Decoupling Common and Unique Representations for Multimodal Self-supervised Learning

Yi Wang[1,2], Conrad M Albrecht[2], Nassim Ait Ali Braham[1,2], Chenying Liu[1], Zhitong Xiong[1], and Xiao Xiang Zhu[1,3]

[1] Data Science in Earth Observation, Technical University of Munich, Germany
{yi4.wang, chenying.liu, zhitong.xiong, xiaoxiang.zhu}@tum.de
[2] Remote Sensing Technology Institute, German Aerospace Center, Germany
[3] Munich Center for Machine Learning, Germany
{conrad.albrecht,nassim.aitalibraham}@dlr.de

Abstract. The increasing availability of multi-sensor data sparks wide interest in multimodal self-supervised learning. However, most existing approaches learn only common representations across modalities while ignoring intra-modal training and modality-unique representations. We propose **De**coupling **C**ommon and **U**nique **R**epresentations (DeCUR), a simple yet effective method for multimodal self-supervised learning. By distinguishing inter- and intra-modal embeddings through multimodal redundancy reduction, DeCUR can integrate complementary information across different modalities. We evaluate DeCUR in three common multimodal scenarios (radar-optical, RGB-elevation, and RGB-depth), and demonstrate its consistent improvement regardless of architectures and for both multimodal and modality-missing settings. With thorough experiments and comprehensive analysis, we hope this work can provide valuable insights and raise more interest in researching the hidden relationships of multimodal representations[4].

Keywords: Self-supervised learning · Multimodal representations

1 Introduction

Self-supervised learning has achieved breakthroughs in machine learning [13] and many other communities [25, 43]. Driven by the success of single-modal representation learning, as well as the great potential that large-scale multi-sensor data bears, multimodal self-supervised learning is gaining increasing attention [1, 31, 35, 48, 52, 53]. A common strategy for existing works is to align different modalities as augmented views and conduct cross-modal contrastive learning, pulling together features of different modalities for the same scene and pushing away those for the different scenes.

While aligning different modalities in a common latent space has shown success in various multimodal scenarios [16, 35, 36, 42], the fact that one modality may hold unique information that can not be extracted from other modalities

[4] https://github.com/zhu-xlab/DeCUR

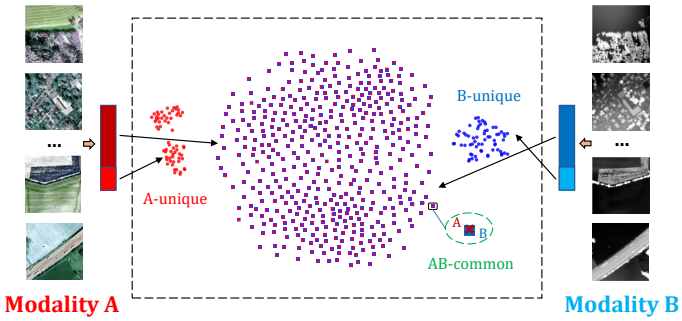


Fig. 1: Decoupled common and unique representations across two modalities visualized by t-SNE [28]. **Each embedding dimension is one data point.** Red and blue circles indicate unique features from modalities A and B; red cross and blue square indicate common features from A and B. The figure shows that common and unique features from different modalities are well separated in the embedding space, and the common features between modalities are well overlapped. Best view in color & zoomed in.

tend to be overlooked. As a result, such modality-unique information is suppressed during training. This forces the model to put potentially orthogonal representations into common feature embeddings, limiting the model's capacity to understand different modalities in detail. To tackle this issue, we introduce the idea of Decoupling Common and Unique Representations (DeCUR) for multimodal self-supervised learning.

Specifically, we separate the feature embedding dimensions into cross-modal common ones and modality-unique ones. During training, we calculate the normalized cross-correlation matrix of the common dimensions and the unique dimensions between two modalities, and drive the matrix of the common dimensions to the identity, while the matrix of the unique dimensions to zero. As a result, common embeddings are aligned across modalities, while modality-unique embeddings are pushed away. In practice, this can be seen as a natural extension of Barlow Twins [57], one self-supervised learning technique that conducts redundancy reduction on the dimension level.

However, simply pushing away unique dimensions will lead to a collapse that these dimensions do not learn any useful information. Therefore, apart from cross-modal learning, we include also intra-modal learning which utilizes all embedding dimensions and drives the cross-correlation matrix between two augmented views of the same modality to the identity. This intra-modal component not only avoids the collapse by letting the unique dimensions learn meaningful representations within one modality, but also enhances cross-modal learning with stronger intra-modal knowledge. Fig. 1 provides a t-SNE [28] visualization of the latent space of the learned representations, where common and unique embeddings of each modality, as well as modality-unique embeddings between modalities, are well separated.

In addition, previous works have shown that different modalities may embed important information in different regions of the feature maps [54, 58]. Inspired by these findings, we introduce a modern deformable attention module [49, 50] in our ConvNet backbones to help the model focus on modality-specific important areas during training. Such deformable attention selects the positions of key and value pairs in self-attention in a data-dependent way, which is both more efficient compared to previous attention modules and well-fitted for our modality-enhancing purpose. In summary, our main contributions are listed as follows:

- We propose DeCUR, a simple yet effective multimodal self-supervised learning method, which decouples common and unique representations across different modalities and enhances both intra- and inter-modal learning. For ConvNet backbones, we adopt a simple adaptation of deformable attention for modality-informative feature learning.
- We evaluate DeCUR with rich experiments and comprehensive analysis covering three important multi-modal scenarios, demonstrating its effectiveness in both multi-modal and modality-missing settings.

2 Related work

Self-supervised learning Self-supervised learning of single modality has been widely studied, which can be categorized into three main types: generative methods such as Autoencoder [41] and Masked Autoencoder [21], predictive methods such as predicting rotation angles [15], and contrastive methods that train joint embedding architectures with or without negative samples. Contrastive methods can be further categorized into four strategies of self-supervision: 1) contrastive learning with negative samples such as CPC [33], SimCLR [8] and MoCo [22]; 2) clustering feature embeddings such as SwAV [5]; 3) knowledge distillation such as BYOL [18], SimSiam [10] and DINO [6]; 4) redundancy reduction such as Barlow Twins [57] and VICReg [3]. The second and third categories usually require common encoders, thus not easily adaptable for modality-specific encoders in multimodal scenarios. While most existing multimodal works are closely related to the first strategy, DeCUR belongs to redundancy reduction as a natural extension of Barlow Twins that does not require numerous negative samples. Specifically, DeCUR's decoupling strategy can be perfectly integrated into a simple correlation-matrix-based loss design in Barlow Twins.

