Exploring Covariate and Concept Shift for Out-of-Distribution Detection

Junjiao Tian
Georgia Institute of Technology
jtian73@gatech.edu

Yen-Chang Hsu
Samsung Research America
yenchang.hsu@samsung.com

Yilin Shen
Samsung Research America
yilin.shen@samsung.com

Hongxia Jin
Samsung Research America
hongxia.jin@samsung.com

Zsolt Kira
Georgia Institute of Technology
zkira@gatech.edu

Abstract

Traditionally the task of out-of-distribution (OOD) detection is associated with epistemic uncertainty estimation. While Bayesian models and generative models have been explored to provide principled uncertainty estimation, deterministic and discriminative models cannot provide such estimation by nature. We propose to characterize the difficulty of OOD detection by the extent of *distribution shift* and theoretically derive two score functions for OOD detection. Additionaly we propose a geometrically-inspired method (Geometric ODIN) to improve OOD detection under distribution shift with only in-distribution data. View project page at https://sites.google.com/view/geometric-decomposition and a long version of the paper at https://arxiv.org/abs/2110.15231.

1 Introduction

Bayesian inference [1, 2, 3] has been regarded as the most principled method of uncertainty modeling [4] because it explicitly models two types of uncertainty: *epistemic uncertainty* and *aleatoric uncertainty* in the form posteriors over parameters and data likelihood respectively. Practically, epistemic uncertainty is often assessed by a model's out-of-distribution (OOD) detection performance [5] or calibration [6, 7], while aleatoric uncertainty can be assessed by in-distribution error detection [5]. Recent attempts [6, 7] to model uncertainty using deterministic models failed to disentangle these two uncertainties due to their non-Bayesian nature. However, it is still possible to capture them empirically in a deterministic model using a combination of density estimation and softmax-entropy [5]. This leaves us the question: **how to approach OOD detection/calibration for *deterministic* (as opposed to Bayesian) and *discriminative* (as opposed to generative) models?** This is arguably the most widely used class of models due to its speed (compared to Bayesian models) and simplicity (compared to generative models). It seems that the conventional association of OOD data with epistemic uncertainty [3] fails under the scope of this type of models and a different perspective is needed to analyze them. A suitable alternative perspective is *distribution shift* [8]. Intuitively, out-of-distribution data refers to data that are sampled from distributions different from the training distribution. There are two dominant shift types: *covariate shift* and *concept shift*. The former usually refers to change in style, e.g., *clean* vs. *noised*, and the latter refers to change in semantics, *dog* vs. *leopard*. We propose to characterize OOD datasets by the degree of shifts in each dimension.

35th Conference on Neural Information Processing Systems (NeurIPS 2021), Sydney, Australia.

While existing works do not distinguish between different types of OOD datasets, these two types of shifts should be examined and addressed separately. We follow [8] for the formal definition of dataset shift, covariate shift and concept shift. Let $X \in \mathcal{R}^D$ denote the covariate which is the input or feature and $Y \in \mathcal{R}$ denote the output label. *Dataset shift* happens when the training joint distribution is not equal to the testing joint distribution $P_{tr}(X, Y) \neq P_{tst}(X, Y)$. Specifically, *covariate shift* appears when $P_{tr}(Y|X) = P_{tst}(Y|X)$ and $P_{tr}(X) \neq P_{tst}(X)$. *Concept shift* appears when $P_{tr}(Y|X) \neq P_{tst}(Y|X)$ and $P_{tr}(X) = P_{tst}(X)$. The first goal of this paper is to derive score functions to reflect the shift in either $P(X)$ or $P(Y|X)$. We denote the score function that reflects changes from $P_{tr}(X)$ to $P_{tst}(X)$ as **covariate shift score function**: $g(x) : \mathcal{R}^D \rightarrow \mathcal{R}$ and the score function that reflects changes from $P_{tr}(Y|X)$ to $P_{tst}(Y|X)$ as **concept shift score function**: $h(y, x) : \mathcal{R} \times \mathcal{R}^D \rightarrow \mathcal{R}$. The second goal is to improve the sensitivity of these scores to their corresponding distribution shift. Specifically, we propose a geometrically inspired **o**ut-of-distribution **d**etection method with only **in**-distribution data (**Geometric ODIN**). We use CIFAR10C [9] and a new CIFAR100 Splits dataset to investigate gradual covariate shift and concept respectively.

