Cross-City Semantic Segmentation (C2Seg) in Multimodal Remote Sensing: Outcome of the 2023 IEEE WHISPERS C2Seg Challenge

Yuheng Liu, Ye Wang, *Graduate Student Member, IEEE*, Yifan Zhang, *Member, IEEE*, Shaohui Mei, *Senior Member, IEEE*, Jiaqi Zou, Zhuohong Li, *Student Member, IEEE*, Fangxiao Lu, Wei He, *Senior Member, IEEE*, Hongyan Zhang, *Senior Member, IEEE*, Huilin Zhao, Chuan Chen, Cong Xia, Hao Li, *Member, IEEE*, Gemine Vivone, *Senior Member, IEEE*, Ronny Hänsch, *Senior Member, IEEE*, Gulsen Taskin, *Senior Member, IEEE*, Jing Yao, *Member, IEEE*, A. K. Qin, *Senior Member, IEEE*, Bing Zhang, *Fellow, IEEE*, Jocelyn Chanussot, *Fellow, IEEE*, and Danfeng Hong, *Senior Member, IEEE*

***Abstract*—Given the ever-growing availability of remote sensing data (e.g., Gaofen in China, Sentinel in the EU, and Landsat in the USA), multimodal remote sensing techniques have been garnering increasing attention and have made extraordinary progress in various Earth observation (EO)-related tasks. The data acquired by different platforms can provide diverse and complementary information. The joint exploitation of multimodal remote sensing has been proven effective in improving the existing methods of land-use/land-cover segmentation in urban environments. To boost technical breakthroughs and accelerate the development of EO applications across cities and regions, one important task is to build novel cross-city semantic segmentation models based on modern artificial intelligence technologies and emerging multimodal remote sensing data. This leads to the development of better semantic segmentation models with high transferability among different cities and regions. The Cross-City Semantic Segmentation contest is organized in conjunction with the 13th Workshop on Hyperspectral Image and Signal Processing: Evolution in Remote Sensing (WHISPERS).**

***Index Terms*—Artificial intelligence (AI), cross-city, deep learning, hyperspectral, land cover, multimodal benchmark datasets, remote sensing, semantic segmentation.**

I. Introduction

REMOTE sensing (RS) presents an essential and prominent approach to acquiring large-scale and high-quality Earth observation (EO) data in a short time, which significantly advances the development of EO techniques. Nevertheless, the traditional expert system-centric RS data analysis has almost reached its potential, and thus, becomes insufficient to facilitate the increasing demand of the big EO data era, particularly when dealing with complex urban scenes on a global scale. Artificial intelligence (AI) techniques [1] provide one promising solution that is capable of discovering potentially valuable knowledge from the vast amount of existing EO data more efficiently, enabling a fast and accurate understanding of the contemporary urban environment.

Given advances in the development of AI models, e.g., deep learning, there have been successful applications for numerous RS and geoscience applications [2], [3], [4], [5], [6], [7], [8], [9], [10], which have been proven to be particularly applicable to urban environments where the types, characteristics, and spatial distributions of surface elements are significantly consistent and similar. However, the capability of adapting to diverse urban environmental differences with highly spatio-temporal and regional change remains limited. In this context, one can envision a possible solution as being twofold: on the one hand, the joint exploitation of multimodal RS data has been proven to help improve the processing ability of cross-city or cross-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 unavoidable trend to mitigate the semantic gap between different urban environments, making it mutually transferable for AI-based RS data analysis.

Existing methods for semantic segmentation of RS images in terms of the design of network architecture, module details, and the use of loss functions have achieved promising and superior performance [11], [12], [13], [14]. However, these models are more often than not well-designed for individual study scenes only. A change to another area will lead to poor model performance, especially for cross-city or cross-region studies. To this end, researchers have started gradually paying more attention to the task of semantic segmentation across regions or cities.

Recently, there has been increasing research interest in the joint use of multimodal RS data to better mine the representation

Manuscript received 19 January 2024; revised 31 March 2024; accepted 4 April 2024. Date of current version 29 April 2024. This work was supported in part by the National Key Research and Development Program of China under Grant 2022YFB3903401, in part by the National Natural Science Foundation of China under Grant 42241109, Grant 42271350, and Grant 62201553. *(Corresponding authors: Hao Li; Danfeng Hong.)*

Please see the Acknowledgment section of this article for the author affiliations.

The data and code used for the contest are publicly available at: https://github.com/danfenghong/Outcome-of-the-2023-IEEE-WHISPERS-C2Seg-Challenge.

Digital Object Identifier 10.1109/JSTARS.2024.3388464

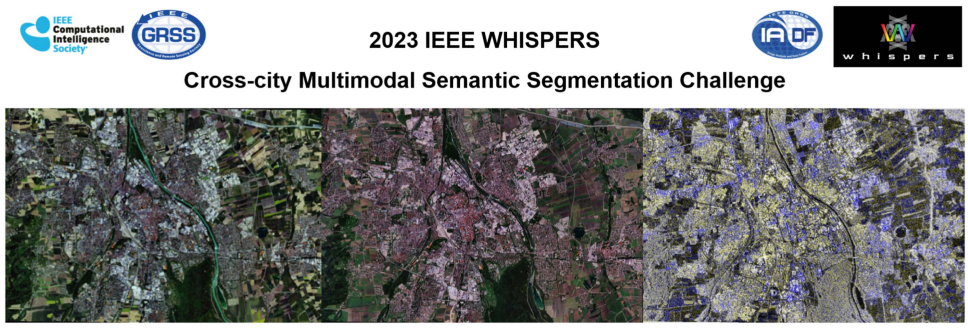

Fig. 1. Contest banner of cross-city multimodal semantic segmentation challenge.

ability of diverse RS modalities for semantic segmentation. By incorporating the complementary information extracted from multimodal RS data, a more robust and reliable model can be built for many RS image analytic tasks (e.g., change detection, LULC classification, etc.). In this context, multimodal RS data fusion is becoming an increasing field to break out of the dilemma induced by unimodal data [15]. For instance, Hong et al. [16] aimed at the semisupervised transfer learning challenge for cross-scene land cover semantic classification in RS via a cross-modal deep network called X-ModelNet. Moreover, Wu et al. [17] proposed a cross-channel reconstruction strategy for more accurate multimodal RS data classification. In [18], Zhao et al. proposed a multiscale progressive network to cascade three subnetworks for gradually segmenting objects into small-scale, large-scale, and another-scale for accurate semantic segmentation of RS data. These aforementioned methods can be unified into a general multimodal deep learning framework for RS-base semantic segmentation on both individual and cross-region environments, e.g., in [19].

To boost technical breakthroughs and accelerate the development of EO applications across cities or regions, it becomes necessary to create multimodal RS benchmark datasets for cross-city land cover segmentation and develop novel AI models with high generalization ability. The cross-city multimodal semantic segmentation (C2Seg) challenge offers a unique and timely opportunity to fill the abovementioned important research gap in the RS community. It is worth noting that the C2Seg dataset is created and available openly and freely in [20], which is used for the contest organization as the Challenge Track 1 in conjunction with the 13th Workshop on Hyperspectral Image and Signal Processing: Evolution in Remote Sensing (WHISPERS), Athens, Greece 2023 as shown in Fig. 1 following a successful edition in last year [21], which is both supported by the IEEE Geoscience and Remote Sensing Society (GRSS) Image Analysis and Data Fusion (IADF) Technical Committee.

The C2Seg Contest received a total of ten highly competitive submissions from international groups, out of which the Top3 teams were announced as the winners of the challenge and presented their solutions in the C2Seg special session during the WHISPERS 2023 conference. In this article, we present a holistic overview of the C2Seg Challenge by elaborating on the methodological design of the Top3 solutions and providing an openly available code base for their implementations

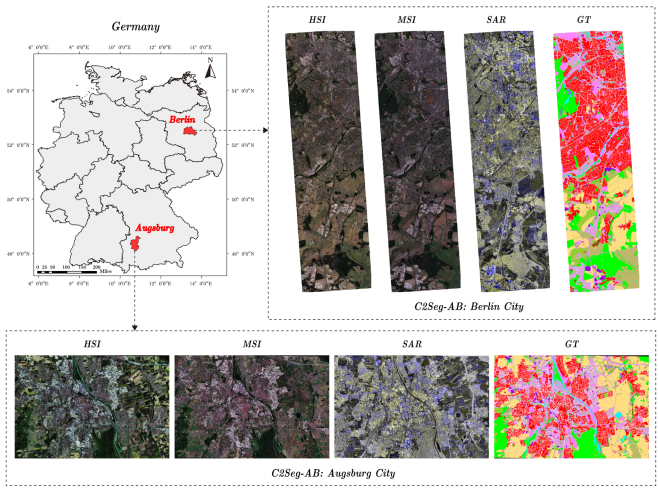

Fig. 2. Visualizing C2Seg-AB datasets for semantic segmentation study scene across Berlin and Augsburg cities in Germany using multimodal RS data. Figure from [20].

together with the C2Seg dataset [20] itself. We hope the lessons learned and the resources provided will lead to a huge impact in promoting the general topic of cross-city multimodal semantic segmentation in the general RS community.