Multimodal self-supervised learning The idea of contrastive self-supervised learning can be easily transferred to multimodal scenarios, as different modalities are naturally the augmented views. Currently, contrastive learning with negative samples has been mostly developed: CLIP [35] for language-image, VATT [1] for video-audio-text, CROMA [14] for radar-optical, and ImageBind [16] for a joint embedding of six different modalities. Different from these methods, we propose to explore the potential of negative-free methods by extending the redundancy reduction loss of Barlow Twins. We also take one step further to decouple common and unique information from different modalities. Meanwhile, we share an

insight with Yang *et al*. [56] and Wang *et al*. [44] that intra-modal representations are important complements to cross-modal representations.

Modality decoupling While not well explored in self-supervised research, modality decoupling has been proven beneficial in multimodal supervised learning. Xiong *et al*. [54, 55] studied multimodal fusion from network architecture, proposing modality separation networks for RGB-D scene recognition. Peng *et al*. [34] investigated modality dominance from the angle of optimization flow, proposing on-the-fly gradient modulation to balance and control the optimization of each modality in audio-visual learning. Zhou *et al*. [59] observed feature redundancy for different supervision tasks, proposing to decompose task-specific and task-shared features for multitask learning in recommendation system. FactorCL [26] studies the decoupling concept in self-supervision, factorizing task-relevant information into shared and unique representations with modality-specific augmentations. Different from the above, we directly perform modality decoupling on the embeddings by separating common and unique dimensions.

3 Methodology

Fig. 2 presents the general structure of DeCUR. As a multimodal extension of Barlow Twins, DeCUR performs self-supervised learning by redundancy reduction in the joint embedding space of augmented views from both intra-/cross-modal perspectives. Here, our main contribution lies in a simple loss design to decouple meaningful modality-unique representations across modalities.

3.1 Decoupling common and unique representations

As shown in Fig. 2, we feed two batches of augmented views of the inputs from each modality into the modality-specific encoders and projectors, producing corresponding embeddings ${Z_{M1}}'$ and ${Z_{M1}}''$ for X_{M1}, and ${Z_{M2}}'$ and ${Z_{M2}}''$ for X_{M2}, respectively. Batch normalization is applied on embeddings such that they are mean-centered along the batch dimension. These embeddings are then used to calculate cross-correlation matrices across/within modalities for optimization.

Cross-correlation matrix Given two embedding vectors $Z^A, Z^B \in \mathbb{R}^K$, the cross-correlation matrices $\mathcal{C}$ between them is formulated as [57]:

$$\mathcal{C}_{ij} = \frac{\sum_b z^A_{b,i} z^B_{b,j}}{\sqrt{\sum_b \left(z^A_{b,i}\right)^2} \sqrt{\sum_b \left(z^B_{b,j}\right)^2}} \tag{1}$$

where b indexes batch samples, and i, j index the dimension of the embedding vectors. $\mathcal{C} \in \mathbb{R}^{K \times K}$ is a square matrix with values ranging from -1 to 1.

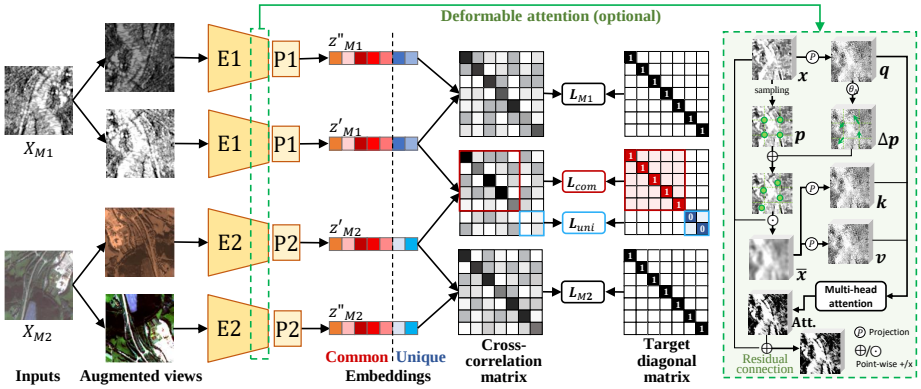


Fig. 2: The structure of DeCUR. $M1$ and $M2$ represent two modalities. Two augmented views from each modality are fed to modality-specific encoders ($E1$, $E2$) and projectors ($P1$, $P2$) to get the embeddings Z. For cross-modal embeddings, the dimensions are separated into common and unique parts. The correlation matrix of the common dimensions is optimized to be close to the identity, while that of the unique ones to zero. For intra-modal embeddings, both common and unique dimensions are used to calculate the correlation matrix which is optimized to be close to the identity. DeCUR optionally adds deformable attention (the green shadowed region on the right side) in the last layers of ConvNet encoders to boost modality-informative learning.

Cross-modal representation decoupling In the cross-modal case, the correlation matrix $\mathcal{C}$ is calculated between two embeddings from different modalities, such as ${Z_{M1}}'$ and ${Z_{M2}}'$ in Fig. 2. While most multimodal self-supervised learning algorithms consider only their common representations, we explicitly consider the existence of modality-unique representations and decouple them during training. Specifically, we separate the total embedding dimension K into K_c and K_u with $K_c + K_u = K$ to store common and unique representations, respectively. The common representations should be identical across modalities (red parts in Fig. 2), while the modality-specific unique representations should be decorrelated (blue parts in Fig. 2).

Specifically, on the one hand, to learn cross-modal common information, a sub-matrix $\mathcal{C}_c \in \mathbb{R}^{K_c \times K_c}$ is generated from only the common dimensions of ${Z_{M1}}'$ and ${Z_{M2}}'$. The redundancy reduction loss for the cross-modal common representations reads:

$$\mathcal{L}_{com} = \sum_i (1 - \mathcal{C}_{c_{ii}})^2 + \lambda_c \cdot \sum_i \sum_{j \neq i} \mathcal{C}_{c_{ij}}^2 \quad , \tag{2}$$

where λ_c is a positive constant trading off the importance of the first invariance term (to make the common embeddings invariant to the input modalities) and the second redundancy reduction term (to decorrelate the embedding vector components and avoid model collapse).

