



# Multimodal Colearning Meets Remote Sensing: Taxonomy, State of the Art, and Future Works

Nhi Kieu, Kien Nguyen , Senior Member, IEEE, Abdullah Nazib , Tharindu Fernando , Clinton Fookes , Senior Member, IEEE, and Sridha Sridharan , Life Senior Member, IEEE

**Abstract**—In remote sensing (RS), multiple modalities of data are usually available, e.g., RGB, multispectral, hyperspectral, light detection and ranging (LiDAR), and synthetic aperture radar (SAR). Multimodal machine learning systems, which fuse these rich multimodal data modalities, have shown better performance compared to unimodal systems. Most multimodal research assumes that all modalities are present, aligned, and noiseless during training and testing time. However, in real-world scenarios, it is common to observe that one or more modalities are missing, noisy, and non-aligned, in either training or testing or both. In addition, acquiring large-scale, noise-free annotations is expensive, as a result, lacking sufficient annotated datasets or having to deal with inconsistent labels are open challenges. These challenges can be addressed under a learning paradigm called multimodal colearning. This article focuses on multimodal colearning techniques for RS data. We first review what data modalities are available in the RS domain and the key benefits and challenges of combining multimodal data in the RS context. We then review the RS tasks that would benefit from multimodal processing including classification, segmentation, target detection, anomaly detection, and temporal change detection. We then dive deeper into technical details by reviewing more than 200 recent efforts in this area and provide a comprehensive taxonomy to systematically review state-of-the-art approaches in four key colearning challenges including missing modalities, noisy modalities, limited modality annotations, and weakly paired modalities. Based on these insights, we propose emerging research directions to inform potential future research in multimodal colearning for RS.

**Index Terms**—Multimodal colearning, multimodal learning, remote sensing (RS), satellite imagery.

## I. INTRODUCTION

THE wide availability of multimodal sensors from satellite platforms such as Landsat and Sentinel, as well as off-the-shelf drone platforms like DJI, has significantly enriched the remote sensing (RS) community with a wealth of multimodal data. The richness in imaging mechanisms, spectral bands, and spatial, radiometric, and temporal resolutions of these sensors can complement each other to provide multidimensional mea-

surements and analysis of the Earth and other objects on it. Advances in machine learning, especially multimodal machine learning, have emerged as a powerful technique for RS applications [1]. By leveraging the complementary nature of multimodal data, multimodal machine learning approaches enable researchers to extract more comprehensive information from RS datasets. Multimodal machine learning techniques can facilitate many RS tasks such as semantic segmentation, target detection, change detection, and land cover mapping [2], [3], [4], [5]. The effective combination of the data from different modalities not only enhances accuracy and lowers false predictions, but also compensates for noisy data due to weather conditions and noise arising from other artifacts and allows machine learning models to tap into richer and more abundant data sources to deal with data scarcity and model generalization.

However, in the RS context, due to the challenges in data collection and airborne platform operation, there are multiple nonideal situations that can heavily impact the availability and quality of data from different modalities. First, one or several modalities can be missing in the training or testing phase of machine learning models [6], [7]. This is a widely occurring scenario in RS since the weather conditions such as clouds and rains can heavily impact optical data. For instance, if a model is designed to process both optical and light detection and ranging (LiDAR) satellite data, it may struggle to perform well when presented with LiDAR-only inputs during testing. Second, one of the input modalities could contain errors, outliers, or irrelevant information due to factors such as sensor inaccuracies, annotation errors, or environmental interference [8], [9]. Models trained on noisy data may produce inaccurate or unreliable predictions. To detect and remove noise, in addition to preprocessing techniques that could rectify the noisy inputs, specific model architectures should also be designed that would induce robustness to noisy inputs and this would increase processing complexity. Third, the lack of annotated data for training in RS, due to the cost and challenge of data collection and flying operation, can hinder the performance and generalization capabilities of machine learning models [10], [11]. When there is a scarcity of annotated training data, models would struggle to learn complex patterns and exhibit poor performance on unseen examples. Fourth, when different modalities [e.g., synthetic aperture radar (SAR) and hyperspectral] are not spatially and/or temporally aligned, it can introduce challenges in multimodal analysis tasks [12]. Besides aligning as a preprocessing step, models would have to be based on representation learning and

Manuscript received 26 November 2023; revised 21 January 2024 and 22 February 2024; accepted 13 March 2024. Date of publication 19 March 2024; date of current version 3 April 2024. (Corresponding author: Kien Nguyen.)

The authors are with the School of Electrical Engineering and Robotics, Queensland University of Technology, Brisbane, QLD 4030, Australia (e-mail: v.kieu@qut.edu.au; nguyentk@qut.edu.au; a.nazib@qut.edu.au; t.warnakulasuriya@qut.edu.au; c.fookes@qut.edu.au; s.sridharan@qut.edu.au).

Digital Object Identifier 10.1109/JSTARS.2024.3378348

Presence of Modalities	Fully Missing	Test	Methods	Reconstruction	AutoEncoder	
				GAN-based		
	Partly Missing	Train		Knowledge Distillation	Teacher-student	
		Test		Hallucination	Prototype	
			Multitask learning	Co-training / Co-learning	Multimodal Fusion	
Noisy Modalities	Data Noise	Adversarial Perturbations		Teacher-student		
		Common Corruptions		Adversarial networks (GAN-based)		
	Label Noise	Mimimization		Noise-aware objective function		
				Semi-supervised learning		
				GConv		
		Correction		Collaborative classifiers		
				Uncertainty-aware networks		
		Supervised label correction				
		Both				
Limited Modality Annotations	Semi-supervised	Inductive	Unsupervised pre-processing		Generative	
					Predictive	GAN-based, CNNs, Siamese, Clustering, Multitask learning
					Contrastive	
		Wrapper Method		Self-training	Pseudo-labels generation	
				Co-training	Multitask learning	
		Intrinsically Semi-supervised		Perturbation	Noise-tolerant objective function	
	Transductive			Manifolds	Multiple lower-dimensional spaces	
				Generative	GAN-based	
		Graph-based		Graph Convolutional Network (G-Conv)		
				Cross-scale graph prototypical network (X-GPN)		
		Global consistent G-Conv				
Weakly-supervised	Coarse-grain level annotations		Points, scribbles, polygons, blocks, image-level		GradCAM, Attention Map	
	Active learning, Zero-shot learning		User prompt			
	Multiple instance learning					
Unsupervised	Feature extractor (CAE, DCAE,...) followed by clustering algorithms (Kmeans, DBSCAN, GMM)					
	GAN-based such as CycleGAN					
Modality Parallelism	Parallel / Strongly paired		Co-registered images facilitate transfer or translation-based learning			
	Non-Parallel / Weakly paired		Weakly-supervised learning: conceptual grounding			
			Unsupervised learning			
			Translation-based methods support registration / multimodal alignment			
Hybrid / Bridge Data		Absence in remote sensing				

Fig. 1. Our proposed taxonomy for multimodal colearning research in RS.

should have architectural robustness to deal with the unpaired or weakly paired multimodal data.

Such nonideal scenarios and the required model complexity, as detailed above, are investigated under the umbrella of multimodal colearning techniques. Multimodal colearning is a hot topic in machine learning, especially in recent deep learning (DL) due to its requirement of very large datasets for training. While there are comprehensive survey papers covering the area of multimodal colearning such as [13]; it has not been explored in the RS context. Multimodal data within the realm of RS are significantly different from multimodal data in a general context. For example, in [13], the main multimodal data sources are vision and language. In the RS context, multimodal data sources are highly spatial, e.g., multispectral, optical and radar with different resolutions and levels of details, which requires unique strategies for multimodal colearning research. We are

missing a comprehensive study of these multimodal colearning techniques in the RS context.

This article aims to thoroughly and systematically investigate the multimodal colearning techniques in RS. It presents the first comprehensive study that not only summarizes the existing multimodal colearning research in RS but also compares and contrasts with the state-of-the-art multimodal colearning research in general in order to propose potential opportunities and future research directions. By reviewing more than 200 related papers, we propose a comprehensive taxonomy for multimodal colearning in the RS context, as shown in Fig. 1. In particular, the taxonomy focuses on four central aspects of multimodal colearning in the RS context: 1) the presence of missing modalities; 2) noisy modalities; 3) limited data annotations; 4) modality parallelism. This taxonomy will guide the review and help set future research directions for multimodal colearning in RS.

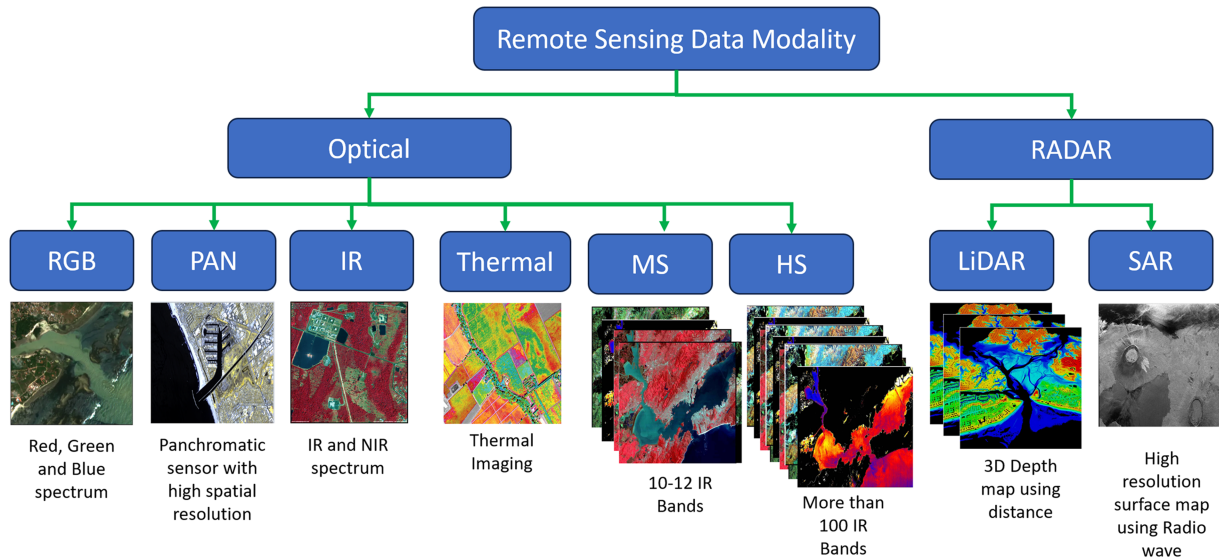


Fig. 2. RS data modalities with two main categories: Optical and radar.

The rest of this article is organized as follows. We provide the background and context for multimodal colearning by reviewing available data modalities, benefits, and challenges in multimodal learning for RS in Section II, and discussing RS tasks and their applications using multimodal data in Section III. After that, we dive deep into the four central challenges of multimodal colearning in the RS context, missing modalities, noisy modalities, limited annotation, and modality parallelism in Sections IV, V, VI, and VII, respectively. Based on these insights, we propose emerging research directions to inform potential future research in Section VIII. Finally, Section IX concludes this article.

## II. MULTIMODAL RS DATA MODALITIES

This section reviews what data modalities are available in the RS domain as illustrated in Fig. 2, the key benefits, and challenges of combining data from multiple modalities in the RS context.

### A. Optical Data Modalities

Optical data modalities are mostly imaging data captured using imaging sensors that capture the visible, and near-infrared (NIR) portion of the electromagnetic spectrum. As these data modalities are human perceptible, their use requires less pre-processing and can directly be used for several purposes such as land use classification, vegetation analysis, environmental monitoring, etc. Despite their ease of use, drastic environmental conditions like cloud and poor illumination can significantly hinder their capacity. Some common types of optical RS sensors are as follows.

1) *RGB*: RGB sensors are inbuilt in most of the RS data sources, hence most multimodal datasets have RGB channels available along with other spectral channels. The RGB channels help to identify the presence of fog, dust, fire, clouds, volcanic ash, etc., with bare eyes. Several such datasets include PASTIS-2 [14], Agricultural-Vision [15], LandCoverNet [16], IDIVAsia [17], etc.

2) *PAN*: Pansharpening is a commonly used technique in RS to improve the spatial and spectral resolution of the images. In RS, panchromatic sensors collect images with high spatial resolution while multispectral/hyperspectral (MS/HS) sensors collect low spatial but high spectral resolution. Combining these two modalities produces images with high spatial as well as spectral resolution, which helps improve downstream tasks. Some of the pansharpening datasets include [18], [19], etc.

3) *IR*: Infrared (IR) imaging is one of the major data sources for RS. The location of the IR spectrum is in between the visible and microwave regions, which is also referred to as the heat region. Based on the wavelength, the IR spectrum is further classified as NIR (0.78–2.0  $\mu\text{m}$ ), short-wave IR (SWIR) (1.4–3.0  $\mu\text{m}$ ), and long-wave IR (LWIR) (8.0–15.0  $\mu\text{m}$ ), etc.

4) *Thermal Imaging*: Thermal imaging refers to the imaging of the IR and thermal radiation of object body. In RS, the thermal IR region of the wavelength, which ranges from 3 to 15  $\mu\text{m}$ , is used for imaging the thermal radiation. Depending on the emissivity of the radiating objects, the thermal imaging sensors can discriminate objects. Common thermal imaging data sources include Landsat-7 and Landsat-8 [20], MODIS [21], ASTER [22], Sentinel-3 SLSTR [23], etc.

5) *Multispectral and Hyperspectral*: Multispectral RS sensors capture 3–10 bands, therefore, each image data contain 3–10 different channels. The hyperspectral, on the other hand, captures hundreds to thousands of bands, and the bands in hyperspectral are much narrower than multispectral. Landsat-8 [20] and USGS Earth Explorer [24] are the two common sources of multispectral and hyperspectral data.

### B. Radar

Radio detection and ranging (RADAR) RS sensors use high-frequency radio waves or microwaves to capture information about the Earth's surface. RADAR sensors are designed to



overcome the limitations of optical sensors for handling particular situations such as cloud cover, darkness, and the need for all-weather capabilities. Radar sensors are commonly found on satellites and aircraft and are used in a variety of applications, including mapping, agriculture, disaster monitoring, and more. Here are some common types of radar RS sensors.

1) *LiDAR*: LiDAR is an RS device that utilizes pulsed laser to detect and measure variable distances (ranges) to the Earth or any other targeting object. Usually, LiDAR devices are mounted on airborne systems to collect data and the collected data produces a precise 3-D map of the Earth's surface. IEEE data fusion challenges [25], [26] are two renowned sources of multimodal data. Other public LiDAR data sources include USGS Earth Explorer, OpenTopography [27], NOAA Digital Coast [28], etc.

2) *SAR*: SAR is a moving radar that emits radio waves to capture the structural properties of the targeting object. Due to the use of radio waves as an energy source, SAR can work in different climate condition that includes dust, haze, cloud, etc. Some of the common SAR data sources are Sentinel-1 [29], ALOS-1 [30], EAR-1,2, ENVISAT [31], etc.

### C. Benefits of Multimodal Processing

In the preceding section, we explored various RS data modalities. Leveraging multimodal data for training DL models offers significant advantages, including heightened accuracy, improved reliability, and enhanced generalization capabilities [32]. Multimodal data are particularly advantageous in scenarios marked by dynamic or extreme conditions, where one modality might miss crucial properties that can be effectively compensated for by other modalities. Harnessing this complementary information empowers models to achieve lower false positive rates and, consequently, superior accuracy. Moreover, multimodal training can mitigate the challenges posed by data scarcity, which is sometimes encountered when acquiring a specific modality of data, especially if it is costly or scarce. By amalgamating multiple modalities, DL models can tap into richer and more abundant data sources, effectively mitigating the impact of data scarcity. Furthermore, multimodal colearning plays a pivotal role in addressing the intricate challenge of domain generalization within the DL community, resulting in improved downstream processing and more versatile models capable of adapting to diverse domains.

### D. Challenges in Multimodal Processing

The process of multimodal colearning presents a multitude of challenges. Integrating data from different sensors can be complex due to variations in sensor characteristics and alignment issues. Furthermore, varying data quality, noise levels, and missing data across modalities require careful handling, necessitating informed decisions on data weighting and preprocessing [33]. Consequently, there is a need for a well-defined and streamlined data preprocessing pipeline before initiating multimodal model training, which can increase operational overhead. The high-dimensional feature spaces often resulting from multimodal data pose challenges, including increased computational complexity and the risk of overfitting, highlighting the importance of

dimensionality reduction techniques and feature selection in colearning. Addressing dependencies and correlations between modalities is crucial for achieving optimal model performance. In addition, annotating multiple modalities can be challenging and costly, particularly for some modalities, making the acquisition of accurate labels reflecting synergistic information a demanding task. Lastly, the utilization of multimodal data can introduce unexpected bias, potentially leading to biased model predictions. Detecting and mitigating this bias before model training represent significant challenges in the realm of multimodal colearning.

## III. MULTIMODAL RS TASKS AND APPLICATIONS

This section discusses multimodal tasks in RS: segmentation (also known as pixelwise classification), target detection, change detection, anomaly detection, and other applications. In each task, we highlight the multimodal nature of them, how data from multiple modalities complements each other, and the applications of these multimodal tasks.

### A. Multimodal Satellite Imagery Segmentation

Image segmentation, also known as pixelwise classification, is a popular task within computer vision where pixels of the image are partitioned or classified into distinct and meaningful regions, segments, or objects. This task has gained significant popularity within the RS community as well due to its myriad of applications ranging from disaster assessment and management to vegetation analysis. The goal of segmentation is to identify and delineate different objects or land cover types within the remotely sensed image, such as buildings, roads, vegetation, water bodies, and more. Compared to segmenting images captured in a terrestrial setting, segmentation of remotely sensed data offers unique challenges. The variation in lighting conditions, poor image resolution, large field of view, and the smaller size of the objects present difficulties. Multimodal satellite imagery segmentation is the setting where multiple satellite sensor data that have been captured from the same or different satellites are integrated to offer a richer data source. This offers several advantages over unimodal satellite image segmentation. Specifically, multimodal data are often helpful in resolving ambiguities that could arise in a unimodal setting. For instance, optical and radar data are useful for distinguishing buildings from trees in monitoring urban areas. Similarly, multimodal sensing could be useful for making the segmentation framework more resilient to adverse atmospheric and weather conditions as different sensors have varying levels of sensitivity to environmental conditions.

