Can multimodal representation learning by alignment preserve modality-specific information?

MACLEAN workshop @ ECML-PKDD 2025

Romain Thoreau[1,3] Jessie Levillain[1,2] Dawa Derksen[1]

[1]CNES [2]INSA Toulouse [3]AgroParisTech - UMR MIA Paris-Saclay

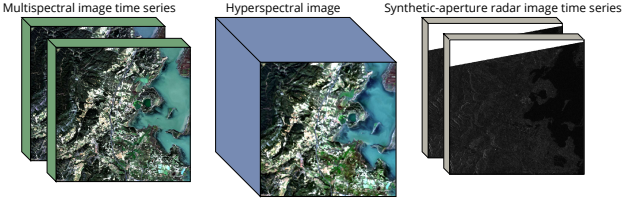
romain.thoreau@agroparistech.fr

September 19, 2025

Overview

Multimodal satellite image analysis

A myriad of sensors acquire massive volumes of data with different **modalities** (*e.g.* optical, radar), and different **resolutions** (*e.g.* spatial, spectral, temporal).



Different sensors usually provide **complementary information** about the Earth's biosphere, *e.g.*

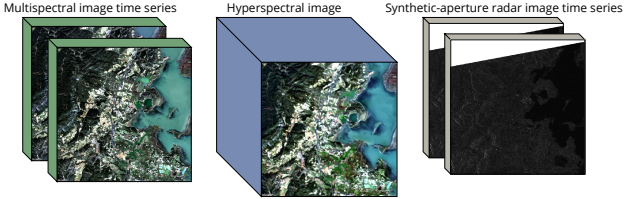
— vegetation carbon concentration estimated from hyperspectral images [Miraglio et al., 2023],

— above-ground biomass estimated from SAR data [Englhart et al., 2011],

— vegetation water deficit estimated from multispectral image time series [Penot and Merlin, 2023].

Multimodal satellite image analysis

A myriad of sensors acquire massive volumes of data with different **modalities** (*e.g.* optical, radar), and different **resolutions** (*e.g.* spatial, spectral, temporal).



Different sensors usually provide **complementary information** about the Earth's biosphere, *e.g.*

— vegetation carbon concentration estimated from hyperspectral images [Miraglio et al., 2023],

— above-ground biomass estimated from SAR data [Englhart et al., 2011],

— vegetation water deficit estimated from multispectral image time series [Penot and Merlin, 2023].

Supervised multimodal representation learning

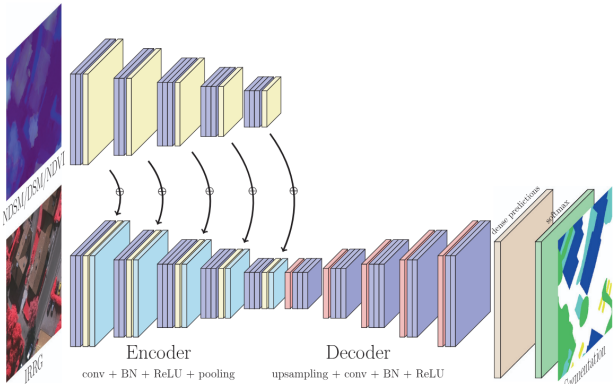


Figure: Illustration of supervised data fusion in remote sensing. Figure from [Audebert et al., 2018].

Self-supervised representation learning: the contrastive framework

Chen et al. **A simple framework for contrastive learning of visual representations.** International conference on machine learning. PmLR, 2020.

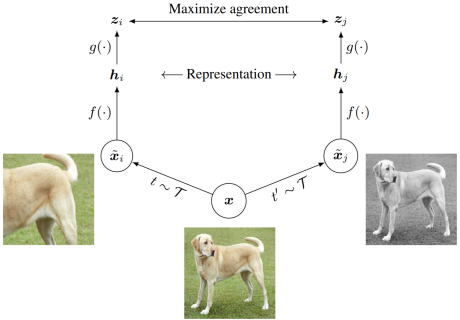


Figure: Illustration of the contrastive framework. Modified from [Chen et al., 2020].

— The projection head $g(\cdot)$ is a small neural network ($g(h_i) = W^{(2)}\mathsf{ReLU}(W^{(1)}h_i)$).

— The loss for a positive pair within a mini-batch is

$$l = -\log \frac{\exp(\mathsf{sim}(z_i, z_j)/\tau)}{\sum_{k=1} \mathbb{1}_{[k \neq i]} \exp(\mathsf{sim}(z_i, z_k)/\tau)}.$$

where sim is the cosine similarity.

