This article has been accepted for publication in IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing. This is the author's version which has not been fully e
content may change prior to final publication. Citation information: DOI 10.1109/JSTARS.2023.3335891

Deep Learning-based Semantic Segmentation of Remote Sensing Images: A Survey

Liwei Huang, Bitao Jiang, Shouye Lv, Yanbo Liu and Ying Fu

***Abstract*—Semantic segmentation of remote sensing images (SSRSI), which aims to assign a category to each pixel in remote sensing images, plays a vital role in a broad range of applications, such as environmental monitoring, urban planning, and land resource utilization. Recently, with the successful application of deep learning in remote sensing, a substantial amount of work has been aimed at developing SSRSI methods using deep learning models. In this survey, we provide a comprehensive review of SSRSI. Firstly, we review the current mainstream semantic segmentation models based on deep learning. Next, we analyze the main challenges faced by SSRSI and comprehensively summarize the current research status of deep learning-based SSRSI, especially some new directions in SSRSI are outlined, including semi-supervised and weakly-supervised SSRSI, unsupervised domain adaption (UDA) in SSRSI, multi-modal data fusion-based SSRSI, and pretrained models for SSRSI. Then, we examine the most widely used datasets and metrics and review the quantitative results and experimental performance of some representative methods of SSRSI. At last, we discuss promising future research directions in this area.**

***Index Terms*—Semantic segmentation, Remote sensing images, Deep learning, Semi-supervised, Weakly-supervised, Unsupervised domain adaptation, Multi-modal fusion, Pretrained models.**

I. Introduction

SEMANTIC image segmentation is essential to many visual understanding systems. It involves labeling a category for each pixel in an image. Semantic segmentation of remote sensing images (SSRSI) achieves the pixel-level classification of remote sensing images by applying semantic segmentation technology to remote sensing and has played a vital role in many applications, e.g., environmental monitoring [1], [2], urban planning [3], [4], and land resource utilization [5], [6].

In the past decade, deep learning has made remarkable progress in remote sensing [7], [8], [9] and has demonstrated much superior performance in many areas (including scene classification [10], objection detection [11], change detection [12], and image fusion [13]). With the rapid development of semantic segmentation technology in computer vision [14], [15], generic deep learning-based models, including FCN [16] and DeepLab [17], [18], have been widely used in SSRSI. At the same time, targeted at the unique issues of remote sensing images, such as complex image backgrounds, large image sizes, and multi-modal characteristics, a large number of deep learning-based remote sensing image semantic segmentation technologies have emerged [19], [20], [21], [22].

Reviews of deep learning methods for SSRSI have been conducted in the past years [21], [22]. Kotaridis et al. [21] conducted a meta-analysis to summarize recent image segmentation studies regarding the algorithms, the software, and the data source. However, this study just performed the statistical analysis of the latest progress and did not involve the technical details of deep learning-based SSRSI. Yuan et al. [22] reviewed recent development in deep neural networks and their applications to SSRSI, but they mainly focused on CNN-based models and did not touch on other advanced methods (e.g., Transformer-based models [23], universal segmentation models [24]) and some important topics (e.g., unsupervised domain adaption (UDA) semantic segmentation).

This survey aims to provide a comprehensive overview and an insightful analysis of models and algorithms for deep learning-based SSRSI. Unlike previous reviews, this survey summarizes some emerging techniques, especially those developed recently, e.g., Transformer-based models, and we thoroughly compare different techniques. In addition, we deeply analyze the unique challenges of SSRSI, which are different from natural image segmentation. Finally, this paper reviews some emerging directions, including UDA semantic segmentation, multi-modal data fusion, and pretrained models for SSRSI, which are not involved in previous reviews.

The remainder of this survey is organized as follows: Section II provides an overview of popular deep neural network architectures of image semantic segmentation. Section III provides a comprehensive overview of the most significant state-of-the-art deep learning-based semantic segmentation models of remote sensing images and some emerging themes. Section IV introduces the commonly used loss functions and metrics of SSRSI. Section V reviews some popular datasets. Section VI reports some semantic segmentation models' quantitative results and experimental performance. Section VII discusses the main challenges and future directions of SSRIS. Finally, Section VIII presents our conclusions.

II. Deep learning architectures in semantic segmentation

This section reviews the most popular deep learning-based semantic segmentation methods widely used in natural and remote sensing images. We first introduce a formal definition of deep learning-based semantic segmentation models. Given the training set $X_i \in \mathbb{R}^{H\times W\times C}$, and the label category $Y_i \in$

Liwei Huang, Bitao Jiang, Shouye Lv, and Yanbo Liu are with the Satellite Information Intelligent Processing and Application Research Laboratory, Beijing Institute of Remote Sensing, Beijing 100854, China (e-mail: dr_huanglw@163.com; jiangbitao@bjir.org.cn; lvshouye@126.com; liuyanbonudt@163.com). Ying Fu is with the Beijing Institute of Technology (email: fuying@bit.edu.com).

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License. For more information, see https://creativecommons.org/licenses/by-nc-nd/4.0/

This article has been accepted for publication in IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing. This is the author's version which has not been fully e
content may change prior to final publication. Citation information: DOI 10.1109/JSTARS.2023.3335891

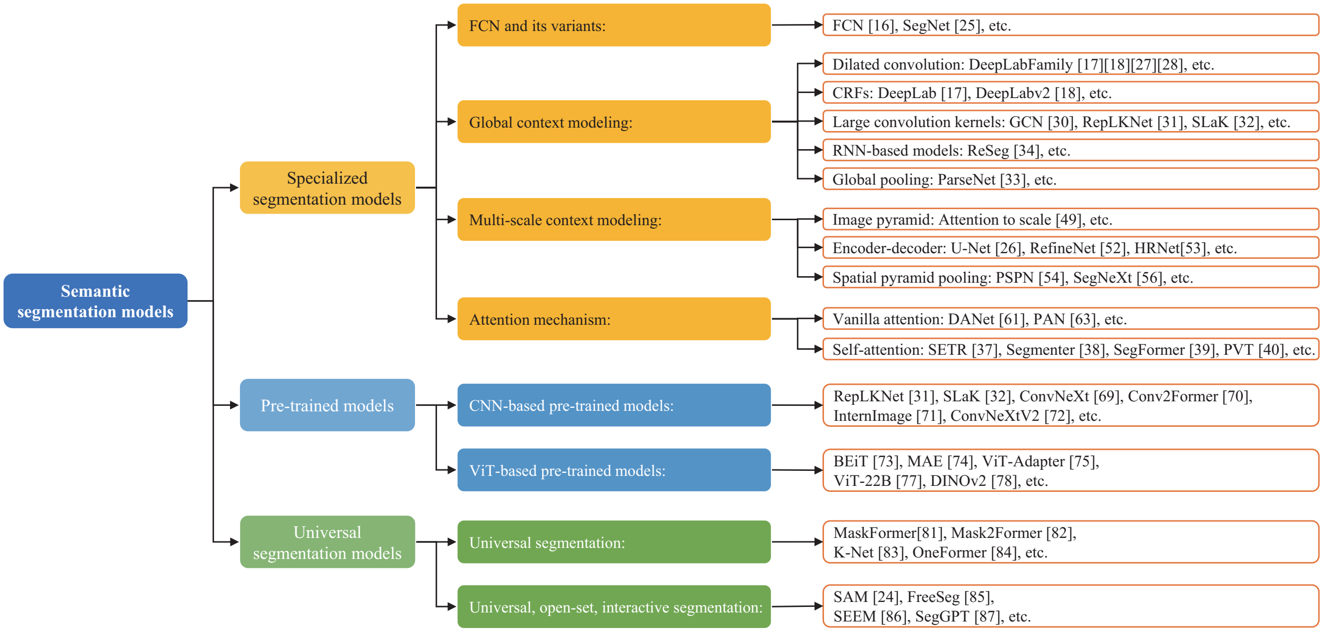
Semantic segmentation models
Specialized segmentation models
FCN and its variants:
FCN [16], SegNet [25], etc.
Global context modeling:
Dilated convolution: DeepLabFamily [17][18][27][28], etc.
CRFs: DeepLab [17], DeepLabv2 [18], etc.
Large convolution kernels: GCN [30], RepLKNet [31], SLaK [32], etc.
RNN-based models: ReSeg [34], etc.
Global pooling: ParseNet [33], etc.
Multi-scale context modeling:
Image pyramid: Attention to scale [49], etc.
Encoder-decoder: U-Net [26], RefineNet [52], HRNet[53], etc.
Spatial pyramid pooling: PSPN [54], SegNeXt [56], etc.
Attention mechanism:
Vanilla attention: DANet [61], PAN [63], etc.
Self-attention: SETR [37], Segmenter [38], SegFormer [39], PVT [40], etc.
Pre-trained models
CNN-based pre-trained models:
RepLKNet [31], SLaK [32], ConvNeXt [69], Conv2Former [70], InternImage [71], ConvNeXtV2 [72], etc.
ViT-based pre-trained models:
BEiT [73], MAE [74], ViT-Adapter [75], ViT-22B [77], DINOv2 [78], etc.
Universal segmentation models
Universal segmentation:
MaskFormer[81], Mask2Former [82], K-Net [83], OneFormer [84], etc.
Universal, open-set, interactive segmentation:
SAM [24], FreeSeg [85], SEEM [86], SegGPT [87], etc.

Fig. 1. The taxonomy of deep learning architectures in semantic segmentation proposed in this paper and some popular deep semantic segmentation models.

$\mathbb{R}^{H\times W\times N}$ corresponding to all pixels in the images, where (H, W) is the size of the image, C is the number of channels of the image, and N is the number of categories of pixels. The semantic segmentation task is to construct a segmentation function f to minimize the loss between the predicted label and the real label of each pixel of the images in the training set.

$$\hat{f} = argmin(\sum_{i} L(f\left(X_i\right), Y_i)), \qquad (1)$$

where $L(\cdot)$ is the loss function. In remote sensing image segmentation tasks, it is usually cross-entropy loss, Dice loss, etc. We can assign a label to each pixel in an unlabeled image through the learned segmentation function $\hat{f}$.

In the early days, models such as FCN [16], SegNet [25], U-Net [26], and DeepLab [17], [18], [27], [28] focused on designing a specialized semantic segmentation model to handle semantic segmentation tasks in a specific scenario. In recent years, some studies have shown that combining a model learned in pretraining tasks with a segmentation network can achieve a performance comparable to or even higher than that of the specialized segmentation network in the downstream semantic segmentation tasks. Recently, with the development of big models, especially the emergence of SAM [24], the universal segmentation models for all segmentation tasks (semantic, instance, panoptic) have made rapid progress, and the research on image segmentation has entered a new stage of development. Fig. 1 shows the taxonomy of deep learning architectures in semantic segmentation proposed in this paper.

A. Specialized architectures for semantic segmentation

Early deep semantic segmentation models were mainly designed for a single task and scenario as a per-pixel classification task. Semantic segmentation often involves some essential techniques.

Global context modeling. Image semantic segmentation needs to model global context and effectively fuse the global context and local context. The main ways to capture global context include dilated convolution, conditional random fields (CRFs), large convolution kernels, global pooling, Recurrent Neural Networks (RNNs), and self-attention mechanisms (mainly Transformer).

Dilated convolution can effectively enlarge the receptive fields without increasing parameters, and has been widely integrated into many semantic segmentation models to capture global context. Representative works include DeepLab family [17], [18], [27], [28], and DenseASPP [29]. But it also brings a grid effect. The second category is utilizing CRFs as a post-processing technique to model the long-range dependence between image pixels, and the most representative work includes DeepLab [17] and DeepLabv2 [18]. These methods can capture local and long-range dependencies within an image to refine the predictions. But they have higher complexity than end-to-end semantic segmentation methods. The third category is to use large convolution kernels to improve the receptive field of the model, such as GCN [30], RepLKNet [31], and SLaK [32]. However, large convolution kernels lead to a large amount of computation. The global pooling captures the global receptive field by performing a pooling operation on each feature map to get a scene-level context vector [33]. However, detailed information related to object boundaries is missing due to the pooling operations. RNN can model the long-term and short-term dependence between pixels and help improve the prediction of the segmentation map [34], [35]. However, the inherent nature of sequential processing makes RNN-based methods difficult to be processed in parallel.

Vision Transformer Models (ViTs) [36] utilize a self-

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License. For more information, see https://creativecommons.org/licenses/by-nc-nd/4.0/

This article has been accepted for publication in IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing. This is the author's version which has not been fully e
content may change prior to final publication. Citation information: DOI 10.1109/JSTARS.2023.3335891

attention mechanism to model the global context by calculating the correlations between two pixels, and has achieved the best performance in semantic segmentation due to its high parallel processing efficiency and strong representation capability. SETR [37] is the first transformer-based model for semantic segmentation. It utilized a pure Transformer as the encoder to model the global dependencies. Segmenter [38] used a mask transformer as the decoder to predict the segmentation map. SegFormer [39] speeded up the model by creating an MLP-based decoder with upsampling operation. However, ViTs output single-scale low-resolution features and have a very high computational cost on large images. Many studies integrated the multi-scale analysis methods into ViTs to generate multi-scale features, including Pyramid Vision Transformer (PVT) [40], Crossformer [41], HRViT [42], U-Netr [43]. Several attention techniques, including shifted window multi-head self-attention [44], spatial-reduction attention [40], spatially separable self-attention [45], cross-shaped window self-attention [46], neighborhood attention [47], dilated neighborhood attention [48] were proposed to overcome the high computational complexity of standard multi-head attention. Overall, Transformers can model global dependencies, but require a large dataset and face very high computational overhead.

Multi-scale context modeling. Semantic segmentation must consider high-level semantic information to improve segmentation accuracy and utilize low-level feature information to retain high-quality spatial details. Therefore, multi-scale context modeling is critical to semantic segmentation. The methods mainly include three categories: image pyramid, encoder-decoder structure, and spatial pyramid pooling.

The first approach is the image pyramid model [49], [50], [51], which resizes an input image into multiple scales and then applies multiple parallel network branches to extract features for each scale input. It also leads to high computational complexity. The second approach is the encoder-decoder structure, which integrates detailed information from multiple scales in the encoder into the decoder through residual connections. The most representative models include U-Net [26], RefineNet [52], HRNet [53], etc. The third approach is spatial pyramid pooling, which typically contains several parallel pooling branches to generate multi-scale features. Representative methods include DeepLabv3 [27], DeepLabv3+ [28], PSPN [54], UPerNet [55], SegNeXt [56]. However, multiple parallel branches increase the computational complexity.

Attention mechanism. To adaptively extract features strongly related to target tasks, various attention mechanism is also widely used in semantic segmentation, including channel attention, spatial attention, channel & spatial attention, and branch attention. At present, the main directions include the combination of the attention mechanism and deep learning models (mainly CNN) and the pure attention-based models (mainly Transformer). We mainly introduce the former here.

Channel attention adaptively recalibrates each channel's weight by generating an attention mask across all channels [57], [58]. Spatial attention uses attention masks across spatial domains to achieve adaptive spatial region selection[59], [60]. Channel & spatial attention predict channel and spatial attention masks separately or generate a joint 3-D channel, height, and width attention mask directly and use them to select important features, such as DANet[61], VAN [62]. Branch attention feeds the features at multiple scales into the attention module to generate multi-branch attention maps to realize a dynamic branch selection, such as Pyramid Attention Network (PAN)[63].

The attention mechanism is still the main component of many state-of-the-art semantic segmentation models due to its ability to adaptively extract important feature information based on task objectives to improve model performance.

B. Pretrained models for semantic segmentation

Pretrained models can alleviate the problem of lacking large-scale annotation data and training resources in downstream tasks. The vanilla pretrained models use the networks pretrained on a large-scale dataset as the starting point for further refinement and perform the best performance. Recently, with the emergence of SAM, many ultra-large-scale pretrained models have been proposed, and they can achieve high image segmentation performance without further finetuning.

CNN-based pretrained models. Semantic segmentation models usually use well-designed backbone networks as encoders. Early research mainly used VGG [64], GoogLeNet [65], ResNet [66], MobileNet [67] and ShuffleNet [68] as backbone networks and used the network pretrained on ImageNet as a starting point. However, due to the small scale of these backbone networks, the effect of pretraining is not obvious. In recent years, with the remarkable progress of large-scale vision transformers, modernized DCNNs migrated the advanced architecture of ViTs to DCNNs. Many improvement methods were proposed to capture long-range dependencies and increase the receptive field utilizing large convolution kernels and deformable convolutions, including ConvNeXt [69], RepLKNet [31], SLaK [32], Conv2Former [70], InternImage [71], ConvNeXtV2 [72] have achieved a comparable or superior performance than ViTs in semantic segmentation. Therefore, ViTs have not entirely taken the role of DCNNs in image semantic segmentation.

ViT-based pretrained models. The pretrained ViTs have been widely used in image semantic segmentation. Most of them are based on the ViT and Swin Transformer [44]. Bao et al. [73] introduced a self-supervised vision representation model named BEiT, which used masked image modeling tasks to pretrain vision Transformers. He et al. [74] developed masked autoencoders (MAE) for self-supervised learning, with an encoder that operated only on the visible subset of patches and a lightweight decoder that reconstructed the original image from the latent representation and mask tokens. Chen et al. [75] proposed the ViT-Adapter, which utilized a pretraining-free adapter to introduce the image-related inductive biases to downstream tasks. Yu et al. [76] hypothesized that the general architecture of the transformer was more critical to the model's performance and proposed a PoolFormer by replacing the attention module in transformers with a simple spatial pooling operator. Dehghani et al. [77] presented a very large vision transformer called ViT-22B, which included 22 billion

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License. For more information, see https://creativecommons.org/licenses/by-nc-nd/4.0/

This article has been accepted for publication in IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing. This is the author's version which has not been fully e
content may change prior to final publication. Citation information: DOI 10.1109/JSTARS.2023.3335891

parameters. Oquab et al. [78] proposed an automatic pipeline to build a dedicated, diverse, and curated image dataset and provided a pretrained visual model, DINOv2, trained with ViT architectures.

However, there are still many underlying challenges for pretrained models. First, the domain discrepancy between the pretraining dataset and the dataset of downstream tasks has been an obstacle to better knowledge transfer. Second, the supervision collapse [79] happens when the pretrained model concentrates on a constrained set of information and ignores components crucial for downstream tasks but have no impact on the pretraining objective. Lastly, the objective divergence between the pretraining tasks and downstream tasks is also a technical challenge for vanilla vision pretraining.

C. Universal architectures for all segmentation tasks

Although universal segmentation is not a new direction in computer vision, it has only begun to develop rapidly in the past two years with the explosion of vision foundation models.

The main objective of universal semantic segmentation achieves multiple segmentation (semantic-, instance-, panoramic-) tasks through a unified model. MaskFormer [80], Mask2Former [81], and K-Net [82] formulated all three tasks to the panoramic segmentation architecture, which can be trained on all three tasks and obtain high performance without changing architecture. However, they must be trained individually on each task to achieve the best performance. OneFormer [83] extended the Mask2Former with a multi-task training setting and can be trained only once and achieve SOTA performance across all three image segmentation tasks.

Recent research [24], [84], [85] has proposed that universal semantic segmentation can achieve open vocabulary semantic segmentation with zero-shot transferability. SegGPT [86] proposed a universal model for semantic segmentation via in-context learning. Qin et al. [84] utilized adaptive prompt learning to facilitate the unified model to capture task-aware and category-sensitive concepts. SAM [24] proposed a promotable segmentation model trained on 11 million images, demonstrating strong zero-shot performance via prompt engineering. SEEM [85] can perform any segmentation tasks in open-set scenarios, and supports visual, textual, and referring region prompts in any arbitrary combination.

With the emergence of SAM, building a universal segmentation model that can handle all segmentation tasks and enable few-shot or zero-shot generalization in an open vocabulary setting has become the crucial research direction.

III. Deep learning-based semantic segmentation in remote sensing images

Thanks to the explosion of deep semantic segmentation, various image segmentation architectures have been widely used in SSRSI and achieved superior performance. However, in contrast to natural scenes, remote sensing images produced from a bird's-eye view have unique issues, such as large size, complex backgrounds, large-scale variation, dense arrangements, and low spatial resolution, these present challenging scientific problems for SSRSI.

High intra-class variance and low inter-class variance. Targets of the same category in remote sensing images, such as roofs, often differ significantly in shape, size, color, and texture, resulting in high intraclass variance. However, targets of different categories, such as road and roof, have more remarkable similarities, resulting in low inter-class variance. This greatly reduces the separability of pixels.

Large-scale variation. Targets in remote sensing images often have large-scale variation, such as airports and aircraft, even for the same category of targets, such as ships, which sizes range from several meters to several hundred meters. Large-scale variation tends to cause small targets to be covered and poses a massive challenge to semantic segmentation.

Class imbalance. The class imbalance problem is severe in remote sensing images. For example, cities' roads cover only a tiny area of urban land. However, other categories, such as impervious surfaces, low vegetation, and trees, account for most of the land. In addition, the background in the remote sensing images occupies most of the image area, leading to foreground-background imbalance. Class imbalance results in severely insufficient classification performance for categories with a small number of pixels.

Large image size. Remote sensing images have a larger size than natural images. For example, the patches in the ISPRS semantic labeling data set (Potsdam) [87] have 6000 × 6000 pixels. It is almost unrealistic to input the whole image into the model for training and testing while cutting or scaling methods lose the objects' global dependency information or detailed information.

Limited labeled data. Due to the large size and complex background of remote sensing images, labeling massive pixel-level samples is labor-intensive and time-consuming. Some complex scenes even required experienced professionals to label samples. Therefore, gathering large-scale annotation data is challenging in many cases, which dramatically affects the performance of the deep semantic segmentation models.

Next, we will focus on reviewing five important research directions in SSRSI, including supervised SSRSI, semi-supervised and weakly-supervised semantic segmentation, UDA semantic segmentation, multi-modal data fusion, and pretrained models for SSRSI. In supervised SSRSI, we discussed how current research addresses the above first four challenges. Other four research directions mainly focus on solving the problem of limited labeled data. The paper lists a simple correspondence (as shown in Fig.2) between five challenging problems and five research directions. But we must clarify that Fig.2 is just a rough division of their correspondence relationships. There are also intersections between them. In particular, the first four challenges are often involved in semi-supervised and weakly-supervised semantic segmentation, UDA semantic segmentation, multi-modal data fusion, and pretrained models for SSRSI.