II. Dataset and Evaluation

In the context of the C2Seg Challenge, we use a new collection of multimodal RS benchmark datasets [20], including hyperspectral, multispectral, and SAR data, for research into cross-city semantic segmentation (i.e., C2Seg). The C2Seg datasets consist of two cross-city scenes as follows.

1) *C2Seg-AB (Fig. 2):* Berlin-Augsburg cities in Germany, which are collected from EnMAP, Sentinel-2, and Sentinel-1 satellite missions on dates as close as possible, and accordingly preprocessed via ESA's SNAP toolbox.
2) *C2Seg-BW (Fig. 3):* Beijing–Wuhan cities in China, which are collected from Gaofen-5, Gaofen-6, and Gaofen-3 satellite missions on dates as close as possible, and preprocessed using the ENVI software.

To generate the reference data for semantic segmentation, we retrieved land use and land cover (LULC)-labeled data from OpenStreetMap (OSM) LULC platform at https://osmlanduse.

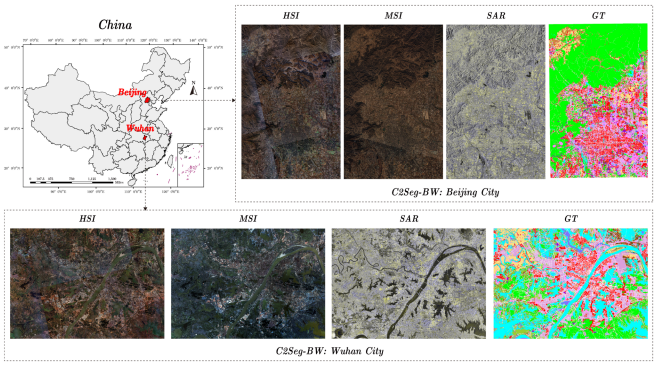

Fig. 3. Visualizing C2Seg-BW datasets for semantic segmentation study scene across Beijing and Wuhan cities in China. Figure from [20].

org/ and considered 12 main classes that are well-defined in OSMLULC. We manually checked annotations and completed the semantic segmentation masks. We 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 [8], [22]. This leads to 13 distinct semantic segmentation categories: 0) Background, 1) Surface water; 2) Street; 3) Urban Fabric; 4) Industrial, commercial, and transport; 5) Mine, dump, and construction sites; 6) Artificial, vegetated areas; 7) Arable land; 8) Permanent crops; 9) Pastures; 10) Forests; 11) Shrub; 12) Open spaces with no vegetation; and 13) Inland wetlands.

For the final evaluation, all submissions were evaluated with the ground truth data by taking the average of both C2Seg datasets (C2Seg-AB and C2Seg-BW). In total, we considered four semantic segmentation metrics, namely, the overall accuracy (OA), the kappa coefficient (Kappa), the F1 score (F1), and the mean intersection over union (mIoU). As C2Seg is a semantic segmentation task, we decided to rank the final results based on the mIoU scores.

III. Top1 Solution: MMGLOTS - Multimodal Global-Local Transformer Segmentor for Remote Sensing Image Segmentation

This section introduces the multimodal global-local transformer segmentor (MMGLOTS), which is the first-place solution of the 2023 C2Seg Challenge. MMGLOTS is composed of three parts: 1) a multimodal semantic feature extractor, 2) a global–local transformer, and 3) a prediction restorer (PR). The overall workflow of the proposed MMGLOTS is shown in Fig. 4.

A. Multimodal Semantic Feature Extractor

The design criteria of the multimodal semantic feature extractor (MMSFE) are based on the observation that multimodal data have different characteristics. For example, the hyperspectral image (HSI) has rich spectral information, while synthetic aperture radar (SAR) can penetrate clouds and dry surface media to some extent. Due to the fact that the multispectral image (MSI) has the highest spatial resolution and provides critical spatial and textural information needed for semantic segmentation, we design an asymmetric feature modeling module, employing a main transformer encoder to extract the spatial and semantic information of MSIs and two weight-shared convolutional neural networks (CNN) encoders to encode the features of HSI and SAR, respectively. Later, a global–local transformer was used to fuse the multimodal RS data for accurate semantic segmentation.

The main transformer encoder adopts the masked image modeling (MIM) pretrained model [1], which is suitable for feature modeling in semantic segmentation tasks. Given an MSI $\mathbf{X_m} \in \mathbb{R}^{H\times W\times C}$, the main transformer encoder first projects the input feature into a latent space $\mathbb{R}^{H\times W\times D}$, where D is the dimension of the latent space. Then, the main transformer encoder adopts a multihead self-attention module to model the spatial and semantic information of the MSI. The output of the main transformer encoder is denoted as $\mathbf{X'_m} \in \mathbb{R}^{H\times W\times D}$.

The HSI encoder and SAR encoder are both composed of a series of convolutional layers. For the sake of simplicity, we denote the output of the HSI encoder and SAR encoder as $\mathbf{X'_h} \in \mathbb{R}^{H\times W\times D}$ and $\mathbf{X'_s} \in \mathbb{R}^{H\times W\times D}$, respectively. These features from different modalities are then fused to form the multimodal semantic feature $\mathbf{X_{mm}} \in \mathbb{R}^{H\times W\times D}$, which is defined as follows:

$$\mathbf{X_{mm}} = \mathbf{X'_m} + \alpha\mathbf{X'_h} + \beta\mathbf{X'_s} \tag{1}$$

where α and β denote the adaptive fusion factors that are learned during the training process, adjusting the contribution of each modality to the final multimodal semantic feature. This multimodal semantic feature $\mathbf{X_{mm}}$ is then fed into the global–local transformer.

B. Global–Local Transformer

The global–local transformer is designed to model the global and local information of the multimodal semantic feature, following the design of the global–local transformer in [23]. The global–local transformer comprises three basic components: 1) Global attention (GA) module, 2) local attention (LA) module, and 2) interaction module. The multimodal features are mainly processed by the LA module, which models the representations in a local context for complexity reduction. The GA module is used to integrate the global dependencies to enhance the local representations. The interaction module is designed to fuse the global and local information and promote the interaction between individual local features, inspired by the shifted window mechanism in [24].

C. Prediction Restorer

It is essential and effective to restore the resolution of encoded features progressively instead of directly upsampling the low-resolution features to the original resolution. MMGLOTS embeds the upsampling process into the global–local transformer, using simple bilinear interpolation to restore the resolution of the multimodal features. The progressive upsampling process restores the resolution of the multimodal features by a factor of two at each stage, which is the most common setting for visual tasks. The restored features are further resized to the original resolution by the PR. For the sake of simplicity, we employ a single convolutional layer and a softmax layer to obtain the final prediction result.

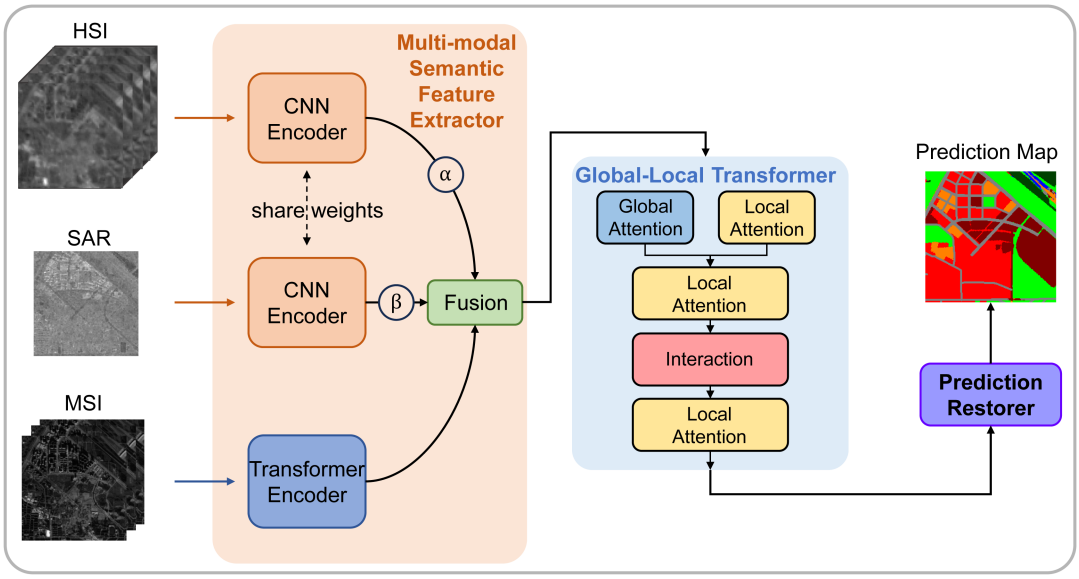

Fig. 4. Overall framework of the Top1 solution - MMGLOTS. It contains three main components: The multimodal semantic feature extractor, the global-local transformer, and the PR.

