Learning Semantic Epistemic Uncertainty via Contrastive Representations for Robust OOD Detection

Wang Ma, Qiang Ji,
Rensselaer Polytechnic Institute
110 8th St
Troy, NY, 12180, USA
maw6@rpi.edu, jiq@rpi.edu

Abstract

Reliable uncertainty quantification and disentanglement are crucial for deploying deep models in high-stakes scenarios. However, conventional models often produce highly correlated estimates of aleatoric uncertainty (AU) and epistemic uncertainty (EU), limiting their interpretability and effectiveness. In particular, high AU—arising from data noise or nuisance factors—can erroneously induce high EU, causing in-distribution examples with high randomness to be misidentified as OOD, which is undesirable. This conflation of semantic novelty with superficial variations compromises the utility of EU as a meaningful indicator of model knowledge. To address this, we propose a novel training framework that leverages contrastive learning to explicitly disentangle semantic epistemic uncertainty from nuisance-induced uncertainty. By treating clean images and their corrupted counterparts as positive pairs, our method encourages the model to learn content-invariant representations while discarding nuisance-specific features. We provide a formal information-theoretic analysis showing that contrastive learning aligns epistemic uncertainty with semantic content by minimizing the mutual information between nuisance variables and learned representations. Experiments on standard benchmarks demonstrate that our method yields epistemic uncertainty estimates that are robust to superficial corruptions yet highly sensitive to true semantic shifts, resulting in state-of-the-art OOD detection performance. These results underscore the importance of disentangled representations for trustworthy uncertainty quantification in safety-critical applications.

1 Introduction

Deep neural networks have achieved remarkable success across various domains, yet their deployment in high-stakes environments like autonomous driving and medical diagnosis requires not only accurate predictions but also reliable estimates of their confidence. Uncertainty quantification (UQ) addresses this need by allowing models to express their own uncertainty. A principled approach, rooted in Bayesian modeling, decomposes the total predictive uncertainty into two fundamental types: **aleatoric uncertainty (AU)**, which captures inherent noise or randomness in the data itself, and **epistemic uncertainty (EU)**, which represents the model's lack of knowledge (Kendall and Gal 2017; Der Kiureghian and Ditlevsen 2009).

Ideally, these two sources of uncertainty should be disentangled, providing distinct and interpretable signals. For example, a model viewing a slightly-corrupted but otherwise familiar image should report high aleatoric uncertainty (due to randomness in the data itself) but relatively low epistemic uncertainty (as the content is familiar). Conversely, an image from a completely novel class should elicit high epistemic uncertainty. However, recent work has shown that in standard training regimes, these two uncertainties are often highly correlated, limiting their practical utility (Mucsányi, Kirchhof, and Oh 2024).

We hypothesize that this entanglement stems from the fact that standard models conflate semantic content with nuisance variations (e.g., blur, noise, lighting and other random corruptions). Epistemic uncertainty, which should ideally track a lack of knowledge about the core *content*, becomes inflated by superficial variations the model was not explicitly trained on.

In this paper, we propose to address this issue by learning a **semantic epistemic uncertainty**. Our core idea is that if a model's representations are invariant to nuisance variations while remaining sensitive to changes in semantic content, its epistemic uncertainty will naturally align with content-level shifts. To achieve this, we employ contrastive learning (Chen et al. 2020; He et al. 2020), which is renowned for learning such invariant representations. We treat clean data as an "anchor" and its corrupted versions as "positive" pairs, forcing the model to learn representations that focus on the shared, underlying content.

Our contributions are threefold:

1. We provide a theoretical framework that formally links contrastive learning to the disentanglement of content and nuisance information in a model's learned representations.
2. We prove that by minimizing nuisance information, our approach aligns the model's observation-level epistemic uncertainty with a more desirable content-level epistemic uncertainty.
3. We empirically validate our approach on standard datasets and tasks, showing that our contrastively trained model produces more meaningful epistemic uncertainty estimates. It correctly identifies corrupted in-distribution data

as having high aleatoric but low epistemic uncertainty, and it significantly outperforms a standard baseline in detecting out-of-distribution data based on its superior epistemic uncertainty estimates.

2 Related Work

Our work is situated at the intersection of three key areas: uncertainty quantification, representation learning, and out-of-distribution detection, (Yang et al. 2024).

Uncertainty Quantification. The estimation and decomposition of predictive uncertainty into aleatoric and epistemic components is a cornerstone of Bayesian deep learning (Depeweg et al. 2018). Practical methods for estimating these quantities often rely on approximations of the Bayesian posterior, such as Monte Carlo Dropout (Gal and Ghahramani 2016) or Deep Ensembles (Lakshminarayanan, Pritzel, and Blundell 2017), which we use in this work. While powerful, these methods do not inherently guarantee that the estimated EU will be semantically meaningful or disentangled from AU. Several works have noted the high correlation and proposed post-hoc recalibration methods (Nayman et al. 2024), but few have addressed it at the training and representation level.

Disentangled Representation Learning. The goal of disentanglement is to learn representations where distinct latent units correspond to distinct, interpretable factors of variation in the data (Bengio, Courville, and Vincent 2013; Locatello et al. 2019). Early work focused on generative models like VAEs, but later research revealed fundamental challenges in achieving unsupervised disentanglement without inductive biases (Locatello et al. 2019). Our work approaches disentanglement from a different perspective: instead of isolating all factors, we aim to specifically separate semantic content from all other nuisance variations for the purpose of robust UQ.

Contrastive Learning. Contrastive learning has emerged as a dominant paradigm in self-supervised learning (Oord, Li, and Vinyals 2018). Methods like SimCLR (Chen et al. 2020) and MoCo (He et al. 2020) learn representations by maximizing agreement between differently augmented "views" of the same image. The inductive bias imposed by data augmentation forces the learned representations to be invariant to these transformations. Recent theoretical work has shown that this process provably isolates content from nuisance factors defined by the augmentations (von Kügelgen et al. 2021). We build directly on this principle, proposing to use data corruptions as a form of augmentation to explicitly learn representations invariant to them, thereby purifying the resulting epistemic uncertainty signal.

3 Preliminaries and Notations

Let $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^N$ denote the training dataset, where $x_i \in \mathcal{X}$ represents an observed input (e.g., an image) and $y_i \in \mathcal{Y}$ is its corresponding label. We postulate a latent variable model where each input x is generated from underlying factors: a *content* variable $c \in \mathcal{C}$ representing the core semantic information relevant to the task, and a *nuisance* variable $n \in \mathcal{N}$ capturing variations irrelevant to the task (e.g., style, lighting, background). Let θ denote the parameters of our

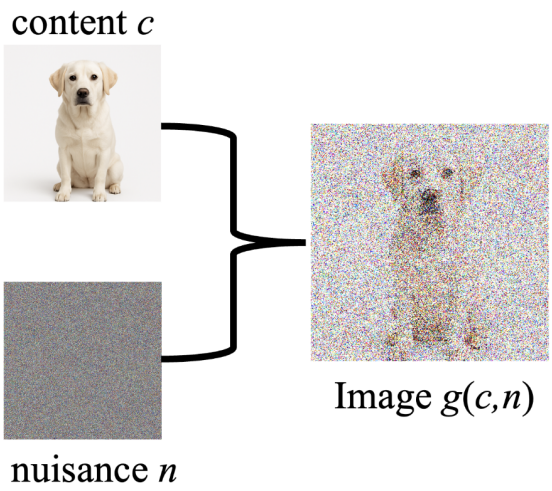

Figure 1: Deterministic data generation illustration.

