SCIENCE CHINA
Information Sciences

April 2023, Vol. 66 140301:1–140301:28
https://doi.org/10.1007/s11432-022-3588-0

• REVIEW •
Special Topic: Artificial Intelligence Innovation in Remote Sensing

From single- to multi-modal remote sensing imagery interpretation: a survey and taxonomy

Xian SUN[1,2,3,4], Yu TIAN[1,2,3,4], Wanxuan LU[1,2], Peijin WANG[1,2], Ruigang NIU[1,2,3,4], Hongfeng YU[1,2] & Kun FU[1,2,3,4*]

[1]*Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing* 100190, *China;*
[2]*Key Laboratory of Network Information System Technology (NIST), Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing* 100190, *China;*
[3]*University of Chinese Academy of Sciences, Beijing* 100190, *China;*
[4]*School of Electronic, Electrical and Communication Engineering, University of Chinese Academy of Sciences, Beijing* 100190, *China*

Received 12 June 2022/Revised 13 August 2022/Accepted 22 September 2022/Published online 27 March 2023

Abstract Modality is a source or form of information. Through various modal information, humans can perceive the world from multiple perspectives. Simultaneously, the observation of remote sensing (RS) is multimodal. We observe the world macroscopically through panchromatic, Lidar, and other modal sensors. Multimodal observation of remote sensing has become an active area, which is beneficial for urban planning, monitoring, and other applications. Despite numerous advancements in this area, there has still not been a comprehensive assessment that provides a systematic overview with a unified evaluation. Accordingly, in this survey paper, we first highlight the key differences between single- and multimodal RS imagery interpretation, then use these differences to guide our research survey of multimodal RS imagery interpretation in a cascaded structure. Finally, some potential future research directions are explored and outlined. We hope that this survey will serve as a starting point for researchers to review state-of-the-art developments and work on multimodal research.

Keywords multimodal, remote sensing, image interpretation, feature fusion, co-learning

Citation Sun X, Tian Y, Lu W X, et al. From single- to multi-modal remote sensing imagery interpretation: a survey and taxonomy. Sci China Inf Sci, 2023, 66(4): 140301, https://doi.org/10.1007/s11432-022-3588-0

1 Introduction

The development of high spatial resolution (HSR) satellites equipped with various sensors has brought a rich data source. Benefiting from this, HSR remote sensing images have the property of multimodal, which provides more challenging issues for the remote sensing and computer vision community. Compared to single-mode observation, it provides more information about sensors, angles, resolution, and time than single-mode observation, which is critical and brings a huge boost to the cutting edge [1–4]. This has given particular urgency to explore how to make full use of the multimodal remote sensing (RS) imagery for Earth observation. It also further declares that the RS imagery interpretation is gradually from single- to multi-model, which provides more observation details to refine the scene information. Therefore, we present a comprehensive and timely survey to help researchers learn and apply the cutting-edge technology of multimodal remote sensing interpretation, and lay a solid foundation for further experiments.

Multimodal RS imagery interpretation (MRSII) is an emerging direction in the communities of Earth observation and computer vision. It is challenging and has greater application value than single-modal. From the perspective of properties, there are four reasons at least.

(1) The image data is multispectral. As shown in Figure 1[1)], the imaging mechanisms and spectral bands of various sensors are different. While spatial image information of the scene is obtained, the

* Corresponding author (email: kunfuiecas@gmail.com)
1) Prasad S. https://hyperspectral.ee.uh.edu/.

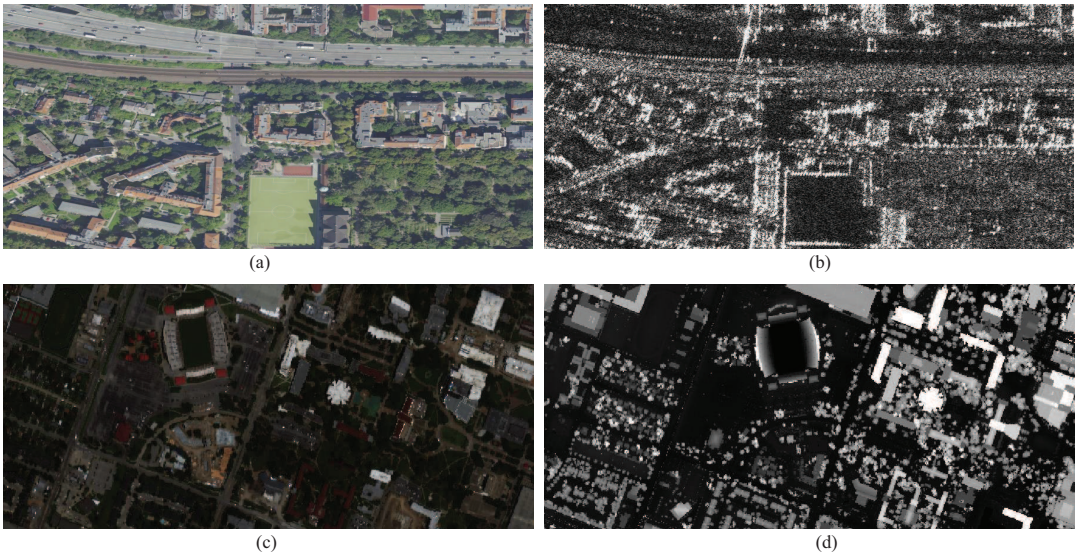

Figure 1 (Color online) Examples of various modal sensor images. (a) Optical image and (b) SAR image are a pair of images in the same scene; (c) hyperspectral image and (d) nDSM are the same area from the Houston dataset.

Table 1 Example of some famous satellites with different modes. Resolution represents the highest resolution of the satellites

Sensor	Resolution (m)	Country/Region	Satellite
Panchromatic	0.31	USA	WV-3, WV-4
	0.41	USA	GeoEye-1
	0.46	USA	WV-1, WV-2
	0.5	China	SuperView-1
		France	Pliades-1, Pliades-2
	1	China	GF-2
	1.5	France	SPOT-6, SPOT-7
	2	China	GF-1, GF-6
Multispectral	1.24	USA	WV-3, WV-4
	1.64	USA	GeoEye-1
	1.8	USA	WV-2
	2	China	SuperView-1
		France	Pliades-1, Pliades-2
	2.5	China	ZY-3 01, ZY-3 02
	4	China	GF-1
		USA	OrbView-3, IKONOS
Hyperspectral	10	China	OVS-1
	30	Chine	ZY-1 02D, GF-5
		USA	EO-1
SAR	1	Japan	ALOS-2
		Korea	Kompsat-5
		China	GF-3
		Germany	TerraSAR-X
	5	Europe	Sentinel-1
		China	HJ-1C
Infrared	10	China	SJ-9B
	20	France	SPOT-5
	30	USA	Landsat-8
	40	China	CBERS-04, GF-4
	150	China	HJ-1B
Laser rangefinder	1	China	ZY-3 02
	3.2	China	GF-7

spectral feature vector of each element reflects the geophysical properties. Due to the above properties, existing pretrain models [5–7] face challenges in channel dimensions and high-level spatial representation.

(2) Multimodal observations over the same area provide complementary information from various perspectives (e.g., three-dimensional space, distance, and altitude). For some Earth observation applications, e.g., urban land planning, 3D-reconstruction, and forest classification, we need multimodal sensors to observe the scene from multiple angles. However, more observation angles bring more computational complexity, which is limited by computation resources.

(3) Table 1 lists the parameters of some famous satellites. Due to the multi-scale property, MRSII requires the system with the adaptive ability to process images of different resolutions. For example, an aircraft may occupy around 400 pixels in a WV-3 image, but only 150 pixels in a Gaofen-2 image. This situation brings an enormous challenge to single-modal models, especially when there are significant scale variations among the same object.

(4) The multimodal images bring a temporal dimension to Earth monitoring, providing new impetus for related researchers. A series of temporal tasks with broad application prospects emerge, e.g., multi-temporal change detection, data fusion, and domain adaptation-based segmentation/detection, ac-

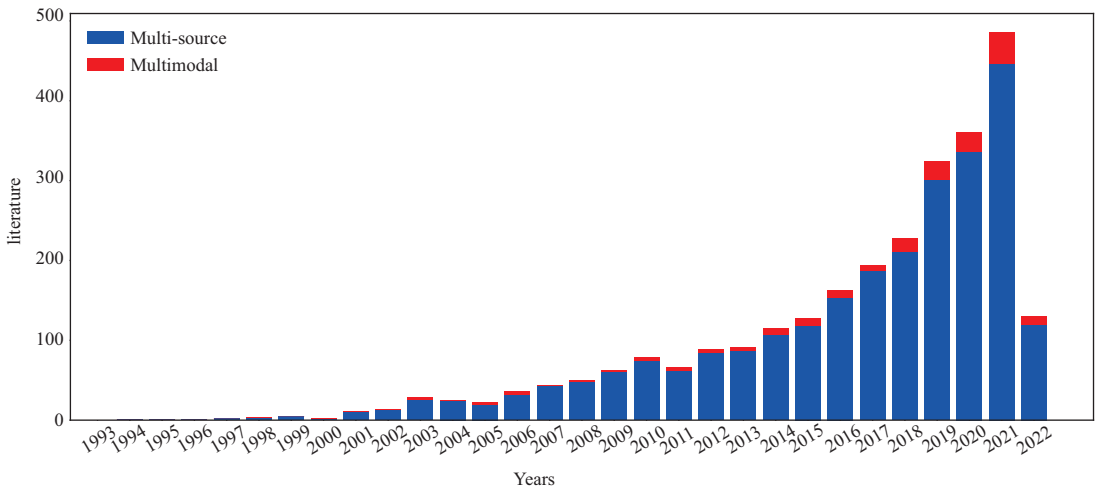

Figure 2 (Color online) Statistics of related literature of multimodal remote sensing. Retrieved data from Web of Science with "multi-source remote sensing" and "multimodal remote sensing" as keywords. The search was conducted on May 2022.

celerating RS imagery interpretation to become multi-dimension and multi-task.

As shown in Figure 2, we retrieve the publications related to MRSII on Web of Science[2)]. From the trend of the number of literature changes in about twenty years, we can find that it is increasing year by year, and MRSII has become a hot topic in remote sensing. Despite over 30 years of research and its importance at the theoretical as well as practical levels, few related surveys have been available. This is an era in which remote sensing imagery interpretation is shifting from single- to multimodal. We hope our work will contribute to the communities of remote sensing and computer vision. The main contributions of this paper are four-fold.

(1) We provide a comprehensive and timely review of the studies on MRSII utilizing multi-platform, multi-sensor data. The exhaustive exposition makes it possible to grasp the whole development process of MRSII, as well as to construct a complete knowledge hierarchy of MRSII.

(2) An easily understandable hierarchical taxonomy is proposed for categorizing MRSII approaches into different tasks: multi-source fusion, multimodal representation, multi-source alignment, cross-modal translation, and co-learning, and then further conducting a more detailed categorization within each task according to the generalizable properties.

(3) We summarize several extensional research topics beyond vanilla MRSII that are emerging lately and discuss recent advances in these topics. These topics are challenging while endowing prominent practical significance to the solution for many realistic imagery interpretation problems.

(4) Based on our summary, we further discuss the applications and future directions of MRSII. We expect that our survey can give a reference for researchers working on multimodal remote sensing imagery interpretation.

2 Taxonomy

At the current time, the fundamental reason why researchers in the field have different opinions on MRSII is that it touches more expansive areas and has fuzzy boundaries. Diverse perspectives lead to different interpretations and taxonomic results. In this study, we refer to [8] and organize MRSII approaches into five major categories (as shown in Figure 3), i.e., multi-source alignment (Section 3), multi-source fusion (Section 4), multimodal representation (Section 5), cross-modal translation (Section 6), and co-learning (Section 7), in light of the core technology to MRSII challenges.

(1) Alignment. Multimodal alignment provides the alignment and matching of different modal information, aiming to find the spatial and temporal connections between modalities. For example, we perform image registration and retrieval between images from different sensors, and retrieval and matching between images and text. These approaches focus on mapping different modalities to a unified semantic space and measuring their similarity by distance.

2) Web of Science. www.webofscience.com/. We retrieve with the keywords multi-source, remote, sensing, or multi-modal, remote, sensing.

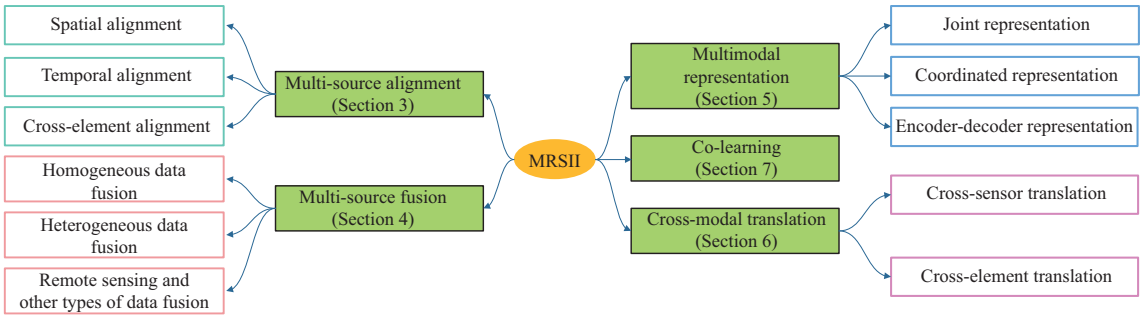

Figure 3 (Color online) The major categories of multimodal remote sensing imagery interpretation (MRSII) in our work, including multi-source alignment, multimodal fusion, multimodal representation, co-learning, and cross-modal translation. They are described and subdivided in the corresponding sections.

(2) Fusion. Multi-source fusion in MRSII aims to join two or more remote sensing data or other observations with complementary information for the same complex scene. By combining the information from them for processing, analysis, and decision making, higher quality data are obtained for object prediction (classification or regression). For example, fusing high-resolution panchromatic images with multispectral images can improve the spatial resolution of multispectral images by several times.

(3) Representation. A fundamental task is to encode the image into a high-level feature space for downstream task analysis. Similarly, in MRSII, the representation is responsible for extracting and abstracting multimodal information into high-level feature vectors. It exploits the complementarity between different modal features and eliminates redundancy to learn better features. For example, to encode for urban classification and 3D building reconstruction, the digital surface model (DSM) and true orthophoto are combined and fed into the same representation space.

(4) Translation. A newly emerging challenge is the translation of information from one modality to another modality. The approaches for this task tend to generate models, and the predicted targets are open-ended or subjective. The generated modality is heterogeneous to the source modality. For example, we use SAR data to generate panchromatic imagery.

(5) Co-learning. One single modal sensor may be scarce for some complex scenarios, so it requires another abundant modality to assist its learning. In some cases where domain adaptation or transfer is needed, inter-modal information can utilize co-learning to aid learning. For example, pre-training the model on resource-rich optical image features and learning on scarce SAR image features can improve performance.

To help illustrate and structure recent work in the emerging research field of MRSII, we further subdivide and summarize each taxonomic class. Different taxonomic classes are not unrelated and in many cases complement each other. They complement each other in a variety of situations, and an excellent multimodal model often requires a combination of more than two techniques. For example, multimodal representation can be used as a backbone model for alignment or translation. In Sections 3–6, we explain these tasks in detail.

3 Multi-source alignment

The multi-source alignment in remote sensing aims to match the original source to the target source, finding corresponding explicit and implicit relationships between heterogeneous data. For example, given two images from different sensors containing the same complex scene, we match or retrieve sub-components of them. Multi-source alignment is an important branch of the MRSII, and relevant work includes image registration [9–11], change detection [12–15], and cross-modal retrieval [16–18].

As shown in Figure 4, we classify multi-source alignment into three types of alignment methods depending on the alignment dimension of the data source: (1) spatial alignment, (2) temporal alignment, and (3) cross-element alignment. Table 2 [9–11, 14, 18–30] lists the different of these methods.

