ISPRS Journal of Photogrammetry and Remote Sensing 186 (2022) 170–189

Contents lists available at ScienceDirect

ISPRS Journal of Photogrammetry and Remote Sensing

journal homepage: www.elsevier.com/locate/isprsjprs

DKDFN: Domain Knowledge-Guided deep collaborative fusion network for multimodal unitemporal remote sensing land cover classification

Yansheng Li [a,*], Yuhan Zhou [a,*], Yongjun Zhang [a,*], Liheng Zhong [b], Jian Wang [b], Jingdong Chen [b]

[a] School of Remote Sensing and Information Engineering, Wuhan University, China
[b] Ant Group, China

ARTICLE INFO

Keywords:
Land cover classification
Deep collaborative network
Domain knowledge incorporation
Multimodal unitemporal remote sensing

ABSTRACT

Land use and land cover maps provide fundamental information that has been used in different types of studies, ranging from public health to carbon cycling. However, the existing remote sensing image classification methods thus far suffer from the insufficient usage of multiple modalities, underconsideration of prior domain knowledge, and poor performance on minority classes. To alleviate these problems, we propose a novel domain knowledge-guided deep collaborative fusion network (DKDFN) with performance boosting for minority categories for land cover classification. More specifically, the DKDFN adopts a multihead encoder and a multibranch decoder structure. The architecture of the encoder probablizes sufficient mining of complementary information from multiple modalities, which are Sentinel-2, Sentinel-1, and SRTM Digital Elevation Data (SRTM) in our case. The multibranch decoder enables land cover classification in a multitask learning setup, performing semantic segmentation and reconstructing multimodal remote sensing indices, which are selected as representatives of domain knowledge. This design incorporates domain knowledge in an effective end-to-end manner. The training stage of our DKDFN is supervised by our proposed asymmetry loss function (ALF), which boosts performance on nearly all categories, especially the categories with a low frequency of occurrence. Ablation studies of the network suggest that our design logic is worth testing in any network with an encoder-decoder structure. The study is conducted in Hunan, China and is verified using a self-labeled multimodal unitemporal remote sensing image dataset. The comparative experiments between DKDFN and 6 state-of-the-art models (U-Net, SegNet, PSPNet, DeepLab, HRNet, MP-ResNet) testify to the superiority of our method and suggest its potential to be applied more widely to map land cover in other geographical areas given the availability of Sentinel-2, Sentinel-1, and SRTM data. The dataset can be downloaded by https://github.com/LauraChow/HunanMultimodalDataset .

1. Introduction

Human activities, as well as environmental changes, exert a profound impact on the distribution of the physical cover of the Earth's surface (Running, 2008). To acquire timely information on land cover and satisfy the demands of policy-makers and landscape planners, an efficient and accurate land cover classification algorithm is significant. Such an algorithm should be able to be applied to various spheres, ranging from socioeconomic to scientific, such as land resource management (Ardila et al., 2011; Ozdarici-Ok et al., 2015; Zhang and Kovacs, 2012), public health (Liang et al., 2010; Xu et al., 2004), climate change studies (Hibbard et al., 2010; Imaoka et al., 2010), and carbon cycling (Ganzeveld et al., 2010; Liu et al., 2011; Poulter et al., 2011). The rise of remote sensing technology probabilizes the automation of land cover mapping by means of providing ground surface observations over space and time, and the follow-up algorithms employing this information complete the mapping from remote sensing imagery to the pixelwise labels of land cover (Gong et al., 2013; Hurskainen et al., 2019; Jun et al., 2014). Given the extensive applications of land cover maps and the availability of remote sensing imagery, land cover mapping has attracted extensive research interest (Li et al., 2020a; Liu et al., 2021; Tong et al., 2020). Although numerous methods have been proposed, off-the-shelf land cover classification methods tend to insufficiently consider many factors, such as domain knowledge and the utilization of multimodalities. Hence, land cover mapping is still facing challenges and deserves further investigation.

* Corresponding authors.

https://doi.org/10.1016/j.isprsjprs.2022.02.013
Received 24 October 2021; Received in revised form 16 February 2022; Accepted 17 February 2022

The characterization of land cover, by its nature, can be regarded as a semantic segmentation task (Li et al., 2021a), to which either traditional machine learning or deep learning algorithms are applied. The most commonly used machine learning methods, such as support vector machines (SVMs) (Cortes and Vapnik, 1995), random forests (RFs) (Breiman, 2001), and XGBoost (Chen and Guestrin, 2016), are limited in feature extraction. As a result, first, handcrafted spectral and textual features need to be retrieved for the following feature classification process. Limited by the representation ability of handcrafted features, machine learning methods often fail to achieve high accuracy and robustness. Deep learning (LeCun et al., 2015; Li et al., 2021b), however, does not require great effort to apply to feature engineering. As a representative data-driven method, deep learning integrates feature extraction and feature classification in an end-to-end manner. Currently, deep learning-based semantic segmentation networks, such as U-Net (Ronneberger et al., 2015), DeepLab V3+ (Chen et al., 2018), and HRNet (Sun et al., 2019), have proven their versatility by achieving state-of-the-art performance in varying domains, including remote sensing research (Li et al., 2020b); and good performance in land cover mapping has been achieved (Calderón-Loor et al., 2021; Hurskainen et al., 2019; Nguyen et al., 2020). Nevertheless, there is still room for improvement since the characteristics of remote sensing data and tasks are under-considered and the traits of the class distribution have not received much attention.

The current deep learning-based remote sensing land cover techniques mainly suffer from the following three challenges: (1) First, incorporating domain knowledge is not trivial. As data-driven methods, deep learning models tend to exhibit unfavorable performance given insufficient labeled training data, which calls for the incorporation of prior domain knowledge that can be integrated into the learning process, posing constraints on or guiding the training; and thus leads to a trustworthy derived model (von Rueden et al., 2019). However, domain knowledge in remote sensing areas suffers from clumsy incorporation strategies (e.g., postprocessing) (Chamorro Martinez et al., 2021; Li et al., 2021c) or does not guarantee extensibility (Chamorro Martinez et al., 2021; El Hajj et al., 2009; Rees et al., 2003; Waldner et al., 2015; Zhou et al., 2021). (2) The second challenge refers to the utilization of multiple modalities. In order for deep learning models to make progress in the interpretation of land cover, they need to be able to interpret and reason about multimodal messages (Baltrušaitis et al., 2017). Currently, the wealth of remote sensing sensors and observation techniques (e.g., active and passive) results in data from different modalities. However, the multimodal data are either not considered in studies (Calderón-Loor et al., 2021; Gong et al., 2013; Li et al., 2020a; Liu et al., 2021; Nguyen et al., 2020; Sica et al., 2019) or the fusing strategy fails to fully mine complementary features across modalities (Amarsaikhan et al., 2010; Belenguer-Plomer et al., 2021; Van Beijma et al., 2014; Wan et al., 2019). (3) The third challenge is tackling imbalanced land cover types, which deteriorates the model performance. In the land cover classification task, it is not easy to figure out which land cover is the most important one; some land covers might account for only a small part of all land covers, but they still serve significant environmental functions. The accuracy of under-represented land covers are important to evaluate the overall model performance. The skewed data problem is a pervasive challenge and is ubiquitous in all learning paradigms, ranging from traditional machine learning to deep learning (Johnson and Khoshgoftaar, 2019; Ortigosa-Hernández et al., 2017; Prati et al., 2015). Commonly, this issue is solved by a reweighting scheme, which is prone to be affected by many factors (Kellenberger et al., 2018); or by introducing a novel loss function, such as Dice coefficient (Milletari et al., 2016). Since the commonly used loss functions (e.g., Dice coefficient) targeting imbalanced datasets are designed for binary classification, they require some transformation to fit in the multiclass problem.

To address the aforementioned problems concerning multimodality data fusion, incorporating prior knowledge, and the cooperative classification of imbalanced land cover categories, we propose a novel deep network called the domain knowledge-guided deep collaborative fusion network (DKDFN). The DKDFN aims to fully exploit the information of the three modalities of optical, SAR, and topography in conjunction and assimilates the generalizable domain knowledge sufficiently in an end-to-end fashion. In addition, the DKDFN is supervised by one new asymmetry loss function (ALF), which promotes the accuracy of nearly all categories, especially the minority categories. More specifically, the DKDFN adopts a multihead encoder and a multibranch decoder structure. By encoding data from various modalities separately and fusing the extracted features with our proposed deep collaborative fusion scheme, the multihead encoder mines the complementary characteristics of three modalities: Sentinel-2, Sentinel-1, and SRTM Digital Elevation Data (SRTM). The multibranch decoder cooperates with the multihead encoder by assimilating a hierarchy of information extracted by the encoder. It also accomplishes land cover classification in a multitask learning setup, completing semantic segmentation with the ALF and the reconstruction of domain knowledge at the same time. We select multimodal remote sensing indices, which have been proposed by remote sensing experts and testified extensively, as the representatives of domain knowledge in this study. This design of our multibranch decoder ensures a highly generalizable and effective solution for incorporating domain knowledge. In the future, more types of domain knowledge can be tested in this framework. In addition, the ALF poses constraints on the semantic segmentation task, both globally and locally, due to its asymmetric structure. The ALF works by supervising the classification of all classes using a robust loss function, serving as a global constraint, while using another loss function designed for imbalanced datasets to supervise rarely occurring categories, serving as local constraints. In summation, complementary characteristics from different modalities can be extracted sufficiently by our multihead encoder, highly generalizable domain knowledge can be assimilated using a a multibranch decoder, and minority categories receive more attention when leveraging the ALF.

To fully verify the effectiveness of the presented DKDFN, our study is conducted in Hunan, a representative province in China. Hunan has an area of approximately 210,000 km^2, and its environmental system is diverse and complex, with three ecoregions lying in it. The experimental results reflect improvements for nearly all classes, with the accuracies of the minority classes increasing by a large margin. Although our study is conducted in Hunan, its implications in terms of the suitability of Sentinel-2, Sentinel-1 observations, and SRTM and the derived improvement level are applicable to other multimodal semantic segmentation tasks in remote sensing areas regardless of the classes of interest or the specific region in the world. To complement the multimodal dataset in the remote sensing area, we also open-source a multimodal dataset in Hunan, China for 2017. The dataset delineates 7 land cover types: cropland, forest, grassland, wetland, water, unused land, and built-up area. All data are preprocessed as 256 by 256 Sentinel-2, Sentinel-1, and SRTM image blocks; and the corresponding reference semantic labels at a 10 m spatial resolution are included. Our dataset covers 32,768,000 pixels.

The main contributions of this paper can be summarized as follows:

a. This paper proposes a new domain knowledge-guided deep collaborative fusion network (i.e., the DKDFN), which collaboratively fuses multimodal data and assimilates highly generalizable domain knowledge (e.g., remote sensing indices) at the same time. Additionally, the idea of modality fusion and knowledge incorporation can be easily extended to any network with an encoder-decoder structure.
b. This paper proposes a new loss function with an asymmetric structure (i.e., the ALF). Different loss functions are used to supervise the channels belonging to minorty and nonminority classes. This design can increase the accuracies of nearly all categories, especially classes with a low frequency of occurrence. This loss is highly flexible and worth testing for any imbalanced dataset.

c. Finally, this paper collects and releases a new multimodal remote sensing image dataset for land cover classification (https://github.com/LauraChow/HunanMultimodalDataset). Specifically, the dataset contains image blocks with pixel-level labels for cropland, forest, grassland, wetland, water, unused land, and built-up area, covering 32,768,000 pixels in Hunan, China.

The remainder of this paper is organized as follows. The related work is detailed in Section 2. Section 3 specifically displays our multimodal dataset. Section 4 introduces the proposed DKDFN. Section 5 reports the experimental results. Finally, Section 6 concludes the paper.

2. Related work

In this section, we briefly review the most relevant works in terms of the aforementioned aspects, including multimodal data classification with deep learning, prior knowledge-based deep learning, and remote sensing image classification with unbalanced categories.

2.1. Multimodal data classification with deep learning

In recent years, multimodal remotely sensed imagery has been generated at an unprecedented rate (e.g., Landsat, MODIS, Sentinel-1, and Sentinel-2 imagery), which benefits land cover mapping by providing affluent spatial and spectral information. A key aspect toward the goal of classification is how to fully utilize the wealth of multimodal data. However, some researchers only consider single modal data (Calderón-Loor et al., 2021; Gong et al., 2013; Li et al., 2020a; Liu et al., 2021; Matikainen et al., 2020; Nguyen et al., 2020; Pan et al., 2020; Sica et al., 2019), which has some limitations since each modality has its own restrictions (e.g., the persistent cloud coverage of optical imagery and the speckle noise of synthetic aperture radar). In many cases, it is true that single modality utilization is able to achieve high accuracy (Phiri et al., 2020; Phiri and Morgenroth, 2017). For example, as reviewed by Phiri et al., most of the Sentinel-2-based methods are able to achieve land cover classification accuracy over 80%. Nevertheless, increasing in accuracy is also reported when Sentinel-2 data is integrated with other modalities. Thus, a single modality solution might benefit from other complementary and meaningful modalities. Studies concerning multimodal fusion have been exploited, among which the fusion of optical and synthetic aperture radar (SAR) imageries is the most frequently discussed combination of modalities (Ghorbanian et al., 2020; Ienco et al., 2019; Liu et al., 2019; Sukawattanavijit et al., 2017; Symeonakis et al., 2018). For instance, the overall accuracy of winter land use using the combination of Sentinel-2 and Sentinel-1 outweighs that of their Sentinel-2 only and Sentinel-1 only counterparts (Denize et al., 2018). However, topographic information, which can be seen as an extra modality and has been reported to greatly enhance land cover separation (Buchner et al., 2020; Hurskainen et al., 2019), seldom cooperates with optical and SAR data. Putting aside the incomplete combination of modalities, the fusion strategy of multimodal data also deserves further consideration. Previous works tend to concatenate features from different modalities at the beginning, regardless of whether these features have undergone a transformation for integration (e.g., the Brovey transform, wavelet-based fusion, the Elhers fusion, and PCA) or a feature extraction process (e.g., spectral, textual, and topographical feature generation); and feed them directly into the classifier (Amarsaikhan et al., 2010; Belenguer-Plomer et al., 2021; Van Beijma et al., 2014; Wan et al., 2019). This type of tactic, together with fusing the results from each individual modality (Bigdeli and Pahlavani, 2016; Shao et al., 2016), can suppress either the intra- or intermodality representation from being efficiently modeled, which has been proven by researchers in the computer science community (Hazirbas et al., 2016; Jiang et al., 2018; Zadeh et al., 2017; Zhang et al., 2020). In other words, the more appropriate strategies of multimodal fusion are methods that allow fusion to occur on multiple layers of a deep model.

2.2. Prior knowledge-based deep learning

The data-driven nature of deep learning methods limits their performance given insufficient labeled training data, which is a common case in remote sensing areas. When sufficient training data are unavailable, prior domain knowledge can be integrated into the learning process, posing constraints on or guiding the training and leading to a trustworthy derived model (von Rueden et al., 2019). The recent growth of research activities in knowledge incorporation has testified to the effectiveness of data- and knowledge-driven combinations. However, there are still some issues in the knowledge-integrated methods in the remote sensing field. For example, the workflow of some knowledge-based methods is not succinct. Two-stage learning (Li et al., 2022) and postprocessing (Chamorro Martinez et al., 2021) are considered in research; however, these approaches complicate the learning process, and the final performance highly relies on the results of the last stage. Moreover, in some studies, the knowledge used does not guarantee generalizability, which means that the methods cannot be easily extended to other areas or globally. Prior expert knowledge of particular sites enhances performance (Chamorro Martinez et al., 2021; El Hajj et al., 2009; Rees et al., 2003; Waldner et al., 2015; Zhou et al., 2021), but domain expert knowledge is unavailable in many regions. Some knowledge is computationally extensive to retrieve (Cui et al., 2021), thus restricting the applications of these methods in case studies. Some researchers selected released maps (e.g., GlobeLand30) in their studies (Li et al., 2020a; Lin et al., 2019), but noisy labels may introduce instability in the learning process. To achieve an efficient and extensible knowledge-based solution, generalizable geographical knowledge needs to be integrated and assist the training process in an end-to-end manner.

