An Entropic Optimal Transport loss for learning deep neural networks under label noise in remote sensing images

Bharath Bhushan Damodaran, Rémi Flamary, Viven Seguy, Nicolas Courty

Research Highlights (Required)

- Studied robustness of deep neural networks under the presence of inaccurately labeled samples in remote sensing images
- Proposed an novel loss function based on entropic optimal transport loss to make deep neural networks tolerant to label noise
- Validated on several benchmarks on two very different applications with different deep learning architectures
- Achieved impressive performance compared to the state-of-the-art methods

Computer Vision and Image Understanding
journal homepage: www.elsevier.com

An Entropic Optimal Transport Loss for Learning Deep Neural Networks under Label Noise in Remote Sensing Images

Bharath Bhushan Damodaran[a,**], Rémi Flamary[b], Viven Seguy[c], Nicolas Courty[a]

[a] *Université de Bretagne Sud, IRISA, UMR 6074, CNRS, Vannes-56000, France*
[b] *Université Côte d'Azur, Lagrange, UMR 7293, CNRS, OCA, Nice-06108, France*
[c] *Kyoto University, Japan*

ABSTRACT

Deep neural networks have established as a powerful tool for large scale supervised classification tasks. The state-of-the-art performances of deep neural networks are conditioned to the availability of large number of accurately labeled samples. In practice, collecting large scale accurately labeled datasets is a challenging and tedious task in most scenarios of remote sensing image analysis, thus cheap surrogate procedures are employed to label the dataset. Training deep neural networks on such datasets with inaccurate labels easily overfits to the noisy training labels and degrades the performance of the classification tasks drastically. To mitigate this effect, we propose an original solution with entropic optimal transportation. It allows to learn in an end-to-end fashion deep neural networks that are, to some extent, robust to inaccurately labeled samples. We empirically demonstrate on several remote sensing datasets, where both scene and pixel-based hyperspectral images are considered for classification. Our method proves to be highly tolerant to significant amounts of label noise and achieves favorable results against state-of-the-art methods.

1. Introduction

Deep learning has been applied with tremendous success on a variety of tasks in remote sensing image analysis. For instance, achievement of state-of-the-art performance in scene classification (Cheng et al., 2018; Anwer et al., 2018), pixel-wise labeling of both multispectral (Huang et al., 2018; Audebert et al., 2018; Maggiori et al., 2017) and hyperspectral datasets (Zhong et al., 2018; Wang et al., 2017), object detection (Kellenberger et al., 2018) and image retrieval (Zhou et al., 2018; Ye et al., 2017; Li et al., 2018), highlights the recent success of deep learning models in remote sensing. But these phenomenal performances is highly dependant on the availability of large collection of datasets with accurate annotations (labels). If either the size of the dataset or the accuracy of the labels is not sufficient (i.e, small scale datasets or inaccurate labels), the performance of the deep learning methods could suffer drastically. The former one can be addressed to some degree by data augmentation strategies, however solving the later case of inaccurate labeling is more difficult.

To address the large scale data requirements of deep learning methods, new datasets have been proposed recently in the remote sensing community (Zhou et al., 2018; Huang et al., 2018; Cheng et al., 2017; Kemker et al., 2017; Wang et al., 2016; Xia et al., 2017). This trend will grow continuously in the coming years, due for instance to the large constellation of the Earth observation satellites. One of the major challenge in collecting this new large scale data is accurate labeling of the samples. Manual expert labeling of such large collection of samples is often not feasible and not cost-effective. Thus, labeling is usually performed by non-experts through crowd sourcing (Snow et al., 2008; Haklay, 2010), keyword query through search engine in the case of images, open street maps, and out-dated classification maps (Kaiser et al., 2017). These

**Corresponding author: Tel.: +33(0)297017266; fax: +0-000-000-0000;
e-mail: bharath-bhushan.damodaran@irisa.fr (Bharath Bhushan Damodaran)

cheap surrogate procedures allows scaling the size of labeled datasets, but at the cost of introducing label noise (i.e. inaccurately labeled samples).

Fig. 1: Examples of images from classes medium residential area (top row) and tennis ground (bottom row).

Even when manual experts are involved in labeling the data samples, they must be provided with sufficient information; otherwise inaccurate labeling may still occur (for instance, during the field survey) (Hickey, 1996). Note that in the some applications, labeling is a subjective task (Smyth et al., 1995) that can again introduce label noise. For instance, figure 1 shows images that have close similarities between two classes Medium residential and Tennis ground. It is easy for a human annotator to miss the Tennis ground and to label it as Medium residential leading to label noise in the dataset. Furthermore, the label noise could occur due to the misregistration of satellite images. Hence in general, large scale datasets might mostly contain inaccurately labeled samples or affected by label noise. In this case, when deep learning methods are employed with conventional loss functions (for instance, categorical cross entropy, mean square error), they will not be robust to label noise, and as a result the classification accuracy decreases significantly (Zhang et al., 2017). This calls for robust approaches to mitigate the impact of label noise on the deep learning methods. This paper exactly lies in this direction, to increase the generalization ability of deep learning models under the label noise scenario.

Recently, it was shown that while training deeper neural networks, models tend to memorize the training data, and this phenomena is more severe when the dataset is affected by the label noise (Zhang et al., 2017). The impact of the label noise in the deep learning models can be partly circumvented by regularization techniques such as drop out layers, and weight regularization. These standard procedures make neural networks robust to some extend, but they are still prone to memorize noisy labels for medium-to-large noise levels. The problem of learning with noisy labels has been long studied in machine learning (Frenay and Verleysen, 2014; Brooks, 2011; Zhu and Wu, 2004; Sáez et al., 2014; Hickey, 1996; Smyth et al., 1995; Natarajan et al., 2013), but still only few works have focused on neural networks. Recently, new approaches have been proposed in the computer vision and machine learning fields to tackle the label noise by cleaning the noisy labels or designing robust loss functions within the deep learning framework (Jiang et al., 2018; Vahdat, 2017; Patrini et al., 2017).

To mitigate the impact of label noise, one category of method relies on estimating the noise transition probability that describes the probability of i^{th} class label being mislabeled to the j^{th} class label, and use it to be robust to label noise (Vahdat, 2017; Natarajan et al., 2013; Patrini et al., 2017). Among those, some of them require a small set of clean labels to estimate the noise transition probability (Vahdat, 2017). The other category of methods proposes to use loss functions which are inherently tolerant to the label noise (Natarajan et al., 2013; van Rooyen et al., 2015; Masnadi-Shirazi, Hamed and Vasconcelos, 2008; Ghosh et al., 2015; Aritra et al., 2017). Though these methods provided satisfactory results, all of them treat the noisy samples independently in their loss functions, discarding potential information about the relationship between the samples in the feature space. Contrary to those methods, our proposal is to consider the local geometric information among the samples in the joint label/feature space through the form of similarity measure between the samples to overcome the impact of noisy labels in deep neural networks.

The primary objective of this paper is to develop a robust approach to tackle the label noise for remote sensing image analysis. The sensitiveness of deep neural networks to label noise has not been well studied in remote sensing image analysis so far as per our knowledge. Hence the first contribution of this article lies in studying the robustness of deep neural networks to label noise, and also to analyse the efficiency of existing robust loss functions for remote sensing classification tasks. The second contribution of this paper is to propose a novel robust solution to tackle the label noise based on optimal transportation theory (Villani, 2009). Indeed we propose to learn a deep learning model which is robust to label noise by fitting the model to the label-features joint distribution of the dataset with respect to the entropy-regularized optimal transport distance. We coin this method as CLEOT for Classification Loss with Entropic Optimal Transport. One major advantage of our approach compared to existing methods is that our method inherently exploits the geometric structure of the underlying data, by considering the similarity between the sample points through the ground cost in optimal transport. A stochastic approximation schemes is proposed to solve the learning problem, and allows the use of our approach within deep learning frameworks. Experiments are conducted on several remote sensing aerial and hyperspectral benchmark datasets, and the results demonstrate that our approach is more robust (tolerant) to high level label noise than current state-of-the-art methods.

The remaining of the paper is organized as follows. Section 2 discusses related works, section 3 defines the label

noise and describes the problem formulation, and section 4 introduces optimal transport. The proposed method is then presented in section 4.2 while experimental datasets and results are explained in section 5. We finally draw some conclusions in section 6.

2. Related works

2.1. Learning with noisy labels

Label noise, and attribute (feature) noise are two types of noise commonly found in machine learning datasets. The label noise is considered as more harmful and difficult to tackle compared to the attribute noise, and can decrease the classification performance significantly (Zhu and Wu, 2004). Learning with noisy labels with shallow learning methods have been widely investigated in the literature (Frenay and Verleysen, 2014; Brooks, 2011; Zhu and Wu, 2004; Sáez et al., 2014; Hickey, 1996; Smyth et al., 1995; Natarajan et al., 2013), but studies in the context of deep learning still remain scarse (but growing recently) (Mnih and Hinton, 2012; Reed et al., 2015; Vahdat, 2017; Hendrycks et al., 2018; Patrini et al., 2017). Among the several methods which have been proposed to robustly train deep neural networks on the datasets with noisy labels, one set of methods approaches the problem from the perspective of cleaning the noisy labels, and use the clean estimated labels for training deep neural networks, or they smoothly reduce the impact of noisy labels by putting smaller weights on noisy label samples, either through directed graphical models (Xiao et al., 2015), conditional random field (Vahdat, 2017), knowledge graph distillation (Li et al., 2017), meta-learning (Ren et al., 2018) or noise-transition matrix estimation (Hendrycks et al., 2018). But those methods require an additional small subset of data with clean labels, or require ground truth of pre-identified noisy labels in order to model the noise in the dataset.

A second kind of methods tries to detect clean instances out of the noisy instances, and use them to update the parameters of the trained neural network (Jiang et al., 2018; Ding et al., 2018). In this category, two deep networks or two stage framework are employed to remove noisy label instances. The last kind of methods design a robust loss function and loss correction approach. The robust loss functions unhinged (van Rooyen et al., 2015), savage (Masnadi-Shirazi, Hamed and Vasconcelos, 2008), sigmoid and ramp (Ghosh et al., 2015) are inherently robust to the label noise with associated theoretical bounds. Most of these method rely on an assumption of symmetric loss function. The loss correction approaches employ the correctness procedure to adjust the loss function to eliminate the influence of the noisy labels by forward and backward correction approach (Patrini et al., 2017) using the estimated noise transition model from the noisy labeled data, adding linear layer on top of a softmax layer (Sukhbaatar et al., 2014; Jacob and Ehud, 2017), using bootstrap approach (Reed et al., 2015) that replaces the noisy labels with a soft or hard combination of noisy labels and their predicted labels.

In remote sensing image analysis, the adverse effect of the label noise is not much studied in literature. The impact of label noise has been recently studied in (Frank et al., 2017; Pelletier et al., 2017a) with shallow classifiers. The feasibility of using online open street map (outdated or mislabeled ground truth) to obtain classification map with deep neural network was studied in (Kaiser et al., 2017), however they didn't consider directly addressing label noise as a specificity of the problem. Some other studies tackle the label noise in the context of shallow classifier (random forest, logistic regression) by selecting clean labeled instances via outlier detection (Pelletier et al., 2017b; Maas et al., 2019), or by using existing noise tolerant logistic regression method (Maas et al., 2016, 2017).

