living planet symposium | BONN 23–27 May 2022

TAKING THE PULSE OF OUR PLANET FROM SPACE

AI4EO: Learning From Human Uncertainty

Christoph Koller[1], Göran Kauermann[2], Xiao Xiang Zhu[1]

[1]German Aerospace Center & Technical University of Munich, [2]Ludwig-Maximilians-Universität München

26.05.2022

Upcoming Topics

I. Uncertainty in Remote Sensing

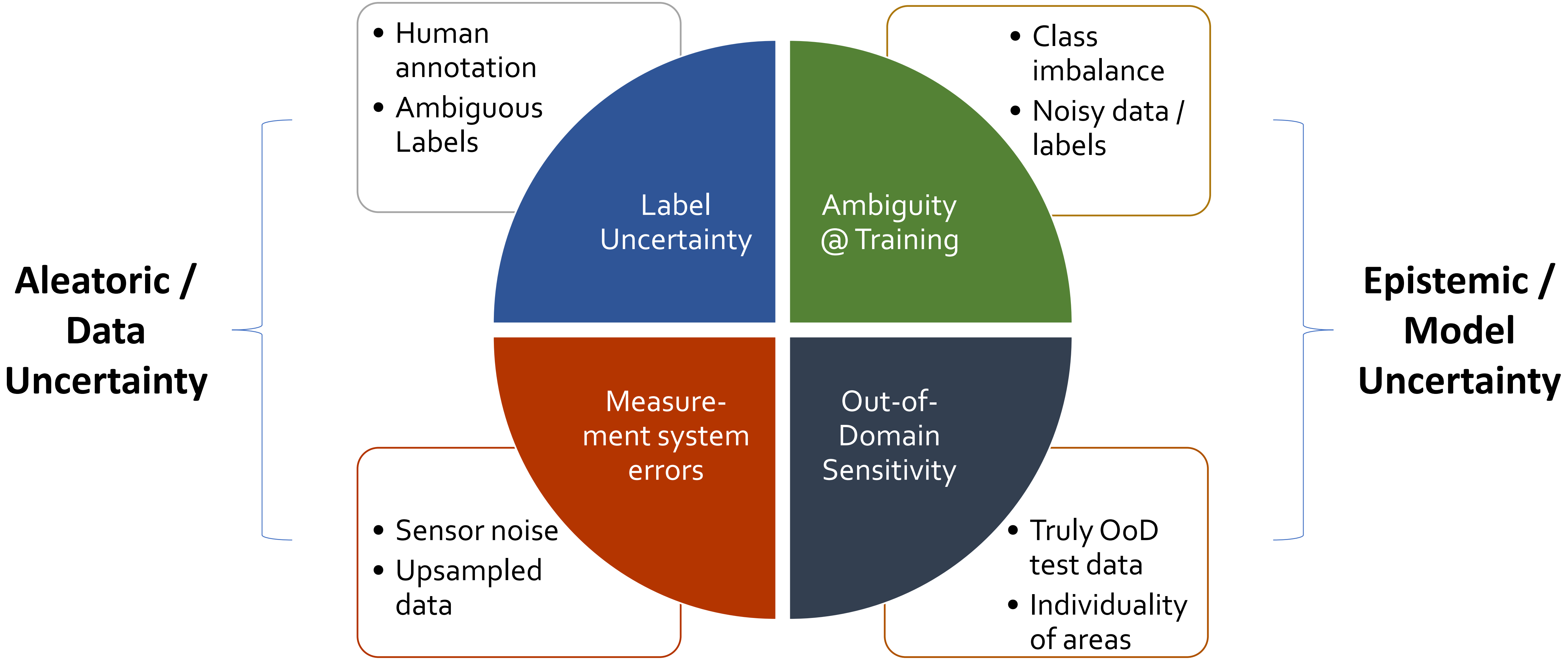
II. Human Label Uncertainty

III. Embedding Uncertainty into Training

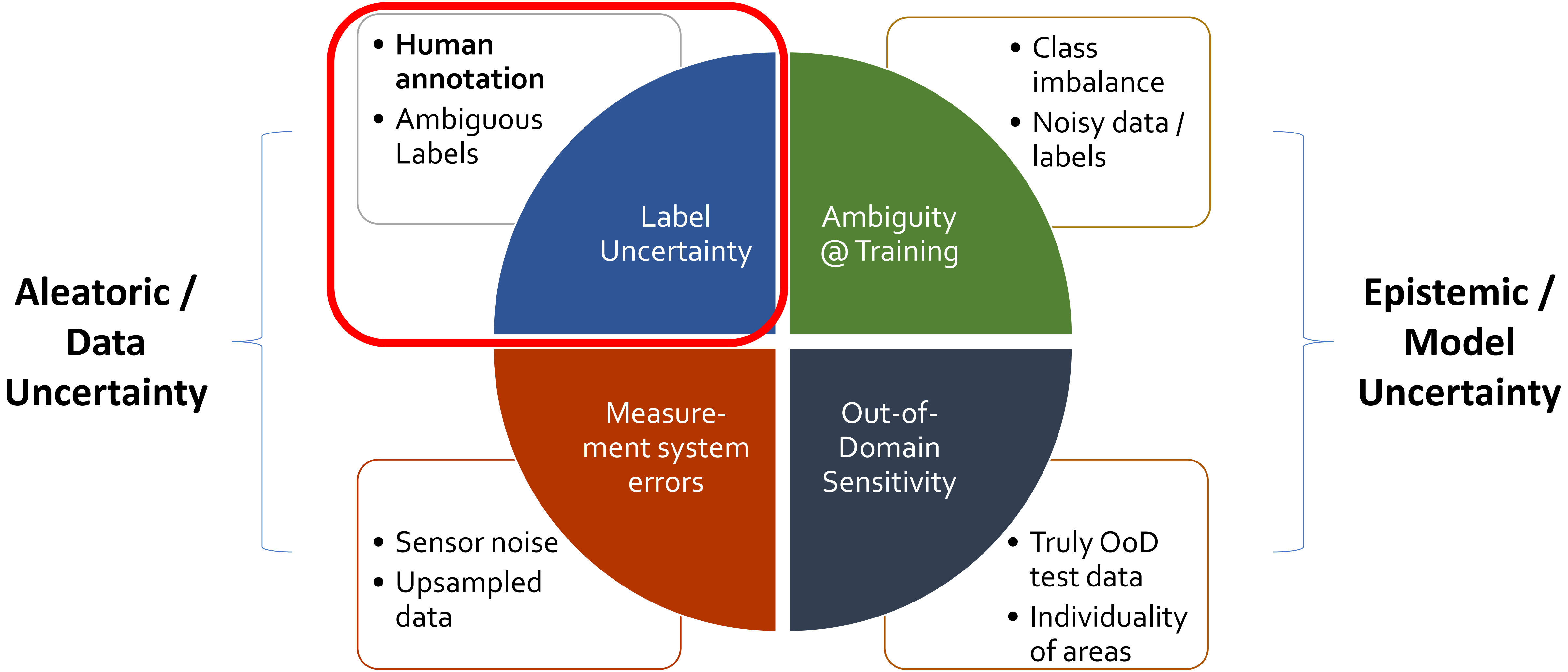
IV. Findings

V. Conclusion

I. Uncertainty in Remote Sensing (Classification)



I. Uncertainty in Remote Sensing (Classification)



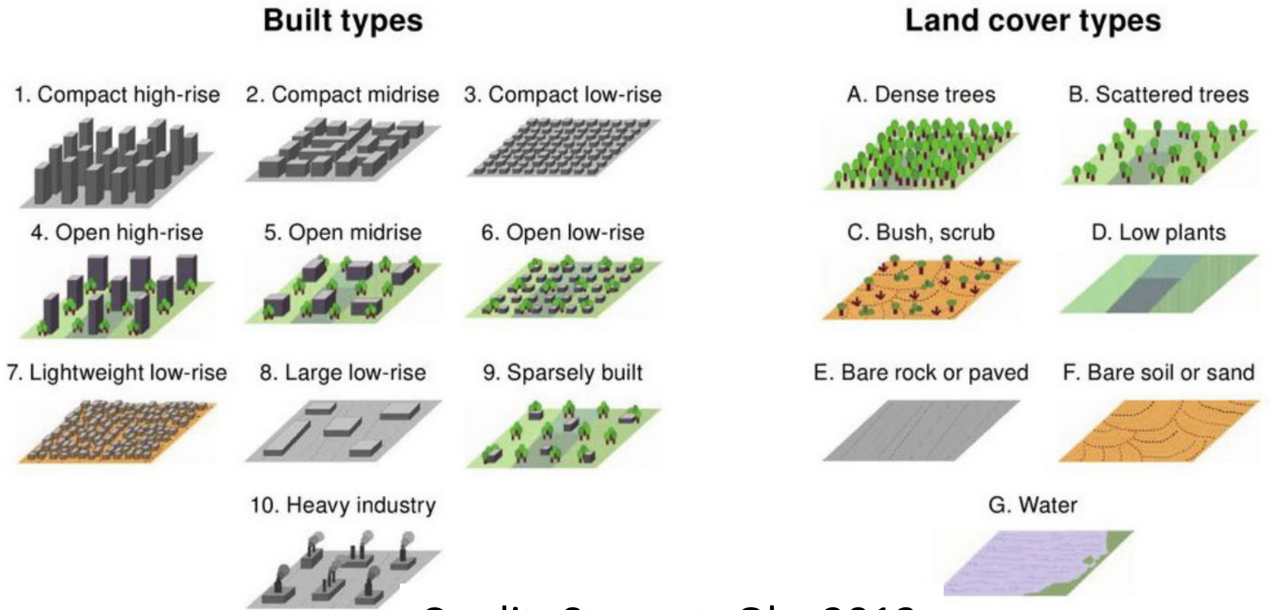
II. Human Label Uncertainty

Studied Data Set:

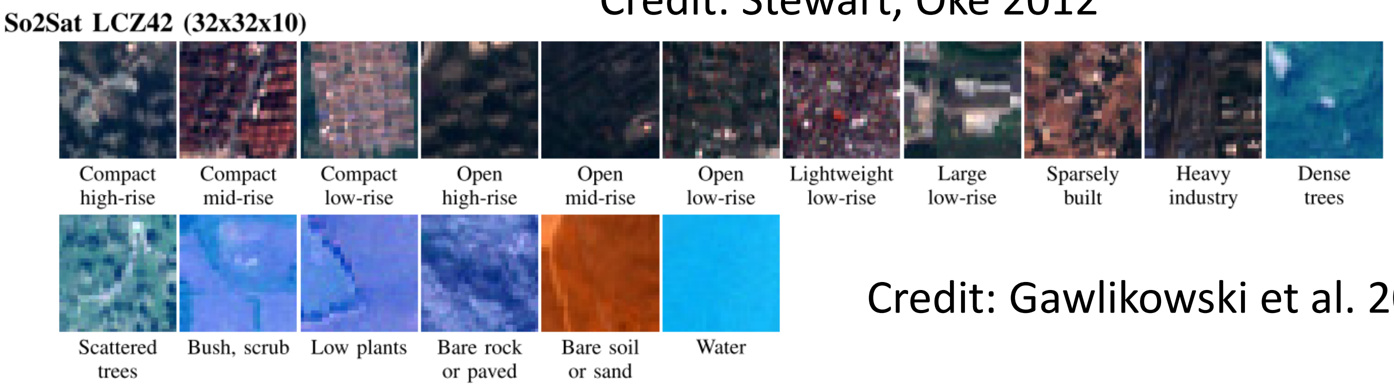
So2Sat LCZ42 (Zhu et al. 2020)

- Sentinel 1&2 patches, 32*32 pixels
- 42 global urban conglomerates
- Local Climate Zones (LCZ) classification scheme (17 classes)
- Ca. 400k labeled patches,

  250k with multiple labels
- Geographical train / test split



Credit: Stewart, Oke 2012



Credit: Gawlikowski et al. 2022

II. Human Label Uncertainty

Label Evaluation Study:

For a subset (European cities + addon areas), **J=10** different voters independently labeled each image patch:

$$V_1^{(i)}, \ldots, V_J^{(i)} \;,\; V_j^{(i)} \in \{1, \ldots, K\} \; \forall\, i = 1, \ldots, n \,. \quad \Rightarrow \quad Y_k^{(i)} = \sum_j \mathbb{1}_{\{V_j^{(i)} = k\}}$$

individual votes observed — ***counts for class k of image i***

One-Hot encoded labels:

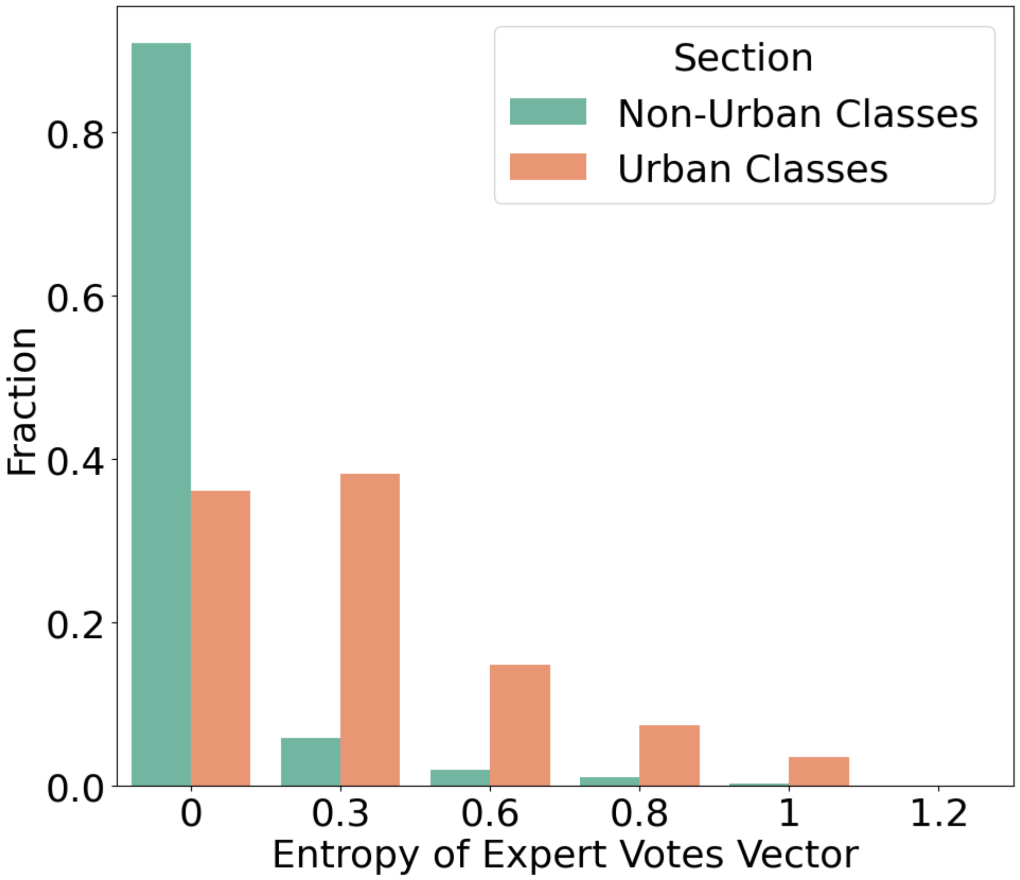
$$Y_{max}^{(i)} := \max_j Y_j^{(i)}$$

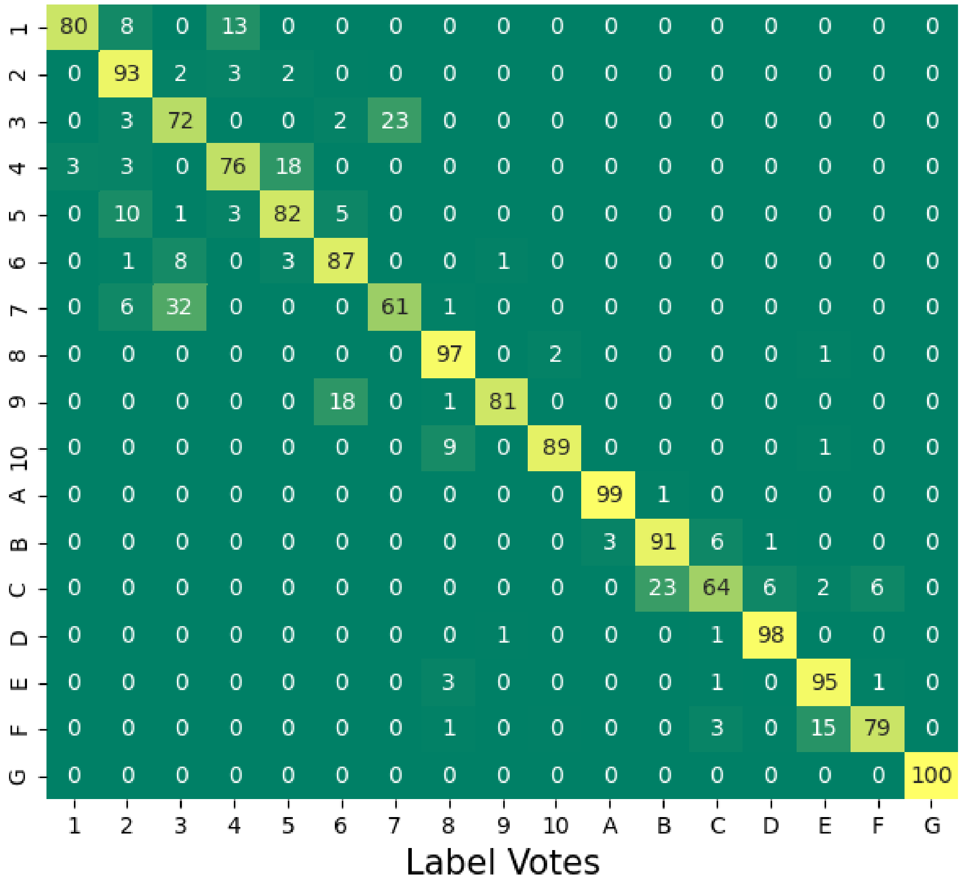
Distributional encoded labels:

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

M = # of votes

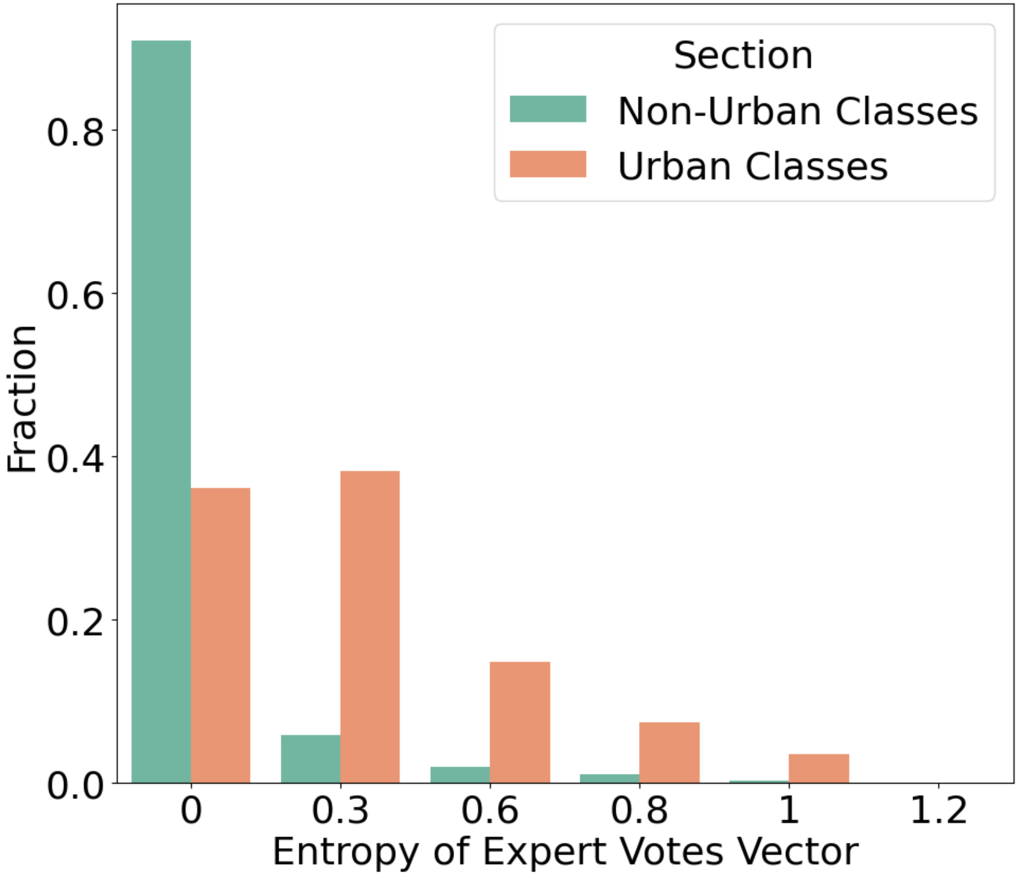
II. Human Label Uncertainty

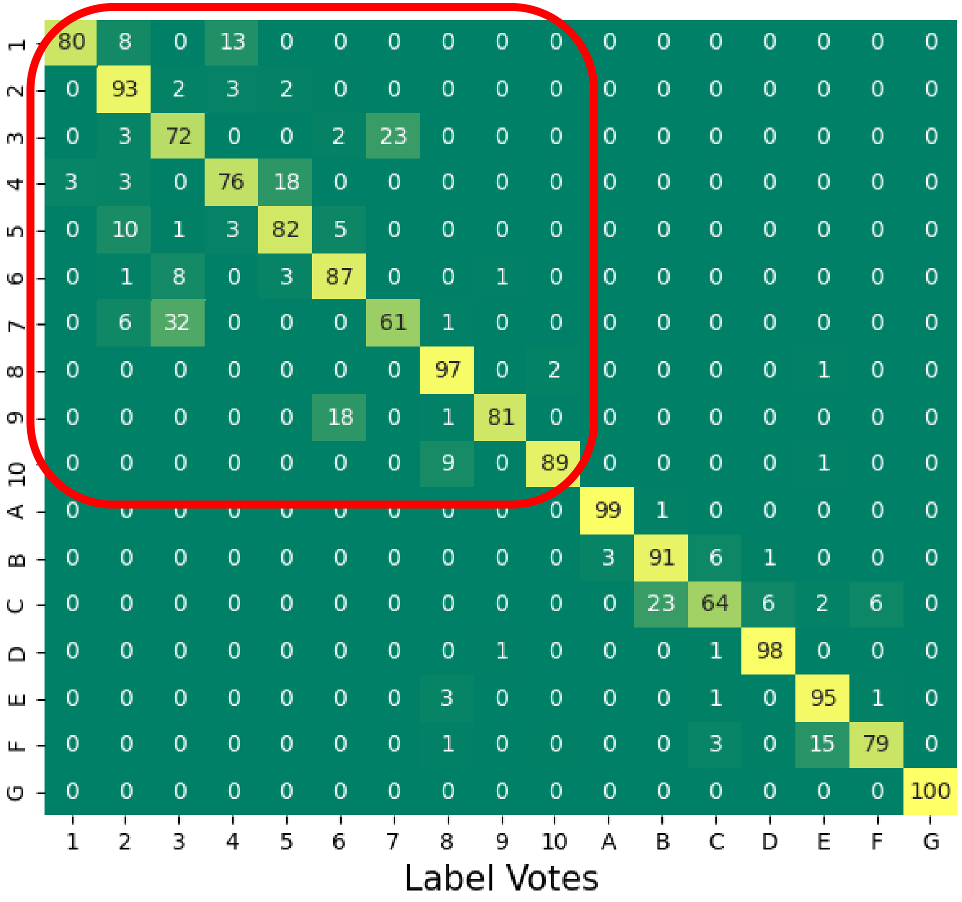




Koller et al. 2022, Submitted

II. Human Label Uncertainty



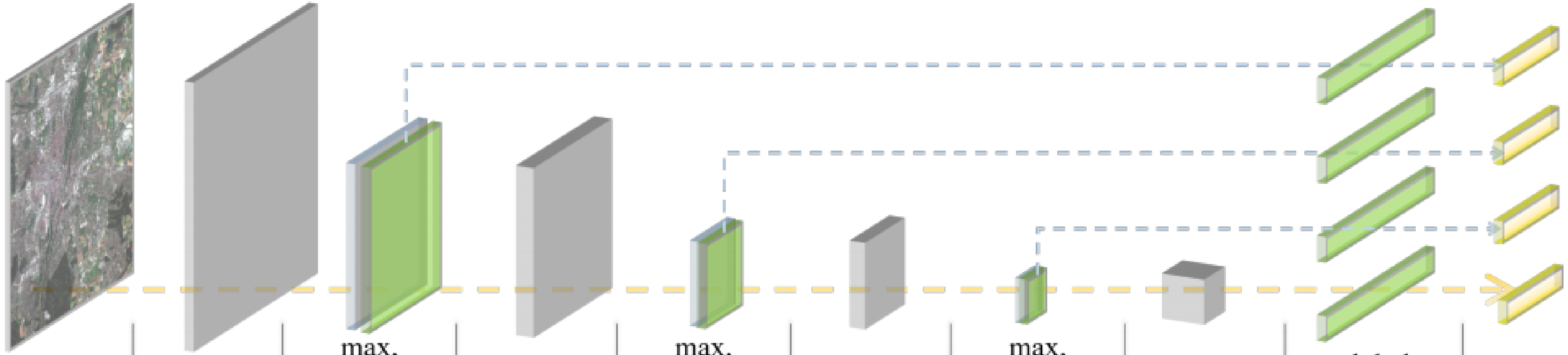


Koller et al. 2022, Submitted

II. Human Label Uncertainty

Model in use:

Sen2LCZ (Qiu et al. 2020), CNN-based Deep Learning model superior over many commonly used image classification models for LCZ42 image classification



input	Conv(3×3) × N	max, average pooling concatenated	Conv(3×3) × N	max, average pooling concatenated	Conv(3×3) × N	max, average pooling concatenated	Conv(3×3) × N	global average pooling	dense	average of softmax
$h \times w \times 10$	$h \times w \times f$	$\frac{h}{2} \times \frac{w}{2} \times 2f$	$\frac{h}{2} \times \frac{w}{2} \times 2f$	$\frac{h}{4} \times \frac{w}{4} \times 4f$	$\frac{h}{4} \times \frac{w}{4} \times 4f$	$\frac{h}{8} \times \frac{w}{8} \times 8f$	$\frac{h}{8} \times \frac{w}{8} \times 8f$	$1 \times 1 \times 8f$, …	1×17, …	1×17

Fig. 2: Architecture of Sen2LCZ-MF. The three light blue lines correspond to the part of multi-level feature fusion. Note that all convolutional layers are followed by batch normalization. The depth $D = 4N + 1$, and the width depends on the filter number of the first block, which is doubled for each subsequent block.

III. Embedding Uncertainty Into Training

Training with distributional labels (Kullback-Leibler (KL) divergence):

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

A Priori Calibration

Post-Hoc Calibration

Label Smoothing:

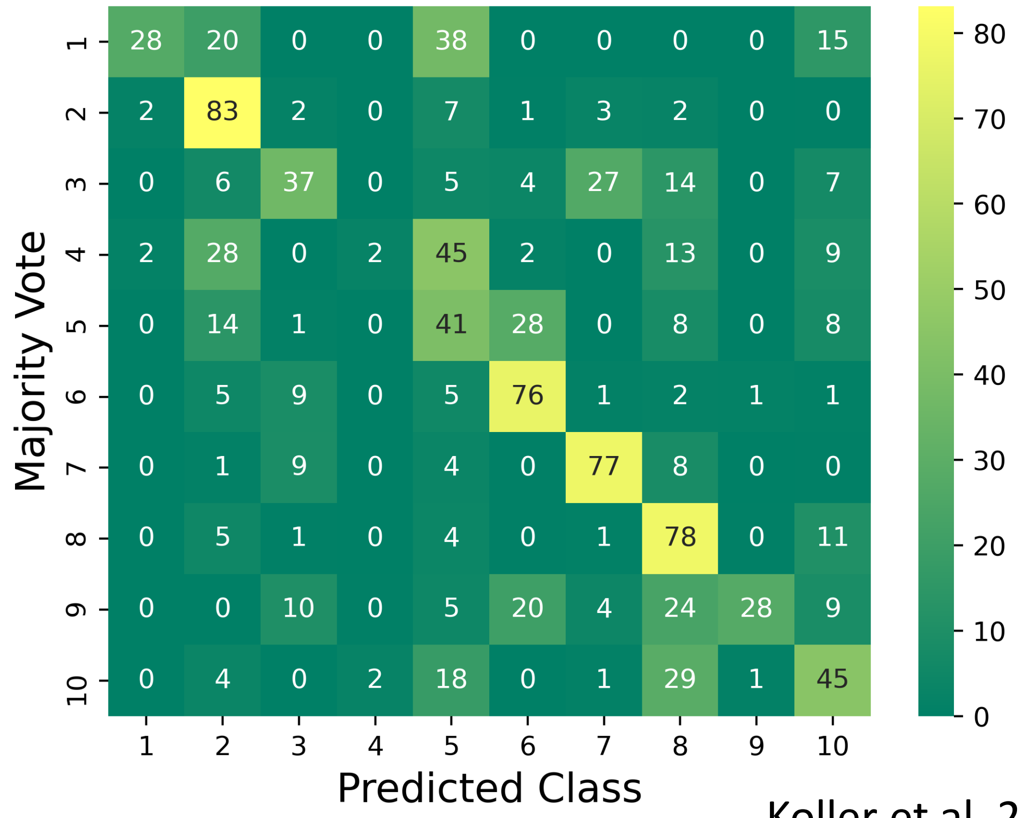
$$y_{\text{smoothed}}^{(i)} = \alpha \cdot u_K + (1 - \alpha) \cdot y^{(i)}$$

Temperature Scaling:

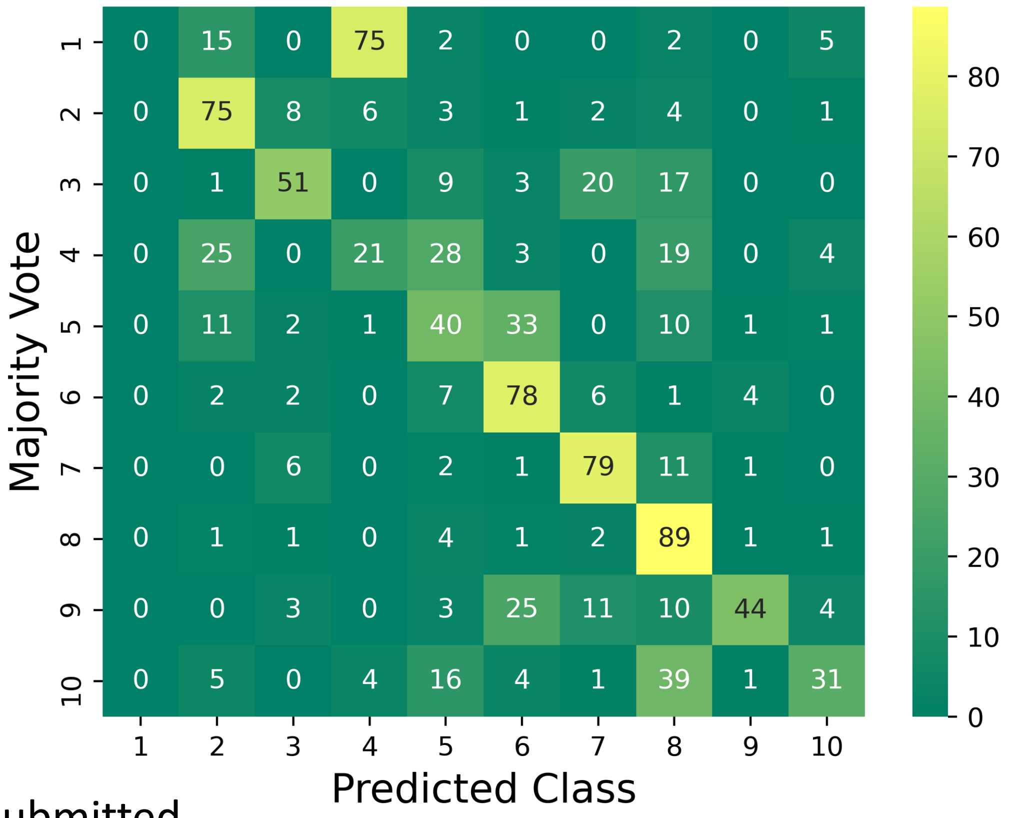
$$\text{softmax}(z^{(i)}) = \frac{\exp(z^{(i)}/T)}{\sum_k \exp(z_k^{(i)}/T)} = \left(\frac{\exp(z_1^{(i)}/T)}{\sum_k \exp(z_k^{(i)}/T)}, \ldots, \frac{\exp(z_K^{(i)}/T)}{\sum_k \exp(z_k^{(i)}/T)}\right)$$

III. Embedding Uncertainty Into Training

Training with One-Hot Encdoded Labels:



Training with Distributional Labels:



Koller et al. 2022, Submitted

III. Embedding Uncertainty Into Training

Table 1: Performance scores on the test set. Accuracy and other related measures were derived with respect to the majority vote. All scores are averaged over 5 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	$\mathbf{50.3 \pm 3.5}$	69.9 ± 2.0	59.4 ± 2.7
Distr.	67.0 ± 2.2	45.8 ± 3.9	$\mathbf{71.0 \pm 0.5}$	58.8 ± 2.4
+ LS	$\mathbf{68.6 \pm 2.3}$	43.4 ± 6.1	69.7 ± 2.1	$\mathbf{60.4 \pm 2.8}$

Koller et al. 2022, Submitted

III. Embedding Uncertainty Into Training

Calibration metrics considered:

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

$$\mathrm{MCE} = \max_{m} |\,\mathrm{acc}(I_m) - \mathrm{conf}(I_m)\,|$$

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

Guo et al. 2017, Nixon et al. 2019

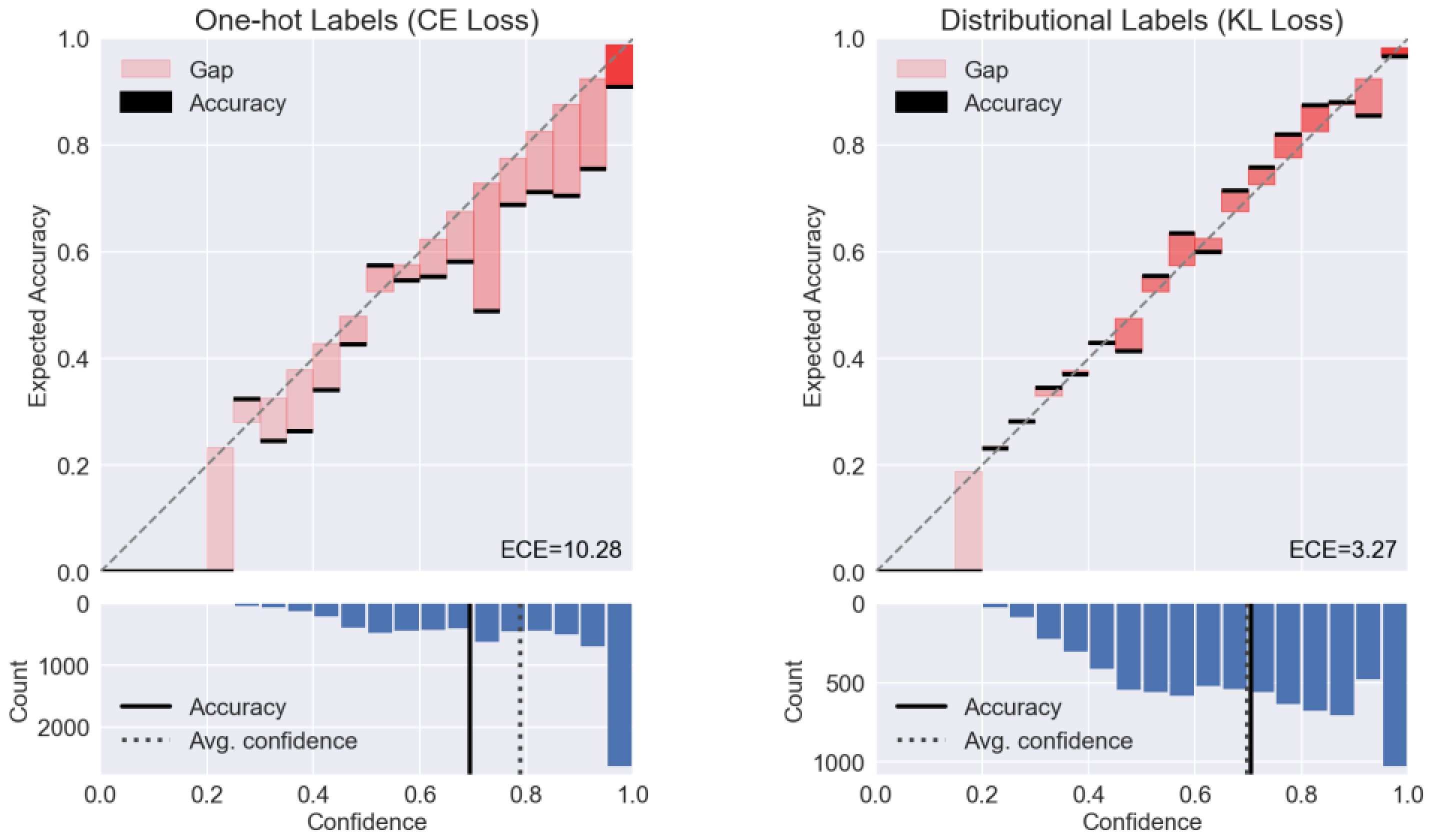
Generalization metrics considered:

$$\mathrm{CE}(y_{\mathrm{max}}^{(i)}, p_\theta(y^{(i)}|x)) = -\sum_{k=1}^{K} y_{\mathrm{max},k}^{(i)} \log(p_\theta(y^{(i)} = k|x)$$

$$\mathrm{CE}(y_{\mathrm{distr}}^{(i)}, p_\theta(y^{(i)}|x)) = -\sum_{k=1}^{K} y_{\mathrm{distr},k}^{(i)} \log(p_\theta(y^{(i)} = k|x)$$

Peterson et al. 2019

III. Embedding Uncertainty Into Training



(a) One-Hot Encoding

(b) Label Distribution Encoding

Koller et al. 2022, Submitted

III. Embedding Uncertainty Into Training

Table 2: 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.

	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

Koller et al. 2022, Submitted

IV. Findings

- Training **objective** with uncertainty is same at the core, but misclassifications aren‘t punished as hard as usual during training

- Natural **calibration** occurs when using labels with distributional form, competitive to off-the-shelf methods which require additional data for hyperparameter tuning

- **Generalization** performance in terms of cross entropy also increases, even when generalizing to one-hot encoded test data

V. Conclusion

- **Label uncertainty** is often neglected in Earth Observation models
- Valuable information is hidden from model if uncertainty associated with the labels is not considered
- **Incorporation** of existing label uncertainty is seamless and requires little changes
- **Generalization** increases, **calibration** occurs naturally

References

- Gawlikowski, J., Saha, S., Kruspe, A., and Zhu, X. X. An advanced dirichlet prior network for out-of-distribution detection in remote sensing. IEEE Transactions on Geoscience and Remote Sensing, 2022.
- Guo, C., Pleiss, G., Sun, Y., and Weinberger, K. Q. On calibration of modern neural networks. In International Conference on Machine Learning, pp. 1321–1330. PMLR, 2017.
- Nixon, J., Dusenberry, M. W., Zhang, L., Jerfel, G., and Tran, D. Measuring calibration in deep learning. In CVPR Workshops, volume 2, 2019.
- Peterson, J. C., Battleday, R. M., Griffiths, T. L., and Russakovsky, O. Human uncertainty makes classification more robust. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 9617–9626, 2019.
- Qiu, C., Tong, X., Schmitt, M., Bechtel, B., and Zhu, X. X. Multilevel feature fusion-based cnn for local climate zone classification from sentinel-2 images: Benchmark results on the so2sat lcz42 dataset. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 13:2793–2806, 2020b.
- Stewart, I. D. and Oke, T. R. Local climate zones for urban temperature studies. Bulletin of the American Meteorological Society, 93(12):1879–1900, 2012.
- Zhu, X. X., Hu, J., Qiu, C., Shi, Y., Kang, J., Mou, L., Bagheri, H., Haberle, M., Hua, Y., Huang, R., et al. So2sat lcz42: a benchmark data set for the classification of global local climate zones [software and data sets]. IEEE Geoscience and Remote Sensing Magazine, 8(3):76–89, 2020

Thank you for your attention!