial View Object Classification with Disjoint Unimodal Feature Extraction and Fully-Connected-Layer Fusion

Mr. Saurav Singh *
Rochester Institute of Technology
Rochester, NY, USA
ss3337@rit.edu

Mr. Manish Sharma *
Rochester Institute of Technology
Rochester, NY, USA
ms8515@rit.edu

Dr. Jamison Heard
Rochester Institute of Technology
Rochester, NY, USA
jrheee@rit.edu

Mr. Jesse D. Lew
New York University
NY, USA
jl8429@nyu.edu

Dr. Eli Saber
Rochester Institute of Technology
Rochester, NY, USA
esseee@rit.edu

Dr. Panos P. Markopoulos
The University of Texas at San Antonio
San Antonio, TX, USA
panagiotis.markopoulos@utsa.edu

*** Authors contributed equally to this research article**

Acknowledgement

Air Force Office of Scientific Research (AFOSR) under award FA9550-20-1-0039

Department of Defense under award HM04761912014

- Special thanks to **Dr. Denise. M. Reeves** for her valuable contribution toward this work.

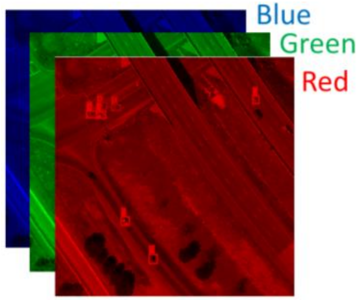
Motivation

Multimodal Data:

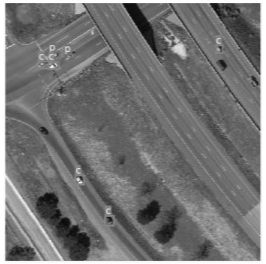
- Information about same phenomenon acquired from different types of sensors.
- Each modality gives optimal information under certain conditions.
- Fusing multi-modal data enhances the discovery of underlying information.[1][2][3]

Sensors:

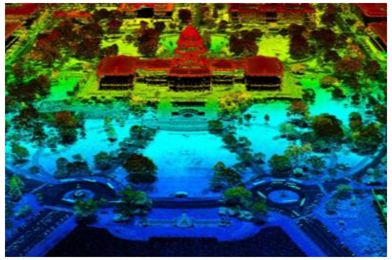
- RGB
- Infrared (IR)
- Light Detection and Ranging (LiDAR)
- Synthetic Aperture Radar (SAR)
- Hyperspectral



