Going Beyond One-Hot Encoding in Classification: Can Human Uncertainty Improve Model Performance in Earth Observation?

Christoph Koller, *Graduate Student Member, IEEE*, Göran Kauermann, and Xiao Xiang Zhu, *Fellow, IEEE*

***Abstract*— Technological and computational advances continuously drive forward the field of deep learning in remote sensing. In recent years, the derivation of quantities describing the uncertainty in the prediction—which naturally accompanies the modeling process—has sparked interest in the remote sensing community. Often neglected in the machine learning setting is the human uncertainty that influences numerous labeling processes. As the core of this work, the task of local climate zone (LCZ) classification is studied by means of a dataset that contains multiple label votes by domain experts for each image. The inherent label uncertainty describes the ambiguity among the domain experts and is explicitly embedded into the training process via distributional labels. We show that incorporating the label uncertainty helps the model to generalize better to the test data and increases model performance. Similar to existing calibration methods, the distributional labels lead to better-calibrated probabilities, which in turn yield more certain and trustworthy predictions. For reproducibility, we provide our code here https://github.com/ChrisKo94/LCZ_LDL and here https://gitlab.lrz.de/ai4eo/WG_Uncertainty/lcz_ldl.**

***Index Terms*— Calibration, classification, human uncertainty, local climate zones (LCZs), uncertainty quantification (UQ).**

I. Introduction

OVER the past years, deep learning has had a tremendous impact on many research fields across almost all domains, and remote sensing is no exception. Deep neural networks have enhanced the precision and accuracy of task-solving models by large margins. Concurrent advancing computational power of modern hardware has enabled such complex models to be trained while using continually shrinking time and resources. While the overall performance can be pushed by ever more complex models, the reliability of the resulting predictions is often neglected. Especially for image classification tasks, where prediction takes the form of a probability distribution over a set of possible classes, it is of crucial importance to rely on the model's confidence in its predictions. A common characteristic regarding the labels is label noise [1], which describes the pollution of the labels from potentially various sources.

Another aspect of label quality is ambiguity among the classes, which can occur naturally in the cases of, for example, multilabel classification or head pose estimation. As a proposed solution to deal with label ambiguity, label distribution learning [2] was introduced. The framework combines the idea of labels taking the form of a distribution across the space of the different possible classes, and a suitable loss function to learn this distribution.

Opening up the learning task to distributional labels can be highly beneficial for many applications. For safety-critical fields such as medical image analysis [3] in particular, predictions are required to be well-calibrated, while there is occasionally inevitable human label uncertainty [4]. These aspects also apply to remote sensing data, but are hardly ever discussed. Label ambiguity is a known but rarely tackled problem, where examples of exception are given by using the ambiguity information for synthetic aperture radar (SAR) image segmentation [5] or the application of label distribution learning toward aerial scene classification [6]. In this work, we focus on the classification of local climate zones (LCZs) from satellite images. LCZs are here adopted to allocate urban conglomerates and their environment around the world into 17 different clusters [7], which is helpful, for e.g., identifying potential urban heat islands (UHIs) or urban planning.

In a recently published work [8] the So2Sat LCZ42 dataset was introduced as a new benchmark dataset for which a label confidence of 85% was stated. We study this dataset, which contains satellite images of European cities as well as additional urban areas around the globe. As a peculiarity, for each image, we are presented with ten label votes from human remote sensing experts. These votes do not coincide for many images, which reflect the inherent human uncertainty about the labels in the data and which show the difficulties entangled with the labeling process. Our novel proposal is to embed this human uncertainty during the training of a

Manuscript received 13 February 2023; revised 1 July 2023 and 13 September 2023; accepted 20 October 2023. Date of publication 12 December 2023; date of current version 15 December 2023. This work was supported in part by the German Federal Ministry of Education and Research (BMBF) in the framework of the International Future AI Lab "AI4EO—Artificial Intelligence for Earth Observation: Reasoning, Uncertainties, Ethics and Beyond" under Grant 01DD20001; in part by the Helmholtz Association under the Joint Research School Munich School for Data Science—MUDS; in part by the German Federal Ministry for Economic Affairs and Climate Action in the framework of the "National Center of Excellence ML4Earth" under Grant 50EE2201C; and in part by the Munich Center for Machine Learning. *(Corresponding author: Xiao Xiang Zhu.)*

Christoph Koller is with the Department of Data Science in Earth Observation, Technical University Munich, 80333 Munich, Germany, and also with the Remote Sensing Technology Institute, German Aerospace Center, 82234 Weßling, Germany.

Göran Kauermann is with the Department of Statistics, LMU Munich, 80539 Munich, Germany.

Xiao Xiang Zhu is with the Department of Data Science in Earth Observation, Technical University Munich, 80333 Munich, Germany (e-mail: xiaoxiang.zhu@tum.de).

Digital Object Identifier 10.1109/TGRS.2023.3336357

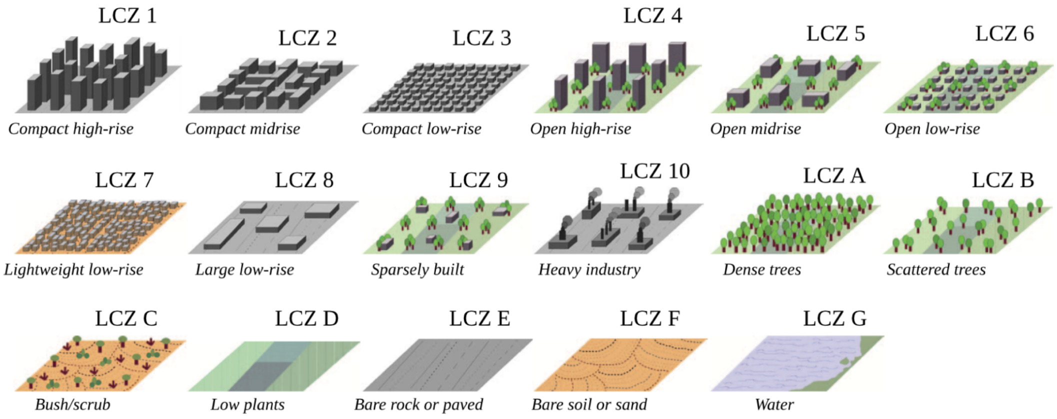


Fig. 1. LCZ classification scheme for the So2Sat LCZ42 dataset as shown by [9]. Classes 1–10 are urban areas, classes A–G are vegetation zones.

neural network classifier and investigate its performance and confidence compared to using traditional one-hot encoded labels. Our experiments send a clear message: Explicitly incorporating the inherent human uncertainty into the training process of the model is highly beneficial for both model performance and the calibration of predictive probabilities.

II. Related Work

Identifying climate zones is a common task when gathering information regarding land cover from satellite images. A popular and widely-used scheme is termed LCZs [7]. The different classes are defined to describe urban areas and their surroundings. Originally published for the evaluation of UHIs [10], [11], the scheme has since then been widely used for various climatological applications [12], [13] and urban plannings [14], [15], [16]. The community-driven project termed World Urban Database and Portal (WUDAPT) [17] targets a global high-quality coverage with the LCZ scheme. So far, significant effort has been shown to reach this goal [18], [19], [20], [21]. With the advancements of deep learning in the remote sensing community, more complex models have been established in order to classify processed satellite imagery into LCZs. Data from both the Landsat [22], [23] and the Sentinel satellite missions [24], [25] have been successfully employed for this task. More recently, a new benchmark dataset [8] was introduced to the community. For a scope of 42 global cities, high-quality satellite image patches are freely provided with manually crafted LCZ labels. Recent works based on this benchmark dataset explore various network architectures [26] or fuse multiple data sources [27], [28], [29]. Deep learning models trained on this benchmark have also been applied to achieve global urban LCZ maps for a better understanding of the global urban morphology [30]. As for land cover classification in general, knowledge graphs have been successfully adapted recently [31]. In particular, for zero-shot classification, where previously unseen data is classified, several works have utilized knowledge graphs [32], [33].

The goal of uncertainty quantification (UQ) in the field of deep learning lies in building a model that not only provides a prediction but also a measure of certainty or confidence [34]. Generally, we can distinguish between epistemic uncertainty, which is caused by the model, and aleatoric uncertainty inherent in the data [35]. Epistemic uncertainty can be reduced by finding a more suitable model or architecture, whereas aleatoric uncertainty is irreducible. Methodological directions to estimate uncertainty quantities include the aggregation of multiple neural networks [36], deterministic networks with distributional assumptions placed on the label space [37], [38], sophisticated use of dropout networks [39], or the Bayesian neural networks [40], [41]. Domain-specific applications in the remote sensing area are still rare [42], [43].

