Causality-Guided Stepwise Intervention and Reweighting for Remote Sensing Image Semantic Segmentation

Shuting Shi, Baohong Li, Laifu Zhang, Kun Kuang, Sensen Wu, Tian Feng, *Member, IEEE*, Yiming Yan, and Zhenhong Du, *Member, IEEE*

***Abstract*— Semantic segmentation is one of the most significant tasks in remote sensing (RS) image interpretation, which focuses on learning global and local information to infer the semantic label of each pixel. Previous studies devise encoder–decoder structured deep learning (DL) models to extract global and local features from RS images with the help of pretraining knowledge to predict semantic labels. However, due to the common heterogeneity between the data for pretraining and the data to be semantically segmented, these models fail to learn general features appropriate to RS datasets. In this article, we propose a novel formulation of the above problem from a causal perspective, where the learned features from pretrained models result from causality and spurious correlations, and only the former carries general information that remains invariant regardless of the exact task and dataset. Based on the above formulation, we propose stepwise intervention and reweighting (SIR). It can reduce the confounding bias introduced by the pretraining knowledge and improve the model's ability to learn general features, making semantic segmentation of RS images benefit more from pretraining. Besides, we conduct a detailed theoretical analysis of our methods and conduct extensive experiments on two widely used public RS datasets. Experimental results demonstrate that applying SIR to encoder–decoder semantic segmentation models achieves performance improvements, proving the effectiveness and application values of the proposed method.**

***Index Terms*— Causal inference, deep learning (DL), remote sensing (RS), semantic segmentation, transfer learning.**

I. Introduction

REMOTE sensing (RS) images cover a large geographical region containing rich semantic information. Therefore, semantic segmentation, i.e., pixel-level classification, becomes one of the most significant tasks in RS image interpretation. It is widely used in various fields of RS, such as disaster prevention and mitigation, land use planning, and environmental monitoring [1], [2], [3], [4].

Unlike scene-level classification tasks focusing on global information that answers what, semantic segmentation tasks also focus on local information that answers where [5]. Previous works on semantic segmentation utilize encoder–decoder deep learning (DL) models to simultaneously extract high-level features containing global information and low-level features containing local information to infer the semantic information. Applying DL-based semantic segmentation methods to RS images brings the problems of limited labeled data and enormous computational costs for training from scratch. A widely used approach to the above problems is transfer learning [6], which takes the feature extraction layers of a model pretrained on a large dataset and applies them as the encoder of the semantic segmentation model, leveraging the well-learned global and local information contained in the transferred features [7]. To ensure that the transferred features have similar global and local information as that in the target dataset to be semantically segmented, either of the following assumptions should be satisfied: 1) the pretrained model can learn *general* features that carry common latent patterns, regardless of the exact task and dataset. 2) The pretraining dataset is *similar* enough to the target dataset. However, the *heterogeneity* between the pretraining and target datasets makes the assumptions usually unsound.

First, most pretrained models, including large models that claim they can understand "anything" [8], utilize massive general datasets for training so that their learned features are ideally considered general since the pretrained model has seen various images. However, experimental results demonstrate that the "general" features are no longer general for the RS images [9], [10] due to the significant differences between RS datasets and general datasets as follows.

1) RS images are captured in a completely different way compared to general images. Unlike general images captured from a camera at a fixed, elevated terrestrial position, RS images are often captured by satellites, or unmanned aerial vehicles (UAVs). Thus, they are much different from general images in many aspects, e.g., having a top-down or oblique perspective, containing large geographic areas, and having various spectral bands.

Manuscript received 13 April 2024; revised 18 June 2024; accepted 19 July 2024. Date of publication 22 July 2024; date of current version 1 August 2024. This work was supported in part by the National Natural Science Foundation of China under Grant 42225605 and Grant 42001323, in part by the National Key Research and Development Program of China under Grant 2021YFB3900902, in part by the Provincial Key Research and Development Program of Zhejiang under Grant 2021C01031, and in part by the Fundamental Research Funds for the Central Universities under Grant 2022FZZX01-05 and Grant 226-2024-00124. *(Shuting Shi and Baohong Li contributed equally to this work.) (Corresponding author: Zhenhong Du.)*

Shuting Shi, Laifu Zhang, Sensen Wu, Yiming Yan, and Zhenhong Du are with the Zhejiang Provincial Key Laboratory of Geographic Information Science, Hangzhou 310028, China, and also with the School of Earth Sciences, Zhejiang University, Hangzhou 310028, China (e-mail: shutingshi@zju.edu.cn; giserfu@zju.edu.cn; wusensengis@zju.edu.cn; yanyiming@zju.edu.cn; duzhenhong@zju.edu.cn).

Baohong Li and Kun Kuang are with the College of Computer Science and Technology, Zhejiang University, Hangzhou 310028, China (e-mail: baohong.li@zju.edu.cn; kunkuang@zju.edu.cn).

Tian Feng is with the School of Software Technology, Zhejiang University, Hangzhou 310028, China (e-mail: t.feng@zju.edu.cn).

Digital Object Identifier 10.1109/TGRS.2024.3432397

2) The importance of spatial information is different. Semantic segmentation models should pay more attention to the spatial information of RS images since they contain much stronger spatial correlations than general images.
3) The semantic density is different. Because of the large geographic areas in RS images, various targets appear in a single image, resulting in high semantic density compared to general images.
4) The semantic information is different. RS images often contain semantic information specific to the geographic domain, such as land use and land cover, which is rare in general images.

Second, although recent studies utilize large-scale RS datasets for pretraining so that the learned features are considered "similar" to other RS data [11], the features are still not similar enough due to noticeable differences among RS datasets as follows.

1) *Regional Heterogeneity:* RS datasets cover diverse geographical regions with unique semantics, leading to considerable dissimilarity between datasets.
2) *Sensor Variations:* RS images are collected using various sensors varying in spatial resolution. Different sensors can capture the same region differently, leading to semantic differences.

As a result, to make the model learn global and local information better from the input images leveraging pretraining knowledge, it is necessary to develop approaches to overcome the heterogeneity between the pretraining and target datasets. Rethinking the two assumptions mentioned earlier, since it is too difficult to obtain large amounts of annotated data similar to the target images, we need to force the pretrained model to learn *general* features that remain invariant regardless of the exact task and dataset [12].

To achieve the above goal, we first present a novel formulation of the above problem from a causal perspective. The essence of machine learning is to learn the correlations in data, where only the causality reflects intrinsic and universal dependency among variables and remains invariant regardless of the exact task and dataset, resulting in the general features. Other correlations, such as confounding bias, are spurious correlations that change with different datasets or tasks, resulting in the data-specific features [13]. In semantic segmentation tasks, we can intuitively consider that the raw input images are the cause, and the pixel-level classification results are the outcome. The different layers of the encoder features can be regarded as mediators on the causal path from the cause to the outcome. Since the encoder is transferred from the pretrained model, the pretraining knowledge affects all encoder features and thus acts as a common cause of every two layers of the encoder features. Such a common cause of two variables is defined as a *confounder* and can introduce spurious correlation, i.e., confounding bias, to the two variables [14]. Therefore, we must remove the spurious correlations and infer the causality to make the model learn general features.

Based on the above causal formulation, we argue that the commonly used fine-tuning process [15] cannot deal with confounding bias, and we propose a novel approach that can reduce confounding bias with some reasonable prior knowledge about the pretraining dataset. Specifically, we use the instrumental variable (IV) theory [16] to illustrate why simply fine-tuning the entire encoder of semantic segmentation models cannot solve confounding bias. Moreover, we propose stepwise intervention and reweighting (SIR) to reduce confounding bias for each encoder layer. SIR stepwise uses the feature of the previous layer as an IV of the current layer for intervention. Furthermore, with prior knowledge about the label distribution of the pretraining dataset, SIR also calculates de-confounding weights and reweights the samples during the learning process of each encoder layer.

We provide theoretical analyses of the proposed SIR and conduct experiments on widely used public datasets. Extensive experimental results show that applying SIR to commonly used encoder–decoder structured semantic segmentation models achieves performance improvements, demonstrating its effectiveness and application values.

In summary, our main contributions are as follows.

1) We present a novel formulation of RS image semantic segmentation from a causal perspective, considering the pretraining knowledge as a confounder in the learning process from the previous encoder layers to the next layer.
2) We propose a novel approach that can reduce confounding bias for each encoder layer. Detailed theoretical analyses guarantee the correctness of our method.
3) We conduct extensive experiments on widely used RS datasets. The results show that applying the proposed approach to encoder–decoder structured semantic segmentation models achieves performance improvements, demonstrating the effectiveness of our method.

II. Related Work

A. DL-Based Semantic Segmentation

Fully convolutional networks (FCNs) [5] is the first DL-based model for image semantic segmentation. It replaces the fully connected layer of convolutional neural networks (CNNs) with the deconvolutional layers, enabling FCNs to generate pixel-level output instead of scene-level output. U-Net [17] introduces skip-connections that combine coarse-grained and fine-grained information based on CNN's encoder–decoder structure. Extensions based on such structures, like U-Net++ [18], SegNet [19], and high-resolution net (HRNet) [20], are proposed after that. Deeplab series [21], [22], [23], [24] introduces dilated convolutions, spatial pyramid pooling, and multigrid strategy to enhance receptive fields and better utilize contextual information. Pan et al. [25] proposed an efficient backbone named DDRNet and designed a new contextual information extractor named deep aggregation pyramid pooling module (DAPPM) to reduce model complexity. Wang et al. [26] proposed a unified multiscale learning (UML) framework that utilizes a multiscale spatial-channel attention mechanism and a multiscale shuffle block to improve the distortion problem in hyperspectral image semantic segmentation. Capsule-vectored neural network (CVNN) [27] leverages capsule networks to overcome the issues of feature redundancy and insufficient labeled samples for RS image interpretation. Wang et al. [28] propose W-Net, a two-branch

multitask coupling framework consisting of a superpixel module and a change detection module, to guarantee the completeness of the features and the edge information in semantic segmentation results.

To address the insufficient ability to extract long-range information in CNN-based semantic segmentation models [29], the transformer structure is introduced into semantic segmentation networks, where transformers can learn global contextual information with the self-attention mechanism. Segmentation transformer (SETR) [30] is the first to introduce vision transformer (ViT) [31] into semantic segmentation networks by replacing the CNN-based encoder with the transformer-based encoder, offering a new perspective for semantic segmentation problems. Pyramid ViT (PVT) [32] introduces a pyramid structure in the transformer, reducing the computations of large feature maps and enabling flexible learning of multiscale and high-resolution features. SegFormer [33] combines a transformer-based encoder with a multilayer perceptron (MLP) decoder to extract local and global information more efficiently. Segmenter [34] uses a mask transformer decoder generating class masks to improve performance further. Cheng et al. [35] conceptualized semantic segmentation as a mask classification task instead of a pixel-level classification task and proposed a transformer-based mask classification model, i.e., MaskFormer. They then propose Mask2Former [36] that uses a transformer decoder with masked attention to improve performance and efficiency. ConvNeXt [37] redesigns the structure of convolutional models to have characteristics similar to transformers, which competes favorably with transformers in terms of accuracy while maintaining the simplicity and efficiency of standard convolutional networks. Guo et al. [38] designed a convolutional attention network named SegNeXt that leverages the attention mechanism and implements it using cheap convolutional operations. Ji et al. [39] proposed PASSNet, a lightweight hybrid model integrating the respective inductive bias from CNNs and global receptive field from transformers, for efficient extraction of both local and global features. Side adapter network (SAN) [40] models the semantic segmentation task as a region recognition problem and leverages the knowledge of a pretrained vision-language model to improve performance.

All the above methods use encoder–decoder structures and usually utilize transfer learning that transfers a model pretrained on a large-scale general dataset as the encoder part. Since the distribution of the RS images differs from those used for pretraining, the above methods suffer from confounding bias caused by the pretraining knowledge.

B. Causal Inference

Causal inference methods aim to remove spurious correlations and learn causality from observational data. These methods mainly focus on addressing the most common spurious correlation caused by confounding bias, which results from common causes of the cause and the outcome [14]. Most of these works utilize different techniques to make the confounding variables independent of the cause, including backdoor adjustment [41], confounder balancing [42], [43], [44], balanced representation learning [45], [46], [47], and generative-model-based methods [48], [49]. Although the above approaches achieve success in many fields, applying them to vision tasks is still challenging because they assume a well-defined causal model of the specific task. Fortunately, previous research formulates a standard causal model of vision tasks that matches the underlying process of generating data [13], which can be expressed as images $\rightarrow$ datasets $\rightarrow$ features $\rightarrow$ labels. Following this formulation, recent works apply causal inference to many subfields of computer vision, including visual question answering [50], image data privacy [51], few-shot classification [52], and weakly-supervised semantic segmentation [53]. Our research is different from previous causality-based computer vision works in the following aspects.

1) The studied problem is different. We study the problem of making semantic segmentation models learn global and local information better from the RS images, leveraging pretraining knowledge. We novelly formulate it as a confounding bias problem where the pretrained model fails to learn general features appropriate to RS datasets because of data heterogeneity. To the best of our knowledge, we are the first to formulate the above problem from a causal perspective.
2) The formulation of the causal model is different. Causal models presented by previous works are designed for their corresponding specific tasks and cannot be applied to our problem. Therefore, we present a novel causal model formulation for semantic segmentation tasks, aligning with the data flow in encoder–decoder structured semantic segmentation modeling.
3) The proposed de-confounding approach is different. Most previous works use backdoor adjustment techniques to address confounding bias, which are practical but inefficient since they involve extra models and learning processes to achieve the adjustment. Our methods address confounding bias in a stepwise fine-tuning way. The entire optimization process only involves parameters of the encoder–decoder structured model itself, and thus, it uses far less computational costs compared to other causality-based works.