2.3. Remote sensing image classification with unbalanced categories

Skewed class problems commonly occur in classification tasks. Due to the difference in number between classes, the algorithms tend to become biased toward the majority values present and do not perform well on the minority values. Oversampling of classes with low occurrence rates (Buda et al., 2018) or undersampling majority classes can be a possible solution, but this approach is not appropriate for multiclass semantic segmentation as minority classes and majority classes tend to be mixed up in a sample and it is not easy to oversample or undersample only part of it. A class reweighting scheme may help to ease the data imbalance problem by disincentivizing ignorance to rare classes, but the exact weights of categories of interests depend on many factors (e.g., data value range, the amount of data, and the training model) and vary between problems to (Kellenberger et al., 2018). Some novel loss functions, such as the Dice coefficient (Milletari et al., 2016), have been proposed to handle the heavy imbalance between the foreground and background and are reported to be superior to sample reweighting. Since the Dice coefficient is originally designed for the binary classification problem and has an unstable gradient, it needs to be adapted to fit this multiclass semantic segmentation task.

3. Study area and data

3.1. Study area

The study is conducted in Hunan, China, as shown in Fig. 1. Hunan is situated between 108° 47′–114° 16′ east longitude and 24° 37′–30° 08′ north latitude with an area of approximately 210,000 km^2. Located in the central part of the Chinese mainland, Hunan borders the divisions of other provinces from southern China, western China, and eastern China, making it a geographically representative province. Diverse landforms and low hills with crisscrossing mountains and valleys characterize Hunan's geographical appearance. Mountains and hills occupy more than 80% and plains occupy less than 20% of the province. According to the Terrestrial Ecoregions of the World (TEOW) (Olson et al., 2001),

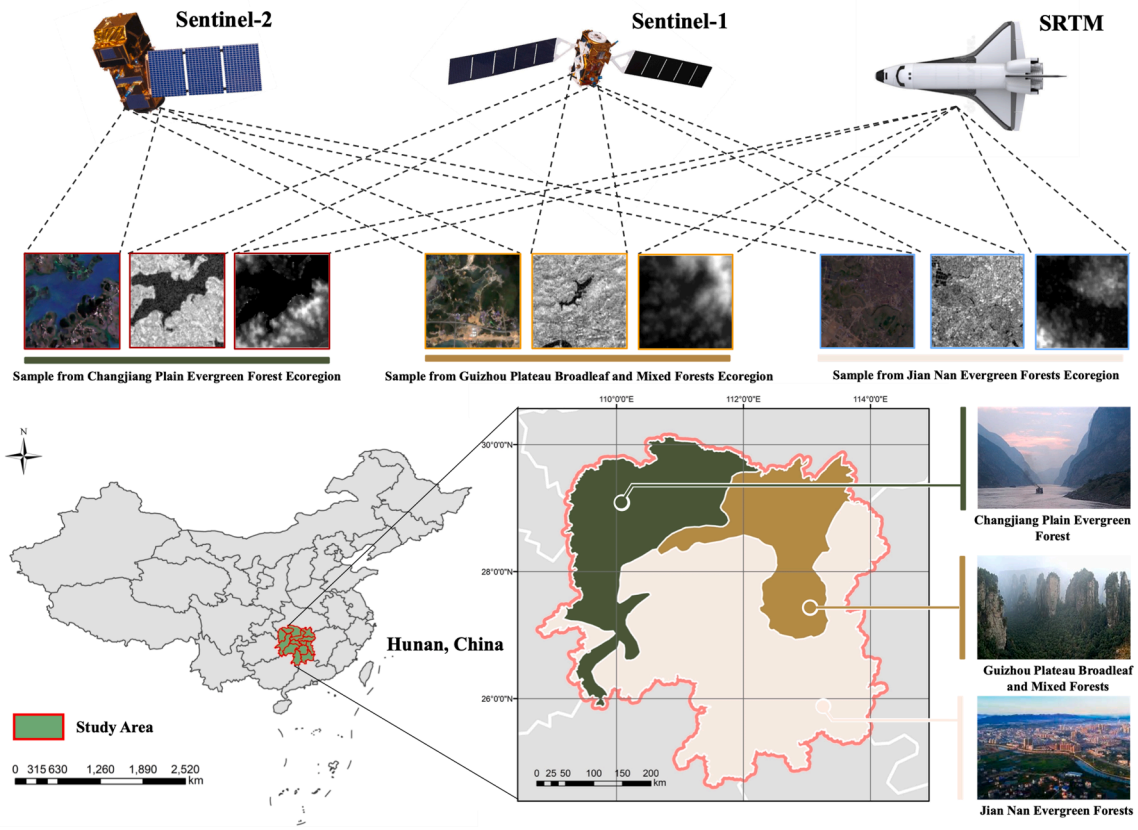

Fig. 1. The geolocation of our study area.

three ecoregions lie in Hunan Province. All of these results illustrate the complex and diverse environmental system of the study area (Fig. 1).

3.2. Data

We build a multimodal dataset for 2017 for the landcover classification task in Hunan. Sentinel-2, Sentinel-1 and SRTM composites are created as the input images. They are selected for their complementary characteristics. The richness of the spectral bands of Sentinel-2 suffices as input information; however, as optical imagery, Sentinel-2 is prone to be affected by weather. Sentinel-1, in contrast, is capable of performing well in all weather conditions thanks to its physical backscattering characteristics (Nghiem et al., 2009). However, shadow and topographic effects limit its efficacy. As a result, SRTM is considered to provide extra topographic information. We expect a triplet of data sources to be instrumental to performance gain in our land cover classification task.

Although sufficient information can be extracted from these 3 modalities, data preprocessing is still pivotal for better data quality.

As clouds and cloud shadows may have negative effects on multispectral images, Sentinel-2 imagery, the primary input data, undergoes careful preprocessing to preclude these harmful regions. For all Sentinel-2 images in 2017, the FMASK algorithm (Zhu et al., 2015) is applied for clouds, cloud shadows, and snow masking. Then, the cloud-free composite of Sentinel-2 is created by temporally aggregating these cloud-free images using the algorithm proposed by (Schmitt et al., 2019). We select 10 bands from Sentinel-2. B1 (Coastal aerosol), B9 (Water vapor), and B10 (SWIR - Cirrus) are excluded for their low correlation to the land cover classification task.

The Sentinel-1 data are preprocessed by the Sentinel-1 Toolbox in the Google Earth Engine (GEE); and the preprocessing includes thermal noise removal, radiometric calibration, terrain correction and decibel conversion. The data are in the dual-polarization (VV and VH). To reduce speckle noise, temporal mean mosaicking is applied to obtain a Sentinel-1 composite (Quin et al., 2014).

The SRTM dataset in our study is composed of the original elevation layer plus a slope layer derived from the DEM data. Although SRTM does not temporally align with Sentinel-2 and Sentinel-1, topographic information can be considered constant to some extent (Lin et al., 2020; Rennó et al., 2008; Sesnie et al., 2008).

All data are resampled to a resolution of 10 m, the finest spatial resolution of the three modalities. We utilize the default resampling strategy nearest neighbor in GEE. The sensors and their corresponding bands in use are listed in Table 1.

We randomly select 256 by 256 image tiles as our experimental data, and the semantic references of the image tiles and manually labeled by domain experts (Fig. 2) primarily using the Sentinel-2 mosaic. For areas with ambiguity, we refer to high-resolution imagery from Google Earth for visual interpretation. The study considers 7 land cover types: cropland, forest, grassland, wetland, water, unused land, and built-up area. The data distribution of these land cover types is imbalanced (Table 2)

Table 1
Sources of the Sentinel-2, Sentinel-1, and SRTM data.

Data Type	Product	Bands	Spatial Resolution	Available time
Sentinel-2	Sentinel-2 MSI	B2, B3, B4, B5, B6, B7, B8, B8A, B11, and B12	10 m and 20 *m*	Jan. 1, 2017–Dec. 31, 2017
Sentinel-1	Sentinel-1 SAR	VV and VH	10 *m*	Jan. 1, 2017–Dec. 31, 2017
SRTM	SRTM Digital Elevation Data Version 4	Elevation and slope	30 *m*	Feb. 11, 2000–Feb. 22, 2000

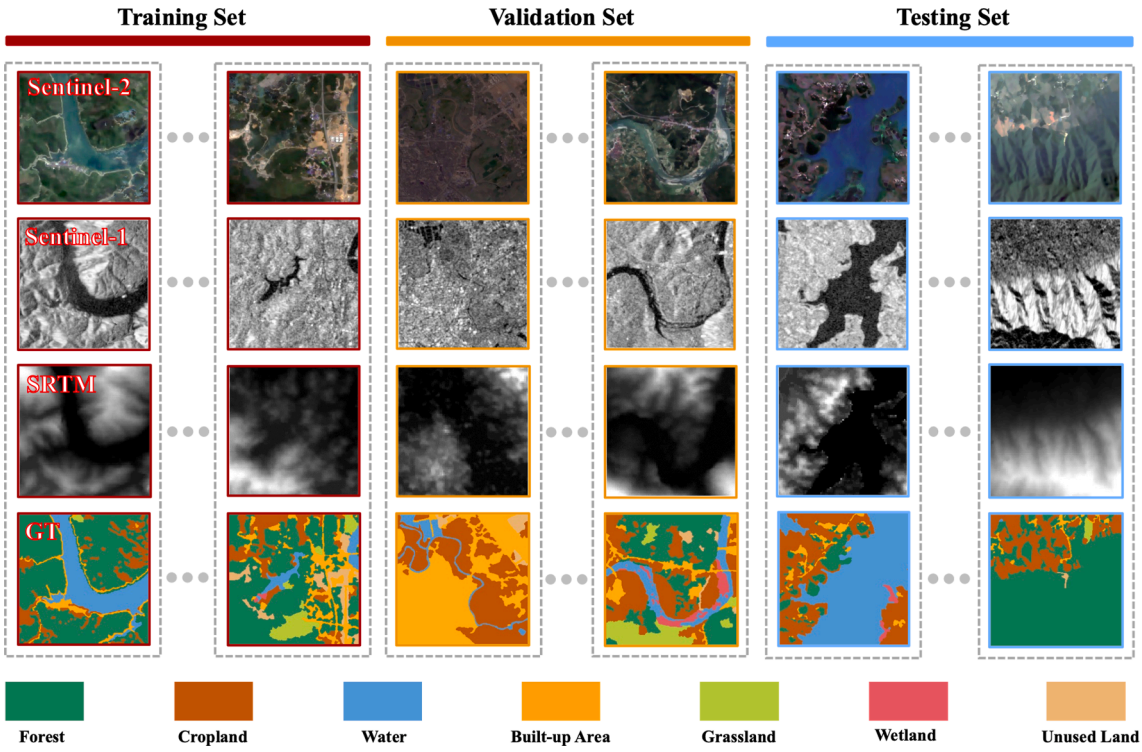

Fig. 2. Overview of our multimodal dataset.

Table 2
Land cover class distribution.

	Forest	Cropland	Water	Built-up Area	Grassland	Wetland	Unused Land
Percentage	42.37%	23.34%	13.35%	10.14%	7.35%	1.89%	1.56%

with grassland, wetland, and unused land accounting for a particularly small part (10.80% in total), showing the difficulty of this land cover classification task. Our dataset covers 32,768,000 pixels and the ratio for the training, validation, and testing tiles is 8:1:1. Training data are used for model learning while validation data are used for model selection. Finally, the model achieving the best accuracy on the validation set is successively employed to perform the land cover classification on the test set. There exists no overlap between the training set, validation set, as well as testing set.

To avoid overfitting and improve the performance of deep neural

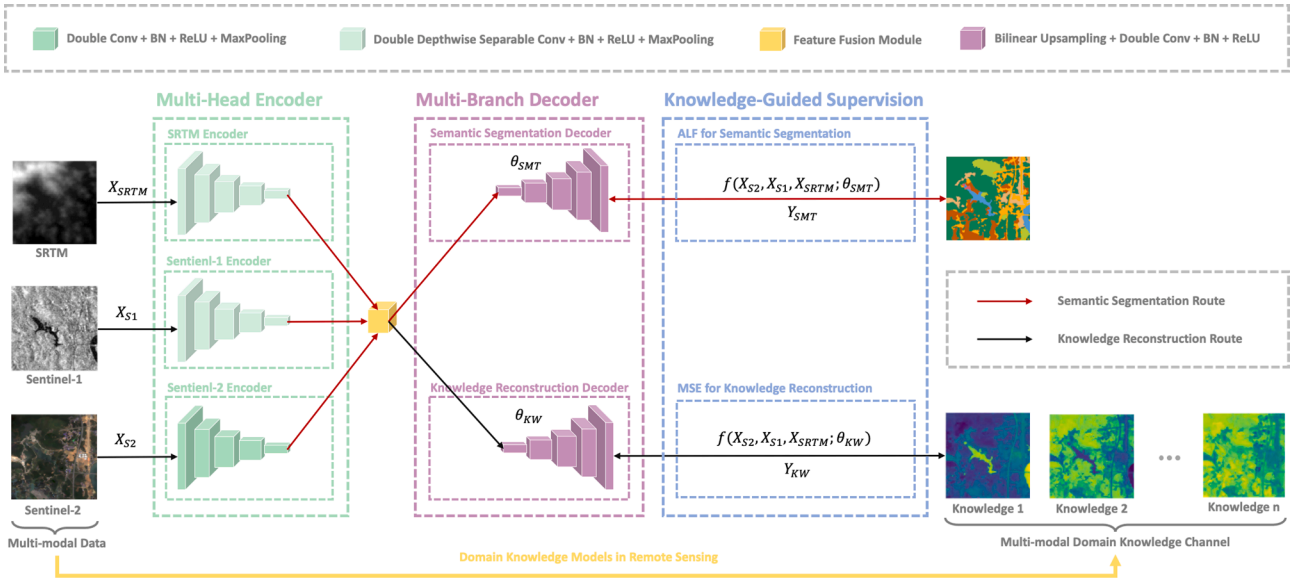

Fig. 3. The overall architecture of our proposed DKDFN. DKDFN is composed of multi-head encoder (greenish part), multi-branch decoder (purple part), and is supervised in a knowledge-guided manner (blue part).

networks, the original training set is augmented using various data augmentation methods including horizontal flipping, vertical flipping, transposition, blurring, random gamma correction, grid distortion, random sized cropping, and shift scale rotation. No data augmentation is employed on the validation set and testing set to ensure a reliable assessment of experimental results.

4. Methodology

In this section, we present the details of our proposed DKDFN. To effectively fuse multimodal data and assimilate domain knowledge, this paper presents a novel deep network whose architecture is shown in Section 4.1. In addition, the corresponding optimization loss function used to train the network is introduced in Section 4.2.

4.1. Dkdfn

The architecture of our proposed DKDFN is illustrated in Fig. 3. The proposed architecture comprises two well-designed parts: a multihead encoder and a multibranch decoder. The multihead encoder aims to collaboratively fuse multimodal data, and the multibranch decoder adaptively assimilates domain knowledge using a multitask learning mechanism.

As shown in Fig. 3, the multihead encoder is for the full absorption of multimodal data. It assesses three modalities, Sentinel-2, Sentinel-1, and SRTM, with the help of three individual encoders and extracts their corresponding feature hierarchies separately. The multilevel features of each modality are then fused by a feature fusion module, so features from all modalities are combined collaboratively to provide information for the decoder. The design of our multihead encoder is able to mine the complementary features from different modalities and fuse them synergistically, thus enabling better classification accuracy.

The multibranch decoder is designed for domain knowledge-guided semantic segmentation, and it consists of two decoding branches addressing two tasks: one branch is for land cover semantic segmentation, and the other branch is for the reconstruction of domain knowledge. In this study, empirical indices from different modalities (e.g., NDVI) are deemed as the approximation of various domain knowledge sources. Benefiting from the flexible framework, one can easily explore more types of domain knowledge (e.g., artificial visual features and quantitative inversion products) in future work. The layers of each decoder jointly fuse features from three modalities. The multibranch decoder gradually upscales the feature maps to the original resolution and finally localizes land cover types of interest in a domain knowledge-guided manner. The design of our multibranch decoder incorporates domain knowledge in an effective way. Knowledge from multiple modalites guides the learning process of the DKDFN, leading to performance improvements.