2 Related Work

Out-of-Distribution (OOD) detection methods can be largely divided into two camps depending on whether they require OOD data during training. [10, 11, 12] leverages anomalous data in training. Our method belongs to the class of methods that do not assume the availability of OOD data during training. [13] uses the maximum softmax probability (MSP) to detect incorrect predictions and OOD data. [14] proposes to use Mahalanobis distance by fitting a Gaussian mixture model (GMM) in the feature space. [5] uses log density of the GMM model instead. [15] uses an energy score as the uncertainty metric. ODIN [16] uses a combination of input processing and post-training tuning to improve OOD detection performance. Generalized ODIN [17] (also [18]) includes an additional network in the last layer to improve OOD detection during training. There are other interesting OOD detection approaches without OOD data such as using contrastive learning with various transformations [19, 20], training a deep ensemble of multiple models [21] and leveraging large pretrained models [22]. They require extended training time, hyperparameter tuning and careful selections of transformations, whereas our method does not introduce any hyperparameters and has negligible influence on standard cross-entropy training time.

3 Method

3.1 Covariate and Concept Score Functions

In this section, we theoretically derive two score functions, $g(x)$ and $h(y, x)$, based on the KL-divergence between a uniform distribution $\mathbf{U}$ and a predicted distribution $\mathbf{P} \in \mathcal{R}^M$, where M is the number of classes. By starting from KL-divergence, we hinge the subsequent derivation of score functions on a physical meaningful uncertainty measure, i.e., how far the predicted distribution is from a uniform distribution. This relationship ensures a natural interpretation of score functions because predictions on distribution shifted data should have larger uncertainty, i.e. smaller distance from uniform. We are specifically interested in softmax-linear models for classification. They typically consist of a feature extractor and a linear layer followed by a softmax activation. Let $\mathbf{f} \in \mathcal{R}^D$ denote a feature vector from the feature extractor[1]. The output of the linear layer, i.e., logits, $\mathbf{l} \in \mathcal{R}^M =< \mathbf{f}, \mathbf{W} >$ is defined as the inner product between the feature vector and a weight matrix in the linear layer. Let $l_i = ||\mathbf{f}||_2 ||\mathbf{w}_i||_2 \cos \phi_i$ denote the ith logit, $P_i = \frac{\exp l_i}{\sum_{j=1}^{M} \exp l_j}$ denote the predicted probability of the ith class. The KL-divergence $\mathbb{KL}(\mathbf{U}||\mathbf{P})$ can be written as following:

$$\mathbb{KL}(\mathbf{U}||\mathbf{P}) = -\sum_{i=1}^{M} \frac{1}{M} \ln MP_i = \ln \sum_{j=1}^{M} \exp l_j - \frac{1}{M} \sum_{i=1}^{M} l_i - \ln M \tag{1}$$

[1] Bold letter indicates vectors

Now we can use the inequality property of Log-Sum-Exp (LSE)[2] functions in Eq. 2 to bound Eq. 1.

$$\max_j l_j \leq \ln \sum_{j=1}^{M} \exp l_j \leq \max_j l_j + \ln M \tag{2}$$

Therefore the KL-divergence (Eq. 1) can be bounded as follows:

$$\mathcal{U} - \ln M \leq \mathbb{KL}(\mathbf{U}||\mathbf{P}) \leq \mathcal{U} \tag{3}$$

where $\mathcal{U} = \max_j l_j - \frac{1}{M}\sum_{i=1}^{M} l_i$. $\mathcal{U}$ can be further decomposed into two multiplicative components by plugging in the definition of logits l_i:

$$\mathcal{U} = \max_j l_j - \frac{1}{M}\sum_{i=1}^{M} l_i = \overbrace{\|\mathbf{f}\|_2}^{g(x)} \underbrace{\left(\max_j \|\mathbf{w}_j\|_2 \cos\phi_j - \frac{1}{M}\sum_{i=1}^{M} \|\mathbf{w}_i\|_2 \cos\phi_i \right)}_{h(y,x)} \tag{4}$$