III. Methodology

A. Problem Formulation

Suppose we have a target dataset $\mathcal{D}_{\text{seg}} = \{\mathbf{x}_i, \mathbf{y}_i\}_{i=1}^{n}$, where n denotes the number of images in $\mathcal{D}_{\text{seg}}$. For a unit i, $\mathbf{x}_i \in \mathcal{X} \subset \mathbb{R}^{h\times w\times c}$ is the pixel-value matrix, and $\mathbf{y}_i \in \mathcal{Y} \subset \mathbb{R}^{h\times w\times d}$ is the semantic-label matrix represented using one-hot encoding, where d denotes the number of semantic classes, and h, w, and c denote the height, width, and channel numbers, respectively. The goal of semantic segmentation is to learn a function $f : \mathcal{X} \rightarrow \mathcal{Y}$ to predict $\mathbf{Y}$ from input $\mathbf{X}$.

A common way to learn f is utilizing deep models to first learn an encoder $\phi : \mathcal{X} \rightarrow \mathcal{R}$ that maps the images into high-level and low-level features, where $\mathcal{R}$ is the feature space, and then learn a decoder $\theta : \mathcal{R} \rightarrow \mathcal{Y}$ that predicts semantic labels of the input images from the learned features.

Following most previous works [7], [23], [24], since training deep models from scratch suffers from limited labeled data and

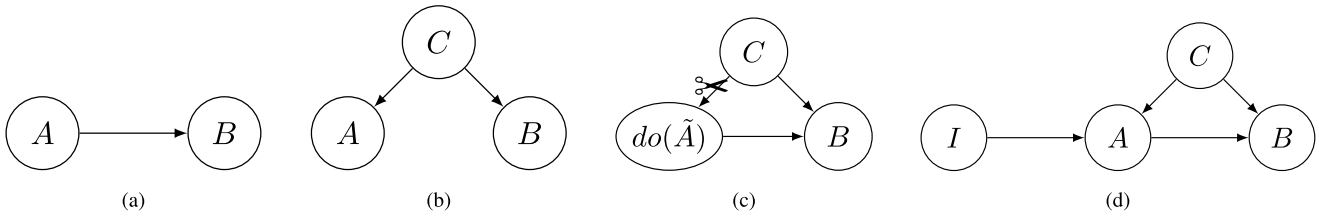


Fig. 1. (a) Causal link. (b) Confounding bias. (c) Intervention to remove confounding bias. (d) IV to remove confounding bias.

enormous computational costs, the feature extraction layers of a model pretrained on a large-scale dataset $\mathcal{D}_{\text{pre}}$ is often used as the encoder. Specifically, the pretrained model also consists of a feature-learning function $\phi_{\text{pre}} : \mathcal{X} \to \mathcal{R}$ and a label-predictive function $g : \mathcal{R} \to \mathcal{L}$, where the label space $\mathcal{L}$ can be different from $\mathcal{Y}$ since the task of the pretrained model may be different, e.g., scene-level classification. Because of the heterogeneity between $\mathcal{D}_{\text{seg}}$ and $\mathcal{D}_{\text{pre}}$, the objective ϕ usually differs from ϕ_{pre} learned by the pretrained model, making the learned features $\mathbf{R}$ not accurately represent the semantic information of the target images, and thus θ also fails to learn accurate semantic labels based on the inaccurate $\mathbf{R}$. Therefore, the main challenge is how to make ϕ_{pre} learned from $\mathcal{D}_{\text{pre}}$ also appropriate to $\mathcal{D}_{\text{seg}}$, i.e., to learn *general* features that remains invariant regardless of the exact task and dataset from the pretrained model, such that the learned features can accurately represent semantic information of the target images as well.

To address the above challenge, we rethink the problem from a causal perspective. That is, because only the causality reflects intrinsic and universal dependency among variables and remains invariant, the key to learning general features is to learn causality and remove spurious correlations in the fine-tuning process of the pretrained encoder.

B. Preliminaries

Before introducing our causal formulation of the problem and the proposed approach, we first briefly introduce causal preliminaries for readers to better understand our motivation.

1) Causality Versus Correlation: Causality means that changing one variable causes a change in another variable, while correlation only indicates that two variables change together. Based on the structural causal model (SCM) [54], the causality between two variables A and B can be represented as the causal graph in Fig. 1(a). In statistical analysis, correlation does not imply causality, mainly due to spurious correlations [14]. Considering the causal graph in Fig. 1(b), a fork structure $A \leftarrow C \rightarrow B$ also introduces a correlation between A and B since a change in C causes changes in both A and B. However, such a correlation is not causality because when we fix C and change A, then B will remain the same.

Formally, the spurious correlation, as shown in Fig. 1(b), is defined as *confounding bias*, and the common cause C is a *confounder* of A and B.

2) De-Confounding Through Intervention: The essence of machine learning is to learn the correlations in data, where only the causality reflects intrinsic and universal dependency among variables and remains invariant regardless of the exact task and dataset. Spurious correlations, such as confounding bias, change with different datasets or tasks. Therefore, the key to obtaining general and stable learning is to infer causality and remove confounding bias from the observational data [13].

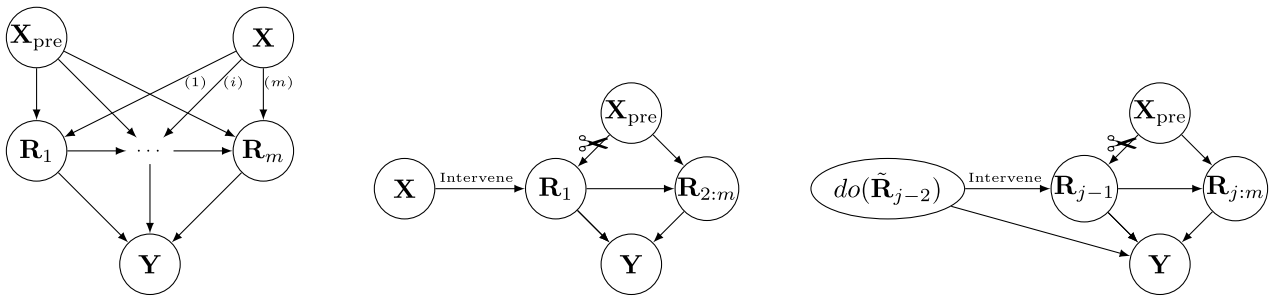
Confounding bias can be effectively removed through interventional studies. An intervention on A, denoted as $do(\tilde{A})$, is to set A to a specific value $\tilde{A}$ manually. After such an intervention, $do(\tilde{A})$ is no longer related to the original direct causes. Thus, all the arrows into A are cut off, as shown in Fig. 1(c). In observational studies, there are various approaches to make C independent of A as a simulation of the postintervention distribution, such as backdoor adjustment and confounder balancing [55]. However, most of them assume the confounder C is well-defined and fully measured, which is easily violated in real-world applications.

3) Intervention by IVs: In cases, where confounders are not well-defined or measured, IVs can be used to simulate interventional studies. A valid IV I of A on B, as shown in Fig. 1(d), should satisfy the following assumptions: 1) I is related to A; 2) I is independent of C; and 3) I is conditional independent of B [16]. With a valid IV, we can learn a function $f_a : \mathcal{I} \to \mathcal{A}$ and use $\tilde{A} = f_a(I)$ as the interventional value of A. Since $\tilde{A}$ is only related to I and independent of C, we can use $\tilde{A}$ instead of A to predict B to remove confounding bias.

C. Motivation

1) Proposed SCM of Semantic Segmentation: Based on the physical data-generation process in encoder–decoder structured semantic segmentation modeling, we propose a novel SCM of semantic segmentation, as shown in Fig. 2(a), where $\mathbf{X}_{\text{pre}}$ denotes the pretraining knowledge, $\mathbf{R}_i$ denotes the output feature of the ith encoder layer, m is the number of encoder layers, and (i) denotes the corresponding causal link is added in the ith step of the proposed methods (as to be clarified in Section III-C2). A detailed explanation of the SCM is as follows.

1) The causal path $\mathbf{X} \rightarrow \mathbf{R}_{1:m} \rightarrow \mathbf{Y}$. In semantic segmentation tasks, the pixel values $\mathbf{X}$ of an input image are the cause, and the semantic labels are the outcome. Most deep semantic segmentation models use images in $\mathcal{D}_{\text{seg}}$ to fine-tune the pretrained encoder to obtain high-level and low-level features and train a decoder to predict semantic labels from these features. Therefore, the causal path between $\mathbf{X}$ and $\mathbf{Y}$ consists of two parts: 1) $\mathbf{X} \rightarrow \mathbf{R}_{1:m}$ that results from the encoder learning different levels of features from input images and 2) $\mathbf{R}_{1:m} \rightarrow \mathbf{Y}$ that results from the decoder learning semantic labels from the different levels of features.
2) $\mathbf{X}_{\text{pre}}$ acts as a confounder of $\mathbf{R}_i$ and $\mathbf{R}_j$. Since the encoder is transferred from a pretrained model, all the features are affected by the pretraining knowledge, i.e., there exist



(a) (b) (c)

Fig. 2. (a) SCM of semantic segmentation. (b) $\mathbf{X}$ is an IV of $\mathbf{R}_1$ on $\mathbf{R}_2$ in the second step. (c) $do(\tilde{\mathbf{R}}_{j-2})$ is an IV of $\mathbf{R}_{j-1}$ on $\mathbf{R}_j$ in the jth step.

causal links between each $\mathbf{R}_i$ and $\mathbf{X}_{\text{pre}}$. Therefore, on the causal path $\mathbf{X} \rightarrow \mathbf{R}_{1:m} \rightarrow \mathbf{Y}$, $\mathbf{X}_{\text{pre}}$ is a confounder of each pair of $\mathbf{R}_i$ and $\mathbf{R}_j$, resulting in noncausal paths like $\mathbf{X} \rightarrow \mathbf{R}_i \leftarrow \mathbf{X}_{\text{pre}} \rightarrow \mathbf{R}_j \rightarrow \mathbf{Y}$. As a result, the pretraining knowledge introduces *confounding bias* into the learning process from the previous layers to the next layer of the encoder.

2) De-Confounding Step by Step: In this article, we propose a novel SIR to reduce the confounding bias introduced by $\mathbf{X}_{\text{pre}}$. SIR concurrently executes two procedures, i.e., SIR. The former constitutes the fundamental framework of SIR, while the latter serves as a module for further de-confounding refinement leveraging additional prior knowledge when available.

a) Stepwise intervention: In the proposed SCM, since $\mathbf{X}_{\text{pre}}$, i.e., the pretraining knowledge, is not directly measured, IVs are needed to reduce the confounding bias introduced by $\mathbf{X}_{\text{pre}}$. Fortunately, before using $\mathbf{X}$ to fine-tune the encoder, $\mathbf{R}_{1:m}$ are all independent of $\mathbf{X}$. It means that if we only fine-tune the first encoder layer and leave the other encoder layers unchanged, only $\mathbf{R}_1$ will be directly affected by $\mathbf{X}$. That is, as shown in Fig. 2(b), $\mathbf{X}$ is directly related to $\mathbf{R}_1$, independent of $\mathbf{X}_{\text{pre}}$, and independent of $\mathbf{R}_{2:m}$ given $\mathbf{R}_1$. Therefore, $\mathbf{X}$ satisfies all requirements for a valid IV of $\mathbf{R}_1$ on $\mathbf{R}_{2:m}$. As a result, just fine-tuning the first layer of the encoder with $\mathbf{X}$ can be regarded as an intervention on $\mathbf{R}_1$ and reduce the confounding bias between $\mathbf{R}_1$ and $\mathbf{R}_{2:m}$.

To reduce the confounding bias between $\mathbf{R}_2$ and $\mathbf{R}_{3:m}$, the intervened $\mathbf{R}_1$ by $\mathbf{X}$, i.e., $do(\tilde{\mathbf{R}}_1)$, can be regarded as an IV of $\mathbf{R}_2$ on $\mathbf{R}_{3:m}$. As stated in Section III-B, intervention on $\mathbf{R}_1$ cuts off $\mathbf{X}_{\text{pre}} \rightarrow \mathbf{R}_1$. Thus, $do(\tilde{\mathbf{R}}_1)$ satisfies that $do(\tilde{\mathbf{R}}_1)$ is independent of $\mathbf{X}_{\text{pre}}$, related to $\mathbf{R}_2$, and independent of $\mathbf{R}_{3:m}$ given $\mathbf{R}_2$. Similarly, for the follow-up layers, after intervention on $\mathbf{R}_{j-2}$ by its intervened previous layer, $do(\tilde{\mathbf{R}}_{j-2})$ can be further used as an IV of $\mathbf{R}_{j-1}$ on $\mathbf{R}_{j:m}$, where $j \in [3, m]$. As a result, the confounding bias between each pair of encoder layers can be reduced step by step.

b) Stepwise reweighting: Although $\mathbf{X}_{\text{pre}}$ cannot be directly measured, with prior knowledge about $\mathcal{D}_{\text{pre}}$ available, we can utilize it to improve the de-confounding performance in each step by direct confounder balancing. Specifically, in each step of SIR, we need to learn a balancing weight w for every sample in $\mathcal{D}_{\text{seg}}$ so that the distribution of the reweighted samples is a simulation of the postintervention distribution. Therefore, the key challenge is calculating w in each step.

