Can Land Cover Classification Models Benefit From Distance-Aware Architectures?

Christoph Koller, *Graduate Student Member, IEEE*, Peter Jung, and Xiao Xiang Zhu, *Fellow, IEEE*

***Abstract*— The quantification of predictive uncertainties helps to understand where the existing models struggle to find the correct prediction. A useful quality control tool is the task of detecting out-of-distribution (OOD) data by examining the model's predictive uncertainty. For this task, deterministic single forward pass frameworks have recently been established as deep learning models and have shown competitive performance in certain tasks. The unique combination of spectrally normalized weight matrices and residual connection networks with an approximate Gaussian process (GP) output layer can here offer the best trade-off between performance and complexity. We utilize this framework with a refined version that adds spectral batch normalization and an inducing points approximation of the GP for the task of OOD detection in remote sensing image classification. This is an important task in the field of remote sensing, because it provides an evaluation of how reliable the model's predictive uncertainty estimates are. By performing experiments on the benchmark datasets Eurosat and So2Sat LCZ42, we can show the effectiveness of the proposed adaptions to the residual networks (ResNets). Depending on the chosen dataset, the proposed methodology achieves OOD detection performance up to 16% higher than previously considered distance-aware networks. Compared with other uncertainty quantification methodologies, the results are on the same level and exceed them in certain experiments by up to 2%. In particular, spectral batch normalization, which normalizes the batched data as opposed to normalizing the network weights by the spectral normalization (SN), plays a crucial role and leads to performance gains of up to 3% in every single experiment. For reproducibility, the code can be found here: https://github.com/ChrisKo94/DUE_Land_Cover.**

***Index Terms*— Distance awareness, land cover classification, out-of-distribution (OOD), spectral normalization (SN), uncertainty quantification.**

Manuscript received 8 November 2023; revised 3 January 2024; accepted 23 January 2024. Date of publication 10 April 2024; date of current version 12 April 2024. This work was supported in part by the Helmholtz Association through the Joint Research School "Munich School for Data Science (MUDS)"; in part by the German Federal Ministry of Education and Research (BMBF) through the Framework of the International Future AI Lab "Artificial Intelligence for Earth Observation (AI4EO): Reasoning, Uncertainties, Ethics and Beyond" under Grant 01DD20001; and in part by the German Federal Ministry of Economics and Technology through the Framework of the "National Center of Excellence ML4Earth" under Grant 50EE2201C. *(Corresponding author: Xiao Xiang Zhu.)*

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

Peter Jung is with the Institute of Optical Sensor Systems, German Aerospace Center, 12489 Berlin, Germany, and also with the Communications and Information Theory Group, Technical University of Berlin, 10623 Berlin, Germany.

Xiao Xiang Zhu is with the Chair of Data Science in Earth Observation, Technical University Munich, 80333 Munich, Germany, and also with the Munich Center for Machine Learning, 80992 Munich, Germany (e-mail: xiaoxiang.zhu.ieee@gmail.com).

Digital Object Identifier 10.1109/LGRS.2024.3375370

I. Introduction

OVER the last decade, the urgency for having a notion of uncertainty along with a model's prediction has arisen in many research communities. Many different techniques have been invented and developed by the machine and deep learning community [1]. Out of the developed approaches, many uncertainty quantification techniques solely focus on cleverly transforming the output of one or multiple deep learning models. In their work, Lakshminarayanan et al. [2] train multiple neural networks in parallel and achieve state-of-the-art performance on many computer vision and regression tasks. The combination of the individual model predictions also yields an expressive notion of model uncertainty. Another famous example is given by Monte Carlo dropout [3], where the model architecture is left unchanged during inference time with the exception of leaving the existing dropout mechanism enabled. Again, the resulting predictions give a sense of model uncertainty accompanying the prediction itself.

More methodologically advanced approaches are covered exemplarily by the broad field of Bayesian neural networks [4]. Here, the individual weights of a neural network architecture are not treated as fixed parameters, but rather as distributions. While tracing the exact likelihood of such networks becomes infeasible and computationally expensive, uncertainties can be modeled much more fine-grained by averaging over the posterior weight distributions. Many extensions of the core idea have been introduced, for example, a sampling-free variational inference approach [5], or a method for inference on the subnetwork level [6].

