Deep Deterministic Uncertainty: A New Simple Baseline

Jishnu Mukhoti[*1,2], Andreas Kirsch[*1], Joost van Amersfoort[1], Philip H.S. Torr[2], Yarin Gal[1]

Abstract

Reliable uncertainty from deterministic single-forward pass models is sought after because conventional methods of uncertainty quantification are computationally expensive. We take two complex single-forward-pass uncertainty approaches, DUQ and SNGP, and examine whether they mainly rely on a well-regularized feature space. Crucially, without using their more complex methods for estimating uncertainty, we find that a single softmax neural net with such a regularized feature-space, achieved via residual connections and spectral normalization, outperforms *DUQ and SNGP's epistemic uncertainty predictions using simple Gaussian Discriminant Analysis* post-training *as a separate feature-space density estimator—without fine-tuning on OoD data, feature ensembling, or input pre-processing. Our conceptually simple* Deep Deterministic Uncertainty (DDU) *baseline can also be used to disentangle aleatoric and epistemic uncertainty and performs as well as Deep Ensembles, the state-of-the art for uncertainty prediction, on several OoD benchmarks (CIFAR-10/100 vs SVHN/Tiny-ImageNet, ImageNet vs ImageNet-O), active learning settings across different model architectures, as well as in large scale vision tasks like semantic segmentation, while being* computationally cheaper.

1. Introduction

Two types of uncertainty are often of interest in ML: *epistemic uncertainty*, which is inherent to the model, caused by a lack of training data, and hence reducible with more data, and *aleatoric uncertainty*, caused by inherent noise or ambiguity in data, and hence irreducible with more data [7]. Disentangling these two is critical for applications such as active learning [16] or detection of out-of-distribution (OoD) samples [24]: in active learning, we wish to avoid inputs with high aleatoric but low epistemic uncertainty, and in OoD detection, we wish to avoid mistaking ambiguous in-distribution (iD) examples as OoD. Disentangling uncertainties is particularly challenging for noisy and ambiguous datasets found in safety-critical applications like autonomous driving [32] and medical diagnosis [11; 13].

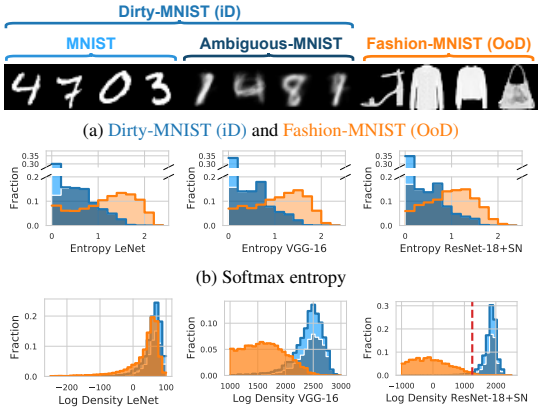


Figure 1. *Disentangling aleatoric and epistemic uncertainty on Dirty-MNIST (iD) and Fashion-MNIST (OoD)* ***(a)*** *requires using* softmax entropy ***(b)*** *and* feature-space density (GMM) ***(c)*** *with a well-regularized feature space (*ResNet-18+SN *vs* LeNet & VGG-16 *without smoothness & sensitivity).* **(b):** Softmax entropy captures aleatoric uncertainty for iD data (Dirty-MNIST), thereby separating unambiguous MNIST samples and Ambiguous-MNIST samples (stacked histogram). However, iD and OoD are confounded: softmax entropy has arbitrary values for OoD, indistinguishable from iD. **(c):** With a well-regularized feature space (DDU with ResNet-18+SN), iD and OoD densities do not overlap, capturing epistemic uncertainty. However, without such feature space (LeNet & VGG-16), feature density suffers from *feature collapse*: iD and OoD densities overlap. Generally, feature-space density confounds unambiguous and ambiguous iD samples as their densities overlap.

Related Work: Most well-known methods of uncertainty quantification in deep learning [1; 10; 15; 40; 66] require multiple forward passes at test time. Amongst these, Deep Ensembles have generally performed best in uncertainty prediction [57], but their significant memory and compute burden at training and test time hinders their adoption in real-life and mobile applications. Consequently, there has been an increased interest in uncertainty quantification using deterministic single forward-pass neural networks which have a smaller footprint and lower latency. Among these approaches, [43] uses Mahalanobis distances to quantify uncertainty by fitting a class-wise Gaussian distribution (with

[*]Equal contribution. [1]OATML, Department of Computer Science, University of Oxford, [2]Torr Vision Group, Department of Engineering Science, University of Oxford. Correspondence to: Jishnu Mukhoti <jishnu.mukhoti@eng.ox.ac.uk>, Andreas Kirsch <andreas.kirsch@cs.ox.ac.uk>.

shared covariance matrices) on the feature space of a pre-trained ResNet encoder. They do not consider the structure of the underlying feature-space however, which might explain why their competitive results require input perturbations, the ensembling of OoD metrics over multiple layers, and fine-tuning on OoD hold-out data.

DUQ & SNGP: Two popular works in single forward-pass uncertainty, DUQ [65] and SNGP [45], propose distance-aware output layers, in the form of RBFs (radial basis functions) or GPs (Gaussian processes), and introduce additional inductive biases in the feature extractor using a Jacobian penalty [19] or spectral normalisation [52], respectively, which encourage smoothness and sensitivity in the latent space. These methods perform well and are almost competitive with Deep Ensembles on OoD benchmarks. However, they require training to be changed substantially, and introduce additional hyper-parameters due to the specialised output layers used at training. Furthermore, DUQ and SNGP cannot disentangle aleatoric and epistemic uncertainty. Particularly, in DUQ, the feature representation of an ambiguous data point, high on aleatoric uncertainty, will be in between two centroids, but due to the exponential decay of the RBF it will seem far from both and thus have uncertainty similar to epistemically uncertain data points that are far from all centroids. In SNGP, the predictive variance is computed using a mean-field approximation of the softmax likelihood, which cannot be disentangled. The variance can also be computed using MC samples of the softmax likelihood which, in theory, can allow disentangling uncertainties (see Eq. (1)), but requires modelling the covariance between the classes, which is not the case in SNGP. We provide a more extensive review of related work in §A.

Contributions: **Firstly**, we investigate the question whether complex methods to estimate uncertainty like in DUQ and SNGP are necessary beyond feature-space regularization that encourages bi-Lipschitzness. When we use spectral normalisation like SNGP does, the short answer is an empirical no. Indeed, with a well-regularized feature space using spectral normalisation, we find that simply performing GDA (Gaussian Discriminant Analysis) *after training* as feature-space density estimator can reliably capture epistemic uncertainty. However, unlike [43], which does not place any constraints on the feature space, we do not require training on "OoD" hold-out data, feature ensembling, and input pre-processing to obtain good performance (see Tab. 1). This results in a conceptually simpler method. Moreover, we find that using a separate covariance matrix for each class improves OoD detection performance as compared to a shared covariance matrix like in [43]. **Secondly**, we investigate how to disentangle aleatoric and epistemic uncertainty. DUQ and SNGP do not address this directly. As we only perform GDA after training, the original softmax layer is trained using cross-entropy as a proper scoring rule [17] and can be temperature-scaled to provide good in-distribution calibration and aleatoric uncertainty. Thus, the combination of using GDA for epistemic uncertainty and the softmax predictive distribution for aleatoric uncertainty after training with feature-space regularisation, e.g. residual connections with spectral normalisation, provides a simple baseline which we call *Deep Deterministic Uncertainty (DDU)*. DDU outperforms regular softmax neural networks, as illustrated in Fig. 1. Furthermore, DDU is competitive with Deep Ensembles [40] and outperforms SNGP and DUQ [45; 65], with no changes to the model architecture beyond spectral normalisation, in several OoD benchmarks and active learning settings. **Finally**, using DeepLab-v3+ [3] on Pascal VOC 2012 [12], we show that DDU improves upon two classic uncertainty methods: MC Dropout [15] and Deep Ensembles, popularly used on the task of semantic segmentation, while being significantly faster to compute.