Following Yue et al. [52], we regard L, i.e., the predicted label by the pretrained model, as a proxy of $\mathbf{X}_{\text{pre}}$, which means that the impact of pretraining knowledge on the semantic segmentation of an image from $\mathcal{D}_{\text{seg}}$ can be reflected by the L value of that image. Therefore, the stepwise reweighting procedure of SIR requires that the marginal distribution of the labels in $\mathcal{D}_{\text{pre}}$, i.e., $\mathbb{P}(L)$, is given. As a result, to obtain the L value as a proxy of $\mathbf{X}_{\text{pre}}$ for each image in $\mathcal{D}_{\text{seg}}$, we first calculate $\mathbb{P}(L \mid \mathbf{X})$ by simply feeding the images in $\mathcal{D}_{\text{seg}}$ into the pretrained model and getting the outputs. The L value l_{p} of each image in $\mathcal{D}_{\text{seg}}$ can then be calculated through

$$l_{\text{p}} = \arg\max_{L} \mathbb{P}(L \mid \mathbf{X}). \tag{1}$$

Proposition 1: Under the assumption that L is a proxy of $\mathbf{X}_{\text{pre}}$ and $\mathbb{P}(L)$ is known, the balancing weights $W_{j-1,j}$ in a certain step j of SIR to address the confounding bias between $\mathbf{R}_{j-1}$ and $\mathbf{R}_j$ are calculated by

$$W_{j-1,j} = \frac{\mathbb{P}(L = l_{\text{p}})}{\mathbb{P}(L = l_{\text{p}} \mid \mathbf{R}_{j-1})}. \tag{2}$$

Proof: Because of confounding bias, the distribution of the original samples is $\mathbb{P}(\mathbf{R}_{j-1}, \mathbf{R}_j, \mathbf{X}_{\text{pre}}) = \mathbb{P}(\mathbf{R}_j \mid \mathbf{R}_{j-1}, \mathbf{X}_{\text{pre}}) \cdot \mathbb{P}(\mathbf{R}_{j-1} \mid \mathbf{X}_{\text{pre}}) \cdot \mathbb{P}(\mathbf{X}_{\text{pre}})$. However, the postintervention distribution is supposed to be $\mathbb{P}_{\text{I}}(\mathbf{R}_{j-1}, \mathbf{R}_j, \mathbf{X}_{\text{pre}}) = \mathbb{P}(\mathbf{R}_j \mid \mathbf{R}_{j-1}, \mathbf{X}_{\text{pre}}) \cdot \mathbb{P}(\mathbf{R}_{j-1}) \cdot \mathbb{P}(\mathbf{X}_{\text{pre}})$. To make the distribution of the reweighted samples as a simulation of the postintervention distribution, $W_{j-1,j}$ needs to satisfy that $W_{j-1,j} \cdot \mathbb{P}(\mathbf{R}_{j-1}, \mathbf{R}_j, \mathbf{X}_{\text{pre}}) = \mathbb{P}_{\text{I}}(\mathbf{R}_{j-1}, \mathbf{R}_j, \mathbf{X}_{\text{pre}})$, i.e.,

$$\begin{aligned}
W_{j-1,j} &= \frac{\mathbb{P}_{\text{I}}(\mathbf{R}_{j-1}, \mathbf{R}_j, \mathbf{X}_{\text{pre}})}{\mathbb{P}(\mathbf{R}_{j-1}, \mathbf{R}_j, \mathbf{X}_{\text{pre}})} \\
&= \frac{\mathbb{P}(\mathbf{R}_j \mid \mathbf{R}_{j-1}, \mathbf{X}_{\text{pre}}) \cdot \mathbb{P}(\mathbf{R}_{j-1}) \cdot \mathbb{P}(\mathbf{X}_{\text{pre}})}{\mathbb{P}(\mathbf{R}_j \mid \mathbf{R}_{j-1}, \mathbf{X}_{\text{pre}}) \cdot \mathbb{P}(\mathbf{R}_{j-1} \mid \mathbf{X}_{\text{pre}}) \cdot \mathbb{P}(\mathbf{X}_{\text{pre}})} \\
&= \frac{\mathbb{P}(\mathbf{R}_{j-1}) \cdot \mathbb{P}(\mathbf{X}_{\text{pre}})}{\mathbb{P}(\mathbf{R}_{j-1} \mid \mathbf{X}_{\text{pre}}) \cdot \mathbb{P}(\mathbf{X}_{\text{pre}})} \\
&= \frac{\mathbb{P}(\mathbf{R}_{j-1}) \cdot \mathbb{P}(\mathbf{X}_{\text{pre}} \mid \mathbf{R}_{j-1}) \cdot \mathbb{P}(\mathbf{X}_{\text{pre}})}{\mathbb{P}(\mathbf{R}_{j-1} \mid \mathbf{X}_{\text{pre}}) \cdot \mathbb{P}(\mathbf{X}_{\text{pre}}) \cdot \mathbb{P}(\mathbf{X}_{\text{pre}} \mid \mathbf{R}_{j-1})} \\
&= \frac{\mathbb{P}(\mathbf{X}_{\text{pre}})}{\mathbb{P}(\mathbf{X}_{\text{pre}} \mid \mathbf{R}_{j-1})}.
\end{aligned}$$

Using $L = l_{\text{p}}$ as a proxy value of $\mathbf{X}_{\text{pre}}$ for each sample, the right-hand side of the above equation is $(\mathbb{P}(L = l_{\text{p}})/\mathbb{P}(L = l_{\text{p}} \mid \mathbf{R}_{j-1}))$. □

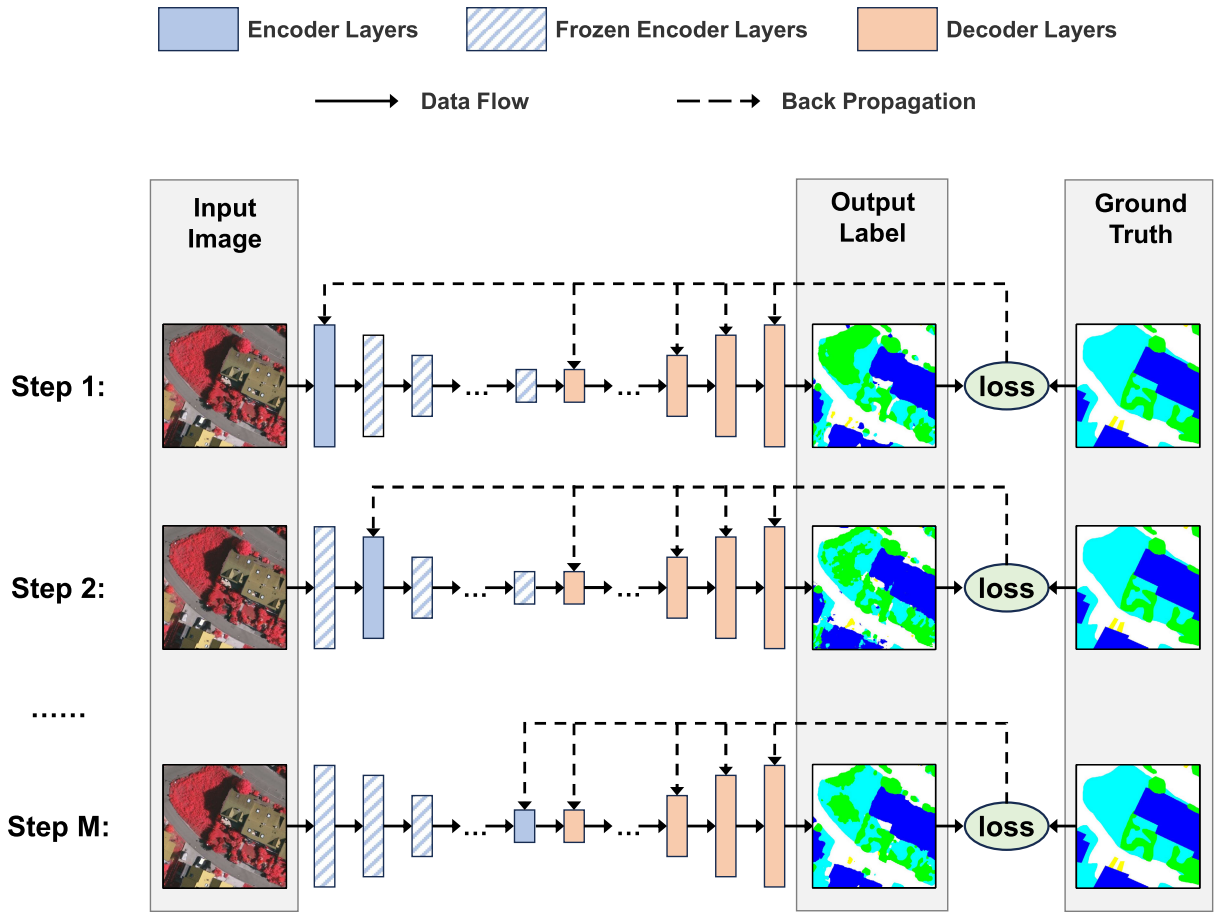


Fig. 3. Stepwise intervention framework of SIR. In the ith step, only the ith encoder layer is fine-tuned and other encoder layers are frozen.

Based on Proposition 1, SIR calculates balancing weights by (2) and reweights the samples in the optimization process of each step to improve the de-confounding performance further. Note that the prior knowledge about $\mathbb{P}(L)$ is necessary for the calculation process of $W_{j-1,j}$ in (2), but it is still possible to use SIR by assuming a uniform distribution in cases where $\mathbb{P}(L)$ is unknown.

To sum up, in SIR, we first fine-tune the first encoder layer with $\mathbf{X}$ to obtain the intervened $\mathbf{R}_1$. Then, we repeat using the output features of the fine-tuned layer (i.e., features after intervention) to calculate the balancing weights and fine-tune the next layer with the reweighted samples until the last encoder layer is fine-tuned. In the scenario, where no prior knowledge about $\mathbb{P}(L)$ is available for calculating the balancing weights, we can either assume a uniform distribution of $\mathbb{P}(L)$ or use an ablation version of SIR without reweighting to reduce the confounding bias.

Note that simply fine-tuning the entire encoder as previous works did cannot address the confounding bias since all encoder layers are simultaneously optimized in a single iteration, which means that $\mathbf{X}$ affects all $\mathbf{R}_{1:m}$. Therefore, $\mathbf{X}$ does not satisfy the IV assumption that it should be independent of $\mathbf{R}_{2:m}$ given $\mathbf{R}_1$. While the proposed SIR makes the IV assumptions satisfied since the causal link between $\mathbf{X}$ and $\mathbf{R}_{1:m}$ is added step by step, as shown in Fig. 2(a). In the ith step of SIR, only $\mathbf{X} \to \mathbf{R}_{1:i}$ exist in the causal graph and $\mathbf{X} \to \mathbf{R}_{i+1:m}$ are not added yet.

D. Implementation

As stated in Section III-A, the label space $\mathcal{L}$ of the pretrained model varies by different tasks. For simplicity, in this article, we introduce SIR with scene-level classification pretrained models whose label space is $\mathcal{L} \subset \mathbb{R}^s$ (represented using one-hot encoding), e.g., VGG [56], ResNet [57], and ViT [31], where s denotes the number of scene-level classes. Note that pretrained models for other tasks are also applicable, as to be clarified in Section III-D2.

1) Implementation of Stepwise Intervention: Stepwise intervention is the basic framework of SIR. It can be implemented as a causality-guided fine-tuning method for encoder–decoder structured semantic segmentation models. The schematic of this process is illustrated in Fig. 3.

Suppose the semantic segmentation model consists of a pretrained encoder with m feature layers and a random-initialized decoder. In the first step of SIR, we use images in $\mathcal{D}_{\text{seg}}$ to fine-tune the first encoder layer and freeze the other encoder layers so that $\mathbf{R}_1$ is intervened by $\mathbf{X}$. That is, only the parameters of the first encoder layer and the entire decoder are optimized through backpropagation in the first step. In the second step, we unfreeze the second encoder layer and freeze the fine-tuned first encoder layer while the other encoder layers remain frozen. We use the features output by the fine-tuned first encoder layer to fine-tune the second encoder layer so that the confounding bias between $\mathbf{R}_1$ and $\mathbf{R}_2$ can be reduced. In the following steps, we repeat this process to fine-tune one