Contrastive to the Bayesian approaches, deterministic methods do not place distributional assumptions over parts of the network and do not rely on sampling from distributions or predictions. Especially, for the downstream task of detecting out-of-distribution (OOD) data, this class of networks has been shown to yield competitive performance [7]. The approaches can generally be divided into generative and discriminative ones. Generative approaches utilize an explicit likelihood from a generative model, whereas discriminative approaches use regularized predictions to form their predictive uncertainties.

Regarding the applicability and usefulness of such uncertainty quantification approaches in the field of remote sensing, little work has been done so far. Landgraf et al. [8] modified the loss function to incorporate predictive uncertainties during training and showed convincing results on semantic segmentation tasks. Regarding OOD detection, Gawlikowski et al. [9] used a Dirichlet prior network that yielded strong performance on a range of remote sensing image classification datasets

with respect to OOD detection. The little attention paid to uncertainty quantification in deep neural networks for remote sensing motivates this work, which is focused on employing so-called distance-aware network architectures by ensuring Lipschitz continuity of the network mapping from the input to the output space.

II. Methodology

We now present the general framework of distance-aware uncertainty quantification as initially presented by Liu et al. [10] and refined by van Amersfoort et al. [11]. The distance awareness here refers to the model's capacity to adequately project distances in the input space onto the output space or prediction of the model. As explained in [10], this is equivalent to the existence of a notion of uncertainty u: $\mathcal{F}(\mathcal{X}) \to \mathbb{R}_+$ satisfying this adequate projection property. This uncertainty measure operates on the output space of the network $\mathcal{F}$, and $\mathcal{X}$ here describes the input space of the network. For a new point $x \in \mathcal{X}$ and the training data $\mathcal{X}_{\text{train}} \subset \mathcal{X}$, we first denote a fitting metric by d: $\mathcal{X} \times \mathcal{X} \to \mathbb{R}_+$, e.g., of the form $d_X(x_1, x_2) = \|x_1 - x_2\|_{\mathcal{X}}, x_1, x_2 \in \mathcal{X}$ with an arbitrary yet suitable norm. With this metric, for the network to be distance-aware, the uncertainty measure u needs to satisfy

$$\forall x \in \mathcal{X}\colon u(x) = v\left(\mathbb{E}_{x' \sim \mathcal{X}_{\text{train}}} d\left(x, x'\right)\right) \tag{1}$$

for a given monotonically increasing function v: $\mathbb{R}_+ \to \mathbb{R}_+$.

A. Hidden Mapping Distance Preservation

Consider a neural network classifier, which yields the well-known unscaled prediction $\text{logit}(x) = g \circ h(x)$, where h describes the hidden mapping, i.e., all layers up until the last hidden representation and g describes the output mapping onto the label space. In this scenario, the distance awareness can be separately defined for both mappings. Regarding the hidden mapping h, Lakshminarayanan et al. [10] state that the distance awareness is here equivalent to fulfilling a bi-Lipschitz constraint of h. This means that there need to exist constants $0 < L_1 \leq 1 \leq L_2$, such that

$$L_1 \cdot d_X\left(x, x'\right) \leq d_X\left(h(x), h\left(x'\right)\right) \leq L_2 \cdot d_X\left(x, x'\right) \tag{2}$$

holds for all $x, x' \in \mathcal{X}$.

Several regularization techniques have been established to enforce this constraint. For example, the constraint can be fulfilled for differentiable functions if the norms of the gradients are bounded. This can be enforced by adding a penalty term $\mathcal{P}_L$ (the subscript L stands for loss) of the form

$$\mathcal{P}_L := \lambda \cdot \left[\|\nabla_x f_\theta(x)\|_2^2 - L_2\right]^2 \tag{3}$$

to the loss function, where $\nabla_x f_\theta(x)$ denotes the gradient of the neural network mapping with respect to x, L_2 is the earlier mentioned upper Lipschitz constant, and λ denotes a hyperparameter steering the impact of the gradient penalty on the loss. This procedure is called two-sided gradient penalty and has been successfully applied to generative adversarial networks (GANs) [12] and radial basis function (RBF) networks [13]. Alternatively, the Lipschitz constant L_2 can be controlled by normalizing the weight matrices of every layer. Specifically, the Lipschitz constant of a fully connected layer $f_k(x) = Wx$ can be represented via the following:

$$\begin{aligned} \|f\|_{\text{Lip}} &= \sup \sigma(\nabla f_k(x)), \quad \text{where} \\ \sigma(A) &:= \max_{x: x \neq 0} \frac{\|Ax\|_2}{\|x\|_2} = \max_{\|x\|_2 \leq 1} \|Ax\|_2. \end{aligned} \tag{4}$$