3.1 Spatial alignment

Spatial alignment is mainly the process of image alignment. It is to find the pixel space mapping relationship between a current image and a reference image from the same complex scene, thus achieving

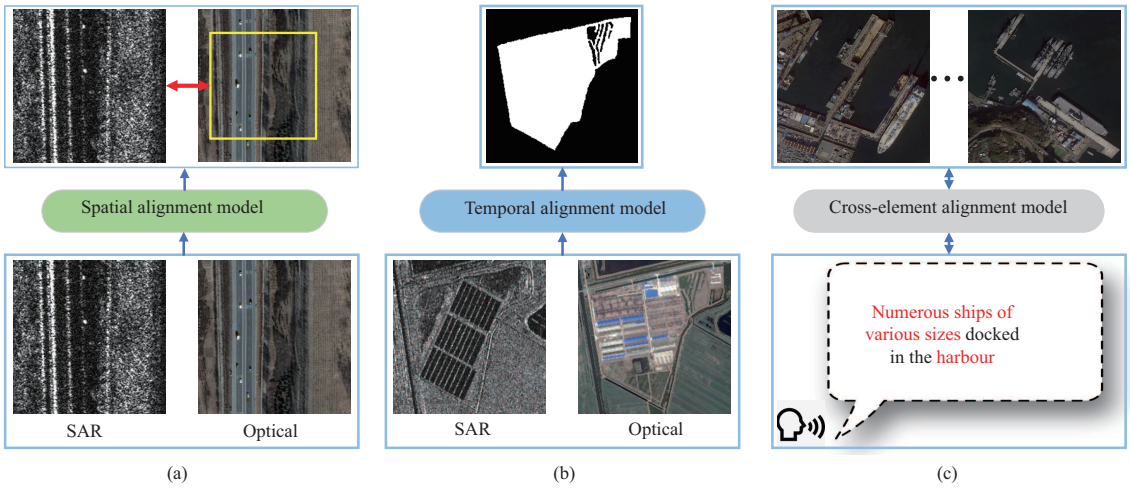

Figure 4 (Color online) The structure of the three frameworks for multimodal alignment. (a) Spatial alignment mainly focuses on the alignment with different modal images in geographic space. (b) Temporal alignment is oriented towards the changes of different ground elements within the same scene. (c) Cross-element alignment primarily addresses the alignment of remote sensing images with non-image modality information.

Table 2 The overview of the three frameworks for multimodal alignment. It is subdivided into three types depending on the alignment dimension of the data source: spatial alignment, temporal alignment, and cross-element alignment.

Category	Data source	Task	Ref.
Spatial alignment	The imagery of different modalities or different resolutions for a short period.	Image fusion, Image registration, Land cover classification	[19, 20] [9–11] [21–23]
Temporal alignment	Long-time series multi-sensor images of complex scenes.	Change detection, Regional planning, Crop classification	[14, 24] [25, 26] [27, 28]
Cross-element alignment	Remote sensing images and other types of modalities.	Image retrieval, Image-text matching	[18, 29, 30] [18, 29, 30]

geometric synchronization of the different source imagery. These images are usually taken by different sensors at different times and viewpoints [19, 20]. Spatial alignment is a significant task that can significantly affect the preprocesses step in MRSII, image fusion, image mosaic, and map updating.

Many types of spatial alignment techniques have been developed in the areas of remote sensing over the past few decades. The spatial alignment frameworks can be categorized into three types based on the training sample type: unsupervised methods, semi-supervised methods, and supervised methods.

Unsupervised methods without any prior training samples and requires direct modeling of the data. They were the first methods applied to multimodal alignment to cluster the same class of features by constructing a series of the paradigms and identifying their potential class rules from these paradigms.

Initially, the application direction of unsupervised methods is image registration, where the alignment of two or more images of the same scene captured by various sensors at different times, is an important prerequisite for various remote sensing applications [31, 32]. Refs. [33–35] employed a maximization of mutual information algorithm and combined it with other feature enhancement methods for accurate registration of different modal satellite data. Naturally, unsupervised spatial alignment also has a wide range of applications in land cover classification [21–23].

Another spatial alignment framework is the semi-supervised method, which utilizes a large amount of unlabeled and labeled data to perform MRSII [36, 37]. In [38], MAPPER was used to produce manifold alignment of optical data and polarimetric SAR data for semi-supervised classification of land cover and local climate. Ref. [39] obtained a linear invertible transformation from a latent space for images with multitemporal, multi-source, multisensor, and multiangular features using a semi-supervised alignment method. Hong et al. [40] proposed a learnable manifold alignment framework to learn a joint graph structure directly from the data. The use of semi-supervised learning for the alignment of multimodal images can reduce the work of annotators and can bring a relatively high accuracy. Therefore, it has received a lot of attention from the remote sensing community.

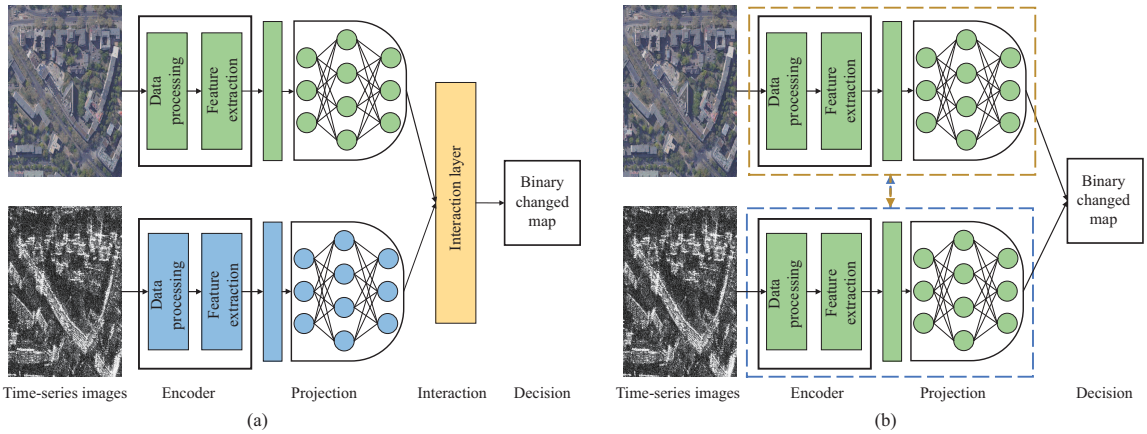

Figure 5 (Color online) There are two main structures for temporal alignment: (a) an asymmetric structure, where the encoder and projection layers differ on each side and complete the feature interaction at the interaction layer; (b) a symmetric structure, where the sub-network structure is the same architecture for different data sources and there is interaction between the individual modal features.

Supervised methods train an optimal model based on the relationship between input and output results from a labeled dataset. In supervised learning, the training data has both features and labels, and through training, the machine can find the connection between features and labels by itself.

Due to the amount of data, some methods [41–43] use a non-deep learning architecture. And as deep learning has evolved, supervised methods have become the mainstream for spatial alignment. Ref. [44–46] designed a generative network to generate the coupled optical and SAR images and used a deep matching network to match them. Zhang et al. [47] and Fan et al. [48] proposed a siamese network for multimodal image registration, which adopts the strategy of maximizing the feature distance between positive and hard negative samples. Mao et al. [49] employed deep forest to train multiple matching models and proposed a multi-scale fusion strategy for few-shot image alignment.

3.2 Temporal alignment

Temporal alignment focuses on long-time series remote sensing imagery analysis. Compared with spatial alignment, it is responsible for finding correspondence between sub-branches or elements of different modal information from the same sub-instance. As shown in Figure 4, given series imagery from different sensors, temporal alignment is oriented towards instances in the scene that changes over time and can be further employed for downstream tasks such as regional planning, crop, and plant classification. Therefore temporal alignment challenges the sensitivity of the algorithm to temporal correlation and spatial variation.

The current temporal alignment is mainly oriented to the alignment of elements between a pair of multimodal images. We divided the methods according to the symmetry of the network structure: symmetric structure, and asymmetric structure. As shown in Figure 5 in the symmetric structure, the sub-network structure is the same architecture for different data sources, and there is an interaction between the various modal features. In the asymmetric structure, the network structure is asymmetric, with differences in the encoder and projection layers on each side.

Symmetric structures prefer to learn and match the properties between different modal information through network architecture. In [50], a graph-based data fusion algorithm is proposed for data-driven semi-unsupervised change detection and biomass estimation in rice crops. Sun et al. [24–26] constructed a robust K-nearest neighbor graph to learn the structure of each image and used graph mapping to compare the graphs within the same image domain. Yang et al. [51] proposed a deep pyramid feature learning network for heterogeneous image change detection.

Asymmetric structures emphasize more on the features between different modal information, and different encoders are used to learn various modal features. Then the decoder is used to fuse and decode them [52–54].

3.3 Cross-element alignment

As artificial intelligence becomes increasingly sophisticated, bringing more new opportunities and challenges to the field of remote sensing. MRSII incorporates more and more fresh elements such as speech, text, OSM, and other non-remote sensing modalities. Cross-element alignment aims to be a global or sub-component alignment between remote sensing images and non-remote sensing modalities. It can be further used for image retrieval and visual question-answering tasks by aligning the modalities.

We divide them into two categories based on the purpose of cross-element alignment: scene enhancement and human-computer interaction. The first category is fused and aligned with non-observed modalities to reduce observation errors and obtain more comprehensive and accurate surface data. The second category is to better facilitate the staff query and search by aligning other modalities with the image for collaborative retrieval and improving image retrieval speed.

There are many non-observed features that can provide an enhanced representation of remote-sensing scenes. They have the same essential purpose as spatial alignment, matching and aligning identical regions in the modalities for downstream tasks. Refs. [55–58] aligned Openstreetmap to remote sensing imagery for building footprint delineation, updating, and urban land use mapping by finding the best match between entities. And Refs. [59–62] combined ground-based data for biomass, vegetation cover, and global ionospheric map estimation, significantly improving the accuracy and confidence of single-modal estimates. In addition, many researchers have aligned GNSS [63–65], GIS [66–68], hydrometeorology [69–71], and other information to achieve applications in traffic statistics, mapping, animal behavior, environmental interactions, etc.

For better human-computer interaction, researchers align speech and text modalities with remote-sensing images. Refs. [17, 72–74] discussed the problem of multi-label cross-modal information retrieval between image and speech-based label annotations in remote sensing through a deep neural network architecture that learns a discriminative shared feature space of input modalities, suitable for semantically consistent information retrieval. Refs. [18, 29, 30] designed a series of image-text matching networks to explore the relevance between remote sensing images and their respective natural language descriptions.

3.4 Related work and challenges

Generalized modal alignment is more focused on cross-element alignment, retrieving and matching sub-component of instances relationships between two or more modalities [8], such as image+text [75–77], video+audio [78–80], and video+text [81–83],

Modal alignment in MRSII focuses more on the alignment of various sensors, in addition to the alignment of different elements. In multi-source alignment, there are still the following challenges. (1) The image scale is too large and contains a much larger number of sub-component of instances than a natural scene. (2) The amount of data in the relevant dataset is too small, which makes it difficult to perform image alignment retrieval by a supervised model and tends to have overfitting problems in the training process. (3) Sub-components of instances are complex and have arbitrary shapes and orientations, even instances of the same area can have distortions or missing due to imaging.

4 Multi-source fusion

Limited by the imaging mechanism of the sensor, the spatial and spectral resolution of remote sensing images are mutually constrained, and a single imaging means cannot obtain remote sensing images with high spatial and spectral resolution. Multi-source fusion is an effective approach to address the limitations of sensor bottlenecks on remote sensing data metrics. It combines data from different indicators or sources through algorithms to obtain richer information than a single data source.

As a big part of MRSII, multi-source data fusion has a long history. The concept of multi-source data fusion was developed in the early 1970s, but the theoretical approach only started in the 1990s. It has made rapid progress and is still a hot research topic. There is a very broad range of applications for multi-source fusion, including natural resource surveys [84–86], precision agriculture [87–89], and urban planning [90–92]. In this section, we provide a detailed review of multi-source fusion by combining levels and categories. Then they are classified according to the type of fusion.

We subdivide multi-source data fusion into three levels by reference to Hall et al. [93]: (1) data-level fusion, (2) feature-level fusion, and (3) decision-level fusion. An overview of three architectures is

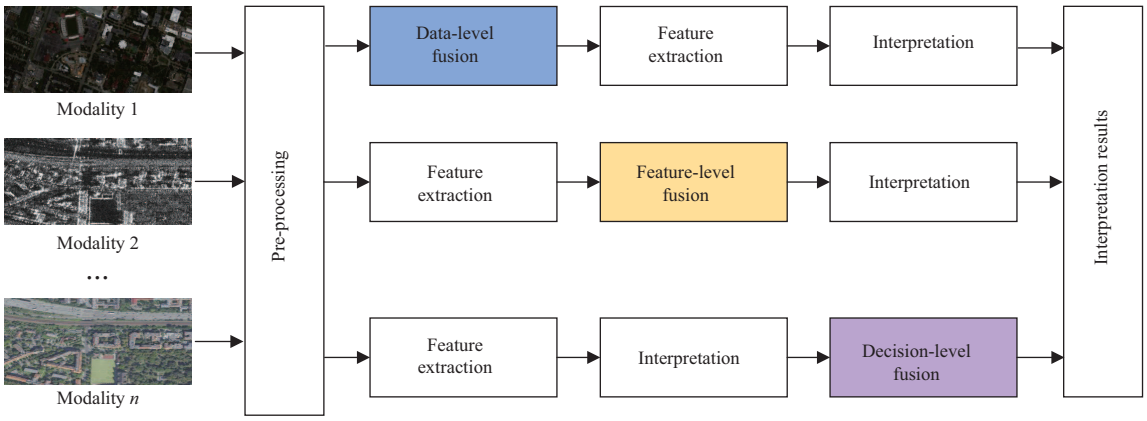

Figure 6 (Color online) The three main levels of multi-source fusion. Data-level fusion is the direct computational processing of multimodal data. Feature-level fusion occurs after the extraction of feature information from pre-processed data. Decision-level fusion needs to filter and classify features, and then perform feature fusion according to categories.

illustrated in Figure 6. Data-level fusion is the direct computational processing of raw sensor data or pre-processed data, which can contain the most original details of the sources. The primary purpose is to improve the quality of the data, i.e., the resolution, contrast, integrity, and other indicators. Feature-level fusion occurs after the extraction of feature information from the target scene (raw sensor data). It fuses the extracted features to generate new features for the subsequent interpretation of complex scenes. Decision-level fusion requires the extraction of the target features from the source image and the filtering and classification of the features, and finally, the fusion of the features according to their category. It mainly addresses the inconsistency of decision results from different data, thus obtaining more reliable decision knowledge from various sensor data. These three fusion strategies are not incompatible but can be used jointly, and the fusion of multiple levels is a cutting-edge research direction.

Based on the types of fused data, we introduce a more straightforward and more explicit classification strategy as shown in Table 3 [94–103]. We classify remote sensing multi-source fusion into three categories: homogeneous data fusion, heterogeneous data fusion, and remote sensing and other types of data fusion.

4.1 Homogeneous data fusion

Homogeneous data fusion refers to data fusion between sensors from the same imaging modality, such as data-level fusion between high-resolution panchromatic images and multispectral images. The primary purpose of this type of method is to improve the resolution of the imagery and to alleviate the interconnection between spatial, spectral, and temporal resolution. Simultaneously, by data-level fusion, noise such as shadows and clouds in the imagery are repaired and filtered to obtain the optimal temporal, spatial and spectral resolution. In addition to panchromatic-multispectral fusion [104–106], this also includes homo-modal fusion [107–109], panchromatic-hyperspectral fusion [110–112], and multispectral-hyperspectral fusion [113–115].

Homogeneous data fusion is a long history issue, and we divide it into two directions: spatial reference and spatio-temporal reference, and the schematic of the fusion methods is shown in Figure 7. The spatial reference approach focuses on the spatially coherent image pair by aligning the images spatially and establishing feature relationships to achieve data fusion. The spatio-temporal reference of fusion is more focused on inferring the high-resolution data of a particular time from a single high-resolution data with multi-temporal low-resolution data. It uses a series of temporal images to construct temporal and spatial dimensions relations and an optimization constraint algorithm to achieve fusion.

