

HUE-OOD: Hybrid Uncertainty–Evidential Dynamics for Out-of-Distribution Detection

Phat Nguyen

521H0126@student.tdtu.edu.vn

Ton Duc Thang University

Research Article

Keywords: Out-of-Distribution Detection, Uncertainty Quantification Evidential Learning, Mahalanobis Distance, Monte Carlo Dropout Rank-based Fusion

Posted Date: September 3rd, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10693574/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License.
[Read Full License](#)

Additional Declarations: No competing interests reported.

HUE-OOD: Hybrid Uncertainty–Evidential Dynamics for Out-of-Distribution Detection

Nguyen Hai Tien Phat

Faculty of Information Technology, Ton Duc Thang University, Ho Chi Minh City,
Vietnam

phathanos0907@gmail.com

<https://orcid.org/0009-0006-9059-9345>

Abstract. Post-hoc out-of-distribution (OOD) detection is attractive because it can augment pretrained classifiers without modifying their training procedure. However, individual detection scores may respond differently to distribution shift, motivating the combination of complementary uncertainty signals. This paper presents HUE-OOD, a post-hoc framework that combines predictive, representational, and evidential information through parameter-free rank-based fusion. Predictive uncertainty is estimated from stochastic forward passes using standard, spatial, and concrete dropout variants. Representational uncertainty is quantified using class-conditional Mahalanobis distance in the penultimate feature space, while evidential uncertainty is obtained by mapping classifier logits to non-negative Dirichlet evidence through a soft-plus transformation. The resulting scores are converted to relative ranks before aggregation, avoiding direct comparison of heterogeneous score scales. We evaluate the framework across 20 OOD benchmarks and 20 backbone architectures and compare it with 20 post-hoc baselines. In the reported experiments, HUE-OOD obtains a mean AUROC of 0.968 ± 0.004 and a mean AUPR of 0.978. The results indicate that combining the three signals can improve detection robustness across different OOD regimes, while ablation experiments quantify the contribution of each component. The method requires only a pretrained classifier and does not introduce additional training or architectural changes.

Keywords: Out-of-Distribution Detection Uncertainty Quantification
Evidential Learning Mahalanobis Distance Monte Carlo Dropout
Rank-based Fusion

1 Introduction

Deep neural networks have achieved strong performance on a wide range of visual recognition tasks, but high predictive accuracy on in-distribution (ID) data does not necessarily imply reliable behavior on samples that differ from the training distribution. In safety-sensitive applications, including autonomous systems, medical image analysis, and biometric recognition, an incorrect prediction accompanied by high confidence can be particularly problematic. OOD detection

therefore aims to identify test inputs for which the predictive behavior of a model may not be supported by the distribution on which the model was trained [1].

A practical requirement in many deployment settings is that OOD detection should be applicable to an already trained classifier. Post-hoc methods address this requirement by deriving an OOD score from an existing model without changing its learned parameters. Representative approaches use the maximum softmax probability (MSP) [1], temperature-scaled confidence and input perturbation [2], energy-based scores [4], feature-space distances [3], feature residuals and norms [6], or activation processing [7]. These methods are attractive because they can often be incorporated into an existing inference pipeline with limited modification.

Nevertheless, the information available in a pretrained network is not restricted to a single scalar confidence value. A test sample can induce different forms of model behavior at different stages of inference. For example, stochastic perturbations may produce unstable predictions, the corresponding latent representation may be distant from class-conditional training features, and the final logits may provide relatively little evidential support for any particular class. Conversely, an OOD sample may appear anomalous according to one signal but comparatively typical according to another. This observation motivates the use of multiple signals rather than assuming that a single uncertainty surrogate is sufficient for all distribution shifts.

In this work, we consider three complementary quantities. The first is *predictive uncertainty*, which measures variation in model outputs under stochastic forward passes. Following the Monte Carlo dropout interpretation of dropout as an approximate Bayesian procedure [11], we use standard, spatial, and concrete dropout mechanisms to obtain multiple stochastic predictions. The second is *representational uncertainty*, expressed through the class-conditional Mahalanobis distance between a test representation and the estimated feature distribution of each class [3]. The third is *evidential uncertainty*, obtained by interpreting transformed classifier logits as non-negative evidence parameters of a Dirichlet distribution. This construction is related to evidential approaches to predictive uncertainty [14, 15], but is used here as a post-hoc scoring mechanism rather than as a replacement for the original training objective.

The three quantities are heterogeneous in both interpretation and numerical scale. Directly adding them therefore requires a normalization or weighting scheme whose behavior can depend on the dataset, architecture, or score distribution. HUE-OOD instead applies a parameter-free rank transformation to each uncertainty score and combines the resulting relative positions. The purpose of this operation is not to claim that rank aggregation is universally optimal, but to provide a simple and scale-insensitive aggregation rule that can be evaluated independently of score magnitudes. This design also avoids introducing fusion parameters that would otherwise require validation data or additional optimization.

The resulting framework, termed *HUE-OOD* (Hybrid Uncertainty–Evidential Dynamics for OOD detection), is designed as a post-hoc detector. It does not re-

quire auxiliary OOD samples, retraining of the backbone, or changes to the classifier architecture. The additional computation consists of stochastic inference for the predictive branch, feature-space distance computation for the representational branch, and a single forward-pass evidential calculation. Consequently, the method can be evaluated as an inference-time extension of an existing classifier.

