This CVPR paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

Beyond What's Shared: Recovering Lost Unique Information from Intermediate Layers to Boost Multimodal Geo-Foundation Models

JangHyeon Lee[1,2] Philipe Ambrozio Dias[2] Yao-Yi Chiang[1] Dalton Lunga[2]
[1]University of Minnesota [2]Oak Ridge National Laboratory

Abstract

Learning general-purpose representations of geographic locations has become essential to geospatial tasks such as population estimation and environmental monitoring. To obtain such representations, multimodal geo-foundation models often use contrastive learning (CL) to align satellite imagery with geo-coordinates, implicitly assuming that cross-modal (shared) information suffices for downstream tasks. However, given the breadth of tasks, task-relevant information may lie beyond the shared space, so retaining modality-specific (unique) features can improve task performance. Prior methods retain unique information through extra training objectives or external models, increasing training complexity. Motivated by the conventional wisdom that earlier layers capture general input features while later layers become task-specific, we hypothesize that intermediate layers in CL models retain more modality-specific structure than the alignment-optimized final layer. Through a trifecta layerwise analysis of modality gap, representation similarity, and mutual information, we validate this trend and find that fusing intermediate (more unique) and final (more shared) representations yields consistent gains on diverse geospatial tasks. Our findings reveal underutilized information diversity in CL models and show that simple layerwise fusion is an efficient path to richer geo-embeddings.

1. Introduction

Learning general-purpose representations of geographic locations is key to modern geospatial analysis [8, 50]. Such representations support a range of tasks including population estimation [2] and environmental monitoring [38, 77]. Recent geo-foundation models, such as GeoCLIP [67] and SatCLIP [27], learn these representations by contrastively aligning satellite imagery with paired geo-coordinates. Through contrastive learning [9, 47], these models learn a shared embedding space that binds location and visual information, enabling broad downstream transfer [24, 40].

However, the underlying assumption in multimodal contrastive learning (CL) warrants closer examination (Fig. 2). CL relies on the *multi-view redundancy* assumption [35, 64], which states that all information relevant to downstream tasks is **shared** between modalities. Under this assumption, maximizing the mutual information between the two modalities is sufficient for learning an optimal representation [54]. As a result, any information that exists only in one modality (**unique**) is considered as *task-irrelevant*.

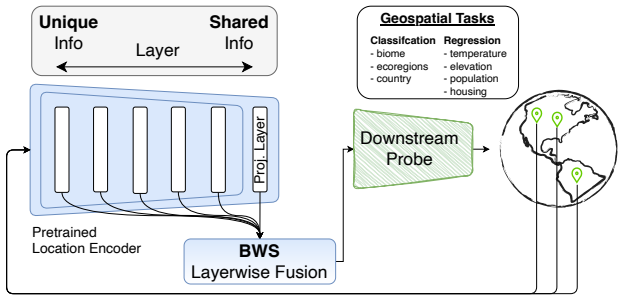


Figure 1. Intermediate layers of pretrained contrastive encoders retain modality-specific (unique) information that becomes less prominent in the final layer of contrastive models. *Beyond What's Shared (BWS)* fuses this underutilized layerwise information with the final representation to exploit the best of both worlds, yielding richer geo-embeddings and consistent gains across tasks.

Yet, given the diversity of geospatial tasks, the multi-view redundancy assumption can be restrictive. For example, some tasks depend on modality-specific information (e.g., spatial context available only from coordinates or surface characteristics visible only from imagery) that is not shared between modalities. Hence, contrastive models may be suboptimal when task-relevant information is not fully shared [54, 68]. This limitation has motivated growing interest in developing contrastive methods that can simultaneously retain unique information [13, 35, 43, 68].

Early work retained unique information in contrastive learning by adding a regularization term that increases the mutual information between the learned representation and its input, preserving modality-specific features [68]. Subsequent methods decomposed multimodal representations into shared and unique components with joint op-

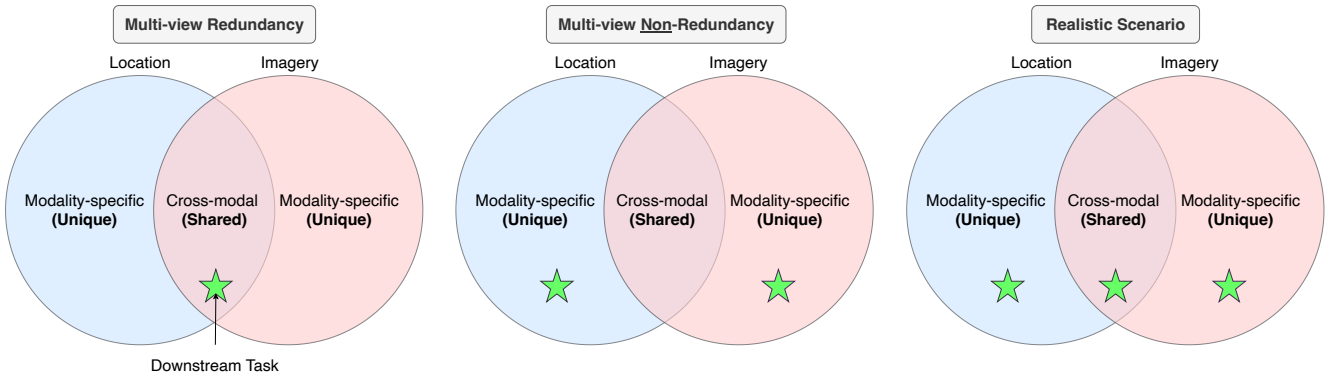


Figure 2. **Decomposition of downstream task-relevant information across modalities. (Left)** Under the *multi-view redundancy* assumption, all task-relevant information lies in the shared (cross-modal) region between modalities. In this case, contrastive learning (CL) is optimal, as it maximizes the lower bound of mutual information between the two modalities. **(Center)** Under the *multi-view non-redundancy* setting, the task-relevant information lies outside the shared region, making CL suboptimal. **(Right)** Given the diversity of geospatial downstream tasks, both cross-modal (shared) and modality-specific (unique) information may be task-relevant.

timization through mutual information bounds [35]. Recently, contrastive representations have been augmented with modality-specific features retrieved from a precomputed feature database of external model embeddings [13].

While these methods can retain modality-specific (unique) information for contrastive models, they introduce additional training complexity and computational overhead. Extra training objectives and mutual information terms increase computational cost and are harder to optimize, while reliance on external models for precomputing databases adds further complexity. This raises a natural question: *can we retain unique information in contrastive models without extra objectives or external components?*

We approach this question by combining insights that neural networks learn hierarchical representations across layers (from general to task-specific) [25, 66, 74, 75] with the Information Bottleneck (IB) theory [26, 55]. From the IB perspective, this layerwise hierarchy arises as the model compresses input information, retaining only what is necessary for the training objective (which, in self-supervised settings, corresponds to the pretext task [5]). Although this layerwise evolution has been observed in unimodal models, its behavior in multimodal contrastive models is underexplored. We hypothesize that multimodal contrastive models also exhibit an information bottleneck: *later layers specialize toward the alignment objective and thus emphasize shared information, while earlier layers, less constrained by this objective, retain more modality-specific structure.*

We test this hypothesis through a layerwise analysis of state-of-the-art multimodal contrastive geo-foundation models. To characterize how modality-specific (unique) and cross-modal (shared) information evolve across depth, we conduct a trifecta of layerwise analyses integrating **(1) geometric, (2) correlational,** and **(3) information-theoretic** perspectives. This multi-perspective framework consistently reveals that later layers concentrate shared information, while intermediate layers retain modality-specific features. Building on this insight, we introduce **Beyond What's Shared (BWS)**, a simple yet effective approach to capture both shared and unique information across layers.