4.1.1 *Spatial reference*

We identify three types of such algorithms for spatial reference: panchromatic sharpening, linear optimization, and deep learning-based. Panchromatic sharpening is a radiometric transformation that can obtain images with the high spatial and spectral resolution by fusing Panoramic and multispectral imagery. Linear optimization mainly focuses on homo-modal fusion by adding linear constraints to obtain the reconstructed image with the optimal solution. Deep learning-based algorithms conduct homogeneous data fusion by modeling the non-linear relationships between images by simulating the structure

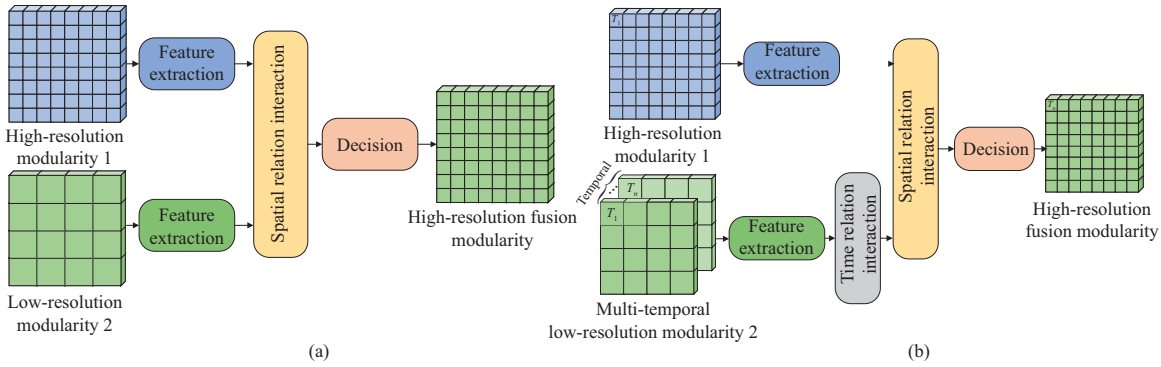

Figure 7 (Color online) The two main types of homogeneous data fusion: (a) spatial reference and (b) spatio-temporal reference. The spatial reference approach focuses on the spatially coherent image pair, and the spatial-temporal reference of fusion is more focused on inferring the high-resolution data of a particular time.

Table 3 The overview of multi-source fusion. Multi-source fusion is further subdivided according to the type of fused data: homogeneous data fusion, heterogeneous data fusion, and remote sensing and other types of data fusion.

Category	Data source	Description	Ref.
Homogeneous data fusion	The same imaging modality	Improve the resolution of the imagery and alleviate the interconnection between spatial, spectral, and temporal resolution.	[94–96]
Heterogeneous data fusion	Different imaging modalities	Leverage the advantages of different sensors to provide rich scene information while reducing noise interference.	[97–99]
Remote sensing and other types of data fusion	Panoramic acquisition data, landscape images, land, atmospheric, hydrological data	Provide comprehensive observations for data sensing of large-scale complex scenes.	[100–103]

of biological neurons.

Popular panchromatic sharpening methods can be divided into two main types: component substitution and multi-resolution analysis. Component substitution models project the image into a newly transformed space, substituting the components containing spatial information with high spatial resolution images and inverting them to obtain the spatially enhanced data. Multi-resolution analysis models decompose all the raw data into different resolution imageries and fuse them, finally inverting them to obtain the fused image.

The pioneers for component substitution (CS) are intensity-hue-saturation (IHS) transform [116–119] and the principal component analysis (PCA) [120–123]. The IHS transform is widely used to fuse images based on its ability to separate the spectral information in the H and S components of the RGB composition while separating out most of the spatial information in the I component. PCA mainly projects the data into a newly transformed space through a linear transformation, the first principal component in the direction of maximum variance, which retains most of the original data information, so the replaced component is the first principal component. An alternative to these methods is the Gram Schmidt method (GS) [124–126]. The essence of this method is the Gram-Schmidt orthogonal method, which transforms a set of linearly independent vectors in unitary space into a set of orthogonal vectors. In addition, there are the brovey transform (BT) [127, 128], tensor factorization [129, 130] etc.

Due to its ability to effectively improve spatial resolution at low computational cost, CS methods are still investigated by the research. Methods based on GS adaptive (GSA), generalized IHS (GIHS) [131], GIHS adaptive (GIHSA) [124], and ratio image-based spectral resampling (RIBSR) [132] have been extensively investigated, addressing the extent of the spectral distortion during multispectral fusion.

Multiresolution analysis (MRA) decomposes the multimodal data into components, which produce a higher resolution image exactly when added back together. The decomposition and fusion of the image to obtain higher resolution images are the core of MRA. Each component ideally decomposes the image into physically meaningful and interpretable parts.

The commonly used MRA methods include high-pass filtering (HPF) [120, 133], wavelets [94, 95, 134], Laplacian pyramids [96, 135], and Curvelet transform [136–138]. MRA methods can maintain spectral information better than CS methods. However, suppose multimodal data are not rigorously aligned. In

that case, spatial distortion may occur in fusion products in the presence of high-pass detail injection, often caused by ringing or aliasing effects, originating shifts, or blur of contours and textures [124].

Linear optimization models format the data fusion problem as its linear optimal solution, assuming the relation between the multi-source data $X_1, \ldots, X_N$ and fused data Z is linear and can be expressed as follows:

$$Z = W_1 X_1 + W_2 X_2 + \cdots + W_N X_N + b, \tag{1}$$

where W_n is the conversion factor and b is bias.

Depending on the principle of the solution, it can be divided into spectral demixing, Bayesian probabilistic, and sparse demixing methods. Spectral demixing decomposes a mixed pixel into a series of constituent spectra (endmembers) and a set of corresponding fractions (abundances) and conducts a linear summation and reconstruction. Bayesian probability theory treats the data to be fused as observed values and the fused data as unobserved true values. It solves the parameter values in the fusion process by calculating the probability of the true values occurring given the observed values and maximizing the probability. Sparse demixing methods decompose the multimodal data into a dictionary matrix and a sparse coefficient matrix and add sparse constraints to solve for the sparse coefficients to obtain the fused data.

Deep learning-based algorithms focus on constructing non-linear relations between images in the same region. The most common approach is based on convolutional neural networks (CNN), which solve the problem of an extremely large number of weights by abandoning global connectivity [139–141]. There are two representative studies. Scarpa et al. [142] designed a light-weight CNN and a target-adaptive usage modality to ensure a great performance also in the case of mismatched data sources. Ref. [115] proposed a 3D-CNN to fuse MS and HS images to obtain a high-resolution hyperspectral image.

4.1.2 *Spatio-temporal reference*

The approach of spatio-temporal referencing is essentially the same as that of spatial referencing in the construction of spatial relation. Therefore, we focus on the research in the construction of temporal relations.

Early studies focused on linear optimization models. The spatial and temporal adaptive reflectivity fusion model (STARFM) proposed by Gao et al. [143] is an effective method for predicting spatio-temporal fusion. And based on STARFM, a series of spatio-temporal reference algorithms have been proposed, e.g., ESTARFM [144], STRUM [145], and USTARFM [146]. Xue et al. [147] proposed a Bayesian statistical algorithm that combines temporal correlation information in the time series, and they treated the fusion problem as an estimation problem with a maximum a posteriori (MAP) estimator to obtain the fused image. For deep learning-based algorithms, the work tends mainly to look for non-linear relations in space, while relatively few temporal relations are constructed.

4.2 Heterogeneous data fusion

Heterogeneous remote sensing data fusion means the fusion between sensors from different imaging modalities, for example, optical-radar, SAR-multispectral, and SAR-hyperspectral data fusion. Due to the excessive differences in imaging mechanisms among different sensors, heterogeneous data fusion is more suitable for feature-level, decision-level fusion, such as feature classification, change detection, and parametric inversion.

Based on the optimization methods, we divide it into feature stacking-based methods, subspace-based methods, and deep learning-based methods. In feature stacking-based methods, we stack the information extracted by the auxiliary sensor onto each pixel of the image to obtain a feature vector containing all modal information. Subspace-based methods project all information into a low-dimensional subspace and then perform feature fusion. And deep learning-based methods learn the non-linear relation between the input and output of the system. These methods can better portray the non-linear relationship between different resolution images and are highly transferable.

Feature stacking-based is the cleanest implementation of heterogeneous fusion. This strategy filters and stacks various source data in the same structure. For example, stacking of height and intensity features extracted from the LiDAR data into the spectral bands of the multi/hyperspectral image and forming an extended feature vector for each pixel in the complex scene [148].

Morphological profiles, attribute profiles, and extinction profiles are widely used for feature extraction and filtering to exploit the discriminatory feature information in the heterogeneous data fully. These methods [149–151] are conceptually simple and computationally efficient, are often used for heterogeneous data fusion, and provide high-quality fusion results.

While feature stacking-based methods can achieve superior fusion results, the direct stacking of spectral, spatial, and elevation features extracted from heterogeneous data increases the feature dimensionality of the samples, thus raising two major challenges for subsequent classification tasks: dimensional catastrophe and high computational complexity.

Subspace-based methods avoid dimensional catastrophes in subsequent classifications and improve computational efficiency. They represent the features from the heterogeneous data features in a low-dimensional subspace to ease the strain on subsequent tasks. In the original subspace model, both the basis of the subspace and the fused features are unknown, so how to estimate them is the core problem of the subspace model.

A lot of early work on subspace-based methods used classical IHS transform-based methods [152] or PCA approaches [153, 154]. These methods can effectively reduce feature dimensionality, improve the signal-to-noise ratio, reduce computational effort, and improve the classification accuracy for heterogeneous data fusion issues.

Deep learning-based. Remote sensing scenes with the complex distribution of multiple categories lead to a non-linear relation between remote sensing data and object samples. Multi-sensor data fusion enhances this non-linear relation, causing the samples to exhibit higher-order nonlinearity in the feature space. Deep learning-based methods can fit non-linear relations between heterogeneous data well, and they have the ability to extract high-order, multi-dimensional, abstract features from data. Features extracted by deep learning are generally invariant to the non-linear distribution of samples in the original space and are robust to complex scenes.

Deep learning-based methods [97–99] can achieve better fusion results and classification accuracy. However, they usually require a large number of labeled samples for training, and the labeled samples of remote sensing scenes are usually difficult to acquire, which to a certain extent limits the application of deep learning methods on heterogeneous data fusion.

4.3 Remote sensing and other types of data fusion

Remote sensing data can also be fused with other types of data for processing and synergistic applications, thus obtaining more data on the characteristics of the resource environment. The fusion of remote sensing data with panoramic acquisition data, landscape images, and land, atmospheric, hydrological data provides more accurate observations for data sensing of large-scale complex scenes. Remote sensing data can provide more accurate initial observations and boundary conditions for scenes, then continuously and automatically adjust them with other data, thus reducing simulation errors to high accuracy, spatially continuous surface data. This fusion approach is an important trend in current development.

Remote sensing and ground-based observation are two important ways of obtaining Earth observation data. Remote sensing can provide large-scale area observations, but the precision of the observations is often difficult to guarantee due to its complex imaging process and susceptibility to environmental interference. Ground-based observations are of high quality, but the sparseness of the observation sites makes it hard to obtain comprehensive observations. Therefore, the fusion of remote sensing and ground-based observation data has attracted the attention of many researchers [100–103]. At the same time, the fusion of remote sensing data with atmospheric and hydrodynamic can further reduce simulation errors and be used for the synergistic analysis of hydrometeorological [155, 156], vegetation [157, 158], and atmospheric information [159, 160].

4.4 Remaining problems

In multimodal fusion, homogeneous data fusion technology has been very mature and has wide applications in life. The optical images we obtain on the Internet are homogenized fused data. And in heterogeneous fusion and other types of data fusion, while fused images have a great improvement over single-modal remote sensing images in terms of spatial dimension and visualization, due to the poor adaptability of models to different sensors and scenes, it is still necessary to test and select different models according to the scene environment in the application. To this end, we focus on the problems

that exist in heterogeneous data fusion, other types of data fusion and hope to bring some insights to the researchers, as follows:

(1) Different sources of data. There is a large inter-class difference in the modal information carried by different sensors. The fusion process requires geospatial alignment and standardized processing between information and the redundancy of multimodal data, which removes redundant information while retaining valid information.

(2) Different observation angles. The angle of observation is different for various modalities. Satellites have different angles of view of the same area, and even with ortho-correction, the two images cannot be identical. Also, other types of data, such as ground-based data, are difficult to align data features when fusing with remote sensing images due to the observation platform on the ground.

(3) Multi-resolution. Multimodal fusion raises the issue of multi-resolution, especially in the fusion of heterogeneous data, which is accentuated by the large differences in data types. The scale of resolution affects the performance of the model, with the model preferring to observe small targets (cars, trees, cottages) when the resolution is too high, and better for large targets (stadiums, roads, tall buildings) when the resolution is low.

(4) Unknown observation scenes. Remote sensing scenes are unpredictable, and current models are often only applicable to a single scene, e.g., urban, forest, desert, and ocean. We often have an unknown situation of the observed scenes, so it is one of the future directions to improve the robustness of the model to apply to unknown observed scenes.

5 Multimodal representation

Using machine learning methods to convert raw data into a mathematical representation that computers can recognize and process for further extraction of useful information for classification or other prediction tasks is a major field of research in representation. In multimodal representation for remote sensing, multimodal data describe a complex scene from different perspectives, and the context information is complementary or supplementary. Thus they carry more excellent information than single-modal data, and it is valuable to exploit the comprehensive semantics provided by several modalities from heterogeneous sources.

The performance of machine learning methods heavily depends on the effective representation of the representation features of the applied data [161]. The representation of single-modal features is relatively advanced for visual [6, 162–164], text [165–167], speech [168, 169], and graph [170–172] modalities and is widely used in realistic applications. However, there are still many difficulties in multimodal feature representation, especially in multimodal representation learning of remote sensing scenes: (1) how to suppress uncontrollable noise from different sensors, (2) how to combine small samples of data from heterogeneous sources, (3) how to handle imaging perspectives between different data sources, (4) how to solve missing data in some modalities.

To facilitate the discussion on how to represent data from different modalities clearly and efficiently, inspired by definition in [8, 173], we classify remote sensing multimodal representation into three frameworks: (1) joint representation, (2) coordinated representation, and (3) encoder-decoder representation. An overview of three architectures is illustrated in Figure 6.

The most common representation learning for remote sensing images is based on CNN. This work tends to transfer learning by taking a deep CNN model that has been pre-trained on a natural scene such as LeNet [174], VGGNet [6], GoogleNet [175], and ResNet [5], thus obtaining better performance by training from scratch. And with the rise of the transformer [176], it is increasingly being used in image representation learning [163, 177, 178] and is a new research hotspot in remote sensing imagery interpretation [179–181]. They can be integrated into multimodal representation learning and trained together with other modal data (e.g., word2vec [165], Glove [166], and BERT [167] for text data and wav2vec [168], PASE [182], and Mockingjay [183]). Through training with representation learning models from other modalities, the performance of multimodal representation learning can be greatly improved.

5.1 Joint representation

Joint representation aims to project various single-modal features into a shared semantic subspace to reduce heterogeneity between modalities and exploit the complementarity between features, thus learning a better representation of the features.

Relevant algorithms represent imagery from various sensors as well as other modal information as feature vectors (tensors), and narrow the heterogeneity gap to obtain complementary feature representations. Sharma et al. [184] and Yang et al. [185] improved the object detection accuracy in remote sensing and unmanned aerial vehicles imagery under various weather conditions by expanding on the capabilities of RGB by learning the characteristic of the infrared sensor. Flynn et al. [186] and Oliveira et al. [187] used aerial video for person detection, by the joint representation of optical images with infrared or thermal camera images and tracking detections over time, and higher detection precision is obtained. Breckon et al. [188] introduced a real-time multimodal object detection algorithm that uses a combination of visible-band, thermal-band, and radar sensing from a deployed network of multiple autonomous platforms (ground and aerial) to detect people and vehicles automatically.

In addition to object detection tasks, multimodal joint representation has a wide range of research applications in other remote sensing imagery interpretation tasks. For the sensing classification task, Audebert et al. [189, 190] investigated early and late joint representation of Lidar and multispectral data, finding the early fusion allows for better joint-feature learning but at the cost of higher sensitivity to missing source and the late fusion makes it possible to recover errors steaming from the ambiguous source. Li et al. [191] proposed a multimodal bilinear fusion network to extract deep semantic feature maps of optical and SAR images, and bilinear integrate the joint representation. Poliyapram et al. [2] proposed a novel deep learning-based end-to-end point-wise LiDAR and optical image multimodal fusion network for 3D segmentation of aerial point clouds by integrating aerial image features. Jeong et al. [192] proposed a multimodal sensor-based semantic 3D mapping system using a 3D Lidar combined with an optic camera.