We conduct an evaluation covering 20 OOD datasets and 20 backbone architectures, including convolutional, transformer-based, and hybrid models. HUE-OOD is compared with 20 post-hoc detection methods representing confidence-, energy-, feature-, activation-, nearest-neighbor-, and gradient-based approaches. The primary evaluation uses AUROC and AUPR, with additional analyses examining individual uncertainty branches, alternative fusion mechanisms, cross-architecture behavior, and computational overhead. In the reported experiments, the proposed method obtains a mean AUROC of 0.968 ± 0.004 and a mean AUPR of 0.978 over the evaluated configurations. These values should be interpreted as aggregate results for the experimental protocol considered in this study, rather than as evidence that the method dominates all possible OOD settings.

The main contributions of this study are summarized as follows:

- We formulate a post-hoc OOD detection framework that combines predictive, representational, and evidential information obtained from a single pre-trained classifier.
- We introduce a parameter-free rank-based fusion procedure for combining heterogeneous uncertainty scores without requiring a learned fusion model or an explicitly tuned weighted sum.
- We provide an empirical analysis of the individual uncertainty branches and their combinations, allowing the contribution of each source to be assessed through controlled ablation experiments.
- We evaluate the proposed detector across multiple OOD regimes, backbone families, and post-hoc baselines, together with statistical and computational analyses intended to characterize both detection performance and practical inference cost.

The remainder of the paper is organized as follows. Section 2 reviews related work on post-hoc OOD detection, uncertainty estimation, and score fusion. Section 3 defines the HUE-OOD formulation and its three uncertainty branches. Section 4 describes the experimental protocol and reports the main results and ablation studies. Section 5 discusses limitations and practical considerations. Section 6 concludes the paper.

2 Related Work

OOD detection has developed through several methodological directions, including confidence-based scoring, feature-space modeling, activation processing, gradient-based analysis, uncertainty quantification, and multi-score aggregation. This section focuses on the aspects most directly related to HUE-OOD.

2.1 Post-hoc OOD Detection

Post-hoc OOD detectors operate on a pretrained model and derive an additional score at inference time. This setting is particularly useful when retraining the original classifier is undesirable or when the detector must be attached to an existing model.

The maximum softmax probability (MSP) is one of the most widely used post-hoc baselines. Hendrycks and Gimpel [1] observed that the maximum predicted probability can provide a useful separation between ID and OOD samples. Although computationally simple, MSP depends exclusively on the final predictive distribution and can remain high for some OOD inputs.

ODIN combines temperature scaling with a small input perturbation to increase the separation between ID and OOD examples [2]. Energy-based detection subsequently proposed using a scalar derived from the log-sum-exp of the logits rather than the maximum softmax probability [4]. These approaches operate primarily at the output layer and therefore provide limited direct information about the geometry of intermediate representations.

Feature-space methods provide a different perspective. Lee et al. [3] proposed a Mahalanobis-distance-based detector that models class-conditional distributions in feature space. For a feature vector $\mathbf{z}$ and class c , the corresponding distance can be written as

$$d_c(\mathbf{z}) = (\mathbf{z} - \boldsymbol{\mu}_c)^\top \boldsymbol{\Sigma}^{-1}(\mathbf{z} - \boldsymbol{\mu}_c), \quad (1)$$

where $\boldsymbol{\mu}_c$ denotes the estimated class mean and $\boldsymbol{\Sigma}$ denotes a covariance estimate. The minimum distance over classes can then be used as an OOD score after selecting the appropriate score orientation. This formulation incorporates feature correlations and therefore differs from confidence measures based only on the output layer.

Several subsequent methods have extended or modified feature-based detection. Relative Mahalanobis distance compares class-specific distances with a background distribution, providing an alternative geometric reference [5]. VIM combines the feature norm with a residual component obtained from a principal subspace [6]. ReAct modifies intermediate activations by clipping large values before computing the detection score [7]. DICE instead uses a sparsification strategy to retain a selected subset of network parameters or activations during inference [8]. Other approaches exploit nearest-neighbor relations in representation space or energy-like functions defined over learned features.

Gradient-based approaches constitute another branch of post-hoc detection. GradNorm uses the magnitude of gradients associated with the predicted class as an OOD signal [9]. Such methods probe sensitivity of the network response rather than relying exclusively on the output probability or latent distance.

These approaches demonstrate that different parts of a pretrained model can provide useful OOD information. However, their scores are generally designed around one primary signal. HUE-OOD does not attempt to replace these individual detectors; instead, it investigates whether information from three distinct sources can be combined using a common, non-parametric aggregation rule.

2.2 Uncertainty Quantification

Uncertainty estimation provides a complementary perspective on OOD detection. A common distinction separates aleatoric uncertainty, which is associated with inherent ambiguity in observations, from epistemic uncertainty, which reflects uncertainty associated with the model or its parameters [13]. In practice, the boundary between these concepts depends on the probabilistic model and estimation procedure. Therefore, in this work we use the terms primarily to describe the operational quantities computed by the detector.

Stochastic Predictive Uncertainty. Dropout was originally introduced as a regularization mechanism [10]. Gal and Ghahramani [11] showed that retaining dropout during inference and averaging multiple stochastic predictions can be interpreted as an approximate Bayesian inference procedure. Given T stochastic forward passes, the predictive mean can be expressed as

$$\bar{\mathbf{p}} = \frac{1}{T} \sum_{t=1}^T \mathbf{p}^{(t)}, \quad (2)$$

where $\mathbf{p}^{(t)}$ is the class-probability vector produced by the t -th stochastic pass. Prediction variability can then be quantified using measures such as predictive entropy, variance, or disagreement.