— Yields invariance to (hopefully) task-irrelevant information.

Self-supervised representation learning: the joint-embedding predictive architecture (JEPA)

Assran et al. **Self-supervised learning from images with a joint-embedding predictive architecture.** Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023.

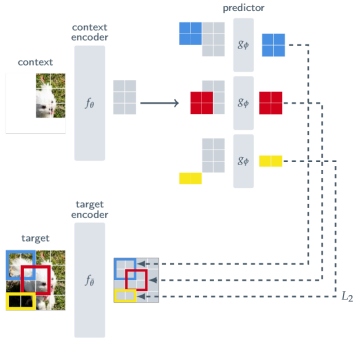


Figure: Illustration of the JEPA framework.
From [Assran et al., 2023].

— Generates only one view, called the *context*.

— The predictor is a ViT that takes the latent representation of the context view and the information of the target position.

— The loss is the L_2 distance between the predicted and target representations.

Self-supervised representation learning in remote sensing: modality alignment

Multimodal learning by alignment in remote sensing: [Pielawski et al., 2020], [Scheibenreif et al., 2022], [Jain et al., 2022], [Feng et al., 2023], [Prexl and Schmitt, 2023], [Astruc et al., 2024], [Marsocci and Audebert, 2025]...

Astruc et al. **AnySat: One Earth Observation Model for Many Resolutions, Scales, and Modalities.** Proceedings of the Computer Vision and Pattern Recognition Conference. 2025.

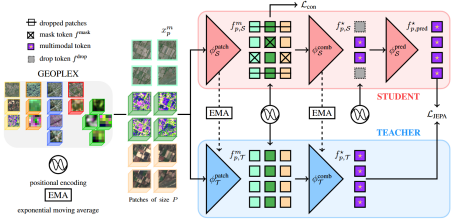


Figure: Illustration of the AnySat learning framework. From [Astruc et al., 2025].

— Combines a contrastive loss and a JEPA loss.

— [JEPA loss] For random patches of the context view, modalities are masked.

Task ≈ predict the targets (all modalities) given the context with missing modalities.

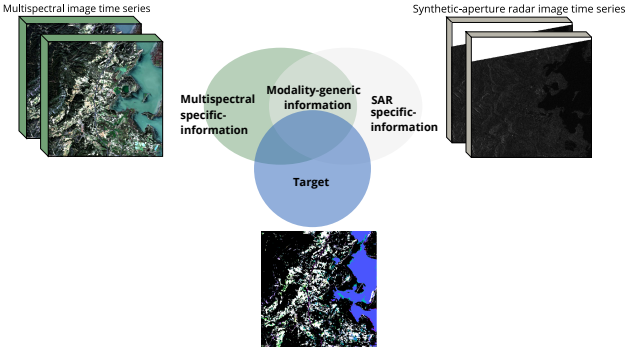
— [Contrastive loss] Modalities are processed as views.

Task ≈ align the representations of the modalities in the latent space.

Motivation of this work

- Contrastive learning is based on the multi-view redundancy assumption:
task-relevant information is shared across the different views of the data [Tian et al., 2020].

- In remote sensing, multimodal data provide complementary and non-redundant information about the downstream task.



Research question

Can multimodal representation learning by alignment preserve modality-specific information?

Multimodal setting & the multi-view non-redundancy assumption

Let us consider two modalities $\boldsymbol{X}_1 \in \mathbb{R}^{N \times D_1}$ and $\boldsymbol{X}_2 \in \mathbb{R}^{N \times D_2}$ that are informative of different targets $\boldsymbol{Y}_1 \in \mathbb{R}^{N \times C_1}$ and $\boldsymbol{Y}_2 \in \mathbb{R}^{N \times C_2}$.

Definition (σ-informativeness)

Data $\boldsymbol{X} \in \mathbb{R}^{N \times D}$ is σ-informative ($0 \leqslant \sigma \leqslant 1$) with respect to the targets $\boldsymbol{Y} \in \mathbb{R}^{N \times C}$ if $\|\hat{\boldsymbol{Y}} - \boldsymbol{Y}\|_F^2 = NC(1-\sigma)$, where $\hat{\boldsymbol{Y}}$ is the prediction of the ordinary least squares (OLS) estimator.

Definition (Multi-view non-redundancy)

Let the data modality $\boldsymbol{X}_i$ be σ_{ij}-informative of targets $\boldsymbol{Y}_j$. For $i \neq j$, $\sigma_{ii} > \sigma_{ji}$.

