

Class Prior-Free Positive-Unlabeled Learning with Taylor Variational Loss for Hyperspectral Remote Sensing Imagery

Hengwei Zhao Xinyu Wang* Jingtao Li Yanfei Zhong

Wuhan University, Wuhan, China

{whu_zhaohw, wangxinyu, JingtaoLi, zhongyanfei}@whu.edu.cn

Abstract

Positive-unlabeled learning (PU learning) in hyperspectral remote sensing imagery (HSI) is aimed at learning a binary classifier from positive and unlabeled data, which has broad prospects in various earth vision applications. However, when PU learning meets limited labeled HSI, the unlabeled data may dominate the optimization process, which makes the neural networks overfit the unlabeled data. In this paper, a Taylor variational loss is proposed for HSI PU learning, which reduces the weight of the gradient of the unlabeled data by Taylor series expansion to enable the network to find a balance between overfitting and underfitting. In addition, the self-calibrated optimization strategy is designed to stabilize the training process. Experiments on 7 benchmark datasets (21 tasks in total) validate the effectiveness of the proposed method. Code is at: <https://github.com/Hengwei-Zhao96/T-HOneCls>.

1. Introduction

Positive-unlabeled learning is aimed at learning a binary classifier from positive and unlabeled data [21, 17, 3]. Due to the lack of negative samples, PU learning is a challenging task, but play an important role in machine learning applications, including product recommendation [16], deceptive reviews detection [30], and medical diagnosis [39].

PU learning in HSI is a powerful tool for environmental monitoring [43, 23]. For example, when mapping the invasive species in complex forestry, PU learning only needs positive labels of invasive species; however, traditional hyperspectral classification [19, 37, 46] requires the various negative classes to be labeled to obtain a discriminative boundary, which is labor-intensive, even impossible, to investigate the negative objects and annotate them in high species richness areas [43].

Few related works have focused on PU learning in HSI. Compared to other tasks, the training data size in HSI is

much smaller [9], and the deep models are more likely to be over-fitting and susceptible to unlabeled data. These characteristics make hyperspectral PU learning a more challenging task.

PU learning methods can be divided into two categories, according to whether the class prior (π_p , i.e., the proportion of positive data) is assumed to be known. (1) Due to the limited supervision information from PU data, most studies assume that the class prior is available [43, 23], but in reality, the class prior is hard to be estimated accurately, especially for HSIs, due to the severe inter-class similarity and intra-class variation. (2) Class prior-free PU learning is a recent research focus of the machine learning community [3, 17], where variational principle-based PU learning [3] is one of the state-of-the-art in theory. It approximates the positive distribution by optimizing the posterior probability, i.e., the classifier, and does not require knowing the class prior. However, the unlabeled data may dominate the optimization process, which makes it difficult for neural networks to find a balance between the underfitting and overfitting of positive data, especially when the variational principle meets limited labeled HSI data (discussed later in Section 3 in detail).

In this paper, a Taylor series expansion-based variational framework—*T-HOneCls*—is proposed to solve the limited labeled hyperspectral PU learning problem without class prior. The contributions of this paper are summarized as follows:

- A novel insight is proposed in terms of the dynamic change of the loss, which demonstrates that the unlabeled data dominating the training process is the bottleneck of the variational principle-based classifier.
- *Taylor variational loss* is proposed to tackle the problem of PU learning without a class prior, which reduces the weight of the gradient of the unlabeled data and simultaneously satisfy the variational principle by Taylor series expansion, to alleviate the problem of unlabeled data dominating the training process.

*Corresponding author.

- *Self-calibrated optimization* is proposed to take advantage of the supervisory signals from the network itself to stabilize the training process and alleviate the potential over-fitting problem caused by limited labeled data with a large pool of unlabeled data.
- Extensive experiments are conducted on 7 benchmark datasets, including 5 hyperspectral datasets (19 tasks in total), CIFAR-10 and STL-10, where the proposed method outperforms other state-of-the-art methods in most cases.

2. Related Works

Deep Learning Based Classification for HSI The methods of HSI classification can be divided into patch-based framework and patch-free framework [46]. The patch-based methods aim to model a mapping function $f_{pb} : R^{S \times S} \rightarrow R$, and first extract the pixels to be classified and their surrounding pixels to build patches with the size $S \times S$, and then use these patches and labels to train a neural network. Different neural networks can be used to model f_{pb} [5, 38, 15, 6]. The patch-free frameworks aim to model a mapping function $f_{pf} : R^{H \times W} \rightarrow R^{H \times W}$ by a fully convolutional neural network [19, 37, 46], and due to the avoidance of redundant computation in patches, the inference time of the patch-free frameworks is improved by hundreds of times [46].

Differing from the above supervised classification methods, which both need positive and negative data, the method proposed in this paper focuses on weakly supervised PU learning and only requires positive data to be labeled.

Positive and Unlabeled Learning Early studies focused on the two-step heuristic approach [8, 12], which first obtain reliable negative samples from the unlabeled data and then train a binary classifier; however, the performance of these two-step heuristic classifiers is limited by whether the selected samples are correct or not. Besides the two-step methods, this weakly supervised task can be tackled by one-step methods, by cost-sensitive based methods [25, 24, 27], label disambiguation based methods [41], and density ratio estimation-based methods [20]. Furthermore, the methods based on risk estimation are some of the most theoretically and practically effective methods [21, 43, 23, 44, 36, 45]. The imbalanced PU learning has attracted attention recently [32, 4]. Specifically, OC loss [44] has been proposed to solve the imbalance problem in HSI. However, most of these methods assume that the true π_p is available in advance, which is difficult to estimate from HSI with inter-class similarity and intra-class variation.

Learning from PU data without a class prior has recently received attention [17, 3, 22]. A convex formulation was proposed in [2]. However, this was based on unbiased

risk estimation, and conflicted with the flexible neural networks [21]. Predictive adversarial networks (PAN) transform the generator in the generative adversarial network into a classifier [17] to learn from PU data. A heuristic mixup technique is proposed in [22]. The vPU [3] is based on the variational principle. However, the performance of these methods is unsatisfactory with limited labeled samples, and the problem of unlabeled data dominating the optimization process still exists with vPU.

Other Weakly Supervised Learning Methods Label noise representation learning and semi-supervised learning are related to this paper.

The problem of PU learning can be regarded as label noise representation learning, if the unlabeled samples are regarded as noisy negative data. The adverse effects of noisy labels can be mitigated in three directions: data, optimization policy, and objective [13]. For the data, the insight is to link the noisy class posterior and clean class posterior by a noise transition matrix [33, 11, 28]. However, the underlying noise transfer pattern is also difficult to estimate. The dynamic optimization process of the deep neural networks is the key to the optimization policy, such as self-training [18] and co-training [14, 40]. However, the noise rate is difficult to estimate. Mitigating noisy labels from the objective function is consistent with the purpose of this paper, and some loss functions that are robust to noisy labels have in fact been proposed [10, 42, 35, 7].