On the other hand, to decouple modality-specific information, a sub-matrix $\mathcal{C}_{\mathrm{u}} \in \mathbb{R}^{K_u \times K_u}$ is generated from the unique dimensions of ${Z_{M1}}'$ and ${Z_{M2}}'$. The redundancy reduction loss for the modality-unique representations reads:

$$\mathcal{L}_{uni} = \sum_i \mathcal{C}_{\mathrm{u}ii}^2 + \lambda_u \cdot \sum_i \sum_{j \neq i} \mathcal{C}_{\mathrm{u}ij}^2 \quad , \tag{3}$$

where λ_u is a positive constant trading off the importance of the first decorrelation term (to decorrelate different modalities) and the second redundancy reduction term (to decorrelate the embedding vector components). However, pure decoupling doesn't ensure the meaningfulness of the unique dimensions, i.e., they could collapse into random decorrelated values. To tackle this issue, we further introduce intra-modal representation enhancement that covers both common and unique dimensions within each modality.

Intra-modal representation enhancing To avoid the collapse of the decoupled unique dimensions in the cross-modal training, as well as to boost intra-modal representations, we introduce intra-modal training that covers all the embedding dimensions. For each modality, a cross-correlation matrix $\mathcal{C}_{\mathrm{M1}}$ (or $\mathcal{C}_{\mathrm{M2}}$) is generated from the full dimensions of the embedding vectors ${Z_{M1}}'$ and ${Z_{M1}}''$ (or ${Z_{M2}}'$ and ${Z_{M2}}''$). The redundancy reduction losses for the intra-modal representations read:

$$\mathcal{L}_{M1} = \sum_i \left(1 - \mathcal{C}_{\mathrm{M1}ii}\right)^2 + \lambda_{M1} \cdot \sum_i \sum_{j \neq i} \mathcal{C}_{\mathrm{M1}ij}^2 \quad , \tag{4}$$

$$\mathcal{L}_{M2} = \sum_i \left(1 - \mathcal{C}_{\mathrm{M2}ii}\right)^2 + \lambda_{M2} \cdot \sum_i \sum_{j \neq i} \mathcal{C}_{\mathrm{M2}ij}^2 \quad , \tag{5}$$

where λ_{M1} and λ_{M2} are positive constants trading off the importance of the invariance term and the redundancy reduction term.

Combining the cross-modal common and unique losses, and the intra-modal losses, the overall training objective of DeCUR reads:

$$\mathcal{L} = \mathcal{L}_{com} + \mathcal{L}_{uni} + \mathcal{L}_{M1} + \mathcal{L}_{M2} \quad . \tag{6}$$

3.2 Deformable attention for modality-informative features

Apart from the DeCUR loss design, we adopt deformable attention to help ConvNet models focus on modality-informative regions. The deformable attention module was proposed in DAT [49] and DAT++ [50] to efficiently model the relations among feature tokens under the guidance of the important regions in the feature maps. The readers are referred to the original papers for technical details, while a brief simplified recap is as follows. Given an input feature map $x \in \mathbb{R}^{H \times W \times C}$, a downsampled grid of points $p \in \mathbb{R}^{H_G \times W_G \times 2}$ is generated as references, where $H_G = H/r$ with r being the downscaling ratio. In parallel, the feature map x is projected to the query tokens $q \in \mathbb{R}^{H \times W \times C}$, where $q = xW_q$.

The query tokens q are fed into a lightweight sub-network θ_{offset} to generate the offsets $\Delta p \in \mathbb{R}^{H_G \times W_G \times 2}$ in order to get final deformed points with $p + \Delta p$. Then the features are sampled from x at the locations of deformed points and interpolated to a feature map $\bar{x} \in \mathbb{R}^{H_G \times W_G \times C}$. This sampled feature map $\bar{x}$ is projected to keys k and values v, where $k = \bar{x}W_k$ and $v = \bar{x}W_v$. Softmax attention is then calculated on flattened queries q and keys k and multiplied with values v. The final output is reshaped back to the same size as the input feature map x.

We adopt the deformable attention module in the last two stages of the encoder to learn regional focus while keeping efficiency. A residual connection from the input feature map to the output of the deformable attention module is added to restrict unexpected influences of the attention module, such as biasing the pretraining towards the pretext task by selecting unexpected deformable points. This is especially helpful in early training, when the model needs to first capture general information. With the training going on, the model then gradually learns detailed modality-specific representations (c.f. Sec. 6).

4 Implementation details

Pretraining datasets We pretrain DeCUR in three multimodal scenarios: SAR-optical, RGB-DEM and RGB-depth. For SAR-optical, we use the SSL4EO-S12 dataset [45] which consists of 251k multi-modal image pairs from multiple seasons with size 264x264. One random season is selected for each image and modality. For RGB-DEM, we conduct pretraining on the training set of GeoNRW dataset [2] which includes aerial RGB images, digital elevation models (DEM) and segmentation maps from the German state North Rhine-Westphalia. We crop the raw 6,942 training scenes to 111k patches with size 250x250. For RGB-depth, we use SUN-RGBD dataset [38] which consists of 10,335 RGBD pairs with various image sizes. We preprocess the depth images to HHA format [20] following [58]. Common data augmentations [18] are selected and used based on the feasibility in each specific modality (c.f. Appendix).

Model architecture As a multimodal extension of Barlow Twins [57], each branch holds a separate backbone and a 3-layer MLP projector (each with output dimension 8192). DeCUR is trained on embedding representations after the projector, whose dimensions are separated into common and unique. We do a light grid search to get the best corresponding ratio. For SAR-optical, the percentage of common dimensions is 87.5%; for RGB-DEM and RGB-depth it is 75%. The backbones are transferred to downstream tasks. We use ResNet-50 [23] for all scenarios, with additional MiT-B2/B5 from SegFormer [51] for RGB-Depth. We adopt deformable attention to the last two stages of the ResNet-50 backbone. We do not adopt deformable attention to MiTs as attention is already integrated.

Optimization We follow the optimization protocol of Barlow Twins [57] and BYOL [18], with default epochs 100 and a batch size of 256 for SAR-optical

Table 1: SAR-optical transfer results (mAP) on BigEarthNet-MM (left: multimodal; middle: SAR-only), and optical-only results on BigEarthNet-S2 compared with SOTA Earth observation foundation models (right). We report both "linear-probing/fine-tuning" scores. *Rand. Init.* represents random initialization, *-cross* represents cross-modal, *BT-SAR* represents Barlow Twins with SAR-only. The same denotations are used in the following. Best scores among self-supervised methods are marked in **bold**. *: SkySense uses a mixed backbone combining ViT and Swin Transformer.

