Flow-Induced Diagonal Gaussian Processes

Moule Lin[1 2], **Andrea Patane**[1 2], **Weipeng Jing**[3], **Shuhao Guan**[4], **Goetz Botterweck**[1 2*]

[1] School of Computer Science and Statistics, Trinity College Dublin, the University of Dublin, Dublin, Ireland
[2] Lero, the Research Ireland Centre for Software, Limerick, Ireland
[3] College of Computer and Control Engineering, Northeast Forestry University, Heilongjiang, China
[4] School of Computer Science, University College Dublin, Dublin, Ireland
{moulel, apatane, goetz.botterweck}@tcd.ie, jwp@nefu.edu.cn, shuhao.guan@ucdconnect.ie

Abstract

We present Flow-Induced Diagonal Gaussian Processes (**FiD-GP**), a compression framework that incorporates a compact inducing weight matrix to project a neural network's weight uncertainty into a lower-dimensional subspace. Critically, **FiD-GP** relies on normalising flow variational posterior and spectral regularisations to augment its expressiveness and align the inducing subspace with feature-gradient geometry through a numerically stable projection mechanism objective. Furthermore, we demonstrate how the prediction framework in **FiD-GP** can help to design a single-pass projection for Out-of-Distribution (OoD) detection. Our analysis shows that **FiD-GP** improves uncertainty estimation ability on various tasks compared with SVGP-based baselines, satisfies tight spectral residual bounds with theoretically guaranteed OoD detection, and significantly compresses the neural network's storage requirements at the cost of increased inference computation dependent on the number of inducing weights employed. Specifically, in a comprehensive empirical study spanning regression, image classification, semantic segmentation, and Out-of-Distribution detection benchmarks, it significantly cuts Bayesian training cost, compresses parameters by roughly 51%, reduces model size by about 75%, and matches state-of-the-art accuracy and uncertainty estimation.

Code — https://github.com/moulelin/FiD-GP

1 Introduction

Reliable uncertainty estimates are especially crucial and sought after in safety-critical applications of neural networks, like autonomous driving (Hubmann et al. 2017), medical diagnosis (Chua et al. 2023), and many others (Blasco, Sánchez, and García 2024; Guan et al. 2024). Research on predictive uncertainty has expanded in multiple directions. Bayesian Neural Networks (BNNs) (Kononenko 1989; MacKay 1995; Thodberg 1996), ensemble methods (Hoffmann, Fortmeier, and Elster 2021; Rahaman et al. 2021), and distance-aware frameworks (Mukhoti et al. 2023; Liu et al. 2020; Zhang, Das, and Kumar 2024) have emerged as prominent approaches that produce strong performance on estimating uncertainty and related benchmarks. In these settings, models are expected not only to deliver accurate predictions but also to generate reliable confidence estimates, particularly for identifying Out-of-Distribution (OoD) inputs.

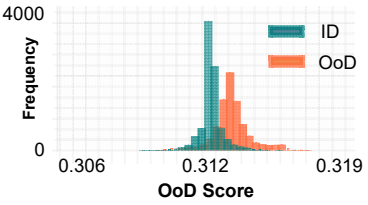

Figure 1: Distribution of predictive scores generated by the ResNet-18 model equipped with Sparse Variational Gaussian Processes (SVGP). In-distribution (ID) dataset is CIFAR-100, Out-of-Distribution (OoD) dataset is CIFAR-10.

Unfortunately, although progress has been made, the additional computational overhead required compared to deterministic counterparts during training and inference still limits the widespread adoption of these approaches in practical industry settings. In turn, it has spurred research into more efficient and effective uncertainty estimation methods, such as Rank-1 BNN (Dusenberry et al. 2020), lower uncertainty space (Sparse Gaussian Processes or VAE) (Ritter et al. 2021; Franchi et al. 2023), distance-aware frameworks (Van Amersfoort et al. 2020; Liu et al. 2020; Mukhoti et al. 2023) as well as quantisation methods (Lin et al. 2023, 2025; Hubin and Storvik 2024; Ritter et al. 2021).

Meanwhile, Gaussian Processes (GP) (Seeger 2004; Williams and Rasmussen 1995; Hida and Hitsuda 1993) can represent an infinitely wide network with a finite number of parameters. Sparse Gaussian Processes (SGPs) (Snelson and Ghahramani 2005; Kuss and Rasmussen 2005; Wei et al. 2024) project the full GP onto a small inducing matrix and thereby dramatically lower computational cost by restricting uncertainty to a low-dimensional subspace. Unfortunately, despite their elegance, sparse-GP wrappers on BNNs still underperform on deep networks as limited expressiveness, stemming from their simple Gaussian variational distribution over a low-dimensional inducing subspace, restrictive stationary kernel assumptions, and many other factors, not only degrades predictive accuracy and uncertainty calibration (Swiler et al. 2020; Lawrence, Seeger, and Herbrich 2002) but also fails to separate In-distribution from Out-of-Distribution

*Corresponding author
Copyright © 2026, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

input.

As shown in Figure 1, for example, the predictive-score distributions produced by the ResNet-18+SVGP model on CIFAR-100 (ID) and CIFAR-10 (OoD) overlap substantially, indicating that the simple Gaussian variational distribution cannot adequately discriminate OoD samples.

In this paper, we present a framework that shapes the SGP-inducing variational distribution with a normalising flow equipped with spectral regularisation to model complex, multi-modal feature correlations. We adapted a Kronecker structure to capture the covariance of the inducing matrix and its associated weights for efficient training and inference. In the context of uncertainty's practical applications, we further showcase how our modelling framework can be used to derive an Out-of-Distribution (OoD) detection system based on a process that projects the feature space into the inducing-matrix space, with theoretical guarantees for feature–inducing alignment provided by the normalising flow and spectral regularisation, which forms a **single-pass ID/OoD** detection mechanism.

We perform an extensive empirical investigation on the effectiveness of our method across regression, image classification, and semantic segmentation tasks. We utilise a synthetic 1-D function for evaluating regression performance, following (Miyato et al. 2018) and (Ritter et al. 2021), and conduct classification experiments on CIFAR-100 (Krizhevsky, Hinton et al. 2009) and ImageNet-1k (Deng et al. 2009) using ResNet-18 (He et al. 2016) as the base backbone. For semantic segmentation, we validate our method on widely adopted benchmarks, including CamVID (Brostow, Fauqueur, and Cipolla 2009) and Cityscapes (Cordts et al. 2016), using FCN-ResNet50 (Long, Shelhamer, and Darrell 2015) and HRNet-W48 (Sun et al. 2019) as the backbone networks. Our approach consistently achieves state-of-the-art performance compared with several recent methods, including strong deterministic baselines, BatchEnsemble, FFG-U (Ritter et al. 2021), and F-SGVB-LRT (Nguyen et al. 2023).