Instead of relying on the final layer for downstream tasks, as is standard for CLIP variants [47], we fuse intermediate (less shared) and final (more shared) representations to produce information-rich priors that better support downstream tasks. Importantly, BWS introduces no auxiliary objectives or extra feature extractors, yet consistently improves performance across diverse geospatial tasks, revealing that contrastive models contain underutilized task-relevant information hidden in their intermediate layers. Our results offer a simple and efficient perspective on retaining modality-specific (unique) information and suggest a general strategy for boosting CL models to a wider variety of downstream tasks. We summarize our contributions:

- **C1** We analyze multimodal contrastive geo-foundation models through the lens of the information bottleneck, investigating whether modality-specific (unique) information can be retained without extra training objectives or external models (§4.1).
- **C2** We conduct a three-way layerwise information analysis, examining representations from *geometric, correlational*, and *information-theoretic* perspectives to identify how cross-modal (shared) and modality-specific (unique) information evolve across network depth (§4.2).
- **C3** By leveraging modality-specific information from intermediate layers, we show that combining intermediate and final representations consistently boosts state-of-the-art geo-foundation models across diverse geospatial tasks, all without auxiliary objectives or extra models (§5.2).

2. Preliminaries

We first revisit how existing vision-location geo-foundation models learn to align satellite imagery and geo-coordinates in a shared embedding space [27, 39, 67].

During pretraining, given N paired samples $\{(\mathbf{I}_i, \mathbf{c}_i)\}_{i=1}^N$, each satellite image $\mathbf{I}_i \in \mathbb{R}^{m \times n \times b}$ (with b spectral bands) is paired with geographic coordinates $\mathbf{c}_i = [\lambda_i, \phi_i]$, where λ_i and ϕ_i denote longitude and latitude. The image encoder f_I and location encoder f_c map their respective inputs into a shared d-dimensional embedding space,

$$f_I : \mathbb{R}^{m \times n \times b} \to \mathbb{R}^d, \qquad f_c : \mathbb{S}^2 \to \mathbb{R}^d, \tag{1}$$

where $\mathbb{S}^2$ represents the Earth's surface. The normalized embeddings $\mathbf{z}_i^I = f_I(\mathbf{I}_i)/\|f_I(\mathbf{I}_i)\|_2$ and $\mathbf{z}_i^c = f_c(\mathbf{c}_i)/\|f_c(\mathbf{c}_i)\|_2$ are aligned using the InfoNCE loss [44]:

$$\mathcal{L}_{\text{CL}} = \frac{1}{2N} \sum_{i=1}^{N} \Bigg[-\log \frac{\exp(\langle \mathbf{z}_i^c, \mathbf{z}_i^I \rangle / \tau)}{\sum_{j=1}^{N} \exp(\langle \mathbf{z}_i^c, \mathbf{z}_j^I \rangle / \tau)} - \log \frac{\exp(\langle \mathbf{z}_i^I, \mathbf{z}_i^c \rangle / \tau)}{\sum_{j=1}^{N} \exp(\langle \mathbf{z}_i^I, \mathbf{z}_j^c \rangle / \tau)} \Bigg], \tag{2}$$

where $\langle \cdot, \cdot \rangle$ denotes cosine similarity and τ is a temperature. This loss is applied only to the final-layer output of each encoder [9, 47]. By maximizing similarity between image and location pairs, this objective preserves shared information while imposing no condition on unique features [4, 68].

After pretraining, the aligned encoders are used to generate embeddings for downstream tasks (e.g., population estimation) [27, 39, 67]. Under the *multi-view redundancy* assumption [35, 64], all task-relevant information is presumed to be shared between modalities. Yet, given the diversity of downstream tasks, task-relevant information may also reside outside the shared embedding space.

Our goal is to retain both shared and unique information to ensure generality across tasks, while doing so in a simple and efficient way that introduces no extra components.

3. Related Work

3.1. Multimodal Contrastive Learning

Multimodal contrastive learning [9, 47] aligns paired samples from different modalities by maximizing their mutual information in a shared embedding space. This approach relies on the *multi-view redundancy* assumption [35, 64], which posits that the shared information between modalities is both necessary and sufficient for downstream tasks [58]. However, given the diversity of real-world tasks, this assumption limits the broad applicability of multimodal representation learning. Many downstream settings instead exhibit *multi-view non-redundancy*, where each modality contains task-relevant unique information that is not shared with others [35, 54, 68]. In such *multi-view non-redundant* cases, contrastive models become inherently suboptimal.

Recent works have begun addressing this limitation in contrastive learning. Intra-modality regularization has been used to reinforce the relationship between each representation and its own input, thereby preserving unique information [68]. Other approaches jointly enforce intra- and inter-modality learning by separating shared and modality-specific features through correlation-based redundancy reduction [70] or by factorizing multimodal representations via mutual information optimization [15, 35].

However, these methods rely on complex mutual information estimation and carefully designed data augmentations, which increase training complexity and sensitivity to estimator quality and augmentation strategies. This motivates the need for an efficient and effective approach that can preserve both shared and unique information.

3.2. Vision-Location Geo-Foundation Models

Vision-location geo-foundation models use contrastive learning (CL) [47] to align geographic coordinates with visual features and produce general purpose geo-embeddings. CSP [39] is an early approach that separately encodes geo-tagged images and coordinates and aligns them in a shared embedding space to capture spatial relationships from unlabeled data. GeoCLIP [67] builds on this idea by treating geo-localization as an image-to-GPS retrieval task, aligning images with continuous coordinate representations rather than discrete spatial cells. SatCLIP [27] further scales contrastive learning to global satellite imagery by aligning a pretrained Sentinel-2 vision encoder [59, 69] with a spherical harmonics location encoder [51], capturing visual-spatial representations of the Earth's surface.

However, such vision-location geo-foundation models compress information into a shared embedding space, where modality-specific features are less represented [54, 68]. This can be suboptimal when downstream tasks depend on information not shared across modalities [35]. RANGE [13] addresses this by retrieving semantically similar image features from a database precomputed with an external model (i.e., SatMAE [11]). However, it only captures unique visual information, leaving location-specific information unaddressed. Instead, we introduce a simple and efficient approach that overcomes these limitations.

3.3. Layerwise Information in Neural Networks

It is well known that representations within deep neural networks vary across layers [10, 28, 75, 76]. Early studies revealed that lower layers encode general features, while higher layers become specialized to the training objective [55, 62, 74]. Analyses of supervised classifiers show that representations become more linearly separable with depth, reflecting the influence of the task supervision to-

ward the later layers [3]. This increasing task alignment enhances in-distribution tasks but also makes the later layers sensitive to data-specific biases and less transferable under distribution shifts [16, 19, 57], whereas intermediate layers often yield more general features than the last layer [65].