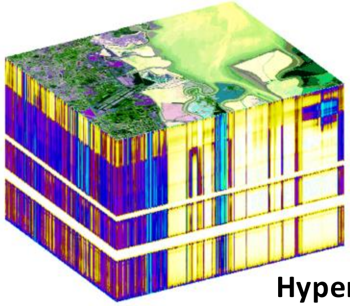
Infrared



LiDAR



Synthetic Aperture Radar



Hyperspectral

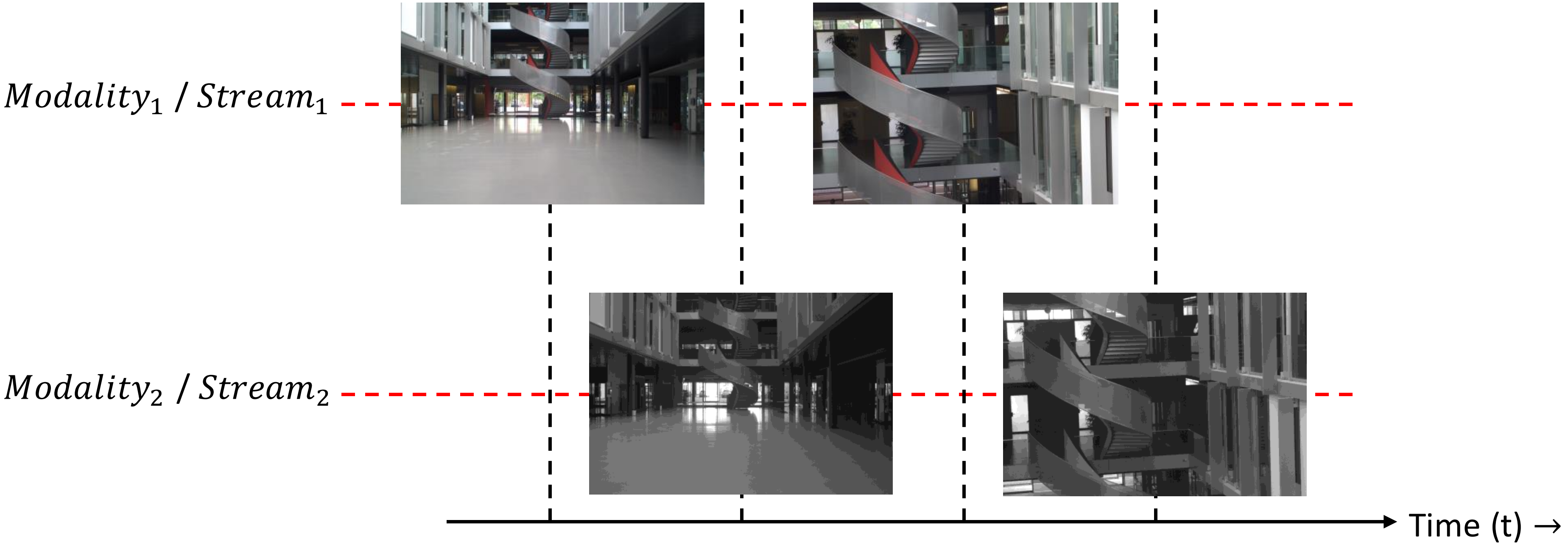
[1] Datta, D., Mallick, P. K., Bhoi, A. K., Ijaz, M. F., Shafi, J., and Choi, J., "Hyperspectral image classification: Potentials, challenges, and future directions," *Computational Intelligence and Neuroscience 2022* (2022).
[2] Yu, J., Chang, H., Lu, K., Zhang, L., and Du, S., "Scene clustering based pseudo-labeling strategy for multi-modal aerial view object classification," *arXiv preprint arXiv:2205.01920* (2022).
[3] Li, J., Hong, D., Gao, L., Yao, J., Zheng, K., Zhang, B., and Chanussot, J., "Deep learning in multimodal remote sensing data fusion: A comprehensive review," *International Journal of Applied Earth Observation and Geoinformation 112, 102926* (2022).

Motivation

Challenges with multimodal data collection in aerial imagery:

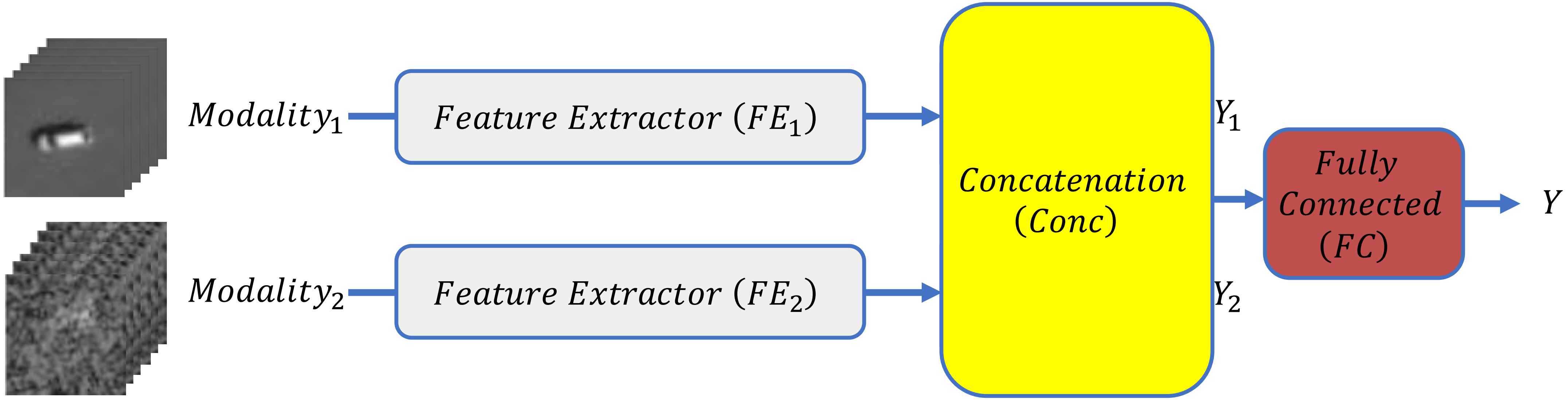
Limited availability of paired multimodal data for training an end-to-end multimodal fusion network [4][5], where paired multimodal data samples simultaneously agree with the following conditions:

- Spatial and temporal correspondence
- Synchronous occurrence/availability



[4] Lahat, D., Adali, T., and Jutten, C., "Multimodal data fusion: an overview of methods, challenges, and prospects," *Proceedings of the IEEE 103(9)*, 1449–1477 (2015).
[5] Zhu, B., Zhou, L., Pu, S., Fan, J., and Ye, Y., "Advances and challenges in multimodal remote sensing image registration," *IEEE Journal on Miniaturization for Air and Space Systems* (2023).

