Collaborative Unpaired Multimodal Representation Learning for Satellite Imagery

Akansh Maurya[1,2] **Rotem Mulayoff**[2] **Sebastian U. Stich**[2]

[1]Universität des Saarlandes [2]CISPA Helmholtz Center for Information Security

Funded by the European Union

European Research Council
Established by the European Commission

Challenge in EO Representation Learning

- **Paired data**: multimodal samples aligned to the same scene; **unpaired data**: independent acquisitions without spatial or temporal correspondence.
- **Paired multimodal samples are rare.** Large satellite archives are **unpaired, unaligned, and distributed**, and data sharing is often restricted due to privacy and policy constraints.

Yet we know multimodal supervision improves in training better classifiers.

Research question

How can institutions with different EO modalities train stronger unimodal models, without the need for paired multimodal data and sharing private data?

Limitations of Existing Approaches

- **Unimodal training:** scalable, but wastes complementary modalities.
- **Paired fusion:** strong performance, but costly paired/co-registered samples acquisition.
- **Missing-modality / pseudo-pairing:** reduces pairing needs, but requires partial supervision or modality-homogeneous encoders.

Our Contribution

- We introduced **Unpaired multimodal learning (UML)**, a framework that enables collaboration among multimodal institutions without any paired data.
- Each modality has a learnable projection layers. A **shared backbone** learns modality-agnostic semantic features from unpaired batches, followed by **post-hoc BN calibration** that adapts the shared model to each modality.
- This method can be extended to **Federated setting**, where modality data remains local to the institution.

Take-aways:

- **No paired data** and **no multimodal inference** required.
- **Scales to heterogeneous modalities.**
- **Consistent gains over unimodal baselines**, especially for weaker modalities and low-data regimes.
- **Privacy-aware**: data remain local, only backbone weights shared.

Dataset and Pre-processing

Multimodal dataset: BigEarthNet-MM, SEN12MS and EuroSAT-S1/RGB

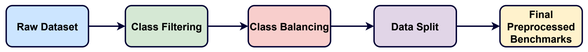

Figure 1. Creating balanced Dataset

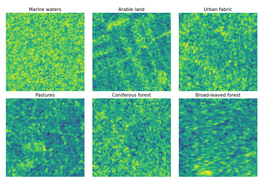

Figure 2. Samples from BGE-MM Sentinel-1 (VV)

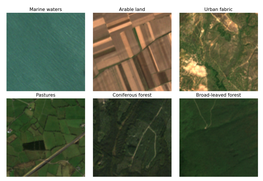

Figure 3. Samples from BGE-MM Sentinel-2 (RGB)

Unpaired vs Paired Multimodal Learning

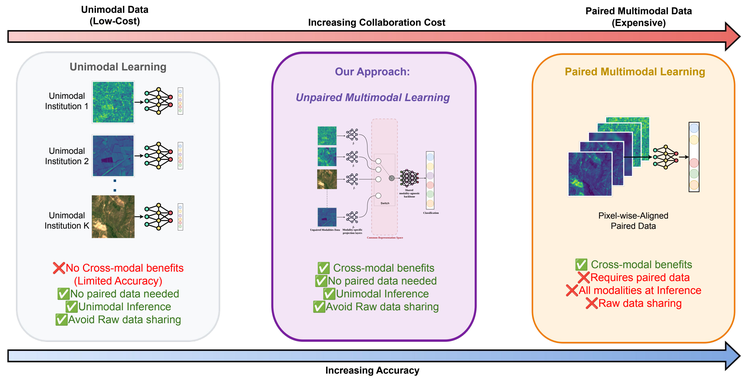

Figure 4. Our framework targets the middle ground, enabling cross-modal benefits without paired data or multimodal inputs at test time.

Proposed Solution

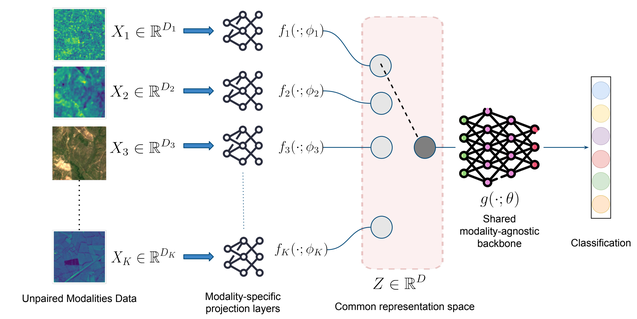

Figure 5. Solution consists of Modality Specific Projection layers and Common Backbone layers + Post-hoc BN calibration step.