predictive model, typically a neural network. The model aims to approximate the true data distribution and provide a predictive distribution $\hat{p}(y|x, \theta)$. When considering uncertainty, particularly in a Bayesian context or using ensembles, we are often interested in the posterior distribution of parameters $p(\theta|\mathcal{D})$. The overall predictive uncertainty for a new input x can be decomposed using the law of total variance for entropy (Depeweg et al. 2018; Hüllermeier and Waegeman 2021):

$$\underbrace{\mathbb{H}\left[\mathbb{E}_{\theta\sim p(\theta|\mathcal{D})}[\hat{p}(y|x,\theta)]\right]}_{\text{Total Predictive Uncertainty}} = \underbrace{\mathbb{I}[y;\theta \mid x,\mathcal{D}]}_{\text{Epistemic Uncertainty (EU)}} + \underbrace{\mathbb{E}_{\theta\sim p(\theta|\mathcal{D})}\left[\mathbb{H}[\hat{p}(y|x,\theta)]\right]}_{\text{Aleatoric Uncertainty (AU)}}. \quad (1)$$

Here, $\mathbb{H}[\cdot]$ denotes the Shannon entropy, $\mathbb{I}[\cdot;\cdot|\cdot]$ denotes the conditional mutual information, and the expectation $\mathbb{E}_{\theta\sim p(\theta|\mathcal{D})}$ is taken over the parameter posterior distribution. Epistemic uncertainty (EU) captures the model's uncertainty due to limited training data (lack of knowledge), while aleatoric uncertainty (AU) captures inherent randomness in the data generating process itself.

3.1 Factorised Data-Generating Process

We assume the following factorized structure for the joint distribution of labels, latent variables, and observations:

$$p(y, c, n, x) = p(y|c, n, x)p(x|c, n)p(c|n)p(n). \quad (2)$$

We make two simplifying assumptions common in disentanglement and robust representation learning literature (Bengio, Courville, and Vincent 2013; Locatello et al. 2019):

1. **Latent Independence:** The content c and nuisance n variables are statistically independent, i.e., $p(c|n) = p(c)$.
2. **Deterministic Generation:** The observation x is generated by a deterministic function $g : \mathcal{C} \times \mathcal{N} \to \mathcal{X}$, such that $p(x|c, n) = \delta(x - g(c, n))$, where $\delta(\cdot)$ is the Dirac delta function.

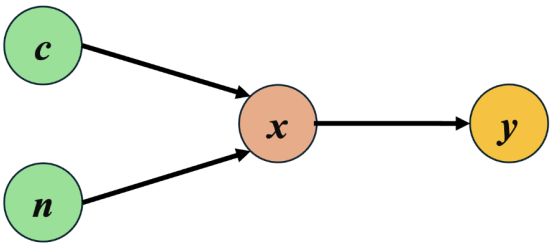

Figure 2: Factorised Data-Generating Process.

Under these assumptions, and noting that x is fully determined by c and n, the dependence of y on x becomes redundant given c and n, i.e., $p(y|c, n, x) = p(y|c, n)$. The factorization simplifies to:

$$p(y, c, n, x) = p(y|c, n)\delta(x - g(c, n))p(c)p(n). \quad (3)$$

Often, it is further assumed that the label y depends only on the content c, i.e., $p(y|c, n) = p(y|c)$, although we do not strictly require this for the subsequent analysis unless specified.

3.2 Model Inference Process

We consider neural network models for classification, parameterized by θ. These models typically consist of an encoder $f_\phi : \mathcal{X} \to \mathcal{Z}$ with parameters ϕ, which maps the input x to a latent representation $z \in \mathcal{Z}$, followed by a classifier head $h_\psi : \mathcal{Z} \to \Delta^{|\mathcal{Y}|-1}$ with parameters ψ, which maps the representation z to a probability distribution over labels. Here, Δ^K denotes the K-dimensional probability simplex, and the full model parameters are $\theta = (\phi, \psi)$. The model's predictive distribution is thus given by:

$$\hat{p}(y|x, \theta) = h_\psi(z; \psi) \quad \text{where} \quad z = f_\phi(x; \phi). \quad (4)$$

Our primary interest lies in learning an encoder f_ϕ that extracts representations z which are informative about the underlying content c while being invariant to nuisance variations n. We hypothesize that such representations facilitate more meaningful uncertainty quantification. In this work, we leverage contrastive learning (Oord, Li, and Vinyals 2018; Chen et al. 2020; He et al. 2020) to train the encoder f_ϕ and investigate its impact on epistemic uncertainty estimation, particularly for challenging tasks like Out-of-Distribution (OOD) detection.

4 Contrastive Training Learns Content-Aligned Epistemic Uncertainty

We now establish theoretically how contrastive learning, specifically through the optimization of the InfoNCE loss (Oord, Li, and Vinyals 2018), encourages the learned representation $z = f_\phi(x)$ to capture content c while discarding nuisance n, and how this property leads to more robust epistemic uncertainty estimates.

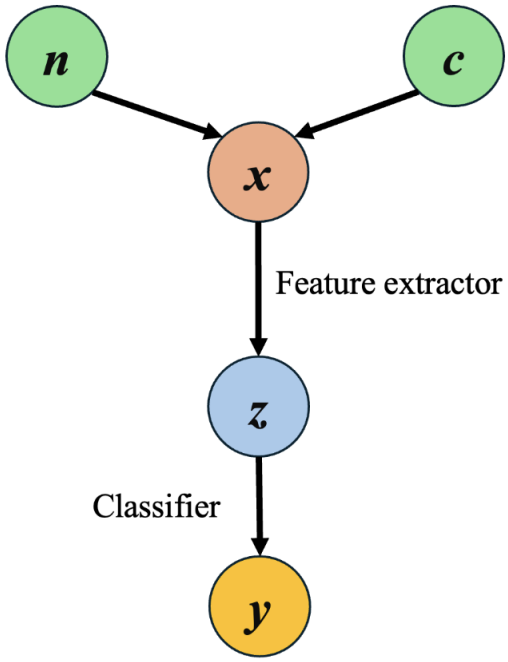

Figure 3: Model Inference Process.

4.1 Contrastive Learning Promotes Content Informativeness and Nuisance Invariance

Contrastive learning aims to learn representations by maximizing agreement between representations of different "views" (augmentations) of the same data point (positive pairs), while minimizing agreement with representations of different data points (negative pairs). We operate under the assumption that the positive pairs, denoted (x_1, x_2), are generated such that they share the same underlying content c but possess independent nuisance variables $n_1, n_2 \sim p(n)$. Formally, $x_1 = g(c, n_1)$ and $x_2 = g(c, n_2)$ for some $c \sim p(c)$ and $n_1, n_2 \overset{\text{i.i.d.}}{\sim} p(n)$. The InfoNCE objective, in its idealized population form, encourages the encoder f_ϕ to learn representations $z = f_\phi(x)$ such that the mutual information between representations of positive pairs is maximized.

This learning process implicitly optimizes the information content of the representation z with respect to the latent factors c and n. The following lemmas formalize this relationship, building upon insights from (Tsai et al. 2020; von Kügelgen et al. 2021).