SAR-optical	1% labels	100% labels
Rand. Init.	58.7	70.1
Supervised	77.0	88.9
SimCLR-cross	77.4/78.7	82.8/89.6
CLIP	77.4/78.7	82.8/89.6
Barlow Twins	78.7/80.3	83.2/89.5
VICReg	74.5/79.0	81.9/89.5
DeCUR (ours)	**79.8**/**81.5**	**86.2**/**89.8**

SAR-only	1% labels	100% labels
Rand. Init.	50.0	54.2
Supervised	67.5	81.9
SimCLR-cross	68.1/70.4	71.7/83.7
CLIP	68.0/70.2	71.7/83.4
Barlow Twins	72.3/73.7	77.8/83.6
VICReg	69.3/71.9	74.1/83.6
BT-SAR	71.2/73.3	77.5/81.6
DeCUR (ours)	**74.4**/**76.0**	**79.5**/**84.0**

Optical-only	Backbone	10% labels
SeCo [29]	RN50	78.6/82.6
SSL4EO [46]	RN50	82.1/86.2
FG-MAE [47]	ViT-S	78.1/85.2
GFM [30]	Swin-B	-/86.3
SpectralGPT [24]	ViT-B	-/87.5
SatMAE [11]	ViT-L	80.3/86.2
CROMA [14]	ViT-L	**85**/**88.3**
SkySense [19]	ViT-L*	**-**/**88.7**
DeCUR (ours)	RN50	83.3/87.2

and RGB-DEM (epochs 200 and batch size 128 for RGB-depth). The trade-off parameters λ of the loss terms are set to 0.0051. Training is distributed across 4 NVIDIA A100 GPUs and takes about 35 hours on SSL4EO-S12, 6 hours on GeoNRW, and 6 hours on SUN-RGBD.

5 Experimental results

We evaluate DeCUR by transferring to three common multimodal tasks: SAR-optical scene classification, RGB-DEM semantic segmentation, and RGB-depth semantic segmentation. We follow common evaluation protocols of frozen encoder and fine-tuning. We report results for full- and limited-label settings, and both multimodal and modality-missing (single-modal) settings.

5.1 SAR-optical scene classification

We pretrain SAR-optical encoders on SSL4EO-S12 [45] and transfer them to BigEarthNet-MM [39], a multimodal multi-label scene classification dataset with 19 classes. Simple late fusion is used for multimodal transfer learning by concatenating the encoded features from both modalities, followed by one classification layer. Mean average precision (mAP, micro) serves as the evaluation metric.

We report multimodal linear probing and fine-tuning results with 1% and 100% training labels in Tab. 1 (left). DeCUR outperforms existing cross-modal SimCLR-like contrastive learning by 2.4%-3.4% in linear probing, while achieving comparable performance on fine-tuning with full labels. Compared to BarlowTwins, we improve by 1.1% and 1.2% on linear evaluation and fine-tuning with 1% labels, and 3.0% and 0.3% with full labels.

We report SAR-only results in Tab. 1 (middle), as it is an essential scenario in practice when optical images are either unavailable or heavily covered by clouds. DeCUR outperforms other methods in most scenarios by a large margin (up

Table 2: RGB-DEM transfer learning results (mIoU) with frozen-encoder and full fine-tuning on GeoNRW (left: multimodal; right: RGB-only).

RGB-DEM	1% labels		100% labels	
	Frozen	Fine-tune	Frozen	Fine-tune
Rand. Init.	14.1	14.1	23.0	23.0
Supervised	22.1	22.1	44.0	44.0
SimCLR-cross	23.0	30.2	35.2	47.3
CLIP	22.8	28.8	35.0	46.7
Barlow Twins	31.2	33.6	43.0	48.4
VICReg	27.4	32.8	38.0	45.1
DeCUR (ours)	**34.7**	**36.6**	**44.7**	**48.9**

RGB-only	1% labels		100% labels	
	Frozen	Fine-tune	Frozen	Fine-tune
Rand. Init.	14.2	14.2	18.5	18.5
Supervised	17.5	17.5	38.8	38.8
SimCLR-cross	20.1	25.9	29.6	42.5
CLIP	20.0	25.7	29.4	42.3
Barlow Twins	29.4	33.4	38.0	45.9
VICReg	23.7	28.7	32.4	41.6
BarlowTwins-RGB	28.6	32.6	36.2	45.7
DeCUR (ours)	**32.2**	**35.7**	**40.8**	**46.7**

to 7.8%), while achieving slightly better performance on fine-tuning with full labels. In addition, DeCUR outperforms single-modal Barlow Twins pretraining by 2.7%-3.2% with 1% labels and 2.0%-2.4% with full labels, indicating that joint multimodal pretraining helps the model better understand individual modalities.

In addition, we compare DeCUR with state-of-the-art Earth observation foundation models on BigEarthNet-S2 with 10% labels in Tab. 1 (right). DeCUR achieves better performance than existing models with comparable model sizes such as SeCo [29], SSL4EO [45] and FG-MAE [47] in both linear probing and fine-tuning. Compared to SOTA large models, DeCUR is only slightly worse than CROMA [14], SkySense [19] and SpectralGPT [24] with much fewer parameters, while even better than GFM [30] and SatMAE [11].

5.2 RGB-DEM semantic segmentation

We pretrain and evaluate RGB-DEM encoders on GeoNRW [2] for semantic segmentation (10 classes). For fair and visible comparison, we use simple fully convolutional networks (FCN) [27] as the segmentation model, which concatenates the last three layer feature maps from both modalities, upsamples and sums them up to generate prediction maps. Similar to the classification task, we report frozen-encoder and full fine-tuning results in Tab. 2 with mean Intersection over Union (mIoU) serving as the evaluation metric.

We present multimodal frozen-encoder and fine-tuning results with 1% and 100% training labels in Tab. 2 (left). Promisingly, DeCUR outperforms other methods in all scenarios by a large margin (up to 11.7% compared to SimCLR). Notably, DeCUR works better than Barlow Twins and VICReg by at least 3.0% when fine-tuning with 1% labels. Meanwhile, we report RGB-only results in Tab. 2 (right). Again DeCUR shows a significant improvement compared to other cross-modal methods in all scenarios (up to 12.1% compared to CLIP). In addition, DeCUR outperforms single-modal Barlow Twins by 3.6%-4.6% with a frozen encoder, and 1.0%-3.1% in fine-tuning.

5.3 RGB-depth semantic segmentation

We pretrain RGB-depth encoders on SUN-RGBD [38] and transfer them to SUN-RGBD and NYU-Depth v2 [32] datasets for semantic segmentation (37

Table 3: RGB-depth fine-tuning results on SUN-RGBD and NYU-Depth v2.