Calibration is a closely related concept, which in the context of classification tasks aims at providing more reliable predictions. In particular, the probabilities derived from a machine learner should adequately describe the certainty inherent in the prediction. Originally proposed for support vector machines [44], Platt scaling describes a parametric method that trains an additional layer to transform the predictions of a classifier into calibrated probabilities. A simplified version termed temperature scaling (TS) has found its way into more recent deep learning-based works [45], [46], [47]. Using only a single parameter, the approach simply scales the softmax probabilities derived from a deep learner in order to limit overconfidence. Another famous technique is termed label smoothing (LS) [48] and helps to overcome overconfident predictions by artificially changing the labels during training. Several recent works [49], [50], [51] have investigated this method since.

As a novelty for the deep learning community, human uncertainty, derived from the labeling stage of the data generation process, has been studied [52]. Specifically, a set of images was labeled by multiple people and therefore received votes for potentially different classes. This rich information was then added to the training process in order to capture human categorization peculiarities [53] or make the underlying classification task more robust [52]. Comparisons to more classic calibration techniques such as LS were drawn in [54]. Though closely linked, the incorporation of this additional information into the labels stands in contrast to the widely studied research direction of investigating label noise [1], [55], [56], which focuses more on identifying mislabeled or anomalous data, or pointing out insufficiently labeled data. In the field of

TABLE I

CITIES AND ADDON AREAS IN THE EVALUATION SUBSET OF THE SO2SAT LCZ42 DATASET

Cities		Addon Areas	
Amsterdam	Berlin	Guangzhou	Islamabad
Cologne	London	Jakarta	Los Angeles
Madrid	Milan	Moscow	Mumbai
Munich	Paris	Nairobi	Riodejaneiro
Rome	Zurich		

TABLE II

MANUAL SPLIT OF THE CITIES AND ADDON AREAS OF THE EVALUATION SET INTO TRAINING DATA AND NONTRAINING DATA, WHICH WAS IN A SECOND STEP SPLIT RANDOMLY BY HALF INTO VALIDATION AND TESTING DATA

	Cities		Addon Areas	
Training Data	Amsterdam Cologne Milan	Berlin London Rome	Guangzhou Jakarta	Islamabad Nairobi
Validation / Test Data	Madrid Paris	Munich Zurich	Los Angeles Mumbai	Moscow Riodejaneiro

remote sensing, the first advances in investigating human uncertainty have been made recently [57], [58].

III. DATA AND METHODOLOGY

A. Satellite Data

As a basis of this work, the So2Sat LCZ42 dataset [8] is analyzed, comprising approximately 400k labeled image patches of size 32 × 32 pixels linked each to an area of 320 × 320 m. The labels follow the classification scheme introduced by [59]: 17 characteristic climate zones, of which ten are defined as urban zones and seven are vegetation zones (see Fig. 1). In the publicly available version,[1] three splits for training, validation, and testing can be chosen: A completely random split of the data, a split at the city level where each city is separated into geographically separated training and testing data, and a third approach that sets aside ten cities from different cultural zones for testing. The labeling process was performed in a manual and labor-intensive process, which is explained in more detail by [8] and largely follows a classic procedure initially used in the World Urban Database (WUDAPT) project [60]. Overall, 13 spectral bands of the Sentinel satellite mission with varying spatial resolutions are available from the Copernicus Hub. For the LCZ data, all four bands with a ground sampling distance (GSD) of 10 m were chosen, as well as the bands with a GSD of 20 m, which were upsampled to 10 m GSD.

As an additional experiment carried out in their work, Zhu et al. [8] launched an evaluation phase to assess the quality of the labeling process. For this, a subset of ten European cities was chosen to serve as reference data for the labelers; nine additional non-European regions were included to ensure a minimum level of class balance. Overall, the evaluation dataset contains roughly 250k satellite image patches, which were cropped out of polygons from homogeneous regions. The entire list of cities, as well as add-on areas for which the label evaluation was performed, is given in Table I.

In order to geographically separate training and testing data, the corresponding datasets were formed by mutually exclusive subsets of the above cities and addon areas. The split on a city level was specifically chosen not to avoid learning the overall similar data distribution, as all images come from European cities, but rather to train on one set of cities and predict as well as evaluate the method on another set of cities. In Table II, the cities in the respective datasets are listed. Note that for the separation of validation and testing data, a random split of the cities and addon areas was performed, halving the entirety of data points into validation and testing data.

Regarding the choice of cities, next to the geographical separation, a strong emphasis was laid on balancing the occurring class frequencies in between the datasets. For the investigated urban classes (LCZs 1–10), the class frequencies of the different datasets relative to the number of samples in the entire dataset can be found in Table III. Note that for the ground-truth label, here the majority vote $\boldsymbol{y}_{\max}$ as later defined in (3) was taken. Images with incomplete label distributions, that is, those having a majority vote from the urban classes and one or more individual votes from the nonurban classes occur very rarely (∼0.1%). The outlier votes have been excluded from the analysis. As can be deduced from Table III, the classes are not perfectly balanced among the training, validation, and test datasets. Yet a significant effort was spent on finding a good split on city level that still retains a moderate level of class imbalance. This imbalance among the datasets adds complexity to the imbalance among the different classes already present in the data.

B. Human Uncertainty

As part of the label evaluation study, a group of ten human remote sensing experts independently cast a label vote for each of the polygons, resulting in a final dataset with ten expert votes for each image. This dataset was made public accompanying this publication and can be downloaded online.[2] Transforming these expert votes into suitable labels for a classification task can be handled in a variety of ways. Thus, let $\boldsymbol{Y} = \boldsymbol{Y}^{(1)}, \dots, \boldsymbol{Y}^{(n)}$ initially be the vote count vectors for the images $i = 1, \dots, n$, where $\boldsymbol{Y}^{(i)} = (Y_1^{(i)}, \dots, Y_K^{(i)})$ stores the vote counts for each of the K LCZ classes for image i where $K = 17$. By indexing the different remote sensing experts via $j = 1, \dots, J$, for image i we receive the expert votes

$$V_1^{(i)}, \dots, V_J^{(i)}, \quad V_j^{(i)} \in \{1, \dots, K\}\ \forall i = 1, \dots, n. \tag{1}$$

We define the votes as a K-dimensional vector

$$\boldsymbol{V}_j^{(i)} = \left(\mathbb{1}_{\{V_j^{(i)}=1\}}, \dots, \mathbb{1}_{\{V_j^{(i)}=K\}}\right), \quad \text{where} \quad \mathbb{1}_{\{V_j^{(i)}=k\}} = 1 \Leftrightarrow V_j^{(i)} = k \tag{2}$$

[1] So2Sat LCZ42 dataset (link to download: https://mediatum.ub.tum.de/1613658).

[2] So2Sat LCZ42 evaluation dataset (link to download: https://mediatum.ub.tum.de/1659039).

TABLE III

RELATIVE CLASS FREQUENCIES OF THE URBAN CLASSES WITHIN THE TRAINING, VALIDATION AND TESTING SET. TOTALS ARE LISTED IN THE RIGHT AND BOTTOM ENTRIES

	Urban LCZ										
	1	2	3	4	5	6	7	8	9	10	$\sum$
Training Data	0.83	0.56	0.87	0.61	0.83	0.74	0.65	0.74	0.65	0.94	43k
Validation Data	0.09	0.22	0.07	0.20	0.08	0.13	0.18	0.13	0.17	0.03	8k
Test Data	0.09	0.22	0.07	0.20	0.08	0.13	0.18	0.13	0.17	0.03	8k
$\sum$	< 1k	10k	2k	1k	11k	18k	3k	9k	1k	3k	59k

and obtain the vote counts via $Y_k^{(i)} = \sum_j \mathbb{1}_{\{V_j^{(i)}=k\}}$. In particular, $Y_k^{(i)} = m$ means that for image i class k received m votes and it holds that $\sum_{k=1}^{K} Y_k^{(i)} = M$, where $M = 10$ represents the number of votes or experts. A common strategy is to rely on the majority vote of the experts that we define for image i as $Y_{\max}^{(i)} := \max_j Y_j^{(i)}$ to be the class which received the most expert votes. The associated one-hot encoded label is denoted as a K-dimensional vector

$$\boldsymbol{y}_{\max}^{(i)} = \left(\mathbb{1}_{\left\{Y_1^{(i)}=Y_{\max}^{(i)}\right\}}, \ldots, \mathbb{1}_{\left\{Y_K^{(i)}=Y_{\max}^{(i)}\right\}}\right), \quad \text{where}$$
$$\mathbb{1}_{\left\{Y_j^{(i)}=Y_{\max}^{(i)}\right\}} = 1 \Leftrightarrow Y_j^{(i)} = Y_{\max}^{(i)}. \tag{3}$$

This simplification gives rise to the question of how much information is lost when relying on the majority label decision. It is argued by [8], that majority voting does help to improve label confidence. Yet part of the uncertainty in the voting process is hidden from the classifier when it is presented with the majority vote $\boldsymbol{y}_{\max}$ in (3). An alternative approach is hence to incorporate the entirety of all votes directly into the label by forming a distributional label. To do so, we directly use the empirical distribution formed by the observed votes and define the soft label for image i in the following discussions via

$$\boldsymbol{y}_{\text{distr}}^{(i)} = \boldsymbol{Y}^{(i)}/M. \tag{4}$$

By doing so, the mode of the distributional label coincides with the mode of $\boldsymbol{y}_{\max}$ (3), but the distributional form allows for a more flexible learning approach, as other classes that were voted for in the evaluation process are also considered.