Multimodal setting & the multi-view non-redundancy assumption

Let us consider two modalities $\boldsymbol{X}_1 \in \mathbb{R}^{N \times D_1}$ and $\boldsymbol{X}_2 \in \mathbb{R}^{N \times D_2}$ that are informative of different targets $\boldsymbol{Y}_1 \in \mathbb{R}^{N \times C_1}$ and $\boldsymbol{Y}_2 \in \mathbb{R}^{N \times C_2}$.

Definition (σ-informativeness)

Data $\boldsymbol{X} \in \mathbb{R}^{N \times D}$ is σ-informative ($0 \leqslant \sigma \leqslant 1$) with respect to the targets $\boldsymbol{Y} \in \mathbb{R}^{N \times C}$ if $\|\hat{\boldsymbol{Y}} - \boldsymbol{Y}\|_F^2 = NC(1-\sigma)$, where $\hat{\boldsymbol{Y}}$ is the prediction of the ordinary least squares (OLS) estimator.

Definition (Multi-view non-redundancy)

Let the data modality $\boldsymbol{X}_i$ be σ_{ij}-informative of targets $\boldsymbol{Y}_j$. For $i \neq j$, $\sigma_{ii} > \sigma_{ji}$.

Multimodal setting & the multi-view non-redundancy assumption

Let us consider two modalities $\boldsymbol{X}_1 \in \mathbb{R}^{N \times D_1}$ and $\boldsymbol{X}_2 \in \mathbb{R}^{N \times D_2}$ that are informative of different targets $\boldsymbol{Y}_1 \in \mathbb{R}^{N \times C_1}$ and $\boldsymbol{Y}_2 \in \mathbb{R}^{N \times C_2}$.

Definition (σ-informativeness)

Data $\boldsymbol{X} \in \mathbb{R}^{N \times D}$ is σ-informative ($0 \leqslant \sigma \leqslant 1$) with respect to the targets $\boldsymbol{Y} \in \mathbb{R}^{N \times C}$ if $\|\hat{\boldsymbol{Y}} - \boldsymbol{Y}\|_F^2 = NC(1-\sigma)$, where $\hat{\boldsymbol{Y}}$ is the prediction of the ordinary least squares (OLS) estimator.

Definition (Multi-view non-redundancy)

Let the data modality $\boldsymbol{X}_i$ be σ_{ij}-informative of targets $\boldsymbol{Y}_j$. For $i \neq j$, $\sigma_{ii} > \sigma_{ji}$.

A theoretical framework to investigate representation learning by alignement

We will build intuition in a linear regime where we consider alignment as a regression task:

$$\min_{\boldsymbol{V}_1,\boldsymbol{V}_2,\boldsymbol{W}_1,\boldsymbol{W}_2,\boldsymbol{Q}_1} \underbrace{\|\boldsymbol{X}_1\boldsymbol{V}_1\boldsymbol{W}_1-\boldsymbol{Y}_1\|_F^2}_{\text{prediction loss 1}}+\underbrace{\|\boldsymbol{X}_2\boldsymbol{V}_2\boldsymbol{W}_2-\boldsymbol{Y}_2\|_F^2}_{\text{prediction loss 2}}+\lambda\underbrace{\|\boldsymbol{X}_1\boldsymbol{V}_1\boldsymbol{Q}_1-\boldsymbol{X}_2\boldsymbol{V}_2\|_F^2}_{\text{alignment loss}} \quad (1)$$

$$\text{subject to } \boldsymbol{V}_2,\boldsymbol{W}_2=\arg\min_{\boldsymbol{V},\boldsymbol{W}}\|\boldsymbol{X}_2\boldsymbol{V}\boldsymbol{W}-\boldsymbol{Y}_2\|_F^2$$

where $\lambda > 0$ controls the trade-off between the alignment and the prediction losses, and

- $\boldsymbol{V}_i \in \mathbb{R}^{D_i \times K}$ are linear encoders that compute the latent representations $\boldsymbol{Z}_i = \boldsymbol{X}_i\boldsymbol{V}_i$, where K is the dimension of the latent space,
- $\boldsymbol{W}_i \in \mathbb{R}^{K \times C_i}$ map the representations to the targets $\boldsymbol{Y}_i$,
- $\boldsymbol{Q}_1 \in \mathbb{R}^{K \times K}$ maps the representations of modality 1 to the representations of modality 2.