TABLE I
RESULTS ON THE TEST SETS ON THE TWO C2SEG DATASETS

Datasets	Methods	OA	mIoU	mF1
C2Seg-AB	U-Net	0.3332	0.0965	0.1541
	SegNet	0.3768	0.0737	0.1179
	Deeplabv3+	0.2306	0.0746	0.1194
C2Seg-BW	U-Net	0.3353	0.1171	0.1755
	SegNet	0.2107	0.0378	0.0649
	Deeplabv3+	0.3784	0.1137	0.1688
Average	U-Net	0.33425	0.1068	0.1648
	SegNet	0.29375	0.05575	0.0914
	Deeplabv3+	0.3045	0.09415	0.1441
	Third Place	0.4347	0.1387	0.2095
	Second Place	**0.5068**	0.1851	0.2631
	MMGLOTS	0.5043	**0.1988**	**0.2835**

The bold values refer to own methods of the authors, which are bold for better comparison.

D. Experiments and Discussion

The MMGLOTS is evaluated on two multimodal remote sensing datasets of the C2Seg Challenge, i.e., the Berlin–Augsburg (C2Seg-AB) and the Beijing–Wuhan (C2Seg-BW) datasets. For further comparisons, we choose some single modal methods as baseline methods, including U-Net [25], SegNet [26], and Deeplabv3+ [27]. Furthermore, the results from the second-place team and the third-place team are also included in the comparison, as shown in Table I.

It can be seen that the MMGLOTS achieves the best performance in terms of mIoU and mF1 by averaging the results on the two datasets, with a slight decrease in OA compared to the second-place team.

Overall, the motivation of the MMGLOTS is straightforward, i.e., to fully exploit the characteristics of multimodal data by an asymmetric feature modeling module, which mainly concentrates on the high spatial resolution modality and regards the other modalities as auxiliary information.

IV. TOP2: MULTIMODAL UNSUPERVISED DOMAIN ADAPTATION FOR REMOTE SENSING IMAGE SEGMENTATION.

This section introduces the multimodal unsupervised domain adaption method of the second-place team in four parts, including constructing the multimodal generator network, the discriminator and adversarial strategy, the loss function, and the postprocessing strategy.

A. Multimodal Generator Network

To enhance the integration of multimodal data [15], we devise the multimodal generator consisting of four specific branches that extract domain-specific information from the HSI, MSI, and SAR, along with a shared branch dedicated to capturing the shared information present in the concatenated multimodal data, as depicted in Fig. 5(b). By default, we employ the Seg-HRNet with HRNet48 [28] for the four branches due to Seg-HRNet's superior ability to maintain resolution and capture semantic information.

The features extracted from the four branches are combined using a feature pyramid network (FPN) decoder to yield the semantic segmentation results. We utilize bilinear upsampling to rescale low-resolution features to align with high-resolution features while preserving the same number of channels. These four features are concatenated and processed through a 1×1 convolution module to generate the predicted segmentation results.

B. Discriminator Network and Adversarial Strategy

According to [29], when the model is pretrained on the source domain and predicts an image from the same domain, the segmentation output will exhibit a high level of confidence, leading to a low entropy in the segmentation result. Conversely, when predicting an image from the target domain, the dissimilarity in

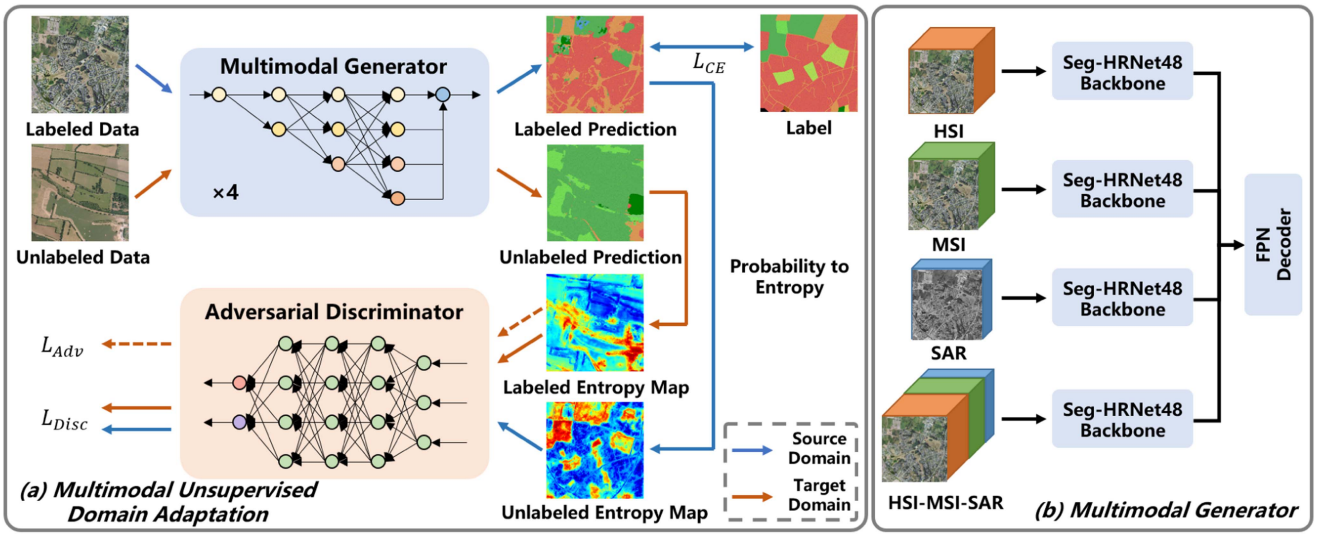

Fig. 5. Overall workflow of the Top2 solution, namely, the multimodal unsupervised domain adaptation network.

distribution between the target and source domains will induce low confidence in the segmentation result, resulting in higher entropy of the segmentation result. Based on this theory, we can use the entropy map of the generator's prediction as input to the discriminator. By comparing the entropy values, the discriminator can classify the prediction into either the source or target domain. In this article, a five-layer convolutional perceptron is constructed for binary classification of the input entropy map, where all convolutional layers use a kernel size of 4, a stride of 2, and a padding of 1. This enables the model to predict whether the original image belongs to the source domain (0) or the target domain (1).

C. Loss Function

The loss function of the multimodal unsupervised domain adaptation method consists of three components: 1) supervised segmentation loss for the multimodal generator using the source domain data, 2) adversarial loss for the multimodal generator using the target domain data, and 3) discriminant loss for the discriminator using the source domain and target domain data.

For the supervised segmentation loss, we employ the conventional cross-entropy loss function [30], [31] to minimize the disparity between the segmentation results and their corresponding reference labels, i.e.,

$$L_{\mathrm{CE}} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{c=1}^{C}\hat{y}_{\mathrm{ic}}\log\left(p_{\mathrm{ic}}\right) \tag{2}$$

where N is the number of labeled samples, C is the number of categories, $\hat{y}_{\mathrm{ic}}$ represents the one-hot encoded label for the ith sample belonging to the cth class, and p_{ic} denotes the predicted probability that the ith sample belongs to the cth class.

The binary cross-entropy function is adopted to compute the generator's adversarial loss, i.e.,

$$L_{\mathrm{Adv}} = \mathrm{BCE}\left(\hat{y}_t, 0\right) \tag{3}$$

where $\hat{y}_t$ denotes the prediction result of the target domain image. During this process, we keep the discriminator parameters fixed and focus on misleading the discriminator by leading it to believe that the prediction result of the target domain image belongs to the source domain.

For the discriminant loss of the discriminator, we also utilize the binary cross-entropy loss. After calculating the adversarial loss, we keep the generator parameters fixed and enhance the discriminator's discriminative ability using the loss function

$$L_{\mathrm{Disc}} = \mathrm{BCE}\left(\hat{y}_s, 0\right) + \mathrm{BCE}\left(\hat{y}_t, 1\right) \tag{4}$$

where $\hat{y}_s$ represents the prediction result of the source domain image.

D. Post-Processing

Inspired by our previous work [32], we further enhance the performance by training three advanced models, including the HRNet 32, ResNeXt 101 [33], and EfficientNet b7 [34]. These models are selected based on their state-of-the-art performance on other benchmark datasets, with each offering distinct network structures that learn variable feature patterns. The predicted results from these models and HRNet 48 are fused.