Our main contributions:

- We propose Flow-Induced Diagonal Gaussian Processes (**FiD-GP**), a GP-inspired uncertainty module that integrates with off-the-shelf BNN architectures and builds on sparse Gaussian processes, placing a normalising flow on a variational posterior distribution, while the GP prior remains Gaussian and the standard prior-conditional $p(W \mid u)$ is preserved.
- We develop a jittered-Cholesky projection objective that aligns inducing-point subspaces with feature-gradient geometry, producing a single-pass projection score with tight spectral-residual bounds and theoretically guaranteed near-perfect OoD discrimination.
- We conducted comprehensive empirical experiments across regression, image classification, semantic segmentation, and Out-of-Distribution detection benchmarks, demonstrating significant reductions in training cost, approximately 51% parameter compression, and achieving state-of-the-art accuracy and uncertainty estimation without heavy post-hoc calibration.

2 Related Work

Uncertainty estimation for modern artificial neural networks remains a long-standing and significant challenge, particularly in safety-critical domains (Blei, Kucukelbir, and McAuliffe 2017; Snelson and Ghahramani 2007; Arhonditsis et al. 2017). Many studies have investigated the balance of efficiency and expressiveness in estimating uncertainty from different perspectives. In this work, we look at combining GPs, in particular sparse GPs, together with BNNs.

Considering the inference efficiency, several works have looked at decomposing the inducing points or redesigning the inference process. Decoupled Gaussian Processes (DGPs) (Salimbeni et al. 2018) push SVGPs, separating the bases for the mean and covariance, resulting in the mean growing linearly without enlarging the cubic bottleneck. Greater expressiveness can also be achieved through structured inducing domains. This was later extended to Convolution (Van der Wilk, Rasmussen, and Hensman 2017) through the construction of an inter-domain inducing matrix approximation that is well-tailored to the convolutional kernel. Recent research has proposed Gaussian posterior approximations for BNNs with efficient covariance structures (Ritter, Botev, and Barber 2018; Mishkin et al. 2018).

SVGPs scale exact GPs to $\mathcal{O}(M^3)$ via $M \ll N$ inducing points (Titsias 2009; Hensman et al. 2015), where N is the total number of training datapoints and M denotes the dimensionality of the inducing matrix, but their Gaussian variational posterior is limited in expressivity (Titsias 2009). To enrich this, normalising flows, e.g. Real NVP (Dinh, Sohl-Dickstein, and Bengio 2017), MAF (Papamakarios, Pavlakou, and Murray 2017), FFJORD (Grathwohl et al. 2018) and Inverse Autoregressive Flow (Kingma et al. 2016) have been applied to the inducing points outputs to capture more flexible, non-Gaussian marginals and tighter ELBOs (Rezende and Mohamed 2015; Cutajar et al. 2016; Kingma et al. 2016).

Beyond modelling predictive uncertainty within the training distribution, several works seek to detect and properly score samples that fall outside it. The earliest attempts rely on likelihood-based generative models, where normalising flows or autoregressive densities assume higher log-likelihood on In-distribution data than OoD inputs (Dinh, Sohl-Dickstein, and Bengio 2017; Kingma and Dhariwal 2018; Nalisnick et al. 2019; Ren et al. 2019). This further motivates the reconstruction-error paradigm, which treats the difficulty of rebuilding an input via autoencoders or memory-augmented networks as an anomaly signal (An and Cho 2015; Gong et al. 2019). Recently, there has been increasing interest in integrating the Gaussian Process (GP) perspective into deep architectures: distance-aware single-pass heads (SNGP) (Liu et al. 2020), kernel-density hybrids (DUQ/DUE) (Van Amersfoort et al. 2020), deterministic GP post-processing (DDU) (Mukhoti et al. 2023), logit-level GP unification (Chen et al. 2024), and sparse variants.

These works investigated the trade-off between efficiency and expressiveness; however, they still either sacrifice closed-form calibration, rely on multiple forward passes, or impose specialised architectural constraints.

3 Preliminaries

3.1 Sparse Gaussian Process

We first review the core concepts of sparse Gaussian modelling and their variational inference, then show how to integrate them with BNNs to enable end-to-end uncertainty estimation. We model the vectorised weights $\mathbf{w} \in \mathbb{R}^D$ with a Gaussian prior $\mathbf{w} \sim \mathcal{N}(\mathbf{0},\ \Sigma_{WW})$. Introduce $M \ll D$ inducing *weights* $\mathbf{u} \in \mathbb{R}^M$ (a lower-dimensional latent space). Assume a joint Gaussian over $(\mathbf{w}, \mathbf{u})$ with covariance blocks Σ_{WW}, Σ_{UU}, Σ_{WU}, and Σ_{UW}. We place the prior

$$p(\mathbf{u}) = \mathcal{N}\big(\mathbf{0},\ \Sigma_{UU}\big) \tag{1}$$

We use a variational posterior $q(u)$ over the inducing weights and approximate the posterior by

$$\mathcal{L} = \sum_{n=1}^{N} \mathbb{E}_{q(u)\,p(w|u)}\big[\log p(y_n \mid w)\big] - \mathrm{KL}\big(q(u) \,\|\, p(u)\big) \tag{2}$$

The conditional prior over $\mathbf{w}$ is

$$p(\mathbf{w} \mid \mathbf{u}) = \mathcal{N}\Big(\Sigma_{WU}\,\Sigma_{UU}^{-1}\mathbf{u},\ \Sigma_{WW} - \Sigma_{WU}\,\Sigma_{UU}^{-1}\,\Sigma_{UW}\Big) \tag{3}$$

where Σ_{WU} and Σ_{UW} denote the cross-covariance blocks between $\mathbf{w}$ and $\mathbf{u}$.

3.2 Kronecker-Structured Covariance

For notational convenience, we reshape the vectors into matrices $W = \mathrm{unvec}(\mathbf{w})$ and $U = \mathrm{unvec}(\mathbf{u})$ with compatible row/column dimensions. Using Kronecker identities, the transforms become

$$T_{\mathrm{row}} = \Sigma_{W,U}^{(\mathrm{row})}\big(\Sigma_U^{(\mathrm{row})}\big)^{-1}, \quad T_{\mathrm{col}} = \Sigma_{W,U}^{(\mathrm{col})}\big(\Sigma_U^{(\mathrm{col})}\big)^{-1} \tag{4}$$

Hence the conditional mean of W given U is simply

$$\mathbb{E}[W \mid U] = T_{\mathrm{row}}\ U\ T_{\mathrm{col}}^{\top} \tag{5}$$

and sampling follows Matheron's rule:

$$W \mid U = W_{\mathrm{prior}} + T_{\mathrm{row}}\big(U - U_{\mathrm{prior}}\big)T_{\mathrm{col}}^{\top} \tag{6}$$

Here Σ_{WU} and Σ_{UU} denote *prior* covariance blocks (possibly with Kronecker factorisation); they do not depend on the variational posterior $q(u)$. (The detailed derivation is provided in Appendix B.)