4.1.1. Multihead encoder for collaboratively fusing multimodal data

The multimodal message interpretation and reasoning ability assists deep learning models in progressing in the classification of land cover. To fully exploit the data of the three modalities, we design a multihead encoder. It consists of three encoders that share the same structure but different convolution types to satisfy different modal characteristics. All three encoders are similar to the original U-Net encoder (Ronneberger et al., 2015). These encoders consist of two connected 3x3 convolutions with a padding of 1. Each convolution layer is followed by a batch normalization layer to reduce internal covariance shifting (Ioffe and Szegedy, 2015), a rectified linear unit (ReLU) (Glorot et al., 2011) to map layer output to a nonlinear space, and a 2x2 max pooling operation with a stride of 2 for downsampling. We double the number of feature channels after performing each downsampling, except for the last layer of each encoder. Although they have the same architecture, different convolution operations are used across modalities. The normal convolution is implemented in the Sentinel-2 encoder while the depthwise separable convolution (Chollet, 2017) is used in the Sentinel-1 and the SRTM encoders. We adopt this design since 10-band Sentinel-2 data may require a relatively complicated architecture while lightweight branches may be appropriate for 2-band Sentinel-1 and 2-band SRTM data to avoid overfitting.

Through the encoding stage, feature hierarchies of the three modalities are extracted. Feature maps belonging to the same level but different modalities need to go through a feature fusion module to serve as the input of the symmetric layer in the multibranch decoder. The feature fusion module takes the feature maps of the three modalities as inputs. Since these features come from different modalities, we concatenate these features and reduce the number of channels by two-thirds using a 1x1 convolution rather than summing up the features directly. Batch normalization and ReLU are performed afterward, balancing the scale of fused features and adding nonlinearity.

The multihead encoder is able to mine the complementary characteristics across modalities due to each well-designed individual encoder, which can extract modality-specific features separately. Features from different modalities are treated equally by the feature fusion module, avoiding modality bias. In this way, sufficient multimodal extraction and fusion are ensured.

4.1.2. Multibranch decoder for adaptively assimilating domain knowledge

Although deep learning has achieved great success in computer vision, it still suffers from some extent of performance degradation after being transferred to the remote sensing field. Incorporating domain knowledge seems to be a promising strategy to guide the training process to achieve a more trustworthy model. To pursue an efficient and extensible knowledge-based solution, generalizable geographical knowledge should be integrated, and the knowledge should assist the training process in an end-to-end manner. To this end, we design our multibranch decoder, which completes land cover semantic segmentation and domain knowledge reconstruction at the same time. This design efficiently incorporates end-to-end domain knowledge. In our case, multimodal remote sensing indices (RSIs) are selected as domain knowledge. They are highly generalizable and can be generated from related data bands.

More specifically, our proposed multibranch decoder is composed of two individual decoders: one for semantic segmentation and the other for domain knowledge reconstruction. These two decoders share the same network architecture, which is symmetric to the structure of each individual encoder. Upsampling is performed with bilinear interpolation. The decoders both take the fused multimodal features from our multihead encoder hierarchically as inputs, but they differ from each other in their output. Densely labeled land cover maps are used to supervise the semantic segmentation process, and task-specific multimodal RSIs are used to supervise the domain knowledge reconstruction decoder. Our multibranch decoder achieves domain knowledge-guided semantic segmentation by completing two tasks together. One process is the semantic segmentation process (Eq. (1)).

$$R\_SMT = f_{SMT}(X_{S2}, X_{S1}, X_{SRTM}; \theta_{SMT}) \tag{1}$$

where R_SMT denotes the output of the semantic segmentation process, which is supervised by our semantic label; θ_{SMT} denotes the parameter weights of the multihead encoder and semantic segmentation decoder; and $f_{SMT}(.;.)$ denotes the semantic segmentation mapping function. The other process is the domain knowledge reconstruction process (Eq. (2)).

$$R\_KW = f_{KW}(X_{S2}, X_{S1}, X_{SRTM}; \theta_{KW}) \tag{2}$$

where R_KW denotes the output of the knowledge reconstruction decoder, which is supervised by our task-specific RSIs; θ_{KW} denotes the parameter weights of the multihead encoder and knowledge reconstruction decoder; and $f_{KW}(.;.)$ denotes the domain knowledge reconstruction mapping function.

To achieve better performance, we carefully select our task-specific

multimodal RSIs. Seven RSIs (Table 3), four from Sentinel-2, two from Sentinel-1, and one from SRTM, are chosen for their reported effectiveness in discriminating the land cover categories of interest. These indices capitalize on the different benefits of bands and are helpful knowledge in promoting accuracies. Furthermore, since these indices come from different modalities, they are able to pick up pixelwise spectral, SAR, and topographic patterns. This information can provide hints to guide our DKDFN during back propogation.

The multibranch decoder is able to efficiently assimilate domain knowledge because it performs land cover classification while also reconstructing helpful domain knowledge, which can reduce the number of candidate functions mapped from the input to the output and result in a more reliable model. The domain knowledge we utilize does not demand expert knowledge and can be easily extended to other research as long as the input bands are accessible. It is also worth noting that although we consider multimodal RSIs in our case, the architecture of our decoder is flexible to reconstruct other types of domain knowledge, such as pixelwise texture features.

4.2. Multitask learning for optimizing the DKDFN

We design the land cover classification task using a multitask learning setup with domain knowledge serving as guidance for the semantic segmentation of land cover types. In the training stage, our proposed DKDFN, which is optimized in a supervised fashion by the ALF and the mean squared error (MSE), predicts the posterior probability and reconstructs the RSIs of each pixel (Eq. (3)).

$$\begin{aligned}\mathscr{L}_{MTL} &= \mathscr{L}_{SMT}(R\_SMT, y\_SMT) + \alpha\bullet\mathscr{L}_{KW}(R\_KW, y\_KW)\\ &= \mathscr{L}_{SMT}(f_{SMT}(X_{S2}, X_{S1}, X_{SRTM}; \theta_{SMT}), y\_SMT)\\ &\quad + \alpha\bullet\mathscr{L}_{KW}(f_{KW}(X_{S2}, X_{S1}, X_{SRTM}; \theta_{KW}), y\_KW)\end{aligned} \tag{3}$$

where $\mathscr{L}_{MTL}$ denotes the total loss of our study. $\mathscr{L}_{MTL}$ is the summation of $\mathscr{L}_{SMT}$, the loss from semantic segmentation; and $\mathscr{L}_{KW}$, the loss from knowledge reconstruction. y_SMT is the semantic label, and y_KW is the multimodal RSI calculated by our input multimodal data. The formula of each channel in y_KW is listed in Table 3. α is the weighting factor adjusting the contribution of $\mathscr{L}_{KW}$.

Table 3
The RSIs of our study.

Domain Knowledge-based Remote Sensing Index	Data Source	Formula
Normalized Difference Built-up Index (NDBI)	Sentinel-2	$(B_8 - B_{12})/(B_8 + B_{12})$
Normalized Difference Vegetation Index (NDVI)	Sentinel-2	$(B_8 - B_4)/(B_8 + B_4)$
Normalized Difference Water Index (NDWI)	Sentinel-2	$(B_3 - B_8)/(B_8 + B_3)$
Bare Surface Index (NDBSI)	Sentinel-2	$(B_4 - B_2)/(B_4 + B_2)$
Normalized Polarization (PoL)	Sentinel-1	$(VH - VV)/(VH + VV)$
Radar Vegetation Index (RVI)	Sentinel-1	$(4*VH)/(VV + VH)$
Terrain Ruggedness Index (TRI)	SRTM	$\sum_{i-1}^{8}\lvert e - e_i\rvert/8$

4.2.1. Asymmetric loss function for semantic segmentation

For performance boosting on minority categories, we propose an ALF. Our ALF, similar to its name, has an asymmetric structure and is able to supervise the semantic segmentation process depending on the situation. In particular, all output channels of each pixel are supervised by the generalized cross entropy (GCE) (Zhang and Sabuncu, 2018), a robust loss function for semantic segmentation. For output channels corresponding to minority classes, the Dice loss (Milletari et al., 2016) is employed as an additional constraint. In other words, the minority channels are supervised by a hybrid of the GCE and Dice loss while other channels are supervised by the GCE only. Eq. (4) presents the general form of our ALF.

$$\begin{aligned}\mathscr{L}_{SMT} &= \begin{cases} GCE(R\_SMT^c, y\_SMT^c) + \beta\bullet Dice(R\_SMT^c, y\_SMT^c), & ifc \in C_{MNRT} \\ GCE(R\_SMT^c, y\_SMT^c), & ifc \in C_{MJRT} \end{cases}\\ &= \begin{cases} \sum_{i=1}^{N} y\_SMT_i^c\left(\frac{1-(R\_SMT_i^c)^q}{q}\right) + \beta\bullet\left(1 - \frac{\sum_{i=1}^{N} R\_SMT_i^c y\_SMT_i^c}{\sum_{i=1}^{N}(R\_SMT_i^c)^2 + \sum_{i=1}^{N}(y\_SMT_i^c)^2}\right), & ifc \in C_{MNRT} \\ \sum_{i=1}^{N} y\_SMT_i^c\left(\frac{1-(R\_SMT_i^c)^q}{q}\right), & ifc \in C_{MJRT} \end{cases}\end{aligned} \tag{4}$$

where R_SMT^c and y_SMT^c denote the cth channel of the model prediction and target, respectively. N is the number of pixels, and $R\_SMT_i^c$ and $y\_SMT_i^c$ represent the prediction and target at the cth channel of pixel i, respectively. C_{MNRT} is the set of minority land cover categories, and C_{MJRT} is the set of majority categories. q is the hyperparameter of the GCE and is empirically set to 0.9. β is the weighting factor that adjusts the contribution of the Dice loss for minority channels.

The detailed calculation process of our ALF is presented in Fig. 4. The forward propagation of the DKDFN presents each pixel using logits, with each channel of the output denoted by z_1 to z_C, where C denotes the number of output channels, which equals the number of classes of interest. Different actions are taken for different channels. First, global constraints are used to supervise all output channels. Specifically, the softmax is used to activate the logits of all channels, resulting in the probabilities of each class. These probabilities are used to calculate the GCE, which poses global constraints on all classes of interest. Then, local constraints, which are constraints for minorities, are applied. The outputs of channels corresponding to minority classes are denoted by z_{mnrt1} to z_{mnrt3} in Fig. 4. The minority classes in our case are grassland, wetland, and unused land. The output logits of these minoirity classes are additionally activated by a sigmoid function. Consequently, the probabilistic output of each minority channel is obtained and becomes the input of the Dice loss together with the target label. This process poses additional local constraints on all minority classes. The loss of each minority channel is the summation of the GCE and Dice loss while the loss of other channels is the GCE only.

It is worth noting that our ALF is identical to accomplishing two types of missions jointly. One mission is global optimization over all categories by minimizing the gap between the probability distribution of the prediction and target. The other mission is local optimization, which

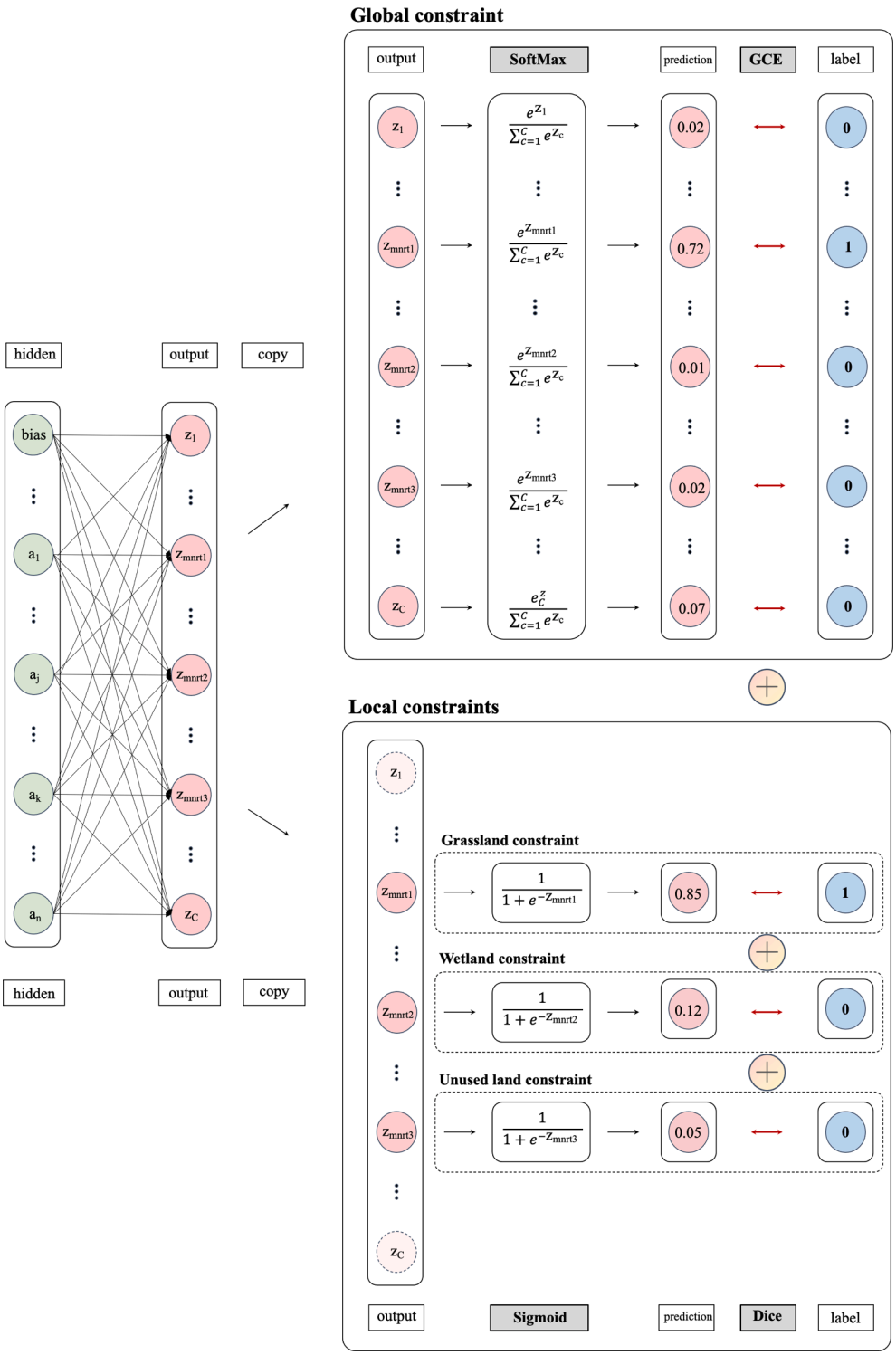

Fig. 4. Calculation process of the ALF. ALF works by posing a normal supervision on all channesls, which can be seen as global constraint (upper part), and also introducing additional constraints on minorities, which can be seen as local constraints (lower part).

is three separate optimizations over minority classes by treating each individual minority class as the foreground and the other classes as the background.

We claim that our ALF is able to boost the performance on minority classes due to its asymmetric structure. The GCE ensures good performance over all categories, and the extra Dice loss targeting each minority takes advantage of the design logic of the Dice loss. In other words, our ALF is capable of transferring the advantage of the Dice loss, which is designed for binary classification problems, to multiclass scenarios.

4.2.2. *MSE loss function for domain knowledge reconstruction*

We adopt the MSE as our loss function for the domain knowledge reconstruction decoder. The loss function is presented in Eq. (5).

$$\mathscr{L}_{KW} = \frac{1}{NC_{KW}} \sum_{i=1}^{N} \sum_{j=1}^{C_{KW}} \left(R\_KW_i^j - y\_KW_i^j\right)^2 \tag{5}$$

where C_{KW} denotes the number of empirical knowledge channels, which is set to 7 in our study. $R\_KW_i^j$ and $y\_KW_i^j$ are the output and the knowledge value of pixel i at the jth channel, respectively.

5. Experimental results and discussion

5.1. Experimental setup and evaluation measures

5.1.1. Experimental setup

In this section, we elaborate on our implementation protocol in detail. The training details are depicted as follows. We train our network for 100 epochs. Minibatch stochastic gradient descent (SGD) with a momentum of 0.9 and a weight decay of 5e − 4 is employed to optimize the objective function with respect to the weights at all network layers. The "poly" learning rate strategy in which the initial rate is set to 0.01 and then multiplied by $(1 - \frac{iter}{\max\_iter})^{power}$ after each iteration with a power of 0.9 is applied. In terms of the weights of the ALF and knowledge reconstruction loss, the best hyperparameter values are chosen using grid search based on the model performance on the validation set, which is 1 for the ALF and 0.75 for knowledge reconstruction.

