SEMI-SUPERVISED SEMANTIC SEGMENTATION NETWORK VIA LEARNING CONSISTENCY FOR REMOTE SENSING LAND-COVER CLASSIFICATION

Bin Zhang[1], Yongjun Zhang[1]*, Yansheng Li[1], Yi Wan[1], Fei Wen[1]

[1] School of Remote Sensing and Information Engineering, Wuhan University, Wuhan, Hubei, China
(bin.zhang, zhangyj, yansheng.li)@whu.edu.cn

Commission II, WG II/6

KEY WORDS: remote sensing, semantic segmentation, semi-supervised learning, convolutional neural network

ABSTRACT:

Current popular deep neural networks for semantic segmentation are almost supervised and highly rely on a large amount of labeled data. However, obtaining a large amount of pixel-level labeled data is time-consuming and laborious. In remote sensing area, this problem is more urgent. To alleviate this problem, we propose a novel semantic segmentation neural network (S4Net) based on semi-supervised learning by using unlabeled data. Our model can learn from unlabeled data by consistency regularization, which enforces the consistency of output under different random transforms and perturbations, such as random affine transform. Thus, the network is trained by the weighted sum of a supervised loss from labeled data and a consistency regularization loss from unlabeled data. The experiments we conducted on DeepGlobe land cover classification challenge dataset verified that our network can make use of unlabeled data to obtain precise results of semantic segmentation and achieve competitive performance when compared to other methods.

1. INTRODUCTION

In remote sensing science and technology, the classification of remote sensing images is one of the most basic research issues, and it is the basis of other remote sensing research and application. In the past, traditional machine learning methods, such as support vector machine, were generally used for classification and recognition of remote sensing images. Traditional machine learning methods generally combine human prior knowledge and intuitive experience to design and select several characteristics and features that are strongly related to the task (LeCun et al., 2015).

In recent years, deep learning has become mainstream in image processing and convolutional neural networks (CNN) have achieved great success (LeCun et al., 2015). With a large number of data sets, the CNN models can be trained by end-to-end to get a more robust feature representation and higher accuracy. Although the currently popular methods can obtain better results, most of the current models are trained by supervised fashion, which needs a large number of labeled data to cooperate with deep networks for learning parameters (Zhang et al., 2016, Zhu et al., 2017, Ball et al., 2017). However, collecting accurately labeled data is extremely time-consuming and laborious, especially accurate pixel-level labeled data. Because labeled data requires a certain amount of expert knowledge and is difficult to obtain for security or privacy considerations (Castrejon et al., 2017). For example, in the field of remote sensing, it is difficult to obtain high-precision, high-quality surface cover data. Therefore, for many practical problems and applications, the lack of resources to create sufficiently large labeled datasets has limited the widespread application of deep learning technologies.

A potential promising approach to solve this problem is *semi-supervised learning* (SSL). Semi-supervised learning is a type of machine learning technology that lies between supervised learning and unsupervised learning. It usually uses a small number of labeled data and a large number of unlabeled data to train a neural network (Chapelle et al., 2009). It has found that combining unlabeled data with a small number of labeled data can significantly improve learning performance. For example, see figure 1, more accurate decision boundaries can be found by using more unlabeled samples. For supervised learning, obtaining data annotations is costly and time-consuming, and is difficult to obtain a large amount of labeled data. While the acquisition of unlabeled data is relatively cheap, so the ap-

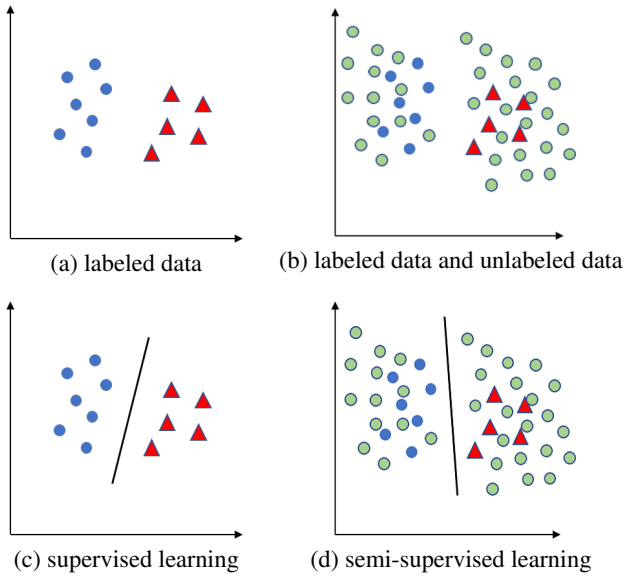


Figure 1. Example of supervised learning and semi-supervised learning. The blue and red dots denote labeled samples and the green dots denote unlabeled samples.

