Contrastive Learning for Urban Land Cover Classification With Multimodal Siamese Network

Rui Liu, *Graduate Student Member, IEEE*, Jing Ling, *Graduate Student Member, IEEE*, Yinyi Lin, *Member, IEEE*, and Hongsheng Zhang, *Senior Member, IEEE*

***Abstract*— The Earth observation era has bestowed dividends upon supervised land cover classification based on deep learning and optical data. However, limitations, such as insufficient spectral information and reduced quality during inclement weather for optical data, coupled with the need for extensive labeled samples, impede accurate classification. This letter harnesses multimodal images with deep contrastive learning to reduce reliance on labeled data and classify land covers. By employing a well-designed contrastive learning method with triangular similarity loss, our model can learn effective multimodal features without labeled samples. Moreover, the learned features are fused at the early feature level and used for the downstream classification task with fewer labeled samples. Experimental results demonstrate the benefits of incorporating multiple modalities, highlighting the potential of combining multimodal image analysis and contrastive learning for land cover classification with limited labeled samples.**

***Index Terms*— Contrastive learning, optical and SAR data, self-supervised learning, urban land cover classification.**

I. Introduction

LAND cover classification stands firmly in the middle in the realm of remote sensing. In this theme, the emergency of the Earth observation era has advanced deep learning with optical images [1], [2] in a supervised manner. When referring to the urban land cover classification, the complexity of artificial landscapes as well as the rainy and cloudy weather in certain tropical and subtropical coastal cities will pose challenges in urban land cover classification, consequently compromising the classification accuracy. Although there is substantial impressive research in mapping the urban land cover [3] at a large scale with index-based methods based on optical data, once deep learning and optical images are utilized, some easily ignored challenges will arise.

Manuscript received 29 April 2024; accepted 7 August 2024. Date of publication 12 August 2024; date of current version 26 August 2024. This work was supported in part by the Research Grants Council (RGC) of Hong Kong, China under Grant HKU27602020 and Grant HKU17613022; in part by the National Natural Science Foundation of China under Grant 42071390; in part by Shenzhen Science and Technology Program under Grant JCYJ20210324124013037; and in part by the Seed Funding for Strategic Interdisciplinary Research Scheme of The University of Hong Kong. *(Corresponding author: Hongsheng Zhang.)*

The authors are with The University of Hong Kong Shenzhen Institute of Research and Innovation, Hong Kong, and also with the Department of Geography, The University of Hong Kong, Hong Kong (e-mail: zhanghs@hku.hk).

Digital Object Identifier 10.1109/LGRS.2024.3442434

The challenges can be illustrated from data and method perspectives. From a data aspect, as the spatial resolution of optical imagery becomes progressively higher, more and more spatial details can be captured, but the details are insufficient to fill the semantic gap in heterogeneous landscapes inside cities, as only visible and infrared wavelengths are captured. Moreover, the clouds, rain, and fog may cause image degradation and further hinder the urban land cover classification task. It might be a solution to consider observations from multiple sources [4], [5] to compensate for the limited information available from a single optical image. From the perspective of methodology, although the index-based method [3] can achieve satisfactory results on a large scale, the classification accuracy is still limited locally. Deep learning has yielded impressive results in land cover classification; however, most current deep learning methods [1], [2] are in a supervised manner, that is, they rely heavily on annotated samples, which will unavoidably entail additional costs, especially when conducting large-scale urban land cover classification mapping. It is also worth noting that they are designed for single-modal optical images only.

With current advancements in Earth observation, remote sensing images can be obtained at diverse spatial and spectral resolutions, offering robust data support. Thus, it is worthwhile to explore strategies for extracting valuable information from these data in a cost-effective manner. Contrastive learning is an ideal approach due to its ability to learn task-independent features without the need for labeled data. It can be transferred to downstream tasks using only a portion of the available labels [6], [7]. Contrastive learning aims to minimize the distance between similar inputs and sometimes also maximize the distance between dissimilar inputs. For example, in [8], the positive pair is the seismic waveform and the augmented from the same earthquake, while the negative pair is the seismic waveform from different earthquake events, and the similarity is measured by supervised contrastive loss. Open self-supervised features are calculated with fewer labeled samples to form the prototype of each class in [9], and the classification is to compare the Euclidean distances across each prototype. As indicated in [10], features from one view are compulsory to be similar to another view on the change mask, and to balance the class imbalance, the focal loss is selected as a similarity indicator. Currently, the prevalent