A. Supervised semantic segmentation of remote sensing images

Under the framework of supervised learning, many semantic segmentation methods have been proposed to effectively solve

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License. For more information, see https://creativecommons.org/licenses/by-nc-nd/4.0/

This article has been accepted for publication in IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing. This is the author's version which has not been fully e
content may change prior to final publication. Citation information: DOI 10.1109/JSTARS.2023.3335891

TABLE I
BRIEF SUMMARY OF SUPERVISED SEMANTIC SEGMENTATION MODELS OF REMOTE SENSING IMAGES AND KEY RELATED WORKS.

Research Issues	Category	Description	Related work
High intra-class variance and low inter-class variance	Effective context modeling and fusion	Multi-scale context modeling with encoder-decoder structure	[88], [89], [90], [91], [92], [93], [94], [95], [96], [97], [98], [99], [100], [101], [102]
		Multi-scale context modeling with spatial pyramid pooling	[101], [103], [104], [105]
		Global context modeling	[106], [107], [88], [108], [109], [110], [111], [102], [112], [113], [114], [115], [116], [117], [118], [119], [120], [121]
		Class-level context modeling	[109], [122]
	Multi-task learning	Combining with boundary detection	[106], [100], [105], [118], [123], [124], [125], [126], [127]
		Combining with change detection	[128]
		Combining with super-resolution	[129]
		Combining with region pixel segmentation	[108], [130]
		Combining with target height prediction	[131]
Large-scale variation	Multi-scale fusion	Fuse features of multiple scales	[110], [111], [132], [100], [133], [134]
	Foreground activation	Enhance the foreground information while suppressing background diversions	[135], [136], [99]
	Multi-scale test	Choose best segmentation result through multi-scale inputs	[137]
Class imbalance	One-stage method	Re-sampling and re-weighting	[100], [115], [135], [136], [138]
	Two-stage method	Decouple the learning of representation and classifier	[121], [139], [140], [141]
Large image size	Global-local pipeline	Two branches for processing downsampled the entire image and the cropped patches	[142], [143], [144], [145], [146], [147]
	Multi-stage refinement	Coarse-to-fine refinement on multi-stage predictions	[148], [149]
	Whole image processing	Lightweight network to process full-size images	[150], [151]

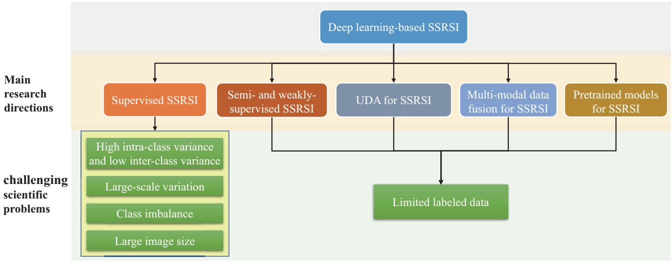
Deep learning-based SSRSI
Main research directions
Supervised SSRSI
Semi- and weakly-supervised SSRSI
UDA for SSRSI
Multi-modal data fusion for SSRSI
Pretrained models for SSRSI
challenging scientific problems
High intra-class variance and low inter-class variance
Large-scale variation
Class imbalance
Large image size
Limited labeled data

Fig. 2. Main research directions and challenging scientific problems for SSRSI.

the challenging problems faced by SSRSI. This section classifies the current techniques according to the research issues. Table I briefly summarizes supervised semantic segmentation methods of remote sensing images and the related works.

1) High intra-class variance and low inter-class variance: High intra-class and low inter-class variance seriously affect the accuracy of SSRSI and cause blurry boundary segmentation. The current main directions include effective context modeling and fusion, and multi-task learning.

Effective context modeling and fusion. Effective modeling and fusion of multi-type and multi-scale context information are the keys to solving the problem of high intra-class and low inter-class variance. Many methods have been proposed to model multi-scale, global, local, and class-level contexts.

The encoder-decoder structure (e.g., U-Net, HRNet) is the most popular architecture for attracting multi-scale feature. The shallow-level features include many spatial details, while deep-level feature contains more semantic information. Simple concatenation or addition method can achieve multi-scale context fusion in encoder-decoder architecture and has been applied in many SSRSI methods [89], [90], [91], [92], [93]. However, these strategies fail to account for the inherent semantic gaps between features at different levels, inadvertently embedding low-level background noise. Various adaptive methods improved multi-scale feature fusion by recalibrating the weights of high-level and low-level features utilizing the attention mechanism [94], [95], [96], [97], [98], [99], gate mechanism [88], [100], and deformable convolution [102].

Many studies of SSRSI considered spatial pyramid pooling a crucial module to model multi-scale features in parallel [103], [104]. In addition, some extensions of spatial pyramid pooling improved the multi-scale feature fusion strategy by adaptive spatial pooling [104], deformable convolution[101], and graph convolution[105].

The above work mainly focuses on the modeling of multi-scale context. In contrast, more research has been proposed to capture global context and class-level context to overcome the problem of high intra-class and low inter-class variance. The global pooling included in many models, such as ParseNet, only can learn the scene-level global context. Some researchers utilized attention mechanisms, including spatial and channel attention, to build a global context module on top of the feature extractor for fine global context modeling [112], [113], [114], [152]. Moreover, Cheng et al. [115] used information entropy as an attention score to enhance valuable global context while weakening the ambiguous representation.

Dilated convolution is also widely used to model global context. Some SSRIS models build backbone networks using dilated convolution instead of standard convolution for feature extraction. In addition, the ASPP proposed in DeepLabv2 was directly used as an essential module to extract multi-scale context in many studies [108], [109], [110]. At the same time, some research improved the parallel stack manner of multi-context aggregation in ASPP by a sequential global-to-local context aggregation [111] and attention mechanism [132].

In recent years, Transformer has been widely used to model the long-range dependency between pixels for SSRSI [116]. In

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License. For more information, see https://creativecommons.org/licenses/by-nc-nd/4.0/

This article has been accepted for publication in IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing. This is the author's version which has not been fully e
content may change prior to final publication. Citation information: DOI 10.1109/JSTARS.2023.3335891

other work, various semantic segmentation models proposed to use swin transformer as an encoder [117], [118], [119], an auxiliary encoder [102], or a decoder [120] for SSRSI. DWin-HRFormer [121] proposed a directional self-attention mechanism to overcome the problem that the network tends to ignore the feature orientation bias.

Most studies use spatial strategies to capture context information, including global pooling, dilated convolution, and spatial pyramid pooling. They did not distinguish pixels from different classes explicitly when calculating the context. Various types of targets around pixels have different contributions to pixel classification. Some work distinguished class-level contexts to further enhance SSRSI [95], [122].

Multi-task learning. Multi-task learning promotes model generalization and robustness by sharing representations between the primary and auxiliary tasks. By combining semantic segmentation tasks with other tasks, such as boundary detection, change detection, super-resolution, region pixel segmentation, target height prediction, multi-task learning can effectively alleviate the problem of high intra-class and low inter-class variance.

Boundary detection is beneficial to mitigate boundary blur and improve segmentation accuracy. The boundary prediction is a two-category task, while the semantic segmentation is a multi-category task in most cases. Most of the existing methods [106], [100], [118], [123], [124] used a shared backbone network to extract features for two tasks at the same time, then designed a boundary-aware module on top of the backbone network to generate a boundary map, finally combined a segmentation loss and a boundary loss for model training. Furthermore, some studies [105], [125], [126] used a framework with two branches in a sequential or parallel manner to process two tasks. However, the dual-stream architecture faces the challenge of high model complexity. Most multi-task methods manually set the weights of different losses, which may lead to a suboptimal training result. some studies [127] adopted the adaptive approach to weighing the segmentation loss and boundary loss for multi-task learning. In addition, combining semantic segmentation with other remote sensing tasks is also an important direction to improve the accuracy of image segmentation, including change detection [128], super-resolution [129], region pixel segmentation [108], [130] and target height prediction [131].

Overall, the problem of high intra-class variance and low inter-class variance leads to weak separability between targets, while current research still lacks systematic theoretical analysis. From the model's perspective, since the pixels of images are not independent and identically distributed, there have been many studies to build a specialized model architecture to model the global, local, and class-level context, as well as multi-scale context. However, modeling multiple contexts at once is still a challenging problem. From the data perspective, utilizing multi-task learning to introduce supervised information from auxiliary tasks to make the learned features more separable is a practical approach. However, reducing negative transfer or destructive interference between multiple tasks remains a significant challenge.

2) Large-scale variation: The long-distance observation of remote sensing images results in the target we are concerned about often containing very few pixels, and their scale changes significantly. To address these problems, researchers have carried out a lot of research on foreground activation [135], [136], [99], multi-scale fusion [110], [111], [100], [132], [133], [134], and multi-scale test [137].

Foreground activation help improve the extraction ability of small targets by enhancing foreground information while suppressing background noise. Zheng et al. [135] proposed a foreground-aware relation network and optimization method for foreground modeling. Ma et al. [136] proposed a foreground activation branch based on a feature pyramid network to activate the small objects. RSSFormer [99] proposed two attention-based modules to suppress background noise and enhance object saliency adaptively.

The most popular technique to adapt to large-scale variation is combining features from different scales and employing low-level features as much as possible to recover spatial details [110], [111], [100]. Merely merging features from the shallow and deep layers may lose the small targets. Adaptively selecting shallow or deep features based on target scale could be a possible solution to this problem [132], [133]. however, the limited learning capacity of each CNN tends to make tradeoffs in segmenting different-scale objects. Hang et al. [134] cascaded three subnetworks for gradually segmenting different-scale objects.

It is impractical to extract features at too many scales in a segmentation network. Meanwhile, there is no guarantee that those pre-defined scales are optimal for a given application scenario. Multi-scale test resizes the original images to various scales and feeds them into a segmentation network, the output segmentation maps are then assembled by average voting or progressive error correction [137].

Foreground activation-based methods belong to the coarse-to-fine approach, significantly improving small target recognition in remote sensing images. Still, the effectiveness of such methods seriously depends on the design of specific foreground activation modules and loss functions. Multi-scale fusion performs feature selection and fusion on a few pre-defined scales. For scenarios with significant variation in target scales, it is challenging to extract appropriate context information for multiple targets of different scales [137]. Although multi-scale test can alleviate this problem, they significantly increase computational complexity through image inputs with multiple scales.

3) Class imbalance: Class imbalance problem seriously affects the accuracy of SSRSI. Anand et al. [153] discussed the impact of class imbalance on neural network training, which showed that the majority class dominated the gradient and guided the model parameter update. The error of the minority category remained at a high level. The directions to solve this problem include one-stage methods [100], [115], [135], [136], [138] and two-stage methods [121], [139], [140], [141].

One-stage methods mainly include re-sampling and re-weighting. The primary purpose of re-sampling is to make each category has the same number of samples [154]. In contrast, re-weight the loss function in training is more com-

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License. For more information, see https://creativecommons.org/licenses/by-nc-nd/4.0/

This article has been accepted for publication in IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing. This is the author's version which has not been fully e
content may change prior to final publication. Citation information: DOI 10.1109/JSTARS.2023.3335891

mon in SSRSI. The typical method is to increase the loss weight of the samples that are difficult to classify and assign them a greater gradient in backpropagation. Therefore, the most important thing in re-weighting is how to measure the difficulty of classifying different samples. Sun et al. [100] proposed an adaptive edge loss function based on online hard example mining loss to alleviate the problem of recognizing tiny objects and sample imbalance. Cheng et al. [115] used entropy to measure the difficulty of sample classification and adopted an entropy matrix to weigh the cross-entropy loss. Zheng et al. [135] proposed a foreground-aware optimization strategy to make the model pay more attention to foreground and hard background examples. Ma et al. [136] applied a small object mining-based network optimization to select effective samples and refine the direction of the optimization. Bai et al. [138] proposed a calibrated focal loss by using a prediction confusion map to measure the classification difficulty.

Two-stage methods aim to improve the long-tail prediction by decoupling representation learning and classifier head. Specifically, the first stage learns the feature extractor and classifier head jointly, and then with the representation fixed, the second stage re-learns the classifier head with a class balancing strategy. Zhang et al. [139] introduced an adaptive calibration function and a distribution alignment strategy to calibrate the classifier. Zhang et al. [121] constructed a distributed alignment module to adjust the biased decision boundaries in the second stage. Cui et al. [140] introduced a region rebalance strategy instead of the pixel rebalance strategy by encouraging features to lie in a more balanced region classification space. Zhong et al. [141] proposed a center collapse regularizer to encourage the network to learn class-equiangular and class-maximally separated structured features.

Compared with image classification task, in which image samples are independent and identically distributed, the strong contextual correlation between pixels results that class frequency in the pixel domain being an unsuitable guide for rebalancing [140], [141]. Thus, one-stage methods in image classification tasks might not always produce performance improvement for semantic segmentation. Recent two-stage methods model the dependency between pixels through region rebalance [140] and neural collapse [141]. However, they mainly focus on the semantic segmentation of natural images, and there is still a lack of pointed studies for remote sensing images with more significant class imbalance issues.

4) Large image size: Deep semantic segmentation models have high memory demand. Current approaches either downsample large-size images or crop them into small patches for separate processing. However, the former removes details, while the latter destroys image context. To overcome these problems, current methods mainly include three categories, including global-local pipeline [142], [143], [144], [145], [146], [147], multi-stage refinement pipeline [148], [149], and the whole image processing pipeline [150], [151].

The framework of the global-local pipeline utilizes two parallel branches for processing downsampled entire image and its cropped local patches separately, then conducting the global-to-local fusion on the predictions. The most representative work is GLNet [142]. Furthermore, some work has improved GLNet. Some work refined segmentation by adaptively distinguishing the importance of different cropped patches [143]. FCtL [144] captured the relevance between the local patch and its various contexts to produce high-quality local segmentation results. Nogueira et al. [145] proposed to select the best size for cropped patches by training a dilated convolution network with distinct patch sizes. Other work investigated how to achieve global and local fusion using the Transformer [146], [147].

Multi-stage refinement pipeline adopts the refinement scheme and performs coarse-to-fine refinement on multi-stage predictions for better segmentation results. CascadePSP [148] proposed a segmentation refinement model, where coarse outputs from any other segmentation models are used as input to refine boundary details in a coarse-to-fine manner. MagNet [149] introduced a multi-scale framework where the output segmentation map will be progressively refined as the image is analyzed from the coarsest to the finest scale.

The whole image processing pipeline focus on inferring the whole image directly. ISDNet [150] directly processed the full-scale and downsampled inputs by integrating shallow and deep networks. The shallow network used full-scale images to enhance spatial detail extraction. The deep network takes the downsampled image to extract high-level semantic information. ElegantSeg [151] processed holistic extra-large image semantic segmentation by extending the tensor storage from GPU to host memory.

In summary, the global-local and multi-stage refinement pipelines require cropping the image into patches. The segmentation model must process both the downsampled full-size images and the cropped patches simultaneously, inevitably leading to complex network architecture and a slow inference speed. The whole image processing pipeline can accelerate the inference speed, but it must face large memory consumption. Therefore, when facing the problem of the large size of remote sensing images, it is necessary to balance multiple perspectives, such as model performance, memory consumption, and inference speed.

B. Semi-supervised and weakly-supervised SSRSI

A critical bottleneck in building deep learning-based semantic segmentation models is that they require massive pixel-level annotation data for model training. Semi-supervised and weakly supervised learning has attracted more and more attention because they can reduce the dependence on high-quality labeled data.

1) Semi-supervised models: Semi-supervised SSRSI aims to leverage both labeled and unlabeled data simultaneously to learn semantic segmentation models. The current main methods comprise three categories, i.e., consistency regularization, self-training and hybrid methods. Table II summarizes some semi-supervised semantic segmentation methods.

Consistency regularization. The consistency regularization methods are based on the low-density assumption that the learned decision function should lie in low-density regions in the input space. Fig.3 shows a simple semi-supervised semantic segmentation framework based on consistent regularization.

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License. For more information, see https://creativecommons.org/licenses/by-nc-nd/4.0/

This article has been accepted for publication in IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing. This is the author's version which has not been fully edited and
content may change prior to final publication. Citation information: DOI 10.1109/JSTARS.2023.3335891

TABLE II
SOME REPRESENTATIVE SEMI-SUPERVISED SSRSI.

Category	Method	Publication	Segmentation Model	Main Mechanism
Consistency regularization	Kang et al. [155]	IEEE J-STARS 2021	DeepLab V3+	Image perturbation
	Zhang et al. [156]	IEEE J-STARS 2022	DeepLabv2 with ResNet	Image perturbation
	Ouali et al. [157]	CVPR 2020	DeepLabv3 +, PSPNet	Feature perturbation
	Chen et al. [158]	CVPR 2021	DeepLabV3+	Network perturbation
Self-training	Sun et al. [159]	IEEE J-STARS 2020	DeepLabv3+ combining boundary attention module	GAN-based
	Zheng et al. [160]	RS 2022	Segmentation network with double-branch encoder	GAN-based
	Kwon et al. [161]	CVPR 2022	DeepLabv3+	Auxiliary task
	Lu et al. [162]	IEEE TGRS 2022	DeepLabv2 with ResNet	Adaptive pseudo label selection
Hybrid method	Wang et al. [163]	RS 2020	U-Net, DeepLabv3, DeepLabv3+	Image perturbation, auxiliary task
	Li et al. [164]	IEEE TGRS 2021	FCN with ResNet-50, U-Net	Feature perturbation, GAN-based
	Wang et al. [165]	IEEE TGRS 2021	U-Net, DeepLabv3, DeepLabv3+	Image perturbation, threshold-based
	Li et al. [166]	P&RS 2021	DeepLabv3 +	Image perturbation, threshold-based
	Chen et al. [167]	JAG 2022	DeepLabv2-based model	Feature perturbation, threshold-based
	Chen et al. [168]	P&RS 2023	U-Net-based model combing dilated convolution and attention	Image perturbation, threshold-based

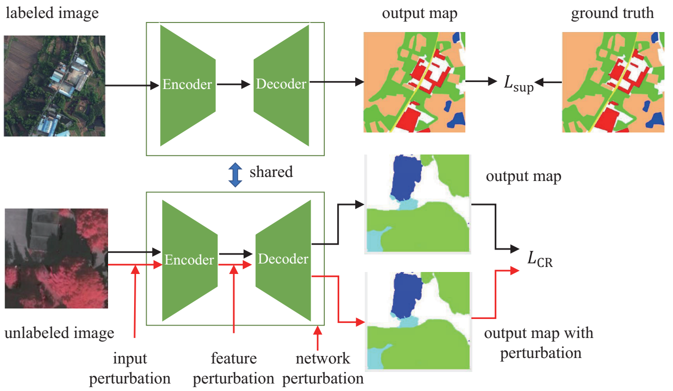

Fig. 3. A simple semi-supervised SSRSI framework based on consistent regularization. The consistency with various perturbations on unlabeled data is carried out at different levels.

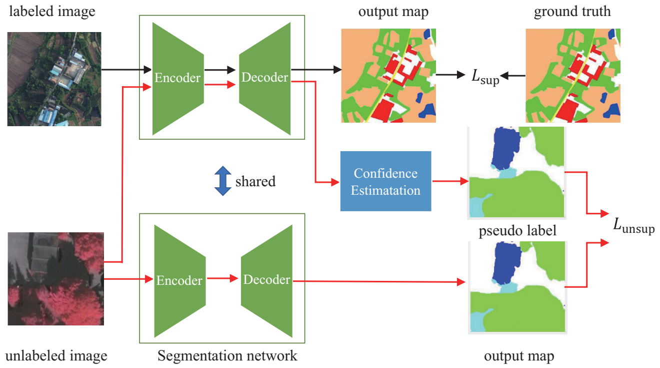

Fig. 4. A simple semi-supervised SSRSI framework based on self-training. The pseudo-labels of unlabeled data with high confidence generated from a model trained on labeled data refine the model iteratively.

It enforces the consistency of the predictions with various perturbations on the same unlabeled examples, e.g., input perturbation, feature perturbation, and network perturbation. Finally, the segmentation model is learned by combining the supervision loss L_{SUP} of the labeled images and the consistency regularization loss L_{CR} of the unlabeled images.

Input perturbation-based approaches leverage random data augmentations on the input images and then perform consistency constraints for the predictions produced from augmented images [155], [166], [156], [169]. Feature perturbation frameworks enforce an invariance of the predictions over small feature perturbations for the same input unlabeled image [157]. Network perturbation-based methods enforce prediction consistency by applying perturbations to the network itself. They are typically implemented by adopting parallel segmentation networks with different structures or the same architecture but with various initializations [158].

Self-training. Self-training, also known as pseudo labeling, has been proposed to solve semi-supervised SSRSI. Fig.4 shows a simple semi-supervised semantic segmentation framework based on self-training. Self-training generates pseudo labels for unlabeled images using a model trained on labeled ones and uses them as supervised signals. The model is further retrained by combining the supervised loss L_{SUP} of labeled images with the pseudo-supervised loss L_{unsup} of unlabeled images. The pseudo labels of unlabeled images inevitably contain some errors. Learning using such labels as supervision causes confirmation bias towards the errors and returns corrupted models consequently. Therefore, the most critical issue is to estimate the confidence of predictions and help select the most effective pseudo labels.

One simplest method that solves this issue is the threshold-based method using an uncertain output map as supervision[190], but their performance depends heavily on hand-tuned thresholds. GANs are an essential solution for correcting pseudo labels of semi-supervised remote sensing image segmentation, which is studied to adaptively select high-confident segmentation predictions on unlabeled images as pseudo segmentation. The semantic segmentation network is used as the generator, while the role of the discriminator is to assess the confidence of predictions [159], [160].