The ten label votes received from each of the ten independent remote sensing experts inherently store a notion of label uncertainty. This uncertainty is visualized in Fig. 2 by means of a confusion matrix between the individual label votes $V_1^{(i)}, \ldots, V_J^{(i)}$ and the majority vote $\boldsymbol{y}_{\max}^{(i)}$ for all images $i = 1, \ldots, n$, as well as a plot showing the entropies of the individual label distributions. In detail, for the given label distribution $\boldsymbol{y}_{\text{distr}}^{(i)}$ [see (4)], we compute the information theoretic (Shannon) entropy of the distribution given for image i via

$$H\left(\boldsymbol{y}_{\text{distr}}^{(i)}\right) = -\sum_{k=1}^{K} \boldsymbol{y}_{\text{distr},k}^{(i)} \log \boldsymbol{y}_{\text{distr},k}^{(i)} \tag{5}$$

and plot the resulting values in a bar plot grouped by the affiliation of the respective majority vote $\boldsymbol{y}_{\max}^{(i)}$ [see (3)] to either the urban or nonurban classes. Whereas zero entropy occurs when all voters agree on a label for an image, maximum entropy is achieved for a uniform distribution across all labels (which cannot occur here because the number of classes exceeds the number of voters). Clearly visible is the higher average entropy of the vote vectors for the urban classes, which corresponds to a higher uncertainty associated with the respective satellite images. We chose to limit our analysis to the urban classes due to the large share of images with the majority vote belonging to nonurban classes that at the same time have zero entropy in the votes.

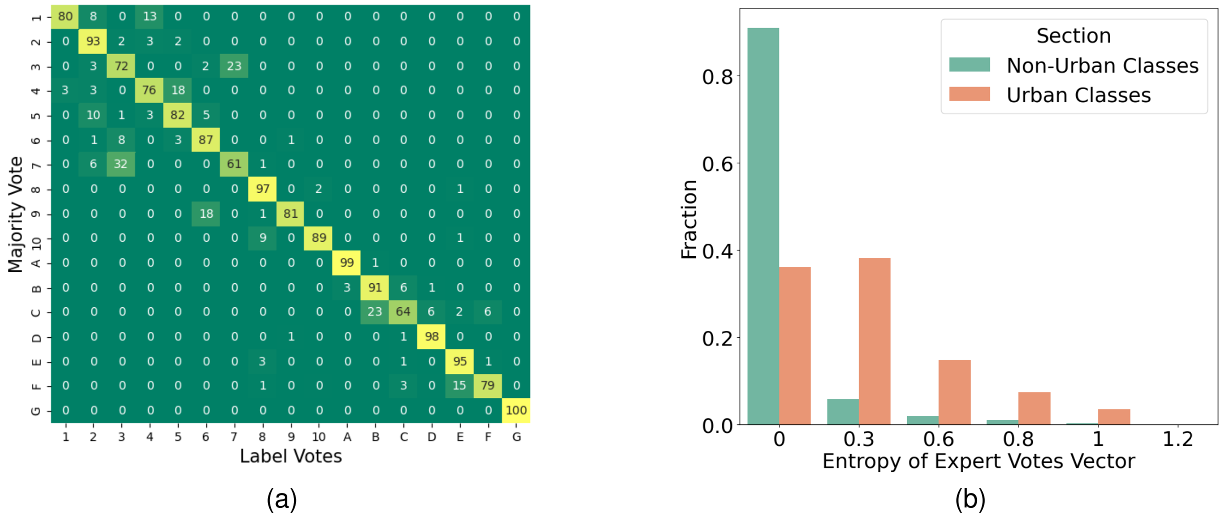
C. Training Process

In the following, let $\{\boldsymbol{x}^{(i)}, \boldsymbol{y}^{(i)}\}_{i=1,\ldots,n} \in (\mathcal{X} \times \mathcal{Y})^n$ be the classification data: $\boldsymbol{x}^{(1)}, \ldots, \boldsymbol{x}^{(n)} \in \mathcal{X}$ are the multispectral LCZ42 image patches, and $\boldsymbol{y}^{(1)}, \ldots, \boldsymbol{y}^{(n)} \in \mathcal{Y}$ are the corresponding labels. Furthermore, let $f_\theta(x)$ be a neural network classifier based on the parameters stored in θ. Given an input $x \in \mathcal{X}$, the predictive distribution of the network is denoted by $p_\theta(y|x)$, and $p_\theta(y = k|x)$ returns the estimated probability of x belonging to class k. For training, typically the cross-entropy loss is used in a classification setting. From an information theory perspective, the cross-entropy defines the amount of additional information needed to approximate a sample from the source distribution.

However, in the literature for the recently presented research field of label distribution learning, the Kullback-Leibler (KL) divergence has been established as a loss function [2]. Given a true distribution, the KL divergence measures the information lost when the target distribution is used as an approximation. It measures the dissimilarity between two probability functions and is therefore also termed relative entropy.

If we now assume for the So2Sat LCZ42 data that the distribution formed by the votes of the ten remote sensing experts is the ground truth label distribution, we can measure the dissimilarity between the predicted and the ground truth distribution with the KL divergence. When implemented as a loss function, it is derived for a batch of data $\{\boldsymbol{x}^{(i)}, \boldsymbol{y}_{\text{distr}}^{(i)}\}_{i=1,\ldots,m}$ via

$$\mathcal{L}_{\text{KL}}\left(f_\theta, \boldsymbol{x}^{(1)}, \ldots, \boldsymbol{x}^{(m)}, \boldsymbol{y}_{\text{distr}}^{(1)}, \ldots, \boldsymbol{y}_{\text{distr}}^{(m)}\right)$$
$$= -\frac{1}{m} \sum_{i=1}^{m} \sum_{k=1}^{K} y_{\text{distr},k}^{(i)} \cdot \log \frac{y_{\text{distr},k}^{(i)}}{p_\theta\left(\boldsymbol{y}^{(i)} = k|\boldsymbol{x}^{(i)}\right)}. \tag{6}$$



Majority Vote \ Label Votes	1	2	3	4	5	6	7	8	9	10	A	B	C	D	E	F	G
1	80	8	0	13	0	0	0	0	0	0	0	0	0	0	0	0	0
2	0	93	2	3	2	0	0	0	0	0	0	0	0	0	0	0	0
3	0	3	72	0	0	2	23	0	0	0	0	0	0	0	0	0	0
4	3	3	0	76	18	0	0	0	0	0	0	0	0	0	0	0	0
5	0	10	1	3	82	5	0	0	0	0	0	0	0	0	0	0	0
6	0	1	8	0	3	87	0	0	1	0	0	0	0	0	0	0	0
7	0	6	32	0	0	0	61	1	0	0	0	0	0	0	0	0	0
8	0	0	0	0	0	0	0	97	0	2	0	0	0	0	1	0	0
9	0	0	0	0	0	18	0	1	81	0	0	0	0	0	0	0	0
10	0	0	0	0	0	0	0	9	0	89	0	0	0	0	1	0	0
A	0	0	0	0	0	0	0	0	0	0	99	1	0	0	0	0	0
B	0	0	0	0	0	0	0	0	0	0	3	91	6	1	0	0	0
C	0	0	0	0	0	0	0	0	0	0	0	23	64	6	2	6	0
D	0	0	0	0	0	0	0	0	1	0	0	0	1	98	0	0	0
E	0	0	0	0	0	0	0	3	0	0	0	0	1	0	95	1	0
F	0	0	0	0	0	0	0	1	0	0	0	0	3	0	15	79	0
G	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	100

(a)

(b)

Fig. 2. (a) Confusion matrix of label votes for the evaluation dataset. (b) Entropies of the voting distributions.

Note that for $y^{(i)}_{\text{distr},k} = 0$ we set $y^{(i)}_{\text{distr},k} \cdot \log y^{(i)}_{\text{distr},k} / (p_\theta(\boldsymbol{y}^{(i)} = k | \boldsymbol{x}^{(i)})) \equiv 0$. This approach is taken to measure the information loss when using the predictive distribution of the neural network to approximate the assumed ground truth distribution over labels. Consequently, training with the KL divergence describes the process of iteratively finding a neural network mimicking human voting behavior.