Joint-feature learning of multi-resolution homogenous data is also an important research direction in joint representation. In various resolution images, the same object has different scales and perceptual fields, and chromatic aberrations exist in the same object due to differences in the imaging methods of different sensors, thus challenging the adaptability and robustness of the model even more. Multi-resolution joint learning has a wide range of applications and research value for tasks such as crop classification [193, 194], object recognition [195, 196], land cover classification [197, 198], and hydrological simulations [199].

5.2 Coordinated representation

Another type of multimodal representation is coordinated representation. In the coordinated representation framework, each modality learns its individual representation alone and then coordinates them through a unified constraint. This type of algorithm emphasizes more on the similarity and complementarity of the elements across modalities. It attempts to learn separate but constrained representations for each modality in a coordinated subspace.

We divide the methods into two categories based on the purpose of coordinated representation: complementary methods and similarity methods. The complementary methods mainly focus on the difference and complementary information between modalities, complementing and enhancing the representation of complex scene information by comparing the difference information. Similarity methods are more concerned with the similarity between different modalities, which expects the distance to modalities related to the same semantics to be as minimum as possible and the distance to different semantics to be as maximum as possible.

Complementary methods enable the coordination space to discover inter-modal variability to complement the fused representation. For example, extracting multispectral images and Lidar features, concatenating and interacting in higher dimensions to obtain complementary fused features for land cover classification [98]. Ref. [1] combined overhead images from Google Maps and ground-based images (side-views) of each urban object from Google Street View to obtain complementary visual information related to urban objects to enhance the understanding of urban land use.

In addition to learning complementarity, coordinated learning of the similarity of the same elements of each modality in the subspace by similarity methods is also an important branch of coordinated representation. Ye et al. [41, 200] performed image registration by conducting a similarity measure based on feature representation of global and local features of SAR images with Lidar data. Uss et al. [201] trained a two-channel patch matching CNN to detect similarities between image patches and measure their mutual displacement. The model has both a higher discriminative power and a more precise localization by testing on real RS images. A deep learning-based matching method by Zhu et al. [202] is to compare

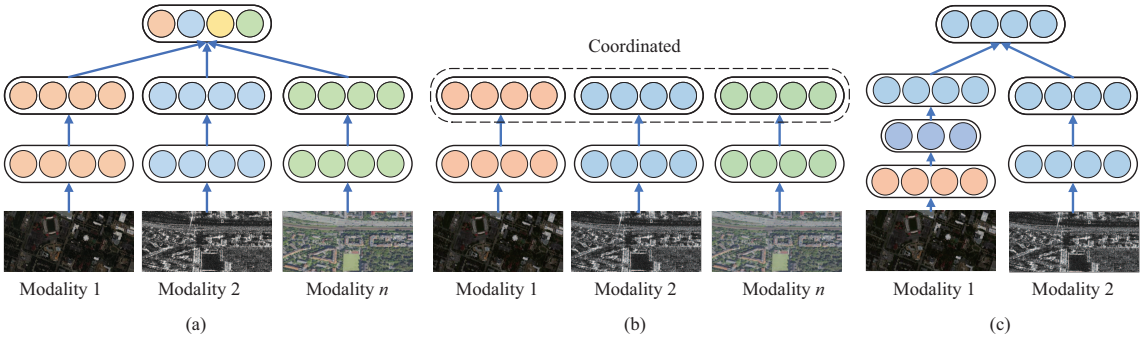

Figure 8 (Color online) Structures of three types frameworks about multimodal representation. (a) Joint representation aims to project various single-modal features into a shared semantic subspace. (b) Coordinated representation learns the representations of each modality individually and coordinates them by uniform constraints. (c) Encoder-decoder representation translates one modality into another and projects them into the same subspace to maintain semantic consistency.

optic and infrared pairs and search for points corresponding to a given point in the reference image in the search window of the target image.

5.3 Encoder-decoder representation

Encoder-decoder representation utilizes the concept of translation. It first converts information from one modality into a feature representation of another modality through an encoder-decoder architecture. It then projects them into the same vector subspace to maintain their semantic consistency. For example, given an optical image, we aim to generate the corresponding SAR features or given the SAR image to generate the corresponding optic features.

This method is mainly used when the data of a modality are relatively complex, and noisy, and the amount of data is small or missing. The majority of multispectral images are affected by clouds, and the generation of corresponding optical features by SAR imagery to recover the affected regions is a current hot topic in multimodal representation learning [203–205]. A generative adversarial network had to be built to combine the SAR images with the simulated optical images in order to reconstruct the corrupted area in [211]. Dai et al. [206] investigated the multi-temporal images to achieve self-training and gated convolutional layers to distinguish cloudy pixels from clean pixels, compensating for the lack of differentiation ability of ordinary convolutional layers.

In addition to interference removal, encoder-decoder representation can be applied to land cover classification, modal transformation, target detection, etc. Hong et al. [207] further improved the performance of land cover classification by using a self-GANs module and mutual GANs module for learning feature representations insensitive to perturbations and eliminating gaps between multimodalities, respectively, to produce more efficient and robust information transfer. And a modal transformation model is proposed by Liu et al. [208] to transform the information of sparse modalities into the feature space of rich modalities, providing a solid foundation for the multi-temporal imagery interpretation task.

5.4 Discussion

Multimodal representation learning has been a widely researched topic, providing a unified feature representation space for other applications, such as modal alignment, and modal transformation.

In this subsection, we classify it into joint representation, coordinated representation, and encoder-decoder representation. From the structures (Figure 8), we can know that joint representation is more suitable for situations where the data from different modalities is balanced and modal interactions are required to predict together during the inference process. In the coordinated representation, the modalities are independent but coordinated with each other, and are more inclined to the situation where there is missing data or single-modal input in the evaluation process. Encoder-decoder representation focuses more on tasks with imbalanced samples or that require additional modality-assisted learning.

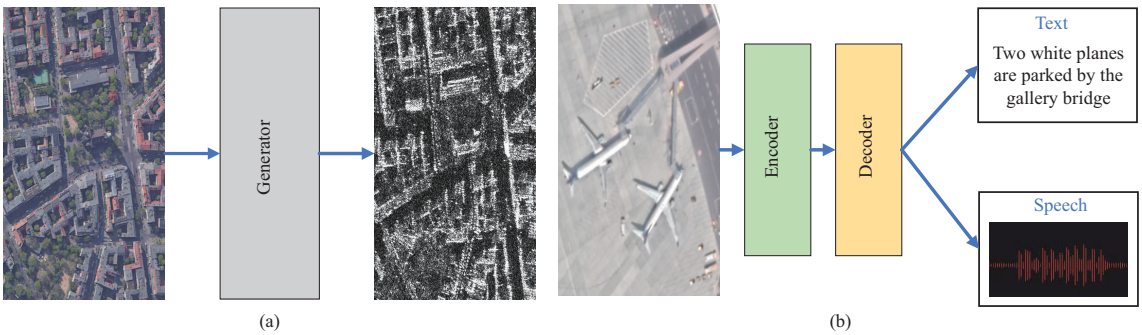

Figure 9 (Color online) Examples of cross-modal translation. (a) Cross-sensor translation is transitions between different image modalities. (b) Cross-element translation is transitions between image and other modalities.

6 Cross-modal translation

Translating information from one modality to another is an excellent challenge in MRSII. Due to the complexity of remote sensing scenes and the variability of sensors, cross-modal translation in remote sensing is more challenging than the intra-class (between different image modalities) and inter-class (between image and other modalities) translation of natural scenes. Cross-modal translation in remote sensing is an emerging topic in remote sensing. With the development of deep learning algorithms and computer hardware, advances have been made in sence image translation [209–212], remote sensing image caption [213–215], etc.

While cross-modal translation is an emerging task in remote sensing, an extensive range of algorithms have been applied to such tasks due to its wide application. Simultaneously, we split the cross-modal translation into two branches: cross-sensor translation and cross-element translation, based on the intra-class and inter-class relationships of the modalities, as shown in Figure 9. Cross-sensor translation focuses on the translation of images between different sensors, e.g., panchromatic and multispectral. Cross-element translation is responsible for the translation between images and other types of modalities, e.g., remote sensing images and text, hydrological data, etc.

6.1 Cross-sensor translation

Remote sensing data has played a more critical role in Earth observation and urban planning in recent years. With abundant data acquired, great challenges still remain due to the following three reasons.

(1) A large part of the data is disturbed by atmospheric factors such as clouds and fog, and such uncontrollable factors significantly reduce the utilization of remote sensing images and increase the difficulty of processing and training. According to [216] for Landsat ETM+ data, there is about 35% of the land area covered by clouds, while the condition is even worse in sea areas.

(2) Some modal images have a relatively small amount of data because of sensor imaging and revisit time. It severely limits the application of deep learning-based algorithms in this area of research.

(3) As a consequence of the trajectory of the sensor carrier, remote sensing data may be missing in some areas or scenes, especially for a specific time phase (season). In remote sensing, for this reason, the difficulty lies in imaging a particular location in a specific period.

Early studies focused on super-resolution reconstruction (SRR), which acquires super-resolution (SR) images from low-resolution (LR) images. And the popular SRR methods are mainly based on traditional algorithms and learning-based algorithms. We divide standard algorithms into interpolation-based and sparse-based representation methods. While interpolation-based methods such as bilinear or bicubic tend to generate overly smooth images with artificial traces, they are still widely used due to their simple implementation. And Refs. [217–219] improved the performance of the model by incorporating a series of optimization strategies with prior knowledge. Sparse-based methods enhance the ability of linear models to recover high-frequency information from prior knowledge, such as wavelet transforms [220], coupled sparse autoencoders [221], and external dictionaries [222], but these methods are computationally complex and require significant computational resources. The learning-based model attempts to capture the co-occurrence prior between image patches. Deep learning is a fundamental learning-based approach. It learns and fits the mapping relations between LR and HR (high resolution) images by building an end-to-end neural network, such as CNN [223–225], GAN [226–228], and attention network [229–231]. Due

to its nonlinear property, it can recover high-frequency information without substantial computational resources. Therefore, SRR based on deep learning has become a research hotspot.

Remote sensing images are affected by various noises during the acquisition process, which blurs the edge details and degrades the quality of the images. Therefore, they need to be denoised to obtain clear, high-quality images. Multimodal-information-based method for remote-sensing image denoising is a novel hot topic by incorporating noise-free reference images of different modalities as prior knowledge into the denoising object function [232, 233]. In addition, researchers have also conducted a series of investigations on the denoising of natural climates. In this regard, cloud removal has become increasingly sophisticated. The existence of clouds is one of the main factors for the missing information in optical images, and how to generate the missing information by other modal images is a significant concern. Huang et al. [203] proposed a removal method based on sparse representation to recover the missing HR information. With the development of GAN networks, more and more researchers adopted GAN for cloud removal and achieved significant improvement, and the reconstructed images are more natural and realistic [204, 205, 211].

Naturally, cross-sensor translation has a wide range of applications in solving rare data. There are two main directions: cross-sensor and cross-regional. Cross-sensor means generating data that is not easily accessible through resource-rich data sources. Refs. [234–236] translated SAR-to-optical for all-weather observation, while simplifying the observation conditions of SAR images. And cross-regional refers to the generation of images from one style of the region to another style of the region to achieve data enhancement. Ji et al. [237] proposed a GAN-based domain adaptation for land cover classification. Peng et al. [238] designed a novel FDANet for building extraction.

6.2 Cross-element translation

The translation of the remote sensing image into other modal information, or using other modal information to caption the semantic information of the remote sensing image and summarize the image content plays a vital role in many fields, such as cross-modal retrieval [17, 72, 74], intelligence generation [213, 239, 240], and Scene Q&A [241–243]. Cross-element translation requires the model to fully understand the complex scene and identify the key components of the scene, and generate a standardized, concise and comprehensive modal information to represent the scene by understanding and analyzing the high-level semantic information.

With the abundance of computing resources and the increase of data, more researchers are turning to image caption in remote sensing. Refs. [244–246] designed a series of attention mechanism-based image captioning methods. Huang et al. [247] considered the problem of missing or omitted features due to large-scale scenes from the perspective of multi-scale feature fusion. A novel word-sentence framework is proposed by Wang et al. [248] to extract the valuable words from the image and generate the well-formed sentence.

6.3 The challenges and differences from nature scenes

The challenges of multimodal translation are mainly in two aspects: the complexity of remote sensing data and the metrics of evaluation.

Remote sensing images often exhibit large scale, high density, and large aspect ratios. In the process of cross-modal translation, there is often information loss. Therefore, it contains several times or even tens of times more information than natural scene images. Especially in cross-element translation, it is difficult for the model to ensure that all the information in a scene is described. In addition, the main challenge of the problem is how to filter and determine the key information in the image and extract it precisely, as well as describe it clearly and in detail.

Another challenge is the evaluation of model performance. Since modal translation is a generation problem, it is difficult to automatically evaluate the generation quality of the model through the evaluation metrics, and even utilizing human judgment will result in subjectivity to some extent. Meanwhile, different from natural scenes, performing cross-sensor translation, such as translation from optical images to SAR images or Lidar data, requires professionals to evaluate, which further increases the difficulty of evaluation.

7 Co-learning

Using resource-rich data to assist resource-poor data for training has been an effective method for solving few-shot learning. In multimodal machine learning, helping one modality through a resource-rich modality to another, especially when the other modality has limited information or lacks labeled data, is noisy input, and has unreliable labels [8].

In this section, we focus on co-learning in MRSII, which includes transfer learning, co-training, and few-shot learning, using the other sensors or modalities to assist one modality for effective learning. By using co-learning to learn features from different modal information, we can achieve a more robust model as a comfortable solution to the challenges of missing samples or noise.

Transfer learning is one of the most commonly used tools in remote sensing imagery interpretation, and using pre-trained models trained in large-scale natural scenes as the backbone of the interpreted model can improve the convergence speed and performance of the model. Cross-sensor transfer learning has also been widely researched and applied. In 2010, Yao et al. [249] introduced MultiSource-TrAdaBoost and TaskTrAdaBoost to transfer knowledge from multiple sources. Liu et al. [250] proposed the multi-kernel joint domain matching, a novel domain adaptation method for unsupervised transfer learning in multimodal data. Ref. [251] applied the knowledge learned from the natural image to the DEM super-resolution problem.

Co-training and few-shot learning are also major research areas in co-learning. Hu et al. [252] designed a co-training classification approach to coping with the unclear observations. Qiu et al. [253] combined the Sentinel-2 and Landsat-8 imagery, as well as the Global Urban Footprint, the OSM, and the Night-time Light data, regarding their relevance for discriminating different LCZ classifications. For few-shot learning, Rostami et al. [3, 254] transferred knowledge from Electro-Optical domains to SAR domains to eliminate the need for huge labeled data in the SAR domains. Ying et al. [255] proposed an effective lightweight CNN that efficiently migrates prior knowledge from optical, hybrid optical, and non-optical domains to the SAR target recognition task.

In MRSII, co-learning is oriented to the case of few or missing target data, which is mainly reflected in two aspects: data source missing and regional missing. Utilizing abundant source data to assist or co-learning with target data can substantially improve the performance of the model, which is a hot research topic in remote sensing.

8 Datasets of MRSII

In this section, we discuss the relevant datasets for MRSII. We classify the different datasets into spatial, temporal, and cross-element based on the dimensionality of the data (some representative datasets summarized in Table 4 [72, 190, 239, 256–262]).

Spatial dataset. The images in spatial datasets are disconnected, and algorithms for these studies focus on basic computer vision tasks, e.g., classification [5, 162, 263], object detection [264–266], segmentation [267–269], and image retrieval [270–272]. With the development of complex networks and brain science, multimodal datasets obtain sustained attention. Researchers utilize multimodal information, e.g., Optic/Lidar [190], Optic/SAR [256], SAR/Lidar [257], to resolve more complex scene understanding, which is an extremely challenging problem for machines. However, existing multimodal datasets do not have enough well-annotated data to support most of the deep-learning-based technology. They are prone to suffer overfitting problems when the data quantity is much too scarce for the number of the model parameters.

Temporal dataset. The temporal datasets focus more on the temporal evolution in the same area, and existing studies contain the comparison of two images and focus on the particular object instance. For example, LEVIR-CD [258] and HRSCD [259] are the fundamental drivers of change detection. Lots of excellent models are implemented on these datasets, which have been applied in daily life. The emergence of CRC [260] and SITSC [261] promotes the progress of crop classification and cover. Correspondingly, Emelyanova et al. [273], Zeebruges [274], and WUDAPT [275] bring the catalyst to data fusion. These datasets establish a precedent for temporal analysis.