Furthermore, we observe that the MSI data effectively captures land details with its relatively high resolution. On the other hand, the HSI data provides stable large-scale land-cover results due to its ample spectral information. Models trained on SAR data demonstrate remarkable performance in detecting water bodies. To fully leverage the strengths of these different modalities, we also train separate models using MSI, HSI, and SAR data and then combine their results to obtain the final land-cover maps, specifically focusing on the classes of "Street," "Water," and "Arable land."

E. Results and Discussion

Fig. 6 presents the segmentation results from different methods on the C2Seg-AB and C2Seg-BW datasets. The top row

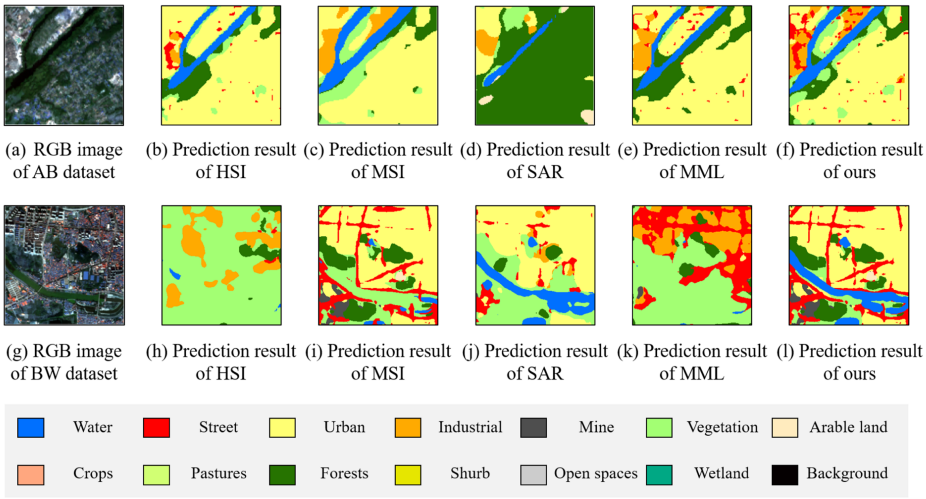

Fig. 6. False color images and predicted segmentation results on the C2Seg-AB and C2Seg-BW datasets of the Top2 solution.

shows the segmentation results on the C2Seg-AB dataset, while the bottom row shows the results on the C2Seg-BW dataset. In Fig. 6(a), a false color image of one test tile ("5_prediction.tif") on the AB dataset is displayed. Fig. 6(b)–(d) shows the prediction results obtained using a single modality (HSI, MSI, or SAR data) on the C2Seg-AB dataset. Fig. 6(e) shows the results obtained using multimodal input, and Fig. 6(f) illustrates the results obtained by our proposed method, which integrates three modalities within the UDA framework. Similarly, Fig. 6(g)–(l) present the corresponding images for a test tile ("4_prediction.tif") on the C2Seg-BW dataset.

As seen in Table I, the proposed method achieves the OA of 0.5068, mIoU of 0.1851, and mF1 of 0.2631 during the test phase, which ranks second among all methods. These results demonstrate the effectiveness and robustness of our method. Moreover, as seen from Fig. 6, our proposed method generates segmentation maps that exhibit more consistent agreement with the RGB image, both for large-scale objects such as "Water," and small-scale objects such as "Street." Furthermore, different modalities display significant variations in their ability to distinguish land covers [35]. For example, MSI data are more effective for road extraction due to its high resolution [36]. In addition, the prediction result using solely HSI data on the C2Seg-BW dataset shows poor performance, primarily because it lacks convergence with numerous bands for training. Similarly, the prediction result using multimodal data also yields unsatisfactory performance. Our method can leverage the strengths and overcome the weaknesses of multimodal data to achieve better segmentation performance.

V. Top3: Multimodal Remote Sensing Network

Inspired by Siamese networks, a multimodal remote sensing network (MRSN) is proposed. This section presents the MRSN consisting of three main parts. We first introduce the overall architecture and then provide detailed illustrations of each component of the model, respectively.

A. Architecture

The overall structure of MRSN is shown in Fig. 7, which consists of three components: 1) data preprocessing, 2) backbone, and 3) decoder module. MRSN is designed with three branches, where each branch corresponds to a single modality.

1) Data Processing: To maximize the utilization of the pretrained parameters of the backbone models, MRSN extracts two types of three-band images from MSI. Red (R), green (G), and blue (B) channels form RGB images that reflect the original visual colors of objects. In addition, near-infrared (Nir), green (G), and blue (B) channels form the GBNir images, which are beneficial for the identification of vegetation, water, and some other objects. To preserve the rich features, SAR images, and HSI images are directly input into the network. Moreover, all bands are normalized using statistical mean and standard deviation to ensure consistent distributions.

2) Backbone: In this challenge, a convolutional neural network (ConvNet) is taken as the backbone. Considering the diverse information in multimodal data, we design three individual backbones that do not share parameters and are trained separately.

Due to the lower spatial resolution and smaller size of images compared to typical computer vision tasks, the fourth stage features of backbones are abandoned. Consequently, only three features of different sizes are reserved for the decoder.

Considering limited computing resources, we ultimately choose ConvNet-small as the most suitable option for this challenge.

3) Decoder: The decoder module (i.e., Uper Net I&II and Fusion I&II in Fig. 7) operates on features from the backbones and the segmentation heads. Considering the rich information contained in HSI, it is designed with a single head and no interaction with other data. The other branches, the RGB branch, GBNir branch, and SAR branch, are decoded with another head. The corresponding features from the backbones are concatenated to produce the logit of this branch.

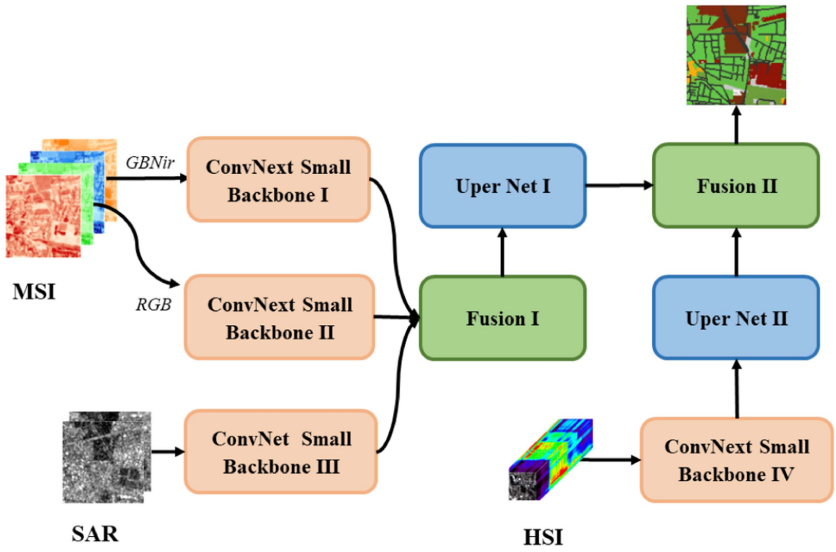

Fig. 7. Overall network architecture of the MRSN proposed for the Top3 solution, which consists of three components: data preprocessing, backbone, and decoder module (i.e., Uper Net I&II and Fusion I&II). MRSN is designed with three branches and each branch corresponds to a single modality.

Therefore, two logits are predicted from HSI and other data. These results are further elementwise added to produce the final interpretation result.

B. Experiment

This section begins by introducing the experiment settings. Next, it presents the results of the experiments. Finally, it illustrates the comparison experiments.

1) Experiment Setting: Since the size of the C2Seg-AB and C2Seg-BW datasets is different, two different experimental settings are adopted for each of these datasets.

C2Seg-AB: We take pretrained ConvNet parameters on ImageNet. To create the validation set, we take 10% of the samples, resulting in 245 training samples. The batch size is set as 24, and training iterations are 8200. The learning rate is initiated as 0.0002 by reducing the learning rate by a factor of 0.5 every 1000 steps.

C2Seg-BW: We take pretrained ConvNet parameters on ImageNet. 6426 samples are selected as the training part and the remaining samples are used for validation. Due to the larger image size and dataset size of C2Seg-BW, batch size is reduced to 6, and training iterations are increased to 40 000 during the training of C2Seg-BW. The decay cycle of the learning rate is also increased to 5000.

We use the AdamW optimizer and a smoothed cross-entropy loss combined with a dice loss for these two datasets. It is worth noting that apart from normalization, no other data enhancement techniques are used.

2) Experiment Result: The validation results of C2Seg-AB and C2Seg-BW are shown in Table II. The term "C2Seg-AB(offline)" refers to the evaluation results from our local training on the C2Seg-AB dataset, while "C2Seg-BW(offline)" denotes the results from our local training on the C2Seg-BW dataset. "C2Seg(online)" represents the final results of our submitted files.