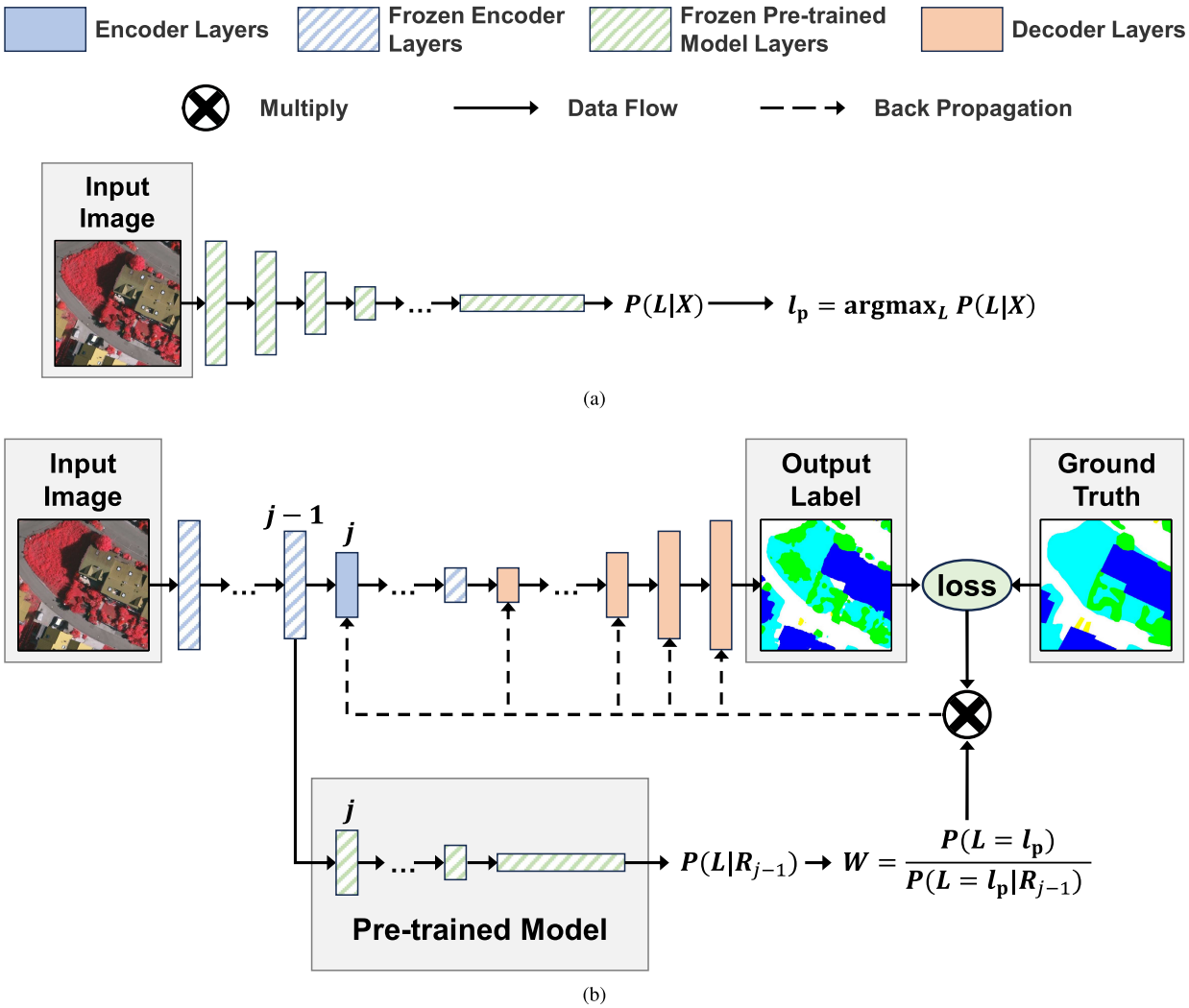


Fig. 4. (a) In SIR, we first calculate $L = l_p$, i.e., the proxy value of $\mathbf{X}_{pre}$, with the frozen pretrained model for images in $\mathcal{D}_{seg}$. (b) In the jth step ($j \in [2, m]$) of SIR, we take the output features of the $(j-1)$th encoder layer as the input to the jth layer of the frozen pretrained model to obtain the balancing weights of the jth step. Then, we fine-tune the jth encoder layer using the samples reweighted by the balancing weights.

single encoder layer at each step until all the encoder layers are fine-tuned. For example, in the ith step ($i \in [2, m]$) of SIR, we freeze all the encoder layers except the ith layer and use the output of the feature by the $(i-1)$th encoder layer to optimize the parameters of the unfrozen parts of the model.

2) Implementation of Stepwise Reweighting: To further improve the de-confounding performance, SIR also conducts a reweighting procedure for each step. With prior knowledge about $\mathbb{P}(L)$, e.g., the pretrained model is trained on a public dataset so that $\mathbb{P}(L)$ is known, we can use L as the proxy of $\mathbf{X}_{pre}$ for stepwise reweighting. That is, we first calculate the L values, denoted as l_p, for images in $\mathcal{D}_{seg}$ before training the model. They can be easily obtained by feeding the images into the frozen scene-level classification pretrained model, as shown in Fig. 4(a). Specifically, we input the images into the frozen pretrained model and get the predicted $\mathbb{P}(L \mid \mathbf{X})$. Then, we obtain l_p by (1). Note that if the pretrained model is not a scene-level classification model, l_p can be a vector or a matrix instead of a scalar. In such cases, we need to further convert l_p to a scalar with, e.g., the trace of l_p if l_p is a matrix, or the norm of l_p if l_p is a vector.

As a result, based on the stepwise intervention framework of SIR, in each step, we further calculate a balancing weight for every image and use the reweighted samples to optimize the parameters of the unfrozen parts. Specifically, in the jth step ($j \in [2, m]$) of SIR, we first get the outputs of the frozen $(j-1)$th encoder layer, i.e., $\mathbf{R}_{j-1}$. Then, we input $\mathbf{R}_{j-1}$ into the frozen scene-level classification pretrained model to get $\mathbb{P}(L \mid \mathbf{R}_{j-1})$, and calculate the balancing weights of the jth step for all training samples by (2). Finally, we reweight the samples and use them to optimize the parameters of the jth encoder layer and the entire decoder.

In each step, we utilize the same loss function to optimize the unfrozen layers in the current step. Note that any loss function designed for semantic segmentation is available. Here, we use the most commonly used pixel-level cross-entropy loss for our implementation. The loss function of the eth step of SIR is

$$\ell_e = -\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{h\times w}\sum_{k=1}^{d} W_{e-1,e} \cdot y_{i,j,k} \cdot \log\left(\hat{y}_{i,j,k}\right) \tag{3}$$

TABLE I
MAIN RESULTS ON THE WHDLD DATASET. THE BEST RESULTS ARE IN BOLD, AND THE SECOND-BEST RESULTS ARE UNDERLINED

Model	PA(%)	MPA(%)	MIoU(%)	IoU(%)					
				Bare soil	Building	Pavement	Road	Vegetation	Water
FCN	81.70	70.55	57.91	36.70	52.52	34.87	53.32	78.65	91.38
U-Net	80.28	66.54	54.97	32.65	50.85	37.37	42.76	77.52	88.65
DLv3+	80.43	70.31	56.30	37.44	45.42	37.59	50.32	76.88	90.18
SETR-Naive	73.96	57.97	45.50	30.60	41.48	31.90	20.70	69.77	78.54
SETR-PUP	73.15	57.47	44.07	33.24	40.61	29.90	14.08	67.63	78.97
SETR-MLA	73.25	57.86	44.96	28.67	39.45	31.88	20.91	68.75	80.12
Segmenter	79.30	68.31	55.73	45.38	48.01	36.19	74.51	40.04	90.23
SegFormer	82.67	69.96	58.33	33.25	54.02	36.58	54.33	78.88	92.88
Mask2Former	79.83	71.65	55.42	**55.39**	56.40	23.84	**77.50**	27.19	92.19
DDRNet	80.72	67.68	56.38	51.23	54.42	34.48	76.43	28.87	92.88
SegNeXt	80.90	70.66	57.53	51.14	53.68	37.90	77.06	34.93	90.49
SAN	65.67	45.69	39.97	36.39	36.01	10.48	68.39	1.940	86.58
SIR+FCN	83.08	72.50	60.10	36.36	55.98	39.22	55.98	79.95	**93.13**
SIR+DLv3+	81.76	72.07	58.72	37.24	52.55	38.16	55.01	77.87	91.51
SIR+SegFormer	**83.51**	**73.17**	**61.19**	39.54	**57.56**	**41.14**	55.79	**80.06**	93.03

TABLE II
MAIN RESULTS ON THE VAIHINGEN DATASET. THE BEST RESULTS ARE IN BOLD, AND THE SECOND-BEST RESULTS ARE UNDERLINED

Model	PA(%)	MPA(%)	MIoU(%)	IoU(%)					
				Imp. Surf.	Building	Low Veg.	Tree	Car	Background
FCN	86.71	80.74	69.23	80.25	86.29	64.76	74.97	54.94	54.20
U-Net	85.30	79.08	65.34	77.56	86.47	58.19	74.62	50.80	44.38
DLv3+	85.34	77.83	66.55	78.04	84.83	61.27	73.50	36.67	64.97
SETR-Naive	77.10	64.90	51.35	66.80	71.66	53.85	61.22	10.48	44.12
SETR-PUP	77.33	57.35	47.47	66.70	71.07	49.87	65.31	37.83	31.47
SETR-MLA	78.71	65.30	52.60	67.99	73.70	54.34	66.43	17.23	35.92
Segmenter	84.35	78.03	66.79	84.56	56.14	64.64	71.39	78.43	45.57
SegFormer	89.34	82.30	74.00	83.69	89.90	70.12	78.98	55.47	65.86
Mask2Former	87.35	78.28	68.88	87.45	68.14	72.14	75.25	**80.66**	29.63
DDRNet	87.56	81.91	71.69	**90.49**	64.23	72.99	74.48	78.42	49.50
SegNeXt	83.77	78.79	65.76	83.41	57.91	63.76	70.81	76.88	41.77
SAN	50.99	38.73	36.13	69.33	12.14	25.33	59.52	44.96	5.500
SIR+FCN	92.34	**88.20**	**81.58**	88.41	92.78	78.31	83.34	65.53	81.13
SIR+DLv3+	87.33	81.64	70.74	80.51	86.59	67.61	75.93	43.43	70.40
SIR+SegFormer	**92.58**	86.80	81.12	87.97	**92.79**	**79.91**	**84.42**	55.58	**86.07**

where $W_{e-1,e}$ denotes the balancing weights of the eth step, $y_{i,j,k}$ denotes the kth dimension of the one-hot encoded ground truth semantic label of the jth pixel in the ith image, and $\hat{y}_{i,j,k}$ denotes the predicted probability of the kth semantic label by the model. Note that in the first step, the balanced weights of all samples are set to 1.

The source code of SIR is available at https://github.com/zjugiser/SIR.

IV. EXPERIMENTS

A. Datasets

The Wuhan Dense Labeling Dataset (WHDLD) [58], [59] contains 4940 RGB images cropped from a large RS image of the urban area in Wuhan, China, with a size of 256 × 256 pixels and the resolution is 2 m. Each pixel in these images is manually labeled into six categories: building, road, pavement, vegetation, bare soil, and water.

The Vaihingen [60] dataset from the International Society for Photogrammetry and RS (ISPRS) semantic segmentation challenge contains 33 aerial images acquired over the city of Vaihingen in Germany, with an average size of approximately 2500 × 2000 pixels and the ground-sampling distance of 0.09 m. Each pixel in these images is manually labeled into six categories: car, tree, low vegetation, building, impervious surfaces, and background. Following previous works [61], [62], [63], we selected seven images for testing (image IDs: 2, 8, 15, 16, 20, 27, and 37) and the remaining for training, and we used a sliding window approach with a window size of 512 × 512 pixels and a stride of 128 pixels for the models to process these images.

B. Experimental Settings

In order to demonstrate the effectiveness of our proposed method, we applied SIR to three representative semantic segmentation models, i.e., FCN [5], DeepLabv3+ (DLv3+) [24], and SegFormer [33]. We compared these models w/SIR plugged in against the following baselines: 1) CNN-based methods: FCN, U-Net [17], DLv3+, DDRNet [25], and SegNeXt [38]; and 2) transformer-based methods: SETR-Naive, SETR-PUP, SETR-MLA [30], Segmenter [34], SegFormer [33], Mask2Former [36], and SAN [40].

We implemented all the models in the PyTorch environment with Python 3.9 on two NVIDIA Titan Xp GPUs. For FCN, U-Net, and DLv3+, we adopted three different encoder backbones pretrained on the ImageNet-1K dataset [64], i.e.,

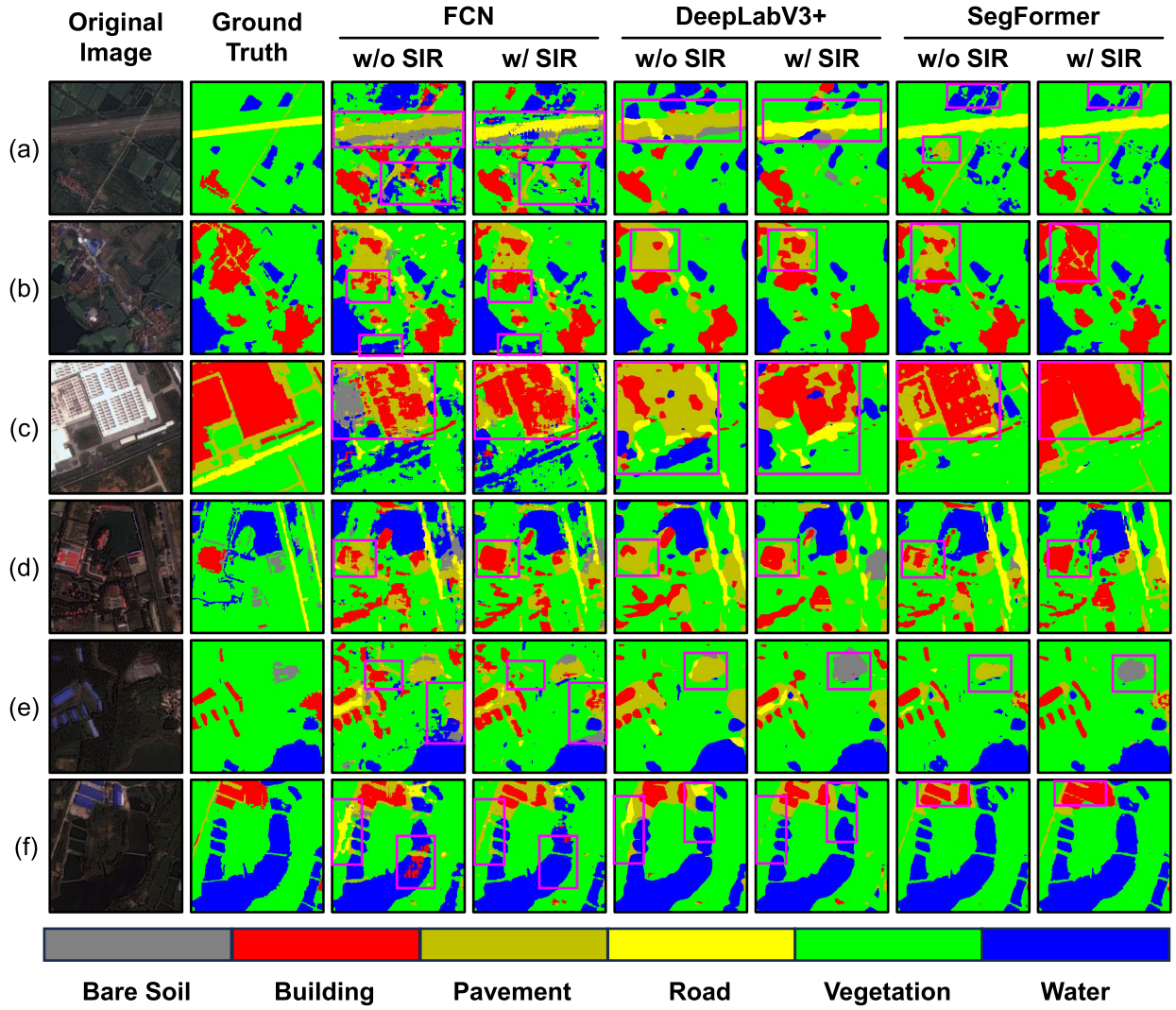


Fig. 5. Visualized semantic segmentation results from the comparison between models without SIR (w/o SIR) and with SIR (w/ SIR) on the WHDLD dataset. In these images, areas of significant difference are delineated by purple boxes, indicating regions where notable variations occur. (a)–(f) Representative images from the test dataset.