All approaches, including our proposed approach and other baselines, are implemented using the PyTorch framework and conducted on a Dell station with 8 Intel Core i7-9700 k processors, 32 GB of RAM, and an NVIDIA GeForce RTX 3090TI.

5.1.2. Evaluation measures

The performance of our proposed DKDFN is comprehensively analyzed. We compute classwise measures, which are the intersection-over-union (IoU) (Eq. (6)), user's accuracy (UA) (Eq. (7)), producer's accuracy (PA) (Eq. (8)), F-score (Eq. (9)), and overall accuracy (OA).

$$\mathrm{IoU} = \frac{|\mathrm{TP}|}{|\mathrm{TP} + \mathrm{FN} + \mathrm{FP}|} \tag{6}$$

$$\mathrm{User's Accuracy(UA)} = \frac{TP}{TP + FP} \tag{7}$$

$$\mathrm{Producer's Accuracy(PA)} = \frac{TP}{TP + FN} \tag{8}$$

$$\mathrm{F-Score} = \frac{2 \bullet PA \bullet UA}{PA + UA} \tag{9}$$

where TP, FP, and FN represent the numbers of pixels that are true positives, false-positives, and false negatives for each class, respectively. In addition, the average of IoU, UA, PA, F-Score are also taken into account to give a holistic understanding of the model performance. Each land cover are assigned with the same weight as we think they are equally important and should be treated fairly.

All of these metrics are included to carry out a comprehensive analysis of our proposed method. IoU is the metric which we employed for the model selection process, so it is the metric on which we mostly focus. PA and UA are included to provide different aspects to understand the results. PA is the map accuracy from the point of view of the map maker (the producer) and corresponds to the probability that a certain land cover of an area on the ground is classified, while UA is the accuracy from the point of view of a map user and essentially tells users how often the class on the map will actually be present on the ground. Since these 2 metrics, PA and UA, are actually biased and tend to contradict with each other, we decide to include F-Score in our evaluation system to present a harmonic mean of PA and UA. We believe F-Score is a more holistic representation of PA and UA, so more concentration is based on F-Score rather than PA and UA. In addition, there exists a correlation between IoU as well as F-Score, so these 2 metrics agree with each other and we discuss their performance together for the sake of brevity. OA is also included to present the essential accuracy information of our derived results.

5.2. Ablation study of the DKDFN

5.2.1. Performance of collaborative fusion with multihead encoders

To test the effectiveness of our proposed multimodal feature fusion strategy, we evaluate the performance of the DKDFN under different schemes.

Table 4 presents the results of our ablation study on the proposed multimodal fusion scheme. We consider two schemes of the feature fusion method: the deep collaborative fusion method (Fig. 5 (b)), which is finally chosen in the DKDFN; and the shallow fusion method (Fig. 5 (a)). To be more specific, the deep collaborative fusion scheme is completed by concatenating the feature maps of each modality derived by each individual encoder first and then reducing the number of concatenated channels by two-thirds using a 1x1 convolution. This process is done identically for each layer, resulting in a hierarchical of fused features that can be utilized in the corresponding layer of decoder; the shallow fusion scheme, however, concatenates all modalities directly at the input level and feeds them into the network.

Table 4 shows that the mIoU and macro F-score increase by 3.96% and 4.19%, respectively, thanks to our deep collaborative fusion scheme. The shallow fusion scheme results in OA of 83.27% and deep collaborative fusion scheme results in OA of 84.38%. The OA increases by 1.11%. Improvements over all land cover types in terms of the IoU and F-score can also be observed, with the extent of the enhancement differing with the land cover classes. For grassland, wetland, and unused land, the deep collaborative fusion scheme considerably boosts performance with the IoU increasing by 5.71%, 7.99%, and 8.43%, respectively; and the F-score increasing by 7.35%, 8.74%, and 9.64%, respectively. A somewhat moderate performance gain is achieved for the water and built-up area categories. Our proposed feature fusion strategy also allows for more accurate detection for cropland and forest, although the improvements are marginal. In addition, the UA of all classes are improved, indicating that the deep collaboratively fused DKDFN perform more accurate prediction than the shallow fused DKDFN. The PAs of four out of seven categories also increase while the PAs of cropland, wetland, and built-up area decline. The spatialized results in

Table 4
Quantitative results of the DKDFN trained by the shallow fusion scheme and deep collaborative fusion scheme.

Fusion Strategy	DKDFN-shallow fusion				DKDFN-deep collaborative fusion			
Metric	IoU	PA	UA	F-score	IoU	PA	UA	F-score
Forest	81.31%	91.68%	87.78%	89.68%	**81.58%**	**92.00%**	**87.80%**	**89.85%**
Cropland	68.83%	**80.18%**	82.95%	81.54%	**69.33%**	79.70%	**84.20%**	**81.88%**
Water	79.22%	88.85%	87.96%	88.40%	**82.22%**	**92.16%**	**88.40%**	**90.24%**
Built-up Area	67.74%	**80.42%**	81.12%	80.76%	**69.57%**	80.11%	**84.09%**	**82.05%**
Grassland	21.93%	31.30%	42.29%	35.97%	**27.65%**	**41.22%**	**45.66%**	**43.32%**
Wetland	31.32%	**58.23%**	40.40%	47.70%	**39.31%**	55.71%	**57.19%**	**56.44%**
Unused Land	28.13%	35.63%	57.20%	43.90%	**36.56%**	**46.79%**	**62.57%**	**53.54%**
Average	54.07%	66.61%	68.52%	66.85%	**58.03%**	**69.67%**	**72.84%**	**71.04%**

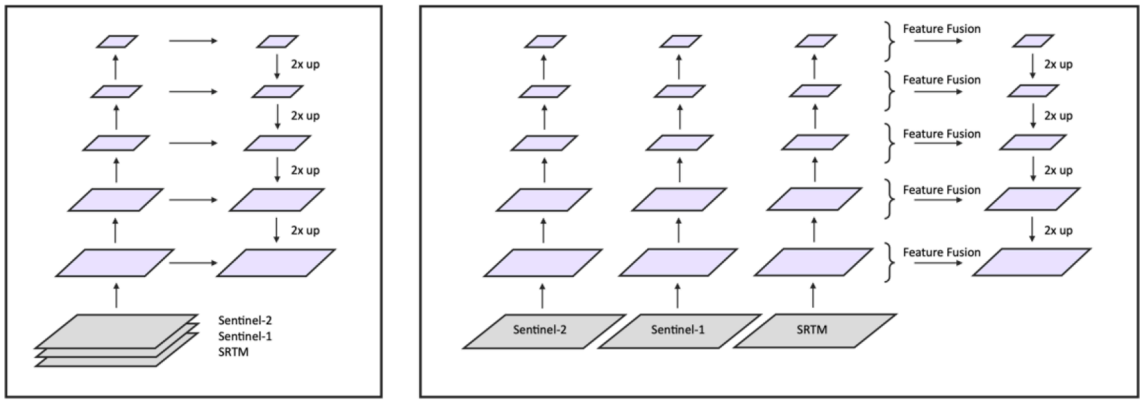

Fig. 5. Two multimodal fusion scheme in the ablation study. (a) is the shallow fusion scheme and (b) is the deep collaborative fusion scheme (ours). The shallow fusion scheme works by concatenating all the channels from three modalities at the input level while the deep collaborative fusion scheme works by extracting modality information separately and fuses the derived information at each layer of the encoder.

Fig. 6 also testifies the superiority of our proposed deep collaborative fusion scheme.

The quantitative and qualitative achievement of the deep collaborative fusion strategy can be attributed to the feature-level information synergy from different modalities over its data-level counterpart. Sentinel-2, Sentinel-1, and SRTM data are considered different modalities, and they share some level of redundancy and complementarity (Baltrušaitis et al., 2017).

The deep collaboratively fused DKDFN is capable of fully exploiting the complementarity between Sentinel-2, Sentinel-1, and SRTM. Since the shallow fusion strategy concatenates modalities at the data level, the three modalities of information are fed into the same encoder sharing the same weights and thus go through the same data transformation process. The three modalities together are projected into a joint space, which may learn a good general representation of all modalities but cannot learn the best representation for each modality as the method fails to consider the difference between Sentinel-2, Sentinel-1, and SRTM. In contrast, our deep collaborative fusion scheme learns separate representations for each modality. The feature extraction process is tailored for Sentinel-2, Sentinel-1, and SRTM, which ensures more sufficient and modality-specific feature mining and leads to better performance. For example, in Fig. 6 (b), the DKDFN adopts a shallow fusion scheme that misclassifies the white dots on rivers as built-up area instead of water. The white dots may be different from surrounding pixels in Sentinel-2, but they are similar to surrounding pixels in Sentinel-1 and SRTM. Since the shallow fusion scheme processes data of three modalities in the same way, the encoder might not be able to considered the favorable information from Sentinel-1 and SRTM thoroughly, which results in commission error. A deep collaborative fusion scheme, however, specializes on each modalities and is capable of completely considering the modality-specific features and thus achieves to the right segmentation result.

5.2.2. Performance of domain Knowledge-Guided decoder

The effectiveness of the domain knowledge-guided decoder is tested by setting the weight of the knowledge reconstruction loss to zero, which deactivates the domain knowledge from guiding the semantic segmentation of land covers during the training process.

The ablation study results are shown in Table 5, from which a promotion of 1.18% in terms of the mIoU and 1.11% in terms of the macro F-score after applying the knowledge-guided decoder can be observed. Also, our OA sees an increasement of 0.57% (i.e., our OA increase from 83.81% to 84.38%). Except for cropland and grassland, our knowledge-guided decoder provides improvements for all land cover categories, especially for wetland and unused land, whose IoUs increase by 4.4% and 2.9%, respectively, and F-score increase by 4.69% and 3.18%, respectively. The impressive improvement of wetland category might be attributed to the RSIs we utilize in this study. We select 7 RSI, among which NDBI, BSI, and Pol are reported with great potential in recognizing wetland (Hird et al., 2017; Mao et al., 2020). With few exceptions, our knowledge-guided decoder mainly improves the UA, meaning it obtains a larger percentage of correctly predicted classes. Fig. 7 shows some spatialized results in our test set with and without a domain knowledge-guided decoder. It can be seen from the red rectangles that the application of our proposed domain knowledge-guided decoder benefit the performance.

To better understand the function of the domain knowledge-guided decoder, the original input band features (Fig. 8 (a)) and the features extracted by our DKDFN with (Fig. 8 (c)) and without domain knowledge (Fig. 8 (b)) information are visualized using t-SNE (van der Maaten and Hinton, 2008). The results are presented in Fig. 8.

Compared with raw data bands, Fig. 8 shows better spatial separability of the DKDFN with and without a knowledge-guided encoder, demonstrating the effectiveness of the learning process. After incorporating domain knowledge by introducing a knowledge-guided decoder, improvements over spatial separability are based on the land cover classes. A clearer margin can be seen between unused land and built-up area, indicating a better separation between these two classes. This can be attributed to the NDBSI and NDBI, two indices that have good sensitivity to bare land, which accounts for most of the unused land and built-up area in our dataset. Additionally, instead of mixing among unused land, forest samples are grouped better thanks to the information provided by domain knowledge. Specifically, Pol may play a part here due to its sensitivity to surface roughness and the vegetation structure, a characteristic that proved useful in discriminating between bare land and forest. However, cropland samples are confused with grassland, which may be due to the lack of indices with the capability to discriminate these two classes. The NDVI and RVI, although they were found to provide great representations of the vegetation structure during testing, may be more sensitive to vegetation dynamics. However, our image source has undergone a temporal aggregation process, which results in a loss of temporal information and restricts the land cover performance, which may benefit from a time series of indices. The extracted feature visualization results agree with the spatialized results.

It can be concluded from the results that domain knowledge, which is a multimodal index in our case, benefits performance. The mapping from original multimodal data bands to the semantic mapping of land covers can be considered a search in space for a set of weights that implement the function. There is a spectrum of largely unexplored possibilities between the input and output. Introducing domain

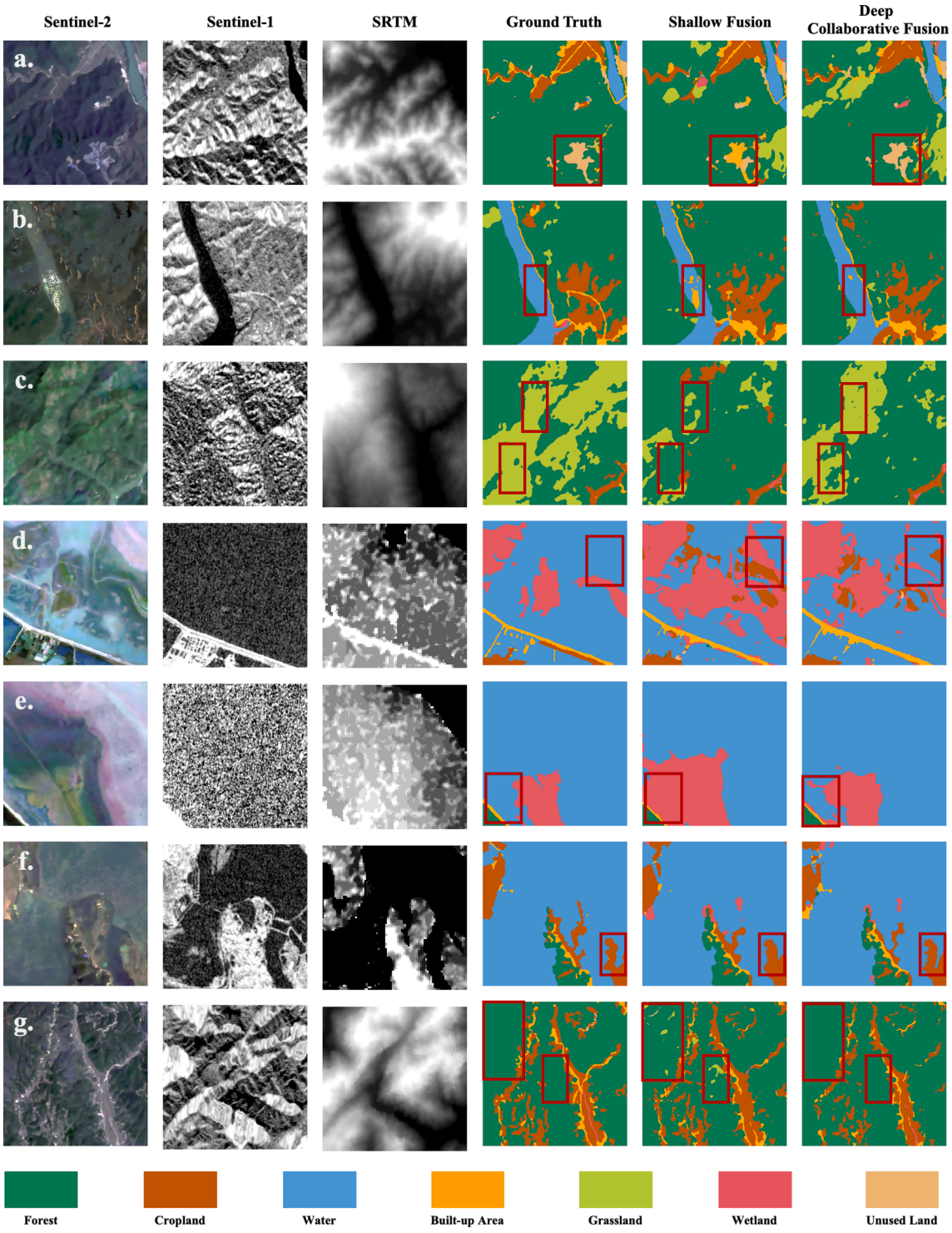

Fig. 6. Qualitative results of the DKDFN adopting different feature fusion schemes. In land cover mapping obtained by the DKDFN using deep collaborative fusion, there are fewer mistakes between built-up area and unused land (the red polygons in Fig. 6 (a)), between built-up area and water (the red polygons in Fig. 6 (b)), between grassland and forest (the red polygons in Fig. 6 (c) and (g)), and between wetland and water (the red polygons in Fig. 6 (d) and (e)). In addition, our deep collaboratively fused DKDFN delineates clearer borders (the red polygons in Fig. 6 (f)).

knowledge reconstruction may be valuable to the learning process by reducing the number of candidate functions mapped from the input to the output (Abu-Mostafa, 1990), in other words, guiding the learning process by directing the network to a solution satisfying the land cover mapping purpose while maintaining the capability of assimilating knowledge from Sentinel-2, Sentinel-1, and SRTM.