This spectral norm is equivalent to the largest singular value of A. Since most common activation functions have Lipschitz norm 1 (or predefined L), the Lipschitz constant of a neural network can be computed by the multiplication of the Lipschitz constants of the individual layers. Miyato et al. [14] made use of the normalization of every layer weight matrix via spectral norm (pointwise division) with respect to the l_2 norm for GANs. This proved to be superior over other weight clipping or regularization techniques (e.g., [15], [16]) and over the gradient penalty earlier introduced.

B. ResNet Hidden Mapping

A key neural network architecture for remote sensing image classification is the class of residual networks (ResNets). Since the now following approach is tailored toward ResNets, it is highly relevant for the remote sensing field. For ResNets [17], there exists a more profound way to control the bi-Lipschitz constraint of the hidden mapping h. The hidden mapping is here given via $h(x) = h_l \circ h_{l-1} \circ \cdots \circ h_1(x)$, where $h_j(x) = x + \gamma_j(x)$, $j = 1, \ldots, l$, denotes the individual residual blocks. If and only if all of the residual mappings $\gamma_j(x)$ $j = 1, \ldots, l$ of the residual blocks are α-Lipschitz for $0 < \alpha < 1$ on $\mathcal{X}$, the bi-Lipschitz condition in (2) holds. Because the concatenation of α-Lipschitz functions is again α-Lipschitz [18], the individual Lipschitz conditions are sufficient for the bi-Lipschitzness of the entire hidden mapping h. For the l residual blocks contained in h, we are then left with the following Lipschitz constants:

$$0 < L_1 = (1 - \alpha)^l < 1 < L_2 = (1 + \alpha)^l < \infty. \tag{5}$$

Given the dataset spans over all of $\mathbb{R}^n$ with respect to the Euclidean norm, the condition $\alpha < 1$ of the bi-Lipschitzness is equivalent to bounding the spectral norm (i.e., largest singular value λ) of the weight matrix W_j of γ_j. In more realistic settings, the spectral norm needs to be approximated. The power iteration presented by Behrmann et al. [20] is most often used in practice for this approximation due to its fast convergence. In order to bound the spectral norm for each layer $j = 1, \ldots, l$, the respective weight matrix of the convolutional layer is updated via

$$W_j = \begin{cases} c \cdot W_j / \hat{\lambda}, & \text{if } c < \hat{\lambda} \\ W_j, & \text{otherwise} \end{cases} \tag{6}$$

where $\hat{\lambda}$ is the approximation for λ obtained by the power iteration. Also, c denotes a hyperparameter to practically control the exact upper bound depending on the input data.

As a novelty compared with a previous study [19], we here follow the convention of [11] and apply the spectral normalization (SN) also to the batch normalization. The Lipschitz constant of the batch normalization operator is given by

TABLE I
EXPERIMENTAL RESULTS ON THE EUROSAT DATASET. AUROC AND AUPR ARE DERIVED FROM A BINARY CLASSIFIER FED WITH ID AND OOD DATA FROM THE TEST SET. AUROC = AREA UNDER RECEIVER OPERATOR CURVE, AUPR = AREA UNDER PRECISION RECALL CURVE, DSM = DEMPSTER–SHAFER METRIC, MSP = 1 − MAXIMUM SOFTMAX PROBABILITY, AND PRED.ENT. = PREDICTIVE ENTROPY. BEST RESULTS BY UNCERTAINTY MEASURE IN BOLD, AND OVERALL BEST RESULTS ARE UNDERLINED