We define the **covariate shift score function** as $g(x) \triangleq \|\mathbf{f}\|_2$ because the norm of a feature vector is the sum of squared activation values and only depends on the input. Intuitively, activation of a neural network on covariate-shifted data should be weaker than in-distribution data. Therefore, $g(x)$ assigns a higher value to in-distribution data than to OOD data. We define the **concept shift score function** as $h(y,x) \triangleq \max_j \|\mathbf{w}_j\|_2 \cos\phi_j - \frac{1}{M}\sum_{i=1}^{M} \|\mathbf{w}_i\|_2 \cos\phi_i$ because it is the difference between the cosine distance of the *predicted* class and the average cosine distance of *all* classes and depends on both the input and final class membership, assigned by the max operator. Intuitively, class assignment should be less obvious under concept shift and the difference should be small. Consequently, $h(y,x)$ assigns a higher value to in-distribution data than to OOD data.In retrospect, our definition of covariate shift and concept shift scores supports existing findings that the features norms correspond to intra-class variance and angles reflect inter-class variation [23]. Intuitively, covariate shift represents non-semantic change within a specific class, i.e., intra-class variance; concept shift represents semantic changes, i.e, inter-class variation. Here we formalize the intuition and observations in [23] as score functions derived analytically from a KL-divergence viewpoint.

More importantly, the **combined score function** $\mathcal{U}$ (Eq. 4) carries a physical meaning: it bounds the KL-divergence between a uniform distribution and the predictive distribution. Intuitively, a small $\mathcal{U}$ indicates large uncertainty because $\mathcal{U}$ upper bounds $\mathbb{KL}(\mathbf{U}||\mathbf{P})$, and a large $\mathcal{U}$ indicates small uncertainty because it also appears in the lower bound. A similar scoring function, $S(x) = \|\mathbf{f}\|_2 \max_j \|\mathbf{w}_j\|_2 \cos\phi_j$, is used in [20], but is only empirically motivated based on observations with limited analytical insights such as its relationship to uncertainty and the functionality of each of components. In contrast, our derivation clearly shows the relationship between these score functions and the KL-divergence, which is as an uncertainty measure, and disentangles their roles.

3.2 Geometric Out-of-Distribution Detection with In-Distribution Data

As derived in Eq. 4, the covariate score is a function of feature norms and the concept score is a function of feature angles. Consequently, improving the sensitivity of feature norms and feature angles to data shifts seems to be the natural next step to improve OOD detection. Therefore, we adopt Geometric Sensitivity Decomposition (GSD) [24] to improve sensitivity to covariate and concept shifts. Specifically, GSD improves sensitivity by extracting sensitive components from norms $\|\mathbf{f}^*\|_2$[3] and angles $|\phi_i^*|$ through a decomposition of them into: an *instance-independent* scalar and an *instance-dependent variance* factor as shown in Eq. 5. Instance-independent scalars $\mathcal{C}_f$ and $\mathcal{C}_\phi$ minimize the loss on the training set and instance-dependent components $\mathbf{f}$ and ϕ_i account *sensitively* for variances in samples.

$$\|\mathbf{f}^*\|_2 = \|\mathbf{f}\|_2 + \mathcal{C}_f, \quad |\phi_i^*| = |\phi_i| - |\mathcal{C}_\phi| \tag{5}$$

[2] Note that the negative LSE function is also defined as *free energy* in [15].

[3] The superscript $*$ denotes the *original* component before decomposition.

With the decomposed components, the *original* logit $l_i^* = \|\mathbf{f}^*\|_2\|\mathbf{w_i}\|_2 \cos\phi_i^*$, can be written as:

$$l_i^* = \|\mathbf{f}^*\|_2\|\mathbf{w_i}\|_2 \cos\phi_i^* \approx l_i = \left(\underbrace{\frac{1}{\cos\mathcal{C}_\phi}}_{\alpha} \|\mathbf{f}\|_2 + \underbrace{\frac{1}{\cos\mathcal{C}_\phi}}_{\alpha} \overbrace{\mathcal{C}_f}^{\beta} \right) \|\mathbf{w_i}\|_2 \cos\phi_i \tag{6}$$

where l_i denote the *new* ith logit. In Eq. 6, the *new*[4] feature $\mathbf{f}$ is a direct output of a feature extractor, and is modified by α and β. Note that the calculation of score functions in Sec. 3.1 only uses the feature and is independent of α and β.

