

Contents lists available at [ScienceDirect](https://www.sciencedirect.com)

Remote Sensing of Environment

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

Cross-city matters: A multimodal remote sensing benchmark dataset for cross-city semantic segmentation using high-resolution domain adaptation networks

Danfeng Hong^a, Bing Zhang^{a,b,*}, Hao Li^c, Yuxuan Li^{a,d}, Jing Yao^a, Chenyu Li^e, Martin Werner^c, Jocelyn Chanussot^f, Alexander Zipf^g, Xiao Xiang Zhu^h

^a Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing 100094, China

^b College of Resources and Environment, University of Chinese Academy of Sciences, Beijing 100049, China

^c Big Geospatial Data Management, Technical University of Munich, Munich 85521, Germany

^d School of Electronic, Electrical and Communication Engineering, University of Chinese Academy of Sciences, Beijing 100049, China

^e School of Mathematics, Southeast University, Nanjing 210096, China

^f Univ. Grenoble Alpes, CNRS, Grenoble INP, GIPSA-Lab, Grenoble 38000, France

^g Institute of Geography, Heidelberg University, Heidelberg 69120, Germany

^h Data Science in Earth Observation, Technical University of Munich, Munich 80333, Germany.

ARTICLE INFO

Edited by Marie Weiss

Keywords:

Cross-city
Deep learning
Dice loss
Domain adaptation
High-resolution network
Land cover
Multimodal benchmark datasets
Remote sensing
Segmentation

ABSTRACT

Artificial intelligence (AI) approaches nowadays have gained remarkable success in single-modality-dominated remote sensing (RS) applications, especially with an emphasis on individual urban environments (e.g., single cities or regions). Yet these AI models tend to meet the performance bottleneck in the case studies across cities or regions, due to the lack of diverse RS information and cutting-edge solutions with high generalization ability. To this end, we build a new set of multimodal remote sensing benchmark datasets (including hyperspectral, multispectral, SAR) for the study purpose of the cross-city semantic segmentation task (called C2Seg dataset), which consists of two cross-city scenes, i.e., Berlin-Augsburg (in Germany) and Beijing-Wuhan (in China). Beyond the single city, we propose a **high-resolution domain adaptation network**, HighDAN for short, to promote the AI model's generalization ability from the multi-city environments. HighDAN is capable of retaining the spatially topological structure of the studied urban scene well in a parallel high-to-low resolution fusion fashion but also closing the gap derived from enormous differences of RS image representations between different cities by means of adversarial learning. In addition, the Dice loss is considered in HighDAN to alleviate the class imbalance issue caused by factors across cities. Extensive experiments conducted on the C2Seg dataset show the superiority of our HighDAN in terms of segmentation performance and generalization ability, compared to state-of-the-art competitors. The C2Seg dataset and the semantic segmentation toolbox (involving the proposed HighDAN) will be available publicly at https://github.com/danfenghong/RSE_Cross-city.

1. Introduction

Remote sensing (RS) is an essential means to acquire large-scale and high-quality Earth observation (EO) data in a concise time, which significantly advances the development of EO techniques. However, the conventional expert system-centric mode has run into bottlenecks and cannot meet the EO demand of the RS big data era well, particularly

when facing complex urban scenes. Artificial Intelligence (AI) techniques provide one viable option that is capable of finding out potentially valuable knowledge from the vast amounts of pluralistic EO data more intelligently, enabling the understanding and monitoring of the contemporary urban environment.

These advanced AI models, e.g., deep learning, have been successfully applied for various RS and geoscience applications, which have

* Corresponding author at: Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing 100094, China.

E-mail addresses: hongdf@aircas.ac.cn (D. Hong), zb@radi.ac.cn (B. Zhang), hao_bgd.li@tum.de (H. Li), liyuxuan231@mails.uicas.ac.cn (Y. Li), yaojing@aircas.ac.cn (J. Yao), lichenyu@seu.edu.cn (C. Li), martin.werner@tum.de (M. Werner), jocelyn.chanussot@grenoble-inp.fr (J. Chanussot), zipf@uni-heidelberg.de (A. Zipf), xiaoxiang.zhu@tum.de (X.X. Zhu).

<https://doi.org/10.1016/j.rse.2023.113856>

Received 13 February 2023; Received in revised form 14 September 2023; Accepted 13 October 2023

Available online 30 October 2023

0034-4257/© 2023 Elsevier Inc. All rights reserved.

been proven to be particularly applicable to the unitary urban environment where the types, characteristics, and spatial distributions of surface elements are significantly consistent and similar. Nevertheless, the ability to address multiple urban environmental issues with highly spatio-temporal and regional change remains limited. The possibly feasible solutions are two-fold. On the one hand, the joint exploitation of multimodal RS data has been proven to be helpful to improve the processing ability of cross-city or regional cases, since the RS data acquired from different platforms or sensors can provide richer and more diverse complementary information. On the other hand, designing more leading-edge AI models with a focus on promoting the generalization ability across cities or regions is an inexorable trend to alleviate the semantic gap between different urban environments, making it mutually transferable for knowledge.

In recent years, enormous efforts have been made to couple or jointly analyze different RS observation sources by the attempts to design advanced fusion and interpretation methods to achieve a more diversified description for the studied urban scene. In particular, a growing body of studies has confirmed the achievement of multimodal AI models in one single urban environment. It should be noted, however, that multi-city-related cases are evolving at a relatively slow speed. This slow progression can be satisfactorily explained by two very likely reasons as follows.

- One refers to the lack of high-quality multimodal RS benchmark datasets for a better understanding of cross-city environments.
- Another is that currently developed methodologies prefer to focus on extreme performance pursuit in one single urban environment rather than improve the model generalization ability, particularly for diverse urban environments (e.g., different cities or regions).

To boost technical breakthroughs and accelerate the development of EO applications across cities or regions, creating a multimodal RS benchmark dataset for cross-city land cover segmentation makes necessary. Just as important, the high generalization ability in terms of methodology development is of paramount importance. This drives us to develop such a model with high transferability between different cities or regions by means of domain adaption (DA) techniques. Numerous experiments will be conducted on the cross-city land cover segmentation dataset to show the superiority of DA-based approaches over those semantic segmentation algorithms that do not consider knowledge transfer across domains. More specifically, our contributions in this paper can be unfolded as follows.

- A new set of multimodal RS benchmark datasets is built for the study purpose of the cross-city semantic segmentation task, named C2Seg for short. C2Seg consists of two subsets, i.e., Berlin-Augsburg (in Germany) dataset collected from EnMAP, Sentinel-2, and Sentinel-1, respectively, Beijing-Wuhan (in China) dataset collected from Gaofen-5, Gaofen-6, and Gaofen-3, respectively. The C2Seg dataset will be available freely and publicly, promoting the research progress on semantic segmentation across cities or regions substantially. To the best of our knowledge, C2Seg is the first benchmark dataset about the cross-city multimodal RS image segmentation task, which considers the three-modality study case, including hyperspectral, multispectral, and synthetic aperture radar (SAR) data acquired from the currently well-known satellite missions. The C2Seg datasets have been utilized for the WHISPERS2023 conference <https://www.ieee-whispers.com/> in the capacity of Challenge 1: Cross-City Multimodal Semantic Segmentation. These datasets are accessible at <https://www.ieee-whispers.com/cross-city-challenge/>, with the training data already made available. Shortly, we plan to make all datasets, including both training and testing data, accessible to the wider research community.
- A high-resolution domain adaptation network (HighDAN) is devised to bridge the gap between RS images from different urban environments utilizing adversarial learning, thereby making it possible to

transfer the learned knowledge from one domain to another effectively and eliminate inter-class variations to a great extent. Further, HighDAN, which is built based on the high-resolution network (HR-Net), is capable of capturing multi-scaled image representations from parallel high-to-low-resolution subnetworks, yielding repetitive information exchange across different resolutions in a highly efficient manner.

- To reduce the impact of the sample number imbalance between classes due to the multi-city studies, the Dice loss is considered and embedded in the proposed HighDAN.

The remaining sections of the paper are organized as follows. Section 2 reviews the related work for semantic segmentation in the land cover classification task systematically from the perspectives of individual study scenes and cross-region (or cross-city) cases. Section 3 introduces the newly-built datasets and correspondingly elaborates on the proposed methodology. Experiments are conducted on the datasets with extensive discussion and analysis in Section 4. Finally, Section 5 makes the conclusion of this paper with some remaining challenges and plausible future solutions.

2. Related work

Over the past decade, deep learning (DL) has been garnering increasing attention in many application fields (Yuan et al., 2020), owing to its powerful ability for data representation and learning. In particular, the ever-perfecting DL techniques for RS enable accurate and automatic land cover mapping. According to different studied scenes, we divide these approaches into individual environments and multi-region (or city) ones, where single-modality and multimodal RS data are further involved.

2.1. Semantic segmentation on individual environments

With the emergence and rapid development of DL, there have been recently numerous semantic segmentation methods successfully developed for RS with a focus on a single studied scene (Yuan et al., 2021). Kampffmeyer et al. (Kampffmeyer et al., 2016) developed deep convolutional neural networks (CNNs) for semantic segmentation in terms of small objects in urban areas, where the uncertainty in CNNs is modeled by Bayesian approximation in Gaussian process (Gal and Ghahramani, 2016). The CNNs-based architecture was also used in (Kemker et al., 2018) for semantic segmentation on multispectral RS images rather than high-resolution RGB images. In this work, synthetic multispectral images are generated for initializing deep CNNs to alleviate the effects of label scarcity. Yi et al. (Yi et al., 2019) proposed a deep residual U-Net (ResUNet) framework, which consists of cascade down-sampling and up-sampling subnetworks, for urban building extraction using very high-resolution (VHR) RS images. Further, Diakogiannis et al. (Diakogiannis et al., 2020) designed an enhanced ResUNet version, ResUNet-a, with atrous convolutions for semantic segmentation of RS images. A multi-scale semantic segmentation network was proposed in (Du et al., 2021) for fine-grained urban functional zone classification using VHR RS images and object-based strategies. Adding to this advancement, Wang et al. (Wang et al., 2022a) introduced a recent breakthrough in the field, unveiling an efficient U-shaped transformer network custom-tailored for the precise execution of semantic segmentation tasks in VHR urban scene images. Concurrently, He and his collaborators (He et al., 2022) incorporated the Swin transformer into the U-Net architecture, further enhancing the capabilities of semantic segmentation in RS applications. In a recent development, as documented in (Wang et al., 2023), a novel approach following the SegFormer (Xie et al., 2021) framework, enriched by the utilization of hypercolumns, has been employed for seismic facies segmentation. Although these DL approaches have provided superior segmentation accuracy over traditional model-driven models on single-modality RS

images, they inevitably meet the performance bottleneck in the complex scene understanding task (due to the lack of diverse modality information).

With the ever-growing availability of RS data sources from well-known spaceborne and airborne missions, e.g., Gaofen in China, Sentinel in the EU, and Landsat in the USA, multimodal RS techniques have been garnering increasing attention and made extraordinary progress in various EO-related tasks. The data acquired by different platforms can provide diverse and complementary information (Dalla Mura et al., 2015). The joint exploitation of different RS data has been therefore proven to be effective in further enhancing our understanding, possibilities, and capabilities in a single urban environment. As the mainstream application, semantic segmentation of multimodal RS images using DL has been widely studied in recent years. Audebert et al. (Audebert et al., 2016) extracted the multi-scaled deep features from multimodal EO data for semantic labeling. Further, the same authors extended their work in (Audebert et al., 2016) by implementing the multi-scale deep fully convolutional networks (FCNs) (Long et al., 2015; Wu et al., 2019) based on SegNet (Badrinarayanan et al., 2017) to process and understand multimodal RS data for land cover segmentation (Audebert et al., 2018). Similar to (Hong et al., 2020a), they also discussed the fusion strategies of different RS modalities, e.g., early, middle, and late fusion. In (Wieland et al., 2019), multi-sensor cloud and shadow segmentation are investigated using CNNs. Wurm et al. (Wurm et al., 2019) proposed to transfer FCNs trained from external datasets for improving the semantic segmentation performance of cross-modal satellite images. Segal et al. (Segal-Rozenhaimer et al., 2020) designed a CNNs-based cloud detection algorithm based on the Deeplab architecture (Chen et al., 2017a) for multimodal satellite images, achieving an effective detection performance improvement. Ren et al. (Ren et al., 2022) proposed a dual-stream high-resolution network (HR-Net) (Sun et al., 2019) for the deep fusion of GF-2 and GF-3 multimodal RS data for land cover classification. In the work by Adriano et al. (Adriano et al., 2021), the authors explored the mapping and evaluation of building damage from a segmentation perspective, leveraging the rich information provided by multimodal and multitemporal RS data, marking a significant advancement in the field of damage assessment.

2.2. Semantic segmentation across regions or cities

Currently developed semantic segmentation networks of RS images in terms of the design of network architecture, module details, and the use of loss functions have reached their performance peak. It is a noticeable phenomenon, however, that these models are more often than not well-designed for individual study scenes. This will lead to poor generalization ability for the model, which can not well match the level of the segmentation performance, particularly in the cases of cross-city or cross-region studies. For this reason, researchers have started gradually paying more attention to the task of semantic segmentation across regions or cities.