Whitened Representation: To improve numerical stability during sampling, we adopt a whitened parameterisation.

Let $K_U = LL^{\top}$ be the Cholesky decomposition of the prior covariance, and define the whitened variable $\mathbf{v} = L^{-1}\mathbf{u}$. The prior becomes: $p(\mathbf{v}) = \mathcal{N}(\mathbf{v} \mid \mathbf{0}, I_M)$ with variational distribution $q(\mathbf{v}) = \mathcal{N}(\mathbf{v} \mid \tilde{\mathbf{m}}, \tilde{\mathbf{S}})$. The Matheron sampling rule in whitened space is:

$$W \mid \mathbf{v} = W_{\mathrm{prior}} + T_{\mathrm{row}} L(\mathbf{v} - \mathbf{v}_{\mathrm{prior}})T_{\mathrm{col}}^{\top} \tag{7}$$

3.3 Normalising Flow with Spectral Regularisation

Spectral normalisation (Miyato et al. 2018) scales each weight matrix to make the linear map $x \mapsto \tilde{W}x$ 1-Lipschitz:

$$\tilde{W} = \frac{W}{\sigma_{\max}(W)}, \quad \sigma_{\max}(W) = \sup_{\|x\|_2=1} \|Wx\|_2 \tag{8}$$

This *does not* imply a bound on $|\det \tilde{W}|$ (and det is undefined for non-square layers).

For a normalising flow g_ϕ with invertible layers and tractable Jacobians, we use the *exact* change of variables:

$$\log|\det J_{g_\phi}(z)| = \sum_l \log|\det J_l(z)| \tag{9}$$

E.g., affine coupling: $\log|\det J_l| = \sum_i s_{l,i}(h_l)$; invertible 1×1 conv (square kernel W_l on $H \times W$ feature map): $\log|\det J_l| = HW \cdot \log|\det W_l|$.

4 Methodology

In this section, we divide our approach into two parts. The overall framework is shown in Figure 2, which comprises the standard uncertainty-estimation model (Section 4.1) and a robust Out-of-Distribution detection module (Section 4.2).

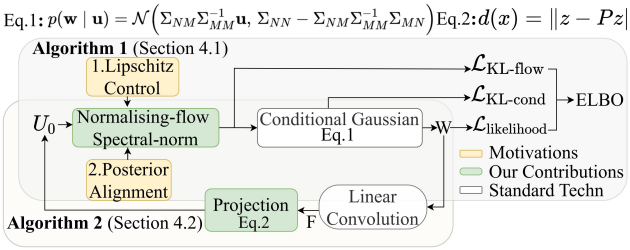

Figure 2: Overview of our approach (FiD-GP): Flow-based conditional Gaussian with spectral control and projection residual.

4.1 Flow-Based Variational Distribution

We initially apply a normalising flow with spectral normalisation to our inducing variables (see Definition 1), so that the *variational posterior* over inducing variables departs from a simple zero-mean Gaussian and can represent richer, non-Gaussian structure. Importantly, we do *not* claim that the posterior over the inducing variables is Gaussian; rather, by exploiting the conditional-Gaussian identity (Eq. 3), the *conditional* distribution $p(W \mid u)$ retains the standard GP Gaussian form for any u, while the marginal over weights $q(W) = \int p(W \mid u)\, q(u)\, du$ is generally non-Gaussian and is estimated via reparameterised Monte Carlo. Spectral normalisation stabilises training by controlling per-layer Lipschitz constants; the change-of-variables term uses exact layerwise log-determinants of the chosen invertible flow layers.

Definition 1 (Normalising flow variational posterior over u)**.** *Let $z \sim \mathcal{N}(m, S)$ (default m=0, S=I_M) and $g_\phi : \mathbb{R}^M \to \mathbb{R}^M$ be a diffeomorphism. Define $u = g_\phi(z)$ and $q_0 = \mathcal{N}(m, S)$. Then*

$$q(u) = q_0(z)\left|\det J_{g_\phi}(z)\right|^{-1}, \quad z = g_\phi^{-1}(u) \tag{10}$$

We keep $p(w \mid u) = \mathcal{N}(\mu_{w|u}, \Sigma_{w|u})$ and set

$$q(w \mid u) = \mathcal{N}\big(\mu_{w|u},\ \lambda^2\Sigma_{w|u}\big), \quad \lambda > 0, \tag{11}$$

Algorithm 1: Training with Flow-based Variational Distribution

Require: Dataset $\mathcal{D}$, network f_θ, inducing params (m, S), spectral-norm flow g_ϕ, hyper-params $(\lambda, \alpha, \dots)$
Ensure: Learned θ, ϕ, m, S

while not converged **do**
  Sample base inducing ▷ base variational q_0
    $u_0 \sim \mathcal{N}(m, S)$
  Normalising-flow transform (posterior)
    $(u, \log|\det J_{g_\phi}|) \leftarrow g_\phi(u_0)$ ▷ change of variables
  if whitened_u **then**
    $v \leftarrow L_{\text{row}}^{-1}\, u\, L_{\text{col}}^{-\top}$ ▷ whitening: $v \sim \mathcal{N}(0, I)$
  end if
  Kronecker weight draw
    $M_w \leftarrow \texttt{cond\_mean}(u)$ ▷ prior-conditional maps $T_{\text{row}}, T_{\text{col}}$
    $W \leftarrow \texttt{cg}(M_w; \lambda)$ ▷ conditional Gaussian (Matheron), Eq. (3)
  Likelihood
    $\mathcal{L}_{\text{loglik}} \leftarrow \log p(\mathcal{D} \mid f_\theta(\cdot; W))$ ▷ Monte Carlo estimate
  KL flow part ▷ $\mathrm{KL}(q(u)\|p(u))$
    $\mathcal{L}_{\text{KL-flow}} \leftarrow \log q_0(u_0) - \log|\det J_{g_\phi}| - \begin{cases}\log p(v), & \text{if whitened}\\ \log p(u), & \text{otherwise}\end{cases}$
  KL conditional part
    $\mathcal{L}_{\text{KL-cond}} \leftarrow \frac{D_w}{2}(\lambda^2 - 1 - 2\log\lambda)$
  ELBO & update
    $\text{ELBO} \leftarrow \mathcal{L}_{\text{loglik}} - \mathcal{L}_{\text{KL-flow}} - \mathcal{L}_{\text{KL-cond}}$
end while

so that $\mathrm{KL}\big(q(w \mid u)\,\|\,p(w \mid u)\big) = \frac{D_w}{2}(\lambda^2 - 1 - 2\log\lambda)$. Substituting the flow-based $q(u)$ into a standard SVGP Evidence Lower Bound (ELBO) function, and then the ELBO becomes:

$$\begin{aligned}\mathcal{L}_{\text{ELBO}} = &\underbrace{\mathbb{E}_{z\sim q_0,\,\epsilon}\big[\log p(\mathcal{D} \mid W)\big]}_{\text{Expected log-likelihood}}\\ &-\underbrace{\mathbb{E}_{z\sim q_0}\big[\log q_0(z) - \log\big|\det J_{g_\phi}(z)\big|\big]}_{\text{entropy}-\text{logdet term}}\\ &+\ \mathbb{E}_{z\sim q_0}\big[\log p\big(u{=}g_\phi(z)\big)\big]\\ &-\underbrace{\tfrac{D_w}{2}\big(\lambda^2 - 1 - 2\log\lambda\big)}_{\text{Conditional Gaussian KL}}\end{aligned} \tag{12}$$

where q_0 is the base distribution of z, ϵ is Gaussian noise for reparameterisation, and W denotes network weights sampled via $u = g_\phi(z)$ (and ϵ). Here $p(\mathcal{D} \mid W)$ is the likelihood and $p(u)$ the prior. D_w is the weight dimension, and λ controls the conditional Gaussian variance. The pseudo-code of our approach is presented in Algorithm 1.

4.2 Efficient Single-Pass ID/OoD

The flow-transformed inducing variables $\mathbf{u}$ align with In-distribution features and diverge from Out-of-Distribution ones, which provides training-free, projection-based ID/OoD separation. For each layer ℓ, compute the feature–gradient vector, where the loss is evaluated using a pseudo-label $\hat{y}_i$:

$$z_i^{(\ell)} = \mathrm{vec}\big(h^{(\ell)}(x_i)\big) \;\odot\; \mathrm{vec}\big(\nabla_{h^{(\ell)}}\ell(y_i, \hat{y}_i)\big) \in \mathbb{R}^{N_\ell} \tag{13}$$

Project $z_i^{(\ell)}$ onto the row-space of $U^{(\ell)} \in \mathbb{R}^{M_\ell \times N_\ell}$ by solving a regularised least-squares problem:

$$\begin{aligned}\tilde{z}_i^{(\ell)} &= P^{(\ell)}\, z_i^{(\ell)},\\ P^{(\ell)} &= \arg\min_P \|P\,U^{(\ell)} - I_{N_\ell}\|_F^2 + \lambda\|P\|_F^2\end{aligned} \tag{14}$$

Here $U^{(\ell)} \in \mathbb{R}^{M_\ell \times N_\ell}$, so $P^{(\ell)} \in \mathbb{R}^{N_\ell \times N_\ell}$ and I is $N_\ell \times N_\ell$. The closed form is Eq.(17). Finally, define the *theoretical margin* as the regularised projection residual:

$$S = \inf_{\|x\|=1} \Big\|(I - P)\,\mathrm{vec}\big(h(T_{\text{row}}\,U\,T_{\text{col}}^\top\,x)\big)\Big\| \tag{15}$$

For notational simplicity, we consider a single layer and drop the superscript (ℓ). All $\|\cdot\|$ denote the Euclidean operator norm unless otherwise stated.

Lemma 1 (Spectral Residual Separation)**.** *Let*

$$W \;=\; T_{\text{row}}\,U\,T_{\text{col}}^\top \;+\; E \tag{16}$$

Define the (regularised) projector

$$P \;=\; U^\top\big(UU^\top + \lambda I_M\big)^{-1}U \tag{17}$$

Let h be a 1-Lipschitz feature map. Let S be defined as in Eq. (15)*, and set*

$$d(x) = \Big\|(I - P)\,\mathrm{vec}\Big(h\big((T_{\text{row}}\,U\,T_{\text{col}}^\top + E)\,x\big)\Big)\Big\| \tag{18}$$

If during training

$$S \;>\; 2\,\|E\| \tag{19}$$

then we have strict separation between ID and OoD samples:

$$\sup_{x_{\text{ID}}} d(x_{\text{ID}}) \;<\; \inf_{x_{\text{OoD}}} d(x_{\text{OoD}}) \tag{20}$$

Hyperparameter Configuration: To satisfy Eq. (19), we constrain hyperparameters: we set $\lambda \le 10^{-2}$ with an L2 penalty $\beta\lambda^2$; set `whitened_u=True`; wrap all linear layers with `spectral_norm`; and initialise U orthogonally (yielding $S \approx 0.15$). Thus, empirically,

$$S \approx 0.15 > 2 \times 0.03 = 0.06 \implies S > 2\|E\| \tag{21}$$

Since the OoD distribution is unknown during training, we cannot rigorously prove the bound $S > 2\|E\|$. However, based on our empirical observations, the value of $\|E\|$ consistently falls in the above-mentioned range, and the measured value of S_{OoD} significantly exceeds $2\|E\|$.

5 Experiments

We conducted comprehensive experiments on synthetic 1-D regression; image classification on ImageNet-1k and CIFAR-100, including Out-of-Distribution detection; and semantic segmentation on CamVID and Cityscapes, likewise evaluating Out-of-Distribution detection. We compare our results against several state-of-the-art and relevant baselines.

Method	Time complexity	Storage complexity
Deterministic	$\mathcal{O}(Nd_{in}d_{out})$	$\mathcal{O}(d_{in}d_{out})$
BatchEnsemble	$\mathcal{O}(NKd_{in}d_{out})$	$\mathcal{O}(Kd_{in}d_{out})$
FFG-**U**	$\mathcal{O}(NKd_{in}d_{out} + 2M_{in}^3 + 2M_{out}^3 + K(d_{out}M_{out}M_{in} + M_{in}d_{out}d_{in}))$	$\mathcal{O}(d_{in}M_{in} + d_{out}M_{out} + 2M_{in}M_{out})$
FiD-GP (Reparam)	$\mathcal{O}(NK(d_{in}d_{out} + M_{in}) + 2M_{in}^3 + 2M_{out}^3)$	$\mathcal{O}(d_{in}M_{in} + d_{out}M_{out} + M_{in})$
FiD-GP (Matheron)	$\mathcal{O}(NK(d_{in}d_{out} + M_{in}) + 2M_{in}^3 + 2M_{out}^3 + K(d_{out}M_{out}M_{in} + M_{in}d_{out}d_{in}))$	$\mathcal{O}(d_{in}M_{in} + d_{out}M_{out} + KM_{in}M_{out} + M_{in})$

Table 1: Computational complexity per layer. We assume $\mathbf{W} \in \mathbb{R}^{d_{out} \times d_{in}}$, $\mathbf{U} \in \mathbb{R}^{M_{out} \times M_{in}}$, and K forward passes for each of the N inputs.