Motivation



Research Questions:

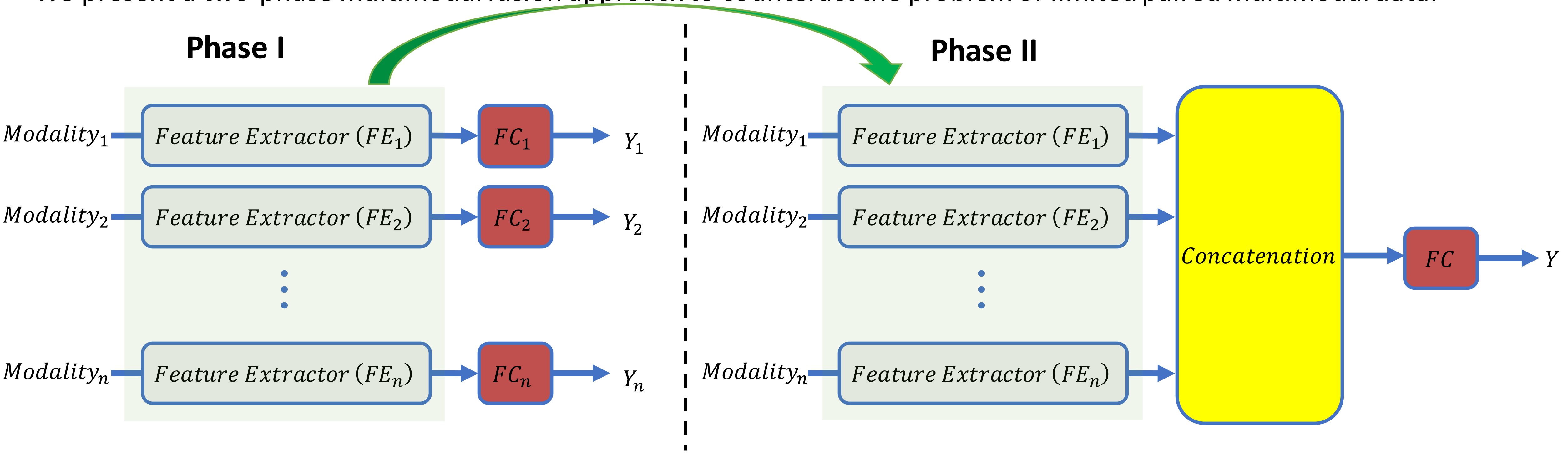
RQ1: Is it computationally advantageous to fuse legacy unimodal pre-trained networks?

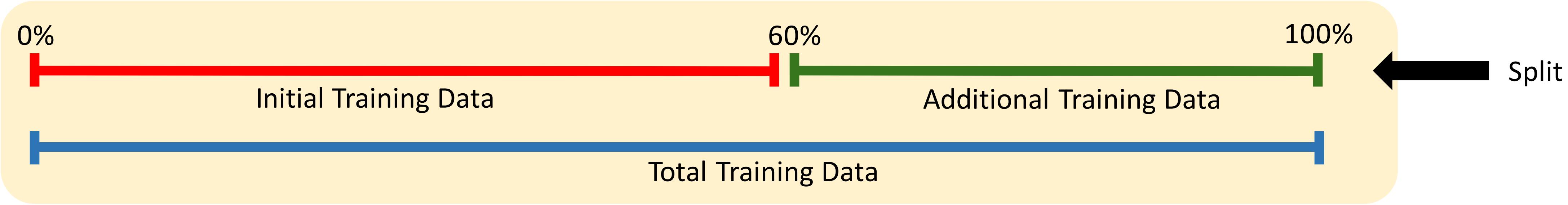
RQ2: What are the efficient approaches to train a fusion network if all paired multimodal data is available?

Methodology

RQ1: Is it computationally advantageous to fuse legacy unimodal pre-trained networks?