Domain adaptation (DA) has been proven to be helpful in reducing the semantic gap between source and target domains (Ganin and Lempitsky, 2015). The DA-related approaches have been recently designed to address the challenge of cross-scene RS image semantic segmentation. For example, Chen et al. (Chen et al., 2017b) proposed a road scene adaptation segmenter by utilizing high-resolution RS images from Google Street View in an unsupervised manner, which is well-designed to solve the problem of dataset biases across different cities effectively. A novel adversarial learning method was presented in (Tsai et al., 2018) for DA in semantic segmentation, where the spatially structural similarity is employed to narrow down the gap between data distribution differences of different domains. Tong et al. (Tong et al., 2020) first pre-trained a deep CNN with a well-annotated Gaofen-2 land cover dataset, and transferred the trained deep model for the unlabeled RS image classification in the target domain. By contrast, Zhu et al. (Zhu et al., 2021) directly learned a transfer network by attempting to align the data

distribution of subdomains with the utilization of a local maximum mean discrepancy for image classification. Li et al. (Li et al., 2022) proposed a few-shot transfer learning (FSTL) method to improve the generalization capability of pre-trained deep CNN on mapping human settlement across countries. Li et al. (Li et al., 2021) reduced the impact of data shift effectively by designing weakly-supervised constraints, making it more suitable for the task of cross-domain RS image semantic segmentation. Moreover, Wang et al. (Wang et al., 2022b) contributed to the field by facilitating domain adaptation (DA) in the context of cross-sensor VHR urban land cover segmentation, with a focus on accommodating both airborne and spaceborne RS images. Further, the same investigators (Ma et al., 2023) extended their work for semantic segmentation in RS by considering local consistency and global diversity to enhance the DA capability.

The joint use of multimodal RS data is capable of better mining the representation ability of diverse RS modalities, further weakening the effects of data shift to some extent when the model is trained on one RS data domain and transferred to another. Hong et al. (Hong et al., 2020b) aimed at the semi-supervised transfer learning challenge for cross-scene land cover semantic classification in RS and accordingly proposed a cross-modal deep network, called X-ModelNet. The same authors in (Hong et al., 2021a) further extended their work with two plug-and-play adversarial modules to enhance the robustness and transferability of cross-region RS image semantic segmentation. Similarly, Ji et al. (Ji et al., 2021) fully aligned the source and target domains in the generative adversarial network (GAN) (Goodfellow et al., 2020) guided image space. The style translation technique is utilized to train an end-to-end deep FCN with a combination of DA and semantic segmentation from the multi-source RS images to identify the different types of land cover elements. Zhao et al. (Zhao et al., 2022) reduced the disparity across scenes by using fractional Fourier fusion and spatial-spectral DA techniques for cross-domain multi-source RS data classification. These aforementioned methods can be unified into a general multimodal deep learning framework for RS image land cover classification (i.e., MDL-RS) on both individual and cross-region environments (Hong et al., 2021b).

There have recently been certain researches developed by attempts to investigate the feasibility and effectiveness of semantic segmentation across regions or cities using multimodal RS images. Yet the inadequate integration among high-performance deep semantic segmentation architectures, DA networks, and the use of multimodal RS data inevitably leads to the performance bottleneck in cross-domain land cover classification. Most importantly, the problems in the lack of multimodal RS benchmark datasets become obstacles to the development of urban RS and further decelerate the technical progress of scientific research in terms of cross-city semantic segmentation.

The follow-up two sections will therefore focus on the solutions to the two above-mentioned difficulties. Accordingly, one creates large-scale multimodal RS benchmark datasets for the study of cross-city semantic segmentation and another brings forth new ideas in the update and upgrade of network architecture and blending between multimodal RS data and DA techniques.

3. C2Seg: a multimodal RS dataset for cross-city semantic segmentation

3.1. Overview

To overcome the difficulty of multimodal RS data shortage and boost the technological innovation of urban scene understanding across cities, we build a new collection of multimodal RS benchmark datasets, including hyperspectral, multispectral, and SAR data, for research into cross-city semantic segmentation (i.e., C2Seg). C2Seg datasets consist of two cross-city scenes as follows.

- C2Seg-AB: Berlin-Augsburg cities in Germany, which are collected from EnMAP, Sentinel-2, and Sentinel-1 satellite missions on the date

as close as possible, and accordingly pre-processed via ESA's SNAP toolbox.

- C2Seg-BW: Beijing-Wuhan cities in China, which are collected from Gaofen-5, Gaofen-6, and Gaofen-3 satellite missions on the date as close as possible, and pre-processed using the ENVI software.

In contrast to certain well-known HR or VHR datasets, such as OpenEarthMap (Xia et al., 2023), it's worth noting that our C2Seg datasets encompass three distinct RS modalities, even though they maintain a GSD of only 10 m. Furthermore, we are committed to fostering research progress in the domain of cross-city semantic segmentation by making the C2Seg datasets openly available for free download. These datasets encompass 13 distinct land use and land cover semantic categories.¹ To the best of our knowledge, this represents a pioneering effort in creating a large-scale benchmark dataset tailored for cross-city multimodal RS semantic segmentation, taking into account three kinds of RS modalities. The C2Seg datasets will be unfolded in detail as follows.

3.2. C2Seg-AB

In C2Seg-AB, the multimodal RS data and labeled semantic categories are prepared across Berlin and Augsburg cities in Germany. C2Seg-AB consists of hyperspectral data from EnMAP, multispectral data from Sentinel-2, and SAR data from Sentinel-1. Fig. 1 visualizes the C2Seg-AB datasets in terms of scene location, image region, and different modalities with ground truth (GT) of semantic segmentation.

1) EnMAP Hyperspectral Data. Before launching the EnMAP satellite, the simulation is the main and widely-used way that obtains the EnMAP-related product, which is synthesized by using the full-chain automatic simulation tool, i.e., EeteS (Segl et al., 2012), on the high-resolution HyMap or HySpex hyperspectral images. The airborne hyperspectral imaging sensors, i.e., HyMap and HySpex, are used to acquire hyperspectral images over Berlin and Augsburg cities and their neighboring areas. Using EeteS, the corresponding EnMAP images can be simulated by HyMap and HySpex at a ground sample distance (GSD) of 30 m, which are openly available from <https://doi.org/10.5880/enmap.2016.002> and <https://mediatum.ub.tum.de/1657312>, respectively. Further, the two hyperspectral images are upsampled to 10 m GSD to keep the identically spatial resolution of all multimodal RS images in the same studied scene. Therefore, the resulting images consist of 2465×811 pixels (Berlin) and 886×1360 pixels (Augsburg), respectively, and they share the same spectral bands (i.e., 242) in the wavelength range of 400 nm to 2500 nm. More details can be found in (Okujeni et al., 2016; Hu et al., 2023).

2) Sentinel-2 Multispectral Data. The Sentinel-2 mission is composed of two twin-orbit satellites (i.e., Sentinel-2 A/B) with a combined revisiting time of approximately five days at the equator, the spatial, spectral, and temporal resolution, therefore, makes Sentinel-2 well-suited for dynamic land cover mapping and monitoring. The Sentinel-2 multispectral sensor covers a total of 13 spectral bands ranging from 10 m to 60 m with different spatial resolutions, and the captured spectral reflectance ranges from visible to NIR and SWIR wavelengths. The best pixels in Sentinel-2 multispectral composite are used in this work, which has been further processed by the SEPAL cloud platform data processing system (sepal.io) of the Food and Agriculture Organization of the United Nations (FAO). Furthermore, the Top of Atmosphere (TOA) reflectance was converted to surface reflectance, and the best pixels were selected from the past three years as of April 2020

¹ They are *Urban Fabric, Industrial/Commercial/Transport Units, Mine/Dump/Construction Sites, Artificial/Non-Agricultural/Vegetated Areas, Surface Water, Street, Arable Land, Permanent Crops, Pastures, Forests, Shrub and/or Herbaceous Vegetation Associations, Open Spaces with Little or Non-Vegetation, and Inland Wetlands.*

using a medoid compositing function, where the radiative transfer models are applied in (Hansen et al., 2008) and were later adapted to Sentinel-2 by FAO. In our case, 4 spectral bands are selected from Sentinel-2, e.g., red, green, blue, and near-infrared (NIR), at a GSD of 10 m by following a geographic reference of WGS84/UTM Zone 32 N.

3) Sentinel-1 SAR Data. The SAR component is acquired by the Sentinel-1 mission, which is a level-1 Ground Range Detected product obtained by the Interferometric Wide Swath mode. The SAR data is characterized by dual-polarized information with VV and VH channels. The SNAP toolbox is specially designed by the European Space Agency (ESA) for pre-processing Sentinel-1 data to obtain an analysis-ready SAR image, which can be available from the link at <https://step.esa.int/main/toolboxes/snap/>. The workflow performed in the SNAP toolbox follows several steps, i.e., precise orbit profile, radiometric calibration, deburst, speckle reduction, and terrain correction. Employing the shuttle radar topography mission, the topographic data are generated well. Different from the Sentinel-2 multispectral image, the Sentinel-1 SAR image is not strictly sampled to the GSD of 10 m. Accordingly, the SAR image is geo-coded to be 10 m GSD via the bilinear interpolation operator. Finally, the SAR images with two channels, i.e., intensities of VV and VH, are aligned with the pixel-wise EnMAP and Sentinel-2 images.

4) Ground Truth of Semantic Segmentation. Herein, we label the GT of semantic segmentation by retrieving land use and land cover (LULC)-labeled data from OpenStreetMap (OSM) LULC platform at <https://osmlanduse.org/> and 12 main classes well-defined in OSMLULC are considered in our case. Accordingly, we manually check the labels within the cities of Berlin and Augsburg and also included the major street network from OSM and appended it to the existing 12 classes, which ensures the granularity and accuracy of the final labeled data. By extending those classes defined in (Schultz et al., 2017), we end up with 13 distinct semantic segmentation features, including urban, industrial, mine, artificial vegetated, arable land, permanent crops, pastures, forests, shrubs, open spaces, inland wetlands, water bodies, and street networks. The elaborately produced LULC maps as GT data (i.e., for the purpose of the semantic segmentation task) in our studied areas are visualized in color (see Fig. 1).

3.3. C2Seg-BW

The C2Seg-BW dataset provides multimodal RS data and labeled semantic categories across Beijing and Wuhan cities in China, as shown in Fig. 2. Similarly, hyperspectral, multispectral, and SAR data are involved in the dataset, which is collected from Gaofen series satellites, such as Gaofen-5, Gaofen-6, and Gaofen-3, respectively. The acquisition dates or satellite perigee passing time of these modality data are late 2019 and early 2020, which ensures that the ground elements remain unchanged as much as possible.

1) Gaofen-5 Hyperspectral Data. The Gaofen-5 hyperspectral data is the level-1 A product collected by the Advanced Hyperspectral Imager (AHSI) (Liu et al., 2019) from the China Center for Resource Satellite Data and Applications (CRESDA). The spatial resolution of the hyperspectral image is around 30 m with a narrow swath width of approximately 60 km, and there are 330 spectral bands ranging from 400 nm to 2500 nm. The spectral resolution in the visible and near-infrared (VNIR) region (i.e., 400 nm to 1000 nm) is about 5 nm, while that in the short-wave infrared (SWIR) region (i.e., 1000 nm to 2500 nm) is about 10 nm.

The hyperspectral images are pre-processed using the ENVI 5.6 software, whose workflow mainly includes radiometric calibration, Fast Line-of-sight Atmospheric Analysis of Spectral Hypercubes (FLAASH) correction, orthorectification, and bands selection. The band selection operation is utilized to massively remove the water vapor absorption, noisy, and bad bands to maintain the image quality. The selected 116 bands are further processed by using the Savitzky-Golay filter. The resulting hyperspectral images are upsampled from 30 m to 10 m GSD and they then consist of 13474×8706 pixels in Beijing and $6225 \times$