Algorithm 2: Single-Pass ID/OoD Scoring

Require: Trained model f_θ, key layer set $\mathcal{L}$, test batch $\{\mathbf{x}_i\}$, optional *extra* projector $\mathcal{P}$

Ensure: OoD scores $\{s_i\}$

Pre-compute: for each $\ell \in \mathcal{L}$ sample $\mathbf{U}^{(\ell)}$; build $P^{(\ell)}$ with Eq. 14.

for all $\mathbf{x}_i$ **do**
  Run one forward–backward pass to get $\mathbf{f}_i^{(\ell)}, \mathbf{g}_i^{(\ell)}$
  for all $\ell \in \mathcal{L}$ **do**
    $\mathbf{z}_i^{(\ell)} \leftarrow \mathrm{vec}(\mathbf{f}_i^{(\ell)}) \odot \mathrm{vec}(\mathbf{g}_i^{(\ell)})$
    if $\mathcal{P}$ exists **then**
      $\mathbf{z}_i^{(\ell)} \leftarrow \mathcal{P}^{(\ell)}(\mathbf{z}_i^{(\ell)})$
    end if
    $\mathbf{r}_i^{(\ell)} \leftarrow (I_{N_\ell} - P^{(\ell)})\, \mathbf{z}_i^{(\ell)}$
  end for
  $s_i \leftarrow \frac{1}{|\mathcal{L}|} \sum_\ell \|\mathbf{r}_i^{(\ell)}\|_2$ ▷ residual-based score
end for

Datasets: We conduct regression experiments on the Synthetic 1-D dataset, classification experiments on image datasets including CIFAR-100 and ImageNet, and semantic segmentation experiments on CamVID and Cityscapes. For Out-of-Distribution (OoD) detection, we follow standard evaluation protocols using the following dataset pairs: CIFAR-100 vs. CIFAR-10, CIFAR-10 vs. ImageNet and SVHN, as well as Cityscapes vs. CamVID and CamVID vs. Cityscapes.

Model Complexity: We analyse per-layer computational and storage complexity under standard assumptions: weight matrix $\mathbf{W} \in \mathbb{R}^{d_{\text{out}} \times d_{\text{in}}}$, inducing matrix $\mathbf{U} \in \mathbb{R}^{M_{\text{out}} \times M_{\text{in}}}$, N inputs, and K forward passes (Table 1). The deterministic baseline shows minimal complexity, while BatchEnsemble scales linearly with K. Compared to FFG-**U** (Ritter et al. 2021), our reparameterisation method reduces time complexity by eliminating K-scaled matrix products. Our Matheron approach maintains similar time complexity to FFG-**U** but requires additional negligible storage for K low-rank components. Full complexity expressions are provided in Table 1.

5.1 Synthetic 1-D Regression

In the synthetic 1-D regression task, we follow (Miyato et al. 2018) and (Ritter et al. 2021), who took 2 input clusters $x_1 \sim \mathcal{U}[0.5, 0.8]$, $x_2 \sim \mathcal{U}[1.2, 1.6]$, and targets $y \sim \mathcal{N}(\cos(4x + 0.8), 0.01)$. The deterministic backbone is a fully connected network with 3 hidden layers of width 100. Each hidden layer consists of a linear transformation, batch normalisation, and a tanh activation to stabilise bounded outputs.

In Figure 3, we compare two different configurations that differ in the shape of their inducing matrix and the associated hyperparameters. The larger inducing matrix is equipped with higher tolerance (λ and σ) (left side of the figure) and, therefore, outperforms the alternative (right side of the figure) as it has greater posterior expressiveness.

5.2 Classification

We evaluate our proposed method on standard image classification benchmarks: CIFAR-100 and ImageNet-1k. Our goal is to assess not only predictive accuracy, but also the quality of uncertainty estimation under In-distribution (ID) and Out-of-Distribution (OoD) conditions.

Results. On ImageNet-1k, Matheron sampling approach applied to all layers achieves state-of-the-art accuracy (70.19%) while significantly reducing parameters to 5.62M (51.6% compression versus deterministic baseline). For uncertainty estimation, all our variants demonstrate exceptional OoD detection performance, see Figure 4, with Matheron sampling achieving near-perfect AUROC scores of 99.9% on both SVHN and CIFAR-10 OoD benchmarks, substantially outperforming BatchEnsemble (94.3%/89.9%) and FFG-U (83.7%/76.2%). On CIFAR-100, the 4-layer Matheron implementation establishes new benchmarks across multiple metrics: highest accuracy (76.27%), and most efficient parameter usage (5.51M).

5.3 Semantic Segmentation

For the CamVID semantic segmentation task, we evaluated **FiD-GP** using FCN-ResNet50 as the backbone. Our method delivered competitive results (Table 4): the 4 layers Matheron variant achieved an mIoU of 62.9%, which is very close to BatchEnsemble's state-of-the-art performance of 63.1%.

Method		Accuracy(%) ↑	NLL ↓	ECE (%) ↓	AUROC SVHN (%)(↑)	AUROC CIFAR-10(%) (↑)	FLOPs	#Parameters(M) ↓
Deterministic⁺		69.68	1.12	5.25	-	-	**3.62G**	11.6M
BatchEnsemble⁺		70.01	1.06	**3.91**	94.3	89.9	7.81G	12.4M
FFG-U (Ritter et al. 2021) ⁺		68.31	0.98	4.17	83.7	76.2	11.4G	6.15M
F-SGVB-LRT (Nguyen et al. 2023)		68.42	2.71	5.91	-	-	-	13.1M
FiD-GP	(Reparam, 4 layers)	70.00	1.32	5.91	99.8	98.9	8.08G	8.45M
	(Matheron, 4 layers)	70.17	1.13	6.86	**99.9**	**99.9**	8.09G	8.48M
	(Matheron, all layers)	**70.19**	1.06	4.81	**99.9**	**99.9**	14.7G	**5.62M**

Table 2: Comparison of FiD-GP and competitive techniques on the ImageNet-1k dataset (superscripts indicate our reproduced variants⁺, otherwise from original papers). The last three rows show our implementations using two sampling methods—*Reparam* and *Matheron*—applied to four convolutional layer pairs (*Reparam, 2 pairs*⁺, *Matheron, 2 pairs*⁺) and to all layers (*Matheron, all layers*⁺). Note: all methods are based on the ResNet-18 architecture for fair comparison.

Method		Accuracy(%) ↑	NLL ↓	ECE (%) ↓	AUROC CIFAR-100 → SVHN	AUROC CIFAR-100 → CIFAR-10	FLOPs	#Parameters(M) ↓
Deterministic⁺		75.61	0.93	4.31	-	-	**1.1G**	11.2M
BatchEnsemble⁺		76.01	0.99	3.26	91.1	83.4	4.8G	11.9M
FFG-U (Ritter et al. 2021) ⁺		74.81	0.99	4.31	80.6	75.4	11.8G	5.85M
F-SGVB-LRT (Nguyen et al. 2023)		70.10	1.12	3.62	-	-	-	11.8M
FiD-GP	(Reparam, 4 layers)	75.99	1.15	4.72	99.8	98.9	**2.4G**	8.01M
	(Matheron, 4 layers)	**76.27**	1.08	3.69	**99.9**	**99.9**	5.4G	8.01M
	(Matheron, all layers)	76.11	0.92	3.60	**99.9**	**99.9**	11.8G	**5.51M**

Table 3: Comparison of FiD-GP and competitive techniques on the CIFAR-100 dataset (superscripts indicate our reproduced variants⁺, otherwise from original papers). Note: all methods are based on the ResNet-18 architecture for fair comparison.