SUN-RGBD	Modal	mIoU	OA
FCN [27]	RGB	27.4	68.2
FCN (CLIP [35])	RGB	30.5	74.2
FCN (DeCUR)	RGB	34.5	75.5
SA-Gate [9]	RGBD	49.4	82.5
SGNet [7]	RGBD	48.6	82.0
ShapeConv [4]	RGBD	48.6	82.2
CMX-B2 [58]	RGBD	49.7	82.8
CMX-B2 (DeCUR)	RGBD	**50.6**	**83.2**

NYUDv2	modal	mIoU	OA
FCN [27]	RGB	29.2	60.0
FCN (CLIP [35])	RGB	30.4	63.3
FCN (DeCUR)	RGB	31.2	63.9
SA-Gate [9]	RGBD	52.4	77.9
ShapeConv [4]	RGBD	51.3	76.4
OMNIVORE [17]	RGBD	54.0	-
CMX-B5 [58]	RGBD	56.9	80.1
CMX-B5 (DeCUR)	RGBD	**57.3**	**80.3**

and 40 classes, respectively). We transfer ResNet50 to simple FCN [27] and MiT-B2/B5 [51] to the recent CMX [58] model. We report single- and multi-modal fine-tuning results with mIoU and overall accuracy (OA) in Tab. 3. As observed, DeCUR helps improve FCN over CLIP by 4.0% mIoU and 1.3% OA on SUN-RGBD, and 0.8% mIoU and 0.6% OA on NYU-Depth v2.

Promisingly, consistent improvements are observed by simply transferring the pretrained backbones to SOTA supervised mutimodal fusion models. Following the published codebase and without tuning any hyperparameter, we push CMX-B2 from 49.7% to 50.6% in mIoU on the SUN-RGBD dataset, and CMX-B5 from 56.9% to 57.3% in mIoU on the NYU-Depth v2 dataset.

6 Ablation studies

Deformable attention We conduct frozen-encoder ablation on DeCUR for the deformable attention with residual connection (RDA) as shown in Tab. 4. While DA without residual connection decreases label-limited performance, the RDA module consistently improves in almost all scenarios.

Table 4: Ablation results (mAP) on the deformable attention module.

Dataset	BigEarthNet-MM		BigEarthNet-SAR		GeoNRW-MM		GeoNRW-RGB	
	1%	100%	1%	100%	1%	100%	1%	100%
w/o. DA	79.4	85.4	73.7	78.3	34.9	43.9	31.4	38.4
with DA	−0.1	-	−0.2	-	−0.6	-	−1.2	-
with RDA	+0.4	+0.8	+0.7	+1.2	−0.2	+0.8	+0.8	+2.4

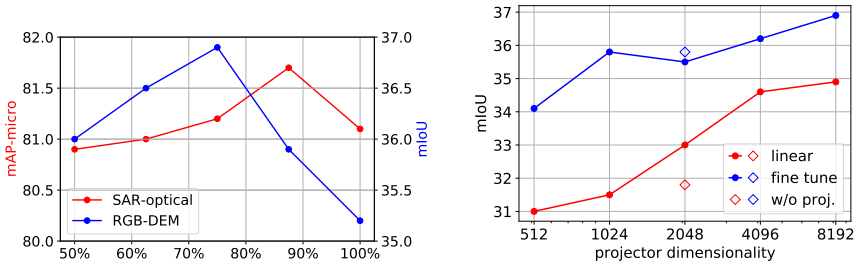
Loss terms The components of the DeCUR loss are ablated in Tab. 5, where fine-tuning is conducted without the RDA module. As shown, when neither intra-modal nor modality-unique information is learned, the performance of the single cross-modal Barlow Twins (w/o intra&decoup.) significantly degrades in downstream tasks. Additionally, solely incorporating modality decoupling (w/o intra) yields unstable effects in the two scenarios. This can be explained by the fact

that without intra-modal training the unique dimensions can be randomly decorrelated and are not ensured meaningful. When intra-modal training is added (w/o decoup.), the model's performance improves consistently in both scenarios, albeit inferior to DeCUR. This underscores the complementary benefits of intra-modal representations to commonly learned cross-modal representations. Together, these results affirm the effectiveness of the total DeCUR loss.

Table 5: Ablation results on loss components, where *intra* and *decoup.* correspond to intra-modal training and modality decoupling, respectively.

	SAR-optical (mAP)	RGB-DEM (mIoU)
DeCUR (ours)	**81.7**	**36.9**
w/o intra&decoup.	80.3	33.6
w/o intra	80.1	34.3
w/o decoup.	81.1	35.2

Decoupling percentage We conduct a simple grid search to find the best ratio between common and unique dimensions for different modality combinations that may have different representation overlaps. As shown in Fig. 3a, the best percentage of common dimensions is 87.5% for SAR-optical and 75% for RGB-DEM. This could be in line with the fact that there is more valid modality-unique information in orthophoto and elevation models than in optical and SAR (when the optical image is almost cloud-free). In both scenarios, the downstream performance increases and decreases smoothly along with the change of the percentage of common dimensions. Interestingly, there is no significant performance drop when decoupling up to 50% unique dimensions, which indicates the sparsity of the common embedding space.



(a) Ablation results on the percentage of common dimensions.

(b) Effect of the projector dimensionality on the GeoNRW dataset.

Fig. 3: Ablation results on the percentage of common dimensions and the projector.

Effect of the projector Inherited from Barlow Twins [57], DeCUR also benefits from the increasing dimensionality of the projector. As can be seen in Fig. 3b, DeCUR keeps improving with all output dimensionality tested. Interestingly, DeCUR works well on the segmentation task even without the projector. Removing the projector gives reasonable downstream performances, while adding it back can further enhance the representations.

Robustness of decoupling percentage The grid search in Fig. 3a was built upon an embedding dimension of 8192. To see how the best percentage changes along with the embedding dimensionality, we repeat the search with an embedding dimension of 512 on SAR-optical and RGB-DEM datasets. As is shown in Fig. 4, the best percentage of common dimensions is interestingly the same for both the small and the big embedding spaces.

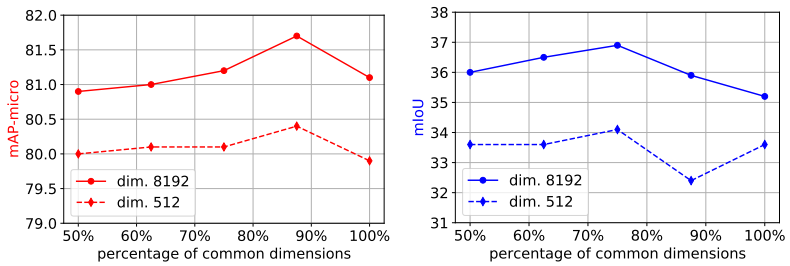


Fig. 4: Ablation on the decoupling percentage with different embedding dimensionalities (left: SAR-optical; right: RGB-DEM).