2.2. Optimal transport

Optimal transport theory provide the Wasserstein distance, that measures the discrepancy between probability distribution in a geometrically sound manner. More recently, optimal transport has found applications in domain adaptation (Courty et al., 2017b,a; Damodaran et al., 2018), generative models (Seguy et al., 2018; Genevay et al., 2017; Arjovsky et al., 2017), data mining (Courty et al., 2018) and image processing (Solomon et al., 2015; Papadakis, 2015).

Among those applications, domain adaptation is the one that is the most related to the problem of noisy labels. It indeed aim at adapting a classifier to better predict on new data whose distribution is different from the training data. In the case of label noise in the training dataset, we want to adapt the classifier to perform well on data using a different noisy dataset for training. One recent approach coined JDOT for joint distribution optimal transport (Courty et al., 2017a) propose to estimate a classifier that minimize the Wasserstein distance between the joint feature/labels distribution and a predicted (with the model) joint distribution on the new data. The approach has been recently extended to the deep learning framework in (Damodaran et al., 2018) and will be described more in detail in section 4.2.

Our proposed method CLEOT is different from Damodaran et. al (DeepJDOT) in several aspects. Both DeepJDOT and CLEOT are extensions of joint distribution optimal transport. In the former, the optimal transport coupling is computed between the samples from source and a different target distribution using a classical optimal transport formulation. In CLEOT we use the JDOT loss on the noisy distribution **itself** together with entropic regularization to promote connections between the neighboring samples, leading to more smoothness in the estimated model. This corresponds to the estimation of new soft labels through the weighted combination given by the optimal transport coupling. Furthermore, the objective function of the DeepJDOT minimizes the Wasserstein distance that is known to be non-differentiable

whereas CLEOT minimizes a smooth entropic formulation for which we can recover the unique gradients with automatic differentiation.

3. Problem formulation and noise model

3.1. Traditional supervised learning

Let $X = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\} \in \mathcal{X}^N$ be the training features/images and $Y = \{\mathbf{y}_1, \ldots, \mathbf{y}_N\} \in \mathcal{Y}^N$ be their associated one-hot encoded class labels ($\mathbf{y}_i \in \{0,1\}^{1\times c}$, c is the number of classes) sampled from the joint distribution $p(\mathbf{x}, \mathbf{y})$. Let $f : X \to Y$ be a neural network model with model parameters θ, which maps the input features into class conditional probabilities $f(\mathbf{x})_j = \hat{p}(\mathbf{y} = j|\mathbf{x})$. The loss function $L(f(\mathbf{x}), \mathbf{y})$ measures the discrepancy (error) between the true label $\mathbf{y}$ and the predicted label distribution $f(\mathbf{x})$ by the neural network. In the standard supervised learning stting, one estimates the parameters θ of f by minimizing the empirical risk on the training set

$$\min_f \quad \frac{1}{N}\sum_{i=1}^{N} L(f(\mathbf{x}_i), \mathbf{y}_i) \tag{1}$$

In this paper we use the cross-entropy defined as : $L(f(\mathbf{x}_i), \mathbf{y}_i) = \sum_{j=1}^{c} -y_{ij}\log(f(\mathbf{x}_i)_j)$, thus eq. (1) can be re-expressed as

$$\min_f \quad \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{c} -y_{ij}\log(f(\mathbf{x}_i)_j), \tag{2}$$

The neural network model f is estimated by minimizing the objective above with respect to its parameters through stochastic optimization procedures. However, minimizing the loss function eq. (2) in certain scenarios can lead to over-fitting. When the dataset is affected by label noise, minimizing the empirical risk can degrade the performance of the neural network. Hence suitable modification of the loss function is necessary to learn a robust neural network model, which is the direction of our proposed method.

In the following subsection, we describe the label noise in the datasets, and how to artificially simulate this noise in two different settings.

3.2. Label noise

Large scale datasets are commonly subjected to label noise (mislabeled samples), especially when using one of the surrogate labeling strategy discussed in the introduction. The occurrence of the label noise in the dataset can be of two types: asymmetric and symmetric label noise.

In the asymmetric label noise, each label $\mathbf{y}$ in the training set is flipped to $\tilde{\mathbf{y}}$ with probability $p(\tilde{\mathbf{y}}|\mathbf{y})$, defining the noise transition matrix, $E_{i,j} = p(\tilde{\mathbf{y}} = j|\mathbf{y} = i)\ \forall i, j$, indicating the probability of i^{th} class label being flipped to j^{th} class label. Thus, the training samples $\{\mathbf{x}_i, \tilde{\mathbf{y}}_i\}$ are observed from the joint distribution

$$p(\mathbf{x}, \tilde{\mathbf{y}}) = \sum_{\mathbf{y}} p(\tilde{\mathbf{y}}|\mathbf{y})p(\mathbf{y}|\mathbf{x})p(\mathbf{x}) = \sum_{\mathbf{y}} E_{i,j}p(\mathbf{y}|\mathbf{x})p(\mathbf{x}), \tag{3}$$

This noise model is realistic and can occur in real world scenario, where non-expert finds difficult to distinguish between the similar fine grained classes. However, the matrix $\mathbf{E}$ is generally unknown in real-world scenarios.

In the symmetric label noise, the label is flipped uniformly across all the classes with probability p_e, irrespective of similarity between the classes. In this case, matrix $\mathbf{E}$ has the entries $1 - p_e$ in the diagonal, and $\frac{p_e}{1-c}$ in the off-diagonal elements. This noise model is much simpler and has a unique parameter.

For both the noise types, learning the classifier f with the loss function mentioned in eq. (1) is not robust and can lead to overfitting to the noisy training labels.

4. Classification Loss with Entropic Optimal Transport (CLEOT)

In this section we first provide an introduction to optimal transport (OT) by discussing unregularized and regularized OT. Next we introduce the joint distribution OT which is starting point of our method. Then we formulate our approach and discuss the numerical resolution of the proposed learning problem.

4.1. Introduction to optimal transport

Optimal transport (see for instance the two monographs by Villani (Villani, 2003; Villani, 2009)) is a theory that allows to compare probability distributions in a geometrically sound manner even when their respective supports do not overlap. OT is hence well-suited to work on empirical distributions and allows to take into account the geometry of the data set in its embedding space. Formally, OT searches for a probabilistic coupling $\mathbf{T} \in \Pi(\mu_1, \mu_2)$ between two distributions μ_1 and μ_2 which yields a minimal total displacement cost wrt. a given cost function $c(\mathbf{x}_1, \mathbf{x}_2)$ measuring the dissimilarity between samples $\mathbf{x}_1$ and $\mathbf{x}_2$ on the support of each distribution μ_1 and μ_2 respectively. Here, $\Pi(\mu_1, \mu_2)$ describes the space of joint probability distributions with marginals μ_1 and μ_2. In a discrete setting (both distributions are empirical) the OT problem becomes:

$$W_C(\mu_1, \mu_2) = \min_{\mathbf{T}\in\Pi(\mu_1,\mu_2)} < \mathbf{T}, \mathbf{C} >_F, \tag{4}$$

where $\langle\cdot,\cdot\rangle_F$ is the Frobenius dot product, $\mathbf{C} \geq 0$ is a ground cost matrix $\in \mathbb{R}^{n_1\times n_2}$ representing the pairwise costs $c(\mathbf{x}_i, \mathbf{x}_j)$, $\mathbf{T}$ is a matrix of size $n_1 \times n_2$ with prescribed marginals, and n_1, n_2 the sizes of the supports of the distributions μ_1 and μ_2 respectively. The minimum of this optimization problem can be used as a measure of discrepency between distributions, and, whenever the cost c is a metric, OT is also a metric and is called the Wasserstein distance.

OT solvers have a super-cubic complexity in the size of the support of the input distributions $n = \max(n_1, n_2)$, which makes OT approaches untractable when dealing with medium to large-scale datasets. In order to speed up OT computation, Cuturi (2013) proposed instead of

the above linear program to solve a regularized version of OT. Regularization is achieved by adding the negative entropy regularization term to the coupling $\mathbf{T}$. Thus, the so-called entropy-regularized Wasserstein distance can be defined as eq. (4) is becomes

$$W_{C,\lambda}(\mu_1,\mu_2) =< \mathbf{T}^*, \mathbf{C} >_F, \tag{5}$$

with

$$\mathbf{T}^* = \underset{\mathbf{T}\in\Pi(\mu_1,\mu_2)}{\operatorname{argmin}} \; < \mathbf{T}, \mathbf{C} >_F + \lambda R(\mathbf{T}) \tag{6}$$

where $R(\mathbf{T}) = \sum_{i,j} \mathbf{T}_{i,j} \log \mathbf{T}_{i,j}$ is the negative entropy of $\mathbf{T}$, and λ is the trade-off between the two terms. When $\lambda = 0$ eq. (5) recovers the original optimal transport problem from eq. (4), and when $\lambda \to \infty$ the resulting divergence has strong links with maximum mean discrepancy as discussed in (Genevay et al., 2017). Efficient computational schemes were proposed with entropic regularization (Cuturi, 2013) as well as stochastic versions using the dual formulation of the problem (Genevay et al., 2016; Arjovsky et al., 2017; Seguy et al., 2018), allowing to tackle middle to large sized problems.

Note that the regularized Wassersein distance is defined in eq. (5) only with the linear term whereas the OT matrix $\mathbf{T}^*$ is optimized with an additional regularization term in eq. (6). This allows for a better approximation of the Wasserstein distance as discussed in Luise et al. (2018), but comes with a slightly more complex problem to minimize when used as objective value as discussed in the following.

4.2. Joint distribution optimal transport

In the context of unsupervised domain adaptation, Courty et al. (2017a) proposed the joint distribution optimal transport (JDOT) method. The idea is to consider the optimal transport problem between distributions on the algebraic product space of features and labels spaces, instead of only considering the feature space distributions.