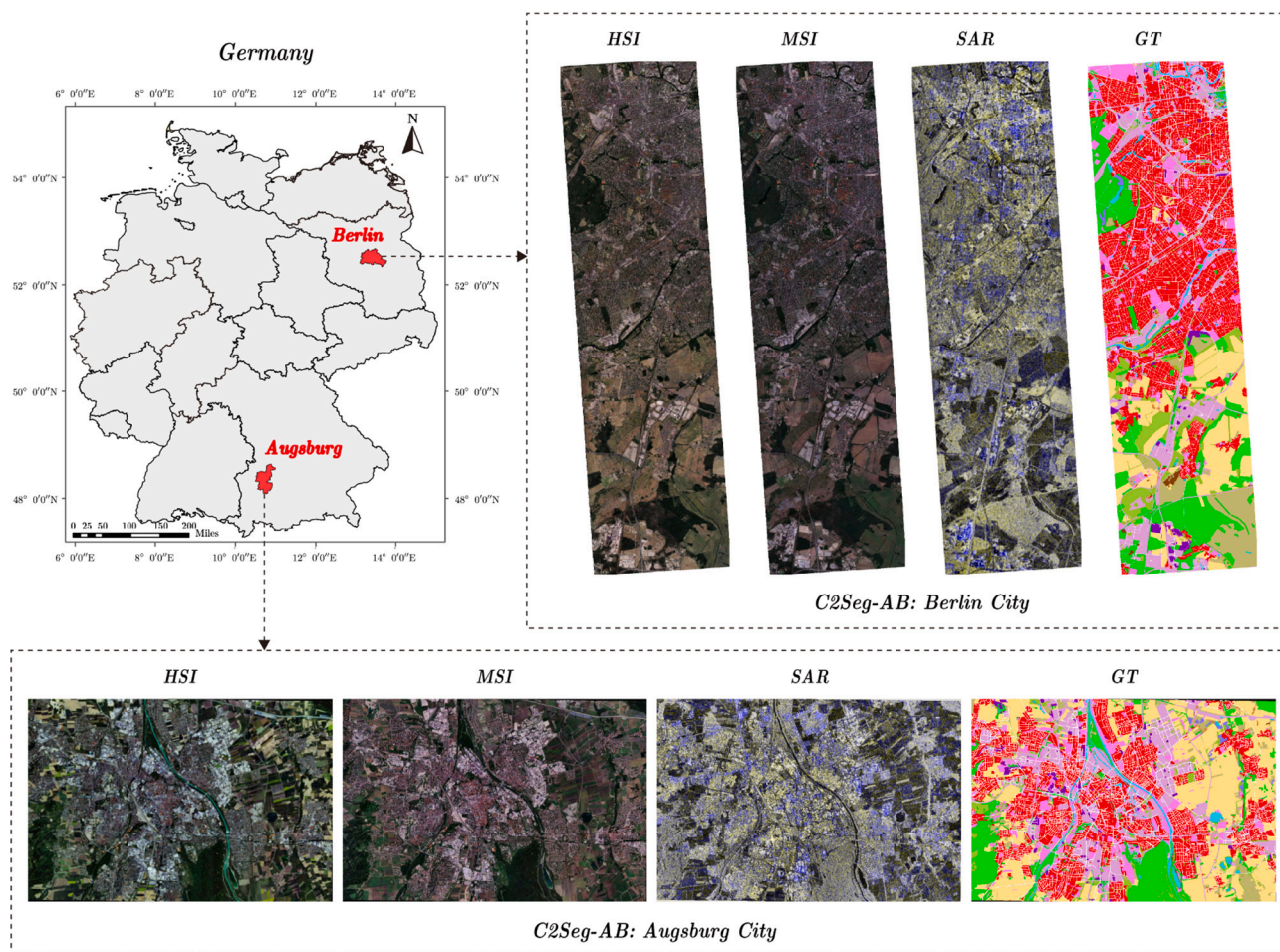


Fig. 1. Visualizing C2Seg-AB datasets for semantic segmentation study scene across Berlin and Augsburg cities in Germany using multimodal RS data.

8670 pixels in Wuhan, respectively, with a geographic reference of WGS1984 Web Mercator (Auxiliary Sphere).

2) Gaofen-6 Multispectral Data. The Gaofen-6 product is acquired by the specially-designed camera to collect the panchromatic and multispectral images with spatial resolutions of 2 m and 8 m simultaneously. The multispectral data are used in this paper and pre-processed on the ENVI platform via the standardized processing flow similar to hyperspectral data. To maintain the consistency of the spatial resolution, the four spectral bands in the multispectral image are resampled to 10 m.

3) Gaofen-3 SAR Data. The Gaofen-3 product is collected under the Wide Fine Stripmap mode, yielding a spatial resolution of 10 m with a swath width of 100 km. The SAR data are prepared by utilizing the functions of de-speckle and terrain correction in the ENVI SARscape Analytics toolbox. The Refined Lee filter (Yommy et al., 2015) with a sliding window size of 5×5 pixels is selected to remove the speckle noises, and the SAR data are corrected employing global digital elevation model (DEM) data in GMTED2010 (Danielson and Gesch, 2011). Similar to Sentinel-1, we adopt the dual-Pol SAR image with HH and HV channels for two studied scenes, and the image size and resolution are the same as those of Gaofen-6 multispectral data.

4) Ground Truth of Semantic Segmentation. Similar to C2Seg-AB, we retrieve LULC-labeled data and major street networks within the cities of Wuhan and Beijing (in China) from OSMLULC and OSM, respectively. Herein, we again classify LULC-labeled data by following the class schema defined in (Schultz et al., 2017), which is based on the widely-accepted Corine Land Cover (CLC) schema (Feranec et al., 2016). However, the availability of OSM data in China is insufficient for semantic labeling. For this reason, we manually map and complete the

LULC features by taking multispectral and hyperspectral images as the reference, making it consistent with the labeling schema used in C2Seg-AB datasets. The labeled data of 13 distinct classes serve as a piece of ground-truth information for the following quantitative analysis of cross-city semantic segmentation tasks throughout this paper.

4. HighDAN: high-resolution domain adaptation network

4.1. A brief recall of HR-net

Convolutional neural networks (CNNs) have been proven to be effective in learning rich representations from images. Many well-known CNNs-based deep network architectures have been put forward successively, such as AlexNet (Krizhevsky et al., 2017), VGGNet (Simonyan and Zisserman, 2015), and GoogleNet (Szegedy et al., 2015). However, there is a potentially common problem in these backbones, i.e., the resolution of the generated feature maps is relatively low when performing the feature extraction by adopting the convolution connection from high resolution to low resolution in series. This inevitably leads to the loss of spatial information. As a result, the traditional solution to this issue is designing an encoder-decoder architecture, i.e., reducing image resolution via the encoder and restoring to high-resolution representations via the decoder. These networks, e.g., U-Net (Ronneberger et al., 2015), SegNet (Badrinarayanan et al., 2017), DeconvNet (Noh et al., 2015), Hourglass (Newell et al., 2016), belong to the member of the encoder-decoder structure in essence. Nevertheless, this kind of deep network architecture tends to generate blurred low-resolution feature maps due to multiple convolution operations. These

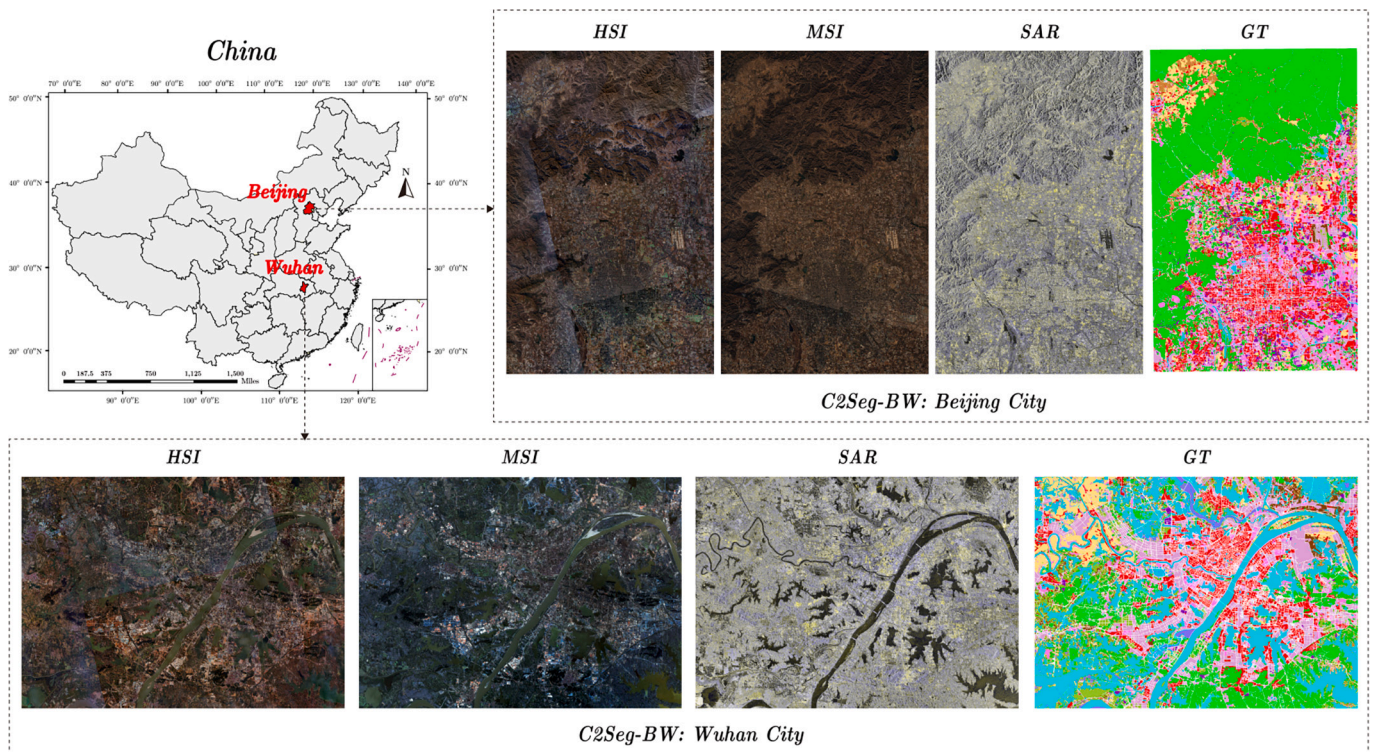


Fig. 2. Visualizing C2Seg-BW datasets for semantic segmentation study scene across Beijing and Wuhan cities in China using multimodal RS data.

feature maps with different resolutions are further integrated into series connections, raising the risk of the loss of edge details and texture information.

To overcome the difficulty mentioned above, HR-Net (Sun et al., 2019) is proposed to generate and maintain high-resolution representations. The HR-Net’s increments lie in three-folds as follows.

- To connect the high-to-low-resolution convolution streams in a parallel fashion instead of previous series connections, as shown in Fig. 3 to visualize their differences.
- To keep high-resolution representations throughout the whole network architecture.
- To exchange the information of feature maps across different resolutions, enabling the compact fusion between high- and low-resolutions to enhance the model’s performance. The fusion strategy mainly consists of 1) identity mapping for feature maps with the same resolutions; 2) bilinear upsampling plus 1×1 convolution for feature maps from low to high-resolutions; 3) 3×3 stride convolution for feature maps from high to low-resolutions. Fig. 4 illustrates the fusion mode in HR-Net.

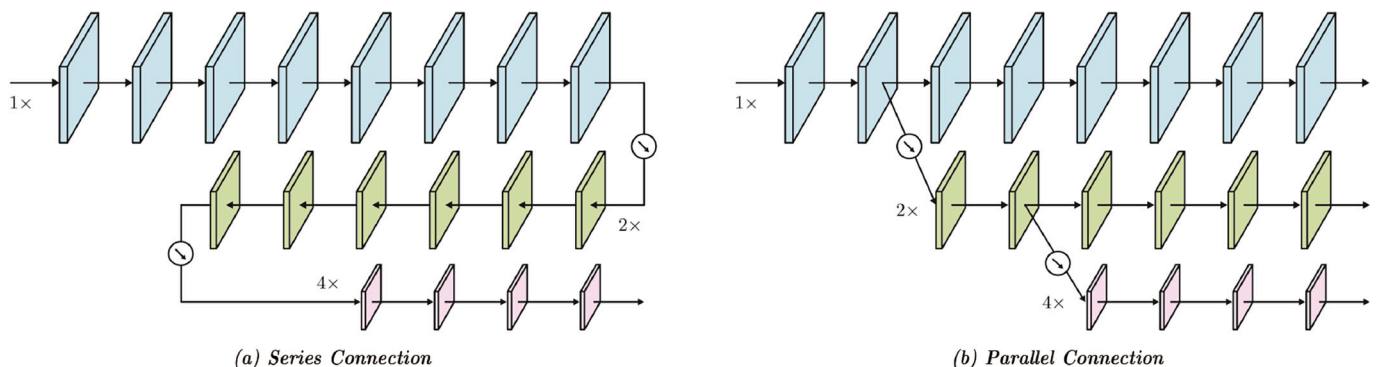


Fig. 3. Visualizing the comparison for connection modes of feature maps with different resolutions: (a) Series Connection and (b) Parallel Connection.

4.2. Method overview of HighDAN

Owing to the advancement and superiority of the HR-Net architecture in terms of learning high-resolution representations from images, we propose a novel multimodal HR-Net backbone (i.e., HighDAN) with unsupervised domain adaptation for the cross-city semantic segmentation task using multimodal RS data. Overall, the HighDAN architecture consists of the multimodal encoder, adversarial domain adaptation, and convolution decoder. The design of the domain adaptation module aims to bridge the gap between the representations of source and target domains in an adversarial learning fashion, thereby fully mining the invariant semantic features from multimodal RS data and transferring them across domains. Embedding Dice loss (Li et al., 2020) into networks, HighDAN is capable of weakening the class imbalance effects that tend to be generated in the case of cross-city image interpretation, e.g., semantic segmentation. An illustrative workflow for HighDAN is given in Fig. 5.

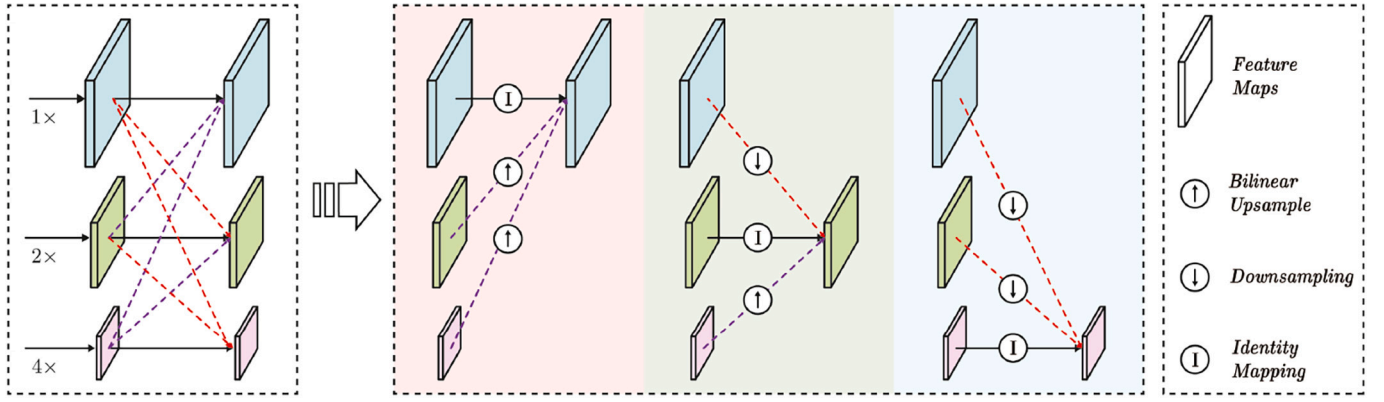


Fig. 4. The fusion strategy in HR-Net for feature maps between different resolutions, including the same resolution fusion, upsampling fusion, and down-sampling fusion.