*Corresponding author

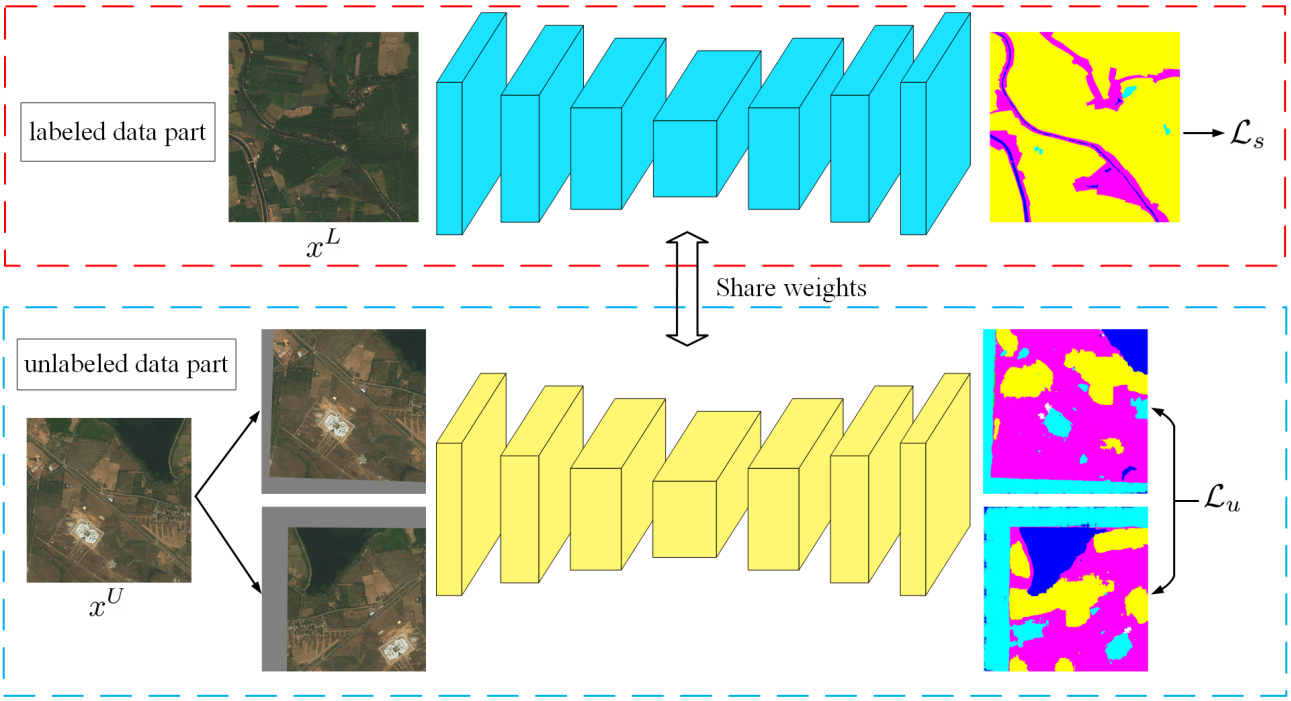


Figure 2. The proposed semi-supervised semantic segmentation framework (S4Net) based on consistency regularization. The UNet is used here with a shared-weights strategy. For the labeled data part, the input images are fed into the network to get predicted outputs, then we can compute supervised loss (such as cross entropy loss). On the other hand, for the unlabeled data part, the input images are augmented and then fed into the network to get two outputs, and their consistency loss is calculated.

plication of semi-supervised learning is more extensive.

To overcome the problem of a large amount of data required for supervised learning, we proposed a semantic segmentation network based on semi-supervised learning, named S4Net in this paper. Specifically, the consistency regularization was introduced to exploit the unlabeled data, which encourages the pixel-level consistency of output under different random transforms and perturbations. Finally, the network was trained by the weighted sum of a supervised loss from labeled data and a consistency regularization loss from unlabeled data. We performed experiments on a public DeepGlobe land cover classification challenge dataset and verified this method can take advantage of unlabeled data and achieve improvements in the context of a small amount of data.

2. RELATED WORK

In this part, the past proposed semi-supervised learning methods for image classification are reviewed. After this, we will discuss related semi-supervised learning works.

Semi-supervised learning (SSL) is somewhere between supervised and unsupervised learning (Chapelle et al., 2009). It can be divided into two categories: transductive learning and inductive learning. It is noted that semi-supervised learning has to rely on some assumptions. The detailed information please refer to the book review (Chapelle et al., 2009). Next, we will review semi-supervised learning methods based on deep learning methods.