Additional Insights: Beyond the above contributions, we also provide additional insights on potential pitfalls for practitioners, which informed the design of DDU. **Firstly**, predictive entropy confounds aleatoric and epistemic uncertainty (Fig. 1b). This can be an issue in active learning in particular. Yet, this issue is often not visible for standard benchmark datasets without aleatoric noise. To examine this failure in more detail, we introduce a new dataset, Dirty-MNIST, which showcases the issue more clearly than artificially curated datasets like MNIST or CIFAR-10. *Dirty-MNIST* is a modified version of MNIST [41] with additional ambiguous digits (Ambiguous-MNIST) having multiple plausible labels and thus higher aleatoric uncertainty (Fig. 1a). **Secondly**, we observe that the softmax entropy of a deterministic model trained with maximum likelihood, while being high for ambiguous points (i.e., with high aleatoric uncertainty), might not be consistent for points with high epistemic uncertainty, i.e. the softmax entropy for an OoD sample might be low, high or anything in between for different models trained on the same data (Fig. 1b). **Thirdly**, we note that feature-space regularization [45] is crucial for the estimation of epistemic uncertainty[1]. Without such regularisation feature-space density alone might not separate iD from OoD data, possibly explaining the limited empirical success of previous approaches which attempt to use feature-space density [59]. This can be seen in Fig. 1c where the feature-space density of a VGG-16 or LeNet model are not able to differentiate iD Dirty-MNIST from OoD Fashion-MNIST while a ResNet-18 with spectral normalization can do so better.

Scope: Our focus is on obtaining a well-regularized feature space using spectral normalization in model architectures with residual connections, following [45]. Note that unsupervised methods using contrastive learning [69] might also obtain such a feature space by training on very large

[1] [58] argue for softmax confidence and entropy in their paper, yet feature-space density performs better in their experiments, too.

datasets, but training on them can be very expensive [63]. We only use GDA for estimating the feature-space density as it is straight-forward to implement and does not require performing expectation maximization or variational inference like other density estimators. Normalizing flows [9] or other more complex density estimators might provide even better density estimates, of course. Yet despite its simplicity, GDA is already sufficient to outperform other more complex approaches and obtain good results. As the amount of training data available grows and feature extractors improve, the quality of feature representations might improve as well. The underlying motiviation of this paper is that simple approaches will remain more applicable than more complex ones as our empirical results suggest.

2. Background

We review concepts for quantifying uncertainty.

Epistemic Uncertainty at point x is a quantity which is high for a previously unseen x, and decreases when x is added to the training set and the model is updated [33]. This conforms with using mutual information in Bayesian models and deep ensembles [37] and feature-space density in deterministic models as surrogates for epistemic uncertainty [59] as we examine below (see Fig. 2a and §F.8).

Aleatoric Uncertainty at point x is a quantity which is high for ambiguous or noisy samples [33]. It does not decrease with more data (see Fig. 2b). Note that aleatoric uncertainty is only meaningful in-distribution, as, by definition, it quantifies the level of ambiguity between the different classes which might be observed[2].

Bayesian Models [48; 55] provide a principled way of measuring uncertainty. Starting with a prior distribution $p(\omega)$ over model parameters ω, they infer a posterior $p(\omega|\mathcal{D})$, given the training data $\mathcal{D}$. The predictive distribution $p(y|x, \mathcal{D})$ for a given input x is computed via marginalisation over the posterior: $p(y|x, \mathcal{D}) = \mathbb{E}_{\omega\sim p(\omega|\mathcal{D})}[p(y|x, \omega)]$. Its predictive entropy $\mathbb{H}[Y|x, \mathcal{D}]$ upper-bounds the epistemic uncertainty, where epistemic uncertainty is quantified as the mutual information $\mathbb{I}[Y; \omega|x, \mathcal{D}]$ (*expected information gain*) between parameters ω and output y [14; 61]:

$$\underbrace{\mathbb{H}[Y|x, \mathcal{D}]}_{\text{predictive}} = \underbrace{\mathbb{I}[Y; \omega|x, \mathcal{D}]}_{\text{epistemic}} + \underbrace{\mathbb{E}_{\mathrm{p}(\omega|\mathcal{D})}[\mathbb{H}[Y|x, \omega]]}_{\text{aleatoric (for iD } x)}. \quad (1)$$

Predictive uncertainty will be high whenever either epistemic uncertainty or aleatoric uncertainty is high. However, the intractability of exact Bayesian inference in deep learning has led to the development of approximate inference methods [1; 15; 27; 28]. In practice, however, these methods are either unable to scale to large datasets and model architectures, suffer from low uncertainty quality, or require expensive Monte-Carlo sampling.

Deep Ensembles [40] average the outputs of an ensemble of neural networks. Uncertainty is estimated as the entropy of the averaged softmax outputs. Despite the high computational overhead at training and test time, Deep Ensembles along with recent extensions [10; 61; 66] form the state-of-the-art in uncertainty quantification in deep learning.

Deterministic Models produce a softmax distribution $p(y|x, \omega)$, and commonly either the *softmax confidence* $\max_c p(y = c|x, \omega)$ or the *softmax entropy* $\mathbb{H}[Y|x, \omega]$ are used as a measure of uncertainty [24]. Popular approaches to improve these metrics include pre-processing of inputs and post-hoc calibration methods [20; 44], alternative objective functions [8; 42], and exposure to outliers [25]. However, these methods are known to suffer from shortcomings like failing under distribution shift [57], requiring significant changes to the training setup, or assuming the availability of OoD samples during training.

Feature-Space Distances [43; 45; 65] and **Feature-Space Density** [46; 59] offer a different approach for estimating uncertainty in deterministic models. Following the definition of epistemic uncertainty above, it decreases when previously unseen samples are added to the training set. Feature-space distance and density methods realise this by estimating distance or density, respectively, to training data in the feature space (see again Fig. 2a). A previously unseen point with high distance (low density), once added to the training data, will have low distance (high density). Hence, they can be used as a proxy for epistemic uncertainty, under important assumptions about the feature space as detailed below. None of these methods, however, is competitive with Deep Ensembles, in uncertainty quantification, potentially for the reasons discussed next.

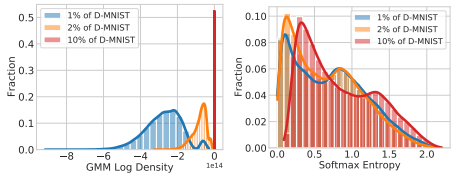
Feature Collapse [65] is why distance and density estimation in the feature space may fail to capture epistemic uncertainty: feature extractors might map the features of OoD inputs to iD regions in the feature space [64].