Lemma 1 (Content Informativeness via Contrastive Learning)**.** *Let $x = g(c, n)$ with $c \sim p(c)$ and $n \sim p(n)$ being independent. Let $z = f_\phi(x)$ be the representation learned by minimizing the InfoNCE loss using positive pairs (x_1, x_2) where $x_1 = g(c, n_1)$ and $x_2 = g(c, n_2)$ share the same content c and have independent nuisances n_1, n_2. Optimizing the InfoNCE objective generally encourages an increase in the mutual information between the representation and the content variable:*

$$\mathbb{I}(c; z) \uparrow .$$

Lemma 2 (Nuisance Invariance via Contrastive Learning)**.** *Under the same assumptions as Lemma 3, optimizing the InfoNCE objective encourages a reduction in the mutual information between the nuisance variable and the representation, conditioned on the content:*

$$\mathbb{I}(n; z \mid c) \searrow 0.$$

Intuition: Lemma 3 arises because the shared information between positive pairs (x_1, x_2) is primarily the content c. By pulling their representations (z_1, z_2) together, the InfoNCE loss implicitly maximizes the information z retains about c. Lemma 4 follows because, for a fixed content c, the representations $z_1 = f_\phi(g(c, n_1))$ and $z_2 = f_\phi(g(c, n_2))$ are encouraged to be similar despite variations in n_1 and n_2. This forces the representation z to become invariant to nuisance variations n, conditional on knowing the content c. Rigorous proofs adapting arguments from prior work are provided in Appendix.

4.2 Contrastive Representations Enable Content-Aligned Epistemic Uncertainty

A key desideratum for reliable uncertainty quantification, especially in safety-critical applications or OOD detection, is that epistemic uncertainty (EU) reflects genuine model ignorance about the task-relevant content c, rather than being inflated by superficial nuisance variations n. We formalize this by defining two related quantities:

Definition 1 (Observation-level and Content-level Epistemic Uncertainty)**.** *Given an input $x = g(c, n)$, the training data $\mathcal{D}$, and the parameter posterior $p(\theta|\mathcal{D})$:*

- *The standard **observation-level epistemic uncertainty** is:*

$$EU_x := \mathbb{I}(y; \theta \mid x, \mathcal{D}) = \mathbb{I}(y; \theta \mid c, n, \mathcal{D}).$$

- *The desired **content-level epistemic uncertainty** is:*

$$EU_c := \mathbb{I}(y; \theta \mid c, \mathcal{D}).$$

Ideally, we want $EU_x \approx EU_c$, meaning the uncertainty estimate for an observation x is determined by its content c, irrespective of the nuisance n. We now demonstrate that representations learned via contrastive methods promote this alignment.

Proposition 1 (Epistemic Uncertainty Alignment via Contrastive Representations)**.** *Let $EU_x = \mathbb{I}(y; \theta \mid c, n, \mathcal{D})$ and $EU_c = \mathbb{I}(y; \theta \mid c, \mathcal{D})$ be the epistemic uncertainties defined above. Assuming the learned representation $z = f_\phi(x)$ captures the relevant information pathways, the absolute difference between these two uncertainty measures can be bounded in terms of the nuisance information retained by the representation:*

$$|EU_x - EU_c| \leq \mathbb{I}(n; z \mid c) \to 0. \tag{5}$$

Proof Sketch: The relationship $|EU_x - EU_c| \leq \mathbb{I}(n; \theta \mid c, \mathcal{D})$ follows from properties of conditional mutual information (see Appendix for details). The crucial step is linking $\mathbb{I}(n; \theta \mid c, \mathcal{D})$ to $\mathbb{I}(n; z \mid c)$, which can be obtained by a Markov chain and data generation process. And since we showed $\mathbb{I}(n; z \mid c)$ approaches 0, then we get that the estimated EU_x is approaching the content-level epistemic uncertainty EU_c.

Implication. Proposition 2 demonstrates that the discrepancy between the standard epistemic uncertainty EU_x and the more desirable content-level epistemic uncertainty EU_c is bounded by the amount of nuisance information encoded in the representation z, conditioned on the content c. According to Lemma 4, optimizing the InfoNCE loss actively minimizes this conditional mutual information $\mathbb{I}(n; z \mid c)$. Therefore, contrastive pre-training naturally encourages the learned epistemic uncertainty (computed using the learned representation z) to align better with the content-level uncertainty EU_c. This provides a theoretical grounding for the empirical observation that contrastively trained models often yield more robust uncertainty estimates, particularly for distinguishing in-distribution nuisance variations from genuine out-of-distribution samples based on content mismatch.

Algorithm 1: Contrastive Training for Semantic Uncertainty (one ensemble member)

Requires : Clean dataset $\mathcal{D}_{\text{clean}} = \{(x, y)\}$;
Nuisance function $\text{Corrupt}(x)$;
Encoder network f_ϕ with parameters ϕ;
Classifier head h_ψ with parameters ψ;
InfoNCE temperature τ;

Output : A trained model consisting of a frozen encoder and a classifier $(f_{\phi_{\text{frozen}}}, h_\psi)$;

```
// Stage 1: Contrastive Training
```

Initialize encoder parameters ϕ **for** *each training epoch* **do**
  for *each batch of clean images* $\{x_i\} \sim \mathcal{D}_{clean}$ **do**
    $x_{\text{anchor}} \leftarrow \{x_i\}$
    $x_{\text{positive}} \leftarrow \{\text{Corrupt}(x_i)\}$
    $z_{\text{anchor}} \leftarrow f_\phi(x_{\text{anchor}})$
    $z_{\text{positive}} \leftarrow f_\phi(x_{\text{positive}})$
    $\mathcal{L}_{\text{contrastive}} \leftarrow \text{Calculate\_InfoNCE\_Loss}(z_{\text{anchor}}, z_{\text{positive}}, \tau)$
    Update ϕ using gradient descent on $\mathcal{L}_{\text{contrastive}}$

$\phi_{\text{frozen}} \leftarrow \phi$

```
// Stage 2: Training Linear Classifier
```

Initialize classifier parameters ψ **for** *each training epoch* **do**
  for *each batch of labeled images* $\{(x_i, y_i)\} \sim \mathcal{D}_{clean}$ **do**
    $z_{\text{frozen}} \leftarrow f_{\phi_{\text{frozen}}}(x_i)$
    $\hat{y}_i \leftarrow h_\psi(z_{\text{frozen}})$
    $\mathcal{L}_{\text{classifier}} \leftarrow \text{CrossEntropyLoss}(\hat{y}_i, y_i)$
    Update ψ using gradient descent on $\mathcal{L}_{\text{classifier}}$

return $(f_{\phi_{frozen}}, h_\psi)$

5 Experiments

We conduct a comprehensive set of experiments to empirically validate our hypothesis that contrastive learning can produce a more disentangled, semantic epistemic uncertainty. We compare our proposed method against a standard supervised baseline across three benchmark datasets: MNIST, CIFAR-10, and CIFAR-100.

