X-ModalNet: A semi-supervised deep cross-modal network for classification of remote sensing data

Danfeng Hong[a,b], Naoto Yokoya[c,d], Gui-Song Xia[e,f,g], Jocelyn Chanussot[h,i], Xiao Xiang Zhu[a,b,*]

[a] *Remote Sensing Technology Institute, German Aerospace Center, 82234 Wessling, Germany*
[b] *Signal Processing in Earth Observation, Technical University of Munich, 80333 Munich, Germany*
[c] *Graduate School of Frontier Sciences, The University of Tokyo, 277-8561 Chiba, Japan*
[d] *Geoinformatics Unit, RIKEN Center for Advanced Intelligence Project, RIKEN, 103-0027 Tokyo, Japan*
[e] *School of Computer Science, Wuhan University, 430072 Wuhan, China*
[f] *Institute of Artificial Intelligence, Wuhan University, 430072 Wuhan, China*
[g] *State Key Laboratory for Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University, 430079 Wuhan, China*
[h] *Univ. Grenoble Alpes, INRIA, CNRS, Grenoble INP, LJK, 38000 Grenoble, France*
[i] *Aerospace Information Research Institute, Chinese Academy of Sciences, 100094 Beijing, China*

ARTICLE INFO

ABSTRACT

This paper addresses the problem of semi-supervised transfer learning with limited cross-modality data in remote sensing. A large amount of multi-modal earth observation images, such as multispectral imagery (MSI) or synthetic aperture radar (SAR) data, are openly available on a global scale, enabling parsing global urban scenes through remote sensing imagery. However, their ability in identifying materials (pixel-wise classification) remains limited, due to the noisy collection environment and poor discriminative information as well as limited number of well-annotated training images. To this end, we propose a novel cross-modal deep-learning framework, called X-ModalNet, with three well-designed modules: self-adversarial module, interactive learning module, and label propagation module, by learning to transfer more discriminative information from a small-scale hyperspectral image (HSI) into the classification task using a large-scale MSI or SAR data. Significantly, X-ModalNet generalizes well, owing to propagating labels on an updatable graph constructed by high-level features on the top of the network, yielding semi-supervised cross-modality learning. We evaluate X-ModalNet on two multi-modal remote sensing datasets (HSI-MSI and HSI-SAR) and achieve a significant improvement in comparison with several state-of-the-art methods.

1. Introduction

Currently operational radar (e.g., Sentinel-1) and optical broadband (multispectral) satellites (e.g., Sentinel-2 and Landsat-8) enable the synthetic aperture radar (SAR) (Kang et al., 2020) and multispectral image (MSI) (Zhang et al., 2019a) openly available on a global scale. Therefore, there has been a growing interest in understanding our environment through remote sensing (RS) images, which is of great benefit to many potential applications such as image classification (Tuia et al., 2015; Han et al., 2018; Srivastava et al., 2019; Cao et al., 2020a), object and change detection (Zhang et al., 2018b, 2019b; Wu et al., 2019; Wu et al., 2020), mineral exploration (Gao et al., 2017a; Hong and Zhu, 2018; Hong et al., 2019b; Yao et al., 2019), multi-modality data analysis (Hong et al., 2019d, 2020a; Hu et al., 2019a; Yang et al., 2019), to name a few. In particular, RS data classification is a fundamental but still challenging problem across computer vision and RS fields. It aims to assign a semantic category to each pixel in a studied urban scene. For example, in Gao et al. (2017b), spectral-spatial information is applied to significantly suppress the influence of noise in dimensionality reduction, and the proposed method is obviously effective in extracting nonlinear features and improving the classification accuracy.

Recently, enormous efforts have been made on developing deep learning (DL)-based approaches (LeCun et al., 2015), such as deep neural networks (DNNs) and convolutional neural networks (CNNs), to parse urban scenes by using street view images. Yet it is less investigated at the level of satellite-borne or aerial images. Bridging advanced learning-based techniques or vision algorithms with RS imagery

---

Table 1
Qualitative comparison of urban scene parsing using street view images and RS images in terms of goal, acquisition perspective, scene covering scale, spatial resolution, feature diversity, data accessibility, and ground truth maps used for training.

Urban Scene Parsing	Street View Images	RS Images
Goal	Pixel-wise Classification	
Perspective	Horizontal	"Bird's"
Scene Scale	Small	Large
Spatial Resolution	High	Low
Feature Diversity	Low	High
Accessibility	Moderate	Easy
Ground Truth Maps	Dense	Sparse

could allow for a variety of new applications potentially conducted on a larger and even a global scale. A qualitative comparison is given in Table 1 to highlight the differences as well as advantages and disadvantages in the classification task using different scene images (e.g., street view or RS images).

1.1. Motivation and objective

We clarify our motivation to answer the following three "why" questions: (1) *Why classify or parse RS images?* (2) *Why use multimodal data?* (3) *Why learn the cross-modal representation?*

- **From Street to Earth Vision**   Remotely sensed imagery can provide a new insight for global urban scene understanding. The data in Earth Vision, on one hand, benefit from a "bird's perspective," providing a structure-related multiview surface information; and, on the other hand, it is acquired on a wider and even global scale.
- **From Unimodal to Multimodal Data**   Limited by the low image resolution and a handful of labeled samples, unimodal RS data are inevitable to meet the bottleneck in performance gain, despite being able to be openly and largely acquired. Therefore, an alternative to maximize the classification accuracy is to jointly leverage the multimodal data.
- **From Multimodal to Crossmodal Learning**   In reality, a large amount of information-rich data, such as hyperspectral imagery (HSI), are hardly collected due to technical limitations of satellite sensors. Thus, only the limited multimodal correspondences can be used to train a model, while one modality is absent in the test phase. This is a typical cross-modality learning (CML) issue.

Fig. 1 illustrates the to-be-solved problem and potential solution, where MSI in magenta or SAR in cyan is freely available at a large and even global scale but they are limited by relatively poor feature representation ability, while the HSI in red is characterized by rich spectral information but fails to be acquired in a large-covered area. This naturally leads to a general but interesting question: *can a limited amount of spectrally discriminative HSI improve the parsing performance of a large amount of low-quality data (SAR or MSI) in the large-scale classification or mapping task?* A feasible solution to the problem is the CML.

Motivated by the above analysis, the CML issue that we aim at tackling can be further generalized to three specific challenges related to computer vision or machine learning.

- RS images acquired from the satellites or airplanes inevitably suffer from various variations caused by environmental conditions (e.g., illumination and topology changes, atmospheric effects) and instrumental configurations (e.g., sensor noise).
- Multimodal RS data are usually characterized by the different properties. Blending multi/ cross-modal representation in a more effective and compacted way is still an important challenge in our case.

- RS images in Earth Vision can provide a larger-scale visual field. This tends to lead to costly labeling and noisy annotations in the process of data preparation.

According to the three factors, our objective can be summarized to develop novel approaches or improve the existing ones, yielding a more discriminative multimodality blending and robust against various variabilities in RS images with the limited number of training annotations.

1.2. Method overview and contributions

Towards the aforementioned goals, a novel cross-modal DL framework is proposed in a semi-supervised fashion, called X-ModalNet, for RS image classification. As illustrated in Fig. 2, a three-stream network is developed to learn the multimodal joint representation in consideration of unlabeled samples, where the network parameters would be shared from the same modalities. Moreover, an interactive learning strategy is modeled across the two modalities to facilitate the information blending more effectively. Prior to the interactive learning (IL) module, we also embed a self-adversarial (SA) module robustly against noise attack, thereby enhance the model's generalization capability. To fully make use of unlabeled samples, we iteratively update pseudo-labels by label propagation (LP) on the graph constructed by high-level hidden representations. Extensive experiments are conducted on two multimodal datasets (HSI-MSI and HSI-SAR), showing the effectiveness and superiority of our proposed X-ModalNet in the RS data classification task.

The main contributions can be highlighted in four-folds:

- To our best knowledge, this is the first time to investigate the HSI-aided CML's case by designing such deep cross-modal network (X-ModalNet) in RS fields for improving the classification accuracy of only using MSI or SAR with the aid of a limited amount of HSI samples.
- According to spatially high resolution of MSI (SAR) as well as spectrally high resolution of HSI, our X-ModalNet is a novel and promising network architecture reasonably, which takes a hybrid network as backbone, that is, CNN for MSI or SAR and DNN for HSI. Such design enables the best full use of high spatial and rich spectral information from MSI or SAR and HSI, respectively.
- We propose two novel plug-and-play modules: SA module and IL module, aiming at improving the robustness and discrimination of the multimodal representation. On the one hand, we modularize the idea of generative adversarial networks (GANs) (Goodfellow et al., 2014a) into the network to generate robust feature representations by simultaneously learning original features and adversarial features in SA module. On the other hand, we design the IL module for better information blending across modalities by interactively sharing the network weights to generate more discriminative and compact features.
- We design an updatable LP mechanism into our proposed end-to-end networks by progressively optimizing pseudo-labels to further find a better decision boundary.
- We validate the superiority and effectiveness of X-ModalNet on two cross-modal datasets with extensive ablation analysis, where we collected and processed the Sentinel-1 SAR data for the second datasets.