Technical translation of our research question

Can we jointly minimize the prediction losses and the alignment loss? In other words, are optimal model parameters independent of λ?

A theoretical framework to investigate representation learning by alignement

We will build intuition in a linear regime where we consider alignment as a regression task:

$$\min_{\boldsymbol{V}_1,\boldsymbol{V}_2,\boldsymbol{W}_1,\boldsymbol{W}_2,\boldsymbol{Q}_1} \underbrace{\|\boldsymbol{X}_1\boldsymbol{V}_1\boldsymbol{W}_1-\boldsymbol{Y}_1\|_F^2}_{\text{prediction loss 1}}+\underbrace{\|\boldsymbol{X}_2\boldsymbol{V}_2\boldsymbol{W}_2-\boldsymbol{Y}_2\|_F^2}_{\text{prediction loss 2}}+\lambda\underbrace{\|\boldsymbol{X}_1\boldsymbol{V}_1\boldsymbol{Q}_1-\boldsymbol{X}_2\boldsymbol{V}_2\|_F^2}_{\text{alignment loss}} \quad (1)$$

$$\text{subject to } \boldsymbol{V}_2,\boldsymbol{W}_2=\arg\min_{\boldsymbol{V},\boldsymbol{W}}\|\boldsymbol{X}_2\boldsymbol{V}\boldsymbol{W}-\boldsymbol{Y}_2\|_F^2$$

where $\lambda > 0$ controls the trade-off between the alignment and the prediction losses, and

- $\boldsymbol{V}_i \in \mathbb{R}^{D_i \times K}$ are linear encoders that compute the latent representations $\boldsymbol{Z}_i = \boldsymbol{X}_i\boldsymbol{V}_i$, where K is the dimension of the latent space,
- $\boldsymbol{W}_i \in \mathbb{R}^{K \times C_i}$ map the representations to the targets $\boldsymbol{Y}_i$,
- $\boldsymbol{Q}_1 \in \mathbb{R}^{K \times K}$ maps the representations of modality 1 to the representations of modality 2.

Technical translation of our research question

Can we jointly minimize the prediction losses and the alignment loss? In other words, are optimal model parameters independent of λ?

Information loss by alignment

Theorem (Information loss by alignment)

Let us consider that data $\boldsymbol{X}_i$ is σ_{ij}-informative of targets $\boldsymbol{Y}_j$, and that $\tilde{Y}_{iK}$ is σ_{iK}-informative of $\boldsymbol{Y}_i$, where $\tilde{Y}_{iK}$ is the projection of $\boldsymbol{Y}_i$ on the top-K eigenspaces of $\boldsymbol{Y}_i^T \boldsymbol{Y}_i$.

Under the assumptions that i) $\sigma_{22} \geqslant \sigma_{2K}$, and ii) $\sigma_{21} < \sigma_{1K}$, then the solution $\boldsymbol{V}_1^*$ of problem (1) does not minimize the prediction loss 1 neither the alignment loss. Furthermore, representations $\boldsymbol{Z}_1$ are σ_{11}^z-informative of $\boldsymbol{Y}_1$, with $\sigma_{11}^z < \sigma_{11}$.

- When $\boldsymbol{X}_2$ is informative enough of $\boldsymbol{Y}_2$ while barely informative of $\boldsymbol{Y}_1$, alignment leads to a fundamental trade-off between the two downstream tasks.
- Task-relevant modality-specific information is lost at the expense of alignment.

Experimental framework

Our experimental design aims to assess the validity of our theoretical results in a non-linear regime:

— We train non-linear encoders f_θ^1 and f_θ^2:

$$\min_{\theta_1,\theta_2} \|f_\theta(\boldsymbol{X}_1)\boldsymbol{W}_1 - \boldsymbol{Y}_1\|_F^2 + \|f_\theta(\boldsymbol{X}_2)\boldsymbol{W}_2 - \boldsymbol{Y}_2\|_F^2 + \lambda\|f_\theta(\boldsymbol{X}_1)\boldsymbol{Q}_1 - f_\theta(\boldsymbol{X}_2)\|_F^2,$$

— for several values of the coefficient $\lambda \in [0, 1]$ that controls the trade-off between the prediction and alignement losses,

— We keep linear classification heads (which is standard [Caron et al., 2021]).

Preliminary experiments in a controlled setting

Experiments with non-linear encoders on a synthetic data set inspired by [Hermann and Lampinen, 2020, Dufumier et al., 2025].