In unimodal self-supervised learning (SSL), a similar layer hierarchy appears. The final projection head is trained to solve the pretext objective by enforcing invariance between augmented views [7, 20, 73]. In doing so, the final projection layer acts as an information bottleneck that filters out input variability unrelated to the pretext task [37, 45]. In contrast, the intermediate or penultimate representations retain more general features and improve downstream performance, including out-of-distribution (OOD) tasks [6, 7].

Whether a similar hierarchical structure exists in multimodal SSL models (e.g., CLIP [47]) is unclear. In CLIP, the InfoNCE loss [44] is applied to the final layer embeddings, optimizing cross-modal similarity between paired data. This objective captures shared information without explicitly preserving modality-specific features. Since earlier layers are not directly constrained by this loss, they may retain modality-specific information from the input that becomes less prominent with depth. We hypothesize that earlier layers encode more modality-specific information, whereas the final layer concentrates shared information. If this holds, combining intermediate and final representations could capture both unique and shared information without additional tasks or extra feature extractors.

4. The Information Bottleneck in Multimodal Contrastive Learning

The information bottleneck (IB) theory [1, 55, 63] provides an intuitive framework for understanding how neural networks progressively compress input information while preserving only what is relevant for the task objective. We formalize this process for multimodal contrastive learning and show that, under the IB principle, later layers become specialized to the alignment objective, whereas earlier layers preserve more modality-specific (unique) information.

4.1. Theoretical Analysis

Contrastive learning as a shared information bottleneck. The traditional IB framework [1, 55, 63] formulates representation learning as a balance between compression and prediction: given an input X, a target Y, and a representation Z, the optimal encoder minimizes

$$\mathcal{L}_{\text{IB}} = I(X;Z) - \beta\, I(Z;Y), \tag{3}$$

where $I(\cdot;\cdot)$ denotes mutual information and β controls the trade-off between compression and prediction. The first term enforces compression of input information and the second preserves task-relevant information through the mutual information between Z and Y.

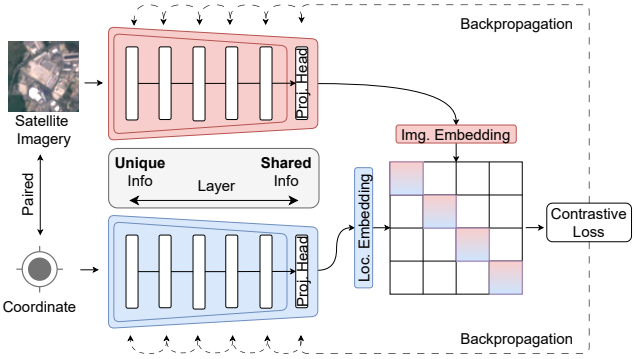


Figure 3. **The information bottleneck in multimodal contrastive models.** In multimodal contrastive learning, information evolves layer by layer, transitioning from modality-specific (unique) to cross-modal (shared) across the network. Intermediate layers retain unique information from the input that becomes less prominent toward the final layer. The projection head (final layer) acts as an information bottleneck that concentrates information relevant to the pretraining objective (cross-modal alignment), potentially underutilizing task-relevant modality-specific structure.

In multimodal contrastive learning, two encoders f_I and f_c produce representations $\mathbf{z}^I = f_I(\mathbf{I})$ and $\mathbf{z}^c = f_c(\mathbf{c})$. Following the multi-view setting [17, 60], each modality acts as supervision for the other. Applying the standard IB principle to this setting, the representation learning process can be interpreted as balancing:

$$\min_{f_I, f_c} \left[I(\mathbf{I};\mathbf{z}^I) + I(\mathbf{c};\mathbf{z}^c) \right] - \beta\, I(\mathbf{z}^I;\mathbf{z}^c), \tag{4}$$

where β controls the trade-off between compression and cross-modal alignment. The first term corresponds to compression of modality-specific inputs, while the second encourages maximizing shared information between the modalities. While InfoNCE (Eq. 2) optimizes the alignment term $I(\mathbf{z}^I;\mathbf{z}^c)$, the compression terms arise implicitly through finite model capacity and the optimization process [45, 55]. This framework offers a useful lens for understanding layerwise information flow: multimodal contrastive training imposes a shared information bottleneck that incentivizes retaining cross-modal structure, while modality-specific information is not explicitly preserved.

Information flow across layers. Eq. 4 defines the information trade-off achieved by the final representations of the two encoders. However, this trade-off is not realized in a single step: deep encoders approximate it progressively through multiple intermediate layers. To capture this hierarchical process, we model each encoder as a Markov chain of latent variables [55].

Let $\mathbf{I}$ and $\mathbf{c}$ denote the paired inputs, and let $\mathbf{z}^{I,(l)} = f_I^{(l)}(\mathbf{I})$ and $\mathbf{z}^{c,(l)} = f_c^{(l)}(\mathbf{c})$ denote the intermediate repre-

sentations at layer l (with L denoting the last layer) of each encoder. Each encoder forms a Markov chain:

$$\begin{aligned}\mathbf{I} \to \mathbf{z}^{I,(1)} \to \mathbf{z}^{I,(2)} \to \cdots \to \mathbf{z}^{I,(L)},\\ \mathbf{c} \to \mathbf{z}^{c,(1)} \to \mathbf{z}^{c,(2)} \to \cdots \to \mathbf{z}^{c,(L)}.\end{aligned} \tag{5}$$

By the Data Processing Inequality (DPI) [12], the mutual information between the input and its representations can only decrease with depth:

$$\begin{aligned}I(\mathbf{I};\mathbf{z}^{I,(1)}) \geq I(\mathbf{I};\mathbf{z}^{I,(2)}) \geq \cdots \geq I(\mathbf{I};\mathbf{z}^{I,(L)}),\\ I(\mathbf{c};\mathbf{z}^{c,(1)}) \geq I(\mathbf{c};\mathbf{z}^{c,(2)}) \geq \cdots \geq I(\mathbf{c};\mathbf{z}^{c,(L)}).\end{aligned} \tag{6}$$

Together with the optimization pressure at the final layer (Eq. 2), which maximizes cross-modal mutual information $I(\mathbf{z}^{I,(L)};\mathbf{z}^{c,(L)})$, this motivates a layerwise information hierarchy. As depth increases, input-specific information decreases (by the DPI) while cross-modal alignment increases (driven by the contrastive objective). We validate this hypothesis empirically in §4.2. Consequently, earlier layers, less constrained by the alignment objective, retain more modality-specific structure, while later layers concentrate on cross-modal (shared) representations (Fig. 3).

4.2. Trifecta Layerwise Representational Analysis

To empirically show how unique and shared information evolve **layer by layer**, we conduct layerwise analyses from three perspectives: *geometric* (modality gap [36]), *correlational* (CKA [29]), and *information-theoretic* (MI [30, 44]).

Each perspective captures a distinct aspect of cross-modal alignment: the geometric view measures spatial separation between modalities, the correlational view captures structural similarity of feature spaces, and the information-theoretic view quantifies the amount of shared information. Together, these views offer a holistic understanding of how unique and shared information are organized across layers.

