A Review of Semi-Supervised Learning Frameworks for Semantic Segmentation of Remote Sensing Images

Zekun Li, Weidong Sun*, Lin Zheng, Yichong Wang

School of Mathematics and Statistics, Changchun University of Science and Technology, Changchun, 130022, China
**Corresponding author: sunweidong@mails.cust.edu.cn*

Abstract: *The purpose of this paper is to review recent advances in semi-supervised learning based on semisupervised learning in the field of semantic segmentation of remotely sensed images, with a special focus on semi-supervised learning methods based on generative and discriminative models. The definition, task and application background of semantic segmentation of remote sensing images are first introduced, followed by an overview of the limitations of traditional supervised learning methods in this field. Then the application of semi-supervised learning framework in semantic segmentation of remote sensing images is discussed in detail, including methods based on generative model and discriminative model. In the generative model, the generative adversarial network (GAN) method and the variational autoencoder (VAE) method are discussed in detail, and their applications in semantic segmentation of remote sensing images are explored. After that, we focus on self-training, pseudo-labeling and consistency training methods in discriminative modeling, discuss their principles, advantages and limitations, and explore the effectiveness of their applications in semantic segmentation of remote sensing images. Finally, the challenges of current research and future directions are summarized to provide reference and outlook for further research in this field.*

Keywords: *Remote Sensing Images; Emantic Segmentation; Semi-supervised Learning; Generative Models; Discriminative Models*

1. Introduction

Semantic segmentation of remote sensing images is a key task for accurate recognition and localization of features in remote sensing images, aiming to accurately classify each pixel into predefined feature categories, such as buildings, water bodies, roads, etc. With the development and wide application of remote sensing technology, there is an urgent need for efficient and accurate feature recognition and segmentation. Traditional supervised learning methods face challenges such as high cost of data labeling, skewed datasets and poor model generalization ability, while semi-supervised learning methods provide an effective solution to these problems by using unlabeled data and a small amount of labeled data to train models. The task of semantic segmentation of remote sensing images involves feature recognition, boundary localization, and monitoring of feature changes, which is widely applied but still faces challenges such as data quality, data volume and sample scarcity, and feature complexity, etc., and the researchers propose techniques such as deep learning-based semantic segmentation models and semi-supervised learning methods to improve accuracy and robustness.

2. A semi-supervised learning approach based on generative modeling

Semi-supervised learning methods based on generative models use generative models to model the distribution of unlabeled data, and then use generative models to generate pseudo-labels or auxiliary labels to expand the labeled dataset in order to improve the performance of supervised learning models[1].In the field of semantic segmentation of remote sensing images, a number of important advances have been made in semi-supervised learning methods based on generative models. In this section, several typical generative model-based semi-supervised learning methods and their applications in semantic segmentation of remote sensing images are reviewed.

2.1. Generative Adversarial Network (GAN) method

2.1.1. Principles of Generative Adversarial Network (GAN) method

Generative Adversarial Network (GAN) is a deep learning model proposed in 2014 by Goodfellow et al. It contains two neural networks, the generator and the discriminator[2]. The generator aims to generate samples that are similar to real data, while the discriminator tries to distinguish between generated and real samples. Through adversarial training, the two networks compete with each other, and eventually the generator is able to generate high-quality samples and the discriminator is better at distinguishing between real and generated samples[3].

The training process of GAN can be represented by the following loss function:

Generator Loss (GL) function:

$$L_G = -IE_{z\sim p(z)}[log\, D\,(G(z))] \tag{1}$$

where$G(z)$ is the random vector in the potential space through which the generator$z$ the generated sample, and$D(G(z))$ is the output of the discriminator for the generated sample. The goal of the generator is to minimize the generator loss, i.e., to make it more difficult for the discriminator to distinguish the generated samples.

Discriminator Loss Function (DLF):

$$L_D = -IE_{x\sim P_{data}(x)}[log\, D\,(x)] - IE_{z\sim p(z)}[log(\,1 - D(z)))] \tag{2}$$

Where x is the true sample, and $P_{data}(x)$ is the distribution of the true data. The goal of the discriminator is to maximize the discriminator loss, i.e., to correctly distinguish between real and generated samples.

2.2. Variational Auto-Encoder (VAE) approach

Variational Auto-Encoder (VAE) is a generative model that incorporates the ideas of auto-encoder and probabilistic inference. It was proposed by Kingma and Welling in 2013[4]. The core idea of VAE is to learn a latent space in which each point corresponds to an interpretable latent variable so that new data samples can be generated by sampling the latent space. The process consists of two main parts: an encoder and a decoder.

1) Encoder (Encoder): the encoder maps the input data to a probability distribution in the latent space. Usually, the encoder maps the input data to the mean and variance parameters of the latent space, expressed as a Gaussian distribution in the latent space.

2) Decoder (Decoder): The decoder maps the points sampled from the potential space back to the data space, generating samples similar to the input data.

The process of training the VAE involves maximizing the edge log likelihood of the observations (ELBO), a value that consists of two components: the reconstruction loss and the KL scatter.