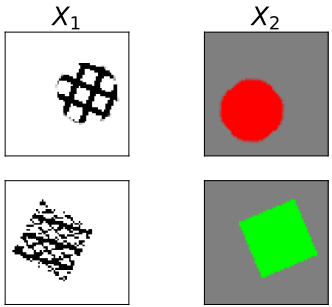


Figure: Three generative factors – shape, texture and color – control the data generation across two modalities: $Y_1 =$ [shape texture], $Y_2 =$ [shape color].

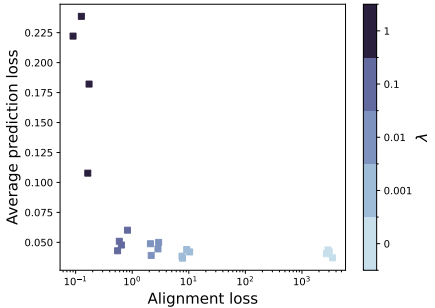


Figure: Prediction VS alignment loss for several values of the coefficient $\lambda \in [0, 1]$ and random network initializations. The lighter the squares, the smaller λ.

Preliminary experiments on real remote sensing data

Experiments with non-linear encoders (a slightly modified) OmniSat architecture on the TreeSatAI-Time-Series dataset [Astruc et al., 2024].

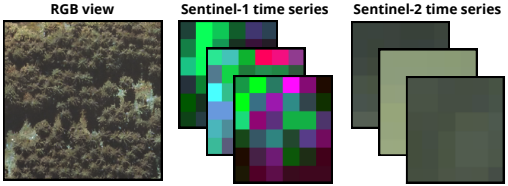


Figure: Random samples of the TreeSatAI-TS dataset.

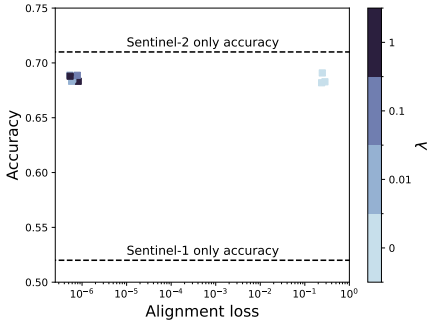


Figure: Train accuracy VS alignment loss for several values of the coefficient λ and random network initializations. Horizontal lines denote the accuracy reached by models trained on a single modality.

Limitations & perspectives

▸ We provide theoretical and empirical evidence that multimodal representation learning by alignment results in a loss of modality-specific information when modalities are non-redundant.

Limitations

— Our theoretical insights are limited to the linear setting.

— We modeled alignment as a regression task:

- $\neq$ NCE loss [Oord et al., 2018] used by contrastive methods,
- informative of multimodal JEPA?

Perspectives

— Which modality-specific features are lost / retained by alignment?

— Should we consider other definitions of alignment, *e.g.* implicit alignment [Shukor and Cord, 2024]?

— Under the *multi-view non-redundancy* assumption, is the disentanglement of modality-generic and modality-specific features possible?

Limitations & perspectives

▸ We provide theoretical and empirical evidence that multimodal representation learning by alignment results in a loss of modality-specific information when modalities are non-redundant.

Limitations

— Our theoretical insights are limited to the linear setting.

— We modeled alignment as a regression task:

- $\neq$ NCE loss [Oord et al., 2018] used by contrastive methods,
- informative of multimodal JEPA?

Perspectives

— Which modality-specific features are lost / retained by alignment?

— Should we consider other definitions of alignment, *e.g.* implicit alignment [Shukor and Cord, 2024]?

— Under the *multi-view non-redundancy* assumption, is the disentanglement of modality-generic and modality-specific features possible?

Limitations & perspectives

▸ We provide theoretical and empirical evidence that multimodal representation learning by alignment results in a loss of modality-specific information when modalities are non-redundant.

Limitations

— Our theoretical insights are limited to the linear setting.

— We modeled alignment as a regression task:

- $\neq$ NCE loss [Oord et al., 2018] used by contrastive methods,
- informative of multimodal JEPA?

Perspectives

— Which modality-specific features are lost / retained by alignment?

— Should we consider other definitions of alignment, *e.g.* implicit alignment [Shukor and Cord, 2024]?

— Under the *multi-view non-redundancy* assumption, is the disentanglement of modality-generic and modality-specific features possible?

Thank you for your attention!

Questions?