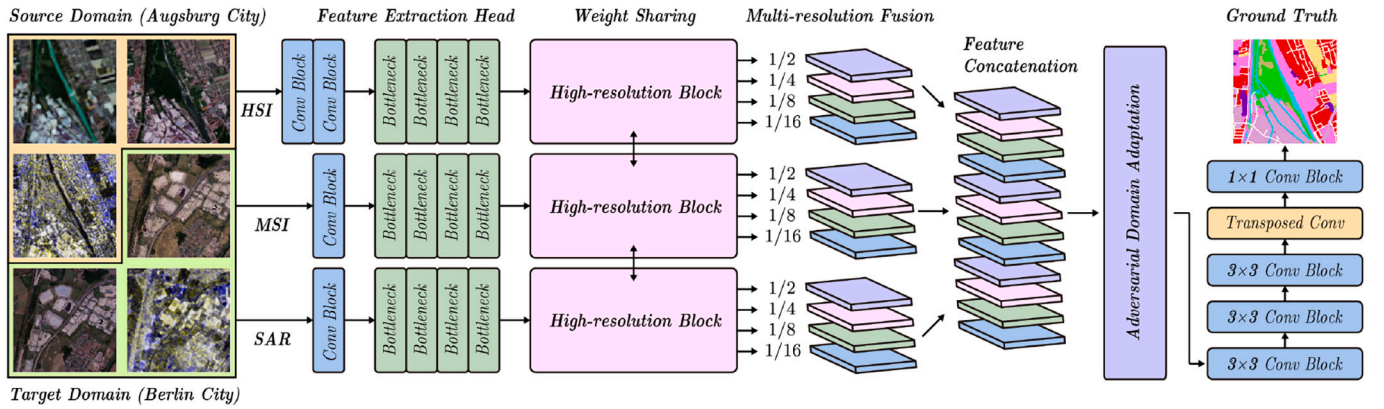


Fig. 5. An illustrative workflow of the proposed HighDAN for cross-city semantic segmentation, which mainly consists of feature extraction head, high-resolution (HR) module, multi-resolution fusion, adversarial domain adaptation, and segmentation head (convolution decoder module).

4.3. Multimodal encoder

The multimodal encoder consists of a feature extraction head and a multimodal high-resolution (HR) subnetwork. As the name suggests, the feature extraction head learns the preliminary representations for different RS modalities by transformations. The head is comprised of the 3×3 convolution block and four bottleneck blocks. Fig. 6 visualizes the feature extraction head: (a) convolution block and (b) bottleneck block. The former convolution block can be formulated as

$$\mathbf{Z}_k = f_{\mathbf{W}_k, \mathbf{B}_k}(\mathbf{X}_k), \quad (1)$$

where k is the index (e.g., 1, 2, ...) for different RS modalities, and \mathbf{X} and \mathbf{Z} denote the input modality image and the feature representations via the convolution block, respectively. The function $f(\cdot)$, i.e., the convolution block, is unfolded as 3×3 convolution operation, batch normalization (BN), and ReLU activation function, which is with respect to the network variables of weights \mathbf{W} and biases \mathbf{B} . Given that hyperspectral data typically possesses a significantly higher dimensionality compared to multispectral and SAR data, it is common practice to employ dimensionality reduction techniques (e.g., PCA) to preprocess the data before feeding it into networks. Additionally, to ensure compatibility with the input dimensions of bottleneck blocks, several extra convolutional layers are utilized for all input data, facilitating seamless integration within the network architecture. The later bottleneck block is expressed by

$$\mathbf{Q}_k = g_{\mathbf{W}_k, \mathbf{B}_k}(\mathbf{Z}_k), \quad (2)$$

where \mathbf{Q} denotes the feature representations via the bottleneck block.

The bottleneck block can be represented as the function $g(\cdot)$ with respect to the to-be-learned network variables: \mathbf{W} and \mathbf{B} , which can be unfolded as 1×1 convolution, BN, 3×3 convolution, BN, 1×1 convolution, and BN in sequence. To provide a further explanation, the multimodal encoder in HighDAN initiates with a three-stream network architecture that takes as input multimodal RS data, including hyperspectral, multispectral, and SAR (see Fig. 5). This architecture is instrumental in elucidating the approach used to effectively combine data from diverse RS modalities.

The multimodal HR subnetwork well inherits attributes of HR-Net that can extract HR image representations. Following the HR-Net, the input RS modality image is firstly downsampled by convolution operations with a 2-stride as the main stem. By gradually adding high-to-low-resolution streams, feature maps with different resolutions are then connected and fused in parallel to acquire diversified resolution representations. The process can be written as

$$\mathbf{V}_k = h_{\mathbf{W}_k, \mathbf{B}_k}(\mathbf{Q}_k), \quad (3)$$

where \mathbf{V}_k denotes the HR representations of the k -th modality via the multimodal HR subnetwork. The function $h(\cdot)$ is defined as the multimodal HR subnetwork by copying the HR module in HR-Net (Sun et al., 2019), which is illustrated in Fig. 6 (c) with HR block. That is, it consists of a multi-resolution group convolution and a multi-scale fusion layer. The former refers to a regular convolution for each resolution stream over different spatial resolutions separately, and the latter aims to perform an interactive fusion of feature maps across scales. It should be noted that the HR module for different RS modalities is shared in terms of network parameters to capture the high-quality multimodal

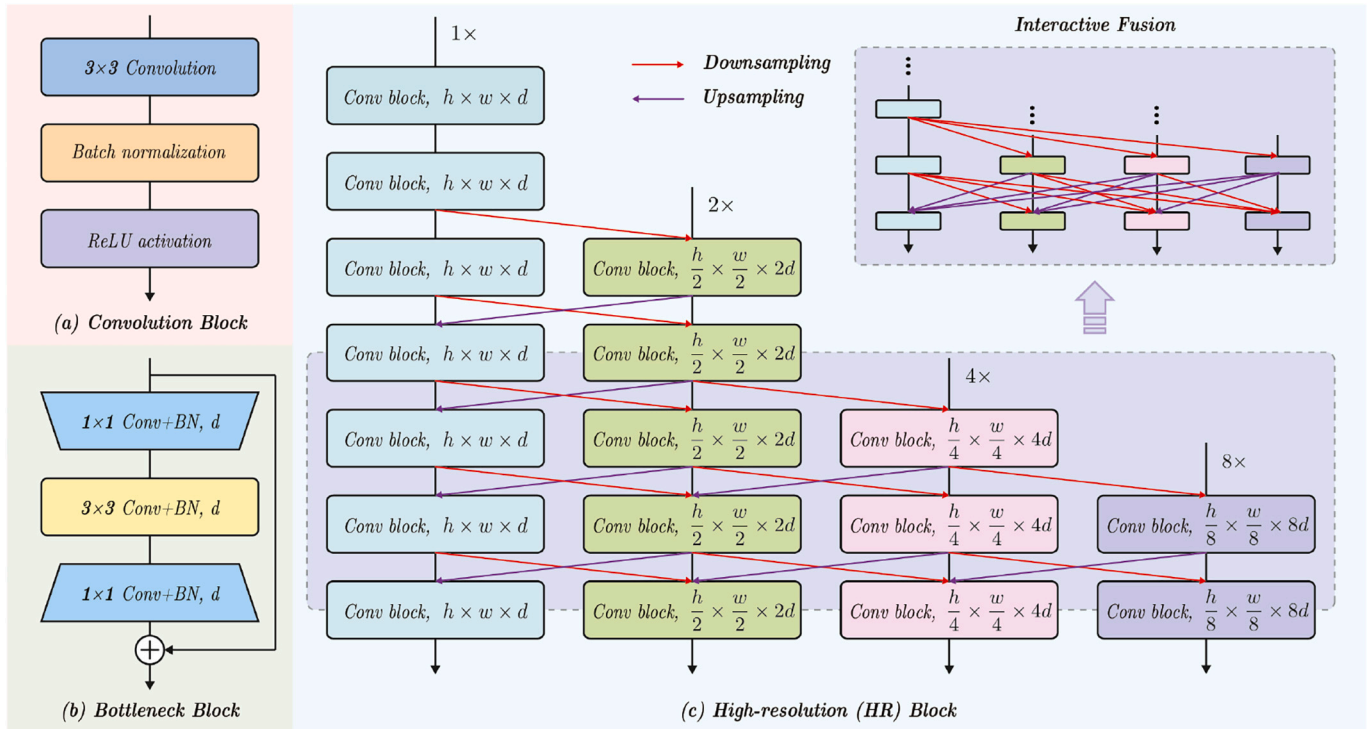


Fig. 6. Clarifying details in the multimodal encoder of HighDAN of (a) convolution block, (b) bottleneck block, and (c) HR block, where (a) and (b) form the feature extraction head and (c) is the main body of multimodal HR subnetwork.

characteristics more steadily. The outputs from each resolution stream are re-scaled to the same resolution as HR representations through bilinear upsampling, achieving the multi-resolution fusion via feature stacking.

4.4. Adversarial domain adaptation

According to the adversarial learning in GAN, the image-to-image translation techniques (Isola et al., 2017) enable the pixel-level alignment and knowledge conversion between source and target domains. This further provides possible and potential solutions to the cross-domain semantic segmentation task. Prior to conducting DA, it is essential to concatenate all feature maps obtained from the various multimodal streams, denoted as $V = \{V_k\}_{k=1}^m$. This consolidation of

feature maps is a crucial step in the process. Inspired by (Yu et al., 2021), we adopt two types of DA modules based on the adversarial learning strategy to align representations of source and target domains at both feature-level and category-level. On the one hand, the feature-level DA module attempts to reduce biases of cross-domain intermediate feature maps (i.e., V) obtained from the multimodal HR encoder. Herein, pixel-wise confidence scores that can reflect the degree of local alignment in different domains are generated from the discriminator, which can be used to reweigh the intermediate features V to correct the representation shift between different domains locally. This yields the aligned representations as A . On the other hand, the category-level DA module aims to enhance global semantic alignment from the label distribution perspective, which is used in the final prediction phase. The global semantic alignment operation can be regarded as a kind of soft constraint

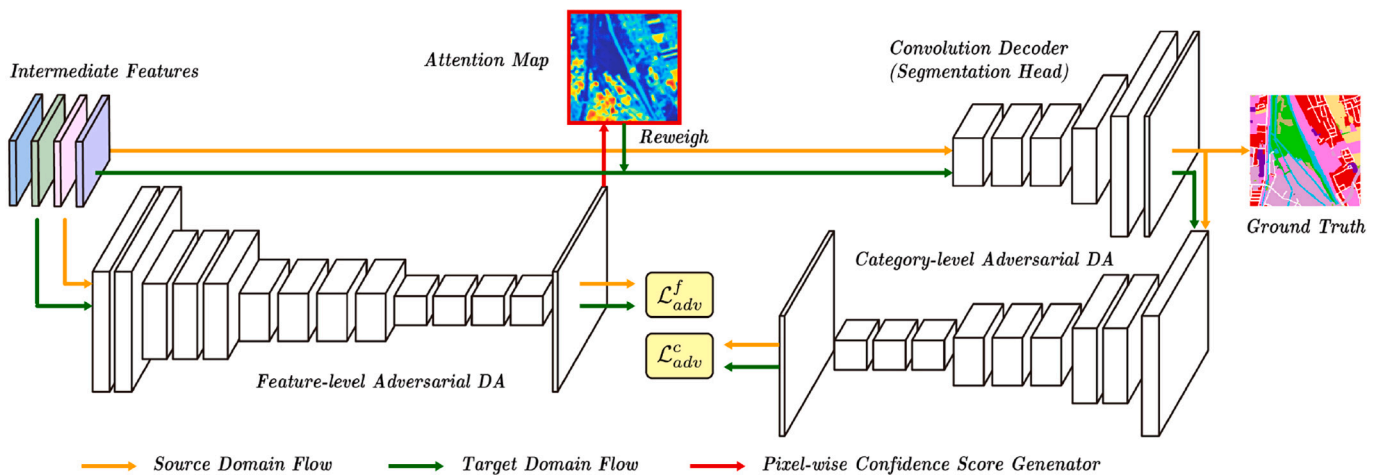


Fig. 7. A diagram of adversarial domain adaptation used in HighDAN, which consists of feature-level adversarial DA and category-level adversarial DA. The pixel-wise attention map is generated by the feature-level adversarial DA and explored for reweighing and correcting the feature representations from the target domain, making it compatible with features from the source domain. The category-level adversarial DA can further refine the segmentation results.

on the category centers, which drives the same category closer to each other in different domains. Visually, Fig. 7 gives the corresponding diagram of adversarial DA used in HighDAN.