The experimental results reveal that our model exhibits high accuracy across various indicators on the local validation set.

TABLE II
VALIDATION RESULTS OF MRSN

Dataset	OA	mIoU	mF1
C2Seg-AB(offline)	0.9347	0.8657	0.9268
C2Seg-BW(offline)	0.9566	0.8547	0.9200
C2Seg(online)	0.4347	0.1387	0.2095

The bold values refer to own methods of the authors, which are bold for better comparison.

TABLE III
VALIDATION RESULTS OF MRSN

Model	mIoU
UperNet	0.7609
MRSN-2B	0.7920
MRSN-3B	0.8090
MRSN	**0.8198**

The bold values refer to own methods of the authors, which are bold for better comparison.

However, a strong change is observed when the model makes predictions on submitted data. This disparity could potentially be attributed to overfitting, a consequence of an excessive number of training iterations.

3) Comparison Experiments: Due to computational limitations, we first test the model's performance on the C2Seg-AB dataset, using a Batch Size of 12 and 2000 iterations. The comparative results of our models are summarized in Table III, where the mIoU metric is recorded for comparison. Notably, the accuracy of MRSN in Table III is slightly lower due to insufficient training, compared to the final results on C2Seg-AB in Table II.

In our comparative experiments, UperNet was used as the baseline for training the dataset. Initially, segmentation was performed using only RGB images. Later, the NIR band and SAR image were incorporated into the model, resulting in a new branch and the creation of the MRSN-2B model. This addition led to an increase in the mIoU metric from 0.7609 to 0.7920. With the subsequent inclusion of HSI, the mIoU

further increased to 0.8090. Following the training of MRSN, a four-branch architecture, the mIoU ultimately peaked at 0.8198. These results underscore the superior performance of our model.

VI. Discussion of the Challenge: The Winners

The Top3 winning teams of the C2Seg challenge have presented their solutions in the C2Seg special session during the WHISPERS 2023 conference. They employed different strategies to tackle the challenge, however, they all focused on the feature mining of different modalities and the fusion of multimodal features.

1) The winning team (see Section III) proposed an MMGLOTS to extract the multimodal semantic features and fuse them with the global–local transformer. The MMGLOTS used the cutting edge MIM pretrained model [37] as the main transformer encoder to extract the spatial and semantic information of the MSI, and two weight-shared CNN encoders to encode the features of HSI and SAR, respectively. It produced the insights that the multimodal data have different characteristics, and it is essential and effective to fully utilize the high spatial resolution modality and regard the other modalities as auxiliary information. Furthermore, the features from different modalities should be fused adaptively so that the model can fully exploit the characteristics of multimodal data. The MMGLOTS achieved the best performance in terms of mIoU and mF1 by averaging the results on the two datasets, with a slight decrease in OA compared to the second-place team.
2) The runner-up team (see Section IV) proposed a multimodal unsupervised domain adaptation method to leverage the strengths and overcome the weaknesses of multimodal data. It has two main components: 1) the multimodal generator network and 2) the discriminator and adversarial strategy. The key idea is to use the entropy map of the generator's prediction as input to the discriminator, and the discriminator can classify the prediction into either the source or target domain by comparing the entropy values. The method achieved the best performance in terms of OA by averaging the results on the two datasets, with a slight decrease in mIoU and mF1 compared to the first-place team.
3) The third-place team (see Section V) proposed an MRSN to fully exploit the characteristics of multimodal data. It used three individual backbones that do not share parameters and are trained separately to extract the features of different modalities. The main focus of the MRSN is to fully extract the features of different modalities and reasonably fuse them. Instead of fusing all features of different modalities in the early stage, the MRSN first fused the features of MSI and SAR and then fused the features of HSI after decoding and upsampling.

VII. Conclusion

In this article, the outcome of the C2Seg challenge is presented, which is organized in conjunction with the 13th WHISPERS and with the support of IEEE GRSS IADF Technical Committee. The methods of the Top3 winning teams in the C2Seg challenge are introduced in detail, and the results of the challenge are presented. Considering the significance of the multimodal remote sensing data in the context of the big EO data era, the C2Seg challenge produces a valuable multimodal RS benchmark dataset for cross-city land cover segmentation. It offers the unique opportunity to help the RS community develop novel models and methods (e.g., RS Foundation models and self-supervised learning) for comprehensive generalization in wide-area RS data analysis.

Acknowledgment

Authors' Affiliations

Yuheng Liu, Ye Wang, Yifan Zhang, and Shaohui Mei are with the School of Electronics and Information, Northwestern Polytechnical University, Xi'an 710129, China (e-mail: hnlyh@mail.nwpu.edu.cn; wy2017263322@mail.nwpu.edu.cn; yifanzhang@nwpu.edu.cn; meish@nwpu.edu.cn).

Jiaqi Zou, Zhuohong Li, Fangxiao Lu, and Wei He are with the State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University, Wuhan 430079, China (e-mail: immortal@whu.edu.cn; ashelee@whu.edu.cn; fangxiaolu@whu.edu.cn; weihe1990@whu.edu.cn).

Hongyan Zhang is with the School of Computer Science, China University of Geosciences, Wuhan 430078, China (e-mail: zhanghongyan@cug.edu.cn).

Huilin Zhao is with the Department of Land Surveying and Geo-Informatics, The Hong Kong Polytechnic University, Hong Kong (e-mail: huilin.zhao@connect.polyu.hk).

Chuan Chen is with the Chair of Cartography and Visual Analytics, TUM School of Engineering and Design, Technical University of Munich, 80333 München, Germany (e-mail: chuan.chen@tum.de).

Cong Xia is with the School of Resource and Environment Engineering, Wuhan 430070, China (e-mail: 265107@whut.edu.cn).

Hao Li is with the Department of Aerospace and Geodesy, Professorship of Big Geospatial Data Management, Technical University of Munich, 80333 München, Germany (e-mail: hao_bgd.li@tum.de).

Gemine Vivone is with the Institute of Methodologies for Environmental Analysis, National Research Council-IMAA, 85050 Tito, Italy (e-mail: gemine.vivone@imaa.cnr.it).

Ronny Hänsch is with the Department SAR Technology German Aerospace Center, Microwaves and Radar Institute, 82234 Wessling, Germany (e-mail: rww.haensch@gmail.com).

Gulsen Taskin is with the Institute of Disaster Management, Istanbul Technical University, 34469 Istanbul, Türkiye (e-mail: gulsen.taskin@itu.edu.tr).

Jing Yao is with the Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing 100094, China (e-mail: yaojing@aircas.ac.cn).

A. K. Qin is with the Department of Computing Technologies, Swinburne University of Technology, Hawthorn, VIC 3122, Australia (e-mail: kqin@swin.edu.au).

Bing Zhang is with the Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing 100094, China, and also with the College of Resources and Environment, University of Chinese Academy of Sciences, Beijing 100049, China (e-mail: zb@radi.ac.cn).

Jocelyn Chanussot is with the Inria, CNRS, Grenoble INP, LJK, Univ. Grenoble Alpes, 38000 Grenoble, France, and also with the Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing 100094, China (e-mail: jocelyn.chanussot@inria.fr).

Danfeng Hong is with the Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing 100094, China, and also with the School of Electronic, Electrical and Communication Engineering, University of Chinese Academy of Sciences, Beijing 100049, China (e-mail: hongdf@aircas.ac.cn).

References

[1] D. Hong et al., "SpectralGPT: Spectral remote sensing foundation model," *IEEE Trans. Pattern Anal. Mach. Intell.*, to be published, doi: 10.1109/TPAMI.2024.3362475.