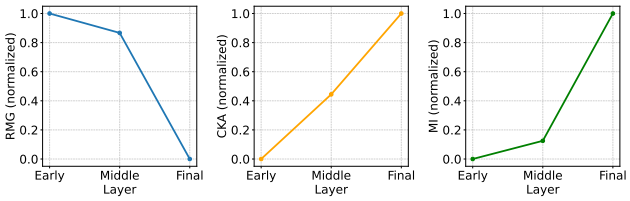


Figure 4. **Layerwise representational analysis. (Left)** The *relative modality gap (RMG)* decreases with depth, indicating that the two modalities become geometrically closer. **(Center)** The *CKA similarity* between the two modalities increases, showing progressively stronger correlation. **(Right)** The *mutual information (MI)* rises across layers, implying stronger shared information.

① **Geometric perspective via relative modality gap.** The modality gap measures the geometric separation between modalities in the shared embedding space [36]. We use the *Relative Modality Gap (RMG)*, which normalizes the cross-modal distance by the average within-modality distances [53]. To analyze layerwise alignment, we fix the final image representations as a reference and compute the RMG between these and each location layer:

$$\text{RMG}^{(l)} = \frac{\frac{1}{N}\sum_i d(\mathbf{z}_i^{I,(L)}, \mathbf{z}_i^{c,(l)})}{\frac{1}{2N(N-1)}\Big(\sum_{i\neq j} d(\mathbf{z}_i^{I,(L)}, \mathbf{z}_j^{I,(L)}) + \sum_{i\neq j} d(\mathbf{z}_i^{c,(l)}, \mathbf{z}_j^{c,(l)})\Big) + \frac{1}{N}\sum_i d(\mathbf{z}_i^{I,(L)}, \mathbf{z}_i^{c,(l)})}. \tag{7}$$

A larger $\text{RMG}^{(l)}$ indicates weaker modality alignment, and a smaller value indicates stronger alignment.

② **Correlational perspective via centered kernel alignment (CKA) similarity.** CKA is a widely used measure of representational similarity in neural networks [29]. To examine how cross-modal alignment evolves across depth, similar to modality gap, we fix the last layer representations of one modality as a reference and compute CKA between each intermediate representation of the other modality and this reference. Larger CKA values indicate greater representational similarity between modalities.

③ **Information-theoretic perspective via mutual information (MI) estimation.** MI measures how much information the representations of one modality contain about the other. To examine how this dependency evolves across depth, we follow the same reference-based comparison used for the modality gap and CKA, and estimate MI with the KSG estimator [30]. Higher MI values indicate stronger cross-modal dependence, and lower values indicate weaker shared information between modalities.

Layerwise trends. We conduct a layerwise analysis on a representative geo-foundation model, SatCLIP [27] (§2). Fig. 4 presents the evolution of geometric, correlational, and information-theoretic similarity across layers, and Fig. 5 visualizes the corresponding embedding distributions via UMAP [41]. Across both quantitative and qualitative analyses, we observe a consistent trend: the modality gap decreases with depth, whereas CKA similarity and mutual information increase. These results support the layerwise information hierarchy described in §4.1, where modality-specific information becomes less prominent with depth as the model focuses on cross-modal alignment. This observation motivates our approach, **BWS** below.

4.3. Fusing Underutilized Layerwise Information

Existing geo-foundation models [27, 39, 67] use the final layer to generate geo-embeddings, which primarily capture cross-modal (shared) information. To leverage the

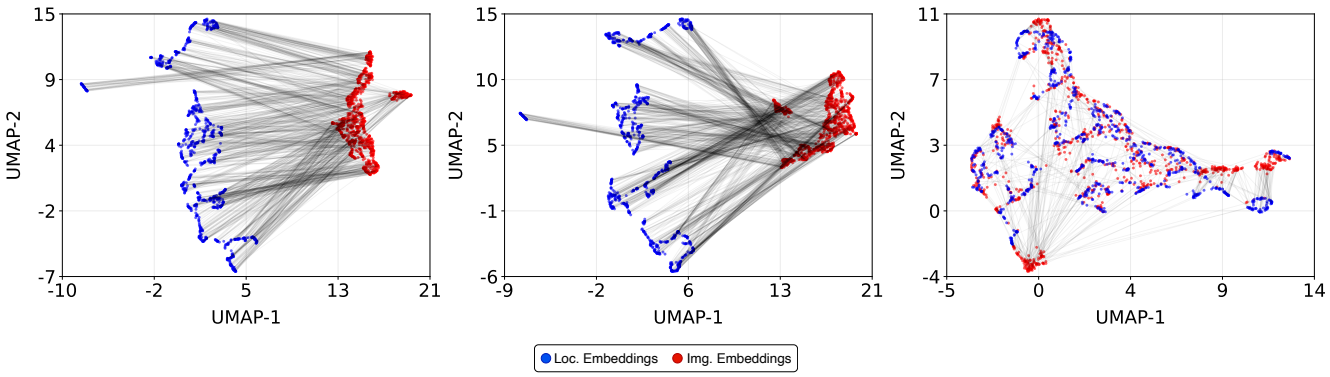


Figure 5. **Layerwise evolution of the modality gap.** UMAP visualizations of location (blue) and image (red) embeddings across network depth. **(Left)** Early layer: a large modality gap indicates weak alignment between modalities, suggesting that representations capture more modality-specific (unique) information. **(Center)** Middle layer: the gap remains clear, implying that cross-modal (shared) structure is still limited. **(Right)** Final layer: the two modalities become closely aligned, indicating increased shared information.

modality-specific information retained in intermediate layers (Fig. 4), we integrate representations across depth, retaining both shared and unique information for information-rich geo-embeddings (Fig. 1). We refer to this strategy as *Beyond What's Shared (BWS)*.

Given a subset of layer representations $\{\mathbf{z}^{(l)}\}_{l\in\mathcal{S}}$ where $\mathcal{S} \subseteq \{1, \ldots, L\}$, we assign depth-dependent weights:

$$\omega_l = \frac{e^{\alpha l}}{\sum_{k\in\mathcal{S}} e^{\alpha k}}, \tag{8}$$

where α controls the emphasis on early or late layers. The fused embedding is obtained by weighted concatenation:

$$\tilde{\mathbf{z}} = \left[\omega_l \mathbf{z}^{(l)}\right]_{l\in\mathcal{S}}. \tag{9}$$

Positive α emphasizes later layers, which are typically more alignment-optimized, while negative α emphasizes earlier layers, which tend to retain more modality-specific structure. When $\alpha = 0$, all layers contribute equally.

5. Experiments

5.1. Experimental Setup

Pretrain and downstream. We use the S2-100k pretraining dataset [27], which consists 100,000 globally distributed Sentinel-2 satellite images (256×256 pixels) paired with their geo-coordinates. For downstream evaluation, we use seven tasks: three classification (biome, ecoregions [14], country [42]) and four regression (temperature [23], elevation, population density [49], California housing price [46]).

Baselines. We compare our approach with state-of-the-art geo-foundation models. We use CSP [39], GeoCLIP [67], SatCLIP [27], and RANGE [13]. CSP, GeoCLIP, and SatCLIP learn shared embeddings by contrastively aligning imagery with geographic coordinates. RANGE augments SatCLIP with a retrieval-augmented strategy that incorporates high-resolution visual features from SatMAE [11].