In spite of the numerous advantages that multimodal sensing offers to satellite imagery segmentation, it also presents several challenges that should be addressed when designing the framework.

- 1) The lower spatial resolution of the target objects in the remotely sensed image could hinder the performance of fine-grained image segmentation. Furthermore, the multimodal data are captured from multiple sensors with different spatial, spectral, and temporal resolutions, as such, the



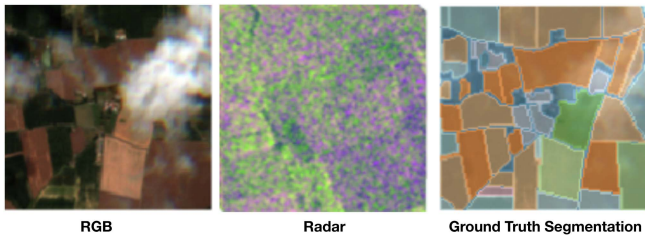


Fig. 3. RGB bands and radar (SAR) modalities of a sample from the PASTIS-R [34] dataset and ground truth segmentation mask.

alignment and integration of distinct data sources while ensuring consistency is complex and challenging.

- 2) As the data are captured from heterogeneous sensors, they capture different representations in the form of different scales, units, and ranges. Therefore, normalization strategies should be carefully defined to bring all the sensor data into the same dynamic range.
- 3) Acquiring a high-quality multimodal RS dataset using satellite or aerial platforms poses several challenges. Apart from the cost associated with obtaining such data, the distinct revisit times of different satellites could constrain the types of modalities that can be incorporated into the multimodal segmentation framework.

Fig. 3 provides visualizations of RGB bands and radar modalities from the PASTIS-R [34] dataset and ground truth segmentation mask. Some popular applications within multimodal RS image segmentation include crop type classification [7], [34], [35], [36], urban area monitoring [37], [38], [39], land use and land cover classification [40], [41], [42], [43], [44], [45], and environment monitoring [46].

### B. Multimodal Satellite Imagery Target Detection

Multimodal target detection attains the identification and localization of single or multiple targets within the multimodal data captured using multiple RS sensor modalities such as optical satellite imagery, IR, radar, and LiDAR. Similar challenges as in multimodal satellite image segmentation, including, the challenges with respect to different spatial, spectral, and temporal resolutions in different sensing platforms, the need for data normalization, challenges in data transmission, storage, and computation due to increased dimensionality, and the scarcity of high-quality multimodal RS datasets exists within the multimodal satellite imagery target detection task. Furthermore, task-specific challenges also exist which could be summarized as follows.

- 1) The variations in the appearance of the target objects across different sensing paradigms pose a challenge in multimodal satellite imagery target detection. Compared to the multimodal target detection with terrestrial data, in satellite imagery, the objects may cover only a few pixels in the entire image, as such the detection algorithm should have sufficient capacity to understand and compensate for these variations.
- 2) In addition to variations across modalities, variations across the appearance of the object across time also pose

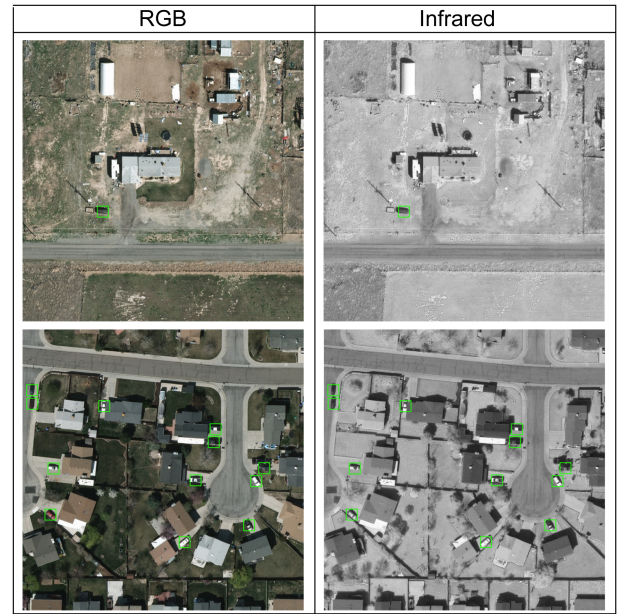


Fig. 4. Examples of multimodal satellite image target data from VEDI dataset [50] and ground truths.

another challenge. The movement of the target objects, changes in appearance, shape or orientation, and changes in capture conditions including atmospheric conditions could render these variations and the machine learning models should possess sufficient intelligence to accommodate these variations.

Example of applications of multimodal satellite imagery-based target detection include the multimodal vehicle detection algorithms proposed in [47] and [48], which could detect 11 classes of vehicles in different backgrounds, including grass, highway, mountains, and urban areas, with the aid of RGB and IR imaging modalities. Furthermore, Sakla et al. [48] investigated the cross-modality knowledge transfer capabilities, which demonstrated the ability to transfer the learned semantics from one modality to another during the modal training such that the trained model could be evaluated even under the unimodal setting. In a different line of work, a combination of satellite attitude and orbit information is leveraged in [49] to filter candidate image patches from the original large-scale satellite image data. Then, a texture analysis on the filtered patches is performed to detect the target objects. This algorithm has been evaluated for aircraft detection from satellite images. In Fig. 4, we provide some sample RGB and IR images that were taken from the areal image-based vehicle detection benchmark named VEDI [50].

### C. Multimodal Satellite Imagery Change Detection

The change detection algorithms are tasked with identifying significant changes in two or more observations that were taken at different times. Visible, IR, SAR, and multispectral and hyperspectral imaging platforms are readily used modalities for change detection. Each of these modalities is capable of capturing the unique characteristics of the observed geospatial location. For instance, the SAR can penetrate clouds and utilizes

longer wavelengths that range from centimeters to meter scales. As such they are capable of accurately capturing the topological information from the Earth's surface. In contrast, hyperspectral imaging could generate information regarding the material properties of the spatial observation. Therefore, the multimodal change detection algorithms could enhance their accuracy by combining them.

In addition to the challenges discussed in the previous applications, multimodal satellite imagery change detection involves several application-specific challenges.

- 1) The change detection algorithms compare and contrast multiple images for detecting changes. However, in satellite image-based change detection, these images should be registered and aligned as they are captured using satellites in different poses. Furthermore, the registration of high-resolution, multimodal satellite images could be computationally expensive and prone to errors, and these errors could lead to false-positive or false-negative change detection.
- 2) Temporal synchronization poses another challenge with respect to multimodal satellite imagery change detection. This is because the multimodal observations are typically captured from different satellites, as such, the distinct observations should be properly synchronized with respect to time as well as accurately calibrated for different atmospheric and other environmental conditions prior to using them in the change detection algorithm.
- 3) Capturing large-scale datasets for model training in the multimodal satellite imagery change detection domain is particularly challenging. In addition to the cost and complexities associated with procuring multiple satellites, change detection is performed across large spatial areas and over long time horizons. Therefore, annotating such large-scale datasets is time-consuming and costly.

The existing applications of multimodal satellite change detection include flood detection and monitoring [51], landslide detection [51], farmland reclaim [52], urban construction monitoring [52], and forest fire monitoring [52], [53]. For instance, Radoi [51] utilized Sentinel-1 and Sentinel-2 satellite image time series for monitoring river fires in the Save River in the South East of Africa. This analysis is conducted using the SEN12-Flood [54] dataset, which contains both SAR and optical images. In a supplementary evaluation, Rambour et al. [54] again used Sentinel-1 SAR and Sentinel-2 visible (B2, B3, B4) bands from Alunu, Valcea district in Romania to detect a landslide that occurred on 15 May 2017. Similarly, Chen et al. [52] used two SAR images that were captured by Radarsat-2 in 2008 and 2009 for analyzing the changes in farmlands. Even though the two images were SAR images, the first image was a single-look image, while the second image was a four-look one, hence, mimicking different modalities. To monitor river bank erosion caused by flooding at the Yellow River in Estuary Chen et al. [52] leveraged an SAR image, which was captured by Radarsat-2 in 2008, and a panchromatic image was captured by Landsat-7 in 2010. The authors also used two multispectral images but with different band configurations that were captured using Landsat-5 (seven bands) and EO-1 ALI (ten bands) to

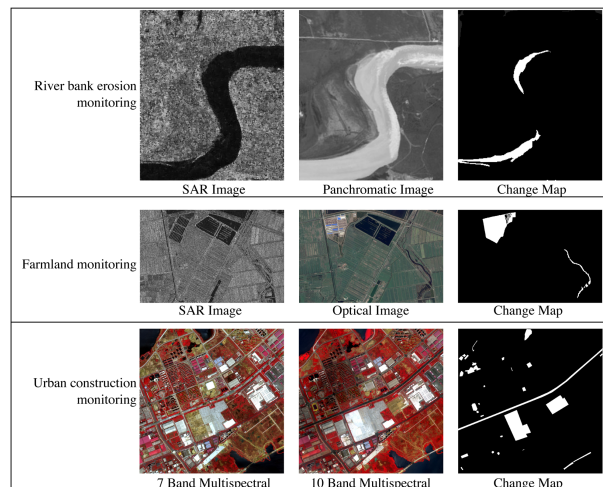


Fig. 5. Samples from the river bank erosion detection, farmland monitoring, and urban construction monitoring datasets used in [52]. Recreated from [52].

detect the impact of forest fires in Texas, USA. The Shuguang dataset used in [52] consisted of SAR and optical aerial images and had been used to monitor changes in Shandong Province, China. Samples from the datasets used in [52] for river bank erosion detection, farmland monitoring, and urban construction monitoring applications are illustrated in Fig. 5.

#### D. Multimodal Satellite Imagery Anomaly Detection

Multimodal satellite imagery provides the required levels of complimentary information to identify unusual or anomalous patterns or objects within the remotely sensed imagery. Anomaly detection is referred to as the process of identifying data points or patterns that deviate significantly from the expected norm, and within the context of RS, this could include but is not limited to unusual objects, events, or changes on the Earth's surface. Atmospheric conditions such as weather conditions, cloud cover, as well as other noise artifacts, and sensor-specific distortions can be mistaken for anomalies as such, distinguishing the fine-grained long-term anomalous changes without being confused by natural variations, artifacts, and distortions is challenging. The complementary information provided by multimodal observations enhances the anomaly detection capabilities. Despite these advantages, multimodal satellite imagery anomaly detection also possesses its own challenges. In addition to the common challenges that exist across all the multimodal satellite imagery analysis applications, including, the variability of resolution, quality, and sensor types across the modalities, and the higher dimensionality of the multimodal imagery that needs complex hardware for processing, the anomaly detection application poses several challenges.

- 1) While a typical satellite imagery covers hundreds of kilometers of the Earth's surface in a single image, only a small proportion of the total pixel data contains anomalous information, leading to class imbalances. This could affect the robustness, generalization, and reliability of the machine learning models as they tend to get biased toward the majority class.



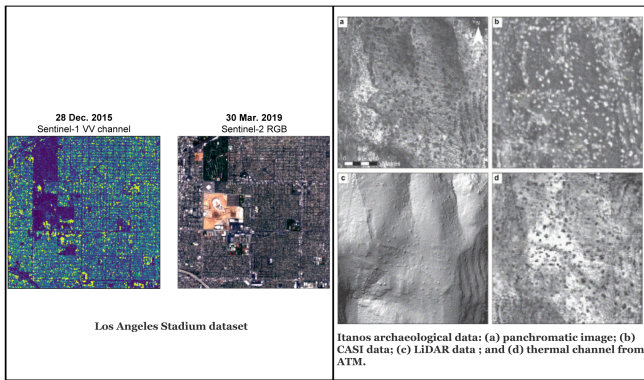


Fig. 6. Samples from the Los Angeles Stadium dataset used in [56] and the archaeological dataset proposed in [55].

- 2) Real-time performance is another challenge, especially when handling higher dimensional multimodal satellite imagery. Satellite imagery-based anomaly detection is typically used in defense and environment monitoring applications, which require instantaneous responses and higher dimensionality of the multimodal data requires substantial time for data downlink across the limited communication bandwidth. A solution is to process the captured multimodal satellite imagery on-board in the satellite but the limited hardware available on the satellite as well as the restricted energy profile limits the computations that can be performed using higher dimensional multimodal data.

There exist numerous applications of multimodal satellite imagery abnormality detection. For instance, Rowlands and Sarris [55] have leveraged a compact airborne spectrographic imager (CASI) that captures visible and NIR portions of the spectrums, an airborne thematic mapper (ATM) to capture soil properties and heat capacity using the short wave IR and the thermal portions of the spectrum and a LiDAR sensor to measure topological properties. Using these diverse modalities, Rowlands and Sarris [55] have investigated the possibility of locating exposed and known buried archaeological remains. In a different line of work, Ziemann et al. [56] have combined SAR imagery from Sentinel-1 and multispectral imagery from both Sentinel-2 and Landsat 8 for the detection of anomalous changes in the environment during the construction of Los Angeles Stadium at Hollywood Park in Inglewood, CA. Rodger et al. [57] have proposed a framework to fuse SAR data captured from Sentinel-1 together with vessel trajectories extracted using the automatic identification system (AIS) for the detection of abnormalities in maritime surveillance. Samples from the datasets used in [56] and [55] are visually illustrated in Fig. 6.

### E. Other Applications

Apart from these popular applications of multimodal RS, there exists an additional set of diverse applications in which multimodal satellite imagery has excelled over their unimodal

counterparts. For instance, Florath et al. [58] have used multispectral data to monitor changes in snow and ice types on the glacier surface. Furthermore, Saad et al. [59] proposed to fuse Sentinel-2 and Landsat 8 data for leveraging a broader set of multispectral bands for the location of prospective hydrothermal mineral deposits. Li et al. [60] designed an unsupervised fusion network to fuse hyperspectral and multispectral data. In a different line of work, Xia et al. [61] fused RGB, hyperspectral, and ground control points (GCPs) measured using a real-time kinematic (RTK) receiver to discriminate between susceptible and resistant weeds in agriculture. Similarly, Quan et al. [62] combined hyperspectral, LiDAR, and GPS data to investigate the relationship between weed comprehensive competition indices and crop parameters. On the other hand, we are transitioning to general multipurpose architectures such as SpectralGPT [63], in which a foundation model serves different use cases by changing downstream heads leveraging features from diverse large-scale datasets.

## IV. PRESENCE OF MISSING MODALITIES

The primary objective of multimodal machine learning is to create models that use information from multiple modalities to achieve higher accuracy than unimodal models. The researcher's focus has led to the development of state-of-the-art multimodal DL models. However, in real-world scenarios, it is not always possible to have all modalities available during training and testing. The prevailing approach to address the incomplete modality problem involves the utilization of multiple streams to encode complementary information from distinct modalities. In addition, diverse fusion strategies are explored for effectively combining features from these streams. Moreover, the integration of advanced loss functions, capable of serving multiple objectives, plays a crucial role in effectively managing missing modality challenges. Designed architectures fall into three typical approaches—reconstruction, knowledge distillation (KD), and multitask learning, as depicted in Fig. 7. On the other hand, Rahate et al. [13] outlined seven cases of missing modality, we streamlined it into three cases: 1) fully missing at test, 2) partially missing at train, and 3) partially missing at test. This helps the analysis generalize to many scenarios. For example, one or more modalities can be partially missing at train while one or more modalities can be fully missing at inference time. Table I shows representative literature addressing missing modalities in RS.

### A. Reconstruction

Missing modality reconstruction is one of the most intuitive methods where given features from available modality, the network will synthesize unavailable signals and parse all features to make the prediction. In other words, in addition to a main downstream task, the network is also trained to translate one type of signal to another. The idea is widely adopted from conventional methods to DL ones. Non-DL methods [65], [66] led the way in proving that supplementary information from missing modality is useful resulting in better performance



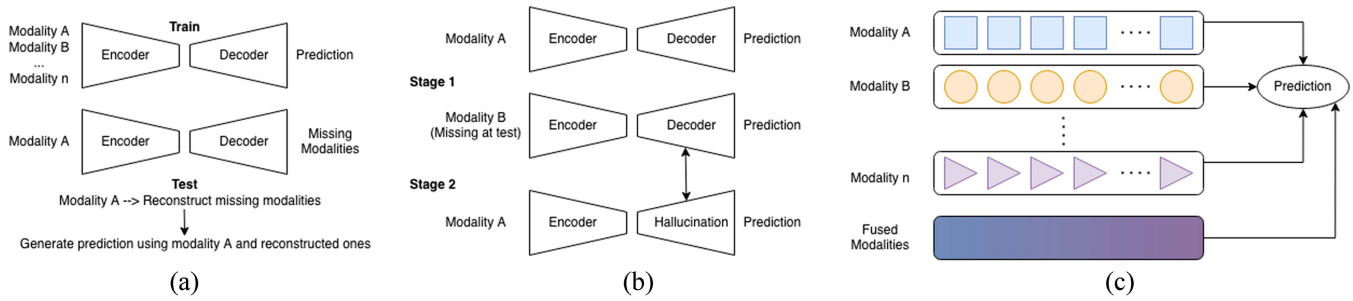


Fig. 7. High-level illustration of three methods dealing with missing modality. The reconstruction approach involves learning to translate a main modality to others so that, at test time, the missing ones are reconstructed given the main modality and all modalities are parsed through the learned backbone. KD methods often have multiple stages, where the first one performs learning using all modalities and later stages introduce hallucination branches in which a main modality learns to mimic the predictive capacity of missing ones. The multitask learning framework proposes a multibranch architecture allowing modality-specific and modality-correlated information to be harnessed simultaneously, which makes dealing with arbitrary combinations of modalities at inference as straightforward as excluding the modality-specific branch. (a) Reconstructed. (b) Knowledge distillation. (c) Multitask learning.