ID Data	OoD Data	Method	Loss ↓	AUROC ↑			AUPR ↑		
				DSM	MSP	Pred.Ent.	DSM	MSP	Pred.Ent.
Non-built classes	Built-up Classes	WideResNet [19]	0.93	0.32	0.65	-	0.32	0.59	-
		WideResNet-SN [19]	0.94	0.17	0.74	-	0.34	0.66	-
		WideResNet-GP [19]	1.10	0.26	0.65	-	0.36	0.60	-
		WideResNet-GP-IP	0.14	**0.77**	0.90	0.90	**0.56**	0.81	0.80
		WideResNet-GP-IP-BN	0.14	0.76	**0.90**	**<u>0.90</u>**	0.54	**0.81**	**<u>0.81</u>**
		WideResNet-SNGP [19]	1.38	0.21	0.63	-	0.35	0.57	-
		WideResNet-SNGP-IP	0.15	0.68	0.83	0.83	0.48	0.69	0.67
		WideResNet-SNGP-IP-BN	0.14	0.54	0.73	0.86	0.53	0.73	0.73
Non-built classes	Built-up Classes	ResNet50 [19]	3.27	0.50	0.62	-	0.55	0.55	-
		ResNet50-SN [19]	4.24	0.43	0.42	-	0.42	0.41	-
		ResNet50-GP [19]	3.82	0.65	0.45	-	0.54	0.44	-
		ResNet50-SNGP [19]	3.58	0.52	0.43	-	0.59	0.43	-

$\max_i |(\eta_i/(\mathrm{Var}(x)_i)^{1/2})|$ [21], with η being the learnable scale parameter and i being the index describing the batch size. This normalization differs from plain SN, because it is input data-driven and does not alter the network mapping by a single hyperparameter.

C. Output Mapping Distance Awareness

The distance awareness of the output mapping g can be handled via modeling g as a Gaussian process (GP) [22] on the hidden mapping output space $\mathbb{H} := \{h(x){:}x \in \mathcal{X}\}$. This process is generally specified by a mean function $m(h)$ and a covariance function $k(h, h')$. Then, an uninformative prior with a mean of 0 and an RBF kernel is placed on the latent process. The likelihood gets built by exposing the GP to all training data; after that the process gets optimized a posteriori. Due to computational and analytical intractability, the spectral normalization + Gaussian process (SNGP) approach [10] proposes the following: 1) a random Fourier feature (RFF) expansion [23] of the initial GP followed by 2) a Laplace approximation for the posterior. We refer the interested reader to [24]. Note that the SNGP approach proved to effectively increase the model's predictive uncertainty quality in an OOD detection setting for remote sensing image classification [19].

Novel to this work is the usage of the deterministic uncertainty estimation (DUE) framework proposed by van Amersfoort et al. [11], which will be investigated by various experiments in the following. For DUE, the GP gets placed directly onto the last layer of the hidden mapping. Then, the K-means algorithm is used to identify m inducing points (defined via the centroids found by the algorithm) in the hidden mapping feature space. The GP is then evaluated only on these points, which keeps the GP nonparametric (as opposed to the RFF expansion in the case of SNGP) [11]. For optimization, the expectation lower bound (ELBO) between the induced GP and the full likelihood is minimized, and the loss gets backpropagated as usual using stochastic gradient descent (SGD).

III. EXPERIMENT

For the experimental section, we investigate different networks for OOD detection in the context of land cover classification for remote sensing images. In particular, we evaluate the predictive uncertainties on unseen classes for which we differentiate between built-up classes and nonbuilt classes and train on one set and predict on another and vice versa. We closely follow the code implementation of van Amersfoort et al. [11], which can be found in [25]. We expand the experiments conducted by Koller et al. [19] by adding another deterministic uncertainty quantification technique, DUE. It is here denoted by GP-IP, which stands for a GP output mapping with the inducing points (IPs) approximation by van Amersfoort et al. [11]. We compare GP-IP to SNGP output mapping as well as SN and GP individually and the aforementioned batch normalization (BN). Two core networks are considered: ResNets with a depth of 50 (ResNet50) and wide ResNets with a depth of 28 (WideResNet). The OOD detection is carried out by a binary classifier defined by the predictive uncertainties, on which the area under the receiver operator curve (AUROC) and under the precision recall curve (AUPR) are derived. The Dempster–Shafer metric (DSM) and 1 − the maximum softmax probability (MSP) have already been previously used as uncertainty metric in [19]; here, the predictive entropy of the softmax prediction (Pred.Ent.) is added specifically for GP-IP approaches.