Implementation details. We use SatCLIP's vision encoder pretrained with MoCo [22] on Sentinel-2 imagery [69], and SatCLIP's location encoder, a SIREN-based MLP [56] with spherical harmonics positional encodings [51]. Both encoders project into a shared embedding space and are optimized with a CLIP-style [47] contrastive loss. Pretraining runs for 500 epochs using Adam (learning rate 10^{-4}, weight decay 0.01) with a batch size of 8k. Data augmentations include random cropping, horizontal and vertical flips, Gaussian blur, and coordinate jitter within 1 km. We use linear probing [13] to purely assess layerwise representation quality for both classification and regression tasks. We report top-1 accuracy for classification and R^2 for regression tasks, following the data splits [13].

5.2. Results

Quantitative. Table 1 shows that BWS consistently improves vision-location geo-foundation models across diverse downstream tasks. These substantial gains highlight that retaining modality-specific (unique) information not explicitly preserved during contrastive alignment is useful for broad downstream tasks. BWS is model agnostic and complements RANGE, which enhances SatCLIP through unique visual information retrieved from similar locations. When combined, RANGE+BWS further improves performance, demonstrating the benefit of leveraging unique information from both imagery and location.

Qualitative. Fig. 6 visualizes the geo-embeddings obtained from SatCLIP [27] using independent component analysis [32]. Each map projects the high-dimensional em-

Table 1. Linear probing results across classification and regression tasks. All values are reported as mean ± std over three random splits. The first three columns correspond to classification and the remaining four to regression tasks. We use $\alpha = 0$ in BWS for simplicity.

Model	Biome	Ecoregions	Country	Temperature	Elevation	Population	Housing
CSP [39]	0.794±0.002	0.751±0.002	0.797±0.002	0.814±0.007	0.390±0.015	0.558±0.003	0.553±0.006
CSP + **BWS**	**0.881**±0.001	**0.834**±0.000	**0.898**±0.001	**0.918**±0.013	**0.659**±0.014	**0.720**±0.003	**0.639**±0.013
GeoCLIP [67]	0.809±0.002	0.739±0.001	0.813±0.000	0.920±0.007	0.608±0.012	0.691±0.001	0.704±0.009
GeoCLIP + **BWS**	**0.910**±0.002	**0.870**±0.002	**0.931**±0.001	**0.950**±0.005	**0.809**±0.009	**0.790**±0.003	**0.723**±0.009
SatCLIP [27]	0.848±0.001	0.779±0.002	0.826±0.001	0.818±0.010	0.668±0.011	0.685±0.002	0.400±0.006
SatCLIP + **BWS**	**0.911**±0.001	**0.865**±0.002	**0.916**±0.001	**0.860**±0.014	**0.788**±0.008	**0.766**±0.002	**0.421**±0.014
RANGE [13]	0.931±0.000	0.894±0.000	0.938±0.002	0.894±0.004	0.842±0.008	0.791±0.002	0.449±0.032
RANGE + **BWS**	**0.945**±0.000	**0.920**±0.001	**0.959**±0.001	**0.897**±0.004	**0.869**±0.006	**0.813**±0.002	**0.457**±0.024

beddings into three independent components shown as RGB channels. SatCLIP, which relies solely on its final-layer representations, primarily captures broad low-frequency spatial patterns. In contrast, our fused variant (SatCLIP+BWS) integrates intermediate and final layers, revealing finer spatial variations in the embeddings without any additional training objectives or external models. These finer details help explain the downstream gains observed in Table 1.

5.3. Ablation Study

① **Controlling for potential confounding factors.** We conduct two control experiments (Table 2). First, we evaluate dimension-matched baselines that output the same representation size as SatCLIP + BWS, showing that BWS still outperforms alternatives. Second, we compare BWS in multimodal versus unimodal settings. Gains are noticeably larger in the multimodal case, indicating that BWS further benefits from leveraging the modality-specific (unique) information retained in intermediate layers, rather than from increased dimensionality or generic feature aggregation.

Table 2. Dimension-controlled and uni/multi-modal ablations. We report average performance for all seven downstream tasks.

Method (same dim.)	Avg. Performance
SatCLIP + GeoCLIP	0.748
SatCLIP + Random Projection	0.750
SatCLIP + **BWS**	**0.788**

Encoder	Model	Avg. Performance *Baseline*	Avg. Performance ***+BWS***
Multimodal Loc.	Spherical Harmonics [51]	0.744	**0.788**
Unimodal Loc.	Spherical Harmonics [51]	0.717	0.731
Multimodal Img.	MoCo-ResNet50 [22]	0.523	**0.700**
Unimodal Img.	MoCo-ResNet50 [22]	0.247	0.301

② **Impact of α.** Table 3 evaluates the weight parameter α, which shifts the contribution toward earlier or later layers. Uniform weighting ($\alpha = 0$) performs best, suggesting that task-relevant information is distributed across layers and that combining diverse representations is most effective. Slight negative values offer a small advantage over positive ones, indicating that modality-specific information in earlier layers is beneficial for downstream tasks.

Table 3. Impact of the weight parameter α in BWS. We report average top-1 classification accuracy (3 tasks), average regression R^2 (4 tasks), and the overall mean for all tasks when varying α from unique-emphasizing (negative, earlier layers) to shared-emphasizing (positive, later layers).

α weight	Classification (↑)	Regression (↑)	Overall (↑)
−2.0	0.877	0.709	0.781
−1.0	0.881	0.716	0.787
0.0	0.881	0.718	0.788
1.0	0.876	0.714	0.783
2.0	0.866	0.708	0.776

③ **Impact of early, middle, and late layers.** Table 4 shows how early, middle, and late layers from SatCLIP contribute to downstream performance when evaluated in isolation. The middle layer performs best across all tasks, likely because it combines much of the modality-specific information preserved in early layers with some of the shared structure that begins to appear later. In the case of SatCLIP, these findings indicate that downstream tasks benefit most from intermediate representations, motivating BWS's strategy of fusing diverse information across layers.

④ **Impact of layer depth.** Table 5 shows that increasing the depth of SatCLIP consistently improves downstream performance when using BWS. Deeper encoders do not add harmful noise. Instead, they introduce a more diverse

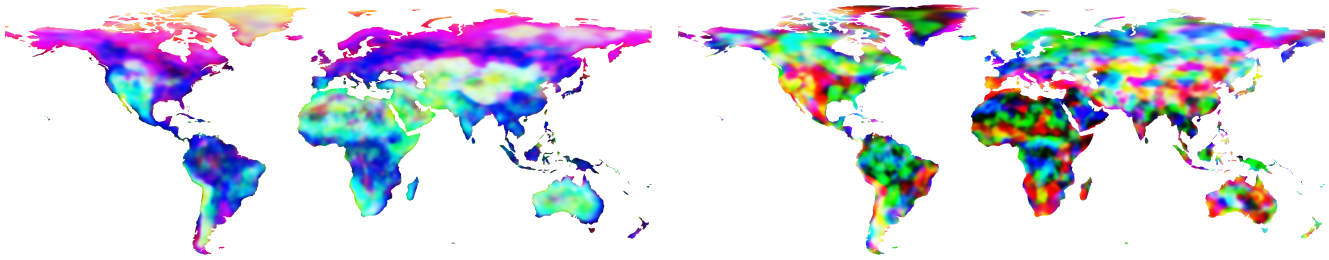


Figure 6. **Visualization of learned geo-embeddings using independent component analysis. (Left)** SatCLIP. **(Right)** SatCLIP+BWS. Three independent components are mapped to RGB color channels, so regions with similar colors have similar embeddings.