Smoothness & Sensitivity can be encouraged to prevent feature collapse by subjecting the feature extractor f_θ, with parameters θ to a *bi-Lipschitz constraint*:

$$K_L \; d_I(\mathrm{x}_1, \mathrm{x}_2) \leq d_F(f_\theta(\mathrm{x}_1), f_\theta(\mathrm{x}_2)) \leq K_U \; d_I(\mathrm{x}_1, \mathrm{x}_2),$$

for all inputs, x_1 and x_2, where d_I and d_F denote metrics for the input and feature space respectively, and K_L and K_U the lower and upper Lipschitz constants [45]. The lower bound ensures *sensitivity* to distances in the input space, and the upper bound ensures *smoothness* in the features, preventing them from becoming too sensitive to input variations, which, otherwise, can lead to poor generalisation and loss of robustness [65]. Methods of enouraging bi-Lipschitzness include: **i)** gradient penalty, by applying a two-sided penalty to the L2 norm of the Jacobian [19], and **ii)** spectral normalisation

[2] If the probability of observing x under the data generating distribution is zero, $p(y|x) = \frac{p(x,y)}{p(x)}$, and hence, the entropy as a measure of aleatoric uncertainty, is not defined.



(a) Epistemic (b) Aleatoric

Figure 2. *Epistemic and aleatoric uncertainty of ResNet-18+SN models trained on increasingly large subsets of DirtyMNIST.* The feature-space density increases while the softmax entropy stays roughly the same, consistent with epistemic and aleatoric uncertainty being reducible and irreducible with more data, respectively. See §F.8 for a discussion on this.

```
# instantiate models
model = create_sensitive_smooth_model()
gda = create_gda()

# train
training_samples, training_labels = load_training_set()
model.fit(training_samples, training_labels)

training_features = model.features(training_samples)
gda.fit(training_features, training_labels)

# test
test_features = model.features(test_sample)

epistemic_uncertainty = -gda.log_density(test_features)

is_ood = epistemic_uncertainty <= ood_threshold
if not is_ood:
  predictions = model.softmax_layer(test_features)
  aleatoric_uncertainty = entropy(predictions)
  return 0, aleatoric_uncertainty

return epistemic_uncertainty, np.log(num_classes)
```

Figure 3. DDU Pseudo-Code

[52] in models with residual connections, like ResNets [22]. [62] provides an in-depth analysis which supports that spectral normalisation leads to bi-Lipschitzness. Compared to the Jacobian gradient penalty [65], spectral normalisation is significantly faster and has more stable training dynamics.

3. Deep Deterministic Uncertainty

As introduced in §1, we discuss the primary components of our proposed DDU in this section.

Ensuring Sensitivity & Smoothness: We ensure sensitivity and smoothness using spectral normalisation in models with residual connections. In addition, we make minor changes to the standard residual block to further encourage sensitivity without sacrificing accuracy (see details in §C.1).

Disentangling Epistemic & Aleatoric Uncertainty: To quantify epistemic uncertainty, we fit a feature-space density estimator after training. We use GDA, a GMM $q(y, z)$ with a single Gaussian component per class, and fit each class component by computing the empirical mean and covariance, per class, of the feature vectors $z = f_\theta(x)$, which are the outputs of the last convolutional layer of the model computed on the training samples x. *Note that we do not require OoD data to fit these and unlike [43] we use a separate covariance matrix for each class.* Fitting a GDA on the feature space, thus requires no further training and only requires a single forward pass through the training set.

Evaluation: At test time, we estimate the epistemic uncertainty by evaluating the marginal likelihood of the feature representation under our density $\mathrm{q}(z) = \sum_y \mathrm{q}(z|y)\,\mathrm{q}(y)$. To quantify aleatoric uncertainty for in-distribution samples, we use the entropy $\mathbb{H}[Y|x, \theta]$ of the softmax distribution $\mathrm{p}(y \mid x, \theta)$. Note that the softmax distribution thus obtained can be further calibrated using temperature scaling [20]. Thus, for a given input, a high feature-space density indicates low epistemic uncertainty (iD), at which point, we can trust the aleatoric estimate from the softmax entropy. The sample can then be either unambiguous (low softmax entropy) or ambiguous (high softmax entropy). Conversely, a low feature density indicates high epistemic uncertainty (OoD), and we cannot trust softmax predictions. A simple Python pseudo-code using a scikit-learn-like API [2] is shown in Fig. 3. A more detailed algorithm and the corresponding computational complexity can be found in §C.

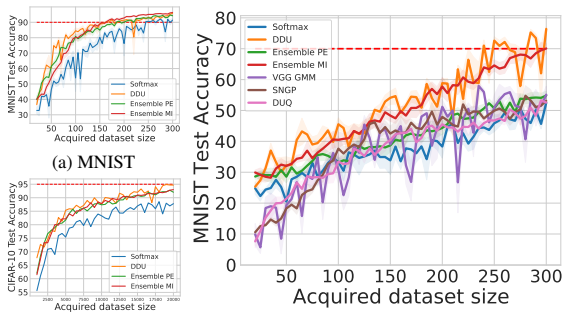
Sanity Check: To verify our claims on DDU's ability to quantify epistemic and aleatoric uncertainty, we train a ResNet-18 model with spectral normalisation (ResNet-18+SN) on increasingly large subsets of DirtyMNIST (1%, 2% and 10% particularly) and plot the feature-space density as well as the softmax entropy for each of these models in Fig. 2. With increasing training set size, feature-space density on the test set increases, following the definition of epistemic uncertainty, whereas softmax entropy remains similar, indicative of aleatoric uncertainty.

4. Experiments

We evaluate DDU's quality of epistemic uncertainty estimation in active learning [4] using MNIST, CIFAR-10 and an ambiguous version of MNIST (Dirty-MNIST). We also test DDU on several OoD detection settings including CIFAR-10 vs SVHN/CIFAR-100/Tiny-ImageNet/CIFAR-10-C, CIFAR-100 vs SVHN/Tiny-ImageNet and ImageNet vs ImageNet-O dataset pairings, where we outperform other deterministic single-forward-pass methods and perform on par with deep ensembles. Finally, we also evaluate DDU on the task of semantic segmentation on Pascal VOC, comparing with a deterministic model, MC Dropout (MCDO) [15] and deep ensembles. In the appendix, we also examine DDU's performance on the real-world QUBIQ challenge in §F.3 and on the well-known Two Moons toy dataset in §F.6. We elaborate on how DDU can disentangle epistemic and aleatoric uncertainty, the setting depicted in Fig. 1, in §F.1.1, and the effect of feature-space regularisation in §F.2.

4.1. Active Learning

We first demonstrate the quality of our uncertainty disentanglement in active learning (AL) [4]. AL aims to train models in a data-efficient manner. Additional training samples are iteratively acquired from a large pool of unlabelled data and labelled with the help of an expert. After each acquisition step, the model is retrained on the newly expanded training set. This is repeated until the model achieves a desirable accuracy—or when a maximum number of samples have been acquired.

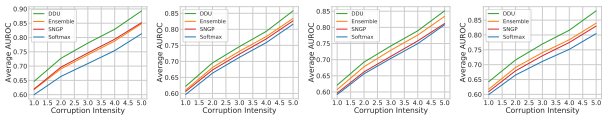


(a) MNIST (b) CIFAR-10 (c) Dirty-MNIST

Figure 4. *Active Learning experiments.* Acquired training set size vs test accuracy. DDU performs on par with Deep Ensembles.