In this setting, the source measure μ_s and the target measures μ_t are measures on the product space $\mathcal{X} \times \mathcal{Y}$, and we note (x^s, y^s), (x^t, y^t) the samples of μ_s and μ_t respectively. The generalized ground cost associated to this space can be naturally expressed as a weighted combination of costs in the input and label spaces, reading

$$d\left(\mathbf{x}_i^s, \mathbf{y}_i^s; \mathbf{x}_j^t, \mathbf{y}_j^t\right) = \alpha c(\mathbf{x}_i^s, \mathbf{x}_j^t) + \beta L(\mathbf{y}_i^s, \mathbf{y}_j^t) \tag{7}$$

for the i-th element of the support of μ_s and j-th element of the support of μ_t. $c(\cdot,\cdot)$ is chosen as a ℓ_2^2 distance and $L(\cdot,\cdot)$ is a classification loss (e.g. hinge or cross-entropy). Parameters α and β are two scalar values weighting the relative contributions of features and label discrepancies. In the unsupervised domain adaptation setting, the labels $\mathbf{y}_j^t$ are unknown and we seek to learn a classifier $f : \mathcal{X} \to \mathcal{Y}$ to estimate the label $f(\mathbf{x}_j^t)$ of each target sample. Hence, with $(x^t, f(\mathbf{x}_j^t))$ the samples from the target distribution, we define the ground loss,

$$d_f\left(\mathbf{x}_i^s, \mathbf{y}_i^s; \mathbf{x}_j^t\right) = \alpha c(\mathbf{x}_i^s, \mathbf{x}_j^t) + \beta L(\mathbf{y}_i^s, f(\mathbf{x}_j^t)) \tag{8}$$

Accounting for the classification loss, JDOT leads to the following minimization problem:

$$\min_f W_{\mathbf{D}_f}(\mu_s,\mu_t) = \min_{f,\mathbf{T}\in\Pi(\mu_s,\mu_t)} < \mathbf{T}, \mathbf{D}_f >_F, \tag{9}$$

where $\mathbf{D}_f$ depends on f and gathers all the pairwise costs $d_f(\cdot,\cdot)$. As a by-product of this optimization problem, samples that share a similar representation and a common label (through classification) are matched, yielding better discrimination. JDOT has been recently extended to deep learning strategies (Damodaran et al., 2018) by computing the optimal transport w.r.t. deep embeddings of the data rather than the original feature space, and also by proposing a large-scale variant of the regularized OT optimization problem.

4.3. Learning with noisy labels using entropy-regularized OT

The main idea of our proposed method is to learn a neural network model f efficiently in the presence of noisy labels. The intuition of proposed method is to leverage the information from the nearby samples to estimate accurate labels. Let $\{\mathbf{x}_i, \tilde{\mathbf{y}}_i\}$, $i,\ j = 1,\dots,N$ be the samples and their associated noisy one-hot labels observed from $p(x,\tilde{y})$. We note $\tilde{\mu}$ the discrete distribution corresponding to these samples. Our proposal is to learn f that yields a discrete distribution $\mu_f = \sum_i \delta_{\mathbf{x}_i, f(\mathbf{x}_i)}$ which minimizes the following problem:

$$\min_f \quad W_{\mathbf{D}_{f,\lambda}}(\tilde{\mu}, \mu_f), \tag{10}$$

which can be reformulated to the following bi-level optimization problem:

$$\min_f \quad \sum_{i,j} \mathbf{T}_{ij}^* L\left(\tilde{\mathbf{y}}_i, f(\mathbf{x}_j)\right) \tag{11}$$

$$\text{s.t. } \mathbf{T}^* = \arg\min_{\mathbf{T}} < \mathbf{T}, \mathbf{D}_f >_F + \lambda R(\mathbf{T}), \tag{12}$$

with

$$D_f^{i,j} = \alpha \|\mathbf{x}_i - \mathbf{x}_j\|^2 + \beta L\left(\tilde{\mathbf{y}}_i, f(\mathbf{x}_j)\right) \tag{13}$$

As we can see from the objective function eq. (11), f will be learned such that each sample classification $f(\mathbf{x}_j)$ needs to be close to every labels $\tilde{\mathbf{y}}_i$ for which T_{ij} is non-zero. This highlights the role of the optimal coupling $\mathbf{T}$ in helping learn a classifier f which is smoother thanks to this averaging process since the OT regularization λ will promote a spread of mass in $\mathbf{T}$. Here the geometry of the dataset is taken into account through the ground metric on the joint feature-label space. This averaging process is even more clear when the classification loss L is linear, which is the case for the cross-entropy loss. Indeed, we have in that case

$$\min_f \quad \sum_{i,j} \mathbf{T}_{ij}^* L\left(\tilde{\mathbf{y}}_i, f(x_j)\right) = \sum_j L\left(\sum_i \mathbf{T}_{ij}^* \tilde{\mathbf{y}}_i, f(\mathbf{x}_j)\right) \tag{14}$$

where $\sum_i \mathbf{T}_{ij}^* \tilde{y}_i$ is as an average of the labels with weights in $\mathbf{T}$, hence a denoised estimate for $\mathbf{y}_j$. Our approach corresponds to learning from labels that have been smoothed

by substituting the each noisy label by a weighted combination of labels where the weights are provided by the optimal couplings estimated w.r.t. the ground cost matrix $\mathbf{D}_f$. We name this approach CLEOT for Classification Loss with Entropic Optimal Transport. This approach is notably motivated by the denoising capacity of the entropy regularized optimal problem, as explored in (Rigollet and Weed, 2018), and where the denoising is conducted directly on the joint distributions.

In order to better interpret our approach, we can look at the limit cases. When $\beta \to 0$ and $\lambda = 0$ the label loss disappears in the OT metric and the OT matrix is the solution between μ and itself. In this case the solution is obviously the identity matrix and the optimization problem wrt f boils down to the classical empirical risk minimization of eq. (1) without cross-terms. When $\lambda = 0$ and $\alpha > 0, \beta > 0$, the OT is performed in the joint distribution sense and will include label information through cost (13), but the solution will be very sparse (a permutation) at the risk of overfitting the sample assignment in $\mathbf{T}^*$. However, when the entropy regularization is included ($\lambda^* > 0$), the probability in $\mathbf{T}^*$ mass is spread-out, as a result the optimal coupling ($\mathbf{T}^*$) will share mass between the samples which have similar features and label representations and perform label smoothing.

This averaging is of particular interest when the labels are corrupted by the noise since we always suppose that the good labels are wining locally in average (or else nothing can be learned anyways). Thus, learning the neural network model (f) with CLEOT (eq. (11)) naturally mitigate the impact of the label noise, and obtains the robust classifier.

Finally we discuss how to solve the optimization problem (10). The authors of JDOT originally proposed to perform alternative optimization on $\mathbf{T}$ and f. This approach works for un-regularized OT and converges to a stationary point. However this does not hold true for regularized OT. When using regularized OT as proposed here, the problem is a bi-level optimization problem (Colson et al., 2007). Bi-level optimization problem that are notoriously difficult to solve. Since the inner problem is a regularized OT problem that is strongly convex, one could solve the problem by using the implicit function theorem as discussed in Luise et al. (2018) on a different application. However, solving the full coupling $\mathbf{T}^*$ is computationally infeasible both in terms of time and memory, because $\mathbf{T}^*$ is a dense matrix and scales quadratically in size to the number of samples. Even if modern solvers have been proposed for regularized OT in the dual (Seguy et al., 2018; Arjovsky et al., 2017; Genevay et al., 2016) or primal (Genevay et al., 2017), they are still computationally intensive and cannot be used properly with alternate optimization. This problem is even aggravated by the necessity to solve the OT problem at each iteration. In order to circumvent these problems, we use a the stochastic optimization scheme by solving the problem on mini-batches, enabling to learn complex deep neural networks on large datasets.

Algorithm 1 Label-noise robust learning

Require: Training features (images) $\mathbf{x}$, noisy labels $\tilde{\mathbf{y}}$; hyperparameters α, β, λ

1: **for** each batch $((\mathbf{x}_1, \tilde{\mathbf{y}}_1), \ldots, (\mathbf{x}_B, \tilde{\mathbf{y}}_B))$ **do**
2: Compute the neural network predictions $f(\mathbf{x}_1), \ldots, f(\mathbf{x}_B)$
3: Compute the OT ground loss as in eq. (13)
4: Solve the OT problem ($\mathbf{T}^*$) in eq. (16) by Sinkhorn iterations
5: Update the neural network parameters by back-propagation
6: **end for**

4.4. Stochastic approximation of proposed method

We propose to approximate the objective function eq. (11) of our proposed method by sampling mini-batches of size m, and minimizing optimization problem:

$$\min_{f} \quad \mathbb{E}\left[\sum_{i,j} T^*_{ij} L(\tilde{\mathbf{y}}_i, f(\mathbf{x}_j)) \right], \tag{15}$$

$$\text{s.t. } \mathbf{T}^* = \arg\min_{\mathbf{T}} < \mathbf{T}, \mathbf{D}_f >_F + \lambda R(\mathbf{T}). \tag{16}$$

where the expectation $\mathbb{E}$ is taken over the randomly sampled mini-batches and eq. (16) is solved only on the mini-batches. As the size m increases, the optimization problem will converge to eq. (11). Still as discussed in Genevay et al. (2017), the expected value over the mini-batches if OT is not equivalent to the full OT and may lead to a different minimum. In practice it has the effect of densifying the equivalent full OT matrix and adding an additional regularization.

In order to optimize the problem above on mini-batches we use the *sinkhorn-autodiff* introduced in (Genevay et al., 2017) that relies on automatic differentiation of the Sinkhorn algorithm that quickly estimates the solution of entropic regularized OT and its gradients (see the psuedo code in algorithm 1). Note that we could have used the approach of Luise et al. (2018) for computing the gradients instead of *sinkhorn-autodiff* but their approach rely on the implicit function theorem which supposes that the inner problem is solved exactly. Since it is difficult to ensure exact convergence of the Sinkhorn, we prefer to perform autodiff on the algorithm with a finite number of iterations which will provide a reasonable gradients even when Sinkhorn has not converged. This stochastic approach has two major advantages: it scales to large datasets, and can be easily integrated into the modern deep learning framework in an end-to-end fashion.

4.5. Illustration on a toy example

For the sake of clarity, we propose an illustration (Figure 2) of the behavior of the method on a simple toy example. It consists in the classical two moons problem, which is a binary classification problem. From a clean version of the dataset (Figure 2.a), containing 400 data samples, labels are randomly flipped with a probability $p = 0.2$

(Figure 2.b). The classifier is a fully connected neural networks that consists in two hidden layers of size 256, with Relu activations. The model is adjusted along 500 epochs. A graphical representation of the decision boundary is given in (Figure 2.c), where one can clearly see that the model is not capable of separating properly the two classes, resulting in a complex boundary that encloses mislabeled samples. Then, three iterations of the proposed approach are represented (one per line). Column (d) shows the coupling $\mathbf{T}^*$ as a graph, i.e. links between samples corresponds to entries of $\mathbf{T}^*$ that are above a given threshold (as $\mathbf{T}^*$ is dense). The width of the connection is proportional to the magnitude of the entry. As expected, most of the connections highlight a geometrical and class label proximity. Labels are then propagated (Column (e)) following eq. (14). The classifier is fine tuned over this new set of fuzzy labels. Column (f) shows the new decision boundary, as well as the corresponding accuracy score (in red). As performances increase, it is worth noting the relative lower complexity of the classifier, that almost correctly classifies clean samples (0.99 of accuracy) after three iterations of CLEOT.

5. Experiments and Results

We evaluate our proposed `CLEOT` method and state-of-the-art (SoA) methods on two remote sensing tasks: Remote sensing (aerial image) scene classification, and pixel-wise labeling of hyperspectral image. The effectiveness of our proposed method is compared with several SoAs which modifies the loss function similar to ours. The considered SoA methods are `Backward` and `Forward` loss correction (Patrini et al., 2017), `Unhinged` (van Rooyen et al., 2015), `Sigmoid` and `Ramp` (Ghosh et al., 2015), `Savage` (Masnadi-Shirazi, Hamed and Vasconcelos, 2008), and `Bootstrap soft` (Reed et al., 2015). The `Unhinged`, `Sigmoid`, `Ramp`, `Savage` loss correction methods did not perform well in their original form. Preliminary experiments show that these methods either do not converge or converge to poor solutions, and sometimes result in premature saturation. In order to make these methods comparable, we stacked the batch-normalization and softmax pooling right before the loss function. This procedure increased the performance of the state-of-the-art methods, compared to the implementation mentioned in (Patrini et al., 2017) and in their respective articles. Thus the performance of the SoA methods can be considered as the strong baseline for our proposed method. A similar procedure is also used for the rest of the methods (including ours) to have uniformity.