VGG16 [56], ResNet50 [57], and EfficientNet-B0 (EN-B0) [65]. For SETR series, segmenter, and SAN, we adopted the ViT-Base [31] encoder pretrained on the ImageNet-1K dataset, ADE20K dataset [66], and COCO stuff dataset [67], respectively. For SegFormer, we adopted a lightweight mix transformer encoder (MiT-B0) pretrained on the Pascal VOC dataset [68]. For Mask2Former, we adopted ResNet50 pretrained on the ADE20K dataset as the encoder. For DDRNet, we used their proposed DDRNet23-slim as the encoder backbone pretrained on the Cityscapes dataset [69]. For SegNeXt, we used their proposed MSCAN-T as the encoder backbone pretrained on the ADE20K dataset.

On the WHDLD dataset, we trained the models for 20 epochs with a batch size of 16. On the Vaihingen dataset, we trained for 20 epochs with a batch size of 8. We used Batch Normalization and the Adam optimizer with a weight decay of 1e-4 to optimize the models in the training process. We set the base learning rate to 1e-3 on the WHDLD dataset and 5e-4 on the Vaihingen dataset. To ensure a fair comparison, we did not apply any data augmentation methods in all experiments on both datasets.

To evaluate the performance of semantic segmentation models on both datasets, we adopted pixel accuracy (PA), mean PA (MPA), intersection over union (IoU), and mean IoU (MIoU) as our evaluation metrics, for which higher values are better. For FCN, U-Net, and DLv3+, we only report the best results among all three choices of pretrained encoders unless otherwise specified. Specifically, we report the results of FCN with EN-B0, U-Net with ResNet50, and DLv3+ with EN-B0 on the WHDLD dataset, and the results of FCN with EN-B0, U-Net with EN-B0, and DLv3+ with EN-B0 on the Vaihingen dataset.

C. Semantic Segmentation Results

In this section, we separately report the semantic segmentation results of different models on the two datasets, including the overall performance and performance on each class, as shown in Tables I and II.

From the results, we have the following observations.

1) SAN and the SETR series perform the worst on all the metrics for the following reasons. Their encoders are complex, and their decoders use the most encoder features to predict

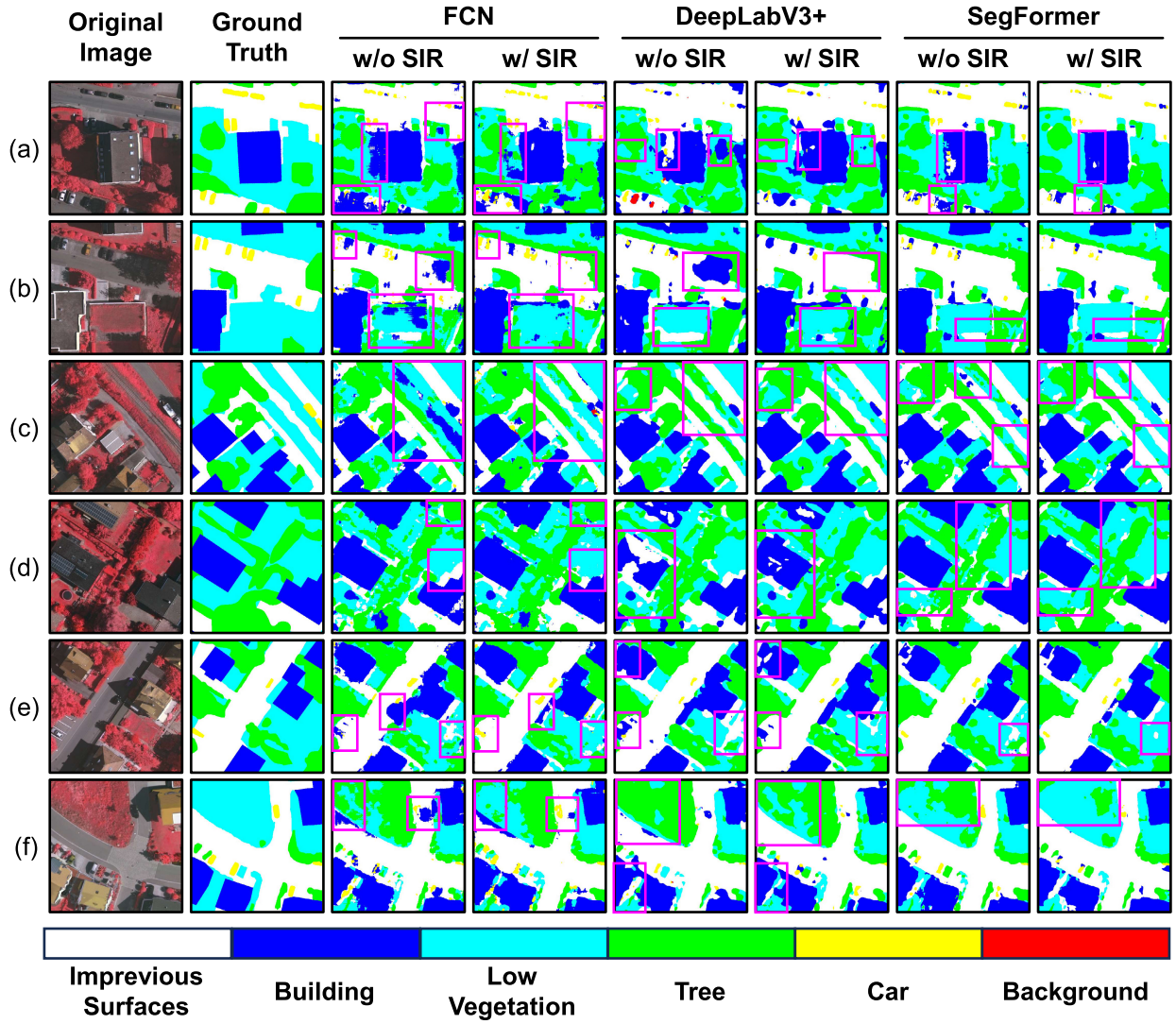


Fig. 6. Visualized semantic segmentation results from the comparison between models w/o SIR and w/ SIR on the Vaihingen dataset. In these images, areas of significant difference are delineated by purple boxes, indicating regions where notable variations occur. (a)–(f) Representative images from the test dataset.

semantic labels, which makes the confounding bias much stronger. The other transformer-based methods perform better than SAN and the SETR series because they use relatively simple decoders, e.g., linear or CNN-based ones. However, they still suffer from confounding bias and thus perform worse than models w/ SIR.

2) The overall performance of the CNN-based models is better than the transformer-based ones except for SegFormer because their encoder size is smaller and easier to train. SegFormer achieves the best performance among all baselines for the same reason, i.e., the pretrained encoder (MiT) is much lighter. Moreover, it outperforms the small CNN-based models because the transformer framework strengthens the ability to extract long-range information. Nevertheless, all these encoder–decoder models follow the proposed SCM and have the confounding bias problem, which strengthens as the decoder uses more encoder features.

3) Applying SIR to semantic segmentation models achieves significant performance improvement compared to their counterparts w/o SIR, and the combination of SIR and SegFormer achieves the best performance. It demonstrates that SIR can reduce confounding bias and improve semantic segmentation performance.

4) On the WHDLD dataset, the overall class-wise performance of SIR is the best among all methods. For example, the combination of SIR and SegFormer achieves the best performance in the building, pavement, and vegetation classes and the second-best in the water class. Similar to the results on the WHDLD dataset, the overall class-wise performance of SIR on the Vaihingen dataset is also the best. For example, the combination of SIR and SegFormer achieves the best performance in the building, low vegetation, tree, and background classes, and the combination of SIR and FCN achieves the second-best in all classes except the car class. The above observations demonstrate that SIR can improve class-wise semantic segmentation performance through de-confounding.

D. Ablation Studies

In this section, we compare the performance of semantic segmentation models trained w/ SIR, w/o SIR, and without fine-tuning (w/o FT) the pretrained encoders (decoder-only training). In order to further evaluate the effectiveness of

TABLE III

PERFORMANCE COMPARISON BETWEEN MODELS W/O FT, W/O SIR, WITHOUT SR (W/O SR), AND W/ SIR ON THE WHDLD DATASET. THE BEST RESULTS AMONG DIFFERENT VERSIONS OF THE SAME MODEL ARE IN BOLD, AND THE SECOND-BEST RESULTS ARE UNDERLINED

Model	BackBone	Version	PA	MPA	MIoU	IoU(%)					
						Bare Soil	Building	Pavement	Road	Vegetation	Water
FCN	VGG16	w/o FT	71.50	54.67	42.77	23.31	34.10	28.31	23.26	65.13	82.48
		w/o SIR	73.14	57.32	44.61	29.86	39.89	20.75	26.22	67.82	83.13
		w/o SR	74.37	65.76	49.72	32.94	43.07	**31.53**	34.38	69.17	87.25
		w/ SIR	**78.16**	**66.14**	**52.24**	**34.72**	**47.82**	31.36	**36.11**	**74.61**	**88.84**
	ResNet50	w/o FT	73.02	54.87	42.49	6.079	41.22	26.83	24.59	70.86	85.36
		w/o SIR	74.03	53.53	44.55	22.39	43.65	27.19	29.62	67.30	77.14
		w/o SR	81.45	69.84	57.51	**38.92**	**56.75**	**38.44**	44.75	77.44	88.77
		w/ SIR	**81.89**	**70.55**	**57.78**	37.60	54.21	37.83	**45.57**	**79.29**	**92.15**
	EN-B0	w/o FT	77.00	63.04	50.48	24.71	46.01	28.93	44.29	71.98	86.99
		w/o SIR	81.70	70.55	57.91	36.70	52.52	34.87	53.32	78.65	91.38
		w/o SR	82.38	72.48	59.51	**37.07**	**56.12**	37.65	54.93	79.22	92.07
		w/ SIR	**83.08**	**72.50**	**60.10**	36.36	55.98	**39.22**	**55.98**	**79.95**	**93.13**
DLv3+	VGG16	w/o FT	68.17	51.52	38.95	20.13	32.10	27.30	13.31	61.65	79.21
		w/o SIR	71.51	49.31	39.32	16.48	37.55	15.26	19.85	64.83	81.94
		w/o SR	70.78	**63.02**	45.01	29.20	43.21	20.97	33.01	68.04	75.63
		w/ SIR	**76.69**	62.99	**50.27**	**30.75**	**45.51**	**30.44**	**34.58**	**72.96**	**87.36**
	ResNet50	w/o FT	74.79	62.80	48.50	30.23	43.10	30.18	31.79	70.49	85.22
		w/o SIR	78.67	63.65	51.79	30.12	49.24	29.23	38.36	75.46	88.33
		w/o SR	80.31	69.53	55.72	**38.32**	51.69	33.09	43.53	**77.13**	90.58
		w/ SIR	**80.31**	**70.05**	**56.27**	37.50	**51.90**	**35.79**	**44.15**	76.99	**91.29**
	EN-B0	w/o FT	78.03	64.11	51.63	32.59	47.34	27.21	42.15	72.92	87.56
		w/o SIR	80.43	70.31	56.30	**37.44**	45.42	37.59	50.32	76.88	90.18
		w/o SR	81.44	70.60	57.79	34.72	**53.75**	**38.47**	51.40	77.34	91.04
		w/ SIR	**81.76**	**72.07**	**58.72**	37.24	52.55	38.16	**55.01**	**77.87**	**91.51**
SegFormer	MiT-B0	w/o FT	79.11	67.01	53.78	27.93	50.66	33.42	48.43	74.25	87.98
		w/o SIR	82.67	69.96	58.33	33.25	54.02	36.58	54.33	78.88	92.88
		w/o SR	82.99	72.07	59.96	35.86	54.00	40.66	**56.78**	79.61	92.83
		w/ SIR	**83.51**	**73.17**	**61.19**	**39.54**	**57.56**	**41.14**	55.79	**80.06**	**93.03**

the stepwise reweighting module in SIR, we also compare the performance of SIR against its ablation version without stepwise reweighting (w/o SR). To validate that our method is not restricted to specific pretrained encoders, we conducted experiments for each possible combination of semantic segmentation models and pretraining encoders.