Data-efficient acquisition relies on acquiring labels for the most informative samples. This can be achieved by selecting points with high epistemic uncertainty [16]. Conversely, repeated acquisition of points with high aleatoric uncertainty is not informative for the model and such acquisitions lead to data inefficiency. AL, therefore, makes an excellent application for evaluating epistemic uncertainty and the ability of models to separate different sources of uncertainty. We evaluate DDU on three different setups: **i)** with clean MNIST samples in the pool set, **ii)** with clean CIFAR-10 samples in the pool set, and **ii)** with Dirty-MNIST, having a 1:60 ratio of MNIST to Ambiguous-MNIST samples, in the pool set. In the first two setups, we compare 3 baselines: **i)** a ResNet-18 with softmax entropy as the acquisition function, **ii)** DDU trained using a ResNet-18 with feature density as acquisition function, and **iii)** a Deep Ensemble of 3 ResNet-18s with the predictive entropy (PE) and mutual information (MI) of the ensemble as the acquisition functions. In the last setup, in addition to the above 3 approaches, we also use **iv)** feature density of a VGG-16 instead of ResNet-18+SN as an ablation to see if feature density of a model without inductive biases performs well, **v)** SNGP and **vi)** DUQ as additional baselines. For MNIST and Dirty-MNIST, we start with an initial training-set size of 20 randomly chosen MNIST points, and in each iteration, acquire the 5 samples with highest reported epistemic uncertainty. For each step, we train the models using Adam [34] for 100 epochs and choose the one with the best validation set accuracy. We stop the process when the training set size reaches 300. For CIFAR-10, we start with 1000 samples and go up to 20000 samples with an acquisition size of 500 samples in each step.

MNIST & CIFAR-10 In Fig. 4(a) and Fig. 4(b), for regular curated MNIST and CIFAR-10 in the pool set, DDU clearly outperforms the deterministic softmax baseline and is competitive with Deep Ensembles. For MNIST, the softmax baseline reaches 90% test-set accuracy at a training-set size of 245. DDU reaches 90% accuracy at a training-set size of 160, whereas Deep Ensemble reaches the same at 185 and 155 training samples with PE and MI as the acquisition functions respectively. Note that DDU is three times faster than a Deep Ensemble, which needs to train three models independently after every acquisition.



(a) WRN-28-10 (b) RN-50 (c) RN-110 (d) DN-121

Figure 5. AUROC vs corruption intensity averaged over all corruption types in CIFAR-10-C for 4 architectures. More details in §4.2 and more ablations in §D in the appendix.

Dirty-MNIST. Real-life datasets often contain observation noise and ambiguous samples. What happens when the pool set contains a lot of such noisy samples having high aleatoric uncertainty? In such cases, it becomes important for models to identify unseen and informative samples with high epistemic uncertainty and not with high aleatoric uncertainty. To study this, we construct a pool set with samples from Dirty-MNIST (see §B). We significantly increase the proportion of ambiguous samples by using a 1:60 split of MNIST to Ambiguous-MNIST (a total of 1K MNIST and 60K Ambiguous-MNIST samples). In Fig. 4(c), for Dirty-MNIST in the pool set, the difference in the performance of DDU and the deterministic softmax model is stark. While DDU achieves a test set accuracy of 70% at a training set size of 240 samples, the accuracy of the softmax baseline peaks at a mere 50%. In addition, all baselines, including SNGP, DUQ and the feature density of a VGG-16, which fail to solely capture epistemic uncertainty, are significantly outperformed by DDU and the MI baseline of the deep ensemble. However, note that DDU also performs better than Deep Ensembles with the PE acquisition function. The difference gets larger as the training set size grows: DDU's feature density and Deep Ensemble's MI solely capture epistemic uncertainty and hence, do not get confounded by iD ambiguous samples with high aleatoric uncertainty.

4.2. OoD Detection

OoD detection is an application of epistemic uncertainty quantification: if we do not train on OoD data, we expect OoD data points to have higher epistemic uncertainty than iD data. We evaluate CIFAR-10 vs SVHN/CIFAR-100/Tiny-ImageNet/CIFAR-10-C, CIFAR-100 vs SVHN/Tiny-ImageNet and ImageNet vs ImageNet-O as iD vs OoD dataset pairs for this experiment [6; 23; 39; 56]. We also evaluate DDU on different architectures: Wide-ResNet-28-10, Wide-ResNet-50-2, ResNet-50, ResNet-110 and DenseNet-121 [22; 30; 71]. The training setup is described in §D.2. In addition to using softmax entropy of a deterministic model (*Softmax*) for both aleatoric and epistemic uncertainty, we also compare with the following **baselines** that do not require training or fine-tuning on OoD data:

- *Energy-based model* [46]: We use the softmax entropy as aleatoric uncertainty and the unnormalized softmax density (the logsumexp of the logits) as epistemic uncertainty

Table 1. *OoD detection performance of different baselines using a Wide-ResNet-28-10 architecture with the CIFAR-10 vs SVHN/CIFAR-100/Tiny-ImageNet and CIFAR-100 vs SVHN/Tiny-ImageNet dataset pairs averaged over 25 runs. SN:* Spectral Normalisation, *JP:* Jacobian Penalty. The best deterministic single-forward pass method and the best method overall are in bold for each metric.

Train Dataset	Method	Penalty	Aleatoric Uncertainty	Epistemic Uncertainty	Accuracy (↑)	ECE (↓)	AUROC SVHN (↑)	AUROC CIFAR-100 (↑)	AUROC Tiny-ImageNet (↑)
CIFAR-10	Softmax	-	Softmax Entropy	Softmax Entropy	95.98 ± 0.02	$\mathbf{0.85 \pm 0.02}$	94.44 ± 0.43	89.39 ± 0.06	88.42 ± 0.05
	Energy-based [46]	-		Softmax Density			94.56 ± 0.51	88.89 ± 0.07	88.11 ± 0.06
	DUQ [65]	JP	Kernel Distance	Kernel Distance	94.6 ± 0.16	1.55 ± 0.08	93.71 ± 0.61	85.92 ± 0.35	86.83 ± 0.12
	SNGP [45]	SN	Predictive Entropy	Predictive Entropy	$\mathbf{96.04 \pm 0.09}$	1.8 ± 0.1	94.0 ± 1.3	91.13 ± 0.15	89.97 ± 0.19
	DDU (ours)	**SN**	**Softmax Entropy**	**GDA Density**	95.97 ± 0.03	$\mathbf{0.85 \pm 0.04}$	$\mathbf{97.86 \pm 0.19}$	$\mathbf{91.34 \pm 0.04}$	$\mathbf{91.07 \pm 0.05}$
	5-Ensemble [40]	-	Predictive Entropy	Predictive Entropy Mutual Information	$\mathbf{96.59 \pm 0.02}$	$\mathbf{0.76 \pm 0.03}$	97.73 ± 0.31 97.18 ± 0.19	$\mathbf{92.13 \pm 0.02}$ 91.33 ± 0.03	90.06 ± 0.03 90.90 ± 0.03
					Accuracy (↑)	ECE (↓)	AUROC SVHN (↑)		AUROC Tiny-ImageNet (↑)
CIFAR-100	Softmax	-	Softmax Entropy	Softmax Entropy	80.26 ± 0.06	4.62 ± 0.06	77.42 ± 0.57		81.53 ± 0.05
	Energy-based [46]	-		Softmax Density			78 ± 0.63		81.33 ± 0.06
	SNGP [45]	SN	Predictive Entropy	Predictive Entropy	80.00 ± 0.11	4.33 ± 0.01	85.71 ± 0.81		78.85 ± 0.43
	DDU (ours)	**SN**	**Softmax Entropy**	**GMM Density**	$\mathbf{80.98 \pm 0.06}$	$\mathbf{4.10 \pm 0.08}$	$\mathbf{87.53 \pm 0.62}$		$\mathbf{83.13 \pm 0.06}$
	5-Ensemble [40]	-	Predictive Entropy	Predictive Entropy Mutual Information	$\mathbf{82.79 \pm 0.10}$	$\mathbf{3.32 \pm 0.09}$	79.54 ± 0.91 77.00 ± 1.54		82.95 ± 0.09 82.82 ± 0.04

Table 2. *OoD detection performance of different baselines using ResNet-50, Wide-ResNet-50-2 and VGG-16 architectures on ImageNet vs ImageNet-O [26].* Best AUROC scores are marked in bold.