2. Related work

2.1. Scene parsing

Most recently, the research on scene parsing has made unprecedented progress, owing to the powerful DNNs (Krizhevsky et al., 2012). Most of these state-of-the-art DL-based frameworks for scene
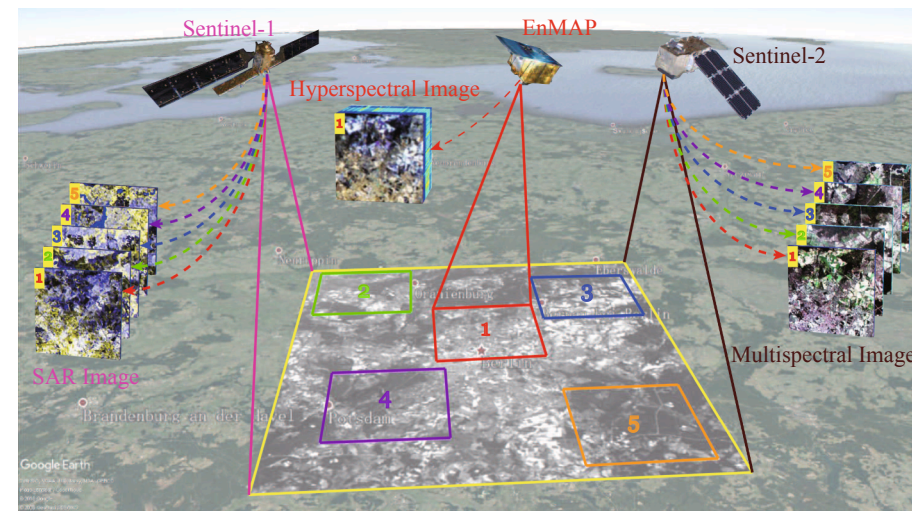
Fig. 1. Our proposed solution (bottom) for the cross-modality learning problem in RS (top). *Top*: Given a large-scale urban area in yellow, both SAR in magenta and MSI in chestnut are openly and largely available with a high spatial resolution but limited by poor feature discrimination power, while the HSI in red is information-rich but only a small-scale available, as shown in area 1 overlapped with SAR (or MSI). *Bottom*: The model is trained on multimodalities (e.g., HSI-MSI or HSI-SAR) with the sparse training labels, and one modality is absent in the proces.s of predicting. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)
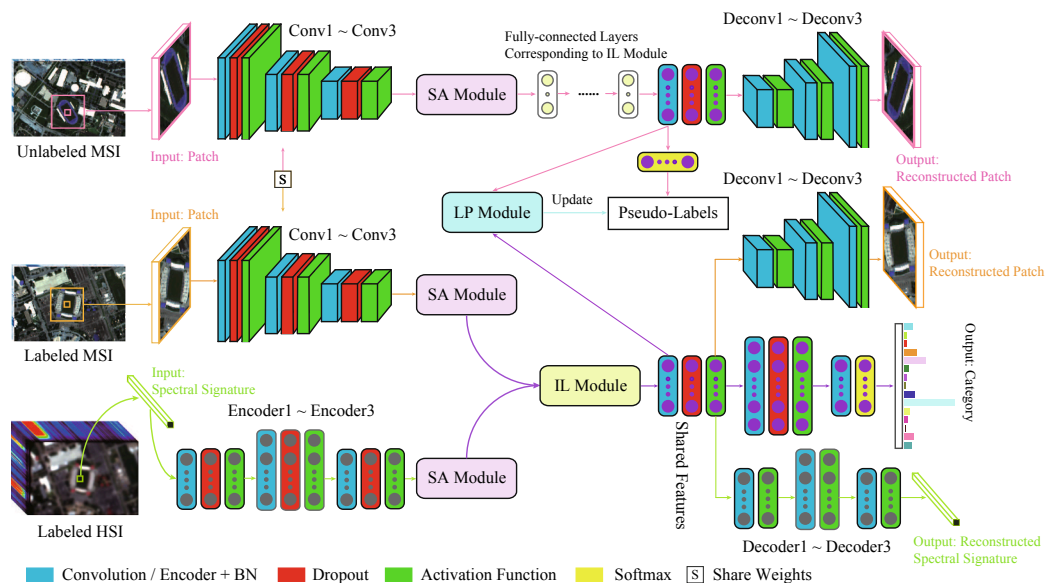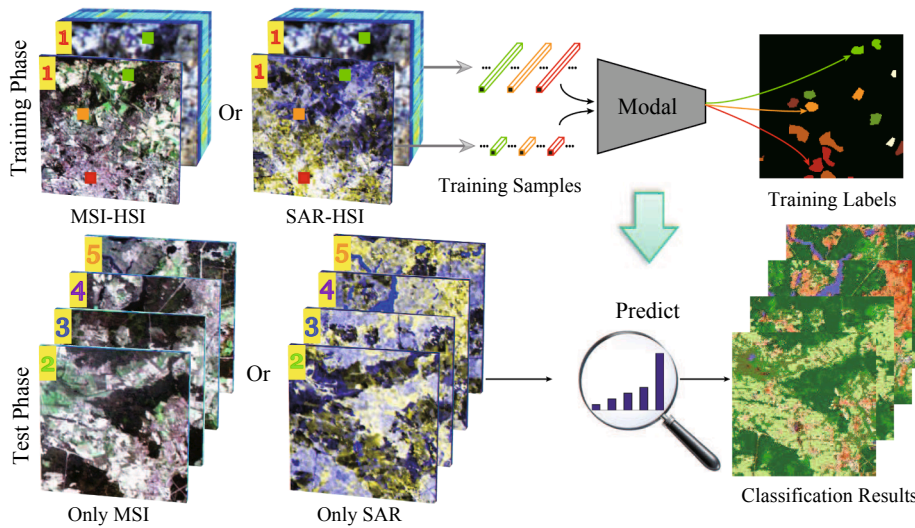

Fig. 2. An overview of the proposed X-ModalNet. It mainly consists of three modules: (a) SA module, (b) IL module, and (c) LP module, installed in a hybrid (MSI or SAR: CNN and HSI: DNN) semi-supervised multimodal DL framework.

parsing (Yu and Koltun, 2015; Noh et al., 2015; Xia et al., 2016; Zhao et al., 2017; Li et al., 2017; Zhang et al., 2018a; Chen et al., 2018; Rasti et al., 2020) are closely associated with two seminal works presented on the prototype of deep CNN: fully convolutional network (Long et al., 2015), DeepLab (Chen et al., 2018). However, a nearly horizontal field of vision makes it difficult to parse a large urban area without extremely diverse training samples. Therefore, RS images might be a feasible and desirable alternative.

We observed that the RS imagery has attracted increasing interest in computer vision field (Lanaras et al., 2015; Xia et al., 2018; Marcos et al., 2018), as it generally holds a diversified and structured source of information, which can be used for better scene understanding and further make a significant breakthrough in global urban motoring and planning (Demir et al., 2018). Chen et al. (2014) fed the vector-based input into a DNN for predicting the category labels in the HSI. They extended their work by training a CNN to achieve a spatial-spectral HSI classification CNN (Chen et al., 2016). Hang et al. (2019) utilized a Cascaded RNN to parse the HSI scenes. Perceptibly, the scene parsing in Earth Vision is normally performed by training an end-to-end network with a vector-based or a patch-based input, as the sparse labels (see Fig. 1) cannot support us to train a FCN-like model. As listed in Table 1, RS images are noisy but low resolution, and are relatively expensive and time-consuming in labeling, limiting the performance improvement. A feasible solution to the issue is to introduce other modalities (e.g., HSI) with more discriminative information, yielding multimodal data analysis.

2.2. Multi/cross-modal learning

Multimodal representation learning related to DNN can be categorized into two aspects (Baltrušaitis et al., 2018).

2.2.1. Joint representation learning

The basic idea is to find a joint space where the discriminative feature representation is expected to be learned over multi-modalities with multilayered neural networks. Although some recent works have attempted to challenge the CML issue by using joint representation learning strategy, e.g., Hong et al. (2019c, 2020a), yet these methods remain limited in data representation and fusion, particularly for heterogeneous data, due to their linearized modeling. A representative work in the multimodal deep learning (MDL) was proposed by Ngiam et al. (2011), in which the high-level features for each modality are extracted using a stacked denoising autoencoder (SDAE) and then jointly learned to a multimodal representation by an additional encoder layer. Silberer and Lapata (2014) extended the work to a semi-supervised version by additionally using a term into loss function that predicts the labels. Similarly, Srivastava et al. utilized the deep belief network (Srivastava and Salakhutdinov, 2012a) and deep Boltzmann machines (Srivastava and Salakhutdinov, 2012b) to explain the multimodal data fusion or learning from the perspective of probabilistic graphical models. In Rastegar et al. (2016), a novel multimodal DL with cross weights (MDL-CW) is proposed to interactively represent the multimodal features for a more effective information blending. Besides, some follow-up work has been successively proposed to learn the joint feature representation more effectively and efficiently (Ouyang et al., 2014; Wang et al., 2014; Peng et al., 2016; Silberer et al., 2017; Luo et al., 2017; Liu et al., 2019).

2.2.2. Coordinated representation learning

It builds the disjunct subnetworks to learn the discriminative features independently for each modality and couples them by enforcing various structured constraints onto the resulting encoder layers. These structures can be measured by similarity (Frome et al., 2013; Feng et al., 2014), correlation (Chandar et al., 2016), and sequentiality (Vendrov et al., 2015), etc.

In recent years, some tentative work has been proposed for multimodal data analysis in RS (Gómez-Chova et al., 2015; Kampffmeyer et al., 2016; Máttyus et al., 2016; Audebert et al., 2016, 2017; Zampieri et al., 2018; Ghosh et al., 2018). Related to ours for scene parsing with multimodal deep networks, an early deep fusion architecture, simply stacking all multi-modalities as input, is used for semantic segmentation of urban RS images (Kampffmeyer et al., 2016). In Audebert et al. (2017), optical and OpenStreetMap (Haklay and Weber, 2008) data are jointly applied with a two-stream deep network for getting a faster and better semantic map. Audebert et al. (2018) parsed the urban scenes under the SegNet-like architecture (Badrinarayanan et al., 2017) by using MSI and Lidar. Similarly, Ghosh et al. (2018) proposed a stacked U-Nets for material segmentation of RS imagery. Nevertheless, these methods are mostly developed with optical (MSI or RGB) or Lidar data for the rough-grained scene parsing (only few categories) and fail to perform sufficiently well in a complex urban scene due to the relatively poor feature representation ability behind the networks, especially in CML (Ngiam et al., 2011).