The problem of semi-supervised learning is to learn from labeled and unlabeled data [31, 26], in the context of binary classification, the labeled data contains positive and negative data. However, PU learning is a more challenging task due to the lack of negative samples.

3. Class Prior-Free PU Learning Framework with Taylor Variational Loss

The proposed PU learning framework (dubbed *T-HOneCls*) is described in this section (Fig. 1). The proposed *Taylor variational loss* is responsible for the task of learning from PU data without a class prior. The *self-calibrated optimization* is proposed to stabilize the training process by taking advantage of the supervisory signals from the network itself with a large pool of unlabeled data.

3.1. Taylor Variational Loss

Preliminaries The spaces of the input and the output are denoted as $X \in R^d$ and $Y \in \{+1, -1\}$, respectively. The joint density of (X, Y) is $p(x, y)$. The marginal distributions of the positive, negative, and unlabeled classes are recorded as $P_p(x) = P(x|y = +1)$, $P_n(x) = P(x|y = -1)$, and $P(x)$, respectively. Let $\mathcal{P} = \{x_i\}_{i=1}^{N_p}$ i.i.d $P_p(x)$

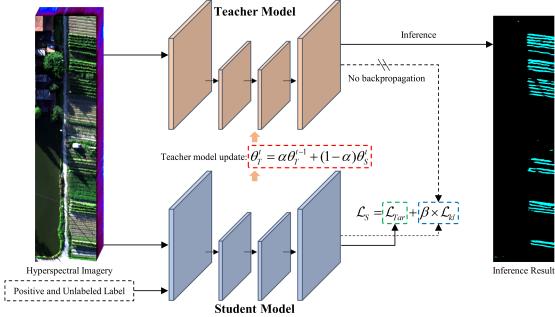


Figure 1: *T-HOneCls*: A Taylor series expansion-based variational framework for HSI PU learning.

and $\mathcal{U} = \{x_i\}_{i=1}^{N_u}$ i.i.d $P(x)$ are the positive and unlabeled dataset, respectively. For simplicity, $f(x; \theta)$ is denoted as $f(x)$, where θ represents the parameters of the neural network. The PU classifier aims to obtain a parametric classifier, i.e., $f(x)$, from the Bayesian classifier, i.e., $f^*(x) = P(y = +1|x)$, from \mathcal{P} and \mathcal{U} .

The estimated positive distribution, i.e., $\hat{P}_p(x)$, can be obtained from the Bayes rule:

$$P_p(x) = \frac{P(y = +1|x)P(x)}{\int P(y = +1|x)P(x)dx} \approx \frac{f(x)P(x)}{E_u[f(x)]} \triangleq \hat{P}_p(x). \quad (1)$$

If a set \mathcal{A} exists and it satisfies the condition of $\int_{\mathcal{A}} P_p(x)dx > 0$ and $f^*(x) = 1, \forall x \in \mathcal{A}$, $P_p(x) = \hat{P}_p(x)$ if and only if $f(x) = f^*(x)$ [3]. The Kullback-Leibler (KL) divergence can be used to estimate the approximate quality of $\hat{P}_p(x)$, and the variational approach can be described as follows:

$$KL(P_p(x) || \hat{P}_p(x)) = \mathcal{L}_{var}(f(x)) - \mathcal{L}_{var}(f^*(x)), \quad (2)$$

where

$$\mathcal{L}_{var}(f(x)) = \log(E_u[f(x)]) - E_p[\log(f(x))]. \quad (3)$$

For completeness of this paper, the proof of Eq. 2 is attached to Appendix 1.

According to the non-negative property of KL divergence, $\mathcal{L}_{var}(f(x))$ is the variational upper bound of $\mathcal{L}_{var}(f^*(x))$, and the minimization of Eq. 2 can be achieved by minimizing Eq. 3, which can be calculated from the empirical averages over \mathcal{P} and \mathcal{U} without a class prior by

$$\hat{\mathcal{L}}_{var}(f(x)) = \log\left(\frac{\sum_{i=1}^{n_u} f(x_i^u)}{n_u}\right) - \frac{\sum_{i=1}^{n_p} \log(f(x_i^p))}{n_p}, \quad (4)$$

where n_p and n_u are the number of positive and unlabeled samples in a batch, respectively. In other words, the classifier can be obtained by minimizing Eq. 4, without π_p .

Theoretical Analysis of Variational Loss The robustness of the variational loss to negative label noise is first analyzed in this subsection, and then a novel insight is proposed to demonstrate that the bottleneck of variational loss is the unlabeled data dominating the training process.

The robustness of variational loss can be obtained by comparing it with cross-entropy loss ($\hat{\mathcal{L}}_{ce}$),

$$\hat{\mathcal{L}}_{ce}(f(x)) = -\frac{\sum_{i=1}^{n_n} \log(1 - f(x_i^n))}{k} - \frac{\sum_{i=1}^{n_p} \log(f(x_i^p))}{k}, \quad (5)$$

where n_n is the number of negative samples in a batch, and $k = n_p + n_n$.

The first characteristic of variational loss is robustness to negative label noise, which can be analyzed from the weight of the gradient. The gradients of the cross-entropy loss and the variational loss are shown in Eq. 6 and Eq. 7, respectively. The unlabeled data are treated as noisy negative data in Eq. 6.

$$\frac{\partial \hat{\mathcal{L}}_{ce} f(x)}{\partial \theta} = \sum_{i=1}^{n_u} \frac{\nabla_{\theta} f(x_i^u)}{k(1 - f(x_i^u))} - \sum_{i=1}^{n_p} \frac{\nabla_{\theta} f(x_i^p)}{k f(x_i^p)}, \quad (6)$$

$$\frac{\partial \hat{\mathcal{L}}_{var} f(x)}{\partial \theta} = \sum_{i=1}^{n_u} \frac{\nabla_{\theta} f(x_i^u)}{\sum_{i=1}^{n_u} f(x_i^u)} - \sum_{i=1}^{n_p} \frac{\nabla_{\theta} f(x_i^p)}{n_p f(x_i^p)}. \quad (7)$$

By calculating the gradient of a batch of data from Eq. 6, the positive data labeled as unlabeled will be given a larger weight if the classifier correctly identifies the sample, and then the neural network will overfit the sample with the wrong label. However, the variational loss treats each unlabeled sample fairly by assigning the same weight $1/\sum_{i=1}^{n_u} f(x_i^u)$, to each unlabeled sample from Eq. 7, which can alleviate the classifier from overfitting these mislabeled positive samples.