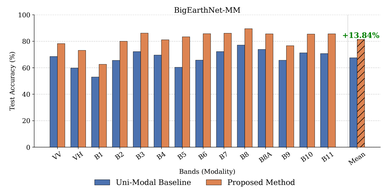

Figure 6. Experiment 1: BigEarthNet-MM (fine-grained)

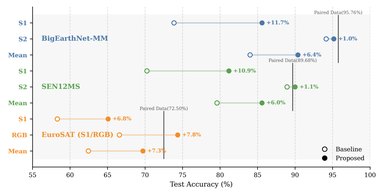

Figure 7. Experiment 2: Bi-modal (S1 vs. S2/RGB)

Algorithm 1 Post-hoc Batch Normalization (BN) Calibration

Require: Trained backbone $g(\cdot;\theta)$, projections $\{f_k(\cdot;\phi_k)\}_{k=1}^K$
Require: Calibration epochs E_{cal}, per-modality datasets $\{\mathcal{D}_k\}_{k=1}^K$, batch size B

```
for k = 1 to K do
    g_k(·;θ_k) ← copy(g(·;θ))                          ▷ independent copy
    Freeze all parameters in g_k and f_k               ▷ weights fixed
    Reset BN running statistics in g_k and f_k to zero
    Set BN layers to accumulate statistics without momentum (with CMA)
    for e = 1 to E_cal do
        for mini-batch {x_i^k}_{i=1}^B ~ D_k do
            Z ← f_k({x_i^k}_{i=1}^B; φ_k)              ▷ batch processing
            g_k(Z; θ_k)                                ▷ forward only, updates BN stats
    Output: Calibrated model h_k = g_k ∘ f_k
```

Benchmarking

We compared to Identifiable Shared Component Analysis (ISCA) [1]—the only prior method for fully unpaired multimodal learning. We also compared our work with standard Domain Adaptation methods with heterogeneous encoders.

Method	BigEarthNet-MM			SEN12MS			EuroSAT S1–RGB		
	S1	S2	Mean	S1	S2	Mean	S1	RGB	Mean
Unimodal Baseline	73.86	94.14	84.00	70.25	88.92	79.59	58.30	66.57	62.44
ISCA [1]	80.76	93.45	87.10	70.67	89.29	79.98	62.75	69.35	66.05
Proposed	**85.59**	**95.16**	**90.38**	**81.17**	**90.00**	**85.59**	**65.07**	**74.35**	**69.71**
Paired Data (Multi-modal inference)	-		95.76	-		89.68	-		72.50

Table 1. Accuracy (%) comparison with unpaired multimodal learning method.

Method	BigEarthNet-MM			SEN12MS			EuroSAT S1–RGB		
	S1	S2	Mean	S1	S2	Mean	S1	RGB	Mean
Unimodal Baseline	73.86	94.14	84.00	70.25	88.92	79.59	58.30	66.57	62.44
DANN [2]	71.31	94.09	82.70	69.71	88.79	79.25	58.67	68.25	63.46
CDAN [3]	77.81	93.76	85.79	71.25	86.18	78.71	58.07	68.92	63.50
MCC [4]	76.95	93.47	85.21	70.00	88.64	79.32	59.67	66.00	62.84
MDD [5]	70.36	94.16	82.26	67.25	87.29	77.27	62.17	66.42	64.30
Proposed	**85.59**	**95.16**	**90.38**	**81.17**	**90.00**	**85.59**	**65.07**	**74.35**	**69.71**

Table 2. Comparison with domain adaptation baselines.

Federated extension

Our approach naturally extends to federated settings by sharing global backbone parameters. Figure 8 and 9 shows consistent gains across local epochs ($L \in \{2, 5, 10, 25, 50\}$). Larger L reduces communication rounds R. Total training budget ($R \times L = 200$) remains fixed.

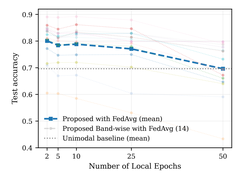

Figure 8. Federated extension of Exp. 1

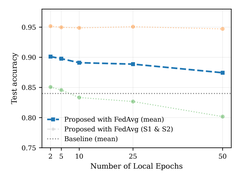

Figure 9. Federated extension of Exp. 2

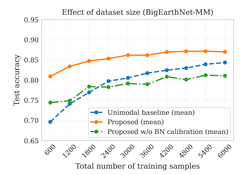

Figure 10. Effect of data size and BN Calibration

Limitation and Future work

- Benefits diminish as per-modality data becomes abundant (Fig. 10).
- Our evaluation is limited to CNN architectures.
- Extreme data imbalance may lead to modality domination effects.
- The method is for the fully unpaired setting and cannot leverage partially paired data.

References

[1] S. Timilsina, S. Shrestha, and X. Fu, "Identifiable shared component analysis of unpaired multimodal mixtures," 2024.

[2] Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. March, and V. Lempitsky, "Domain-adversarial training of neural networks," *Journal of Machine Learning Research*, vol. 17, no. 59, pp. 1–35, 2016.

[3] M. Long, Z. Cao, J. Wang, and M. I. Jordan, "Conditional adversarial domain adaptation," *Advances in neural information processing systems*, vol. 31, 2018.

[4] Y. Jin, X. Wang, M. Long, and J. Wang, "Minimum class confusion for versatile domain adaptation," in *Computer Vision - ECCV 2020* (A. Vedaldi, H. Bischof, T. Brox, and J. Frahm, eds.), vol. 12366 of *Lecture Notes in Computer Science*, pp. 464–480, Springer, 2020.

[5] J. Li, E. Chen, Z. Ding, L. Zhu, K. Lu, and H. T. Shen, "Maximum density divergence for domain adaptation," *IEEE transactions on pattern analysis and machine intelligence*, vol. 43, no. 11, pp. 3918–3930, 2020.