Model	Accuracy (↑)		ECE (↓)		AUROC (↑)				
	Deterministic	3-Ensemble	Deterministic	3-Ensemble	Softmax Entropy	Energy-based Model	DDU	3-Ensemble PE	3-Ensemble MI
ResNet-50	74.8 ± 0.05	76.01	2.08 ± 0.11	2.07	51.42 ± 0.61	55.76 ± 0.81	$\mathbf{71.29 \pm 0.08}$	60.3	62.43
Wide-ResNet-50-2	76.75 ± 0.11	77.58	1.18 ± 0.07	1.22	52.71 ± 0.23	57.13 ± 0.4	$\mathbf{73.12 \pm 0.19}$	60.45	64.81
VGG-16	72.48 ± 0.02	73.54	2.62 ± 0.11	2.59	50.67 ± 0.22	52.04 ± 0.23	54.32 ± 0.14	58.74	**60.56**

without regularisation to avoid feature collapse.

- *DUQ [65] & SNGP [45]*: We compare with the state-of-the-art deterministic methods for uncertainty quantification including DUQ and SNGP. For SNGP, we use the exact predictive covariance computation and we use the entropy of the average of the MC softmax samples as uncertainty. For DUQ, we use the closest kernel distance. Note that for CIFAR-100, DUQ's one-vs-all objective did not converge during training and hence, we do not include the DUQ baseline for CIFAR-100.
- *5-Ensemble*: We use an ensemble of 5 networks and compute the predictive entropy of the ensemble as both epistemic and aleatoric uncertainty and mutual information as epistemic uncertainty.

Results: Table 1 presents the AUROC for Wide-ResNet-28-10 models on CIFAR-10 vs SVHN/CIFAR-100/Tiny-ImageNet and CIFAR-100 vs SVHN/Tiny-ImageNet along with their respective test set accuracy and ECE post temp-scaling (additional calibration scores in §F.5 and comparison with more baselines in §F.4). The equivalent results for other architectures: ResNet-50/110 and DenseNet-121 can be found in Tab. 5, Tab. 6 and Tab. 7 in the appendix. Note that for DDU, post-hoc calibration with temperature scaling [20], is simple as it does not affect the GMM density. We also plot the AUROC averaged over corruption types vs corruption intensity for CIFAR-10 vs CIFAR-10-C in Fig. 5, with AUROC plots per corruption type in Fig. 10, Fig. 11, Fig. 12 and Fig. 13 of the appendix. Finally, in Tab. 2, we present AUROC for models trained on ImageNet.

For OoD detection, *DDU outperforms all other deterministic single-forward-pass methods, DUQ, SNGP and the energy-based model approach from [46], on CIFAR-10 vs SVHN/CIFAR-100/Tiny-ImageNet, CIFAR-10 vs CIFAR-10-C and CIFAR-100 vs SVHN/Tiny-ImageNet, often performs on par with state-of-the-art Deep Ensembles—and even performing better in a few cases*. This holds true for all the architectures we experimented on. Similar observations can be made on ImageNet vs ImageNet-O as well. Importantly, the great performance in OoD detection comes without compromising on the single-model test set accuracy in comparison to other deterministic methods.

Ablations: Additional ablations for the CIFAR-10/100 experiments are detailed in §E: Tab. 8 and 9. These tables along with observations in Tab. 2, show that *the feature density of a VGG-16 (i.e. without residual connections and spectral normalisation) is unable to beat a VGG-16 ensemble, whereas a Wide-ResNet-28-10 with spectral normalisation outperforms its corresponding ensemble in almost all the cases*. This result further validates the importance of having a regularized feature space on the model to obtain smoothness and sensitivity. Also note that, even without spectral normalisation, a Wide-ResNet-28 has residual connections built into its model architecture, which can be a contributing factor towards good performance as residual connections make the model sensitive to changes in the input space. Finally, we also provide an ablation using LDA [43], which uses a shared covariance matrix over all classes, instead of GDA with covariance matrices per class. The resulting AUROC for Wide-ResNet-28-10 trained on CIFAR-10/100 and for Wide-ResNet-50-2 and ResNet-50 trained on ImageNet in Tab. 10 in §E. LDA only outperforms GDA when using SVHN as an OoD dataset. In all other cases, GDA obtains significantly higher AUROC, thereby indicating the advantage of modeling density using individual covariance matrices per class.

4.3. Semantic Segmentation

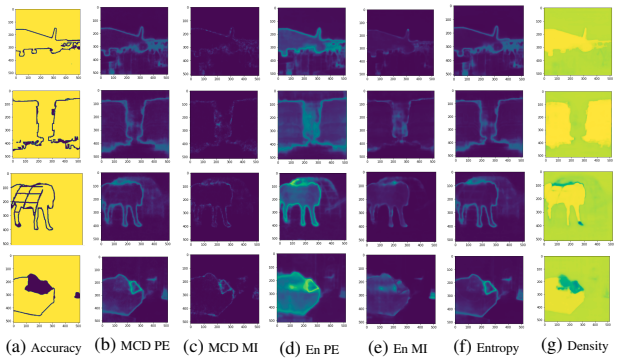
In this section, we apply DDU to the task of semantic segmentation on Pascal VOC 2012 [12], comparing with a vanilla softmax model, MC Dropout and deep ensembles. Semantic segmentation [47], classifies every pixel of a given to one of a fixed set of classes. Since different classes can have

different levels of representation in a segmentation dataset, it forms a classic example of a problem with class imbalance, thereby requiring reliable estimates of epistemic uncertainty. Furthermore, due to the computationally heavy nature of semantic segmentation, classic uncertainty quantification approaches like MC Dropout and deep ensemblesare rendered infeasible in real-world applications.

Pixel-Independent Class-Wise Means and Covariances: As each pixel has a corresponding prediction in semantic segmentation, it is natural to ask if the Gaussian means and covariance matrices need to be computed per pixel. To examine this, in Fig. 9 of §D.3, we plot the L2 distances between feature space means of all pairs of classes obtained from a DeepLab-v3+ [3] model with a ResNet-101 backbone for two "distant" pixels. We observe that pixels of the same class are much closer in the feature space than pixels of different classes, irrespective of their location in the image. In spirit of a new simple baseline, we thus compute the Gaussian means and covariances per class, taking each pixel as a separate data point.