The second characteristic of variational loss is the problem of the unlabeled data dominating the optimization process, which makes it difficult for neural networks to find a balance between the underfitting and overfitting of positive data. This phenomenon can be demonstrated by studying the dynamic changes of the positive part of the variational loss (dubbed positive loss) (Fig. 2). As shown in Fig. 2b, although the total loss ($\hat{\mathcal{L}}_{var}(f(x))$) decreases as the training progresses, the positive loss shows an increasing trend in the early training stage (Fig. 2a). In other words, the unlabeled data dominate the optimization process. This phenomenon leads to sub-optimal F1-scores and an erratic training process (Fig. 2c). The number of iterations is uncertain when the unlabeled data dominate training, which leading to a significantly large standard deviation of F1-score in Fig. 2c.

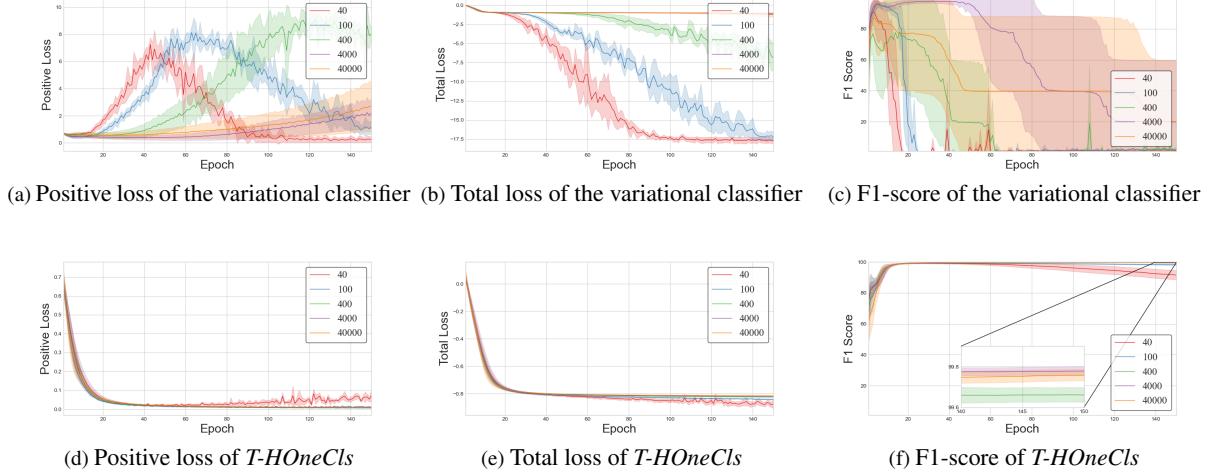


Figure 2: The curves of loss and F1-score of the variational classifier and *T-HOneCls* with different positive samples in the training stage (taking the cotton in the HongHu dataset as an example). The first row show the curves of the variational classifier, and the second row show the curves of the classifier proposed in this paper. The less positive class training data, the faster the variational model collapses.

Although the positive loss will decrease when the number of positive data is small, F1-score will not steadily increase, which indicates that the network has changed from underfitting to overfitting of positive data, rapidly. The smaller the number of positive training samples, the more obvious the instability in the training process, which can be shown in Fig. 2c.

One of the potential factors for training instability is the large weight given to the gradient of the unlabeled data. A simple example is illustrated: the flexible neural networks can very easily overfit to the training data, which makes $f(x_i^u)$ keep going to 0, and causes the weight of the gradient of the unlabeled samples to keep increasing. Based on the above analyses, a new loss function is designed in the following.

Taylor Series Expansion for Variational Loss The Taylor series expansion is introduced into the variational principle to reduce the weight of the gradient of the unlabeled data and simultaneously satisfy the variational principle, that is, the loss should be greater than or equal to the variational upper bound (\mathcal{L}_{var}).

If a given $h(x)$ is differentiable at $x = x_0$ to order o , the Taylor series of $h(x)$ is:

$$h(x) = \sum_{i=0}^{\infty} \frac{h^{(i)}(x_0)}{i!} (x - x_0)^i, \quad (8)$$

where the i -th order derivative of $h(x)$ at x_0 is $h^{(i)}(x_0)$. If the $h(x)$ is defined as $h(x) = \log(x)$, then we set $x_0 = 1$,

and for $\forall i \geq 1$,

$$h^{(i)}(x_0 = 1) = (-1)^{i-1}(i-1)!, \quad (9)$$

then the $\log(E_u[f(x)])$ can be expressed as

$$\log(E_u[f(x)]) = \sum_{i=1}^{\infty} -\frac{(1 - E_u[f(x)])^i}{i}. \quad (10)$$

If the finite terms are reserved, the variational loss can be approximated as

$$\mathcal{L}_{Tar}(f(x)) = \sum_{i=1}^o -\frac{(1 - E_u[f(x)])^i}{i} - E_p[\log(f(x))], \quad (11)$$

where $o \in \mathcal{N}_+$ denotes the order of the Taylor series. The *Taylor variational loss* can be calculated from the empirical averages over \mathcal{P} and \mathcal{U} by

$$\hat{\mathcal{L}}_{Tar}(f(x)) = \sum_{i=1}^o -\frac{\sigma_u^i}{i} - \frac{\sigma_p}{n_p}, \quad (12)$$

where $\sigma_u = 1 - \frac{1}{n_u} \sum_{i=1}^{n_u} f(x_i^u)$ and $\sigma_p = \sum_{i=1}^{n_p} \log(f(x_i^p))$.

The proposed *Taylor variational loss* can effectively alleviate the problem of training instability. It is obvious that

$$\mathcal{L}_{Tar}(f(x)) \geq \mathcal{L}_{var}(f(x)). \quad (13)$$

The effectiveness of the *Taylor variational loss* can be further illustrated from the weight of the gradient of the unlabeled data. The detailed proof is as follows:

If we let

$$\hat{\mathcal{L}}_{Tar-u}(f(x)) = \sum_{i=1}^o -\frac{\sigma_u^i}{i}, \quad (14)$$

and then,

$$\frac{\partial \hat{\mathcal{L}}_{Tar-u} f(x)}{\partial \theta} = \frac{1}{n_u} \sum_{i=1}^o \sigma_u^{i-1} \sum_{i=1}^{n_u} \nabla_{\theta} f(x_i^u). \quad (15)$$

Given that $0 < \sum_{i=1}^{n_u} f(x_i^u) < n_u$, then

$$\frac{\partial \hat{\mathcal{L}}_{Tar-u} f(x)}{\partial \theta} = \frac{1 - \sigma_u^o}{\sum_{i=1}^{n_u} f(x_i^u)} \sum_{i=1}^{n_u} \nabla_{\theta} f(x_i^u). \quad (16)$$

More proof of Eq. 16 can be found in Appendix 2.