Different dropout mechanisms perturb the network in different ways. Standard dropout removes individual activations, spatial dropout can remove entire feature maps, and concrete dropout provides a continuous relaxation for learning dropout probabilities [12]. These mechanisms need not provide independent uncertainty estimates in a strict statistical sense; rather, they induce different stochastic perturbation patterns. HUE-OOD uses this diversity to construct a predictive uncertainty signal based on the observed variation of predictions.

Evidential Uncertainty. Evidential deep learning represents class probabilities through a Dirichlet distribution whose concentration parameters are related to non-negative evidence [14]. If e_c denotes the evidence associated with class c , the corresponding concentration parameter can be written as

$$\alpha_c = e_c + 1, \quad S = \sum_{c=1}^C \alpha_c. \quad (3)$$

A commonly used uncertainty quantity is related to the inverse of the total evidence or total concentration. In HUE-OOD, classifier logits ℓ_c are transformed into non-negative evidence using

$$e_c = \text{softplus}(\ell_c), \quad (4)$$

followed by Eq. 3. This construction does not imply that the pretrained classifier was explicitly trained under a Dirichlet likelihood. Rather, it provides a

post-hoc evidential interpretation of the available logits. Accordingly, we treat the resulting quantity as an *evidential uncertainty score* rather than as a calibrated posterior probability.

This distinction is important because post-hoc transformations of logits cannot, by themselves, establish that the resulting Dirichlet model is a correctly specified probabilistic model. The purpose of the evidential branch is instead to provide an additional score that captures information not explicitly represented by the feature-distance branch or stochastic prediction variability.

Representational Uncertainty. The Mahalanobis formulation in Eq. 1 can also be interpreted as a measure of representational conformity. A sample close to one or more estimated class-conditional feature distributions has greater geometric similarity to the training representation, whereas a large distance indicates lower conformity under the assumed Gaussian model.

This interpretation differs from predictive uncertainty. Two samples can have similar output probabilities while occupying different locations in feature space. Conversely, stochastic predictions can vary for a sample whose deterministic representation is relatively close to a class centroid. These observations motivate treating the feature-space and predictive signals as distinct quantities rather than assuming that one can be substituted for the other.

2.3 Score Fusion and Rank Aggregation

Combining multiple OOD scores introduces an additional methodological problem: heterogeneous scores may have different numerical ranges, orientations, and distributions. A weighted linear combination,

$$s(\mathbf{x}) = \sum_{k=1}^K w_k s_k(\mathbf{x}), \quad (5)$$

requires both a common score orientation and an appropriate choice of weights. Normalization can reduce scale differences, but normalization statistics themselves may depend on the evaluation distribution. A learned fusion model can provide greater flexibility, but it introduces additional parameters and may require ID/OOD validation data.

Rank aggregation provides a simple alternative. For a score s_k and a set of N samples, let $r_k(\mathbf{x})$ denote the normalized rank of sample $\mathbf{x}$ under score k . The fused score can then be defined as

$$s_{\text{rank}}(\mathbf{x}) = \frac{1}{K} \sum_{k=1}^K r_k(\mathbf{x}). \quad (6)$$

Because rank depends on relative ordering rather than absolute magnitude, this operation is invariant to strictly monotonic transformations of an individual score. It can consequently reduce sensitivity to differences in score scale.

However, rank transformation also discards magnitude information. HUE-OOD therefore treats rank fusion as an empirical design choice whose utility must be established experimentally rather than as a universally superior aggregation principle.

Rank aggregation has been studied extensively in information retrieval, ensemble methods, and preference aggregation [29, 30]. Its use in OOD detection is less standardized than weighted score fusion, motivating an explicit comparison against alternative aggregation strategies in our experiments.

2.4 Multi-signal OOD Detection

Several existing OOD detectors already combine information from more than one network quantity. For example, VIM incorporates both feature norm and residual information [6], while other methods combine logits with feature statistics, activation transformations, or nearest-neighbor relationships. These methods demonstrate the practical value of exploiting multiple aspects of the learned representation.

The distinction of HUE-OOD is the particular decomposition used for fusion. Rather than combining multiple statistics from a single representation space, the framework explicitly constructs three branches: stochastic predictive variation, feature-space conformity, and a post-hoc evidential quantity derived from the classifier logits. The objective is to examine whether these signals provide complementary information under a common rank-based aggregation procedure.

We deliberately avoid claiming that the three branches are statistically independent. Complementarity in this paper is an empirical notion: a branch is considered useful if it provides additional discriminatory information when added to the other branches, as assessed through correlation analysis, controlled ablations, and performance comparisons. This definition avoids conflating complementary information with formal probabilistic independence.

2.5 Positioning of HUE-OOD

HUE-OOD is positioned as a lightweight post-hoc framework rather than a new training paradigm. Its design combines three properties that are usually evaluated separately: stochastic predictive variability, representational distance, and post-hoc evidential information. The fusion stage is deliberately non-parametric so that the detector does not depend on learned fusion weights.

Table 1 summarizes the conceptual positioning of the method relative to representative categories of OOD detection.

The positioning in Table 1 should not be interpreted as a strict taxonomy. In particular, several existing methods exploit multiple intermediate quantities even when their final detector is represented by a single scalar. The purpose of the table is to identify the primary sources explicitly combined by HUE-OOD.

Overall, the related literature suggests three relevant observations. First, post-hoc OOD detection can benefit from information beyond the maximum softmax probability. Second, uncertainty estimation provides multiple operational

Table 1. Conceptual positioning of representative post-hoc OOD detection approaches. The table describes the primary information source rather than claiming that individual methods use exclusively one signal.