The coupling issue is another problem of SSRSI methods based on self-training, i.e., the teacher (trained on labeled data) and student networks (re-trained on labeled data and pseudo-labeled data) easily generate similar predictions on the same input. Many methods have been proposed to decouple their predictions as well as alleviate overfitting on noisy pseudo labels, such as injecting strong data augmentations on unlabeled images [170], repetitively producing pseudo

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License. For more information, see https://creativecommons.org/licenses/by-nc-nd/4.0/

This article has been accepted for publication in IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing. This is the author's version which has not been fully e
content may change prior to final publication. Citation information: DOI 10.1109/JSTARS.2023.3335891

labels for each training minibatch [158], adaptive choosing a threshold for filtering incorrect labels [162]. Furthermore, some studies [171], [161] introduced auxiliary tasks to correct pseudo labels. Ke et al. [171] proposed a flaw detector to estimate the prediction confidence and help correct the unreliable pixels in the predictions. Kwon et al. [161] introduced a new auxiliary task called error localization, which enabled semi-supervised learning to be robust against inaccurate pseudo-labels by disregarding label noises.

Hybrid method. There has been increasing research on combining self-training with consistency regularization. Consistency regularization can help improve the teacher network, thus helping self-training to generate high-quality pseudo labels and further enhance segmentation models. Wang et al. [163] performed image perturbation by color jittering and random flipping, then employed the consistency regularization trained model for the average update of pseudo labels. Li et al. [164] proposed a GAN-based consistency self-training framework, which used the semantic segmentation network as the generator to produce predictions and sent them with random perturbations to the discriminator to assess the confidence. Wang et al. [165] randomly pasted part of the labeled image into the unlabeled image as a strong perturbation, generating pseudo labels for weakly augmented data. Li et al. [166] combined self-training and consistency regularization for cross-domain SSRSI, which learned the transfer invariance and rotation consistency under image perturbation and generated pseudo labels for unlabeled data. Chen et al. [167] leveraged the feature perturbation to conduct regularization constraints, and provided the one-hot pseudo supervisions for further self-training. SemiRoadExNet [168] proposed a GAN-based semi-supervised road extraction network, which leveraged the potential information of low-confidence pixels in pseudo labels by entropy maps generated by the segmentation network.

Overall, consistency regularization extracts knowledge from unlabeled data by learning consistent predictions under various perturbations. However, the problem of high intra-class and low inter-class variance in remote sensing images may result in insignificant low-density regions, making the predictions of consistency regularization methods potentially incorrect. Self-training relies on the quality of pseudo-labels and the metrics for selecting high-confidence pseudo-labeled samples. In SSRSI, combining these two methods and other methods, such as contrastive learning, is a more promising direction.

2) Weakly-supervised models: Weak annotation is far easier to collect than fully pixel-level annotations. Weakly-supervised semantic segmentation (WSSS) approaches used weak annotations to reduce the dependence on fully annotated data. Most existing methods of WSSS follow a two-step pipeline, i.e., generating pseudo labels and training segmentation models. This section introduces the related research according to the categories of weakly-supervised labels, which are image-level labels [172], [173], [174], [175], bounding boxes [176], [177], point annotation [178], [179], [180], and scribble annotation [180], [181]. Table III compares the characteristics of some representative WSSS methods.

Image-level WSSS. Image-level WSSS is a challenging issue since the image labels indicate only the existence of object categories and do not inform accurate object locations that are essential for semantic segmentation. Recent Image-level WSSS approaches commonly rely on Class Activation Maps (CAMs) [187] as seeds to generate pseudo-ground truths. However, CAMs identify class-specific discriminative image regions, making these initial labels usually incomplete and noisy. Some studies have proposed targeted methods to address this issue. Cao et al. [174] proposed a coarse-to-fine pixel-level label generation method to alleviate the local high response property of CAMs and the potential label noise problem by object-based label extraction and noisy label correction. Li et al. [153] designed a confidence area selection module and a low-to-high loss function to obtain reliable supervision information from the coarse labels. Fang et al. [175] used the adversarial climbing strategy to optimize CAMs. Zeng et al. [182] proposed a framework directly transferring the scene classification model to perform semantic segmentation.

Bounding boxes-level WSSS. Bounding boxes provide information about individual objects and their location. They can help remove the irrelevant regions and focus on the foreground regions. Bounding boxes-level WSSS is mainly focusing on the semantic segmentation of natural images. Dai et al. [176] adopted a recursive training procedure to iterate between automatically generating region proposals and training segmentation networks. Song et al. [177] proposed a filling rate-guided adaptive loss to help the model ignore the incorrectly labeled pixels in the pixel-level proposal produced from the bounding box supervision.

Point-level WSSS. A point label is an instance-wise point, which roughly points out the center location of an object but does not indicate the object's scope. Lian et al. [178] presented a road seeds estimation model to extract pseudo ground truths from point labels. NFANet [179] fully utilized the high similarity between pixels in water bodies and performed water extraction through the features of adjacent pixels and point labels. Chen et al. [185] used low-resolution land cover products (LCP) as WS information to obtain an accurate high-resolution LCP.

Scribble-level WSSS. A scribble roughly provides an object's location and extension with a single stroke. Maggiolo et al. [180] utilized scribbled annotation to build a weakly-supervised semantic segmentation model and used a fully connected CRF to generate the pseudo ground truth by modeling long-range dependencies. Wei et al. [181] proposed a weakly-supervised road surface extraction method, which introduced a road label propagation algorithm to propagate semantic information from sparse scribbles to unlabeled pixels.

Compared with fully pixel-level annotations, weak-supervised annotations have the characteristics of fast labeling and low time cost. However, weak annotations usually lack vital information, including shape, texture, and edges, making the generated pseudo labels incomplete and noisy. WSSS shows limited performance in obtaining the comprehensiveness of semantic information and segmentation accuracy.

C. Unsupervised Domain Adaptation for SSRSI

In remote sensing application scenarios, training data (source domain) and test data (target domain) often have the

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License. For more information, see https://creativecommons.org/licenses/by-nc-nd/4.0/

This article has been accepted for publication in IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing. This is the author's version which has not been fully e
content may change prior to final publication. Citation information: DOI 10.1109/JSTARS.2023.3335891

TABLE III
SOME REPRESENTATIVE WEAKLY-SUPERVISED SSRSI.

Category	Method	Publication	Segmentation Model	Pseudo Label Generation Method
Image-level	Cao et al. [174]	P&RS 2022	U-Net	A coarse-to-fine pixel-level label generation method
	Li et al. [173]	P&RS 2022	A multi-scale convolutional network with resolution-preserving	A confidence area selection module with a low-to-high loss function
	Fang et al. [175]	IEEE J-STARS 2022	DeepLabv3+	A random walk strategy
	Zeng et al. [182]	IEEE TGRS 2023	A U-Net-like segmentation model	An end-to-end model without generating pseudo labels
Bounding box-level	Dai et al. [176]	CVPR 2015	FCN combined with CRF	Multiscale combinatorial grouping [183]
	Song et al. [177]	CVPR 2019	DeepLab	Unsupervised dense CRF
Point-level	Lian et al.[200]	IEEE TGRS 2021	Stacked hourglass network [184]	A road seeds estimation model combined with SVM
	Lu et al. [179]	P&RS 2022	A U-Net-like segmentation model	A feature aggregation module and post-processing method
	Chen et al. [185]	P&RS 2023	VGG-16+MLP	Label propagation
Scribble-level	Maggiolo et al.[203]	IEEE TGRS 2021	Hypercolumn network [186], U-Net	A fully connected CRF
	Wei et al. [181]	IEEE TGRS 2021	DeepLabv3+	A road label propagation algorithm

problem of domain drift. For example, the urban and rural scenes have different class distributions. The urban scenes with high population densities contain many artificial objects, such as buildings and roads. In contrast, the rural scenes include more natural elements, such as forests and water. Model performance may decline rapidly under domain shifts. UDA semantic segmentation, aiming to improve the generalization ability for transferring knowledge from the source domain to the target domain, has recently gained increasing attention. Most UDA methods focus on photometric alignment to align the source and target images in the input, feature, and output spaces. Fig.5 shows a generic domain alignment framework, often achieved through a Siamese architecture with two streams. Each stream corresponds to a segmentation model for processing the source and target domain images. The parameters of the two models can be shared, partially shared, or domain-specific. Generally speaking, the Siamese network can be trained by combining two loss items. One is the cross-entropy loss L_{CE} corresponding to the supervisory signals in the source domain, and the other is the adaptive loss L_{Ada} to measure the distance between the source samples and the target samples. Table IV compares the characteristics of some UDA methods for SSRSI. Next, we detail the alignment solutions between source and target domains at different levels (i.e. input, feature, and output levels).

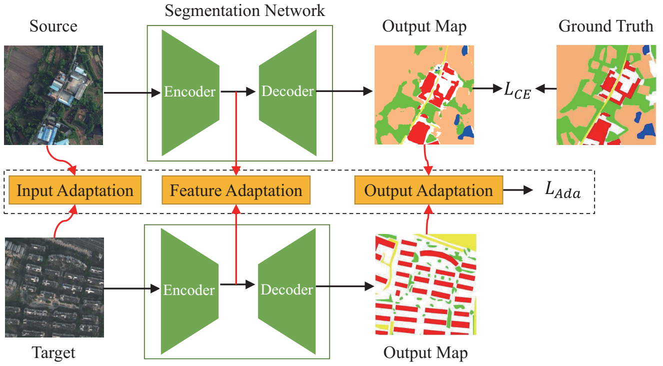

Fig. 5. A generic domain alignment framework. A Siamese architecture with one branch per domain is adopted, where domain alignment is often performed at input-, feature-, and output-level.

1) Input-level adaptation: Due to the high-level semantic similarity in scene content and layout between the images from the source domain and the target domain, many studies of input-level adaptation use image translation or style transfer methods to map data from one domain to another while maintaining semantic consistency between images. Then, the semantic segmentation models can be trained using translated images with source domain labels.

A considerable number of studies [188], [189], [190] used GAN architecture to address the input space's domain adaptation in SSRSI. The most typical method is CycleGAN [191], which is used to make a bidirectional image-to-image translation between the source and target domains [188], [191]. However, remote sensing images contain a lot of complex and heterogeneous structures, it isn't easy to wholly maintain the semantic consistency of images translated by GAN. Tasar et al. [192] proposed a GAN-based UDA framework named ColorMapGAN, which transformed the colors of the training images into the colors of the test images without doing any structural changes. Other studies adopted traditional image processing methods for image translation, such as the Wallis filter method [193].

Most of existing research focuses on the domain adaptation of single source and target domains, hindering their further expansion. Some researchers have extended domain adaptation to scenarios with multi-source and multi-target domains. Tasar et al. [194] proposed a DAugNet for multi-source, multi-target, and life-long domain adaptation of satellite images in an unsupervised manner. In [195], a StandardGAN was proposed to deal with the multi-source domain adaptation problem. They standardized each source and target domain using GANs so that all the data had similar distributions.

2) Feature-level adaptation: The feature-level domain adaption methods solve the problem of domain drift by minimizing the distribution of the source domain and target domain in the feature space. In general, they learn domain invariant features by forcing a feature extractor to adjust the feature distribution of the source domain and target domain data. Therefore, choosing a proper divergence measure is the core of

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License. For more information, see https://creativecommons.org/licenses/by-nc-nd/4.0/

This article has been accepted for publication in IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing. This is the author's version which has not been fully e
content may change prior to final publication. Citation information: DOI 10.1109/JSTARS.2023.3335891

these methods. Widely used measures include Maximum Mean Discrepancy(MMD) [196], Correlation Alignment (CORAL) [197], Contrastive Domain Discrepancy(CDD) [198], Wasserstein Distance [199], etc. In SSRSI, some work [200], [201], [202] has proposed various metrics based on covariance to measure the domain divergence.

Adversarial learning uses a discriminator as a measure, which is another effective way to align features in different domains [203]. Lu et al. [204] proposed an adversarial learning framework for cross-domain road detection. A global-local alignment operation was introduced to adjust the weight of the adversarial loss according to the recognition difficulty of each pixel. MBATA-GAN [205] combined cross-attention and self-attention to learn transferable features between the source domain and target domain in a framework of adversarial learning. Wang et al. [206] proposed a two-stage UDA framework. The first stage performed global-local alignment by adversarial learning, while the second stage used self-training to align category features.

3) Output-level adaptation: Prediction maps of the segmentation network can retain useful semantic information, some researchers utilize output-level alignment to realize domain adaptation. At present, the main methods include adversarial learning [207], [208], [209] and self-training [210], [211], and their combination [212], [213], [214], [215], [216], [217].

Adversarial learning implements the output-level domain adaptation using a domain discriminator to distinguish whether the segmentation predictions were from the source or the target domain. Zheng et al. [207] performed local feature alignment using the adaptive weight on the predictive entropy map in the target domain to guide the adversarial learning. Chen et al. [208] proposed a region and category adaption domain discriminator to measure the differences in regions and categories. Chen et al. [209] designed a category-certainty attention module to reduce the attention of the discriminator on category-level aligned features and increase the attention on category-level unaligned features.

Self-training generates pseudo-labels for the target domain images using the model trained in the source domain and iteratively refining the segmentation model using the most confident pseudo-labels. Tong et al. [210] proposed a pseudo-labeling and retrieval-based sample selection scheme. The patches with high confidence were assigned pseudo labels and employed as the queries to retrieve related samples from the source data. The retrieved results were used for fine-tuning the segmentation model. DAFormer [218] used Transformer as the backbone and proposed three training strategies, including rare class sampling, thing-class ImageNet feature loss, and a learning rate warm-up method for UDA segmentation. Li et al. [211] utilized a DAFormer for UDA segmentation, and presented a local dynamic quality strategy to improve the quality of the pseudo-labels.

The main idea to combine adversarial learning with self-training is to use adversarial learning to achieve domain alignment in the output space while using a self-training strategy to generate pseudo labels for samples in the target domain, then further train the segmentation model using high-quality pseudo labels [213], [214], [216]. Yan et al. [212] proposed a triplet adversarial domain adaptation method. The discriminator took a triplet of segmentation maps as input and decided whether two segmentation maps were from the same domain or not. Zhang et al. [215] proposed a two-stage cross-domain adaptation framework. The first stage used GAN to align the source domain to the easy-to-adapt target domain. The second stage used the pseudo label generated in the first stage to adapt the segmentation model to the hard-to-adapt target patches. MemoryAdaptNet [217] explored memory mechanisms to store, extract, and update variant domain-level prototype information in adversarial learning and self-training frameworks. Ma et al. [219] introduced local consistency and global diversity metric on the basis of the framework of adversarial learning and self-training learning to improve the segmentation performance.

4) Multi-level adaptation: Multi-level adaptation often achieves better alignment results by simultaneously aligning domains at multiple levels. Ji et al. [220] aligned the source and target images at the image-, feature-, and output-level by CycleGAN, adversarial learning, and mean teacher model, respectively. Shi et al. [221] used CycleGAN and adversarial learning to realize the input- and feature-level domain adaptation, respectively. Li et al. [222] performed input-level domain adaption by projecting the source and target domains into a color space with normalized distribution and implemented feature-level alignment using adversarial learning. Self-training was adopted to enhance the domain adaption at output space. Liu et al. [223] proposed a bispace adversarial learning strategy to minimize domain discrepancy in feature space and output space. Xu et al. [224] proposed a class-aware domain alignment approach, including adaptive category selection and adaptive category alignment, to model the intra-class compactness and inter-class separability. Chen et al. [225] developed a class-aware domain adaptation method that separately executed class-specific and global domain alignment on feature and output spaces by a joint local and global adversarial adaptation framework. MDANet [226] achieved distribution alignment in input, feature, and output levels, and combined contrastive learning and memory mechanisms to extract and utilize domain invariant features.

The image-level domain adaptation methods achieve domain invariance by reducing the cross-domain discrepancy in the image layout and structure. The potential issue is that the performance heavily relies on the quality of translated images. Pixel-level flaws could significantly influence the accuracy. The feature-level domain adaptation methods can realize feature alignment in hidden space with adversarial strategies or divergence measures. However, aligning high-dimensional features brings a high computational burden regarding large-scale and complex remote sensing scenes. The output-level domain adaptation can reach efficient cross-domain distribution alignment on the low-dimensional output space. But including only high-level semantic information in output space affects the performance of UDA semantic segmentation. The multi-level domain adaptation method is equivalent to a class of multi-task learning methods. Learning the weights of different tasks is important as the number of tasks increases.

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License. For more information, see https://creativecommons.org/licenses/by-nc-nd/4.0/

This article has been accepted for publication in IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing. This is the author's version which has not been fully e
content may change prior to final publication. Citation information: DOI 10.1109/JSTARS.2023.3335891

TABLE IV
SOME REPRESENTATIVE UDA METHODS OF SSRSI.

Category	Method	Publication	Segmentation Model	Domain alignment
Image-level	Benjdira et al. [188]	RS 2019	U-Net	CycleGAN
	Wittich et al. [189]	P&RS 2021	FCN	GAN
	Cai et al. [190]	IEEE TGRS 2022	DeepLabv2	CycleGAN
	Tasar et al. [192]	IEEE TGRS 2020	U-Net	ColorMapGAN
	Tasar et al. [194]	CVPRW2020	U-Net	GAN
	Tasar et al. [195]	IEEE TGRS 2020	U-Net	GAN
Feature-level	Zhang et al. [200]	IEEE GRSL2020	U-Net	Feature covariance alignment
	Wu et al. [201]	IEEE TGRS 2022	DeepLabv2	Covariance regularization
	Iqbal et al. [203]	P&RS 2020	U-Net	GAN
	Lu et al. [204]	P&RS 2020	FCN	GAN
	Liu et al. [202]	IEEE GRSL2022	FCN	GAN
	Ma et al. [205]	IEEE TGRS 2023	DeepLabv2 with ResNet-101	GAN
	Wang et al. [206]	IEEE J-STARS 2023	Deeplabv3+ with ResNet-101	GAN
Output-level	Zheng et al. [207]	IEEE TGRS 2021	DeepLabv2 with ResNet-101	GAN
	Chen et al. [208]	IEEE TGRS 2022	DeepLabv2	GAN
	Chen et al. [209]	IEEE TGRS 2022	DeepLabv2 with ResNet-101	GAN
	Tong et al. [210]]	RSE 2020	A hybrid model combining patch classification and pixel segmentation	Self-training
	DAFormer [218]	CVPR2022	DAFormer	self-training
	Li et al. [211]	RS 2022	DAFormer	Self-training
	Yan et al. [212]	IEEE TGRS 2019	DeepLabv3	Adversarial learning and self-training
	Zhang et al. [213]	IEEE TGRS 2021	An FCN-like model with ResNet and feature pyramid	Adversarial learning and self-training
	Yan et al. [214]	IEEE TGRS 2019	DeepLabv3	Adversarial learning and self-training
	Zhang et al. [215]	IEEE TGRS 2021	An FCN-like model with ResNet	Adversarial learning and self-training
	Yao et al. [216]	IEEE TGRS 2021	DeepLabv2 with a pretrained ResNet	Adversarial learning and self-training
	Zhu et al. [217]	IEEE TGRS 2023	DeepLabv2 with ResNet-101	Adversarial learning and self-training
Multi-level	Peng et al. [193]	IEEE TGRS 2021	U-Net with attention mechanism	(1) Image-level: Wallis filter (2) Feature-level: adversarial learning (3) Output-level: mean teacher model
	Ji et al. [220]	IEEE TGRS 2020	Modified FCN	(1) Image-level: CycleGAN (2) Feature-level: adversarial learning (3) Output-level: mean teacher model
	Shi et al. [221]	IEEE GRSL2020	FCN with ASPP	(1) Image-level: CycleGAN (2) Feature-level: adversarial learning
	Li et al. [222]	IEEE TGRS 2022	DeepLabv2	(1) Image-level: Colorspace Mapping (2) Feature-level: adversarial learning (3) Output-level: self-training
	Liu et al. [223]	IEEE TGRS 2020	A two-beaches network with U-Net and wavelet transform	(1) Feature-level: adversarial learning (2) Output-level: adversarial learning
	Xu et al. [224]	IEEE TGRS 2022	DeepLabv2 with ResNet-101	(1) Feature-level: adversarial learning (2) Output-level: category alignment
	Chen et al. [225]	IEEE J-STARS 2020	Modified DeepLabv2	(1) Feature-level: adversarial learning (2) Output-level: adversarial learning
	Chen et al. [226]	IEEE TGRS 2023	Deeplabv3+ with ResNet-101	(1) Image-level: Wallis filter (2) Feature-level: adversarial learning (3) Output-level: self-training

D. Multi-modal data fusion for SSRSI

Remote sensing data are often multimodal, e.g., optical (multi- and hyperspectral), Lidar, and synthetic aperture radar (SAR). In addition, sensors often carry auxiliary data, such as digital surface models (DSMs). Remote sensing data of different modalities can provide information from different perspectives for the targets. Fusing multi-modal data has become a new direction to improve the performance of semantic segmentation algorithms. Generally speaking, designing a deep semantic segmentation network for multi-modal fusion must solve three critical problems. What to fuse: Which modal data need to be fused, and how to represent them; When to fuse: At which stage shall multimodal data be fused; How to fuse: What kind of fusion methods should be used to realize information fusion. In this section, we introduce the existing research on these three aspects. Table V shows representative methods based on multi-modal data fusion.

1) What to fuse:

Fusion of multispectral image and auxiliary data. A DSM provides elevation data of objects in a remote sensing image, facilitating the segmentation of tall objects, such as buildings and trees. Utilizing DSMs to extract additional features can further improve the semantic segmentation of multispectral images [228], [229], [232], [248], [233], [235], [236], [238]. In addition, other indicators such as Normalized DSM(NDSM) [230], [231], Digital Elevation Model (DEM) [123], [227], [237], Normalized Difference Vegetation Index (NDVI) [230], [231], [234] from the near-infrared and red channels, Normalized Difference Water Index [231] using near-infrared and green channels, are utilized to fuse with the multispectral images.