According to Eq. 16, as with the variational loss, the *Taylor variational loss* also assigns the same weight to each unlabeled sample, but the weight of the unlabeled sample in the *Taylor variational loss* is less than that in variational loss if the finite terms are reserved, as shown in Eq. 17, which prevents the gradients of the unlabeled samples from being given too much weight and then avoids the unlabeled samples dominating the optimization process of the neural network.

$$\frac{1}{\sum_{i=1}^{n_u} f(x_i^u)} - \frac{\sigma_u^o}{\sum_{i=1}^{n_u} f(x_i^u)} < \frac{1}{\sum_{i=1}^{n_u} f(x_i^u)}. \quad (17)$$

As o gets larger, the weight of the gradient of the unlabeled samples in *Taylor variational loss* is convergent to that of variational loss for a given classifier.

3.2. Self-calibrated Optimization

Self-calibrated optimization is aimed at improving the performance of the classifier from the optimization process by using additional supervisory signals from the neural network itself. Specifically, *KL-Teacher* is proposed to utilize the memorization ability of the neural network, to stabilize the training process and alleviate the overfitting problem with a large pool of unlabeled data.

The memorization ability [1] of the neural network can also be observed when using variational-based loss to train the neural network. As the number of training epochs increases, the F1-score of the test set will first rise and then decrease until convergence, as shown by the curves of the F1-score in Fig. 2c, especially when the number of labeled samples is limited (40 labeled samples).

In order to capture the supervisory signal brought by the memorization ability of the neural network, two neural networks with the same architecture are used, with one being

the teacher network (T) and the other the student network (S). The weights of the teacher network (θ_T^t , where t is the number of iterations) are updated by the exponential moving average (EMA) of the student network, as follows:

$$\theta_T^t = \alpha \theta_T^{t-1} + (1 - \alpha) \theta_S^t. \quad (18)$$

Due to the utilization of the EMA, the teacher network acts as an “F1-score filter” and can obtain more stable classification results, which is demonstrated in Section 4.

A consistency loss (\mathcal{L}_{kl}) based on KL divergence is used to force the teacher network and the student network to have the same output, which can be used as an additional supervisory signal to alleviate the overfitting problem of the student network from a large pool of unlabeled data:

$$\mathcal{L}_{kl} = KL(p_T || p_S) + KL(p_S || p_T), \quad (19)$$

where p_T and p_S are the probabilistic outputs of the teacher network and the student network, respectively. The objective function of the student network is:

$$\mathcal{L}_S = \mathcal{L}_{Tar} + \beta \mathcal{L}_{kl}. \quad (20)$$

The output of the teacher network is used as the final classification result.

A detailed description of the training of *T-HOneCls* is provided in Appendix 3. More ablation experiments about EMA and \mathcal{L}_{kl} can be found in Section 4.

4. Experimental Results and Analysis

4.1. Experimental Settings

Datasets 7 challenging datasets were used, including 3 UAV hyperspectral datasets (HongHu, LongKou, and HanChuan, 15 tasks in total) [47], 2 HSI classification datasets (Indian Pines and Pavia University, 4 tasks in total) and 2 RGB datasets (CIFAR-10 and STL-10). More detailed information can be found in Appendix 4.

PU learning on UAV hyperspectral datasets is a challenging task. These UAV datasets mainly contain visually indistinct crops, and have strong inter-class similarity and intra-class variation. The UAV HSI along with the ground truth and spectral curves as an example are shown in Appendix 4. It can be seen that the spectral curves of the vegetation are very similar. In particular, there are shadows in the HanChuan dataset, which significantly increase the intra-class variability. In UAV datasets, some ground objects with very high textural and spectral similarity were selected for classification. For 5 HSI datasets, only 100 positive samples for each class were used to simulate the situation of limited training samples to train the neural network.

CIFAR-10 and STL-10 were used to verify the effectiveness of the proposed \mathcal{L}_{Tar} compared with other state-of-the-art PU learning methods.

Class	Class prior-based classifiers		Label noise representation learning				Class prior-free classifiers		
	nnPU [21]	OC Loss [44]	MSE Loss [10]	GCE Loss [42]	SCE Loss [35]	TCE Loss [7]	PAN [17]	vPU [3]	T-HOneCls
Cotton	99.44(0.32)	99.44(0.25)	17.08(8.25)	18.39(4.80)	96.34(2.36)	20.11(6.31)	16.66(1.40)	1.86(0.48)	98.15(0.35)
Rape	82.06(0.71)	81.81(1.23)	96.32(0.72)	96.69(0.72)	97.35(0.18)	97.64(0.12)	77.89(10.17)	8.31(1.10)	97.81(0.16)
Chinese cabbage	0.00(0.00)	88.06(2.89)	93.61(0.55)	94.06(0.60)	93.78(0.63)	94.19(0.43)	92.31(1.34)	24.89(1.22)	94.25(0.70)
Cabbage	54.20(49.50)	89.79(1.27)	99.20(0.21)	99.10(0.18)	99.12(0.20)	99.30(0.08)	98.18(0.28)	34.84(2.51)	99.37(0.07)
Tuber mustard	23.99(0.21)	23.57(0.22)	95.23(0.66)	96.05(0.56)	95.50(0.87)	96.60(0.11)	92.17(1.79)	23.28(1.19)	97.38(0.35)
Macro F1	51.94	76.53	80.29	80.86	96.42	81.57	75.44	18.64	97.39
Macro F1 of supervised binary classifier						75.62			

Table 1: The F1-scores for the HongHu dataset

Class	Class prior-based classifiers		Label noise representation learning				Class prior-free classifiers		
	nnPU [21]	OC Loss [44]	MSE Loss [10]	GCE Loss [42]	SCE Loss [35]	TCE Loss [7]	PAN [17]	vPU [3]	T-HOneCls
Strawberry	89.16(1.49)	89.52(1.54)	33.69(5.71)	34.56(2.53)	92.44(0.96)	77.69(18.03)	30.95(0.88)	9.40(0.97)	94.58(1.28)
Cowpea	59.66(3.63)	58.97(3.56)	46.55(3.39)	46.27(2.38)	70.98(7.69)	56.82(3.09)	43.95(1.08)	12.83(1.00)	90.31(1.13)
Soybean	43.63(3.14)	42.34(1.06)	97.42(0.94)	97.26(1.06)	97.19(1.11)	98.55(0.59)	86.74(4.51)	38.73(2.36)	99.13(0.28)
Watermelon	11.76(0.36)	12.23(0.46)	94.02(0.74)	93.79(0.98)	93.45(0.94)	92.67(0.84)	91.99(0.45)	54.77(2.43)	92.99(0.90)
Road	0.00(0.00)	89.40(4.34)	76.54(4.98)	74.53(3.88)	85.71(1.84)	86.29(2.06)	61.56(1.93)	25.02(1.63)	91.73(1.06)
Water	95.25(0.81)	94.90(0.63)	87.52(9.20)	92.12(5.26)	96.97(0.49)	94.15(4.70)	73.08(24.40)	1.43(0.98)	98.37(0.32)
Macro F1	49.91	64.56	72.62	73.09	89.46	84.36	64.71	23.70	94.52
Macro F1 of supervised binary classifier						66.96			