1) Reconstruction Loss (Reconstruction Loss): represents the difference between the data generated by the decoder and the original input data, usually using the cross entropy or mean square error as the reconstruction loss function.

2) KL Divergence (Kullback-Leibler Divergence): a measure of the difference between the potential spatial distribution generated by the encoder and the standard normal distribution, used to ensure the smoothness of the potential space.

The training process of VAE can be represented as follows:

$$L_{VAE} = IE_{q(z|x)}[log\, p\,(x|z)] - KL(q(z|x))||p(x) \tag{3}$$

Where $q(z|x)$ is the potential spatial distribution of the encoder output given the input data$x$ conditioned on the latent spatial distribution of the encoder output, and$p(x|z)$ is the distribution of the data generated by the decoder, and $p(x)$ is the standard normal distribution, and KL denotes the KL dispersion.

3. Semi-supervised learning methods based on discriminative models

Semi-supervised learning methods based on discriminative models aim to improve the performance of supervised learning models by exploiting the relationship between labeled and unlabeled data, and labeling or assisting in labeling the unlabeled data through discriminative models to expand the labeled dataset. In the field of semantic segmentation of remote sensing images, semi-supervised learning methods based on discriminative models have made a series of important advances. Several typical semi-supervised learning methods based on discriminative models and their applications in semantic segmentation of remote sensing images will be introduced in the following.

3.1. Self-training methods

Self-training (Self-training) is a semi-supervised learning method, whose basic idea is to increase the number and diversity of training data by training alternately and iteratively between labeled and unlabeled data, using the model to make predictions on the unlabeled data, and adding high-confidence prediction results as pseudo-labels to the training data in order to improve the model's performance. Self-training methods are usually based on discriminative models and are suitable for a variety of supervised learning tasks, including image classification, target detection, semantic segmentation, etc.

Self-training methods are also widely used in remote sensing image semantic segmentation tasks. Semantic segmentation of remote sensing images refers to assigning each pixel in a remote sensing image to a predefined semantic category, such as buildings, roads, vegetation, etc., which is one of the important tasks in the field of remote sensing image processing. Traditional supervised learning methods usually require a large amount of labeled data to train the model, but labeling remote sensing image data is very time-consuming and labor-intensive, resulting in a relative scarcity of labeled data, so semi-supervised learning methods such as self-training become an effective way to solve this problem.

The process of the self-training method in semantic segmentation of remote sensing images includes data preparation, model initialization and training, using the model to predict unlabeled data, generating pseudo-labels, expanding the training data and re-training the model, and iterative training. This method improves model performance and generalization ability by effectively using unlabeled data, reduces labeling cost, and provides new ideas and methods for remote sensing image processing.

3.2. Consistency Training Methods

3.2.1. Principles of Coherence Training

Consistency Training (CT) is a semi-supervised learning method that is based on discriminative models and improves model performance by utilizing unlabeled data. The core idea of Consistency Training is to improve the generalization performance of the model by exploiting the consistency of data. Its basic assumption is that if two inputs have similar outputs in the model, then they should also be similar in the input space. Consistency training achieves effective utilization of unlabeled data by constructing additional loss functions on unlabeled data, comparing the model's predictions under different inputs, and encouraging the model to make similar inputs have similar outputs.

3.2.2. Consistency Training Algorithm

The core of the consistency training algorithm is to ensure that the model produces consistent outputs under different input conditions by introducing a consistency loss function between labeled and unlabeled data. This means that for similar inputs, the model should generate similar outputs, i.e., be consistent in terms of model outputs.

The continuity and consistency of the data is exploited to improve the generalization ability of the model. By maximizing the consistency of the model's output across conditions, the model's learning for unlabeled data can be strengthened, thus improving the model's performance.

The consistency loss function plays a key role in consistency training.

$$L_{connsistency} = ||f(x) - f(T(x))|| \tag{4}$$

Where x is the labeled data, andT(x) is the data that has been transformed to x data after some transformation, andf(·) is the output of the model. The consistency loss function measures the degree of consistency of the model's output under different conditions, and by minimizing this loss function, the model can be made to produce similar outputs for similar inputs, thus improving the model's generalization ability[4].

3.2.3. Application of consistency training in semantic segmentation of remote sensing images

In the remote sensing image semantic segmentation task, consistency training is an important method to help the model better utilize a large amount of unlabeled data, so as to improve the performance and generalization ability of the model. Specifically, the applications of consistency training in semantic segmentation of remote sensing images include the following aspects: enhancing the model generalization ability, reducing the need for labeled data, improving the model adaptation and dealing with the problem of uneven data distribution[5].

4. Comparison and analysis of methods

4.1. Introduction to assessment indicators

In order to objectively assess the performance of deep learning-based semantic segmentation methods for remote sensing images, the segmentation models need to be evaluated using the evaluation criteria recognized in this field. The commonly used evaluation metrics are intersection ratio, average intersection ratio, and accuracy rate.