2.3. Semi-supervised learning

Considering the fact that the labeling cost is very expensive, particularly for RS images, the use of unlabeled samples has gathered increasing attention as a feasible solution to further improve the classification performance of RS data. There have been many non-DL-based semi-supervised learning approaches in a variety of RS-related applications, such as regression-based multitask learning (Hong et al., 2019a; Hong, 2019), manifold alignment (Tuia et al., 2014, 2019b), factor analysis (Zhao et al., 2019). Yet this topic is less investigated by using the DL-based approaches. Cao et al. (2020b) integrated CNNs and active learning to better utilize the unlabeled samples for hyperspectral image classification. Riese et al. (2020) developed a semi-supervised shallow network – self-organizing map framework – to classify and estimate physical parameters from MSI and HSI. Nevertheless, how to embed the semi-supervised techniques into deep networks more effectively remains challenging.
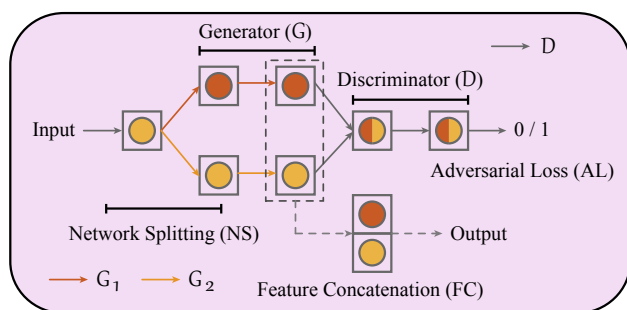
3. The proposed X-ModalNet

The CML's problem setting drives us to develop a robust and discriminative network for pixel-wise classification of RS images in complex scenes. Fig. 2 illustrates the architecture overview of the X-ModalNet, which is built upon a three-stream deep architecture. The *IL* module is designed for highly compact feature blending before feeding the features of each modality into joint representation, and we also equip with the *SA* module and an iterative LP mechanism to improve the robustness and the generalization ability of the proposed X-ModalNet, particularly in the presence of noisy samples.
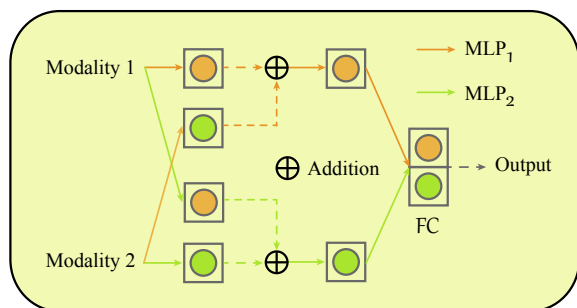
3.1. Network architecture

The bimodal deep autoencoder (DAE) in Ngiam et al. (2011) is a well-known work in MDL, and we advance it to the proposed X-ModalNet for classification of RS imagery. The differences and improvements mainly lie in four aspects.

3.1.1. Hybrid network architecture

Similarly to Cangea et al. (2017), we propose a hybrid-stream network architecture in a bimodal DAE fashion, including two CNN-streams on the labeled MSI (SAR) and unlabeled one, and a DNN-stream on HSI, to exploit high spatial information of MSI/SAR data and high spectral information of HSI more effectively. Since hyperspectral imaging enables discrimination between spectrally similar classes (high-spectral resolution) but its swath width from space is narrow compared to multispectral or SAR ones (high-spatial resolution). More specifically, we take the patches centered by pixels as the input of CNN-streams for labeled and unlabeled MSIs (SARs), and the spectral

(a) Self-Adversarial (SA) module

Fig. 3. An illustration for three proposed modules in the X-ModalNet: (a) SA module, (b) IL module, and (c) LP module. The arrowed solid lines denote the to-be-learned parameters, and their colors mean the different streams in (a) or modalities in (b). Note that MLP is the abbreviation of multi-layer perception (Pal and Mitra, 1992). For example to see modality 1 in (b), modality 1 reaches the hidden layer (yellow) through the parameters and meanwhile modality 2 reaches the hidden layer (magenta) through the .same parameters. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

(b) Interactive learning (IL) module

(c) Label propagation (LP) module

signatures of the corresponding pixels as the input of DNN-stream for labeled HSI. Moreover, the reconstructed patches (CNN-streams) and spectral signatures (DNN-stream) of all pixels as well as the one-hot encoded labels can be regarded as the network outputs.

3.1.2. Self-adversarial module

Due to the environmental factors (e.g., illumination, physically and chemically atmospheric effects) and instrumental errors, it is inevitable to have some distortions in RS imaging. These noisy images tend to generate attacked samples, thereby hurting the network performance (Szegedy et al., 2013; Goodfellow et al., 2014b; Melis et al., 2017). Unlike the previous adversarial training approaches (Donahue et al., 2016; Biggio and Roli, 2018) that generate adversarial samples in the first place and then feed them into a new network for training, we learn the adversarial information in the feature-based level rather than the sample-based one, with an end-to-end learning process. This might lead to a more robust feature representation in accord with the learned network parameters. As illustrated in Fig. 3(a), given a vector-based feature input of the module, the network is first split into two streams (NS). It is well-known that the discriminator in GANs enables the generation of adversarial examples to fool the networks. Inspired by it, we assume that in our SA module, one stream extracts or generates the high-level features of the input, while another one correspondingly learns the adversarial features by allowing for an adversarial loss on the top layer (AL). In this process, the discriminator can be well regarded as a constraint to achieve the function. In addition, this has been also proven to be effective by the reference Yu et al. (2019) to a great extent. Finally, the features represented from the two subnetworks are concatenated as the module output (FC) in order to generate more robust feature representations by simultaneously considering the original features and its adversarial features into the network training. Moreover, the superiority of our SA module mainly lies in that the parameters in the module is an end-to-end trainable in the whole X-ModalNet, which can make the learned adversarial features more suitable for our classification tasks. By contrary, if we select to first generate adversarial samples by using an independent GAN and feed them into the classification network together with existing real samples, then the

generated adversarial samples could bring the uncertainty for the classification performance improvement. The main reason is that the adversarial samples are generated by an independent GAN, which might be applicable to the GAN but might not be applicable to the classification network because they are trained separately.

3.1.3. Interactive learning module

We found that in the layer of multimodal joint representation, massive connections occur in variables from the same modality but few neurons across the modalities are activated, even if each modality passes through multiple individual hidden layers before being fed into the joint layer. Different from the *hard-interactive mapping learning* in Rastegar et al. (2016) and Nie et al. (2018) that additionally learns the weights across the different modalities, we propose a *soft-interactive learning strategy* that directly copies the weights learned from one modality to another one without additional computational cost and information loss, then fuses them on the top layer only with a simple addition operation, as illustrated in Fig. 3(b). This would be capable of learning the inter-modality corrections both effectively and efficiently by reducing the gap between the modalities, yielding a smooth multi-stream networks blending.

3.1.4. Label propagation module

Beyond the supervised learning, we also consider the unlabeled samples by incorporating the label propagation (see Fig. 3(c)) into the networks to further improve the model's generalization. The main workflow in the LP module is detailed as follows:

- We first train a classifier on the training set (SVMs used in our case) and predict unlabeled samples by using the trained classifier. These predicted results (pseudo-labels) can be regarded as the network ground truth of unlabeled data stream, which is further considered with real labels into the network training for a multitask learning.
- Next, we start to train our networks until convergence occurs. We call this process as one-round network training. Once one-round network training has been completed, the high-level features extracted from the top of the network (see Fig. 2) are used to update

Table 2
Network configuration in each layer of X-ModalNet. FC, Conv, and BN are abbreviations of fully connected, convolution, and batch normalization, respectively. The symbols of '↔' and '–' represent the parameter sharing and no operations, respectively. Moreover, d_1 and d_2, denote the dimensions of MSI/ SAR and HSI, and C is the number of class. Please note that the reconstruction happens after passing through the first block of prediction module.

	X-ModalNet				
Pathway	Labeled MSI/ SAR (d_1)		Unlabeled MSI/ SAR (d_1)		Labeled HSI (d_2)
Feature Extractor	5 × 5 Conv + BN + Dropout Tanh (32)	↔	5 × 5 Conv + BN + Dropout Tanh (32)	–	FC Encoder + BN + Dropout Tanh (160)
	3 × 3 Conv + BN + Dropout Tanh (64)	↔	3 × 3 Conv + BN + Dropout Tanh (64)	–	FC Encoder + BN + Dropout Tanh (64)
SA Module	FC Encoder + BN + Dropout Tanh (128)	↔	FC Encoder + BN + Dropout Tanh (128)	–	FC Encoder + BN + Dropout Tanh (128)
	FC Encoder + BN + Dropout Tanh (64)	↔	FC Encoder + BN + Dropout Tanh (64)	–	FC Encoder + BN + Dropout Tanh (64)
IL Module	FC Encoder + BN + Dropout Tanh (64)	↔	FC Encoder + BN + Dropout Tanh (64)	–	FC Encoder + BN + Dropout Tanh (64)
	FC Encoder + BN + Dropout Tanh (64)	↔	FC Encoder + BN + Dropout Tanh (64)	–	FC Encoder + BN + Dropout Tanh (64)
Prediction	FC Encoder + BN + Dropout Tanh (128)	↔	FC Encoder + BN + Dropout Tanh (128)	–	FC Encoder + BN + Dropout Tanh (128)
	FC Encoder + BN + Dropout Tanh (256)		FC Encoder + Softmax		FC Encoder + BN + Dropout Tanh (256)
	FC Encoder + BN + Dropout Tanh (64)				FC Encoder + BN + Dropout Tanh (64)
	FC Encoder + Softmax Tanh (C)				FC Encoder + Softmax Tanh (C)
Reconstruction	FC Encoder + BN Tanh (64)	↔	FC Encoder + BN Tanh (64)	–	FC Encoder + BN Tanh (64)
	3 × 3 Conv + BN Tanh (32)	↔	3 × 3 Conv + BN Tanh (32)	–	FC Encoder + BN Tanh (160)
	5 × 5 Conv + BN Sigmoid (d_1)	↔	5 × 5 Conv + BN Sigmoid (d_1)	–	FC Encoder + BN Sigmoid (d_2)

the pseudo-labels using the graph-based LP (Zhu et al., 2005). The LP algorithm consists of the following two steps.

– Step 1: construct similarity matrix. The similarity matrix $\mathbf{S}$ between any two samples (Hong et al., 2015), e.g., x_i and x_j, either labeled or unlabeled, is computed by

$$\mathbf{S}_{i,j} = (\exp - \frac{||i - x_j||^2}{\sigma^2}), \tag{1}$$

where σ is a hyperparameter determined from the range of [0.001, 0.01, 0.1, 1, 10, 100] by cross-validation on the training set.