7 Discussion

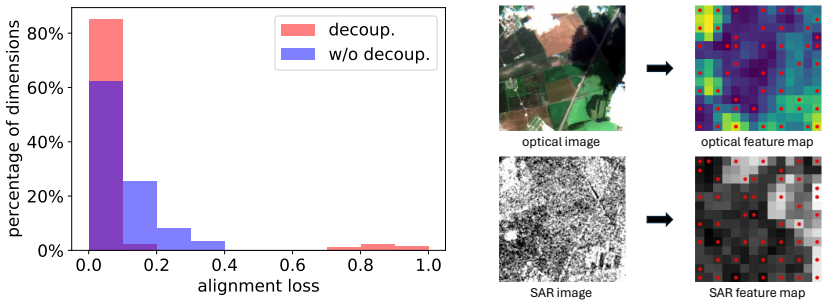
In this section, we demonstrate an explainability analysis to interpret the multimodal representations learned by DeCUR. We illustrate SAR-optical analysis on the SSL4EO-S12 dataset as an example scenario here.

Cross-modal representation alignment To examine that each modality contains unique information that is difficult to integrate into a common space, we calculate the cross-modal alignment of every embedding dimension. This is done by counting the on-diagonal losses of the cross-correlation matrix $\mathcal{C}$:

$$\mathcal{L}_i = (1 - \mathcal{C}_{ii})^2 \quad , \tag{7}$$

where i is the i_{th} embedding dimension. The closer $\mathcal{L}_i$ to 0, the better the alignment of the two modalities in this dimension. We count the loss for all dimensions and plot the histogram of one random batch for both DeCUR and cross-modal

Barlow Twins. The former explicitly decouples unique dimensions, while the latter assumes that all dimensions are common. As shown in Fig. 5a, the alignment loss remains high for a certain number of dimensions with cross-modal Barlow Twins. On the contrary, by allowing the decorrelation of several dimensions (the loss of which moves to 1), the misalignment of common dimensions decreases. These results are aligned with the visualization by t-SNE mentioned above in Fig. 1, where modality-unique dimensions are well separated, while common dimensions are perfectly overlapped in DeCUR.



(a) Representation alignment analysis. **(b)** Feature maps with deformable points.

Fig. 5: Cross-modal representation alignment (left) and DA visualization (right).

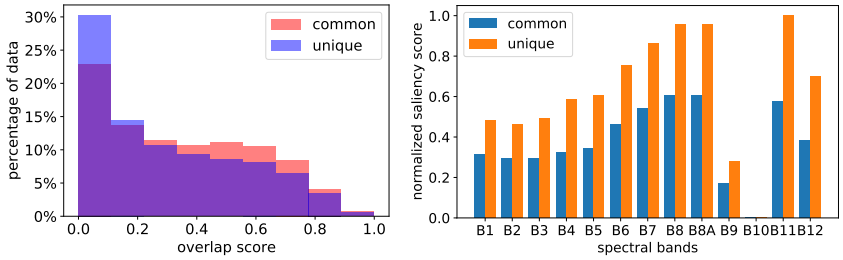
Deformable feature visualization We visualize the first principle component [12] of the feature maps with deformable points learned by the deformable attention module in Fig. 5b. For illustration purposes, we visualize in color for optical and in grey for SAR. It shows that the two modalities make different focuses in this scene: the model learns to ignore clouds in the optical image, while on the contrary paying more attention to such regions in the SAR image as radar signal can go through the clouds. The deformable points in general follow the feature attention, but spread more sparsely compared to natural vision. This is due to two facts: 1) satellite imagery is not object-centric and information is spread out, and 2) the residual connection restricts the deforming capacity.

Spatial and spectral saliency statistics We count spatial and spectral saliency statistics over the whole dataset. For preparation, we average the common and unique dimensions as two single values. Next, one-time backpropagation is performed w.r.t the corresponding output target (0 for common and 1 for unique) to get GradCAM [37] and Integrated Gradients [40]. One "common" and one "unique" saliency map are generated for each modality with both visualization methods.

Then, we use GradCAM saliency maps to count spatial statistics, calculating one overlap score for the common area and one for the unique area between

the modalities, respectively. Going through the whole dataset, we can get a histogram in Fig. 6a. The histogram shows a trend of unique scores being more towards 0 than common scores, indicating that the interesting areas of modality-unique representations tend to be more orthogonal than common representations which tend to overlap.

On the other hand, We average the importance scores of each band in the Integrated Gradients saliency maps to get spectral saliency for both common and unique representations. We normalize the scores, count statistics over the whole dataset, and plot the histograms in Fig. 6b. The figure confirms the bigger influence of the spectral information on optical-unique representations. Meanwhile, the band-wise importance distribution is promisingly consistent with the domain knowledge: 1) near-infrared bands (B5-B8A, including vegetation red edge) are very important; 2) water vapor (B9) and cirrus (B10) are strongly related to the atmosphere and thus less important for land surface monitoring; etc.



(a) SAR-optical spatial saliency statistics. **(b)** Spectral saliency statistics for 13 optical bands.

Fig. 6: Spatial saliency statistics (left) and spectral saliency statistics (right).

8 Conclusion

We propose DeCUR, a multimodal self-supervised learning method that learns both cross-modal common and modality-unique representations. Extensive experiments on three common multimodal scenarios prove the effectiveness of DeCUR, suggesting the great potential of modality-decoupling.

One limitation of DeCUR is it simplifies the multimodal situation by allocating the same common-unique ratio across the dataset. Future work could consider more complex scenarios where one modality may contain more unique information than the other in different scenes. Another limitation is the grid search for the best percentage of common dimensions, which can be costly on a huge dataset. While a general percentage of around 80% can achieve reasonable performance in our tested scenarios, a more efficient discovering strategy is to be explored. Other directions for future work include the exploration of adaptive decoupling strategies, and integrating modality decoupling in unified foundation models with more than two modalities.