The source code of our proposed method and SoA methods will be published here once upon acceptance.

In the next subsections, for each dataset, we first present the data, then detail the label noise simulations and implementation details, and finally present and discuss the results.

5.1. Aerial Image Labeling

We have considered four diverse publicly available remote sensing aerial scene classification datasets: *NWPU-RESIS45* (Cheng et al., 2017), *NWPU-19* (Cheng et al., 2017), *PatternNet* (Zhou et al., 2018), *AID* (Xia et al., 2017). The description of each dataset is provided below followed by a description of the label noise applied to them.

5.1.1. Datasets

NWPU-RESIS45. This dataset consists of 31'500 remote sensing images covering 45 scene classes. Each class contains 700 images with a size of 256×256 in the red green blue (RGB) color space. The spatial resolution of this dataset varies from about 30 m to 0.2 m per pixel. This dataset was extracted by the experts in the field of remote sensing image interpretation, from Google Earth (Google Inc.). Additional details of this dataset can be found in (Cheng et al., 2017).

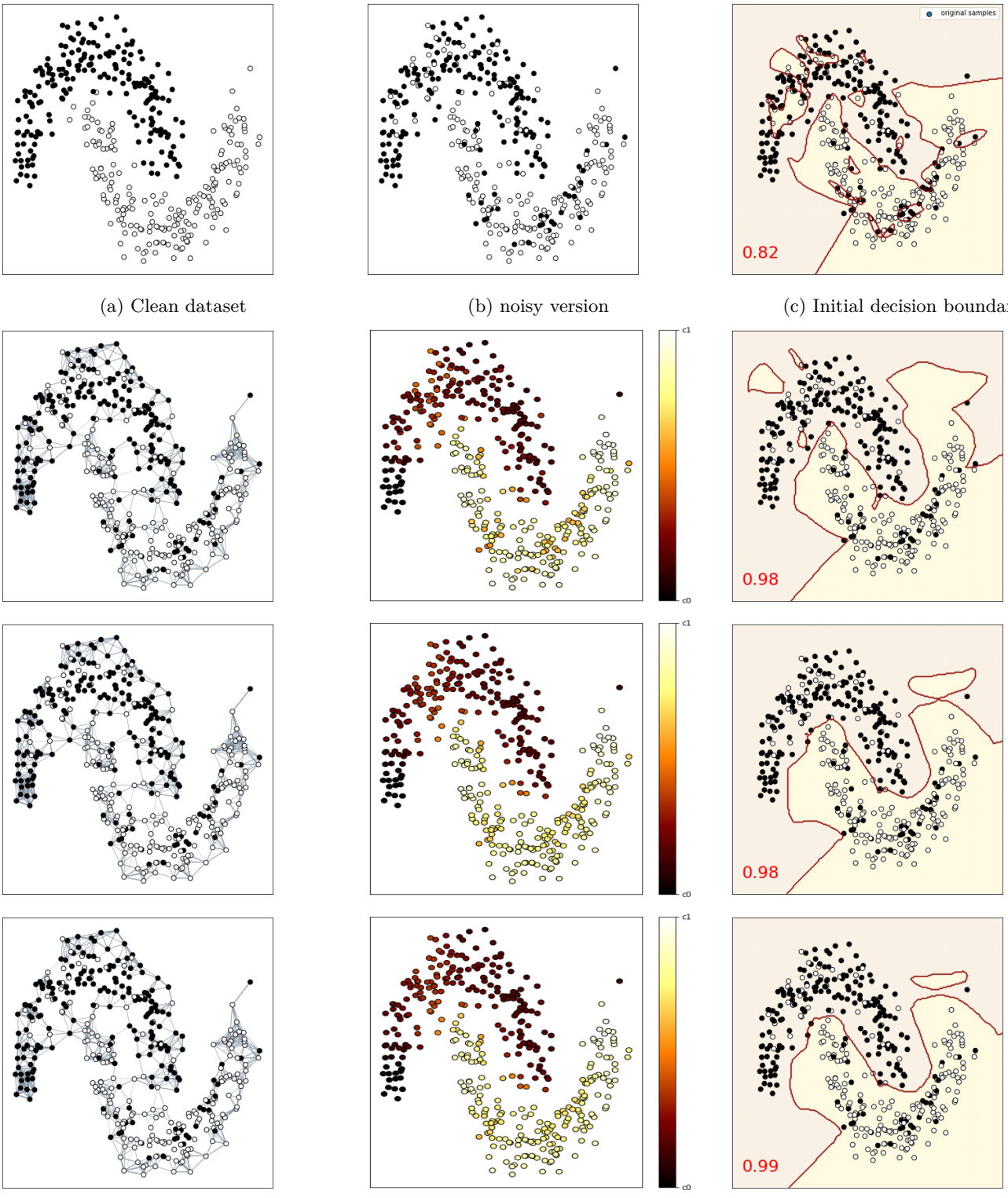
NWPU-19. This dataset is a subset of NWPU-RESIS45 dateset, which consists of 13'300 remote sensing images divided into 19 scene classes. The number of samples per class, and its size and spatial configuration are similar to NWPU-RESIS45.

PatternNet. This is a large-scale high resolution remote sensing dataset collected for remote sensing image retrieval. Here, we have used it for classification task. It contains 38 classes, and each class consists of 800 images of size 256×256 pixels, totals to 30'400 image scenes. The images in PatternNet are collected from Google Earth imagery or via the Google Map API for US cities. The images are of higher spatial resolution than the NWPU-RESIS45 dataset, the highest spatial resolution is around 0.062 m and lowest is around 4.69 m. For further information, please see (Zhou et al., 2018).

AID. This dataset is made up of 10'000 images covering 30 scene classes. Unlike the above datasets, the number of images in this dataset varies a lot with different aerial scene types, from 220 to 420 sample images. The spatial resolution is varied from 0.5 m to 8 m, and the size of each aerial image is fixed to 600×600 pixels. Similar to above datasets, this dataset is also collected from Google Earth at different time and seasons, over different countries and regions around the world. For more details, please see (Xia et al., 2017).

5.1.2. Label noise simulation

In order to evaluate our proposed method, we artificially simulate the (asymmetric) label noise in the above datasets, to meet the requirements of real world scenarios. We carefully inspected the samples, and flipped labels according to the noise probability to the visually similar classes. The class permutations that were selected are reported in Table 1.



(a) Clean dataset (b) noisy version (c) Initial decision boundary

(d) Self-coupling (e) Label propagation (f) Decision boundary

Fig. 2: Illustration of CLEOT on a toy example along three iterations of the method. First row depicts (a) clean dataset, (b) noisy datasets. Labels are permuted with a probability $p = 0.2$ (c) Initial classification by the neural network. Then column (d) shows the self coupling T as a graph (e) label propagation step where each label value is computed as eq. (14) (f) new classification boundaries after learning over the propagated labels. On these last images, the classification accuracy is written on the bottom left corner in red.

Table 1: Table of the selected label noise on the Aerial Image Labeling datasets. The classes which are flipped according to the label noise is described below. $a \rightarrow b$ indicates the label a is flipped into class label b, and $a \leftrightarrow b$ indicates that the labels are flipped in both direction.

Dataset	Label noise
NWPU-RESIS45 14/45 classes impacted	baseball diamond → medium residential, beach → river, dense residential ↔ medium residential, intersection → freeway, mobile home park ↔ dense residential, overpass ↔ intersection, tennis court → medium residential, runway → freeway, thermal power station → cloud, wetland → lake, rectangular farm land → meadow, church → palace, commercial area → dense residential
NWPU-19 7/19 classes impacted	baseball diamond → medium residential, beach → river, dense residential → medium residential, intersection → freeway, mobile home park → dense residential, overpass → intersection, tennis court → medium residential.
PatternNet 11/38 classes impacted	cemetery → christmas tree farm, harbor ↔ ferry terminal, dense residential → coastal home, overpass ↔ intersection, parking space → parking lot, runway mark → parking space, coastal home ↔ sparse residential, swimming pool → coastal home
AID 12/30 classes impacted	bareland → desert, centre → storage tank, church → centre, storage tank; dense residential → medium residential, desert → bareland, industrial → medium residential, meadow → farm land, medium residential → dense residential, play ground → meadow, school; resort → medium residential, school → medium residential, play ground; stadium → play ground

5.1.3. Model

We employed pre-trained VGG16 architecture, replacing the last layer with two MLPs that map to 512 hidden neurons before predicting the classes with $l_2 = 1e^{-3}$ regularization, respectively. The dropout layer with $p = 0.5$ is inserted before the last MLP and the batch normalization is applied before the softmax operator. To have uniformity, all the methods follow similar architecture design. During the training, the network is fine-tuned by freezing the weights of the VGG16 layers. We optimized the SoA methods for 300 epochs using the SGD optimizer ($lr = 0.01$) with momentum ($m = 0.9$) using the mini-batch size of 128. The proposed `CLEOT` method is also optimized as above, but with different learning rate ($lr = 0.1$) and mini-batch size (50 samples per class). The hyper-parameters of `CLEOT` method are set as $\alpha = 1$, $\beta = 0.005$, and $\lambda = 0.005$ experimentally for all the datasets. Additionally, we have used early stopping criterion to terminate the training process, if the validation loss did not decrease for 25 epochs. This allows to prevent over-fitting to the noisy labels for all the methods. Furthermore, we retained the model weights with best validation loss. For our proposed method, $\mathbf{x}$ in (11), and (13) refers to the VGG16 features extracted before the two MLPs.

For all the datasets, from the available number of samples we partitioned 80% of samples for training, 10% samples for validation and the remaining 10% samples for evaluating the performance. All the methods are trained with the noisy labeled training and validation samples, and evaluated with the clean testing label instances.

5.1.4. Results

Table 2 presents the classification performance of our proposed and SoA methods on the four aerial scene classification datasets with different noise levels. We have also included `Cross entropy` loss function, which is the baseline for all the approaches. The noise level 0 indicates that the methods are trained with the clean labeled training and validation samples, and it can be considered as the gold standard. The impact of label noise is varied and analyzed in the range of $p_e = \{0, 0.2, 0.4, 0.6, 0.8\}$. The amount of actual noise depends on the number of the classes affected by the label noise in the dataset.