approach involves utilizing the augmented viewpoint [11], [12] or generated image [13] as a positive sample pair alongside the original image, while other images remain as negative pairs [14], [15]. Owing to the revisiting period of satellite, the augmented view could be images from different times [16]. However, data from a single modality cannot describe the properties of the surface thoroughly [17], and identifying good views for the modality is time-consuming [13]. Multimodal data can serve as a natural positive pair, as objects observed across different modalities possess inherent similarities in their intrinsic qualities. When applied, the process of generating different views can be omitted. For example, the optical, mask, and SAR images are combined jointly for contrastive learning [18]. The optical and digital surface model is used in [19], while the optical and SAR data are proven to be more effective than contrastive learning based solely on the optical data [20], [21].

Although the aforementioned methods have demonstrated improvements, there are still several considerations regarding their limitations. First, practical applications often face challenges in collecting large amounts of unlabeled data, particularly multimodal data for contrastive learning. Second, as most current multimodal contrastive learning methods are based on bimodal data, it raises the question of how to extend the learning to the scenarios with an increased number of modalities and whether incorporating more modalities necessarily leads to better pretask results. To narrow the gaps, we collect multiple modalities of data, including Gaofen-5 hyperspectral image, Landsat-8 multispectral image, and ALOS-2 SAR image, and an urban test site is chosen in their common area to test our hypothesis. Furthermore, we have developed a novel contrastive learning approach for multiple modalities, drawing inspiration from the Simsiam [22] and incorporating a triplet loss. The method can effectively generalize to accommodate an expanded range of modal data. Moreover, an application-oriented urban land cover classification task based on early fusion is used to evaluate the effectiveness of the proposed contrastive learning method. To summarize, our contributions are as follows.

1) We designed a multimodal contrast learning framework and demonstrated experimentally that more modalities are beneficial for contrast learning.
2) We prove that the contrastive learning method enables achieving better classification accuracy with only small training samples toward practical remote sensing applications.

II. Method

A. Problem Definition

Given an input image $I = \{(X^H, X^M, X^S)\} \in R^{C\times H\times W}$, where H, M, and S mean hyperspectral, multispectral, and SAR images, respectively. Then, the labeled samples can be denoted as $I_L = \{(x_i^H, x_i^M, x_i^S, y_i)\}_{i=1}^N$, where N is the number of labeled samples; then, the unlabeled samples can be denoted as $I_{unL} = \{(x_u^H, x_u^M, x_u^S) \in I\}_{u=N}^{H\times W}$; typically, N is much smaller than $H \times W$. Let f be the function which can map the inputs into lower dimension features, and then, the features

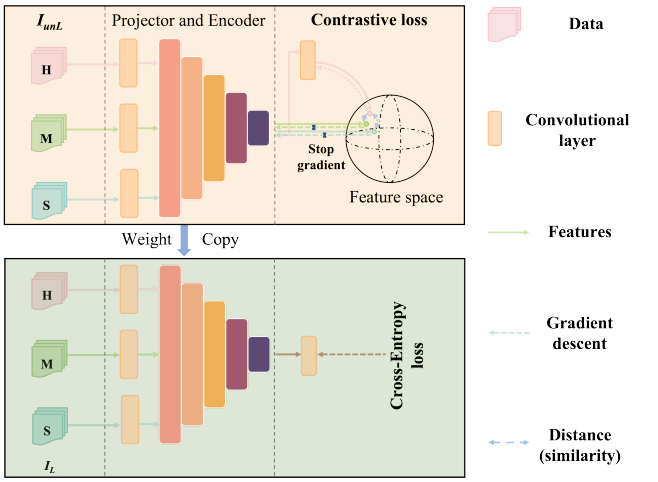

Fig. 1. Framework of multimodal Siamese network for contrastive learning and downstream classification.

are fed into function g to obtain land cover types. Our goal is to learn the function $g \circ f$ with as little I_L as possible, which is possible when f can be learned from I_{unL} and it can force samples from the same class to be closer in feature space. Indeed, the goal of contrastive learning is to learn such a function f. Once f is well learned, g can be learned via supervised learning with fewer labeled samples. Next, f and g will be introduced separately and how contrastive learning is utilized to learn f will be described.