Fusion of multispectral image and LiDAR data [236], [239], [240]. Multispectral images can provide high spatial resolution and rich spectral and textural information under good

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License. For more information, see https://creativecommons.org/licenses/by-nc-nd/4.0/

This article has been accepted for publication in IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing. This is the author's version which has not been fully e
content may change prior to final publication. Citation information: DOI 10.1109/JSTARS.2023.3335891

TABLE V
SOME REPRESENTATIVE REMOTE SENSING IMAGE SEMANTIC SEGMENTATION METHODS BASED ON MULTI-MODAL DATA FUSION. "MULTI" DENOTES MULTISPECTRAL IMAGE, "AUX" DENOTES AUXILIARY DATA, "HYPER" DENOTES HYPERSPECTRAL IMAGE.

Category	Method	Publish	What to fuse	How to fuse	When to fuse
Multi and Aux	Marmanis et al. [227]	ISPRS Annals 2016	RGB+DEM	Concatenation	Late
	Sherrah et al. [228]	arXiv 2016	RGB+DSM	Concatenation	Late
	Kampffmeyer et al. [229]	CVPRW2016	RGB+DSM	Concatenation	Early
	Audebert et al. [230]	IEEE TGRS 2016	IRRG+DSM/NDSM/NDVI	Averaging	Middle and Late
	Volpi et al. [231]	IEEE TGRS 2016	RGB+NDVI/NDWI/NDSM	Concatenation	Early
	Zhang et al. [232]	RS 2017	IRRGB+DSM	Summation	Middle
	Marmanis et al. [123]	P&RS 2018	RGB+DEM	Concatenation	Late
	Cao et al. [233]	IEEE GRSL2019	IRRG+DSM	Concatenation	Middle
	Zheng et al. [234]	IEEE TGRS 2021	RGB+ DSM/NDVI	Adaptive fusion (gate)	Middle
	Zhou et al. [235]	IEEE TGRS 2021	RGB+DSM	Adaptive fusion (gate)	Middle
	Zhao et al. [236]	Neurocomputing 2022	RGB+DSM	Adaptive fusion (attention)	Middle
	Liu et al. [237]	IEEE TGRS 2023	RGB+DEM	Summation	Middle
	Zhou et al. [238]	IEEE TGRS 2023	RGB+DSM	Adaptive fusion (attention, gate)	Middle
Multi and LiDAR	Liu et al. [239]	CVPRW2017	RGB+LiDAR	Adaptive fusion (CRF)	Late
	Audebert et al. [240]	P&RS 2018	RGB+LiDAR	Summation	Early and Late
	Sun et al. [241]	P&RS 2018	RGB+LiDAR	Concatenation	Middle
Multi and SAR	Li et al. [242]	IEEE J-STARS 2020	RGBIR+SAR	Adaptive fusion (attention)	Middle
	Li et al. [243]	JAG 2022	RGBIR+SAR	Adaptive fusion (attention)	Middle
	Ren et al. [244]	JAG 2022	RGBIR+SAR	Adaptive fusion (attention)	Middle
	Kang et al. [245]	IEEE J-STARS 2022	RGBIR+SAR	Adaptive fusion (gate)	Middle
	He et al. [246]	P&RS 2022	RGB+DSM,RGBIR+SAR	Adaptive fusion (GCN)	Middle
Multi and Hyper	Hong et al. [247]	IEEE TGRS 2020	Multi+Hyper	Adaptive fusion (GAN)	Middle

illumination and fair weather. However, optical cameras are strongly affected by the level of illumination. LiDAR senses the environment by using its emitted pulses of laser light. Therefore, LiDAR is only marginally affected by external light conditions. Furthermore, LiDAR provides accurate range measurements.Thus, it is easy to see that fusing the data of LiDAR and multispectral images could give an enhanced total.

Fusion of multispectral image and SAR [241], [242], [243], [244]. SAR sensors penetrate particular ground objects in all weather conditions, and SAR images provide rich geometric information on ground objects. Taking SAR images as complementary data in optical image processing can prevent the weather's interference to a certain extent. Fusing multispectral and SAR images for SSRSI is a promising approach to improving segmentation accuracy.

Fusion of multispectral image and hyperspectral image [247]. Hyperspectral image collects the electromagnetic spectrum from the visible to the near-infrared wavelength and has more abundant spectral characteristics. However, hyperspectral images have the disadvantage of low spatial resolution. Combining multispectral and hyperspectral images for semantic segmentation can benefit from each technology's unique advantages [204].

2) How to fuse: This section summarizes typical fusion operations in neural networks. We restrict our discussion to two sensing modalities for simplicity, though more still apply.

Addition or Average [237], [230], [232]. This operation adds the feature maps element-wise or calculates the average mean of the feature maps to obtain fused feature.

Concatenation [123], [227], [228], [229], [231], [248], [233]. This method usually stacks the feature maps from different modalities along their channels.

Adaptive fusion. The above two modes ignore the heterogeneity between different modalities and are prone to generate redundant features and noise information. Therefore, it is better to select complementary information for fusion adaptively. Among them, the most representative strategies include attention mechanism [236], [238], [242], [243], [244], gate mechanism [234], [235], [238], [245], Graph neural network (GCN) [246], GAN [247], conditional random field [239].

3) When to fuse: Deep neural networks represent features hierarchically and offer various choices to combine multimodal data at early, middle, or late stages. Next, we will briefly introduce the early, middle, and late fusions.

Early fusion. This strategy concatenates the images of two modalities along the channels into a multichannel tensor and feeds the tensor into a one-stream neural network for training [229], [231].

Late fusion. This method merges features from different modalities at the last layer of the neural network before the final prediction layer or integrates the prediction results from different modalities to generate the final segmentation map [123], [227], [228], [230], [239].

Middle fusion. This method is the compromise of early and late fusion. The middle fusion uses a dual-stream architecture to extract features from the different modalities. The feature fusion is carried out at different feature levels of the semantic segmentation model. Therefore, there is a lot of work to explore the appropriate fusion level for the best results [237], [232], [233], [241].

Early fusion learns the joint features of multiple modalities at an early stage, fully exploiting the information of the raw data. Meantime, early fusion has low computation requirements as it jointly processes the multiple sensing modalities early. However, early fusion is sensitive to spatial-temporal data misalignments caused by differences in imaging principles, imaging time, imaging conditions, and resolution. Only its domain-specific network must be trained when introducing a new modality without affecting other branches. Late fusion requires multiple prediction networks to predict segmented

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License. For more information, see https://creativecommons.org/licenses/by-nc-nd/4.0/

This article has been accepted for publication in IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing. This is the author's version which has not been fully e
content may change prior to final publication. Citation information: DOI 10.1109/JSTARS.2023.3335891

maps from different modalities, thus requiring significant computational complexity and memory consumption. In addition, it discards rich intermediate features that may be highly beneficial when fused. The middle fusion approach has high flexibility. Nevertheless, given a network architecture, finding the optimal way to fuse the middle layer isn't easy. The optimal fusion architecture should be found automatically. Neural network structure search and regularization methods can potentially solve the problem, but there is a lack of pointed research.

E. Pretrained models for SSRSI

Applying pretrained models to semantic segmentation is an important way to alleviate the problem of missing large-scale labeled data and has been the most important direction in remote sensing image processing in the past two years. Combining pretrained models with a segmentation network and fine-tuning the model on a small-scale dataset can achieve the highest segmentation performance. Table VI lists the current representative remote sensing pretrained models.

The three elements that make up a foundation model include a large-scale dataset, a pretrained model, and the pretraining method. Since 2021, several million-scale publicly available large-scale datasets have been built in remote sensing, including FMoW [249], SeCo [250], Million-AID [251], Levir-KR [252], GeoPile [253], SSL4EO-S12 [254] and SSL4EO-L [255]. In the early days, pretrained model research typically deals with ResNet50 [250], [252], ViT Large [256], [257], Swin Base [253], [258], ViTAE-B [259], and ViTAEv2-S [260], the number of parameters of them was relatively small compared to those in computer vision. Recently, with the rapid development of vision foundation models, the size of remote sensing pretrained models has expanded to billon-scale [261].

The training methods for pretrained models include supervised learning and self-supervised learning (SSL). Wang et al. [260] investigated different supervised pretrained models on the Million-AID dataset. However, the performance of supervised pretraining depends on the domain difference between source and target data. SSL does not require human annotation in the pretraining process, which can also ensure a small domain gap between the pretraining dataset and the downstream task dataset. Therefore, SSL has become the main method of remote sensing foundation models. Currently, SSL methods in remote sensing mainly include contrastive learning (CS) [250], self-distillation [253], and masked image modeling (MIM) [256], [257], [258], [259], [262], [261]. Contrastive learning constructs positive and negative pairs to learn both invariant and distinguishable visual features [250], [254], [255]. SeCo [250] leveraged the seasonal changes to enforce consistency between positive samples. Self-distillation is learning to predict relationships between multiple views of an unlabeled image. GFM [253] leveraged the teacher-student paradigm for continual pretraining. The masked image modeling randomly masks parts of an image and learns to reconstruct the masked part. Currently, some studies [259], [261] directly utilize MAE for pretraining. However, the random mask strategy in MAE may remove the tokens of critical regions and ignore many small objects. It is inappropriate to directly apply this strategy to self-supervised learning on remote sensing data. Some optimized mask strategies have been proposed, including PIMask [258] and adaptive masking token strategies [262]. Moreover, SatMAE [256] leveraged temporal or multi-spectral information in RS images to improve self-supervised pretraining with MAE. Scale-MAE [257] improved MAE with a GSD-based positional encoding to model scale-specific information.

In computer vision, the training dataset of the largest pretrained model (ViT-22B [77]) has reached 4B images, and the model parameters have reached 22B. However, the largest remote sensing pretrained model (ViT-G12 [261]) still has a significant gap in the size of the dataset and model parameters, so there is still great room for the development of remote sensing pretrained models.

IV. Loss functions and metrics for SSRSI

A. Loss functions

Loss functions, which aim to measure the dissimilarity between the ground truth and the predicted segmentation, play an essential role in SSRSI. The commonly used loss functions can be divided into three categories: distribution-based Loss, region-based Loss, and compound Loss. This section will introduce the loss functions commonly used in SSRSI.

1) Distribution-based loss: The distribution-based loss is a pixel-by-pixel measurement of the distance between the predicted and true values. The most fundamental function in this category is cross entropy, all other functions are derived from cross entropy. For an image with k categories and N pixels, cross-entropy is defined as follows:

$$L_{CE} = -\frac{1}{N}\sum_{n=1}^{N}\sum_{i=0}^{k} y_n^i \log\left(p_n^i\right) \tag{2}$$

where y_n^i is the ground truth binary indicator of class label i of pixel n, and p_n^i is the corresponding predicted segmentation probability.

Obviously, cross entropy calculates losses on individual pixel, and it works best in scenarios with equal data distribution among classes. It often fails to achieve good results due to class imbalance issue. As a result, a number of enhanced loss functions, such as Weighted Cross Entropy (WCE), have been suggested to weight different pixels. WCE is a commonly used extension of CE, which is defined by

$$L_{WCE} = -\frac{1}{N}\sum_{n=1}^{N}\sum_{i=0}^{k} \beta_i y_n^i \log\left(p_n^i\right) \tag{3}$$

where β_i is the weight for class i.

Focal loss is another extension of CE, which adapts WCE to focus on hard examples by reducing the loss assigned to well-classified examples. It can also deal with foreground-background imbalance. It is defined by

$$L_{Focal} = -\frac{1}{N}\sum_{n=1}^{N}\sum_{i=0}^{k} \beta_i (1-p_n^i)^{\gamma} y_n^i \log\left(p_n^i\right) \tag{4}$$

where $(1-p_n^i)^{\gamma}$ is a modulating factor with tunable focusing parameter $\gamma \geq 0$.

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License. For more information, see https://creativecommons.org/licenses/by-nc-nd/4.0/

This article has been accepted for publication in IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing. This is the author's version which has not been fully e
content may change prior to final publication. Citation information: DOI 10.1109/JSTARS.2023.3335891

TABLE VI
SOME REPRESENTATIVE REMOTE SENSING PRETRAINED MODELS FOR SEMANTIC SEGMENTATION.

Category	Method	Pretrained model	Parameters	Pretrained dataset	Pretraining method
Supervised learning	GeoKR [252]	ResNet50	24M	Levir-KR	Geographical knowledge as supervision
	RSP-ViTAEv2-S [260]	ViTAEv2-S	18.8M	Million-AID	Scene classification
Contrastive SSL	SeCo [250]	ResNet50	24M	SeCo	Time-level CS
	Wang et al. [254]	ViT-S	26M	SSL4EO-S12	MoCo v2
	Stewart et al. [255]	ResNet-50	24M	SSL4EO-L	SimCLR v1 and MoCo v2
MIM-based SSL	SatMAE [256]	ViT Large	307M	FMoW	SatMAE
	RingMo [258]	Swin Base	88M	A dataset of 2M images	MIM with PIMask strategy
	RVSA-ViTAE-B [259]	ViTAE-B	89M	Million-AID	MAE
	Scale-MAE [257]	ViT Large	307M	FMoW	Scale-MAE
	AST [262]	AST	463M	A dataset of 1M images	MIM with adaptive masking strategy
	ViT-G12 [261]	ViT-G12	2.4B	Million-AID	MAE
Self-distillation SSL	GFM [253]	Swin Base	88M	GeoPile	Teacher-student paradigm

2) Region-based loss: The goal of region-based loss is to minimize the mismatch or maximize the overlap regions between the segmentation map and the ground truth map. Considering the classifications of other pixels in the image is necessary, which can highlight foreground information and alleviate the problem of class imbalance. Representative loss functions include Dice Loss, IoU (Jaccard) loss, Tversky Loss, and Focal Tversky (fTversky) Loss.

Dice Loss is inspired from Dice Coefficient. Dice loss function is formulated as follows:

$$L_{Dice} = 1 - \frac{2\sum_{n=0}^{N}\sum_{i=0}^{k} y_n^i p_n^i}{\sum_{n=0}^{N}\sum_{i=0}^{k} y_n^i + \sum_{n=0}^{N}\sum_{i=0}^{k} p_n^i} \tag{5}$$

Intersection over Union (IoU) loss, similar to Dice loss. It is defined by

$$L_{IoU} = 1 - \frac{\sum_{n=0}^{N}\sum_{i=0}^{k} y_n^i p_n^i}{\sum_{n=0}^{N}\sum_{i=0}^{k} (y_n^i + p_n^i - y_n^i p_n^i)} \tag{6}$$

To achieve a better trade-off between precision and recall, Tversky loss adapts the Dice loss to emphasize false negatives, it is defined by

$$\begin{aligned} L_{Tversky} = 1 - {}_{(\sum_{n=0}^{N}\sum_{i=0}^{k} y_n^i p_n^i)} / \, (\sum_{n=0}^{N}\sum_{i=0}^{k} y_n^i p_n^i \\ + \alpha \sum_{n=0}^{N}\sum_{i=0}^{k} (1-y_n^i) p_n^i + (1-\alpha)\sum_{n=0}^{N}\sum_{i=0}^{k} y_n^i (1-p_n^i)) \end{aligned} \tag{7}$$

where α is a hyper-parameter that controls the tradeoff between false negatives and false positives.

Similar to the Focal loss, fTversky loss utilizes a γ modifier to leverage hard examples. It is defined as below:

$$L_{fTversky} = (L)^{\gamma} \tag{8}$$

3) Compound loss: The compound loss type is a combination of above two types, thereby leveraging pixel- and region-level losses. We present two typical combinations.

Combo loss is defined as a weighted sum of Dice loss and a cross entropy. It makes an effort to employ cross-entropy for curve smoothing while simultaneously utilizing Dice loss to solve class imbalance problem. It's defined as:

$$L_{Combo} = \alpha L_{CE} + (1-\alpha)\ L_{Dice} \tag{9}$$

To alleviate class imbalance problem and force the model to learn from hard segmentation pixels better, combining Dice loss and focal loss was proposed, it is defined by

$$L_{DiceFocal} = \alpha L_{Dice} + (1-\alpha)\ L_{Focal} \tag{10}$$

It is worth noting that there are other loss functions[263], but they are basic variations or combinations of cross entropy loss and Dice loss. They are either used to solve class imbalance problem, or to improve segmentation performance by paying more attention to hard-to-classify pixels during training. Although many loss functions have been proposed, current research [263], [264] has shown that there is no one that performs better in all situations. It is necessary to choose the appropriate loss function according to our objectives.

B. Evaluation Metrics

Generally speaking, the quality of a model can be evaluated from several perspectives, including quantitative accuracy, training efficiency, memory requirements, etc. This paper mainly introduces the quantitative metrics used to assess the accuracy of SSRSI.

1) Pixel accuracy: Pixel Accuracy (PA), corresponding to Overall Accuracy (OA), indicates the proportion of correctly classified pixels in the image to the total number of pixels. Pixel accuracy can be defined as:

$$PA = \frac{\sum_{i=1}^{K} q^{i,i}}{\sum_{i=1}^{K}\sum_{j=1}^{K} q^{i,j}}, \tag{11}$$

where $q^{i,j}$ represents the number of pixels classified to the semantic category j while the actual semantic category is i, so $q^{i,i}$ represents the number of true positive samples in all pixels with semantic category i.

2) Mean pixel accuracy: Mean Pixel Accuracy (MPA) is an extension of pixel accuracy, representing the average accuracy of each category of pixels. MPA can be defined as:

$$MPA = \frac{1}{K+1}\sum_{i=1}^{K}\frac{q^{i,i}}{\sum_{j=1}^{K} q^{i,j}}, \tag{12}$$

3) Mean Intersection over Union: Intersection over Union (IoU) is defined as the coincidence area between the predicted segmentation map and ground truth, divided by the union area of the predicted segmentation map and ground truth. Mean

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License. For more information, see https://creativecommons.org/licenses/by-nc-nd/4.0/

This article has been accepted for publication in IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing. This is the author's version which has not been fully e
content may change prior to final publication. Citation information: DOI 10.1109/JSTARS.2023.3335891

Intersection over Union (mIoU) refers to the mean value of the Intersection over Union of all categories. Therefore, mIoU can be defined as:

$$mIoU = \frac{1}{K+1}\sum_{i=1}^{K}\frac{q^{i,i}}{\sum_{j=1}^{K} q^{i,j} + \sum_{j=1}^{K} q^{j,i} - q^{i,i}}, \quad (13)$$

4) Precision, Recall, and F1-score: Precision and recall are important indicators for many classical image segmentation models. Similarly, precision and recall can be defined at the category or global level.

$$Precision = \frac{TP}{TP+FP}, \quad (14)$$

$$Recall = \frac{TP}{TP+FN}, \quad (15)$$

where TP represents the true positive part of the image pixel, FP represents the false positive part of the image pixel, and FN represents the false negative part. F1-score is used to measure the accuracy by combining precision and recall, which can be defined as:

$$F1 = \frac{2 * Precision * Recall}{Precision + Recall}, \quad (16)$$

5) Dice coefficient: Dice coefficient is another popular indicator of image segmentation. It can be defined as twice the overlapping area of the predicted segmentation map and the real segmentation map, divided by the union of pixels in the two images. Therefore, Dice coefficient is very similar to IoU:

$$Dice = \frac{2 * |A \cap B|}{|A| + |B|}, \quad (17)$$

where A represents the predicted segmentation map, and B represents the real segmentation map.

V. Datasets for SSRSI

This section investigates the datasets most commonly used for training and testing the SSRSI models based on deep learning (shown in Table VII). Firstly, the representative data sets of several important research fields are summarized, including fully supervised semantic segmentation, semi-supervised and weakly supervised semantic segmentation, UDA semantic segmentation, and multi-modal remote sensing image semantic segmentation. Finally, we summarize several representative datasets in cloud detection and road detection. The overall situation has the following characteristics.

1) Remote sensing image segmentation datasets have developed rapidly in recent years. In terms of data scale, datasets including iSAID [265], SEN12MS [266], Houston 2018 dataset [267], MDAS [268], and 38-Cloud [269] are already in the same order of magnitude as natural image semantic segmentation datasets (such as COCO, ADE20K). On the richness of data types, current datasets cover multiple data types such as multispectral, SAR, hyperspectral, and LiDAR.

2) There are relatively few standard datasets in semi-supervised and weakly supervised research. Semi-supervised research can be carried out directly using fully-supervised datasets, while weakly supervised research must build specialized datasets, standard datasets in this field are still relatively lacking.

3) There aren't many datasets available for UDA research. Early, Some UDA methods have been developed by combining two public datasets. However, directly utilizing combined datasets may result in insufficient common categories and inconsistent annotation granularity. Recently, the LoveDA dataset [270] that distinguished between real urban and rural scenes has facilitated research in the UDA field.

4) There are currently a lot of datasets available in the multimodal data fusion field. Under the guidance of competitions such as the IEEE GRSS data fusion contest, many semantic segmentation datasets fusing various data types have been constructed in recent years.

5) The construction of standard data sets in cloud and road detection is relatively rich. High-resolution datasets covering Landsat, Sentinel, Gaofen, and other satellites have been constructed in cloud detection. It has been developed in road detection for several years, and several standard datasets have also been built. Due to the limited length of the paper, more datasets in these two fields can be referred to [271], [272].

VI. Performance Review

In this section, we will discuss the performance of SSRSI methods in different research fields. It is worth noting that some works are based on non-standard datasets, and do not adequately describe the experimental settings. Most importantly, many publications do not provide source code for their model implementations. It is challenging to compare various methods on a unified standard. Therefore, this survey mainly focuses on analyzing the overall situation of various technology fields rather than comparing a single technology.

A. Performance comparison of supervised SSRSI

Supervised learning methods occupy the majority of remote sensing image semantic segmentation models. Table VIII summarizes the performances of several deep learning-based segmentation models on ISPRS 2D semantic segmentation dataset. We can see that since attention-based methods, especially Transformer-based models, have been widely used in SSRSI, the performance of SSRSI has improved to some extent and achieved the best prediction accuracy. At the same time, the models based on multi-task learning can effectively enhance the model's ability by integrating other related tasks. Figure 6 provide segmentation results for ISPRS 2D semantic segmentation dataset to compare the performance of five methods, including SDFCN, ResUNet-a, ABCNet, RSSFormer, and UNetFormer.