– Step 2: propagate labels over all samples. Before carrying out LP, a label transfer matrix ($\mathbf{P}$), e.g., from the sample i to the sample j, is defined as

$$\mathbf{P}_{i,j} := \mathbf{P}(i \rightarrow j) = \frac{\mathbf{S}_{i,j}}{\sum_{k=1}^{N} \mathbf{S}_{i,k}}, \tag{2}$$

where N is the number of samples. Assume that given M labeled and $N - M$ unlabeled samples with C categories, a soft label matrix $\mathbf{Y} \in \mathbb{R}^{N \times C}$ is constructed, which consists of a labeled matrix $\mathbf{Y}_l \in \mathbb{R}^{M \times C}$ and a unlabeled matrix $\mathbf{Y}_u \in \mathbb{R}^{(N-M) \times C}$ obtained by one-hot encoding. Our goal is to update the matrix $\mathbf{Y}$, we then have the update rule in the t-th ($t \geq 1$) iteration as follows: (1) update $\mathbf{Y}^t$ by $\mathbf{PY}^{t-1}$; (2) reset $\mathbf{Y}_l^t$ in $\mathbf{Y}^t$ using the original $\mathbf{Y}_l$ as $\mathbf{Y}_l^t = \mathbf{Y}_l$; (3) repeat the steps (1) and (2) until convergence. We re-feed these updated pseudo-labels, i.e., $\mathbf{Y}_u$ into the next-round network training. The workflow is run repeatedly until the pseudo-labels are not changed any more. Note that we experimentally found that three to four repetitions are usually enough, leading to the model convergence.

3.2. Objective function

Let $\mathbf{x}_{SA}$ and $\mathbf{z}_{SA}$ be the input and output of the SA module, and then we have

$$\mathbf{z}_{SA}^1 = G_1(\mathbf{x}_{SA}),$$
$$\mathbf{z}_{SA}^2 = G_2(\mathbf{x}_{SA}),$$
$$\mathbf{z}_{SA} = [\mathbf{z}_{SA}^1, \ \mathbf{z}_{SA}^2], \tag{3}$$

where G is the generative subnetwork that consists of several encoder, normalization (BN) (Ioffe and Szegedy, 2015) and dropout (Srivastava et al., 2014) layers (see Fig. 3). Given the inputs of two modalities $\mathbf{x}_{IL}^1$ and $\mathbf{x}_{IL}^2$ in the IL module, its output ($\mathbf{z}_{IL}$) can be formulated by

$$\mathbf{z}_{IL}^{11} = MLP_1(\mathbf{x}_{IL}^1), \mathbf{z}_{IL}^{22} = MLP_2(\mathbf{x}_{IL}^2), \mathbf{z}_{IL}^{12} = MLP_2(\mathbf{x}_{IL}^1), \mathbf{z}_{IL}^{21} = MLP_1(\mathbf{x}_{IL}^2), \mathbf{z}_{IL}$$
$$= [\mathbf{z}_{IL}^{11} + \mathbf{z}_{IL}^{12}, \ \mathbf{z}_{IL}^{22} + \mathbf{z}_{IL}^{21}], \tag{4}$$

where MLP, namely multi-layer perception, holds a same structure with G in Eq. (3), as illustrated in Fig. 3. We define the different modalities as $\mathbf{x}_i$ where $i \in \{o, t, u\}$ stands for the first modality, the second modality, the unlabeled samples, and the corresponding l-th hidden layer as $\mathbf{z}_i^{(l)}$. Accordingly, the network parameters can be updated by jointly optimizing the following overall loss function.

$$L = L_l + L_{pl} + L_{rec} + L_{adv}, \tag{5}$$

where L_l is the cross-entropy loss for labeled samples while L_{pl} for pseudo-labeled samples. In addition to the two loss functions that connect the input data with labels (or pseudo-labels), we consider the reconstruction loss (L_{rec}) for each modality as well as unlabeled samples.

$$L_{rec} = ||\mathbf{x}_o - \hat{\mathbf{x}}_o||_2^2 + ||\mathbf{x}_t - \hat{\mathbf{x}}_t||_2^2 + ||\mathbf{x}_u - \hat{\mathbf{x}}_u||_2^2, \tag{6}$$

where $\hat{\mathbf{x}}_i$ denotes the reconstructed data of $\mathbf{x}_i$. For the adversarial loss (L_{adv}), it acts on the SA module formulated based on GANs as

$$L_{adv} = L_{adv}^o + L_{adv}^t + L_{adv}^u,$$
$$L_{adv}^i = \max_{D_i} \mathbb{E}[log(D_i(\mathbf{z}_i^r)) + log(1 - D_i(\mathbf{z}_i^f))], \tag{7}$$

where D_i represents the discriminator in adversarial training. Linking with Eq. (3), $\mathbf{z}_i^r = (\mathbf{z}_{SA}^1)_i$ and $\mathbf{z}_i^f = (\mathbf{z}_{SA}^2)_i$ are a real/ fake pair of data representation on the last layers of SA module.
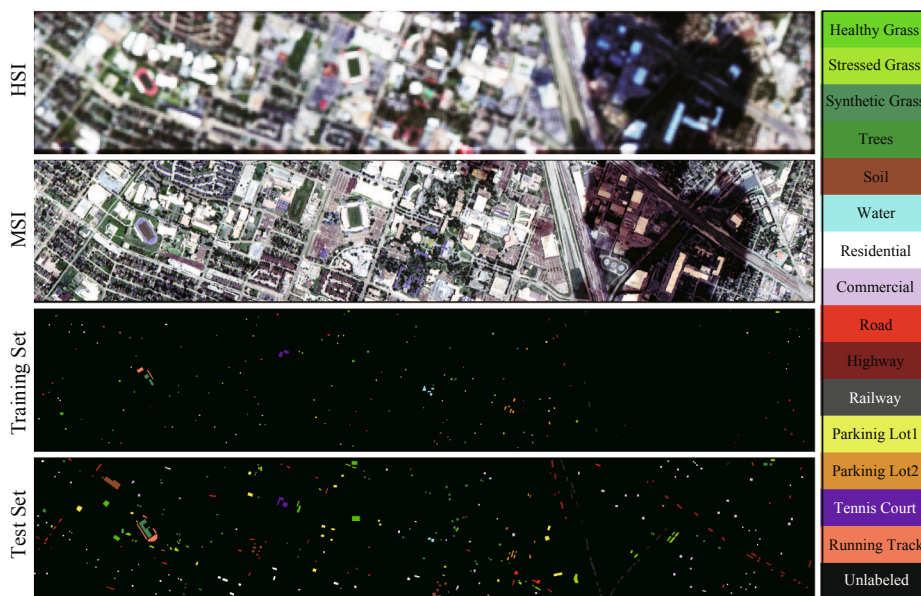
(a) HSI-MSI datasets
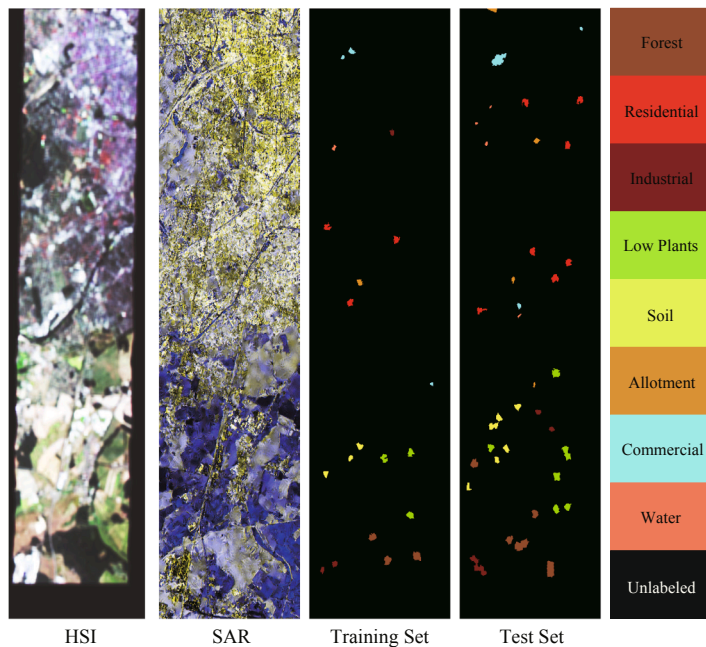
(b) HSI-SAR datasets

Fig. 4. Exemplary datasets for HSI-MSI and HSI-SAR: false-color images and corresponding training and test labels.

3.3. Model architecture

The X-ModalNet starts with a *feature extractor*: two convolution layers with 5×5 and 3×3 convolutional kernels for MSI or SAR pathway and two fully-connected layers for HSI pathway, and then passes through the *SA* module with two fully-connected layers. Following it, an *IL* module with two fully-connected layers is connected over the previous outputs. In the end, four fully-connected layers with an additional soft-max layer are applied to bridge the hidden layers with one-hot encoded labels. Table 2 details the network configuration for each layer in X-ModalNet.

4. Experiments

4.1. Data description

We evaluate the performance of the X-ModalNet on two different datasets. Fig. 4 shows the false-color images for both datasets as well as the corresponding training and test ground truth maps, while scene categories and the number of training and test samples are detailed in Table 3. **There are two things particularly noteworthy in our CML's setting:** (1) *vector (or patch)-based* input due to the sparse groundtruth maps; (2) we assume that *the HSI is present only in the process of training and it is absent in the test phase.*

Table 3
The number of training and test samples on two datasets.