Because $\cos\mathcal{C}_\phi$ and $\mathcal{C}_f$ are instance-independent, we can parametrize them separately from the main network. Unlike GSD which parametrizes them as instance-independent scalars, inspired by [17, 18], we make $\alpha(\mathbf{f})$ and $\beta(\mathbf{f})$ instance-dependent scalars and use a single linear layer to learn them. To enforce numerical constraints, i.e., $0 < \alpha < 1$ and $\beta > 0$, $\alpha(\mathbf{f})$ uses a *sigmoid activation* and $\beta(\mathbf{f})$ uses a *softplus activation*. Finally, the relaxed output is:

$$P(Y = i|x) = \frac{\exp l_i}{\sum_{j=1}^{M} \exp l_j} = \frac{\exp\left(\left(\frac{1}{\alpha(\mathbf{f})}\|\mathbf{f}\|_2 + \frac{\beta(\mathbf{f})}{\alpha(\mathbf{f})}\right)\|\mathbf{w_i}\|_2\cos\phi_i\right)}{\sum_{j=1}^{M}\exp\left(\left(\frac{1}{\alpha(\mathbf{f})}\|\mathbf{f}\|_2 + \frac{\beta(\mathbf{f})}{\alpha(\mathbf{f})}\right)\|\mathbf{w_j}\|_2\cos\phi_j\right)} \tag{7}$$

Now the *new* predicted norm $\|\mathbf{f}\|_2$ and angle ϕ_i are more sensitive to input changes because they encode variances in samples as shown in Eq. 5. Therefore, including β (related to norms) improves sensitivity to covariate shift and including α (related to angles) improves sensitivity to concept shift. Note that, under this construction, Generalized ODIN [17] is a special case of our proposed method. Generalized ODIN only includes the $\alpha(\mathbf{f})$ which only improves angle sensitivity but not norm sensitivity. Unlike [17]'s probabilistic perspective[5], our model builds on a geometric perspective and captures both covariate and concept shifts by improving norm and angle sensitivity. The new model can be trained identically as the vanilla network without additional hyperparameter tuning.

3.3 CIFAR100 Splits

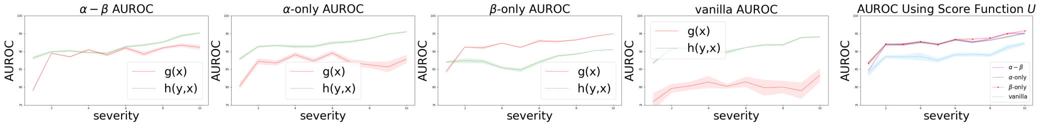
Table 1: **CIFAR100 Concept Shift Splits** Small group numbers indicate less conceptual similarity to CIFAR10 classes. The similarity is calculated using inner product between the Glove embeddings of a CIFAR100 class and a CIFAR10 class. For each CIFAR100 class, the largest similarity to each CIFAR10 class is taken as the overall similarity to CIFAR10. The average shows average similarity to CIFAR10 and the standard deviation shows in-group variance.

	1	2	3	4	5	6	7	8	9	10	AVE.	STD.
CIFAR10	airplane	automobile	bird	cat	deer	dog	frog	horse	ship	truck		
Group 9	cattle	shrew	motorcycle	squirrel	snake	trout	sea	tractor	bus	pickup	24.96	2.39
Group 8	bear	elephant	leopard	camel	lizard	rabbit	beaver	spider	raccoon	orchid	21.99	0.44
Group 7	lion	mountain	crab	bicycle	turtle	beetle	train	mouse	snail	otter	20.18	1.14
Group 6	possum	shark	forest	pine	dinosaur	boy	porcupine	wolf	road	butterfly	17.79	0.32
Group 5	girl	rocket	man	tiger	bee	tank	whale	baby	kangaroo	dolphin	16.26	0.44
Group 4	willow	worm	chimpanzee	skunk	cup	mushroom	oak	cockroach	crocodile	hamster	14.64	0.55
Group 3	castle	can	bridge	lobster	house	bed	fox	maple	pear	woman	12.65	0.63
Group 2	palm	streetcar	pepper	keyboard	bottle	seal	rose	couch	caterpillar	goldfish	10.18	0.51
Group 1	flatfish	apple	orange	plate	table	tulip	bowl	television	skyscraper	ray	8.95	0.25
Group 0	wardrobe	lamp	plain	lawnmower	chair	poppy	clock	cloud	sunflower	telephone	7.5	0.97

Special CIFAR100 Splits for Gradual Concept Shift While it is natural to associate covariate shift with increasing degrees of image corruption, finding a dataset to benchmark gradual concept shift is not straightforward because concept shift is traditionally thought as binary: overlapping or non-overlapping. However, not all non-overlapping labels are the same. For example, *pickup truck* (CIFAR100) is much closer to *truck* (CIFAR10) than *sunflowers* (CIFAR100) is semantically. To create this gradual concept/semantic shift, we propose to divide the CIFAR100 dataset into 10 sub-datasets with increasing conceptual difference from CIFAR10 classes. Specially, we use Glove