Method family	Logits	Features	Stochastic predictions
MSP	✓	–	–
Energy	✓	–	–
ODIN	✓	–	–
Mahalanobis	–	✓	–
VIM	✓	✓	–
ReAct	✓	✓	–
GradNorm	✓	✓	–
MC-dropout approaches	✓	–	✓
HUE-OOD	✓	✓	✓

views of model behavior, including stochastic prediction variation and evidential quantities. Third, combining heterogeneous scores requires an aggregation rule that does not introduce unnecessary dependence on score scale. HUE-OOD builds on these observations by defining an explicit three-branch detector and evaluating rank-based fusion under a common experimental protocol.

3 Problem Formulation and Method Overview

Let f_θ denote a pretrained classifier with parameters θ , and let $\mathbf{x}$ be a test input. The model produces a feature representation $\mathbf{z} = f_\theta^{\text{feat}}(\mathbf{x})$ and a vector of logits $\ell = f_\theta^{\text{logit}}(\mathbf{x})$ for C classes.

The OOD detection problem considered in this study is to construct a scalar score $S(\mathbf{x})$ for which ID and OOD samples exhibit separable score distributions. No OOD samples are assumed to be available during the construction of the detector. The ID training data are used only where required to estimate feature statistics for the representational branch.

HUE-OOD computes three quantities,

$$U_{\text{pred}}(\mathbf{x}), \quad U_{\text{repr}}(\mathbf{x}), \quad U_{\text{evid}}(\mathbf{x}), \quad (7)$$

corresponding respectively to predictive stochastic variability, representational distance, and post-hoc evidential uncertainty.

After aligning their score directions so that larger values correspond to greater OOD evidence, each quantity is transformed into a normalized rank. The final detector score is

$$S_{\text{HUE}}(\mathbf{x}) = \frac{1}{3} [R_{\text{pred}}(\mathbf{x}) + R_{\text{repr}}(\mathbf{x}) + R_{\text{evid}}(\mathbf{x})]. \quad (8)$$

The detector therefore contains no learned fusion parameter. Its behavior is determined by the three branch definitions and the ranking operation. The

detailed formulation, implementation choices, and computational analysis are presented in Section 4.

4 Mathematical Notation and Methodology

This section formalizes HUE-OOD as a post-hoc OOD detection framework. The method operates on a pretrained classifier and does not update its learned parameters. It consists of three scoring branches that characterize different observable aspects of model behavior: stochastic predictive variability, feature-space conformity, and a post-hoc evidential quantity derived from the classifier logits. These scores are subsequently combined using a rank-based aggregation rule.

Throughout this section, all quantities are defined for a fixed pretrained model and a fixed inference protocol. Unless otherwise stated, statistics estimated from ID data are computed without using OOD samples.

4.1 Preliminaries and Notation

Let

$$\mathcal{D}_{\text{ID}} = \{(\mathbf{x}_i, y_i)\}_{i=1}^N, \quad y_i \in \{1, \dots, K\}, \quad (9)$$

denote the in-distribution training set containing N samples and K semantic classes.

Let f_θ denote a pretrained classifier with fixed parameters θ . We decompose the network into a feature extractor and a classification head,

$$f_\theta = g_{\theta_g} \circ h_{\theta_h}, \quad (10)$$

where

$$\mathbf{h} = h_{\theta_h}(\mathbf{x}) \in \mathbb{R}^D \quad (11)$$

is the penultimate representation and

$$\boldsymbol{\ell} = g_{\theta_g}(\mathbf{h}) \in \mathbb{R}^K \quad (12)$$

is the vector of class logits. For a linear classification head,

$$\boldsymbol{\ell} = \mathbf{W}\mathbf{h} + \mathbf{b}, \quad \mathbf{W} \in \mathbb{R}^{K \times D}, \quad \mathbf{b} \in \mathbb{R}^K. \quad (13)$$

The corresponding softmax distribution is

$$p_k(\mathbf{x}) = \frac{\exp(\ell_k)}{\sum_{j=1}^K \exp(\ell_j)}, \quad k = 1, \dots, K. \quad (14)$$

The predicted class is

$$\hat{y}(\mathbf{x}) = \arg \max_k p_k(\mathbf{x}). \quad (15)$$

HUE-OOD defines three scalar quantities,

$$U_{\text{pred}}(\mathbf{x}), \quad U_{\text{repr}}(\mathbf{x}), \quad U_{\text{evid}}(\mathbf{x}), \quad (16)$$

where larger values are defined to indicate stronger OOD evidence.

The three branches are intentionally distinguished by the information they use:

- U_{pred} measures variation in predictive distributions under stochastic perturbations of the network at inference time;
- U_{repr} measures the distance between the test representation and the estimated ID class-conditional feature distributions;
- U_{evid} is a post-hoc evidential score obtained from the classifier logits through a non-negative transformation.

The distinction should not be interpreted as implying statistical independence among the three quantities. They are generated by the same pretrained model and may therefore exhibit non-zero correlations. In this study, their complementarity is an empirical property assessed through correlation and ablation analyses.