TABLE I  
REPRESENTATIVE METHODS DEALING WITH MISSING MODALITY IN RS

No	Pub	Datasets	Missing			Approaches			Modality	Backbone
			Fully Test	Partially Train	Partially Test	RC	KD	MT		
[81]	IGARSS 2017	Vaihingen, Potsdam	✓		✓		✓		RGB-NIR, Depth	CNNs
[77]	CBMI 2018	Urban Mapper 3D	✓	✓		✓			RGB, Depth	SegNet
[82]	ICCVW 2019	MS-PAN, Houston, Indian Pines	✓				✓		MS-PAN, HS, Depth	CNNs, C-GAN
[83]	TGRS 2021	PoDelta, Libourne	✓				✓		RGB, SAR	AlexNet
[12]	JPRS 2021	MSAW	✓				✓		RGB, SAR	DeepLabv3
[84]	TGRS 2023	Houston, Augsburg	✓				✓		HS, MS, Depth, RGB	CNNs, Transformers
[85]	JPRS 2023	URFC	✓					✓	RGB, User visit	ResNet, SPP, LSTM
[86]	NRCP 2023	PAN-MS, Pavia	✓				✓		MS-PAN	CNNs
[87]	TGRS 2023	MSAW	✓				✓		RGB, SAR	CNNs
[64]	2023	DFC2013, LULC, DFC2013	✓					✓	RGB, SAR, Depth	Transformers, BiLSTM

Three main groups are reconstruction (RC), KD, and multitask (MT). Downstream tasks include semantic segmentation, object detection, and classification encoded by the color of dataset name.

compared to a unimodal model. They often involve classic matrix/kernel completion using probabilistic methods, which define an early form of feature reconstruction category. With the development of deep neural networks, reconstruction gets better performance using fully connected and convolutional neural network (CNN) [67], [68], [69], in which autoencoder architecture [70] is the core mechanism for missing modality imputation. The cascaded residual autoencoder [68] still served as a strong foundation for more modern approaches [69], [71]. Nonetheless, disparate modalities often possess inherent differences that pose challenges in encoding their information using similar encoders. For instance, while CNN excels at capturing spatial features, it encounters difficulties in grasping long-range dependencies found in natural language or sequencelike modalities. To address this, encoder branches can be tailored to enhance modality-specific performance [6], [72], [73]. For example, Wang et al. [6] demonstrated that multimodal feature distribution alignment could be used to learn shared features through optimization, while an auxiliary task, such as modality classification, could yield modality-specific features. In addition to autoencoder architecture, generative adversarial network (GAN) [74], comprising a discriminator and generator, also have emerged as a suitable and extensively utilized method

across various disciplines for tackling the missing modality challenge [75], [76], [77], [78], [79], [80]. One of the notable methods among them is the SMIL framework, in which Ma et al. [79] utilized meta-learning to optimize flexibility and efficiency in a severe missing modality setting. In such a setting, the authors observed that missing modalities could be either in training, in testing, or in both, and most training data could have incomplete modalities. As such, a Bayesian meta-learning-based solution is devised to uniformly achieve these two diverse objectives.

Remote sensing: In the study by Bischke et al. [77], a generative approach utilizing GANs was employed to address the issue of missing depth information in the context of building footprint detection. The research highlights two key findings: first, models utilizing partial depth features outperform those relying solely on RGB data; and second, the introduction of reconstructed depth data leads to improved model performance, even though it falls short of the performance achieved when all modalities are available. However, it is important to note that reconstruction methods exhibit limited flexibility due to their explicit translation mechanism, which restrains the exploration of diverse input modality combinations, particularly when dealing with more than two modalities. Furthermore, challenges

arise in capturing features across modalities within one-to-one translation relationships.

### B. Knowledge Distillation

The KD [88] approach aims to transfer knowledge from one process to another. In its early stage, this approach focuses on harnessing useful features from closely related datasets [89]. Subsequently, in the context of addressing missing modality challenges, KD has evolved to encompass the transfer of knowledge from one modality to another(s), which is achieved through the implementation of a distillation loss or hallucination network. A classic method is the teacher–student paradigm, wherein one or more teachers attempt to learn valuable features from the data derivatives [71], [83], [90], [91] or related datasets [12]. The student network can be simultaneously trained, leveraging additional information from the teacher(s), and later deployed for inference. Both distillation loss and hallucination streams share the common objective of encouraging the student network to mimic the behavior of the teacher(s). Instead of using actual ground truth labels in the downstream task, soft labels, or logits are employed. Nonetheless, the utilization of the hallucination method might introduce inefficiencies, as it often needs one or more dedicated network branches to achieve its objective. This becomes particularly pronounced when dealing with more than two modalities and different combinations of them are collected [81], [84], [92]. In contrast, distillation loss incorporates the knowledge transfer process into the main downstream task. In addition, architectures utilizing hallucination streams often require multistage training rather than an end-to-end approach [82], [93], [94]. Variations of CNN architecture dominate in this category. Although recurrent neural networks (RNNs) and bi-LSTM are expected to work better with sequence data; limited literature explored those options for KD.

Remote sensing: Pande et al. [82] integrated GANs into a teacher–student training paradigm. On the other hand, the works of Wei et al. [84] and Li et al. [83] underscored the utility of hallucination branches. Differing from prior methodologies, Kampffmeyer et al. [81] presented an approach capable of handling three modalities, although it introduces more one-to-one translation branches (i.e., RGB-IR and RGB-Depth, assuming RGB as the dominant modality). Fernando et al. [7] presented a KD framework in which a teacher model with multiple modalities (i.e., multispectral and radar) and time-series input is employed to provide guidance to a lightweight student network that operates in a single modality. The student network only receives a single frame from the multispectral time-series data. Even though a separate hallucination network is not used, a feature-level distillation loss is used to guide the student to generate a rich feature vector that is similar to the multimodal teacher with time-series input. Similarly to the reconstruction category, the aforementioned methods in KD face challenges in terms of flexibility and efficiency, due to the presence of hallucination streams and a multistage training protocol. Some KD studies exclusively rely on CNNs, which possess known limitations in capturing expansive receptive fields and facilitating effective cross-modal interactions. Both reconstruction and KD techniques often require all modalities

during training; hence, imposing further limitations on their adaptability.

### C. Multitask Learning

Unlike the reconstruction method, multitask learning does not require the missing modality to be reconstructed explicitly. On the other hand, it does not require dedicated branch(es) for translation/hallucinationlike KD. However, multitask learning architectures also have multiple branches, each of which is associated with distinct loss to encourage the network to learn valuable modality-specific and -correlated representations across different modalities. Each modality can have its dedicated branch, and a fusion component is employed, which can be a separate branch [64], [112] or adopt immediate/late fusion techniques [113], [114]. Fusion strategies can have a significant impact on model performance depending on the nature of the data [115]. Transformers [116] with cross-attention mechanism is promising to provide a generalizable solution for multimodal fusion [117], [118], exploiting cross-modal features. In contrast to reconstruction, which can be computationally expensive, multitask learning benefits from the capabilities of Transformers to learn correlated information, thereby circumventing the explicit generation of missing modalities [64], [114], [119], [120]. On the other hand, in multitask learning, engineering appropriate loss functions are crucial, as neglecting this aspect may lead the network to exhibit biases toward dominant modalities while disregarding complementary ones. These auxiliary losses are also called regularizers [114], [119].

Remote sensing: Multitask learning emerges as a more flexible alternative, tolerating missing modalities during training and enabling the leveraging of any available information. Consequently, this approach can be generalized to various combinations of modalities, making it highly promising for addressing missing modality challenges. Even though multitask learning has gained recognition in medical image processing [6], [93], [114], [119], [120], [121] and dense prediction task—semantic segmentation in other domains [91], [94], its potential in RS remains largely unexplored. One of the pioneers in RS proposing a multitask architecture was [64], which combined masked autoencoders (MAE) [122] and contrastive learning to learn modality-specific and modality-correlated features simultaneously. As shown in Fig. 8, it outperforms MultiViT in all modality-incomplete cases, some of which by a large margin. The performance differences are about 30% when tested with missing DSM (SAR, RGB) and RGB only. A similar pattern is observed in the LULC dataset. On the other hand, the majority of existing literature addresses missing modalities during inference, with limited work investigating incomplete modalities at both the training and testing stages. Hence, applying multitask learning to RS to flexibly deal with missing modalities is an appealing research direction.

### D. Discussion

In order to deal with missing modality, there are three popular frameworks: reconstruction, KD, and multitask learning. Table I summarizes some representative studies for the presence of

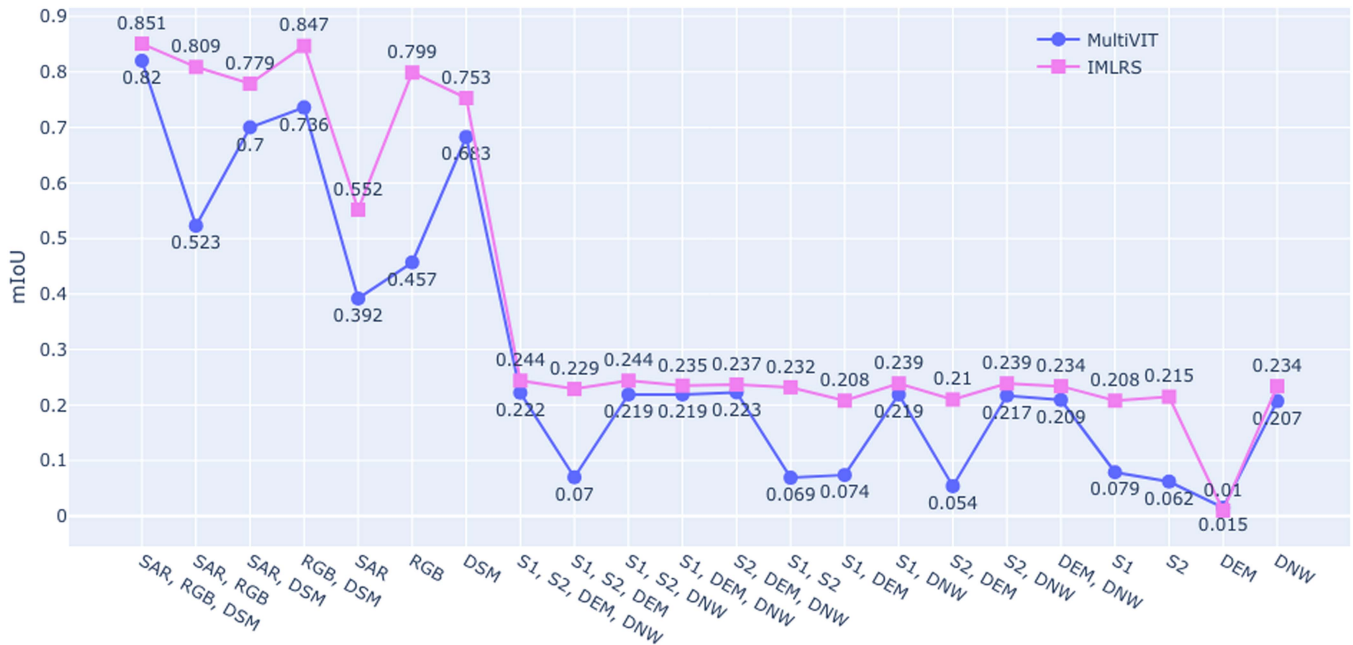


Fig. 8. Effectiveness of multitask learning in situations where one or more modalities are missing at inference. The IMLRS approach is proposed in [64]. MultiViT is the baseline without a random modality combination strategy. The initial seven data points represent models’ performance on DFC2023 dataset, the remaining points reflect results evaluated on LULC dataset.

modalities in RS with various downstream tasks (e.g., semantic segmentation, building extraction, scene classification, etc.). Existing literature is conducted with different use cases, which makes it hard to have direct comparisons between them. However, from an architectural design standpoint, multitask learning exhibits considerable promise. This is particularly true with the advancement of attention mechanisms, which offer notable advantages in terms of effectiveness, efficiency, and flexibility in contrast to the other two approaches.

Notably, the performance gain of the same architecture can significantly vary when datasets change (see Fig. 9), as observed in instances such as Hall Net (Vaihingen, Potsdam), DH-ADNet (PoDelta, Libourne), MSH-Net, and social sensing (URFC-A, URFC-B). Fig. 9 also shows that proposed approaches generally enhance the performance of unimodal models by 2%–4% in cases of missing nondominant modalities. However, C-GAN distillation and social sensing show no improvement. The authors of the former paper pointed out that there is limited correlated information between MS and PAN bands. Although the performance gap diminishes when employing Indian Pines and Houston datasets, the proposed student networks exhibit lower accuracy compared to unimodal branches. Similarly, the authors of the Social Sensing paper acknowledge the advantages of modality-specific classifiers over their proposed architecture. These outcomes offer valuable guidance for future research in the field of multimodal colearning in RS. On the other hand, while some models remain robust in missing dominant modality scenarios (e.g., DH-ADNet, DiffAlignOp, MSH-Net), others are either not evaluated or are no longer effective. Besides, the majority of proposed models aim to close the gap with full modality baseline solutions, meanwhile, C-GAN distillation and social sensing have the most significant gaps. Only

MSH-Net manages to surpass the full modality baseline in specific datasets. There are some factors that need to be considered when assessing the relative performance gains. The variation in evaluation metrics across reviewed papers, although they are popular in the field (e.g., IoU, Accuracy, F1), introduces some level of ambiguity. The effectiveness of a model could be sensitive to the choice of evaluation metrics; for instance, DiffAlignOp appears more effective when judged based on IoU rather than F1 Score. In addition, the use of different baselines and datasets further complicates the interpretation of performance gain presented in Fig. 9.

## V. NOISY MODALITIES

Multimodal colearning can also improve model robustness in noisy conditions which can broadly be categorized into label noise and data noise. Noise modeling includes noise estimation and noise reduction in which uncertainty modeling plays an important role. On the other hand, denoising autoencoders, regularization, and noise-aware loss function are typical approaches in noise reduction. In addition, employing a weighted fusion scheme based on modalities’ estimated noise level or an ensemble method, where multiple predictors are combined, can strengthen model robustness toward noise. Representative methods are presented in Table II and its relative performance gain is visualized in Fig. 10.

### A. Data Noise

Collected signals in RS are often susceptible to environment inferences, device failures, and nonstationary objects, resulting in noise. Traditional DL models struggle to differentiate between



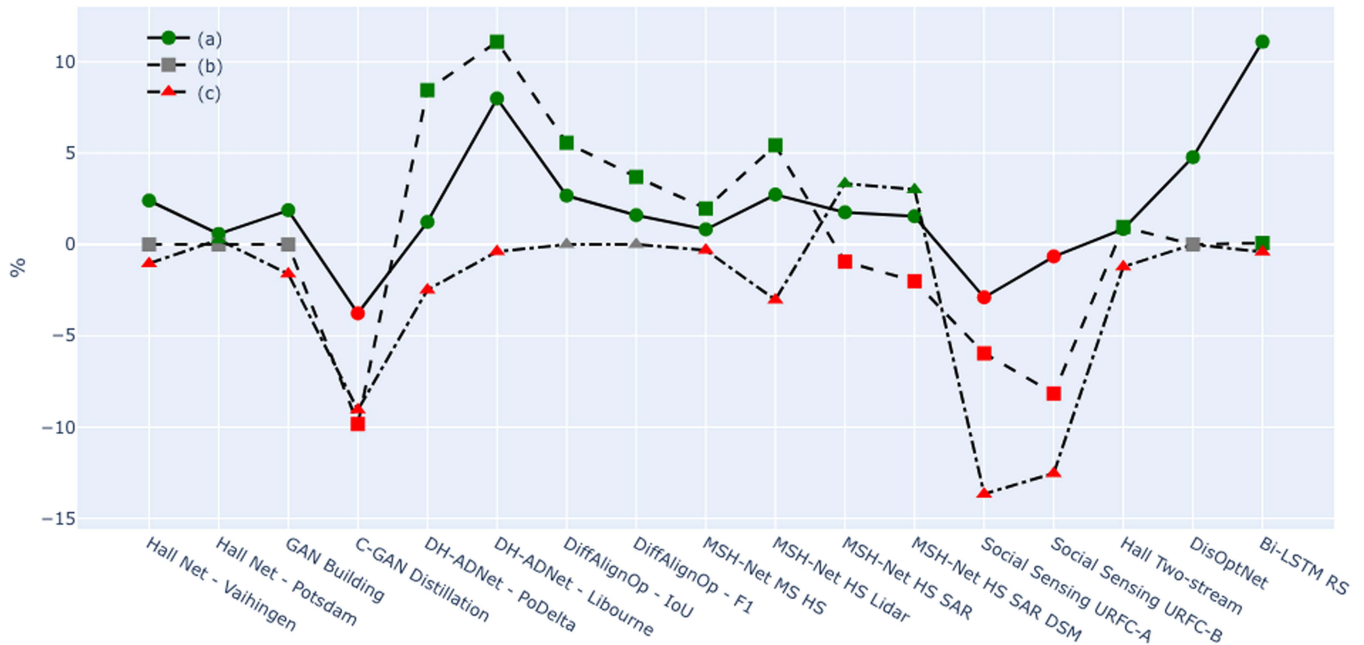


Fig. 9. Relative performance gain analysis. (a) Difference between the proposed method missing nondominant modality and the unimodal baseline. (b) Difference between the proposed method missing dominant modality and the unimodal baseline. (c) Difference between the proposed method missing nondominant modality and the full modality model. Green indicates positive values—the proposed approaches are better than selected baseline models in reviewed papers. Red color represents negative numbers—the proposed approaches cannot surpass selected baselines models. Gray color means no relevant result reported for a given paper. Hall Net [81], GAN Building [77], C-GAN Distillation [82], DH-ADNet [83], DiffAlignOp [12], MSH-Net [84], Social Sensing URFC-A [85], Hall Two-stream [86], DisOptNet [87], Bi-LSTM RS [64]. Table I provides more information about datasets and backbones of each reviewed method.