Architecture, Training and Evaluation Metrics: As mentioned above, we evaluate DDU on Pascal VOC 2012 and compare to a vanilla softmax model, MC Dropout with 5 forward passes at test time, and a deep ensembles with 3 members. We use DeepLab-v3+ with a ResNet-101 backbone as the model architecture. Further training details are in §D.3. Finally, to evaluate the uncertainty estimates, we use metrics proposed in [53]: $p(\text{accurate}|\text{certain})$, $p(\text{uncertainty}|\text{inaccurate})$ and PAVPU. $p(\text{accurate}|\text{certain})$ computes the probability of the model being accurate given that it is confident. Similarly, $p(\text{uncertainty}|\text{inaccurate})$ measures probability of the model being uncertain given that it is inaccurate and $PAVPU$ computes the probability of the model being confident on accurate predictions and uncertain on inaccurate ones. Ideally, high values for these metrics indicate better uncertainty estimates in segmentation. Furthermore, note that these metrics can be computed at different thresholds of uncertainty (defining if a model is certain or not).

Results and Discussion: In Fig. 7, we present the above 3 metrics for all segmentation baselines evaluated on the Pascal VOC validation set. We also report the val set accuracy and runtime of a single forward pass in Tab. 3. Finally, we visualise uncertainty estimates from each baseline in Fig. 6. Firstly, from Tab. 3, it is clear that DDU has the runtime of a deterministic model which is significantly faster than both MC Dropout and deep ensembles. Also note that DDU's mIoU is the same as that of the vanilla softmax model. Secondly, from Fig. 7, we see that DDU consistently performs better on all 3 evaluation metrics compared to the other baselines. Finally, Fig. 6 qualitatively validates that DDU's feature-space density captures epistemic uncertainty while the softmax entropy captures aleatoric uncertainty. For DDU,



(a) Accuracy (b) MCD PE (c) MCD MI (d) En PE (e) En MI (f) Entropy (g) Density

Figure 6. *Visualisation of uncertainty baselines on four PASCAL VOC validation samples (**rows**).* **Columns**: **(a)** shows pixel-wise accuracy; **(b), (c)** predictive entropy (PE) and mutual information (MI) obtained for MC Dropout (MCD); **(d), (e)** for deep ensembles; **(f)** per-pixel softmax entropy, *the aleatoric estimate of DDU*; and **(g)** feature density, *the epistemic component of DDU*. For all but (g): the brighter, the more uncertain, whereas DDU's density (g) captures certainty: hence, the brighter, the more certain.

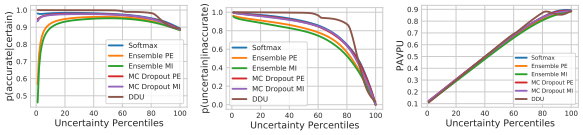


Figure 7. $p(\text{accurate}|\text{certain})$, $p(\text{uncertain}|\text{inaccurate})$ and PAVPU evaluated on PASCAL VOC validation set. DDU outperforms all other baselines.

Table 3. *Pascal VOC val set mIoU and runtime in milliseconds averaged over 10 forward passes.* For MC Dropout, we perform 5 stochastic forward passes.

Baseline	Softmax	MC Dropout	Deep Ensemble	**DDU**
mIoU	78.53	78.61	78.47	78.53
Runtime (ms)	275.48 ± 1.91	1576.75 ± 1.56	875.87 ± 0.79	$\mathbf{263.83 \pm 2.79}$

for the first two samples (first two rows, Fig. 6g), the epistemic uncertainty is not high and only aleatoric uncertainty is captured along edges of objects. However, for the last sample (4$^{\text{th}}$ row, Fig. 6g), the epistemic uncertainty is high for a relatively large patch on the image which is inaccurately predicted by the model as well. Note that only DDU's feature density is significantly lower for that entire region, whereas softmax entropy does not capture high uncertainty there and is only high along the edges. These observations are in line with [33]: aleatoric uncertainty is high on edges of objects as they correspond to regions of high ambiguity and noise; on the other hand, epistemic uncertainty is high for regions of the image which are previously unseen.

5. Additional Insights

We conclude with a discussion on potential pitfalls of predictive entropy in general and softmax entropy of deterministic models in particular. Proofs for all statements are provided in §G.

Potential Pitfalls of *Predictive Entropy*: Conceptually, *predictive entropy confounds epistemic and aleatoric uncer-*

tainty. Since ensembling can also be interpreted as Bayesian Model Averaging [21; 68], with each ensemble member approximating a sample from a posterior, eq. (1) can be applied to ensembles to disentangle epistemic and aleatoric uncertainty. Both mutual information $\mathbb{I}[Y; \omega \mid x, \mathcal{D}]$ and predictive entropy $\mathbb{H}[Y \mid x, \mathcal{D}]$ could be used to detect OoD samples. However, previous empirical findings show predictive entropy outperforming mutual information [50]. Indeed, much of the recent literature only focuses on predictive entropy for OoD detection (see §A.1). We explain these findings using the following observation:

Observation 5.1. If we *know* that *either* aleatoric *or* epistemic uncertainty is *low* for a sample, predictive entropy is a good measure of the uncertainty type which is high.

Thus, predictive entropy, as an upper-bound of mutual information, can separate iD and OoD data better when datasets are curated and have low aleatoric uncertainty. However, as seen in eq. (1), predictive entropy can be high for both iD ambiguous samples (high aleatoric) as well as OoD samples (high epistemic) (see Fig. 15) and might *not* be an effective measure for OoD detection when used with datasets that are not curated with ambiguous samples, like Dirty-MNIST, as seen in our active learning results.

Potential Pitfalls of *Softmax Entropy*: The softmax entropy for deterministic models trained with maximum likelihood can be *inconsistent*. In fact, the mechanism underlying the estimation of Deep Ensemble epistemic uncertainty requires it to be so:

Proposition 5.2. *Let x_1 and x_2 be points such that x_1 has **higher** epistemic uncertainty than x_2 under the ensemble: $\mathbb{I}[Y_1; \omega \mid x_1, \mathcal{D}] > \mathbb{I}[Y_2; \omega \mid x_2, \mathcal{D}] + \delta, \delta \geq 0$. Further assume both have similar predictive entropy $\mid \mathbb{H}[Y_1 \mid x_1, \mathcal{D}] - \mathbb{H}[Y_2 \mid x_2, \mathcal{D}] \mid \leq \epsilon$, $\epsilon \geq 0$. Then, there exist sets of ensemble members Ω with $\mathrm{p}(\Omega \mid \mathcal{D}) > 0$, such that for all softmax models $\omega \in \Omega$ the softmax entropy of x_1 is **lower** than the softmax entropy of x_2: $\mathbb{H}[Y_1 \mid x_1, \omega] < \mathbb{H}[Y_2 \mid x_2, \omega] - (\delta - \epsilon)$.*

If a sample is assigned higher epistemic uncertainty (in the form of mutual information) by a Deep Ensemble than another sample, it will necessarily be assigned lower softmax entropy by at least one of the ensemble's members. As a result, a priori, we cannot know whether a softmax model preserves the order or not, and *the empirical observation that the mutual information of an ensemble can quantify epistemic uncertainty well implies that the softmax entropy of a deterministic model might not*. We see this in Fig. 1b, 15 and §G.1.3 where softmax entropy for OoD samples can be high, low or anywhere in between. While *not all* model architectures might behave like this, when the mutual information of a Deep Ensemble works well empirically, Proposition 5.2 holds.