Dataset	HSI-MSI			HSI-SAR		
No.	Class	Training	Test	Class	Training	Test
1	Healthy Grass	537	699	Forest	1437	3249
2	Stressed Grass	61	1154	Residential	961	2373
3	Synthetic Grass	340	357	Industrial	623	1510
4	Tree	209	1035	Low Plants	1098	2681
5	Soil	74	1168	Soil	728	1817
6	Water	22	303	Allotment	260	747
7	Residential	52	1203	Commercial	451	1313
8	Commercial	320	924	Water	144	256
9	Road	76	1149	–	–	–
10	Highway	279	948	–	–	–
11	Railway	33	1185	–	–	–
12	Parking Lot1	329	904	–	–	–
13	Parking Lot2	20	449	–	–	–
14	Tennis Court	266	162	–	–	–
15	Running Track	279	381	–	–	–
	Total	2832	12,197	Total	5702	13,946

4.1.1. Homogeneous HSI-MSI dataset

The HSI scene that has been widely used in many works (Hong et al., 2017, 2020b) consists of 349×1905 pixels with 144 spectral bands in the wavelength range from 380 nm to 1050 nm at a ground sampling distance (GSD) of 10 m (low spatial-resolution), while the aligned MSI with the dimensions of $349 \times 1905 \times 8$ is obtained at a GSD of 2.5 m (high spatial-resolution).

Spectral simulation is performed to generate the low-spectral re-solution MSI by degrading the reference HSI in the spectral domain using the MS spectral response functions of Sentinel-2 as filters. Using this, the MSI consists of 349×1905 pixels with eight spectral bands at a GSD of 2.5 m.

Spatial simulation is performed to generate the low-spatial resolution HSI by degrading the reference HSI in the spatial domain using an isotropic Gaussian point spread function, thus yielding the HSI with the dimensions of $349 \times 1905 \times 144$ at a GSD of 10 m by upsampling to the MSI's size.

4.1.2. Heterogeneous HSI-SAR dataset

The EnMap benchmark HSI covering the Berlin urban area is freely available from the website[1]. This image consists of 797×220 pixels with a GSD of 30 m, and 244 spectral bands ranging from 400 nm to 2500 nm. According to the geographic coordinates, we downloaded the same scene of SAR image from the Sentinel-1 satellite, with the size of 1723×476 pixels at a GSD of 13 m and four polarimetric bands (Yamaguchi et al., 2005). The used SAR image is dual-polarimetric SAR data collected by interferometric wide swath mode. It is organized as a commonly used four-component PolSAR covariance matrix (four bands) (Yamaguchi et al., 2005). Note that we upsample the HSI to the same size with the SAR image by the nearest-neighbor interpolation.

4.2. Implementation details

Our approach is implemented on the Tensorflow framework (Abadi et al., 2016). The network configuration, to our knowledge, always plays a critical role in a practical DL system. The model is trained on the training set, and the hyper-parameters are determined using a grid search on the validation set[2]. In the training phase, we adopt the Adam optimizer with the "poly" learning rate policy (Chen et al., 2018). The

---

[1] http://doi.org/10.5880/enmap.2016.002
[2] Ten replications are conducted to randomly split the original training set into the new training and validation sets with the percentage of 8:2 to de-termine the network's hyperparameters.

Table 4
Quantitative performance comparison with baseline models on the HSI-MSI dataset. The best one is shown in bold.

Methods	Pixel Acc. (%)	mIoU (%)
Baseline	70.51	57.84
CCA (Hardoon et al., 2004)	73.01	64.72
Unimodal DAE (Chen et al., 2014)	72.85 ± 1.2	62.75 ± 0.3
Bimodal DAE (Ngiam et al., 2011)	75.43 ± 0.6	67.67 ± 0.1
Bimodal SDAE (Silberer et al., 2017)	79.51 ± 1.7	69.62 ± 0.3
MDL-CW (Rastegar et al., 2016)	83.27 ± 1.0	74.60 ± 0.3
Corr-AE (Feng et al., 2014)	80.49 ± 1.2	70.85 ± 0.2
CorrNet (Chandar et al., 2016)	82.66 ± 0.8	73.48 ± 0.2
X-ModalNet	**88.23 ± 0.7**	**80.31 ± 0.2**

Table 5
Quantitative performance comparison with baseline models on the HSI-SAR datasets. The best one is shown in bold.

Methods	Pixel Acc. (%)	mIoU (%)
Baseline	43.91	18.70
CCA (Hardoon et al., 2004)	36.66	12.04
Unimodal DAE (Chen et al., 2014)	51.51 ± 0.5	29.32 ± 0.2
Bimodal DAE (Ngiam et al., 2011)	56.04 ± 0.5	34.13 ± 0.2
Bimodal SDAE (Silberer et al., 2017)	59.27 ± 0.6	37.78 ± 0.2
MDL-CW (Rastegar et al., 2016)	62.51 ± 0.8	42.15 ± 0.1
Corr-AE (Feng et al., 2014)	60.59 ± 0.5	39.12 ± 0.3
CorrNet (Chandar et al., 2016)	64.65 ± 0.7	44.25 ± 0.3
X-ModalNet	**71.38 ± 1.0**	**54.02 ± 0.3**

current learning rate can be updated by multiplying the base one with $(1 - \frac{iter}{maxIter})^{power}$, where the base learning rate and power are set to 0.0005 and 0.98, respectively. We use the DAE to pretrain the sub-networks for each modality to greatly reduce the training time of the model and find a better local optimum easier. Also, the momentum is set to 0.9.

To facilitate network training and reduce overfitting, BN and dropout techniques are orderly used for all DL-based methods prior to the activation functions. The model training ends up with 150 epochs for the heterogeneous HSI-MSI dataset and 200 epochs for the hetero-geneous HSI-SAR dataset with a minibatch size of 300. Both labeled and unlabeled samples in SAR or MSI share the same network parameters in the process of model optimization.

In the experiments, we found that when the unlabeled samples, from neither training nor test sets, are selected at an approximated scale with the test set, the final classification results are similar to that di-rectly using test set. We have to admit, however, that the full use of unlabeled samples enable further improvement in classification per-formance, but we have to make a trade-off between the limited per-formance improvement and exponentially increasing cost in data sto-rage, transmission, and computation. Moreover, we expect to see the performance gain when using these proposed modules, thereby de-monstrating their effectiveness and superiority. As a result, we, for simplicity, select the test set as the unlabeled set for all semi-supervised compared methods for a fair comparison.

Furthermore, two commonly used indices: *Pixel-wise Accuracy* (Pixel Acc.) and *mean Intersection over Union* (mIoU) are calculated to quan-titatively evaluate the parsing performance by collecting all pixel-wise predictions of the test set. Due to random initialization, both metrics show the average accuracy and the variation of the results out of 10 runs.

4.3. Comparison with state-of-the-art

Several state-of-the-art baselines closely related to our task (CML) are selected for comparison; they are.

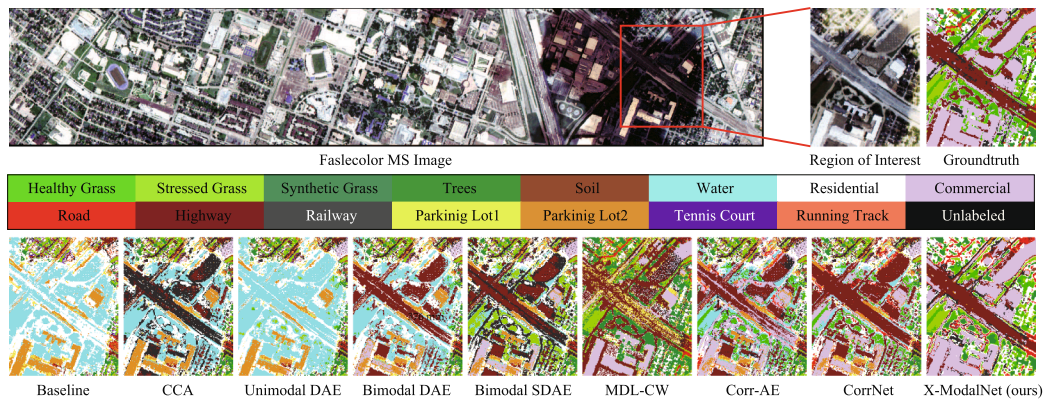(1) **Baseline:** We train a linear SVM classifier directly using original

Healthy Grass	Stressed Grass	Synthetic Grass	Trees	Soil	Water	Residential	Commercial
Road	Highway	Railway	Parkinig Lot1	Parkinig Lot2	Tennis Court	Running Track	Unlabeled

Fig. 5. Classification maps of ROI on HSI-MSI datasets. The ground truth in this highlighted area is manually labelled.

pixel-based MSI or SAR features. Note that the hyperparameters in SVM are determined by 10-fold cross-validation on the training set.

(2) **Canonical Correlation Analysis (CCA)** (Hardoon et al., 2004): We learn a shared latent subspace from two modalities on the training set, and project the test samples from any one of the two modalities into the subspace. This is a typical cross-modal feature learning. Finally, the learned features are fed into a linear SVM. We used the code from the website[3].

(3) **Unimodal DAE** (Chen et al., 2014): This is a classical deep autoencoder. We train a DAE on the target-modality (MSI or SAR) in a unsupervised way, and finely tune it using labels. The hidden representation of the encoder layer is used for final classification. The code we used is from the website[4].

(4) **Bimodal DAE** (Ngiam et al., 2011): As a DL's pioneer to multimodal application, it learns a joint feature representation over the encoder layers generated by AEs for each modality.

(5) **Bimodal SDAE** (Silberer et al., 2017): This is a semi-supervised version for Bimodal DAE by considering the reconstruct loss of all unlabeled samples for each modality and adding an additional soft-max layer over the encoder layer for those limited labeled data.

(6) **MDL-CW** (Rastegar et al., 2016): A end-to-end multimodal network is trained with cross weights acted on the two-stream subnetworks for more effective information blending.

(7) **Corr-AE** (Feng et al., 2014): A coupled AEs are first used to learn a shared high-level feature representation by enforcing similarity constraint between the encoder layers of two modalities. The learned features are then fed into a classifier.

(8) **CorrNet** (Chandar et al., 2016): Similar to Corr-AE, AE is responsible for extracting features of each modality, while CCA serves as a link with the features by maximizing their correlations. The code is available from the website[5].