Cross-element dataset. Cross-element datasets are mainly interested in the interaction and transformation of RS images with other types of elements, such as image/audio [16, 72, 244], image/text [18, 214, 239], and image/atmospheric data [262, 276]. RS image caption [214, 239, 277], cross-modal re-

Table 4 Taxonomy of MRSII dataset. The different datasets are divided into spatial, temporal, and cross-element based on the dimensionality of the data

Type	Description	Task	Modality	Detail	Ref.
Spatial	It focuses on basic computer vision tasks, and the images in the spatial dataset are disconnected.	Segmentation	SAR + Optical	It covers 120 square kilometers of multiple overlapping areas and has over 48000 construction labels.	[256]
			Lidar + SAR	It includes 31 field measurement sites in six wetlands within the riparian transition zone and surface waters of the Utikuma regional study area.	[257]
			Lidar + Optical	These datasets cover Vaihingen and Potsdam, semantically labeling images on six categories.	[190]
Temporal	The temporal datasets focus on the temporal evolution in the same area.	Change detection	Optical	It consists of large-scale bitemporal optical images with 637 image pairs.	[258]
			Optical	It consists of optical images and contains several regions of France.	[259]
		Classification	Optical	It utilizes multi-temporal Landsat TM and ETM+ images to map crop residue cover.	[260]
			Optical	It consists of multi-temporal Sentinel-2 satellite images for crop classification.	[261]
Cross-element	Cross-element datasets are interested in the interaction and transformation of RS images with other type data.	Retrieval	Image + Audio	It consists of a total of 68170 audio with American pronunciation and 13634 images of remote sensing images.	[72]
		Image caption	Image + Text	It is composed of a total of 24333 sentences and 10921 images of remote sensing images.	[239]
		Segmentation	Image + Atmospheric	It utilizes Landsat-8 and atmospheric measurements for coral reef studies.	[262]

trieval [16, 18, 72, 244], and atmospheric data observation [262, 276] are all dependent on these datasets. Since RS images are large-scale and contain too much content, it is often difficult for other types of elements to describe and align the critical information of the whole scene. Therefore, the main problem of current cross-element datasets is still how to represent the information between different elements in a reasonable and detailed way in an effective and homogeneous representation.

9 Applications

9.1 Land use classification

Land use classification (Figure 10(a)) is the earliest application of MRSII. Different land covers have similar spectral features in satellite images, and their classification by a single modality will inevitably

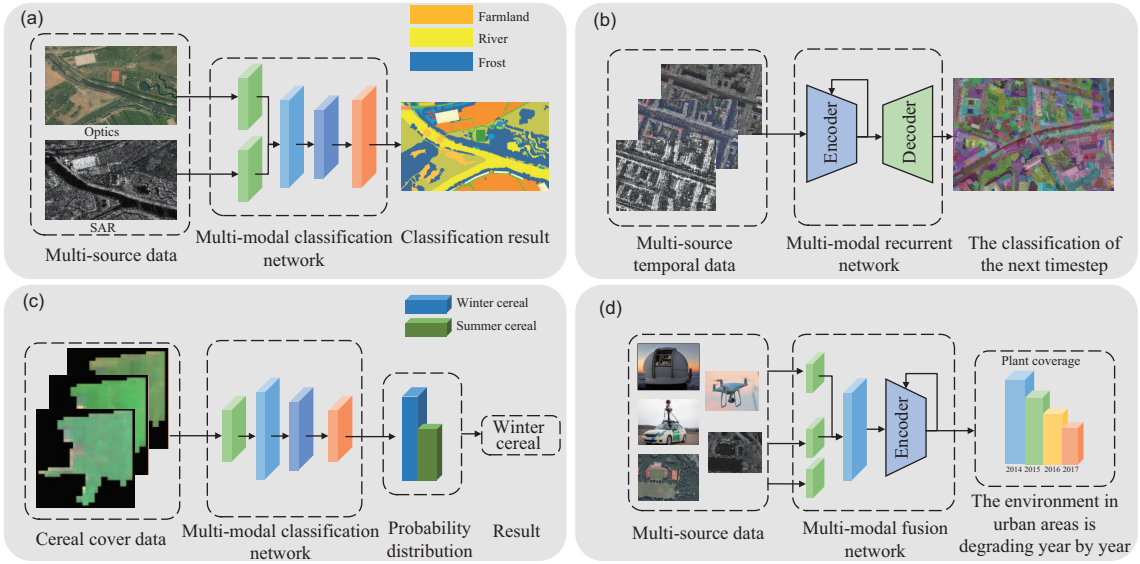

Figure 10 (Color online) Four main applications of MRSII. (a) Land use classification; (b) urban planning; (c) agriculture; and (d) ecology.

lead to some discriminative errors. By analyzing them through multimodal data, they can be enhanced in three aspects: (1) resolution; (2) spectral; (3) temporal. MRSII effectively improves the resolution of the same area and reduces the number of mixed pixels; the high spectral resolution improves the fidelity and accuracy of the spectral dimension information; the temporal information is further supplemented from the different characteristics of land cover types in temporal information.

Chen et al. [278] fused Landsat-8 data with MODIS, HJ-1A, and ASTER DEM data to improve land cover classification accuracy. A study applied Sentinel-1, Sentinel-2, and Landsat-8 data to solve the lack of spatial continuity due to cloud cover [279]. Ref. [280] further investigated the comparison of the effects of different levels (data-level, feature-level, and decision-level) of fused data.

9.2 Urban planning

Through the analysis of multimodal data, the observation interval of the same area is greatly reduced. Therefore, the multimodality of the data provides the possibility to observe the change and development of the urban from multiple perspectives, and through the observation of historical data, the development of the urban can be effectively planned and predicted (Figure 10(b)).

Ref. [281] proposed an unsupervised deep convolutional coupling network for change detection based on two heterogeneous images. A recent study explored the use of convolutional autoencoder and commonality autoencoder to eliminate most of the redundancy in two heterogeneous images (optical and SAR) to obtain a more consistent feature representation [282]. Another study designed an edge-preservation neural network (EPUNet) to automatically update the existing building databases to their current status with minimal manual intervention [54].

9.3 Agriculture and ecology

The monitoring of multimodal satellite images is of great political and economic importance in both agricultural and ecological directions (Figures 10(c) and (d)). Many crops are often similar in appearance at the same moment and need to be observed by satellite image time series to improve the classification accuracy. In ecology, multimodal imagery also has great potential for applications such as the estimation of ecological variables, ecosystem dynamics monitoring, and disturbance detection in ecosystems [283,284].

Garnot et al. [27, 28] proposed to extract temporal features using a bespoke neural architecture based on self-attention and designed a lightweight temporal self-attention for large-scale agricultural parcel classification. A study explored a DeepCropMapping through integrating multi-temporal and multispectral remote sensing data for large-scale dynamic corn and soybean mapping [285]. He et al. [286] combined remote sensing data on fine particulate matter (PM2.5) concentration, surface temperature (LST), and vegetation cover (VC) to assess urban environmental changes in China at the national scale, among urban

agglomerations, and in rapidly urbanizing areas. Hilker et al. [287] and Tran et al. [288] used STAARCH to fuse Landsat and MODIS reflectance data to map forest disturbances.

10 Future directions

As data sources increase, MRSII offers the feasibility of high resolution, hyperspectral, and long temporal observations. Meanwhile, it also brings more tasks and challenges to the field of remote sensing. Below, we propose some potential research directions in different perspectives.

Multimodal image restoration has been attracting increasing interest from researchers, due to a variety of intriguing applications. Unlike single-modal image restoration, this task prefers to obtain complementary information from heterogeneous images for image recovery, which requires proper modeling of the dependencies among different modalities and has a highly significant role in denoising tasks such as cloud removal [204, 289, 290].

3D scene reconstruction and multi-view interpretation. Automatic 3D reconstruction of scene models from satellite images is still a challenging research topic with many interesting applications such as scene modeling, urban simulation, and path planning. When modeling a complex remote sensing scene, it is necessary to observe the scene from multiple perspectives, while involving the analysis of various data sources. Compared with the indoor reconstruction of natural scenes, large-scale remote sensing scenes are more complex (especially for complex urban areas) and therefore pose a great challenge.

In recent years, this task has started to emerge, and Huang et al. [291, 292] constructed a series of relevant datasets and applied pose estimation methods to the reconstruction algorithm, making great breakthroughs and progress.

Land use classification and detection. Although recent years have witnessed considerable progress of MRSII in land use classification and detection, it is difficult to be established in all practical scenarios because previous datasets are often not representative. For now, most methods lack robustness and universality, and they are over-designed for specific categories and datasets, weakening their applicability to other more general scenarios. An ideal multimodal decoding framework should be able to handle a variety of learning tasks with different data complexity and data sources. Therefore, how to improve the robustness and universality of methods is a hot topic of the current task.

Heterogeneous image time series change detection. Currently, the heterogeneous image change detection task only considers bi-temporal remote sensing images. While in practical applications, we often need to analyze a series of long time-series images to infer the change and development of the scene in that time period, which is extremely useful for urban development, planning, and natural environment protection.

Scene prediction and complementary is an emerging research direction. It predicts the future development of a scene or complements the elements of the intermediate moments by feature extraction and modeling of a long time series of a scene. This task offers the possibility of regional development prediction and historicity analysis.

Cross-element analysis. Due to the large scale and complexity of remote sensing images, it poses great challenges to cross-element analysis. The key to the task is to extract key instances in complex scenes and align or transform them with other modalities. Therefore, the task mainly involves the contents related to multi-source alignment (Section 3) and cross-modal translation (Section 6), mainly involving the research directions of remote sensing image-speech (text) alignment, remote sensing scene description, and remote sensing scene Q&A.

11 Conclusion

Utilizing multi-source data for large-scale scene observation and interpretation is crucial to the further development of the field of remote sensing and computer vision. To our best knowledge, this study is the first survey that describes the advances of the field of multimodal remote sensing and proposes a succinct and understandable taxonomy to group all MRSII methods. Through in-depth analysis, it analyzes them in three directions: spatial, temporal, and cross-element, and reveals the intrinsic relationships among the mainstream methods. MRSII has recently become an active research area; therefore, we hope this survey can help researchers, as a starting point, to review state-of-the-art developments and provide a

systematic and unprecedented survey for them.

Acknowledgements This work was supported by National Key R&D Program of China (Grant No. 2021YFB3900504) and National Natural Science Foundation of China (Grant Nos. 61725105, 62171436).

References

1 Srivastava S, Vargas-Munoz J E, Tuia D. Understanding urban landuse from the above and ground perspectives: a deep learning, multimodal solution. Remote Sens Environ, 2019, 228: 129–143
2 Poliyapram V, Wang W, Nakamura R. A point-wise LiDAR and image multimodal fusion network (PMNet) for aerial point cloud 3D semantic segmentation. Remote Sens, 2019, 11: 2961
3 Rostami M, Kolouri S, Eaton E, et al. Deep transfer learning for few-shot SAR image classification. Remote Sens, 2019, 11: 1374
4 Xu F, Hu C, Li J, et al. Special focus on deep learning in remote sensing image processing. Sci China Inf Sci, 2020, 63: 140300
5 He K, Zhang X, Ren S, et al. Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016. 770–778
6 Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. 2014. ArXiv:1409.1556
7 Sun K, Xiao B, Liu D, et al. Deep high-resolution representation learning for human pose estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019. 5693–5703
8 Baltrusaitis T, Ahuja C, Morency L P. Multimodal machine learning: a survey and taxonomy. IEEE Trans Pattern Anal Mach Intell, 2018, 41: 423–443
9 Uss M L, Vozel B, Lukin V V, et al. Multimodal remote sensing image registration with accuracy estimation at local and global scales. IEEE Trans Geosci Remote Sens, 2016, 54: 6587–6605
10 Fan J, Wu Y, Li M, et al. SAR and optical image registration using nonlinear diffusion and phase congruency structural descriptor. IEEE Trans Geosci Remote Sens, 2018, 56: 5368–5379
11 Wang S, Quan D, Liang X, et al. A deep learning framework for remote sensing image registration. ISPRS J Photogrammetry Remote Sens, 2018, 145: 148–164
12 Zhu Z. Change detection using landsat time series: a review of frequencies, preprocessing, algorithms, and applications. ISPRS J Photogrammetry Remote Sens, 2017, 130: 370–384
13 Mou L, Bruzzone L, Zhu X X. Learning spectral-spatial-temporal features via a recurrent convolutional neural network for change detection in multispectral imagery. IEEE Trans Geosci Remote Sens, 2018, 57: 924–935
14 Saha S, Bovolo F, Bruzzone L. Unsupervised deep change vector analysis for multiple-change detection in VHR images. IEEE Trans Geosci Remote Sens, 2019, 57: 3677–3693
15 Yan J, Wang L, Song W, et al. A time-series classification approach based on change detection for rapid land cover mapping. ISPRS J Photogrammetry Remote Sens, 2019, 158: 249–262
16 Guo M, Zhou C, Liu J. Jointly learning of visual and auditory: a new approach for RS image and audio cross-modal retrieval. IEEE J Sel Top Appl Earth Observations Remote Sens, 2019, 12: 4644–4654
17 Chen Y, Lu X, Wang S. Deep cross-modal image-voice retrieval in remote sensing. IEEE Trans Geosci Remote Sens, 2020, 58: 7049–7061
18 Yuan Z, Zhang W, Fu K, et al. Exploring a fine-grained multiscale method for cross-modal remote sensing image retrieval. IEEE Trans Geosci Remote Sens, 2022, 60: 1–19
19 Zitová B, Flusser J. Image registration methods: a survey. Image Vision Computing, 2003, 21: 977–1000
20 Moigne J L. Introduction to remote sensing image registration. In: Proceedings of IEEE International Geoscience and Remote Sensing Symposium (IGARSS), 2017. 2565–2568
21 Shi X, Deng Z, Ding X, et al. Land cover classification combining Sentinel-1 and Landsat 8 imagery driven by Markov random field with amendment reliability factors. J Wireless Com Network, 2020, 2020: 87
22 Ma L, Crawford M M, Zhu L, et al. Centroid and covariance alignment-based domain adaptation for unsupervised classification of remote sensing images. IEEE Trans Geosci Remote Sens, 2018, 57: 2305–2323
23 Gao G, Gu Y. Tensorized principal component alignment: a unified framework for multimodal high-resolution images classification. IEEE Trans Geosci Remote Sens, 2019, 57: 46–61
24 Sun Y, Lei L, Li X, et al. Patch similarity graph matrix-based unsupervised remote sensing change detection with homogeneous and heterogeneous sensors. IEEE Trans Geosci Remote Sens, 2020, 59: 4841–4861
25 Sun Y, Lei L, Guan D, et al. Iterative robust graph for unsupervised change detection of heterogeneous remote sensing images. IEEE Trans Image Process, 2021, 30: 6277–6291
26 Sun Y, Lei L, Li X, et al. Nonlocal patch similarity based heterogeneous remote sensing change detection. Pattern Recognition, 2021, 109: 107598
27 Garnot V S F, Landrieu L, Giordano S, et al. Satellite image time series classification with pixel-set encoders and temporal self-attention. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020. 12325–12334
28 Garnot V S F, Landrieu L. Lightweight temporal self-attention for classifying satellite images time series. In: Proceedings of International Workshop on Advanced Analytics and Learning on Temporal Data, 2020. 171–181
29 Abdullah T, Bazi Y, Al Rahhal M M, et al. TextRS: deep bidirectional triplet network for matching text to remote sensing images. Remote Sens, 2020, 12: 405
30 Cheng Q, Zhou Y, Fu P, et al. A deep semantic alignment network for the cross-modal image-text retrieval in remote sensing. IEEE J Sel Top Appl Earth Observations Remote Sens, 2021, 14: 4284–4297
31 Yan L, Wang Z, Liu Y, et al. Generic and automatic Markov random field-based registration for multimodal remote sensing image using grayscale and gradient information. Remote Sens, 2018, 10: 1228
32 Xiang Y, Tao R, Wan L, et al. OS-PC: combining feature representation and 3-D phase correlation for subpixel optical and SAR image registration. IEEE Trans Geosci Remote Sens, 2020, 58: 6451–6466
33 Cole-Rhodes A A, Johnson K L, Lemoigne J, et al. Multiresolution registration of remote sensing imagery by optimization of mutual information using a stochastic gradient. IEEE Trans Image Process, 2003, 12: 1495–1511
34 Chen H-M, Varshney P K, Arora M K. Performance of mutual information similarity measure for registration of multitemporal remote sensing images. IEEE Trans Geosci Remote Sens, 2003, 41: 2445–2454