Objective Mismatch: The predictive probability induced by a feature-density estimator will generally not be well-calibrated as there is an objective mismatch. This was overlooked in previous research on uncertainty quantification for deterministic models [22; 43; 45; 59; 65]. Specifically, a mixture model $\mathrm{q}(y, z) = \sum_y \mathrm{q}(z \mid y)\, \mathrm{q}(y)$, using one component per class, cannot be optimal for both feature-space density and predictive distribution estimation as there is an *objective mismatch* [54, Ex. 4.20, p. 145]:

Proposition 5.3. *For an input x, let $z = f_\theta(x)$ denote its feature representation in a feature extractor f_θ with parameters θ. Then the following hold:*

1. *A discriminative classifier $\mathrm{p}(y \mid z)$, e.g. a softmax layer, is well-calibrated in its predictions when it maximises the conditional log-likelihood $\log \mathrm{p}(y \mid z)$;*
2. *A feature-space density estimator $\mathrm{q}(z)$ is optimal when it maximises the marginalised log-likelihood $\log \mathrm{q}(z)$;*
3. *A mixture model $\mathrm{q}(y, z) = \sum_y \mathrm{q}(z \mid y)\, \mathrm{q}(y)$ might not maximise both objectives, conditional log-likelihood and marginalised log-likelihood, at the same time. In the specific instance that a GMM with one component per class does maximise both, the resulting model must be a GDA (but the opposite does not hold).*

Hence, importantly, DDU uses *both* a discriminative classifier (softmax layer) to capture aleatoric uncertainty for iD samples and a separate feature-density estimator to capture epistemic uncertainty even on a model trained using conditional log-likelihood, i.e. the usual cross-entropy objective. Figure 17 and §G.2.3 provide additional intuitions.

6. Conclusion

Deep Deterministic Uncertainty (DDU) can outperform state-of-the-art deterministic single-pass uncertainty methods in active learning and OoD detection by fitting a GDA for feature-space density estimation after training a model with residual connections and spectral normalization [43; 45] while performing as well as deep ensembles in several settings. Hence, DDU provides a very simple method to obtain good epistemic and aleatoric uncertainty estimates and might be taken into consideration as an alternative to deep ensembles without requiring the complexities or computational cost of the current state-of-the-art. Reliable uncertainty quantification is an important requirement to make deep neural nets safe for deployment. Thus, we hope our work will contribute to increasing safety, reliability and trust in AI.

Acknowledgements This project was supported by the UKRI grant: Turing AI Fellowship EP/W002981/1 and EPSRC/MURI grant: EP/N019474/1. We would also like to thank the Royal Academy of Engineering and FiveAI. AK is supported by the UK EPSRC CDT in Autonomous Intelligent Machines and Systems (grant reference EP/L015897/1). JvA is grateful for funding by the EPSRC (grant reference EP/N509711/1) and Google-DeepMind.

References

[1] Charles Blundell, Julien Cornebise, Koray Kavukcuoglu, and Daan Wierstra. Weight uncertainty in neural network. In *International Conference on Machine Learning*, pages 1613–1622. PMLR, 2015. 1, 3

[2] Lars Buitinck, Gilles Louppe, Mathieu Blondel, Fabian Pedregosa, Andreas Mueller, Olivier Grisel, Vlad Niculae, Peter Prettenhofer, Alexandre Gramfort, Jaques Grobler, Robert Layton, Jake VanderPlas, Arnaud Joly, Brian Holt, and Gaël Varoquaux. API design for machine learning software: experiences from the scikit-learn project. In *ECML PKDD Workshop: Languages for Data Mining and Machine Learning*, pages 108–122, 2013. 4

[3] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587*, 2017. 2, 7, 14

[4] David A Cohn, Zoubin Ghahramani, and Michael I Jordan. Active learning with statistical models. *Journal of artificial intelligence research*, 4:129–145, 1996. 4

[5] Thomas M Cover. *Elements of information theory*. John Wiley & Sons, 1999. 27

[6] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 5

[7] Armen Der Kiureghian and Ove Ditlevsen. Aleatory or epistemic? does it matter? *Structural safety*, 31(2):105–112, 2009. 1

[8] Terrance DeVries and Graham W Taylor. Learning confidence for out-of-distribution detection in neural networks. *arXiv preprint arXiv:1802.04865*, 2018. 3, 13

[9] Laurent Dinh, David Krueger, and Yoshua Bengio. Nice: Non-linear independent components estimation, 2015. 3

[10] Michael Dusenberry, Ghassen Jerfel, Yeming Wen, Yian Ma, Jasper Snoek, Katherine Heller, Balaji Lakshminarayanan, and Dustin Tran. Efficient and scalable bayesian neural nets with rank-1 factors. In *International conference on machine learning*, pages 2782–2792. PMLR, 2020. 1, 3

[11] Andre Esteva, Brett Kuprel, Roberto A Novoa, Justin Ko, Susan M Swetter, Helen M Blau, and Sebastian Thrun. Dermatologist-level classification of skin cancer with deep neural networks. *nature*, 542(7639):115–118, 2017. 1

[12] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88(2):303–338, 2010. 2, 6

[13] Angelos Filos, Sebastian Farquhar, Aidan N Gomez, Tim GJ Rudner, Zachary Kenton, Lewis Smith, Milad Alizadeh, Arnoud de Kroon, and Yarin Gal. A systematic comparison of bayesian deep learning robustness in diabetic retinopathy tasks. *arXiv preprint arXiv:1912.10481*, 2019. 1, 17

[14] Yarin Gal. *Uncertainty in Deep Learning*. PhD thesis, University of Cambridge, 2016. 3

[15] Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*, pages 1050–1059, 2016. 1, 2, 3, 4

[16] Yarin Gal, Riashat Islam, and Zoubin Ghahramani. Deep bayesian active learning with image data. In *International Conference on Machine Learning*, pages 1183–1192. PMLR, 2017. 1, 5

[17] Tilmann Gneiting and Adrian E Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American statistical Association*, 102(477):359–378, 2007. 2, 28

[18] Henry Gouk, Eibe Frank, Bernhard Pfahringer, and Michael J Cree. Regularisation of neural networks by enforcing lipschitz continuity. *Machine Learning*, 110(2):393–416, 2021. 13

[19] Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron C Courville. Improved training of wasserstein gans. In *NeurIPS*, 2017. 2, 3

[20] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. *arXiv preprint arXiv:1706.04599*, 2017. 3, 4, 6

[21] Bobby He, Balaji Lakshminarayanan, and Yee Whye Teh. Bayesian Deep Ensembles via the Neural Tangent Kernel. In *Advances in neural information processing systems*, 2020. 8

[22] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 4, 5, 8, 17, 18

[23] Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. *arXiv preprint arXiv:1903.12261*, 2019. 5

[24] Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. *arXiv preprint arXiv:1610.02136*, 2016. 1, 3, 13

[25] Dan Hendrycks, Mantas Mazeika, and Thomas Dietterich. Deep anomaly detection with outlier exposure. In *International Conference on Learning Representations*, 2018. 3, 13

[26] Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Song. Natural adversarial examples. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15262–15271, 2021. 6, 16

[27] José Miguel Hernández-Lobato and Ryan Adams. Probabilistic backpropagation for scalable learning of bayesian neural networks. In *International Conference on Machine Learning*, pages 1861–1869. PMLR, 2015. 3

[28] Geoffrey E Hinton and Drew Van Camp. Keeping the neural networks simple by minimizing the description length of the weights. In *Proceedings of the sixth annual conference on Computational learning theory*, pages 5–13, 1993. 3

[29] Yen-Chang Hsu, Yilin Shen, Hongxia Jin, and Zsolt Kira. Generalized odin: Detecting out-of-distribution image without learning from out-of-distribution data. In *CVPR*, 2020. 13