B. Multimodal Siamese Network

Multimodal Siamese network is the desired function f. As depicted in Fig. 1, it is parameterized by the individual projection convolutions, a shared weight encoder, and a predictor. Given unlabeled samples I_{unL}, each element inside I_{unL} is data from three modalities, named sample pair. Contrastive learning would force each sample pair and similar sample pairs to be close in the feature space by gradually updating the network parameters until convergence.

As the images from H, M, and S have different numbers of bands, they may need to be processed by individual encoders, which will burden the computation. The projection convolution is dedicated to projecting H, M, and S into the same dimensional space, which will release the demand of individual encoders for each modality. Three convolutions of 16 kernels of size 3×3 for each are used as projectors here. After projection, the features from different modalities are fed into the shared encoder to gain high-level semantic representations. Theoretically, the encoder could be any image classification model, regardless of a CNN or a transformer. Here, slimed MobileNetv2 is chosen due to its fewer parameters and computational load. The predictor and stop gradient serve to prevent the model from falling into collapse. The component of the predictor is a fully connected layer, batch normalization (BN) layer, and ReLU, followed by another fully connected layer.

The multimodal contrastive learning is executed as follows. After multimodal inputs are passed through the encoder, respectively, the features can be denoted as z_k, where $k \in \{H, W, S\}$. Given a feature from a specific modality,

like H in Fig. 1, it will be passed through the predictor, giving rise to p_k. The similarity will be calculated with features from the remaining modalities. Here, cosine similarity is used, where

$$\text{sim}\left(\overrightarrow{m_1}, \overrightarrow{m_2}\right) = \frac{\overrightarrow{m_1} \cdot \overrightarrow{m_2}}{\left\|\overrightarrow{m_1}\right\|_2 \left\|\overrightarrow{m_2}\right\|_2}. \tag{1}$$

Since the objective is to maximize (1), therefore, when optimizing the parameters using gradient descent, the loss function should be

$$\text{loss} = -\frac{\sum_{k\in\{H,M,S\}} \sum_{j\in\{H,M,S\}, j\neq k} \text{sim}\left(p_k, z_j\right)}{\binom{n}{2}}. \tag{2}$$

It is worth noting that Fig. 1 only shows how the modality H updates the parameters of the projector, encoder, and predictor. When backpropagation, the predictor of the remaining modalities will act as the predictor in Fig. 1 and the gradients from the other two modalities will also be stopped. Through this asynchronous updating, there is no need for negative sample pairs, and the goal of contrastive learning is achieved. Equation (2) applies equally to a wide range of modalities, and if more projection convolutions were utilized, the method could be easily extended for a wider range of modalities.

C. Classification Based on Early Fusion

Since the aim is land use classification, once the multimodal Siamese network converges, the next step is to learn function g with limited labeled samples I_L. Features z_H, z_M, and z_S from projection convolution are added first. After the encoder, g will be used to make the prediction. Owing to the limited size of I_L, to prevent overparameterization of g and ensure that supervised learning can effectively update its parameters, the complexity of g should be carefully constrained. Here, g is parameterized as

$$g(z_H, z_M, z_S) = \text{linear}(f(z_H + z_M + z_S)). \tag{3}$$

The same as most classifications based on supervised learning, the loss function is the cross-entropy loss.