35 Fan X F, Rhody H, Saber E. A Spatial-feature-enhanced MMI algorithm for multimodal airborne image registration. IEEE Trans Geosci Remote Sens, 2010, 48: 2580–2589
36 Gross W, Espinosa N, Becker M, et al. Improving linear classification using semi-supervised invertible manifold alignment. In: Proceedings of IEEE International Geoscience and Remote Sensing Symposium, 2018. 3551–3554
37 Pournemat A, Adibi P, Chanussot J. Semisupervised charting for spectral multimodal manifold learning and alignment. Pattern Recognition, 2021, 111: 107645
38 Hu J, Hong D, Zhu X X. MIMA: MAPPER-induced manifold alignment for semi-supervised fusion of optical image and polarimetric SAR data. IEEE Trans Geosci Remote Sens, 2019, 57: 9025–9040
39 Devis T, Michele V, Maxime T, et al. Semisupervised manifold alignment of multimodal remote sensing images. IEEE Trans Geosci Remote Sens, 2014, 52: 7708–7720
40 Hong D, Yokoya N, Ge N, et al. Learnable manifold alignment (LeMA): a semi-supervised cross-modality learning framework for land cover and land use classification. ISPRS J Photogrammetry Remote Sens, 2019, 147: 193–205
41 Ye Y, Shan J, Bruzzone L, et al. Robust registration of multimodal remote sensing images based on structural similarity. IEEE Trans Geosci Remote Sens, 2017, 55: 2941–2958
42 Li Z, Zhang H, Huang Y. A rotation-invariant optical and SAR image registration algorithm based on deep and Gaussian features. Remote Sens, 2021, 13: 2628
43 Ye Y, Yang C, Zhu B, et al. Improving co-registration for sentinel-1 SAR and sentinel-2 optical images. Remote Sens, 2021, 13: 928
44 Quan D, Wang S, Liang X, et al. Deep generative matching network for optical and SAR image registration. In: Proceedings of IEEE International Geoscience and Remote Sensing Symposium, 2018. 6215–6218
45 Zhang J, Ma W, Wu Y, et al. Multimodal remote sensing image registration based on image transfer and local features. IEEE Geosci Remote Sens Lett, 2019, 16: 1210–1214
46 Xiang Y, Tao R, Wang F, et al. Automatic registration of optical and SAR images via improved phase congruency model. IEEE J Sel Top Appl Earth Observations Remote Sens, 2020, 13: 5847–5861
47 Zhang H, Ni W, Yan W, et al. Registration of multimodal remote sensing image based on deep fully convolutional neural network. IEEE J Sel Top Appl Earth Observations Remote Sens, 2019, 12: 3028–3042
48 Fan R, Hou B, Liu J, et al. Registration of multiresolution remote sensing images based on L2-siamese model. IEEE J Sel Top Appl Earth Observations Remote Sens, 2020, 14: 237–248
49 Mao S, Yang J, Gou S, et al. Multi-scale fused SAR image registration based on deep forest. Remote Sens, 2021, 13: 2227
50 Jimenez-Sierra D A, Benítez-Restrepo H D, Vargas-Cardona H D, et al. Graph-based data fusion applied to: change detection and biomass estimation in rice crops. Remote Sens, 2020, 12: 2683
51 Yang M, Jiao L, Liu F, et al. DPFL-Nets: deep pyramid feature learning networks for multiscale change detection. IEEE Trans Neural Netw Learn Syst, 2022, 33: 6402–6416
52 Xue D, Lei T, Jia X, et al. Unsupervised change detection using multiscale and multiresolution gaussian-mixture-model guided by saliency enhancement. IEEE J Sel Top Appl Earth Observations Remote Sens, 2020, 14: 1796–1809
53 Chen Y, Bruzzone L. Self-supervised change detection in multi-view remote sensing images. 2021. ArXiv:2103.05969
54 Guo H, Shi Q, Marinoni A, et al. Deep building footprint update network: a semi-supervised method for updating existing building footprint from bi-temporal remote sensing images. Remote Sens Environ, 2021, 264: 112589
55 Kaiser P, Wegner J D, Lucchi A, et al. Learning aerial image segmentation from online maps. IEEE Trans Geosci Remote Sens, 2017, 55: 6054–6068
56 Zampieri A, Charpiat G, Tarabalka Y. Coarse to fine non-rigid registration: a chain of scale-specific neural networks for multimodal image alignment with application to remote sensing. 2018. ArXiv:1802.09816
57 Kocur-Bera K, Dawidowicz A. Land use versus land cover: geo-analysis of national roads and synchronisation algorithms. Remote Sens, 2019, 11: 3053
58 Zhong Y, Su Y, Wu S, et al. Open-source data-driven urban land-use mapping integrating point-line-polygon semantic objects: a case study of Chinese cities. Remote Sens Environ, 2020, 247: 111838
59 Corona P, Fattorini L, Franceschi S, et al. Mapping by spatial predictors exploiting remotely sensed and ground data: a comparative design-based perspective. Remote Sens Environ, 2014, 152: 29–37
60 Chen P, Yao W, Zhu X. Combination of ground- and space-based data to establish a global ionospheric grid model. IEEE Trans Geosci Remote Sens, 2014, 53: 1073–1081
61 Zhang R, Zhou X, Ouyang Z, et al. Estimating aboveground biomass in subtropical forests of China by integrating multisource remote sensing and ground data. Remote Sens Environ, 2019, 232: 111341
62 Babaeian E, Paheding S, Siddique N, et al. Estimation of root zone soil moisture from ground and remotely sensed soil information with multisensor data fusion and automated machine learning. Remote Sens Environ, 2021, 260: 112434
63 Handcock R, Swain D, Bishop-Hurley G, et al. Monitoring animal behaviour and environmental interactions using wireless sensor networks, GPS collars and satellite remote sensing. Sensors, 2009, 9: 3586–3603
64 McRoberts R E, Chen Q, Walters B F, et al. The effects of global positioning system receiver accuracy on airborne laser scanning-assisted estimates of aboveground biomass. Remote Sens Environ, 2018, 207: 42–49
65 Carlá T, Tofani V, Lombardi L, et al. Combination of GNSS, satellite InSAR, and GBInSAR remote sensing monitoring to improve the understanding of a large landslide in high alpine environment. Geomorphology, 2019, 335: 62–75
66 Jat M K, Garg P K, Khare D. Monitoring and modelling of urban sprawl using remote sensing and GIS techniques. Int J Appl Earth Observation Geoinf, 2008, 10: 26–43
67 Bachagha N, Wang X, Luo L, et al. Remote sensing and GIS techniques for reconstructing the military fort system on the Roman boundary (Tunisian section) and identifying archaeological sites. Remote Sens Environ, 2020, 236: 111418
68 Manzoni M, Monti-Guarnieri A, Molinari M E. Joint exploitation of spaceborne SAR images and GIS techniques for urban coherent change detection. Remote Sens Environ, 2021, 253: 112152
69 Moradkhani H. Hydrologic remote sensing and land surface data assimilation. Sensors, 2008, 8: 2986–3004
70 Khan S I, Hong Y, Wang J, et al. Satellite remote sensing and hydrologic modeling for flood inundation mapping in lake victoria basin: implications for hydrologic prediction in ungauged basins. IEEE Trans Geosci Remote Sens, 2010, 49: 85–95
71 Cimini D, Pierdicca N, Pichelli E, et al. On the accuracy of integrated water vapor observations and the potential for mitigating electromagnetic path delay error in InSAR. Atmos Meas Tech, 2012, 5: 1015–1030
72 Mao G, Yuan Y, Lu X Q. Deep cross-modal retrieval for remote sensing image and audio. In: Proceedings of the 10th IAPR Workshop on Pattern Recognition in Remote Sensing (PRRS), 2018. 1–7
73 Chaudhuri U, Banerjee B, Bhattacharya A, et al. CMIR-NET: a deep learning based model for cross-modal retrieval in

remote sensing. Pattern Recognition Lett, 2020, 131: 456–462
74 Ning H, Zhao B, Yuan Y. Semantics-consistent representation learning for remote sensing image-voice retrieval. IEEE Trans Geosci Remote Sens, 2022, 60: 1–14
75 Zhou N, Fan J. Automatic image-text alignment for large-scale web image indexing and retrieval. Pattern Recognition, 2015, 48: 205–219
76 Wehrmann J, Kolling C, Barros R C. Adaptive cross-modal embeddings for image-text alignment. In: Proceedings of the AAAI Conference on Artificial Intelligence, 2020. 12313–12320
77 Zhang F, Xu M, Mao Q, et al. Joint attribute manipulation and modality alignment learning for composing text and image to image retrieval. In: Proceedings of the 28th ACM International Conference on Multimedia, 2020. 3367–3376
78 Sargin M E, Yemez Y, Erzin E, et al. Audiovisual synchronization and fusion using canonical correlation analysis. IEEE Trans Multimedia, 2007, 9: 1396–1403
79 Halperin T, Ephrat A, Peleg S. Dynamic temporal alignment of speech to lips. In: Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2019. 3980–3984
80 Wang J, Fang Z, Zhao H. AlignNet: a unifying approach to audio-visual alignment. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2020. 3309–3317
81 Bojanowski P, Lajugie R, Grave E, et al. Weakly-supervised alignment of video with text. In: Proceedings of the IEEE International Conference on Computer Vision, 2015. 4462–4470
82 Song Y C, Naim I, Mamun A A, et al. Unsupervised alignment of actions in video with text descriptions. In: Proceedings of the 25th International Joint Conference on Artificial Intelligence, 2016. 2025–2031
83 Wang X, Zhu L, Yang Y. T2VLAD: global-local sequence alignment for text-video retrieval. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021. 5079–5088
84 Walker J J, de Beurs K M, Wynne R H, et al. Evaluation of Landsat and MODIS data fusion products for analysis of dryland forest phenology. Remote Sens Environ, 2012, 117: 381–393
85 Ward D P, Petty A, Setterfield S A, et al. Floodplain inundation and vegetation dynamics in the Alligator Rivers region (Kakadu) of northern Australia assessed using optical and radar remote sensing. Remote Sens Environ, 2014, 147: 43–55
86 Zhao Y, Huang B, Song H. A robust adaptive spatial and temporal image fusion model for complex land surface changes. Remote Sens Environ, 2018, 208: 42–62
87 Gevaert C M, Suomalainen J, Tang J, et al. Generation of spectral-temporal response surfaces by combining multispectral satellite and hyperspectral UAV imagery for precision agriculture applications. IEEE J Sel Top Appl Earth Observations Remote Sens, 2015, 8: 3140–3146
88 Maimaitijiang M, Ghulam A, Sidike P, et al. Unmanned aerial system (UAS)-based phenotyping of soybean using multi-sensor data fusion and extreme learning machine. ISPRS J Photogrammetry Remote Sens, 2017, 134: 43–58
89 Kimm H, Guan K, Jiang C, et al. Deriving high-spatiotemporal-resolution leaf area index for agroecosystems in the U.S. Corn Belt using Planet Labs CubeSat and STAIR fusion data. Remote Sens Environ, 2020, 239: 111615
90 Im J, Lu Z, Rhee J, et al. Impervious surface quantification using a synthesis of artificial immune networks and decision/regression trees from multi-sensor data. Remote Sens Environ, 2012, 117: 102–113
91 Liu L, Coops N C, Aven N W, et al. Mapping urban tree species using integrated airborne hyperspectral and LiDAR remote sensing data. Remote Sens Environ, 2017, 200: 170–182
92 Cao R, Tu W, Yang C, et al. Deep learning-based remote and social sensing data fusion for urban region function recognition. ISPRS J Photogrammetry Remote Sens, 2020, 163: 82–97
93 Hall D L, Llinas J. An introduction to multisensor data fusion. Proc IEEE, 1997, 85: 6–23
94 Pradhan P S, King R L, Younan N H, et al. Estimation of the number of decomposition levels for a wavelet-based multiresolution multisensor image fusion. IEEE Trans Geosci Remote Sens, 2006, 44: 3674–3686
95 Palsson F, Sveinsson J R, Ulfarsson M O, et al. Model-based fusion of multi- and hyperspectral images using PCA and wavelets. IEEE Trans Geosci Remote Sens, 2014, 53: 2652–2663
96 Schmitt M, Zhu X X. Data fusion and remote sensing: an ever-growing relationship. IEEE Geosci Remote Sens Mag, 2016, 4: 6–23
97 Moosavi V, Talebi A, Mokhtari M H, et al. A wavelet-artificial intelligence fusion approach (WAIFA) for blending Landsat and MODIS surface temperature. Remote Sens Environ, 2015, 169: 243–254
98 Chen Y, Li C, Ghamisi P, et al. Deep fusion of remote sensing data for accurate classification. IEEE Geosci Remote Sens Lett, 2017, 14: 1253–1257
99 Li H, Ghamisi P, Soergel U, et al. Hyperspectral and LiDAR fusion using deep three-stream convolutional neural networks. Remote Sens, 2018, 10: 1649
100 Vörösmarty C J, Willmott C J, Choudhury B J, et al. Analyzing the discharge regime of a large tropical river through remote sensing, ground-based climatic data, and modeling. Water Resour Res, 1996, 32: 3137–3150
101 Chatterjee A, Michalak A M, Kahn R A, et al. A geostatistical data fusion technique for merging remote sensing and ground-based observations of aerosol optical thickness. J Geophys Res, 2010, 115: D20207
102 Tian J, Chen D. A semi-empirical model for predicting hourly ground-level fine particulate matter (PM2.5) concentration in southern Ontario from satellite remote sensing and ground-based meteorological measurements. Remote Sens Environ, 2010, 114: 221–229
103 Li F, Zhang X, Kondragunta S, et al. A preliminary evaluation of GOES-16 active fire product using Landsat-8 and VIIRS active fire data, and ground-based prescribed fire records. Remote Sens Environ, 2020, 237: 111600
104 Alparone L, Aiazzi B, Baronti S, et al. Multispectral and panchromatic data fusion assessment without reference. photogramm eng remote Sens, 2008, 74: 193–200
105 Li Z, Leung H. Fusion of multispectral and panchromatic images using a restoration-based method. IEEE Trans Geosci Remote Sens, 2009, 47: 1482–1491
106 Zhang L P, Shen H F, Gong W, et al. Adjustable model-based fusion method for multispectral and panchromatic images. IEEE Trans Syst Man Cybern B, 2012, 42: 1693–1704
107 Chanussot J, Mauris G, Lambert P. Fuzzy fusion techniques for linear features detection in multitemporal SAR images. IEEE Trans Geosci Remote Sens, 1999, 37: 1292–1305
108 Jeon B, Landgrebe D A. Decision fusion approach for multitemporal classification. IEEE Trans Geosci Remote Sens, 1999, 37: 1227–1233
109 Dai X, Khorram S. Data fusion using artificial neural networks: a case study on multitemporal change analysis. Comput Environ Urban Syst, 1999, 23: 19–31