2.1 Semi-supervised learning for image classification

One of the most simple methods is Pseudo-labeling (Lee, 2013), which is widely used in practice, likely because of its simplicity and generality. The class which has the maximum probability was used as the label of samples. The π-model and temporal ensembling (Laine, Aila, 2016) proposed a method based on consistency regularization that takes advantage of the stochastic and minimizes the difference between the predictions under different random transforms and perturbations to input samples (Sajjadi et al., 2016). Different from the π-model, Mean Teacher (Tarvainen, Valpola, 2017) used a more stable predicted output by using an exponential moving average of network parameters. Instead of using the randomness of the network, Virtual Adversarial Training (VAT) (Miyato et al., 2018) directly used as target a small perturbation to input which would most significantly affect the output of the prediction function inspired by adversarial training. Instead of adding perturbations to each single training sample, Smooth Neighbors on Teacher Graphs (SNTG) (Luo et al., 2018) encouraged neighbors to get similar predictions while the non-neighbors are pushed apart from each other. The Co-Training method (Qiao et al., 2018) can learn multiple neural networks from different views and use adversarial examples to force differences between different views. Inspired by the mixup method (Zhang et al., 2018), Interpolation Consistency Training (ICT) (Verma et al., 2019) proposed a semi-supervised learning method by enforcing the output at an interpolation of unlabeled samples to be consistent with the interpolation of the output at those samples' outputs. Instead of using the class which has the maximum predicted probability as labels, Deep Label Propagation (Iscen et al., 2019) used the transductive label propagation method to obtain pseudo labels according to the manifold assumption. The MixMatch (Berthelot et al., 2019) combined ideas and components from the current dominant paradigms for semi-supervised learning.

2.2 Semi-supervised learning for semantic segmentation

Though substantial recent progress has been made in developing semi-supervised algorithms in image classification task for comparatively small datasets, many of these methods do not scale readily to the semantic segmentation task of real-world applications. Some works have been proposed for semi-supervised semantic segmentation task in recent years. Hong et al. (Hong et al., 2015) proposed a decoupled network to learn classification and segmentation networks separately by exploiting unlabelled samples with image-level labels and pixel-wise annotations. Souly et el. (Souly et al., 2017) proposed to use a GAN architecture for semi-supervised semantic segmentation. In this architecture, generated data, unlabeled data, and labeled data were fed to a discriminator to get class confidences and generate confidence maps for each class as well as a label for fake data. Hung et al. (Hung et al., 2018) also proposed an adversarial network for semi-supervised semantic segmentation. The difference is they design a fully convolutional discriminator to discover trustworthy regions of unlabeled samples that facilitate the training process for segmentation. Kalluri et al. (Kalluri et al., 2019) devised a universal segmentation model, which can be jointly trained across different datasets with different categories.

3. METHOD

In this section, we first formulate the semi-supervised learning problem, and then we present our semi-supervised semantic segmentation framework, denoted as S4Net.

3.1 Overview

In the context of supervised learning, all the input data is labeled data $\mathcal{D}_L = \{(x_i^L, y_i^L)\}_{i=1}^{N_L}$ and the neural network is usually trained by minimizing a supervised loss term:

$$\mathcal{L}_s(X_L, Y_L; \theta) = \sum_{i=1}^{N_L} \ell_s(f_\theta(x_i^L), y_i^L) \tag{1}$$

where the supervised loss ℓ_s is usually formulated as the cross entropy loss and $f_\theta(\cdot)$ denotes the neural network with parameters θ.

However, for the context of semi-supervised learning, one can access a number of labeled samples $\mathcal{D}_L = \{(x_i^L, y_i^L)\}_{i=1}^{N_L}$ and unlabeled samples $\mathcal{D}_U = \{x_i^U\}_{i=1}^{N_U}$, where $y_i^L \in cardinal(C)$ and C is the number of classes. N_L and N_U are the number of labeled and unlabeled samples with $N_L \ll N_U$. The goal of semi-supervised learning is to get a better model by using all labeled data and unlabeled data than supervised learning. Thus, the loss function is formulated as the weighted sum of a supervised loss $\mathcal{L}_s$ from labeled data and a regularization loss $\mathcal{L}_u$ from unlabeled data or both labeled and unlabeled data:

$$\mathcal{L} = \mathcal{L}_s + \lambda \mathcal{L}_u \tag{2}$$

where λ is a hyperparameter, which quantified the importance of the regularization loss.