4.2 Predictive Uncertainty via Stochastic Forward Passes

Stochastic Prediction Model. The predictive branch estimates the sensitivity of the model’s output distribution to stochastic perturbations induced by dropout. Let ξ denote a random dropout configuration. A stochastic forward pass can then be represented as

$$\mathbf{p}(\mathbf{x}; \xi) = \text{softmax}(f_{\theta}(\mathbf{x}; \xi)). \quad (17)$$

For a particular dropout strategy s , we generate T stochastic predictions,

$$\mathcal{P}_s(\mathbf{x}) = \left\{ \mathbf{p}_s^{(t)}(\mathbf{x}) \right\}_{t=1}^T. \quad (18)$$

We consider three stochastic perturbation mechanisms:

1. standard Bernoulli dropout, which applies independent activation-level masks;
2. spatial dropout, which applies structured masks to feature maps where supported by the architecture;
3. concrete dropout, implemented as a continuous stochastic relaxation of Bernoulli masking.

The three mechanisms are used as alternative perturbation operators rather than as three independently trained Bayesian models. In particular, no dropout probability is optimized during the post-hoc procedure. The probabilities used for inference are fixed according to the experimental configuration described in Section 4.

This distinction is important for the concrete-dropout branch. Concrete dropout was originally introduced as a differentiable formulation that can learn dropout probabilities during training. In HUE-OOD, the continuous relaxation is instead used only as an inference-time stochastic perturbation. Consequently, we do not interpret this branch as reproducing the posterior approximation obtained by a separately trained concrete-dropout model.

Predictive Variability. For each strategy s , the empirical mean prediction is

$$\bar{\mathbf{p}}_s(\mathbf{x}) = \frac{1}{T} \sum_{t=1}^T \mathbf{p}_s^{(t)}(\mathbf{x}). \quad (19)$$

We quantify stochastic variability using the mean of the per-class empirical variances,

$$u_s(\mathbf{x}) = \frac{1}{K} \sum_{k=1}^K \text{Var}_t \left[p_{s,k}^{(t)}(\mathbf{x}) \right], \quad (20)$$

where

$$\text{Var}_t \left[p_{s,k}^{(t)}(\mathbf{x}) \right] = \frac{1}{T-1} \sum_{t=1}^T \left(p_{s,k}^{(t)}(\mathbf{x}) - \bar{p}_{s,k}(\mathbf{x}) \right)^2. \quad (21)$$

The quantity in Eq. 20 is bounded because the probabilities lie in $[0, 1]$. It measures prediction disagreement under the selected stochastic perturbation mechanism. It should therefore be interpreted as an operational measure of stochastic predictive variability rather than as a complete or uniquely defined measure of epistemic uncertainty.

ID-Based Scale Normalization. The three stochastic mechanisms can produce uncertainty values with different empirical scales. To reduce this effect, we optionally normalize each branch using an ID reference statistic. Let

$$\mu_s^{\text{ID}} = \frac{1}{|\mathcal{D}_{\text{ref}}|} \sum_{\mathbf{x} \in \mathcal{D}_{\text{ref}}} u_s(\mathbf{x}), \quad (22)$$

where $\mathcal{D}_{\text{ref}} \subseteq \mathcal{D}_{\text{ID}}$ is an ID reference subset.

The normalized branch score is

$$\tilde{u}_s(\mathbf{x}) = \frac{u_s(\mathbf{x})}{\mu_s^{\text{ID}} + \epsilon}, \quad (23)$$

where $\epsilon > 0$ is a small numerical constant.

The resulting predictive uncertainty score is

$$U_{\text{pred}}(\mathbf{x}) = \frac{1}{3} \sum_{s=1}^3 \tilde{u}_s(\mathbf{x}). \quad (24)$$

The ID normalization is not required for the subsequent rank transformation in a strict mathematical sense, because strictly positive multiplicative rescaling does not change ranks within an individual branch. It is retained here because it provides a common intermediate representation and makes the branch statistics easier to inspect and compare. An ablation without this normalization is therefore included in Section 4.

In the default experimental configuration, $T = 10$ stochastic forward passes are used per strategy. The sensitivity of the predictive branch to the number of stochastic samples is evaluated separately.

4.3 Representational Uncertainty via Mahalanobis Distance

The representational branch evaluates the conformity of a test feature to the class-conditional feature distributions estimated from ID data.

For class k , let

$$\mathcal{I}_k = \{i : y_i = k\}, \quad N_k = |\mathcal{I}_k|. \quad (25)$$

The class centroid is estimated as

$$\boldsymbol{\mu}_k = \frac{1}{N_k} \sum_{i \in \mathcal{I}_k} \mathbf{h}_i, \quad \mathbf{h}_i = h_{\theta_h}(\mathbf{x}_i). \quad (26)$$

We use a pooled within-class covariance estimate,

$$\boldsymbol{\Sigma}_w = \frac{1}{N - K} \sum_{k=1}^K \sum_{i \in \mathcal{I}_k} (\mathbf{h}_i - \boldsymbol{\mu}_k)(\mathbf{h}_i - \boldsymbol{\mu}_k)^\top. \quad (27)$$

The pooled covariance is regularized as

$$\boldsymbol{\Sigma}_\lambda = \boldsymbol{\Sigma}_w + \lambda \mathbf{I}, \quad \lambda > 0, \quad (28)$$

where λ controls numerical stabilization. The regularization parameter is fixed before evaluation and is not selected using OOD samples.

For a test representation $\mathbf{h}$, the class-conditional Mahalanobis distance is

$$M_k(\mathbf{x}) = (\mathbf{h} - \boldsymbol{\mu}_k)^\top \boldsymbol{\Sigma}_\lambda^{-1} (\mathbf{h} - \boldsymbol{\mu}_k). \quad (29)$$

The representational score is defined as the distance to the closest estimated class distribution,

$$U_{\text{repr}}(\mathbf{x}) = \min_{k \in \{1, \dots, K\}} M_k(\mathbf{x}). \quad (30)$$

Large values indicate that the representation is relatively far from all estimated class centroids under the covariance metric. This interpretation is conditional on the adequacy of the feature statistics and the shared covariance approximation.