What should be noting is that the direct self-reconstruction from modalities cannot produce competitive performance with our knowledge reconstruction decoder. However, this does not mean that the multimodal reconstruction deserves no further testing. The multimodal reconstruction in pretraining stage, the deriving of classification label from input channels could be the possible solutions. Since it is not the focus of our current study, we only provide preliminary experimental results and leave the remaining research to our future work.

Table 5
Quantitative results of the ablation study of the knowledge-guided decoder.

Fusion Strategy	DKDFN-no-kw				DKDFN			
Metric	IoU	PA	UA	F-score	IoU	PA	UA	F-score
Forest	81.06%	89.42%	**89.65%**	89.53%	**81.58%**	**92.00%**	87.80%	**89.85%**
Cropland	**69.73%**	**80.92%**	83.45%	**82.16%**	69.33%	79.70%	**84.20%**	81.88%
Water	81.06%	91.16%	87.97%	89.53%	**82.22%**	**92.16%**	**88.40%**	**90.24%**
Built-up Area	68.68%	**80.92%**	81.94%	81.42%	**69.57%**	80.11%	**84.09%**	**82.05%**
Grassland	**28.88%**	**46.33%**	43.40%	**44.81%**	27.65%	41.22%	**45.66%**	43.32%
Wetland	34.91%	**58.32%**	46.52%	51.75%	**39.31%**	55.71%	**57.19%**	**56.44%**
Unused Land	33.66%	44.04%	58.81%	50.36%	**36.56%**	**46.79%**	**62.57%**	**53.54%**
Average	56.85%	**70.15%**	70.24%	69.93%	**58.03%**	69.67%	**72.84%**	**71.04%**

5.2.3. *Performance of the asymmetric loss function*

The effectiveness of the ALF is tested by substituting our proposed ALF with the GCE to supervise the semantic segmentation process. In other words, by setting β, the weighting factor that adjusts the contribution of the Dice loss for minority channels, to 0, the baseline semantic segmentation task is supervised by the GCE only without additional supervision for minority classes.

Fig. 9 presents per-class gain and loss in terms of the IoU, F-score, PA, and UA, respectively. Each bar group corresponds to a class. The detailed quantitative results are presented in Table 6.

Our proposed ALF is beneficial in terms of the IoU and F-score metric for five out of seven classes, particularly for the minority classes, which are grassland, wetland, as well as unused land since their percentage of the dataset is lower than 10%. The highest increment is observed for the wetland category with the IoU and F-score increasing by 7.14% and 7.81%, respectively. Both grassland and unused land increase the IoU and F-score, ranging from 3%∼4% for grassland and 4%∼5% for unused land. The promotion in all minority classes is consistent with the design logic of the Dice coefficient (Milletari et al., 2016), an objective function for optimization under imbalanced foreground and background classes. The difference in the promotion level may be due to the different difficulties in discriminating each minority classes. However, the ALF causes minor losses for forest and cropland, the IoUs and F-scores of which declined by less than 1% for cropland and less than 0.5% for forest. The OA of DKDFN-no-ALF is 84.04% and the OA of DKDFN is 84.38%.

Fig. 10 shows the spatialized results for the DKDFN with and without applying the proposed ALF. Compared with the DKDFN without the ALF, the land cover mapping results of the DKDFN are more consistent with the reference satellite images. The spatialized results in Fig. 10 also show the ability of the ALF to improve the accuracies of minority classes. The phenomenon presented in the qualitative results agrees with the quantitative results of the DKDFN with and without the ALF.

Thanks to the asymmetric structure of our ALF, which performs extra supervision for minority channels, the performance on minority classes significantly improves without much accuracy loss for the majority classes.

To further prove the effectiveness of our ALF, we compare the experimental results of DKDFN under different loss function constraints. Specifically, we train DKDFN using dice coefficient only, GCE only, a hybrid of Dice coefficient and GCE, and our proposed ALF. The quantitative results are shown in Fig. 11, with our proposed ALF exhibiting the highest accuracy. Our ALF outperforms other loss function constraints. It is superior in its mIoU, which is 1.78% higher than the second best one, hybrid of Dice and GCE. Also, five out of seven classes show the best results in terms of IoU and F-Score. In particular, great rise in IoU and F-Score of all minorities can be observed, with IoU of all minorities outperform the second best one in a range of 3.03% to 4.32% and F-Score in a range of 3.2% to 4.78%. Cropland and forest achieve their highest accuracy when DKDFN is supervised by a hybrid of Dice coefficient and GCE, though the advantage is not obvious. It is also worth noting that DKDFN trained by Dice coefficient do not learn to discriminate grassland and wetland, two of three minorities, with IoU and F-Score of both are zero.

In addition, from Fig. 11, we can clearly observe that IoU and F-Score show the same trend under different loss configurations. Obviously, Cropland and forest, two majorities, reach the best score under the supervision of a hybrid of dice coefficient and GCE, while other land covers perform the best under the supervision of ALF.

These phenomena could be explained by the working scheme of different loss function. When using alone, dice coefficient bias toward the majorities and leaded to insufficient learning of minorities. With the combination of GCE, dice coefficient promotes the accuracies of majorities, as they account for a larger part of the training data. Our ALF, however, promote the performance by focusing on the minorities.

5.3. *Performance on different combination of modalities*

To test the contribution of each modality, we train DKDFN under different modality configuration, with Sentinel-2 as the basic modality and Sentinel-1 and SRTM combined in an incremental way. Note that deep collaborative fusion is employed to be the feature fusion scheme. Fig. 12. presents IoU, F-Score, PA, and UA in terms of each land cover category under different modality combination and presents the quantitative results. The OA of just using Sentinel-2 data is 83.54%. After adding Sentinel-2 data, OA becomes 83.40%. The OA is 84.38% when three modalies are utilized.

The IoU performance under different modality combination is consistent with that of F-Score. For wetland, water, unused land, and built-up area, both IoU and F-Score show an upward trend with more modalities were set as the input, and reach the highest when Sentinel-2, Sentinel-1, and SRTM are combined. Especially for wetland and unused land, their performance incline by 9.07% and 5.14% in terms of IoU, and 10.01% and 5.73% in terms of F-Score, respectively. Forest show a performance drop when Sentinel-1 is added, but it achieves the best score with the help of SRTM. For cropland and grassland, the combination of Sentinel-2 and Sentinel-1 allows for more accurate detections compared to Sentinel-2 only and three modalities configuration, however, these improvements are marginal.

Generally speaking, the inclusion of other modalities in addition to Sentinel-2 brings improvements to the classification of land covers. However, the combination of Sentinel-1 is reported with decline in IoU and F-Score when dealing with forest. This phenomenon might due to the geographical characteristics of forest area in study area. The mountainous environment leads to steeper topography with unstable slopes, which contribute to a loss in Sentinel-1 coherence (Frey et al., 2012). Also, there are areas where Sentinel-1 data is not received at the sensor because of the effects of steep topography on the radar image (Robson et al., 2015). These drawbacks of Sentinel-1 data can be addressed by DEM and performance gain of forest can be observed after the inclusion of SRTM.

Cropland and grassland exhibit a different pattern in comparison with other land covers. Instead of reaching the highest performance when three modalities are combined, these two categories perform the best when Sentinel-2 and Sentinel-1 are fed into the network. This is attributable to the fragmented geometric appearance of grassland and cropland. They benefit from Sentinel-1 since Sentinel-1 is the data

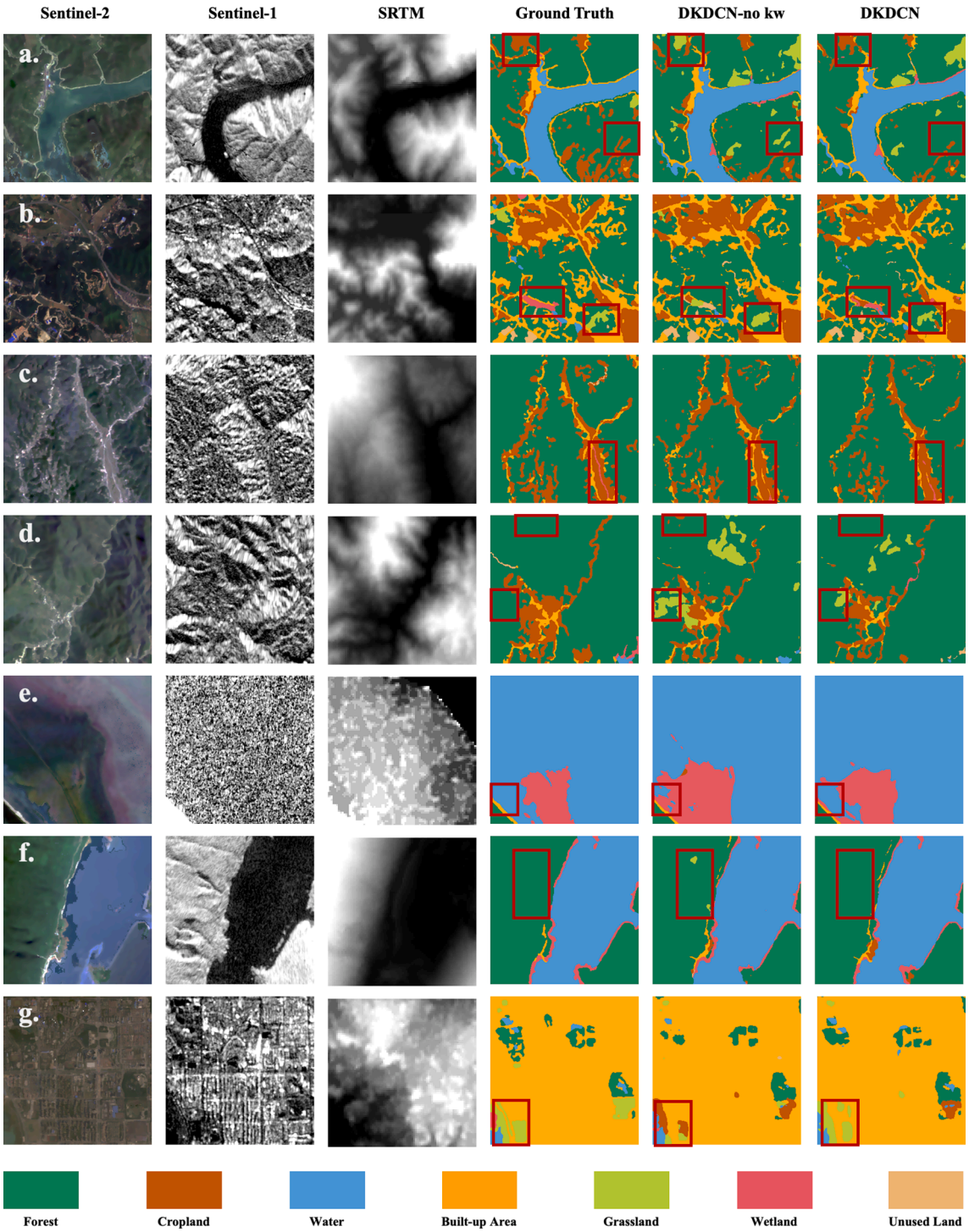

Fig. 7. Qualitative results of the DKDFN with and without a knowledge-guided decoder. The detection of wetlands (the red polygons in Fig. 7 (b), (c) and (e)) and forests (the red polygons in Fig. 7 (a), (d) and (f)) improves. Clearer delineation of cropland (the red polygons in Fig. 7 (a)) and grassland (the red polygons in Fig. 7 (b) and (g)) can also be observed, which agrees with the quantitative results of these two land covers in terms of the UA.

source with the highest spatial resolution, which helps delineate the detail of tiny size land covers. However, the high fragmentation of these two land covers leads to decline in performance after the consideration of SRTM, whose spatial resolution is reported coarser than the nominal 30 m (Grohmann, 2018). Some tiny cropland and grassland become unrecognizable if their size is less than the grid resolution.

We also analyse the model performance under different landscape characteristics to understand whether multimodal co-registration causes problems. Specifically, we divide our test set into flat group and mountainous group (two groups have no overlap) manually and calculate the model performance (Fig. 13, Table 7) under different modality combinations. Note that the image patches in flat group contain no mountainous areas but the image patches in mountainous group might contain some flat areas.

In terms of mIoU (Fig. 13), for all modality combinations, flat group exhibits better accuracy than mountainous group, which means that flat areas might be easier to delineate. With the addition of other modalities, the performance of flat group and mountainous group both improve,

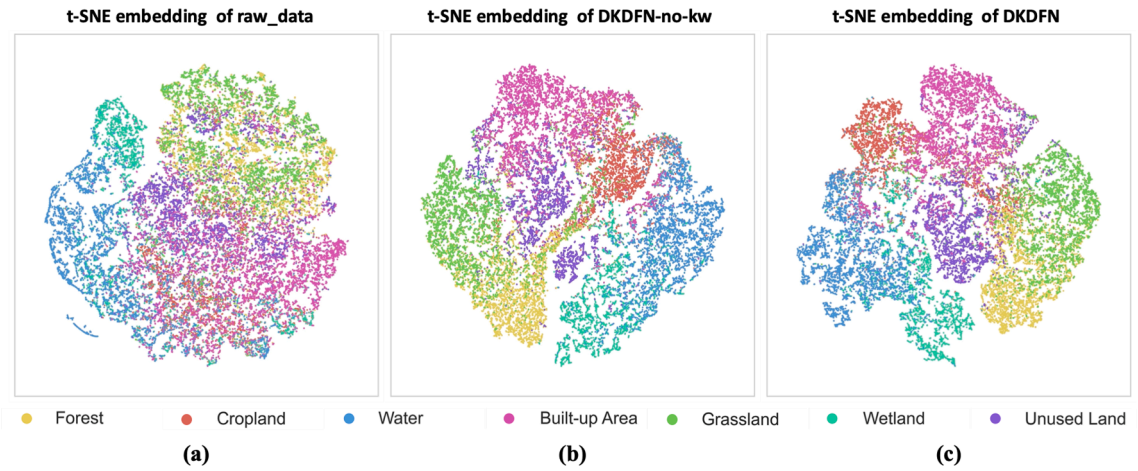

Fig. 8. t-SNE embedding of raw data bands (a) and DKDFN with (c) and without (b) knowledge guided decoder.

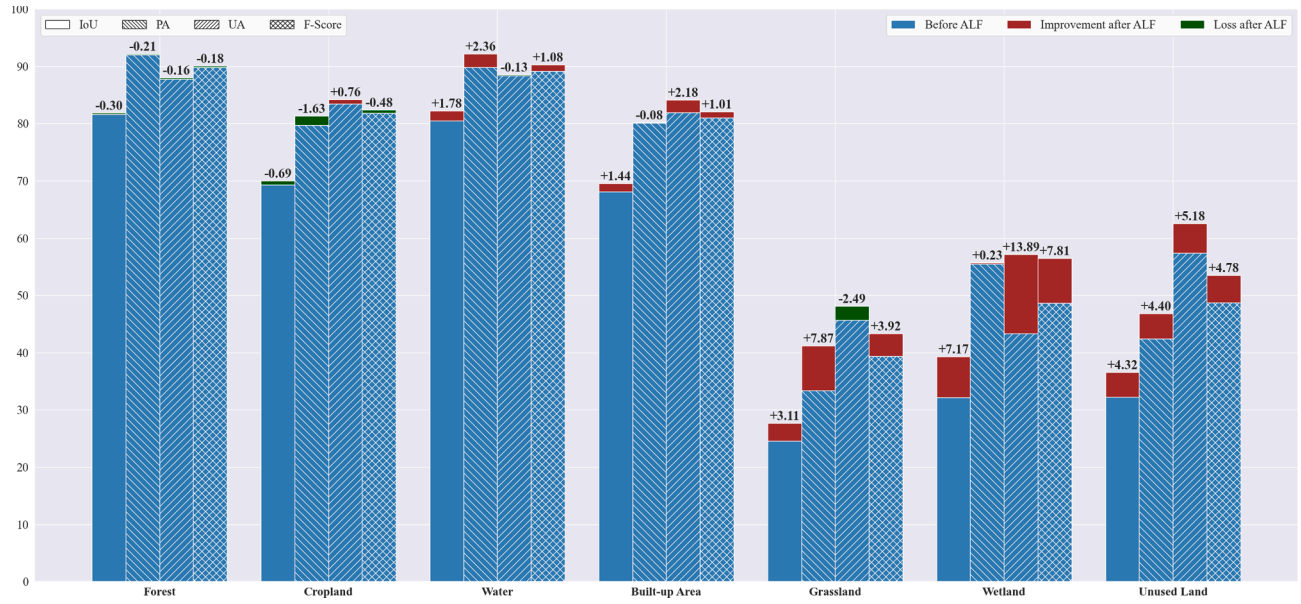

Fig. 9. IoU, F-score, PA, and UA for all land covers before and after employing the ALF. Bars with different texture refer to different evaluation metric of an individual land cover. The blueish parts correspond to the accuracy without our ALF, the reddish parts correspond to the accuracy gain after applying ALF while the greenish parts correspond to the accuracy loss after applying ALF. The quantitative relative improvement in percentage is at the top of each bar. Grassland, wetland, as well as unused land are the three minorities we consider in this study.