4.5. Convolution decoder

Given the aligned feature representations A via DA, a segmentation head in the form of the convolution decoder is further applied on A to progressively reconstruct feature maps consistent with the size of semantic labels, which can be formulated by

$$U = T_{W,B}(A), \quad (4)$$

where U denotes the predicted semantic label map, and the function $T(\cdot)$ represents the decoder module that consists of convolution, BN, ReLU activation function, and $2\times$ upsampling operation.

Algorithm 1. A flowchart of the proposed HighDAN.

Input: Different RS modality data (\mathbf{X}_k) including labeled (source domain, \mathbf{X}_k^s) and unlabeled (target domain, \mathbf{X}_k^t) samples, and semantic segmentation labels (\mathbf{Y}) corresponding to labeled samples.

Output: Model, predicted segmentation results $\hat{\mathbf{Y}}$

- 1 **Step 1 (Data preparation):** Band-wise normalization for \mathbf{X}_k , and feed the normalized data into networks with corresponding labels.
- 2 **Step 2 (Model Training):**
- 3 **for** $t = 1$ **to** T **do**
- 4 Part 1: Multimodal Encoder
- 5 1) Feature Extraction Head:
- 6 Convolution block: Feature representations \mathbf{Z}_k obtained by Eq. (1);
- 7 Bottleneck block: Feature representations \mathbf{Q}_k obtained by Eq. (2);
- 8 2) Multimodal HR Head: HR representations \mathbf{V}_k obtained by Eq. (3).
- 9 Part 2: Adversarial Domain Adaptation
- 10 1) Feature Concatenation: $\mathbf{V} = \{\mathbf{V}_k\}_{k=1}^m$;
- 11 2) Feature Alignment: Aligned representations \mathbf{A} via feature-level DA on \mathbf{V} .
- 12 Part 3: Convolution Decoder
- 13 1) Predict semantic labels U by Eq. (4);
- 14 2) Compute the overall loss by Eq. (5);
- 15 3) Gradient backpropagation and update networks.
- 16 **end**
- 17 **Step 3 (Model Inference):**
- 18 1) Normalize multimodal testing data in a band-wise fashion;
- 19 2) Feed the normalized data into learned networks;
- 20 3) Obtain the predicted labels and output the final segmentation results.

4.6. Model training

A flowchart illustrating the proposed HighDAN model is outlined in Algorithm 1, with step-by-step procedures provided for clarity. Let $X \in \mathbb{R}^{hw \times N}$ and $Y \in \mathbb{R}^{l \times N}$ be the input images and the ground truth (GT) of semantic segmentation labels with hw and l dimensions, respectively, by N pixels. Then, x_i and y_i are denoted to be the corresponding i -th element (or pixel). With these definitions, the network concerning the to-be-updated parameters of W and B is trained by optimizing the following objective function. The overall loss \mathcal{L} in objective function is

$$\mathcal{L} = \mathcal{L}_{seg} + \lambda \mathcal{L}_{adv}^f + \mu \mathcal{L}_{adv}^c, \quad (5)$$

where λ and μ are defined as the penalty parameters to balance different terms in the training phase, and we set them to be both 0.5 empirically

and experimentally. More specifically, the three terms are detailed in the following.

The first term in Eq. (5) is the segmentation loss, which consists of multi-class cross-entropy loss and Dice loss, i.e.,

$$\mathcal{L}_{seg} = \mathcal{L}_{MCE} + \mathcal{L}_{Dice}. \quad (6)$$

\mathcal{L}_{MCE} calculates the loss for each pixel equally, and \mathcal{L}_{Dice} can alleviate the negative effects due to the imbalanced training samples, e.g.,

$$\mathcal{L}_{Dice} = 1 - \frac{2 \sum_{i=1}^N y_i \hat{y}_i}{\sum_{i=1}^N y_i + \sum_{i=1}^N \hat{y}_i}, \quad (7)$$

where \hat{y}_i denotes the predicted semantic label in the i -th pixel.

The second term in Eq. (5) is the feature-level adversarial loss. Unlike the vanilla GAN that utilizes the classic cross-entropy loss to train the discriminator, the least square loss in (Mao et al., 2017) is exploited in our DA task to avoid the gradient vanishing issue. Suppose the input modality data X^s is from the source domain and X^t is from the target

domain, the generator E_f and discriminator D_f can be alternatively optimized by minimizing

$$\begin{aligned} \mathcal{L}_{adv}^f(D_f) &= \mathbb{E}_{X^s} \left[(D_f(V^s) - 0)^2 \right] + \mathbb{E}_{X^t} \left[(D_f(V^t) - 1)^2 \right], \\ \mathcal{L}_{adv}^f(E_f) &= \mathbb{E}_{X^t} \left[(D_f(E_f(X^t)) - 0)^2 \right], \end{aligned} \quad (8)$$

where V^s and V^t are the feature maps (e.g., using Eq. (3)) extracted from the source domain and target domain, respectively, via multimodal HR encoder module (collectively known as the generator E_f in our case). To ensure the stability of feature maps of the target domain, we optimize V^t by using the updated rule of $V_{new}^t = V^t + V^t \odot \alpha$, where α denotes the attention map.

The third term in Eq. (5) is the category-level adversarial loss. The analogy to the second term, the adversary is performed at the category

level to improve the global adaptation ability in networks. We thus have the following adversarial loss:

$$\begin{aligned} \mathcal{L}_{adv}^c(D_c) &= \mathbb{E}_{X^s} [(D_c(U^s) - 0)^2] + \mathbb{E}_{X^t} [(D_c(U^t) - 1)^2], \\ \mathcal{L}_{adv}^c(P_c) &= \mathbb{E}_{X^t} [(D_c(P_c(X^t)) - 0)^2], \end{aligned} \quad (9)$$

where U^s and U^t are the output's decoder maps (e.g., using Eq. (4)) of the source domain and target domain via the proposed HighDAN, that is P_c as well.

5. Experiments

5.1. Experimental preparation

5.1.1. Implementation details

The proposed HighDAN is implemented on the PyTorch platform, and all deep models are trained using CPU with i7-6850K, RAM with 128GB, and GPU with 11GB NVIDIA GTX1080Ti. The Adam (Kingma and Ba, 2014) is selected as the network optimizer with the iterations of 6000 epochs for C2Seg-AB and 10,000 epochs for C2Seg-BW, respectively. The learning rates of the segmentation network and discriminator are both 0.0001 with a batch size of 16. By cropping the whole scene images with the sliding window at certain intervals, we collect 273 (or 7140) and 140 (or 850) images with the size of 128×128 (or 256×256) as a source domain for training and as a target domain for testing, respectively, on C2Seg-AB (or C2Seg-BW) datasets.

5.1.2. Network configuration

To enable the reconstruction of the proposed semantic segmentation network, we particularize the HighDAN architecture layer by layer. HighDAN successively starts with convolution blocks, and four bottleneck blocks are connected. Behind it, three feature encoding modules

Table 1

Layer-wise network configuration of the proposed HighDAN. Conv, BN, HR, and Num are abbreviations of convolution, batch normalization, high resolution, and number, respectively.

	Hyperspectral	Multispectral	SAR	Output Dimension
Convolution Block	3×3 Conv BN ReLU	3×3 Conv BN ReLU	3×3 Conv BN ReLU	64
Bottleneck Module	3×3 Conv BN ReLU	– – –	– – –	48
Feature Encoding 1	Bottleneck Block*4	Bottleneck Block*4	Bottleneck Block*4	48/96
Feature Encoding 2	Basic HR Sum Fusion	Basic HR Sum Fusion	Basic HR Sum Fusion	48/96/192
Feature Encoding 3	Basic HR Block*4 Sum Fusion	Basic HR Block*4 Sum Fusion	Basic HR Block*4 Sum Fusion	48/96/ 192/384
Fusion Layer	Feature Concatenation			720*3
Adversarial Module	Output the attention scores to reweigh feature maps			720*3
Convolution Decoder Module	3×3 Conv BN ReLU			256
	3×3 Conv BN ReLU			128
	3×3 Conv BN ReLU Transposed Convolution			64
	1×1 Conv			Num of Class

are adopted, each consisting of four basic HR blocks. The convolution decoder module is finally added with the combination of four decoding blocks. Between the two modules, an adversarial block and a concatenation-based fusion layer are embedded. For more details, the layer-wise network configuration of HighDAN is listed in Table 1.

5.1.3. Evaluation metrics

We evaluate the cross-city semantic segmentation performance qualitatively and quantitatively in terms of three metrics in common use: overall accuracy (OA), mean intersection over union (mIoU), and mean F1 score (mF1). OA, also known as pixel accuracy (PA), collects each pixel prediction:

$$OA = \frac{\sum_{i=1}^l p_{ii}}{\sum_{i=1}^l \sum_{j=1}^l p_{ij}}, \quad (10)$$

where i, j , and l represent the real value, predicted value, and the total number of classes, respectively, and the p_{ij} denotes the number of pixels that predict the i -th class as the j -th class. mIoU computes the intersection and union of two sets, which is defined by

$$mIoU = \frac{1}{l} \sum_{i=1}^l \frac{p_{ii}}{\sum_{j=1}^l p_{ij} + \sum_{j=1}^l p_{ji} - p_{ii}}. \quad (11)$$

mF1 score is the harmonic mean of precision (P) and recall (R), which is given by

$$mF1 = \frac{1}{l} \times \frac{2 \times P \times R}{P + R}, \quad (12)$$

where

$$P = \frac{1}{\sum_{i=1}^l \sum_{j=1}^l p_{ij} + p_{ii}}, \quad R = \frac{1}{\sum_{i=1}^l \sum_{j=1}^l p_{ji} + p_{ii}}. \quad (13)$$

5.1.4. Comparison with state-of-the-art models

We select current state-of-the-art (SOTA) semantic segmentation models for qualitative and quantitative performance comparison using multimodal RS data in the cross-city case. They are DeepLabv3 (Chen et al., 2017a), SegNet (Badrinarayanan et al., 2017), FastFCN (Wu et al., 2019), AdaptSeg (Tsai et al., 2018), deep subdomain adaptation network (DSAN) (Zhu et al., 2021), Dual-stream HR-Net (DualHR) (Ren et al., 2022), SegFormer (Xie et al., 2021), and our proposed HighDAN. The (Xie et al., 2021; Wu et al., 2019; Badrinarayanan et al., 2017; Chen et al., 2017a; Ren et al., 2022) models fail to consider data shifts between different domains, while the rest effectively embed the DA strategy into networks. It is worth noting that we prioritize using the same network configurations (given in the original literature) for compared approaches. Further, the relevant parameters can be slightly adjusted, making it applicable to the segmentation experiments of multimodal RS data.

5.2. Quantitative evaluation on C2Seg datasets

Tables 2 and 3 quantify the cross-city semantic segmentation performance by comparing current SOTA deep models with our HighDAN in terms of pixel-wise OA, mIoU, mF1, and F1 scores for each class as well as the model's computational complexity (FLOPs) and parameters on C2Seg datasets (C2Seg-AB and C2Seg-BW, respectively).

By and large, the cross-city segmentation performance of deep networks without the consideration of data shifts across domains (e.g., DeepLabv3, SegNet) is inferior to that of those models that effectively embed the DA strategy into networks. SegNet shows comparable performance with DeepLabv3 in terms of OA, mIoU, and mF1 on C2Seg-AB Datasets, while SegNet and DeepLabv3 hold similar segmentation accuracies on C2Seg-BW Datasets. For those DA-guided segmentation

Table 2

Quantitative performance comparison of deep semantic segmentation networks in terms of OA, mIoU, mF1, and F1 scores for each class as well as the model's computational complexity (FLOPs) and parameters on C2Seg-AB datasets. The symbol ‘-’ denotes no pixels correctly identified. The best result is marked in bold.

Class Name	DeepLab	SegNet	FastFCN	AdaptSeg	DSAN	DualHR	SegFormer	HighDAN
Surface water	0.31	18.53	24.24	22.96	12.20	38.49	37.45	52.06
Street network	9.54	3.00	13.00	13.61	5.80	20.98	0.79	21.35
Urban fabric	56.58	58.08	60.75	66.00	71.66	62.73	69.84	72.52
Industrial, commercial, and transport	35.86	39.88	41.48	42.99	26.15	48.13	57.98	62.37
Mine, dump, and construction sites	-	5.72	6.30	6.90	15.96	17.72	22.32	21.95
Artificial vegetated areas	25.35	0.33	16.21	19.11	24.38	18.01	26.56	33.41
Arable land	53.28	58.34	48.01	48.47	55.85	61.14	53.91	65.92
Permanent crops	-	-	-	-	-	-	-	1.17
Pastures	1.69	34.91	8.02	0.83	21.49	0.23	35.97	37.87
Forests	62.85	34.20	51.47	53.43	65.72	53.52	70.87	74.35
Shrub	0.59	1.60	0.39	11.23	17.64	2.34	5.06	14.51
Open spaces with no vegetation	-	-	-	-	-	-	-	-
Inland wetlands	-	-	-	-	-	-	-	-
OA (%)	42.46	43.40	43.51	44.53	47.68	48.05	53.40	57.66
mIoU (%)	12.64	12.65	13.33	14.30	16.33	16.49	20.16	24.76
mF1 (%)	18.93	19.59	20.76	21.96	24.37	24.87	29.29	35.19
FLOPs (B) / GFLOPs	56.42	30.53	38.58	38.34	53.67	37.46	5.82	40.11
Params (M)	72.68	29.48	56.70	44.86	107.44	15.33	28.97	16.55

Table 3

Quantitative performance comparison of deep semantic segmentation networks in terms of OA, mIoU, mF1, and F1 scores for each class as well as the model's computational complexity (FLOPs) and parameters on C2Seg-BW datasets. The symbol ‘-’ denotes no pixels correctly identified. The best result is marked in bold.