5.1 Experimental Setup

Datasets and Data Splits. For each benchmark, we define four categories of test data to evaluate model performance and uncertainty under different distribution shifts:

- **Clean In-Distribution (ID)**: The standard, unmodified test set of the dataset (e.g., CIFAR-10 test set).
- **Nuisance Shift (Seen)**: ID images corrupted by a set of nuisance transformations (e.g., blur, noise from CIFAR-10-C ,(Hendrycks and Dietterich 2019)), with a fixed severity. These specific corruption types are used during the training phase.
- **Nuisance Shift (Unseen)**: ID images corrupted by a *different*, held-out set of nuisance transformations at a higher severity, representing a novel distributional shift in the nuisance variable.
- **Semantic Shift (OOD)**: A true out-of-distribution dataset with different semantic content. We use FashionMNIST, KMNIST, and EMNIST for MNIST; SVHN for CIFAR-10; and SVHN for CIFAR-100.

Models and Training.

Our experiments are based on a **ResNet-18**, (He et al. 2016), architecture. For each experiment, we train a deep ensemble of 5 models to obtain robust uncertainty estimates. We compare two training paradigms:

- **Standard (Baseline)**: A ResNet-18 model trained with a standard cross-entropy loss on a mixture of Clean ID and Nuisance Shift (Seen) data.
- **Contrastive (Ours)**: Following algorithm 1, we first train the ResNet-18 encoder using a contrastive loss, where clean images serve as anchors and their Nuisance Shift (Seen) counterparts act as positive samples. Subsequently, the encoder is frozen, and a linear classifier is trained on top using the labeled Clean ID data.

Evaluation. We evaluate models on: (1) Test accuracy; (2) Mean uncertainty scores (Aleatoric, Epistemic, and Total); and (3) OOD detection performance, measured by AUROC and AUPR (Hendrycks and Gimpel 2016; Davis and Goadrich 2006; Hendrycks and Dietterich 2019).

5.2 Test Accuracy under Nuisance and Semantic Shifts

We evaluate the test accuracy of our contrastive method against a standard supervised baseline across multiple benchmark datasets: MNIST, CIFAR-10, and CIFAR-100. Table 1 summarizes these results clearly.

In general, the test accuracy of our contrastive learning method exhibits mixed performance when compared to the standard supervised baseline. Specifically, on CIFAR-10, our contrastively trained model achieves a slightly higher accuracy (0.886) compared to the baseline (0.878) on Clean In-Distribution (ID) data, as well as on the Nuisance Shift (Seen) set (0.864 vs. 0.822). However, the accuracy on Nuisance Shift (Unseen) data is lower (0.635) compared to the baseline (0.698). On MNIST, the accuracy is comparable between methods on Clean ID and Nuisance Shift (Seen) sets, but noticeably lower for our model on Nuisance Shift (Unseen) data (0.388 vs. 0.572). Finally, on CIFAR-100, our model has a moderately lower Clean ID accuracy (0.713 vs. 0.787), but achieves competitive performance on both Nuisance Shift (Seen) and (Unseen) datasets.

The primary reason for the occasionally lower accuracy observed in our contrastive method, particularly notable on CIFAR-100 and Nuisance Shift (Unseen) sets, is the nature of the contrastive training stage. Our contrastive encoder training exclusively utilizes self-supervised objectives, treating clean and corrupted images as positive pairs without any direct label supervision. This self-supervised approach intentionally emphasizes content invariance over label-specific discriminative information, thereby potentially sacrificing classification performance.

This inherent trade-off highlights an avenue for future work: integrating label information during the contrastive training stage, (Khosla et al. 2020), such as employing supervised or semi-supervised contrastive learning techniques. Incorporating labels could enhance both semantic disentanglement and accuracy, leading to robust and precise uncertainty quantification suitable for high-stakes, real-world scenarios.

5.3 Uncertainty Under Nuisance and Semantic Shifts

We then analyze how the uncertainty estimates from each model behave across the different data splits. Table 1 summarizes the accuracy and mean uncertainty scores.

Across all three datasets, the **aleatoric uncertainty (AU)** behaves as expected for both models, increasing monotonically as the data quality degrades from Clean ID to Nuisance Shift (Seen) and Nuisance Shift (Unseen). This correctly reflects the increasing randomness and inherent difficulty in the input data.

The primary distinction emerges in the **epistemic uncertainty (EU)**. For the **Standard** baseline, the EU consistently rises when moving from Clean ID to Nuisance Shift (Seen) data (e.g., from 0.063 to 0.076 on CIFAR-10; 0.300 to 0.380 on CIFAR-100). This indicates the model perceives familiar corruptions as a source of model uncertainty (knowledge), confounding nuisance with novelty.

In stark contrast, our **Contrastive** model demonstrates remarkable robustness. On CIFAR-100, its EU remains almost constant between Clean ID (0.170) and Nuisance Shift (Seen) (0.186). On CIFAR-10, the EU (0.073 vs 0.086) is also much more stable than the baseline. This provides strong evidence that our training method successfully learns to be invariant to seen nuisances, attributing them correctly to data noise (high AU) rather than model ignorance (low EU). Crucially, for novel shifts, the EU of our model did not increase as sharply as baseline model for Nuisance (Unseen) and for Semantic (OOD) the EU of our model increases significantly, demonstrating its consistency to non-content shifts and sensitivity to content-level distribution shifts.

In one word, based on the EU results, our model can identify Nuisance (Unseen) and Semantic (OOD), while treating the first as ID data, and the second as OOD data, which is desired. But baseline model fails in identifying Nuisance (Unseen) and Semantic (OOD), meaning it will treat low-quality ID data as OOD, hurting the model's ability.

Table 1: Accuracy and Mean Uncertainty Scores across all datasets. Our contrastive method maintains low epistemic uncertainty (EU) on familiar nuisance shifts while correctly elevating it for unseen and semantic shifts, unlike the standard baseline which shows increased EU even for familiar corruptions.

Dataset	Model	Clean ID	Nuisance Shift (Seen)	Nuisance Shift (Unseen)	Semantic Shift (OOD)
			Test Accuracy (↑)		
CIFAR-10	Standard	0.878	0.822	**0.698**	—
	Contrastive (Ours)	**0.886**	**0.864**	0.635	—
MNIST	Standard	**0.997**	**0.996**	**0.572**	—
	Contrastive (Ours)	0.996	0.994	0.388	—
CIFAR-100	Standard	0.787	0.683	0.395	—
	Contrastive (Ours)	0.713	**0.687**	**0.397**	—
			Mean Aleatoric Uncertainty (AU)		
CIFAR-10	Standard	0.313	0.509	0.869	1.317
	Contrastive (Ours)	0.294	0.336	0.496	0.929
MNIST	Standard	0.014	0.038	0.420	0.529*
	Contrastive (Ours)	0.029	0.145	0.194	0.487*
CIFAR-100	Standard	0.615	0.991	1.607	2.296
	Contrastive (Ours)	1.994	1.985	2.223	2.490
			Mean Epistemic Uncertainty (EU)		
CIFAR-10	Standard	0.063	0.076	0.115	0.112
	Contrastive (Ours)	0.073	0.086	0.164	**0.266**
MNIST	Standard	0.005	0.006	0.133	0.247*
	Contrastive (Ours)	0.008	0.041	0.152	**0.296***
CIFAR-100	Standard	0.300	0.380	0.540	0.547
	Contrastive (Ours)	0.170	0.186	0.346	**0.459**

*For MNIST, OOD uncertainty is averaged across FashionMNIST, KMNIST, and EMNIST. (Xiao, Rasul, and Vollgraf 2017; Cohen et al. 2017; Clanuwat et al. 2018; Netzer et al. 2011; Krizhevsky, Hinton et al. 2009; LeCun et al. 2002)

5.4 Out-of-Distribution Detection

Distinguishing Distributional Shifts from Clean Data A key test for a well-estimated uncertainty model is its ability to distinguish shifted ID data and true OOD data. Table 2 shows this comparison.