To make use of unlabeled data, the consistency regularization (Sajjadi et al., 2016, Laine, Aila, 2016, Tarvainen, Valpola, 2017) was usually introduced to exploit the potential data manifolds:

$$\mathcal{L}_u(X_U; \theta) = \sum_{i=1}^{N_U} \ell_u \left(f_\theta \left(x_i^U \right), f_{\bar{\theta}} \left(\tilde{x}_i^U \right) \right) \tag{3}$$

where $\tilde{x}_i$ refers to an example x_i that is applied to a random perturbation. In image classification, the random flip, random crop and random noise are usually used as the random perturbation. The network parameter $\bar{\theta}$ is either equal to the original parameter θ or any other transformation of it, such as the exponential moving average over the update of the network. The consistency regularization term $\mathcal{L}_u$ often uses mean squared error (squared L2 norm) or Kullback-Leibler divergence, which encourages the pixel-level consistency of the output under different random transforms and perturbations.

Algorithm 1: Mini-batch training for semi-supervised semantic segmentation

Data: labeled samples $\mathcal{D}_L = \{(x_i^L, y_i^L)\}_{i=1}^{N_L}$, unlabeled samples $\mathcal{D}_U = \{x_i^U\}_{i=1}^{N_U}$
Require: neural network with parameters θ
Require: random perturbation function φ
for *t in [1, number of epochs]* **do**
  for *each minibatch B* **do**
    get labeled samples x^L and unlabeled samples x^U from B
    compute supervised loss $\mathcal{L}_s$ using Equation (1)
    get two random perturbations φ_1, φ_2
    perform random perturbation on unlabeled samples $\tilde{x}_1^U = \varphi_1(x^U)$, $\tilde{x}_2^U = \varphi_2(x^U)$
    get two outputs by feeding perturbation samples to network $f_\theta(\tilde{x}_1^U)$, $f_\theta(\tilde{x}_2^U)$
    perform inverse transform to get two outputs $\varphi_1^{-1}(f_\theta(\tilde{x}_1^U))$, $\varphi_2^{-1}(f_\theta(\tilde{x}_2^U))$
    compute semi-supervised consistency regularization loss $\mathcal{L}_u$ using Equation (4)
    compute total loss $\mathcal{L} = \mathcal{L}_s + \lambda \mathcal{L}_u$
    update θ using optimizer, e.g., SGD
  end
end

Description	Output size	
input	H × W	encoder
conv, 7×7, 64, stride 2	H/2 × W/2	
max pool, 3×3, stride 2	H/4 × W/4	
ResBlock×3	H/4 × W/4	
ResBlock×4	H/8 × W/8	
ResBlock×6	H/16 × W/16	
ResBlock×3	H/32 × W/32	
conv, 3×3, 192	H/32 × W/32	decoder
Transposeconv, 4×4, 128 Concat conv, 3×3, 128	H/16 × W/16	
Transposeconv, 4×4, 96 Concat conv, 3×3, 96	H/8 × W/8	
Transposeconv, 4×4, 64 Concat conv, 3×3, 64	H/4 × W/4	
Transposeconv, 4×4, 48 conv, 3×3, 48	H/2 × W/2	
Transposeconv, 4×4, 32 conv, 3×3, 32 conv, 1×1, C	H × W	

Table 1. U-Net based ResNet50 encoder.

3.2 Semi-supervised semantic segmentation framework (S4Net)

Figure 2 shows the proposed semi-supervised semantic segmentation framework. We adopt the UNet (Ronneberger et al., 2015) with ResNet-50 (He et al., 2016) model pre-trained on the ImageNet dataset as our segmentation baseline network. The decoder network uses 3×3 convolutions and strided 4×4 transposed convolutions to recover the original input size. The detailed network structure is shown in Table 1. For labeled samples, we fed them to the network to compute supervised loss $\mathcal{L}_s$, such as cross entropy loss. For unlabeled samples x^U, one can get two different transformed samples $\tilde{x}_1^U, \tilde{x}_2^U$ by performing two random perturbations φ_1, φ_2, namely $\tilde{x}_1^U = \varphi_1(x^U)$, $\tilde{x}_2^U = \varphi_2(x^U)$. Here we use the random affine transformation as the random perturbation. Then, feeding them to the network can get two outputs $f_\theta(\tilde{x}_1^U)$, $f_\theta(\tilde{x}_2^U)$. Different from the classification task, to compute the pixel-level consistency of two outputs, we have to perform the inverse transform to put every pixel to the original location. We denote two inverse transforms of the random perturbations as $\varphi_1^{-1}, \varphi_2^{-1}$. Thus, we can get inverse transformed outputs $\varphi_1^{-1}(f_\theta(\tilde{x}_1^U))$, $\varphi_2^{-1}(f_\theta(\tilde{x}_2^U))$ and the semi-supervised consistency regularization loss can compute as follows:

$$\mathcal{L}_u(X_U;\theta) = \sum_{i=1}^{N_U} \ell_u \left(\varphi_1^{-1}\left(f_\theta\left(\tilde{x}_i^1\right)\right), \varphi_2^{-1}\left(f_{\bar{\theta}}\left(\tilde{x}_i^2\right)\right)\right) \quad (4)$$

Here we used mean squared error as the consistency regularization loss. For the affine transformation, we used the translation in the range [-0.2, 0.2] factor of both height and width, scaling in the range [0.75, 1.25] and rotation in the range $[-15°, 15°]$. As mentioned above, the algorithm flow of the proposed semi-supervised semantic segmentation framework is shown in Algorithm 1.

