Network Inversion for Uncertainty-Aware Out-of-Distribution Detection

Anonymous Author(s)
Affiliation
Address
email

Abstract

Out-of-distribution (OOD) detection and uncertainty estimation (UE) are critical components for building safe machine learning systems, especially in real-world scenarios where unexpected inputs are inevitable. However the two problems have, until recently, separately been addressed. In this work, we propose a novel framework that combines network inversion with classifier training to simultaneously address both OOD detection and uncertainty estimation. For a standard n-class classification task, we extend the classifier to an (n+1)-class model by introducing a "garbage" class, initially populated with random gaussian noise to represent outlier inputs. After each training epoch, we use network inversion to reconstruct input images corresponding to all output classes that initially appear as noisy and incoherent and are therefore excluded to the garbage class for retraining the classifier. This cycle of training, inversion, and exclusion continues iteratively till the inverted samples begin to resemble the in-distribution data more closely, with a significant drop in the uncertainty, suggesting that the classifier has learned to carve out meaningful decision boundaries while sanitising the class manifolds by pushing OOD content into the garbage class. During inference, this training scheme enables the model to effectively detect and reject OOD samples by classifying them into the garbage class. Furthermore, the confidence scores associated with each prediction can be used to estimate uncertainty for both in-distribution and OOD inputs. Our approach is scalable, interpretable, and does not require access to external OOD datasets or post-hoc calibration techniques while providing a unified solution to the dual challenges of OOD detection and uncertainty estimation.

1 Introduction

The increasing deployment of machine learning models in high-stakes, real-world applications—such as autonomous driving, medical diagnosis, and financial decision-making—has underscored the importance of model reliability and robustness. A key limitation of modern neural networks is their tendency to produce overconfident predictions Suhail and Sethi [2025] even on inputs that lie far outside the training distribution. This makes it crucial to develop models capable of both out-of-distribution (OOD) detection—the ability to identify inputs that fall outside the training distribution—and uncertainty estimation (UE)—the ability to quantify confidence in predictions to ensure safe decision-making under distributional shift.

Both capabilities are vital for trustworthiness in deployment scenarios where the data encountered during inference may deviate from the training distribution in subtle or unexpected ways. Although these two problems are inherently linked, most existing approaches treat them separately, often relying on post-hoc calibration techniques or auxiliary OOD datasets, which may not always be available.

Submitted to 39th Conference on Neural Information Processing Systems (NeurIPS 2025). Do not distribute.

In this work, we propose a novel framework that leverages network inversionSuhail and Sethi [2024], not only to detect OOD inputs but also to estimate prediction uncertainty, unifying the two objectives in a single training procedure. By extending a standard (n+1)-class model with an auxiliary garbage class, and iteratively refining the model using inverted reconstructions, we encourage the network to carve out clean decision boundaries while isolating ambiguous or anomalous regions. Unlike prior approaches, our method requires no external OOD datasets or post-hoc calibration, offering a simple and interpretable solution to ensure robustness in classification under distributional shift.

2 Prior Work

Inversion attempts to reconstruct inputs that elicit desired outputs or internal activations of a neural network. Early studies on multilayer perceptrons applied gradient-based inversion to visualize decision rules, but these often yielded noisy or adversarial-like images Kindermann and Linden [1990], Jensen et al. [1999], Saad and Wunsch [2007]. Evolutionary search and constrained optimization were explored as alternatives Wong [2017]. Later work introduced prior-based regularization, including smoothness constraints and pretrained generative models, to improve realism and interpretability of reconstructions Mahendran and Vedaldi [2014], Yosinski et al. [2015], Mordvintsev et al. [2015], Nguyen et al. [2016, 2017]. The connection to adversarial examples has been emphasized, as unconstrained inversion can converge to adversarial artifacts Szegedy et al. [2014], Goodfellow et al. [2015]. In contrast, adversarially robust classifiers tend to produce more human-aligned features Tsipras et al. [2019], Engstrom et al. [2019], enabling more interpretable reconstructions Santurkar et al. [2019]. Recent advances include learning surrogate loss landscapes to stabilize inversion Liu et al. [2022], and generative methods that conditionally reconstruct inputs likely to produce a given output Suhail and Sethi [2024]. Alternative formulations recast inversion into logical reasoning frameworks, encoding classifiers into CNF constraints for deterministic reconstruction Suhail [2024].