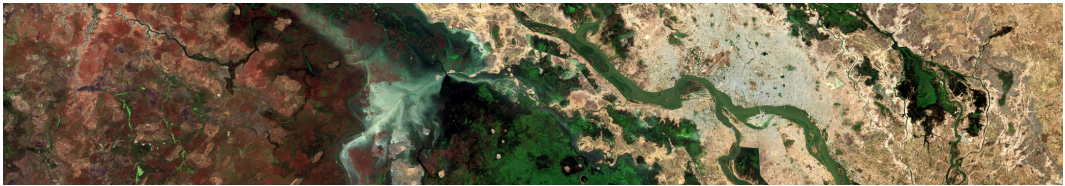


Figure: Credit: ESA, modified Sentinel data (2024), processed by ESA.

Code and data: `https://github.com/Romain3Ch216/alg_maclean_25`

Theorem (Solutions to the learning problem (1))

The objective function of the problem (1) *is minimized for* $\boldsymbol{V}_1^*, \boldsymbol{V}_2^*, \boldsymbol{W}_1^*, \boldsymbol{W}_2^*, \boldsymbol{Q}_1^*$ *such that:*

$$\text{span}(\boldsymbol{V}_1^*) = \text{span}\big(\boldsymbol{P}_{\boldsymbol{X}_1^T\boldsymbol{X}_1}\boldsymbol{D}_{\boldsymbol{X}_1^T\boldsymbol{X}_1}^{-\frac{1}{2}}(\boldsymbol{P}_{\boldsymbol{H}_1})_{\cdot,1:K}\big)$$

$$\text{span}(\boldsymbol{V}_2^*) = \text{span}\big(\boldsymbol{P}_{\boldsymbol{X}_2^T\boldsymbol{X}_2}\boldsymbol{D}_{\boldsymbol{X}_2^T\boldsymbol{X}_2}^{-\frac{1}{2}}(\boldsymbol{P}_{\boldsymbol{H}_2})_{\cdot,1:K}\big)$$

$$\boldsymbol{W}_1^* = (\boldsymbol{V}_1^{*T}\boldsymbol{X}_1^T\boldsymbol{X}_1\boldsymbol{V}_1^*)^{-1}\boldsymbol{V}_1^{*T}\boldsymbol{X}_1^T\boldsymbol{Y}_1$$

$$\boldsymbol{W}_2^* = (\boldsymbol{V}_2^{*T}\boldsymbol{X}_2^T\boldsymbol{X}_2\boldsymbol{V}_2^*)^{-1}\boldsymbol{V}_2^{*T}\boldsymbol{X}_2^T\boldsymbol{Y}_2$$

$$\boldsymbol{Q}_1^* = (\boldsymbol{V}_1^{*T}\boldsymbol{X}_1^T\boldsymbol{X}_1\boldsymbol{V}_1^*)^{-1}\boldsymbol{V}_1^{*T}\boldsymbol{X}_1^T\boldsymbol{X}_2\boldsymbol{V}_2^*$$

where $\text{span}(\boldsymbol{M})$ stands for the linear subspace generated by the columns of $\boldsymbol{M}$, and:

$$\begin{cases} \boldsymbol{H}_1 := \boldsymbol{D}_{\boldsymbol{X}_1^T\boldsymbol{X}_1}^{-\frac{1}{2}}\boldsymbol{P}_{\boldsymbol{X}_1^T\boldsymbol{X}_1}^T\boldsymbol{A}\boldsymbol{P}_{\boldsymbol{X}_1^T\boldsymbol{X}_1}\boldsymbol{D}_{\boldsymbol{X}_1^T\boldsymbol{X}_1}^{-\frac{1}{2}} \\ \boldsymbol{A} := \boldsymbol{X}_1^T(\boldsymbol{Y}_1\boldsymbol{Y}_1^T + \lambda\boldsymbol{Z}_2\boldsymbol{Z}_2^T)\boldsymbol{X}_1 \end{cases} \qquad \begin{cases} \boldsymbol{Z}_2 := \boldsymbol{X}_2\boldsymbol{V}_2^* \\ \boldsymbol{P}_{\boldsymbol{H}_2} = (\boldsymbol{P}_{\boldsymbol{X}_2\boldsymbol{X}_2^T})_{\cdot,1:D}^T(\boldsymbol{P}_{\boldsymbol{Y}_2\boldsymbol{Y}_2^T})_{\cdot,1:D} \end{cases}$$

Sketch of the proof of Theorem 1 (analog to the proof of Theorem 1 in [Balestriero and LeCun, 2024])

— First, we express optimal parameters $\boldsymbol{W}_1^*$ and $\boldsymbol{Q}_1^*$ as functions of $\boldsymbol{V}_1$.