III. Result and Discussion

A. Experimental Setup

1) Implementation Details: The parameters of f are randomly initialized, after contrastive learning is finished, to validate the effectiveness, in the classification task, f is initialized in two ways, by loading the parameters from contrastive learning and randomly initialized, while g has to be randomly initialized. The batch size is set to be 512. The input is a patch of size 13 centered at pixels in I. The output dimension of the encoder is 2048 and the output dimension for the projector is 512. AdamW optimizer is used for all experiments, with a learning rate of 0.002 for contrastive learning and 100 training epochs. For the classification model, the epoch is 60. For the learning rate, it would be 0.009 for finetuning and 0.02 for training from scratch. All labeled samples are randomly divided equally into five folds, four of which are used for training and one for testing. To objectively validate the model performance, a fivefold cross-validation was used for all classifications. Gaofen-5, Landsat, and ALOS-2 data are multimodal data, and Kwai Tsing is chosen as the study area due to its complex terrestrial landscape. The Gaofen-5 underwent ortho rectification and band selection, resulting in 288 remaining bands. The optical images underwent radiometric calibration and atmospheric correction. ALOS-2 data underwent radiometric calibration, spatial filtering, and terrain calibration for further processing. All data were registered under the WGS 84, UTM 50N coordinate system using Landsat as the base map. Within the study area, eight land cover types were defined, and a total of 2422 uniformly distributed sample points were manually annotated based on high-resolution imagery. These samples included 330 bare lands, 442 buildings, 224 grasslands, 201 ground, 527 roads, 24 vegetation, 251 water, and 203 other samples, which are numbered from 1 to 7, except others.

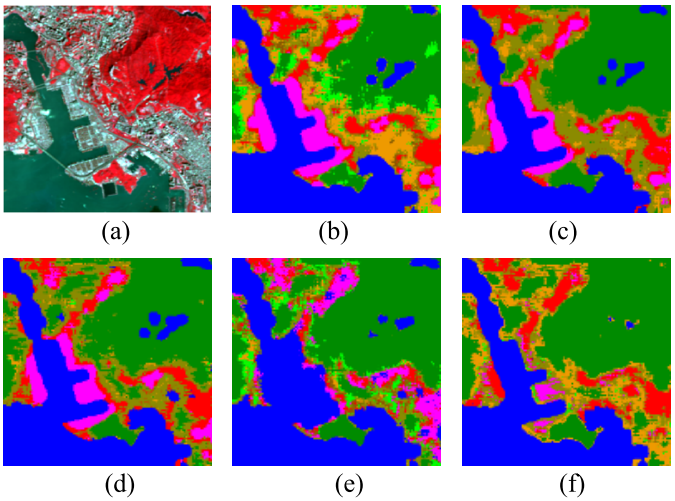

Fig. 2. Classifications of HMS contrastive learning under different ratios of training samples. The classes are bare land, buildings, grass, ground, road, vegetation, water, and others. (a) Image. (b) Ratio 1.0. (c) Ratio 0.4. (d) Ratio 0.2. (e) Ratio 0.1. (f) Ratio 0.05.

2) Comparison: Seven types of contrast, listed as unimodal Gaofen-5 (H), Landsat (M), ALOS-2 (S), double modalities, the Gaofen-5 and Landsat (HM), the Gaofen-5 and ALOS-2 (HS), the Landsat and ALOS-2 (MS), and all three modalities (HMS) are experimented to test our hypothesis. For classification, 5%, 10%, 20%, and 40% training samples are used for finetuning or training from scratch in classification. In addition, SOTA models, such as CLIP [23], DeCUR [24], and SimCLR [25], are compared.

3) Evaluation Metrics: Three commonly used metrics are utilized; they are overall accuracy (OA), average accuracy (AA), and mean intersection over union (mIoU).

B. Results and Discussion

Fig. 2 shows the classification results of three modality models, with the training ratios ranging from 5% to 100%. Generally, seawater, vegetation, and grassland are correctly categorized across all ratios of training data. However, for terrestrial water, underestimation tends to occur when the number of training samples is small, and overestimation occurs when the number of samples is large. More regrettably, the