When the conventional `Cross entropy` loss function is considered, the classification accuracy drops to few percentage of points (3-5%) initially and decreases drastically (above 15%) as the magnitude of label noise increases with all the four datasets. This shows that regularization techniques such as weight regualizers, dropout and early stopping criteria can circumvents label noise only up to some degree and is inefficient in high level label noise. This emphasizes need for the inclusion of robust loss functions while training the deep neural networks in remote sensing image analysis. Next, when the performance of existing SoA methods are considered, they showed robust performance and did not degrade the performance compared to the clean training set under the noise level 0.2 but decreases about 4% on the mid noise level, however they outperformed the conventional `Cross entropy` loss function. Further, it is noted that under the high level label noise, the SoA methods are similar or less than the `Cross entropy` loss function. Thus, SoA methods are still limited to tackle the complex noise scenarios.

Lastly, when the performance of the proposed `CLEOT` method is analyzed, one could see that `CLEOT` achieves better or similar performance to the SoA methods in the low-level noise, and achieves impressive performance on the higher noise levels. For instance, on average our method decreases only 2.6%, 10.6% compared to clean training set with 0.4, and 0.6 noise level, where as the best SoA decreases about 4%, and 22.5% respectively. Further, `Forward` and `Backward` methods are inferior to the the robust loss functions (`Unhinged, Sigmoid`, etc), which is contrary w.r.t to the observation in Patrini et al. (2017). This reveals that methods that perform well on machine learning datasets not necessarily achieve better performance in remote sensing datasets, thus new methods has to be designed specific to remote sensing datasets. Lastly,

Table 2: The average classification accuracies and standard deviation of SoA methods and proposed `CLEOT` method on remote sensing aerial scene classification datasets. The accuracy measures are averaged over 5 runs. and the best accuracies are reported in **bold**.

Method	NWPU-45					NWPU-19				
	0.0	0.2	0.4	0.6	0.8	0.0	0.2	0.4	0.6	0.8
Cross entropy	82.93±0.09	80.53±0.30	75.63±1.04	67.80±0.40	62.05±0.04	89.75±0.30	84.95±0.49	78.19±2.08	68.45±0.95	62.25±0.17
Unhinged	82.81±0.21	82.13±0.14	78.38±0.57	63.07±0.26	61.01±0.01	90.07±0.20	87.36±0.46	79.22±0.77	66.65±0.85	61.22±0.41
Sigmoid	71.74±0.40	68.08±0.18	65.76±0.50	57.10±0.06	56.61±0.31	89.69±0.06	88.73±0.19	84.37±0.09	66.19±0.29	59.62±0.11
Ramp	82.99±0.10	**82.26±0.20**	78.81±0.26	62.97±0.16	60.91±0.32	**90.77±0.28**	86.69±0.18	78.62±0.77	67.25±0.60	60.70±0.44
Savage	76.85±0.15	75.13±0.11	69.96±0.14	59.56±0.03	58.08±0.07	90.20±0.18	89.01±2.80	81.13±0.41	67.21±0.20	60.62±0.12
Bootstrap soft	82.98±0.17	80.65±0.47	75.82±0.88	67.39±0.86	62.22±0.21	89.74±0.18	85.19±0.56	79.64±0.95	69.20±1.50	62.03±0.05
Backward $\hat{E}$	82.79±0.14	80.65±0.51	75.96±0.72	68.67±0.75	62.45±0.52	89.73±0.43	85.20±0.37	78.17±1.01	68.72±1.60	62.06±0.08
Forward $\hat{E}$	**83.06±0.11**	80.87±0.53	74.97±1.02	68.12±1.16	62.56±0.16	89.97±0.32	85.37±1.04	78.89±1.28	69.07±0.95	62.31±0.24
CLEOT	82.41±0.27	81.54±0.18	**80.84±0.45**	**76.07±0.35**	**70.14±0.33**	89.98±0.38	**89.17±0.30**	**86.26±0.63**	**78.04±0.42**	**68.08 ± 0.53**
	PatternNet					AID				
	0.0	0.2	0.4	0.6	0.8	0.0	0.2	0.4	0.6	0.8
Cross entropy	97.68±0.15	94.82±0.30	89.11±0.55	79.76±1.36	73.68±0.34	86.94±0.51	82.92±0.33	73.80±1.03	65.56±1.18	57.80±0.84
Unhinged	**97.76±0.16**	97.55±0.05	95.31±0.19	73.62±0.59	71.23±0.14	87.64±0.19	**86.33±0.19**	78.67±0.29	65.93±0.52	57.20±0.16
Sigmoid	96.63±0.03	96.31±0.31	94.63±0.24	73.17±0.39	70.58±0.04	85.41±0.26	84.71±0.25	82.05 ±0.17	60.96±0.44	56.18±0.08
Ramp	97.73±0.07	**97.56±0.02**	**95.44±0.16**	72.94±0.47	71.37±0.16	**87.74±0.22**	86.24±0.23	78.37±0.56	66.04±0.59	57.21±0.27
Savage	96.82±0.05	96.41±0.03	93.94±0.11	73.16±0.13	70.72±0.01	83.65±0.10	85.73±0.21	**82.28±0.27**	62.88±0.50	56.55±0.48
Bootstrap soft	97.62±0.13	94.45±0.39	88.88±0.79	79.13±0.73	73.39±0.48	87.03±0.40	82.54±0.78	73.75±0.82	65.24±1.09	58.00±0.45
Backward $\hat{E}$	97.60±0.10	94.76±0.31	89.07±0.70	79.89±0.36	73.47±0.32	86.87±0.52	82.63±0.59	74.03±0.56	65.71±1.16	57.90±0.22
Forward $\hat{E}$	97.67±0.06	94.43±0.78	89.16±1.01	79.44±0.51	73.21±0.62	86.91±0.41	82.30±1.08	73.59±0.76	64.91±0.74	58.43±0.59
CLEOT	97.29±0.04	96.77±0.09	94.51±0.15	**83.75±0.20**	**79.84 ± 0.22**	87.02±0.63	85.39±1.12	79.19±0.94	**71.76±0.66**	**63.23±0.42**

it is noted from table 2 that, among the SoA methods there is no single best method which consistently performs better across the datasets, and noise levels. Thus, there exists dilemma in choice of method among the existing SoA for the underlying real world task. On contrary, our method consistently outperforms across different datasets, and noise levels.

5.2. Hyperspectral image classification

Next, we evaluate our proposed method on the pixel-wise labeling task of hyperspectral datasets. For this, we have chosen three hyperspectral datasets from three different type of sensors covering agricultural and urban cover settings.

5.2.1. Datasets

Pavia University. The first hyperspectral data considered here was collected over the University of Pavia, Italy by the ROSIS airborne hyperspectral sensor in the framework of the HySens project managed by DLR (German national aerospace agency). The ROSIS sensor collects images in 115 spectral bands in the spectral range from 0.43 to 0.86 µm with a spatial resolution of 1.3 m/pixel. After the removal of noisy bands, 103 bands were selected for experiments. This data contains 610×340 pixels with nine classes of interest, which covers the urban materials. The total number of available labeled ground truth samples is 42'776.

Chikusei. The airborne hyperspectral dataset was taken by Headwall Hyperspec-VNIR-C imaging sensor over agricultural and urban areas in Chikusei, Ibaraki, Japan. The hyperspectral dataset has 128 bands in the spectral range from 363 nm to 1018 nm. The scene consists of 2517×2335 pixels and the ground sampling distance was 2.5 m. Ground truth of 19 classes was collected via a field survey and visual inspection using high-resolution color images obtained by Canon EOS 5D Mark II together with the hyperspectral data. The number of labeled reference samples is 77'592. For additional details, please refer (Yokoya and Iwasaki, 2016).

GRSS_DFC_2018. The last hyperspectral dataset used were acquired over the University of Houston campus and its neighborhood on February 2017 by an ITRES CASI 1500 imaging sensor. This dataset contains 48 spectral bands covering the spectral range of 380 nm to 1050 nm with 1 m ground sampling distance. The scene consists of 601×2384 pixels representing 20 urban land use/cover classes, and contains 50'4856 labeled reference samples. The details of this dataset can be found in [1].

5.2.2. Label noise simulation

For the hyperspectral datasets, it is difficult to find similar classes with visual inspection due to the high dimensionality of the data. So we measure class similarity using the Jeffries-Matusita distance and Transformed divergence measure (Richards et al., 1999) and visual interpret the spectral signatures of some training samples for the most similar classes. According, the class labels are flipped as defined in Table 3.

5.2.3. Model

We used a recent state-of-the-art hyperspectral image classification framework named spectral-spatial residual residual network (SSRN) (Zhong et al., 2018). It consecutively extracts spectral and spatial features for pixel wise classification of hyperspectral image. The spectral feature learning consists of two 3-D convolutional layer, and two residual blocks. Following the spectral features, spatial

[1] http://www.grss-ieee.org/community/technical-committees/data-fusion

Table 3: Table of the selected label noise on the Hyperspectral image classification datasets. The classes which are flipped according to the label noise is described below. $a \rightarrow b$ indicates the label a is flipped into class label b, and $a \leftrightarrow b$ indicates that the labels are flipped in both direction.

Dataset	Label noise
Pavia University 7/9 classes impacted	meadows ↔ trees, gravel ↔ self building blocks, bare soil → meadows, bitumen ↔ asphalt. Out of 9 classes, 7 classes are impacted by the label noise.
Chikusei 10/19 classes impacted	baresoil (park) → baresoil (farm); baresoil (farm) → baresoil (park), rowcrops; weeds → grass, rowcrops, forest; forest → rice(grown), weeds; grass → weeds, rowcrops; rice(grown) → forest, weeds; rowcrops → baresoil(farm), weeds, grass; plastic home → asphalt, manmade(dark); manmade(dark) → plastic house;paved ground → baresoil(farm)
GRSS_DFC_2018 10/20 classes impacted	healthy grass → stressed grass; stressed grass → bare earth; evergreen trees → deciduous trees; deciduous trees → residential buildings; residential buildings → roads, sidewalks; non-residential buildings → sidewalks; roads → major thoroughfares, sidewalks; sidewalks → major thoroughfares, crosswalks; crosswalks → major thoroughfares; major thoroughfares → highways

features are extracted using 3-D convolutional layer and two spatial residual blocks. Average pooling layer is added on top of the spectral-spatial feature volume, and followed by the fully connected layer with softmax activation function. Dropout layer (p=0.5) is added after the average pooling layer and batchnormalization layer is stacked before the softmax activation. Please refer to (Zhong et al., 2018) for additional details of the SSRN architecture.

We trained the SSRN architecture using SGD momentum optimizer with lr = 0.01 and m = 0.9 for 600 epochs using the batch size of 128 for the SoA methods, and 256 for the `CLEOT`. As with aerial scene classification we also employed the eary stopping criterion to avoid overfitting, and terminate the training process, if the validation loss did not decrease for 15 epochs. All the methods follow similar training procedure. The hyperparameters of our proposed method `CLEOT` are set to $\alpha = 1, \beta = 0.05$ experimentally for all the datasets, and the entropic regularizer is set to $\lambda = 0.05$ for the PaviaU and Chikusei datasets, and $\lambda = 1$ for the remaining dataset. In the case of pixel-wise labeling task of hyperspectral image, $\mathbf{x}$ in (11) and (13) refers to the input image pixel.