— Second, plugging $\boldsymbol{W}_1^*$ and $\boldsymbol{Q}_1^*$ into the objective function yields the following optimization problem:

$$\max_{\boldsymbol{V}_1} \quad Tr\big(\boldsymbol{V}_1^T \boldsymbol{X}_1^T (\boldsymbol{Y}_1 \boldsymbol{Y}_1^T + \lambda \boldsymbol{X}_2 \boldsymbol{V}_2 (\boldsymbol{X}_2 \boldsymbol{V}_2)^T) \boldsymbol{X}_1 \boldsymbol{V}_1\big) \tag{2}$$
$$\text{subject to } \boldsymbol{V}_1^T \boldsymbol{X}_1^T \boldsymbol{X}_1 \boldsymbol{V}_1 = \boldsymbol{I}_K$$

— Third, we solve problem (2) by solving the following generalized eigenvalue problem:

$$\boldsymbol{A}\boldsymbol{V}_1' = \boldsymbol{B}\boldsymbol{V}_1'\boldsymbol{\Lambda} \tag{3}$$

where $\boldsymbol{A} = \boldsymbol{X}_1^T (\boldsymbol{Y}_1 \boldsymbol{Y}_1^T + \lambda \boldsymbol{X}_2 \boldsymbol{V}_2 (\boldsymbol{X}_2 \boldsymbol{V}_2)^T) \boldsymbol{X}_1$, $\boldsymbol{B} = \boldsymbol{X}_1^T \boldsymbol{X}_1$, and the columns of $\boldsymbol{V}_1' \in \mathbb{R}^{D \times D}$ are the eigenvectors of $\boldsymbol{A}$ and the diagonal elements of $\boldsymbol{\Lambda}$ its eigenvalues. Then, we can simply take the top K eigenvectors of V_1' to build $\boldsymbol{V}_1$, *i.e.* $\boldsymbol{V}_1 = (\boldsymbol{V}_1')_{\cdot,1:K}$.

Sketch of the proof of Theorem 2

The idea is to show that if $\boldsymbol{V}_1^\star$ is the optimal solution of the prediction task 1 and the alignment task, then $\sigma_{21} \geqslant \sigma_{1K}$, which contradicts assumption ii). The proof is divided into five steps:

1. We assume that $\boldsymbol{V}_1^\star$ is the optimal solution of the prediction task 1 and the alignment task, and show that, as a result, the intersection of the top-K eigenspaces of $\boldsymbol{Z}_2\boldsymbol{Z}_2^T$ and $\boldsymbol{Y}_1\boldsymbol{Y}_1^T$ is of dimension K,
2. We show that $\sigma_{22} \geqslant \sigma_{2K}$ implies that the intersection of the top-K eigenspaces of $\boldsymbol{X}_2\boldsymbol{X}_2^T$ and $\boldsymbol{Y}_2\boldsymbol{Y}_2^T$ is of dimension K,
3. From step 2, we show that the intersection of the top-K eigenspaces of $\boldsymbol{Z}_2\boldsymbol{Z}_2^T$ and $\boldsymbol{X}_2\boldsymbol{X}_2^T$ is of dimension K,
4. From step 1 and step 3, we deduce that the intersection of the top-K eigenspaces of $\boldsymbol{Y}_1\boldsymbol{Y}_1^T$ and $\boldsymbol{X}_2\boldsymbol{X}_2^T$ is of dimension K.
5. Finally, we show that this implies that $\sigma_{21} \geqslant \sigma_{1K}$, which contradicts assumption ii).

Related work: theoretical analysis using information theory

Definition (Multi-view redundancy)

$\exists \epsilon > 0$ such that $I(X_1; Y|X_2) \leqslant \epsilon$ and $I(X_2; Y|X_1) \leqslant \epsilon$.

Definition (Multi-view non-redundancy)

$\exists \epsilon > 0$ such that $I(X_1; Y|X_2) > \epsilon$ or $I(X_2; Y|X_1) > \epsilon$.

Theorem (Suboptimality of standard CL [Liang et al., 2023])

When there is multi-view non-redundancy, given optimal representations $\{Z_1, Z_2\}$ *that satisfy* $I(Z_1; Y|X_2) = I(Z_2; Y|X_1) = 0$*, we have that*

$$I(Z_1, Z_2; Y) < I(X_1, X_2; Y).$$

Theorem ([Dufumier et al., 2025])

Under the multiview redundancy assumption, cross-modal contrastive learning methods are limited to only learn the redundant information.