D. Calibration and Generalization

Neural network classifiers are often prone to overconfidence, namely that predicted probabilities for a class overestimate the percentage of times the algorithm actually yields a correct prediction [46]. In this circumstance, derived uncertainty quantities are not reliable, because the underlying probabilities are ill-defined in the first place. Therefore, one can speak of a frequentist notion of uncertainty when referring to calibration [36]. Deviations from the perfectly calibrated model can be measured with different error rates, of which the expected calibration error (ECE) is the most prominent one. It can be visualized in a so-called reliability diagram [61], [62], which displays the accuracy versus the corresponding confidence in a single plot, averaged over predefined intervals. Following the notation of [46], we denote these bins by $B_m = ((m-1)/M, (m/M)], m = 1, \ldots, M$ and define the indices of the images whose confidences $\hat{p}^{(i)}$ fall into the respective bin by $I_m, m = 1, \ldots, M$. The needed quantities are then calculated by

$$\mathrm{acc}(I_m) = \sum_{i \in I_M} \mathbb{1}_{\left\{\hat{y}^{(i)} = \boldsymbol{y}^{(i)}_{\max}\right\}} \quad \text{and } \mathrm{conf}(I_m) = \sum_{i \in I_m} \hat{p}^{(i)} \tag{7}$$

and displayed in a 2-D plot with respect to the earlier defined intervals. Here, $\hat{\boldsymbol{y}}^{(i)}$ is the one-hot encoded predicted class for input $\boldsymbol{x}^{(i)}$, $\hat{p}^{(i)}$ the corresponding predicted probability, and one defines $\boldsymbol{y}^{(i)}_{\max}$ as in (3). The ECE is then derived via

$$\mathrm{ECE} = \sum_{m=1}^{M} \frac{|I_m|}{n} |\mathrm{acc}(I_m) - \mathrm{conf}(I_m)|. \tag{8}$$

A simpler version of the ECE, which only considers the maximum gap between confidence and accuracy out of all bins considered, is the maximum calibration error (MCE). Given the predictions $\hat{\boldsymbol{y}}^{(i)}, i = 1, \ldots, n$ and corresponding confidences $\hat{p}^{(i)}, i = 1, \ldots, n$, the computation is as follows:

$$\mathrm{MCE} = \max_m |\mathrm{acc}(I_m) - \mathrm{conf}(I_m)|. \tag{9}$$

While the MCE gives a first indication whether the evaluated classifier is severely miscalibrated for a certain bin, the same downside appears as for the ECE, namely that the included metrics are not considered class-specific. For this purpose, Nixon et al. [63] introduced the static calibration error (SCE), which measures the accuracies and confidences on a class level. For class k and within bin m, these are depicted by $\mathrm{acc}(m, k)$ and $\mathrm{conf}(m, k)$, leading to the formula for the SCE given by

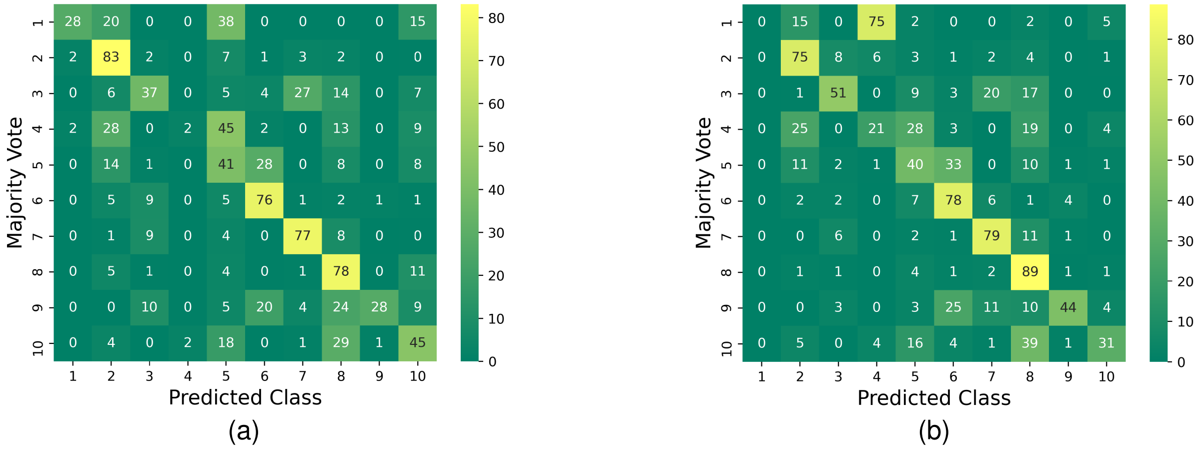
$$\mathrm{SCE} = \frac{1}{K} \sum_{k=1}^{K} \sum_{m=1}^{M} \frac{n_{mk}}{n} |\mathrm{acc}(m, k) - \mathrm{conf}(m, k)| \tag{10}$$

where n_{bk} denotes the number of samples belonging to class k (with respect to the majority vote) within bin b. Here, the class-specific scores are also weighted with respect to the class share within the respective bin.

Presented as a straightforward and effective calibration strategy, Guo et al. [46] introduced TS. The technique can be seen as a simplification to Platt scaling [44] which scales the logits predicted by the classifier by a constant parameter. So given the predicted logit of the neural classifier $f_\theta(x)$ for image i via $z^{(i)}$, the corresponding softmax prediction of the scaled logit is given via

$$\mathrm{softmax}\left(z^{(i)}\right) = \frac{\exp\left(z^{(i)}/T\right)}{\sum_k \exp\left(z^{(i)}_k/T\right)}. \tag{11}$$

As described by [46], we optimized the parameter T (termed temperature) with respect to the negative log-likelihood on the validation dataset. Adding to that method, LS represents another quick and easy to implement off-the-shelf calibration



	1	2	3	4	5	6	7	8	9	10
1	28	20	0	0	38	0	0	0	0	15
2	2	83	2	0	7	1	3	2	0	0
3	0	6	37	0	5	4	27	14	0	7
4	2	28	0	2	45	2	0	13	0	9
5	0	14	1	0	41	28	0	8	0	8
6	0	5	9	0	5	76	1	2	1	1
7	0	1	9	0	4	0	77	8	0	0
8	0	5	1	0	4	0	1	78	0	11
9	0	0	10	0	5	20	4	24	28	9
10	0	4	0	2	18	0	1	29	1	45

(a)

	1	2	3	4	5	6	7	8	9	10
1	0	15	0	75	2	0	0	2	0	5
2	0	75	8	6	3	1	2	4	0	1
3	0	1	51	0	9	3	20	17	0	0
4	0	25	0	21	28	3	0	19	0	4
5	0	11	2	1	40	33	0	10	1	1
6	0	2	2	0	7	78	6	1	4	0
7	0	0	6	0	2	1	79	11	1	0
8	0	1	1	0	4	1	2	89	1	1
9	0	0	3	0	3	25	11	10	44	4
10	0	5	0	4	16	4	1	39	1	31

(b)

Fig. 3. Exemplary pairwise confusion between predicted class and respective majority vote for the test set. (a) One-hot encoding. (b) Label distribution encoding.

method. As opposed to TS, the method is not applied post-hoc, but uses a hyperparameter α to scale the labels before training. Given the label $y^{(i)}$ for image i, the scaled version is then received via

$$y^{(i)}_{\text{smoothed}} = \alpha \cdot u_K + (1-\alpha) \cdot y^{(i)} \tag{12}$$

where u_K denotes the uniform distribution over the K classes.

The hyperparameter is not directly optimized, but empirically chosen. As a third calibration option, we utilized Monte Carlo Dropout [39]. Given the already trained networks, we left the dropout layers of Sen2LCZ active during prediction and by doing so created a set of 20 unique predictions. After averaging the softmax vectors of the individual predictions, we proceeded with deriving the calibration and generalization metrics as before. Lastly, a deep ensemble [36] with $k = 5$ ensemble members was considered. As proposed by the authors, five identical models were trained in parallel. Afterward, the calibration and generalization metrics were calculated based on the average of the five predictions.

Further measures of generalization as shown in Table IV are namely are the cross-entropies between the predictive distribution of the network $p_\theta(y|x)$ and the one-hot encoded label based on the majority vote $y_{\max}$ or the distributional label y_{distr}. They are termed "CE One-hot" and "CE Distr" and given for image i via

$$\begin{aligned}&\text{CE}\left(y^{(i)}_{\max}, p_\theta\left(y^{(i)}|x\right)\right)\\ &\quad= -\sum_{k=1}^{K} y^{(i)}_{\max,k} \log\left(p_\theta\left(y^{(i)}=k|x\right)\right) \quad \text{and}\end{aligned} \tag{13}$$

$$\begin{aligned}&\text{CE}\left(y^{(i)}_{\text{distr}}, p_\theta\left(y^{(i)}|x\right)\right)\\ &\quad= -\sum_{k=1}^{K} y^{(i)}_{\text{distr},k} \log\left(p_\theta\left(y^{(i)}=k|x\right)\right)\end{aligned} \tag{14}$$

respectively.

IV. Experiments

A. Setup and Settings

For the classification task, we use the model introduced by [26] (Sen2LCZ-Net), which is a modified convolutional neural network (CNN) using intermediate deep feature representations at multiple stages of the network. These representations are then averaged and pooled at the end before being transformed into the logit space. In particular, we used a network depth of 17 (following from the use of four convolutional layers in each block), a width of 16, a dropout rate of 0.2 (at the end of the second and third block), and activated multilevel feature fusion and double-pooling. Class weights did not lead to improved results, hence they were discarded since also the class imbalance differed largely between the training, validation, and testing set. The Nesterov Adam optimizer implementation of Keras [64] was used for training. An early stopping mechanism was installed, which monitored the validation loss with a patience of 20 epochs. Weights were saved after every epoch if and only if the validation loss decreased. The model has been shown to be superior over many state-of-the-art neural network architectures in extensive benchmark tests on the So2Sat LCZ42 dataset by the authors.