TABLE II  
REPRESENTATIVE METHODS DEALING WITH NOISY MODALITIES; DOWNSTREAM TASKS INCLUDE CLOUD REMOVAL (CR), SEMANTIC SEGMENTATION (SS), IMAGE MATCHING (IM), CLASSIFICATION (CL), CHANGE DETECTION (CD), AND ROAD EXTRACTION (RE)

Category		No	Pub	Task	Dataset	Modality	Backbone
Data Noise	Adversarial	[95]	JRS 2021	CR	RADARSAT-2, Rapideye	RGB, SAR	GAN
		[96]	IGARSS 2018	CR	SEN1-2	SAR, MS	cGAN
		[97]	JRS 2020	CR	GF2-3, GRSS 2001, SEN1-2	RGB, SAR	GAN
		[98]	GRSL 2021	CR	RICE	RGB	GAN, Residual
		[99]	TGRS 2023	SS	Berlin, Augsburg, YRE	MS, HSI, SAR, DSM	GAN
	Denoising	[100]	PRSSIS 2022	IM	Daedalus, QuickBird, SEN1-2	RGB-NIR, SAR, DSM	Dilated Gaussian Conv
		[8]	JAS 2023	IM	Custom	RGB-NIR, SAR	PCA, CNN
Label Noise	Minimisation	[9]	JIF 2023	CR	SEN12MS-CR	RGB, SAR	ResBlocks, Attention
		[101]	GRSL 2020	CL	RESISC45, TSX, MSTAR	SAR	ResNet-18
		[102]	TGRS 2020	CL	RESISC45, AID	RGB	ResNet-18
		[103]	JSTARS 2021	CL	RESISC45, AID	RGB	ResNet-18
		[104]	JRS 2021	SS	Vaihingen	RG-NIR	SAE
		[105]	JIS 2022	SS	Pavia, Salinas, Houston	HSI	G-Conv
	Correction	[106]	TC 2020	CL	RSI-CB256, AID, PatternNet, BUD	RGB	CNNs
		[107]	TGRS 2021	SS	SII-3, Maricopa	SAR	Kmeans, MRF
		[108]	IJRS 2021	SS	OSM, Kutupalong	RGB	UNet
	Combined	[109]	JPRS 2023	CD	ZY3, GaodeMap	RGB-NIR	ResNet-50
		[110]	GRSL 2021	CL	RESISC45, PatternNet	RGB	ResNet-18
[111]		TGRS 2020	RE	Mass-, Cheng-, Zimbabwe-Roads	RGB	VGG19	

noisy and clean signals, leading to decreased performance. However, completely detecting and removing noisy samples is complicated, costly, and wasteful. Therefore, for modern models, it is essential to learn useful features even in the presence of noisy signals. In the early stage, student–teacher networks showed promising results in handling noisy unimodal data, a concept later extended to multiview learning. Subsequently, multimodal

colearning methods emerged, leveraging abundant and complementary features from multiple modalities while tolerating noise to some extent. There are two types of data noises—adversarial perturbations and common corruptions. Gaussian noise and environmental variations such as cloud, fog, rain, motion, etc., are common sources of noise. SAR signals are particularly susceptible to randomly distributed speckle noise. Adversarial networks

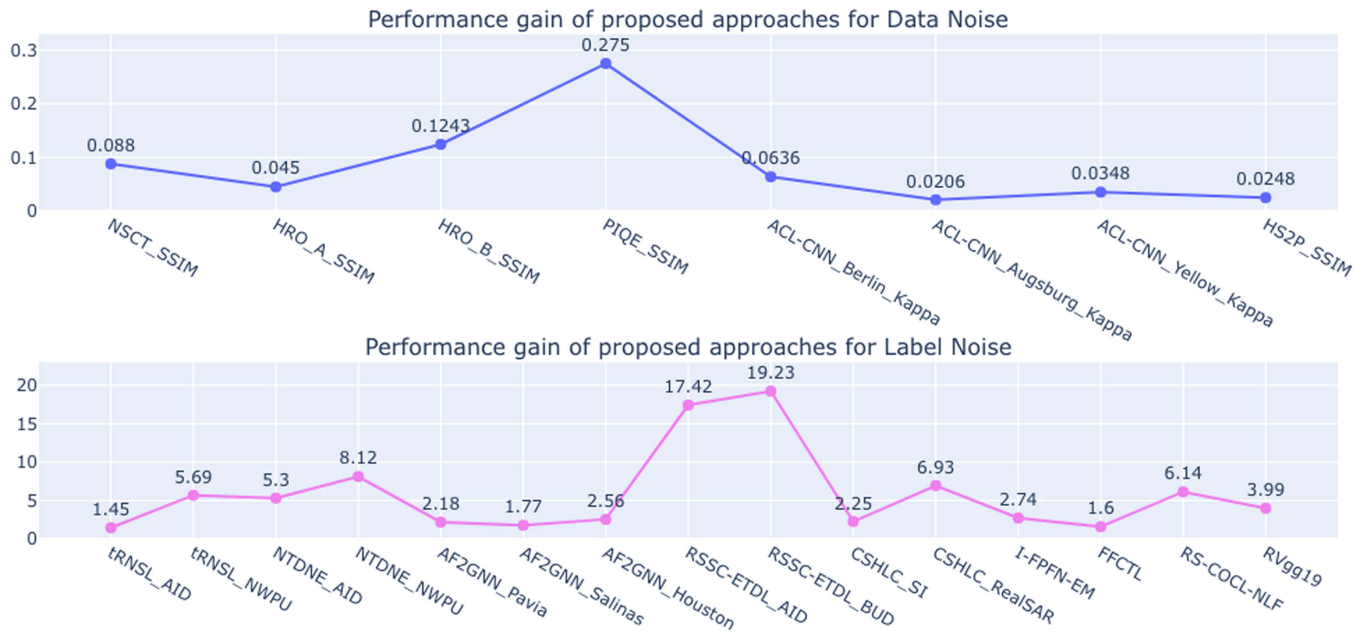


Fig. 10. Relative performance gain analysis. Approaches dealing with data noise: NSCT [95], HRO [97], PIQE [98], ACL-CNN [99], and HS2P [9]. Approaches dealing with label noise: tRNSL [102], NTDNE [103], AF2GNN [105], RSSC-ETDL [106], CSHLC [107], I-PPFN-EM [108], FFCTL [109], RS-COCL-NLF [110], and RVgg19 [111]. Table II provides more information about datasets, modalities, and backbone of each reviewed methods.

are the leading DL-based methods dealing with noisy data [13]. Especially in the data imbalance scenario, [123] proposed a deep multimodal fusion GAN (DMGAN) to solve the task of recognizing faculty homepages (i.e., binary classification from text, image, and HTML layout). Kong et al. [95] developed a GAN-based network with an extensive loss function to remove speckle noise and fuse optical and SAR images simultaneously. GAN-based models have demonstrated their potential in cloud removal for RS applications in several studies [96], [97], [98]. Recently, attention has also been directed toward denoising SAR images [8], [9], [100] overcoming the sensitivity of gradient-based descriptors to noise. While Han et al. [8] took advantage of a PCA-based approach to filter out noise, Zhu et al. [100] proposed a structural descriptor SFOC with dilated Gaussian convolution to resist noisy signals. On the other hand, Li et al. [9] integrated denoising into a fusion network using a hierarchical spatial-spectral structure in which residual blocks with channel attention mechanism (RBCA) is an important component.

The interpretation of Fig. 10 has a few limitations. First, the methods were evaluated on different datasets (details in Table II), making direct comparisons challenging. Second, each method was benchmarked against distinct baselines selected by the authors, introducing variability in assessments. Despite different evaluation metrics being employed, the use of both SSIM structural similarity index measure (SSIM) and Kappa, which are important factors in assessing model’s robustness toward data noise, facilitates a relative comparison. Their values both range from  $-1$  to  $1$  also allows for meaningful performance gain interpretation. Fig. 10 reveals subtle improvements, except for PIQE, underscoring the complexity of the challenge and the need for sophisticated approaches to enhance the resilience of these networks toward data noise. Generalizability emerges as

another crucial concern, as evidenced by the varying performance gains of HRO and ACL-CNN across different datasets. Future research directions should explore strategies to improve the generalizability of these methods, ensuring consistent performance across diverse scenarios.

### B. Label Noise

Label noise refers to incorrect ground truth annotations provided to the network for optimizing model parameters, resulting in a significant decrease in model performance. Large-scale dataset is vulnerable to label noise because of human errors (e.g., lack of expertise, subjective judgment, inconsistency, etc.) leading to misclassifications and omissions of true labels. In the context of multimodal colearning, we assume a single set of ground truth labels is used for each downstream task. Therefore, while most studies discussed in this section are conducted in an unimodal setting, the proposed approaches can be extended to handle multimodal inputs. Methods dealing with label noise can be categorized into two groups—label-noise-minimization and label-noise-correction. The former aims to suppress the impact of noisy labels, while the latter enhances the training process by progressively replacing incorrect labels. It is worth noting that there is a more extensive body of research in label-noise-minimization compared to label-noise-correction.

Methods for label-noise minimization, such as [101], [102], and [103], employ dedicated loss functions to mitigate the impact of noisy labels. Another approach to mitigating the detrimental effects of noisy data is semisupervised learning, as demonstrated in [104]. In addition to well-known CNNs, graph convolution (GConv) models have shown promise. For instance, Ding et al. [105] introduced the AF2GNN model,

TABLE III  
REPRESENTATIVE METHODS USING SEMISUPERVISED REGIME IN RS.

No	Pub	Datasets	Semi-supervised						Modality	Backbone
			Inductive					Trans		
			UP	Wrapper method		Intrinsically semisupervised				
	ST	CT	PT	MF	GT					
[125]	TGRS 2022	UCMerced, AID, RESISC45	✓						HIS	VGG, KMeans++
[10]	JSTARS 2020	Despeckling, Limagne	✓	✓					SAR	UNet
[126]	CFEED 2022	Custom		✓					RGB	MBCConv, DBiFPN
[127]	JRS 2023	Ilmajoki		✓					UAV	Unbiased Teacher v2
[104]	JRS 2021	Vaihingen		✓					NIR	SAE
[128]	JRS 2021	MiniSAR, Toronto			✓				RGB, SAR	Faster RCNN
[129]	GRSL 2021	Pavia, IP			✓				MSI, HSI	CNNs with SE [130]
[131]	TGRS 2020	MSTAR, OpenSARShip				✓	✓		SAR	ResNet101
[132]	JRS 2022	Vaihingen, Potsdam				✓			RGB	VGG16, FCN
[133]	JSP 2023	Vaihingen, Potsdam						✓	RGB	ResNet50, Attention, GAN
[134]	TGRS 2022	UCMerced, AID, RESISC45						✓	RGB	Siamese, GAN
[135]	TNNLS 2022	Indian Pines, Pavia, KSC						✓	HSI	G-Conv, FCN
[136]	TIM 2021	Indian Pines, Pavia, Houston						✓	HSI	G-Conv, Clustering

Methods include unsupervised preprocessing (UP), self-training (ST), cotraining (CT), perturbations (PT), manifolds (MF), and generative (GT). Trans is short for transductive. Downstream tasks include semantic segmentation, object detection, and classification encoded by the color of dataset name.

which reduces noise using adaptive filters and aggregator fusion. Li et al. [106] utilized multiview learning to monitor uncertain labels and employs the adaptive multifeature collaborative representation classifier (AMF-CRC) for iterative correction of weak samples. However, it is worth noting that training multiple CNNs, as done in this approach, can be computationally intensive. Another innovative approach is presented by [107], which progressively refines segmentation output using a novel constrained smoothing and hierarchical label correction (CSHLC) scheme. This scheme incorporates techniques such as pixel group counting comparison (PGCC) and gray similarity comparison (GSC) coupled with Markov random fields. Ahmed et al. [108] proposed a multitask architecture where the network learns to approximate error matrices while extracting building information. Nevertheless, a clean set of true labels calibrated by the author is required for the label correction learning process, which is costly to apply to other large-scale datasets. A promising research direction is the combination of noise minimization and correction, as explored in recent studies [109], [110], [111]. Most recent approaches commonly involve modeling uncertainty using probabilistic methods to address label noise [124].

Fig. 10 shows relative performance gain of discussed approaches. The primary evaluation metric chosen for comparison is overall accuracy (OA), with intersection over union (IoU) used as a substitute when data are unavailable. To facilitate a more intuitive comparison, both metrics are converted to percentages. Similar to the relative performance gains in the presence of data noise, there are variations in the performance gains across datasets. The extent of improvement is also influenced by the selection of baselines, a factor discussed in the previous section. Authors nominate baseline models based on the specific challenges they aim to tackle and prevailing methods at the time of publication. As can be seen from Fig. 10, the reviewed papers contribute meaningfully to the development of models resilient

to label noise, even though a majority of them exhibit less than a 10% improvement.

## VI. LIMITED MODALITY ANNOTATIONS

Acquiring annotations for dense prediction tasks is expensive and time-consuming, especially with very high-resolution RS data. This section discusses techniques to deal with limited or no data label scenarios. Specifically, we will review methods dealing with three levels of data annotations involved in the training process. 1) Semisupervised learning where only a small portion of data annotations are available. 2) Weakly supervised learning where coarse grain level labels are available to support dense prediction downstream tasks. 3) Unsupervised learning where no labels are available at all. Representative research is summarized in Tables III and IV, with associated relative performance gains in Fig. 14.

### A. Semisupervised Learning

In semisupervised learning, the goal is to effectively learn from a limited portion of data that includes ground truth information. According to the taxonomy presented in [150], semisupervised learning methods can be inductive or transductive in which the former is more popular with typical train and inference phrases. Inductive methods are further categorized into three main groups that are 1) unsupervised preprocessing, 2) wrapper method, and 3) intrinsically semisupervised. Fig. 11 provides a simplified illustration of typical methods.

1) *Unsupervised Preprocessing*: The first group, also known as self-supervised learning, involves the use of self-generated signals to learn valuable representations (pretraining) before performing downstream tasks. Within this category, there are three subcategories as outlined in [151]. 1) Generative methods aim to reconstruct the original signal from noisy/missing input. Dalsasso et al. [10] proposed a despeckling framework to



TABLE IV  
REPRESENTATIVE METHODS WEAKLY SUPERVISED AND UNSUPERVISED LEARNING; SEMANTIC SEGMENTATION (SS), DOMAIN ADAPTATION (DA), SEISMIC ANALYSIS (SA), SR, IMAGE REGISTRATION (IR), CHANGE DETECTION (CD), HYPERSPECTRAL UNMIXING (HU)

Category	No	Pub	Task	Dataset	Modality	Backbone	Note
Weakly-supervised	[137]	GRSL 2021	SS	Vaihingen, Zurich	RGB	VGG16	Point, line, polygon
	[138]	FB Research	SS	SA-1B	RGB	MAE ViT	User prompt, zero shot learning
	[11]	JRS 2023	DA	MRSSC	RGB, SWI, INF, SAR	DANN, AFN, MCD	Image level label
	[139]	JRS 2023	SS	Midwestern US	RGB	UNet	Point, image level label
	[140]	JRSE 2023	SS	GF1, ZY3	RGB-NIR	CNNs	Block level label
Unsupervised	[141]	GEO 2018	SA	Liziba	Prestack seismic	DCAE	Clustering: Kmeans, SOM
	[142]	GEO 2019	SA	Songliao Basin	Poststack seismic	CNNs	Kmeans KL divergence loss
	[143]	GRSL 2019	SA	SCSN, KNET, KiKNet	Earthquake waveforms	DCAE	Clustering, KL loss
	[144]	GRL 2020	SR	HR-CinCGAN	RGB	CGAN, EDSR	Generative-based method
	[145]	TGRS 2023	SS	Trento, MUUFL, Houston	HSI, LiDAR	Graph Fusion Network	Laplacian and tSNE loss
	[146]	TNNLS 2020	CD	Texas, Cali, Italy, France	RGB-NIR, VV-VH, MS	DCAE	Cycle-consistency loss Code correlation loss
	[147]	TGRS 2023	HU	SIM2, Muffle, Houston	HSI, LiDAR	Linear, SE	Sparsity regularization, Minimum Volume Constraint loss
	[148]	TGRS 2022	IR	WHU, GF1, ZY3, Landsat-8	RGB, SAR	CNN, SE, Residual	Novel loss function modelling structural similarity
	[149]	TGRS 2023	SS	Houston, LCZ HK, Augsburg	MS, HSI, SAR, LiDAR	Graph subspace learning	Latent subspace alignment, Laplacian

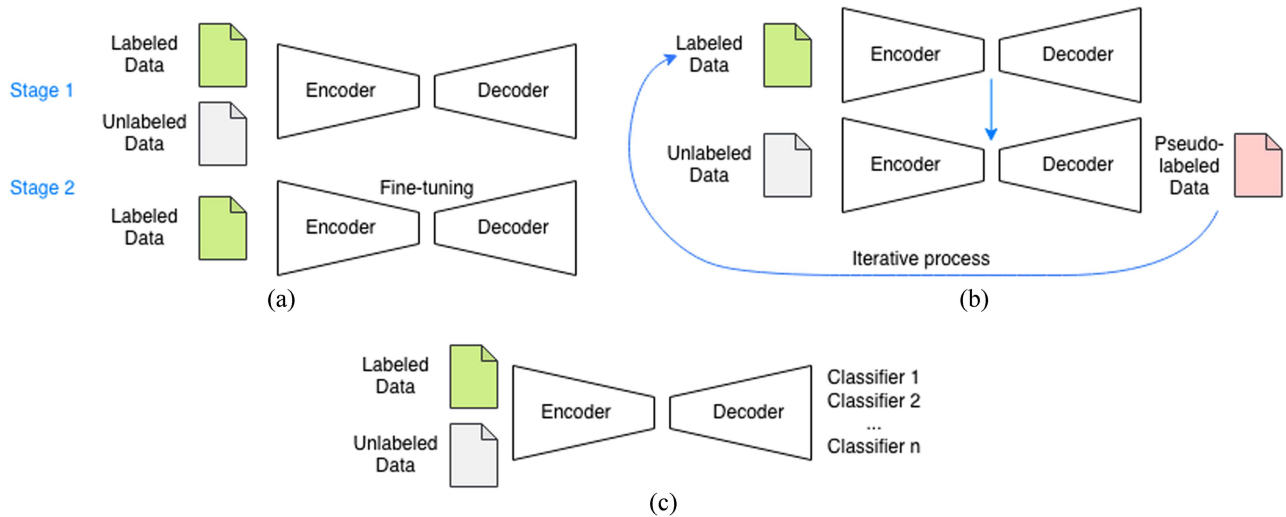


Fig. 11. High-level illustration of typical methods in the semisupervised learning area. Unsupervised preprocessing approaches first use labeled and unlabeled data for some auxiliary tasks that do not require annotations then the backbone is fine-tuned for a main downstream task using labeled data. Self-training employs an iterative process where labeled data is used to train the network then learned parameters produce pseudolabels for unlabeled data. Cotraining is an end-to-end framework with multiple classifiers that can use both labeled and unlabeled data. Multiple objectives are optimized simultaneously forcing the network to learn useful feature representations from the inputs. (a) Unsupervised preprocessing. (b) Self-training. (c) Cotraining.

reconstruct SAR images from noisy acquisitions. 2) Predictive methods focus on predicting derived properties from the original inputs, such as the position of a patch in a larger image or the order of a frame in a video. 3) Contrastive learning seeks to make the feature space as separable as possible using positive and negative examples. Clustering is similar to contrastive learning, where alike samples are pulled together and dissimilar ones are pushed apart [125]. Generative and predictive models excel at capturing local descriptors, while contrastive learning is better suited for capturing global relationships. However, negative sampling in contrastive learning can be computationally expensive. Consequently, there is a recent trend in multitask

learning that combines these three techniques to create more robust networks.