Assran, M., Duval, Q., Misra, I., Bojanowski, P., Vincent, P., Rabbat, M., LeCun, Y., and Ballas, N. (2023).
Self-supervised learning from images with a joint-embedding predictive architecture.
In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15619–15629.

Astruc, G., Gonthier, N., Mallet, C., and Landrieu, L. (2024).
OmniSat: Self-supervised modality fusion for Earth observation.
ECCV.

Astruc, G., Gonthier, N., Mallet, C., and Landrieu, L. (2025).
Anysat: One earth observation model for many resolutions, scales, and modalities.
In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 19530–19540.

Audebert, N., Le Saux, B., and Lefèvre, S. (2018).
Beyond rgb: Very high resolution urban remote sensing with multimodal deep networks.
ISPRS journal of photogrammetry and remote sensing, 140:20–32.

Balestriero, R. and LeCun, Y. (2024).
How learning by reconstruction produces uninformative features for perception.
In *Proceedings of the 41st International Conference on Machine Learning*, pages 2566–2585.

Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., and Joulin, A. (2021).
Emerging properties in self-supervised vision transformers.
In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9650–9660.

Chen, T., Kornblith, S., Norouzi, M., and Hinton, G. (2020).
A simple framework for contrastive learning of visual representations.
In *International conference on machine learning*, pages 1597–1607. PmLR.

Dufumier, B., Castillo-Navarro, J., Tuia, D., and Thiran, J.-P. (2025).
What to align in multimodal contrastive learning?

Englhart, S., Keuck, V., and Siegert, F. (2011).
Aboveground biomass retrieval in tropical forests—the potential of combined x-and l-band sar data use.
Remote sensing of environment, 115(5):1260–1271.

Feng, Z., Song, L., Yang, S., Zhang, X., and Jiao, L. (2023).
Cross-modal contrastive learning for remote sensing image classification.
IEEE Transactions on Geoscience and Remote Sensing, 61:1–13.

Hermann, K. and Lampinen, A. (2020).
What shapes feature representations? exploring datasets, architectures, and training.
Advances in Neural Information Processing Systems, 33:9995–10006.

Jain, U., Wilson, A., and Gulshan, V. (2022).
Multimodal contrastive learning for remote sensing tasks.
arXiv preprint arXiv:2209.02329.

Liang, P. P., Deng, Z., Ma, M. Q., Zou, J. Y., Morency, L.-P., and Salakhutdinov, R. (2023).
Factorized contrastive learning: Going beyond multi-view redundancy.
Advances in Neural Information Processing Systems, 36:32971–32998.

Marsocci, V. and Audebert, N. (2025).
Cross-sensor self-supervised training and alignment for remote sensing.
IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing.


Miraglio, T., Coops, N. C., Wallis, C. I., Crofts, A. L., Kalacska, M., Vellend, M., Serbin, S. P., Arroyo-Mora, J. P., and Laliberté, E. (2023).
Mapping canopy traits over quebec using airborne and spaceborne imaging spectroscopy.
Scientific Reports, 13(1):17179.


Oord, A. v. d., Li, Y., and Vinyals, O. (2018).
Representation learning with contrastive predictive coding.
arXiv preprint arXiv:1807.03748.


Penot, V. and Merlin, O. (2023).
Estimation du stress hydrique par télédétection spatiale.


Pielawski, N., Wetzer, E., Öfverstedt, J., Lu, J., Wählby, C., Lindblad, J., and Sladoje, N. (2020).
Comir: Contrastive multimodal image representation for registration.
Advances in neural information processing systems, 33:18433–18444.


Prexl, J. and Schmitt, M. (2023).
Multi-modal multi-objective contrastive learning for sentinel-1/2 imagery.
In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2136–2144.


Scheibenreif, L., Mommert, M., and Borth, D. (2022).
Contrastive self-supervised data fusion for satellite imagery.
ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences, 3:705–711.


Shukor, M. and Cord, M. (2024).
Implicit multimodal alignment: On the generalization of frozen llms to multimodal inputs.
In *Advances in Neural Information Processing Systems (NeurIPS)*.


Tian, Y., Sun, C., Poole, B., Krishnan, D., Schmid, C., and Isola, P. (2020).
What makes for good views for contrastive learning?
Advances in neural information processing systems, 33:6827–6839.