Table 4. Overall downstream performance is highest for middle layers, with early layers outperforming late layers.

Layer Group	Classification (↑)	Regression (↑)	Overall (↑)
Early	0.864	0.695	0.768
Middle	0.873	0.711	0.780
Late	0.841	0.690	0.755

set of intermediate representations that BWS can combine. Early layers provide stronger modality-specific signals, while later layers capture more shared information, and additional layers expand the range of these complementary features. This diversity proves useful for downstream tasks, which explains the steady improvements.

Table 5. Overall downstream performance increases with encoder layer depth when using BWS method.

Encoder Depth	Classification (↑)	Regression (↑)	Overall (↑)
3 layers	0.886	0.719	0.790
5 layers	0.893	0.729	0.799
10 layers	0.910	0.737	0.810
20 layers	0.926	0.750	0.826

5.4. Method Simplicity

Our layerwise method (BWS) adds virtually no computational overhead. BWS introduces no new parameters or additional forward passes, so the FLOPs remain identical to the underlying encoder. In contrast, methods that retain unique information through auxiliary objectives [35, 68] require additional mutual information estimation, increasing training cost and sensitivity to estimator quality [33] and augmentation design [61]. RANGE [13] also uses an external model (SatMAE [11]) to precompute a feature database. BWS does not require such extra components.

6. Discussion & Conclusion

Rethinking how to retain unique information. Our findings suggest a practical way to retain modality-specific (unique) information for multimodal contrastive models. Our layerwise analysis shows that multimodal contrastive models exhibit a layerwise information hierarchy in which intermediate layers capture more modality-specific (unique) features while later layers encode shared ones. Exploiting these intermediate representations provides a simple and efficient way to enjoy both unique and shared information without auxiliary tasks or external resources.

Leveraging layerwise information. Neural networks are widely understood to form hierarchical representations, with earlier layers encoding general features and later layers refining these features toward the training objective [75]. This progression aligns with the information bottleneck view, in which representations become increasingly compressed to retain only task-relevant information [55]. Our results indicate that multimodal contrastive models exhibit a similar behavior. Since the contrastive loss is applied at the projection layer, the final layer intuitively contains the most shared information. Earlier layers, by contrast, preserve modality-specific (unique) information originating from the input. This layerwise transition is evident in our analysis: intermediate layers capture more modality-specific information, whereas the final layer emphasizes shared information. Fusing these layers yields consistent gains across geospatial tasks, suggesting that the multi-view redundancy assumption [35, 64] may not hold in general. We further test our layerwise method across diverse settings (e.g., different modalities and downstream tasks) in the supplementary.

Limitations & Future Work. Our work focuses on two modalities (vision-location). Whether a similar layerwise pattern holds for three or more modalities remains an open question, particularly as multimodal interactions are not well-defined beyond two variables [34, 72]. Future work should explore how layerwise information emerges in high-order multimodal settings and across broader domains.

Acknowledgments

This manuscript has been authored by UT-Battelle, LLC under Contract No. DE-AC05-00OR22725 with the U.S. Department of Energy. The United States Government retains and the publisher, by accepting the article for publication, acknowledges that the United States Government retains a non-exclusive, paid-up, irrevocable, world-wide license to publish or reproduce the published form of this manuscript, or allow others to do so, for United States Government purposes. The Department of Energy will provide public access to these results of federally sponsored research in accordance with the DOE Public Access Plan (http://energy.gov/downloads/doe-public-access-plan). This work was supported by the Laboratory Directed Research and Development Program of Oak Ridge National Laboratory, managed by UTBattelle, LLC for the U.S. Department of Energy.

References

[1] Alessandro Achille and Stefano Soatto. Emergence of invariance and disentanglement in deep representations. *Journal of Machine Learning Research*, 19(50):1–34, 2018. 4

[2] Mohit Agarwal, Mimi Sun, Chaitanya Kamath, Arbaaz Muslim, Prithul Sarker, Joydeep Paul, Hector Yee, Marcin Sieniek, Kim Jablonski, Swapnil Vispute, et al. General geospatial inference with a population dynamics foundation model. *arXiv preprint arXiv:2411.07207*, 2024. 1

[3] Guillaume Alain and Yoshua Bengio. Understanding intermediate layers using linear classifier probes. *arXiv preprint arXiv:1610.01644*, 2016. 4

[4] Antonio Almudévar, José Miguel Hernández-Lobato, Sameer Khurana, Ricard Marxer, and Alfonso Ortega. Aligning multimodal representations through an information bottleneck. *arXiv preprint arXiv:2506.04870*, 2025. 3

[5] Randall Balestriero, Mark Ibrahim, Vlad Sobal, Ari Morcos, Shashank Shekhar, Tom Goldstein, Florian Bordes, Adrien Bardes, Gregoire Mialon, Yuandong Tian, et al. A cookbook of self-supervised learning. *arXiv preprint arXiv:2304.12210*, 2023. 2

[6] Daniel Bolya, Po-Yao Huang, Peize Sun, Jang Hyun Cho, Andrea Madotto, Chen Wei, Tengyu Ma, Jiale Zhi, Jathushan Rajasegaran, Hanoona Rasheed, et al. Perception encoder: The best visual embeddings are not at the output of the network. *arXiv preprint arXiv:2504.13181*, 2025. 4

[7] Florian Bordes, Randall Balestriero, Quentin Garrido, Adrien Bardes, and Pascal Vincent. Guillotine regularization: Why removing layers is needed to improve generalization in self-supervised learning. *arXiv preprint arXiv:2206.13378*, 2022. 4

[8] Christopher F Brown, Michal R Kazmierski, Valerie J Pasquarella, William J Rucklidge, Masha Samsikova, Chenhui Zhang, Evan Shelhamer, Estefania Lahera, Olivia Wiles, Simon Ilyushchenko, et al. Alphaearth foundations: An embedding field model for accurate and efficient global mapping from sparse label data. *arXiv preprint arXiv:2507.22291*, 2025. 1

[9] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020. 1, 3

[10] Yixiong Chen, Alan Yuille, and Zongwei Zhou. Which layer is learning faster? a systematic exploration of layer-wise convergence rate for deep neural networks. In *The Eleventh International Conference on Learning Representations*, 2023. 3

[11] Yezhen Cong, Samar Khanna, Chenlin Meng, Patrick Liu, Erik Rozi, Yutong He, Marshall Burke, David Lobell, and Stefano Ermon. Satmae: Pre-training transformers for temporal and multi-spectral satellite imagery. *Advances in Neural Information Processing Systems*, 35:197–211, 2022. 3, 6, 8

[12] Thomas M Cover. *Elements of information theory*. John Wiley & Sons, 1999. 5

[13] Aayush Dhakal, Srikumar Sastry, Subash Khanal, Adeel Ahmad, Eric Xing, and Nathan Jacobs. Range: Retrieval augmented neural fields for multi-resolution geo-embeddings. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 24680–24689, 2025. 1, 2, 3, 6, 7, 8