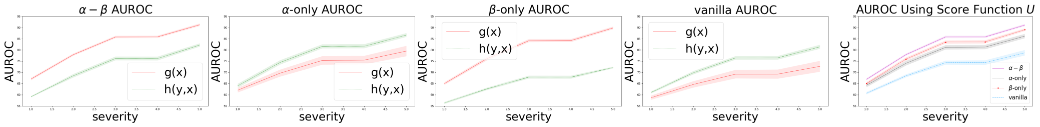
[4] Even though both $\mathbf{f}^*$ and $\mathbf{f}$ are outputs directly from the feature extractor, we use *original* and *new* to indicate whether GSD is applied.

[5] $\alpha(\mathbf{f})$ is interpreted as $P(d_{in}|x)$, the probability of x being in-distribution.



(a) α-β**,** α**-only,** β**-only and vanilla models using score functions** $g(x)$ **and** $h(y,x)$ (b) **All using** $\mathcal{U}$

Figure 1: **Capturing Concept Shift (CIFAR100 Splits)** All results are averaged over 5 runs. Modeling concept shift (both α-only and α-β models) yields the best performance as shown in Fig. 1b.



(a) α-β**,** α**-only,** β**-only and vanilla models using score functions** $g(x)$ **and** $h(y,x)$ (b) **All using** $\mathcal{U}$

Figure 2: **Capturing Covariate Shift (Motion Blur)** All results are averaged over 5 runs. Modeling covariate shift (α-β model) yields the best performance as shown in Fig. 2b.

word embeddings [25] trained on the entire wikipedia2014 and Gigaword5 [26] to measure semantic closeness (inner product) between CIFAR100 and CIFAR10 classes. The result is 10 subdatasets split from CIFAR100. We list the splits in Tab. 1.

4 Experiments

The $g(x)$ **score captures covariate shift.** We compare the α-β, α-only, β-only variants and the vanilla model on CIFAR10C [9] corrupted by motion blur in Fig. 2 with increasing degrees of noise. From 2a, we observe that 1) as covariate shift severity increases, OOD detection becomes easier because AUROC increases with increasing severity. 2) in practice, covariate shift does not happen in isolation from concept shift and vice versa because AUROC using either $g(x)$ or $h(y,x)$ increases. 3) the vanilla model is more sensitive to the concept shift component because $h(y,x) > g(x)$ in the vanilla model plot even though covariate shift is the dominant distribution shift in this example. 4) when sensitivity to both covariate and concept shift is improved , the α-β model becomes more sensitive to the covariate shift component. This suggests that the dominant shift in this example is indeed covariate shift. From Fig. 2b, we observe that the α-β model outperforms the β-only model using the combined score function $\mathcal{U}$. This suggests that improving sensitivity to both shifts and using $\mathcal{U}$ yield the best OOD detection. **The** $h(y,x)$ **score captures concept shift.** Following the previous section, we benchmark the α-β, α-only, β-only variants and the vanilla model on the newly created CIFAR100 Splits. From Fig. 1a, we observe that 1) as concept shift severity increases, OOD detection becomes easier because AUROC increases with increasing severity. 2) both concept shift and covariate shit are present because AUROC using either $h(y,x)$ or $g(x)$ increases. 3) the vanilla model is dominantly more sensitive to the concept shift component because concept shift is the dominant distribution shift in CIFAR100 Splits by construction and vanilla ResNet is more sensitive to concept shift. The same behavior is also observed on CIFAR10C. 4) when sensitivity to both shifts is improved, the α-β model is still more sensitive to concept shift ($h(y,x) > g(x)$). This reconfirms that the dominant shift type is indeed concept shift. From Fig. 1b, we observe that all three variants perform similarly and all outperform the vanilla model. Combined with previous observations that the dataset has strong concept shift and the vanilla model is already very sensitive to concept shift, improving sensitivity to the covariate shift component yields equally good performance as improving sensitivity to both shifts. We provide results on standard benchmarks in Appendix 6.1.