Adding to the experimental setup presented by Koller et al. [19], we conduct OOD detection on another remote sensing image classification benchmark dataset, namely, on the So2Sat LCZ42 [26] dataset. This dataset

TABLE II

EXPERIMENTAL RESULTS ON THE SO2SAT LCZ42 DATASET. AUROC AND AUPR ARE DERIVED FROM A BINARY CLASSIFIER FED WITH ID AND OOD DATA FROM THE TEST SET. AUROC = AREA UNDER RECEIVER OPERATOR CURVE, AUPR = AREA UNDER PRECISION RECALL CURVE, DSM = DEMPSTER–SHAFER METRIC, MSP = 1 − MAXIMUM SOFTMAX PROBABILITY, AND PRED.ENT. = PREDICTIVE ENTROPY. BEST RESULTS BY UNCERTAINTY MEASURE IN BOLD, AND OVERALL BEST RESULTS ARE UNDERLINED

ID Data	OoD Data	Method	Loss ↓	AUROC ↑			AUPR ↑		
				DSM	MSP	Pred.Ent.	DSM	MSP	Pred.Ent.
Built-up classes	Non-built Classes	WideResNet-GP-IP	0.12	0.81	0.87	0.87	0.78	0.83	0.85
		WideResNet-GP-IP-BN	0.11	**0.82**	0.86	0.87	**0.80**	0.82	0.84
		WideResNet-SNGP-IP	0.11	0.80	0.87	0.88	0.75	0.82	0.84
		WideResNet-SNGP-IP-BN	0.11	0.82	**0.88**	**0.89**	0.79	**0.85**	**0.87**
Non-built classes	Built-up Classes	WideResNet-GP-IP	0.09	0.84	0.92	0.93	0.78	0.92	0.93
		WideResNet-GP-IP-BN	0.08	**0.86**	**0.93**	**0.94**	**0.81**	**0.93**	**0.94**
		WideResNet-SNGP-IP	0.08	0.81	0.93	0.93	0.73	0.92	0.92
		WideResNet-SNGP-IP-BN	0.10	0.84	0.92	0.93	0.79	0.92	0.93

contains labeled Sentinel-2 imagery of 42 urban conglomerates around the world. The labels follow the local climate zone (LCZ) scheme from [27]. We split the 17 classes into built-up (seven) and nonbuilt (ten) classes for the OOD detection task. This split is widely used, and a label evaluation study performed by Zhu et al. [26] showed little to no label ambiguity between the two label sets. We use the training set of the publicly available cultural split of the dataset and randomly separate it into training and testing with 70% and 30%, respectively. The EuroSat dataset [28] also contains labeled Sentinel-2 patches, with a ten-class land cover labeling scheme. It was split similarly, with only two of the ten classes being specified as built-up classes. For additional details and class descriptions regarding the Eurosat dataset, see [19] for details.

Regarding the core network and the SNGP approach, we use the same hyperparameters as initially stated by Koller et al. [19] with the exception of having a width of $k = 5$ for the WideResNet (WRN). Although a width of 1 outperformed the full WRN network in the previous work [19], we saw stronger performance with a width of 5 in the case of the DUE approach. The width is a key parameter for the WideResNet, as it controls how many filters are trained in parallel for each convolution. We chose a width of 5, because it clearly outperformed networks with less width and was still running in a reasonable amount of time. The strong performance may be due to the nature of the GP output mapping: before, the GP was evaluated completely; now, a limited number of IPs keep the output mapping more flexible and allow for a more complex hidden mapping.