Acknowledgement

The main work of Y. Wang, C. Liu, and C. Albrecht was funded by the Helmholtz Association through the Framework of Helmholtz AI, grant ID: `ZT-I-PF-5-01` - *Local Unit Munich Unit @Aeronautics, Space and Transport (MASTr)*. The compute was supported by the Helmholtz Association's Initiative and Networking Fund on the HAICORE@FZJ partition. The work of N. Ait Ali Braham was supported by the European Commission through the project "EvoLand" under the Horizon 2020 Research and Innovation program (Grant Agreement No. 101082130). The work of X. Zhu was supported by the German Federal Ministry of Education and Research (BMBF) in the framework of the international future AI lab "AI4EO – Artificial Intelligence for Earth Observation: Reasoning, Uncertainties, Ethics and Beyond" (grant number: 01DD20001) and by the Munich Center for Machine Learning. The work of Z. Xiong was supported by the German Federal Ministry for the Environment, Nature Conservation, Nuclear Safety and Consumer Protection (BMUV) based on a resolution of the German Bundestag (grant number: 67KI32002B; Acronym: *EKAPEx*). Y. Wang's work on rebuttal and camera-ready paper preparation was supported by the European Commission through the project "ThinkingEarth—Copernicus Foundation Models for a Thinking Earth" under the Horizon 2020 Research and Innovation program (Grant Agreement No. 101130544).

References

1. Akbari, H., Yuan, L., Qian, R., Chuang, W.H., Chang, S.F., Cui, Y., Gong, B.: Vatt: Transformers for multimodal self-supervised learning from raw video, audio and text. Advances in Neural Information Processing Systems **34**, 24206–24221 (2021) 1, 3
2. Baier, G., Deschemps, A., Schmitt, M., Yokoya, N.: Geonrw (2020). https://doi.org/10.21227/s5xq-b822, https://dx.doi.org/10.21227/s5xq-b822 7, 9
3. Bardes, A., Ponce, J., LeCun, Y.: Vicreg: Variance-invariance-covariance regularization for self-supervised learning. arXiv preprint arXiv:2105.04906 (2021) 3
4. Cao, J., Leng, H., Lischinski, D., Cohen-Or, D., Tu, C., Li, Y.: Shapeconv: Shape-aware convolutional layer for indoor rgb-d semantic segmentation. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 7088–7097 (2021) 10
5. Caron, M., Misra, I., Mairal, J., Goyal, P., Bojanowski, P., Joulin, A.: Unsupervised learning of visual features by contrasting cluster assignments. Advances in neural information processing systems **33**, 9912–9924 (2020) 3
6. Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., Joulin, A.: Emerging properties in self-supervised vision transformers. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 9650–9660 (2021) 3
7. Chen, L.Z., Lin, Z., Wang, Z., Yang, Y.L., Cheng, M.M.: Spatial information guided convolution for real-time rgbd semantic segmentation. IEEE Transactions on Image Processing **30**, 2313–2324 (2021) 10

8. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A simple framework for contrastive learning of visual representations. In: International conference on machine learning. pp. 1597–1607. PMLR (2020) 3
9. Chen, X., Lin, K.Y., Wang, J., Wu, W., Qian, C., Li, H., Zeng, G.: Bi-directional cross-modality feature propagation with separation-and-aggregation gate for rgb-d semantic segmentation. In: European Conference on Computer Vision. pp. 561–577. Springer (2020) 10
10. Chen, X., He, K.: Exploring simple siamese representation learning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 15750–15758 (2021) 3
11. Cong, Y., Khanna, S., Meng, C., Liu, P., Rozi, E., He, Y., Burke, M., Lobell, D., Ermon, S.: Satmae: Pre-training transformers for temporal and multi-spectral satellite imagery. Advances in Neural Information Processing Systems **35**, 197–211 (2022) 8, 9
12. Dunteman, G.H.: Principal components analysis, vol. 69. Sage (1989) 13
13. Ericsson, L., Gouk, H., Loy, C.C., Hospedales, T.M.: Self-supervised representation learning: Introduction, advances, and challenges. IEEE Signal Processing Magazine **39**(3), 42–62 (2022) 1
14. Fuller, A., Millard, K., Green, J.: Croma: Remote sensing representations with contrastive radar-optical masked autoencoders. Advances in Neural Information Processing Systems **36** (2024) 3, 8, 9
15. Gidaris, S., Singh, P., Komodakis, N.: Unsupervised representation learning by predicting image rotations. arXiv preprint arXiv:1803.07728 (2018) 3
16. Girdhar, R., El-Nouby, A., Liu, Z., Singh, M., Alwala, K.V., Joulin, A., Misra, I.: Imagebind: One embedding space to bind them all. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 15180–15190 (June 2023) 1, 3
17. Girdhar, R., Singh, M., Ravi, N., van der Maaten, L., Joulin, A., Misra, I.: Omnivore: A single model for many visual modalities. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 16102–16112 (June 2022) 10
18. Grill, J.B., Strub, F., Altché, F., Tallec, C., Richemond, P., Buchatskaya, E., Doersch, C., Avila Pires, B., Guo, Z., Gheshlaghi Azar, M., et al.: Bootstrap your own latent-a new approach to self-supervised learning. Advances in neural information processing systems **33**, 21271–21284 (2020) 3, 7
19. Guo, X., Lao, J., Dang, B., Zhang, Y., Yu, L., Ru, L., Zhong, L., Huang, Z., Wu, K., Hu, D., et al.: Skysense: A multi-modal remote sensing foundation model towards universal interpretation for earth observation imagery. arXiv preprint arXiv:2312.10115 (2023) 8, 9
20. Gupta, S., Girshick, R., Arbeláez, P., Malik, J.: Learning rich features from rgb-d images for object detection and segmentation. In: Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part VII 13. pp. 345–360. Springer (2014) 7
21. He, K., Chen, X., Xie, S., Li, Y., Dollár, P., Girshick, R.: Masked autoencoders are scalable vision learners. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 16000–16009 (2022) 3
22. He, K., Fan, H., Wu, Y., Xie, S., Girshick, R.: Momentum contrast for unsupervised visual representation learning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 9729–9738 (2020) 3

23. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016) 7
24. Hong, D., Zhang, B., Li, X., Li, Y., Li, C., Yao, J., Yokoya, N., Li, H., Jia, X., Plaza, A., et al.: Spectralgpt: Spectral foundation model. arXiv preprint arXiv:2311.07113 (2023) 8, 9
25. Krishnan, R., Rajpurkar, P., Topol, E.J.: Self-supervised learning in medicine and healthcare. Nature Biomedical Engineering **6**(12), 1346–1352 (2022) 1
26. Liang, P.P., Deng, Z., Ma, M.Q., Zou, J.Y., Morency, L.P., Salakhutdinov, R.: Factorized contrastive learning: Going beyond multi-view redundancy. Advances in Neural Information Processing Systems **36** (2024) 4
27. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3431–3440 (2015) 9, 10
28. Van der Maaten, L., Hinton, G.: Visualizing data using t-sne. Journal of machine learning research **9**(11) (2008) 2
29. Manas, O., Lacoste, A., Giró-i Nieto, X., Vazquez, D., Rodriguez, P.: Seasonal contrast: Unsupervised pre-training from uncurated remote sensing data. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 9414–9423 (2021) 8, 9
30. Mendieta, M., Han, B., Shi, X., Zhu, Y., Chen, C.: Towards geospatial foundation models via continual pretraining. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 16806–16816 (2023) 8, 9
31. Mu, N., Kirillov, A., Wagner, D., Xie, S.: Slip: Self-supervision meets language-image pre-training. In: European Conference on Computer Vision. pp. 529–544. Springer Nature Switzerland Cham (2022) 1
32. Nathan Silberman, Derek Hoiem, P.K., Fergus, R.: Indoor segmentation and support inference from rgbd images. In: ECCV (2012) 9
33. Oord, A.v.d., Li, Y., Vinyals, O.: Representation learning with contrastive predictive coding. arXiv preprint arXiv:1807.03748 (2018) 3
34. Peng, X., Wei, Y., Deng, A., Wang, D., Hu, D.: Balanced multimodal learning via on-the-fly gradient modulation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8238–8247 (2022) 4
35. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PMLR (2021) 1, 3, 10
36. Scheibenreif, L., Hanna, J., Mommert, M., Borth, D.: Self-supervised vision transformers for land-cover segmentation and classification. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1422–1431 (2022) 1
37. Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Grad-cam: Visual explanations from deep networks via gradient-based localization. In: Proceedings of the IEEE international conference on computer vision. pp. 618–626 (2017) 13
38. Song, S., Lichtenberg, S.P., Xiao, J.: Sun rgb-d: A rgb-d scene understanding benchmark suite. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 567–576 (2015) 7, 9
39. Sumbul, G., De Wall, A., Kreuziger, T., Marcelino, F., Costa, H., Benevides, P., Caetano, M., Demir, B., Markl, V.: Bigearthnet-mm: A large-scale, multimodal,

multilabel benchmark archive for remote sensing image classification and retrieval [software and data sets]. IEEE Geoscience and Remote Sensing Magazine **9**(3), 174–180 (2021) 8
40. Sundararajan, M., Taly, A., Yan, Q.: Axiomatic attribution for deep networks. In: International conference on machine learning. pp. 3319–3328. PMLR (2017) 13
41. Vincent, P., Larochelle, H., Lajoie, I., Bengio, Y., Manzagol, P.A., Bottou, L.: Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. Journal of machine learning research **11**(12) (2010) 3
42. Wang, L., Luc, P., Recasens, A., Alayrac, J.B., Oord, A.v.d.: Multimodal self-supervised learning of general audio representations. arXiv preprint arXiv:2104.12807 (2021) 1
43. Wang, Y., Albrecht, C.M., Braham, N.A.A., Mou, L., Zhu, X.X.: Self-supervised learning in remote sensing: A review. arXiv preprint arXiv:2206.13188 (2022) 1
44. Wang, Y., Albrecht, C.M., Zhu, X.X.: Self-supervised vision transformers for joint sar-optical representation learning. In: IGARSS 2022-2022 IEEE International Geoscience and Remote Sensing Symposium. pp. 139–142. IEEE (2022) 4
45. Wang, Y., Braham, N.A.A., Xiong, Z., Liu, C., Albrecht, C.M., Zhu, X.X.: SSL4EO-S12: A large-scale multi-modal, multi-temporal dataset for self-supervised learning in earth observation. arXiv preprint arXiv:2211.07044 (2022) 7, 8, 9
46. Wang, Y., Braham, N.A.A., Xiong, Z., Liu, C., Albrecht, C.M., Zhu, X.X.: Ssl4eo-s12: 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) 8
47. Wang, Y., Hernández, H.H., Albrecht, C.M., Zhu, X.X.: Feature guided masked autoencoder for self-supervised learning in remote sensing. arXiv preprint arXiv:2310.18653 (2023) 8, 9
48. Wei, L., Xie, L., Zhou, W., Li, H., Tian, Q.: Mvp: Multimodality-guided visual pre-training. In: European Conference on Computer Vision. pp. 337–353. Springer Nature Switzerland Cham (2022) 1
49. Xia, Z., Pan, X., Song, S., Li, L.E., Huang, G.: Vision transformer with deformable attention. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 4794–4803 (2022) 3, 6
50. Xia, Z., Pan, X., Song, S., Li, L.E., Huang, G.: Dat++: Spatially dynamic vision transformer with deformable attention. arXiv preprint arXiv:2309.01430 (2023) 3, 6
51. Xie, E., Wang, W., Yu, Z., Anandkumar, A., Alvarez, J.M., Luo, P.: Segformer: Simple and efficient design for semantic segmentation with transformers. Advances in Neural Information Processing Systems **34**, 12077–12090 (2021) 7, 10
52. Xiong, Z., Wang, Y., Zhang, F., Stewart, A.J., Hanna, J., Borth, D., Papoutsis, I., Saux, B.L., Camps-Valls, G., Zhu, X.X.: Neural plasticity-inspired foundation model for observing the earth crossing modalities. arXiv preprint arXiv:2403.15356 (2024) 1
53. Xiong, Z., Wang, Y., Zhang, F., Zhu, X.X.: One for all: Toward unified foundation models for earth vision. arXiv preprint arXiv:2401.07527 (2024) 1
54. Xiong, Z., Yuan, Y., Wang, Q.: Msn: Modality separation networks for rgb-d scene recognition. Neurocomputing **373**, 81–89 (2020) 3, 4
55. Xiong, Z., Yuan, Y., Wang, Q.: Ask: Adaptively selecting key local features for rgb-d scene recognition. IEEE Transactions on Image Processing **30**, 2722–2733 (2021) 4

56. Yang, J., Duan, J., Tran, S., Xu, Y., Chanda, S., Chen, L., Zeng, B., Chilimbi, T., Huang, J.: Vision-language pre-training with triple contrastive learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 15671–15680 (2022) 4
57. Zbontar, J., Jing, L., Misra, I., LeCun, Y., Deny, S.: Barlow twins: Self-supervised learning via redundancy reduction. In: International Conference on Machine Learning. pp. 12310–12320. PMLR (2021) 2, 3, 4, 7, 12
58. Zhang, J., Liu, H., Yang, K., Hu, X., Liu, R., Stiefelhagen, R.: Cmx: Cross-modal fusion for rgb-x semantic segmentation with transformers. arXiv preprint arXiv:2203.04838 (2022) 3, 7, 10
59. Zhou, J., Yu, Q., Luo, C., Zhang, J.: Feature decomposition for reducing negative transfer: A novel multi-task learning method for recommender system. arXiv preprint arXiv:2302.05031 (2023) 4