The minimum over classes is used because an ID sample should be close to at least one class-conditional distribution. Conversely, an input that is far from every estimated class provides stronger geometric evidence of a distributional mismatch.

The Mahalanobis branch should not be interpreted as estimating a calibrated probability of OOD membership. It is a distance-based conformity score. Its contribution to the final detector is assessed empirically rather than assumed from the Gaussian modeling assumption.

4.4 Post-Hoc Evidential Score from Classifier Logits

The third branch derives an evidential score from the logits of the pretrained classifier. The construction is motivated by evidential classification, where class-specific non-negative evidence determines the concentration parameters of a Dirichlet distribution [14].

For a vector of positive concentration parameters $\boldsymbol{\alpha}$, the Dirichlet density is

$$\text{Dir}(\mathbf{p} \mid \boldsymbol{\alpha}) = \frac{\Gamma(\alpha_0)}{\prod_{k=1}^K \Gamma(\alpha_k)} \prod_{k=1}^K p_k^{\alpha_k - 1}, \quad (31)$$

where

$$\alpha_0 = \sum_{k=1}^K \alpha_k. \quad (32)$$

In an explicitly trained evidential model, the concentration parameters may be expressed as

$$\alpha_k = e_k + 1, \quad e_k \geq 0. \quad (33)$$

HUE-OOD does not train the classifier to produce evidential parameters. Instead, it constructs a post-hoc evidence-like quantity from the existing logits. Specifically,

$$e_k = \text{softplus}(\ell_k) = \log(1 + \exp(\ell_k)). \quad (34)$$

The associated concentration-like quantity is therefore

$$\alpha_k^{\text{post}} = e_k + 1, \quad (35)$$

and the corresponding total concentration is

$$A(\mathbf{x}) = \sum_{k=1}^K \alpha_k^{\text{post}} = K + \sum_{k=1}^K \text{softplus}(\ell_k). \quad (36)$$

We define the evidential OOD score as

$$U_{\text{evid}}(\mathbf{x}) = \frac{K}{A(\mathbf{x})} = \frac{K}{K + \sum_{k=1}^K \text{softplus}(\ell_k)}. \quad (37)$$

Larger values correspond to lower aggregate evidence under this particular post-hoc construction.

Interpretation and Scope of the Evidential Branch. Equation 37 requires careful interpretation. The transformation from logits to non-negative evidence does not establish that the resulting Dirichlet distribution is the posterior predictive distribution of the original classifier. In particular, the classifier was not necessarily optimized under an evidential objective, and the logit-to-evidence mapping introduces assumptions that are not implied by the standard softmax model.

Accordingly, we refer to U_{evid} as a *post-hoc evidential score* rather than a calibrated epistemic posterior. The term “epistemic” is used operationally to indicate that the score is intended to characterize uncertainty associated with the strength of model output evidence.

An additional consideration is that softmax probabilities are invariant to adding the same constant to every logit, whereas Eq. 34 is not. Thus, the evidential score contains information about the absolute logit scale that is discarded by the softmax. This property is deliberate, because the branch is intended to exploit logit magnitude, but it also makes the score sensitive to the logit parameterization of the pretrained classifier. Cross-architecture experiments are therefore necessary to determine whether this signal transfers reliably.

The evidential branch requires only the deterministic logits from one forward pass and introduces no additional model parameters or training steps.

4.5 Score Orientation and Rank Transformation

The three branch scores differ in numerical scale and statistical distribution. Moreover, the Mahalanobis score is naturally an OOD-oriented distance, whereas the predictive and evidential quantities are uncertainty measures whose orientation must be explicitly defined.

Before fusion, all scores are transformed such that larger values correspond to stronger OOD evidence. Let

$$V_m(\mathbf{x}), \quad m \in \{\text{pred}, \text{repr}, \text{evid}\}, \quad (38)$$

denote the resulting OOD-oriented branch scores.

For each branch, scores are ranked over a reference population. Let $\mathcal{R}$ denote the fixed ID reference set used for the rank transformation. For a test batch $\mathcal{B}$, the ranking population is

$$\mathcal{P} = \mathcal{R} \cup \mathcal{B}. \quad (39)$$

For branch m , let $r_m(\mathbf{x})$ denote the rank of $\mathbf{x}$ in $\mathcal{P}$, with the largest OOD-oriented score assigned the largest rank. The normalized rank is

$$R_m(\mathbf{x}) = \frac{r_m(\mathbf{x}) - 1}{|\mathcal{P}| - 1}. \quad (40)$$

Hence,

$$0 \leq R_m(\mathbf{x}) \leq 1. \quad (41)$$

Ties are resolved using a deterministic average-rank convention. The reference set is fixed before evaluating the corresponding test protocol.

This formulation makes the fusion invariant to strictly monotonic transformations of each individual branch score within a ranking population. It therefore reduces dependence on arbitrary numerical units. However, the transformation also discards absolute score magnitude. Rank fusion should consequently be viewed as a robustness-oriented aggregation strategy rather than as a universally optimal estimator.