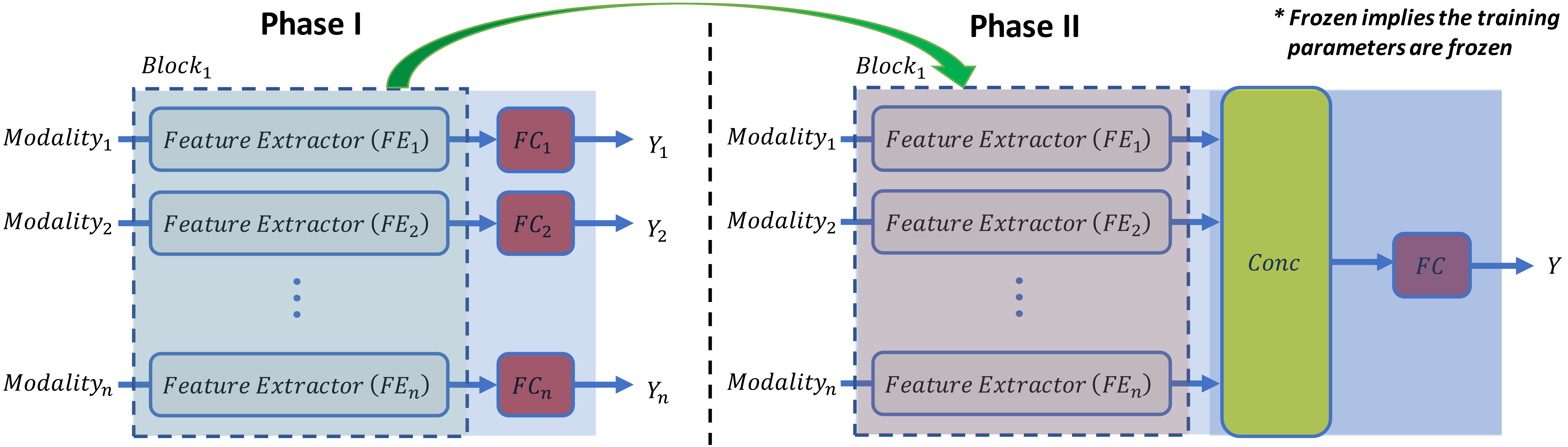
We present a two-phase multimodal fusion approach to counteract the problem of limited paired multimodal data.





Methodology

RQ2: What are the efficient approaches to train a fusion network if all paired multimodal data is available?



*** Frozen implies the training parameters are frozen***

Configuration	Block 1 Initialization Parameters	Block 1 Frozen
Unimodal	Random	No
Joint Stream	Random	No
Non-Frozen *	Unimodal weights	No
Frozen *	Unimodal weights	Yes

NTIRE-21 Dataset [6]

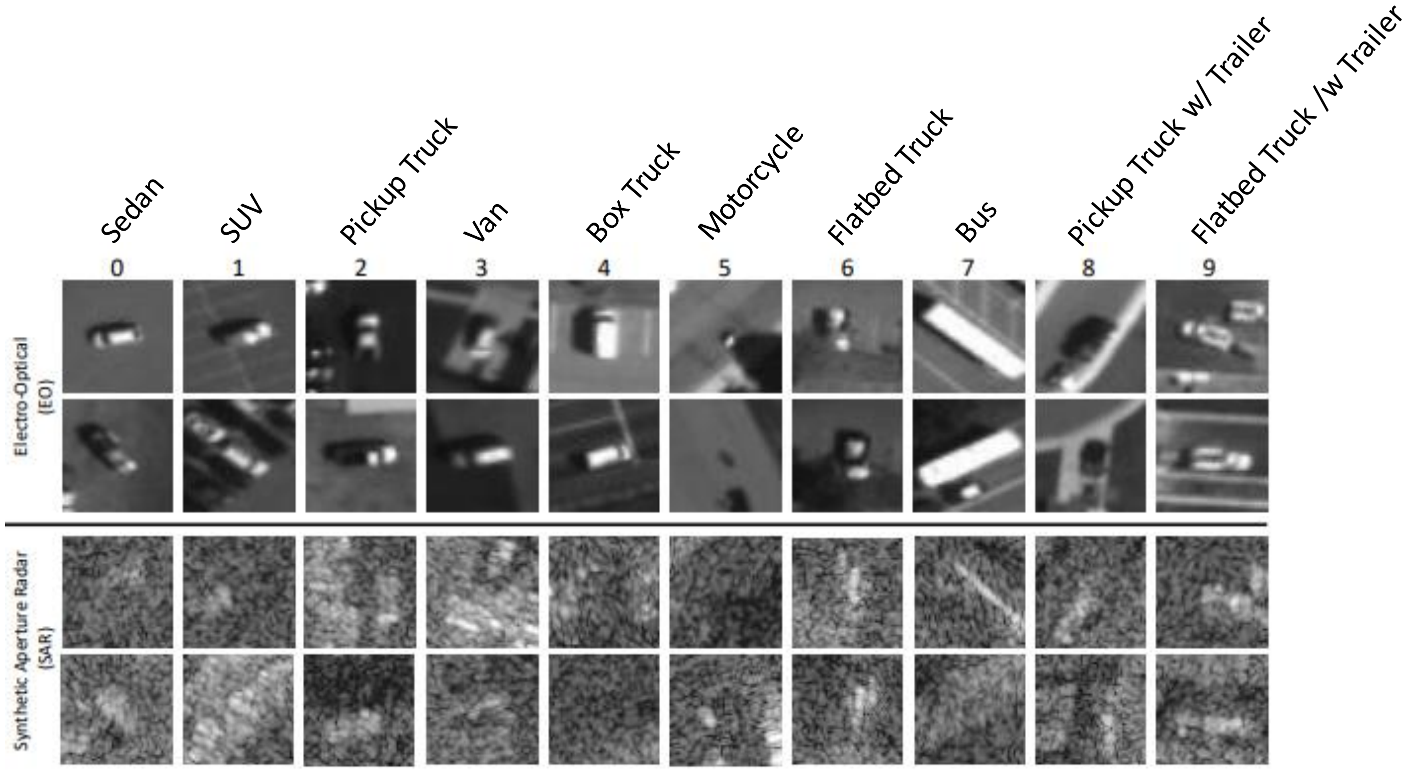
Multimodal dataset from NTIRE 2021 Multi-modal Aerial View Object Classification Challenge includes:

- Electro-Optical (EO) Images
- Synthetic Aperture Radar (SAR) Images

Classes: 10

Samples per Class: 625

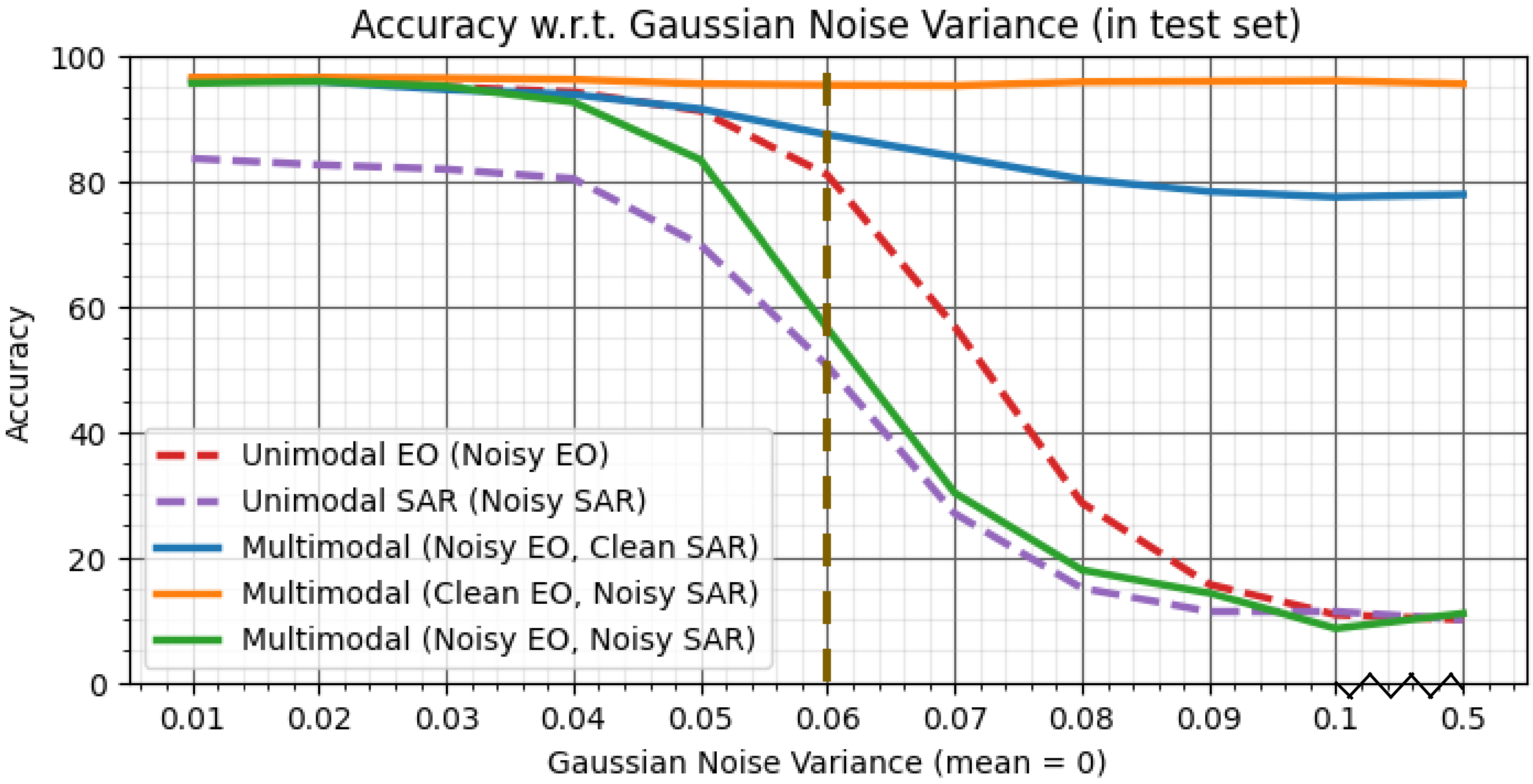
Training / Testing: 5250 / 1000



[6] J. Liu et al., "NTIRE 2021 Multi-modal Aerial View Object Classification Challenge," *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 2021*, pp. 588-595, doi: 10.1109/CVPRW53098.2021.00071.

Results: Performance of Fusion on Noisy Data

Training data: Gaussian Noise with $\mu = 0$ and $\sigma^2 = 0.02$
Test data: Incremental Gaussian Noise with $\mu = 0$