The number of IPs plays a crucial role in the DUE approach. We experimented with multiple values ranging from 10 to 200 and found 50 to be the best-performing number. Note that the number of IPs is interdependent on the batch size and on the number of classes. Experimentally, we found a batch size of 64 fitting for our layout. This is because a larger batch size not only scales the computational effort of a WRN exponentially but also requires more IPs, which has a similar effect. The additional hyperparameters were chosen as described in [19]. The predictive uncertainties for the SNGP approaches are computed via the DSM and the maximum softmax probability (see [19] for details). For the inducing point (-IP) methods, we follow the original authors' approach and additionally use the predictive entropy of the softmax vector [11]. Regarding the computational cost, we observed an increased computational demand from the GP output layer of roughly 1.3–1.5 times the demand of the core network itself. Due to very fast convergence, the spectral (batch) normalization has little to no effect on the computational cost.

IV. DISCUSSION

The performance gain of DUE over previous methods is immediately visible in Table I: all combinations using the Gaussian process with IPs (GP-IPs) beat previous baselines by a large margin for the task of detecting OOD data within the Eurosat dataset. Again, the DSM performs poorly, whereas the other two uncertainty metrics achieve more or less similar OOD detection results. Overall, the predictive entropy seems to work best. Interestingly, SN seems to hinder the model from effectively detecting OOD data, particularly for the area under the precision recall curve (AUPR). Despite the high value for c, the unrestricted hidden mapping performs slightly better than the restricted one. On the other hand, the data-driven spectral batch normalization has positive effects on the performance in most settings.

For the So2Sat LCZ42 dataset, an overall higher OOD performance can be observed, and the gap between the AUROC and the AUPR has narrowed; see Table II. This can be partially attributed to the larger size of the dataset. But, the more clear class separation could potentially also play a role. Adding to that, the patch size is only 25% of the size of the Eurosat patches. Due to poor performance on the Eurosat dataset, which lay far from the performance of the DUE approach, we did not perform LCZ42 experiments for the original SNGP approach. The balance between built-up and nonbuilt classes is relatively even, which allowed us to perform the OOD detection task in both ways. Interestingly, the detection of built-up classes as OOD samples works a lot better than

detecting nonbuilt classes when having seen only built-up classes during training. Overall, the earlier findings regarding spectral (batch) normalization can also be applied here.

In general, we achieve similar performance results as in [9] for the LCZ42 dataset. Note, however, that the DUE approach does not need any kind of OOD data during training. This result is interesting, since the approach taken in [9] constructs a framework that explicitly maximizes the loss toward OOD data, and the framework incorporates OOD data for this maximization already during training. The approach of DUE [11] and similar approaches, on the other hand, do not incorporate this data. Instead, they design the network to automatically map points further away, which lie further from the training data in the input space. This design seems to work reasonably well for land cover classification applications; we, therefore, leave further exploration of this important model architecture class for future research.

V. Conclusion

A model's ability to express the uncertainty about its prediction can be measured by means of OOD detection. Deterministic approaches with single forward passes achieved superior performance on machine learning benchmark datasets. By constructing such a network architecture to be distance-aware, unseen data points at test time are automatically mapped further away from the training data. This is then reflected by a higher uncertainty in the prediction. The SNGP approach [10] performed very well with little changes to the existing ResNets, and clear advantages of the approach could be shown for the field of remote sensing image classification [19]. Building upon the refined approach of [11], we introduced spectral batch normalization and a GP output layer, which is only realized on a limited number of IPs.

We then revisited the OOD detection task and showed strong results over two land cover benchmark datasets. In particular, for the Eurosat dataset, a performance gain in OOD detection of more than 10% points was observed. Even higher performance numbers were achieved on the So2Sat LCZ42 dataset. Overall, the IPs approximation works more reliably than the previous approximation using Laplace and RFF expansion. We believe this is due to the flexibility provided by the nonparametric properties of the GP and the reduced complexity due to the limited amount of IPs. Similarly, the data-driven spectral batch normalization shows more positive effects than the plain SN. These results emphasize the effectiveness of the proposed deterministic single forward pass uncertainty quantification framework for remote sensing image classification.

References

[1] J. Gawlikowski et al., "A survey of uncertainty in deep neural networks," *Artif. Intell. Rev.*, vol. 56, no. Suppl. 1, pp. 1513–1589, Oct. 2023, Art. no. 1589.

[2] 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.

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

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

[5] J. Schmitt and S. Roth, "Sampling-free variational inference for neural networks with multiplicative activation noise," in *Proc. DAGM German Conf. Pattern Recognit.*, 2021, pp. 33–47.