While partitioning the ground truth reference samples into training, validation and testing subsets, we followed the conventional protocol in the hyperspectral remote sensing community to train classifier with small number of training samples. Accordingly, we used 20% of samples for training, 10% samples for validation and remaining 70% samples for testing purpose for the Pavia University, and Chikusei datasets. Where as for the GRSS_DFC_2018 dataset, we used 10% samples for training, 10% samples for validation and remaining 80% samples for evaluation. The training and validation samples are impacted by the label noise, and clean testing labeled samples is used for evaluation.

5.2.4. Results

Table 4 presents the classification performance of SoA methods and `CLEOT` for the three hyperspectral datasets. The experiments are conducted with different noise levels (see Table 4 for noise levels) to effectively access the robustness of SoA and proposed methods. The noise level $p_e = 0$ indicates the clean labeled training and validation samples, which is an upper bound for all the methods.

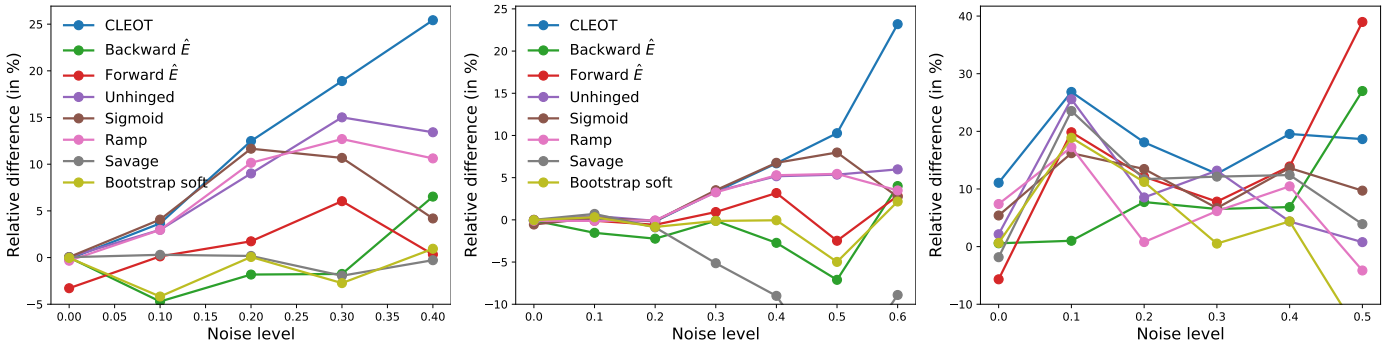
Figure 3 shows the relative difference (in %) of methods with respect to the baseline (`Cross entropy`) method. The proposed `CLEOT` consistently outperformed the existing SoA methods with large margin (about 15-20%) in the higher level label noise (except in GRSS_DFC_2018). However our method still has the large performance margin with other methods. On the lower level label noise, fig. 3 reveals there is no significant difference between the best SoA and `CLEOT` method. However our method has several distinct advantages over the SoA, `CLEOT` (i) converges faster than the best SoA method, which is beneficial for very large scale remote sensing datasets, (ii) consistently performs better irrespective of noise level, whereas the best SoA varies with respect to the noise level, for instance with Pavia University, `Sigmoid` outperforms in mid level noise, but in higher level noise `Unhinged` outperformed the `Sigmoid` loss function, (iii) monotonically degrades the classification accuracy as the noise level increases with complex dataset, whereas SoAs do not follow this trend, thus existing methods are not as reliable. This also reveals that the best SoAs might be more sensitive to the neural network initialization under label noise. Thus, our proposed method can be considered as a alternative candidate to train robust deep neural networks for remote sensing image analysis.

As observed with aerial scene classification, the the classification accuracy of loss `Cross entropy` decreases as the noise level increases. The magnitude of decrease in accuracy is dependent on the amount classes affected by the label noise, and also the nature of the datasets. It is noted that on the *Chikusei* dataset, `Cross entropy` is very robust compared to other datasets, this might be due to the large patches of homogeneous landscapes in the dataset as well as the appearance of label noise at pixel level. In future work, we will consider more complicated noise model for the hyperspectral datasets, where the label noise could appears as spatially correlated clusters of pixels.

Table 4: The average classification accuracies and standard deviation of SoA methods and our proposed `CLEOT` method on the pixel-wise labeling task of hyperspectral datasets. The accuracy measures are computed over three runs, and the best accuracies are reported in **bold**.

Method	PaviaU					GRSS_DFC_2018					
	0.0 (0.0)	0.1 (0.19)	0.2 (0.37)	0.3 (0.57)	0.4	0.0	0.1	0.2	0.3	0.4	0.5
Cross entropy	99.93±0.02	95.62±0.38	85.31±0.60	78.01±1.78	65.13±1.49	84.93±2.27	68.66±16.13	76.71±6.37	79.54±16.57	71.54±14.54	44.33±11.52
Unhinged	99.97±0.02	98.58±0.01	94.91±0.01	93.02±0.02	78.55±0.01	87.07±1.48	94.23±1.03	85.24±3.85	92.76±0.65	75.90±2.21	45.11±4.45
Sigmoid	**99.98±0.00**	**99.68±0.05**	97.56±0.01	88.69±0.01	69.31±0.02	90.33±6.72	84.87±7.01	90.18±4.09	86.23±3.79	85.18±5.31	54.04±6.11
Ramp	99.59±0.49	98.58±0.17	96.04±0.65	90.70±2.02	75.75±4.65	92.32±3.11	85.91±8.29	77.50±3.32	85.70±3.69	82.01±8.13	40.19±8.92
Savage	99.97±0.01	95.91±0.86	86.07±0.38	76.06±0.08	64.84±1.33	83.09±10.68	92.20±3.27	88.42±4.15	91.69±1.43	83.96±3.50	48.23±15.16
Bootstrap soft	99.89±0.01	91.44±3.28	85.96±2.34	75.27±1.18	66.07±2.25	85.57±13.96	87.55±5.59	87.93±5.96	80.07±3.81	75.92±12.42	28.37±7.38
Backward $\hat{E}$	99.94±0.02	90.91±3.52	84.08±0.64	76.26±3.39	71.65±6.74	85.51±7.97	69.67±25.91	84.44±5.94	86.08±4.64	78.40±11.54	71.13±12.40
Forward $\hat{E}$	96.65±4.65	95.74±0.26	87.64±1.05	84.05±3.24	65.46±5.09	79.26±9.11	88.50±0.50	88.84±3.04	87.35±6.44	85.44±5.28	**83.30±4.46**
CLEOT	99.80±0.12	99.28±0.24	**98.42±0.21**	**96.91±0.50**	**91.31±0.23**	**96.01±0.59**	**95.50±0.56**	**94.81±0.50**	**92.18±1.51**	**91.08±0.15**	62.98±1.65

Method	chikusei						
	0.0	0.1	0.2	0.3	0.4	0.5	0.6
Cross entropy	**99.99±0.01**	99.27±0.04	98.98±0.02	95.92±0.17	92.40±0.42	86.37±0.31	61.31±0.26
Unhinged	99.87±0.18	99.74±0.16	99.45±0.09	99.19±0.01	97.57±0.22	91.72±0.37	67.29±7.45
Sigmoid	99.44±0.01	99.51±0.01	99.42±0.01	99.23±0.02	99.06±0.02	94.35±0.06	64.02±0.03
Ramp	99.88±0.15	99.16±0.64	99.50±0.01	99.17±0.13	97.69±0.75	91.81±0.96	64.79±7.15
Savage	99.99±0.02	**99.98±0.00**	98.67±0.01	90.78±1.45	83.39±3.81	66.87±2.46	52.40±0.05
Bootstrap soft	99.95±0.04	99.59±0.16	98.73±0.31	95.79±1.44	92.35±0.48	81.39±7.73	63.48±4.44
Backward $\hat{E}$	99.92±0.10	97.74±1.43	97.34±1.24	95.81±0.90	89.68±5.60	79.26±4.32	65.31±9.60
Forward $\hat{E}$	99.99±0.01	99.17±0.01	98.98±0.04	96.84±0.05	95.58±0.06	83.89±0.06	64.15±0.04
CLEOT	99.88±0.01	99.41±0.13	**99.59±0.12**	**99.24±0.01**	**99.10±0.02**	**96.64±1.20**	**84.50±1.35**



(a) Pavia University (b) Chikusei (c) GRSS_DFC_2018

Fig. 3: Relative difference in performance of our proposed method and state-of-the-art methods with respect to `cross entropy` loss function for the three hyperspectral datasets: (a) Pavia University, (b) Chikusei, (c) GRSS_DFC_2018.

5.3. Analysis of the results

We now provide an analysis of our approach through interpretation of the confusion matrices and by looking at actual examples.

First, we interpret the impact of label noise through confusion matrix and our proposed method to the best performing SoA (`Savage`)[2], and baseline method (`CCE`) on NWPU19 dataset with clean and noisy labels (with noise ratio of 0.4). Though, the label noise is simulated in the paper, the existence of label noise is very realistic in remote sensing, because of the close (spectral) similarities between classes or presence of two objects in the same image. For example, there is a possibility for the class `tennis ground` to co-exist with `medium residential area`, and the classes (intersection, freeway, and overpass) are known to be very similar between each other. Even with clean labeled data, confusion matrix in figure 4 (a) reveals that traditional CNN confuses among those classes, this is due to the fact that class `tennis ground` also looks like `medium residential area` as shown in figure 1, and the human annotator can also make similar kind of mistake. Further, the confusion matrix visualized in Figure 4 shows that class `tennis ground` is the most difficult one to correctly classify. Our proposed `CLEOT` additionally misclassifies 13 samples compared to `CCE` on clean data. On the other hand `CCE`, and `savage` misclassifies majority of the noisy samples. Further, detailed analysis of the misclassified samples reveals that different methods has a complementary behavior to each other, suggesting that ensemble methods could tackle the label noise effectively.

To analysis our results qualitatively, we also visualized images with respect to the predictions of three methods (`CCE, Savage, CLEOT`) as shown in Figure 5. In the top line of Figure 5 we visualize images that are not correctly predicted by all methods including when trained on clean data, showing the difficulty of CNN on those classes. A close inspection reveals that a human can label the second

[2]We choose the method which performed the best across the noise level

image with a baseball diamond correctly, and could confuse it with the class in the two other images (residential, tennis court). While such behavior exists in clean label case, and if the dataset is affected by the label noise, the deep neural network might find more difficult to predict the classes correctly, yielding a strong need for accurate labeled annotations. In the middle row we provide three images that are not correctly predicted by all the three methods under label noise (noisy `CCE`, `Savage`, and `CLEOT`), but are correctly classified by the clean `CCE`. Finally, the bottom row presents images that are correctly classified by `CLEOT`, but failed with noisy `CCE`, and `Savage` loss functions. The visual inspection reveals all the three images are quite easy to classify and not confusing since the class of interest have large spatial extent in the image. However, the `Savage` and `CCE` failed to correctly predict the classes. We can see that `CLEOT` has a behavior similar to human inspection on this case and predicted the classes correctly. This illustrates that exploiting knowledge from the nearby samples in the feature space while training helps making neural networks less sensitive to label noise.