4. EXPERIMENTS

4.1 Dataset

To verify our method, we consider using the land cover classification dataset on DeepGlobe Challenge [1] (Demir et al., 2018). This dataset offers 1,146 high-resolution sub-meter satellite images and each image has a size of 2448×2448 pixels. The whole dataset is split into training, validation and test set, each with 803, 171 and 172 images. The mask images are RGB images with 7 classes, see figure 3. The unknown class is ignored in the evaluation stage.

It is worth noting that we only use the training set as the experimental data, and randomly divide 100, 503, and 200 images as labeled data, unlabeled data, and validation dataset.

4.2 Implementation Details

Our implementation used the PyTorch framework and an NVIDIA Titan X GPU was used to accelerate training. We used stochastic gradient descent (SGD) with a mini-batch size of 6 to train our model, including 4 labeled samples and 2 unlabeled samples. The weight decay was set to 0.0001 and the momentum was set to 0.9. Cosine annealing strategy was used as the learning rate policy. The initial learning rate started from 0.01 and the models were trained for a total of 100,000 steps.

[1] https://competitions.codalab.org/competitions/18468

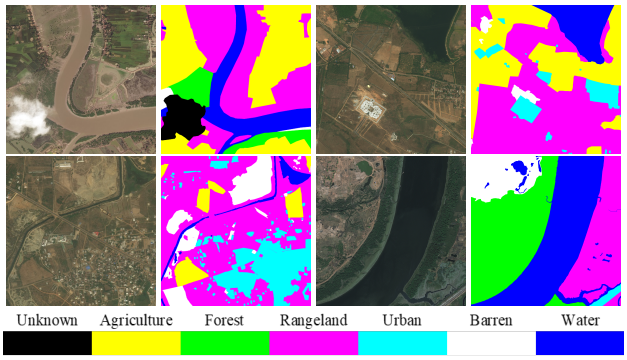


Figure 3. Some examples of land cover classification dataset on DeepGlobe Challenge.

	Mean IoU
Hung et al. (Hung et al., 2018)	55.1
Baseline	62.1
Fully supervised	66.8
S4Net(Ours)	65.2

Table 2. The results on validation dataset.

For the weight of semi-supervised consistency regularization loss component λ, we used a sigmoid-shaped ramp-up curve function $e^{-5(1-x)^2}$ in the first 80,000 steps. The maximum of λ is 2.0. For the data augmentation strategy, we used the random horizontal and vertical flip. And finally, the crop size is 512×512.

For evaluation, the mean intersection over union (mIoU) is calculated as the evaluation metric. The IoU is defined as the size of the intersection divided by the size of the union of two sets.