2) *Wrapper Method:* For the wrapper method, self-training and cotraining are the two dominant architectures. Self-training model carries an iterative process in which the model is initially trained with the available annotations, which are typically limited in quantity. It then generates pseudolabels for additional unlabeled data, incorporating them into the training set. This process is repeated iteratively as demonstrated in studies [10], [104], [126], [127]. On the other hand, cotraining is usually conducted in an end-to-end manner, where there could be multiple classifiers benefiting one another [128], [129]. The idea could

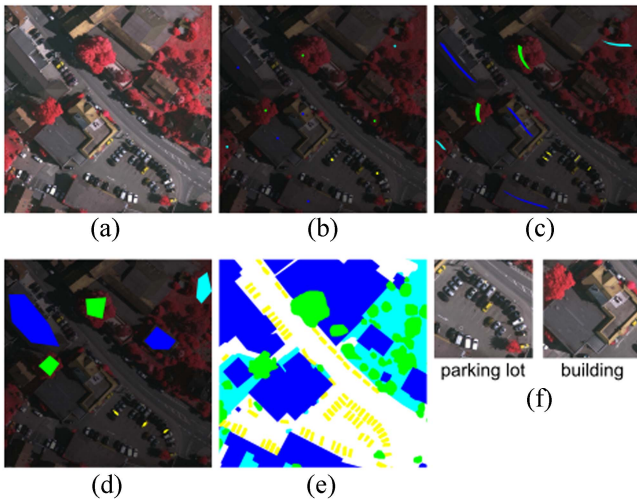


Fig. 12. Examples of different levels of annotations that can be used in weakly supervised learning. (a) Sample patch from the Vaihingen dataset, (b) point-, (c) scribble-, (d) polygon-level annotations, (e) dense pixel true label, and (f) image/block level annotations.

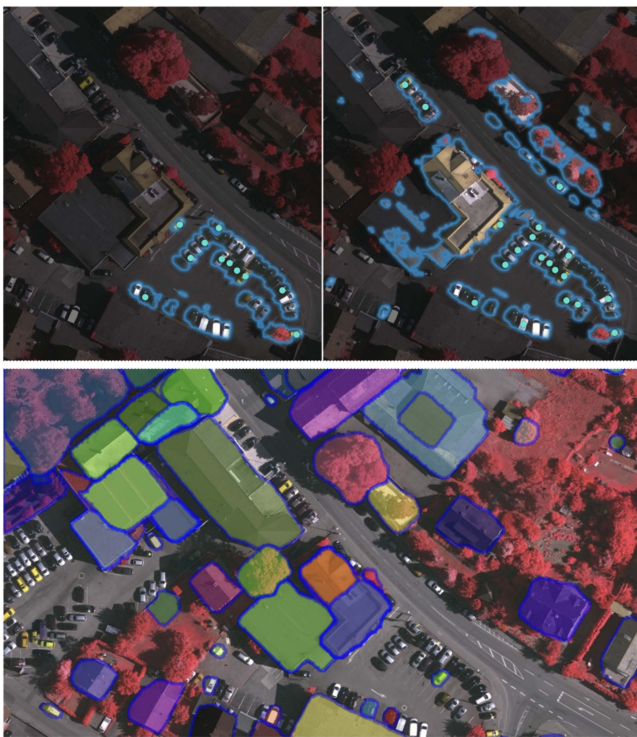


Fig. 13. SAM zero-shot learning on a Vaihingen sample. Object detection is good at local areas but becomes noisy when expanding to a wider region of an image. Scale variations significantly affect model performance resulting in missing objects in large image patches.

be considered as multitask learning; however, unlike the multitask method in self-supervised (unsupervised preprocessing), auxiliary tasks do not have to use the same dataset. For example, Liao et al. [128] proposed FDDA network, where labeled SAR images are used for detection, while at the same time, unlabeled SAR images are leveraged for reconstruction and optical

data is used to perform domain adaption with respect to SAR data.

3) *Intrinsically Semisupervised*: Intrinsically semisupervised methods are those that directly incorporate unlabeled data into their objectives. These methods are often categorized into three popular subcategories, as defined in [150]: perturbation, manifold, and generative. With perturbation-based methods, the learning process assumes that model predictions should be robust to noise. The objective function ensures that encoded feature spaces are minimally affected by a small amount of noise, as seen in studies such as [131] and [132]. Manifold methods, on the other hand, operate under the assumption that some small variations in the input can yield different results. Multiple lower dimensional manifolds are constructed given the input space, where data points belonging to the same group are situated on the same lower-dimensional manifold [131], [152]. In contrast to discriminative methods, generative approaches learn useful features by modeling data-generating distribution, in which GAN [74] and variational autoencoders (VAE) [153] are the fundamentals. Recent studies [133] and [134] have demonstrated the effectiveness of GAN in RS semantic segmentation and scene classification, respectively, in a semisupervised learning manner.

Graph-based models: Graph-based models represent a class of transductive methods that connect both labeled and unlabeled data points through similarity measurements. These algorithms aim to maximize the matching between predictions and ground truth labels for labeled data points. As a result, unlabeled data points will inherit the same label with similar data points determined by the graph. Xi et al. [135] introduced a cross-scale graph prototypical network (X-GPN), which utilizes multiple branches of graph convolutional network (GCN) and uses a self-branch attention addition (SBAA) component. This design enables the learning of multiscale features in HSI images. Spectral-spatial graphs can embed both spectral signatures and spatial proximity of individual pixels or image patches. However, its complexity grows exponentially when learning from RS images which are very high in resolution and have a large number of spectral bands. Moreover, model performance can strongly depend on local features. To overcome this challenge, Ding et al. [136] proposed a global consistent GCN (GCGCN) method to better capture long-range dependencies of HSI data. While these approaches have shown promise in unimodal data settings, adapting them to a multimodal learning landscape remains an open challenge.

## B. Weakly Supervised Learning

Weakly supervised learning seeks to leverage sparse annotations to achieve dense predictions. For instance, in RS semantic segmentation, instead of having detailed pixel-level labels, more economical annotations like point markers, scribbles, polygons, or image-level labels can be used as shown in Fig. 12. An example of this is the feature and spatial relational regularization (FESTA) approach introduced in [137], which encodes and regularizes relations between pixels in both spectral and spatial domains. Another avenue involves human feedback, where the model refines its predictions based on user guidance or prompts.

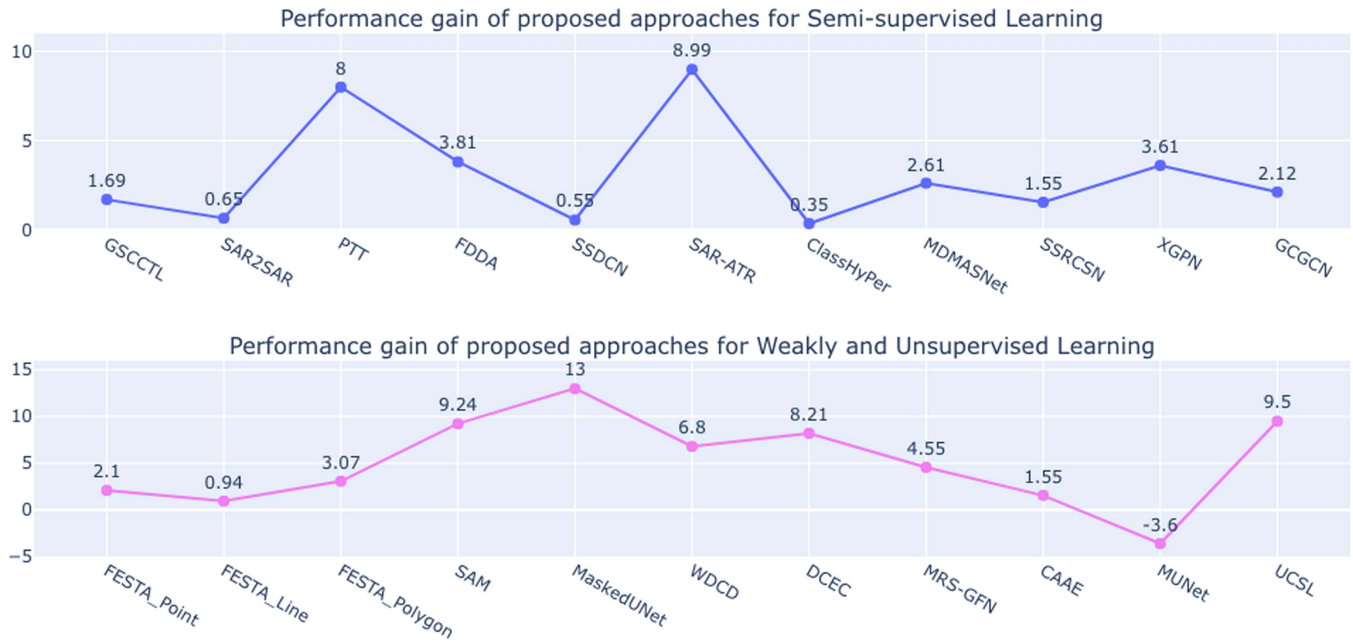


Fig. 14. Relative performance gain analysis. Semisupervised learning methods: GSCCTL [125], SAR2SAR [10], PTT [126], FDDA [128], SSDCN [129], SAR-ATR [131], ClassHyPer [132], MDMASNet [133], SSRCSN [134], XGPN [135], and GCGCN [136]. Weakly and unsupervised learning methods: FESTA [137], SAM [138], MaskedUNet [139], WDCD [139], DCEC [142], MRS-GFN [145], CAAE [146], MUNet [148], and UCSL [149]. Results (OA or F1) are aggregated wherever data is available (across different datasets or parameters). Tables III and IV provide information about datasets, modalities, and backbone of reviewed methods.

This approach also opens the door to active learning and zero-shot learning scenarios. The segment anything model (SAM) described in [138], originally trained on a large image dataset, can adapt to RS data with a few coarse-grained user-provided labels. However, through our experiment, large RS image scale variations and long-range dependencies are still challenging despite SAM advances as seen in Fig. 13. Furthermore, these approaches are primarily designed for unimodal learning and have yet to harness the potential of multimodal data. Multiple instance learning (MIL) is another type of weakly supervised learning where data points are grouped into bags, each sharing the same label [154]. Promising results have been demonstrated by applying MIL in a multimodal DL pipeline for tasks like cancer prediction, as shown in [155]. In the context of RS, research on weakly supervised learning, particularly in the realm of multimodal data, remains scarce. A notable exception is the work by [11], which primarily focused on domain adaptation algorithms to transfer knowledge from the RGB modality to others (SWI, INF, SAR). However, the reported quantitative results assessing the effectiveness of weakly supervised learning are still limited. Gradient weighted class activation map (Grad-CAM) [156] and attention map play an important role in constructing dense prediction. As pointed out in [157], this is an appealing research direction given that there is room for improvement to achieve satisfactory performance. Wang et al. [139] proposed applying UNet to transfer point- and image-level to dense prediction output, even with as few as 100 labels. Meanwhile, Li et al. [140] introduced a global convolutional pooling layer combined with a VGG backbone to generate cloud masks based on block-level annotations. However, both of these works only produce binary segmentation maps and operate in

an unimodal setting. Future research in RS should investigate weakly supervised learning in more complex tasks, such as multiclass semantic segmentation with multimodal data.

### C. Unsupervised Learning

In the realm of unsupervised learning, the objective is to acquire compact feature representations and perform tasks like clustering or pattern recognition without the need for explicit labels [158]. Various backbones can be employed as a feature extractor [e.g., cascaded autoencoders (CAE), deep cascaded autoencoders (DCAE)] followed by a clustering method such as K-Means, as observed in [141], [142], and [143]. CycleGAN-based framework has been utilized for unsupervised learning because it does not require training set pairing, for example, Niu et al. [144] proposed CinCGAN to perform precise rock image superresolution (SR). However, CycleGAN is not the most ideal fit for geophysical tasks due to its extremely time-consuming training process and instability. The combination of a CAE and lightweight clustering algorithms like KMeans is more suitable. When it comes to time series tasks, RNN, long short-term memory networks (LSTM), and transformers provide a high-performance framework. However, the majority of existing works supporting unsupervised learning within the RS domain are conducted in an unimodal setting. A recent paper by Du et al. [145] is one of the rare research in RS evaluating the effectiveness of unsupervised learning in the context of multimodal. While feature extraction and fusion are carried out in an unsupervised manner, pixel-level classification still relies on classifiers like support vector machines (SVM). Nevertheless, it alleviates the need for acquiring large training sets with



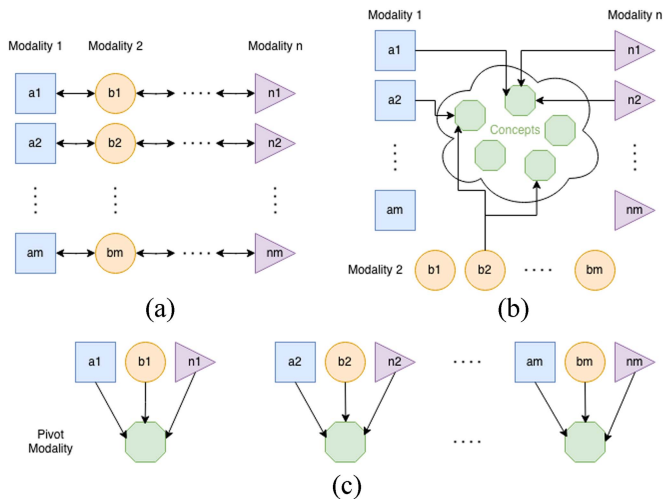


Fig. 15. Three types of modality parallelism. (a) Parallel/Strongly paired. (b) Non-parallel/Weakly paired. (c) Hybrid/Bridge.

dense annotations. In another recent study, Yao et al. [149] proposed a novel framework unsupervised common subspace learning (UCSL), which was able to construct useful multimodal representation without the need for a label. Unsupervised learning has been gaining attention in RS benefiting different downstream tasks such as semantic segmentation [145], change detection [146], hyperspectral unmixing [147], image registration [148]. However, change detection tasks dominate the unsupervised learning landscape, and dense prediction task is underexplored.

## VII. MODALITY PARALLELISM

In an ideal multimodal colearning scenario, for every observed instance, there is associated information in all modalities. However, real-world applications often do not conform to this ideal scenario. For instance, considering an action recognition application, we can lack pixel-level annotations for every frame describing the object's action. Instead, we have access to a short text description of the scene and recorded audio. These textual, visual, and auditory cues are not directly linked at the basic unit level (e.g., word, pixel, sound), but they can converge on a higher level concept, such as describing a specific action. Rahate et al. [13] suggested that there are three types of modality parallelism—parallel/strongly paired, nonparallel/weakly paired, and hybrid/bridge, as illustrated in Fig. 15. This classification is particularly relevant in domains involving text-audio-video modalities, such as vision-language tasks, where elements from frames, textual descriptions, and audio do not always have a one-to-one mapping. In contrast, with RS, multimodal inputs are typically assumed to be coregistered through geospatial information, a critical factor in most downstream tasks.

### A. Parallel/Strongly Paired Data

This is the most desirable type of data in multimodal colearning, in which every observation has a direct link from one

modality to another. Rahate et al. [13] provided many examples of visual language (VL) tasks and translation-based colearning methods that benefit from parallel multimodal data. In RS, most studies about multimodal colearning, especially those that are presented in this article, assume multimodal inputs (e.g., MS, HSI, LiDAR, and VHR) are geographically coregistered. For dense prediction tasks like semantic segmentation and change detection, it is intuitive to demand a strongly paired dataset because the learning process is optimized at a pixel level. For a coarser task like object detection, it remains challenging if multimodal information is not aligned because the network still needs to localize objects of interest based on the height and spectral signature of a given location. In general, multiple-instance learning in RS requires strongly paired data. However, creating strongly paired multimodal inputs is challenging. There is a dedicated research area for RS image registration [160], [161], [162]. Since modalities are strongly paired, we can apply transfer or translation-based learning so that given an arbitrary combination of modalities at inference, the network can derive the missing ones as discussed in Section IV.

### B. Nonparallel/Weakly Paired Data

Nonparallel or weakly paired data refers to multimodal inputs that are not aligned at a fine-grained level. In other words, the elements of multimodal inputs do not have a one-to-one correspondence. Instead, they can be connected over some coarser conceptual level, which is closely related to weakly supervised learning presented in Section VI-B. Learning multimodal representations from nonparallel data can facilitate the establishment of stronger semantic connections between modalities. To deal with nonparallel data, we need to perform multimodal alignment or conceptual grounding in addition to solving downstream tasks. Rahate et al. [13] gave significant credit to attention mechanism [116] for implicitly aligning modalities in visual question answering (VQA) tasks through phrase grounding. The task of scene classification in RS bears similarities to VQA, with more relaxed data requirements compared to semantic segmentation and change detection, as discussed earlier. However, unlike various VQA tasks, RS does not have many large-scale datasets with meaningful high-level textual descriptions. The first dataset published recently benchmarking visual grounding in RS was [163]. One of a few studies where multimodal inputs were aligned at the area level rather than individual pixels was [85]. It combined sensed images and social activities signature. Zheng et al. [12] proposed a DiffAlignOp component that encourages learning from registration-free data. However, each modality still has its own set of annotations, which is still costly to obtain. There are a few studies that learn to perform the SR task using weakly paired data. Farooq et al. [164] used cycle GANs to produce superresolved images from low-resolution ones. The authors in [165] and [166] attempted to leverage unregistered temporal images, albeit with different objectives; the former tackled the SR task, while the latter addressed relative radiometric normalization. Several recent studies have pushed the boundaries further by learning from unregistered images in an unsupervised manner. the authors in [167], [168], and [169]

demonstrated the potential of this approach. On the other hand, instead of attempting to align actual multimodal inputs, learning to generate one modality from another and then create coregistered images for downstream tasks is another method [170]; however, the error carried forward from the translation process is concerning.