[14] Eric Dinerstein, David Olson, Anup Joshi, Carly Vynne, Neil D Burgess, Eric Wikramanayake, Nathan Hahn, Suzanne Palminteri, Prashant Hedao, Reed Noss, et al. An ecoregion-based approach to protecting half the terrestrial realm. *BioScience*, 67(6):534–545, 2017. 6

[15] Benoit Dufumier, Javiera Castillo-Navarro, Devis Tuia, and Jean-Philippe Thiran. What to align in multimodal contrastive learning? *arXiv preprint arXiv:2409.07402*, 2024. 3

[16] Utku Evci, Vincent Dumoulin, Hugo Larochelle, and Michael C Mozer. Head2toe: Utilizing intermediate representations for better transfer learning. In *International Conference on Machine Learning*, pages 6009–6033. PMLR, 2022. 4

[17] Marco Federici, Anjan Dutta, Patrick Forré, Nate Kushman, and Zeynep Akata. Learning robust representations via multi-view information bottleneck. *arXiv preprint arXiv:2002.07017*, 2020. 4

[18] Letian Fu, Gaurav Datta, Huang Huang, William Chung-Ho Panitch, Jaimyn Drake, Joseph Ortiz, Mustafa Mukadam, Mike Lambeta, Roberto Calandra, and Ken Goldberg. A touch, vision, and language dataset for multimodal alignment. *arXiv preprint arXiv:2402.13232*, 2024. 1

[19] Gianluca Guglielmo and Marc Masana. Leveraging intermediate representations for better out-of-distribution detection. *arXiv preprint arXiv:2502.12849*, 2025. 4

[20] Kartik Gupta, Thalaiyasingam Ajanthan, Anton van den Hengel, and Stephen Gould. Understanding and improving the role of projection head in self-supervised learning. *arXiv preprint arXiv:2212.11491*, 2022. 4

[21] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 1

[22] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF con-*

ference on computer vision and pattern recognition, pages 9729–9738, 2020. 6, 7

[23] Josh Hooker, Gregory Duveiller, and Alessandro Cescatti. A global dataset of air temperature derived from satellite remote sensing and weather stations. *Scientific data*, 5(1):1–11, 2018. 6

[24] Umangi Jain, Alex Wilson, and Varun Gulshan. Multimodal contrastive learning for remote sensing tasks. *arXiv preprint arXiv:2209.02329*, 2022. 1

[25] Jiachen Jiang, Jinxin Zhou, and Zhihui Zhu. Tracing representation progression: Analyzing and enhancing layer-wise similarity. *arXiv preprint arXiv:2406.14479*, 2024. 2

[26] Kenji Kawaguchi, Zhun Deng, Xu Ji, and Jiaoyang Huang. How does information bottleneck help deep learning? In *International conference on machine learning*, pages 16049–16096. PMLR, 2023. 2

[27] Konstantin Klemmer, Esther Rolf, Caleb Robinson, Lester Mackey, and Marc Rußwurm. Satclip: Global, general-purpose location embeddings with satellite imagery. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 4347–4355, 2025. 1, 3, 5, 6, 7

[28] Yunyong Ko, Dongwon Lee, and Sang-Wook Kim. Not all layers are equal: A layer-wise adaptive approach toward large-scale dnn training. In *Proceedings of the ACM Web Conference 2022*, pages 1851–1859, 2022. 3

[29] Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited. In *International conference on machine learning*, pages 3519–3529. PMLR, 2019. 5

[30] Alexander Kraskov, Harald Stögbauer, and Peter Grassberger. Estimating mutual information. *Physical Review E—Statistical, Nonlinear, and Soft Matter Physics*, 69(6): 066138, 2004. 5

[31] Alexandre Lacoste, Nils Lehmann, Pau Rodriguez, Evan Sherwin, Hannah Kerner, Björn Lütjens, Jeremy Irvin, David Dao, Hamed Alemohammad, Alexandre Drouin, et al. Geobench: Toward foundation models for earth monitoring. *Advances in Neural Information Processing Systems*, 36: 51080–51093, 2023. 1

[32] Te-Won Lee. Independent component analysis. In *Independent component analysis: Theory and applications*, pages 27–66. Springer, 1998. 6

[33] Cheuk Ting Leung, Rohan Ghosh, and Mehul Motani. Towards robust scale-invariant mutual information estimators. *Transactions on Machine Learning Research*, 2025. 8

[34] Paul Pu Liang, Yun Cheng, Xiang Fan, Chun Kai Ling, Suzanne Nie, Richard Chen, Zihao Deng, Nicholas Allen, Randy Auerbach, Faisal Mahmood, et al. Quantifying & modeling multimodal interactions: An information decomposition framework. *Advances in Neural Information Processing Systems*, 36:27351–27393, 2023. 8

[35] Paul Pu Liang, Zihao Deng, Martin Q Ma, James Y Zou, Louis-Philippe Morency, and Ruslan Salakhutdinov. Factorized contrastive learning: Going beyond multi-view redundancy. *Advances in Neural Information Processing Systems*, 36:32971–32998, 2023. 1, 2, 3, 8

[36] Victor Weixin Liang, Yuhui Zhang, Yongchan Kwon, Serena Yeung, and James Y Zou. Mind the gap: Understanding the modality gap in multi-modal contrastive representation learning. *Advances in Neural Information Processing Systems*, 35:17612–17625, 2022. 5

[37] Jiajun Ma, Tianyang Hu, and Wenjia Wang. Deciphering the projection head: Representation evaluation self-supervised learning. *arXiv preprint arXiv:2301.12189*, 2023. 4

[38] Gengchen Mai, Weiming Huang, Jin Sun, Suhang Song, Deepak Mishra, Ninghao Liu, Song Gao, Tianming Liu, Gao Cong, Yingjie Hu, et al. On the opportunities and challenges of foundation models for geospatial artificial intelligence. *arXiv preprint arXiv:2304.06798*, 2023. 1

[39] Gengchen Mai, Ni Lao, Yutong He, Jiaming Song, and Stefano Ermon. Csp: Self-supervised contrastive spatial pre-training for geospatial-visual representations. In *International Conference on Machine Learning*, pages 23498–23515. PMLR, 2023. 3, 5, 6, 7

[40] Gengchen Mai, Weiming Huang, Jin Sun, Suhang Song, Deepak Mishra, Ninghao Liu, Song Gao, Tianming Liu, Gao Cong, Yingjie Hu, et al. On the opportunities and challenges of foundation models for geoai (vision paper). *ACM Transactions on Spatial Algorithms and Systems*, 10(2):1–46, 2024. 1

[41] Leland McInnes, John Healy, and James Melville. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018. 5

[42] Jose Morales-Arilla and Shreyas Gadgin Matha. Glocal: A global development dataset of subnational administrative areas. *Scientific Data*, 11(1):851, 2024. 6

[43] Min Namgung, Yijun Lin, JangHyeon Lee, and Yao-Yi Chiang. Less is more: Multimodal region representation via pairwise inter-view learning. *arXiv preprint arXiv:2505.18178*, 2025. 1

[44] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018. 3, 4, 5

[45] Zhuo Ouyang, Kaiwen Hu, Qi Zhang, Yifei Wang, and Yisen Wang. Projection head is secretly an information bottleneck. *arXiv preprint arXiv:2503.00507*, 2025. 4