Uncertainty quantification (UQ) has emerged as a cornerstone of reliable AI systems, particularly in domains where overconfident false predictions can lead to critical failures. Post-hoc methods are attractive because they can be retrofitted to pretrained deterministic classifiers without requiring retraining. Monte Carlo Dropout (MC Dropout) Gal and Ghahramani [2016] introduces stochasticity during inference to approximate Bayesian model averaging, while temperature scaling Guo et al. [2017] improves calibration with a single scalar parameter applied to logits. More recently, auxiliary prediction heads or meta-models have been explored. Evidential Deep Learning Sensoy et al. [2018] reformulates classification into the prediction of Dirichlet parameters, providing both predictive means and uncertainty. Direct Epistemic Uncertainty Prediction (DEUP) Jain et al. [2022] learns a secondary model to estimate generalization error from data embeddings. Later, Shen et al. [2023] proposes evidential meta-models that generate Dirichlet distributions from classifier logits.

Bayesian neural networks (BNNs) Neal [1996] Blundell et al. [2015] and related variational inference techniques offer a more principled alternative by maintaining posterior distributions over network weights. Ensemble learning remains one of the most empirically effective strategies for UQ. Deep Ensembles Lakshminarayanan et al. [2017] aggregate predictions from independently trained networks and consistently achieve strong calibration and robustness under distributional shift. Domain-specific strategies include test-time augmentation to approximate prediction variance, uncertainty-aware segmentation masks to enhance interpretability Jungo et al. [2020], and Bayesian approximations adapted to volumetric imaging Kwon et al. [2020]. Shen et al. [2023] proposed evidential meta-models trained on classifier embeddings to predict Dirichlet distributions, enabling decomposition into epistemic and aleatoric uncertainty. Jain et al. [2022] generalized this idea with DEUP to out-of-distribution and low-data regimes, while Bala et al. [2025] introduced BAY-MED, a Dirichlet meta-model for breast cancer classification that demonstrates robustness to OOD samples.

Recent work in Ansari et al. [2022] proposed Autoinverse, a framework for neural network inversion that prioritizes solutions near reliable training samples, using embedded regularization and predictive uncertainty minimization to improve robustness. Later Lu et al. [2023] introduced a semantically coherent OOD detection (SCOOD) approach by combining uncertainty-aware optimal transport with dynamic cost modeling and inter-cluster enhancements. While Chen et al. [2024] developed a Gaussian process-based model that operates solely on in-distribution data. Similarly, Charpentier et al. [2020] presents PostNet, which employs normalizing flows to model posterior distributions over predicted probabilities, allowing reliable uncertainty estimation and effective OOD discrimination—even without OOD supervision.