### C. Hybrid/Bridge Data

This final category involves a pivot modality that acts as an intermediary bridging unaligned modalities. There are two subcategories that are 1–1 and 1-many. This category is typically divided into two subcategories: 1–1 and 1-many, with common applications found in the VL domain. One illustrative example is multilingual video searching, where visual cues serve as the pivot modality connecting different languages. Microsoft Research Multimodal Aligned Recipe Corpus dataset [171] was created using this approach. Different recipes of the same dish are aligned based on the relationship between video transcripts and textual instructions. While this hybrid data alignment approach has found applications in fields like VL processing, there is a notable absence of relevant datasets or studies conducted in a hybrid data manner within the realm of RS.

## VIII. EMERGING RESEARCH DIRECTIONS

From the analysis and insights during the review, we identify the following promising future directions for multimodal colearning in the RS context: developing explainable AI (XAI) techniques to make multimodal colearning interpretable and transparent, developing onboard processing techniques with limited resources to allow multimodal colearning to be effectively deployed real-time in satellites or drone platforms, leveraging the power of recent generative AI to generate large-scale synthetic data for RS tasks, and leveraging large pretrained models and foundation models to improve generalization and zero-shot learning in the RS context.

### A. XAI in Multimodal Colearning

Within the realm of XAI model transparency methods aim to make the decision-making process of the DL models transparent by explaining the structure of the model. This could involve explaining the design choices such as the neural network architecture, activation functions, number of epochs, batch sizes, and other hyperparameter values, and/or visualizing model internal functionality via obtaining decision boundaries, and intermediate representations.

Model interpretability is the paradigm that involves developing explainable models or the process of incorporating explainability into a complex model. Compared to classical machine learning methods such as decision trees or linear models, DL models are inherently complex. The use of multimodality data further aggravates this complexity, therefore, special mechanisms should be leveraged to explain the sophisticated models. Among these techniques, feature attribution [172], [173] and visualization of model decisions [174], [175] are common within RS literature. Feature attribution techniques could help

identify the most influential features within the feature space for a particular decision. Perturbation-based methods including local interpretable model-agnostic explanations (LIME) [176], Shapley additive explanation (SHAP) [177], and partial dependence plots (PDPs) [178] fall under this category. Furthermore, back-propagation or gradient-based methods also generate feature attribution maps while analyzing the information propagated through the forward and backward passes of a particular network. Grad-CAM [179] and its variants and Deep-LIFT [180] are widely used methods under this category. Researchers within the RS community have also utilized various visualization techniques such as heatmaps, saliency maps, and overlays of model predictions to understand the salient regions within the remotely sensed image [181].

Despite the emerging interest, to the best of our knowledge, none of the existing multimodal colearning approaches have investigated the use of XAI to explain model decisions. As elaborated earlier, the use of multiple information sources leads to complex multibranch network architectures and sophisticated feature fusion strategies. Therefore, off-the-shelf XAI frameworks such as feature attribution methods may not accomplish the intended explainability in such intricate settings. Further investigations into the applicability of existing XAI methods and the design of novel explainability paradigms to better handle multimodal colearning frameworks are required.

### B. Onboard Processing in Multimodal Colearning

In RS applications, one of the major bottlenecks for real-time decisions is the time-consuming data down-linking process over limited communication bandwidth. As such, processing the captured data onboard of the satellite is preferred instead of transmitting the raw data. The advent of low-power embedded graphic processing units (GPUs) such as the NVIDIA Jetson series, and Intel Movidius Vision Processing Unit (VPU) series has propelled this domain even further. When considering the deployment of DL models in a resource-constrained onboard evaluation setting literature has readily used model compression techniques. These techniques encompass the methods used to reduce the size and computational complexity of a DL model while maintaining its performance. Among these techniques model pruning, KD, automated machine learning (AutoML), and neural architecture search (NAS) are common.

Model pruning techniques start with a large model and iteratively remove unimportant weights, neurons, or entire layers from this network to reduce its size and computational requirements. Numerous works within the RS literature have leveraged model pruning for designing lightweight models for different applications, and these applications range from land use classification [182], [183] to target detection [183].

In KD, the complex and computationally exhaustive larger network (called teacher) is replaced using a lightweight student network. Specifically, the student network learns to mimic the behavior of the teacher and this training process helps the transfer of knowledge from a complex model to a simpler one. Feature distillation [184], cross-modality distillation [7], and

response-based KD [185] are some methods proposed for KD in RS applications.

AutoML encompasses various aspects within the machine learning pipeline, including, data preprocessing, feature selection, model selection, and hyperparameter tuning, and the aim of AutoML is the automation of these processes. As such, the concept of AutoML simplifies the machine learning workflow and makes it accessible to nonmachine learning experts. The AutoML workflow can also be used for model compression which would enable onboard processing. For instance, Koh et al. [186] have leveraged AutoML to define a complete high-performance end-to-end machine learning pipeline for RS image-based plant phenotyping.

NAS is a subset of AutoML specifically focused on automating the design and selection of compact network architecture that meets predefined performance criteria. Numerous NAS-based architecture for efficient computing has been proposed for RS applications which extend from satellite image classification [187], [188], [189], [190] to object detection [191].

When reviewing the literature, it is clear that the application of model compression techniques for multimodal RS applications is scarce. Therefore, research that investigates the applicability of advanced model compression techniques such as AutoML or NAS is warranted.

### C. Generative AI for Multimodal Colearning

A predominant application of generative AI in RS is to increase the spatial resolution of the captured images and remove noise. Typical RS images have low spatial resolution due to orbit altitude, sensor characteristics, and data transmission and storage constraints. Therefore, the researchers have readily leveraged generative AI to synthesize high-resolution imagery from low-resolution sensor data [192], [193]. Furthermore, RS data can become noisy due to atmospheric interference and illumination conditions. The data-denoising ability of generative AI could significantly improve the quality of the extracted information and has been leveraged in some RS applications [194].

The generative AI models have the ability to hallucinate missing information in the input data as such they have been used for pansharpening in RS. Specifically, in pansharpening high spatial resolution panchromatic images are fused with low spatial resolution multispectral imagery to create a single high-resolution image. The generative models could hallucinate the missing spatial and spectral information in the input information sources when generating the high-resolution output. The applications of generative AI for pansharpening include [195], [196], [197]. Similarly, recent introductions within generative AI such as diffusion models [198], [199] could augment or bootstrap the data labeling process via performing conditional generations, for instance, text-to-image generation. Furthermore, generative AI models could generate complementary synthetic data and artificially induce variations within the dataset such as seasonal variations, day/night variations, and sensor type variations [198], [200].

Traditional denoising methods often focus on addressing specific types of noise through iterative removal processes, leading

to inefficiencies when confronted with mixed noise scenarios. On the other hand, GAN-based approaches exhibit superior performance in generic image denoising. However, they may struggle with accurately recovering specific small objects and complex scenes, necessitating additional detailed supplementation tasks. The diversity inherent in diffusion models is a strength, but optimization of the inverse diffusion direction is crucial for enhanced outcomes. The current randomness in the diffusion model generation process poses challenges for SR reconstruction. While diffusion models help mitigate issues like oversmoothing (common in multiple convolution-based methods) and mode collapse (associated with the unstable training of GAN-based methods), there is still room for improvement in these areas. In addition, future research should focus on accelerating these processes to enable real-time applications.

Despite the recent interest, the utilization of generative AI in RS-based multimodal colearning is still under-explored. For instance, the literature on multimodal RS imagery denoising and synthetic multimodal RS image generation is scarce. Therefore, further investigation into the utility of generative AI for multimodal colearning is recommended.

### D. Large Pretrained Models and Foundation Models

Foundation models capture a broader concept within the realm of pretrained models and the purpose is to provide a large-scale starting point or building block for designing more specialized models. In contrast to LLMs which are purposely built for a specific downstream task, the foundation models are more general purpose and are typically pretrained on a broad set of unlabeled data. The primary goal of foundation models is to facilitate the transferring of the learned knowledge during the pretraining into specialized downstream tasks.

The success of foundation models has seeped into the RS domain. One of the notable developments is the introduction of “Prithvi”—a temporal vision transformer pretrained by a team of scientists from IBM and NASA [201]. This model has been pretrained using a large unlabeled data corpus obtained from Harmonized Landsat Sentinel 2 (HLS) data and a MAE based self-supervised learning strategy has been employed for pretraining. Similarly, numerous works have applied masked image modeling to pretrain large vision transformer models which have been derived based on the inspiration from large language models (LLMs). The model introduced in [202] named RingMo has been pretrained on a large RS dataset captured using both satellite and aerial platforms. Along these lines, there have been works that extend the limits of pretrained vision transformer parameter space from 100 million [203] to billion [204]. As such in the future, the domain of LLM-inspired pretrained models and foundation models will only become larger and more robust.

We expect the success of large pretrained models and foundation models in the RS domain will help the formulation of multimodal foundation models which can be pretrained using a variety of modalities and could extract contextual information across the modalities. The introduction of Microsoft’s UniLM [205], Google’s VideoBERT [206], and All-In-One model [207] demonstrates this capability across video, text, and



image platforms. As such we expect to see the introduction of multimodal foundation models for RS in the near future.

## IX. CONCLUSION

Thanks to the advancements in RS devices and platforms, we now have unprecedented access to a vast amount of multimodal data. Modalities such as RGB, NIR-IR, multispectral, hyperspectral, and LiDAR have revolutionized machine learning in RS by providing abundant and complementary features for a wide range of downstream tasks surpassing the capabilities of unimodal methods. Multimodal colearning leveraging both modality-specific and modality-correlated information, offers a promising solution to enhance model robustness in nonideal scenarios. In this article, we focus on four main challenges under the multimodal colearning regime: 1) missing one or more modalities at train and/or test, 2) noisy data and label, 3) limited annotated dataset, and 4) nonaligned modalities. All common downstream tasks in RS such as semantic segmentation, change detection, and object detection are vulnerable to these challenges, which hinders superior performance in critical applications (e.g., disaster monitoring and management and land usage analysis). Therefore, this article provides a comprehensive taxonomy and a systematic review of state-of-the-art approaches, drawing insights from more than 200 papers in the field. Future research directions are suggested based on current development and shortcomings encouraging experts and practitioners to improve the accuracy and reliability of multimodal DL in RS.

## REFERENCES

- [1] J. Li et al., "Deep learning in multimodal remote sensing data fusion: A comprehensive review," *Int. J. Appl. Earth Observ. Geoinf.*, vol. 112, 2022, Art. no. 102926.
- [2] D. Hong et al., "More diverse means better: Multimodal deep learning meets remote-sensing imagery classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 5, pp. 4340–4354, May 2021.
- [3] Q. He, X. Sun, W. Diao, Z. Yan, F. Yao, and K. Fu, "Multimodal remote sensing image segmentation with intuition-inspired hypergraph modeling," *IEEE Trans. Image Process.*, vol. 32, pp. 1474–1487, 2023.
- [4] S. Chirakkal, F. Bovolo, A. R. Misra, L. Bruzzone, and A. Bhattacharya, "A general framework for change detection using multimodal remote sensing data," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 10665–10680, 2021.
- [5] C. Persello et al., "2023 IEEE GRSS data fusion contest: Large-scale fine-grained building classification for semantic urban reconstruction," *IEEE Geosci. Remote Sens. Mag.*, vol. 11, no. 1, pp. 94–97, Mar. 2023.
- [6] H. Wang, Y. Chen, C. Ma, J. Avery, L. Hull, and G. Carneiro, "Multimodal learning with missing modality via shared-specific feature modelling," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 15878–15887.
- [7] T. Fernando, C. Fookes, H. Gammulle, S. Denman, and S. Sridharan, "Towards on-board panoptic segmentation of multispectral satellite images," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5402312.
- [8] S. Han, X. Liu, J. Dong, and H. Liu, "Remote sensing multimodal image matching based on structure feature and learnable matching network," *Appl. Sci.*, vol. 13, no. 13, Jun. 2023, Art. no. 7701.
- [9] Y. Li, F. Wei, Y. Zhang, W. Chen, and J. Ma, "HS2P: Hierarchical spectral and structure-preserving fusion network for multimodal remote sensing image cloud and shadow removal," *Inf. Fusion*, vol. 94, pp. 215–228, Jun. 2023.
- [10] E. Dalsasso, L. Denis, and F. Tupin, "SAR2SAR: A semi-supervised despeckling algorithm for SAR images," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 4321–4329, 2021.
- [11] K. Liu, J. Yang, and S. Li, "Remote-sensing cross-domain scene classification: A dataset and benchmark," *Remote Sens.*, vol. 14, no. 18, 2023, Art. no. 4635.
- [12] Z. Zheng, A. Ma, L. Zhang, and Y. Zhong, "Deep multisensor learning for missing-modality all-weather mapping," *ISPRS J. Photogramm. Remote Sens.*, vol. 174, pp. 254–264, Apr. 2021.
- [13] A. Rahate, R. Walambe, S. Ramanna, and K. Kotecha, "Multimodal co-learning: Challenges, applications with datasets, recent advances and future directions," *Inf. Fusion*, vol. 81, pp. 203–239, 2022.
- [14] V. S. F. Garnot and L. Landrieu, "Panoptic segmentation of satellite image time series with convolutional temporal attention networks," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 4852–4861.
- [15] M. T. Chiu et al., "Agriculture-vision: A large aerial image database for agricultural pattern analysis," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 2825–2835.
- [16] H. Alemohammad and K. Booth, "Landcovernet: A global benchmark land cover classification training dataset," in *Proc. AI Earth Sci. Workshop, NeurIPS*, 2020. [Online]. Available: <https://arxiv.org/abs/2012.03111>
- [17] R. Remelgado et al., "A crop type dataset for consistent land cover classification in central asia," *Sci. Data*, vol. 7, no. 1, Jul. 2020, Art. no. 250.
- [18] X. Meng et al., "A large-scale benchmark data set for evaluating pansharpening performance: Overview and implementation," *IEEE Geosci. Remote Sens. Mag.*, vol. 9, no. 1, pp. 18–52, Mar. 2021.
- [19] G. Vivone, M. D. Mura, A. Garzelli, and F. Pacifici, "A benchmarking protocol for pansharpening: Dataset, preprocessing, and quality assessment," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 6102–6118, 2021.
- [20] Landsat Science | A joint NASA/USGS Earth observation program, "Landsat 8 | landsat science," Jan. 2023. Accessed: Oct. 24, 2023. [Online]. Available: <https://landsat.gsfc.nasa.gov/satellites/landsat-8>
- [21] NASA, "MODIS (Moderate resolution imaging spectroradiometer) data," 2024. Accessed: Jan. 18, 2024. [Online]. Available: <https://modis.gsfc.nasa.gov/>
- [22] NASA, "ASTER (Advanced spaceborne thermal emission and reflection radiometer)," 2024. Accessed: Jan. 18, 2024. [Online]. Available: <https://asterweb.jpl.nasa.gov/>
- [23] European Space Agency, "Sentinel-3 SLSTR (Sea and land surface temperature radiometer)," 2024. Accessed: Jan. 18, 2024. [Online]. Available: [https://www.esa.int/Applications/Observing\\_the\\_Earth/Copernicus/Sentinel-3](https://www.esa.int/Applications/Observing_the_Earth/Copernicus/Sentinel-3)
- [24] USGS-U. S. Geological Survey, "EarthExplorer," Oct. 2023. Accessed: Oct. 24, 2023. [Online]. Available: <https://earthexplorer.usgs.gov/>
- [25] B. L. Saux, N. Yokoya, R. Hänsch, and M. Brown, "Data fusion contest 2019 (DFC 2019)," *IEEE Dataport*, 2019, 10.21227/c6tm-vw12.
- [26] S. Prasad, B. L. Saux, N. Yokoya, and R. Hansch, "2018 IEEE GRSS data fusion challenge—fusion of multispectral LiDAR and hyperspectral data," *IEEE Dataport*, 2018, 10.21227/jnh9-nz89.
- [27] "OpenTopography - Find topography data," Oct. 2023. Accessed: Oct. 24, 2023. [Online]. Available: <https://opentopography.org/data>
- [28] "NOAA: Data access viewer," Oct. 2023. Accessed: Oct. 24, 2023. [Online]. Available: <https://coast.noaa.gov/dataviewer/>
- [29] S. Online, "Sentinel-1 - Data products - Sentinel online - Sentinel online," Oct. 2023. Accessed: Oct. 24, 2023. [Online]. Available: <https://sentinels.copernicus.eu/web/sentinel/missions/sentinel-1/data-products>
- [30] European Space Agency, "ALOS-1 - Earth online," Oct. 2023. Accessed: Oct. 24, 2023. [Online]. Available: <https://earth.esa.int/eogateway/missions/alos>
- [31] "EnviSat - Earth online," Oct. 2023. Accessed: Oct. 24, 2023. [Online]. Available: <https://earth.esa.int/eogateway/missions/envisat>
- [32] D. Hong, J. Chanussot, and X. X. Zhu, "An overview of multimodal remote sensing data fusion: From image to feature, from shallow to deep," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2021, pp. 1245–1248.
- [33] M. D. Mura, S. Prasad, F. Pacifici, P. Gamba, J. Chanussot, and J. A. Benediktsson, "Challenges and opportunities of multimodality and data fusion in remote sensing," *Proc. IEEE*, vol. 103, no. 9, pp. 1585–1601, Sep. 2015.
- [34] V. S. F. Garnot, L. Landrieu, and N. Chehata, "Multi-modal temporal attention models for crop mapping from satellite time series," *ISPRS J. Photogrammetry Remote Sens.*, vol. 187, pp. 294–305, 2022.
- [35] Y. Yuan, L. Lin, Z.-G. Zhou, H. Jiang, and Q. Liu, "Bridging optical and SAR satellite image time series via contrastive feature extraction for crop classification," *ISPRS J. Photogrammetry Remote Sens.*, vol. 195, pp. 222–232, 2023.
- [36] F. Weilandt et al., "Early crop classification via multi-modal satellite data fusion and temporal attention," *Remote Sens.*, vol. 15, no. 3, 2023, Art. no. 799.