4.3.1. Results on the homogeneous datasets

Table 4 shows the quantitative performance comparison in terms of Pixel Acc. and mIoU. Limited by the feature diversity, the baseline yields a poor classification performance, while there is a performance improvement (about 2%) in the unimodal DAE due to the powerful learning ability of DL-based techniques. For the homogeneous HSI-MSI correspondences, the linearized CCA is more likely to catch the shared features and obtains the better classification results. The features can be better fused over the hidden representations of two modalities. Therefore, the bimodal DAE improves the performance by 2% on the basis of CCA's. The accuracy of bimodal SDAE can further increase to around 79%, since it aims at training an end-to-end multimodal network to

generate more discriminative features. Different from previous strategies, Corr-AE and CorrNet couple two subnetworks by enforcing the structural measurement on hidden layers, such as Euclidean similarity and correlation, which allows a more effective pixel-wise classification. The MDL-CW with learning cross weights can facilitate the multimodal information fusion, thus achieving better classification results than Corr-AE and CorrNet. As expected, X-ModalNet outperforms these state-of-the-art methods, demonstrating its superiority and effectiveness with a large improvement of at least 6% Pixel Acc. and mIoU over CorrNet (the second best method).

4.3.2. Results on the heterogeneous datasets

Similar to the former dataset, we evaluate the performance for the Heterogeneous HSI-SAR scene quantitatively. Two assessment indices (Pixel Acc. and mIoU) for different algorithms are summarized in Table 5. There is a basically consistent trend in performance improvement of different algorithms. That is, the performance of X-ModalNet is significantly superior to that of others, and the methods with the hyperspectral information perform better than those without one, such as Baseline and Unimodal DAE. It is worth noting that the proposed X-ModalNet brings increments of about 9% Pixel Acc. and 10% mIoU on the basis of CorrNet. Moreover, the CCA fails to linearly represent the heterogeneous data, leading to a worse parsing result and even lower than the baseline. Additionally, the gap (or heterogeneity) between SAR and optical data can be effectively reduced by mutually learning weights. This might explain the case that the MDL-CW observably exceeds most compared methods without such interactive module (nearly 20% over baseline), e.g., Bimodal DAE and its semi-supervised version (Bimodal SDAE) as well as CorrNet.

4.4. Visual comparison

Apart from quantitative assessment, we also make a visual comparison by highlighting a salient region overshadowed by the cloud on the Houston2013 datasets. As shown in Fig. 5, our method is capable of identifying various materials more effectively, particularly for the material *Commercial* in the upper-right of the predicted maps. Besides, a trend can be figured out, that is, the methods with the input of multimodalities achieve more smooth parsing results compared to those with the input of single modalities.

Similarly, we visually show the classification maps of those comparative algorithms in a region of interest in the EnMap datasets, as shown in Fig. 6. We can see that our X-ModalNet shows a more competitive and realistic parsing result, especially in classifying *Soil* and *Plants*, which is more approaching to the real scene.

4.5. Ablation studies

We analyze the performance gain of X-ModalNet by step-wise

---
[3] https://github.com/CommonClimate/CCA
[4] http://deeplearning.stanford.edu/wiki/index.php/UFLDL_Tutorial
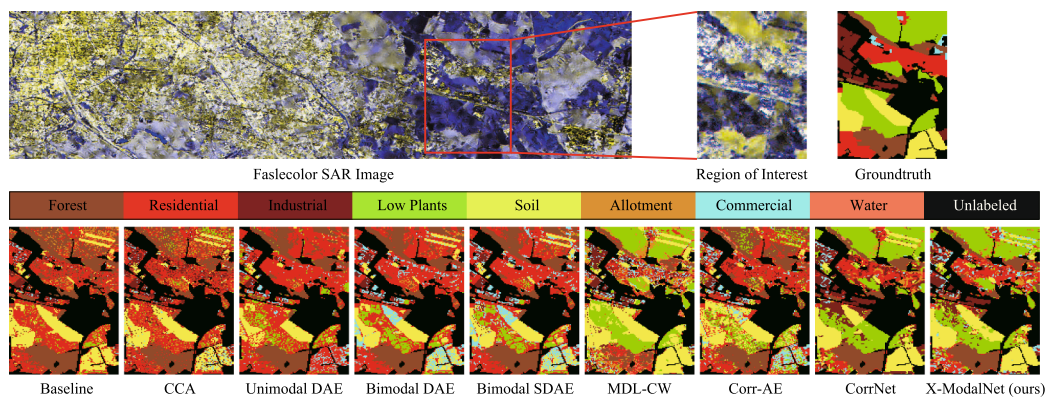[5] https://github.com/apsarath/CorrNet

Fig. 6. Classification maps of ROI on HSI-SAR datasets. The OpenStreetMap (Haklay and Weber, 2008) is used as the ground truth generato.r for this area.

Table 6
Ablation analysis of the X-ModalNet with a combination of different modules in term of Pixel Acc. on two datasets. Moveover, importance analysis in the presence and absence of BN and dropout operations is discussed as well.

Methods	BN	Dropout	IL	LP	SA	Pixel Acc. (%)	
						HSI-MSI	HSI-SAR
X-ModalNet	✓	✓	✗	✗	✗	83.14 ± 0.9	64.44 ± 1.1
X-ModalNet	✓	✓	✓	✗	✗	85.07 ± 0.8	68.73 ± 0.8
X-ModalNet	✓	✓	✓	✓	✗	86.58 ± 1.0	70.19 ± 0.8
X-ModalNet	✓	✓	✓	✓	✓	88.23 ± 0.7	71.38 ± 1.0
X-ModalNet	✗	✗	✓	✓	✓	80.33 ± 0.5	62.47 ± 0.7
X-ModalNet	✗	✓	✓	✓	✓	81.94 ± 0.6	64.10 ± 0.9
X-ModalNet	✓	✗	✓	✓	✓	85.10 ± 0.6	67.34 ± 0.8

adding the different components (or modules). Table 6 lists a progressive performance improvement by gradually embedding different modules, while Fig. 7 correspondingly visualizes the learned features in the latent space (top encoder layer). It is clear to observe that successively adding each component into the X-ModalNet is conducive to a more discriminative feature generation.

We also investigate the importance of dropout and BN techniques in avoiding overfitting and improving network performance. As can be seen in Table 6, turning off the dropout would hinder X-ModalNet from generalizing well, yielding a performance degradation. What is worse is that the classification accuracy without BN reduces sharply. This could result from low-efficiency gradient propagation, thereby hurting the learning ability of the network. Moreover, we can observe from Table 6 that the classification performance without any proposed modules is limited, only yielding about 83.14% and 64.44% Pixel Acc. on the two datasets. It is worth noting that the results achieve an obvious improvement (around 2%~3%) after plugging the IL module. By introducing the semi-supervised mechanism, our LP module can bring increments of 1.5% and 2% Pixel Acc. on the basis of only using the IL module for HSI-MSI and HSI-SAR, respectively. Remarkably, when adding the SA module over the IL and LP modules in networks, our X-ModalNet behaves superiorly and obtains a further dramatic improvement in classification accuracies. These, to a great extent, demonstrate the effectiveness and superiority of several proposed modules as well as their positive effects on the classification performance.
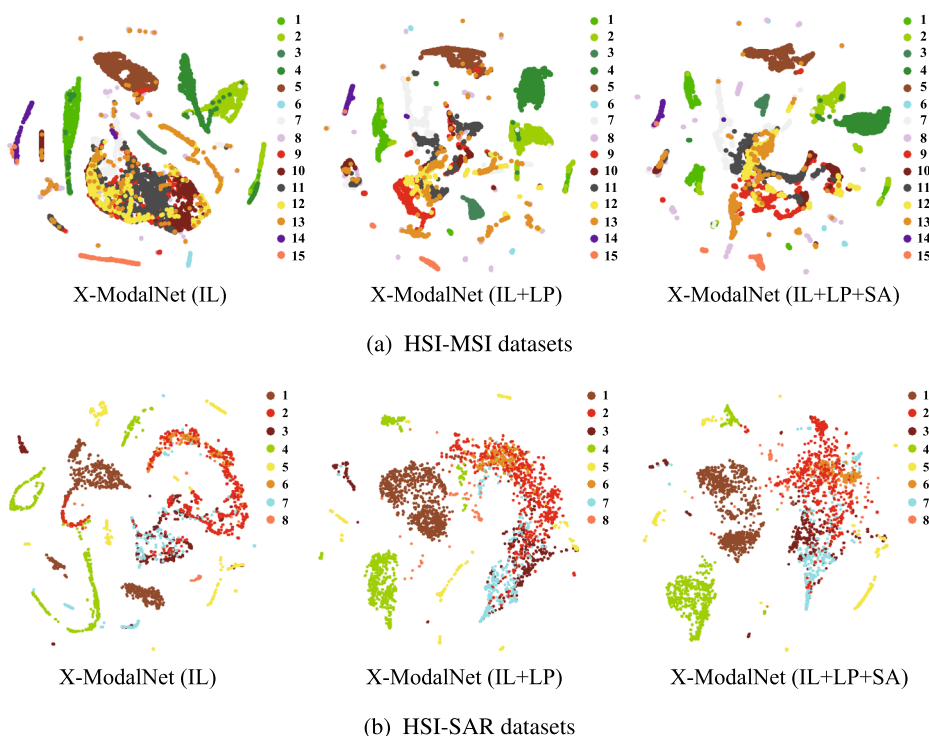

(a) HSI-MSI datasets

(b) HSI-SAR datasets

Fig. 7. t-SNE visualization of the learned multimodal features in the latent space using X-ModalNet with different modules on the two different datasets.
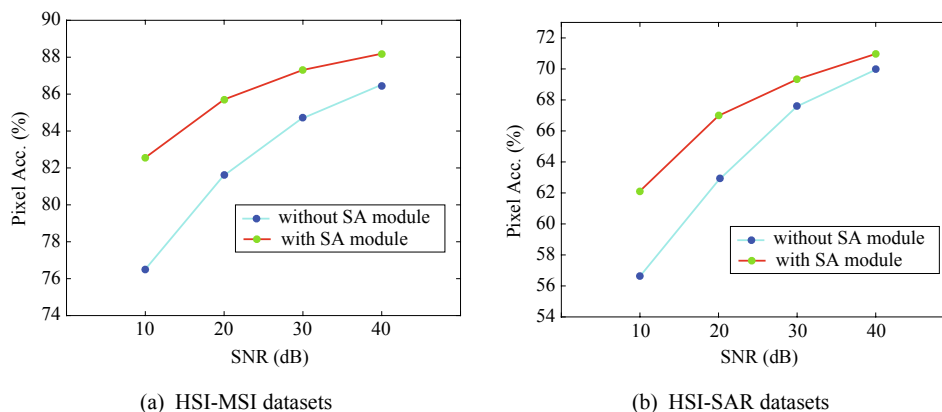
(a) HSI-MSI datasets

(b) HSI-SAR datasets

Fig. 8. Resistance analysis to noise attack using the proposed X-ModalNet with and without *SA* module on the two datasets.