5 Conclusion

In this work, we propose to characterize the spectrum of out-of-distribution data from the perspective of dataset shift, specifically covariate and concept shift. This categorization provides a different perspective to study OOD detection for *deterministic* and *discriminative* models. At representation level, we derive two score functions that represent and capture each shift separately. At modeling level, inspired by these score functions, we propose a geometrically-inspired method, Geometric ODIN, to improve a model's sensitivity to both shift.

References

[1] Didier Dubois, Henri Prade, and Philippe Smets. Representing partial ignorance. *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans*, 26(3):361–377, 1996.

[2] Stefan Depeweg, Jose-Miguel Hernandez-Lobato, Finale Doshi-Velez, and Steffen Udluft. Decomposition of uncertainty in bayesian deep learning for efficient and risk-sensitive learning. In *International Conference on Machine Learning*, pages 1184–1193. PMLR, 2018.

[3] Alex Kendall and Yarin Gal. What uncertainties do we need in bayesian deep learning for computer vision? *arXiv preprint arXiv:1703.04977*, 2017.

[4] Eyke Hüllermeier and Willem Waegeman. Aleatoric and epistemic uncertainty in machine learning: An introduction to concepts and methods. *Machine Learning*, 110(3):457–506, 2021.

[5] Jishnu Mukhoti, Andreas Kirsch, Joost van Amersfoort, Philip HS Torr, and Yarin Gal. Deterministic neural networks with appropriate inductive biases capture epistemic and aleatoric uncertainty. *arXiv preprint arXiv:2102.11582*, 2021.

[6] 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.

[7] 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. *arXiv preprint arXiv:2006.10108*, 2020.

[8] Jose G Moreno-Torres, Troy Raeder, Rocío Alaiz-Rodríguez, Nitesh V Chawla, and Francisco Herrera. A unifying view on dataset shift in classification. *Pattern recognition*, 45(1):521–530, 2012.

[9] Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. *Proceedings of the International Conference on Learning Representations*, 2019.

[10] Dan Hendrycks, Mantas Mazeika, and Thomas Dietterich. Deep anomaly detection with outlier exposure. *arXiv preprint arXiv:1812.04606*, 2018.

[11] Sunil Thulasidasan, Sushil Thapa, Sayera Dhaubhadel, Gopinath Chennupati, Tanmoy Bhattacharya, and Jeff Bilmes. A simple and effective baseline for out-of-distribution detection using abstention. 2020.

[12] Abhijit Guha Roy, Jie Ren, Shekoofeh Azizi, Aaron Loh, Vivek Natarajan, Basil Mustafa, Nick Pawlowski, Jan Freyberg, Yuan Liu, Zach Beaver, et al. Does your dermatology classifier know what it doesn't know? detecting the long-tail of unseen conditions. *arXiv preprint arXiv:2104.03829*, 2021.

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

[14] Kimin Lee, Kibok Lee, Honglak Lee, and Jinwoo Shin. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. *Advances in neural information processing systems*, 31, 2018.

[15] Weitang Liu, Xiaoyun Wang, John D Owens, and Yixuan Li. Energy-based out-of-distribution detection. *arXiv preprint arXiv:2010.03759*, 2020.

[16] Shiyu Liang, Yixuan Li, and Rayadurgam Srikant. Enhancing the reliability of out-of-distribution image detection in neural networks. *arXiv preprint arXiv:1706.02690*, 2017.

[17] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10951–10960, 2020.

[18] Engkarat Techapanurak, Masanori Suganuma, and Takayuki Okatani. Hyperparameter-free out-of-distribution detection using softmax of scaled cosine similarity. *arXiv preprint arXiv:1905.10628*, 2019.

[19] 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.

[20] Jihoon Tack, Sangwoo Mo, Jongheon Jeong, and Jinwoo Shin. Csi: Novelty detection via contrastive learning on distributionally shifted instances. *arXiv preprint arXiv:2007.08176*, 2020.

[21] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. *arXiv preprint arXiv:1612.01474*, 2016.

[22] Stanislav Fort, Jie Ren, and Balaji Lakshminarayanan. Exploring the limits of out-of-distribution detection. *arXiv preprint arXiv:2106.03004*, 2021.

[23] Weiyang Liu, Zhen Liu, Zhiding Yu, Bo Dai, Rongmei Lin, Yisen Wang, James M Rehg, and Le Song. Decoupled networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2771–2779, 2018.

[24] Junjiao Tian, Dylan Yung, Yen-Chang Hsu, and Zsolt Kira. A geometric perspective towards neural calibration via sensitivity decomposition. In *NeurIPS*, 2021.