B. Performance comparison of semi-supervised and weakly-supervised SSRSI

There are few semi-supervised and weakly-supervised methods in SSRSI. Table IX and Table X show the performance of several of the semi-supervised and weakly-supervised methods of SSRSI, respectively. We can find that semi-supervised

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License. For more information, see https://creativecommons.org/licenses/by-nc-nd/4.0/

This article has been accepted for publication in IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing. This is the author's version which has not been fully e
content may change prior to final publication. Citation information: DOI 10.1109/JSTARS.2023.3335891

TABLE VII
SOME STANDARD DATASETS OF SSRSI TASK

Category	Dataset	GSD	Year	Sensor	Modality	Semantic Category	Image Size	Quantity
Fully-supervised	ISPRS Vahingen [87]	0.09	2013	Airborne	Multispectral, DSM	6	2000×3000	33
	ISPRS Potsdam [87]	0.05	2013	Airborne	Multispectral, DSM	6	6000×6000	38
	Zurich Summer [273]	0.6	2015	QuickBird	Multispectral	8	1000×1150	20
	Zeebruges [274]	0.05	2015	Airborne	Multispectral, LiDAR	8	10000×10000	7
	DeepGlobe [275]	0.5	2018	WorldView-2	Multispectral	7	2448×2448	1146
	AIRS [276]	0.075	2019	Airborne	Multispectral	2	10000×10000	1047
	iSAID [265]	4	2019	JL-1, GF-2	Multispectral	16	512×512	10468
	LandCoverNet [277]	10	2020	Sentinel-2	Multispectral	7	256×256	9000
	LandCover.ai [278]	0.05-0.25	2020	Airborne	Multispectral	3	9000×9500, 4200×4700	41
	GID [210]	4	2020	GF-2	Multispectral	5	6800×7200	150
	FUSAR-MAP [279]	3	2021	GF-3	SAR	4	1024×1024	610
Semi-supervised	MiniFrance [280]	0.5	2021	Airborne	Multispectral	14	10000×10000	2121
	FloodNet Dataset [281]		2021	Airborne	Multispectral	10	4000×3000	2343
Weakly-supervised	WDCD [282]	8,16	2020	GaoFen-1	Multispectral	2	250×250	206384
	SEN12MS [266]	10	2020	Sentinel-1, Sentinel-2, MODIS	Multispectral, SAR	17	256×256	541986
UDA	Inria Aerial Image Labeling [283]	0.3	2017	Airborne	Multispectral	2	5000×5000	180
	LoveDA [270]	0.3	2021	Spaceborne	Multispectral	7	1024×1024	5987
Multi-modal	Houston 2018 [267]	0.05-1	2018	Airborne	LiDAR, Multispectral, Hyperspectral	20	601×2384	504712
	US3D [284]	0.3	2019	WorldView-3, Airborne	Multispectral, LiDAR	6	1024×1024	2783
	IEEE GRSS data fusion contest 2021	10	2021	Sentinel-1, Sentinel-2, Landsat 8, Suomi NPP	Multispectral, SAR, Infrared	4	800×800	98
	UBC [285]	0.5-2	2023	SuperView-1, Gaofen-2, Gaofen-3, Gaofen-7, WorldView 1 and 2	Multispectral, SAR, DSM	61	600×600	800
	MDAS [268]	0.25-30	2023	Sentinel-1, Sentinel-2, DLR 3K, HySpex, Open street map	SAR, Multispectral, Hyperspectral, DSM, GIS	14	1371×888	108000
Cloud detection	L7_Irish [286]	30	2012	Landsat 7	Multispectral	4	7000×6000	166
	38-Cloud [269]	30	2018	Landsat 8	Multispectral	2	384×384	17601
	WHU Cloud Dataset [287]	30	2021	Landsat 8	Multispectral	3	512×512	859
	KappaSet [288]	10	2021	Sentinel-2	Multispectral	6	512×512	9251
	GF1_WHU [289]	16	2017	Gaofen-1	Multispectral	4	17000×16000	108
	Levir_CS [290]	16	2021	Gaofen-1	Multispectral	2	1320×1200	4168
	AIR-CD [101]	4	2021	Gaofen-2	Multispectral	2	7300×6908	34
	HRC_WHU [291]	0.5-15	2019	Google Earth	Multispectral	2	1280×720	150
Road detection	Massachusetts roads [292]	1	2013	Airborne	Multispectral	3	1500×1500	1171
	SpaceNet [293]	0.3	2018	WorldView-3	Multispectral	2	3000×3000	2517
	DeepGlobe road extraction [294]	0.5	2018	WorldView-3	Multispectral	2	1024×1024	8470
	RoadNet [295]	0.21	2019	Google Earth	Multispectral	3		20

learning methods can improve the performance of semantic segmentation models by integrating unlabeled data. In addition, the mIoU values of all semi-supervised deep semantic segmentation models are less than 80%, indicating that deep semantic segmentation models' performance is still limited in small-scale labeled datasets. The segmentation accuracy of all weakly-supervised semantic segmentation approaches is still far behind that of supervised learning models, indicating that it is still challenging to attain high segmentation accuracy when employing weak supervision signals. In addition, the lack of standard benchmarks and datasets hinders the progress of weakly-supervised semantic segmentation research.

C. Performance comparison of UDA SSRSI

Table XI summarizes the experimental results of some representative UDA methods of SSRSI. We can find that the models trained only with source domain datasets have very low mIoU values in the target domain in most studies, indicating that domain drifts greatly reduce the performance of semantic segmentation models. At the same time, domain adaptation has dramatically improved the performance of all methods, which shows that domain adaptation is necessary for the distribution discrepancy between training sets and test sets. It is worth noting that different methods used different segmentation models. Hence, it isn't easy to compare the performance of UDA methods directly through the results in their papers.

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License. For more information, see https://creativecommons.org/licenses/by-nc-nd/4.0/

This article has been accepted for publication in IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing. This is the author's version which has not been fully e
content may change prior to final publication. Citation information: DOI 10.1109/JSTARS.2023.3335891

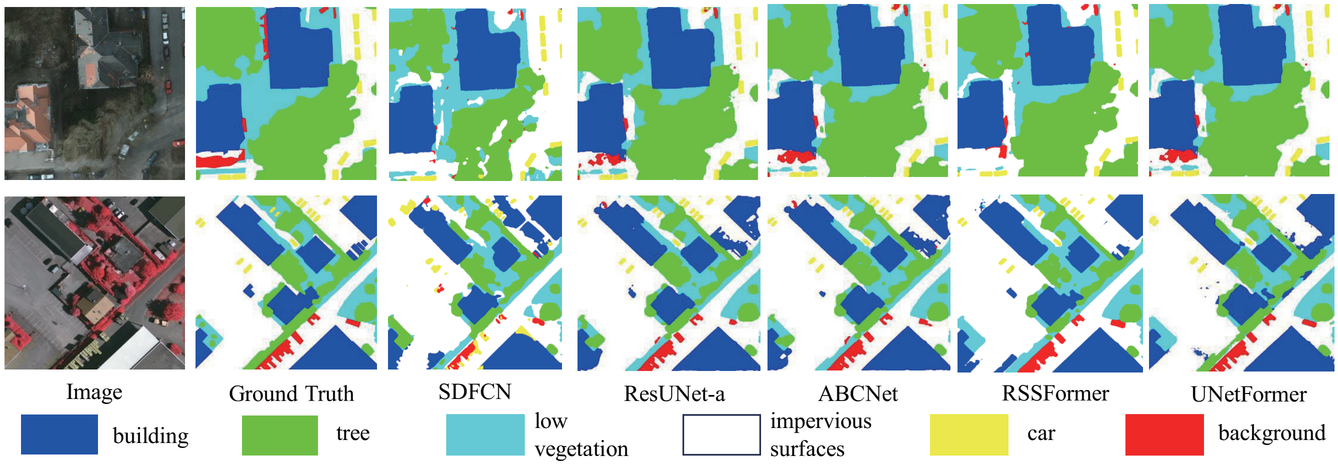

Fig. 6. Visualization of results on the Potsdam (top) and Vaihingen (bottom) datasets.

TABLE VIII
PERFORMANCE OF SOME SEGMENTATION MODELS ON ISPRS 2D SEMANTIC SEGMENTATION DATASET, IN TERMS OF MEAN INTERSECTION OVER UNION (MIOU), AND OVERALL ACCURACY (OA).

Category	Method	Dataset	mIoU(%)	OA(%)
CNN-based models	RiFCN [89]	Potsdam	–	88.3
	ScasNet [111]	Vahingen	83.9	–
		Potsdam	87.78	–
	SDFCN [90]	Vahingen	62.38	87.79
		Potsdam	70.41	86.92
	TreeUNet [91]	Vahingen	–	90.4
		Potsdam	–	90.7
	SBSS-MS [137]	Potsdam	87.68	–
Attention-based DCNNs	S-RA-FCN [113]	Vahingen	–	88.59
		Potsdam	–	89.23
	LANet [94]	Vahingen	–	89.83
		Potsdam	–	90.84
	ABCNet [152]	Vahingen	81.3	90.7
		Potsdam	86.5	91.3
	HMANet [122]	Vahingen	82.87	90.98
		Potsdam	87.28	92.21
	SAPNet [98]	Vahingen	–	89.7
		Potsdam	–	91.8
	RSSFormer [99]	Vahingen	–	90.84
		Potsdam	–	91.25
Transformer-based models	DC-Swin [119]	Vahingen	83.22	91.63
		Potsdam	87.56	92
	UNetFormer [120]	Vahingen	82.7	91
		Potsdam	87.5	92
Multi-Task methods	SDNF [108]	Vahingen	–	92.2
		Potsdam	–	92.6
	ResUNet-a [106]	Potsdam	–	91.5

TABLE IX
PERFORMANCE OF SOME SEMI-SUPERVISED SEMANTIC SEGMENTATION MODELS.

Category	Method	Dataset	Labeled data	mIoU(%)
CR	Zhang et al. [156]	Inria	1/8, Semi	73.26
Self-training	Sun et al. [159]	Vahingen	1/4, Full	76.22
			1/4, Semi	77.53
	Lu et al. [162]	Vahingen	1/4, Semi	65.34
Hybrid method	Wang et al. [163]	Vahingen	1/4, Semi	68.29
			1/4, Semi	77.39
	Wang et al. [165]	Vahingen	1/8, Full	57.91
			1/8, Semi	64.56
	Chen et al. [168]	Massachusetts	1/5, Full	37.59
			1/5, Semi	54.66

TABLE X
PERFORMANCE OF SOME WEAKLY -SUPERVISED SEMANTIC SEGMENTATION MODELS.

Category	Method	Dataset	mIoU(%)
Image-level	Gao et al. [174]	Gaofen-2	80.6
	Fang et al. [175]	Vahingen	77.5
		Potsdam	83.1
Point-level	Lian et al. [178]	Massachusetts	83.1
	Lu et al. [179]	Gaofen	79.1
Scribble-based	Wei et al. [181]	DeepGlobe	57.82

D. Performance comparison of multi-modal data fusion for SSRSI

Table XII shows the performance of the semantic segmentation models integrating the data of different modalities. We can find that all methods have achieved high overall accuracy. In addition, most methods show the experimental results before and after multi-modal data fusion. Through comparison, we can find that the accuracy of semantic segmentation of multispectral images can be improved by fusing auxiliary, LiDAR, and SAR images. However, we also see that fusing SAR images can achieve greater improvement than auxiliary and LiDAR data, possibly due to the higher data diversity between multispectral images and SAR images.

E. Performance comparison of pretrained models for SSRSI

Table XIII shows the performance of some representative remote sensing pretrained models in SSRSI. We can see that all pretrained models based on large-scale remote sensing datasets achieved the highest segmentation accuracy currently, indicating the superiority of the pretrained models. Secondly, ViT-G12 reached the highest OA value, which shows that the large scale of the pretrained models has improved the performance of SSRSI. Finally, most remote sensing pretrained models adopt a universal model architecture and pretrained methods, which outperform most specialized semantic segmentation methods. This demonstrates that pretrained models have broad application prospects in SSRSI.

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License. For more information, see https://creativecommons.org/licenses/by-nc-nd/4.0/

This article has been accepted for publication in IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing. This is the author's version which has not been fully e
content may change prior to final publication. Citation information: DOI 10.1109/JSTARS.2023.3335891

TABLE XI
COMPARISON RESULTS OF SOME REPRESENTATIVE UDA METHODS OF SSRSI.

Category	Method	Source Domain	Target Domain	mIoU(%)	
				Source only	Adaptation
Image-level	Benjdira et al. [188]	Potsdam	Vaihingen	17	30
	Wittich et al. [189]	Potsdam	Vaihingen	60.8	65.9
	Cai et al. [190]	Potsdam	Vaihingen	13.6	42.2
		Vaihingen	Potsdam	10.5	40.8
Feature-level	Wu et al. [201]	LoveDA urban	LoveDA rural	32.02	45.17
		LoveDA rural	LoveDA urban	31.86	46.36
	Lu et al. [204]	SpaceNet	DeepGlobe	17.69	32.05
		DeepGlobe	SpaceNet	21.24	22.9
	Liu et al. [202]	Potsdam	Vaihingen	33	64.3
	Ma et al. [205]	Potsdam	Vaihingen	37.13	63.5
		Vaihingen	Potsdam	12.11	48.42
	Wang et al. [206]	Potsdam	Vaihingen	31.04	53.33
		Vaihingen	Potsdam	28.49	50.94
Output-level	Zheng et al. [207]	Potsdam	Vaihingen	24.01	40.35
		Vaihingen	Potsdam	23.97	37.32
	Chen et al. [208]	Vaihingen	Potsdam	21.78	49.6
	Chen et al. [209]	Potsdam	Vaihingen	25.93	45.91
	Li et al. [211]	Potsdam	Vaihingen	–	63.23
		Vaihingen	Potsdam	–	56.01
	Yan et al. [212]	Potsdam	Vaihingen	29.2	56.8
	Yan et al. [214]	Potsdam	Vaihingen	29.2	60.1
	Zhang et al. [215]	Potsdam	Vaihingen	31.04	52.03
		Vaihingen	Potsdam	28.49	47.87
	Zhu et al. [217]	Potsdam	Vaihingen	43.58	56.05
		Vaihingen	Potsdam	41.22	49.82
Multi-level	Ji et al. [220]	Potsdam	Vaihingen	26.7	43.7
	Li et al. [222]	Potsdam	Vaihingen	25.68	54.34
	Liu et al. [223]	Potsdam	Vaihingen	12.98	43.3
		Vaihingen	Potsdam	12.82	39.61
	Xu et al. [224]	Potsdam	Vaihingen	34.19	51.84
		Vaihingen	Potsdam	26.31	42.85

TABLE XII
COMPARISON RESULTS OF SOME REPRESENTATIVE METHODS OF SSRSI BASED ON MULTI-MODAL DATA FUSION.

Method	What to fuse	Dataset	OA(%)	
			Before fusion	After fusion
Marmanis et al. [227]	RGB+DEM	Vaihingen	–	88.5
Sherrah et al. [228]	RGB+DSM	Potsdam	87.28	87.42
Audebert et al. [230]	IRRG+DSM /NDSM/NDVI	Vaihingen	89.4	89.8
Volpi et al. [231]	RGB+NDVI /NDWI/NDSM	Vaihingen	–	87.83
		Potsdam	–	89.86
Marmanis et al. [123]	RGB+DEM	Vaihingen	89.4	90.3
		Potsdam	–	86.2
Cao et al. [233]	IRRG+DSM	Vaihingen	–	91.5
Zheng et al. [234]	RGB+ DSM/NDVI	Vaihingen	90.2	92
		Potsdam	90.3	92.2
Zhou et al. [235]	RGB+DSM	Vaihingen	76.97	89.27
		Potsdam	83.97	85.16
Liu et al. [239]	RGB+LiDAR	Potsdam	85.5	88.4
Audebert et al. [240]	RGB+LiDAR	Vaihingen	90.2	91.1
		Potsdam	90	90.6
Sun et al. [241]	RGB+LiDAR	Potsdam	80.62	90.65
Li et al. [242]	RGBIR+SAR	PoDelta	87.72	93.61
Ren et al. [244]	RGBIR+SAR	GF-2, GF-3	80.58	89.19
Kang et al. [245]	RGBIR+SAR	GID, GF-3	72.16	79.05
		SpaceNet6	98.31	98.6
He et al. [246]	RGB+DSM	Vaihingen	90.4	92.9
	RGBIR+SAR	MSAW	89.3	93.7

TABLE XIII
SOME REPRESENTATIVE REMOTE SENSING PRETRAINED MODELS FOR SEMANTIC SEGMENTATION

Method	Segmenatation dataset	OA(%)	mIOU(%)
GeoKR+FCN [252]	Potsdam	-	70.4
SatMAE+UperNet [256]	SpaceNet	-	78.51
RingMo+UperNet [258]	Potsdam	91.74	-
RVSA-ViTAE-B+UperNet [259]	Potsdam	91.22	-
Scale-MAE+UperNet [257]	Potsdam	-	78.9
RSP-ViTAEv2-S+UperNet [260]	Potsdam	91.64	-
AST+UperNet [262]	Potsdam	91.72	-
ViT-G12+UperNet [261]	Potsdam	92.58	-

VII. PROBLEMS AND PROSPECTS

SSRSI based on deep learning has made significant progress. However, we also realize that in many real application scenarios, there are still many areas for improvement in the performance and efficiency of various deep semantic segmentation models. Based on the analysis of current technologies, this section discusses several challenging problems and possible research directions.

A. Universal segmentation models

With the outbreak of SAM, building a universal segmentation model that can be applied to all segmentation (semantic, instance, panoramic) tasks and achieve open vocabulary segmentation has become a current research hotspot in computer vision, including SAM [24], SegGPT [86], and SEEM [85] have achieved strong segmentation capabilities and generalization ability. However, the performance of current universal segmentation models is not ideal in some complex and practical scenarios [296], [297], [298]. In many cases, using the tips of fine design is still necessary. Especially in remote sensing, due to the complex background and the presence of a large number of low-contrast, small, and irregular targets, the current universal segmentation models have not achieved ideal results [296]. Moreover, no universal segmentation model is specifically designed for remote sensing image segmentation. There is an urgent need to build universal segmentation models suitable for remote sensing images.

B. Few-shot semantic segmentation

Compared with natural images, the imaging of remote sensing images depends on airborne or spaceborne platforms. The image quality is easily affected by weather, imaging angle, spatial resolution, and other factors. Acquiring large-scale, high-quality remote sensing images is very costly. At the same time, pixel-level semantic annotation requires much time and labor. Transfer learning is an important way to achieve few-shot semantic segmentation in the future. Many remote sensing pretrained models, such as RingMo [258], SatMAE [256], Scale-MAE [257], and ViT-G12 [261] have achieved comparable results with specially designed models in semantic segmentation tasks. Moreover, including semi-supervised, weakly-supervised, and unsupervised semantic

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License. For more information, see https://creativecommons.org/licenses/by-nc-nd/4.0/

This article has been accepted for publication in IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing. This is the author's version which has not been fully e
content may change prior to final publication. Citation information: DOI 10.1109/JSTARS.2023.3335891

segmentation, can somewhat reduce the dependence on labeled data. However, from the results in section V, we can find that most methods, in the case of no or a small amount of labeled data, cannot achieve satisfactory results. It is necessary to improve further the abilities of semi-supervised, weakly-supervised, and unsupervised learning.

C. More efficient segmentation model

The resolution of current remote sensing satellites is constantly improving, the width of most RS images exceeds 10000 pixels. Processing such large-scale images poses a considerable challenge to deep semantic segmentation models. Meantime, semantic segmentation methods based on deep learning often have high computational complexity. For example, the model complexity of SETR [37] has reached 318.3M. It is necessary to develop more efficient segmentation models. One direction is to use lightweight backbone networks for model training and inference. However, lightweight backbone networks have smaller representation capabilities. The other direction is to use pretrained models on a large-scale dataset as the backbone network of the segmentation model. It can reduce the calculation amount of the model in the training stage and improve the segmentation ability. Finally, knowledge distillation can build a large teacher model with high performance and then train a small student model to match the teacher model for model inference, which can also improve the efficiency in the inference stage.

D. Domain generalization of semantic segmentation models

Many UDA semantic segmentation methods have been proposed to solve the problem of domain drift. However, we often face some complex scenes we have never seen before. Generalizing the models trained in several different but related domains to new scenes is a challenging problem, that is, domain generalization [299]. The research on domain generalization is very little in SSRSI. One stream for domain generalization tasks is to learn domain invariant features. The goal is to reduce the representation discrepancy between multiple source domains in a specific feature space and enable the learned model to generalize unseen domains. The second stream is feature disentanglement, which decomposes a feature representation into understandable sub-features.

E. More effective multimodal fusion methods

Currently, the means of earth observation are increasingly diversified. SAR, hyperspectral, and other means are developing rapidly. The semantic segmentation research using multi-source data fusion has advanced significantly. However, the results in section V show that the performance improvement of most deep learning-based segmentation methods using multi-modal data is still limited. One reason is the lack of diversity among multimodal data and the inability to effectively extract complementary information, fusing diverse data sources and adopting adaptive fusion methods is an effective way. On the other hand, due to the inexplicability of neural networks, the complementary information extracted by the deep semantic segmentation models from multi-modal data is often unclear, affecting the further improvement of multimodal data fusion. By tracking the training process of the models, explaining the models through visualization methods can provide directions for further improvement of semantic segmentation models.

F. More abundant application modes

Although remote sensing image segmentation has been applied in land use/land cover mapping, road surface extraction, building extraction, green plastic cover, and other fields, due to the complexity of remote sensing application scenarios, the accuracy and reliability of the current deep semantic segmentation methods are difficult to ensure fully autonomous operation in most real-world applications. To enable the application of the semantic segmentation models in more scenarios, we must constantly enrich the application modes of SSRSI in combination with the real scene particularity and task requirements. For example, to meet the high-reliability requirements of environmental monitoring, we can use the machine for preliminary processing to ensure a high recall. Then, professionals can conduct secondary confirmation.