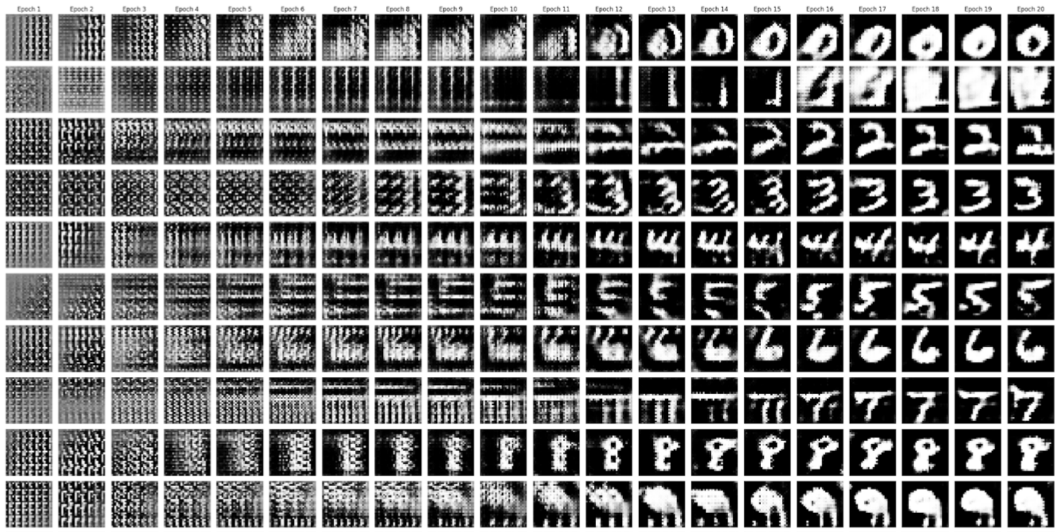

Figure 1: Inverted Samples across epochs for different classes, beginning to resemble the training data as OODs are excluded into garbage class.

3 Methodology

Our unified training approach integrates out-of-distribution (OOD) detection and uncertainty estimation (UE) into a single framework using network inversion and an auxilary garbage class. For an n-class classification task, we extend the classifier to an (n+1)-class model by introducing an additional "garbage" class designed to absorb anomalous inputs. This garbage class is initially populated with random Gaussian noise, representing OOD samples.

Between successive training epochs, we perform network inversion as in Suhail and Sethi [2024] to reconstruct samples from the input space of the classifier for all output classes. Formally, we train a conditional generator $\mathcal{G}_\phi : \mathcal{Z} \times \mathbb{R}^K \to \mathcal{X}$, parameterized by ϕ, to invert the classifier's behavior by optimizing it to minimize a composite loss

$$\mathcal{L}_{\text{Inv}} = \alpha \cdot \mathcal{L}_{\text{KL}} + \beta \cdot \mathcal{L}_{\text{CE}} + \gamma \cdot \mathcal{L}_{\text{Cosine}}$$

where, $\mathcal{L}_{\text{KL}}$ is the KL Divergence loss, $\mathcal{L}_{\text{CE}}$ is the Cross Entropy loss, and $\mathcal{L}_{\text{Cosine}}$ is the Cosine Similarity loss. The hyperparameters α, β, γ control the contribution of each individual loss term defined as:

$$\mathcal{L}_{\text{KL}} = \sum_i P(i) \log \frac{P(i)}{Q(i)}, \quad \mathcal{L}_{\text{CE}} = -\sum_i y_i \log(\hat{y}_i), \quad \mathcal{L}_{\text{Cosine}} = \frac{1}{N(N-1)} \sum_{i \neq j} \cos(\theta_{ij})$$

where $\mathcal{L}_{\text{KL}}$ represents the KL Divergence between the input distribution P and the output distribution Q, y_i is the set encoded label, $\hat{y}_i$ is the predicted label from the classifier, and $\cos(\theta_{ij})$ is the cosine similarity between the features of generated images i and j in a batch of N.

Given the vastness of the input space, during early training stages, these reconstructions tend to be visually incoherent and do not resemble real data, reflecting the model's incomplete or uncertain understanding of the class manifolds. These reconstructions are assigned to the garbage class and added to the training set for the subsequent epochs. In subsequent epochs the classifier is trained using a weighted cross-entropy loss to account for the class imbalance introduced by addition of garbage samples.

By iteratively repeating this cycle of training, inversion, and exclusion, the model gradually learns to refine the decision boundaries while pushing anomalous content into the garbage class. As the training progresses, inverted samples in Fig 1 begin to look like training data, indicating that the classifier has effectively carved out the in-distribution manifold while isolating outliers into the garbage class.