[25] Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014.

[26] Courtney Napoles, Matthew R Gormley, and Benjamin Van Durme. Annotated gigaword. In *Proceedings of the Joint Workshop on Automatic Knowledge Base Construction and Web-scale Knowledge Extraction (AKBC-WEKEX)*, pages 95–100, 2012.

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

Table 2: **AUROC ↑ for Near and Far OOD detection**. Results are averaged over 5 runs. Our $\alpha(x)$-only model is the same as Generalized ODIN [17] without its input processing.* denotes results from [5].

AUROC↑	Score Functions	ID:CIFAR100			ID:CIFAR10		
		Near CIFAR10	Near CIFAR100C	Far SVHN	Near CIFAR100	Near CIFAR10C	Far SVHN
Vanilla Wide-ResNet-28-10 [27]	MSP [13]	80.68±0.34	70.38±1.37	77.37±2.25	88.93±0.37	70.58±0.59	93.66±1.79
	Energy [15]	**80.74±0.45**	70.79±0.85	79.48±2.91	88.84±0.44	70.60±0.52	94.39±2.30
	Mahanobis [14]	78.27±1.51	69.16±0.48	84.37±4.56	87.63±1.19	68.38±0.50	93.08±6.02
	GMM Density [5]	71.67±0.57	73.15±0.49	85.67±2.02	90.98±0.24	75.71±0.80	97.68±0.38
DUQ [6]	Kernel Distance	–	–	–	85.92±0.35*	–	93.71±0.61*
SNGP [7]	SoftMax Entropy	–	–	85.71±0.81*	91.13±0.15*	–	94.0±1.30*
DDU [5]	GMM Density	73.05±0.51	73.31±0.45	87.32±0.77	90.69±0.42	76.00±0.00	97.12±1.21
5-Ensemble [21]	SoftMax Entropy	–	–	79.54±0.91*	92.13±0.02*	–	97.73±0.31*
Ours: $\alpha(x)$-only	$h(y, x)$	79.24±0.37	**73.60±0.55**	83.75±3.30	92.28±0.15	77.00±0.00	97.56±0.90
Ours: α-β	$h(y, x)$	79.42±0.39	72.95±0.12	88.69±2.61	91.29±0.07	73.80±0.45	98.42±0.21
	$g(x)$	61.06±0.31	65.50±2.18	89.76±3.34	89.08±0.24	76.80±1.10	99.40±0.14
	$\mathcal{U}$	71.48±0.26	71.72±1.09	**93.80±2.42**	**92.31±0.21**	**78.00±0.71**	**99.54±0.08**

6 Appendix

6.1 Out-of-Distribution Detection Results

The α-β model is the best. In Tab. 2, we present OOD detection results against state-of-the-art methods on existing near and far OOD categorization. For near OOD detection under strong concept shift, CIFAR10 (ID) vs. CIFAR100 (OOD), both our α-only and α-β variants achieve the the best performance. This demonstrates that $\alpha(x)$ improves the sensitivity of angles and hence the sensitivity to concept-shifted data. For near OOD detection under strong covariate shift, CIFAR10 (ID) vs. CIFAR10C (OOD), the α-β variant achieves the the best performance. This suggests that $\beta(x)$ improves the sensitivity of norms and hence the sensitivity to covariate-shifted data. In CIFAR100 (ID) vs. CIFAR10/CIFAR100C (OOD) experiments, the performance of the α-only and α-β variants are within variance and is close to some other compared methods, we can not make clear observations from those experiments[6]. For far OOD detection, CIFAR10/CIFAR100 (ID) vs. SVHN (OOD), the α-β model achieves state-of-the-art performance. This reconfirms that $\beta(x)$ improves sensitivity to covariate-shifted data, because SVHN has both covariate and concept shifts compared to the CIFAR datasets, and the α-β model outperforms the α-only variant, which only improves on concept shift, by a noticeable margin. In terms of score functions, the best performing one for the α-β model is $\mathcal{U}$, which is a product of $g(x)$ and $h(y, x)$ (Sec. 3.1), while that of the α-only model is $h(y, x)$. This shows that depending on which component is more sensitive, different scoring functions are preferred. When the sensitivity of both norms and angles are improved, as in the α-β variant, the combined score function $\mathcal{U}$ performs well under different distribution shifts.

[6] Other confounding factors could contribute to the close performance. Prior works either omit comparisons under these settings [5] or report only marginal improvement [19]