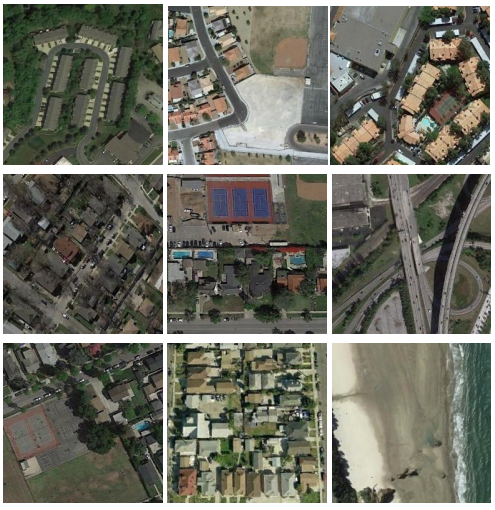


Fig. 5: Top row indicates the difficult samples (**true label: medium residential, baseball diamond, tennis court**) which are incorrectly predicted by clean `CCE` loss, and robust loss functions (**pred label: tennis court, tennis court, intersection**). The second row indicates the samples (**true label: dense residential, tennis court, overpass**) which are correctly predicted by the clean `CCE`, but failed with noisy `CCE`, `Savage`, `CLEOT` loss functions under the impact of label noise (**pred label: medium residential, medium residential, intersection**). The Last row indicates the samples (**true label: tennis court, dense residential, beach**) correctly predicted by our proposed method, failed by the noisy `CCE`, `Savage` loss functions (**pred label: medium residential, medium residential,river**).

5.4. Computational time analysis

Table 5 reports the computational time analysis of our proposed `CLEOT` method and the SoA methods with one dataset each from aerial image labeling task and pixel-wise labeling of hyperspectral image. As expected, computational time of our proposed method is higher than the existing methods. The high computational time is due to the number of sinkhorn iterations required to approximate the optimal transport distance. The computational time of `CLEOT` can be reduced, if the number of sinkhorn iterations are decreased. This might slightly decrease the accuracy of our proposed method. For example, when the sinkhorn iterations is decreased to 15, the computational time is reduced to 173 Secs and the classification accuracy is reduced by 2%, but still our method outperform the existing methods. The forward and backward method requires trained crossentropy method to estimate the class transition matrix, thus resulting in higher computational time. The computational burden of our proposed method is reasonable with the existing state of art methods.

Table 5: Computational run time (in Secs) of the `CLEOT` and existing methods computed with one aerial image labeling dataset and hyperspectral dataset over the single run. The computational time are measured for fixed number of epochs (100 epochs) for all the methods. The numbers in brackets for `CLEOT` indicates the number of fixed sinkhorn iterations.

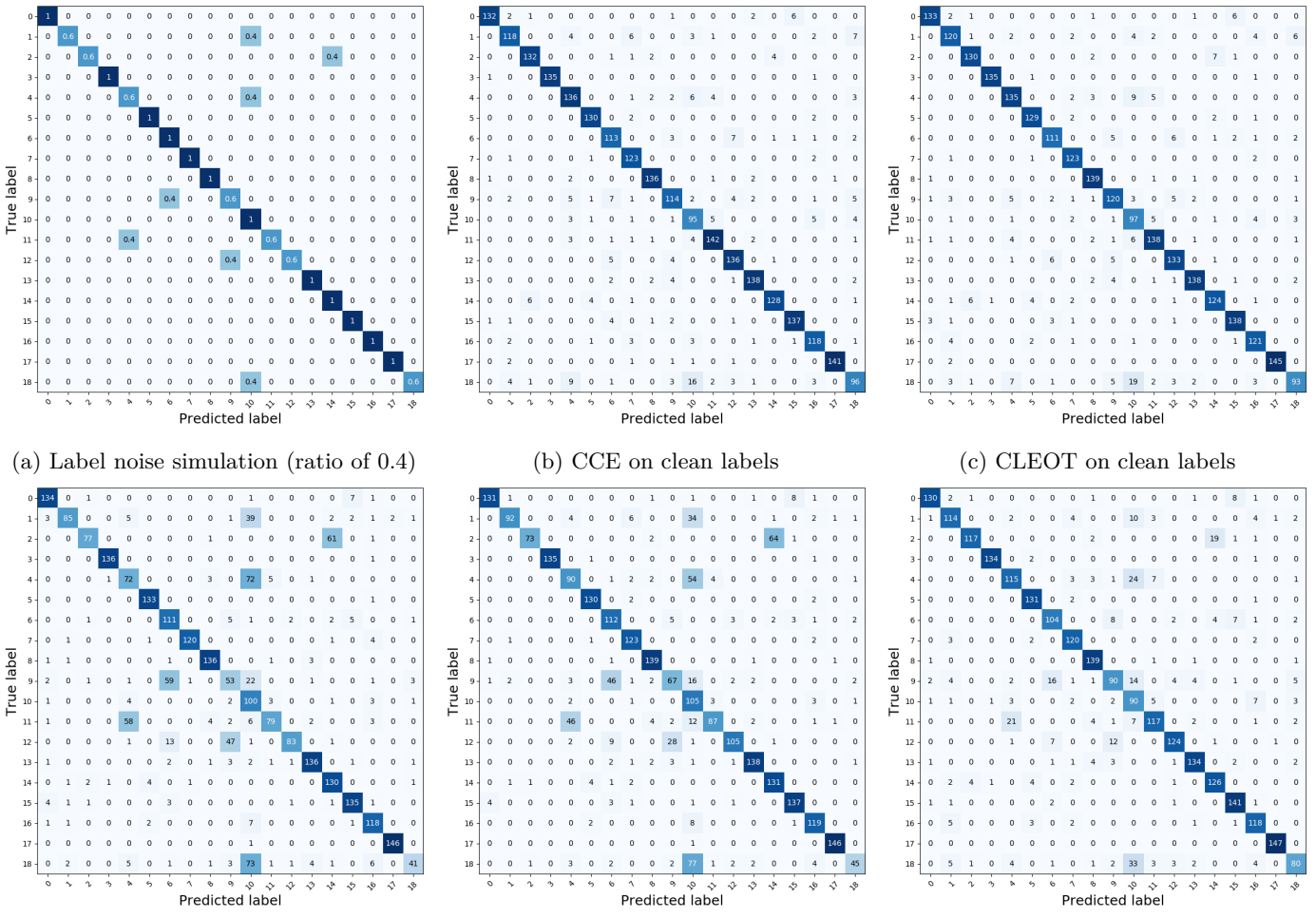
Method	NWPU-19 Comp.time (in Secs)	PaviaU Comp.time (in Secs)
Crossentropy	114	586
Unhinged	108	582
Sigmoid	108	585
Ramp	110	591
Savage	115	599
Bootstrap soft	108	586
Backward $\hat{E}$	221	1164
Forward $\hat{E}$	223	1165
CLEOT (50)	245	778
CLEOT (15)	173	674

6. Conclusion

In this paper, we proposed the `CLEOT` method to learn robust deep neural networks under label noise in remote sensing. The proposed method leverages on the geometric structure of underlying data, and uses optimal transport with entropic regularization to regularize the classification model. We evaluated the robustness of `CLEOT` on two very different applications, one focusing on image scene classification, the second one on pixel-wise classification of hyperspectral images with different deep learning architectures. Our proposed approach performed better than competing state of the art approaches and has shown strong robustness in the presence of significant amount of label noise. Future works will consider other regularization schemes of the optimal transport problem, and use an embedding metric in the definition of the cost matrix $\mathbf{D}_f$ instead of relying to the distance in the input space.

Acknowledgments

This work benefited from the support of Region Bretagne grant and OATMIL ANR-17-CE23-0012 project of



(a) Label noise simulation (ratio of 0.4)

(b) CCE on clean labels

(c) CLEOT on clean labels

(d) CCE on noisy labels

(e) Savage on noisy labels

(f) CLEOT on noisy labels

Fig. 4: Confusion matrix of different methods on NWPU19. First row indicates the confusion matrix with the clean labels and second row with noisy labels. The class names are in the following order (starting index from '0'): airplane, baseball diamond, beach, chaparral, dense residential, forest, freeway, golf course, harbor, intersection, medium residential, mobile homepark, overpass, parking lot, river, runway, sparse residential, storage tanks, tennis court.

the French National Research Agency (ANR), and from CNRS PEPS 3IA DESTOPT. The authors would like to thank Prof. Paolo Gamba for providing Pavia University dataset, and the National Center for Airborne Laser Mapping and the Hyperspectral Image Analysis Laboratory at the University of Houston for acquiring and providing the GRSS_DFC_2018 data, and the IEEE GRSS Image Analysis and Data Fusion Technical Committee. We also gratefully acknowledge the support of NVIDIA Corporation with the donation of the Titan Xp GPU used for this research.

References

Anwer, R.M., Khan, F.S., van de Weijer, J., Molinier, M., Laaksonen, J., 2018. Binary patterns encoded convolutional neural networks for texture recognition and remote sensing scene classification. ISPRS Journal of Photogrammetry and Remote Sensing 138, 74–85.

Aritra, G., Himanshu, K., PS, S., 2017. Robust Loss Functions under Label Noise for Deep Neural Networks, in: AAAI Conference on Artificial Intelligence.

Arjovsky, M., Chintala, S., Bottou, L., 2017. Wasserstein generative adversarial networks, in: ICML, pp. 214–223.

Audebert, N., Le Saux, B., Lefèvre, S., 2018. Beyond RGB: Very high resolution urban remote sensing with multimodal deep networks. ISPRS Journal of Photogrammetry and Remote Sensing 140, 20–32.

Brooks, J.P., 2011. Support vector machines with the ramp loss and the hard margin loss. Operations Research 59, 467–479.

Cheng, G., Han, J., Lu, X., 2017. Remote Sensing Image Scene Classification: Benchmark and State of the Art. Proceedings of the IEEE 105, 1865–1883.

Cheng, G., Yang, C., Yao, X., Guo, L., Han, J., 2018. When Deep Learning Meets Metric Learning: Remote Sensing Image Scene Classification via Learning Discriminative CNNs. IEEE Transactions on Geoscience and Remote Sensing 56, 2811–2821.

Colson, B., Marcotte, P., Savard, G., 2007. An overview of bilevel optimization. Annals of operations research 153, 235–256.

Courty, N., Flamary, R., Ducoffe, M., 2018. Learning wasserstein embeddings. ICLR, arXiv:1710.07457 .

Courty, N., Flamary, R., Habrard, A., Rakotomamonjy, A., 2017a. Joint distribution optimal transportation for domain adaptation, in: NIPS.

Courty, N., Flamary, R., Tuia, D., Rakotomamonjy, A., 2017b. Optimal transport for domain adaptation. IEEE TPAMI 39, 1853–1865.

Cuturi, M., 2013. Sinkhorn distances: Lightspeed computation of optimal transportation, in: NIPS, pp. 2292–2300.

Damodaran, B.B., Kellenberger, B., Flamary, R., Tuia, D., Courty, N., 2018. DeepJDOT: Deep Joint Distribution Optimal Transport for Unsupervised Domain Adaptation, in: The European Conference on Computer Vision (ECCV), pp. 447–463.