Table 2: The F1-scores for the HanChuan dataset

Training Details As for hyperspectral datasets, following [44], this paper used FreeOCNet as the fully convolutional neural network. As shown in Appendix 5, FreeOCNet includes an encoder, decoder, and lateral connection. More details about FreeOCNet can be found in [44]. In order to make a fair comparison, all the methods used the same network and the same common hyperparameters. If not specified, the order of the Taylor expansion in *T-HOneCls* is 2, and $\alpha = 0.99$. $\beta = 0.5$ in the HongHu, LongKou, Indian Pines and Pavia University datasets, and $\beta = 0.2$ in the HanChuan dataset. As for RGB datasets, 7-layer CNN was used for CIFAR-10 and STL-10. The settings of these common hyperparameters are listed in Appendix 4. The experiments were conducted using an NVIDIA RTX 3090 GPU.

Metrics The F1-score were selected as the metric to measure the performance in HSI datasets. The precision and recall are shown in Appendix 6 as supplements. The macro F1-score is the average of the F1-scores over the selected classes, which can measure the robustness of a classifier on different ground objects. The overall accuracy (OA) were selected as the metric in RGB datasets. Without special instructions, all the experiments were repeated five times, and the mean and standard deviation values are reported.

Methods There were three types of comparison algorithms in HSI datasets. Firstly, the proposed method—*T-HOneCls*—is compared with the class prior based classifiers, i.e., nnPU [21] and OC Loss [44]. The class priors

were estimated by the KMPE [29]. Methods of label noise representation learning were also compared, i.e., MSE Loss [10], GCE Loss [42], SCE Loss [35], TCE Loss [7]. What is more, the proposed method was also compared with the state-of-the-art class prior-free PU classifiers from the machine learning community, i.e., PAN [17] and vPU [3]. As a supplement, unlabeled data is used as negative class to illustrate that the performance of supervised binary classifier is limited in one-class scenarios.

As for RGB datasets, the proposed $\mathcal{L}_{T\text{ar}}$ is compared with other state-of-the-art PU learning methods: nnPU [21], PUET [36], DistPU [45], P3MIX [22] and \mathcal{L}_{var} [3].

4.2. Results on Hyperspectral Datasets

The results of hyperspectral data are listed in Table 1–Table 4. Limited by the number of pages, the distribution maps are shown in Appendix 6.

From the macro F1-score, *T-HOneCls* achieves the best results in all UAV datasets, which fully demonstrates the robustness of the proposed algorithm. A more detailed analysis follows: 1) It is clear that, without the limitation of the class prior, the macro F1-score of *T-OneCls* is significantly higher than that of the class prior-based methods. The class prior estimation for cotton is accurate, and the best F1-score for the cotton is obtained by the class prior-based methods; however, the F1-score drops when the estimated class prior is inaccurate (e.g., tuber mustard). 2) Compared with the label noise representation learning methods, *T-HOneCls* achieves a better F1-score in 17 of the 19 tasks, which indicates the necessity for developing a PU algorithm instead

Class	Class prior-based classifiers		Label noise representation learning				Class prior-free classifiers		
	nnPU [21]	OC Loss [44]	MSE Loss [10]	GCE Loss [42]	SCE Loss [35]	TCE Loss [7]	PAN [17]	vPU [3]	T-HOneCls
Corn	98.54(2.24)	99.67(0.11)	99.44(0.27)	99.16(0.25)	98.50(0.87)	98.82(0.70)	97.16(2.10)	8.54(1.03)	99.70(0.12)
Sesame	10.97(24.52)	75.95(2.78)	99.77(0.07)	99.77(0.09)	99.78(0.03)	99.79(0.09)	99.73(0.04)	67.99(13.73)	99.82(0.07)
Broad-leaf soybean	84.69(1.11)	88.02(0.26)	81.98(2.84)	87.29(1.67)	87.03(3.36)	74.94(3.48)	58.23(6.90)	4.47(0.25)	92.64(0.89)
Rice	0.00(0.00)	99.70(0.39)	98.94(0.24)	99.19(0.14)	99.16(0.24)	98.78(0.84)	98.63(0.40)	34.94(1.28)	99.50(0.16)
Macro F1	48.55	90.84	95.03	96.35	96.12	93.09	88.44	28.98	97.92
Macro F1 of supervised binary classifier						90.49			

Table 3: The F1-scores for the LongKou dataset

Class	Class prior-based classifiers		Label noise representation learning				Class prior-free classifiers		
	nnPU [21]	OC Loss [44]	MSE Loss [10]	GCE Loss [42]	SCE Loss [35]	TCE Loss [7]	PAN [17]	vPU [17]	T-HOneCls
India Pines-2	42.30(0.73)	43.14(0.96)	85.30(1.19)	86.16(2.19)	86.89(0.77)	88.60(1.45)	82.54(1.45)	8.44(1.72)	93.40(0.50)
India Pines-11	63.35(1.01)	62.88(0.46)	75.95(2.64)	77.04(2.30)	83.65(1.34)	83.03(1.73)	65.22(3.69)	3.40(0.62)	91.86(1.14)
Pavia University-2	89.17(2.60)	90.75(0.80)	93.52(1.24)	91.29(1.45)	92.38(2.54)	90.41(1.14)	89.92(3.49)	10.74(2.32)	95.01(1.04)
Pavia University-8	0.00(0.00)	82.63(3.46)	90.90(0.67)	91.27(1.46)	88.67(1.46)	92.05(0.77)	87.08(1.59)	37.46(2.20)	91.89(1.81)

Table 4: The F1-scores for the Indian Pines and Pavia University datasets

of directly applying the label noise representation learning methods to HSI. 3) Compared with the recent class prior-free methods proposed by the machine learning community, *T-HOneCls* obtains a better F1-score on all tasks.

Another conclusion is that the proposed *T-HOneCls* can balance the precision and recall. As shown in Appendix 6, most other methods cannot obtain high precision and recall at the same time, that is, these methods cannot find a balance between the overfitting and underfitting of the training data. This balance was found by *T-HOneCls*, and a good F1-score was obtained by *T-HOneCls* in all tasks.