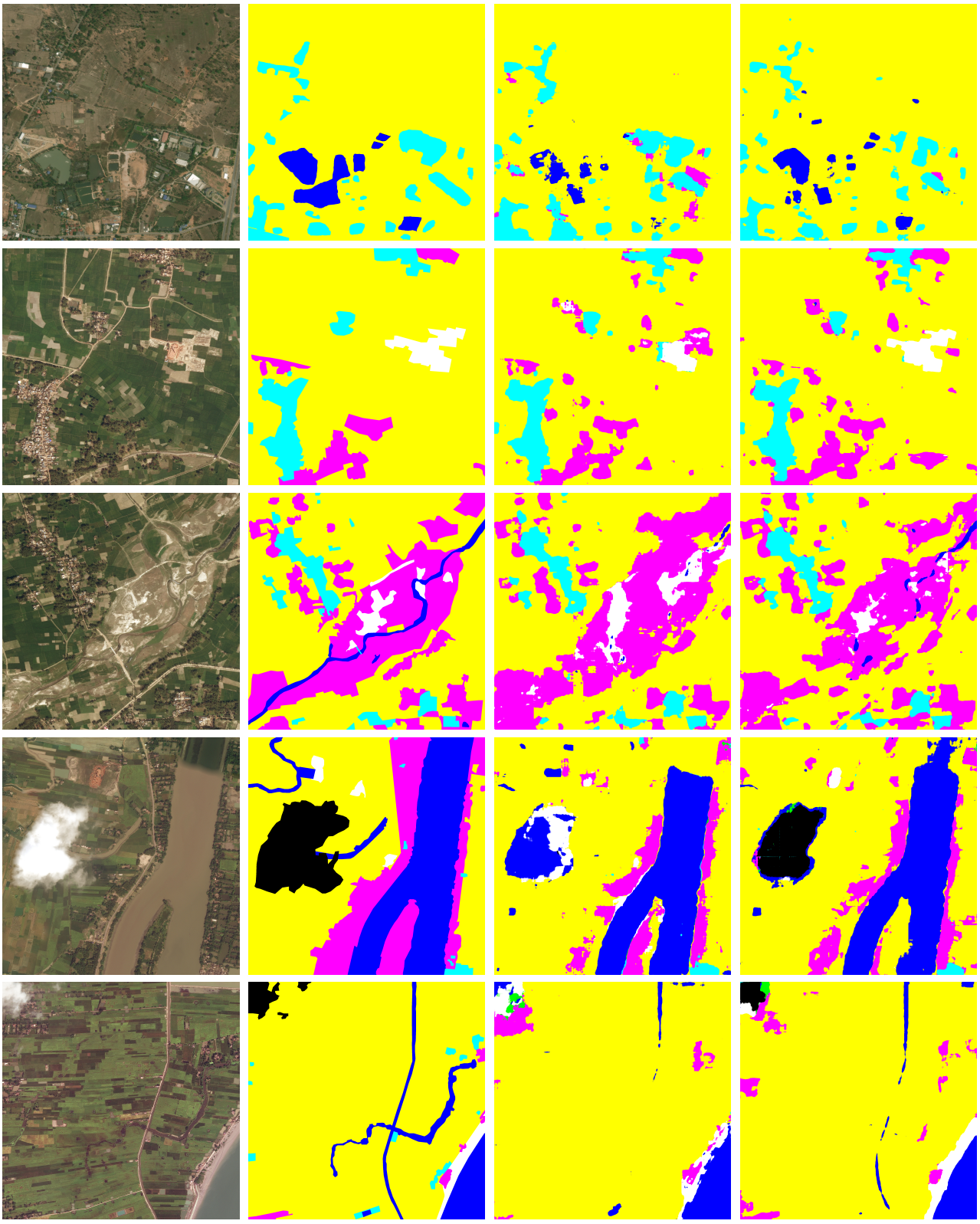
$$IoU = \frac{|R_g \cap R_p|}{|R_g \cup R_p|} = \frac{|R_g \cap R_p|}{|R_g| + |R_p| - |R_g \cap R_p|} \quad (5)$$

where R_g and R_p are the set of label pixels and the set of prediction pixels. $\cap$ and $\cup$ denote the intersection and union operations, respectively. $|\cdot|$ denotes the number of pixels in the set. The mIoU can be obtained by averaging the per-class IoU.

4.3 Experimental Results

To evaluate our method, we trained UNet in a supervised way as the baseline. And we also compared with Hung et al.'s (Hung et al., 2018) method, in which they used a generative adversarial network to determine the confidence maps of unlabeled data output. The experimental results were shown in Table 2. As we can see, the baseline method can achieve 62.1 mIoU. However, Hung et al.'s method got a worse result. We suspected that the reason is that the training process is unstable for the generative adversarial network when the number of unlabeled data is much larger than that of labeled data. We also trained UNet both on labeled data and unlabeled data to get the upper bound of semi-supervised learning and achieved 66.8 mIoU. The experimental result showed that our method can achieve 4.7 mIoU improvement compared with the baseline method.

The detailed per-class performance of our method and other methods on the validation dataset were presented in Table 3. Similarly, Hung et al.'s method got worse results, especially forest land, rangeland. We find that our semi-supervised method outperforms supervised baseline methods by a significant margin, for example getting 4.22% and 23.92% improvement for



(a) Image (b) GT (c) baseline (d) S4Net(Ours)

Figure 4. The results of the baseline method and our method.

forest land and rangeland. For barren land and water class, our method can get 57.97 and 81.10 mIoU and achieve better performance compared with fully supervised results. Thus, we believe that our proposed method takes advantage of unlabeled data.

We also visualized some results of the validation dataset for qualitative comparison, as illustrated in Figure 4. As we can see, our semi-supervised method can do better for details than the baseline method. And our semi-supervised method can achieve better integrity and correctness. For example, our

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
https://doi.org/10.5194/isprs-annals-V-2-2020-609-2020 | © Authors 2020. CC BY 4.0 License.