LS was performed with a smoothing parameter of 0.1. TS was implemented as described by [46] via tuning the scaling parameter on the validation set with respect to minimizing the negative log-likelihood. This minimization was performed using the Adam optimizer implementation of Keras [64] with a learning rate of 0.01 and a maximum of 10k iterations (more iterations did not improve the results significantly). It is worth noting that TS scales the logits, but does not change the accuracy of the model, being the reason why we did not include the metric in Table V. Regarding the hyperparameters of Sen2LCZ, we set a batch size of 64 and an initial learning rate of 2×10^{-3} which was gradually reduced by a factor of 0.5 every five epochs.

B. Results

Sen2LCZ-Net was trained on the training set utilizing different labels and loss functions plus additional off-the-shelf calibration methods. The categorical cross-entropy loss incorporates the one-hot encoded labels based on the majority vote of the expert votes, and the KL divergence takes the distribution formed by the empirical distribution of the expert votes. All models were trained using the same training and

TABLE IV

Cross-Entropies Between Predicted Softmax Probabilities and Labels on the Test Set as Well as Calibration Errors, Averaged Over Five Runs. CE = Cross Entropy, LS = Label Smoothing, TS = Temperature Scaling, MC-Drop = Monte Carlo Dropout. Binning Was Performed Using 20 Equally-Sized Bins. Ensembles Treated Separately as They Were Not Performed Multiple Times

	CE One-hot ↓	CE Distr. ↓	ECE ↓	MCE ↓	SCE ↓
One-hot	1.12 ± 0.05	1.38 ± 0.07	9.79 ± 3.18	23.14 ± 3.97	**1.03 ± 0.50**
+ LS	1.05 ± 0.01	1.23 ± 0.03	7.33 ± 2.62	19.80 ± 4.82	1.11 ± 0.25
+ TS	1.00 ± 0.13	1.17 ± 0.07	4.15 ± 2.37	15.88 ± 10.60	1.44 ± 0.22
+ LS & TS	1.02 ± 0.03	1.18 ± 0.02	**3.21 ± 0.96**	**11.30 ± 4.26**	1.32 ± 0.03
+ MC-Drop	1.12 ± 0.05	1.37 ± 0.06	9.57 ± 3.11	23.10 ± 4.22	1.04 ± 0.49
+ LS & MC-Drop	1.05 ± 0.01	1.23 ± 0.03	7.11 ± 2.41	33.22 ± 26.60	1.12 ± 0.24
Distr.	1.06 ± 0.07	1.21 ± 0.07	5.80 ± 1.07	15.57 ± 4.22	1.21 ± 0.20
+ LS	0.98 ± 0.03	1.08 ± 0.02	8.31 ± 2.46	17.32 ± 4.84	1.73 ± 0.06
+ TS	0.96 ± 0.09	1.07 ± 0.07	5.89 ± 2.50	15.37 ± 3.94	1.72 ± 0.15
+ LS & TS	**0.95 ± 0.04**	**1.05 ± 0.04**	4.21 ± 1.38	15.35 ± 1.95	1.58 ± 0.10
One-hot ensemble	0.84	1.03	4.83	16.70	1.74
Distr. ensemble	0.89	1.01	7.96	23.12	1.68

validation data. Identical hyperparameters were used (after grid search), and the same stopping and convergence criteria were set. In particular, the same input in the form of satellite imagery was used in both models.

Exemplary confusion matrices of the predictions on the test set are presented for both models in Fig. 3. Here, the ground truth label is set to be the majority vote of the experts, and the predicted class is defined via the highest predicted probability of the respective network $\hat{p} = \max_k p_\theta(y = k|x)$. First note that for most of the part, both models have the same pitfalls of misclassifying certain classes. This holds in particular for class 3 (compact low-rise) being falsely classified as class 7 (lightweight low-rise), class 4 (open high-rise) being misclassified as class 2 (compact midrise), as well as class 5 (open midrise) being classified as class 6 (open low-rise). Adding to this claim is the strong confusion with class 8 (large low-rise), which is often falsely predicted for images from various classes (with respect to the majority vote).

In a similar manner, images for which experts agreed on class 9 (sparsely built) are often falsely classified as classes 6 and 8. The results regarding the majority-voted class 1 seem rather arbitrary, however, as they largely depend on the data distribution of the few test samples. In particular, there are only around 30 samples of class 1 (with respect to the majority vote) in the test set, of which none could be correctly classified using the distributional approach. Note, however, that the high confusion with class 4 is also visible to some extent in the human label uncertainty. It could furthermore be the case that those particular samples where this confusion occurs in the data have mostly ended up in the test set. Of particular interest is furthermore the comparison between the model confusion and the confusion among voters. With one in every three votes deviating from the majority vote, experts chose to vote for class 3 in cases where the majority of experts settled for class 7. This large confusion is only slightly reflected in model confusion.

The opposite case is matched rather closely in the predictions of both models on the test set. With more than one in every six votes deviating from the majority vote, the combinations of classes 4 and 5, as well as 9 and 6, are also found to be confusing for the model in both training cases. However, in the cases of classes 4 and 5, it seems to be significantly less confusing for the model trained with distributional labels. Further deviations of the model confusion from the human confusion can be partially attributed to the previously mentioned arbitrariness of the very low number of affected samples in the data. All other model confusions cannot be directly linked to the confusion in the voting process and therefore result from difficulties entirely attributable to the model.

TABLE V

Performance Scores on the Test Set. Accuracy and Other Related Measures Were Derived With Respect to the Majority Vote. All Scores Are Averaged Over Five Runs. OA = Overall Accuracy, MAA = Macro Avg. Accuracy, WAA = Weighted Avg. Accuracy, κ = Kappa Score, LS = Label Smoothing

	OA ↑	MAA ↑	WAA ↑	κ ↑
One-hot	68.4 ± 5.5	42.9 ± 6.4	69.6 ± 2.2	60.2 ± 6.4
+ LS	67.5 ± 2.4	**50.3 ± 3.5**	69.9 ± 2.0	59.4 ± 2.7
Distr.	67.0 ± 2.2	45.8 ± 3.9	**71.0 ± 0.5**	58.8 ± 2.4
+ LS	**68.6 ± 2.3**	43.4 ± 6.1	69.7 ± 2.1	**60.4 ± 2.8**

We can deduce the results of the trained models in Table V. For the distributional label approach, accuracy was again measured with respect to majority voting, which helps to explain the observed minor differences in performance between the two model configurations. Although the distributional approach performs on average better than the regular approach in 3 out of 4 metrics, the macro average accuracy lags behind by a large margin. Regarding the deviance of predictions from the true labels, Table IV shows the cross-entropy between the two distributions on the identical test set as well as the introduced calibration error rates (based on 20 identically-sized bins). When trained with distributional labels, the model can fit better toward the test label distributions when predicting