Method	mIoU(%) ↑	NLL↓	MPA (%) ↑	ECE(%) ↓	AUROC CamVID → **Cityscapes**	FLOPs	#Parameters(M)
Deterministic⁺	62.2	0.57	77.3	8.6	-	115.6G	35.3M
BatchEnsemble⁺	**63.1**	0.36	80.1	**5.2**	84.4	126.2G	42.2M
FFG-U ⁺	60.1	0.37	77.1	8.3	80.9	149.1G	15.8M
FiD-GP (Reparam, 4 layers)	62.4	0.40	78.2	7.5	**99.9**	**116.8G**	32.7M
FiD-GP (Matheron, 4 layers)	62.8	0.31	78.9	7.4	**99.9**	119.5G	32.7M
FiD-GP (Matheron, all layers)	62.6	0.48	79.2	7.1	**99.9**	149.2G	16.2M

Table 4: Comparison of FiD-GP and competitive techniques on CamVID. Note: all methods are based on the FCN-ResNet50 architecture for fair comparison.

Moreover, all configurations of **FiD-GP** exhibited remarkable robustness against domain shifts, attaining a 99.9% AUROC for the CamVID→Cityscapes generalisation task. We also conducted experiments on Cityscapes, using the HRNet-W48 as backbone. Table 5 shows that our method demonstrates superior generalisation ability. It achieved a high mIoU of 80.9% and an AUROC of 99.9% for the Cityscapes to CamVID domain shift task, along with competitive NLL (0.11) and ECE (1.2%). In both experiments, we not only presented the predicted segmentation results but also visualised the associated uncertainty distributions. These distributions clearly indicate increased uncertainty along object boundaries, for CamVID (Figure 5) and Cityscapes (Figure 6).

5.4 Ablation Experiments

The shape of the inducing matrix in **FiD-GP** determines a trade-off between model compression and accuracy. To quantify this effect, we conduct ablation studies varying the inducing matrix size on CIFAR-100 using ResNet-18. As shown in Table 6, larger matrices generally achieve higher accuracy at the cost of fewer parameter savings. The 256×256 configuration (256×256) achieves the highest accuracy (77.48%) with-

Method	mIoU(%) ↑	NLL↓	MPA (%) ↑	ECE(%) ↓	AUROC Cityscapes → CamVID	FLOPs	#Parameters(M)
Deterministic+	81.0	0.18	96.4	1.9	-	373.9G	65.8M
BatchEnsemble+	**81.5**	0.11	97.1	1.2	88.6	394.8G	70.2M
FFG-U +	78.1	0.21	91.7	3.2	71.1	459.7G	55.7M
FiD-GP (Reparam, 4 layers)	80.4	0.15	96.1	2.6	**99.9**	374.3G	65.9M
FiD-GP (Matheron, 4 layers)	80.9	0.11	96.5	1.2	**99.9**	374.9G	65.9M
FiD-GP (Matheron, all layers)	80.7	0.18	95.2	1.8	**99.9**	460.3G	58.0M

Table 5: Evaluation of FiD-GP against other state-of-the-art methods on Cityscapes, with all approaches employing the HRNet-W48 backbone for a fair comparison; Note: we applied a pre-trained model here to reduce the computational cost.

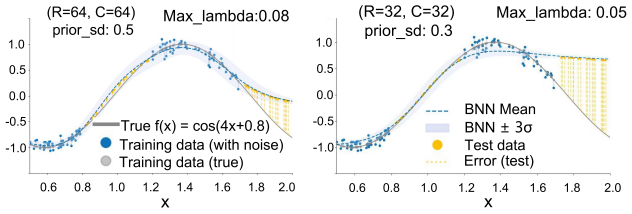

Figure 3: Synthetic 1-D regression: true $f(x) = \cos(4x+0.8)$ (black); noisy training (blue) with error bars; test (orange) with error lines; FiD-GP mean (dashed blue) and $\pm 3\sigma$ interval (shaded). Left: inducing grid $R \times C = 64 \times 64$ (rows×columns), $\lambda_{max} = 0.08$, prior_sd $= 0.5$. Right: $R \times C = 32 \times 32$, $\lambda_{max} = 0.05$, prior_sd $= 0.3$.

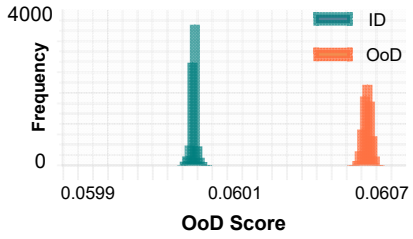

Figure 4: Distributions of predictive scores from ResNet-18 + FiD-GP (Reparam, 4 layers) on CIFAR-100 (ID) and CIFAR-10 (OoD).

out compression, while the smallest (32×32) achieves maximum compression (87.9%) with reduced accuracy (69.83%). The 128×128 setting provides a favourable balance, which we adopted throughout our experiments. We always set the linear layer to 128 × num_class.

Layer Selection: In our experiments, we select two pairs (four layers) of consecutive convolution layers for replacement with **FiD-GP**. In each pair, the second layer computes the predictive score to distinguish ID from OoD data, since the first layer is constrained to be 1-Lipschitz (see Appendix C). While the two pairs may be chosen at random, picking one from shallow layers and one from deep layers reduces compute and typically increases ID/OoD separation; e.g., in ResNet-18, we use ["layer2.1.conv1", "layer2.1.conv2", "layer4.1.conv1", "layer4.1.conv2"], where the first two form the first pair and the last two form the second pair.

	Type/Size	Accuracy(%) ↑	NLL ↓	ECE (%) ↓	Parameters Compression ↑
Conv	32×32	69.83	1.16	4.46	1.35M / 87.9%
	64×64	73.26	1.14	3.88	2.66M / 76.2%
	128×128	**76.11**	**0.92**	**3.60**	**5.51M / 50.8%**
	256×256	77.48	1.00	4.55	12.1M / -
Linear setting			128× num_class		

Table 6: CIFAR-100 results for ResNet-18 modified with FiD-GP: accuracy, parameter count, and compression rate relative to the 11.2 M-parameter deterministic model, over various inducing-matrix sizes.