Class Name	DeepLab	SegNet	FastFCN	AdaptSeg	DSAN	DualHR	SegFormer	HighDAN
Surface water	50.42	37.87	45.39	59.57	-	60.51	78.49	78.37
Street network	16.30	6.17	2.38	17.05	-	0.29	0.05	0.58
Urban fabric	33.75	34.86	38.44	25.06	42.08	0.76	30.90	40.04
Industrial, commercial, and transport	2.48	1.87	27.63	32.77	25.68	24.19	20.38	43.67
Mine, dump, and construction sites	1.70	1.36	0.86	1.94	1.22	0.26	2.52	1.67
Artificial vegetated areas	8.97	2.23	8.83	8.99	9.55	4.19	10.79	9.28
Arable land	-	16.23	-	9.11	25.37	-	18.01	0.43
Permanent crops	-	1.57	-	0.30	-	-	1.77	0.10
Pastures	-	-	-	-	-	-	-	-
Forests	1.68	13.31	4.14	32.98	46.85	32.70	38.72	47.22
Shrub	0.72	-	-	-	-	0.22	10.48	0.26
Open spaces with no vegetation	-	-	-	-	-	0.33	0.01	-
Inland wetlands	-	-	-	-	-	0.01	0.01	-
OA (%)	19.51	15.94	21.22	29.26	18.55	31.97	33.56	39.58
mIoU (%)	5.45	5.17	5.96	8.92	7.09	6.17	10.89	11.92
mF1 (%)	8.92	8.88	9.82	14.44	11.60	9.53	16.32	17.69
FLOPs (B) / GFLOPs	213.01	122.13	154.32	144.58	214.67	149.86	23.30	160.45
Params (M)	72.68	29.48	56.70	44.86	107.44	15.33	28.97	16.55

networks, the adversarial DA methods (e.g., FastFCN, AdaptSeg) show competitive results compared to DSAN based on the local maximum mean discrepancy. Although FastFCN and AdaptSeg perform moderately lower than DSAN at an average decrease of 3%~4% OAs, 2%~3% mIoUs, and 2%~4% mF1s, respectively, yet their F1 scores for each category are holistically comparable to DSANs' and the main differences lie in certain special categories, e.g., *Pastures*, *Forests*, *Shrub*, etc. on the C2Seg-AB datasets. It is important to note that when confronted with more complex and extensive datasets e.g., C2Seg-BW, the generalization capability of DSAN appears to be somewhat constrained in comparison to FastFCN and AdaptSeg.

Furthermore, the HR-Net backbone architecture can offer greater potential for extracting a wealth of semantic information from multimodal RS data in comparison with the CNNs-based backbone in the semantic segmentation task. For example, DualHR brings increments of 13% OA based on DSAN on C2Seg-BW datasets, but the performance is basically identical to those on C2Seg-AB datasets, compared to DSAN. However, it is essential to note that transformer-based methods (i.e., SegFormer) consistently demonstrate competitive and stable performance on both C2Seg datasets, achieving the second-highest results across all evaluation indices. Not unexpectedly, the proposed HighDAN achieves the best segmentation performance by 4.26%, 4.60%, and 5.90% gains in OA, mIoU, and mF1 (cf. SegFormer) on C2Seg-AB

datasets, while there is also a nearly similar trend, even higher performance (e.g., over 6% OA increase), on C2Seg-BW datasets. A more noteworthy point to demonstrate the superiority of HighDAN lies in that HighDAN obtains the highest F1 scores in many dominated categories, e.g., *Surface water*, *Street network*, *Urban fabric*, *Arable land*, *Forests*, etc. on either C2Seg-AB or C2Seg-BW datasets. We have to admit, however, that C2Seg is a very challenging semantic segmentation dataset. It is observed that some categories are hardly identified, that is, the segmentation results for certain classes are 0% and few are approximately close to 0%.

5.3. Visual comparison on C2Seg datasets

Figs. 8 and 9 visualize the segmentation maps of eight different algorithms in terms of 13 semantic categories for the whole scenes of Berlin city and Wuhan city on C2Seg datasets. There is a more significant visual difference between predicted segmentation results and GT (on both Berlin and Wuhan scenes) in DeepLab and SegNet. On the one hand, *Pastures* are prone to be wrongly classified as *Arable land*, while *Inland Wetlands* are heavily identified to be *Forests* in the Wuhan scene. On the other hand, *Urban fabric* and *Industrial, commercial, and transport* are easily confused due to their similar spectral characteristics and functions. Compared to the first two methods, FastFCN has visible

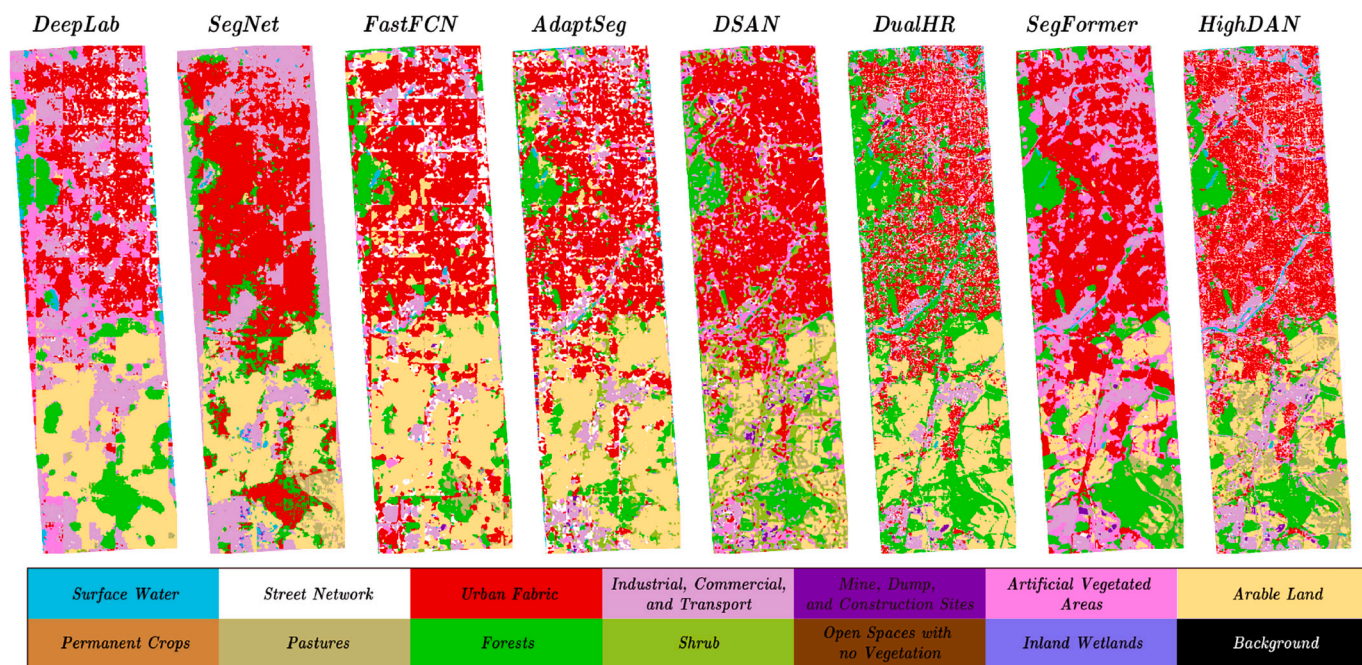


Fig. 8. Visualization of semantic segmentation results obtained by successively using DeepLab, SegNet, FastFCN, AdaptSeg, DSAN, DualHR, SegFormer, and our proposed HighDAN on C2Seg-AB (testing set: Berlin) datasets.

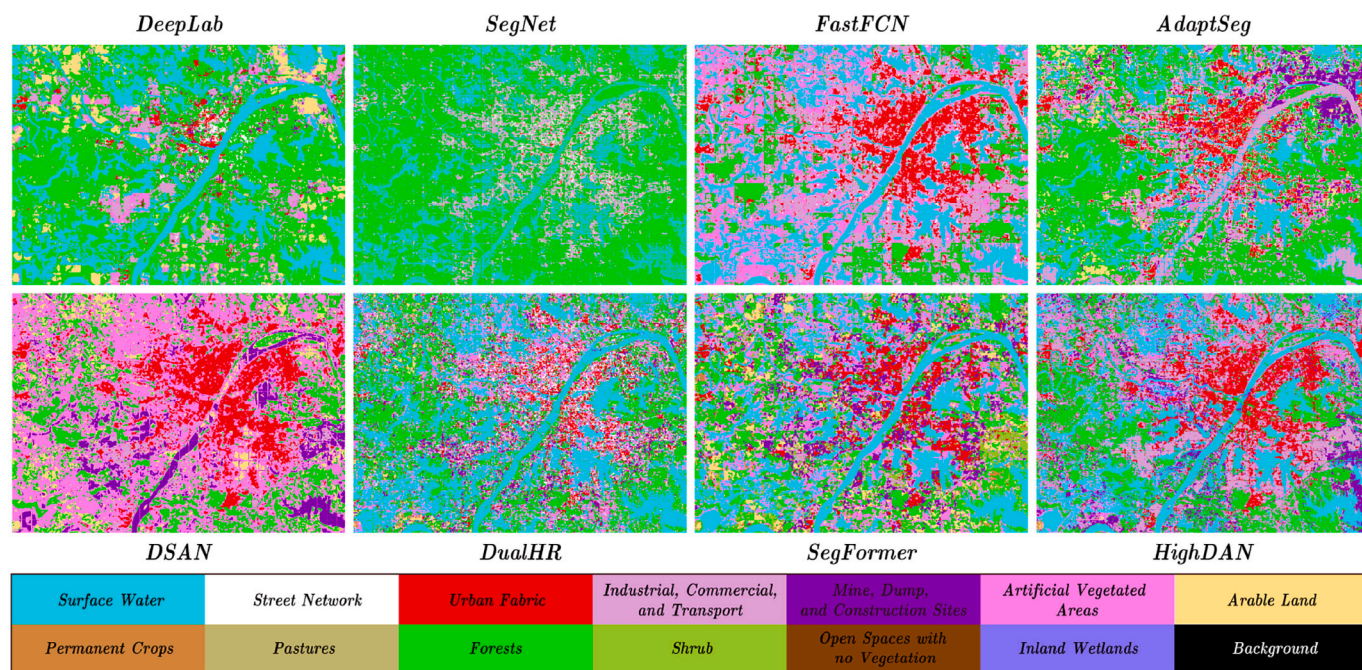


Fig. 9. Visualization of semantic segmentation results obtained by successively using DeepLab, SegNet, FastFCN, AdaptSeg, DSAN, DualHR, SegFormer, and our proposed HighDAN on C2Seg-BW (testing set: Wuhan) datasets.

advantages in discriminating the semantic category of *Urban Fabric* and *Artificial vegetated areas*, while *AdaptSeg* is capable of identifying *Arable Land* more accurately (despite the over-recognition of *Shrub* and *Pastures* being *Arable Land*). We have to admit, however, that the ability of *AdaptSeg* to classify urban-related semantic elements remains limited. *DSAN* is a good recognizer for urban-related and vegetation semantic categories, which can well distinguish *Urban fabric* and *Industrial, commercial, and transport* as well as *Forests* and *Arable land*. In the family of HR-Net, *DualHR* is sensitive to capturing water bodies from a big urban scene but fails to detect urban accurately, while the proposed *HighDAN*

visually shows, as expected, comparatively realistic segmentation maps closer to GT (cf. *SegFormer*). In particular, water bodies, urban, and forests have nearly identical semantic segmentation profiles to those in GT. There is, notwithstanding, considerable room for improvement in *HighDAN*, to further enhance the identification and recognition ability in *Arable Land*, *Street Network*, and *Inland Wetlands*.