	Agriculture land	Forest land	Rangeland	Urban land	Barren land	Water
Hung et al. (Hung et al., 2018)	80.80	53.66	12.97	69.57	45.51	67.95
Baseline	84.54	63.72	23.16	73.29	51.22	76.95
Fully supervised	87.10	72.63	32.95	77.00	51.69	79.35
S4Net(Ours)	83.85	**66.41**	**28.70**	73.35	**57.97**	**81.10**

Table 3. The detailed per-class performance of our method and other methods.
The **bold** font means the performance of our method is lager than the baseline method and Hung et al.'s method.

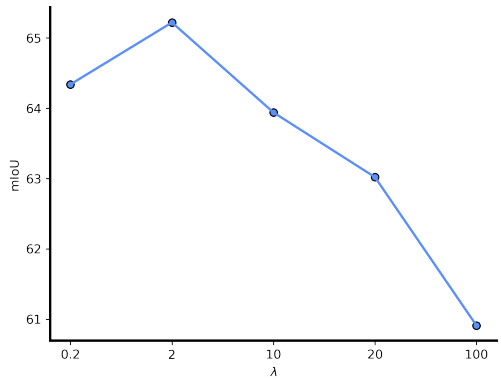


Figure 5. The results of different choices for λ based on mIoU on validation dataset.

method can get a better result for the fourth image, where the clouds in the image are correctly classified.

4.4 Hyperparameter analysis

In this section, the weight of semi-supervised consistency regularization loss λ will be analyzed. Since it is not possible to try all possible values, based on previous literature (Laine, Aila, 2016, Verma et al., 2019), we used five different λ choices here: 0.2, 2, 10, 20 and 100. The implementation detail and setting are same as the previous experiments. We evaluate the results of different λ choices based on mIoU on the validation dataset and the experimental results are shown in Figure 5. As shown in Figure 5, the reported result significantly more than the other four values when the weight value equal to 2.0. Thus, we use 2.0 as the default weight of semi-supervised consistency regularization loss.

4.5 Evaluation robustness of the proposed method

	Number of labeled data		
	20	50	100
Baseline	51.52±3.47	59.29±3.43	63.51±0.97
S4Net(Ours)	53.47±2.82	61.70±4.32	66.42±1.75

Table 4. The results on validation dataset using three different number of labeled data.

To evaluate the robustness of the proposed method, we considered three different numbers of labeled samples. In detail, same as previous experiments, 200 images from 803 were selected as the validation set according, and the data containing labeled and unlabeled data were divided from the remaining 603 images. Then we run three times to calculate the mean and standard deviation by using different random seed.

The results obtained are recorded in Table 4. Under three different settings, the proposed method is superior to the baseline method. Even in the case where the number of labeled data is very small, that is, the training dataset containing 20 labeled data (the rest are unlabeled data), the proposed method can still obtain considerable results. We also observed that when the number of labeled data was reduced from 50 to 20, there was a larger decrease in accuracy. We attribute this to the fact that the network cannot get enough valid and correct signals from the training data species when there is a small number of labeled data. However, we can mitigate this problem by utilizing large amounts of unlabeled data through semi-supervised learning.

5. CONCLUSION

In this work, a novel semi-supervised semantic segmentation framework (S4Net) was proposed via enforcing consistency regularization for remote sensing images. The proposed method can make use of unlabeled data to improve performance by encouraging the pixel-level consistency of output under different random transforms and perturbations. The experiments show that this method is promising and can bring higher accuracy when there are fewer labeled samples. Especially in remote sensing application scenarios, such as pan-sharpening (Zhang et al., 2019a) and super-resolution (Zhang et al., 2019b), where accurately labeled data is difficult to obtain, semi-supervised learning can play a greater role.

In the future, we will continue working for semi-supervised to make use of unlabeled data and future research should consider more diverse transformations or perturbations. For example, we can introduce adversarial perturbation to augment training samples.

ACKNOWLEDGEMENTS

This work was supported in part by the National Key Research and Development Program of China under Grant 2018YFB0505003; the National Natural Science Foundation of China under Grant 41971284.

REFERENCES

Ball, J. E., Anderson, D. T., Chan, C. S., 2017. Comprehensive survey of deep learning in remote sensing: theories, tools, and challenges for the community. *Journal of Applied Remote Sensing*, 11(4), 042609.

Berthelot, D., Carlini, N., Goodfellow, I., Papernot, N., Oliver, A., Raffel, C. A., 2019. Mixmatch: A holistic approach to semi-supervised learning. *Advances in Neural Information Processing Systems*, 5050–5060.

Castrejon, L., Kundu, K., Urtasun, R., Fidler, S., 2017. Annotating object instances with a polygon-rnn. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 5230–5238.