VIII. Conclusion

Deep learning has made explosive development in the past decade, immensely stimulating the research and application of deep learning in SSRSI. This paper studied the challenges of SSRSI compared with natural image semantic segmentation and reviewed the research status of SSRSI based on deep learning by providing an in-depth analysis of existing methods. Then we described the datasets and metrics commonly used in semantic segmentation and reviewed the quantitative results and experimental performance of some representative models of SSRSI. Finally, we discussed several potential directions for future research. We hope that this survey can provide valuable insights to researchers and inspire more progress in the future.

References

[1] T. Blaschke, S. Lang, E. Lorup, J. Strobl, and P. Zeil, "Object-oriented image processing in an integrated gis/remote sensing environment and perspectives for environmental applications," *Environmental information for planning, politics and the public*, vol. 2, pp. 555–570, 2000.

[2] L. Matikainen and K. Karila, "Segment-based land cover mapping of a suburban area—comparison of high-resolution remotely sensed datasets using classification trees and test field points," *Remote Sensing*, vol. 3, no. 8, pp. 1777–1804, 2011.

[3] Q. Zhang and K. C. Seto, "Mapping urbanization dynamics at regional and global scales using multi-temporal dmsp/ols nighttime light data," *Remote Sensing of Environment*, vol. 115, no. 9, pp. 2320–2329, 2011.

[4] X. Lu, Y. Yuan, and X. Zheng, "Joint dictionary learning for multispectral change detection," *IEEE transactions on cybernetics*, vol. 47, no. 4, pp. 884–897, 2016.

[5] G. Moser, S. B. Serpico, and J. A. Benediktsson, "Land-cover mapping by markov modeling of spatial–contextual information in very-high-resolution remote sensing images," *Proceedings of the IEEE*, vol. 101, no. 3, pp. 631–651, 2012.

[6] N. Kussul, M. Lavreniuk, S. Skakun, and A. Shelestov, "Deep learning classification of land cover and crop types using remote sensing data," *IEEE Geoscience and Remote Sensing Letters*, vol. 14, no. 5, pp. 778–782, 2017.

[7] X. X. Zhu, D. Tuia, L. Mou, G.-S. Xia, L. Zhang, F. Xu, and F. Fraundorfer, "Deep learning in remote sensing: A comprehensive review and list of resources," *IEEE geoscience and remote sensing magazine*, vol. 5, no. 4, pp. 8–36, 2017.

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License. For more information, see https://creativecommons.org/licenses/by-nc-nd/4.0/

This article has been accepted for publication in IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing. This is the author's version which has not been fully e
content may change prior to final publication. Citation information: DOI 10.1109/JSTARS.2023.3335891

[8] L. Zhang, L. Zhang, and B. Du, "Deep learning for remote sensing data: A technical tutorial on the state of the art," *IEEE Geoscience and remote sensing magazine*, vol. 4, no. 2, pp. 22–40, 2016.

[9] L. Ma, Y. Liu, X. Zhang, Y. Ye, G. Yin, and B. A. Johnson, "Deep learning in remote sensing applications: A meta-analysis and review," *ISPRS journal of photogrammetry and remote sensing*, vol. 152, pp. 166–177, 2019.

[10] A. B. Hamida, A. Benoit, P. Lambert, and C. B. Amar, "3-d deep learning approach for remote sensing image classification," *IEEE Transactions on geoscience and remote sensing*, vol. 56, no. 8, pp. 4420–4434, 2018.

[11] K. Li, G. Wan, G. Cheng, L. Meng, and J. Han, "Object detection in optical remote sensing images: A survey and a new benchmark," *ISPRS journal of photogrammetry and remote sensing*, vol. 159, pp. 296–307, 2020.

[12] C. Zhang, P. Yue, D. Tapete, L. Jiang, B. Shangguan, L. Huang, and G. Liu, "A deeply supervised image fusion network for change detection in high resolution bi-temporal remote sensing images," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 166, pp. 183–200, 2020.

[13] Z. Shao and J. Cai, "Remote sensing image fusion with deep convolutional neural network," *IEEE journal of selected topics in applied earth observations and remote sensing*, vol. 11, no. 5, pp. 1656–1669, 2018.

[14] A. Garcia-Garcia, S. Orts-Escolano, S. Oprea, V. Villena-Martinez, and J. Garcia-Rodriguez, "A review on deep learning techniques applied to semantic segmentation," *arXiv preprint arXiv:1704.06857*, 2017.

[15] S. Minaee, Y. Boykov, F. Porikli, A. Plaza, N. Kehtarnavaz, and D. Terzopoulos, "Image segmentation using deep learning: A survey," *IEEE transactions on pattern analysis and machine intelligence*, vol. 44, no. 7, pp. 3523–3542, 2021.

[16] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3431–3440.

[17] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Semantic image segmentation with deep convolutional nets and fully connected crfs," *arXiv preprint arXiv:1412.7062*, 2014.

[18] Chen, Liang-Chieh and Papandreou, George and Kokkinos, Iasonas and Murphy, Kevin and Yuille, Alan L, "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs," *IEEE transactions on pattern analysis and machine intelligence*, vol. 40, no. 4, pp. 834–848, 2017.

[19] R. Kemker, C. Salvaggio, and C. Kanan, "Algorithms for semantic segmentation of multispectral remote sensing imagery using deep learning," *ISPRS journal of photogrammetry and remote sensing*, vol. 145, pp. 60–77, 2018.

[20] M. D. Hossain and D. Chen, "Segmentation for object-based image analysis (obia): A review of algorithms and challenges from remote sensing perspective," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 150, pp. 115–134, 2019.

[21] I. Kotaridis and M. Lazaridou, "Remote sensing image segmentation advances: A meta-analysis," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 173, pp. 309–322, 2021.

[22] X. Yuan, J. Shi, and L. Gu, "A review of deep learning methods for semantic segmentation of remote sensing imagery," *Expert Systems with Applications*, vol. 169, p. 114417, 2021.

[23] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[24] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo *et al.*, "Segment anything," *arXiv preprint arXiv:2304.02643*, 2023.

[25] V. Badrinarayanan, A. Kendall, and R. Cipolla, "Segnet: A deep convolutional encoder-decoder architecture for image segmentation," *IEEE transactions on pattern analysis and machine intelligence*, vol. 39, no. 12, pp. 2481–2495, 2017.

[26] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*. Springer, 2015, pp. 234–241.

[27] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam, "Rethinking atrous convolution for semantic image segmentation," *arXiv preprint arXiv:1706.05587*, 2017.

[28] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 801–818.

[29] M. Yang, K. Yu, C. Zhang, Z. Li, and K. Yang, "Denseaspp for semantic segmentation in street scenes," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 3684–3692.

[30] C. Peng, X. Zhang, G. Yu, G. Luo, and J. Sun, "Large kernel matters–improve semantic segmentation by global convolutional network," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 4353–4361.

[31] X. Ding, X. Zhang, J. Han, and G. Ding, "Scaling up your kernels to 31x31: Revisiting large kernel design in cnns," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 11 963–11 975.

[32] S. Liu, T. Chen, X. Chen, X. Chen, Q. Xiao, B. Wu, M. Pechenizkiy, D. Mocanu, and Z. Wang, "More convnets in the 2020s: Scaling up kernels beyond 51x51 using sparsity," *arXiv preprint arXiv:2207.03620*, 2022.

[33] W. Liu, A. Rabinovich, and A. C. Berg, "Parsenet: Looking wider to see better," *arXiv preprint arXiv:1506.04579*, 2015.

[34] F. Visin, M. Ciccone, A. Romero, K. Kastner, K. Cho, Y. Bengio, M. Matteucci, and A. Courville, "Reseg: A recurrent neural network-based model for semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 2016, pp. 41–48.

[35] W. Byeon, T. M. Breuel, F. Raue, and M. Liwicki, "Scene labeling with lstm recurrent neural networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3547–3555.

[36] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[37] S. Zheng, J. Lu, H. Zhao, X. Zhu, Z. Luo, Y. Wang, Y. Fu, J. Feng, T. Xiang, P. H. Torr *et al.*, "Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 6881–6890.

[38] R. Strudel, R. Garcia, I. Laptev, and C. Schmid, "Segmenter: Transformer for semantic segmentation," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 7262–7272.

[39] E. Xie, W. Wang, Z. Yu, A. Anandkumar, J. M. Alvarez, and P. Luo, "Segformer: Simple and efficient design for semantic segmentation with transformers," *Advances in Neural Information Processing Systems*, vol. 34, pp. 12 077–12 090, 2021.

[40] W. Wang, E. Xie, X. Li, D.-P. Fan, K. Song, D. Liang, T. Lu, P. Luo, and L. Shao, "Pyramid vision transformer: A versatile backbone for dense prediction without convolutions," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 568–578.

[41] W. Wang, W. Chen, Q. Qiu, L. Chen, B. Wu, B. Lin, X. He, and W. Liu, "Crossformer++: A versatile vision transformer hinging on cross-scale attention," *arXiv preprint arXiv:2303.06908*, 2023.

[42] J. Gu, H. Kwon, D. Wang, W. Ye, M. Li, Y.-H. Chen, L. Lai, V. Chandra, and D. Z. Pan, "Multi-scale high-resolution vision transformer for semantic segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 12 094–12 103.

[43] A. Hatamizadeh, Y. Tang, V. Nath, D. Yang, A. Myronenko, B. Landman, H. R. Roth, and D. Xu, "Unetr: Transformers for 3d medical image segmentation," in *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, 2022, pp. 574–584.

[44] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 10 012–10 022.

[45] X. Chu, Z. Tian, Y. Wang, B. Zhang, H. Ren, X. Wei, H. Xia, and C. Shen, "Twins: Revisiting the design of spatial attention in vision transformers," *Advances in Neural Information Processing Systems*, vol. 34, pp. 9355–9366, 2021.

[46] X. Dong, J. Bao, D. Chen, W. Zhang, N. Yu, L. Yuan, D. Chen, and B. Guo, "Cswin transformer: A general vision transformer backbone with cross-shaped windows," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 12 124–12 134.

[47] A. Hassani, S. Walton, J. Li, S. Li, and H. Shi, "Neighborhood attention transformer," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 6185–6194.

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License. For more information, see https://creativecommons.org/licenses/by-nc-nd/4.0/

This article has been accepted for publication in IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing. This is the author's version which has not been fully e
content may change prior to final publication. Citation information: DOI 10.1109/JSTARS.2023.3335891

[48] A. Hassani and H. Shi, "Dilated neighborhood attention transformer," *arXiv preprint arXiv:2209.15001*, 2022.

[49] L.-C. Chen, Y. Yang, J. Wang, W. Xu, and A. L. Yuille, "Attention to scale: Scale-aware semantic image segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 3640–3649.

[50] D. Eigen and R. Fergus, "Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 2650–2658.

[51] H. Ding, X. Jiang, B. Shuai, A. Q. Liu, and G. Wang, "Context contrasted feature and gated multi-scale aggregation for scene segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 2393–2402.

[52] G. Lin, A. Milan, C. Shen, and I. Reid, "Refinenet: Multi-path refinement networks for high-resolution semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1925–1934.

[53] K. Sun, B. Xiao, D. Liu, and J. Wang, "Deep high-resolution representation learning for human pose estimation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 5693–5703.

[54] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 2881–2890.

[55] T. Xiao, Y. Liu, B. Zhou, Y. Jiang, and J. Sun, "Unified perceptual parsing for scene understanding," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 418–434.

[56] M.-H. Guo, C.-Z. Lu, Q. Hou, Z. Liu, M.-M. Cheng, and S.-M. Hu, "Segnext: Rethinking convolutional attention design for semantic segmentation," *arXiv preprint arXiv:2209.08575*, 2022.

[57] Z. Qin, P. Zhang, F. Wu, and X. Li, "Fcanet: Frequency channel attention networks," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 783–792.

[58] Q. Wang, B. Wu, P. Zhu, P. Li, W. Zuo, and Q. Hu, "Eca-net: Efficient channel attention for deep convolutional neural networks," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 11 534–11 542.

[59] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7794–7803.

[60] J. Dai, H. Qi, Y. Xiong, Y. Li, G. Zhang, H. Hu, and Y. Wei, "Deformable convolutional networks," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 764–773.

[61] J. Fu, J. Liu, H. Tian, Y. Li, Y. Bao, Z. Fang, and H. Lu, "Dual attention network for scene segmentation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 3146–3154.

[62] M.-H. Guo, C.-Z. Lu, Z.-N. Liu, M.-M. Cheng, and S.-M. Hu, "Visual attention network," *arXiv preprint arXiv:2202.09741*, 2022.

[63] H. Li, P. Xiong, J. An, and L. Wang, "Pyramid attention network for semantic segmentation," *arXiv preprint arXiv:1805.10180*, 2018.

[64] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Communications of the ACM*, vol. 60, no. 6, pp. 84–90, 2017.

[65] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[66] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[67] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, "Mobilenets: Efficient convolutional neural networks for mobile vision applications," *arXiv preprint arXiv:1704.04861*, 2017.

[68] X. Zhang, X. Zhou, M. Lin, and J. Sun, "Shufflenet: An extremely efficient convolutional neural network for mobile devices," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 6848–6856.

[69] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, and S. Xie, "A convnet for the 2020s," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 11 976–11 986.

[70] Q. Hou, C.-Z. Lu, M.-M. Cheng, and J. Feng, "Conv2former: A simple transformer-style convnet for visual recognition," *arXiv preprint arXiv:2211.11943*, 2022.

[71] W. Wang, J. Dai, Z. Chen, Z. Huang, Z. Li, X. Zhu, X. Hu, T. Lu, L. Lu, H. Li *et al.*, "Internimage: Exploring large-scale vision foundation models with deformable convolutions," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 14 408–14 419.

[72] S. Woo, S. Debnath, R. Hu, X. Chen, Z. Liu, I. S. Kweon, and S. Xie, "Convnext v2: Co-designing and scaling convnets with masked autoencoders," *arXiv preprint arXiv:2301.00808*, 2023.

[73] H. Bao, L. Dong, S. Piao, and F. Wei, "Beit: Bert pre-training of image transformers," *arXiv preprint arXiv:2106.08254*, 2021.

[74] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick, "Masked autoencoders are scalable vision learners," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 16 000–16 009.

[75] Z. Chen, Y. Duan, W. Wang, J. He, T. Lu, J. Dai, and Y. Qiao, "Vision transformer adapter for dense predictions," *arXiv preprint arXiv:2205.08534*, 2022.

[76] W. Yu, M. Luo, P. Zhou, C. Si, Y. Zhou, X. Wang, J. Feng, and S. Yan, "Metaformer is actually what you need for vision," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 10 819–10 829.

[77] M. Dehghani, J. Djolonga, B. Mustafa, P. Padlewski, J. Heek, J. Gilmer, A. Steiner, M. Caron, R. Geirhos, I. Alabdulmohsin *et al.*, "Scaling vision transformers to 22 billion parameters," *arXiv preprint arXiv:2302.05442*, 2023.

[78] M. Oquab, T. Darcet, T. Moutakanni, H. Vo, M. Szafraniec, V. Khalidov, P. Fernandez, D. Haziza, F. Massa, A. El-Nouby *et al.*, "Dinov2: Learning robust visual features without supervision," *arXiv preprint arXiv:2304.07193*, 2023.

[79] M. Oquab, L. Bottou, I. Laptev, and J. Sivic, "Learning and transferring mid-level image representations using convolutional neural networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 1717–1724.

[80] B. Cheng, A. Schwing, and A. Kirillov, "Per-pixel classification is not all you need for semantic segmentation," *Advances in Neural Information Processing Systems*, vol. 34, pp. 17 864–17 875, 2021.

[81] B. Cheng, I. Misra, A. G. Schwing, A. Kirillov, and R. Girdhar, "Masked-attention mask transformer for universal image segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 1290–1299.

[82] W. Zhang, J. Pang, K. Chen, and C. C. Loy, "K-net: Towards unified image segmentation," *Advances in Neural Information Processing Systems*, vol. 34, pp. 10 326–10 338, 2021.

[83] J. Jain, J. Li, M. T. Chiu, A. Hassani, N. Orlov, and H. Shi, "Oneformer: One transformer to rule universal image segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 2989–2998.

[84] J. Qin, J. Wu, P. Yan, M. Li, R. Yuxi, X. Xiao, Y. Wang, R. Wang, S. Wen, X. Pan *et al.*, "Freeseg: Unified, universal and open-vocabulary image segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 19 446–19 455.

[85] X. Zou, J. Yang, H. Zhang, F. Li, L. Li, J. Gao, and Y. J. Lee, "Segment everything everywhere all at once," *arXiv preprint arXiv:2304.06718*, 2023.

[86] X. Wang, X. Zhang, Y. Cao, W. Wang, C. Shen, and T. Huang, "Seggpt: Segmenting everything in context," *arXiv preprint arXiv:2304.03284*, 2023.

[87] ISPRS 2D Semantic Labeling Contest. [Online]. Available: https://www.isprs.org/education/benchmarks/UrbanSemLab/semantic-labeling.aspx.

[88] Z. Zheng, X. Zhang, P. Xiao, and Z. Li, "Integrating gate and attention modules for high-resolution image semantic segmentation," *Ieee Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 14, pp. 4530–4546, 2021.

[89] L. Mou and X. X. Zhu, "Rifcn: Recurrent network in fully convolutional network for semantic segmentation of high resolution remote sensing images," *arXiv preprint arXiv:1805.02091*, 2018.

[90] G. Chen, X. Zhang, Q. Wang, F. Dai, Y. Gong, and K. Zhu, "Symmetrical dense-shortcut deep fully convolutional networks for semantic segmentation of very-high-resolution remote sensing images," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 11, no. 5, pp. 1633–1644, 2018.

[91] K. Yue, L. Yang, R. Li, W. Hu, F. Zhang, and W. Li, "Treeunet: Adaptive tree convolutional neural networks for subdecimeter aerial image segmentation," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 156, pp. 1–13, 2019.

[92] R. Li, S. Zheng, C. Duan, J. Su, and C. Zhang, "Multistage attention resu-net for semantic segmentation of fine-resolution remote sensing images," *IEEE Geoscience and Remote Sensing Letters*, vol. 19, pp. 1–5, 2021.

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License. For more information, see https://creativecommons.org/licenses/by-nc-nd/4.0/

This article has been accepted for publication in IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing. This is the author's version which has not been fully e
content may change prior to final publication. Citation information: DOI 10.1109/JSTARS.2023.3335891

[93] H.-F. Zhong, Q. Sun, H.-M. Sun, and R.-S. Jia, "Nt-net: A semantic segmentation network for extracting lake water bodies from optical remote sensing images based on transformer," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–13, 2022.

[94] L. Ding, H. Tang, and L. Bruzzone, "Lanet: Local attention embedding to improve the semantic segmentation of remote sensing images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 59, no. 1, pp. 426–435, 2020.

[95] S. Xiang, Q. Xie, and M. Wang, "Semantic segmentation for remote sensing images based on adaptive feature selection network," *IEEE Geoscience and Remote Sensing Letters*, vol. 19, pp. 1–5, 2021.

[96] R. Li, S. Zheng, C. Zhang, C. Duan, J. Su, L. Wang, and P. M. Atkinson, "Multiattention network for semantic segmentation of fine-resolution remote sensing images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–13, 2021.

[97] Q. Zhao, J. Liu, Y. Li, and H. Zhang, "Semantic segmentation with attention mechanism for remote sensing images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–13, 2021.

[98] X. Li, F. Xu, F. Liu, X. Lyu, Y. Tong, Z. Xu, and J. Zhou, "A synergistical attention model for semantic segmentation of remote sensing images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 61, pp. 1–16, 2023.

[99] R. Xu, C. Wang, J. Zhang, S. Xu, W. Meng, and X. Zhang, "Rssformer: Foreground saliency enhancement for remote sensing land-cover segmentation," *IEEE Transactions on Image Processing*, vol. 32, pp. 1052–1064, 2023.

[100] X. Sun, M. Xia, and T. Dai, "Controllable fused semantic segmentation with adaptive edge loss for remote sensing parsing," *Remote Sensing*, vol. 14, no. 1, p. 207, 2022.

[101] Q. He, X. Sun, Z. Yan, and K. Fu, "Dabnet: Deformable contextual and boundary-weighted network for cloud detection in remote sensing images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–16, 2021.

[102] X. He, Y. Zhou, J. Zhao, D. Zhang, R. Yao, and Y. Xue, "Swin transformer embedding unet for remote sensing image semantic segmentation," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–15, 2022.

[103] B. Yu, L. Yang, and F. Chen, "Semantic segmentation for high spatial resolution remote sensing images based on convolution neural network and pyramid pooling module," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 11, no. 9, pp. 3252–3261, 2018.

[104] J. Zhang, S. Lin, L. Ding, and L. Bruzzone, "Multi-scale context aggregation for semantic segmentation of remote sensing images," *Remote Sensing*, vol. 12, no. 4, p. 701, 2020.

[105] R. Niu, X. Sun, Y. Tian, W. Diao, Y. Feng, and K. Fu, "Improving semantic segmentation in aerial imagery via graph reasoning and disentangled learning," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–18, 2021.

[106] F. I. Diakogiannis, F. Waldner, P. Caccetta, and C. Wu, "Resunet-a: A deep learning framework for semantic segmentation of remotely sensed data," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 162, pp. 94–114, 2020.

[107] F. Waldner and F. I. Diakogiannis, "Deep learning on edge: Extracting field boundaries from satellite images with a convolutional neural network," *Remote sensing of environment*, vol. 245, p. 111741, 2020.

[108] L. Mi and Z. Chen, "Superpixel-enhanced deep neural forest for remote sensing image semantic segmentation," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 159, pp. 140–152, 2020.

[109] G. Deng, Z. Wu, C. Wang, M. Xu, and Y. Zhong, "Ccanet: Class-constraint coarse-to-fine attentional deep network for subdecimeter aerial image semantic segmentation," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–20, 2021.