Table 6
Quantitative results of the ablation study of the ALF.

Fusion Strategy	DKDFN-no-ALF				DKDFN			
Metric	IoU	PA	UA	F-score	IoU	PA	UA	F-score
Forest	**81.88%**	**92.21%**	**87.96%**	**90.03%**	81.58%	92.00%	87.80%	89.85%
Cropland	**70.02%**	**81.33%**	83.43%	**82.36%**	69.33%	79.70%	**84.20%**	81.88%
Water	80.44%	89.80%	**88.53%**	89.16%	**82.22%**	**92.16%**	88.40%	**90.24%**
Built-up Area	68.13%	**80.19%**	81.91%	81.04%	**69.57%**	80.11%	**84.09%**	**82.05%**
Grassland	24.54%	33.35%	**48.15%**	39.40%	**27.65%**	**41.22%**	45.66%	**43.32%**
Wetland	32.14%	55.48%	43.30%	48.63%	**39.31%**	**55.71%**	**57.19%**	**56.44%**
Unused Land	32.24%	42.39%	57.39%	48.76%	**36.56%**	**46.79%**	**62.57%**	**53.54%**
Average	55.63%	67.82%	70.09%	68.48%	**58.03%**	**69.67%**	**72.84%**	**71.04%**

showing that they both benefit from the information brought by other modalities. However, it is worth noting that flat group shows a higher performance increment (+5.0%) than the mountainous group (+2.41%) and their performance gap is becoming bigger with more modalities added.

The detailed IoU score of each land cover can be found in Table 7. It can be clearly seen from this table that, nearly all land covers in flat group patches show performance increase, demonstrating the effectiveness of our proposed method. But, in mountainous group, cropland, water, as well as grassland achieve the highest score with only Sentinel-2

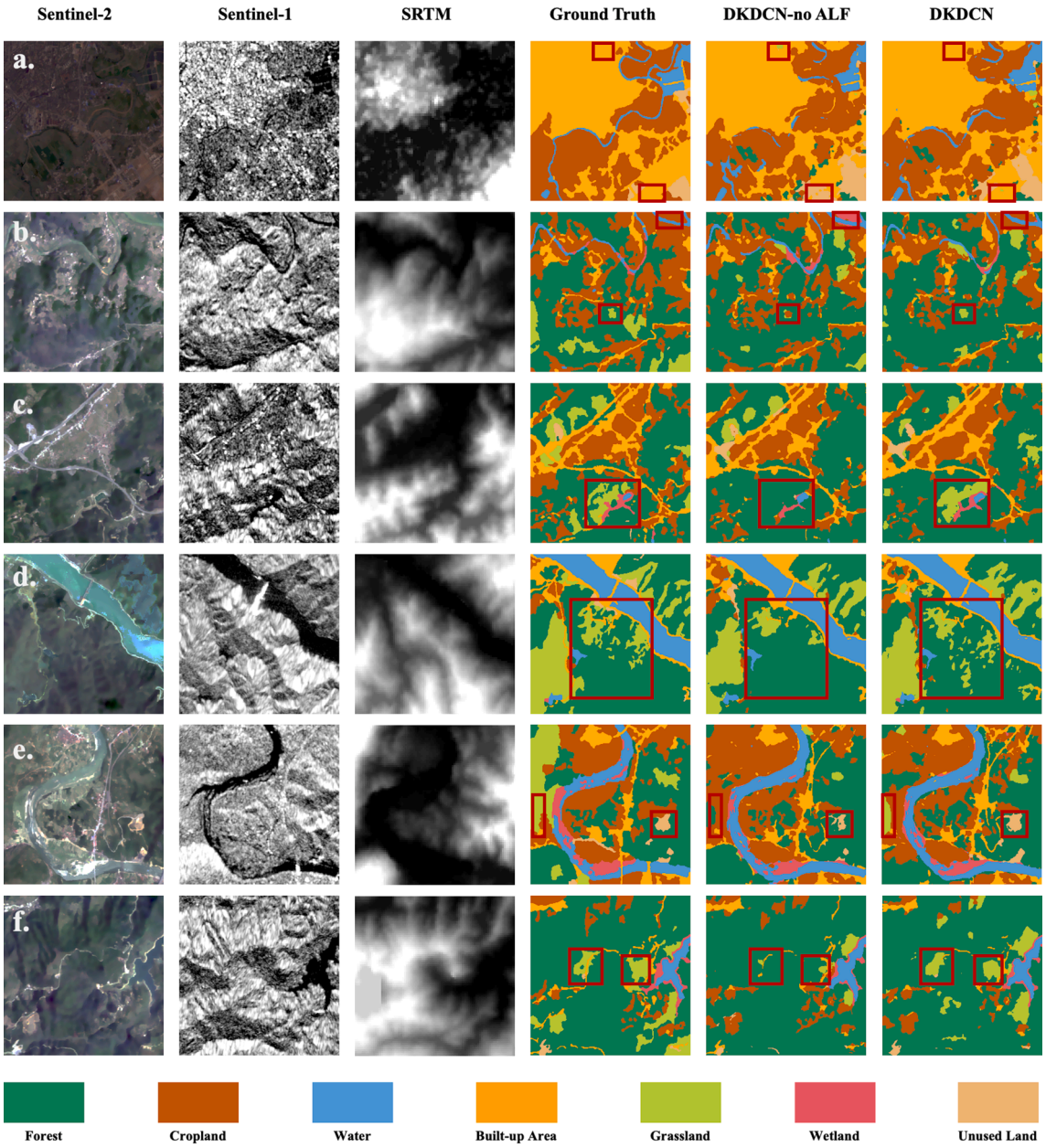

Fig. 10. Spatialized results of the ablation study of the ALF. More specifically, thanks to the ALF, the misclassifications between built-up area and water (the red polygons in Fig. 10 (a)), unused land and built-up area (the red polygons in Fig. 10 (a)), wetland and cropland (the red polygons in Fig. 10 (b)), grassland and wetland (the red polygons in Fig. 10 (b)), grassland and forest (the red polygons in Fig. 10 (c), (d), (e), and (g)), unused land and forest (the red polygons in Fig. 10 (e)) decrease.

set as the input. This phenomenon might due to the land cover distribution and pattern in mountainous group. The land cover dominates mountainous group is forest, which accounts for a much greater proportion than other land covers; Cropland, water, and grassland are the following land covers but they do not appear in high percentage and tend to present in small patches; Built-up area, unused land, as well as wetland seldom appear, even if they appear, they are in the flat areas in mountainous group. Since the co-registration error mainly affect boundary area rather than the center region of a land cover object, the fragmented cropland, water, and grassland are influenced and result in worse accuracy; The boundary of forest might be influenced, too, but the accuracy gain in center region of forest might be bigger than the accuracy loss in boundary region and together lead to a better performance; Built-up area, unused land, and wetland are not influenced severely because they often appear in the flat areas, which are not that prone to co-registration error.

In conclusion, our proposed method is beneficial in employing multimodal information, which is demonstrated by the performance gain after adding more modalities. Overall, our methodology results in the best performance when all modalities are considered, not matter in flat areas or mountainous areas. However, what should be pay attention to is that, in mountainous region the performance of some kinds of land covers that exhibits fragmented spatial pattern deteriorates after the consideration of extra modalities. Maybe co-registration strategy should be taken into account under this situation. Also, for high resolution imagery, co-registration needs to be considered since the co-registration error is more severe in high resolution images compared with moderate resolution images.

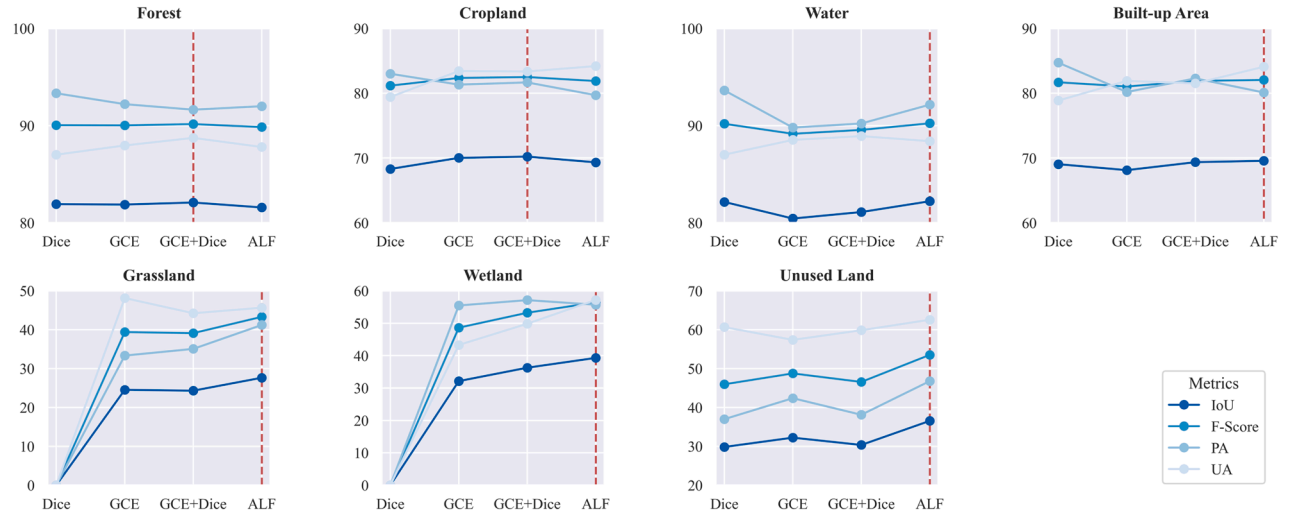

Fig. 11. IoU and F-Score trend for all land covers under different loss constraints. The dashed red lines referred to the loss function striking the best performance on IoU and F-Score.

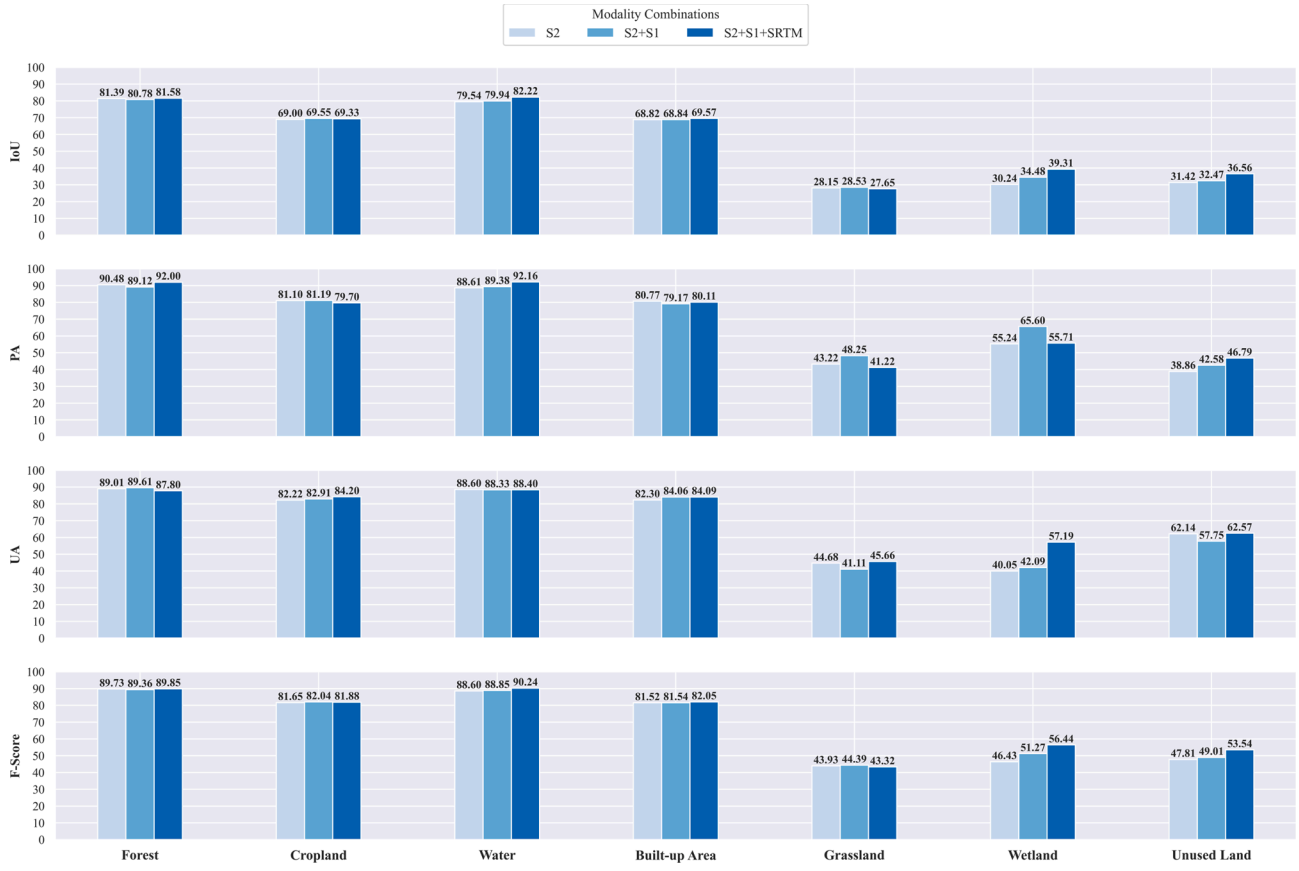

Fig. 12. Land cover performance under different modality combinations. The light blue bars refer to the results of just using Sentinel-2 data; the medium blue bars refer to the results of using Sentinel-2 and Sentinel-1 data; the deep blue bars refer to the results of using Sentinel-2, Sentinel-1, and SRTM data. The quantitative absolute accuracy in percentage is denoted at the top of each bar.

5.4. Comparative experiments with state-of-the-art networks

In this section, we give a full comparison with the state-of-the-art methods. To the best of our knowledge, there do not exist any open-sourced remote sensing-oriented multi-modal semantic segmentation networks for three modalities in literature. With this consideration, the baselines mainly include the state-of-the-art deep networks for semantic segmentation, which have been frequently used as baselines in the remote sensing field. More specifically, we consider the following networks: U-Net (Ronneberger et al., 2015), SegNet (Badrinarayanan et al., 2017), PSPNet (Zhao et al., 2017), DeepLab V3+ (Chen et al., 2018), HRNet (Sun et al., 2019), and MP-ResNet (Ding et al., 2021). More specifically, U-Net adopts a U-shaped network with a contracting path and an expanding path for the precise location and classification of

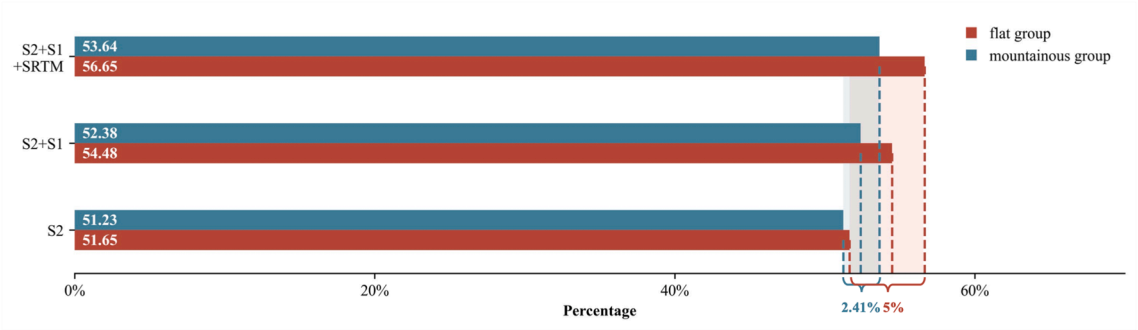

Fig. 13. mIoU of different modality combinations in flat group and mountainous group.