4.3. Results on CIFAR-10 and STL-10

The experimental results on RGB datasets show that $\mathcal{L}_{T_{ar}}$ is not limited to hyperspectral data, and $\mathcal{L}_{T_{ar}}$ also performs well in other PU learning tasks. The OA of $\mathcal{L}_{T_{ar}}$ is better than that of other state-of-the-art PU learning methods (Table 5), and the curves of loss and OA can also prove the effectiveness of the proposed $\mathcal{L}_{T_{ar}}$ (Fig. 3).

4.4. Ablation Experiments Analysis

Analysis of the Training Process and Training Samples

The curves of *T-HOneCls* for the positive class and the total loss of the different positive training samples of cotton in the HongHu dataset are shown in Fig. 2d and Fig. 2e, respectively. The curves of the F1-score are also shown (Fig. 2f). The variational loss using fewer training samples will lead to the gradient domination optimization process of unlabeled samples at the beginning of the training, which makes the loss of positive class rise at the beginning of the training. Although the loss of the positive samples decreases as the training progresses, for example, 40, 100, or 400, the F1-score is unstable, and determining the optimal training epoch is very challenging without using additional data. The total loss of cotton of vPU shows large reduction

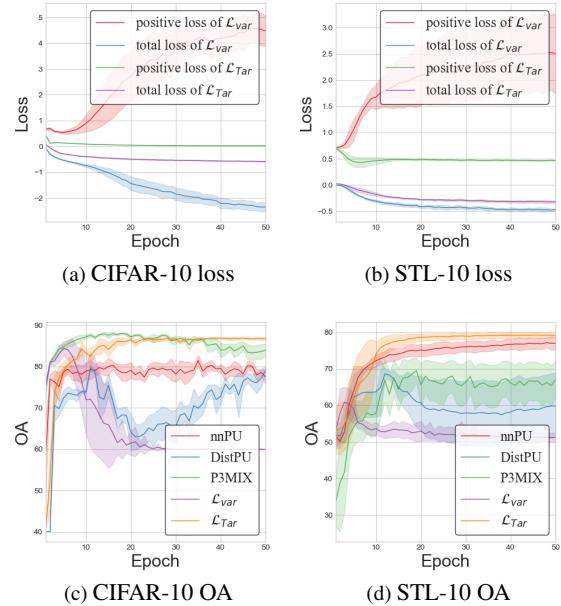


Figure 3: The curves of loss and OA on CIFAR-10 and STL-10 datasets.

Datasets	nnPU [21]	PUET [36]	DistPU [45]	P3MIX [22]	\mathcal{L}_{var} [3]	\mathcal{L}_{tar}
CIFAR-10	77.53(2.04)	75.60(0.10)	79.15(1.12)	83.99(1.68)	60.00(0.00)	86.76(0.35)
STL-10	76.98(1.91)	75.67(0.22)	59.83(10.03)	67.05(5.58)	51.26(1.46)	79.17(0.71)

Table 5: The OA of different methods on CIFAR-10 and STL-10 datasets. Definitions of classes ('Positive' vs 'Negative') are as follows: CIFAR-10: '0,1,8,9' vs '2,3,4,5,6,7'. STL-10: '0,2,3,8,9' vs '1,4,5,6,7'.

in Fig. 2, however, the F1 (1.86) is very poor, which is be-

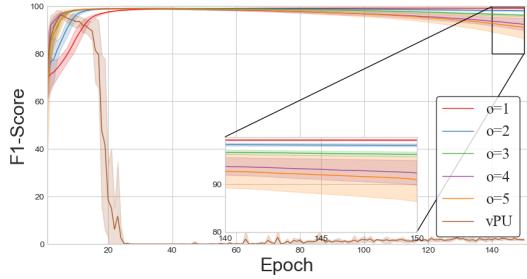


Figure 4: The F1-score curves (cotton in the HongHu dataset) for the different order of the Taylor series.

cause vPU overfits the noisy negative data (i.e., unlabeled data). These shortcomings can be solved by the proposed \mathcal{L}_{Tar} due to the reduction of the weight of the gradient of unlabeled data. More analysis can be found in Appendix 7.

Analysis of the Order of the Taylor Series One of the contributions of this paper is that we point out that the reason for the poor performance of variational loss is that the gradient of the unlabeled data is given too much weight, which can be tackled by the proposed *Taylor variational loss*. The order of the Taylor expansion is analyzed as a hyperparameter in this subsection, and the F1-score curves of cotton in the HongHu dataset are shown in Fig. 4 as an example. Five other ground objects were also analyzed, and the results are displayed in Appendix 8. As shown in Fig. 4, the neural networks converge to a poor result with variational loss. An empirical conclusion can be obtained from the order analysis: the higher the order of the Taylor expansion, the faster the neural network converges. However, the rapid convergence of the neural network can lead to overfitting. In other words, the classification results will rise first and then decline with the progress of the training.

Analysis of KL-Teacher This subsection analyzes the advantages of the proposed self-calibrated optimization. Three ground objects from the three datasets were selected as examples to demonstrate the advantages of self-calibration optimization. The F1-score curves of cowpea in the HanChuan dataset are shown in Fig. 5 and other classes are shown in Appendix 9.

It can be seen from Table 6 that the training is failed, if \mathcal{L}_{var} with self-calibrated optimization is used. It can be seen from Fig. 5 that the F1-score fluctuates greatly when only stochastic gradient descent is used to optimize the *Taylor variational loss*. The EMA has the function of an “F1-score filter”, which makes the F1-score of the teacher model more stable. The EMA allows the teacher model to lag behind the student model, and due to the memorization ability of the neural network, the F1-score of the lagged neu-

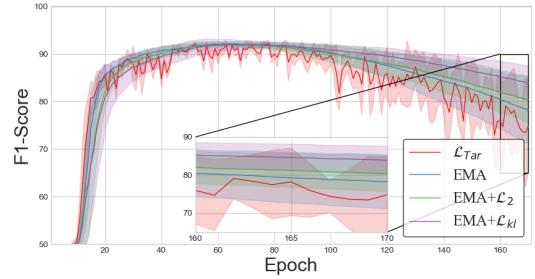


Figure 5: The F1-score curves (cowpea in the HanChuan dataset, $o=5$) for the different components of *KL-Teacher*.