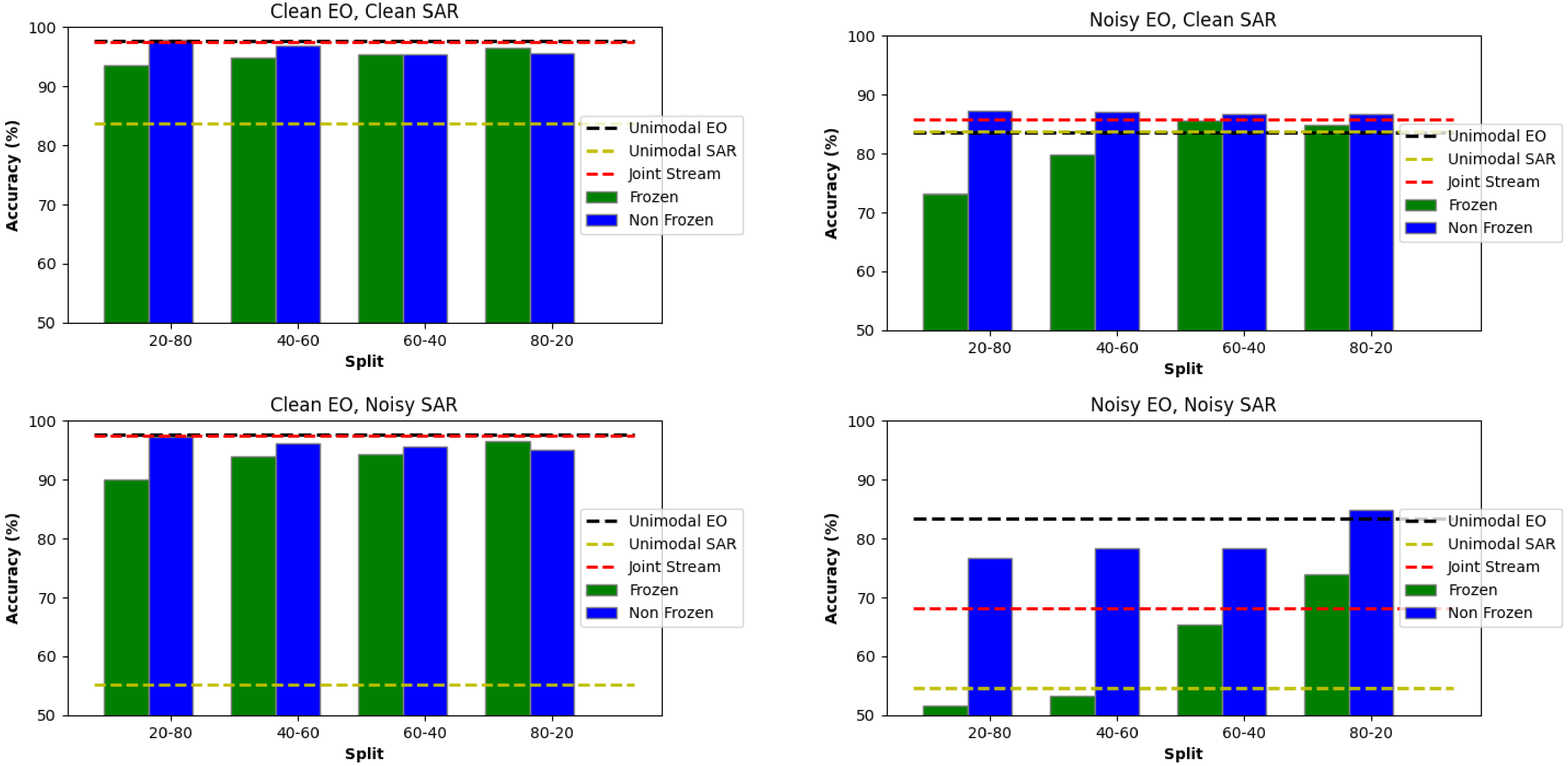


EO is the Dominant Modality because the network performance is affected more by the presence of noise in the EO modality.