[2] M. Wieland, Y. Li, and S. Martinis, "Multi-sensor cloud and cloud shadow segmentation with a convolutional neural network," *Remote Sens. Environ.*, vol. 230, 2019, Art. no. 111203.
[3] X. He, Y. Chen, L. Huang, D. Hong, and Q. Du, "Foundation model-based multimodal remote sensing data classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, pp. 1–17, 2023.
[4] Y. Yi, Z. Zhang, W. Zhang, C. Zhang, W. Li, and T. Zhao, "Semantic segmentation of urban buildings from VHR remote sensing imagery using a deep convolutional neural network," *Remote Sens.*, vol. 11, no. 15, 2019, Art. no. 1774.
[5] F. I. Diakogiannis, F. Waldner, P. Caccetta, and C. Wu, "ResUNet-a: A deep learning framework for semantic segmentation of remotely sensed data," *ISPRS J. Photogrammetry Remote Sens.*, vol. 162, pp. 94–114, 2020.
[6] D. Hong, J. Yao, C. Li, D. Meng, N. Yokoya, and J. Chanussot, "Decoupled-and-coupled networks: Self-supervised hyperspectral image super-resolution with subpixel fusion," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, pp. 1–12, 2023.
[7] H. Li, B. Herfort, W. Huang, M. Zia, and A. Zipf, "Exploration of openstreetmap missing built-up areas using Twitter hierarchical clustering and deep learning in Mozambique," *ISPRS J. Photogrammetry Remote Sens.*, vol. 166, pp. 41–51, 2020.
[8] H. Li, J. Zech, D. Hong, P. Ghamisi, M. Schultz, and A. Zipf, "Leveraging openstreetmap and multimodal remote sensing data with joint deep learning for wastewater treatment plants detection," *Int. J. Appl. Earth Observation Geoinformation*, vol. 110, 2022, Art. no. 102804.
[9] M. Zhou et al., "A general spatial-frequency learning framework for multimodal image fusion," *IEEE Trans. Pattern Anal. Mach. Intell.*, to be published, doi: 10.1109/TPAMI.2024.3368112.
[10] H. Li, J. Wang, J. M. Zollner, G. Mai, N. Lao, and M. Werner., "Rethink geographical generalizability with unsupervised self-attention model ensemble: A case study of openstreetmap missing building detection in africa," in *Proc. 31st Int. Conf. Adv. Geographic Inf. Syst.*, pp. 1–9.
[11] X. He, Y. Zhou, J. Zhao, D. Zhang, R. Yao, and Y. Xue, "Swin transformer embedding UNet for remote sensing image semantic segmentation," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–15, 2022.
[12] Y. Li, D. Hong, C. Li, J. Yao, and J. Chanussot, "HD-Net: High-resolution decoupled network for building footprint extraction via deeply supervised body and boundary decomposition," *ISPRS J. Photogrammetry Remote Sens.*, vol. 209, pp. 51–65, 2024.
[13] X. Yang et al., "An attention-fused network for semantic segmentation of very-high-resolution remote sensing imagery," *ISPRS J. Photogrammetry Remote Sens.*, vol. 177, pp. 238–262, 2021.
[14] R. Hang, S. Xu, P. Yuan, and Q. Liu, "AANet: An ambiguity-aware network for remote-sensing image change detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, pp. 1–11, 2024.
[15] D. Hong, C. Li, B. Zhang, N. Yokoya, J. A. Benediktsson, and J. Chanussot, "Multimodal artificial intelligence foundation models: Unleashing the power of remote sensing Big Data in earth observation," *Innov. Geosci.*, vol. 2, no. 1, 2024, Art. no. 100055.
[16] D. Hong, N. Yokoya, G.-S. Xia, J. Chanussot, and X. X. Zhu, "X-ModalNet: A semi-supervised deep cross-modal network for classification of remote sensing data," *ISPRS J. Photogrammetry Remote Sens.*, vol. 167, pp. 12–23, 2020.
[17] X. Wu, D. Hong, and J. Chanussot, "Convolutional neural networks for multimodal remote sensing data classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–10, 2021.
[18] R. Hang, P. Yang, F. Zhou, and Q. Liu, "Multiscale progressive segmentation network for high-resolution remote sensing imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–12, 2022.
[19] D. Hong, J. Yao, D. Meng, Z. Xu, and J. Chanussot, "Multimodal GANs: Toward crossmodal hyperspectral–multispectral image segmentation," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 6, pp. 5103–5113, Jun. 2021.
[20] D. Hong et al., "Cross-city matters: A multimodal remote sensing benchmark dataset for cross-city semantic segmentation using high-resolution domain adaptation networks," *Remote Sens. Environ.*, vol. 299, 2023, Art. no. 113856.
[21] G. Vivone, A. Garzelli, Y. Xu, W. Liao, and J. Chanussot, "Panchromatic and hyperspectral image fusion: Outcome of the 2022 whispers hyperspectral pansharpening challenge," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 16, pp. 166–179, 2022.
[22] M. Schultz, J. Voss, M. Auer, S. Carter, and A. Zipf, "Open land cover from openstreetmap and remote sensing," *Int. J. Appl. Earth Observation Geoinformation*, vol. 63, pp. 206–213, 2017.
[23] Y. Liu, Y. Zhang, Y. Wang, and S. Mei, "Rethinking transformers for semantic segmentation of remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5617515.
[24] Z. Liu et al., "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 10012–10022.
[25] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervention*, 2015, pp. 234–241.
[26] V. Badrinarayanan, A. Kendall, and R. Cipolla, "SegNet: A deep convolutional encoder-decoder architecture for image segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 12, pp. 2481–2495, Dec. 2017.
[27] Liang-Chieh Chen, Y. G. Zhu, F. Papandreou Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 801–818.
[28] J. Wang et al., "Deep high-resolution representation learning for visual recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 10, pp. 3349–3364, Oct. 2021.
[29] T.-H. Vu, H. Jain, M. Bucher, M. Cord, and P. Pérez, "ADVENT: Adversarial entropy minimization for domain adaptation in semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 2512–2521.
[30] J. Zou, W. He, and H. Zhang, "LESSFormer: Local-enhanced spectral-spatial transformer for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–16, 2022.
[31] H. Zhang, J. Zou, and L. Zhang, "EMS-GCN: An end-to-end mix-hop superpixel-based graph convolutional network for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5526116.
[32] Z. Li et al., "The outcome of the 2021 IEEE GRSS data fusion contest–track MSD: Multitemporal semantic change detection," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 15, pp. 1643–1655, 2022.
[33] S. Xie, R. Girshick, P. Dollár, Z. Tu, and K. He, "Aggregated residual transformations for deep neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit*, 2017, pp. 5987–5995.
[34] M. Tan and Q. Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," in *Proc. Int. Conf. Mach. Learn.*, 2019, pp. 6105–6114.
[35] Z. Li, H. Zhang, F. Lu, R. Xue, G. Yang, and L. Zhang, "Breaking the resolution barrier: A low-to-high network for large-scale high-resolution land-cover mapping using low-resolution labels," *ISPRS J. Photogrammetry Remote Sens.*, vol. 192, pp. 244–267, 2022.
[36] Z. Li, W. He, M. Cheng, J. Hu, G. Yang, and H. Zhang, "SinoLC-1: The first 1 m resolution national-scale land-cover map of China created with a deep learning framework and open-access data," *Earth Syst. Sci. Data*, vol. 15, no. 11, pp. 4749–4780, 2023.
[37] H. Bao, L. Dong, S. Piao, and F. Wei, "BEit: BERT pre-training of image transformers," in *Proc. Int. Conf. Learn. Representations*, 2022, pp. 1–18.

Yuheng Liu received the B.E. degree in electronic and information engineering and the M.E. degree in information and communication engineering from Northwestern Polytechnical University, Xi'an, China, in 2021 and 2024, respectively.

His research interests include remote sensing, image processing, and deep learning.

Ye Wang (Graduate Student Member, IEEE) received the B.S. degree in automation from the Xi'an University of Science and Technology, Xi'an, China, in 2017, and the M.S. degree in biomedical engineering, in 2020, from Northwestern Polytechnical University, Xi'an, China, where she is currently working toward the Ph.D. degree in information and communication engineering with the School of Electronics and Information.

Her research interests include remote sensing, hyperspectral object tracking, image processing, and deep learning.

Yifan Zhang (Member, IEEE) received the B.S. degree in electronics and information technology and the M.S. and Ph.D. degrees in signal and information processing from Northwestern Polytechnical University, Xi'an, China, in 2001, 2004, and 2007, respectively.

From 2007 to 2010, she worked as a Postdoctoral Researcher with the Vision Laboratory, Department of Physics, University of Antwerp, Antwerp, Belgium. She is currently a Professor with the School of Electronics and Information, Northwestern Polytechnical University. Her research interests include hyperspectral image analysis, image fusion, and image restoration.

Shaohui Mei (Senior Member, IEEE) received the B.S. degree in electronics and information engineering and the Ph.D. degree in signal and information processing from Northwestern Polytechnical University, Xi'an, China, in 2005 and 2011, respectively.

He was a Visiting Student with the University of Sydney, Camperdown, NSW, Australia, from 2007 to 2008. He is currently a Professor with the School of Electronics and Information, Northwestern Polytechnical University. His research interests include hyperspectral remotesensing image processing and applications, intelligent signal and information acquisition and processing, video processing, and pattern recognition.

Dr. Mei was the recipient of the First prize of Natural Science Award of Shaanxi Province in 2022, the Excellent Doctoral Dissertation Award of Shaanxi Province in 2014, the Best Paper Award of the IEEE International Symposium on Intelligent Signal Processing and Communication Systems (ISPACS) in 2017, the Best Reviewer for the IEEE JOURNAL OF SELECTED TOPICS IN APPLIED EARTH OBSERVATIONS AND REMOTE SENSING (JSTARS) in 2019, and the IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING (TGRS) in 2022. He serves as the Associate Editor for IEEE TGRS and IEEE JSTARS, the Guest Editor for *Remote Sensing*, and the Reviewer for more than 30 international famous academic journals.