4.6 Parameter-Free Rank-Based Fusion

Let the three normalized ranks be

$$R_{\text{pred}}(\mathbf{x}), \quad R_{\text{repr}}(\mathbf{x}), \quad R_{\text{evid}}(\mathbf{x}). \quad (42)$$

The HUE-OOD score is defined as their arithmetic mean,

$$S_{\text{HUE}}(\mathbf{x}) = \frac{1}{3} [R_{\text{pred}}(\mathbf{x}) + R_{\text{repr}}(\mathbf{x}) + R_{\text{evid}}(\mathbf{x})]. \quad (43)$$

By construction,

$$0 \leq S_{\text{HUE}}(\mathbf{x}) \leq 1. \quad (44)$$

The method contains no learned fusion coefficients. Each branch contributes equally after rank transformation.

The choice of equal weighting is deliberate for two reasons. First, it avoids estimating branch-specific coefficients from OOD validation data. Second, it provides a reproducible baseline for evaluating whether the three information sources collectively provide useful discrimination. Nevertheless, equal weighting is not assumed to be optimal. Weighted rank fusion and other aggregation strategies are therefore evaluated as comparative alternatives.

Properties of the Rank Aggregation. The proposed aggregation has several useful properties.

Monotonic transformation invariance. If $q_m(\cdot)$ is a strictly increasing function, then

$$R_m(\mathbf{x}; V_m) = R_m(\mathbf{x}; q_m(V_m)), \quad (45)$$

provided the ranking population is unchanged. Thus, the fusion does not depend on the numerical units of an individual branch.

Bounded output. Because every branch rank lies in $[0, 1]$, the final score also lies in $[0, 1]$. This provides a convenient common representation, although it does not imply that the score is a calibrated OOD probability.

No learned fusion parameters. The fusion rule requires no coefficient estimation and therefore does not introduce a meta-classifier. This property simplifies cross-dataset and cross-backbone evaluation.

Information loss. Rank transformation removes absolute magnitude information. Consequently, two samples that differ substantially in raw uncertainty may become neighbors in rank space if their ordering is similar. This is an inherent trade-off of the proposed approach and motivates comparison with magnitude-preserving alternatives.

Dependence between branches. The three ranks are not assumed to be independent. Therefore, the average in Eq. 43 should not be interpreted as a formally optimal variance-reducing estimator. Any improvement arising from the combination must be established empirically through ablation and correlation analyses.

4.7 Reference-Set Construction and Thresholding

Rank-based scoring requires a reference population. We use an ID-only reference set

$$\mathcal{R} = \left\{ \mathbf{x}_j^{(r)} \right\}_{j=1}^{N_{\text{ref}}}, \quad \mathcal{R} \subset \mathcal{D}_{\text{ID}}, \quad (46)$$

with $N_{\text{ref}} = 1000$ in the default configuration. The reference samples are selected before evaluation using a fixed random seed.

The reference set is not an OOD validation set. Its role is to establish an ID score distribution against which test samples can be ranked. The reference scores are computed once and retained throughout evaluation.

Because test samples are appended to the ranking population in the formulation of Eq. 39, the exact normalized rank of a test sample can depend on the composition of the evaluated batch. This dependence is an intrinsic consequence of transductive ranking and must be considered when interpreting absolute score values. For this reason, AUROC and AUPR are used as the primary threshold-independent evaluation metrics, while fixed-threshold deployment experiments use a separate calibration protocol described below.

For threshold-based detection, let $\mathcal{D}_{\text{cal}}$ denote an ID-only calibration subset that is not used to fit the backbone or the Mahalanobis statistics. A threshold τ is selected according to a predefined operating criterion,

$$\hat{y}_{\text{OOD}}(\mathbf{x}) = \mathbb{I}[S_{\text{HUE}}(\mathbf{x}) > \tau]. \quad (47)$$

For example, τ may be selected to achieve a target ID false-positive rate. The threshold-selection procedure is kept separate from the OOD test data to prevent evaluation leakage.

Importantly, the bounded range of S_{HUE} does not imply that the score has a probabilistic interpretation. A threshold such as $\tau = 0.8$ means that the sample occupies a relatively high position under the specified ranking protocol; it does not mean that the sample has an 80% probability of being OOD.

4.8 Complete HUE-OOD Algorithm

The complete procedure can be summarized as follows.

1. Train or obtain a pretrained classifier f_θ using only the ID training protocol.
2. Extract penultimate features from the ID training set and estimate the class centroids $\{\boldsymbol{\mu}_k\}_{k=1}^K$ and the regularized pooled covariance $\boldsymbol{\Sigma}_\lambda$.
3. Construct the ID reference set $\mathcal{R}$ and compute the reference distributions required for the predictive, representational, and evidential branches.
4. For each test sample, perform T stochastic forward passes for each selected dropout strategy and compute U_{pred} .
5. Perform a deterministic forward pass and compute the feature vector $\mathbf{h}$. Evaluate Eq. 30 to obtain U_{repr} .
6. Use the deterministic logits $\boldsymbol{\ell}$ to compute the post-hoc evidential score U_{evid} according to Eq. 37.
7. Orient the three scores consistently so that larger values represent stronger OOD evidence.
8. Convert the three scores into normalized ranks relative to the specified reference population.
9. Compute the final HUE-OOD score using Eq. 43.
10. If a binary OOD decision is required, apply a threshold determined from the independent ID calibration protocol.

4.9 Overview of the HUE-OOD Inference Procedure

The HUE-OOD inference procedure is designed as a post-hoc scoring pipeline operating on a fixed pretrained classifier. Given an input $\mathbf{x}$, the classifier provides three types of information without modifying its learned parameters: stochastic predictive outputs obtained from repeated dropout-enabled forward passes, a penultimate-layer representation $\mathbf{h} = h(\mathbf{x})$, and a deterministic logit vector $\mathbf{l} = g(\mathbf{h})$. These quantities are used to construct three scores that characterize different aspects of the model response.