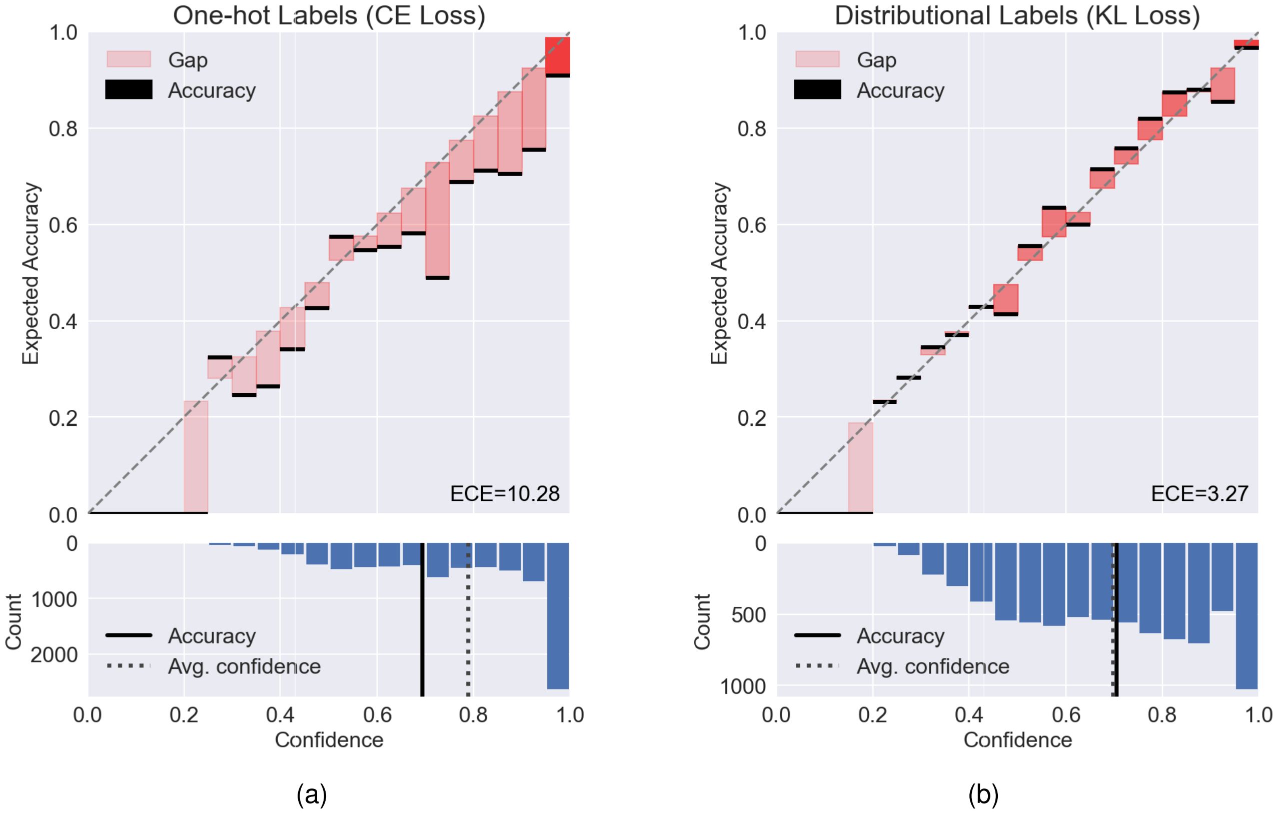


Fig. 4. Exemplary reliability diagrams on the test set. Both visualizations were created using [65]. (a) One-hot encoding. (b) Label distribution encoding.

on the test set by a large margin. This holds true even though the model with distributional labels was trained with the KL divergence as loss.

While LS helps to improve the performance in both settings, TS leads to mixed calibration performance results. The additional use of Monte Carlo Dropout was enforced via averaging the confidences of multiple predictions with activated dropout. While this enabled to bind the ECE to a reasonable level, the generalization performance was nearly unaffected. The use of an ensemble of classifiers, on the other hand, strongly impacted the generalization performance but worsened the calibration ability of the model. Note that we considered the ensemble results separately, as they were not averaged over multiple runs. This would have increased the overall computational demand by a multiple of the number of runs. Overall, all calibration errors benefit from the uncertainty-guided approach, although the overall best errors are achieved with conventional calibration techniques.

Yet more unforeseen is the cross-entropy with respect to the one-hot encoded labels. Although the model trained with the distributional labels uses a different loss function and different labels, it can better approximate the one-hot encoded test labels and underlines the strong generalization performance of the human uncertainty models. While the additional off-the-shelf calibration methods help to marginally improve the cross-entropies, the ECE is negatively affected by individual calibration methods in the distributional setting. An explanation for this could be the fact that the optimization of the temperature is prone to overfitting on the validation dataset, leading to poor generalization when already considering human uncertainty within the training process. The overall better cross-entropies of the models including human label uncertainty underline a crucial aspect of training with distributional labels: If chosen appropriately, they keep the model more flexible when generalizing to the test data. When explicitly modeling label distributions, all classes with a nonzero entry in the distribution of the label are changing the value of the loss (as compared to the logit of the ground truth class in the case of cross entropy training); therefore the training is more flexible and less prone to overconfidence in its predictions.

This phenomenon is particularly visible in Fig. 4, which displays reliability diagrams [46] of a single run for the two modeling approaches in an uncalibrated setting. Clearly visible is the significantly lower ECE for the model trained with the distributional labels, especially for the regions with higher confidence. In these regions, the confidence is more meaningful and is reflected better by the accuracy, resulting in a better-calibrated model when using distributional labels. Since calibration can be treated as the frequentist notion of uncertainty [36], we can deduce that by incorporating the voting uncertainty into the training process, the resulting model is better calibrated and hence yields a predictive distribution that contains a more feasible sense of uncertainty.

Moreover, the lower part of Fig. 4 shows the overconfidence of the one-hot encoded model, where the average confidence in the predictions exceeds the overall accuracy by a large margin. On the contrary, the accuracy is almost precisely met by the average confidence when the model has been trained with distributional labels. Note, however, that here the accuracy was again calculated by means of the majority vote when modeling the distributional labels. This can lead to underconfidence (as opposed to overconfidence in the one-hot case) because a set of images can be classified correctly even with relatively low average confidence if the label distributions are multimodal.

V. Discussion

Generally speaking, for every deep learning model it is desirable to find a suitable calibration technique in order to avoid under- or overconfidence of the model in its predictions. Especially for the field of remote sensing, this has often been paid little attention to. Traditional methods such as TS [44], [62] or LS [48] train additional hyperparameters using the validation dataset for this purpose. The approach we take here is vastly different, though it achieves similar goals, and is closely related to the works of [52]. We claim that incorporating human uncertainty about the labels helps to overcome not only the poor calibration of the predicted probabilities but also the inability to generalize well on cities not seen during training. Given the special structure of the studied satellite dataset, the label votes cast by ten remote sensing experts are used to form an empirical distribution over the classes that serve as an approximation of the inherent human label uncertainty. This distribution is then embedded as a distributional label for training. Due to its theoretical properties, KL divergence is implemented as a loss function. In order to explicitly measure the impact of human uncertainty, the model architecture, as well as affiliated hyperparameters, remained unchanged.

The questions we pose while forming the distributional labels are whether this approximation (a) is sufficient and (b) has a potential benefit for modeling purposes and downstream tasks. The latter can be partially answered by the findings of this work: we see a direct improvement in terms of calibration regarding the predictive distribution when incorporating the distributional labels for the uncalibrated baseline model. Here, the ECE was on average almost halved. Also with regard to the overall quality of the approximation, the model benefited from the label distributions. An improvement of approximately 10% on average in the cross-entropy could be seen at test time. This holds for the cross-entropy both between the predicted probabilities and the ground truth label distributions and between the probabilities and the one-hot encoded labels. We see huge potential in downstream tasks that benefit from well-calibrated probabilities, for example in the presence of ambiguous or hand-crafted labels.

As for the first question posed, one has to bear in mind that labeling images is a labor-intensive process. This holds especially true when there is a need for experts in the field of remote sensing. A natural question is therefore whether it would be more beneficial to have a larger number of images or individual votes overall. We see this question as a very promising research direction and will leave it for future work. Adding to that, we would like to note that we see potential of the approach also for more fine-grained applications in remote sensing such as pixel-level segmentation. In these tasks, label ambiguity is a much more complex problem, as the transition between classes is not always indicated by a clear line. Similarly, a patch that has a single label in a classification setting can have multiple labels in a segmentation setting. A better-calibrated model with more reliable uncertainty estimates would therefore be particularly helpful to represent the underlying ambiguity. To the best of our knowledge, no large-scale dataset with multiple human annotations exists in this domain. Although this would come with a much higher annotation cost, we see many benefits in creating such a dataset for the aforementioned reasons.

In conclusion, we see clear benefits of the proposed approach, namely stronger generalization performance and better calibrated predictive probabilities, hence more trustworthy resulting uncertainties. At the same time, the approach is currently limited to datasets with multiple annotations per image, which are very costly to obtain. Note also that ideally the labeling should be done by domain experts so that the resulting labels reflect reliable human label uncertainty.

VI. Conclusion

In this work, we examine the impact of employing human uncertainty in the classification process of satellite images into LCZs. We apply our methodology to the So2Sat LCZ42 data, covering several large European cities as well as additional areas from all over the world. Ten expert label votes are supplied for each satellite image, which comprise a notion of uncertainty within the classification task. Forming a label distribution from these votes allows us to directly implement the human uncertainty linked to the voting process into the network training. To do so, we employ the framework of label distribution learning, which enables the model to better adapt to the uncertainty rooted in the labels.

When label uncertainty is incorporated into training, an improvement in the generalization performance of the model can be measured for the remote sensing dataset studied. The overall loss when generalizing to the test data was reduced on average by a margin of approximately 10%. Off-the-shelf calibration methods such as LS or TS lead to improved performance competitive to the uncertainty-guided approach, yet while requiring additional data for hyperparameter tuning. The ECE, a key measure of calibration quality, is almost halved in the majority of experiments by means of the embedded human label uncertainty when no further calibration technique is applied. The improved calibration and generalization performance of the uncertainty-guided approach can be directly converted to a more feasible notion of predictive uncertainty. This is because the predictive probabilities of the network when generalizing to the test data, yield more reliable estimates of the label distributions associated with the test data. We see promising applications in the remote sensing field that could benefit from the explicit incorporation of human (label) uncertainty into the training process.

References

[1] B. Frenay and M. Verleysen, "Classification in the presence of label noise: A survey," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 25, no. 5, pp. 845–869, May 2014.