- [37] P. Ghamisi, B. Höfle, and X. X. Zhu, "Hyperspectral and LiDAR data fusion using extinction profiles and deep convolutional neural network," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 10, no. 6, pp. 3011–3024, Jun. 2017.
- [38] N. Audebert, B. L. Saux, and S. Lefèvre, "Semantic segmentation of earth observation data using multimodal and multi-scale deep networks," in *Proc. 13th Asian Conf. Comput. Vis.*, 2016, pp. 180–196.
- [39] Y. Chen, C. Li, P. Ghamisi, X. Jia, and Y. Gu, "Deep fusion of remote sensing data for accurate classification," *IEEE Geosci. Remote Sens. Lett.*, vol. 14, no. 8, pp. 1253–1257, Aug. 2017.
- [40] Y. Xu et al., "Advanced multi-sensor optical remote sensing for urban land use and land cover classification: Outcome of the 2018 IEEE GRSS data fusion contest," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 12, no. 6, pp. 1709–1724, Jun. 2019.
- [41] S. Srivastava, J. E. Vargas-Munoz, and D. Tuia, "Understanding urban landuse from the above and ground perspectives: A deep learning, multimodal solution," *Remote Sens. Environ.*, vol. 228, pp. 129–143, 2019.
- [42] E. Capliez et al., "Multi-sensor temporal unsupervised domain adaptation for land cover mapping with spatial pseudo labelling and adversarial learning," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5405716.
- [43] L. Bergamasco, F. Bovolo, and L. Bruzzone, "A dual-branch deep learning architecture for multisensor and multitemporal remote sensing semantic segmentation," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 16, pp. 2147–2162, 2023.
- [44] D. Hong et al., "Cross-city matters: A multimodal remote sensing benchmark dataset for cross-city semantic segmentation using high-resolution domain adaptation networks," *Remote Sens. Environ.*, vol. 299, Dec. 2023, Art. no. 113856.
- [45] J. Yao, B. Zhang, C. Li, D. Hong, and J. Chanussot, "Extended vision transformer (ExViT) for land use and land cover classification: A multimodal deep learning framework," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5514415.
- [46] J. Radoux, A. Bourdouxhe, W. Coos, M. Dufrière, and P. Defourny, "Improving ecotope segmentation by combining topographic and spectral data," *Remote Sens.*, vol. 11, no. 3, 2019, Art. no. 354.
- [47] M. Sharma et al., "YOLOrs: Object detection in multimodal remote sensing imagery," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 1497–1508, 2021.
- [48] W. Sakla, G. Konjevod, and T. N. Mundhenk, "Deep multi-modal vehicle detection in aerial ISR imagery," in *Proc. IEEE Winter Conf. Appl. Comput. Vis.*, 2017, pp. 916–923.
- [49] L. Yu, Q. Yang, and L. Dong, "Aircraft target detection using multimodal satellite-based data," *Signal Process.*, vol. 155, pp. 358–367, 2019.
- [50] S. Razakarivony and F. Jurie, "Vehicle detection in aerial imagery: A small target detection benchmark," *J. Vis. Commun. Image Representation*, vol. 34, pp. 187–203, 2016.
- [51] A. Radoi, "Multimodal satellite image time series analysis using GAN-based domain translation and matrix profile," *Remote Sens.*, vol. 14, no. 15, 2022, Art. no. 3734.
- [52] H. Chen, N. Yokoya, C. Wu, and B. Du, "Unsupervised multimodal change detection based on structural relationship graph representation learning," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5635318.
- [53] A. Farasin, G. Nini, P. Garza, and C. Rossi, "Unsupervised burned area estimation through satellite tiles: A multimodal approach by means of image segmentation over remote sensing imagery," in *Proc. MACLEAN, PKDD/ECML*, 2019, pp. 1–10.
- [54] C. Rambour, N. Audebert, E. Koeniguer, B. L. Saux, M. Crucianu, and M. Datcu, "Flood detection in time series of optical and SAR images," *Int. Arch. Photogrammetry, Remote Sens. Spatial Inf. Sci.*, vol. 43, no. B2, pp. 1343–1346, 2020.
- [55] A. Rowlands and A. Sarris, "Detection of exposed and subsurface archaeological remains using multi-sensor remote sensing," *J. Archaeological Sci.*, vol. 34, no. 5, pp. 795–803, 2007.
- [56] A. Ziemann, C. X. Ren, and J. Theiler, "Multi-sensor anomalous change detection at scale," *Proc. SPIE*, vol. 10986, 2019, Art. no. 1098615.
- [57] M. Rodger and R. Guida, "Classification-aided SAR and AIS data fusion for space-based maritime surveillance," *Remote Sens.*, vol. 13, no. 1, 2020, Art. no. 104.
- [58] J. Florath, S. Keller, R. A.-d. Rio, S. Hinz, G. Staub, and M. Weinmann, "Glacier monitoring based on multi-spectral and multi-temporal satellite data: A case study for classification with respect to different snow and ice types," *Remote Sens.*, vol. 14, no. 4, 2022, Art. no. 845.
- [59] S. Saad et al., "Fusion of multispectral remote-sensing data through GiS-based overlay method for revealing potential areas of hydrothermal mineral resources," *Minerals*, vol. 12, no. 12, 2022, Art. no. 1577.
- [60] J. Li, K. Zheng, J. Yao, L. Gao, and D. Hong, "Deep unsupervised blind hyperspectral and multispectral data fusion," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, 2022, Art. no. 6007305.
- [61] F. Xia, Z. Lou, D. Sun, H. Li, and L. Quan, "Weed resistance assessment through airborne multimodal data fusion and deep learning: A novel approach towards sustainable agriculture," *Int. J. Appl. Earth Observ. Geoinf.*, vol. 120, 2023, Art. no. 103352.
- [62] L. Quan et al., "Multimodal remote sensing application for weed competition time series analysis in maize farmland ecosystems," *J. Environ. Manage.*, vol. 344, 2023, Art. no. 118376.
- [63] D. Hong et al., "SpectralGPT: Spectral foundation model," *TPAMI*, to be published, doi: [10.48550/arXiv.2311.07113](https://doi.org/10.48550/arXiv.2311.07113).
- [64] Y. Chen, M. Zhao, and L. Bruzzone, "Incomplete multimodal learning for remote sensing data fusion," *arXiv:2304.11381*.
- [65] R. Salakhutdinov and G. Hinton, "Deep Boltzmann machines," in *Proc. Artif. Intell. Statist.*, 2009, pp. 448–455.
- [66] W. Shao, X. Shi, and P. S. Yu, "Clustering on multiple incomplete datasets via collective kernel learning," in *Proc. IEEE 13th Int. Conf. Data Mining*, 2013, pp. 1181–1186.
- [67] I. Sheikh, R. Chakraborty, and S. K. Koppurapu, "Audio-visual fusion for sentiment classification using cross-modal autoencoder," in *Proc. 32nd Conf. Neural Inf. Process. Syst.*, 2018, pp. 1–4.
- [68] L. Tran, X. Liu, J. Zhou, and R. Jin, "Missing modalities imputation via cascaded residual autoencoder," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 4971–4980.
- [69] J. Zhao, R. Li, and Q. Jin, "Missing modality imagination network for emotion recognition with uncertain missing modalities," in *Proc. 59th Annu. Meeting Assoc. Comput. Linguistics 11th Int. Joint Conf. Natural Lang. Process.*, 2021, pp. 2608–2618.
- [70] D. Bank, N. Koenigstein, and R. Giryes, "Autoencoders," in *Machine Learning*. Cambridge, MA, USA: Academic, Mar. 2020, pp. 193–208.
- [71] Q. Wang, L. Zhan, P. Thompson, and J. Zhou, "Multimodal learning with incomplete modalities by knowledge distillation," in *Proc. 26th ACM SIGKDD Int. Conf. Knowl. Discov. Data Mining*, 2020, pp. 1828–1838.
- [72] C. Zhang et al., "M3Care: Learning with missing modalities in multimodal healthcare data," in *Proc. 28th ACM SIGKDD Conf. Knowl. Discov. Data Mining*, 2022, pp. 2418–2428.
- [73] S. Woo, S. Lee, Y. Park, M. A. Nugroho, and C. Kim, "Towards good practices for missing modality robust action recognition," in *Proc. AAAI Conf. Artif. Intell.*, 2022, pp. 2776–2784.
- [74] I. J. Goodfellow et al., "Generative adversarial networks," *Commun. ACM*, vol. 63, pp. 139–144, Jun. 2020.
- [75] W. Hao, Z. Zhang, and H. Guan, "CMCGAN: A uniform framework for cross-modal visual-audio mutual generation," in *Proc. 32nd Conf. Artif. Intell., 13th Innov. Appl. Artif. Intell. Conf., 8th Symp. Educ. Adv. Artif. Intell.*, 2017, pp. 6886–6893.
- [76] Y.-H. H. Tsai, P. P. Liang, A. Zadeh, L.-P. Morency, and R. Salakhutdinov, "Learning factorized multimodal representations," in *Proc. Int. Conf. Learn. Representations*, May 2019, doi: [10.48550/arXiv.1806.06176](https://doi.org/10.48550/arXiv.1806.06176).
- [77] B. Bischke, P. Helber, F. König, D. Borth, and A. Dengel, "Overcoming missing and incomplete modalities with generative adversarial networks for building footprint segmentation," in *Proc. Int. Conf. Content-Based Multimedia Indexing*, 2018, pp. 1–6.
- [78] C. Du, C. Du, and H. He, "Multimodal deep generative adversarial models for scalable doubly semi-supervised learning," *Inf. Fusion*, vol. 68, pp. 118–130, Apr. 2021.
- [79] M. Ma, J. Ren, L. Zhao, S. Tulyakov, C. Wu, and X. Peng, "SMIL: Multimodal learning with severely missing modality," in *Proc. AAAI Conf. Artif. Intell.*, 2021, pp. 2302–2310.
- [80] L. Cai, Z. Wang, H. Gao, D. Shen, and S. Ji, "Deep adversarial learning for multi-modality missing data completion," in *Proc. 24th Int. Conf. Knowl. Discov. Data Mining*, 2018, pp. 1158–1166.
- [81] M. Kampffmeyer, A.-B. Salberg, and R. Jenssen, "Urban land cover classification with missing data using deep convolutional neural networks," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2017, pp. 5161–5164.
- [82] S. Pande, A. Banerjee, S. Kumar, B. Banerjee, and S. Chaudhuri, "An adversarial approach to discriminative modality distillation for remote sensing image classification," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshop*, 2019, pp. 4571–4580.

- [83] X. Li, L. Lei, Y. Sun, and G. Kuang, "Dynamic-hierarchical attention distillation with synergetic instance selection for land cover classification using missing heterogeneity images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, Jan. 2021, Art. no. 4400816.
- [84] S. Wei, Y. Luo, X. Ma, P. Ren, and C. Luo, "MSH-Net: Modality-shared hallucination with joint adaptation distillation for remote sensing image classification using missing modalities," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, Apr. 2023, Art. no. 4402615.
- [85] R. Cao et al., "Deep learning-based remote and social sensing data fusion for urban region function recognition," *ISPRS J. Photogramm. Remote Sens.*, vol. 163, pp. 82–97, 2023.
- [86] S. Kumar, B. Banerjee, and S. Chaudhuri, "Improved landcover classification using online spectral data hallucination," *Neurocomputing*, vol. 439, pp. 316–326, 2023.
- [87] J. Kang et al., "DisOptNet: Distilling semantic knowledge from optical images for weather-independent building segmentation," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2023, Art. no. 4706315.
- [88] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," in *Proc. NIPS Deep Learn. Representation Learn. Workshop*, 2015.
- [89] Z. Ding, M. Shao, and Y. Fu, "Missing modality transfer learning via latent low-rank constraint," *IEEE Trans. Image Process.*, vol. 24, no. 11, pp. 4322–4334, Jul. 2015.
- [90] S.-I. Seo, S. Na, and J. Kim, "HMTL: Heterogeneous modality transfer learning for audio-visual sentiment analysis," *IEEE Access*, vol. 8, pp. 140426–140437, 2020.
- [91] S. Wei, Y. Luo, and C. Luo, "MMANet: Margin-aware distillation and modality-aware regularization for incomplete multimodal learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 20039–20049.
- [92] N. Garcia, P. Morerio, and V. Murino, "Modality distillation with multiple stream networks for action recognition," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 103–118.
- [93] Y. Wang et al., "ACN: Adversarial co-training network for brain tumor segmentation with missing modalities," in *Proc. Med. Image Comput. Comput. Assist. Interv.*, 2021, pp. 410–420.
- [94] H. Maheshwari, Y.-C. Liu, and Z. Kira, "Missing modality robustness in semi-supervised multi-modal semantic segmentation," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis.*, 2023, pp. 1020–1030.
- [95] Y. Kong, F. Hong, H. Leung, and X. Peng, "A fusion method of optical image and SAR image based on Dense-UGAN and GramSchmidt transformation," *Remote Sens.*, vol. 13, no. 21, Oct. 2021, Art. no. 4274.
- [96] C. Grohnfeldt, M. Schmitt, and X. Zhu, "A conditional generative adversarial network to fuse SAR and multispectral optical data for cloud removal from Sentinel-2 images," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2018, pp. 1726–1729.
- [97] J. Gao, Q. Yuan, J. Li, H. Zhang, and X. Su, "Cloud removal with fusion of high resolution optical and SAR images using generative adversarial networks," *Remote Sens.*, vol. 12, no. 1, Jan. 2020, Art. no. 191.
- [98] Y. Zhao, S. Shen, J. Hu, Y. Li, and J. Pan, "Cloud removal using multimodal GAN with adversarial consistency loss," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, Jul. 2021, Art. no. 8015605.
- [99] Y. Gao, M. Zhang, W. Li, X. Song, X. Jiang, and Y. Ma, "Adversarial complementary learning for multisource remote sensing classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, Mar. 2023, Art. no. 5505613.
- [100] B. Zhu, J. Zhang, T. Tang, and Y. Ye, "SFOC: A novel multi-directional and multi-scale structural descriptor for multimodal remote sensing image matching," *Int. Arch. Photogrammetry, Remote Sens. Spatial Inf. Sci.*, vol. 43, pp. 113–120, May 2022.
- [101] Z. Huang, C. O. Dumitru, Z. Pan, B. Lei, and M. Datcu, "Classification of large-scale high-resolution SAR images with deep transfer learning," *IEEE Geosci. Remote Sens. Lett.*, vol. 18, no. 1, pp. 107–111, Jan. 2020.
- [102] J. Kang, R. Fernandez-Beltran, P. Duan, X. Kang, and A. J. Plaza, "Robust normalized softmax loss for deep metric learning-based characterization of remote sensing images with label noise," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 10, pp. 8798–8811, Oct. 2020.
- [103] J. Kang, R. Fernandez-Beltran, X. Kang, J. Ni, and A. Plaza, "Noise-tolerant deep neighborhood embedding for remotely sensed images with label noise," *IEEE J. Sel. Topics Appl. Earth Obs. Remote Sens.*, vol. 14, pp. 2551–2562, 2021.
- [104] E. Protopapadakis, A. Doulamis, N. Doulamis, and E. Maltezos, "Stacked autoencoders driven by semi-supervised learning for building extraction from near infrared remote sensing imagery," *Remote Sens.*, vol. 13, no. 3, Jan. 2021, Art. no. 371.
- [105] Y. Ding et al., "AF2GNN: Graph convolution with adaptive filters and aggregator fusion for hyperspectral image classification," *Inform. Sci.*, vol. 602, pp. 201–219, Jul. 2022.
- [106] Y. Li, Y. Zhang, and Z. Zhu, "Error-tolerant deep learning for remote sensing image scene classification," *IEEE Trans. Cybern.*, vol. 51, no. 4, pp. 1756–1768, Apr. 2021.
- [107] R. Shang et al., "SAR image segmentation based on constrained smoothing and hierarchical label correction," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, May 2021, Art. no. 5102216.
- [108] N. Ahmed, R. M. Rahman, M. S. G. Adnan, and B. Ahmed, "Dense prediction of label noise for learning building extraction from aerial drone imagery," *Int. J. Remote Sens.*, vol. 42, pp. 8906–8929, 2021.
- [109] Y. Cao and X. Huang, "A full-level fused cross-task transfer learning method for building change detection using noise-robust pretrained networks on crowdsourced labels," *Remote Sens. Environ.*, vol. 284, Jan. 2023, Art. no. 113371.
- [110] Q. Li, Y. Chen, and P. Ghamisi, "Complementary learning-based scene classification of remote sensing images with noisy labels," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, 2021, Art. no. 8021105.
- [111] P. Li et al., "Robust deep neural networks for road extraction from remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 7, pp. 6182–6197, Jul. 2020.
- [112] R. Hinami, J. Liang, S. Satoh, and A. Hauptmann, "Multimodal co-training for selecting good examples from Webly labeled video," 2018, *arXiv:1804.06057*.
- [113] J.-C. Hou, S.-S. Wang, Y.-H. Lai, Y. Tsao, H.-W. Chang, and H.-M. Wang, "Audio-visual speech enhancement using multimodal deep convolutional neural networks," *IEEE Trans. Emerg. Topics Comput. Intell.*, vol. 2, no. 2, pp. 117–128, Apr. 2018.
- [114] Y. Ding, X. Yu, and Y. Yang, "RFNet: Region-aware fusion network for incomplete multi-modal brain tumor segmentation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 3955–3964.
- [115] M. Ma, J. Ren, L. Zhao, D. Testuggine, and X. Peng, "Are multimodal transformers robust to missing modality?," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 18156–18165.
- [116] A. Vaswani et al., "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 6000–6010.
- [117] A. Jaegle et al., "Perceiver IO: A general architecture for structured inputs & outputs," in *Proc. Int. Conf. Learn. Representations*, Jan. 2022.
- [118] N. Kieu, K. Nguyen, S. Sridharan, and C. Fookes, "General-purpose multimodal transformer meets remote sensing semantic segmentation," in *Proc. Int. Conf. Learn. Representations*, 2023, pp. 1–8.
- [119] Y. Zhang et al., "mmFormer: Multimodal medical transformer for incomplete multimodal learning of brain tumor segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Interv.*, 2022, pp. 107–117.
- [120] H. Ting and M. Liu, "Multimodal transformer of incomplete MRI data for brain tumor segmentation," *IEEE J. Biomed. Health Inform.*, vol. 28, no. 1, pp. 89–99, Jan. 2024.
- [121] Y. Diao, F. Li, and Z. Li, "Joint learning-based feature reconstruction and enhanced network for incomplete multi-modal brain tumor segmentation," *Comput. Biol. Med.*, vol. 163, Jul. 2023, Art. no. 107234.
- [122] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick, "Masked autoencoders are scalable vision learners," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 16000–16009.
- [123] Q. Li, G. Yu, J. Wang, and Y. Liu, "A deep multimodal generative and fusion framework for class-imbalanced multimodal data," *Multimed. Tools Appl.*, vol. 79, no. 33, pp. 25023–25050, Sep. 2020.
- [124] D. Ji, D. Oh, Y. Hyun, O.-M. Kwon, and M.-J. Park, "How to handle noisy labels for robust learning from uncertainty," *Neural Netw.*, vol. 143, pp. 209–217, Nov. 2021.
- [125] H. Song and W. Yang, "GSCCTL: A general semi-supervised scene classification method for remote sensing images based on clustering and transfer learning," *Int. J. Remote Sens.*, vol. 43, no. 15/16, pp. 5976–6000, Aug. 2022.
- [126] W. Zha, L. Hu, C. Duan, and Y. Li, "Semi-supervised learning-based satellite remote sensing object detection method for power transmission towers," *Energy Rep.*, vol. 9, pp. 15–27, Sep. 2023.
- [127] E. Khoramshahi et al., "A novel deep multi-image object detection approach for detecting alien barleys in oat fields using RGB UAV images," *Remote Sens.*, vol. 15, no. 14, Jul. 2023, Art. no. 3582.
- [128] L. Liao, L. Du, and Y. Guo, "Semi-supervised SAR target detection based on an improved faster R-CNN," *Remote Sens.*, vol. 14, no. 1, Dec. 2021, Art. no. 143.