[110] Y. Su, J. Cheng, H. Bai, H. Liu, and C. He, "Semantic segmentation of very-high-resolution remote sensing images via deep multi-feature learning," *Remote Sensing*, vol. 14, no. 3, p. 533, 2022.

[111] Y. Liu, B. Fan, L. Wang, J. Bai, S. Xiang, and C. Pan, "Semantic labeling in very high resolution images via a self-cascaded convolutional neural network," *ISPRS journal of photogrammetry and remote sensing*, vol. 145, pp. 78–95, 2018.

[112] L. Mou, Y. Hua, and X. X. Zhu, "A relation-augmented fully convolutional network for semantic segmentation in aerial scenes," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 12 416–12 425.

[113] ——, "Relation matters: Relational context-aware fully convolutional network for semantic segmentation of high-resolution aerial images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 58, no. 11, pp. 7557–7569, 2020.

[114] H. Li, K. Qiu, L. Chen, X. Mei, L. Hong, and C. Tao, "Scattnet: Semantic segmentation network with spatial and channel attention mechanism for high-resolution remote sensing images," *IEEE Geoscience and Remote Sensing Letters*, vol. 18, no. 5, pp. 905–909, 2020.

[115] X. Cheng, X. He, M. Qiao, P. Li, S. Hu, P. Chang, and Z. Tian, "Enhanced contextual representation with deep neural networks for land cover classification based on remote sensing images," *International Journal of Applied Earth Observation and Geoinformation*, vol. 107, p. 102706, 2022.

[116] Z. Xu, W. Zhang, T. Zhang, Z. Yang, and J. Li, "Efficient transformer for remote sensing image segmentation," *Remote Sensing*, vol. 13, no. 18, p. 3585, 2021.

[117] L. Gao, H. Liu, M. Yang, L. Chen, Y. Wan, Z. Xiao, and Y. Qian, "Stransfuse: Fusing swin transformer and convolutional neural network for remote sensing image semantic segmentation," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 14, pp. 10 990–11 003, 2021.

[118] C. Zhang, W. Jiang, Y. Zhang, W. Wang, Q. Zhao, and C. Wang, "Transformer and cnn hybrid deep neural network for semantic segmentation of very-high-resolution remote sensing imagery," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–20, 2022.

[119] L. Wang, R. Li, C. Duan, C. Zhang, X. Meng, and S. Fang, "A novel transformer based semantic segmentation scheme for fine-resolution remote sensing images," *IEEE Geoscience and Remote Sensing Letters*, vol. 19, pp. 1–5, 2022.

[120] L. Wang, R. Li, C. Zhang, S. Fang, C. Duan, X. Meng, and P. M. Atkinson, "Unetformer: A unet-like transformer for efficient semantic segmentation of remote sensing urban scene imagery," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 190, pp. 196–214, 2022.

[121] Z. Zhang, X. Huang, and J. Li, "Dwin-hrformer: A high-resolution transformer model with directional windows for semantic segmentation of urban construction land," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 61, pp. 1–14, 2023.

[122] R. Niu, X. Sun, Y. Tian, W. Diao, K. Chen, and K. Fu, "Hybrid multiple attention network for semantic segmentation in aerial images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–18, 2021.

[123] D. Marmanis, K. Schindler, J. D. Wegner, S. Galliani, M. Datcu, and U. Stilla, "Classification with an edge: Improving semantic image segmentation with boundary detection," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 135, pp. 158–172, 2018.

[124] Z. Xu, W. Zhang, T. Zhang, and J. Li, "Hrcnet: High-resolution context extraction network for semantic segmentation of remote sensing images," *Remote Sensing*, vol. 13, no. 1, p. 71, 2020.

[125] Y. Chong, X. Chen, and S. Pan, "Context union edge network for semantic segmentation of small-scale objects in very high resolution remote sensing images," *IEEE Geoscience and Remote Sensing Letters*, vol. 19, pp. 1–5, 2020.

[126] B. Sui, Y. Cao, X. Bai, S. Zhang, and R. Wu, "Bibed-seg: Block-in-block edge detection network for guiding semantic segmentation task of high-resolution remote sensing images," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 2023.

[127] A. Li, L. Jiao, H. Zhu, L. Li, and F. Liu, "Multitask semantic boundary awareness network for remote sensing image segmentation," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–14, 2021.

[128] M. Papadomanolaki, M. Vakalopoulou, and K. Karantzalos, "A deep multitask learning framework coupling semantic segmentation and fully convolutional lstm networks for urban change detection," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 59, no. 9, pp. 7651–7668, 2021.

[129] Q. Zhang, G. Yang, and G. Zhang, "Collaborative network for super-resolution and semantic segmentation of remote sensing images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–12, 2021.

[130] X. Zheng, Q. Ma, L. Huan, X. Xie, H. Xiong, and J. Gong, "Semantic-aware region loss for land-cover classification," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 2023.

[131] W. Liu, W. Zhang, X. Sun, Z. Guo, and K. Fu, "Hecr-net: Height-embedding context reassembly network for semantic segmentation in aerial images," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 14, pp. 9117–9131, 2021.

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License. For more information, see https://creativecommons.org/licenses/by-nc-nd/4.0/

This article has been accepted for publication in IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing. This is the author's version which has not been fully e
content may change prior to final publication. Citation information: DOI 10.1109/JSTARS.2023.3335891

[132] R. Liu, L. Mi, and Z. Chen, "Afnet: Adaptive fusion network for remote sensing image semantic segmentation," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 59, no. 9, pp. 7871–7886, 2020.

[133] C. Peng, K. Zhang, Y. Ma, and J. Ma, "Cross fusion net: A fast semantic segmentation network for small-scale semantic information capturing in aerial scenes," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–13, 2021.

[134] R. Hang, P. Yang, F. Zhou, and Q. Liu, "Multiscale progressive segmentation network for high-resolution remote sensing imagery," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–12, 2022.

[135] Z. Zheng, Y. Zhong, J. Wang, and A. Ma, "Foreground-aware relation network for geospatial object segmentation in high spatial resolution remote sensing imagery," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 4096–4105.

[136] A. Ma, J. Wang, Y. Zhong, and Z. Zheng, "Factseg: Foreground activation-driven small object semantic segmentation in large-scale remote sensing imagery," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–16, 2021.

[137] Y. Cai, L. Fan, and Y. Fang, "Sbss: Stacking-based semantic segmentation framework for very high-resolution remote sensing image," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 61, pp. 1–14, 2023.

[138] H. Bai, J. Cheng, Y. Su, S. Liu, and X. Liu, "Calibrated focal loss for semantic labeling of high-resolution remote sensing images," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 15, pp. 6531–6547, 2022.

[139] S. Zhang, Z. Li, S. Yan, X. He, and J. Sun, "Distribution alignment: A unified framework for long-tail visual recognition," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 2361–2370.

[140] J. Cui, Y. Yuan, Z. Zhong, Z. Tian, H. Hu, S. Lin, and J. Jia, "Region rebalance for long-tailed semantic segmentation," *arXiv preprint arXiv:2204.01969*, 2022.

[141] Z. Zhong, J. Cui, Y. Yang, X. Wu, X. Qi, X. Zhang, and J. Jia, "Understanding imbalanced semantic segmentation through neural collapse," *arXiv preprint arXiv:2301.01100*, 2023.

[142] W. Chen, Z. Jiang, Z. Wang, K. Cui, and X. Qian, "Collaborative global-local networks for memory-efficient segmentation of ultra-high resolution images," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 8924–8933.

[143] T. Wu, Z. Lei, B. Lin, C. Li, Y. Qu, and Y. Xie, "Patch proposal network for fast semantic segmentation of high-resolution images," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 07, 2020, pp. 12 402–12 409.

[144] Q. Li, W. Yang, W. Liu, Y. Yu, and S. He, "From contexts to locality: Ultra-high resolution image segmentation via locality-aware contextual correlation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 7252–7261.

[145] K. Nogueira, M. Dalla Mura, J. Chanussot, W. R. Schwartz, and J. A. Dos Santos, "Dynamic multicontext segmentation of remote sensing images based on convolutional networks," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 57, no. 10, pp. 7503–7520, 2019.

[146] L. Ding, J. Zhang, and L. Bruzzone, "Semantic segmentation of large-size vhr remote sensing images using a two-stage multiscale training architecture," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 58, no. 8, pp. 5367–5376, 2020.

[147] L. Ding, D. Lin, S. Lin, J. Zhang, X. Cui, Y. Wang, H. Tang, and L. Bruzzone, "Looking outside the window: Wide-context transformer for the semantic segmentation of high-resolution remote sensing images. arxiv 2021," *arXiv preprint arXiv:2106.15754*.

[148] H. K. Cheng, J. Chung, Y.-W. Tai, and C.-K. Tang, "Cascadepsp: Toward class-agnostic and very high-resolution segmentation via global and local refinement," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 8890–8899.

[149] C. Huynh, A. T. Tran, K. Luu, and M. Hoai, "Progressive semantic segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 16 755–16 764.

[150] S. Guo, L. Liu, Z. Gan, Y. Wang, W. Zhang, C. Wang, G. Jiang, W. Zhang, R. Yi, L. Ma *et al.*, "Isdnet: Integrating shallow and deep networks for efficient ultra-high resolution segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 4361–4370.

[151] W. Chen, Y. Li, B. Dang, and Y. Zhang, "Elegantseg: End-to-end holistic learning for extra-large image semantic segmentation," *arXiv preprint arXiv:2211.11316*, 2022.

[152] R. Li, S. Zheng, C. Zhang, C. Duan, L. Wang, and P. M. Atkinson, "Abcnet: Attentive bilateral contextual network for efficient semantic segmentation of fine-resolution remotely sensed imagery," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 181, pp. 84–98, 2021.

[153] R. Anand, K. G. Mehrotra, C. K. Mohan, and S. Ranka, "An improved algorithm for neural network classification of imbalanced training sets," *IEEE Transactions on Neural Networks*, vol. 4, no. 6, pp. 962–969, 1993.

[154] M. Buda, A. Maki, and M. A. Mazurowski, "A systematic study of the class imbalance problem in convolutional neural networks," *Neural networks*, vol. 106, pp. 249–259, 2018.

[155] J. Kang, Z. Wang, R. Zhu, X. Sun, R. Fernandez-Beltran, and A. Plaza, "Picoco: Pixelwise contrast and consistency learning for semisupervised building footprint segmentation," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 14, pp. 10 548–10 559, 2021.

[156] B. Zhang, Y. Zhang, Y. Li, Y. Wan, H. Guo, Z. Zheng, and K. Yang, "Semi-supervised deep learning via transformation consistency regularization for remote sensing image semantic segmentation," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 2022.

[157] Y. Ouali, C. Hudelot, and M. Tami, "Semi-supervised semantic segmentation with cross-consistency training," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 12 674–12 684.

[158] X. Chen, Y. Yuan, G. Zeng, and J. Wang, "Semi-supervised semantic segmentation with cross pseudo supervision," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 2613–2622.

[159] X. Sun, A. Shi, H. Huang, and H. Mayer, "Bas ˆ{4} net: Boundary-aware semi-supervised semantic segmentation network for very high resolution remote sensing images," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 13, pp. 5398–5413, 2020.

[160] Y. Zheng, M. Yang, M. Wang, X. Qian, R. Yang, X. Zhang, and W. Dong, "Semi-supervised adversarial semantic segmentation network using transformer and multiscale convolution for high-resolution remote sensing imagery," *Remote Sensing*, vol. 14, no. 8, p. 1786, 2022.

[161] D. Kwon and S. Kwak, "Semi-supervised semantic segmentation with error localization network," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 9957–9967.

[162] X. Lu, L. Jiao, F. Liu, S. Yang, X. Liu, Z. Feng, L. Li, and P. Chen, "Simple and efficient: A semisupervised learning framework for remote sensing image semantic segmentation," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–16, 2022.

[163] J. Wang, C. HQ Ding, S. Chen, C. He, and B. Luo, "Semi-supervised remote sensing image semantic segmentation via consistency regularization and average update of pseudo-label," *Remote Sensing*, vol. 12, no. 21, p. 3603, 2020.

[164] J. Li, B. Sun, S. Li, and X. Kang, "Semisupervised semantic segmentation of remote sensing images with consistency self-training," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–11, 2021.

[165] J.-X. Wang, S.-B. Chen, C. H. Ding, J. Tang, and B. Luo, "Ranpaste: Paste consistency and pseudo label for semisupervised remote sensing image semantic segmentation," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–16, 2021.

[166] Y. Li, T. Shi, Y. Zhang, W. Chen, Z. Wang, and H. Li, "Learning deep semantic segmentation network under multiple weakly-supervised constraints for cross-domain remote sensing image semantic segmentation," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 175, pp. 20–33, 2021.

[167] J. Chen, B. Sun, L. Wang, B. Fang, Y. Chang, Y. Li, J. Zhang, X. Lyu, and G. Chen, "Semi-supervised semantic segmentation framework with pseudo supervisions for land-use/land-cover mapping in coastal areas," *International Journal of Applied Earth Observation and Geoinformation*, vol. 112, p. 102881, 2022.

[168] H. Chen, Z. Li, J. Wu, W. Xiong, and C. Du, "Semiroadexnet: A semi-supervised network for road extraction from remote sensing imagery via adversarial learning," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 198, pp. 169–183, 2023.

[169] M. Tang, K. Georgiou, H. Qi, C. Champion, and M. Bosch, "Semantic segmentation in aerial imagery using multi-level contrastive learning with local consistency," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2023, pp. 3798–3807.

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License. For more information, see https://creativecommons.org/licenses/by-nc-nd/4.0/

This article has been accepted for publication in IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing. This is the author's version which has not been fully e
content may change prior to final publication. Citation information: DOI 10.1109/JSTARS.2023.3335891

[170] L. Yang, W. Zhuo, L. Qi, Y. Shi, and Y. Gao, "St++: Make self-training work better for semi-supervised semantic segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 4268–4277.

[171] Z. Ke, D. Qiu, K. Li, Q. Yan, and R. W. Lau, "Guided collaborative training for pixel-wise semi-supervised learning," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIII 16*. Springer, 2020, pp. 429–445.

[172] J. Ahn and S. Kwak, "Learning pixel-level semantic affinity with image-level supervision for weakly supervised semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 4981–4990.

[173] 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 Journal of Photogrammetry and Remote Sensing*, vol. 192, pp. 244–267, 2022.

[174] Y. Cao and X. Huang, "A coarse-to-fine weakly supervised learning method for green plastic cover segmentation using high-resolution remote sensing images," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 188, pp. 157–176, 2022.

[175] F. Fang, D. Zheng, S. Li, Y. Liu, L. Zeng, J. Zhang, and B. Wan, "Improved pseudomasks generation for weakly supervised building extraction from high-resolution remote sensing imagery," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 15, pp. 1629–1642, 2022.

[176] J. Dai, K. He, and J. Sun, "Boxsup: Exploiting bounding boxes to supervise convolutional networks for semantic segmentation," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1635–1643.

[177] C. Song, Y. Huang, W. Ouyang, and L. Wang, "Box-driven class-wise region masking and filling rate guided loss for weakly supervised semantic segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 3136–3145.

[178] R. Lian and L. Huang, "Weakly supervised road segmentation in high-resolution remote sensing images using point annotations," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–13, 2021.

[179] M. Lu, L. Fang, M. Li, B. Zhang, Y. Zhang, and P. Ghamisi, "Nfanet: A novel method for weakly supervised water extraction from high-resolution remote-sensing imagery," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–14, 2022.

[180] L. Maggiolo, D. Marcos, G. Moser, S. B. Serpico, and D. Tuia, "A semisupervised crf model for cnn-based semantic segmentation with sparse ground truth," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–15, 2021.

[181] Y. Wei and S. Ji, "Scribble-based weakly supervised deep learning for road surface extraction from remote sensing images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–12, 2021.

[182] X. Zeng, T. Wang, Z. Dong, X. Zhang, and Y. Gu, "Superpixel consistency saliency map generation for weakly supervised semantic segmentation of remote sensing images," *IEEE Transactions on Geoscience and Remote Sensing*, 2023.

[183] P. Arbeláez, J. Pont-Tuset, J. T. Barron, F. Marques, and J. Malik, "Multiscale combinatorial grouping," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 328–335.

[184] A. Newell, K. Yang, and J. Deng, "Stacked hourglass networks for human pose estimation," in *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part VIII 14*. Springer, 2016, pp. 483–499.

[185] Y. Chen, G. Zhang, H. Cui, X. Li, S. Hou, J. Ma, Z. Li, H. Li, and H. Wang, "A novel weakly supervised semantic segmentation framework to improve the resolution of land cover product," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 196, pp. 73–92, 2023.

[186] B. Hariharan, P. Arbeláez, R. Girshick, and J. Malik, "Hypercolumns for object segmentation and fine-grained localization," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 447–456.

[187] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2921–2929.

[188] B. Benjdira, Y. Bazi, A. Koubaa, and K. Ouni, "Unsupervised domain adaptation using generative adversarial networks for semantic segmentation of aerial images," *Remote Sensing*, vol. 11, no. 11, p. 1369, 2019.

[189] D. Wittich and F. Rottensteiner, "Appearance based deep domain adaptation for the classification of aerial images," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 180, pp. 82–102, 2021.

[190] Y. Cai, Y. Yang, Y. Shang, Z. Chen, Z. Shen, and J. Yin, "Iterdanet: Iterative intra-domain adaptation for semantic segmentation of remote sensing images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–17, 2022.

[191] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2223–2232.

[192] O. Tasar, S. Happy, Y. Tarabalka, and P. Alliez, "Colormapgan: Unsupervised domain adaptation for semantic segmentation using color mapping generative adversarial networks," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 58, no. 10, pp. 7178–7193, 2020.

[193] D. Peng, H. Guan, Y. Zang, and L. Bruzzone, "Full-level domain adaptation for building extraction in very-high-resolution optical remote-sensing images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–17, 2021.

[194] O. Tasar, A. Giros, Y. Tarabalka, P. Alliez, and S. Clerc, "Daugnet: Unsupervised, multisource, multitarget, and life-long domain adaptation for semantic segmentation of satellite images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 59, no. 2, pp. 1067–1081, 2020.

[195] O. Tasar, Y. Tarabalka, A. Giros, P. Alliez, and S. Clerc, "Standardgan: Multi-source domain adaptation for semantic segmentation of very high resolution satellite images by data standardization," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2020, pp. 192–193.

[196] A. Rozantsev, M. Salzmann, and P. Fua, "Beyond sharing weights for deep domain adaptation," *IEEE transactions on pattern analysis and machine intelligence*, vol. 41, no. 4, pp. 801–814, 2018.

[197] B. Sun and K. Saenko, "Deep coral: Correlation alignment for deep domain adaptation," in *Computer Vision–ECCV 2016 Workshops: Amsterdam, The Netherlands, October 8-10 and 15-16, 2016, Proceedings, Part III 14*. Springer, 2016, pp. 443–450.

[198] G. Kang, L. Jiang, Y. Yang, and A. G. Hauptmann, "Contrastive adaptation network for unsupervised domain adaptation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 4893–4902.

[199] X. Liu, Y. Han, S. Bai, Y. Ge, T. Wang, X. Han, S. Li, J. You, and J. Lu, "Importance-aware semantic segmentation in self-driving with discrete wasserstein training," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 07, 2020, pp. 11 629–11 636.

[200] Z. Zhang, K. Doi, A. Iwasaki, and G. Xu, "Unsupervised domain adaptation of high-resolution aerial images via correlation alignment and self training," *IEEE Geoscience and Remote Sensing Letters*, vol. 18, no. 4, pp. 746–750, 2020.

[201] L. Wu, M. Lu, and L. Fang, "Deep covariance alignment for domain adaptive remote sensing image segmentation," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–11, 2022.

[202] Y. Liu, X. Kang, Y. Huang, K. Wang, and G. Yang, "Unsupervised domain adaptation semantic segmentation for remote-sensing images via covariance attention," *IEEE Geoscience and Remote Sensing Letters*, vol. 19, pp. 1–5, 2022.

[203] J. Iqbal and M. Ali, "Weakly-supervised domain adaptation for built-up region segmentation in aerial and satellite imagery," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 167, pp. 263–275, 2020.

[204] X. Lu, Y. Zhong, Z. Zheng, and J. Wang, "Cross-domain road detection based on global-local adversarial learning framework from very high resolution satellite imagery," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 180, pp. 296–312, 2021.

[205] X. Ma, X. Zhang, Z. Wang, and M.-O. Pun, "Unsupervised domain adaptation augmented by mutually boosted attention for semantic segmentation of vhr remote sensing images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 61, pp. 1–15, 2023.

[206] L. Wang, P. Xiao, X. Zhang, and X. Chen, "A fine-grained unsupervised domain adaptation framework for semantic segmentation of remote sensing images," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 2023.

[207] A. Zheng, M. Wang, C. Li, J. Tang, and B. Luo, "Entropy guided adversarial domain adaptation for aerial image semantic segmentation," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–14, 2021.

[208] X. Chen, S. Pan, and Y. Chong, "Unsupervised domain adaptation for remote sensing image semantic segmentation using region and category

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License. For more information, see https://creativecommons.org/licenses/by-nc-nd/4.0/

This article has been accepted for publication in IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing. This is the author's version which has not been fully e
content may change prior to final publication. Citation information: DOI 10.1109/JSTARS.2023.3335891

adaptive domain discriminator," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–13, 2022.

[209] J. Chen, J. Zhu, Y. Guo, G. Sun, Y. Zhang, and M. Deng, "Unsupervised domain adaptation for semantic segmentation of high-resolution remote sensing imagery driven by category-certainty attention," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–15, 2022.

[210] X.-Y. Tong, G.-S. Xia, Q. Lu, H. Shen, S. Li, S. You, and L. Zhang, "Land-cover classification with high-resolution remote sensing images using transferable deep models," *Remote Sensing of Environment*, vol. 237, p. 111322, 2020.