Table 7
Land cover performance under different modality combinations in flat group and mountainous group.

	S2		S2 + S1		S2 + S1 + SRTM	
	Flat Group	Mountainous Group	Flat Group	Mountainous Group	Flat Group	Mountainous Group
Forest	0.6960	0.8250	**0.7251**	0.8271	0.7232	**0.8370**
Cropland	0.7428	**0.5780**	0.7455	0.5680	**0.7496**	0.5466
Water	0.7907	**0.8838**	0.7947	0.8768	**0.8189**	0.8748
Built-up Area	0.7305	0.5236	0.7342	0.5316	**0.7362**	**0.5525**
Grassland	0.0967	**0.3220**	0.1230	0.3200	**0.1731**	0.3045
Wetland	0.2974	0.1431	0.3530	0.2330	**0.4074**	**0.2641**
Unused Land	0.2611	0.3106	0.3381	0.3101	**0.3567**	**0.3755**
Average	0.5165	0.5123	0.5448	0.5238	**0.5665**	**0.5364**

Table 8
Comparison results between our proposed method and existing methods (IoU).

	UNet	SegNet	PSPNet	DeepLab	HRNet	MPResNet	DKDFN (Ours)
Forest	**81.87%**	81.50%	78.77%	79.39%	79.93%	79.96%	81.58%
Cropland	69.21%	67.43%	67.72%	65.22%	68.81%	66.67%	**69.33%**
Water	79.39%	78.32%	78.65%	75.38%	77.85%	78.78%	**82.22%**
Built-up Area	68.75%	65.88%	63.72%	63.70%	67.16%	63.22%	**69.57%**
Grassland	26.19%	18.41%	20.63%	18.60%	16.79%	19.81%	**27.65%**
Wetland	30.17%	22.61%	26.98%	19.77%	25.66%	29.03%	**39.31%**
Unused Land	30.81%	27.16%	29.88%	22.14%	28.35%	25.58%	**36.56%**
Average	55.20%	51.62%	52.34%	49.17%	52.08%	51.86%	**58.03%**

Table 9
Comparison results between our proposed method and existing methods (PA).

	UNet	SegNet	PSPNet	DeepLab	HRNet	MPResNet	DKDFN (Ours)
Forest	91.35%	**93.16%**	90.52%	90.36%	90.53%	92.49%	92.00%
Cropland	80.86%	80.88%	78.95%	75.00%	**81.29%**	78.64%	79.70%
Water	88.56%	88.06%	89.04%	87.10%	87.13%	88.66%	**92.16%**
Built-up Area	81.02%	77.09%	76.40%	**81.63%**	77.83%	76.79%	80.11%
Grassland	38.01%	23.50%	30.03%	28.07%	25.57%	24.89%	**41.22%**
Wetland	**57.60%**	39.02%	50.63%	39.87%	56.03%	51.54%	55.71%
Unused Land	39.51%	33.56%	37.56%	29.40%	35.71%	30.87%	**46.79%**
Average	68.13%	62.18%	64.73%	61.63%	64.87%	63.41%	**69.67%**

Table 10
Comparison results between our proposed method and existing methods (UA).

	UNet	SegNet	PSPNet	DeepLab	HRNet	MPResNet	DKDFN (Ours)
Forest	**88.74%**	86.69%	85.85%	86.73%	87.22%	85.51%	87.80%
Cropland	82.76%	80.21%	82.64%	83.35%	81.76%	81.41%	**84.20%**
Water	**88.46%**	87.63%	87.08%	84.84%	87.96%	87.60%	88.40%
Built-up Area	81.94%	81.91%	79.34%	74.36%	83.04%	78.15%	**84.09%**
Grassland	45.71%	45.95%	39.73%	35.53%	32.83%	**49.27%**	45.66%
Wetland	38.78%	34.96%	36.61%	28.17%	32.14%	39.93%	**57.19%**
Unused Land	58.33%	58.75%	59.38%	47.25%	57.92%	59.89%	**62.57%**
Average	69.24%	68.01%	67.23%	62.89%	66.12%	68.82%	**72.84%**

Table 11
Comparison results between our proposed method and existing methods (F-score).

	UNet	SegNet	PSPNet	DeepLab	HRNet	MPResNet	DKDFN (Ours)
Forest	**90.02%**	89.80%	88.12%	88.50%	88.84%	88.86%	89.85%
Cropland	81.79%	80.54%	80.75%	78.95%	81.52%	80.00%	**81.88%**
Water	88.50%	87.84%	88.04%	85.95%	87.54%	88.12%	**90.24%**
Built-up Area	81.47%	79.42%	77.84%	77.82%	80.35%	77.46%	**82.05%**
Grassland	41.50%	31.09%	34.20%	31.36%	28.74%	33.07%	**43.32%**
Wetland	46.35%	36.87%	42.49%	33.01%	40.84%	44.99%	**56.44%**
Unused Land	47.10%	42.71%	46.01%	36.24%	44.18%	40.74%	**53.54%**
Average	68.1%	64.03%	65.35%	61.69%	64.57%	64.74%	**71.04%**

Table 12
Comparison results between our proposed method and existing methods (OA).

	UNet	SegNet	PSPNet	DeepLab	HRNet	MPResNet	DKDFN (Ours)
OA	84.02%	82.85%	81.95%	80.85%	82.20%	82.32%	**84.38%**

target classes. SegNet achieves pixelwise segmentation with an encoder-decoder structure using memorized max pooling indices. PSPNet fully exploits global contextual information through a pyramid pooling module in the scene parsing task. DeepLab V3 + is a model combining the advantages of the spatial pyramid pooling module and encoder-decoder structure. HRNet maintains high-resolution representations throughout the entire segmentation process. MP-ResNet learns semantic context through its parallel multiscale branches. All of these networks were mainly used as baselines in the mapping tasks in remote sensing field, for example, cropland mapping (Zhang, et al., 2020), building extraction (Liu et al., 2019), as well as land cover mapping (Ding et al., 2021).

Note that all networks take the three modalities as their input. The quantitative results of these networks, together with those our proposed DKDFN, are presented in Table 8, Table 9, Table 10, Table 11, and Table 12 with the IoU, PA, UA, F-score, and OA, respectively, as the measures. These tables clearly show that our proposed DKDFN outperforms the state-of-the-art methods. As for IoU as well as F-Score, our proposed DKDFN achieves the best results not only in mIoU and macro F-Score, but also in each individual land cover, except for forest. However, we find the decline in forest is marginal (i.e., 0.29% lower than UNet in forest IoU and 0.17% lower than UNet in forest F-Score). In terms of PA and UA, our DKDFN attains the best performance in average PA and average UA. Also, our model reaches the highest PA in 3 land covers (i.e., water, grassland, and unused land) and obtains the highest UA in 4 land covers (i.e., cropland, built-up area, wetland, and unused land). We also get the best OA, as Table 12 suggests.

6. Conclusion and future perspectives

This paper proposes a new deep learning network, DKDFN, for land cover mapping. The network can fully exploit information from three modalities including Sentinel-2, Sentinel-1, and SRTM using a multihead encoder. For the purpose of incorporating domain knowledge into our network, a knowledge-guided decoder is designed to guide the semantic segmentation process. To boost the performance on minority classes, an ALF is created to supervise our network. The effectiveness of each design is tested by ablation studies and comparative experiments. Our proposed DKDFN achieves promising results and outperforms the state-of-the-art approaches in terms of overall measures such as average of IoU, average of UA, average of PA , macro F-Score, and OA. As for performance under individual land covers, the proposed DKDFN exhibits advantages over nearly all of the individual classes in terms of IoU and F-Score. As for PA and UA for individual class, we are able to achieve the best performance for more than 3 out of 7 land cover categories. We also provide a new multimodal land cover dataset, which contributes to the research into land cover mapping.

Overall, our network can be easily extended to other networks with an encoder-decoder structure. Furthermore, our network can be transferred to other semantic segmentation tasks utilizing different modalities. In future work, we will exploit more modalities and domain knowledge to improve the performance of our proposed DKDFN under different spatial and temporal conditions.

The effectiveness of our DKDFN has already been testified. Improvements brought by the multihead encoder, domain knowledge-guided decoder, and ALF have been evaluated quantitatively and qualitatively. Nevertheless, some designs in our work demand further developments. For instance, nighttime remote sensing data might be utilized as additional modality in our future work. Moreover, we only consider multimodal indices for knowledge reconstruction. However, other useful domain knowledge, such as texture features and co-occurrence relationships between land covers, may provide some guidance in feature extraction and introduce more stability in our land cover mapping task. Finally, it is worthwhile to further test our DKDFN given different spatial and temporal configurations.

CRediT authorship contribution statement

Yansheng Li: Methodology, Funding acquisition, Investigation, Project administration, Writing – original draft. **Yuhan Zhou:** Data curation, Writing – original draft, Methodology, Validation. **Yongjun Zhang:** Writing – review & editing, Supervision, Project administration. **Liheng Zhong:** Writing – review & editing, Validation. **Jian Wang:** Writing – review & editing, Validation. **Jingdong Chen:** Writing – review & editing, Validation.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China under Grant 41971284 ; the State Key Program of the National Natural Science Foundation of China under Grant 42030102; the Foundation for Innovative Research Groups of the Natural Science Foundation of Hubei Province under Grant 2020CFA003.

References

Abu-Mostafa, Y.S., 1990. Learning from hints in neural networks. J. Complexity 6 (2), 192–198.