[6] E. Daxberger, E. Nalisnick, J. U. Allingham, J. Antorán, and J. M. Hernández-Lobato, "Bayesian deep learning via subnetwork inference," in *Proc. Int. Conf. Mach. Learn.*, 2021, pp. 2510–2521.

[7] J. Postels et al., "On the practicality of deterministic epistemic uncertainty," 2021, *arXiv:2107.00649*.

[8] S. Landgraf, M. Hillemann, K. Wursthorn, and M. Ulrich, "U-CE: Uncertainty-aware cross-entropy for semantic segmentation," 2023, *arXiv:2307.09947*.

[9] 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, 2022, Art. no. 3140324.

[10] J. Liu, Z. Lin, S. Padhy, D. Tran, T. B. Weiss, and B. Lakshminarayanan, "Simple and principled uncertainty estimation with deterministic deep learning via distance awareness," in *Proc. Adv. Neural Inf. Process. Syst.*, 2020, pp. 7498–7512.

[11] J. van Amersfoort, L. Smith, A. Jesson, O. Key, and Y. Gal, "On feature collapse and deep kernel learning for single forward pass uncertainty," 2021, *arXiv:2102.11409*.

[12] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. Courville, "Improved training of Wasserstein GANs," 2017, *arXiv:1704.00028*.

[13] J. Van Amersfoort, L. Smith, Y. W. Teh, and Y. Gal, "Uncertainty estimation using a single deep deterministic neural network," in *Proc. Int. Conf. Mach. Learn.*, 2020, pp. 9690–9700.

[14] T. Miyato, T. Kataoka, M. Koyama, and Y. Yoshida, "Spectral normalization for generative adversarial networks," 2018, *arXiv:1802.05957*.

[15] M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein generative adversarial networks," in *Proc. Int. Conf. Mach. Learn.*, Jul. 2017, pp. 214–223.

[16] G.-J. Qi, "Loss-sensitive generative adversarial networks on Lipschitz densities," *Int. J. Comput. Vis.*, vol. 128, no. 5, pp. 1118–1140, May 2020.

[17] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[18] P. L. Bartlett, S. N. Evans, and P. M. Long, "Representing smooth functions as compositions of near-identity functions with implications for deep network optimization," 2018, *arXiv:1804.05012*.

[19] C. Koller, P. Jung, and X. Xiang Zhu, "Exploring distance-aware uncertainty quantification for remote sensing image classification," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, Jul. 2023, pp. 5692–5695.

[20] J. Behrmann, W. Grathwohl, R. T. Chen, D. Duvenaud, and J.-H. Jacobsen, "Invertible residual networks," in *Proc. Int. Conf. Mach. Learn.*, 2019, pp. 537–582.

[21] H. Gouk, E. Frank, B. Pfahringer, and M. J. Cree, "Regularisation of neural networks by enforcing Lipschitz continuity," *Mach. Learn.*, vol. 110, no. 2, pp. 393–416, Feb. 2021.

[22] C. K. Williams and C. E. Rasmussen, *Gaussian Processes for Machine Learning*. Cambridge, MA, USA: MIT Press, 2006.

[23] A. Rahimi and B. Recht, "Random features for large-scale kernel machines," *Adv. Neural Inf. Process. Syst.*, pp. 1–8, 2007.

[24] J. Zhe Liu et al., "A simple approach to improve single-model deep uncertainty via distance-awareness," 2022, *arXiv:2205.00403*.

[25] J. Van Amersfoort. (2021). *Code for on Feature Collapse and Deep Kernel Learning for Single Forward Pass Uncertainty*. [Online]. Available: https://github.com/y0ast/DUE

[26] 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. (replaces Newsletter)*, vol. 8, no. 3, pp. 76–89, Sep. 2020.

[27] 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.

[28] P. Helber, B. Bischke, A. Dengel, and D. Borth, "EuroSAT: A novel dataset and deep learning benchmark for land use and land cover classification," *IEEE J. Sel. Topics Appl. Earth Obser. Remote Sens.*, vol. 12, no. 7, pp. 2217–2226, Jul. 2019.