[30] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017. 5

[31] Haiwen Huang, Zhihan Li, Lulu Wang, Sishuo Chen, Bin Dong, and Xinyu Zhou. Feature space singularity for out-of-distribution detection. In *Proceedings of the Workshop on Artificial Intelligence Safety 2021 (SafeAI 2021)*, 2021. 20, 21

[32] Yu Huang and Yue Chen. Autonomous driving with deep learning: A survey of state-of-art technologies. *arXiv preprint*

arXiv:2006.06091, 2020. 1

[33] Alex Kendall and Yarin Gal. What uncertainties do we need in bayesian deep learning for computer vision? In *Advances in neural information processing systems*, pages 5574–5584, 2017. 3, 7, 22, 23

[34] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015. 5

[35] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. In *ICLR*, 2014. 12

[36] Andreas Kirsch, Clare Lyle, and Yarin Gal. Unpacking information bottlenecks: Unifying information-theoretic objectives in deep learning. *arXiv preprint arXiv:2003.12537*, 2020. 27

[37] Andreas Kirsch, Joost van Amersfoort, and Yarin Gal. Batchbald: Efficient and diverse batch acquisition for deep bayesian active learning. In *Advances in Neural Information Processing Systems*, 2019. 3

[38] Agustinus Kristiadi, Matthias Hein, and Philipp Hennig. Being bayesian, even just a bit, fixes overconfidence in relu networks. In *ICML*, 2020. 13

[39] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009. 5, 17

[40] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Advances in neural information processing systems*, pages 6402–6413, 2017. 1, 2, 3, 6, 13, 15, 21

[41] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998. 2, 17, 18

[42] Kimin Lee, Honglak Lee, Kibok Lee, and Jinwoo Shin. Training confidence-calibrated classifiers for detecting out-of-distribution samples. In *International Conference on Learning Representations*, 2018. 3, 13

[43] Kimin Lee, Kibok Lee, Honglak Lee, and Jinwoo Shin. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. In *NeurIPS*, 2018. 1, 2, 3, 4, 6, 8, 12, 16, 19

[44] Shiyu Liang, Yixuan Li, and R Srikant. Enhancing the reliability of out-of-distribution image detection in neural networks. In *International Conference on Learning Representations*, 2018. 3, 13

[45] Jeremiah Zhe Liu, Zi Lin, Shreyas Padhy, Dustin Tran, Tania Bedrax-Weiss, and Balaji Lakshminarayanan. Simple and principled uncertainty estimation with deterministic deep learning via distance awareness. In *NeurIPS*, 2020. 2, 3, 6, 8, 12, 13, 15, 21

[46] Weitang Liu, Xiaoyun Wang, John Owens, and Yixuan Li. Energy-based out-of-distribution detection. *Advances in Neural Information Processing Systems*, 33, 2020. 3, 5, 6, 12, 15, 16

[47] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440, 2015. 6

[48] David JC MacKay. *Bayesian methods for adaptive models*. PhD thesis, California Institute of Technology, 1992. 3

[49] Wesley J Maddox, Pavel Izmailov, Timur Garipov, Dmitry P Vetrov, and Andrew Gordon Wilson. A simple baseline for bayesian uncertainty in deep learning. In *Advances in Neural Information Processing Systems*, pages 13153–13164, 2019. 20, 21

[50] Andrey Malinin and Mark JF Gales. Predictive uncertainty estimation via prior networks. In *NeurIPS*, 2018. 8, 13

[51] Andrey Malinin, Bruno Mlodozeniec, and Mark John Francis Gales. Ensemble distribution distillation. *ArXiv*, abs/1905.00076, 2020. 13

[52] Takeru Miyato, Toshiki Kataoka, Masanori Koyama, and Yuichi Yoshida. Spectral normalization for generative adversarial networks. In *International Conference on Learning Representations*, 2018. 2, 4, 17

[53] Jishnu Mukhoti and Yarin Gal. Evaluating bayesian deep learning methods for semantic segmentation. *arXiv preprint arXiv:1811.12709*, 2018. 7

[54] Kevin P Murphy. *Machine learning: a probabilistic perspective*. MIT press, 2012. 8, 29, 30

[55] Radford M Neal. *Bayesian learning for neural networks*, volume 118. Springer Science & Business Media, 2012. 3

[56] Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, and Andrew Y Ng. Reading digits in natural images with unsupervised feature learning. 2011. 5

[57] Yaniv Ovadia, Emily Fertig, Jie Ren, Zachary Nado, David Sculley, Sebastian Nowozin, Joshua Dillon, Balaji Lakshminarayanan, and Jasper Snoek. Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift. In *Advances in Neural Information Processing Systems*, pages 13991–14002, 2019. 1, 3

[58] Tim Pearce, Alexandra Brintrup, and Jun Zhu. Understanding softmax confidence and uncertainty, 2021. 2

[59] Janis Postels, Hermann Blum, Cesar Cadena, Roland Siegwart, Luc Van Gool, and Federico Tombari. Quantifying aleatoric and epistemic uncertainty using density estimation in latent space. *arXiv preprint arXiv:2012.03082*, 2020. 2, 3, 8, 12

[60] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations*, 2015. 17, 18

[61] Lewis Smith and Yarin Gal. Understanding Measures of Uncertainty for Adversarial Example Detection. In *UAI*, 2018. 3

[62] Lewis Smith, Joost van Amersfoort, Haiwen Huang, Stephen Roberts, and Yarin Gal. Can convolutional resnets approximately preserve input distances? a frequency analysis perspective, 2021. 4

[63] Chen Sun, Abhinav Shrivastava, Saurabh Singh, and Abhinav Gupta. Revisiting unreasonable effectiveness of data in deep learning era. In *Proceedings of the IEEE international conference on computer vision*, pages 843–852, 2017. 3

[64] Joost van Amersfoort, Lewis Smith, Andrew Jesson, Oscar Key, and Yarin Gal. Improving deterministic uncertainty estimation in deep learning for classification and regression. *arXiv preprint arXiv:2102.11409*, 2021. 3

[65] Joost van Amersfoort, Lewis Smith, Yee Whye Teh, and Yarin Gal. Uncertainty estimation using a single deep deterministic neural network. In *International Conference on Machine Learning*, pages 9690–9700. PMLR, 2020. 2, 3, 4, 6, 8, 12,

13, 15, 17

[66] Yeming Wen, Dustin Tran, and Jimmy Ba. Batchensemble: an alternative approach to efficient ensemble and lifelong learning. In *International Conference on Learning Representations*, 2019. 1, 3

[67] Yeming Wen, Dustin Tran, and Jimmy Ba. Batchensemble: an alternative approach to efficient ensemble and lifelong learning. In *International Conference on Learning Representations*, 2020. 20, 21

[68] Andrew Gordon Wilson and Pavel Izmailov. Bayesian deep learning and a probabilistic perspective of generalization. *arXiv preprint arXiv:2002.08791*, 2020. 8

[69] Jim Winkens, Rudy Bunel, Abhijit Guha Roy, Robert Stanforth, Vivek Natarajan, Joseph R Ledsam, Patricia MacWilliams, Pushmeet Kohli, Alan Karthikesalingam, Simon Kohl, et al. Contrastive training for improved out-of-distribution detection. *arXiv preprint arXiv:2007.05566*, 2020. 2, 12

[70] Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*, 2017. 18

[71] Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. *arXiv preprint arXiv:1605.07146*, 2016. 5