110 McKeown D M, Cochran S D, Ford S J, et al. Fusion of HYDICE hyperspectral data with panchromatic imagery for cartographic feature extraction. IEEE Trans Geosci Remote Sens, 1999, 37: 1261–1277
111 Hardie R C, Eismann M T, Wilson G L. MAP estimation for hyperspectral image resolution enhancement using an auxiliary sensor. IEEE Trans Image Process, 2004, 13: 1174–1184
112 Cetin M, Musaoglu N. Merging hyperspectral and panchromatic image data: qualitative and quantitative analysis. Int J Remote Sens, 2009, 30: 1779–1804
113 Zehtabian A, Ghassemian H. An adaptive pixon extraction technique for multispectral/hyperspectral image classification. IEEE Geosci Remote Sens Lett, 2015, 12: 831–835
114 Yokoya N, Grohnfeldt C, Chanussot J. Hyperspectral and multispectral data fusion: a comparative review of the recent literature. IEEE Geosci Remote Sens Mag, 2017, 5: 29–56
115 Palsson F, Sveinsson J R, Ulfarsson M O. Multispectral and hyperspectral image fusion using a 3-D-convolutional neural network. IEEE Geosci Remote Sens Lett, 2017, 14: 639–643
116 Haydn R. Application of the IHS color transform to the processing of multisensor data and image enhancement. In: Proceedings of the International Symposium on Remote Sensing of Arid and Semi-Arid Lands, Cairo, 1982
117 Carper W, Lillesand T, Kiefer R. The use of intensity-hue-saturation transformations for merging spot panchromatic and multispectral image data. Photogrammetric Engin Remote Sens, 1990, 56: 459–467
118 Ehlers M. Multisensor image fusion techniques in remote sensing. ISPRS J Photogrammetry Remote Sens, 1991, 46: 19–30
119 Ling Y, Ehlers M, Usery E L, et al. FFT-enhanced IHS transform method for fusing high-resolution satellite images. ISPRS J Photogrammetry Remote Sens, 2007, 61: 381–392
120 Chavez P, Sides S C, Anderson J A, et al. Comparison of three different methods to merge multiresolution and multispectral data- landsat tm and spot panchromatic. Photogrammetric Engin Remote Sens, 1991, 57: 295–303
121 Shettigara V K. A generalized component substitution technique for spatial enhancement of multispectral images using a higher resolution data set. Photogrammetric Engin Remote Sens, 1992, 58: 561–567
122 Licciardi G, Khan M M, Chanussot J. Fusion of hyperspectral and panchromatic images: a hybrid use of indusion and nonlinear PCA. In: Proceedings of the 19th IEEE International Conference on Image Processing, 2012. 2133–2136
123 Shahdoosti H R, Ghassemian H. Combining the spectral PCA and spatial PCA fusion methods by an optimal filter. Inf Fusion, 2016, 27: 150–160
124 Aiazzi B, Baronti S, Selva M. Improving component substitution pansharpening through multivariate regression of MS +Pan data. IEEE Trans Geosci Remote Sens, 2007, 45: 3230–3239
125 Maurer T. How to pan-sharpen images using the Gram-Schmidt pan-sharpen method—a recipe. In: Proceedings of International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, 2013. 239–244
126 Yilmaz V, Yilmaz C S, Güngör O, et al. A genetic algorithm solution to the Gram-Schmidt image fusion. Int J Remote Sens, 2020, 41: 1458–1485
127 Tu T M, Lee Y C, Chang C P, et al. Adjustable intensity-hue-saturation and Brovey transform fusion technique for IKONOS/QuickBird imagery. Opt Eng, 2005, 44: 116201
128 Du Q, Younan N H, King R, et al. On the performance evaluation of pan-sharpening techniques. IEEE Geosci Remote Sens Lett, 2007, 4: 518–522
129 Dian R, Li S. Hyperspectral image super-resolution via subspace-based low tensor multi-rank regularization. IEEE Trans Image Process, 2019, 28: 5135–5146
130 Xu H, Qin M, Chen S, et al. Hyperspectral-multispectral image fusion via tensor ring and subspace decompositions. IEEE J Sel Top Appl Earth Observations Remote Sens, 2021, 14: 8823–8837
131 Tu T M, Huang P S, Hung C L, et al. A fast intensity-hue-saturation fusion technique with spectral adjustment for IKONOS imagery. IEEE Geosci Remote Sens Lett, 2004, 1: 309–312
132 Chen Z, Pu H, Wang B, et al. Fusion of hyperspectral and multispectral images: a novel framework based on generalization of pan-sharpening methods. IEEE Geosci Remote Sens Lett, 2014, 11: 1418–1422
133 Gangkofner U G, Pradhan P S, Holcomb D W. Optimizing the high-pass filter addition technique for image fusion. Photogramm Eng Remote Sens, 2008, 74: 1107–1118
134 Nunez J, Otazu X, Fors O, et al. Multiresolution-based image fusion with additive wavelet decomposition. IEEE Trans Geosci Remote Sens, 1999, 37: 1204–1211
135 Aiazzi B, Alparone L, Barducci A, et al. Multispectral fusion of multisensor image data by the generalized laplacian pyramid. In: Proceedings of IEEE 1999 International Geoscience and Remote Sensing Symposium, 1999. 1183–1185
136 Nencini F, Garzelli A, Baronti S, et al. Remote sensing image fusion using the curvelet transform. Inf Fusion, 2007, 8: 143–156
137 Choi M, Kim R Y, Nam M R, et al. Fusion of multispectral and panchromatic satellite images using the curvelet transform. IEEE Geosci Remote Sens Lett, 2005, 2: 136–140
138 Dong L, Yang Q, Wu H, et al. High quality multi-spectral and panchromatic image fusion technologies based on Curvelet transform. Neurocomputing, 2015, 159: 268–274
139 Masi G, Cozzolino D, Verdoliva L, et al. Pansharpening by convolutional neural networks. Remote Sens, 2016, 8: 594
140 Wei Y, Yuan Q, Shen H, et al. Boosting the accuracy of multispectral image pansharpening by learning a deep residual network. IEEE Geosci Remote Sens Lett, 2017, 14: 1795–1799
141 Yang J, Fu X, Hu Y, et al. PanNet: a deep network architecture for pan-sharpening. In: Proceedings of the IEEE International Conference on Computer Vision, 2017. 5449–5457
142 Scarpa G, Vitale S, Cozzolino D. Target-adaptive CNN-based pansharpening. IEEE Trans Geosci Remote Sens, 2018, 56: 5443–5457
143 Gao F, Masek J, Schwaller M, et al. On the blending of the Landsat and MODIS surface reflectance: predicting daily Landsat surface reflectance. IEEE Trans Geosci Remote Sens, 2006, 44: 2207–2218
144 Zhu X, Chen J, Gao F, et al. An enhanced spatial and temporal adaptive reflectance fusion model for complex heterogeneous regions. Remote Sens Environ, 2010, 114: 2610–2623
145 Gevaert C M, García-Haro F J. A comparison of STARFM and an unmixing-based algorithm for Landsat and MODIS data fusion. Remote Sens Environ, 2015, 156: 34–44
146 Xie D, Zhang J, Zhu X, et al. An improved STARFM with help of an unmixing-based method to generate high spatial and temporal resolution remote sensing data in complex heterogeneous regions. Sensors, 2016, 16: 207
147 Xue J, Leung Y, Fung T. A Bayesian data fusion approach to spatio-temporal fusion of remotely sensed images. Remote Sens, 2017, 9: 1310

148 Pedergnana M, Marpu P R, Mura M D, et al. Classification of remote sensing optical and LiDAR data using extended attribute profiles. IEEE J Sel Top Signal Process, 2012, 6: 856–865
149 Chini M, Pierdicca N, Emery W J. Exploiting SAR and VHR optical images to quantify damage caused by the 2003 Bam earthquake. IEEE Trans Geosci Remote Sens, 2008, 47: 145–152
150 Pedergnana M, Marpu P R, Mura M D, et al. A novel technique for optimal feature selection in attribute profiles based on genetic algorithms. IEEE Trans Geosci Remote Sens, 2013, 51: 3514–3528
151 Ghamisi P, Benediktsson J A, Phinn S. Land-cover classification using both hyperspectral and LiDAR data. Int J Image Data Fusion, 2015, 6: 189–215
152 Rasti B, Ghamisi P. Remote sensing image classification using subspace sensor fusion. Inf Fusion, 2020, 64: 121–130
153 Rasti B, Ulfarsson M O, Sveinsson J R. Hyperspectral feature extraction using total variation component analysis. IEEE Trans Geosci Remote Sens, 2016, 54: 6976–6985
154 Rasti B, Ghamisi P, Gloaguen R. Hyperspectral and LiDAR fusion using extinction profiles and total variation component analysis. IEEE Trans Geosci Remote Sens, 2017, 55: 3997–4007
155 McCabe M F, Wood E F, Wójcik R, et al. Hydrological consistency using multi-sensor remote sensing data for water and energy cycle studies. Remote Sens Environ, 2008, 112: 430–444
156 Awange J L, Schumacher M, Forootan E, et al. Exploring hydro-meteorological drought patterns over the Greater Horn of Africa (1979–2014) using remote sensing and reanalysis products. Adv Water Resources, 2016, 94: 45–59
157 Teillet P. Effects of spectral, spatial, and radiometric characteristics on remote sensing vegetation indices of forested regions. Remote Sens Environ, 1997, 61: 139–149
158 Babst F, Esper J, Parlow E. Landsat TM/ETM+ and tree-ring based assessment of spatiotemporal patterns of the autumnal moth (Epirrita autumnata) in northernmost Fennoscandia. Remote Sens Environ, 2010, 114: 637–646
159 Guanter L, Richter R, Kaufmann H. On the application of the MODTRAN4 atmospheric radiative transfer code to optical remote sensing. Int J Remote Sens, 2009, 30: 1407–1424
160 Bloom A A, Worden J, Jiang Z, et al. Remote-sensing constraints on South America fire traits by Bayesian fusion of atmospheric and surface data. Geophys Res Lett, 2015, 42: 1268–1274
161 Bengio Y, Courville A, Vincent P. Representation learning: a review and new perspectives. IEEE Trans Pattern Anal Mach Intell, 2013, 35: 1798–1828
162 Krizhevsky A, Sutskever I, Hinton G E. ImageNet classification with deep convolutional neural networks. Commun ACM, 2017, 60: 84–90
163 Liu Z, Lin Y, Cao Y, et al. Swin transformer: hierarchical vision transformer using shifted windows. 2021. ArXiv:2103.14030
164 Zhang L, Lan M, Zhang J, et al. Stagewise unsupervised domain adaptation with adversarial self-training for road segmentation of remote-sensing images. IEEE Trans Geosci Remote Sens, 2022, 60: 1–13
165 Mikolov T, Chen K, Corrado G, et al. Efficient estimation of word representations in vector space. 2013. ArXiv:1301.3781
166 Pennington J, Socher R, Manning C D. Glove: global vectors for word representation. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2014. 1532–1543
167 Devlin J, Chang M W, Lee K, et al. Bert: pre-training of deep bidirectional transformers for language understanding. 2018. ArXiv:1810.04805
168 Schneider S, Baevski A, Collobert R, et al. wav2vec: unsupervised pre-training for speech recognition. 2019. ArXiv:1904.05862
169 Xu Q, Baevski A, Likhomanenko T, et al. Self-training and pre-training are complementary for speech recognition. In: Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2021. 3030–3034
170 Perozzi B, Al-Rfou R, Skiena S. Deepwalk: online learning of social representations. In: Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2014. 701–710
171 Grover A, Leskovec J. node2vec: scalable feature learning for networks. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016. 855–864
172 Wang H, Wang J, Wang J, et al. GraphGAN: graph representation learning with generative adversarial nets. In: Proceedings of the AAAI Conference on Artificial Intelligence, 2018
173 Guo W, Wang J, Wang S. Deep multimodal representation learning: a survey. IEEE Access, 2019, 7: 63373–63394
174 LeCun Y, Bottou L, Bengio Y, et al. Gradient-based learning applied to document recognition. Proc IEEE, 1998, 86: 2278–2324
175 Szegedy C, Liu W, Jia Y, et al. Going deeper with convolutions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015. 1–9
176 Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need. In: Proceedings of Advances in Neural Information Processing Systems, 2017. 5998–6008
177 Dosovitskiy A, Beyer L, Kolesnikov A, et al. An image is worth 16×16 words: transformers for image recognition at scale. In: Proceedings of International Conference on Learning Representations, 2020
178 Yuan L, Chen Y, Wang T, et al. Tokens-to-Token ViT: training vision transformers from scratch on ImageNet. 2021. ArXiv:2101.11986
179 Bazi Y, Bashmal L, Rahhal M M A, et al. Vision transformers for remote sensing image classification. Remote Sens, 2021, 13: 516
180 He X, Chen Y, Lin Z. Spatial-spectral transformer for hyperspectral image classification. Remote Sens, 2021, 13: 498
181 Chen H, Qi Z, Shi Z. Remote sensing image change detection with transformers. IEEE Trans Geosci Remote Sens, 2022, 60: 1–14
182 Pascual S, Ravanelli M, Serrà J, et al. Learning problem-agnostic speech representations from multiple self-supervised tasks. In: Proceedings of Interspeech 2019, 2019. 161–165
183 Liu A T, Yang S W, Chi P H, et al. Mockingjay: unsupervised speech representation learning with deep bidirectional transformer encoders. In: Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2020. 6419–6423
184 Sharma M, Dhanaraj M, Karnam S, et al. YOLOrs: object detection in multimodal remote sensing imagery. IEEE J Sel Top Appl Earth Observations Remote Sens, 2020, 14: 1497–1508
185 Yang D, Liu X, He H, et al. Air-to-ground multimodal object detection algorithm based on feature association learning. Int J Adv Robotic Syst, 2019, 16: 172988141984299
186 Flynn H, Cameron S. Multi-modal people detection from aerial video. In: Proceedings of the 8th International Conference on Computer Recognition Systems, 2013. 815–824

187 de Oliveira D C, Wehrmeister M A. Towards real-time people recognition on aerial imagery using convolutional neural networks. In: Proceedings of IEEE 19th International Symposium on Real-Time Distributed Computing, 2016. 27–34
188 Breckon T P, Gaszczak A, Han J, et al. Multi-modal target detection for autonomous wide area search and surveillance. In: Proceedings of SPIE—International Society for Optics and Photonics, 2013. 889913
189 Audebert N, Le Saux B, Lefèvre S. Semantic segmentation of earth observation data using multimodal and multi-scale deep networks. In: Proceedings of Asian Conference on Computer Vision, 2016. 180–196
190 Audebert N, Le Saux B, Lefévre S. Beyond RGB: very high resolution urban remote sensing with multimodal deep networks. ISPRS J Photogrammetry Remote Sens, 2018, 140: 20–32
191 Li X, Lei L, Sun Y, et al. Multimodal bilinear fusion network with second-order attention-based channel selection for land cover classification. IEEE J Sel Top Appl Earth Observations Remote Sens, 2020, 13: 1011–1026
192 Jeong J, Yoon T S, Park J B. Multimodal sensor-based semantic 3D mapping for a large-scale environment. Expert Syst Appl, 2018, 105: 1–10
193 Farooq A, Jia X, Hu J, et al. Multi-resolution weed classification via convolutional neural network and superpixel based local binary pattern using remote sensing images. Remote Sens, 2019, 11: 1692
194 Li Z, Chen G, Zhang T. A CNN-transformer hybrid approach for crop classification using multitemporal multisensor images. IEEE J Sel Top Appl Earth Observations Remote Sens, 2020, 13: 847–858
195 Zhou M, Jing M, Liu D, et al. Multi-resolution networks for ship detection in infrared remote sensing images. Infrared Phys Tech, 2018, 92: 183–189
196 Wang Y, Wang C, Zhang H, et al. Automatic ship detection based on RetinaNet using multi-resolution Gaofen-3 imagery. Remote Sens, 2019, 11: 531
197 Bergado J R, Persello C, Stein A. Recurrent multiresolution convolutional networks for VHR image classification. IEEE Trans Geosci Remote Sens, 2018, 56: 6361–6374
198 Robinson C, Hou L, Malkin K, et al. Large scale high-resolution land cover mapping with multi-resolution data. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019. 12726–12735
199 Wirion C, Bauwens W, Verbeiren B. Location- and time-specific hydrological simulations with multi-resolution remote sensing data in urban areas. Remote Sens, 2017, 9: 645
200 Ye Y, Bruzzone L, Shan J, et al. Fast and robust matching for multimodal remote sensing image registration. IEEE Trans Geosci Remote Sens, 2019, 57: 9059–9070
201 Uss M, Vozel B, Lukin V, et al. Efficient discrimination and localization of multimodal remote sensing images using CNN-based prediction of localization uncertainty. Remote Sens, 2020, 12: 703
202 Zhu R, Yu D, Ji S, et al. Matching RGB and infrared remote sensing images with densely-connected convolutional neural networks. Remote Sens, 2019, 11: 2836
203 Huang B, Li Y, Han X, et al. Cloud removal from optical satellite imagery with SAR imagery using sparse representation. IEEE Geosci Remote Sens Lett, 2015, 12: 1046–1050
204 Meraner A, Ebel P, Zhu X X, et al. Cloud removal in Sentinel-2 imagery using a deep residual neural network and SAR-optical data fusion. ISPRS J Photogrammetry Remote Sens, 2020, 166: 333–346
205 Zhao Y, Shen S, Hu J, et al. Cloud removal using multimodal GAN with adversarial consistency loss. IEEE Geosci Remote Sens Lett, 2022, 19: 1–5
206 Dai P, Ji S, Zhang Y. Gated convolutional networks for cloud removal from bi-temporal remote sensing images. Remote Sens, 2020, 12: 3427
207 Hong D, Yao J, Meng D, et al. Multimodal GANs: toward crossmodal hyperspectral-multispectral image segmentation. IEEE Trans Geosci Remote Sens, 2020, 59: 5103–5113
208 Liu X, Hong D, Chanussot J, et al. Modality translation in remote sensing time series. IEEE Trans Geosci Remote Sens, 2022, 60: 1–14
209 Sun L, Mi X, Wei J, et al. A cloud detection algorithm-generating method for remote sensing data at visible to short-wave infrared wavelengths. ISPRS J Photogrammetry Remote Sens, 2017, 124: 70–88
210 Ao D, Dumitru C O, Schwarz G, et al. Dialectical GAN for SAR image translation: from Sentinel-1 to TerraSAR-X. Remote Sens, 2018, 10: 1597
211 Gao J, Yuan Q, Li J, et al. Cloud removal with fusion of high resolution optical and SAR images using generative adversarial networks. Remote Sens, 2020, 12: 191
212 Fu S L, Xu F, Jin Y-Q. Reciprocal translation between SAR and optical remote sensing images with cascaded-residual adversarial networks. Sci China Inf Sci, 2021, 64: 122301
213 Shi Z, Zou Z. Can a machine generate humanlike language descriptions for a remote sensing image? IEEE Trans Geosci Remote Sens, 2017, 55: 3623–3634
214 Lu X, Wang B, Zheng X, et al. Exploring models and data for remote sensing image caption generation. IEEE Trans Geosci Remote Sens, 2017, 56: 2183–2195
215 Shen X, Liu B, Zhou Y, et al. Remote sensing image captioning via variational autoencoder and reinforcement learning. Knowledge-Based Syst, 2020, 203: 105920
216 Ju J, Roy D P. The availability of cloud-free Landsat ETM+ data over the conterminous United States and globally. Remote Sens Environ, 2008, 112: 1196–1211
217 Ling F, Du Y, Li X, et al. Interpolation-based super-resolution land cover mapping. Remote Sens Lett, 2013, 4: 629–638
218 Pignol F, Colone F, Martelli T. Lagrange-polynomial-interpolation-based keystone transform for a passive radar. IEEE Trans Aerosp Electron Syst, 2017, 54: 1151–1167
219 Zhang Y, Fan Q, Bao F, et al. Single-image super-resolution based on rational fractal interpolation. IEEE Trans Image Process, 2018, 27: 3782–3797
220 Chavez-Roman H, Ponomaryov V. Super resolution image generation using wavelet domain interpolation with edge extraction via a sparse representation. IEEE Geosci Remote Sens Lett, 2014, 11: 1777–1781
221 Shao Z, Wang L, Wang Z, et al. Remote sensing image super-resolution using sparse representation and coupled sparse autoencoder. IEEE J Sel Top Appl Earth Observations Remote Sens, 2019, 12: 2663–2674
222 Hou B, Zhou K, Jiao L. Adaptive super-resolution for remote sensing images based on sparse representation with global joint dictionary model. IEEE Trans Geosci Remote Sens, 2017, 56: 2312–2327
223 Chang Y, Luo B. Bidirectional convolutional LSTM neural network for remote sensing image super-resolution. Remote Sens, 2019, 11: 2333
224 Gu J, Sun X, Zhang Y, et al. Deep residual squeeze and excitation network for remote sensing image super-resolution.