1) Main Results: As shown in Tables III and IV and Figs. 5 and 6, we compare the performance of models w/ SIR against models w/o FT the pretrained encoders, baseline models w/o SIR, and models with only the stepwise intervention framework of SIR (w/o SR). Some observations and conclusions are as follows.

1) Fine-tuning is necessary because the performance of semantic segmentation models w/o FT the pretrained encoders is notably poor. For example, on the WHDLD dataset, the combination of FCN and EN-B0 w/o FT exhibits a 4.70% decrease in PA, a 7.51% decrease in MPA, and a 7.43% decrease in MIoU compared to the fine-tuned results. On the Vaihingen dataset, the combination of FCN and EN-B0 w/o FT exhibits a 6.25% decrease in PA, a 20.25% decrease in MPA, and an 18.55% decrease in MIoU compared to the fine-tuned results. The results demonstrate that the heterogeneity between pretraining and target datasets can introduce bias in the transferred features, validating that our motivation is reasonable.

2) Applying the stepwise intervention framework of SIR alone to the baseline models shows performance improvements across all evaluation metrics. For example, on the WHDLD dataset, applying stepwise intervention to the combination of FCN and ResNet50 improves the PA by 7.42%, MPA by 16.31%, and MIoU by 12.96%. On the Vaihingen dataset, applying stepwise intervention to the combination of FCN and EN-B0 improves the PA by 4.70%, MPA by 7.01%, and MIoU by 10.71%. The above observations demonstrate that the stepwise intervention framework of SIR can improve the semantic segmentation performance of the baseline models.

3) Applying SIR to the baseline models shows further performance improvements compared to applying the stepwise intervention framework alone. For example, on the WHDLD dataset, applying SIR to the combination of FCN and VGG16 further improves the PA by 3.79% and MIoU by 2.52%. On the Vaihingen dataset, applying SIR to the combination of DLv3+ and VGG16 further improves the MPA by 12.70% and MIoU by 9.98%. It demonstrates that the stepwise reweighting module of SIR can further improve the de-confounding performance through direct confounder balancing.

2) Illustrative Visualized Results: In order to visually highlight the performance improvement achieved by applying SIR to the baseline models, we also provide some visualized examples.

The results show that applying SIR to the baseline models shows clear improvements in the accuracy of predicted results in the selected regions. For example, on the WHDLD dataset, the baseline models have difficulty distinguishing between

TABLE IV

PERFORMANCE COMPARISON BETWEEN MODELS W/O FT, W/O SIR, W/O SR, AND W/ SIR ON THE VAIHINGEN DATASET. THE BEST RESULTS AMONG DIFFERENT VERSIONS OF THE SAME MODEL ARE IN BOLD, AND THE SECOND-BEST RESULTS ARE UNDERLINED

Model	BackBone	Version	PA	MPA	MIoU	IoU(%)					
						Imp. Surf.	Building	Low Veg.	Tree	Car	Background
FCN	VGG16	w/o FT	76.32	59.41	48.88	64.92	71.21	51.93	59.69	6.654	38.84
		w/o SIR	79.62	60.15	50.33	71.17	76.39	54.28	64.46	16.13	19.57
		w/o SR	81.16	69.50	56.91	71.98	77.08	56.58	68.82	23.04	43.94
		w/ SIR	**87.49**	**79.02**	**69.08**	**80.52**	**85.58**	**69.44**	**76.73**	**29.29**	**72.93**
	ResNet50	w/o FT	84.35	64.28	55.15	76.65	80.56	61.74	**74.64**	37.24	0.001
		w/o SIR	85.24	70.59	62.09	77.89	83.01	61.90	74.56	44.13	31.07
		w/o SR	85.89	74.53	65.13	79.58	84.36	64.47	73.81	50.26	38.30
		w/ SIR	**86.03**	**77.54**	**67.00**	**80.78**	**85.39**	**64.81**	71.54	**54.27**	**45.21**
	EN-B0	w/o FT	80.46	60.49	50.68	70.67	76.89	53.21	70.76	32.57	0.000
		w/o SIR	86.71	80.74	69.23	80.25	86.29	64.76	74.97	54.94	54.20
		w/o SR	91.41	87.79	79.94	86.58	91.87	76.31	81.94	61.09	**81.83**
		w/ SIR	**92.34**	**88.20**	**81.58**	**88.41**	**92.78**	**78.31**	**83.34**	**65.53**	81.13
DLv3+	VGG16	w/o FT	67.35	49.72	37.84	54.74	59.01	36.94	52.09	6.739	17.54
		w/o SIR	74.20	53.86	42.60	62.18	66.71	41.97	63.44	2.523	18.76
		w/o SR	78.74	59.03	49.20	69.25	75.89	54.23	63.77	11.25	20.81
		w/ SIR	**81.80**	**71.73**	**59.18**	**72.63**	**77.34**	**57.57**	**69.94**	**30.44**	**47.16**
	ResNet50	w/o FT	78.36	57.30	47.33	68.12	72.42	50.15	67.66	24.97	23.14
		w/o SIR	80.12	71.01	58.22	69.93	72.05	59.29	67.65	35.22	45.20
		w/o SR	82.97	72.78	61.30	75.37	80.43	57.59	70.54	39.84	44.03
		w/ SIR	**84.58**	**78.80**	**66.30**	**75.60**	**81.90**	**60.94**	**74.61**	**40.51**	**64.21**
	EN-B0	w/o FT	81.76	72.53	60.63	71.80	77.97	57.22	70.59	35.77	50.43
		w/o SIR	85.34	77.83	66.55	78.04	84.83	61.27	73.50	36.67	64.97
		w/o SR	86.06	78.98	68.05	78.83	85.75	64.80	73.92	**44.31**	60.68
		w/ SIR	**87.33**	**81.64**	**70.74**	**80.51**	**86.59**	**67.61**	**75.93**	43.43	**70.40**
SegFormer	MiT-B0	w/o FT	83.77	74.63	64.87	74.72	80.17	59.93	73.68	42.32	58.39
		w/o SIR	89.34	82.30	74.00	83.69	89.90	70.12	78.98	55.47	65.86
		w/o SR	90.70	85.29	78.40	86.54	91.41	75.32	78.55	**57.39**	81.19
		w/ SIR	**92.58**	**86.80**	**81.12**	**87.97**	**92.79**	**79.91**	**84.42**	55.58	**86.07**

TABLE V

EXPERIMENTAL RESULTS OF PRETRAINING WITH RS DATASETS

Model	BackBone	Version	PA	MPA	MIoU	IoU(%)					
						Imp. Surf.	Building	Low Veg.	Tree	Car	Background
FCN	V → W	w/o FT	39.18	33.61	19.64	2.020	11.69	18.50	2.328	28.19	55.13
		w/o SIR	82.18	68.35	57.59	34.29	53.15	35.73	51.74	78.29	92.33
		w/o SR	**82.58**	71.33	58.95	36.64	**54.26**	34.60	56.35	78.96	92.90
		w/ SIR	82.35	**73.86**	**59.81**	**36.79**	53.88	**39.48**	**56.52**	**79.31**	**92.90**
	W → V	w/o FT	69.97	53.04	39.27	62.02	64.98	35.35	66.74	4.263	2.253
		w/o SIR	86.23	81.41	67.13	79.64	86.40	62.70	74.56	48.48	50.99
		w/o SR	90.29	84.66	77.59	85.55	90.18	72.73	**80.24**	57.21	79.61
		w/ SIR	**90.44**	**86.91**	**78.81**	**85.83**	**91.01**	**72.93**	80.02	**62.83**	**80.27**
DLv3+	V → W	w/o FT	73.95	58.68	45.97	28.65	41.47	31.26	26.42	68.50	79.53
		w/o SIR	80.88	67.61	56.10	34.87	48.99	**35.50**	50.43	76.76	90.08
		w/o SR	81.12	70.78	57.25	**37.80**	53.05	34.04	51.10	77.12	90.39
		w/ SIR	**82.00**	**71.33**	**58.40**	37.56	**54.64**	34.47	**54.11**	**78.20**	**91.43**
	W → V	w/o FT	80.26	62.41	52.60	69.97	75.92	53.18	69.86	25.03	21.63
		w/o SIR	85.22	77.65	67.03	77.75	83.88	62.88	73.51	41.15	62.99
		w/o SR	86.19	79.65	67.68	78.60	**85.24**	65.06	**74.75**	40.45	62.01
		w/ SIR	**86.26**	**83.15**	**70.46**	**78.61**	84.90	**66.18**	74.30	**48.17**	**70.61**
SegFormer	V → W	w/o FT	77.19	61.96	50.44	27.94	45.37	31.30	40.99	71.99	85.07
		w/o SIR	82.48	69.09	58.34	36.85	53.24	37.08	52.30	78.45	92.15
		w/o SR	82.64	71.65	59.17	34.62	**55.77**	**37.88**	55.45	79.11	92.18
		w/ SIR	**82.99**	**72.95**	**60.13**	**39.60**	53.51	37.87	**57.60**	**79.31**	**92.89**
	W → V	w/o FT	79.81	66.89	56.09	70.53	73.55	51.53	69.58	33.72	37.65
		w/o SIR	88.64	83.84	73.73	82.84	88.97	69.42	77.22	53.51	70.41
		w/o SR	90.25	85.05	76.95	84.98	91.16	72.34	79.98	53.82	79.42
		w/ SIR	**91.21**	**88.10**	**79.46**	**86.63**	**91.52**	**76.55**	**80.93**	**60.59**	**80.55**

the pavement, building, road, and bare soil classes. However, after applying SIR to these models, the results are corrected, which demonstrates the effectiveness of SIR. On the Vaihingen dataset, the baseline models cannot distinguish well between

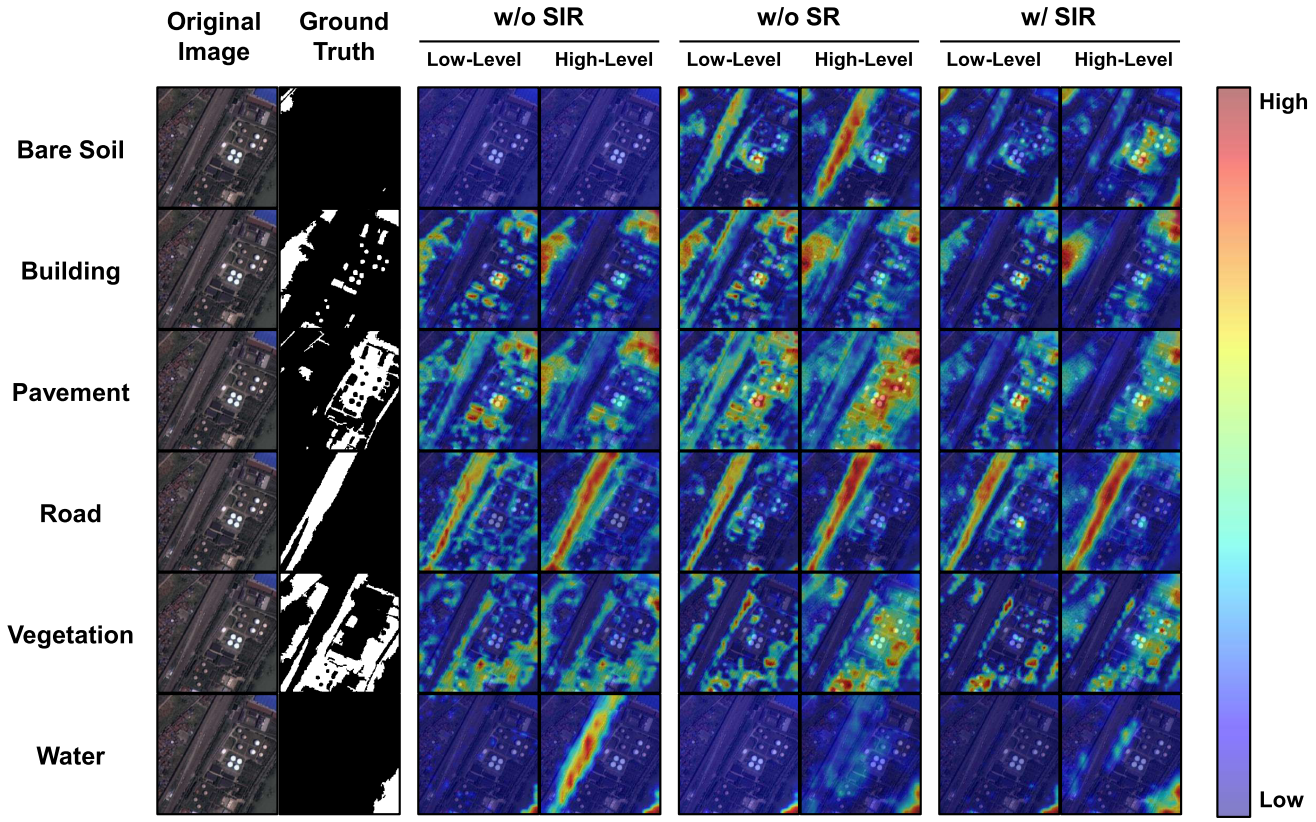


Fig. 7. Grad-CAM of an example image from the WHDLD dataset. The low-level and high-level column titles denote the low-level and high-level features, respectively. As illustrated in the colorbar, the gradient from blue to red signifies the transition from low to high semantic attention levels.