[2] X. Geng, "Label distribution learning," *IEEE Trans. Knowl. Data Eng.*, vol. 28, no. 7, pp. 1734–1748, Jul. 2016.

[3] S. K. Zhou et al., "A review of deep learning in medical imaging: Imaging traits, technology trends, case studies with progress highlights, and future promises," *Proc. IEEE*, vol. 109, no. 5, pp. 820–838, May 2021.

[4] Y. Dgani, H. Greenspan, and J. Goldberger, "Training a neural network based on unreliable human annotation of medical images," in *Proc. IEEE 15th Int. Symp. Biomed. Imag. (ISBI)*, Apr. 2018, pp. 39–42.

[5] F. Wang, Y. Wu, P. Zhang, Q. Zhang, and M. Li, "Unsupervised SAR image segmentation using ambiguity label information fusion in triplet Markov fields model," *IEEE Geosci. Remote Sens. Lett.*, vol. 14, no. 9, pp. 1479–1483, Sep. 2017.

[6] J. Luo, Y. Wang, Y. Ou, B. He, and B. Li, "Neighbor-based label distribution learning to model label ambiguity for aerial scene classification," *Remote Sens.*, vol. 13, no. 4, p. 755, Feb. 2021.

[7] I. D. Stewart and T. R. Oke, "Local climate zones for urban temperature studies," *Bull. Amer. Meteorological Soc.*, vol. 93, no. 12, pp. 1879–1900, Dec. 2012.

[8] X. X. Zhu et al., "So2Sat LCZ42: A benchmark data set for the classification of global local climate zones [software and data sets]," *IEEE Geosci. Remote Sens. Mag.*, vol. 8, no. 3, pp. 76–89, Sep. 2020.

[9] T. Samsonov and K. Trigub, "Towards computation of urban local climate zones (LCZ) from openstreetmap data," in *Proc. 14th Int. Conf. GeoComputation*, Leeds, U.K., 2017, pp. 4–7.

[10] G. Thomas, A. P. Sherin, S. Ansar, and E. J. Zachariah, "Analysis of urban heat island in Kochi, India, using a modified local climate zone classification," *Proc. Environ. Sci.*, vol. 21, pp. 3–13, Jan. 2014.

[11] P. J. Alexander and G. Mills, "Local climate classification and Dublin's urban heat island," *Atmosphere*, vol. 5, no. 4, pp. 755–774, 2014.

[12] I. D. Stewart, T. R. Oke, and E. S. Krayenhoff, "Evaluation of the 'local climate zone' scheme using temperature observations and model simulations," *Int. J. Climatol.*, vol. 34, no. 4, pp. 1062–1080, Mar. 2014.

[13] J. Quanz, S. Ulrich, D. Fenner, A. Holtmann, and J. Eimermacher, "Micro-scale variability of air temperature within a local climate zone in Berlin, Germany, during summer," *Climate*, vol. 6, no. 1, p. 5, Jan. 2018.

[14] N. G. R. Perera and R. Emmanuel, "A 'local climate zone' based approach to urban planning in Colombo, Sri Lanka," *Urban Climate*, vol. 23, pp. 188–203, Mar. 2018.

[15] E. Lelovics, J. Unger, T. Gál, and C. Gál, "Design of an urban monitoring network based on local climate zone mapping and temperature pattern modelling," *Climate Res.*, vol. 60, no. 1, pp. 51–62, May 2014.

[16] J. Yang et al., "Local climate zone ventilation and urban land surface temperatures: Towards a performance-based and wind-sensitive planning proposal in megacities," *Sustain. Cities Soc.*, vol. 47, May 2019, Art. no. 101487.

[17] G. Mills, J. Ching, L. See, B. Bechtel, and M. Foley, "An introduction to the WUDAPT project," in *Proc. 9th Int. Conf. Urban Climate*, Toulouse, France, 2015, pp. 20–24.

[18] O. Danylo, L. See, B. Bechtel, D. Schepaschenko, and S. Fritz, "Contributing to WUDAPT: A local climate zone classification of two cities in Ukraine," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 9, no. 5, pp. 1841–1853, May 2016.

[19] B. Bechtel et al., "Mapping local climate zones for a worldwide database of the form and function of cities," *ISPRS Int. J. Geo-Inf.*, vol. 4, no. 1, pp. 199–219, Feb. 2015.

[20] B. Bechtel et al., "Generating WUDAPT level 0 data—Current status of production and evaluation," *Urban Climate*, vol. 27, pp. 24–45, Mar. 2019.

[21] O. Brousse, A. Martilli, M. Foley, G. Mills, and B. Bechtel, "WUDAPT, an efficient land use producing data tool for mesoscale models? Integration of urban LCZ in WRF over Madrid," *Urban Climate*, vol. 17, pp. 116–134, Sep. 2016.

[22] C. Yoo, D. Han, J. Im, and B. Bechtel, "Comparison between convolutional neural networks and random forest for local climate zone classification in mega urban areas using Landsat images," *ISPRS J. Photogramm. Remote Sens.*, vol. 157, pp. 155–170, Nov. 2019.

[23] M.-L. Verdonck, A. Okujeni, S. van der Linden, M. Demuzere, R. De Wulf, and F. Van Coillie, "Influence of neighbourhood information on 'local climate zone' mapping in heterogeneous cities," *Int. J. Appl. Earth Observ. Geoinf.*, vol. 62, pp. 102–113, Oct. 2017.

[24] J. Hu, P. Ghamisi, and X. Zhu, "Feature extraction and selection of Sentinel-1 dual-Pol data for global-scale local climate zone classification," *ISPRS Int. J. Geo-Inf.*, vol. 7, no. 9, p. 379, Sep. 2018.

[25] C. Qiu, L. Mou, M. Schmitt, and X. X. Zhu, "Local climate zone-based urban land cover classification from multi-seasonal Sentinel-2 images with a recurrent residual network," *ISPRS J. Photogramm. Remote Sens.*, vol. 154, pp. 151–162, Aug. 2019.

[26] C. Qiu, X. Tong, M. Schmitt, B. Bechtel, and X. X. Zhu, "Multilevel feature fusion-based CNN for local climate zone classification from Sentinel-2 images: Benchmark results on the So2Sat LCZ42 dataset," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 13, pp. 2793–2806, 2020.

[27] C. Qiu, L. Mou, M. Schmitt, and X. X. Zhu, "Fusing multiseasonal Sentinel-2 imagery for urban land cover classification with multibranch residual convolutional neural networks," *IEEE Geosci. Remote Sens. Lett.*, vol. 17, no. 10, pp. 1787–1791, Oct. 2020.

[28] J. Gawlikowski, M. Schmitt, A. Kruspe, and X. X. Zhu, "On the fusion strategies of Sentinel-1 and Sentinel-2 data for local climate zone classification," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, Sep. 2020, pp. 2081–2084.

[29] P. Feng, Y. Lin, G. He, J. Guan, J. Wang, and H. Shi, "A dynamic end-to-end fusion filter for local climate zone classification using SAR and multi-spectrum remote sensing data," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, Sep. 2020, pp. 4231–4234.

[30] X. X. Zhu et al., "The urban morphology on our planet—Global perspectives from space," *Remote Sens. Environ.*, vol. 269, Feb. 2022, Art. no. 112794.

[31] X. Hao et al., "Construction and application of a knowledge graph," *Remote Sens.*, vol. 13, no. 13, p. 2511, 2021.

[32] Y. Li, Z. Zhu, J.-G. Yu, and Y. Zhang, "Learning deep cross-modal embedding networks for zero-shot remote sensing image scene classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 12, pp. 10590–10603, Dec. 2021.

[33] Y. Li, D. Kong, Y. Zhang, Y. Tan, and L. Chen, "Robust deep alignment network with remote sensing knowledge graph for zero-shot and generalized zero-shot remote sensing image scene classification," *ISPRS J. Photogramm. Remote Sens.*, vol. 179, pp. 145–158, Sep. 2021.

[34] J. Gawlikowski et al., "A survey of uncertainty in deep neural networks," 2021, *arXiv:2107.03342*.

[35] E. Hüllermeier and W. Waegeman, "Aleatoric and epistemic uncertainty in machine learning: An introduction to concepts and methods," *Mach. Learn.*, vol. 110, no. 3, pp. 457–506, Mar. 2021.

[36] B. Lakshminarayanan, A. Pritzel, and C. Blundell, "Simple and scalable predictive uncertainty estimation using deep ensembles," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 1–12.

[37] A. Malinin and M. Gales, "Predictive uncertainty estimation via prior networks," in *Proc. 32nd Int. Conf. Neural Inf. Process. Syst.*, 2018, pp. 7047–7058.

[38] M. Sensoy, L. Kaplan, and M. Kandemir, "Evidential deep learning to quantify classification uncertainty," in *Proc. 32nd Int. Conf. Neural Inf. Process. Syst.*, 2018, pp. 3183–3193.

[39] Y. Gal and Z. Ghahramani, "Dropout as a Bayesian approximation: Representing model uncertainty in deep learning," in *Proc. Int. Conf. Mach. Learn.*, 2016, pp. 1050–1059.