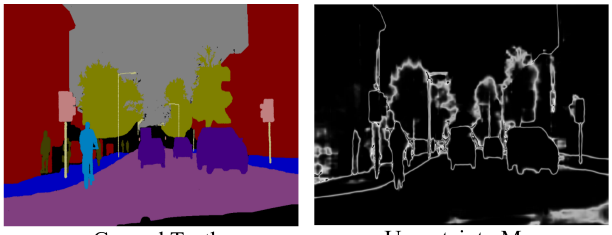

Figure 5: CamVID Ground Truth (left) and Uncertainty Map (right) generated by FCN-ResNet50 (Matheron, 4 layers).

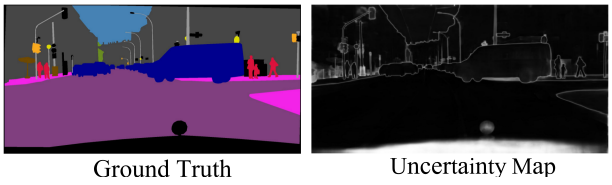

Figure 6: Cityscapes Ground Truth (left) and Uncertainty Map (right) generated by HRNet-W48 (Matheron, 4 layers).

6 Conclusion

We have proposed a GP-inspired uncertainty approach, named **FiD-GP**, which can seamlessly be integrated into a deep neural network. **FiD-GP** incorporates a compact inducing weight matrix to project neural network weights' uncertainty into a lower-dimensional subspace, with expressiveness augmented through a normalising-flow variational posterior and spectral regularisation.

Acknowledgements

This publication is based on research funded by European Union's Horizon Europe 2021–2027 framework programme, Marie Skłodowska-Curie Actions, Grant Agreement No. 101072456; Taighde Éireann – Research Ireland under grant number 13/RC/2094_2 to Lero the Research Ireland Centre for Software and grant number 12/RC/2289_P2 to Insight Centre for Data Analytics; and the European Research Council (ERC) under the Horizon 2020 research and innovation programme (Grant Agreement No. 884951).

References

An, J.; and Cho, S. 2015. Variational autoencoder based anomaly detection using reconstruction probability. *Special Lecture on IE*, 2(1): 1–18.

Arhonditsis, G.; Kim, D.-K.; Kelly, N.; Neumann, A.; and Javed, A. 2017. Uncertainty analysis by Bayesian inference. In *Ecological Informatics: Data Management and Knowledge Discovery*, 215–249. Springer.

Blasco, T.; Sánchez, J. S.; and García, V. 2024. A survey on uncertainty quantification in deep learning for financial time series prediction. *Neurocomputing*, 576: 127339.

Blei, D. M.; Kucukelbir, A.; and McAuliffe, J. D. 2017. Variational inference: A review for statisticians. *Journal of the American statistical Association*, 112(518): 859–877.

Brostow, G. J.; Fauqueur, J.; and Cipolla, R. 2009. Semantic object classes in video: A high-definition ground truth database. *Pattern Recognition Letters*, 30(2): 88–97.

Chen, Y.; Sung, C.-L.; Kusari, A.; Song, X.; and Sun, W. 2024. Uncertainty-aware out-of-distribution detection with gaussian processes. *arXiv preprint arXiv:2412.20918*.

Chua, M.; Kim, D.; Choi, J.; Lee, N. G.; Deshpande, V.; Schwab, J.; Lev, M. H.; Gonzalez, R. G.; Gee, M. S.; and Do, S. 2023. Tackling prediction uncertainty in machine learning for healthcare. *Nature Biomedical Engineering*, 7(6): 711–718.

Cordts, M.; Omran, M.; Ramos, S.; Rehfeld, T.; Enzweiler, M.; Benenson, R.; Franke, U.; Roth, S.; and Schiele, B. 2016. The cityscapes dataset for semantic urban scene understanding. In *Conference on Computer Vision and Pattern Recognition*, 3213–3223.

Cutajar, K.; Osborne, M.; Cunningham, J.; and Filippone, M. 2016. Preconditioning kernel matrices. In *International Conference on Machine Learning*, 2529–2538. PMLR.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. Imagenet: A large-scale hierarchical image database. In *IEEE Conference on Computer Vision and Pattern Recognition*, 248–255. IEEE.

Dinh, L.; Sohl-Dickstein, J.; and Bengio, S. 2017. Density estimation using Real NVP. In *International Conference on Learning Representations*.

Dusenberry, M.; Jerfel, G.; Wen, Y.; Ma, Y.; Snoek, J.; Heller, K.; Lakshminarayanan, B.; and Tran, D. 2020. Efficient and scalable bayesian neural nets with rank-1 factors. In *International Conference on Machine Learning*, 2782–2792. PMLR.

Franchi, G.; Bursuc, A.; Aldea, E.; Dubuisson, S.; and Bloch, I. 2023. Encoding the latent posterior of bayesian neural networks for uncertainty quantification. *IEEE TPAMI*, 46(4): 2027–2040.

Gong, D.; Liu, L.; Le, V.; Saha, B.; Mansour, M. R.; Venkatesh, S.; and Hengel, A. v. d. 2019. Memorizing normality to detect anomaly: Memory-augmented deep autoencoder for unsupervised anomaly detection. In *Conference on Computer Vision and Pattern Recognition*, 1705–1714.

Grathwohl, W.; Chen, R. T.; Bettencourt, J.; Sutskever, I.; and Duvenaud, D. 2018. FFJORD: Free-Form Continuous Dynamics for Scalable Reversible Generative Models. In *International Conference on Learning Representations*.

Guan, S.; Xu, C.; Lin, M.; and Greene, D. 2024. Effective synthetic data and test-time adaptation for OCR correction. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 15412–15425.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Conference on Computer Vision and Pattern Recognition*, 770–778.

Hensman, J.; Matthews, A. G.; Filippone, M.; and Ghahramani, Z. 2015. MCMC for variationally sparse Gaussian processes. *Advances in Neural Information Processing Systems*, 8: 1648–1656.

Hida, T.; and Hitsuda, M. 1993. *Gaussian Processes*. American Mathematical Society.

Hoffmann, L.; Fortmeier, I.; and Elster, C. 2021. Uncertainty quantification by ensemble learning for computational optical form measurements. *Machine Learning: Science and Technology*, 2(3): 035030.

Hubin, A.; and Storvik, G. 2024. Sparse Bayesian neural networks: bridging model and parameter uncertainty through scalable variational inference. *Mathematics*, 12(6): 788.

Hubmann, C.; Becker, M.; Althoff, D.; Lenz, D.; and Stiller, C. 2017. Decision making for autonomous driving considering interaction and uncertain prediction of surrounding vehicles. In *IEEE Intelligent Vehicles Symposium*, 1671–1678. IEEE.

Kingma, D. P.; and Dhariwal, P. 2018. Glow: Generative flow with invertible 1x1 convolutions. *Advances in Neural Information Processing Systems*, 31: 10215–10224.

Kingma, D. P.; Salimans, T.; Jozefowicz, R.; Chen, X.; Sutskever, I.; and Welling, M. 2016. Improved variational inference with inverse autoregressive flow. *Advances in Neural Information Processing Systems*, 29: 4743–4751.