Class	Order	\mathcal{L}	Self-calibrated optimization			F1-score
			EMA	\mathcal{L}_2	\mathcal{L}_{kl}	
Cotton	2	\mathcal{L}_{var}	✓		✓	0.00
		\mathcal{L}_{Tar}				97.51
		\mathcal{L}_{Tar}	✓			97.58
		\mathcal{L}_{Tar}	✓	✓		97.61
		\mathcal{L}_{Tar}	✓		✓	98.15
	5	\mathcal{L}_{Tar}				72.01
		\mathcal{L}_{Tar}	✓			84.27
		\mathcal{L}_{Tar}	✓	✓		81.25
		\mathcal{L}_{Tar}	✓		✓	91.01
Broad-leaf soybean	2	\mathcal{L}_{var}	✓		✓	0.12
		\mathcal{L}_{Tar}				90.74
		\mathcal{L}_{Tar}	✓			91.22
		\mathcal{L}_{Tar}	✓	✓		91.42
		\mathcal{L}_{Tar}	✓		✓	92.64
	5	\mathcal{L}_{Tar}				81.06
		\mathcal{L}_{Tar}	✓			81.61
		\mathcal{L}_{Tar}	✓	✓		81.78
		\mathcal{L}_{Tar}	✓		✓	82.79
Cowpea	2	\mathcal{L}_{var}	✓		✓	4.00
		\mathcal{L}_{Tar}				88.87
		\mathcal{L}_{Tar}	✓			88.59
		\mathcal{L}_{Tar}	✓	✓		88.78
		\mathcal{L}_{Tar}	✓		✓	90.31
	5	\mathcal{L}_{Tar}				74.78
		\mathcal{L}_{Tar}	✓			78.20
		\mathcal{L}_{Tar}	✓	✓		80.38
		\mathcal{L}_{Tar}	✓		✓	83.90

Table 6: Analysis of KL-Teacher

ral network is better than that of the student network at the later stage of training. The use of consistency loss can promote the output of the student model to approximate that of the teacher model with a large pool of unlabeled data, so as to alleviate the overfitting problem. If L2 loss (\mathcal{L}_2) is regarded as the consistency loss, it is equivalent to Mean-Teacher [34] being used. However, according to the results listed in Table 6, \mathcal{L}_{kl} can more effectively alleviate the overfitting of the student model.

5. Conclusion

In this paper, we have focused on tackling the problem of limited labeled HSI PU learning without class-prior. The

proposed *Taylor variational loss* is responsible for the task of learning from limited labeled PU data without a class prior. The *self-calibrated optimization* proposed in this paper is used to stabilize the training process. The extensive experiments (7 datasets, 21 tasks in total) demonstrated the superiority of the proposed method.

Acknowledgements: This work was supported by National Key Research and Development Program of China under Grant No.2022YFB3903502, National Natural Science Foundation of China under Grant No.42325105, 42071350, 42101327, and LIESMARS Special Research Funding.

References

- [1] Devansh Arpit, Stanisław Jastrzudebski, Nicolas Ballas, David Krueger, Emmanuel Bengio, Maxinder S. Kanwal, Tegan Maharaj, Asja Fischer, Aaron Courville, Yoshua Bengio, and Simon Lacoste-Julien. A closer look at memorization in deep networks. In *Proceedings of the International Conference on Machine Learning*, pages 233–242, 2017.
- [2] Shizhen Chang, Bo Du, and Liangpei Zhang. Positive unlabeled learning with class-prior approximation. In *Proceedings of the International Joint Conference on Artificial Intelligence*, 2021.
- [3] Hui Chen, Fangqing Liu, Yin Wang, Liyue Zhao, and Hao Wu. A variational approach for learning from positive and unlabeled data. In *Advances in Neural Information Processing Systems*, volume 33, pages 14844–14854. Curran Associates, Inc., 2020.
- [4] Xiuhua Chen, Chen Gong, and Jian Yang. Cost-sensitive positive and unlabeled learning. *Information Sciences*, 558:229–245, 2021.
- [5] Yushi Chen, Zhouhan Lin, Xing Zhao, Gang Wang, and Yanfeng Gu. Deep learning-based classification of hyperspectral data. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 7(6):2094–2107, 2014.
- [6] J. Feng, H. Yu, L. Wang, X. Cao, X. Zhang, and L. Jiao. Classification of hyperspectral images based on multiclass spatial-spectral generative adversarial networks. *IEEE Transactions on Geoscience and Remote Sensing*, 57(8):5329–5343, 2019.
- [7] Lei Feng, Senlin Shu, Zhuoyi Lin, Fengmao Lv, Li Li, and Bo An. Can cross entropy loss be robust to label noise? In *Proceedings of the International Joint Conference on Artificial Intelligence*, 2021.
- [8] Giles M. Foody, Ajay Mathur, Carolina Sanchez-Hernandez, and Doreen S Boyd. Training set size requirements for the classification of a specific class. *Remote Sensing of Environment*, 104(1):1–14, 2006.
- [9] Pedram Ghamisi, Javier Plaza, Yushi Chen, Jun Li, and Antonio J Plaza. Advanced spectral classifiers for hyperspectral images: A review. *IEEE Geoscience and Remote Sensing Magazine*, 5(1):8–32, 2017.
- [10] Aritra Ghosh, Himanshu Kumar, and P. S. Sastry. Robust loss functions under label noise for deep neural networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 1919–1925. AAAI Press, 2017.
- [11] Jacob Goldberger and Ehud Ben-Reuven. Training deep neural-networks using a noise adaptation layer. In *International Conference on Learning Representations*, 2017.
- [12] Tieliang Gong, Guangtao Wang, Jieping Ye, Zongben Xu, and Ming Lin. Margin based pu learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*. AAAI Press, 2018.
- [13] Bo Han, Quanming Yao, Tongliang Liu, Gang Niu, Ivor W. Tsang, James T. Kwok, and Masashi Sugiyama. A Survey of Label-noise Representation Learning: Past, Present and Future. *arXiv e-prints*, 2020.
- [14] Bo Han, Quanming Yao, Xingrui Yu, Gang Niu, Miao Xu, Weihua Hu, Ivor Tsang, and Masashi Sugiyama. Co-teaching: Robust training of deep neural networks with extremely noisy labels. In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.
- [15] R. Hang, Q. Liu, D. Hong, and P. Ghamisi. Cascaded recurrent neural networks for hyperspectral image classification. *IEEE Transactions on Geoscience and Remote Sensing*, 57(8):5384–5394, 2019.
- [16] Cho-Jui Hsieh, Nagarajan Natarajan, and Inderjit S Dhillon. Pu learning for matrix completion. In *Proceedings of the International Conference on International Conference on Machine Learning*, pages 2445–2453, 2015.
- [17] Wpeng hu, Ran Le, Bing Liu, Feng Ji, Jinwen Ma, Dongyan Zhao, and Rui Yan. Predictive adversarial learning from positive and unlabeled data. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35:7806–7814, 05 2021.
- [18] Lu Jiang, Zhengyuan Zhou, Thomas Leung, Li-Jia Li, and Li Fei-Fei. MentorNet: Learning data-driven curriculum for very deep neural networks on corrupted labels. In *Proceedings of the International Conference on Machine Learning*, volume 80, pages 2304–2313, 2018.
- [19] L. Jiao, M. Liang, H. Chen, S. Yang, H. Liu, and X. Cao. Deep fully convolutional network-based spatial distribution prediction for hyperspectral image classification. *IEEE Transactions on Geoscience and Remote Sensing*, 55(10):5585–5599, 2017.
- [20] Masahiro Kato, Takeshi Teshima, and Junya Honda. Learning from positive and unlabeled data with a selection bias. In *International Conference on Learning Representations*, 2019.
- [21] Ryuichi Kiryo, Gang Niu, Marthinus C du Plessis, and Masashi Sugiyama. Positive-unlabeled learning with non-negative risk estimator. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.
- [22] Changchun Li, Ximing Li, Lei Feng, and Jihong Ouyang. Who is your right mixup partner in positive and unlabeled learning. In *International Conference on Learning Representations*, 2022.
- [23] Jingtao Li, Xinyu Wang, Hengwei Zhao, Xin Hu, and Yanfei Zhong. Detecting pine wilt disease at the pixel level