The results reveal two critical patterns. First, when detecting **Nuisance Shift (Seen)** data, our contrastive model consistently yields a lower AUROC score than the baseline (e.g., 0.539 vs 0.579 EU-AUROC on CIFAR-10). This is a desirable property, confirming that our model correctly perceives these familiar nuisance variations as being "in-distribution" from a semantic standpoint and does not flag them as novel.

Second, for detecting true **Semantic Shift (OOD)**, our model is dramatically superior. On CIFAR-100, the EU-based AUROC for our model is **0.959**, a massive improvement over the baseline's 0.757. Similarly, on CIFAR-10, we see a jump from 0.743 to **0.876**. This demonstrates that the epistemic uncertainty learned via our contrastive method is a far more reliable and pure signal for detecting genuine, content-based novelty. While the baseline's Total or Aleatoric uncertainty sometimes achieves high AUROC, it does so by confounding data noise with semantic shifts, a problem our method mitigates. Commonly, epistemic uncertainty should be more related to OOD detection performance, the good performances of AU and EU of baseline model mostly come from the correlation of AU and EU (Mucsányi, Kirchhof, and Oh 2024), which is not desired, and one of the motivation of our methods is to reduce this kind of correlations.

Distinguishing Between Degraded ID data and OOD Data Finally, we test the model's ability to perform more nuanced distinctions, as shown in Table 3. A key challenge for robust models is to separate true semantic OOD samples from merely corrupted in-distribution ones.

Firstly, we tested whether the model can identify Nuisance (Seen) from Nuisance (Unseen), we see our method consistently gets lower AUROC score, which means our model treat Nuisance (Seen) and Nuisance (Unseen) similar, meaning our method concentrates more on the content part.

The results of "Degraded ID vs. OOD" are striking. When tasked with distinguishing **Nuisance (Seen) vs. Semantic OOD**, our model's epistemic uncertainty provides a powerful separative signal. For example, On CIFAR-100, it achieves an AUROC of **0.941**, whereas the baseline model is close to random chance at 0.675. This is a critical result: the baseline model, when presented with a familiar corrupted image, is almost as uncertain as when shown a true OOD image. Our model, however, has learned to be confident about the content of the familiar corruption (low EU) while being appropriately uncertain about the true OOD image (high EU), enabling clear separation.

Similarly, when distinguishing **Nuisance (Unseen) vs. Semantic OOD**, our model consistently maintains a stronger signal (e.g., EU-AUROC of 0.704 vs 0.482 on CIFAR-100), demonstrating its superior ability to disentangle content novelty from nuisance variations even in challenging scenarios.

6 Conclusion and Future Work

In this paper, we addressed the prevalent issue that traditional estimation of epistemic uncertainty will be polluted by nuisance factors. We argued that for uncertainty to be interpretable and reliable, a model's epistemic uncertainty (EU) should be grounded in semantic content, reflecting true model ignorance rather than superficial nuisance variations in the data. To this end, we proposed a training framework that uses contrastive learning to build representations that are invariant to known data corruptions. Our theoretical analysis provides a formal justification, showing that this approach minimizes the influence of nuisance variables on the final epistemic uncertainty estimate.

Our comprehensive experiments across MNIST, CIFAR-10, and CIFAR-100 provide strong empirical validation of

Table 2: OOD Detection vs. Clean ID Data (AUROC). Our method is intentionally less sensitive to seen nuisance shifts but vastly superior at detecting true semantic shifts using epistemic uncertainty (EU), highlighted in bold.

Dataset	**Uncertainty**	**vs. Nuisance (Seen)**(↓)		**vs. Nuisance (Unseen)**(↓)		**vs. Semantic (OOD)**(↑)	
		Standard	Contrastive	Standard	Contrastive	Standard	Contrastive
CIFAR-10	Total	0.618	**0.536**	0.775	**0.672**	**0.925**	0.890
	Aleatoric	0.623	**0.535**	0.782	**0.659**	**0.936**	0.883
	Epistemic	0.579	**0.539**	0.697	**0.691**	0.743	**0.876**
MNIST	Total	**0.712**	0.788	0.922	**0.883**	**0.973**	0.954
	Aleatoric	**0.713**	0.788	0.923	**0.876**	**0.972**	0.950
	Epistemic	**0.698**	0.790	0.917	**0.899**	0.974	**0.978**
CIFAR-100	Total	0.628	**0.502**	0.807	**0.607**	**0.942**	0.704
	Aleatoric	0.634	**0.497**	0.810	**0.564**	**0.950**	0.637
	Epistemic	0.587	**0.542**	**0.726**	0.797	0.757	**0.959**

Table 3: Detection Performance Between Degraded and OOD Data (AUROC). Our model's epistemic uncertainty provides a significantly stronger signal for distinguishing true semantic OOD from both seen and unseen nuisance shifts.

Dataset	**Uncertainty**	**Nuisance (Seen) vs. (Unseen)**(↓)		**Nuisance (Seen) vs. Semantic OOD**(↑)		**Nuisance (Unseen) vs. Semantic OOD**(↑)	
		Standard	Contrastive	Standard	Contrastive	Standard	Contrastive
CIFAR-10	Total	0.677	**0.639**	0.859	**0.871**	0.701	**0.774**
	Aleatoric	0.678	**0.627**	**0.868**	0.864	0.720	**0.784**
	Epistemic	0.683	**0.656**	0.683	**0.850**	0.523	**0.703**
MNIST*	Total	0.851	**0.634**	0.701	**0.766**	0.649	**0.799**
	Aleatoric	0.848	**0.604**	0.689	**0.788**	0.590	**0.817**
	Epistemic	0.855	**0.681**	0.707	**0.714**	0.752	**0.757**
CIFAR-100	Total	0.700	**0.606**	**0.862**	0.703	**0.689**	0.581
	Aleatoric	0.695	**0.568**	**0.870**	0.631	**0.715**	0.560
	Epistemic	**0.656**	0.765	0.675	**0.941**	0.482	**0.704**

*For MNIST, OOD uncertainty is averaged across FashionMNIST, KMNIST, and EMNIST.

our claims. The results consistently demonstrate that our contrastively trained model learns a more disentangled and meaningful form of uncertainty. As per our core expectation, the model correctly identifies low-quality in-distribution data (Nuisance Shifts) as being semantically familiar, evidenced by two key findings: (1) its epistemic uncertainty remains low and stable, comparable to that on clean data, and (2) its ability to distinguish these nuisance-shifted samples from clean data is near random chance, the desired outcome for nuisance invariance. Conversely, the model's EU proves to be a powerful and pure signal for detecting true, content-based novelty. This is most evident in the semantic OOD detection tasks, where our method dramatically outperforms the standard baseline (e.g., achieving an EU-based AUROC of 0.959 vs. 0.757 on CIFAR-100). Furthermore, our model excels at the critical task of separating true OOD samples from merely corrupted ones—a scenario where the baseline model fails.