In addition to the scene-wide segmentation visualization, we also provide detailed segmentation results in sub-regions, as shown in *Figs. 10 and 11* corresponding to *Figs. 8 and 9*, respectively. The visual comparison of local semantic segmentation results highlights the

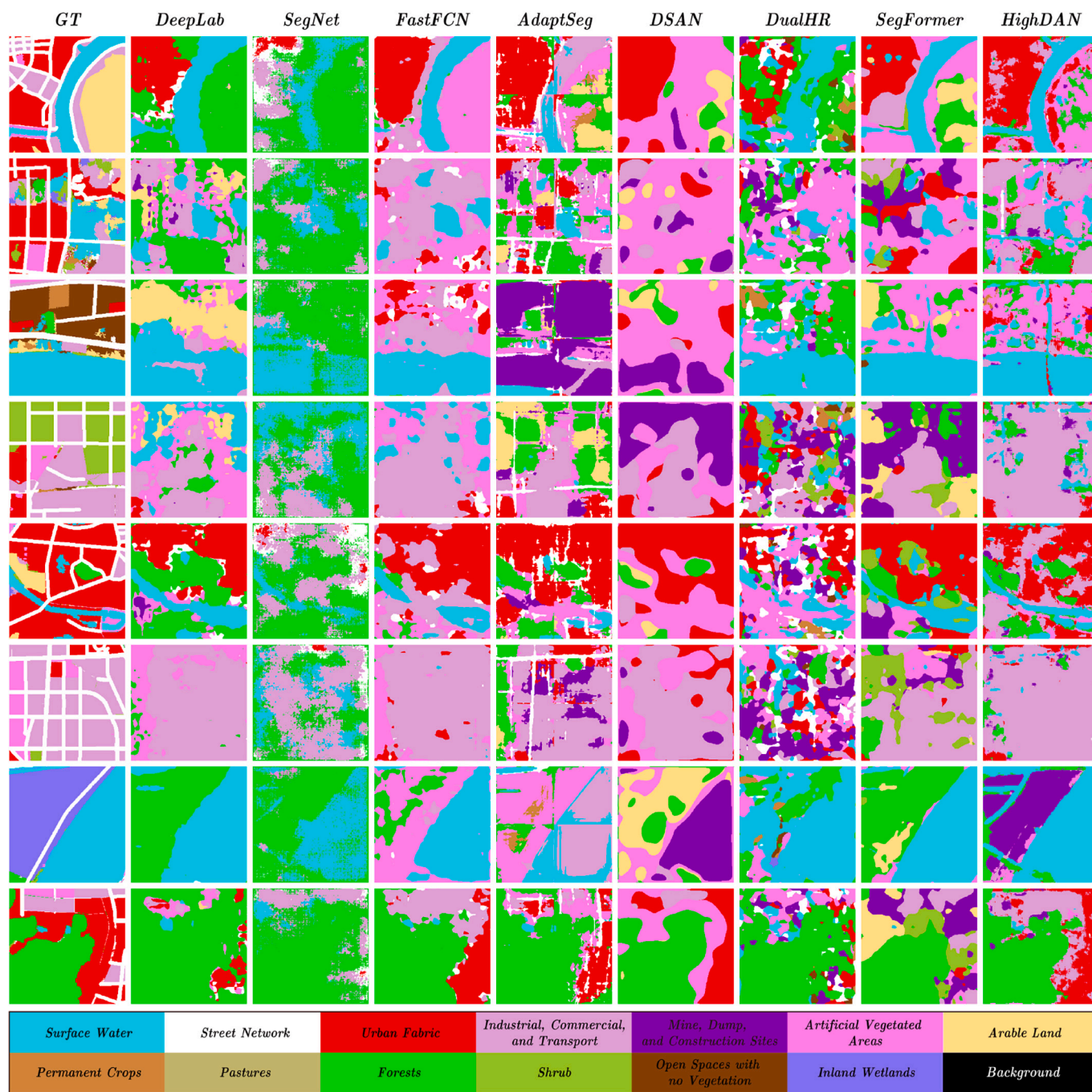


Fig. 10. Visualizing semantic segmentation results of sub-regions corresponding to Fig. 8 for all compared models.

advantages of the proposed HighDAN in terms of preserving fine-grained details of objects in RS images. Further, HighDAN is capable of effectively capturing small-scale features and details of the objects. This was particularly evident in the cases of man-made objects with intricate shapes and textures, where HR-Net-based models (i.e., DualHR, HighDAN) are apt to segment the objects without losing important details. In comparison, the baseline methods, such as DeepLab and SegNet, yield segmentation results with a severe loss of detailed information, which shows their limitations in capturing tiny and irregular objects or structures. While other compared methods have demonstrated some improvement in identifying semantic categories with varying shapes, their ability in recognition accuracy and boundary segmentation remains limited. Yet the visual analysis also reveals that our HighDAN can effectively adapt to changes in imaging conditions and variability in

object appearance across domains or cities, resulting in improved segmentation accuracy and robustness. It should be noted, however, that some categories are almost entirely misclassified in certain sub-images, such as *Artificial vegetated areas*, *Open spaces with no vegetation*, *Inland wetlands*, *Shrub*. To sum up, these observations highlight the need for continued exploration and optimization of semantic segmentation methods in the aspects of HR feature extraction and DA enhancement.

To further assess the effectiveness of our proposed HighDAN model in extracting class-related semantic information, we visualize class activation maps (CAMs) (Zhou et al., 2016) on the C2Seg-AB datasets, as shown in Fig. 12. These visualizations demonstrate that HighDAN excels in capturing high-level semantic information with precise class activation, even for small classes, e.g., *Street Network*. This capability underscores the model's proficiency in semantic segmentation tasks.



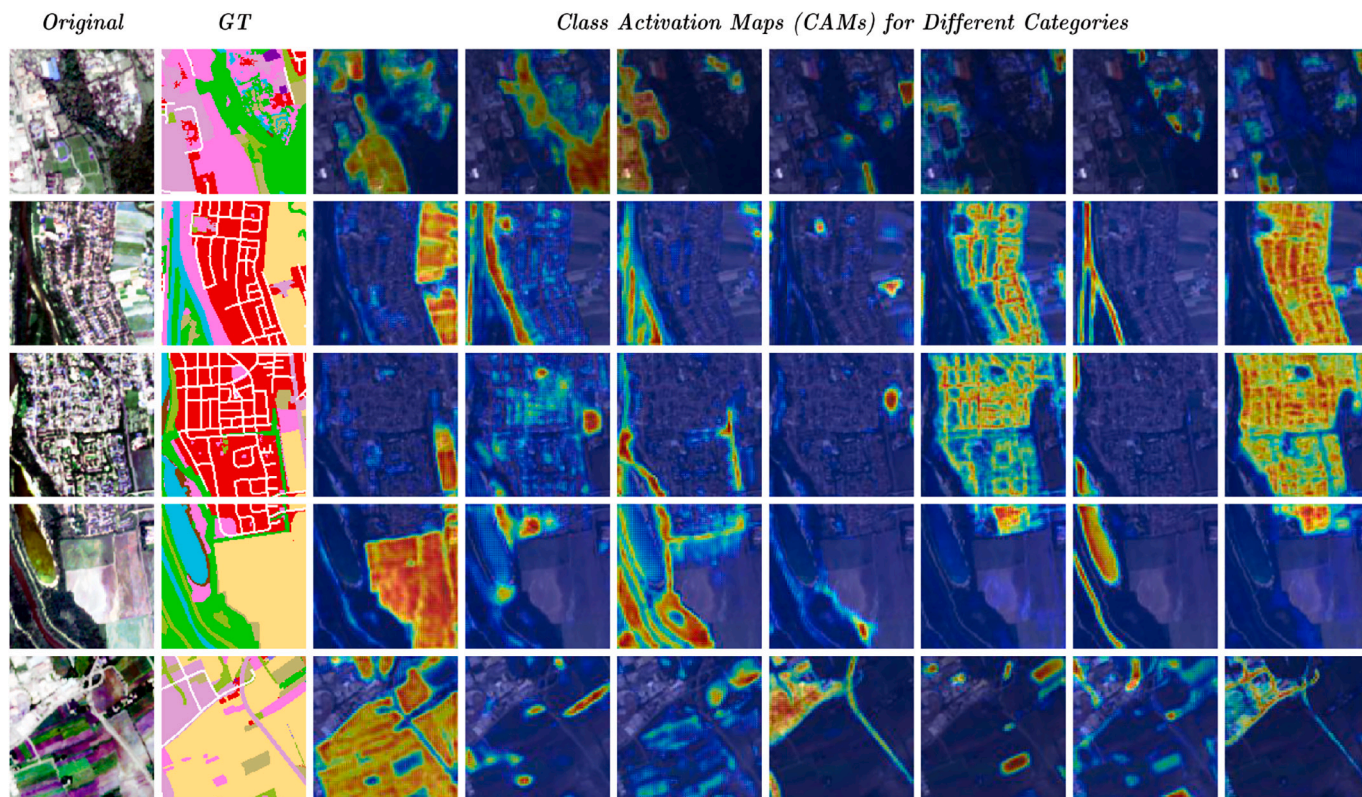


Fig. 12. Visualization of class activation maps (CAMs) using HighDAN on the C2Seg-AB datasets.

Table 4

Ablation analysis on C2Seg-AB datasets, where ‘Bottleneck’ and ‘HR’ represent the bottleneck feature extraction module and the high-resolution module, while ‘Feature DA’ and ‘Category DA’ denote feature-level domain adaptation module and category-level domain adaptation module, respectively. The best results are marked in bold.

Model	Bottleneck	HR	Feature DA	Category DA	Dice Loss	OA (%)	mIoU (%)	mF1 (%)
SegNet	✗	✗	✗	✗	✗	43.40	12.65	19.59
HR-Net	✗	✓	✗	✗	✗	48.53	15.86	23.35
HR-Net	✓	✓	✗	✗	✗	49.85	18.95	27.80
HR-Net	✓	✓	✗	✗	✓	50.59	19.90	29.04
HighDAN	✓	✓	✗	✓	✓	53.74	20.62	29.84
HighDAN	✓	✓	✓	55	✓	52.73	21.51	31.09
HighDAN	✓	✓	✓	✓	✗	56.44	24.03	34.32
HighDAN	✓	✓	✓	✓	✓	57.66	24.76	35.19

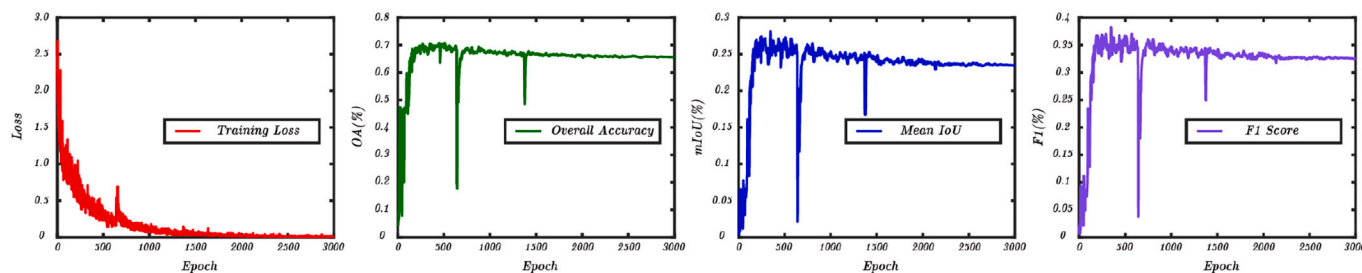


Fig. 13. Performance analysis of the proposed HighDAN in robustness in terms of training loss, OA, mIoU, and mF1 score on the individual C2Seg-AB datasets.

(53.91%, 31.81%, 21.74%).

In addition, we presented the results, which included training loss, OA, mIoU, and mF1, for individual datasets using an 8:2 training and testing ratio, specifically focusing on C2Seg-AB. This comprehensive evaluation process allowed us to assess the performance and robustness of the proposed HighDAN model. Fig. 13 illustrates that the training loss

of the model exhibits a consistent decrease throughout the training process, indicating the model’s stability and robust convergence during learning. As expected, there is a similar trend in segmentation performance (i.e., OA, mIoU, mF1) across individual C2Seg-AB datasets.

6. Conclusion

Fast monitoring and understanding of urban environments are inseparable from explosively developing RS techniques. The success of RS enables the accurate identification and detection of materials of interest in complex urban scenes. As a primary and indispensable research topic, the semantic segmentation of RS images has long dominated the overwhelming role in the land use land cover classification of urban environments. However, these well-designed and dedicated segmentation methodologies are, for the most part, applicable only to one single city case. This severely hinders the application deployments across cities or regions, since urban planning and management, e.g., policy-making, land use, spatial layout, information transfer, etc., have to accommodate multi-city studies.

For the reason mentioned above, we in this paper focus on investigating cross-city semantic segmentation and provide solutions accordingly. The solutions are two-fold. On the one hand, we build a multimodal RS benchmark dataset (i.e., C2Seg) to solve the issue of insufficient discriminative information by only using single modality RS data for cross-city semantic segmentation. On the other hand, we propose a cutting-edge deep network architecture, HighDAN for short, by embedding the adversarial learning-based DA's idea into HR-Net with Dice Loss (to reduce the effects of the class imbalance), making it largely possible to break the semantic segmentation performance bottleneck in terms of accuracy and generalization ability from cross-city studies. Extensive experiments conducted on the C2Seg datasets demonstrate that our HighDAN achieves the best segmentation performance, which beats other SOTA competitors in almost all important indices. Moreover, we will also release the C2Seg benchmark datasets and the corresponding source codes, contributing to the interpretation research of urban environments across cities.

In future work, we aim to extend the C2Seg datasets in a wide range of cities on a national scale and even a global scale for the better study of cross-city semantic segmentation. In particular, the development of hyperspectral RS, especially concerning its application on a large scale, is indeed an issue that warrants urgent attention and exploration, due to certain inherent imaging constraints associated with hyperspectral RS technology. Furthermore, more advanced AI models should be developed and made accessible by further considering explicit and explainable knowledge embedding, e.g., geometric priors, climate characteristics, and urban morphological properties, to guide deep networks to learn more accurate segments and promote the model's generalization ability across cities.

CRedit authorship contribution statement

Danfeng Hong: Conceptualization, Methodology, Writing – original draft. **Bing Zhang:** Conceptualization, Methodology, Supervision. **Hao Li:** Data curation, Writing – original draft. **Yuxuan Li:** Data curation, Software, Validation. **Jing Yao:** Methodology, Visualization, Investigation. **Chenyu Li:** Software, Validation. **Martin Werner:** Software, Writing – review & editing. **Jocelyn Chanussot:** Conceptualization, Writing – review & editing. **Alexander Zipf:** Data curation, Validation. **Xiao Xiang Zhu:** Data curation, Writing – review & editing.

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.

Data availability

Data will be made available on request.