4.6. Robustness to noises

Neural networks have shown their vulnerability to adversarial samples generated by slight perturbation, e.g., imperceptible noises. To study the effectiveness of our SA module against noise or perturbation attack, we simulate the corrupted input by adding Gaussian white noises with different signal-to-noise-ratios (SNRs) ranging from 10 dB to 40 dB at a 10 dB interval. Fig. 8 shows a quantitative comparison in term of Pixel Acc. before and after triggering the *SA* module.

5. Conclusion

In this paper, we investigate the cross-modal classification task by utilizing multimodal satellite or aerial images (RS data). In reality, the HSI is only able to be collected in a locally small area due to the limitations of the imaging system, while MSI and SAR are openly available on a global scale. This motivate us to learn to transfer the HSI knowledge into large-scale MSI or SAR by training the model on both modalities and predict only on one modality. To address the CML's issue in RS, we propose a novel DL-based model X-ModalNet, with two well-designed components (IL and SA modules) to effectively learn a more discriminative feature representation and robustly resist the noise attack, respectively, and with an iteratively updating LP mechanism for further improving the network performance. In the future work, we would like to introduce the physical mechanism of spectral imaging into the network learning for earth observation tasks.

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.

References

Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., Devin, M., Ghemawat, S., Irving, G., Isard, M., 2016. Tensorflow: a system for large-scale machine learning. In: OSDI. vol. 16. pp. 265–283.

Audebert, N., Saux, B.L., Lefèvre, S., 2016. Semantic segmentation of earth observation data using multimodal and multi-scale deep networks. In: Proc. ACCV. Springer, pp. 180–196.

Audebert, N., Saux, B.L., Lefèvre, S., 2017. Joint learning from earth observation and openstreetmap data to get faster better semantic maps. In: Proc. CVPR Workshop. IEEE, pp. 1552–1560.

Audebert, N., Saux, B.L., Lefèvre, S., 2018. Beyond rgb: Very high resolution urban remote sensing with multimodal deep networks. ISPRS J. Photogramm. Remote Sens. 140, 20–32.

Badrinarayanan, V., Kendall, A., Cipolla, R., 2017. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. IEEE Trans. Pattern Anal. Mach. Intell. 12, 2481–2495.

Baltrušaitis, T., Ahuja, C., Morency, L., 2018. Multimodal machine learning: a survey and taxonomy. IEEE Trans. Pattern Anal. Mach, Intell.

Biggio, B., Roli, F., 2018. Wild patterns: ten years after the rise of adversarial machine learning. Pattern Recognit.

Cangea, C., Veličković, P., Liò, P., 2017. Xflow: 1d–2d cross-modal deep neural networks for audiovisual classification. arXiv preprint arXiv:1709.00572.

Cao, X., Yao, J., Fu, X., Bi, H., Hong, D., 2020a. An enhanced 3-dimensional discrete wavelet transform for hyperspectral image classification. IEEE Geosci. Remote Sens. Lett. https://doi.org/10.1109/LGRS.2020.2990407.

Cao, X., Yao, J., Xu, Z., Meng, D., 2020b. Hyperspectral image classification with convolutional neural network and active learning. IEEE Trans. Geosci. Remote Sens. doi:10.1109/TGRS.2020.2964627.

Chandar, S., Khapra, M., Larochelle, H., Ravindran, B., 2016. Correlational neural networks. Neural Comput 28 (2), 257–285.

Chen, Y., Lin, Z., Zhao, X., Wang, G., Gu, Y., 2014. Deep learning-based classification of hyperspectral data. IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens. 7 7 (6), 2094–2107.

Chen, Y., Jiang, H., Li, C., Jia, X., Ghamisi, P., 2016. Deep feature extraction and classification of hyperspectral images based on convolutional neural networks. IEEE Trans. Geosci. Remote Sens. 54 (10), 6232–6251.

Chen, L., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.L., 2018. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. IEEE Trans. Pattern Anal. Mach. Intell. 40 (4), 834–848.

Demir, I., Koperski, K., Lindenbaum, D., Pang, G., Huang, J., Basu, S., Hughes, F., Tuia, D., Raskar, R., 2018. Deepglobe 2018: A challenge to parse the earth through satellite images. In: Proc. CVPR Workshop.

Donahue, J., Krähenbühl, P., Darrell, T., 2016. Adversarial feature learning. arXiv preprint arXiv:1605.09782.

Feng, F., Wang, X., Li, R., 2014. Cross-modal retrieval with correspondence autoencoder. In: Proc. ACMMM. ACM, pp. 7–16.

Frome, A., Shlens, G.S.C.J., s. Bengio, Dean, J., Mikolov, T., 2013. Devise: A deep visual-semantic embedding model. In: Proc. NIPS. pp. 2121–2129.

Gao, L., Yao, D., Li, Q., Zhuang, L., Zhang, B., Bioucas-Dias, J., 2017a. A new low-rank representation based hyperspectral image denoising method for mineral mapping. Remote Sens. 9 (11), 1145.

Gao, L., Zhao, B., Jia, X., Liao, W., Zhang, B., 2017b. Optimized kernel minimum noise fraction transformation for hyperspectral image classification. Remote Sens. 9 (6), 548.

Ghosh, A., Ehrlich, M., Shah, S., Davis, L., Chellappa, R., 2018. Stacked u-nets for ground material segmentation in remote sensing imagery. In: Proc. CVPR Workshop, pp. 257–261.

Gómez-Chova, L., Tuia, D., Moser, G., Camps-Valls, G., 2015. Multimodal classification of remote sensing images: a review and future directions. Proc. IEEE 103 (9),

1560–1584.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y., 2014a. Generative adversarial nets. In: Proc. NIPS, pp. 2672–2680.

Goodfellow, I., Shlens, J., Szegedy, C., 2014b. Explaining and harnessing adversarial examples. arXiv:1412.6572.

Haklay, M., Weber, P., 2008. Openstreetmap: User-generated street maps. IEEE Pervasive Comput. 7 (4), 12–18.

Han, X., Huang, X., Li, J., Li, Y., Yang, M., Gong, J., 2018. The edge-preservation multi-classifier relearning framework for the classification of high-resolution remotely sensed imagery. ISPRS J. Photogramm. Remote Sens. 138, 57–73.

Hang, R., Liu, Q., Hong, D., Ghamisi, P., 2019. Cascaded recurrent neural networks for hyperspectral image classification. IEEE Trans. Geosci. Remote Sens. 57 (8), 5384–5394.

Hardoon, D., Szedmak, S., Shawe-Taylor, J., 2004. Canonical correlation analysis: an overview with application to learning methods. Neural Comput. 16 (12), 2639–2664.

Hong, D., 2019. Regression-induced representation learning and its optimizer: a novel paradigm to revisit hyperspectral imagery analysis. Ph.D. thesis. Technische Universität München.

Hong, D., Zhu, B., 2018. SULoRA: Subspace unmixing with low-rank attribute embedding for hyperspectral data analysis. IEEE J. Sel. Topics Signal Process. 12 (6), 1351–1363.

Hong, D., Liu, W., Su, J., Pan, Z., Wang, G., 2015. A novel hierarchical approach for multispectral palmprint recognition. Neurocomputing 151, 511–521.

Hong, D., Yokoya, N., Zhu, X., 2017. Learning a robust local manifold representation for hyperspectral dimensionality reduction. IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens. 10 (6), 2960–2975.

Hong, D., Yokoya, N., Chanussot, J., Xu, J., Zhu, X.X., 2019a. Learning to propagate labels on graphs: an iterative multitask regression framework for semi-supervised hyper-spectral dimensionality reduction. ISPRS J. Photogramm. Remote Sens. 158, 35–49.

Hong, D., Yokoya, N., Chanussot, J., Zhu, X., 2019b. An augmented linear mixing model to address spectral variability for hyperspectral unmixing. IEEE Trans. Image Process. 28 (4), 1923–1938.

Hong, D., Yokoya, N., Chanussot, J., Zhu, X.X., 2019c. CoSpace: Common subspace learning from hyperspectral-multispectral correspondences. IEEE Trans. Geosci. Remote Sens. 57 (7), 4349–4359.

Hong, D., Yokoya, N., Ge, N., Chanussot, J., Zhu, X., 2019d. Learnable manifold align-ment (LeMA): a semi-supervised cross-modality learning framework for land cover and land use classification. ISPRS J. Photogramm. Remote Sens. 147, 193–205.

Hong, D., Chanussot, J., Yokoya, N., Kang, J., Zhu, X., 2020a. Learning shared cross-modality representation using multispectral-lidar and hyperspectral data. IEEE Geosci. Remote Sens. Lett. doi: 10.1109/LGRS.2019.2944599.

Hong, D., Wu, X., Ghamisi, P., Chanussot, J., Yokoya, N., Zhu, X., 2020b. Invariant at-tribute profiles: a spatial-frequency joint feature extractor for hyperspectral image classification. IEEE Trans. Geosci. Remote Sens. 58 (6), 3791–3808.

Hu, J., Hong, D., Wang, Y., Zhu, X., 2019a. A comparative review of manifold learning techniques for hyperspectral and polarimetric sar image fusion. Remote Sens. 11 (6), 681.

Hu, J., Hong, D., Zhu, X., 2019b. MIMA: Mapper-induced manifold alignment for semi-supervised fusion of optical image and polarimetric sar data. IEEE Trans. Geosci. Remote Sens. 57 (11), 9025–9040.