During inference, this training procedure equips the classifier to identify and reject out-of-distribution (OOD) inputs by assigning them to the garbage class. Additionally, the softmax confidence scores corresponding to class predictions can be used to assess the model's uncertainty. Low softmax confidence on in-distribution predictions indicates ambiguous or uncertain inputs, while high confidence in the garbage class suggests a strong belief that the input is OOD. We quantify uncertainty using the softmax confidence values across all $n + 1$ output classes by capturing how sharply peaked or spread out the model's predictive distribution is. The uncertainty estimate for a prediction $\mathbf{p}$ is given by:

$$\mathrm{UE}(\mathbf{p}) = 1 - \frac{\sum_{i=1}^{n+1} \left(p_i - \frac{1}{n+1}\right)^2}{\sum_{i=1}^{n+1} \left(\delta_{i,k} - \frac{1}{n+1}\right)^2} \tag{1}$$

where $k = \arg\max_i p_i$ and $\delta_{i,k}$ is the Kronecker delta. The resulting score ranges from 0 to 1, providing an interpretable measure of confidence by computing the squared distance between the predicted vector $\mathbf{p}$ and the uniform distribution, normalized by the maximum possible distance under a one-hot prediction.

4 Quantitative Results

We evaluate the effectiveness of our approach to uncertainty-aware out-of-distribution detection across four benchmark image classification datasets: MNIST [Deng, 2012], FashionMNIST [Xiao et al., 2017], SVHN, and CIFAR-10 [Krizhevsky et al.]. To assess OOD detection performance, we follow a one-vs-rest evaluation strategy: the model is trained exclusively on one dataset and evaluated on the remaining three as OOD sources.

Table 1: Accuracy for both in and out-of-distribution datasets.

Train \ Test	**MNIST**	**FMNIST**	**SVHN**	**CIFAR-10**
MNIST	99.1	89.5	99.1	99.4
FMNIST	85.2	92.6	96.3	95.7
SVHN	93.6	94.9	89.4	87.6
CIFAR-10	97.8	95.7	88.2	85.5

Table 1 presents the accuracy for uncertainty-aware OOD detection across all pairs of datasets. Each row corresponds to a model trained on one of the datasets and diagonal entries represent the in-distribution (ID) performance measured on the standard test set of the training dataset. Off-diagonal entries indicate OOD detection performance, where the accuracy represents how well the model distinguishes out-of-distribution samples by correctly classifying them into the garbage class. High values across both diagonal and off-diagonal entries demonstrate that the model maintains strong classification performance on ID data while reliably identifying OOD inputs.

We also observe that while the majority of OOD samples are correctly assigned to the garbage class, a small percentage of the samples can still be misclassified into in-distribution classes. However, a significant finding is that the least confidently classified in-distribution sample is still more confidently classified compared to the most confidently misclassified out-of-distribution sample, suggesting the existence of a clear threshold.

5 Conclusion

In conclusion, our unified framework seamlessly integrates out-of-distribution (OOD) detection and uncertainty estimation (UE) by extending the classification model with a garbage class and leveraging network inversion for inverted sample generation. Through iterative training and inversion cycles, the model learns to delineate in-distribution data from anomalous inputs while progressively refining its class boundaries. This approach enables robust OOD rejection and provides interpretable uncertainty estimates based on softmax confidence distributions. Future work can also consider the use n garbage classes—one for each of the in-distribution classes—for fine-grained separation of OOD samples and weighted individual OOD sample contribution to the loss while retraining the classifier based on uncertainty.

References

Navid Ansari, Hans-Peter Seidel, Nima Vahidi Ferdowsi, and Vahid Babaei. Autoinverse: Uncertainty aware inversion of neural networks. In *Advances in Neural Information Processing Systems(NeurIPS)*, 2022.

Gouranga Bala, Abhimanyu Chauhan, and Amit Sethi. Bay-med: Bayesian approximation for post-hoc uncertainty in medical imaging. In *2025 IEEE 22nd International Symposium on Biomedical Imaging (ISBI)*, pages 1–4, 2025. doi: 10.1109/ISBI60581.2025.10981251.