A limitation of our current approach is that the contrastive pre-training of the encoder is entirely self-supervised and does not leverage class labels. This may potentially limit the discriminative power of the learned features for the downstream classification task, which could contribute to the slightly lower clean accuracy observed on CIFAR-100. This opens up a clear avenue for future work: exploring supervised or semi-supervised contrastive learning techniques, (Khosla et al. 2020). By integrating label information into the contrastive objective, it may be possible to enhance classification performance while retaining the robust, semantic uncertainty properties demonstrated in this work. Ultimately, this research paves the way for building more reliable models where the reported uncertainties are more closely and interpretably aligned with their intended real-world meaning.

References

Bengio, Y.; Courville, A.; and Vincent, P. 2013. Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, 35(8): 1798–1828.

Chen, T.; Kornblith, S.; Norouzi, M.; and Hinton, G. 2020. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, 1597–1607. PmLR.

Clanuwat, T.; Bober-Irizar, M.; Kitamoto, A.; Lamb, A.; Yamamoto, K.; and Ha, D. 2018. Deep Learning for Classical Japanese Literature. .

Cohen, G.; Afshar, S.; Tapson, J.; and Van Schaik, A. 2017. EMNIST: Extending MNIST to handwritten letters. In *2017 international joint conference on neural networks (IJCNN)*, 2921–2926. IEEE.

Davis, J.; and Goadrich, M. 2006. The relationship between Precision-Recall and ROC curves. In *Proceedings of the 23rd international conference on Machine learning*, 233–240.

Depeweg, S.; Hernandez-Lobato, J.-M.; Doshi-Velez, F.; and Udluft, S. 2018. Decomposition of uncertainty in Bayesian deep learning for efficient and risk-sensitive learning. In *International conference on machine learning*, 1184–1193. PMLR.

Der Kiureghian, A.; and Ditlevsen, O. 2009. Aleatory or epistemic? Does it matter? *Structural safety*, 31(2): 105–112.

Gal, Y.; and Ghahramani, Z. 2016. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*, 1050–1059. PMLR.

He, K.; Fan, H.; Wu, Y.; Xie, S.; and Girshick, R. 2020. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 9729–9738.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep Residual Learning for Image Recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Hendrycks, D.; and Dietterich, T. 2019. Benchmarking neural network robustness to common corruptions and perturbations. *arXiv preprint arXiv:1903.12261*.

Hendrycks, D.; and Gimpel, K. 2016. A baseline for detecting misclassified and out-of-distribution examples in neural networks. *arXiv preprint arXiv:1610.02136*.

Hüllermeier, E.; and Waegeman, W. 2021. Aleatoric and epistemic uncertainty in machine learning: An introduction to concepts and methods. *Machine learning*, 110(3): 457–506.

Kendall, A.; and Gal, Y. 2017. What uncertainties do we need in bayesian deep learning for computer vision? In *Advances in neural information processing systems*, 5574–5584.

Khosla, P.; Teterwak, P.; Wang, C.; Sarna, A.; Tian, Y.; Isola, P.; Maschinot, A.; Liu, C.; and Krishnan, D. 2020. Supervised contrastive learning. *Advances in neural information processing systems*, 33: 18661–18673.

Krizhevsky, A.; Hinton, G.; et al. 2009. Learning multiple layers of features from tiny images.(2009).

Lakshminarayanan, B.; Pritzel, A.; and Blundell, C. 2017. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Advances in neural information processing systems*, 6402–6413.

LeCun, Y.; Bottou, L.; Bengio, Y.; and Haffner, P. 2002. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11): 2278–2324.

Locatello, F.; Bauer, S.; Lucic, M.; Raetsch, G.; Gelly, S.; Schölkopf, B.; and Bachem, O. 2019. Challenging common assumptions in the unsupervised learning of disentangled representations. In *international conference on machine learning*, 4114–4124. PMLR.

Mucsányi, B.; Kirchhof, M.; and Oh, S. J. 2024. Benchmarking uncertainty disentanglement: Specialized uncertainties for specialized tasks. *Advances in neural information processing systems*, 37: 50972–51038.

Nayman, N.; Noy, A.; Halperin, T.; Avidan, S.; and Gal, Y. 2024. The Unreasonable Effectiveness of Post-hoc Uncertainty Quantification. *arXiv preprint arXiv:2402.19460*.

Netzer, Y.; Wang, T.; Coates, A.; Bissacco, A.; Wu, B.; Ng, A. Y.; et al. 2011. Reading digits in natural images with unsupervised feature learning. In *NIPS workshop on deep learning and unsupervised feature learning*, volume 2011, 7. Granada.

Oord, A. v. d.; Li, Y.; and Vinyals, O. 2018. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*.

Tsai, Y.-H. H.; Wu, Y.; Salakhutdinov, R.; and Morency, L.-P. 2020. Self-supervised learning from a multi-view perspective. *arXiv preprint arXiv:2006.05576*.

von Kügelgen, J.; Sharma, Y.; Gresele, L.; Brendel, W.; Schölkopf, B.; Besserve, M.; and Locatello, F. 2021. Self-Supervised Learning with Data Augmentations Provably Isolates Content from Style. In Ranzato, M.; Beygelzimer, A.; Dauphin, Y.; Liang, P.; and Vaughan, J. W., eds., *Advances in Neural Information Processing Systems*, volume 34, 16451–16467. Curran Associates, Inc.

Wang, T.; and Isola, P. 2020. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *International conference on machine learning*, 9929–9939. PMLR.

Xiao, H.; Rasul, K.; and Vollgraf, R. 2017. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*.

Yang, J.; Zhou, K.; Li, Y.; and Liu, Z. 2024. Generalized out-of-distribution detection: A survey. *International Journal of Computer Vision*, 132(12): 5635–5662.

A Proofs of Theoretical Results

Lemma 3 (Content Informativeness via Contrastive Learning)**.** *Let $x = g(c, n)$ with $c \sim p(c)$ and $n \sim p(n)$ being independent. Let $z = f_\phi(x)$ be the representation learned by minimizing the InfoNCE loss using positive pairs (x_1, x_2) where $x_1 = g(c, n_1)$ and $x_2 = g(c, n_2)$ share the same content c and have independent nuisances n_1, n_2. Optimizing the InfoNCE objective generally encourages an increase in the mutual information between the representation and the content variable:*

$$\mathbb{I}(c; z) \uparrow . \tag{6}$$

Proof. By the variational lower bound (Oord, Li, and Vinyals 2018), the InfoNCE loss satisfies

$$I(z_1; z_2) \geq \log N - L_{\text{InfoNCE}},$$

where $z_i = f_\phi(x_i)$ for $i = 1, 2$. Since $x_1 = g(c, n_1)$ and $x_2 = g(c, n_2)$ share only the content c and have independent nuisances, we have the Markov chain

$$z_1 \leftarrow x_1 \leftarrow c \rightarrow x_2 \rightarrow z_2.$$

Given the content variable c, the generation processes for x_1 and x_2 are independent. This means that z_1 and z_2 are conditionally independent given c. By the data-processing inequality,

$$I(z_1; z_2) \leq I(c; z_2) = I(c; z).$$

Combining these two inequalities yields Equation (6). Hence, as L_{InfoNCE} decreases, $I(c; z)$ must increase, proving the lemma. □

A.1 Lemma 2 (Nuisance Invariance via Contrastive Learning)