TABLE VI
COMPARISON OF TRAINING AND INFERENCE TIME BETWEEN MODELS W/ SIR, W/O SR, AND W/O SIR ON THE WHDLD DATASET

Model	Version	Training Time↓	FPS↑
FCN	w/o SIR	24.40s	155.2
	w/o SR	32.44s	158.1
	w/ SIR	39.84s	156.9
DLv3+	w/o SIR	22.71s	139.3
	w/o SR	28.54s	141.2
	w/ SIR	32.97s	139.4
SegFormer	w/o SIR	26.42s	135.0
	w/o SR	34.72s	135.2
	w/ SIR	42.29s	134.4

TABLE VII
COMPARISON OF TRAINING AND INFERENCE TIME BETWEEN MODELS W/ SIR, W/O SR, AND W/O SIR ON THE VAIHINGEN DATASET

Model	Version	Training Time↓	FPS↑
FCN	w/o SIR	20.46s	151.3
	w/o SR	25.15s	152.4
	w/ SIR	30.36s	150.1
DLv3+	w/o SIR	19.70s	134.8
	w/o SR	22.56s	129.2
	w/ SIR	25.73s	132.8
SegFormer	w/o SIR	23.65s	129.0
	w/o SR	29.47s	133.1
	w/ SIR	33.93s	132.9

the impervious surfaces and building classes, as well as the tree and low vegetation classes, which are also improved by applying SIR.

3) Pretraining With RS Images: In Section I, we have discussed the heterogeneity problem even when pretraining and target datasets are both RS data. Therefore, we conducted experiments using RS datasets for pretraining. Specifically, we pretrained with the WHDLD dataset and fine-tuned it on the Vaihingen dataset, as well as pretrained with the Vaihingen dataset and fine-tuned it on the WHDLD dataset. From the results reported in Table V, we have the following observations and conclusions.

1) Small-scale RS pretraining datasets offer less knowledge than large-scale general pretraining datasets w/o FT. Compared to results obtained from pretraining on large-scale general datasets w/o FT, the performance of models pretrained on small-scale RS datasets w/o FT is noticeably inferior. For example, comparing the FCN and EN-B0 combination pretrained on the Vaihingen dataset versus pretrained on the ImageNet-1K dataset, the performance on the WHDLD dataset significantly degraded: PA decreases by 37.82%, MPA by 29.43%, and MIoU by 30.84%. Similar noticeable decreases are also observed in the Vaihingen dataset. This observation demonstrates that the scale of pretraining data is crucial.

2) Similar to observations in the main results, applying SIR to the baseline model enhances performance. For example, applying SIR to the combination of FCN and EN-B0 improves the MPA by 5.51% and MIoU by 2.22% on the WHDLD dataset, and improves the PA by 4.21%, MPA by 5.5%, and MIoU by 11.68% on the Vaihingen dataset. This demonstrates the heterogeneity among RS datasets, and our method can effectively address this problem.

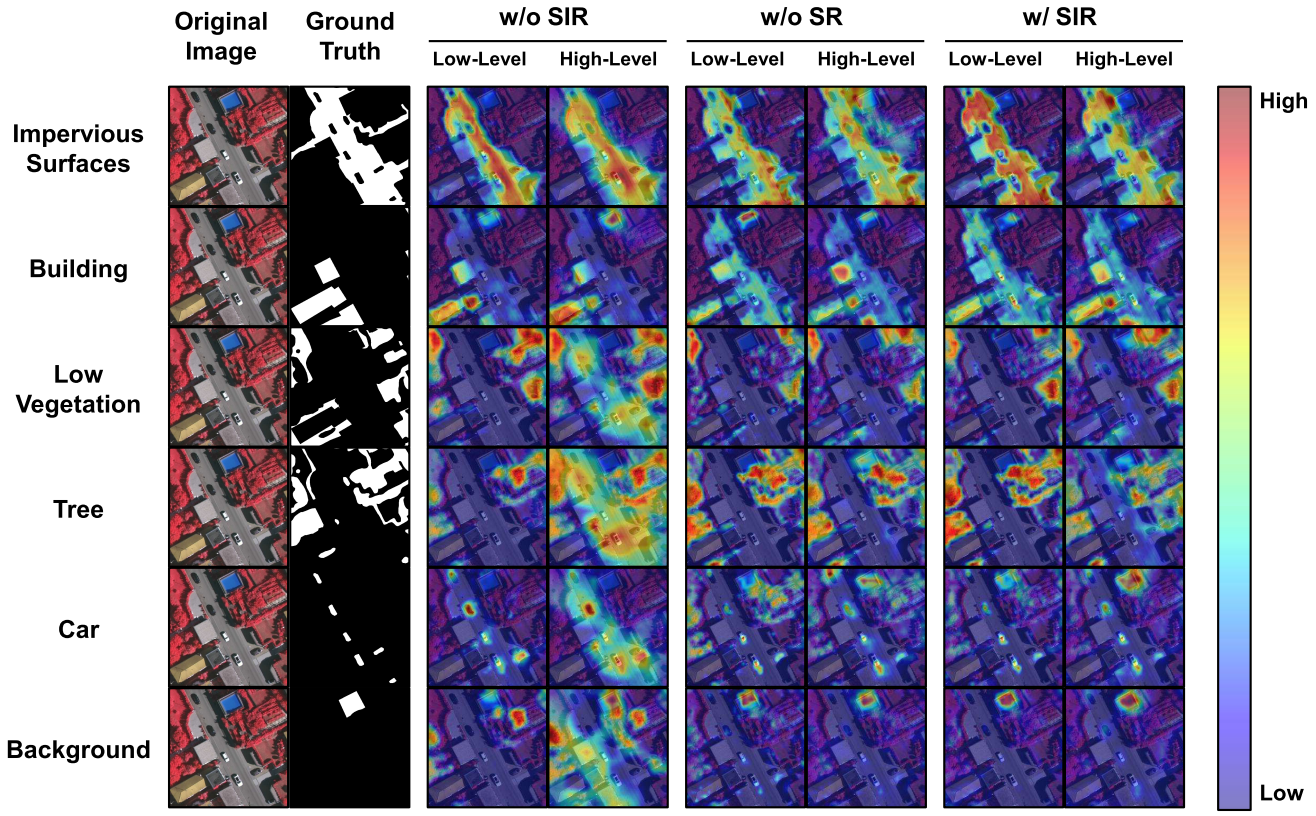


Fig. 8. Grad-CAM of an example image from the Vaihingen dataset. The low-level and high-level column titles denote the low-level and high-level features, respectively. As illustrated in the colorbar, the gradient from blue to red signifies the transition from low to high semantic attention levels.

E. Efficiency

In this section, we aim to evaluate the training and inference efficiency of SIR. To evaluate training efficiency, we recorded the average training time per epoch for models w/o SIR, with the stepwise intervention framework of SIR alone (w/o SR), and w/ SIR. To evaluate inference efficiency, following previous works [63], [70], [71], we calculated the frames per second (FPS) of each model, which is the number of images processed by a model per second.

As reported in Tables VI and VII, because the stepwise intervention framework of SIR needs individual loss computation at each step, the training time of models w/o SR increases by around 30% on the WHDLD dataset and around 25% on the Vaihingen dataset compared to the baseline models w/o SIR. Due to the incorporation of a frozen pretrained model within the stepwise reweighting module for weight computation, the training time for models w/ SIR additionally increases by around 20% on the WHDLD dataset and around 15% on the Vaihingen dataset compared to the models w/o SR. Nonetheless, upon completion of model training, since SIR does not alter the parameter count of the baseline model, we observe no significant difference in FPS among models w/ SIR, w/o SR, and w/o SIR.

F. Feature Visualization

In this section, we evaluate the ability of SIR to address the confounding bias between encoder layers and to learn different levels of features better. Here, we take the combination of FCN and EN-B0 as an example and use Grad-CAM [72] to visualize the low-level and high-level features of the standalone baseline model (w/o SIR), model with only the stepwise intervention framework of SIR (w/o SR), and model w/ SIR, respectively. Grad-CAM can produce a heat map that highlights the critical regions in an image for a specific feature layer to predict a target semantic label. As the FCN model takes the output features of three encoder layers for semantic segmentation in total, we take the first encoder layer of FCN (i.e., the third feature layer of EN-B0) as the low-level feature and the last encoder layer of FCN (i.e., the seventh feature layer of EN-B0) as the high-level feature to visualize. The feature visualization results on the WHDLD and Vaihingen datasets are shown in Figs. 7 and 8, respectively.

From the results, we have the following observations.

1) For the baseline, i.e., the standalone FCN model, the low-level features are acceptable. However, the high-level features seriously mislead the semantic segmentation performance of the water class on the WHDLD dataset and the building, low vegetation, tree, car, and background classes on the Vaihingen dataset. For example, on the WHDLD dataset, except for the bare soil class, in which both the low-level and high-level features fail to find any information from the image, the highlighted regions of the low-level features are basically where the corresponding semantic labels are located. However, when predicting the water class, the high-level features mistakenly overemphasize the road regions but downplay the correct water regions. On the Vaihingen dataset, the high-level features make even more mistakes, i.e., when predicting

the low vegetation, tree, car, and background classes, the high-level features show an excessive focus on the impervious surfaces regions, resulting in incorrect predictions. The above observations demonstrate that the baseline model suffers from the confounding bias between each level of features introduced by the pretraining knowledge as the high-level features fail to learn correct patterns from the well-learned low-level features.

2) For SIR and its ablation version, the highlights of the low-level features more accurately reflect the regions of the corresponding semantic classes. For example, on the WHDLD dataset, the low-level features of the ablation model w/o SR detect the correct bare soil region, which the baseline model fails to do. Moreover, the low-level features of SIR further remove the noises induced by the road regions and show a more precise localization. On the Vaihingen dataset, SIR and its ablation version add more helpful information on the tree class to the low-level features and remove noises in predicting the background class.

3) The high-level features learned by SIR and its ablation version show a more coherent pattern than the baseline models and maintain consistency with the low-level features. For example, on the WHDLD dataset, the high-level features of the ablation model w/o SR learn more helpful information about the pavement and vegetation classes. When predicting the pavement and building classes, they reduce the noises of the low-level features induced by the road class. SIR further makes the prediction of the bare soil and water class more accurate, particularly in terms of localization and scale. On the Vaihingen dataset, the high-level features of SIR and its ablation version learn more helpful information based on the low-level features and are less susceptible to noises induced by the impervious surfaces class when predicting the low vegetation, tree, car, and background classes. The above observations demonstrate that SIR can better learn both the low-level and high-level features by addressing the confounding bias between each encoder layer, thus predicting semantic labels more precisely.

V. Conclusion and Future Work

In this article, we study the challenge of making semantic segmentation of RS images benefit more from pretraining, where the pretrained model fails to learn general features appropriate to RS datasets because of data heterogeneity. We present a novel formulation of the above problem from a causal perspective, considering the pretraining knowledge as a confounder in the learning process from the previous layers to the next layer of the encoder. Based on the causal formulation, we propose a novel method, namely, SIR, to reduce the confounding bias and improve semantic segmentation performance. Extensive experimental results show that applying SIR to encoder–decoder structured semantic segmentation models achieves performance improvements and does make the models learn low-level and high-level features better, demonstrating the effectiveness of our method.

The main limitation of the proposed methods is that SIR needs prior knowledge about the data used for pretraining. However, when such knowledge is unavailable, we can still use an ablation version of SIR that only utilizes the stepwise intervention framework or make reasonable assumptions on the pretraining dataset. Another limitation is that SIR increases model training time. However, once trained, the inference speed of the model remains unaffected because SIR modifies the training process without introducing new parameters. Future research could explore the extension of the proposed causality-guided problem formulation to a broader range of semantic segmentation applications, such as training domain-specific large vision models.

Acknowledgment

The Vaihingen Dataset was provided by the German Society for Photogrammetry, Remote Sensing and Geoinformation (DGPF) [60]: http://www.ifp.uni-stuttgart.de/dgpf/DKEP-Allg.html.

References

[1] P. Wei, D. Chai, T. Lin, C. Tang, M. Du, and J. Huang, "Large-scale rice mapping under different years based on time-series Sentinel-1 images using deep semantic segmentation model," *ISPRS J. Photogramm. Remote Sens.*, vol. 174, pp. 198–214, Apr. 2021.

[2] Q. Shen, J. Huang, M. Wang, S. Tao, R. Yang, and X. Zhang, "Semantic feature-constrained multitask Siamese network for building change detection in high-spatial-resolution remote sensing imagery," *ISPRS J. Photogramm. Remote Sens.*, vol. 189, pp. 78–94, Jul. 2022.

[3] J. Ge, H. Tang, N. Yang, and Y. Hu, "Rapid identification of damaged buildings using incremental learning with transferred data from historical natural disaster cases," *ISPRS J. Photogramm. Remote Sens.*, vol. 195, pp. 105–128, Jan. 2023.

[4] X. Zhang, W. Yu, M.-O. Pun, and W. Shi, "Cross-domain landslide mapping from large-scale remote sensing images using prototype-guided domain-aware progressive representation learning," *ISPRS J. Photogramm. Remote Sens.*, vol. 197, pp. 1–17, Mar. 2023.

[5] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 3431–3440.

[6] J. Yosinski, J. Clune, Y. Bengio, and H. Lipson, "How transferable are features in deep neural networks?" in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 27, 2014, pp. 3320–3328.

[7] Q. Yuan et al., "Deep learning in environmental remote sensing: Achievements and challenges," *Remote Sens. Environ.*, vol. 241, May 2020, Art. no. 111716.