Ding, Y., Wang, L., Fan, D., Gong, B., 2018. A Semi-Supervised Two-Stage Approach to Learning from Noisy Labels. arXiv preprint arXiv:1802.02679 .

Frank, J., Rebbapragada, U., Bialas, J., Oommen, T., Havens, T., Frank, J., Rebbapragada, U., Bialas, J., Oommen, T., Havens, T.C., 2017. Effect of Label Noise on the Machine-Learned Classification of Earthquake Damage. Remote Sensing 9, 803.

Frenay, B., Verleysen, M., 2014. Classification in the Presence of Label Noise: A Survey. IEEE Transactions on Neural Networks and Learning Systems 25, 845–869.

Genevay, A., Cuturi, M., Peyré, G., Bach, F., 2016. Stochastic optimization for large-scale optimal transport, in: NIPS, pp. 3432–3440.

Genevay, A., Peyré, G., Cuturi, M., 2017. Sinkhorn-autodiff: Tractable wasserstein learning of generative models. arXiv preprint arXiv:1706.00292 .

Ghosh, A., Manwani, N., Sastry, P., 2015. Making risk minimization tolerant to label noise. Neurocomputing 160, 93–107.

Haklay, M., 2010. How Good is Volunteered Geographical Information? A Comparative Study of OpenStreetMap and Ordnance Survey Datasets. Environment and Planning B: Planning and Design 37, 682–703.

Hendrycks, D., Mazeika, M., Wilson, D., Gimpel, K., 2018. Using Trusted Data to Train Deep Networks on Labels Corrupted by Severe Noise, in: Advances in Neural Information Processing Systems.

Hickey, R.J., 1996. Noise modelling and evaluating learning from examples. Artificial Intelligence 82, 157–179.

Huang, B., Lu, K., Audebert, N., Khalel, A., Tarabalka, Y., Malof, J., Boulch, A., Saux, B.L., Collins, L., Bradbury, K., Lefèvre, S., El-Saban, M., 2018. Large-scale semantic classification: outcome of the first year of Inria aerial image labeling benchmark, in: IEEE International Geoscience and Remote Sensing Symposium – IGARSS 2018.

Jacob, G., Ehud, B.R., 2017. Training deep neural-networks using a noise adaptation layer, in: In ICLR.

Jiang, L., Zhou, Z., Leung, T., Li, L.J., Fei-Fei, L., 2018. MentorNet: Learning Data-Driven Curriculum for Very Deep Neural Networks on Corrupted Labels, in: Proceedings of the 35 th International Conference on Machine Learning.

Kaiser, P., Wegner, J.D., Lucchi, A., Jaggi, M., Hofmann, T., Schindler, K., 2017. Learning Aerial Image Segmentation From Online Maps. IEEE Transactions on Geoscience and Remote Sensing 55, 6054–6068.

Kellenberger, B., Marcos, D., Tuia, D., 2018. Detecting mammals in UAV images: Best practices to address a substantially imbalanced dataset with deep learning. Remote Sensing of Environment 216, 139–153.

Kemker, R., Salvaggio, C., Kanan, C., 2017. High-Resolution Multispectral Dataset for Semantic Segmentation. arXiv preprint arXiv:1703.01918 .

Li, Y., Yang, J., Song, Y., Research, Y., Cao, L., Luo, J., Li, L.J., 2017. Learning from Noisy Labels with Distillation, in: IEEE International Conference on Computer Vision.

Li, Y., Zhang, Y., Huang, X., Zhu, H., Ma, J., 2018. Large-Scale Remote Sensing Image Retrieval by Deep Hashing Neural Networks. IEEE Transactions on Geoscience and Remote Sensing 56, 950–965.

Luise, G., Rudi, A., Pontil, M., Ciliberto, C., 2018. Differential properties of sinkhorn approximation for learning with wasserstein distance. arXiv preprint arXiv:1805.11897 .

Maas, A., Rottensteiner, F., Heipke, C., 2016. Using label noise robust logistic regression for automated updating of topographic geospatial databases, in: ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences, Volume III-7.

Maas, A., Rottensteiner, F., Heipke, C., 2017. Classification under label noise based on outdated maps, in: ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences, Volume IV-1/W1.

Maas, A.E., Rottensteiner, F., Heipke, C., 2019. A label noise tolerant random forest for the classification of remote sensing data based on outdated maps for training. Computer Vision and Image Understanding 188, 102782.

Maggiori, E., Tarabalka, Y., Charpiat, G., Alliez, P., 2017. High-Resolution Aerial Image Labeling With Convolutional Neural Networks. IEEE Transactions on Geoscience and Remote Sensing 55, 7092–7103.

Masnadi-Shirazi, Hamed and Vasconcelos, N., 2008. On the Design of Loss Functions for Classification: Theory, Robustness to Outliers, and SavageBoost, in: Proceedings of the 21st International Conference on Neural Information Processing Systems, pp. 1049–1056.

Mnih, V., Hinton, G., 2012. Learning to label aerial images from noisy data, in: Proceedings of the 29th International Coference on International Conference on Machine Learning, Omnipress, USA. pp. 203–210. URL: http://dl.acm.org/citation.cfm?id=3042573.3042603.

Natarajan, N., Dhillon, I.S., Ravikumar, P., Tewari, A., 2013. Learning with Noisy Labels, in: 26th International Conference on Neural Information Processing Systems, Nevada. pp. 1196–1204.

Papadakis, N., 2015. Optimal Transport for Image Processing. Ph.D. thesis. URL: https://hal.archives-ouvertes.fr/tel-01246096.

Patrini, G., Rozza, A., Menon, A., Nock, R., Qu, L., 2017. Making Deep Neural Networks Robust to Label Noise: a Loss Correction Approach, in: 2017 IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI.

Pelletier, C., Valero, S., Inglada, J., Champion, N., Marais Sicre, C., Dedieu, G., Pelletier, C., Valero, S., Inglada, J., Champion, N., Marais Sicre, C., Dedieu, G., 2017a. Effect of Training Class Label Noise on Classification Performances for Land Cover Mapping with Satellite Image Time Series. Remote Sensing 9, 173.

Pelletier, C., Valero, S., Inglada, J., Dedieu, G., Champion, N., 2017b. Filtering mislabeled data for improving time series classification, in: 2017 9th International Workshop on the Analysis of Multitemporal Remote Sensing Images (MultiTemp), IEEE. pp. 1–4.

Reed, S., Lee, H., Anguelov, D., Szegedy, C., Erhan, D., Rabinovich, A., 2015. Training Deep Neural Networks on Noisy Labels with Bootstrapping, in: Workshop at International conference on learning representations.

Ren, M., Zeng, W., Yang, B., Urtasun, R., 2018. Learning to Reweight Examples for Robust Deep Learning, in: Proceedings of the 35 th International Conference on Machine Learning.

Richards, J.A.J.A., Jia, X., Gessner, W., 1999. Remote sensing digital image analysis : an introduction. Springer-Verlag.

Rigollet, P., Weed, J., 2018. Entropic optimal transport is maximum-likelihood deconvolution. arXiv preprint arXiv:1809.05572 .

van Rooyen, B., Menon, A.K., Williamson, R.C., 2015. Learning with symmetric label noise: the importance of being unhinged, in: Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 1, MIT Press. pp. 10–18.

Sáez, J.A., Galar, M., Luengo, J., Herrera, F., 2014. Analyzing the presence of noise in multi-class problems: alleviating its influence with the One-vs-One decomposition. Knowledge and Information Systems 38, 179–206.

Seguy, V., Damodaran, B, B., Flamary, R., Courty, N., Rolet, A., Blondel, M., 2018. Large-scale optimal transport and mapping estimation. ICLR, arXiv preprint arXiv:1711.02283 .

Smyth, P., Fayyad, U.M., Burl, M.C., Perona, P., Baldi, P., 1995. Inferring Ground Truth from Subjective Labelling of Venus Images, in: Advances in Neural Information Processing sustems, pp. 1085–1092.

Snow, R., O'Connor, B., Jurafsky, D., Ng, A.Y., 2008. Cheap and fast—but is it good?: evaluating non-expert annotations for natural language tasks, in: Proceedings of the Conference on Empirical Methods in Natural Language Processing, Association for Com-

putational Linguistics. pp. 254–263.

Solomon, J., de Goes, F., Peyré, G., Cuturi, M., Butscher, A., Nguyen, A., Du, T., Guibas, L., 2015. Convolutional wasserstein distances: efficient optimal transportation on geometric domains. ACM Transactions on Graphics 34, 66:1–66.

Sukhbaatar, S., Bruna, J., Paluri, M., Bourdev, L., Fergus, R., 2014. Training Convolutional Networks with Noisy Labels. arXiv preprint arXiv:1406.2080 .

Vahdat, A., 2017. Toward Robustness against Label Noise in Training Deep Discriminative Neural Networks, in: 31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA.

Villani, C., 2003. Topics in optimal transportation. 58, American Mathematical Soc.

Villani, C., 2009. Optimal transport: old and new. Grundlehren der mathematischen Wissenschaften, Springer.

Wang, L., Zhang, J., Liu, P., Choo, K.K.R., Huang, F., 2017. Spectral–spatial multi-feature-based deep learning for hyperspectral remote sensing image classification. Soft Computing 21, 213–221.

Wang, S., Bai, M., Mattyus, G., Chu, H., Luo, W., Yang, B., Liang, J., Cheverie, J., Fidler, S., Urtasun, R., 2016. TorontoCity: Seeing the World with a Million Eyes. arXiv preprint arXiv:1612.00423 .

Xia, G.S., Hu, J., Hu, F., Shi, B., Bai, X., Zhong, Y., Zhang, L., Lu, X., 2017. AID: A Benchmark Data Set for Performance Evaluation of Aerial Scene Classification. IEEE Transactions on Geoscience and Remote Sensing 55, 3965–3981.

Xiao, T., Xia, T., Yang, Y., Huang, C., Wang, X., 2015. Learning from massive noisy labeled data for image classification, in: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE. pp. 2691–2699.

Ye, D., Li, Y., Tao, C., Xie, X., Wang, X., 2017. Multiple Feature Hashing Learning for Large-Scale Remote Sensing Image Retrieval. ISPRS International Journal of Geo-Information 6, 364.

Yokoya, N., Iwasaki, A., 2016. Airborne hyperspectral data over Chikusei. Technical Report. Space Application Laboratory, University of Tokyo. Japan.

Zhang, C., Bengio, S., Hardt, M., Recht, B., Vinyals, O., 2017. Understanding deep learning requires rethinking generalization, in: International Conference on Learning Representations (ICLR).

Zhong, Z., Li, J., Luo, Z., Chapman, M., 2018. Spectral–Spatial Residual Network for Hyperspectral Image Classification: A 3-D Deep Learning Framework. IEEE Transactions on Geoscience and Remote Sensing 56, 847–858.

Zhou, W., Newsam, S., Li, C., Shao, Z., 2018. PatternNet: A benchmark dataset for performance evaluation of remote sensing image retrieval. ISPRS Journal of Photogrammetry and Remote Sensing .

Zhu, X., Wu, X., 2004. Class Noise vs. Attribute Noise: A Quantitative Study. Artificial Intelligence Review 22, 177–210.