The intersection and concatenation ratio is the ratio of the intersection and concatenation between the predicted and labeled regions of an image. The average intersection and concatenation ratio is the value obtained by averaging the intersection and concatenation ratios for each category.

$$loU = \frac{TP}{TP + FP + FN} \tag{5}$$

$$mloU = \frac{1}{k}\sum_{i=1}^{k} loU_i \tag{6}$$

Precision rate is used to evaluate the percentage of correct classifications in segmented images.

$$Precision = \frac{TP}{TP + FP} \tag{7}$$

In the formulaTP denotes the number of pixels correctly categorized as a target, andFP denotes the number of pixels where the background is judged as a target.

4.2. Comparative analysis of experimental results

Comparison of the effect of different methods in semantic segmentation of remote sensing images are shown in table 1.

Table 1: Comparison of the effect of different methods in semantic segmentation of remote sensing images

methodologies	principle	vantage	drawbacks
GAN	The generator produces realistic remote sensing images and the discriminator distinguishes between the generated and real images.	It can effectively utilize unlabeled data to improve the generalization ability and performance of the model;	The training process is more complex and prone to training instability and lack of diversity in the generated images;
VAE	Learning the hidden variable representation of the data distribution to generate the data i.e. learning the latent representation of the remotely sensed image and generating a realistic image.	The ability to learn potential representations of the data improves the generalization of the model;	Generated images may be of low quality and sometimes difficult to distinguish from real images; the generator may ignore subtle features of the data;
self-training	Using the existing labeled data to train the initial model, using the model to predict the unlabeled data, taking the sample with higher confidence as pseudo-label, merging the pseudo-label with the labeled data, and train the model again.	Simple and easy to implement; can effectively utilize untagged data;	May be affected by the initial model, resulting in inaccurate learned pseudo-labels; need to set thresholds or other strategies to filter high-confidence pseudo-labels, the process is more cumbersome;
Coherence training	Imposing consistency constraints in the input space using data augmentation techniques to improve model stability and performance.	Strong theoretical foundation, can improve the robustness and generalization ability of the model; effective for small samples and unbalanced datasets.	Need to choose appropriate consistency constraints and data augmentation strategies, otherwise it may lead to model performance degradation; Higher requirements on model robustness may require longer training time and large computational resources.

5. Summary and prospect

Semi-supervised learning framework for semantic segmentation of remote sensing images is a key research direction, which aims to solve the problems of sample dependency, network model migration ability and balance between segmentation accuracy and processing speed in current deep learning semantic segmentation methods. With the rapid development of image semantic segmentation methods in recent years, semantic segmentation of remote sensing images has attracted much attention. However, there are certain differences in remote sensing images taken by different satellites, and remote sensing images have complex features and different structures, which are particularly limited in terms of data. In this paper, we outline, summarize and analyze the related researches based on semi-supervised learning in the field of semantic segmentation of remote sensing images, and briefly introduce the characteristics of each type of methods, with special focus on the semi-supervised learning methods based on generative model and discriminative model. The following descriptions are made for the deficiencies and possible future development trends in this field:

(1) This approach is expected to reduce the cost of data labeling, improve the efficiency of model training, and increase the generalization ability of the model.

(2) Using a semi-supervised learning framework, it is possible to explore how models can be trained with a small amount of labeled data and a large amount of unlabeled data, thus reducing the dependence on labeled data. This requires an in-depth study on how to effectively utilize the information from unlabeled data to improve the performance and generalization ability of the model.

(3) Under the semi-supervised learning framework, work needs to be done to improve the migration learning capability of the network model to adapt to different datasets and remote sensing images in different environments. This includes exploring how to dynamically adapt the dataset during model training to improve the robustness and generalization ability of the model.

(4) Under the semi-supervised learning framework, there is a need to continue to investigate how to balance the relationship between segmentation accuracy and processing speed. This may involve designing new network structures, optimization algorithms and hardware acceleration techniques to achieve fast and accurate segmentation of remote sensing images.

Summarizing the above outlook, the semi-supervised learning framework for semantic segmentation of remote sensing images has great potential to provide a more effective and efficient solution for remote sensing image decoding. However, realizing this goal requires in-depth research in both theory and practice to overcome the challenges of current approaches and achieve further breakthroughs.

References

[1] Zhuang F, Qi Z, Duan K, et al. A Comprehensive Survey on Transfer Learning[J]. Proceedings of the IEEE, 2021, 109(1): 43-76.

[2] Locatello F, Tschannen M, Bauer S, et al. Disentangling factors of variations using few labels [C]//ICLR. Proceedings of the 8th International Conference on Learning Representations. Addis Ababa: ICLR, 2020.

[3] Dittadi A, Träuble F, Locatello F, et al. On the transfer of disentangled representations in realistic settings[C]//ICLR. Proceedings of the 9th International Conference on Learning Representations. addis Ababa: ICLR, 2021.

[4] Tschannen M, Bachem O, Lucic M. Recent Advances in Autoencoder-Based Representation Learning[J]. 2018.DOI:10.48550/arXiv.1812.05069.

[5] Husnain M, Missen M M S, Mumtaz S, et al. Visualization of high-dimensional data by pairwise fusion matrices using t-SNE[J]. Symmetry, 2019, 11(1): 107.