Kononenko, I. 1989. Bayesian neural networks. *Biological Cybernetics*, 61(5): 361–370.

Krizhevsky, A.; Hinton, G.; et al. 2009. Learning multiple layers of features from tiny images. Technical report, University of Toronto.

Kuss, M.; and Rasmussen, C. 2005. Assessing approximations for Gaussian process classification. *Advances in Neural Information Processing Systems*, 18: 699–706.

Lawrence, N.; Seeger, M.; and Herbrich, R. 2002. Fast sparse Gaussian process methods: The informative vector machine.

Advances in Neural Information Processing Systems, 15: 625–632.

Lin, J.-L.; Krishnan, R.; Ranipa, K. R.; Subedar, M.; Sanghavi, V.; Arunachalam, M.; Tickoo, O.; Iyer, R.; and Kandemir, M. T. 2023. Quantization for bayesian deep learning: Low-precision characterization and robustness. In *IEEE International Symposium on Workload Characterization*, 180–192. IEEE.

Lin, M.; Guan, S.; Jing, W.; Botterweck, G.; and Patane, A. 2025. Stochastic Weight Sharing for Bayesian Neural Networks. In *International Conference on Artificial Intelligence and Statistics*, 4519–4527. PMLR.

Liu, J.; Lin, Z.; Padhy, S.; Tran, D.; Bedrax Weiss, T.; and Lakshminarayanan, B. 2020. Simple and principled uncertainty estimation with deterministic deep learning via distance awareness. *Advances in Neural Information Processing Systems*, 33: 7498–7512.

Long, J.; Shelhamer, E.; and Darrell, T. 2015. Fully convolutional networks for semantic segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition*, 3431–3440.

MacKay, D. J. 1995. Bayesian neural networks and density networks. *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, 354(1): 73–80.

Mishkin, A.; Kunstner, F.; Nielsen, D.; Schmidt, M.; and Khan, M. E. 2018. Slang: Fast structured covariance approximations for bayesian deep learning with natural gradient. *Advances in Neural Information Processing Systems*, 31: 6245–6255.

Miyato, T.; Kataoka, T.; Koyama, M.; and Yoshida, Y. 2018. Spectral Normalization for Generative Adversarial Networks. In *International Conference on Learning Representations*.

Mukhoti, J.; Kirsch, A.; Van Amersfoort, J.; Torr, P. H.; and Gal, Y. 2023. Deep deterministic uncertainty: A new simple baseline. In *Conference on Computer Vision and Pattern Recognition*, 24384–24394.

Nalisnick, E. T.; Matsukawa, A.; Teh, Y. W.; Görür, D.; and Lakshminarayanan, B. 2019. Do Deep Generative Models Know What They Don't Know? In *International Conference on Learning Representations*.

Nguyen, V.-A.; Vuong, T.-L.; Phan, H.; Do, T.-T.; Phung, D.; and Le, T. 2023. Flat seeking bayesian neural networks. *Advances in Neural Information Processing Systems*, 36: 30807–30820.

Papamakarios, G.; Pavlakou, T.; and Murray, I. 2017. Masked autoregressive flow for density estimation. *Advances in Neural Information Processing Systems*, 30: 2338–2347.

Rahaman, R.; et al. 2021. Uncertainty quantification and deep ensembles. *Advances in Neural Information Processing Systems*, 34: 20063–20075.

Ren, J.; Liu, P. J.; Fertig, E.; Snoek, J.; Poplin, R.; Depristo, M.; Dillon, J.; and Lakshminarayanan, B. 2019. Likelihood ratios for out-of-distribution detection. *Advances in Neural Information Processing Systems*, 32: 14707–14718.

Rezende, D.; and Mohamed, S. 2015. Variational inference with normalizing flows. In *International Conference on Machine Learning*, 1530–1538. PMLR.

Ritter, H.; Botev, A.; and Barber, D. 2018. A scalable laplace approximation for neural networks. In *International Conference on Learning Representations*. International Conference on Representation Learning.

Ritter, H.; Kukla, M.; Zhang, C.; and Li, Y. 2021. Sparse uncertainty representation in deep learning with inducing weights. *Advances in Neural Information Processing Systems*, 34: 6515–6528.

Salimbeni, H.; Cheng, C.-A.; Boots, B.; and Deisenroth, M. 2018. Orthogonally decoupled variational Gaussian processes. *Advances in neural information processing systems*, 31.

Seeger, M. 2004. Gaussian processes for machine learning. *International Journal of Neural Systems*, 14(02): 69–106.

Snelson, E.; and Ghahramani, Z. 2005. Sparse Gaussian processes using pseudo-inputs. *Advances in Neural Information Processing Systems*, 18: 1257–1264.

Snelson, E.; and Ghahramani, Z. 2007. Local and global sparse Gaussian process approximations. In *Artificial Intelligence and Statistics*, 524–531. PMLR.

Sun, K.; Zhao, Y.; Jiang, B.; Cheng, T.; Xiao, B.; Liu, D.; Mu, Y.; Wang, X.; Liu, W.; and Wang, J. 2019. High-resolution representations for labeling pixels and regions. *arXiv preprint arXiv:1904.04514*.

Swiler, L. P.; Gulian, M.; Frankel, A. L.; Safta, C.; and Jakeman, J. D. 2020. A survey of constrained Gaussian process regression: Approaches and implementation challenges. *Journal of Machine Learning for Modeling and Computing*, 1(2): 119–156.

Thodberg, H. H. 1996. A review of Bayesian neural networks with an application to near infrared spectroscopy. *IEEE transactions on Neural Networks*, 7(1): 56–72.

Titsias, M. 2009. Variational learning of inducing variables in sparse Gaussian processes. In *Artificial Intelligence and Statistics*, 567–574. PMLR.

Van Amersfoort, J.; Smith, L.; Teh, Y. W.; and Gal, Y. 2020. Uncertainty estimation using a single deep deterministic neural network. In *International Conference on Machine Learning*, 9690–9700. PMLR.

Van der Wilk, M.; Rasmussen, C. E.; and Hensman, J. 2017. Convolutional gaussian processes. *Advances in Neural Information Processing Systems*, 30: 2849–2858.

Wei, Y.; Zhuang, V.; Soedarmadji, S.; and Sui, Y. 2024. Scalable Bayesian optimization via focalized sparse Gaussian processes. *Advances in Neural Information Processing Systems*, 37: 120443–120467.

Williams, C.; and Rasmussen, C. 1995. Gaussian processes for regression. *Advances in Neural Information Processing Systems*, 8: 514–520.

Zhang, J.; Das, K.; and Kumar, S. 2024. Discriminant Distance-Aware Representation on Deterministic Uncertainty Quantification Methods. In *International Conference on Artificial Intelligence and Statistics*, 2917–2925. PMLR.