Charles Blundell, Julien Cornebise, Koray Kavukcuoglu, and Daan Wierstra. Weight uncertainty in neural network. In Francis Bach and David Blei, editors, *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 1613–1622, Lille, France, 07–09 Jul 2015. PMLR. URL `https://proceedings.mlr.press/v37/blundell15.html`.

Bertrand Charpentier, Daniel Zügner, and Stephan Günnemann. Posterior network: Uncertainty estimation without ood samples via density-based pseudo-counts. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Yang Chen, Chih-Li Sung, Arpan Kusari, Xiaoyang Song, and Wenbo Sun. Uncertainty-aware out-of-distribution detection with gaussian processes. *arXiv:2412.20918*, 2024.

Li Deng. The mnist database of handwritten digit images for machine learning research. *IEEE Signal Processing Magazine*, 29(6):141–142, 2012.

Logan Engstrom, Brandon Tran, Dimitris Tsipras, Ludwig Schmidt, and Aleksander Madry. A rotation and a translation suffice: Fooling CNNs with simple transformations, 2019. URL `https://openreview.net/forum?id=BJfvknCqFQ`.

Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In Maria Florina Balcan and Kilian Q. Weinberger, editors, *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pages 1050–1059, New York, New York, USA, 20–22 Jun 2016. PMLR. URL `https://proceedings.mlr.press/v48/gal16.html`.

Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples, 2015. URL `https://arxiv.org/abs/1412.6572`.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. *CoRR*, abs/1706.04599, 2017. URL `http://arxiv.org/abs/1706.04599`.

Moksh Jain, Salem Lahlou, Hadi Nekoei, Victor I Butoi, Paul Bertin, Jarrid Rector-Brooks, Maksym Korablyov, and Yoshua Bengio. DEUP: Direct epistemic uncertainty prediction, 2022. URL `https://openreview.net/forum?id=Jep2ykGUdS`.

C.A. Jensen, R.D. Reed, R.J. Marks, M.A. El-Sharkawi, Jae-Byung Jung, R.T. Miyamoto, G.M. Anderson, and C.J. Eggen. Inversion of feedforward neural networks: algorithms and applications. *Proceedings of the IEEE*, 87(9):1536–1549, 1999. doi: 10.1109/5.784232.

Alain Jungo, Fabian Balsiger, and Mauricio Reyes. Analyzing the quality and challenges of uncertainty estimations for brain tumor segmentation. *Frontiers in Neuroscience*, 14:282, 2020. ISSN 1662-453X. doi: 10.3389/fnins.2020.00282. URL `https://doi.org/10.3389/fnins.2020.00282`.

J Kindermann and A Linden. Inversion of neural networks by gradient descent. *Parallel Computing*, 14(3):277–286, 1990. ISSN 0167-8191. doi: https://doi.org/10.1016/0167-8191(90)90081-J. URL `https://www.sciencedirect.com/science/article/pii/016781919090081J`.

Alex Krizhevsky, Vinod Nair, and Geoffrey Hinton. Cifar-10 (canadian institute for advanced research). URL `http://www.cs.toronto.edu/~kriz/cifar.html`.

Yongchan Kwon, Joong-Ho Won, Beom Joon Kim, and Myunghee Cho Paik. Uncertainty quantification using bayesian neural networks in classification: Application to biomedical image segmentation. *Comput. Stat. Data Anal.*, 142(C), February 2020. ISSN 0167-9473. doi: 10.1016/j.csda.2019.106816. URL `https://doi.org/10.1016/j.csda.2019.106816`.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 6405–6416, Red Hook, NY, USA, 2017. Curran Associates Inc. ISBN 9781510860964.

Ruoshi Liu, Chengzhi Mao, Purva Tendulkar, Hao Wang, and Carl Vondrick. Landscape learning for neural network inversion, 2022. URL `https://arxiv.org/abs/2206.09027`.

Fan Lu, Kai Zhu, Wei Zhai, Kecheng Zheng, and Yang Cao. Uncertainty-aware optimal transport for semantically coherent out-of-distribution detection. In *Computer Vision and Pattern Recognition (CVPR)*, 2023.