Lemma 4 (Nuisance Invariance via Contrastive Learning)**.** *Under the same assumptions as Lemma 3, optimizing the InfoNCE objective encourages a reduction in the mutual information between the nuisance variable and the representation, conditioned on the content:*

$$\mathbb{I}(n; z \mid c) \searrow 0.$$

Proof. Let (z_1, z_2) be the representations of two views (x_1, x_2) sharing the same content c but independent nuisances. (Wang and Isola 2020) show that minimizing the InfoNCE objective enforces

$$\mathbb{E}_{c,n_1,n_2}\left[\|z_1 - z_2\|^2\right] \to 0$$

as $L_{\text{InfoNCE}} \to 0$. Consequently, for each fixed c, the distributions $p(z \mid c, n)$ and $p(z \mid c)$ converge in KL divergence:

$$D_{\text{KL}}\big(p(z \mid c, n) \| p(z \mid c)\big) \to 0.$$

Taking expectation over (c, n),

$$I(n; z \mid c) = \mathbb{E}_{c,n}\big[D_{\text{KL}}(p(z \mid c, n) \| p(z \mid c))\big] \to 0,$$

which completes the proof. □

A.2 Proposition 1 (Alignment of Observation- and Content-level Epistemic Uncertainty)

Proposition 2 (Epistemic Uncertainty Alignment via Contrastive Representations)**.** *Let $EU_x = \mathbb{I}(y; \theta \mid c, n, \mathcal{D})$ and $EU_c = \mathbb{I}(y; \theta \mid c, \mathcal{D})$ be the epistemic uncertainties defined above. Assuming the learned representation $z = f_\phi(x)$ captures the relevant information pathways, the absolute difference between these two uncertainty measures can be bounded in terms of the nuisance information retained by the representation:*

$$|EU_x - EU_c| \leq \mathbb{I}(n; z \mid c) \to 0. \tag{7}$$

Proof. Starting from the chain rule of conditional mutual information,

$$I(y; \theta \mid c, n, D) = I(y; \theta \mid c, D) - I(y; n \mid c, D) + I(y; n \mid c, \theta, D).$$

By definition, $\text{EU}_x = I(y; \theta \mid c, n, D)$ and $\text{EU}_c = I(y; \theta \mid c, D)$, so

$$\text{EU}_x - \text{EU}_c = I(y; n \mid c, \theta, D) - I(y; n \mid c, D).$$

Since $0 \leq I(y; n \mid c, \theta, D) \leq I(y; n \mid c, D)$

$$\big|\text{EU}_x - \text{EU}_c\big| \leq I(y; n \mid c, D).$$

Under the assumption that $(c, n) \to z \to y$ form a Markov chain conditioned on c, the data-processing inequality gives

$$I(y; n \mid c, D) \leq I(z; n \mid c, D) \leq I(n; z \mid c).$$

This establishes Equation (7). By Lemma 2, $I(n; z \mid c) \to 0$ under contrastive pre-training, completing the proof. □

B Additional Experimental Results

B.1 Impact of Pretrained Encoder Initialization

To further investigate the role of encoder initialization in our framework, we conduct an additional experiment on CIFAR-10 where we initialize the encoder using a pretrained ResNet-18 model (trained on ImageNet), instead of training it from scratch as done in our main contrastive setting. The rest of the contrastive training and classifier fine-tuning process remains unchanged.

Table 4 summarizes the classification accuracy and uncertainty statistics under this pretrained setup. Consistent with our main findings, we observe that the contrastive-learned model still maintains low epistemic uncertainty (EU) on seen nuisance shifts (no big gap with clean ID), and exhibits higher EU when encountering true semantic OOD data (more far away from shifted ID). The overall trends of uncertainty behavior remain similar, though the pretrained encoder yields modest improvements in clean accuracy and robustness on unseen nuisance shifts.

But we also see the accuracy of standard model is much higher than contrastive-trained model, which is expected since the standard model can use the label information. With the pre-trained model having basic vision ability, the standard model will learn to classify better with label information, while our contrastive-trained model did not use.

Table 5 summarizes the similar conclusion with the original setting under this pretrained setup. Consistent with our

Table 4: Accuracy and Mean Uncertainty Scores on CIFAR-10 with pre-trained encoder. Our contrastive method maintains low epistemic uncertainty (EU) on familiar nuisance shifts while correctly elevating it for unseen and semantic shifts, unlike the standard baseline which shows increased EU even for familiar corruptions.

Dataset	Model	Clean ID	Nuisance Shift (Seen)	Nuisance Shift (Unseen)	Semantic Shift (OOD)
			Test Accuracy *(↑)*		
CIFAR-10	Standard	**0.958**	**0.916**	**0.783**	—
	Contrastive (Ours)	0.907	0.881	0.669	—
			Mean Aleatoric Uncertainty (AU)		
CIFAR-10	Standard	0.0824	0.204	0.569	1.2731
	Contrastive (Ours)	0.249	0.286	0.498	0.9158
			Mean Epistemic Uncertainty (EU)		
CIFAR-10	Standard	0.039	0.060	0.114	0.156
	Contrastive (Ours)	0.066	0.079	0.173	**0.276**

Table 5: OOD Detection vs. Clean ID Data (AUROC) on CIFAR-10 with pre-trained encoder. Our method is intentionally less sensitive to seen nuisance shifts but vastly superior at detecting true semantic shifts using epistemic uncertainty (EU), highlighted in bold.

Dataset	Uncertainty	**vs. Nuisance (Seen)**(↓)		**vs. Nuisance (Unseen)**(↓)		**vs. Semantic (OOD)**(↑)	
		Standard	Contrastive	Standard	Contrastive	Standard	Contrastive
CIFAR-10	Total	0.630	**0.534**	0.824	**0.699**	**0.985**	0.905
	Aleatoric	0.634	**0.532**	0.832	**0.689**	**0.990**	0.899
	Epistemic	0.614	**0.536**	0.778	**0.711**	0.889	**0.893**

main findings, we observe that our contrastive-trained model works better on not distinguishing shifted ID data and our semantic epistemic uncertainty works slightly better on identifying true OOD. This means, even the contrastive-trained model does not have a good classification performance, but it still works better on identifying low-quality ID and true OOD.

These results indicate that using a pretrained encoder does not undermine the disentanglement effect achieved by contrastive training. On the contrary, it can complement our approach by injecting better general visual priors, suggesting an effective path toward improving uncertainty quality.

B.2 AUPR Metrics for OOD and Degraded Detection

In the main paper (Tables 2 and 3), we reported AUROC as the primary metric for evaluating OOD and distributional shift detection. Here, we supplement those results with additional evaluation using the Area Under the Precision-Recall Curve (AUPR), which provides a complementary view, especially under class imbalance.

Table 6 reports AUPR scores corresponding to the same scenarios evaluated in Table 2. Table 7 provides AUPR values for the degraded-vs-OOD settings from Table 3.

We observe that the trends in AUPR mirror those in AUROC: our contrastively trained model consistently yields lower AUPR scores for nuisance shift detection (desirable) and higher AUPR scores for semantic OOD detection (also desirable). And our contrastive-trained model has the best improvement on CIFAR-100 with semantic epistemic uncertainty. This confirms the robustness and consistency of our model's epistemic uncertainty in distinguishing content-level novelty from nuisance-level shifts, validating our findings across multiple evaluation metrics.