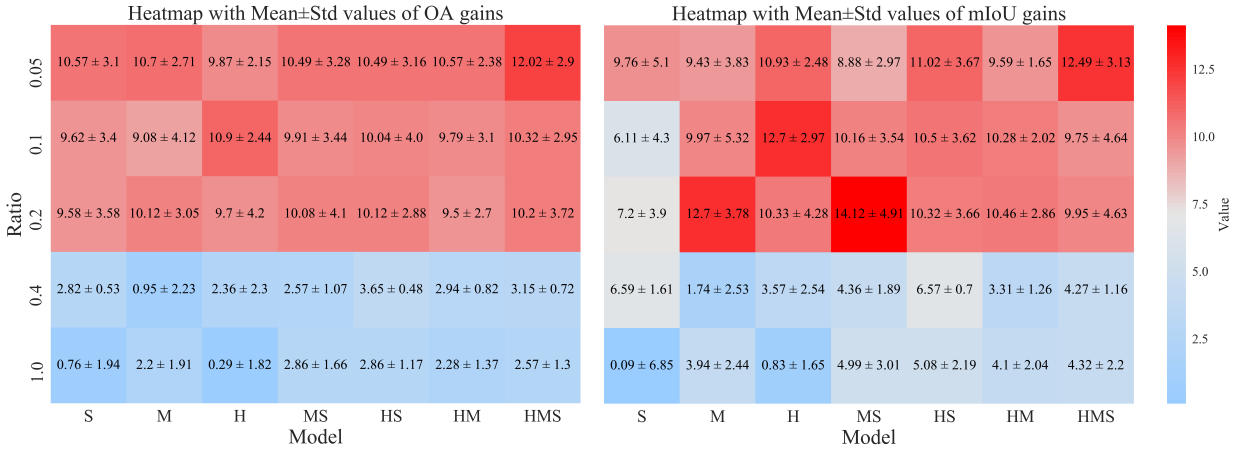

Fig. 3. Performance gains based on contrastive learning initialized models versus randomly initialized models, denoted as the average gains and standard variation of OA and mIoU.

TABLE I

CLASSIFICATION PERFORMANCE OF FIVEFOLD CROSS VALIDATION ON THREE MODALITIES HMS DATA

Types	Randomly initialized f					Contrastive learning weights of f				
Metric	Ratio					Ratio				
	5%	10%	20%	40%	100%	5%	10%	20%	40%	100%
OA	51.37± 4.26	59.74± 1.53	68.08± 3.06	84.29± 0.63	93.32± 1.70	63.39± 3.38	70.07± 3.15	78.28± 2.76	87.44± 1.25	95.90± 0.69
AA	49.81± 5.94	60.46± 3.06	68.40± 2.14	84.89± 0.90	93.24± 1.35	64.56± 3.83	70.62± 2.90	77.99± 3.30	87.67± 1.31	95.74± 0.55
mIoU	37.33± 5.62	47.90± 2.91	56.66± 3.39	74.76± 1.02	87.87± 2.59	49.82± 4.54	57.65± 3.04	66.61± 3.66	79.03± 1.84	92.18± 1.21

model fails to distinguish some artificial categories no matter what proportion of training data it is given, for example, the sea-spanning bridges are not recognized on all images.

The gain of the classification model initialized by contrastive learning compared to the randomly initialized classification results is visualized. As shown in Fig. 3, all numbers in the figure are displayed as percentages. The first number is the mean of the gain, followed by the standard deviation. In general, regardless of the proportion of training data, models that are initialized with contrastive learning tend to yield better classification results than randomly initialized models. When the ratio of training data is below 40%, the gain is even more prominent across all modality combinations, both in terms of overall (OA) and category correctness (mIoU). This proves the necessity of contrastive learning in terms of lower data amount. Another interesting finding is, considering contrastive learning from double modalities, the learning of the heterogeneous SAR data against the optical data, such as the hyperspectral and multispectral image. This may indicate the great potential of SAR data for multimodal contrastive learning.

With the training ratio increasing, contrastive learning may *lose* its edge to some extent, sometimes to the detriment of downstream classification tasks in our case. As shown in Fig. 3, when the ratio of training data changes from 20% to 40%, there is an obvious decrease in the classification performance gain, and the decrease is much clearer when the ratio continues to increase. When the ratio reaches 100%, the models initialized by contrastive learning may introduce performance degradation when there is only SAR or hyperspectral data. Fortunately, as the number of modalities increases, the degradation is relatively small. Given the success of contrastive learning in hyperspectral classification, the failure case of contrastive learning in our experiments may be triggered by a contradiction between the complexity of the classification model and the inadequately labeled data. However, when there are more modalities, the degradation seems to disappear, as the mean of OA is around 2.5%–3% and mIoU is around 3.3%–6.6%, both with small standard deviations. The increase in the number of modalities looks to lead to similar results to the inducement of more data, i.e., both can extend the benefits of contrastive learning.