[8] A. Kirillov et al., "Segment anything," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, Oct. 2023, pp. 4015–4026.

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

[10] Z. Yue, P. Zhou, R. Hong, H. Zhang, and Q. Sun, "Few-shot learner parameterization by diffusion time-steps," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2024, pp. 23263–23272.

[11] Y. Wei, Y. Zheng, and Q. Yang, "Transfer knowledge between cities," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Aug. 2016, pp. 1905–1914.

[12] Y. Tian, Y. Wang, D. Krishnan, J. B. Tenenbaum, and P. Isola, "Rethinking few-shot image classification: A good embedding is all you need?" in *Computer Vision—ECCV 2020*. Glasgow, U.K.: Springer, 2020, pp. 266–282.

[13] P. Cui and S. Athey, "Stable learning establishes some common ground between causal inference and machine learning," *Nature Mach. Intell.*, vol. 4, no. 2, pp. 110–115, Feb. 2022.

[14] M. A. Hernán and J. M. Robins, *Causal Inference: What If*. Boca Raton, FL, USA: CRC Press, 2020.

[15] S. Kornblith, J. Shlens, and Q. V. Le, "Do better ImageNet models transfer better?" in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 2656–2666.

[16] J. D. Angrist, G. W. Imbens, and D. B. Rubin, "Identification of causal effects using instrumental variables," *J. Amer. Stat. Assoc.*, vol. 91, no. 434, pp. 444–455, Jun. 1996.

[17] 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*. Munich, Germany: Springer, Oct. 2015, pp. 234–241.

[18] Z. Zhou et al., "UNet++: Redesigning skip connections to exploit multiscale features in image segmentation," *IEEE Trans. Med. Imag.*, vol. 39, no. 6, pp. 1856–1867, Dec. 2019.

[19] V. Badrinarayanan, A. Kendall, and R. Cipolla, "SegNet: A deep convolutional encoder-decoder architecture for image segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 12, pp. 2481–2495, Dec. 2017.

[20] K. Sun, B. Xiao, D. Liu, and J. Wang, "Deep high-resolution representation learning for human pose estimation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2019, pp. 5693–5703.

[21] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Semantic image segmentation with deep convolutional nets and fully connected CRFs," in *Proc. Int. Conf. Learn. Represent.*, 2015.

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

[23] L. C. Chen, G. Papandreou, and I. Kokkinos, "DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 4, pp. 834–848, Jun. 2017.

[24] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder–decoder with atrous separable convolution for semantic image segmentation," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 801–818.

[25] H. Pan, Y. Hong, W. Sun, and Y. Jia, "Deep dual-resolution networks for real-time and accurate semantic segmentation of traffic scenes," *IEEE Trans. Intell. Transp. Syst.*, vol. 24, no. 3, pp. 3448–3460, Mar. 2023.

[26] X. Wang, K. Tan, P. Du, C. Pan, and J. Ding, "A unified multiscale learning framework for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 4508319.

[27] X. Wang, K. Tan, P. Du, B. Han, and J. Ding, "A capsule-vectored neural network for hyperspectral image classification," *Knowl.-Based Syst.*, vol. 268, May 2023, Art. no. 110482.

[28] X. Wang et al., "Double U-Net (W-Net): A change detection network with two heads for remote sensing imagery," *Int. J. Appl. Earth Observ. Geoinf.*, vol. 122, Aug. 2023, Art. no. 103456.

[29] I. Bello, B. Zoph, Q. Le, A. Vaswani, and J. Shlens, "Attention augmented convolutional networks," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 3286–3295.

[30] S. Zheng et al., "Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 6881–6890.

[31] A. Dosovitskiy, "An image is worth 16×16 words: Transformers for image recognition at scale," in *Proc. Int. Conf. Learn. Represent.*, 2021.

[32] W. Wang et al., "Pyramid vision transformer: A versatile backbone for dense prediction without convolutions," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 568–578.

[33] E. Xie et al., "SegFormer: Simple and efficient design for semantic segmentation with transformers," in *Proc. Adv. Neural Inf. Process. Sys. (NIPS)*, vol. 34, Dec. 2021, pp. 12077–12090.

[34] R. Strudel, R. Garcia, I. Laptev, and C. Schmid, "Segmenter: Transformer for semantic segmentation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 7262–7272.

[35] B. Cheng, A. Schwing, and A. Kirillov, "Per-pixel classification is not all you need for semantic segmentation," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 34, 2021, pp. 17864–17875.

[36] B. Cheng, I. Misra, A. G. Schwing, A. Kirillov, and R. Girdhar, "Masked-attention mask transformer for universal image segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 1290–1299.

[37] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, and S. Xie, "A ConvNet for the 2020s," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2022, pp. 11976–11986.

[38] 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," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 35, 2022, pp. 1140–1156.

[39] R. Ji, K. Tan, X. Wang, C. Pan, and L. Xin, "PASSNet: A spatial–spectral feature extraction network with patch attention module for hyperspectral image classification," *IEEE Geosci. Remote Sens. Lett.*, vol. 20, pp. 1–5, 2023.

[40] M. Xu, Z. Zhang, F. Wei, H. Hu, and X. Bai, "Side adapter network for open-vocabulary semantic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 2945–2954.

[41] J. Pearl and D. Mackenzie, *The Book of Why: The New Science of Cause and Effect*. New York, NY, USA: Basic Books, 2018.

[42] F. Li, K. L. Morgan, and A. M. Zaslavsky, "Balancing covariates via propensity score weighting," *J. Amer. Stat. Assoc.*, vol. 113, no. 521, pp. 390–400, Jan. 2018.

[43] S. Athey, G. W. Imbens, and S. Wager, "Approximate residual balancing: Debiased inference of average treatment effects in high dimensions," *J. Roy. Stat. Soc. B, Stat. Methodol.*, vol. 80, no. 4, pp. 597–623, Sep. 2018.

[44] K. Kuang et al., "Treatment effect estimation via differentiated confounder balancing and regression," *ACM Trans. Knowl. Discovery Data*, vol. 14, no. 1, pp. 1–25, Feb. 2020.

[45] F. Johansson, U. Shalit, and D. Sontag, "Learning representations for counterfactual inference," in *Proc. Int. Conf. Mach. Learn.*, 2016, pp. 3020–3029.

[46] U. Shalit, F. D. Johansson, and D. Sontag, "Estimating individual treatment effect: Generalization bounds and algorithms," in *Proc. Int. Conf. Mach. Learn.*, 2017, pp. 3076–3085.

[47] A. Wu et al., "Learning decomposed representations for treatment effect estimation," *IEEE Trans. Knowl. Data Eng.*, vol. 35, no. 5, pp. 4989–5001, May 2023.

[48] C. Louizos, U. Shalit, J. M. Mooij, D. Sontag, R. Zemel, and M. Welling, "Causal effect inference with deep latent-variable models," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 6446–6456.

[49] J. Yoon, J. Jordon, and M. Van Der Schaar, "GANITE: Estimation of individualized treatment effects using generative adversarial nets," in *Proc. Int. Conf. Learn. Represent.*, 2018.

[50] V. Agarwal, R. Shetty, and M. Fritz, "Towards causal VQA: Revealing and reducing spurious correlations by invariant and covariant semantic editing," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 9687–9695.

[51] Q. Tian, K. Kuang, K. Jiang, F. Liu, Z. Wang, and F. Wu, "Confounder-GAN: Protecting image data privacy with causal confounder," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 35, 2022, pp. 32789–32800.

[52] Z. Yue, H. Zhang, Q. Sun, and X.-S. Hua, "Interventional few-shot learning," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 33, 2020, pp. 2734–2746.

[53] D. Zhang, H. Zhang, J. Tang, X.-S. Hua, and Q. Sun, "Causal intervention for weakly-supervised semantic segmentation," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 33, 2020, pp. 655–666.

[54] J. Pearl, *Causality*. Cambridge, U.K.: Cambridge Univ. Press, 2009.

[55] R. Guo, L. Cheng, J. Li, P. R. Hahn, and H. Liu, "A survey of learning causality with data: Problems and methods," *ACM Comput. Surv.*, vol. 53, no. 4, pp. 1–37, 2020.

[56] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. Int. Conf. Learn. Represent.*, 2015.

[57] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[58] Z. Shao, K. Yang, and W. Zhou, "Performance evaluation of single-label and multi-label remote sensing image retrieval using a dense labeling dataset," *Remote Sens.*, vol. 10, no. 6, p. 964, Jun. 2018.

[59] Z. Shao, W. Zhou, X. Deng, M. Zhang, and Q. Cheng, "Multilabel remote sensing image retrieval based on fully convolutional network," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 13, pp. 318–328, 2020.

[60] M. Cramer, "The DGPF-test on digital airborne camera evaluation overview and test design," *Photogrammetrie-Fernerkundung-Geoinf.*, vol. 2010, no. 2, pp. 73–82, May 2010.

[61] N. Audebert, B. Le Saux, and S. Lefèvre, "Beyond RGB: Very high resolution urban remote sensing with multimodal deep networks," *ISPRS J. Photogramm. Remote Sens.*, vol. 140, pp. 20–32, Jun. 2018.

[62] MMS Contributors. (2020). *MMSegmentation: OpenMMLab Semantic Segmentation Toolbox and Benchmark*. [Online]. Available: https://github.com/open-mmlab/mmsegmentation

[63] X. He, Y. Zhou, J. Zhao, D. Zhang, R. Yao, and Y. Xue, "Swin transformer embedding UNet for remote sensing image semantic segmentation," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 4408715.

[64] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 248–255.
[65] M. Tan and Q. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in *Proc. Int. Conf. Mach. Learn.*, 2019, pp. 6105–6114.
[66] B. Zhou et al., "Semantic understanding of scenes through the ADE20K dataset," *Int. J. Comput. Vis.*, vol. 127, no. 3, pp. 302–321, Mar. 2019.
[67] H. Caesar, J. Uijlings, and V. Ferrari, "COCO-stuff: Thing and stuff classes in context," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2018, pp. 1209–1218.
[68] M. Everingham et al., "The PASCAL visual object classes challenge: A retrospective," *Int. J. Comput. Vis.*, vol. 111, no. 1, pp. 98–136, Jan. 2015.
[69] M. Cordts et al., "The cityscapes dataset for semantic urban scene understanding," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 3213–3223.
[70] K. Chen et al., "MMDetection: Open MMLab detection toolbox and benchmark," 2019, *arXiv:1906.07155*.
[71] W. Zhang, J. Pang, K. Chen, and C. C. Loy, "K-Net: Towards unified image segmentation," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 13, 2021, pp. 10326–10338.
[72] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 618–626.

Shuting Shi received the B.S. degree from the School of Earth Sciences, Zhejiang University, Hangzhou, China, in 2022, where she is currently pursuing the M.S. degree.

Her main research interests include semantic segmentation of remote sensing images, spatial–temporal analysis, and deep learning.

Baohong Li received the B.S. degree from the College of Computer Science and Technology, Zhejiang University, Hangzhou, China, in 2022, where he is currently pursuing the Ph.D. degree.

His main research interests include causal inference, trustworthy machine learning, and data mining.

Laifu Zhang is currently pursuing the Ph.D. degree in remote sensing and geographical information systems with Zhejiang University, Hangzhou, China.

His research interests include nighttime light, urban remote sensing, downscaling, deep learning, and high-resolution data fusion and reconstruction.

Kun Kuang received the Ph.D. degree in computer science from Tsinghua University, Beijing, China, in 2019.

He was a Visiting Scholar with Prof. Susan Athey's Group, Stanford University, Stanford, CA, USA. He is currently an Associate Professor with the College of Computer Science and Technology, Zhejiang University, Hangzhou, China. He has published more than 100 papers in prestigious conferences and journals in data mining and machine learning, including *Cell Patterns*, IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, ICML, NeurIPS, KDD, ICDE, WWW, MM, DMKD, and *Engineering*. His main research interests include causal inference, machine learning, and data mining.

Dr. Kuang received the ACM SIGAI China Rising Star Award in 2022.

Sensen Wu received the Ph.D. degree in cartography and geographic information systems from Zhejiang University, Hangzhou, China, in 2018.

He is currently working as an Associate Professor with the School of Earth Sciences, Zhejiang University. His research interests include spatial–temporal analysis, remote sensing, and deep learning.

Tian Feng (Member, IEEE) received the B.Sc. degree from Zhejiang University (ZJU), Hangzhou, China, in 2012, and the Ph.D. degree from the Singapore University of Technology and Design (SUTD), Singapore, in 2017.

He was a Visiting Ph.D. Student with the University of Massachusetts Boston, Boston, MA, USA, from 2015 to 2016. He is currently a Research Associate Professor with the School of Software Technology, ZJU. His research interests include computer graphics, computer vision, and human–computer interaction.

Yiming Yan received the Ph.D. degree in remote sensing and geographic information systems from Zhejiang University, Hangzhou, China, in 2021.

He is currently working as a Technician with the School of Earth Sciences, Zhejiang University. His research interests include spatial–temporal analysis and big data mining.

Zhenhong Du (Member, IEEE) received the Ph.D. degree in cartography and geographic information science from Zhejiang University, Hangzhou, China, in 2010.

From 2013 to 2019, he was an Associate Professor with Zhejiang University, where he is currently a Professor with the School of Earth Sciences. He is also the Vice-Dean of the School of Earth Sciences and the Director of the Institute of Geography and Spatial Information, Zhejiang University. He is also the Director of the Zhejiang Provincial Key Laboratory of Geographic Information System, Zhejiang University. His research interests include remote sensing and geographic information science, spatial–temporal big data, artificial intelligence, and big data and Earth systems.