Results: Accuracy vs Split

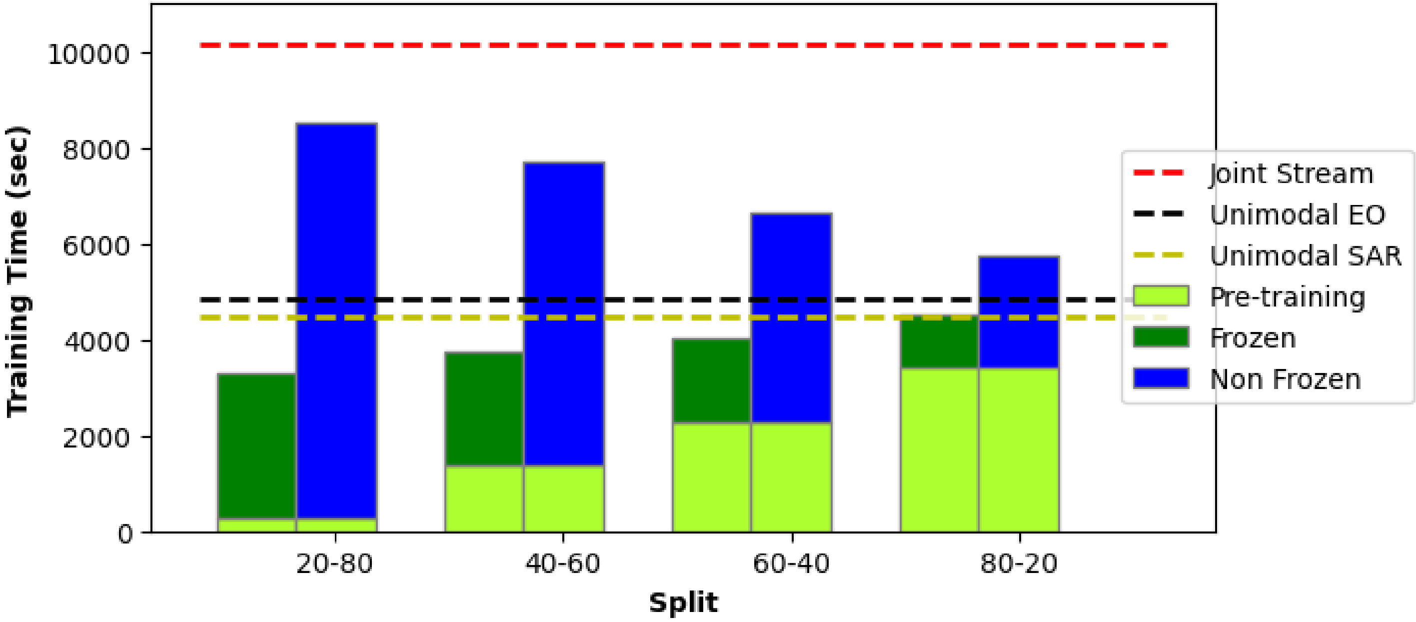
Split: Unimodal training data – Multimodal fusion training data