Table I documents the performance of contrastive learning on classification tasks based on three modalities under different proportions of training data. The first is that when all training data are used for classification, there is not much difference between the models weighted by contrastive learning and random. They have more than 90% classification accuracy, in terms of OA and AA, and the small standard deviation also implies a more stable classification performance. What is impressive is that when the ratio of the training sample is below 40%, the network pretrained based on contrastive learning still achieves average results of 70% OA, with an average of 10% higher than the classification results of the randomly initialized network. Moreover, in most cases, the standard deviation from contrastive learning is close to the randomly initialized network, which means that the

TABLE II
COMPARISON WITH THE STATE-OF-THE-ART MODEL ON FULL TRAINING SAMPLES WITH THREE MODALITIES

Method	IoU per category							OA	mIoU
	1	2	3	4	5	6	7		
SimCLR	84.56±3.80	91.18±0.94	89.69±3.29	77.61±7.88	100±0.00	88.42±2.84	95.31±2.22	94.90±0.70	90.80±1.52
CLIP	85.88±4.73	92.08±2.62	89.49±3.62	75.39±8.93	100±0.00	87.04±2.73	93.47±3.90	94.65±1.44	90.37±2.70
DeCUR	84.42±2.95	91.05±1.43	91.03±5.57	70.97±8.49	100±0.00	87.36±1.63	93.52±4.34	94.36±0.65	89.74±1.52
Ours	90.16±5.38	94.91±1.03	90.54±3.07	78.29±3.69	98.95±2.11	90.52±1.56	94.08±2.73	95.90±0.69	92.18±1.21

classification results of the contrastive learning initialized model show similar robustness; it is also worth noting that when the modalities increase, the standard deviation becomes smaller for contrastive learning initialized models. Therefore, when it is difficult to obtain large amounts of labeled data, contrastive learning may achieve better results to some extent, as long as the modality is sufficient. Table II summarizes the comparisons between our method and three SOTA methods. In terms of categories, our method achieves better results than the other methods in the categories of bare land, building, ground, and vegetation and achieves at least 1% and 1.38% higher overall performance metrics in OA and mIoU.

IV. CONCLUSION

This letter proposes a framework for contrastive learning of multimodal data and validates it in real remote sensing application scenarios. In contrastive learning, images from different modalities are coupled to form the positive pairs; as the image bands vary from modalities, projection convolution is first utilized to normalize the image bands. Then, there follows a Siamese encoder to extract features from all modalities. The predictor is used to asymmetrically update the parameters of the contrastive learning model by minimizing the closed-loop multimodal similarity loss function. Experiments show that contrastive learning does work when there is less labeled data. However, as the amount of training data rises, it may not always reveal better results than a randomly initialized model, especially in a single modality. However, there is a reduction in the degradation when the number of modalities increases. The limitations of this letter lie in less labeled samples in the stringent application scenario setting. Moreover, data augmentation is not introduced into contrastive learning, and more diverse contrastive learning function f, *as well as classification function* g, are not considered, which can be the future direction.

REFERENCES

[1] R. Liu, L. Mi, and Z. Chen, "AFNet: Adaptive fusion network for remote sensing image semantic segmentation," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 9, pp. 7871–7886, Sep. 2021.

[2] B. Ren, B. Liu, B. Hou, Z. Wang, C. Yang, and L. Jiao, "SwinTFNet: Dual-stream transformer with cross attention fusion for land cover classification," *IEEE Geosci. Remote Sens. Lett.*, vol. 21, 2024, Art. no. 2501505.

[3] S. Xu and S. Fina, "National-scale imperviousness mapping and detection of urban land changes," *ISPRS J. Photogramm. Remote Sens.*, vol. 202, pp. 369–384, Aug. 2023.