[46] R Kelley Pace and Ronald Barry. Sparse spatial autoregressions. *Statistics & Probability Letters*, 33(3):291–297, 1997. 6

[47] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 1, 2, 3, 4, 6

[48] Mariia Rizhko and Joshua S Bloom. Astrom3: A self-supervised multimodal model for astronomy. *The Astronomical Journal*, 170(1):28, 2025. 1

[49] Esther Rolf, Jonathan Proctor, Tamma Carleton, Ian Bolliger, Vaishaal Shankar, Miyabi Ishihara, Benjamin Recht, and Solomon Hsiang. A generalizable and accessible approach to machine learning with global satellite imagery. *Nature communications*, 12(1):4392, 2021. 6

[50] Esther Rolf, Konstantin Klemmer, Caleb Robinson, and Hannah Kerner. Position: Mission critical–satellite data is

a distinct modality in machine learning. In *Forty-first International Conference on Machine Learning*, 2024. 1

[51] Marc Rußwurm, Konstantin Klemmer, Esther Rolf, Robin Zbinden, and Devis Tuia. Geographic location encoding with spherical harmonics and sinusoidal representation networks. *arXiv preprint arXiv:2310.06743*, 2023. 3, 6, 7

[52] Michael Schmitt, Lloyd Haydn Hughes, Chunping Qiu, and Xiao Xiang Zhu. Sen12ms–a curated dataset of georeferenced multi-spectral sentinel-1/2 imagery for deep learning and data fusion. *arXiv preprint arXiv:1906.07789*, 2019. 1

[53] Simon Schrodi, David T Hoffmann, Max Argus, Volker Fischer, and Thomas Brox. Two effects, one trigger: On the modality gap, object bias, and information imbalance in contrastive vision-language models. *arXiv preprint arXiv:2404.07983*, 2024. 5

[54] Ravid Shwartz Ziv and Yann LeCun. To compress or not to compress—self-supervised learning and information theory: A review. *Entropy*, 26(3):252, 2024. 1, 3

[55] Ravid Shwartz-Ziv and Naftali Tishby. Opening the black box of deep neural networks via information. *arXiv preprint arXiv:1703.00810*, 2017. 2, 3, 4, 8

[56] Vincent Sitzmann, Julien Martel, Alexander Bergman, David Lindell, and Gordon Wetzstein. Implicit neural representations with periodic activation functions. *Advances in neural information processing systems*, 33:7462–7473, 2020. 6

[57] Oscar Skean, Md Rifat Arefin, Dan Zhao, Niket Patel, Jalal Naghiyev, Yann LeCun, and Ravid Shwartz-Ziv. Layer by layer: Uncovering hidden representations in language models. *arXiv preprint arXiv:2502.02013*, 2025. 4

[58] Stefano Soatto and Alessandro Chiuso. Visual representations: Defining properties and deep approximations. *arXiv preprint arXiv:1411.7676*, 2014. 3

[59] Francois Spoto, Omar Sy, Paolo Laberinti, Philippe Martimort, Valerie Fernandez, Olivier Colin, Bianca Hoersch, and Aime Meygret. Overview of sentinel-2. In *2012 IEEE international geoscience and remote sensing symposium*, pages 1707–1710. IEEE, 2012. 3

[60] Karthik Sridharan and Sham M Kakade. An information theoretic framework for multi-view learning. In *Colt*, number 114, pages 403–414, 2008. 4

[61] Yonglong Tian, Chen Sun, Ben Poole, Dilip Krishnan, Cordelia Schmid, and Phillip Isola. What makes for good views for contrastive learning? *Advances in neural information processing systems*, 33:6827–6839, 2020. 8

[62] Naftali Tishby and Noga Zaslavsky. Deep learning and the information bottleneck principle. In *2015 ieee information theory workshop (itw)*, pages 1–5. Ieee, 2015. 3

[63] Naftali Tishby, Fernando C Pereira, and William Bialek. The information bottleneck method. *arXiv preprint physics/0004057*, 2000. 4

[64] Christopher Tosh, Akshay Krishnamurthy, and Daniel Hsu. Contrastive learning, multi-view redundancy, and linear models. In *Algorithmic Learning Theory*, pages 1179–1206. PMLR, 2021. 1, 3, 8

[65] Arnas Uselis and Seong Joon Oh. Intermediate layer classifiers for ood generalization. *arXiv preprint arXiv:2504.05461*, 2025. 4

[66] Betty Van Aken, Benjamin Winter, Alexander Löser, and Felix A Gers. How does bert answer questions? a layer-wise analysis of transformer representations. In *Proceedings of the 28th ACM international conference on information and knowledge management*, pages 1823–1832, 2019. 2

[67] Vicente Vivanco Cepeda, Gaurav Kumar Nayak, and Mubarak Shah. Geoclip: Clip-inspired alignment between locations and images for effective worldwide geolocalization. *Advances in Neural Information Processing Systems*, 36:8690–8701, 2023. 1, 3, 5, 6, 7

[68] Haoqing Wang, Xun Guo, Zhi-Hong Deng, and Yan Lu. Rethinking minimal sufficient representation in contrastive learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16041–16050, 2022. 1, 3, 8

[69] Yi Wang, Nassim Ait Ali Braham, Zhitong Xiong, Chenying Liu, Conrad M Albrecht, and Xiao Xiang Zhu. Ssl4eos12: A large-scale multimodal, multitemporal dataset for self-supervised learning in earth observation [software and data sets]. *IEEE Geoscience and Remote Sensing Magazine*, 11(3):98–106, 2023. 3, 6

[70] Yi Wang, Conrad M Albrecht, Nassim Ait Ali Braham, Chenying Liu, Zhitong Xiong, and Xiao Xiang Zhu. Decoupling common and unique representations for multimodal self-supervised learning. In *European Conference on Computer Vision*, pages 286–303. Springer, 2024. 3

[71] Zifeng Wang, Zhenbang Wu, Dinesh Agarwal, and Jimeng Sun. Medclip: Contrastive learning from unpaired medical images and text. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing. Conference on Empirical Methods in Natural Language Processing*, page 3876, 2022. 1

[72] Paul L Williams and Randall D Beer. Nonnegative decomposition of multivariate information. *arXiv preprint arXiv:1004.2515*, 2010. 8

[73] Yihao Xue, Eric Gan, Jiayi Ni, Siddharth Joshi, and Baharan Mirzasoleiman. Investigating the benefits of projection head for representation learning. *arXiv preprint arXiv:2403.11391*, 2024. 4

[74] Jason Yosinski, Jeff Clune, Yoshua Bengio, and Hod Lipson. How transferable are features in deep neural networks? *Advances in neural information processing systems*, 27, 2014. 2, 3

[75] Matthew D Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *European conference on computer vision*, pages 818–833. Springer, 2014. 2, 3, 8

[76] Chiyuan Zhang, Samy Bengio, and Yoram Singer. Are all layers created equal? *Journal of Machine Learning Research*, 23(67):1–28, 2022. 3

[77] Xiao Xiang Zhu, Devis Tuia, Lichao Mou, Gui-Song Xia, Liangpei Zhang, Feng Xu, and Friedrich Fraundorfer. Deep learning in remote sensing: A comprehensive review and list of resources. *IEEE geoscience and remote sensing magazine*, 5(4):8–36, 2017. 1