Amarsaikhan, D., Blotevogel, H.H., van Genderen, J.L., Ganzorig, M., Gantuya, R., Nergui, B., 2010. Fusing high-resolution SAR and optical imagery for improved urban land cover study and classification. Fusing high-resolution SAR and optical imagery for improved urban land cover study and classification. 1 (1), 83–97.
Ardila, J.P., Tolpekin, V.A., Bijker, W., Stein, A., 2011. Markov-random-field-based super-resolution mapping for identification of urban trees in VHR images. ISPRS J. Photogramm. Remote Sens. 66 (6), 762–775.
Badrinarayanan, V., Kendall, A., Cipolla, R., 2017. SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation. IEEE Trans. Pattern Anal. Mach. Intell. 39 (12), 2481–2495.
Baltrusaitis, T., Ahuja, C., Morency, L.-P., 2019. Multimodal Machine Learning: A Survey and Taxonomy. IEEE Trans. Pattern Anal. Mach. Intell. 41 (2), 423–443.
Belenguer-Plomer, M.A., Tanase, M.A., Chuvieco, E., Bovolo, F., 2021. CNN-based burned area mapping using radar and optical data. Remote Sens. Environ. 260, 112468. https://doi.org/10.1016/j.rse.2021.112468.
Bigdeli, B., Pahlavani, P., 2016. High resolution multisensor fusion of SAR, optical and LiDAR data based on crisp vs. fuzzy and feature vs. decision ensemble systems. Int. J. Appl. Earth Obs. Geoinf. 52, 126–136.
Breiman, L., 2001. Random forests. Mach. Learn. 45, 5–32.
Buchner, J., Yin, H.e., Frantz, D., Kuemmerle, T., Askerov, E., Bakuradze, T., Bleyhl, B., Elizbarashvili, N., Komarova, A., Lewińska, K.E., Rizayeva, A., Sayadyan, H., Tan, B., Tepanosyan, G., Zazanashvili, N., Radeloff, V.C., 2020. Land-cover change in the Caucasus Mountains since 1987 based on the topographic correction of multi-temporal Landsat composites. Remote Sens. Environ. 248, 111967. https://doi.org/10.1016/j.rse.2020.111967.
Buda, M., Maki, A., Mazurowski, M.A., 2018. A systematic study of the class imbalance problem in convolutional neural networks. Neural Networks 106, 249–259.
Calderón-Loor, M., Hadjikakou, M., Bryan, B.A., 2021. High-resolution wall-to-wall land-cover mapping and land change assessment for Australia from 1985 to 2015. Remote Sens. Environ. 252, 112148. https://doi.org/10.1016/j.rse.2020.112148.
Chamorro Martinez, J.A., Cué La Rosa, L.E., Feitosa, R.Q., Sanches, I.D., Happ, P.N., 2021. Fully convolutional recurrent networks for multidate crop recognition from multitemporal image sequences. ISPRS J. Photogramm. Remote Sens. 171, 188–201.
Chen, L.C., Zhu, Y., Papandreou, G., Schroff, F., Adam, H., 2018. Encoder-decoder with atrous separable convolution for semantic image segmentation. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 801–818.
Chen, T., Guestrin, C., 2016. XGBoost: A scalable tree boosting system. Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining 13-17-Augu, 785–794.
Chollet, F., 2017. Xception: Deep learning with depthwise separable convolutions. Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017 2017-Janua, 1800-1807.
Cortes, C., Vapnik, V., 1995. Support-vector networks. Mach. Learn. 20 (3), 273–297.
Cui, W., He, X., Yao, M., Wang, Z., Hao, Y., Li, J., Wu, W., Zhao, H., Xia, C., Li, J., Cui, W., 2021. Knowledge and spatial pyramid distance-based gated graph attention network for remote sensing semantic segmentation. Remote Sensing 13 (7), 1312.
Denize, J., Hubert-Moy, L., Corgne, S., Betbeder, J., & Pottier, E., 2018, July. Identification of winter land use in temperate agricultural landscapes based on Sentinel-1 and 2 Times-Series. In IGARSS 2018-2018 IEEE International Geoscience and Remote Sensing Symposium (pp. 8271-8274). IEEE.
Ding, L., Zheng, K., Lin, D., Chen, Y., Liu, B., Li, J., Bruzzone, L., 2022. MP-ResNet: Multipath Residual Network for the Semantic Segmentation of High-Resolution PolSAR Images. IEEE Geosci. Remote Sensing Lett. 19, 1–5.
El Hajj, M., Bégué, A., Guillaume, S., Martiné, J.-F., 2009. Integrating SPOT-5 time series, crop growth modeling and expert knowledge for monitoring agricultural practices — The case of sugarcane harvest on Reunion Island. Remote Sens. Environ. 113 (10), 2052–2061.
Frey, H., Paul, F., Strozzi, T., 2012. Compilation of a glacier inventory for the western Himalayas from satellite data: Methods, challenges, and results. Remote Sens. Environ. 124, 832–843.
Ganzeveld, L., Bouwman, L., Stehfest, E., van Vuuren, D.P., Eickhout, B., Lelieveld, J., 2010. Impact of future land use and land cover changes on atmospheric chemistry-climate interactions. J. Geophys. Res. Atmos. 115 (D23) https://doi.org/10.1029/2010JD014041.
Ghorbanian, A., Kakooei, M., Amani, M., Mahdavi, S., Mohammadzadeh, A., Hasanlou, M., 2020. Improved land cover map of Iran using Sentinel imagery within Google Earth Engine and a novel automatic workflow for land cover classification using migrated training samples. ISPRS J. Photogramm. Remote Sens. 167, 276–288.
Glorot, X., Bordes, A., Bengio, Y., 2011. Deep sparse rectifier neural networks. J. Mach. Learn. Res. 15, 315–323.
Gong, P., Wang, J., Yu, L.e., Zhao, Y., Zhao, Y., Liang, L.u., Niu, Z., Huang, X., Fu, H., Liu, S., Li, C., Li, X., Fu, W., Liu, C., Xu, Y., Wang, X., Cheng, Q.u., Hu, L., Yao, W., Zhang, H., Zhu, P., Zhao, Z., Zhang, H., Zheng, Y., Ji, L., Zhang, Y., Chen, H., Yan, A. n., Guo, J., Yu, L., Wang, L., Liu, X., Shi, T., Zhu, M., Chen, Y., Yang, G., Tang, P., Xu, B., Giri, C., Clinton, N., Zhu, Z., Chen, J., Chen, J., 2013. Finer resolution observation and monitoring of global land cover: First mapping results with Landsat TM and ETM+ data. Int. J. Remote Sens. 34 (7), 2607–2654.
Grohmann, C.H., 2018. Evaluation of TanDEM-X DEMs on selected Brazilian sites: Comparison with SRTM, ASTER GDEM and ALOS AW3D30. Remote Sens. Environ. 212, 121–133.
Hazirbas, C., Ma, L., Domokos, C., on, D.C.-A. conference, 2016, undefined, 2017. Fusenet: Incorporating depth into semantic segmentation via fusion-based cnn architecture. Springer 10111 LNCS, 213–228.
Hibbard, K., Janetos, A., Van Vuuren, D.P., Pongratz, J., Rose, S.K., Betts, R., Herold, M., Feddema, J.J., 2010. Research priorities in land use and land-cover change for the Earth system and integrated assessment modelling. Int. J. Climatol. 30, 2118–2128.
Hird, J.N., DeLancey, E.R., McDermid, G.J., Kariyeva, J., 2017. Google earth engine, open-access satellite data, and machine learning in support of large-area probabilisticwetland mapping. Remote Sensing 9 (12), 1315.
Hurskainen, P., Adhikari, H., Siljander, M., Pellikka, P.K.E., Hemp, A., 2019. Auxiliary datasets improve accuracy of object-based land use/land cover classification in heterogeneous savanna landscapes. Remote Sens. Environ. 233, 111354. https://doi.org/10.1016/j.rse.2019.111354.
Ienco, D., Interdonato, R., Gaetano, R., Ho Tong Minh, D., 2019. Combining Sentinel-1 and Sentinel-2 Satellite Image Time Series for land cover mapping via a multi-source deep learning architecture. ISPRS J. Photogramm. Remote Sens. 158, 11–22.
Imaoka, K., Kachi, M., Fujii, H., Murakami, H., Hori, M., Ono, A., Igarashi, T., Nakagawa, K., Oki, T., Honda, Y., Shimoda, H., 2010. Global change observation mission (GCOM) for monitoring carbon, water cycles, and climate change. Proc. IEEE 98 (5), 717–734.
Ioffe, S., Szegedy, C., 2015. Batch normalization: Accelerating deep network training by reducing internal covariate shift. 32nd International Conference on Machine Learning, ICML.
Jiang, J., Zheng, L., Luo, F., & Zhang, Z., 2018. Rednet: Residual encoder-decoder network for indoor rgb-d semantic segmentation. arXiv preprint arXiv:1806.01054.
Johnson, J.M., Khoshgoftaar, T.M., 2019. Survey on deep learning with class imbalance. J. Big Data 6 (1), 1–54.
Jun, C., Ban, Y., Li, S., 2014. Open access to Earth land-cover map. Nature 514 (7523).
Kellenberger, B., Marcos, D., Tuia, D., 2018. Detecting mammals in UAV images: Best practices to address a substantially imbalanced dataset with deep learning. Remote Sens. Environ. 216, 139–153.
LeCun, Y., Bengio, Y., Hinton, G., 2015. Deep learning. Nature 521 (7553), 436–444.
Li, W., Chen, K., Chen, H., Shi, Z., 2022. Geographical Knowledge-Driven Representation Learning for Remote Sensing Images. IEEE Trans. Geosci. Remote Sens. 60, 1–16.
Li, W., Dong, R., Fu, H., Wang, J., Yu, L.e., Gong, P., 2020a. Integrating Google Earth imagery with Landsat data to improve 30-m resolution land cover mapping. Remote Sens. Environ. 237, 111563. https://doi.org/10.1016/j.rse.2019.111563.
Li, Y., Chen, W., Zhang, Y., Tao, C., Xiao, R., Tan, Y., 2020b. Accurate cloud detection in high-resolution remote sensing imagery by weakly supervised deep learning. Remote Sens. Environ. 250, 112045. https://doi.org/10.1016/j.rse.2020.112045.
Li, Y., Shi, T.e., Zhang, Y., Chen, W., Wang, Z., Li, H., 2021a. Learning deep semantic segmentation network under multiple weakly-supervised constraints for cross-domain remote sensing image semantic segmentation. ISPRS J. Photogramm. Remote Sens. 175, 20–33.
Li, Y., Zhang, Y., Zhu, Z., 2021b. Error-Tolerant Deep Learning for Remote Sensing Image Scene Classification. IEEE Trans. Cybern. 51 (4), 1756–1768.
Li, Y., Kong, D., Zhang, Y., Tan, Y., Chen, L., 2021c. Robust deep alignment network with remote sensing knowledge graph for zero-shot and generalized zero-shot remote sensing image scene classification. ISPRS J. Photogramm. Remote Sens. 179, 145–158.
Liang, L.u., Xu, B., Chen, Y., Liu, Y., Cao, W., Fang, L., Feng, L., Goodchild, M.F., Gong, P., Li, W., 2010. Combining spatial-temporal and phylogenetic analysis approaches for improved understanding on global H5N1 transmission. PLoS ONE 5 (10), e13575.
Lin, C., Du, P., Samat, A., Li, E., Wang, X., Xia, J., 2019. Automatic Updating of Land Cover Maps in Rapidly Urbanizing Regions by Relational Knowledge Transferring from GlobeLand30. Remote Sensing 11 (12), 1397. https://doi.org/10.3390/rs11121397.
Lin, Y., Zhang, H., Lin, H., Gamba, P.E., Liu, X., 2020. Incorporating synthetic aperture radar and optical images to investigate the annual dynamics of anthropogenic impervious surface at large scale. Remote Sens. Environ. 242, 111757. https://doi.org/10.1016/j.rse.2020.111757.
Liu, H., Gong, P., Wang, J., Wang, X.i., Ning, G., Xu, B., 2021. Production of global daily seamless data cubes and quantification of global land cover change from 1985 to 2020 - iMap World 1.0. Remote Sens. Environ. 258, 112364. https://doi.org/10.1016/j.rse.2021.112364.
Liu, J., Vogelmann, J.E., Zhu, Z., Key, C.H., Sleeter, B.M., Price, D.T., Chen, J.M., Cochrane, M.A., Eidenshink, J.C., Howard, S.M., Bliss, N.B., Jiang, H., 2011. Estimating California ecosystem carbon change using process model and land cover disturbance data: 1951–2000. Ecol. Model. 222 (14), 2333–2341.
Liu, S., Qi, Z., Li, X., Yeh, A.G.O., 2019. Integration of convolutional neural networks and object-based post-classification refinement for land use and land cover mapping with optical and SAR data. Remote Sensing 11 (6), 690.
Mao, D., Wang, Z., Du, B., Li, L., Tian, Y., Jia, M., Zeng, Y., Song, K., Jiang, M., Wang, Y., 2020. National wetland mapping in China: A new product resulting from object-based and hierarchical classification of Landsat 8 OLI images. ISPRS J. Photogramm. Remote Sens. 164, 11–25.
Matikainen, L., Karila, K., Litkey, P., Ahokas, E., Hyyppä, J., 2020. Combining single photon and multispectral airborne laser scanning for land cover classification. ISPRS J. Photogramm. Remote Sens. 164, 200–216.
Milletari, F., Navab, N., Ahmadi, S.A., 2016. V-Net: Fully convolutional neural networks for volumetric medical image segmentation. Proceedings - 2016 4th International Conference on 3D Vision, 3DV 2016 565–571.
Nghiem, S.V., Balk, D., Rodriguez, E., Neumann, G., Sorichetta, A., Small, C., Elvidge, C. D., 2009. Observations of urban and suburban environments with global satellite scatterometer data. ISPRS J. Photogramm. Remote Sens. 64 (4), 367–380.
Nguyen, L.H., Joshi, D.R., Clay, D.E., Henebry, G.M., 2020. Characterizing land cover/land use from multiple years of Landsat and MODIS time series: A novel approach using land surface phenology modeling and random forest classifier. Remote Sens. Environ. 238, 111017. https://doi.org/10.1016/j.rse.2018.12.016.
Olson, D.M., Dinerstein, E., Wikramanayake, E.D., Burgess, N.D., Powell, G.V.N., Underwood, E.C., D'amico, J.A., Itoua, I., Strand, H.E., Morrison, J.C., Loucks, C.J.,

Allnutt, T.F., Ricketts, T.H., Kura, Y., Lamoreux, J.F., Wettengel, W.W., Hedao, P., Kassem, K.R., 2001. Terrestrial ecoregions of the world: A new map of life on Earth. Bioscience 51 (11), 933–938.
Ortigosa-Hernández, J., Inza, I., Lozano, J.A., 2017. Measuring the class-imbalance extent of multi-class problems. Pattern Recogn. Lett. 98, 32–38.
Ozdarici-Ok, A., Ok, A., Schindler, K., 2015. Mapping of agricultural crops from single high-resolution multispectral images-data-driven smoothing vs. parcel-based smoothing. Remote Sens. 7 (5), 5611–5638.
Pan, S., Guan, H., Chen, Y., Yu, Y., Nunes Gonçalves, W., Marcato Junior, J., Li, J., 2020. Land-cover classification of multispectral LiDAR data using CNN with optimized hyper-parameters. ISPRS J. Photogramm. Remote Sens. 166, 241–254.
Phiri, D., Morgenroth, J., 2017. Developments in Landsat Land Cover Classification Methods: A Review. Remote Sens. 9 (9), 967. https://doi.org/10.3390/rs9090967.
Phiri, D., Simwanda, M., Salekin, S., Nyirenda, V.R., Murayama, Y., Ranagalage, M., 2020. Sentinel-2 data for land cover/use mapping: A review. Remote Sens. 12 (14), 2291.
Poulter, B., Frank, D.C., Hodson, E.L., Zimmermann, N.E., 2011. Impacts of land cover and climate data selection on understanding terrestrial carbon dynamics and the CO 2 airborne fraction. Biogeosciences 8 (8), 2027–2036.
Prati, R.C., Batista, G.E.A.P.A., Silva, D.F., 2015. Class imbalance revisited: a new experimental setup to assess the performance of treatment methods. Knowl. Inf. Syst. 45 (1), 247–270.
Quin, G., Pinel-Puyssegur, B., Nicolas, J.-M., Loreaux, P., 2014. MIMOSA: An automatic change detection method for sar time series. IEEE Trans. Geosci. Remote Sens. 52 (9), 5349–5363.
Rees, W.G., Williams, M., Vitebsky, P., 2003. Mapping land cover change in a reindeer herding area of the Russian Arctic using Landsat TM and ETM+ imagery and indigenous knowledge. Remote Sens. Environ. 85 (4), 441–452.
Rennó, C.D., Nobre, A.D., Cuartas, L.A., Soares, J.V., Hodnett, M.G., Tomasella, J., Waterloo, M.J., 2008. HAND, a new terrain descriptor using SRTM-DEM: Mapping terra-firme rainforest environments in Amazonia. Remote Sens. Environ. 112 (9), 3469–3481.
Robson, B.A., Nuth, C., Dahl, S.O., Hölbling, D., Strozzi, T., Nielsen, P.R., 2015. Automated classification of debris-covered glaciers combining optical, SAR and topographic data in an object-based environment. Remote Sens. Environ. 170, 372–387.
Ronneberger, O., Fischer, P., Brox, T., 2015. U-net: Convolutional networks for biomedical image segmentation. In: Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). Springer Verlag, pp. 234–241.
Running, S.W., 2008. Climate change: Ecosystem disturbance, carbon, and climate. Science 321 (5889), 652–653.
Schmitt, M., Hughes, L.H., Qiu, C., Zhu, X.X., 2019. AGGREGATING CLOUD-FREE SENTINEL-2 IMAGES with GOOGLE EARTH ENGINE. ISPRS Ann. Photogramm. Remote Sens. Spatial Inf. Sci. 4, 145–152.
Sesnie, S.E., Gessler, P.E., Finegan, B., Thessler, S., 2008. Integrating Landsat TM and SRTM-DEM derived variables with decision trees for habitat classification and change detection in complex neotropical environments. Remote Sens. Environ. 112 (5), 2145–2159.
Shao, Z., Fu, H., Fu, P., Yin, L.i., 2016. Mapping Urban Impervious Surface by Fusing Optical and SAR Data at the Decision Level. Remote Sensing 8 (11), 945. https://doi.org/10.3390/rs8110945.
Sica, F., Pulella, A., Nannini, M., Pinheiro, M., Rizzoli, P., 2019. Repeat-pass SAR interferometry for land cover classification: A methodology using Sentinel-1 Short-Time-Series. Remote Sens. Environ. 232, 111277. https://doi.org/10.1016/j.rse.2019.111277.
Sukawattanavijit, C., Chen, J., Zhang, H., 2017. GA-SVM Algorithm for Improving Land-Cover Classification Using SAR and Optical Remote Sensing Data. IEEE Geosci. Remote Sens. Lett. 14 (3), 284–288.
Sun, K., Xiao, B., Liu, D., Wang, J., 2019. Deep high-resolution representation learning for human pose estimation. Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition 2019-June, 5686–5696.
Symeonakis, E., Higginbottom, T.P., Petroulaki, K., Rabe, A., 2018. Optimisation of savannah land cover characterisation with optical and SAR data. Remote Sensing 10 (4), 499.
Tong, X.-Y., Xia, G.-S., Lu, Q., Shen, H., Li, S., You, S., Zhang, L., 2020. Land-cover classification with high-resolution remote sensing images using transferable deep models. Remote Sens. Environ. 237, 111322. https://doi.org/10.1016/j.rse.2019.111322.
van Beijma, S., Comber, A., Lamb, A., 2014. Random forest classification of salt marsh vegetation habitats using quad-polarimetric airborne SAR, elevation and optical RS data. Remote Sens. Environ. 149, 118–129.
van der Maaten, L., Hinton, G., 2008. Visualizing data using t-SNE. J. Mach. Learn. Res. 9, 2579–2625.
von Rueden, L., Mayer, S., Beckh, K., Georgiev, B., Giesselbach, S., Heese, R., ... & Schuecker, J., 2019. Informed Machine Learning–A Taxonomy and Survey of Integrating Knowledge into Learning Systems. arXiv preprint arXiv:1903.12394.
Waldner, F., Canto, G.S., Defourny, P., 2015. Automated annual cropland mapping using knowledge-based temporal features. ISPRS J. Photogramm. Remote Sens. 110, 1–13.
Wan, L., Xiang, Y., You, H., 2019. A Post-Classification Comparison Method for SAR and Optical Images Change Detection. IEEE Geosci. Remote Sens. Lett. 16 (7), 1026–1030.
Xu, B., Gong, P., Biging, G., Liang, S., Seto, E., Spear, R., 2004. Snail density prediction for schistosomiasis control using IKONOS and ASTER images. Photogramm. Eng. Remote Sens. 70 (11), 1285–1294.
Zadeh, A., Chen, M., Poria, S., Cambria, E., Morency, L.-P., 2017. Tensor Fusion Network for Multimodal Sentiment Analysis. EMNLP 2017 - Conference on Empirical Methods in Natural Language Processing, Proceedings 1103–1114.
Zhang, C., Kovacs, J.M., 2012. The application of small unmanned aerial systems for precision agriculture: A review. Precis. Agric. 13 (6), 693–712.
Zhang, C., Yang, Z., He, X., Deng, L.i., 2020. Multimodal Intelligence: Representation Learning, Information Fusion, and Applications. IEEE J. Sel. Top. Sign. Proces. 14 (3), 478–493.
Zhang, Z., Sabuncu, M.R., 2018. Generalized cross entropy loss for training deep neural networks with noisy labels. Advances in Neural Information Processing Systems 2018-Decem, 8778–8788.
Zhao, H., Shi, J., Qi, X., Wang, X., Jia, J., 2017. Pyramid scene parsing network. Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017 2017-Janua.
Zhou, S., Kuester, T., Bochow, M., Bohn, N., Brell, M., Kaufmann, H., 2021. A knowledge-based, validated classifier for the identification of aliphatic and aromatic plastics by WorldView-3 satellite data. Remote Sens. Environ. 264, 112598. https://doi.org/10.1016/j.rse.2021.112598.
Zhu, Z., Wang, S., Woodcock, C.E., 2015. Improvement and expansion of the Fmask algorithm: cloud, cloud shadow, and snow detection for Landsats 4–7, 8, and sentinel 2 images. Remote Sens. Environ. 159, 269–277.