Acknowledgements

The authors would like to thank Ms. Zhu Han and Ms. Luyang Cai for pre-processing the Gaofen data used in this paper. This work was supported by the National Key Research and Development Program of China under Grant 2022YFB3903401, the National Natural Science Foundation of China under Grant 42241109, Grant 42271350, and Grant 62201553, the MIAI@Grenoble Alpes (ANR-19-P3IA-0003), the Klaus Tschira Stiftung (KTS) Heidelberg, and the AXA Research Fund.

References

- Yuan, Q., Shen, H., Li, T., Li, Z., Li, S., Jiang, Y., Xu, H., Tan, W., Yang, Q., Wang, J., et al., 2020. Deep learning in environmental remote sensing: achievements and challenges. *Remote Sens. Environ.* 241, 111716.
- Yuan, X., Shi, J., Gu, L., 2021. A review of deep learning methods for semantic segmentation of remote sensing imagery. *Expert Syst. Appl.* 169, 114417.
- Kampffmeyer, M., Salberg, A.-B., Jenssen, R., 2016. Semantic segmentation of small objects and modeling of uncertainty in urban remote sensing images using deep convolutional neural networks. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp. 1–9.
- Gal, Y., Ghahramani, Z., 2016. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In: *Proceedings of the International Conference on Machine Learning (ICML)*. PMLR, pp. 1050–1059.
- Kemker, R., Salvaggio, C., Kanan, C., 2018. Algorithms for semantic segmentation of multispectral remote sensing imagery using deep learning. *ISPRS J. Photogramm. Remote Sens.* 145, 60–77.
- Yi, Y., Zhang, Z., Zhang, W., Zhang, C., Li, W., Zhao, T., 2019. Semantic segmentation of urban buildings from vhr remote sensing imagery using a deep convolutional neural network. *Remote Sens.* 11, 1774.
- Diakogiannis, F.I., Waldner, F., Caccetta, P., Wu, C., 2020. Resunet-a: A deep learning framework for semantic segmentation of remotely sensed data. *ISPRS J. Photogramm. Remote Sens.* 162, 94–114.
- Du, S., Du, S., Liu, B., Zhang, X., 2021. Mapping large-scale and fine-grained urban functional zones from vhr images using a multi-scale semantic segmentation network and object based approach. *Remote Sens. Environ.* 261, 112480.
- Wang, L., Li, R., Zhang, C., Fang, S., Duan, C., Meng, X., Atkinson, P.M., 2022. Unetformer: a unet-like transformer for efficient semantic segmentation of remote sensing urban scene imagery. *ISPRS J. Photogramm. Remote Sens.* 190, 196–214.
- He, X., Zhou, Y., Zhao, J., Zhang, D., Yao, R., Xue, Y., 2022. Swin transformer embedding unet for remote sensing image semantic segmentation. *IEEE Trans. Geosci. Remote Sens.* 60, 1–15.
- Wang, Z., Wang, Q., Yang, Y., Liu, N., Chen, Y., Gao, J., 2023. Seismic facies segmentation via a segformer-based specific encoder–decoder–hypercolumns scheme. *IEEE Trans. Geosci. Remote Sens.* 61, 1–11.
- Xie, E., Wang, W., Yu, Z., Anandkumar, A., Alvarez, J.M., Luo, P., 2021. Segformer: simple and efficient design for semantic segmentation with transformers. *Adv. Neural Inf. Proces. Syst.* 34, 12077–12090.
- Dalla Mura, M., Prasad, S., Pacifici, F., Gamba, P., Chanussot, J., Benediktsson, J.A., 2015. Challenges and opportunities of multimodality and data fusion in remote sensing. *Proc. IEEE* 103, 1585–1601.
- Audebert, N., Saux, B.L., Lefèvre, S., 2016. Semantic segmentation of earth observation data using multimodal and multi-scale deep networks. In: *Proceedings of the Asian Conference on Computer Vision (ACCV)*. Springer, pp. 180–196.
- Long, J., Shelhamer, E., Darrell, T., 2015. Fully convolutional networks for semantic segmentation. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3431–3440.
- Wu, H., Zhang, J., Huang, K., Liang, K., Yu, Y., 2019. Fastfcn: Rethinking dilated convolution in the backbone for semantic segmentation.
- Badrinarayanan, V., Kendall, A., Cipolla, R., 2017. Segnet: a deep convolutional encoder-decoder architecture for image segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* 39, 2481–2495.
- Audebert, N., Le Saux, B., Lefèvre, S., 2018. Beyond rgb: very high resolution urban remote sensing with multimodal deep networks. *ISPRS J. Photogramm. Remote Sens.* 140, 20–32.
- Hong, D., Gao, L., Hang, R., Zhang, B., Chanussot, J., 2020. Deep encoder-decoder networks for classification of hyperspectral and lidar data. *IEEE Geosci. Remote Sens.* 19, 1–5.
- Wieland, M., Li, Y., Martinis, S., 2019. Multi-sensor cloud and cloud shadow segmentation with a convolutional neural network. *Remote Sens. Environ.* 230, 111203.
- Wurm, M., Stark, T., Zhu, X.X., Weigand, M., Taubenböck, H., 2019. Semantic segmentation of slums in satellite images using transfer learning on fully convolutional neural networks. *ISPRS J. Photogramm. Remote Sens.* 150, 59–69.
- Segal-Rozenhaimer, M., Li, A., Das, K., Chirayath, V., 2020. Cloud detection algorithm for multi-modal satellite imagery using convolutional neural networks (cnn). *Remote Sens. Environ.* 237, 111446.
- Chen, L.-C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.L., 2017. Deeplab: semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE Trans. Pattern Anal. Mach. Intell.* 40, 834–848.

- Ren, B., Ma, S., Hou, B., Hong, D., Chanussot, J., Wang, J., Jiao, L., 2022. A dual-stream high resolution network: deep fusion of gf-2 and gf-3 data for land cover classification. *Int. J. Appl. Earth Obs. Geoinf.* 112, 102896.
- Sun, K., Xiao, B., Liu, D., Wang, J., 2019. Deep high-resolution representation learning for human pose estimation. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5693–5703.
- Adriano, B., Yokoya, N., Xia, J., Miura, H., Liu, W., Matsuoka, M., Koshimura, S., 2021. Learning from multimodal and multitemporal earth observation data for building damage mapping. *ISPRS J. Photogramm. Remote Sens.* 175, 132–143.
- Ganin, Y., Lempitsky, V., 2015. Unsupervised domain adaptation by backpropagation. In: *Proceedings of the International Conference on Machine Learning (ICML)*. PMLR, pp. 1180–1189.
- Chen, Y.-H., Chen, W.-Y., Chen, Y.-T., Tsai, B.-C., Wang, Y.-C., Frank, Sun, M., 2017. No more discrimination: Cross city adaptation of road scene segmenters. In: *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 1992–2001.
- Tsai, Y.-H., Hung, W.-C., Schuster, S., Sohn, K., Yang, M.-H., Chandraker, M., 2018. Learning to adapt structured output space for semantic segmentation. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7472–7481.
- 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.
- Zhu, Y., Zhuang, F., Wang, J., Ke, G., Chen, J., Bian, J., Xiong, H., He, Q., 2021. Deep subdomain adaptation network for image classification. *IEEE Trans. Neural Netw. Learn. Syst.* 32, 1713–1722.
- Li, H., Herfort, B., Lautenbach, S., Chen, J., Zipf, A., 2022. Improving openstreetmap missing building detection using few-shot transfer learning in sub-saharan africa. *Trans. GIS* 26, 3125–3146.
- Li, Y., Shi, T., Zhang, Y., Chen, W., Wang, Z., Li, H., 2021. 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.
- Wang, J., Ma, A., Zhong, Y., Zheng, Z., Zhang, L., 2022. Cross-sensor domain adaptation for high spatial resolution urban land-cover mapping: from airborne to spaceborne imagery. *Remote Sens. Environ.* 277, 113058.
- Ma, A., Zheng, C., Wang, J., Zhong, Y., 2023. Domain adaptive land-cover classification via local consistency and global diversity. *IEEE Trans. Geosci. Remote* 61, 1–17.
- Hong, D., Yokoya, N., Xia, G.-S., Chanussot, J., Zhu, X.-X., 2020. X-modalnet: a semi-supervised deep cross-modal network for classification of remote sensing data. *ISPRS J. Photogramm. Remote Sens.* 167, 12–23.
- Hong, D., Yao, J., Meng, D., Xu, Z., Chanussot, J., 2021. Multimodal gans: toward crossmodal hyperspectral–multispectral image segmentation. *IEEE Trans. Geosci. Remote Sens.* 59, 5103–5113.
- Ji, S., Wang, D., Luo, M., 2021. Generative adversarial network-based full-space domain adaptation for land cover classification from multiple-source remote sensing images. *IEEE Trans. Geosci. Remote Sens.* 59, 3816–3828.
- Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y., 2020. Generative adversarial networks. *Commun. ACM* 63, 139–144.
- Zhao, X., Zhang, M., Tao, R., Li, W., Liao, W., Philips, W., 2022. Cross-domain classification of multisource remote sensing data using fractional fusion and spatial-spectral domain adaptation. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* 15, 5721–5733.
- Hong, D., Gao, L., Yokoya, N., Yao, J., Chanussot, J., Du, Q., Zhang, B., 2021. More diverse means better: multimodal deep learning meets remote-sensing imagery classification. *IEEE Trans. Geosci. Remote Sens.* 59, 4340–4354.
- Xia, J., Yokoya, N., Adriano, B., Broni-Bediako, C., 2023. Openearthmap: A benchmark dataset for global high-resolution land cover mapping. In: *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 6254–6264.
- Segl, K., Guanter, L., Rogass, C., Kuester, T., Roessner, S., Kaufmann, H., Sang, B., Mogulsky, V., Hofer, S., 2012. EeteS—The EnMAP end-to-end simulation tool. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* 5, 522–530.
- Okujeni, A., van der Linden, S., Hostert, P., 2016. Berlin-urban-gradient dataset 2009-an enmap preparatory flight campaign.
- Hu, J., Liu, R., Hong, D., Camero, A., Yao, J., Schneider, M., Kurz, F., Segl, K., Zhu, X.-X., 2023. Mdas: a new multimodal benchmark dataset for remote sensing. *Earth Syst. Sci. Data* 15, 113–131.
- Hansen, M.C., Roy, D.P., Lindquist, E., Adusei, B., Justice, C.O., Altstatt, A., 2008. A method for integrating modis and landsat data for systematic monitoring of forest cover and change in the Congo basin. *Remote Sens. Environ.* 112, 2495–2513.
- Schultz, M., Voss, J., Auer, M., Carter, S., Zipf, A., 2017. Open land cover from openstreetmap and remote sensing. *Int. J. Appl. Earth Obs. Geoinf.* 63, 206–213.
- Liu, Y.-N., Sun, D.-X., Hu, X.-N., Ye, X., Li, Y.-D., Liu, S.-F., Cao, K.-Q., Chai, M.-Y., Zhang, J., Zhang, Y., et al., 2019. The advanced hyperspectral imager: aboard china's gaofen-5 satellite. *IEEE Geosci. Remote Sens. Mag.* 7, 23–32.
- Yommy, A.S., Liu, R., Wu, S., 2015. Sar image despeckling using refined lee filter. In: *2015 7th International Conference on Intelligent Human-Machine Systems and Cybernetics*, volume 2. IEEE, pp. 260–265.
- Danielson, J.J., Gesch, D.B., 2011. Global multi-resolution terrain elevation data 2010 (GMTED2010). US Department of the Interior, US Geological Survey Washington, DC, USA.
- Feranec, J., Soukup, T., Hazeu, G., Jaffrain, G., 2016. European landscape dynamics: CORINE land cover data. CRC Press.
- Krizhevsky, A., Sutskever, I., Hinton, G.E., 2017. Imagenet classification with deep convolutional neural networks. *Commun. ACM* 60, 84–90.
- Simonyan, K., Zisserman, A., 2015. Very deep convolutional networks for large-scale image recognition. In: *Proceedings of the International Conference on Learning Representations (ICLR)*.
- Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A., 2015. Going deeper with convolutions. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1–9.
- Ronneberger, O., Fischer, P., Brox, T., 2015. U-net: convolutional networks for biomedical image segmentation. In: *International Conference on Medical Image Computing and Computer-assisted Intervention (MICCAI)*. Springer, pp. 234–241.
- Noh, H., Hong, S., Han, B., 2015. Learning deconvolution network for semantic segmentation. In: *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 1520–1528.
- Newell, A., Yang, K., Deng, J., 2016. Stacked hourglass networks for human pose estimation. In: *Proceedings of the European Conference on Computer Vision*. Springer, pp. 483–499.
- Li, X., Sun, X., Meng, Y., Liang, J., Wu, F., Li, J., 2020. Dice loss for data-imbalanced nlp tasks. In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 465–476.
- Isola, P., Zhu, J.-Y., Zhou, T., Efros, A.A., 2017. Image-to-image translation with conditional adversarial networks. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1125–1134.
- Yu, F., Zhang, M., Dong, H., Hu, S., Dong, B., Zhang, L., 2021. Dast: Unsupervised domain adaptation in semantic segmentation based on discriminator attention and self-training. In: *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, pp. 10754–10762.
- Mao, X., Li, Q., Xie, H., Lau, R.Y., Wang, Z., Smolley, S., Paul, 2017. Least squares generative adversarial networks. In: *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 2794–2802.
- Kingma, D., Ba, J., 2014. Adam: A method for stochastic optimization.
- Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., Torralba, A., 2016. Learning deep features for discriminative localization. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2921–2929.