[4] Y. Lin, H. Zhang, G. Li, T. Wang, L. Wan, and H. Lin, "Improving impervious surface extraction with shadow-based sparse representation from optical, SAR, and LiDAR data," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 12, no. 7, pp. 2417–2428, Jul. 2019.

[5] J. Ling, H. Zhang, and Y. Lin, "Improving urban land cover classification in cloud-prone areas with polarimetric SAR images," *Remote Sens.*, vol. 13, no. 22, p. 4708, Nov. 2021.

[6] Y. Wang, C. M. Albrecht, N. A. A. Braham, L. Mou, and X. X. Zhu, "Self-supervised learning in remote sensing: A review," *IEEE Geosci. Remote Sens. Mag.*, vol. 10, no. 4, pp. 213–247, Dec. 2022.

[7] C. Tao, J. Qi, M. Guo, Q. Zhu, and H. Li, "Self-supervised remote sensing feature learning: Learning paradigms, challenges, and future works," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5610426.

[8] R. U. Murshed, K. Noshin, M. A. Zakaria, M. F. Uddin, A. F. M. S. Amin, and M. E. Ali, "Real-time seismic intensity prediction using self-supervised contrastive GNN for earthquake early warning," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, 2024, Art. no. 5909119.

[9] C. Qiu, A. Yu, X. Yi, N. Guan, D. Shi, and X. Tong, "Open self-supervised features for remote-sensing image scene classification using very few samples," *IEEE Geosci. Remote Sens. Lett.*, vol. 20, 2023, Art. no. 2500505.

[10] M. Hu, C. Wu, and L. Zhang, "HyperNet: Self-supervised hyperspectral spatial–spectral feature understanding network for hyperspectral change detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5543017.

[11] N. Wang, H. Bi, F. Li, C. Xu, and J. Gao, "Self-distillation-based polarimetric image classification with noisy and sparse labels," *Remote Sens.*, vol. 15, no. 24, p. 5751, Dec. 2023.

[12] Z. Kuang, H. Bi, and F. Li, "Complex-valued self-supervised PolSAR image classification integrating attention mechanism," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, Jul. 2023, pp. 5958–5961.

[13] J. Bayrooti, N. Goodman, and A. Tamkin, "Multispectral contrastive learning with viewmaker networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2023, pp. 440–448.

[14] C. Tao, J. Qi, W. Lu, H. Wang, and H. Li, "Remote sensing image scene classification with self-supervised paradigm under limited labeled samples," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, pp. 1–5, 2022.

[15] S. Mao et al., "Adaptive self-supervised SAR image registration with modifications of alignment transformation," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5203715.

[16] K. Ayush et al., "Geography-aware self-supervised learning," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 10161–10170.

[17] Y. Chen, M. Zhao, and L. Bruzzone, "A novel approach to incomplete multimodal learning for remote sensing data fusion," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, 2024, Art. no. 5404914.

[18] K. Cha, J. Seo, and Y. Choi, "Contrastive multiview coding with electro-optics for SAR semantic segmentation," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, pp. 1–5, 2022.

[19] Y. Zhang, Y. Zhao, Y. Dong, and B. Du, "Self-supervised pretraining via multimodality images with transformer for change detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5402711.

[20] P. Jain, B. Schoen-Phelan, and R. Ross, "Self-supervised learning for invariant representations from multi-spectral and SAR images," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 15, pp. 7797–7808, 2022.

[21] Y. Wang, C. M. Albrecht, and X. X. Zhu, "Self-supervised vision transformers for joint SAR-optical representation learning," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, Jul. 2022, pp. 139–142.

[22] X. Chen and K. He, "Exploring simple Siamese representation learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 15745–15753.

[23] A. Radford et al., "Learning transferable visual models from natural language supervision," in *Proc. Int. Conf. Mach. Learn.*, 2021, pp. 8748–8763.

[24] Y. Wang, C. M. Albrecht, N. A. A. Braham, C. Liu, Z. Xiong, and X. X. Zhu, "Decoupling common and unique representations for multimodal self-supervised learning," 2023, *arXiv:2309.05300*.

[25] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *Proc. Int. Conf. Mach. Learn.*, 2020, pp. 1597–1607.