Chapelle, O., Scholkopf, B., Zien, A., 2009. Semi-supervised learning (chapelle, o. et al., eds.; 2006)[book reviews]. *IEEE Transactions on Neural Networks*, 20(3), 542–542.

Demir, I., Koperski, K., Lindenbaum, D., Pang, G., Huang, J., Basu, S., Hughes, F., Tuia, D., Raska, R., 2018. Deepglobe 2018: A challenge to parse the earth through satellite images. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, IEEE, 172–17209.

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

Hong, S., Noh, H., Han, B., 2015. Decoupled deep neural network for semi-supervised semantic segmentation. *Advances in neural information processing systems*, 1495–1503.

Hung, W.-C., Tsai, Y.-H., Liou, Y.-T., Lin, Y.-Y., Yang, M.-H., 2018. Adversarial learning for semi-supervised semantic segmentation. *Proceedings of the British Machine Vision Conference (BMVC)*, BMVC Press, 65.

Iscen, A., Tolias, G., Avrithis, Y., Chum, O., 2019. Label propagation for deep semi-supervised learning. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 5070–5079.

Kalluri, T., Varma, G., Chandraker, M., Jawahar, C., 2019. Universal semi-supervised semantic segmentation. *Proceedings of the IEEE International Conference on Computer Vision*, 5259–5270.

Laine, S., Aila, T., 2016. Temporal ensembling for semi-supervised learning. *arXiv preprint arXiv:1610.02242*.

LeCun, Y., Bengio, Y., Hinton, G., 2015. Deep learning. *nature*, 521(7553), 436.

Lee, D.-H., 2013. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. *Workshop on challenges in representation learning, ICML*, 3, 2.

Luo, Y., Zhu, J., Li, M., Ren, Y., Zhang, B., 2018. Smooth neighbors on teacher graphs for semi-supervised learning. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 8896–8905.

Miyato, T., Maeda, S.-i., Koyama, M., Ishii, S., 2018. Virtual adversarial training: a regularization method for supervised and semi-supervised learning. *IEEE transactions on pattern analysis and machine intelligence*, 41(8), 1979–1993.

Qiao, S., Shen, W., Zhang, Z., Wang, B., Yuille, A., 2018. Deep co-training for semi-supervised image recognition. *Proceedings of the european conference on computer vision (ECCV)*, 135–152.

Ronneberger, O., Fischer, P., Brox, T., 2015. U-net: Convolutional networks for biomedical image segmentation. *International Conference on Medical image computing and computer-assisted intervention*, Springer, 234–241.

Sajjadi, M., Javanmardi, M., Tasdizen, T., 2016. Regularization with stochastic transformations and perturbations for deep semi-supervised learning. *Advances in Neural Information Processing Systems*, 1163–1171.

Souly, N., Spampinato, C., Shah, M., 2017. Semi supervised semantic segmentation using generative adversarial network. *Proceedings of the IEEE International Conference on Computer Vision*, 5688–5696.

Tarvainen, A., Valpola, H., 2017. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. *Advances in neural information processing systems*, 1195–1204.

Verma, V., Lamb, A., Kannala, J., Bengio, Y., Lopez-Paz, D., 2019. Interpolation consistency training for semi-supervised learning. *Proceedings of the 28th International Joint Conference on Artificial Intelligence*, IJCAI'19, AAAI Press, 3635–3641.

Zhang, H., Moustapha, C., Yann, N. D., David, L.-P., 2018. mixup: Beyond Empirical Risk Minimization. *6th International Conference on Learning Representations, ICLR 2018*.

Zhang, L., Zhang, L., Du, B., 2016. Deep learning for remote sensing data: A technical tutorial on the state of the art. *IEEE Geoscience and Remote Sensing Magazine*, 4(2), 22–40.

Zhang, Y., Liu, C., Sun, M., Ou, Y., 2019a. Pan-Sharpening Using an Efficient Bidirectional Pyramid Network. *IEEE Transactions on Geoscience and Remote Sensing*, 57(8), 5549–5563.

Zhang, Y., Zheng, Z., Luo, Y., Zhang, Y., Wu, J., Peng, Z., 2019b. A CNN-Based Subpixel Level DSM Generation Approach via Single Image Super-Resolution. *Photogrammetric Engineering & Remote Sensing*, 85(10), 765–775.

Zhu, X. X., Tuia, D., Mou, L., Xia, G.-S., Zhang, L., Xu, F., Fraundorfer, F., 2017. Deep learning in remote sensing: A comprehensive review and list of resources. *IEEE Geoscience and Remote Sensing Magazine*, 5(4), 8–36.