Remote Sens, 2019, 11: 1817
225 Lu T, Wang J, Zhang Y, et al. Satellite image super-resolution via multi-scale residual deep neural network. Remote Sens, 2019, 11: 1588
226 Haut J M, Fernandez-Beltran R, Paoletti M E, et al. A new deep generative network for unsupervised remote sensing single-image super-resolution. IEEE Trans Geosci Remote Sens, 2018, 56: 6792–6810
227 Lei S, Shi Z, Zou Z. Coupled adversarial training for remote sensing image super-resolution. IEEE Trans Geosci Remote Sens, 2019, 58: 3633–3643
228 Xiong Y, Guo S, Chen J, et al. Improved SRGAN for remote sensing image super-resolution across locations and sensors. Remote Sens, 2020, 12: 1263
229 Zhang D, Shao J, Li X, et al. Remote sensing image super-resolution via mixed high-order attention network. IEEE Trans Geosci Remote Sens, 2020, 59: 5183–5196
230 Salvetti F, Mazzia V, Khaliq A, et al. Multi-image super resolution of remotely sensed images using residual attention deep neural networks. Remote Sens, 2020, 12: 2207
231 Zhang S, Yuan Q, Li J, et al. Scene-adaptive remote sensing image super-resolution using a multiscale attention network. IEEE Trans Geosci Remote Sens, 2020, 58: 4764–4779
232 Liu P, Wang M, Wang L, et al. Remote-sensing image denoising with multi-sourced information. IEEE J Sel Top Appl Earth Observations Remote Sens, 2019, 12: 660–674
233 Feng X, Zhang W, Su X, et al. Optical remote sensing image denoising and super-resolution reconstructing using optimized generative network in wavelet transform domain. Remote Sens, 2021, 13: 1858
234 Enomoto K, Sakurada K, Wang W, et al. Image translation between SAR and optical imagery with generative adversarial nets. In: Proceedings of IEEE International Geoscience and Remote Sensing Symposium, 2018. 1752–1755
235 Reyes M F, Auer S, Merkle N, et al. SAR-to-optical image translation based on conditional generative adversarial networks-optimization, opportunities and limits. Remote Sens, 2019, 11: 2067
236 Zhang Q, Liu X, Liu M, et al. Comparative analysis of edge information and polarization on SAR-to-optical translation based on conditional generative adversarial networks. Remote Sens, 2021, 13: 128
237 Ji S, Wang D, Luo M. Generative adversarial network-based full-space domain adaptation for land cover classification from multiple-source remote sensing images. IEEE Trans Geosci Remote Sens, 2020, 59: 3816–3828
238 Peng D, Guan H, Zang Y, et al. Full-level domain adaptation for building extraction in very-high-resolution optical remote-sensing images. IEEE Trans Geosci Remote Sens, 2022, 60: 1–17
239 Qu B, Li X, Tao D, et al. Deep semantic understanding of high resolution remote sensing image. In: Proceedings of 2016 International Conference on Computer, Information and Telecommunication Systems (CITS), 2016. 1–5
240 Wang B, Lu X, Zheng X, et al. Semantic descriptions of high-resolution remote sensing images. IEEE Geosci Remote Sens Lett, 2019, 16: 1274–1278
241 Lobry S, Murray J, Marcos D, et al. Visual question answering from remote sensing images. In: Proceedings of IEEE International Geoscience and Remote Sensing Symposium, 2019. 4951–4954
242 Lobry S, Marcos D, Murray J, et al. RSVQA: visual question answering for remote sensing data. IEEE Trans Geosci Remote Sens, 2020, 58: 8555–8566
243 Zheng X, Wang B, Du X, et al. Mutual attention inception network for remote sensing visual question answering. IEEE Trans Geosci Remote Sens, 2022, 60: 1–14
244 Lu X, Wang B, Zheng X. Sound active attention framework for remote sensing image captioning. IEEE Trans Geosci Remote Sens, 2019, 58: 1985–2000
245 Wu S, Zhang X, Wang X, et al. Scene attention mechanism for remote sensing image caption generation. In: Proceedings of 2020 International Joint Conference on Neural Networks (IJCNN), 2020. 1–7
246 Zhao R, Shi Z, Zou Z. High-resolution remote sensing image captioning based on structured attention. IEEE Trans Geosci Remote Sens, 2022, 60: 1–14
247 Huang W, Wang Q, Li X. Denoising-based multiscale feature fusion for remote sensing image captioning. IEEE Geosci Remote Sens Lett, 2020, 18: 436–440
248 Wang Q, Huang W, Zhang X, et al. Word-sentence framework for remote sensing image captioning. IEEE Trans Geosci Remote Sens, 2021, 59: 10532–10543
249 Yao Y, Doretto G. Boosting for transfer learning with multiple sources. In: Proceedings of 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2010. 1855–1862
250 Liu W, Qin R. A MultiKernel domain adaptation method for unsupervised transfer learning on cross-source and cross-region remote sensing data classification. IEEE Trans Geosci Remote Sens, 2020, 58: 4279–4289
251 Xu Z, Chen Z, Yi W, et al. Deep gradient prior network for DEM super-resolution: Transfer learning from image to DEM. ISPRS J Photogrammetry Remote Sens, 2019, 150: 80–90
252 Hu T, Huang X, Li J, et al. A novel co-training approach for urban land cover mapping with unclear Landsat time series imagery. Remote Sens Environ, 2018, 217: 144–157
253 Qiu C, Schmitt M, Mou L, et al. Feature importance analysis for local climate zone classification using a residual convolutional neural network with multi-source datasets. Remote Sens, 2018, 10: 1572
254 Rostami M, Kolouri S, Eaton E, et al. Sar image classification using few-shot cross-domain transfer learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, 2019
255 Ying Z, Xuan C, Zhai Y, et al. TAI-SARNET: deep transferred atrous-inception CNN for small samples SAR ATR. Sensors, 2020, 20: 1724
256 Shermeyer J, Hogan D, Brown J, et al. SpaceNet 6: multi-sensor all weather mapping dataset. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, 2020. 196–197
257 Montgomery J, Brisco B, Chasmer L, et al. SAR and lidar temporal data fusion approaches to boreal wetland ecosystem monitoring. Remote Sens, 2019, 11: 161
258 Chen H, Shi Z. A spatial-temporal attention-based method and a new dataset for remote sensing image change detection. Remote Sens, 2020, 12: 1662
259 Daudt R C, Saux B L, Boulch A, et al. Multitask learning for large-scale semantic change detection. Comput Vision Image Understanding, 2019, 187: 102783
260 Zheng B, Campbell J B, de Beurs K M. Remote sensing of crop residue cover using multi-temporal Landsat imagery. Remote Sens Environ, 2012, 117: 177–183
261 Garnot V S F, Landrieu L, Giordano S, et al. Satellite image time series classification with pixel-set encoders and temporal

self-attention. In: Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020
262 Wei J, Lee Z, Garcia R, et al. An assessment of Landsat-8 atmospheric correction schemes and remote sensing reflectance products in coral reefs and coastal turbid waters. Remote Sens Environ, 2018, 215: 18–32
263 Zhang W, Tang P, Zhao L. Remote sensing image scene classification using CNN-CapsNet. Remote Sens, 2019, 11: 494
264 Ren S, He K, Girshick R, et al. Faster R-CNN: towards real-time object detection with region proposal networks. IEEE Trans Pattern Anal Mach Intell, 2017, 39: 1137–1149
265 Yang X, Sun H, Fu K, et al. Automatic ship detection in remote sensing images from Google Earth of complex scenes based on multiscale rotation dense feature pyramid networks. Remote Sens, 2018, 10: 132
266 Tian Z, Shen C, Chen H, et al. FCOS: fully convolutional one-stage object detection. In: Proceedings of 2019 IEEE/CVF International Conference on Computer Vision, Seoul, 2019. 9626–9635
267 Long J, Shelhamer E, Darrell T. Fully convolutional networks for semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015. 3431–3440
268 Chen L C, Papandreou G, Kokkinos I, et al. DeepLab: semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs. IEEE Trans Pattern Anal Mach Intell, 2017, 40: 834–848
269 Mi L, Chen Z. Superpixel-enhanced deep neural forest for remote sensing image semantic segmentation. ISPRS J Photogrammetry Remote Sens, 2020, 159: 140–152
270 Yang Y, Newsam S. Bag-of-visual-words and spatial extensions for land-use classification. In: Proceedings of the 18th SIGSPATIAL International Conference on Advances in Geographic Information Systems, 2010. 270–279
271 Xia G S, Hu J, Hu F, et al. AID: a benchmark data set for performance evaluation of aerial scene classification. IEEE Trans Geosci Remote Sens, 2017, 55: 3965–3981
272 Zhou W, Newsam S, Li C, et al. PatternNet: a benchmark dataset for performance evaluation of remote sensing image retrieval. ISPRS J Photogrammetry Remote Sens, 2018, 145: 197–209
273 Emelyanova I V, McVicar T R, van Niel T G, et al. Assessing the accuracy of blending Landsat-MODIS surface reflectances in two landscapes with contrasting spatial and temporal dynamics: a framework for algorithm selection. Remote Sens Environ, 2013, 133: 193–209
274 Campos-Taberner M, Romero-Soriano A, Gatta C, et al. Processing of extremely high-resolution LiDAR and RGB data: outcome of the 2015 IEEE GRSS data fusion contest-part A: 2-D contest. IEEE J Sel Top Appl Earth Observations Remote Sens, 2016, 9: 5547–5559
275 Ching J, Mills G, Bechtel B, et al. WUDAPT: an urban weather, climate, and environmental modeling infrastructure for the anthropocene. Bull Am Meteorol Soc, 2018, 99: 1907–1924
276 Cadiou E, Mammez D, Dherbecourt J B, et al. Atmospheric boundary layer CO_2 remote sensing with a direct detection LIDAR instrument based on a widely tunable optical parametric source. Opt Lett, 2017, 42: 4044–4047
277 Zhang X, Wang Q, Chen S, et al. Multi-scale cropping mechanism for remote sensing image captioning. In: Proceedings of IEEE International Geoscience and Remote Sensing Symposium, 2019. 10039–10042
278 Chen B, Huang B, Xu B. Multi-source remotely sensed data fusion for improving land cover classification. ISPRS J Photogrammetry Remote Sens, 2017, 124: 27–39
279 Carrasco L, O'Neil A, Morton R, et al. Evaluating combinations of temporally aggregated Sentinel-1, Sentinel-2 and Landsat 8 for land cover mapping with Google Earth Engine. Remote Sens, 2019, 11: 288
280 Piramanayagam S, Saber E, Schwartzkopf W, et al. Supervised classification of multisensor remotely sensed images using a deep learning framework. Remote Sens, 2018, 10: 1429
281 Liu J, Gong M, Qin K, et al. A deep convolutional coupling network for change detection based on heterogeneous optical and radar images. IEEE Trans Neural Netw Learn Syst, 2016, 29: 545–559
282 Wu Y, Li J, Yuan Y, et al. Commonality autoencoder: learning common features for change detection from heterogeneous images. IEEE Trans Neural Netw Learn Syst, 2022, 33: 4257–4270
283 Zhu X, Cai F, Tian J, et al. Spatiotemporal fusion of multisource remote sensing data: literature survey, taxonomy, principles, applications, and future directions. Remote Sens, 2018, 10: 527
284 Li J, Li Y F, He L, et al. Spatio-temporal fusion for remote sensing data: an overview and new benchmark. Sci China Inf Sci, 2020, 63: 140301
285 Xu J, Zhu Y, Zhong R, et al. DeepCropMapping: a multi-temporal deep learning approach with improved spatial generalizability for dynamic corn and soybean mapping. Remote Sens Environ, 2020, 247: 111946
286 He C, Gao B, Huang Q, et al. Environmental degradation in the urban areas of China: evidence from multi-source remote sensing data. Remote Sens Environ, 2017, 193: 65–75
287 Hilker T, Wulder M A, Coops N C, et al. A new data fusion model for high spatial- and temporal-resolution mapping of forest disturbance based on Landsat and MODIS. Remote Sens Environ, 2009, 113: 1613–1627
288 Tran T V, de Beurs K M, Julian J P. Monitoring forest disturbances in Southeast Oklahoma using Landsat and MODIS images. Int J Appl Earth Observation GeoInf, 2016, 44: 42–52
289 Singh P, Komodakis N. Cloud-GAN: cloud removal for sentinel-2 imagery using a cyclic consistent generative adversarial networks. In: Proceedings of IEEE International Geoscience and Remote Sensing Symposium, 2018. 1772–1775
290 Zhang Q, Yuan Q, Zeng C, et al. Missing data reconstruction in remote sensing image with a unified spatial-temporal-spectral deep convolutional neural network. IEEE Trans Geosci Remote Sens, 2018, 56: 4274–4288
291 Huang H, Kuhn A, Michelini M, et al. 3D urban scene reconstruction and interpretation from multisensor imagery. In: Proceedings of Multimodal Scene Understanding, 2019. 307–340
292 Liu Y, Xue F, Huang H. UrbanScene3D: a large scale urban scene dataset and simulator. 2021. ArXiv:2107.04286