Aravindh Mahendran and Andrea Vedaldi. Understanding deep image representations by inverting them, 2014. URL `https://arxiv.org/abs/1412.0035`.

Alexander Mordvintsev, Christopher Olah, and Mike Tyka. Inceptionism: Going deeper into neural networks, 2015. URL `https://research.googleblog.com/2015/06/inceptionism-going-deeper-into-neural.html`.

Radford M. Neal. *Bayesian Learning for Neural Networks*, volume 118 of *Lecture Notes in Statistics*. Springer, New York, NY, 1 edition, 1996. ISBN 978-0-387-94724-2. doi: 10.1007/978-1-4612-0745-0.

Anh Nguyen, Alexey Dosovitskiy, Jason Yosinski, Thomas Brox, and Jeff Clune. Synthesizing the preferred inputs for neurons in neural networks via deep generator networks, 2016. URL `https://arxiv.org/abs/1605.09304`.

Anh Nguyen, Jeff Clune, Yoshua Bengio, Alexey Dosovitskiy, and Jason Yosinski. Plug & play generative networks: Conditional iterative generation of images in latent space, 2017. URL `https://arxiv.org/abs/1612.00005`.

Emad W. Saad and Donald C. Wunsch. Neural network explanation using inversion. *Neural Networks*, 20(1):78–93, 2007. ISSN 0893-6080. doi: https://doi.org/10.1016/j.neunet.2006.07.005. URL `https://www.sciencedirect.com/science/article/pii/S0893608006001730`.

Shibani Santurkar, Dimitris Tsipras, Brandon Tran, Andrew Ilyas, Logan Engstrom, and Aleksander Madry. Image synthesis with a single (robust) classifier, 2019. URL `https://arxiv.org/abs/1906.09453`.

Murat Sensoy, Lance Kaplan, and Melih Kandemir. Evidential deep learning to quantify classification uncertainty, 2018. URL `https://arxiv.org/abs/1806.01768`.

Maohao Shen, Yuheng Bu, Prasanna Sattigeri, Soumya Ghosh, Subhro Das, and Gregory Wornell. Post-hoc uncertainty learning using a dirichlet meta-model. In *Proceedings of the Thirty-Seventh AAAI Conference on Artificial Intelligence and Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence and Thirteenth Symposium on Educational Advances in Artificial Intelligence*, AAAI'23/IAAI'23/EAAI'23. AAAI Press, 2023. ISBN 978-1-57735-880-0. doi: 10.1609/aaai.v37i8.26167. URL `https://doi.org/10.1609/aaai.v37i8.26167`.

Pirzada Suhail. Network inversion of binarised neural nets. In *The Second Tiny Papers Track at ICLR 2024*, 2024. URL `https://openreview.net/forum?id=zKcB0vb7qd`.

Pirzada Suhail and Amit Sethi. Network inversion of convolutional neural nets. In *Muslims in ML Workshop co-located with NeurIPS 2024*, 2024. URL `https://openreview.net/forum?id=f9sUu7U1Cp`.

Pirzada Suhail and Amit Sethi. Network inversion for generating confidently classified counterfeits, 2025. URL `https://arxiv.org/abs/2503.20187`.

Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks, 2014. URL https://arxiv.org/abs/1312.6199.

Dimitris Tsipras, Shibani Santurkar, Logan Engstrom, Alexander Turner, and Aleksander Madry. Robustness may be at odds with accuracy, 2019. URL https://arxiv.org/abs/1805.12152.

Eric Wong. Neural network inversion beyond gradient descent. In *WOML NIPS*, 2017. URL https://api.semanticscholar.org/CorpusID:208231247.

Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms, 2017. URL https://arxiv.org/abs/1708.07747.

Jason Yosinski, Jeff Clune, Anh Nguyen, Thomas Fuchs, and Hod Lipson. Understanding neural networks through deep visualization, 2015. URL https://arxiv.org/abs/1506.06579.