Ioffe, S., Szegedy, C., 2015. Batch normalization: accelerating deep network training by reducing internal covariate shift. arXiv:1502.03167.

Kampffmeyer, M., Salberg, A., Jenssen, R., 2016. Semantic segmentation of small objects and modeling of uncertainty in urban remote sensing images using deep convolu-tional neural networks. In: Proc. CVPR Workshop, pp. 1–9.

Kang, J., Hong, D., Liu, J., Baier, G., Yokoya, N., Demir, B., 2020. Learning convolutional sparse coding on complex domain for interferometric phase restoration. IEEE Trans. Neural Netw. Learn. Syst. doi:10.1109/TNNLS.2020.2979546.

Krizhevsky, A., Sutskever, I., Hinton, G., 2012. Imagenet classification with deep con-volutional neural networks. In: Proc. NIPS, pp. 1097–1105.

Lanaras, C., Baltsavias, E., Schindler, K., 2015. Hyperspectral super-resolution by coupled spectral unmixing. In: Proc. ICCV, pp. 3586–3594.

LeCun, Y., Bengio, Y., Hinton, G., 2015. Deep learning. Nature 521 (7553), 436.

Li, X., Jie, Z., Wang, W., Liu, C., Yang, J., Shen, X., Lin, Z., Chen, Q., Yan, S., Feng, J., 2017. Foveanet: Perspective-aware urban scene parsing. In: Proc. ICCV, pp. 784–792.

Liu, X., Deng, C., Chanussot, J., Hong, D., Zhao, B., 2019. Stfnet: A two-stream con-volutional neural network for spatiotemporal image fusion. IEEE Trans. Geosci. Remote Sens. 57 (9), 6552–6564.

Long, J., Shelhamer, E., Darrell, T., 2015. Fully convolutional networks for semantic segmentation. In: Proc. CVPR, pp. 3431–3440.

Luo, Z., Zou, Y., Hoffman, J., Fei-Fei, L., 2017. Label efficient learning of transferable representations acrosss domains and tasks. In: Proc. NIPS, pp. 165–177.

Marcos, D., Tuia, D., Kellenberger, B., Zhang, L., Bai, M., Liao, R., Urtasun, R., 2018. Learning deep structured active contours end-to-end. In: Proc. CVPR, pp. 8877–8885.

Máttyus, G., Wang, S., Fidler, S., Urtasun, R., 2016. Hd maps: Fine-grained road seg-mentation by parsing ground and aerial images. In: Proc. ICCV, pp. 3611–3619.

Melis, M., Demontis, A., Biggio, B., Brown, G., Fumera, G., Roli, F., 2017. Is deep learning safe for robot vision? adversarial examples against the icub humanoid. In: Proc. ICCV, pp. 751–759.

Ngiam, J., Khosla, A., Kim, M., Nam, J., Lee, H., Ng, A., 2011. Multimodal deep learning. In: Proc. ICML, pp. 689–696.

Nie, X., Feng, J., Yan, S., 2018. Mutual learning to adapt for joint human parsing and pose estimation. In: Proc. ECCV, pp. 502–517.

Noh, H., Hong, S., Han, B., 2015. Learning deconvolution network for semantic seg-mentation. In: Proc. ICCV, pp. 1520–1528.

Ouyang, W., Chu, X., Wang, X., 2014. Multi-source deep learning for human pose esti-mation. In: Proc. CVPR, pp. 2329–2336.

Pal, S., Mitra, S., 1992. Multilayer perceptron, fuzzy sets, and classification. IEEE Trans. Neural Netw. 3 (5), 683–697.

Peng, Y., Huang, X., Qi, J., 2016. Cross-media shared representation by hierarchical learning with multiple deep networks. In: Proc. IJCAI, pp. 3846–3853.

Rastegar, S., Soleymani, M., Rabiee, H., Shojaee, S.M., 2016. MDL-CW: A multimodal deep learning framework with cross weights. In: Proc. CVPR, pp. 2601–2609.

Rasti, B., Hong, D., Hang, R., Ghamisi, P., Kang, X., Chanussot, J., Benediktsson, J., 2020. Feature extraction for hyperspectral imagery: The evolution from shallow to deep (overview and toolbox). IEEE Geosci. Remote Sens. Mag. doi: 10.1109/MGRS.2020.2979764.

Riese, F., Keller, S., Hinz, S., 2020. Supervised and semi-supervised self-organizing maps for regression and classification focusing on hyperspectral data. Remote Sens. 12 (1), 7.

Silberer, C., Lapata, M., 2014. Learning grounded meaning representations with auto-encoders. In: Proc. ACL. vol. 1. pp. 721–732.

Silberer, C., Ferrari, V., Lapata, M., 2017. Visually grounded meaning representations. IEEE Trans. Pattern Anal. Mach. Intell. 39 (11), 2284–2297.

Srivastava, N., Salakhutdinov, R., 2012a. Learning representations for multimodal data with deep belief nets. In: Proc. ICML Workshop. vol. 79.

Srivastava, N., Salakhutdinov, R., 2012b. Multimodal learning with deep boltzmann machines. In: Proc. NIPS, pp. 2222–2230.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., Salakhutdinov, R., 2014. Dropout: a simple way to prevent neural networks from overfitting. J. Mach. Learn. Res. 15 (1), 1929–1958.

Srivastava, S., Vargas-Mu noz, J., Tuia, D., 2019. Understanding urban landuse from the above and ground perspectives: a deep learning, multimodal solution. Remote Sens. Environ. 228, 129–143.

Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., Fergus, R., 2013. Intriguing properties of neural networks. arXiv:1312.6199.

Tuia, D., Volpi, M., Trolliet, M., Camps-Valls, G., 2014. Semisupervised manifold align-ment of multimodal remote sensing images. IEEE Trans. Geosci. Remote Sens. 52 (12), 7708–7720.

Tuia, D., Flamary, R., Courty, N., 2015. Multiclass feature learning for hyperspectral image classification: sparse and hierarchical solutions. ISPRS J. Photogramm. Remote Sens. 105, 272–285.

Vendrov, I., Kiros, R., Fidler, S., Urtasun, R., 2015. Order-embeddings of images and language. arXiv:1511.06361.

Wang, W., Ooi, B.C., Yang, X., Zhang, D., Zhuang, Y., 2014. Effective multi-modal re-trieval based on stacked auto-encoders. Proc. VLDB 7 (8), 649–660.

Wu, X., Hong, D., Tian, J., Chanussot, J., Li, W., Tao, R., 2019. ORSIm Detector: A novel object detection framework in optical remote sensing imagery using spatial-fre-quency channel features. IEEE Trans. Geosci. Remote Sens. 57 (7), 5146–5158.

Wu, X., Hong, D., Chanussot, J., Xu, Y., Tao, R., Wang, Y., 2020. Fourier-based rotation-invariant feature boosting: an efficient framework for geospatial object detection. IEEE Geosci. Remote Sens. Lett. 17 (2), 302–306.

Xia, F., Wang, P., Chen, L., Yuille, A.L., 2016. Zoom better to see clearer: Human and object parsing with hierarchical auto-zoom net. In: Proc. ECCV. Springer, pp. 648–663.

Xia, G., Bai, X., Ding, J., Zhu, Z., Belongie, S., Luo, J., Datcu, M., Pelillo, M., Zhang, L., 2018. Dota: A large-scale dataset for object detection in aerial images. In: Proc. CVPR.

Yamaguchi, Y., Moriyama, T., Ishido, M., Yamada, H., 2005. Four-component scattering model for polarimetric sar image decomposition. IEEE Trans. Geosci. Remote Sens. 43 (8), 1699–1706.

Yang, M., Rosenhahn, B., Murino, V., 2019. Introduction to multimodal scene under-standing. Multimodal Scene Understanding Elsevier, 1–7.

Yao, J., Meng, D., Zhao, Q., Cao, W., Xu, Z., 2019. Nonconvex-sparsity and nonlocal-smoothness-based blind hyperspectral unmixing. IEEE Trans. Image Process. 28 (6), 2991–3006.

Yu, F., Koltun, V., 2015. Multi-scale context aggregation by dilated convolutions. arXiv:1511.07122.

Yu, N., Davis, L., Fritz, M., 2019. Attributing fake images to gans: Learning and analyzing gan fingerprints. In: Proc. ICCV, pp. 7556–7566.

Zampieri, A., Charpiat, G., Girard, N., Tarabalka, Y., 2018. Multimodal image alignment through a multiscale chain of neural networks with application to remote sensing. In: Proc. ECCV.

Zhang, H., Dana, K., Shi, J., Zhang, Z., Wang, X., Tyagi, A., Agrawal, A., 2018a. Context encoding for semantic segmentation. In: Proc. CVPR.

Zhang, Z., Vosselman, G., Gerke, M., Tuia, D., Yang, M., 2018b. Change detection be-tween multimodal remote sensing data using siamese cnn. arXiv preprint arXiv:1807.09562.

Zhang, B., Zhang, M., Kang, J., Hong, D., Xu, J., Zhu, X., 2019a. Estimation of pmx concentrations from landsat 8 oli images based on a multilayer perceptron neural network. Remote Sens. 11 (6), 646.

Zhang, Z., Vosselman, G., Gerke, M., Persello, C., Tuia, D., Yang, M., 2019b. Detecting building changes between airborne laser scanning and photogrammetric data. Remote Sens. 11 (20), 2417.

Zhao, H., Shi, J., Qi, X., Wang, X., Jia, J., 2017. Pyramid scene parsing network. In: Proc. CVPR, pp. 2881–2890.

Zhao, B., Sveinsson, J., Ulfarsson, M., Chanussot, J., 2019. (semi-) supervised mixtures of factor analyzers and deep mixtures of factor analyzers dimensionality reduction al-gorithms for hyperspectral images classification. In: Proc. IGARSS. IEEE, pp. 887–890.

Zhu, X., Lafferty, J., Rosenfeld, R., 2005. Semi-supervised learning with graphs. Ph.D. thesis. Carnegie Mellon University, Language Technologies Institute, School of Computer Science.