Jiaqi Zou received the B.S. degree in surveying and mapping engineering from the China University of Mining and Technology, Beijing, China, in 2020. She is currently working toward the Ph.D. degree in photogrammetry and remote sensing with State Key Laboratory of Information Engineering in Surveying, Mapping, and Remote Sensing, Wuhan University, Wuhan, China.

Her research interests include hyperspectral image classification, image processing, and deep learning.

Ms. Zou was the recipient of the second place in the Cross-City Multimodal Semantic Segmentation Challenge of the 2023 IEEE WHISPERS, and the second place in the Semisupervised Learning for Land Cover Classification Challenge of the 2022 IEEE Geoscience and Remote Sensing Society Data Fusion Contest.

Zhuohong Li (Student Member, IEEE) received the B.S. degree in communication engineering, in 2020, from Wuhan University, Wuhan, China, where he is currently working toward the Ph.D. degree in photogrammetry and remote sensing with the State Key Laboratory of Information Engineering in Surveying, Mapping, and Remote Sensing.

His research interests include land-cover mapping and large-scale Earth observation application.

Mr. Li was the recipient of the first-place prize in the Multitemporal Semantic Change Detection Track of the 2021 IEEE Geoscience and Remote Sensing Society (GRSS) Data Fusion Contest, the second place in the Semisupervised Learning for Land Cover Classification Challenge of the 2022 IEEE GRSS Data Fusion Contest, the second place in the Cross-City Multimodal Semantic Segmentation Challenge of the 2023 IEEE WHISPERS, and the highlight paper award in the IEEE/CVF Conference on Computer Vision and Pattern Recognition 2024.

Fangxiao Lu received the B.S. degree, in 2020, from the School of Geomatics and Geodesy, Wuhan University, Wuhan, China, where he is currently working toward the Ph.D. degree in photogrammetry and remote sensing with State Key Laboratory of Information Engineering in Surveying, Mapping, and Remote Sensing.

His research interests intelligent remote sensing interpretation.

Mr. Lu was the recipient of the first-place prize in the Multitemporal Semantic Change Detection Track of the 2021 IEEE Geoscience and Remote Sensing Society (GRSS) Data Fusion Contest, the second place in the Semisupervised Learning for Land Cover Classification Challenge of the 2022 IEEE GRSS Data Fusion Contest, and the second place in the Cross-City Multimodal Semantic Segmentation Challenge of the 2023 IEEE Workshop on Hyperspectral Image and Signal Processing: Evolution in Remote Sensing.

Wei He (Senior Member, IEEE) received the B.S. degree from the School of Mathematics and Statistics and the Ph.D degree in surveying, mapping, and remote sensing from the State Key Laboratory of Information Engineering in Surveying, Mapping, and Remote Sensing (LIESMARS), Wuhan University, Wuhan, China, in 2012 and 2017, respectively.

From 2018 to 2020, he was a Researcher with the Geoinformatics Unit, RIKEN Center for Advanced Intelligence Project, Tokyo, Japan, where he was a Research Scientist, from 2020 to 2021. He is currently a Full Professor with LIESMARS, Wuhan University. His research interests include image quality improvement, remote sensing image processing, and low rank representation and deep learning.

Hongyan Zhang (Senior Member, IEEE) received the B.S. degree in geographic information systems and the Ph.D. degree in photogrammetry and remote sensing from Wuhan University, Wuhan, China, in 2005 and 2010, respectively.

From 2010 to 2022, he was a Researcher with the State Key Laboratory of Information Engineering in Surveying, Mapping, and Remote Sensing, Wuhan University. He is currently the Dean of the School of Computer Science, China University of Geosciences, Wuhan. He is also a Young Chang-Jiang Scholar appointed by the Ministry of Education of China. He has authored/coauthored more than 110 research papers. His research interests include high-dimensional data intelligent processing and agricultural remote sensing.

Dr. Zhang was recipient of the first place in the Data Fusion Contest of 2021 and 2019 organized by the IEEE Image Analysis and Data Fusion Technical Committee. He serves as an Associate Editor for IEEE JOURNAL OF SELECTED TOPICS IN APPLIED EARTH OBSERVATIONS AND REMOTE SENSING, *Photogrammetric Engineering and Remote Sensing* and *Computers and Geosciences*. He is a Reviewer for more than 40 international academic journals, including *Remote Sensing of Environment*, *ISPRS Journal of Photogrammetry and Remote Sensing*, IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING, and IEEE TRANSACTIONS ON IMAGE PROCESSING.

Huilin Zhao received the B.S. and M.S. degrees from the School of Resource and Environmental Engineering, Wuhan University of Technology, Wuhan, China, in 2020 and 2023, respectively. He is currently working toward the Ph.D. degree with the Department of Land Surveying and Geo-informatics, Hong Kong Polytechnic University, Hongkong.

His research interests include the application of deep learning in remote sensing image interpretation, knowledge graph, and semantic segmentation.

Chuan Chen received the B.S. degree from Wuhan University, Wuhan, China, in 2019, and the M.S. degree, in 2022, from the Technical University of Munich, Munich, Germany, where he is currently working toward the Ph.D. degree with the Chair of Chair of Cartography and Visual Analytics.

His research interests include semantic segmentation of remote sensing image, knowledge graph, and responsible GIS.

Cong Xia received the B.S. and M.S. degrees from the School of Resource and Environmental Engineering, Wuhan University of Technology, Wuhan, China, in 2020 and 2023, respectively.

His research interests include remote sensing image semantic segmentation.

Hao Li (Member, IEEE) received the double B.Sc. degrees in cartography and computer science from Wuhan University, Wuhan, China in 2015, the M.Sc. degree in geomatic engineering from the University of Stuttgart, Stuttgart, Germany, in 2018, and the Dr. rer. nat degree in Geoinformatics from the GIScience research group, Heidelberg University, Germany in 2022. He is currently working toward the Habilitation degree in Geoinformatics with the Technical University of Munich, Munich, Germany.

Since 2022, he has been a Postdoc Researcher with the Professorship of Big Geospatial Data Management, Technical University of Munich. His research interests include volunteered geographic information, geo-semantics, remote sensing, and geospatial Big Data management.

Dr. Li was the recipient of the ISPRS Best Poster Award (TC IV) in the XXIV ISPRS Congress in Nice in 2022, and the ACM SIGSPATIAL GIS Cup Runners-Up Award in 2023.

Gemine Vivone (Senior Member, IEEE) received the B.Sc. (summa cum laude), the M.Sc. (summa cum laude), and the Ph.D. (highest rank) degrees in information engineering from the University of Salerno, Salerno, Italy, in 2008, 2011, and 2014, respectively.

In 2014, he joined the North Atlantic Treaty Organization Science & Technology Organization Centre for Maritime Research and Experimentation, La Spezia, Italy, as a Scientist. In 2019, he was an Assistant Professor with the University of Salerno, Fisciano, Italy. In 2019, he was a Visiting Professor with the Grenoble Institute of Technology (INPG), Grenoble, France. He is currently a senior Researcher with the National Research Council, Rome, Italy. His research interests include image fusion, statistical signal processing, deep learning, and classification and tracking of remotely sensed images.

Dr. Vivone was the recipient of the IEEE Geoscience and Remote Sensing Society (GRSS) Early Career Award in 2021, the Symposium Best Paper Award at IEEE International Geoscience and Remote Sensing Symposium (IGARSS) in 2015, and the Best Reviewer Award of the IEEE Transactions on Geoscience and Remote Sensing in 2017. Moreover, he is listed in the World's Top 2. He is an ex-officio member of the IEEE GRSS Administrative Committee, a Co-Chair of the IEEE GRSS Image Analysis and Data Fusion Technical Committee, a member of the IEEE Task Force on "Deep Vision in Space," and he was the Leader of the Image and Signal Processing Working Group of the IEEE Image Analysis and Data Fusion Technical Committee from 2020 to 2021. He is currently the Editor in Chief for IEEE GEOSCIENCE AND REMOTE SENSING eNEWSLETTER, an Area Editor for *Elsevier Information Fusion*, and Associate Editor for IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING, IEEE JOURNAL OF SELECTED TOPICS IN APPLIED EARTH OBSERVATIONS AND REMOTE SENSING, and IEEE GEOSCIENCE AND REMOTE SENSING LETTERS. Moreover, he is an Advisory Board Member for *ISPRS Journal of Photogrammetry and Remote Sensing*, and an Editorial Board Member for *Nature Scientific Reports* and *MDPI Remote Sensing*. He served as a Guest Associate Editor for several Special Issues.