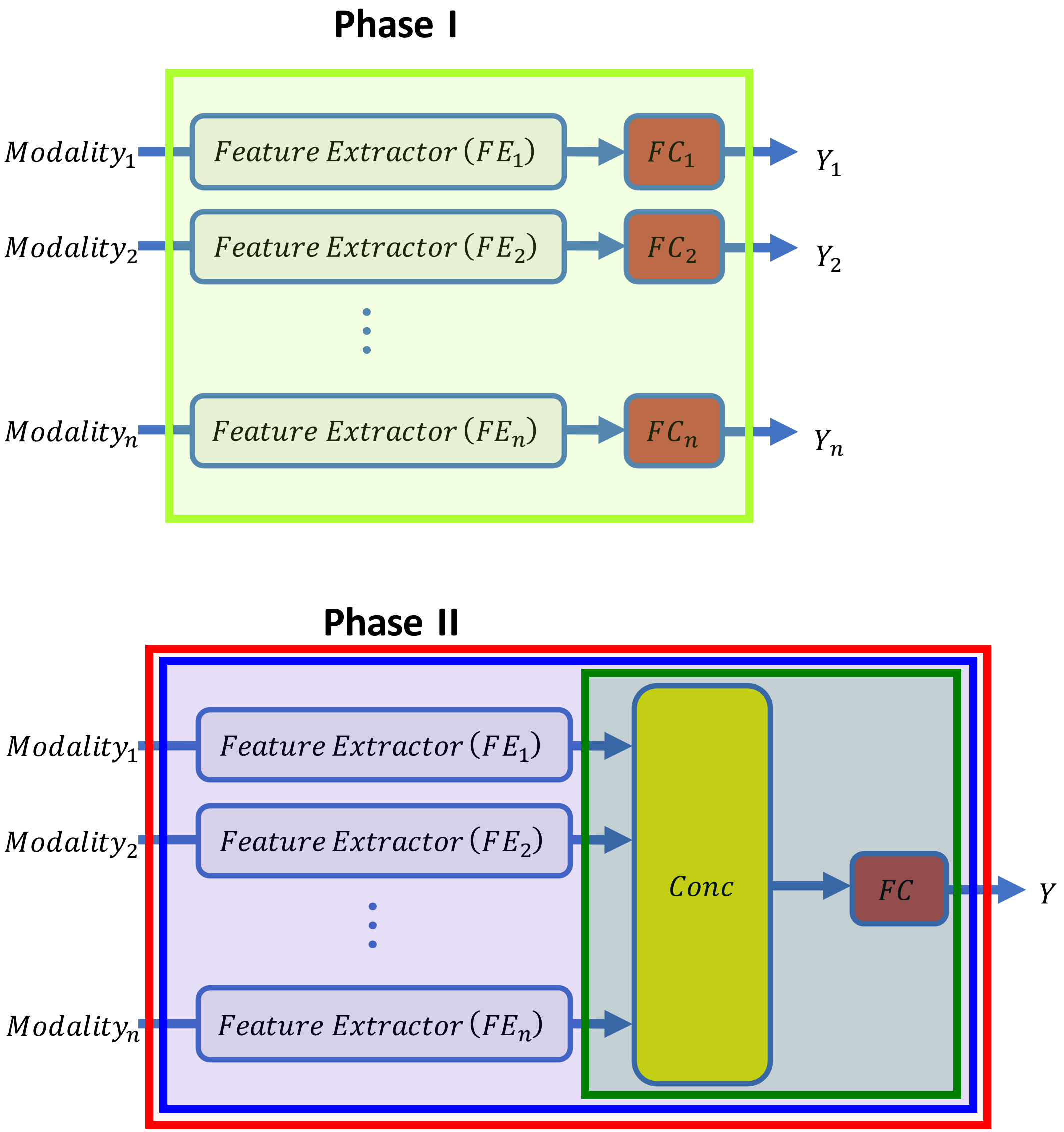
Noise in the test dataset: Gaussian Noise with $\mu = 0$ and $\sigma^2 = 0.06$



Frozen and Non-Frozen configurations performs better when there is significant noise in the Dominant Modality (EO)

Results: Training Time





Conclusions

- Proposed a two-phase multimodal network training method that provides a way to **fuse legacy unimodal networks** trained on unpaired data from different modalities into a multimodal network.
- The **training time** of the multimodal network with the proposed method is **significantly less** than joint stream end-to-end training of the multimodal network; however, there is a **small yet acceptable drop in the performance accuracy**.
- Enhances the usability of the legacy unimodal networks while transitioning to the multimodal sensing paradigm and would benefit industries such as satellite surveillance, and autonomous vehicles.

RQ1: Is it computationally advantageous to fuse legacy unimodal pre-trained networks? **YES!**

- **Less data required** for fusion training since **80-20 split** generally performs the best.
- **Training time** is significantly **reduced**.

RQ2: What are the efficient approaches to train a fusion network if all paired multimodal data is available? **Depends!**

- **Least training time:** Frozen
- **Most robust to noisy data:** Non-Frozen
- **Highest accuracy:** Joint Stream

Thank You!!

Any questions

Please feel free to reach out to us at ss3337@rit.edu