[211] W. Li, H. Gao, Y. Su, and B. M. Momanyi, "Unsupervised domain adaptation for remote sensing semantic segmentation with transformer," *Remote Sensing*, vol. 14, no. 19, p. 4942, 2022.

[212] L. Yan, B. Fan, H. Liu, C. Huo, S. Xiang, and C. Pan, "Triplet adversarial domain adaptation for pixel-level classification of vhr remote sensing images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 58, no. 5, pp. 3558–3573, 2019.

[213] L. Zhang, M. Lan, J. Zhang, and D. Tao, "Stagewise unsupervised domain adaptation with adversarial self-training for road segmentation of remote-sensing images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–13, 2021.

[214] L. Yan, B. Fan, S. Xiang, and C. Pan, "Cmt: Cross mean teacher unsupervised domain adaptation for vhr image semantic segmentation," *IEEE Geoscience and Remote Sensing Letters*, vol. 19, pp. 1–5, 2021.

[215] B. Zhang, T. Chen, and B. Wang, "Curriculum-style local-to-global adaptation for cross-domain remote sensing image segmentation," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–12, 2021.

[216] X. Yao, Y. Wang, Y. Wu, and Z. Liang, "Weakly-supervised domain adaptation with adversarial entropy for building segmentation in cross-domain aerial imagery," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 14, pp. 8407–8418, 2021.

[217] J. Zhu, Y. Guo, G. Sun, L. Yang, M. Deng, and J. Chen, "Unsupervised domain adaptation semantic segmentation of high-resolution remote sensing imagery with invariant domain-level prototype memory," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 61, pp. 1–18, 2023.

[218] L. Hoyer, D. Dai, and L. Van Gool, "Daformer: Improving network architectures and training strategies for domain-adaptive semantic segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 9924–9935.

[219] A. Ma, C. Zheng, J. Wang, and Y. Zhong, "Domain adaptive land-cover classification via local consistency and global diversity," *IEEE Transactions on Geoscience and Remote Sensing*, 2023.

[220] S. Ji, D. Wang, and M. Luo, "Generative adversarial network-based full-space domain adaptation for land cover classification from multiple-source remote sensing images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 59, no. 5, pp. 3816–3828, 2020.

[221] L. Shi, Z. Wang, B. Pan, and Z. Shi, "An end-to-end network for remote sensing imagery semantic segmentation via joint pixel-and representation-level domain adaptation," *IEEE Geoscience and Remote Sensing Letters*, vol. 18, no. 11, pp. 1896–1900, 2020.

[222] J. Li, S. Zi, R. Song, Y. Li, Y. Hu, and Q. Du, "A stepwise domain adaptive segmentation network with covariate shift alleviation for remote sensing imagery," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–15, 2022.

[223] W. Liu, F. Su, X. Jin, H. Li, and R. Qin, "Bispace domain adaptation network for remotely sensed semantic segmentation," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–11, 2022.

[224] Q. Xu, X. Yuan, and C. Ouyang, "Class-aware domain adaptation for semantic segmentation of remote sensing images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–17, 2022.

[225] J. Chen, G. Chen, B. Fang, J. Wang, and L. Wang, "Class-aware domain adaptation for coastal land cover mapping using optical remote sensing imagery," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 14, pp. 11 800–11 813, 2021.

[226] J. Chen, P. He, J. Zhu, Y. Guo, G. Sun, M. Deng, and H. Li, "Memory-contrastive unsupervised domain adaptation for building extraction of high-resolution remote sensing imagery," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 61, pp. 1–15, 2023.

[227] D. Marmanis, J. D. Wegner, S. Galliani, K. Schindler, M. Datcu, and U. Stilla, "Semantic segmentation of aerial images with an ensemble of cnss," *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences, 2016*, vol. 3, pp. 473–480, 2016.

[228] Sherrah, Jamie, "Fully convolutional networks for dense semantic labelling of high-resolution aerial imagery," *arXiv preprint arXiv:1606.02585*, 2016.

[229] J. Sherrah, "Fully convolutional networks for dense semantic labelling of high-resolution aerial imagery," *arXiv preprint arXiv:1606.02585*, 2016.

[230] N. Audebert, B. Le Saux, and S. Lefèvre, "Semantic segmentation of earth observation data using multimodal and multi-scale deep networks," in *Computer Vision–ACCV 2016: 13th Asian Conference on Computer Vision, Taipei, Taiwan, November 20-24, 2016, Revised Selected Papers, Part I 13*. Springer, 2017, pp. 180–196.

[231] M. Volpi and D. Tuia, "Dense semantic labeling of subdecimeter resolution images with convolutional neural networks," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 55, no. 2, pp. 881–893, 2016.

[232] W. Zhang, H. Huang, M. Schmitz, X. Sun, H. Wang, and H. Mayer, "Effective fusion of multi-modal remote sensing data in a fully convolutional network for semantic labeling," *Remote Sensing*, vol. 10, no. 1, p. 52, 2017.

[233] Z. Cao, K. Fu, X. Lu, W. Diao, H. Sun, M. Yan, H. Yu, and X. Sun, "End-to-end dsm fusion networks for semantic segmentation in high-resolution aerial images," *IEEE Geoscience and Remote Sensing Letters*, vol. 16, no. 11, pp. 1766–1770, 2019.

[234] X. Zheng, X. Wu, L. Huan, W. He, and H. Zhang, "A gather-to-guide network for remote sensing semantic segmentation of rgb and auxiliary image," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–15, 2021.

[235] W. Zhou, J. Jin, J. Lei, and J.-N. Hwang, "Cegfnet: Common extraction and gate fusion network for scene parsing of remote sensing images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–10, 2021.

[236] J. Zhao, D. Zhang, B. Shi, Y. Zhou, J. Chen, R. Yao, and Y. Xue, "Multi-source collaborative enhanced for remote sensing images semantic segmentation," *Neurocomputing*, vol. 493, pp. 76–90, 2022.

[237] X. Liu, Y. Peng, Z. Lu, W. Li, J. Yu, D. Ge, and W. Xiang, "Feature-fusion segmentation network for landslide detection using high-resolution remote sensing images and digital elevation model data," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 61, pp. 1–14, 2023.

[238] S. Zhou, Y. Feng, S. Li, D. Zheng, F. Fang, Y. Liu, and B. Wan, "Dsm-assisted unsupervised domain adaptive network for semantic segmentation of remote sensing imagery," *IEEE Transactions on Geoscience and Remote Sensing*, 2023.

[239] Y. Liu, S. Piramanayagam, S. T. Monteiro, and E. Saber, "Dense semantic labeling of very-high-resolution aerial imagery and lidar with fully-convolutional neural networks and higher-order crfs," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2017, pp. 76–85.

[240] N. Audebert, B. Le Saux, and S. Lefèvre, "Beyond rgb: Very high resolution urban remote sensing with multimodal deep networks," *ISPRS journal of photogrammetry and remote sensing*, vol. 140, pp. 20–32, 2018.

[241] Y. Sun, X. Zhang, Q. Xin, and J. Huang, "Developing a multi-filter convolutional neural network for semantic segmentation using high-resolution aerial imagery and lidar data," *ISPRS journal of photogrammetry and remote sensing*, vol. 143, pp. 3–14, 2018.

[242] X. Li, L. Lei, Y. Sun, M. Li, and G. Kuang, "Multimodal bilinear fusion network with second-order attention-based channel selection for land cover classification," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 13, pp. 1011–1026, 2020.

[243] X. Li, G. Zhang, H. Cui, S. Hou, S. Wang, X. Li, Y. Chen, Z. Li, and L. Zhang, "Mcanet: A joint semantic segmentation framework of optical and sar images for land use classification," *International Journal of Applied Earth Observation and Geoinformation*, vol. 106, p. 102638, 2022.

[244] B. Ren, S. Ma, B. Hou, D. Hong, J. Chanussot, J. Wang, and L. Jiao, "A dual-stream high resolution network: Deep fusion of gf-2 and gf-3 data for land cover classification," *International Journal of Applied Earth Observation and Geoinformation*, vol. 112, p. 102896, 2022.

[245] W. Kang, Y. Xiang, F. Wang, and H. You, "Cfnet: A cross fusion network for joint land cover classification using optical and sar images," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 15, pp. 1562–1574, 2022.

[246] Q. He, X. Sun, W. Diao, Z. Yan, D. Yin, and K. Fu, "Transformer-induced graph reasoning for multimodal semantic segmentation in remote sensing," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 193, pp. 90–103, 2022.

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License. For more information, see https://creativecommons.org/licenses/by-nc-nd/4.0/

This article has been accepted for publication in IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing. This is the author's version which has not been fully e
content may change prior to final publication. Citation information: DOI 10.1109/JSTARS.2023.3335891

[247] D. Hong, J. Yao, D. Meng, Z. Xu, and J. Chanussot, "Multimodal gans: Toward crossmodal hyperspectral–multispectral image segmentation," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 59, no. 6, pp. 5103–5113, 2020.

[248] W. Sun and R. Wang, "Fully convolutional networks for semantic segmentation of very high resolution remotely sensed images combined with dsm," *IEEE Geoscience and Remote Sensing Letters*, vol. 15, no. 3, pp. 474–478, 2018.

[249] G. Christie, N. Fendley, J. Wilson, and R. Mukherjee, "Functional map of the world," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 6172–6180.

[250] O. Manas, A. Lacoste, X. Giró-i Nieto, D. Vazquez, and P. Rodriguez, "Seasonal contrast: Unsupervised pre-training from uncurated remote sensing data," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 9414–9423.

[251] Y. Long, G.-S. Xia, S. Li, W. Yang, M. Y. Yang, X. X. Zhu, L. Zhang, and D. Li, "On creating benchmark dataset for aerial image interpretation: Reviews, guidances, and million-aid," *IEEE Journal of selected topics in applied earth observations and remote sensing*, vol. 14, pp. 4205–4230, 2021.

[252] W. Li, K. Chen, H. Chen, and Z. Shi, "Geographical knowledge-driven representation learning for remote sensing images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–16, 2021.

[253] M. Mendieta, B. Han, X. Shi, Y. Zhu, C. Chen, and M. Li, "Gfm: Building geospatial foundation models via continual pretraining," *arXiv preprint arXiv:2302.04476*, 2023.

[254] Y. Wang, N. A. A. Braham, Z. Xiong, C. Liu, C. M. Albrecht, and X. X. Zhu, "Ssl4eo-s12: A large-scale multi-modal, multi-temporal dataset for self-supervised learning in earth observation," *arXiv preprint arXiv:2211.07044*, 2022.

[255] A. J. Stewart, N. Lehmann, I. A. Corley, Y. Wang, Y.-C. Chang, N. A. A. Braham, S. Sehgal, C. Robinson, and A. Banerjee, "Ssl4eo-l: Datasets and foundation models for landsat imagery," *arXiv preprint arXiv:2306.09424*, 2023.

[256] Y. Cong, S. Khanna, C. Meng, P. Liu, E. Rozi, Y. He, M. Burke, D. Lobell, and S. Ermon, "Satmae: Pre-training transformers for temporal and multi-spectral satellite imagery," *Advances in Neural Information Processing Systems*, vol. 35, pp. 197–211, 2022.

[257] C. J. Reed, R. Gupta, S. Li, S. Brockman, C. Funk, B. Clipp, S. Candido, M. Uyttendaele, and T. Darrell, "Scale-mae: A scale-aware masked autoencoder for multiscale geospatial representation learning," *arXiv preprint arXiv:2212.14532*, 2022.

[258] X. Sun, P. Wang, W. Lu, Z. Zhu, X. Lu, Q. He, J. Li, X. Rong, Z. Yang, H. Chang *et al.*, "Ringmo: A remote sensing foundation model with masked image modeling," *IEEE Transactions on Geoscience and Remote Sensing*, 2022.

[259] D. Wang, Q. Zhang, Y. Xu, J. Zhang, B. Du, D. Tao, and L. Zhang, "Advancing plain vision transformer towards remote sensing foundation model," *IEEE Transactions on Geoscience and Remote Sensing*, 2022.

[260] D. Wang, J. Zhang, B. Du, G.-S. Xia, and D. Tao, "An empirical study of remote sensing pretraining," *IEEE Transactions on Geoscience and Remote Sensing*, 2022.

[261] K. Cha, J. Seo, and T. Lee, "A billion-scale foundation model for remote sensing images," *arXiv preprint arXiv:2304.05215*, 2023.

[262] Q. He, X. Sun, Z. Yan, B. Wang, Z. Zhu, W. Diao, and M. Y. Yang, "Ast: Adaptive self-supervised transformer for optical remote sensing representation," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 200, pp. 41–54, 2023.

[263] S. Jadon, "A survey of loss functions for semantic segmentation," in *2020 IEEE conference on computational intelligence in bioinformatics and computational biology (CIBCB)*. IEEE, 2020, pp. 1–7.

[264] H. Xu, H. He, Y. Zhang, L. Ma, and J. Li, "A comparative study of loss functions for road segmentation in remotely sensed road datasets," *International Journal of Applied Earth Observation and Geoinformation*, vol. 116, p. 103159, 2023.

[265] S. Waqas Zamir, A. Arora, A. Gupta, S. Khan, G. Sun, F. Shahbaz Khan, F. Zhu, L. Shao, G.-S. Xia, and X. Bai, "isaid: A large-scale dataset for instance segmentation in aerial images," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2019, pp. 28–37.

[266] M. Schmitt, L. H. Hughes, C. Qiu, and X. X. Zhu, "Sen12ms–a curated dataset of georeferenced multi-spectral sentinel-1/2 imagery for deep learning and data fusion," *arXiv preprint arXiv:1906.07789*, 2019.

[267] Y. Xu, B. Du, L. Zhang, D. Cerra, M. Pato, E. Carmona, S. Prasad, N. Yokoya, R. Hänsch, and B. Le Saux, "Advanced multi-sensor optical remote sensing for urban land use and land cover classification: Outcome of the 2018 ieee grss data fusion contest," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 12, no. 6, pp. 1709–1724, 2019.

[268] J. Hu, R. Liu, D. Hong, A. Camero, J. Yao, M. Schneider, F. Kurz, K. Segl, and X. X. Zhu, "Mdas: A new multimodal benchmark dataset for remote sensing," *Earth System Science Data*, vol. 15, no. 1, pp. 113–131, 2023.

[269] S. Mohajerani, T. A. Krammer, and P. Saeedi, "Cloud detection algorithm for remote sensing images using fully convolutional neural networks," *arXiv preprint arXiv:1810.05782*, 2018.

[270] J. Wang, Z. Zheng, A. Ma, X. Lu, and Y. Zhong, "Loveda: A remote sensing land-cover dataset for domain adaptive semantic segmentation," *arXiv preprint arXiv:2110.08733*, 2021.

[271] Z. Li, H. Shen, Q. Weng, Y. Zhang, P. Dou, and L. Zhang, "Cloud and cloud shadow detection for optical satellite imagery: Features, algorithms, validation, and prospects," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 188, pp. 89–108, 2022.

[272] Z. Chen, L. Deng, Y. Luo, D. Li, J. M. Junior, W. N. Gonçalves, A. A. M. Nurunnabi, J. Li, C. Wang, and D. Li, "Road extraction in remote sensing data: A survey," *International Journal of Applied Earth Observation and Geoinformation*, vol. 112, p. 102833, 2022.

[273] M. Volpi and V. Ferrari, "Semantic segmentation of urban scenes by learning local class interactions," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2015, pp. 1–9.

[274] M. Campos-Taberner, A. Romero-Soriano, C. Gatta, G. Camps-Valls, A. Lagrange, B. Le Saux, A. Beaupere, A. Boulch, A. Chan-Hon-Tong, S. Herbin *et al.*, "Processing of extremely high-resolution lidar and rgb data: outcome of the 2015 ieee grss data fusion contest–part a: 2-d contest," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 9, no. 12, pp. 5547–5559, 2016.

[275] I. Demir, K. Koperski, D. Lindenbaum, G. Pang, J. Huang, S. Basu, F. Hughes, D. Tuia, and R. D. Raskar, "A challenge to parse the earth through satellite images. arxiv 2018," *arXiv preprint arXiv:1805.06561*, 2018.

[276] Q. Chen, L. Wang, Y. Wu, G. Wu, Z. Guo, and S. Waslander, "Aerial imagery for roof segmentation: A large-scale dataset towards automatic mapping of buildings. arxiv 2018," *arXiv preprint arXiv:1807.09532*, 2018.

[277] H. Alemohammad and K. Booth, "Landcovernet: A global benchmark land cover classification training dataset," *arXiv preprint arXiv:2012.03111*, 2020.

[278] A. Boguszewski, D. Batorski, N. Ziemba-Jankowska, T. Dziedzic, and A. Zambrzycka, "Landcover. ai: Dataset for automatic mapping of buildings, woodlands, water and roads from aerial imagery," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 1102–1110.

[279] X. Shi, S. Fu, J. Chen, F. Wang, and F. Xu, "Object-level semantic segmentation on the high-resolution gaofen-3 fusar-map dataset," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 14, pp. 3107–3119, 2021.

[280] J. Castillo-Navarro, B. Le Saux, A. Boulch, N. Audebert, and S. Lefèvre, "Semi-supervised semantic segmentation in earth observation: The minifrance suite, dataset analysis and multi-task network study," *Machine Learning*, pp. 1–36, 2021.

[281] M. Rahnemoonfar, T. Chowdhury, A. Sarkar, D. Varshney, M. Yari, and R. R. Murphy, "Floodnet: A high resolution aerial imagery dataset for post flood scene understanding," *IEEE Access*, vol. 9, pp. 89 644–89 654, 2021.

[282] Y. Li, W. Chen, Y. Zhang, C. Tao, R. Xiao, and Y. Tan, "Accurate cloud detection in high-resolution remote sensing imagery by weakly supervised deep learning," *Remote Sensing of Environment*, vol. 250, p. 112045, 2020.

[283] E. Maggiori, Y. Tarabalka, G. Charpiat, and P. Alliez, "Can semantic labeling methods generalize to any city? the inria aerial image labeling benchmark," in *2017 IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*. IEEE, 2017, pp. 3226–3229.

[284] M. Bosch, K. Foster, G. Christie, S. Wang, G. D. Hager, and M. Brown, "Semantic stereo for incidental satellite images," in *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 2019, pp. 1524–1532.

[285] X. Huang, L. Ren, C. Liu, Y. Wang, H. Yu, M. Schmitt, R. Hänsch, X. Sun, H. Huang, and H. Mayer, "Urban building classification (ubc)-a dataset for individual building detection and classification from satellite imagery," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 1413–1421.

[286] P. L. Scaramuzza, M. A. Bouchard, and J. L. Dwyer, "Development of the landsat data continuity mission cloud-cover assessment algorithms,"

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License. For more information, see https://creativecommons.org/licenses/by-nc-nd/4.0/

This article has been accepted for publication in IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing. This is the author's version which has not been fully e
content may change prior to final publication. Citation information: DOI 10.1109/JSTARS.2023.3335891

IEEE Transactions on Geoscience and Remote Sensing, vol. 50, no. 4, pp. 1140–1154, 2011.

[287] S. Ji, P. Dai, M. Lu, and Y. Zhang, "Simultaneous cloud detection and removal from bitemporal remote sensing images using cascade convolutional neural networks," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 59, no. 1, pp. 732–748, 2020.

[288] M. Domnich, I. Sünter, H. Trofimov, O. Wold, F. Harun, A. Kostiukhin, M. Järveoja, M. Veske, T. Tamm, K. Voormansik *et al.*, "Kappamask: Ai-based cloudmask processor for sentinel-2," *Remote Sensing*, vol. 13, no. 20, p. 4100, 2021.

[289] Z. Li, H. Shen, H. Li, G. Xia, P. Gamba, and L. Zhang, "Multi-feature combined cloud and cloud shadow detection in gaofen-1 wide field of view imagery," *Remote sensing of environment*, vol. 191, pp. 342–358, 2017.

[290] X. Wu, Z. Shi, and Z. Zou, "A geographic information-driven method and a new large scale dataset for remote sensing cloud/snow detection," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 174, pp. 87–104, 2021.

[291] Z. Li, H. Shen, Q. Cheng, Y. Liu, S. You, and Z. He, "Deep learning based cloud detection for medium and high resolution remote sensing images of different sensors," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 150, pp. 197–212, 2019.

[292] V. Mnih, *Machine learning for aerial image labeling*. University of Toronto (Canada), 2013.

[293] A. Van Etten, D. Lindenbaum, and T. M. Bacastow, "Spacenet: A remote sensing dataset and challenge series," *arXiv preprint arXiv:1807.01232*, 2018.

[294] I. Demir, K. Koperski, D. Lindenbaum, G. Pang, J. Huang, S. Basu, F. Hughes, D. Tuia, and R. Raskar, "Deepglobe 2018: A challenge to parse the earth through satellite images," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2018, pp. 172–181.

[295] Y. Liu, J. Yao, X. Lu, M. Xia, X. Wang, and Y. Liu, "Roadnet: Learning to comprehensively analyze road networks in complex urban scenes from high-resolution remotely sensed images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 57, no. 4, pp. 2043–2056, 2018.

[296] W. Ji, J. Li, Q. Bi, W. Li, and L. Cheng, "Segment anything is not always perfect: An investigation of sam on different real-world applications," *arXiv preprint arXiv:2304.05750*, 2023.

[297] L. Tang, H. Xiao, and B. Li, "Can sam segment anything? when sam meets camouflaged object detection," *arXiv preprint arXiv:2304.04709*, 2023.

[298] G.-P. Ji, D.-P. Fan, P. Xu, M.-M. Cheng, B. Zhou, and L. Van Gool, "Sam struggles in concealed scenes–empirical study on" segment anything"," *arXiv preprint arXiv:2304.06022*, 2023.

[299] J. Wang, C. Lan, C. Liu, Y. Ouyang, T. Qin, W. Lu, Y. Chen, W. Zeng, and P. Yu, "Generalizing to unseen domains: A survey on domain generalization," *IEEE Transactions on Knowledge and Data Engineering*, 2022.

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License. For more information, see https://creativecommons.org/licenses/by-nc-nd/4.0/