C Experimental Settings

C.1 Contrastive Model Settings

We use a contrastive training pipeline designed to disentangle semantic content from nuisance factors. Each encoder in the ensemble is based on a ResNet-18 backbone without pretraining. The projection head is a two-layer MLP with batch normalization and ReLU, outputting 128-dimensional L2-normalized features. Our contrastive loss follows the NT-Xent formulation with a temperature of 0.07.

The dataset used is MNIST/CIFAR-10/CIFAR-100, with a training-validation split of 90/10. For contrastive training, each image is transformed into two augmented views: one clean and one corrupted using a randomly sampled nuisance type (e.g., blur, noise, brightness). All augmentations are applied before normalization. Corruptions follow CIFAR-10-C style and are applied at severity level 2. The encoder is trained using AdamW with a cosine learning rate schedule for 1500 epochs. Batch size is 1024, and mixed-precision training is enabled.

After the encoder is trained, we freeze it and train an MLP

Table 6: OOD Detection vs. Clean ID Data (AUPR). Our method is intentionally less sensitive to seen nuisance shifts but vastly superior at detecting true semantic shifts using epistemic uncertainty (EU), highlighted in bold.

Dataset	Uncertainty	vs. Nuisance (Seen)(↓)		vs. Nuisance (Unseen)(↓)		vs. Semantic (OOD)(↑)	
		Standard	Contrastive	Standard	Contrastive	Standard	Contrastive
CIFAR-10	Total	0.586	**0.529**	0.653	**0.653**	0.911	**0.943**
	Aleatoric	0.592	**0.526**	0.655	**0.627**	0.915	**0.935**
	Epistemic	0.539	**0.535**	**0.612**	0.701	0.823	**0.939**
MNIST	Total	**0.676**	0.771	0.931	**0.865**	**0.973**	0.955
	Aleatoric	**0.681**	0.771	0.931	**0.840**	**0.972**	0.947
	Epistemic	**0.648**	0.765	0.922	**0.892**	**0.974**	0.962
CIFAR-100	Total	0.643	**0.498**	0.827	**0.590**	**0.974**	0.801
	Aleatoric	0.653	**0.492**	0.829	**0.548**	**0.978**	0.750
	Epistemic	**0.556**	0.565	**0.686**	0.845	0.827	**0.985**

Table 7: Detection Performance Between Degraded and OOD Data (AUPR). Our model's epistemic uncertainty provides a significantly stronger signal for distinguishing true semantic OOD from both seen and unseen nuisance shifts.

Dataset	Uncertainty	Nuisance (Seen) vs. (Unseen)(↓)		Nuisance (Seen) vs. Semantic OOD(↑)		Nuisance (Unseen) vs. Semantic OOD(↑)	
		Standard	Contrastive	Standard	Contrastive	Standard	Contrastive
CIFAR-10	Total	0.682	**0.620**	0.926	**0.932**	0.818	**0.877**
	Aleatoric	0.678	**0.597**	**0.936**	0.923	0.843	**0.880**
	Epistemic	0.776	**0.665**	0.776	**0.923**	0.686	**0.827**
MNIST*	Total	0.876	**0.620**	0.631	**0.740**	0.592	**0.776**
	Aleatoric	0.870	**0.560**	0.619	**0.774**	0.523	**0.794**
	Epistemic	0.881	**0.722**	**0.670**	0.632	**0.770**	0.707
CIFAR-100	Total	0.683	**0.594**	**0.928**	0.804	**0.834**	0.724
	Aleatoric	0.677	**0.558**	**0.932**	0.756	**0.847**	0.707
	Epistemic	**0.631**	0.808	0.784	**0.977**	0.675	**0.804**

*For MNIST, OOD uncertainty is averaged across FashionMNIST, KMNIST, and EMNIST.

classifier head on top using clean data. The MLP classifier includes one hidden layer (dimension 512) with dropout (rate 0.5), batch normalization, and ReLU. Training is performed with AdamW for 100 epochs using a ReduceLROnPlateau scheduler. Early stopping is applied with patience of 100 epochs.

Each ensemble consists of 10 such encoder-classifier pairs, trained with different seeds for diversity.

C.2 Baseline Model Settings

The baseline model follows standard supervised training using ResNet-18 without pretraining. To better reflect robustness in real-world scenarios, we train on a mixed batch of clean and corrupted images. Corruptions used are the same as in the contrastive setting and applied at severity level 2.

Training is conducted using SGD with momentum (0.9), initial learning rate 0.01, and weight decay 1e-4. Learning rate milestones are set at epochs 30, 60, and 80. Each model is trained for up to 100 epochs with early stopping (patience 15), and gradients are clipped at a maximum norm of 1.0.

The validation and test sets include both clean and corrupted versions of the same CIFAR-10 images, simulating mixed-domain evaluation. Models are evaluated individually and as ensembles using average logits.

C.3 Evaluation Protocol for Uncertainty and OOD Detection

To evaluate uncertainty and out-of-distribution (OOD) detection performance, we design a unified evaluation pipeline applicable to both our proposed contrastive model and the supervised baseline ensemble. Each model consists of an ensemble of 10 independently trained classifiers. During evaluation, we subsample $M = 8$ ensemble members and perform uncertainty-based inference over $K = 5$ randomized runs, ensuring robust and fair estimation.

Datasets. We use the CIFAR-10/MNIST/CIFAR-100 test set as the in-distribution (ID) dataset. To simulate various distribution shifts, we evaluate on:

- **ID corruption (Nuisance, Seen):** Gaussian noise, brightness, and other corruptions used during contrastive training.
- **OOD corruption (Nuisance, Unseen):** Severe perturbations such as shot noise, impulse noise, defocus blur, glass blur, and snow.
- **SVHN/Fashion-MNIST/EMNIST/KMNIST** as natural semantic shifts representing entirely different label spaces.

Uncertainty Estimation. For each ensemble subset, we collect logits $\{f_m(x)\}_{m=1}^{M}$ for input x and compute the predictive probability $p(y|x) = \frac{1}{M}\sum_{m=1}^{M} \text{softmax}(f_m(x))$. The following uncertainty scores are calculated:

- **Total Uncertainty (TU):** Predictive entropy of $p(y|x)$.
- **Aleatoric Uncertainty (AU):** Mean entropy of individual model predictions $\text{AU} = 1/M \sum_{m=1}^{M} \mathbb{H}[p_m(y|x)]$.
- **Epistemic Uncertainty (EU):** Defined as $\text{TU} - \text{AU}$, reflecting model disagreement.
- **Max Probability (MP):** The maximum softmax value in $p(x)$, used as a confidence proxy.

OOD Detection. We evaluate OOD detection by comparing uncertainty distributions between an ID dataset (e.g., CIFAR-10 clean) and each OOD or corrupted dataset. For each uncertainty metric, we compute the Area Under the Receiver Operating Characteristic curve (AUROC) and the Area Under the Precision-Recall curve (AUPR). Both "Clean vs. OOD" and "Corrupt (ID) vs. OOD" comparisons are performed.

Aggregation and Visualization. All metrics are averaged across the $K = 5$ subsampling runs.

Implementation. The evaluation code is implemented in PyTorch and reuses trained checkpoints. For fair comparison, both contrastive and baseline models use the same uncertainty computation function and data processing pipeline. All experiments are run on a single GPU (RTX 4090) with batch size 1024.