[40] C. Blundell, J. Cornebise, K. Kavukcuoglu, and D. Wierstra, "Weight uncertainty in neural network," in *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 1613–1622.

[41] W. J. Maddox, P. Izmailov, T. Garipov, D. P. Vetrov, and A. G. Wilson, "A simple baseline for Bayesian uncertainty in deep learning," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 32, 2019, pp. 13153–13164.

[42] J. Gawlikowski, S. Saha, A. Kruspe, and X. X. Zhu, "An advanced Dirichlet prior network for out-of-distribution detection in remote sensing," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–19, 2022.

[43] M. Rußwurm, M. Ali, X. X. Zhu, Y. Gal, and M. Körner, "Model and data uncertainty for satellite time series forecasting with deep recurrent models," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, Sep. 2020, pp. 7025–7028.

[44] J. Platt, "Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods," *Adv. Large Margin Classifiers*, vol. 10, no. 3 pp. 61–74, Mar. 1999.

[45] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," 2015, *arXiv:1503.02531*.

[46] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, "On calibration of modern neural networks," in *Proc. Int. Conf. Mach. Learn.*, 2017, pp. 1321–1330.

[47] A. S. Mozafari, H. S. Gomes, W. Leão, S. Janny, and C. Gagné, "Attended temperature scaling: A practical approach for calibrating deep neural networks," 2018, *arXiv:1810.11586*.

[48] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2818–2826.

[49] R. Müller, S. Kornblith, and G. Hinton, "When does label smoothing help?" 2019, *arXiv:1906.02629*.

[50] M. Lukasik, S. Bhojanapalli, A. Menon, and S. Kumar, "Does label smoothing mitigate label noise?" in *Proc. Int. Conf. Mach. Learn.*, 2020, pp. 6448–6458.

[51] L. Yuan, F. E. Tay, G. Li, T. Wang, and J. Feng, "Revisiting knowledge distillation via label smoothing regularization," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 3902–3910.

[52] J. Peterson, R. Battleday, T. Griffiths, and O. Russakovsky, "Human uncertainty makes classification more robust," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 9617–9626.

[53] R. M. Battleday, J. C. Peterson, and T. L. Griffiths, "Capturing human categorization of natural images by combining deep networks and cognitive models," *Nature Commun.*, vol. 11, no. 1, pp. 1–14, Oct. 2020.

[54] C.-B. Zhang et al., "Delving deep into label smoothing," *IEEE Trans. Image Process.*, vol. 30, pp. 5984–5996, 2021.

[55] C. E. Brodley and M. A. Friedl, "Identifying mislabeled training data," *J. Artif. Intell. Res.*, vol. 11, pp. 131–167, Aug. 1999.

[56] D. Rolnick, A. Veit, S. Belongie, and N. Shavit, "Deep learning is robust to massive label noise," 2017, *arXiv:1705.10694*.

[57] C. García Rodríguez, J. Vitrià, and O. Mora, "Uncertainty-based human-in-the-loop deep learning for land cover segmentation," *Remote Sens.*, vol. 12, no. 22, p. 3836, Nov. 2020.

[58] M. Rußwurm, S. Wang, and D. Tuia, "Humans are poor few-shot classifiers for Sentinel-2 land cover," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, Jul. 2022, pp. 4859–4862.

[59] I. Stewart, "Local climate zones: Origins, development, and application to urban heat island studies," in *Proc. Annu. Meeting Amer. Assoc. Geographers*, Seattle, WA, USA, 2011, pp. 12–16.

[60] J. Ching et al., "WUDAPT: An urban weather, climate, and environmental modeling infrastructure for the anthropocene," *Bull. Amer. Meteorological Soc.*, vol. 99, no. 9, pp. 1907–1924, Sep. 2018.

[61] M. H. DeGroot and S. E. Fienberg, "The comparison and evaluation of forecasters," *J. Roy. Stat. Soc., Ser. D*, vol. 32, nos. 1–2, pp. 12–22, 1983.

[62] A. Niculescu-Mizil and R. Caruana, "Predicting good probabilities with supervised learning," in *Proc. 22nd Int. Conf. Mach. Learn.*, 2005, pp. 625–632.

[63] J. Nixon, M. W. Dusenberry, L. Zhang, G. Jerfel, and D. Tran, "Measuring calibration in deep learning," in *Proc. CVPR Workshops*, 2019, vol. 2, no. 7, pp. 1–4.

[64] F. Chollet. (2015). *Keras*. [Online]. Available: https://keras.io

[65] M. Hollemans. (2020). *Reliability Diagrams*. Accessed: Jan. 10, 2022. [Online]. Available: https://github.com/hollance/reliability-diagrams

Christoph Koller (Graduate Student Member, IEEE) received the bachelor's degree in mathematics from Technical University Munich (TUM), Munich, Germany, in 2017, and the master's degree in statistics from LMU Munich, Munich, in 2020. He is currently pursuing the Ph.D. degree with the Chair of Data Science in Earth Observation, Remote Sensing Institute, German Aerospace Center (DLR), Weßling, Germany.

His research interests include uncertainty quantification, calibration, and out-of-distribution detection for machine and deep learning models in the context of remote sensing image classification.

Göran Kauermann received the diploma degree in economic mathematics and the Doctorate of Natural Sciences (Dr.rer.nat.) degree in statistics from Technical University Berlin, Berlin, Germany, in 1991 and 1994, respectively.

He was a Visiting Scholar/Professor at the Department of Statistics, University of Chicago, Chicago, IL, USA, and the Department of Statistics, University of New South Wales, Sydney, NSW, Australia, in 1996 and 2005, respectively. He took the role of an Assistant Professor at the Ludwig-Maximilians-University (LMU) Munich, Munich, Germany, from 1998 to 2000. After his habilitation in statistics in 2000, he served as a Senior Lecturer at the Department of Statistics, the University of Glasgow, Glasgow, U.K., before becoming a Full Professor of statistics at the Department of Economics and Business Administration, University of Bielefeld, Bielefeld, Germany, in 2003. He served as the Dean of the Faculty of Mathematics, Computer Science and Statistics, LMU Munich, from 2019 to 2021. Since 2011, he holds the Chair of Statistics—in Economics, Business Administration and Social Sciences, Department of Statistics, LMU Munich. He has been the Speaker of the Elite Master Program Data Science, LMU Munich, since 2016. His research interests include semi- and nonparametric models, generalized linear models and network data analysis.

Prof. Kauermann served as the Chair of the Deutsche Arbeitsgemeinschaft Statistik ("German Working Group for Statistics") (DAG) from 2005 to 2013 and was elected as a reviewer for Statistics and Economics at the DAG. He served as an (Associate) Editor of the *AStA—Advances in Statistical Analysis, Statistical Modelling*: An International Journal, the *Biometrical Journal*, and the *Journal of the Royal Statistical Society, Series C (Applied Statistics)*.

Xiao Xiang Zhu (Fellow, IEEE) received the M.Sc., Dr.-Ing., and Habilitation degrees in signal processing from the Technical University of Munich (TUM), Munich, Germany, in 2008, 2011 and 2013, respectively.

She was the founding Head of the Department "EO Data Science," Remote Sensing Technology Institute, German Aerospace Center (DLR), Weßling, Germany. From 2019 to 2022, she was a Co-Coordinator of the Munich Data Science Research School (www.mu-ds.de) and the Head of the Helmholtz Artificial Intelligence—Research Field "Aeronautics, Space and Transport." Since May 2020, she has been the PI and the Director of the International Future AI Lab "AI4EO—Artificial Intelligence for Earth Observation: Reasoning, Uncertainties, Ethics and Beyond," Munich, Germany. Since October 2020, she has been the Director of the Munich Data Science Institute (MDSI), TUM. She is the Chair Professor for Data Science in Earth Observation, TUM. She was a Guest Scientist or Visiting Professor at the Italian National Research Council (CNR-IREA), Naples, Italy; Fudan University, Shanghai, China; the University of Tokyo, Tokyo, Japan; and the University of California, Los Angeles, CA, USA, in 2009, 2014, 2015, and 2016, respectively. She is currently a Visiting AI Professor at ESA's Phi-Lab, Frascati, Italy. Her main research interests are remote sensing and Earth observation, signal processing, machine learning and data science, with their applications in tackling societal grand challenges, e.g., global urbanization, UN's SDGs and climate change.

Dr. Zhu has been a member of Young Academy (Junge Akademie/Junges Kolleg) at the Berlin-Brandenburg Academy of Sciences and Humanities, the German National Academy of Sciences Leopoldina, and the Bavarian Academy of Sciences and Humanities. She serves on the scientific advisory board in several research organizations, among others the German Research Center for Geosciences (GFZ, from 2020 to 2023) and Potsdam Institute for Climate Impact Research (PIK). She is an Associate Editor of IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING, *Pattern Recognition* and serves as the Area Editor responsible for special issues of *IEEE Signal Processing Magazine*.