- from high spatial and spectral resolution uav-borne imagery in complex forest landscapes using deep one-class classification. *International Journal of Applied Earth Observation and Geoinformation*, 112:102947, 2022.
- [24] Wenkai Li, Qinghua Guo, and Charles Elkan. A positive and unlabeled learning algorithm for one-class classification of remote-sensing data. *IEEE Transactions on Geoscience and Remote Sensing*, 49(2):717–725, 2011.
 - [25] Wenkai Li, Qinghua Guo, and Charles Elkan. One-class remote sensing classification from positive and unlabeled background data. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 14:730–746, 2021.
 - [26] Yuyuan Liu, Yu Tian, Yuanhong Chen, Fengbei Liu, Vasileios Belagiannis, and Gustavo Carneiro. Perturbed and strict mean teachers for semi-supervised semantic segmentation. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 4248–4257, 2022.
 - [27] Ying Lu and Le Wang. How to automate timely large-scale mangrove mapping with remote sensing. *Remote Sensing of Environment*, 264:112584, 2021.
 - [28] Michal Lukasik, Srinadh Bhojanapalli, Aditya Menon, and Sanjiv Kumar. Does label smoothing mitigate label noise? In *Proceedings of the International Conference on Machine Learning*, volume 119, pages 6448–6458, 2020.
 - [29] Harish Ramaswamy, Clayton Scott, and Ambuj Tewari. Mixture proportion estimation via kernel embeddings of distributions. In *Proceedings of The International Conference on Machine Learning*, volume 48, pages 2052–2060, 2016.
 - [30] Yafeng Ren, Donghong Ji, and Hongbin Zhang. Positive unlabeled learning for deceptive reviews detection. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 488–498, 2014.
 - [31] Kihyuk Sohn, David Berthelot, Nicholas Carlini, Zizhao Zhang, Han Zhang, Colin A Raffel, Ekin Dogus Cubuk, Alexey Kurakin, and Chun-Liang Li. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. In *Advances in Neural Information Processing Systems*, volume 33, pages 596–608. Curran Associates, Inc., 2020.
 - [32] Guangxin Su, Weitong Chen, and Miao Xu. Positive-unlabeled learning from imbalanced data. In *Proceedings of the International Joint Conference on Artificial Intelligence*, pages 2995–3001, 2021.
 - [33] Sainbayar Sukhbaatar, Joan Bruna, Manohar Paluri, Lubomir Bourdev, and Rob Fergus. Training convolutional networks with noisy labels. In *International Conference on Learning Representations*, 2015.
 - [34] Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.
 - [35] Yisen Wang, Xingjun Ma, Zaiyi Chen, Yuan Luo, Jinfeng Yi, and James Bailey. Symmetric cross entropy for robust learning with noisy labels. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 322–330, 2019.
 - [36] Jonathan Wilton, Abigail Koay, Ryan Ko, Miao Xu, and Nan Ye. Positive-unlabeled learning using random forests via recursive greedy risk minimization. In *Advances in Neural Information Processing Systems*, 2022.
 - [37] Y. Xu, B. Du, and L. Zhang. Beyond the patchwise classification: Spectral-spatial fully convolutional networks for hyperspectral image classification. *IEEE Transactions on Big Data*, 6(3):492–506, 2020.
 - [38] Y. Xu, L. Zhang, B. Du, and F. Zhang. Spectral–spatial unified networks for hyperspectral image classification. *IEEE Transactions on Geoscience and Remote Sensing*, 56(10):5893–5909, 2018.
 - [39] P. Yang, X. L. Li, J. P. Mei, C. K. Kwoh, and S. K. Ng. Positive-unlabeled learning for disease gene identification. *Bioinformatics*, 28(20):2640–7, 2012.
 - [40] Xingrui Yu, Bo Han, Jiangchao Yao, Gang Niu, Ivor Tsang, and Masashi Sugiyama. How does disagreement help generalization against label corruption? In *Proceedings of the International Conference on Machine Learning*, volume 97, pages 7164–7173, 2019.
 - [41] Chuang Zhang, Dexin Ren, Tongliang Liu, Jian Yang, and Chen Gong. Positive and unlabeled learning with label disambiguation. In *Proceedings of the International Joint Conference on Artificial Intelligence*, pages 4250–4256, 2019.
 - [42] Zhilu Zhang and Mert Sabuncu. Generalized cross entropy loss for training deep neural networks with noisy labels. In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.
 - [43] Hengwei Zhao, Yanfei Zhong, Xinyu Wang, Xin Hu, Chang Luo, Mark Boitt, Rami Piironen, Liangpei Zhang, Janne Heiskanen, and Petri Pellikka. Mapping the distribution of invasive tree species using deep one-class classification in the tropical montane landscape of kenya. *ISPRS Journal of Photogrammetry and Remote Sensing*, 187:328–344, 2022.
 - [44] Hengwei Zhao, Yanfei Zhong, Xinyu Wang, and Hong Shu. One-class risk estimation for one-class hyperspectral image classification. *IEEE Transactions on Geoscience and Remote Sensing*, pages 1–1, 2023.
 - [45] Yunrui Zhao, Qianqian Xu, Yangbangyan Jiang, Peisong Wen, and Qingming Huang. Dist-pu: Positive-unlabeled learning from a label distribution perspective. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 14441–14450, 2022.
 - [46] Zhuo Zheng, Yanfei Zhong, Ailong Ma, and Liangpei Zhang. Fpga: Fast patch-free global learning framework for fully end-to-end hyperspectral image classification. *IEEE Transactions on Geoscience and Remote Sensing*, 58(8):5612–5626, 2020.
 - [47] Yanfei Zhong, Xin Hu, Chang Luo, Xinyu Wang, Ji Zhao, and Liangpei Zhang. Whu-hi: Uav-borne hyperspectral with high spatial resolution (h2) benchmark datasets and classifier for precise crop identification based on deep convolutional neural network with crf. *Remote Sensing of Environment*, 250:112012, 2020.