- [129] W. Chen, X. Zheng, and X. Lu, "Semisupervised spectral degradation constrained network for spectral super-resolution," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, May 2021, Art. no. 5506205.
- [130] J. Hu, L. Shen, S. Albanie, G. Sun, and E. Wu, "Squeeze-and-excitation networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 8, pp. 2011–2023, Aug. 2020.
- [131] C. Wang et al., "Semisupervised learning-based SAR ATR via self-consistent augmentation," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 6, pp. 4862–4873, Jun. 2021.
- [132] Y. He, J. Wang, C. Liao, B. Shan, and X. Zhou, "ClassHyPer: ClassMix-Based hybrid perturbations for deep semi-supervised semantic segmentation of remote sensing imagery," *Remote Sens.*, vol. 14, no. 4, Feb. 2022, Art. no. 879.
- [133] L. Zhang et al., "MDMASNet: A dual-task interactive semi-supervised remote sensing image segmentation method," *Signal Process.*, vol. 212, Nov. 2023, Art. no. 109152.
- [134] W. Miao, J. Geng, and W. Jiang, "Semi-supervised remote-sensing image scene classification using representation consistency siamese network," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, Jan. 2022, Art. no. 5616614.
- [135] B. Xi et al., "Semisupervised cross-scale graph prototypical network for hyperspectral image classification," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 34, no. 11, pp. 9337–9351, Nov. 2022.
- [136] Y. Ding, Y. Guo, Y. Chong, S. Pan, and J. Feng, "Global consistent graph convolutional network for hyperspectral image classification," *IEEE Trans. Instrum. Meas.*, vol. 70, Feb. 2021, Art. no. 5501516.
- [137] Y. Hua, D. Marcos, L. Mou, X. X. Zhu, and D. Tuija, "Semantic segmentation of remote sensing images with sparse annotations," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, Jan. 2021, Art. no. 8006305.
- [138] A. Kirillov et al., "Segment anything," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2023, pp. 3992–4003.
- [139] S. Wang, W. Chen, S. M. Xie, G. Azzari, and D. B. Lobell, "Weakly supervised deep learning for segmentation of remote sensing imagery," *Remote Sens.*, vol. 12, no. 2, 2023, Art. no. 207.
- [140] Y. Li, W. Chen, Y. Zhang, C. Tao, R. Xiao, and Y. Tan, "Accurate cloud detection in high-resolution remote sensing imagery by weakly supervised deep learning," *Remote Sens. Environ.*, vol. 250, 2023, Art. no. 112045.
- [141] F. Qian, M. Yin, X.-Y. Liu, Y. J. Wang, C. Lu, and G.-M. Hu, "Unsupervised seismic facies analysis via deep convolutional autoencoders," *Geophysics*, vol. 83, no. 3, pp. A39–A43, May 2018.
- [142] Y. Duan, X. Zheng, L. Hu, and L. Sun, "Seismic facies analysis based on deep convolutional embedded clustering," *Geophysics*, vol. 84, no. 6, pp. IM87–IM97, Nov. 2019.
- [143] S. M. Mousavi, W. Zhu, W. Ellsworth, and G. Beroza, "Unsupervised clustering of seismic signals using deep convolutional autoencoders," *IEEE Geosci. Remote Sens. Lett.*, vol. 16, no. 11, pp. 1693–1697, Nov. 2019.
- [144] Y. Niu, Y. D. Wang, P. Mostaghimi, P. Swietojanski, and R. T. Armstrong, "An innovative application of generative adversarial networks for physically accurate rock images with an unprecedented field of view," *Geophys. Res. Lett.*, vol. 47, no. 23, Dec. 2020, Art. no. e2020GL089029.
- [145] X. Du, X. Zheng, X. Lu, and A. A. Doudkin, "Multisource remote sensing data classification with graph fusion network," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 12, pp. 10062–10072, Dec. 2021.
- [146] L. T. Luppino et al., "Code-aligned autoencoders for unsupervised change detection in multimodal remote sensing images," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 35, no. 1, pp. 60–72, Jan. 2024.
- [147] Z. Han, D. Hong, L. Gao, J. Yao, B. Zhang, and J. Chanussot, "Multimodal hyperspectral unmixing: Insights from attention networks," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2023, Art. no. 5524913.
- [148] Y. Ye, T. Tang, B. Zhu, C. Yang, B. Li, and S. Hao, "A multi-scale framework with unsupervised learning for remote sensing image registration," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, Apr. 2022, Art. no. 5622215.
- [149] J. Yao, D. Hong, H. Wang, H. Liu, and J. Chanussot, "UCSL: Toward unsupervised common subspace learning for cross-modal image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5514212.
- [150] E. J. van Engelen and H. H. Hoos, "A survey on semi-supervised learning," *Mach. Learn.*, vol. 109, no. 2, pp. 373–440, Feb. 2020.
- [151] 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.
- [152] T. Long, P. Mettes, H. T. Shen, and C. G. M. Snoek, "Searching for Actions on the Hyperbole," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 1138–1147.
- [153] D. P. Kingma and M. Welling, "Auto-encoding variational Bayes," in *Proc. Int. Conf. Learn. Representations*, Apr. 2014.
- [154] Z.-H. Zhou, "A brief introduction to weakly supervised learning," *Nat. Sci. Rev.*, vol. 5, no. 1, pp. 44–53, Jan. 2018.
- [155] R. J. Chen et al., "Pan-cancer integrative histology-genomic analysis via multimodal deep learning," *Cancer Cell*, vol. 40, no. 8, pp. 865–878, Aug. 2022.
- [156] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "GRAD-CAM: Visual explanations from deep networks via gradient-based localization," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2016, pp. 618–626.
- [157] L. Khelifi and M. Mignotte, "Deep learning for change detection in remote sensing images: Comprehensive review and meta-analysis," *IEEE Access*, vol. 8, pp. 126385–126400, 2020.
- [158] S. Yu and J. Ma, "Deep learning for geophysics: Current and future trends," *Rev. Geophys.*, vol. 59, no. 3, 2023, Art. no. e2021RG000742.
- [159] T. Baltrusaitis, C. Ahuja, and L.-P. Morency, "Multimodal machine learning: A survey and taxonomy," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 2, pp. 423–443, Feb. 2019.
- [160] Y. Ye, J. Shan, L. Bruzzone, and L. Shen, "Robust registration of multimodal remote sensing images based on structural similarity," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 5, pp. 2941–2958, May 2017.
- [161] S. Cui, M. Xu, A. Ma, and Y. Zhong, "Modality-free feature detector and descriptor for multimodal remote sensing image registration," *Remote Sens.*, vol. 12, no. 18, Sep. 2020, Art. no. 2937.
- [162] X. Zhang, C. Leng, Y. Hong, Z. Pei, I. Cheng, and A. Basu, "Multimodal remote sensing image registration methods and advancements: A survey," *Remote Sens.*, vol. 13, no. 24, Dec. 2021, Art. no. 5128.
- [163] Y. Zhan, Z. Xiong, and Y. Yuan, "RSVG: Exploring data and models for visual grounding on remote sensing data," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5604513.
- [164] M. Farooq, M. N. Dailey, A. Mahmood, J. Moonrinta, and M. Ekpanyapong, "Human face super-resolution on poor quality surveillance video footage," *Neural Comput. Appl.*, vol. 33, no. 20, pp. 13505–13523, Oct. 2021.
- [165] A. B. Molini, D. Valsesia, G. Fracastoro, and E. Magli, "DeepSUM: Deep neural network for super-resolution of unregistered multitemporal images," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 5, pp. 3644–3656, May 2020.
- [166] A. Moghimi, T. Celik, A. Mohammadzadeh, and H. Kusetogullari, "Comparison of keypoint detectors and descriptors for relative radiometric normalization of bitemporal remote sensing images," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 4063–4073, 2021.
- [167] Y. Qu, H. Qi, C. Kwan, N. Yokoya, and J. Chanussot, "Unsupervised and unregistered hyperspectral image super-resolution with mutual dirichlet-Net," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, May 2021, Art. no. 5507018.
- [168] S. Fu, F. Xu, and Y.-Q. Jin, "Reciprocal translation between SAR and optical remote sensing images with cascaded-residual adversarial networks," *Sci. China Inf. Sci.*, vol. 64, no. 2, pp. 1–15, Feb. 2021.
- [169] K. Zheng, L. Gao, D. Hong, B. Zhang, and J. Chanussot, "NonRegSRNet: A nonrigid registration hyperspectral super-resolution network," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, Dec. 2021, Art. no. 5520216.
- [170] C. Ren, X. Wang, J. Gao, X. Zhou, and H. Chen, "Unsupervised change detection in satellite images with generative adversarial network," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 12, pp. 10047–10061, Dec. 2021.
- [171] A. S. Lin et al., "A recipe for creating multimodal aligned datasets for sequential tasks," in *Proc. 58th Annu. Meeting Assoc. Comput. Linguistics*, 2020, pp. 4871–4884.
- [172] A. Pérez-Suay et al., "Interpretability of recurrent neural networks in remote sensing," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2020, pp. 3991–3994.
- [173] I. Kakogeorgiou and K. Karantzalos, "Evaluating explainable artificial intelligence methods for multi-label deep learning classification tasks in remote sensing," *Int. J. Appl. Earth Observ. Geoinf.*, vol. 103, 2021, Art. no. 102520.
- [174] C.-Y. Hsu and W. Li, "Explainable GeoAI: Can saliency maps help interpret artificial intelligence's learning process? An empirical study on natural feature detection," *Int. J. Geographical Inf. Sci.*, vol. 37, no. 5, pp. 963–987, 2023.

- [175] X. Guo et al., “Visual explanations with detailed spatial information for remote sensing image classification via channel saliency,” *Int. J. Appl. Earth Observ. Geoinf.*, vol. 118, 2023, Art. no. 103244.
- [176] M. T. Ribeiro, S. Singh, and C. Guestrin, ““Why should I trust you?” Explaining the predictions of any classifier,” in *Proc. 22nd ACM Int. Conf. Knowl. Discov. Data Mining*, 2016, pp. 1135–1144.
- [177] S. M. Lundberg and S.-I. Lee, “A unified approach to interpreting model predictions,” in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 4768–4777.
- [178] A. Goldstein, A. Kapelner, J. Bleich, and E. Pitkin, “Peeking inside the black box: Visualizing statistical learning with plots of individual conditional expectation,” *J. Comput. Graphical Statist.*, vol. 24, no. 1, pp. 44–65, 2015.
- [179] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, “Grad-CAM: Visual explanations from deep networks via gradient-based localization,” in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 618–626.
- [180] A. Shrikumar, P. Greenside, and A. Kundaje, “Learning important features through propagating activation differences,” in *Proc. Int. Conf. Mach. Learn.*, 2017, pp. 3145–3153.
- [181] O. Hall, M. Ohlsson, and T. Rönkvallsson, “A review of explainable ai in the satellite data, deep machine learning, and human poverty domain,” *Patterns*, vol. 3, no. 10, 2022, Art. no. 100600.
- [182] D. Browne, M. Giering, and S. Prestwich, “PulseNetOne: Fast unsupervised pruning of convolutional neural networks for remote sensing,” *Remote Sens.*, vol. 12, no. 7, 2020, Art. no. 1092.
- [183] Y. Lei, D. Wang, S. Yang, J. Shi, D. Tian, and L. Min, “Network collaborative pruning method for hyperspectral image classification based on evolutionary multi-task optimization,” *Remote Sens.*, vol. 15, no. 12, 2023, Art. no. 3084.
- [184] L. Gu, Q. Fang, Z. Wang, E. Popov, and G. Dong, “Learning lightweight and superior detectors with feature distillation for onboard remote sensing object detection,” *Remote Sens.*, vol. 15, no. 2, 2023, Art. no. 370.
- [185] B. Grabowski et al., “Squeezing nnU-nets with knowledge distillation for on-board cloud detection,” 2023, *arXiv:2306.09886*.
- [186] J. C. O. Koh, G. Spangenberg, and S. Kant, “Automated machine learning for high-throughput image-based plant phenotyping,” *Remote Sens.*, vol. 13, no. 5, 2021, Art. no. 858.
- [187] Z. Zhang, S. Liu, Y. Zhang, and W. Chen, “RS-DARTS: A convolutional neural architecture search for remote sensing image scene classification,” *Remote Sens.*, vol. 14, no. 1, 2021, Art. no. 141.
- [188] C. Peng, Y. Li, R. Shang, and L. Jiao, “RSBNet: One-shot neural architecture search for a backbone network in remote sensing image recognition,” *Neurocomputing*, vol. 537, pp. 110–127, 2023.
- [189] G. Liu, Y. Li, Y. Chen, R. Shang, and L. Jiao, “Pol-NAS: A neural architecture search method with feature selection for PolSAR image classification,” *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 15, pp. 9339–9354, 2022.
- [190] C. Peng, Y. Li, L. Jiao, and R. Shang, “Efficient convolutional neural architecture search for remote sensing image scene classification,” *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 7, pp. 6092–6105, Jul. 2020.
- [191] P. Gudzius, O. Kurasova, V. Darulis, and E. Filatovas, “AutoML-based neural architecture search for object recognition in satellite imagery,” *Remote Sens.*, vol. 15, no. 1, 2022, Art. no. 91.
- [192] J. Liu, Z. Yuan, Z. Pan, Y. Fu, L. Liu, and B. Lu, “Diffusion model with detail complement for super-resolution of remote sensing,” *Remote Sens.*, vol. 14, no. 19, 2022, Art. no. 4834.
- [193] K. Jiang, Z. Wang, P. Yi, G. Wang, T. Lu, and J. Jiang, “Edge-enhanced GAN for remote sensing image superresolution,” *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 8, pp. 5799–5812, Aug. 2019.
- [194] J. Zhang, Z. Cai, F. Chen, and D. Zeng, “Hyperspectral image denoising via adversarial learning,” *Remote Sens.*, vol. 14, no. 8, 2022, Art. no. 1790.
- [195] J. Ma, W. Yu, C. Chen, P. Liang, X. Guo, and J. Jiang, “Pan-GAN: An unsupervised pan-sharpening method for remote sensing image fusion,” *Inf. Fusion*, vol. 62, pp. 110–120, 2020.
- [196] W. Diao, F. Zhang, J. Sun, Y. Xing, K. Zhang, and L. Bruzzone, “ZeR-GAN: Zero-reference GAN for fusion of multispectral and panchromatic images,” *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 34, no. 11, pp. 8195–8209, Nov. 2023.
- [197] Z. H. Cao, S. Q. Cao, X. Wu, J. M. Hou, R. Ran, and L.-J. Deng, “DDRF: Denoising diffusion model for remote sensing image fusion,” 2023, *arXiv:2304.04774*.
- [198] A. Sebaq and M. ElHelw, “RSDiff: Remote sensing image generation from text using diffusion model,” 2023, *arXiv:2309.02455*.
- [199] Y. Zhang et al., “mmFormer: Multimodal medical transformer for incomplete multimodal learning of brain tumor segmentation,” in *Proc. Med. Image Comput. Comput. Assist. Interv.*, 2022, pp. 107–117.
- [200] Z. Yuan et al., “Efficient and controllable remote sensing fake sample generation based on diffusion model,” *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5615012.
- [201] J. Jakubik et al., “Prithvi-100 M,” Aug. 2023, doi: [10.57967/hf/0952](https://doi.org/10.57967/hf/0952).
- [202] X. Sun et al., “RingMo: A remote sensing foundation model with masked image modeling,” *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2022, Art. no. 5612822.
- [203] D. Wang et al., “Advancing plain vision transformer toward remote sensing foundation model,” *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2022, Art. no. 5607315.
- [204] K. Cha, J. Seo, and T. Lee, “A billion-scale foundation model for remote sensing images,” 2023, *arXiv:2304.05215*.
- [205] L. Dong et al., “Unified language model pre-training for natural language understanding and generation,” in *Proc. Adv. Neural Inf. Process. Syst.*, 2019, pp. 13063–13075.
- [206] C. Sun, A. Myers, C. Vondrick, K. Murphy, and C. Schmid, “VideoBERT: A joint model for video and language representation learning,” in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 7464–7473.
- [207] J. Wang et al., “All in one: Exploring unified video-language pre-training,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 6598–6608.