Ronny Hänsch (Senior Member, IEEE) received the diploma in computer science and the Ph.D. degree from the TU Berlin, Berlin, Germany, in 2007 and 2014, respectively.

He is a currently a Scientist with the Microwave and Radar Institute, German Aerospace Center, Cologne, Germany, where he leads the Machine Learning Team in the Signal Processing Group, SAR Technology Department. He continues to lecture with the Computer Vision and Remote Sensing Group, TU Berlin. He has extensive experience in organizing remote sensing community competitions, and serves as the Geoscience and Remote Sensing Society (GRSS) representative within SpaceNet, and was the technical lead of the SpaceNet 8 Challenge. His research interests include computer vision and machine learning with a focus on remote sensing (in particular SAR processing and analysis).

Dr. Hänsch served as an Editor for the Geoscience and Remote Sensing Society (GRSS) eNewsletter from 2021 to 2024, as the Editor in Chief of the *Geoscience and Remote Sensing Letters* since 2024 and as Associate Editor of the *ISPRS Journal of Photogrammetry* and *Remote Sensing* since 2022. He served as a (co-)chair of the IEEE GRSS Image Analysis and Data Fusion technical committee from 2017 to 2023 and has been serving as the co-chair of the ISPRS working group on Image Orientation and Sensor Fusion since 2016. He served as an organizer of the CVPR Workshop EarthVision from 2017 to 2024 and the IGARSS Tutorial on Machine Learning in Remote Sensing from 2017 to 2024.

Gulsen Taskin (Senior Member, IEEE) received the B.S., M.S., and Ph.D. degrees in computational science and engineering from Istanbul Technical University, Istanbul, Turkey, in 2001, 2003, and 2011, respectively.

She was a Visiting Scholar with the School of Electrical and Computer Engineering and School of Civil Engineering, Purdue University, West Lafayette, IN, USA, from 2008 to 2009 and 2016 to 2017. She is currently an Associate Professor with the Institute of Disaster Management, Istanbul Technical University. Her research interests include hyperspectral image analysis, machine learning, explainable AI, and sensitivity analysis.

Dr. Taskin has served as an Associate Editor for IEEE GEOSCIENCE AND REMOTE SENSING LETTERS journal and has reviewed several other journals.

Jing Yao (Member, IEEE) received the Ph.D. degree in mathematics from Xi'an Jiaotong University, Xi'an, China, in 2021.

From 2019 to 2020, he was a Visiting Student with the Signal Processing in Earth Observation, Technical University of Munich, Munich, Germany, and with the Remote Sensing Technology Institute, German Aerospace Center, Oberpfaffenhofen, Germany. Since 2021, he has been an Assistant Professor with the Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing, China. His research interests include hyperspectral and multimodal remote sensing image analysis, mainly including optimization and deep learning-based methods for image processing and interpretation tasks.

Dr. Yao was the recipient of the Jose Bioucas Dias Award for recognizing an outstanding paper at WHISPERS in 2021.

A. K. Qin (Senior Member, IEEE) received the B.Eng. degree in automatic control from Southeast University, Nanjing, China, in 2001, and the Ph.D. degree in computer science and engineering from Nanyang Technology University, Singapore, in 2007.

From 2007 to 2017, he was with the University of Waterloo, Waterloo, ON, Canada; INRIA Grenoble Rhône-Alpes, Montbonnot-Saint-Martin, France; and RMIT University, Melbourne, VIC, Australia. In 2017, he joined the Swinburne University of Technology, Hawthorn, VIC, Australia, where he is currently a Professor. He is also currently the Director of Swinburne Intelligent Data Analytics Lab and the Deputy Director of Swinburne Space Technology and Industry Institute, Hawthorn, VIC, Australia. His research interests include machine learning, evolutionary computation, computer vision, remote sensing, services computing, and edge computing.

Dr. Qin was the recipient of the 2012 IEEE Transactions on Evolutionary Computation Outstanding Paper Award and the 2022 IEEE Transactions on Neural Networks and Learning Systems Outstanding Associate Editor Award. He is currently the Chair of the IEEE Computational Intelligence Society (CIS) Neural Networks Task Force on "Deep Vision in Space," the Vice-Chair of the IEEE CIS Emergent Technologies Task Force on "Multitask Learning and Multitask Optimization," the Vice-Chair of the IEEE CIS Neural Networks Task Force on "Deep Edge Intelligence." He served as the General Co-Chair of the 2022 IEEE International Joint Conference on Neural Networks and as the Chair of the IEEE CIS Neural Networks Technical Committee during the 2021 to 2022 term.

Bing Zhang (Fellow, IEEE) received the B.S. degree in geography from Peking University, Beijing, China, in 1991, and the M.S. and Ph.D. degrees in remote sensing from the Institute of Remote Sensing Applications, Chinese Academy of Sciences, Beijing, China, in 1994 and 2003, respectively.

He is currently a Full Professor and the Deputy Director of the Aerospace Information Research Institute, China Academy of Sciences. He has authored or coauthored more than 400 journal papers. His research focuses on hyperspectral remote sensing technology and applications.

Dr. Zhang is a Fellow of the Geographical Society of China. He is also a Highly Cited Researcher (Clarivate Analytics). He was the recipient more than ten important prizes from international institutions and the Chinese government, including the IEEE Geoscience and Remote Sensing Society Regional Leader Award, the National Science and Technology Advance Award of China, the Outstanding Scientific Achievement Award of Chinese Academy of Sciences, etc., for his creative achievements.

Jocelyn Chanussot (Fellow, IEEE) received the M.Sc. degree in electrical engineering from the Grenoble Institute of Technology (Grenoble INP), Grenoble, France, in 1995, and the Ph.D. degree from the Université de Savoie, Annecy, France, in 1998.

From 1999 to 2023, he was with Grenoble INP, where he was a Professor of Signal and Image Processing. He is currently a Research Director with INRIA, Grenoble, France. He has been a Visiting Scholar with Stanford University, Stanford, CA, USA; KTH (Sweden), Stockholm, Sweden; and National University of Singapore, Singapore. Since 2013, he is an Adjunct Professor of the University of Iceland, Reykjavik, Iceland. From 2015 to 2017, he was a Visiting Professor with the University of California, Los Angeles, CA, USA. He holds the AXA chair in remote sensing and is an Adjunct Professor with the Chinese Academy of Sciences, Aerospace Information research Institute, Beijing, China. His research interests include image analysis, hyperspectral remote sensing, data fusion, machine learning and artificial intelligence.

Dr. Chanussot is a Fellow of ELLIS, a Fellow of AAIA, a Member of the Institut Universitaire de France from 2012 to 2017. He is an Associate Editor for the IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING, IEEE TRANSACTIONS ON IMAGE PROCESSING, and PROCEEDINGS OF THE IEEE. He was the Editor in Chief of the IEEE JOURNAL OF SELECTED TOPICS IN APPLIED EARTH OBSERVATIONS AND REMOTE SENSING from 2011 to 2015. In 2014, he served as a Guest Editor for the *IEEE Signal Processing Magazine*. He is the founding President of IEEE Geoscience and Remote Sensing French chapter from 2007–2010 which received the 2010 IEEE Geoscience and Remote Sensing Society Chapter Excellence Award. He was the Vice-President of the IEEE Geoscience and Remote Sensing Society, in charge of meetings and symposia from 2017 to 2019. He is a Highly Cited Researcher (Clarivate Analytics/Thomson Reuters, since 2018).

Danfeng Hong (Senior Member, IEEE) received the Dr.-Ing degree (summa cum laude) from the Signal Processing in Earth Observation, Technical University of Munich, Munich, Germany, in 2019.

He was a Research Scientist and led a Spectral Vision Working Group with the Remote Sensing Technology Institute, German Aerospace Center, Oberpfaffenhofen, Germany. He was also an Adjunct Scientist with GIPSA-lab, Grenoble INP, CNRS, University Grenoble Alpes, Grenoble, France. Since 2022, he has been a Full Professor with the Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing, China. His research interests include artificial intelligence, multimodal remote sensing, foundation models, hyperspectral imaging, and large-scale Earth observation.

Dr. Hong was the recipient of the Best Reviewer Award of the IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING (TGRS) in 2021 and 2022, the Best Reviewer Award of the IEEE JOURNAL OF SELECTED TOPICS IN APPLIED EARTH OBSERVATIONS AND REMOTE SENSING in 2022, the Jose Bioucas Dias Award for recognizing the outstanding paper at WHISPERS in 2021, the Remote Sensing Young Investigator Award in 2022, the IEEE Geoscience and Remote Sensing Society Early Career Award in 2022, and a Highly Cited Researcher (Clarivate Analytics) in 2022 and 2023. He is an Associate Editor for IEEE TGRS and the Editorial Board Member of *Information Fusion* and ISPRS JOURNAL OF PHOTOGRAMMETRY AND REMOTE SENSING.