First, the predictive branch estimates stochastic prediction variability from multiple forward passes under the specified dropout mechanisms. This branch therefore measures instability of the predictive output under the chosen stochastic perturbations. Second, the representational branch computes the Mahalanobis distance between the test representation and the closest class-conditional centroid estimated from ID training features. This score quantifies the degree to which the representation conforms to the feature-space statistics observed during training. Third, the evidential branch transforms the deterministic logits into non-negative evidence using the softplus function and derives an inverse-total-evidence score. Importantly, this quantity is treated as a post-hoc evidential surrogate rather than as a calibrated posterior epistemic uncertainty estimate, since the underlying classifier was not trained with an evidential objective.

Because the resulting branch scores generally have different numerical scales and orientations, each score is first transformed so that larger values consistently correspond to stronger OOD evidence. The oriented scores are subsequently converted into normalized ranks with respect to a fixed ID reference set. The rank transformation discards the original score magnitudes and retains their relative ordering. The three normalized ranks are then averaged without learned coefficients to obtain the final HUE-OOD score. Consequently, the fusion stage does not require an auxiliary OOD validation set or a learned meta-classifier.

For binary OOD decision making, the resulting score can be thresholded using an ID-only validation protocol. The threshold-selection procedure should be specified independently of the OOD test sets to avoid evaluation leakage. Performance metrics such as AUROC and AUPR are computed directly from the continuous HUE-OOD score and therefore do not require a manually selected operating threshold. The complete procedure consists of precomputed ID statistics, stochastic inference for the predictive branch, single-pass feature and logit extraction, score orientation, rank transformation, and rank aggregation. No additional model training or architectural modification is required.

4.10 Computational Complexity

Let C_f denote the computational cost of one forward pass through the pretrained classifier, and let C_M denote the cost of evaluating the regularized Mahalanobis distances for one feature vector.

The predictive branch requires approximately

$$C_{\text{pred}} = 3TC_f, \quad (48)$$

when all three stochastic strategies use T forward passes.

The representational branch requires one deterministic feature extraction and evaluation against K class centroids. If the inverse covariance matrix is precomputed, the distance calculation requires approximately

$$C_{\text{repr}} = O(KD^2) \quad (49)$$

under a dense matrix representation. In practice, the exact cost depends on whether the quadratic form is implemented using a Cholesky factor, precomputed inverse, or another numerically stable linear-algebra routine.

The evidential branch requires only the logits from a deterministic forward pass and therefore has computational cost dominated by the classifier itself,

$$C_{\text{evid}} = C_f + O(K). \quad (50)$$

The final rank aggregation over B test samples and a reference set of size N_{ref} can be implemented using standard sorting procedures. Its complexity is approximately

$$C_{\text{rank}} = O(3(B + N_{\text{ref}}) \log(B + N_{\text{ref}})). \quad (51)$$

Thus, the dominant computational component is expected to be stochastic forward inference. The exact wall-clock overhead is architecture- and hardware-dependent and is therefore reported empirically rather than treated as a hardware-independent property of HUE-OOD.

4.11 Implementation Considerations

Several implementation choices are important for reproducibility.

First, dropout layers are activated during stochastic inference, whereas batch-normalization layers, when present, remain in evaluation mode so that their accumulated running statistics are not updated by test samples. This separation prevents stochastic inference from simultaneously changing the normalization statistics.

Second, the feature extractor used for Mahalanobis modeling must be identical between ID-statistics estimation and test-time inference. Changing the feature layer, normalization operation, or preprocessing pipeline changes the geometry on which the Mahalanobis model is defined.

Third, the covariance regularization parameter λ , the number of Monte Carlo samples T , and the reference-set size N_{ref} are treated as explicit experimental parameters. They are not implicitly optimized using OOD test samples.

Fourth, all preprocessing operations applied to ID data must also be applied to OOD test data under the corresponding benchmark protocol. Otherwise, the detector may exploit preprocessing differences rather than distributional characteristics learned by the classifier.

Finally, the reported computational overhead must be measured using a specified hardware configuration, batch size, precision, and implementation backend. A statement such as “less than 5 ms” is therefore treated as an empirical measurement for the corresponding experimental environment, rather than as a universal complexity guarantee.

4.12 Methodological Assumptions and Scope

HUE-OOD relies on several assumptions that delimit the interpretation of the method.

First, the representational branch assumes that class-conditional feature statistics estimated from ID data provide a useful approximation of the local geometry relevant to OOD detection. The shared covariance model may be restrictive for highly multimodal or strongly non-Gaussian feature distributions.

Second, the predictive branch assumes that stochastic dropout perturbations provide informative variation in model predictions. This assumption can depend on whether dropout layers are present in the pretrained architecture and on how their inference-time behavior is implemented.

Third, the evidential branch is explicitly heuristic. The softplus mapping does not convert a conventional classifier into a fully trained evidential network. Its usefulness must therefore be established empirically rather than inferred solely from the Dirichlet interpretation.

Fourth, rank fusion preserves relative ordering but discards score magnitude. It can therefore be beneficial when heterogeneous score scales are problematic, but it may be less effective when absolute uncertainty magnitude contains important discriminative information.

These assumptions motivate the experimental design in Section 4. In particular, we evaluate individual branches, pairwise combinations, the complete three-branch model, alternative fusion rules, sensitivity to T , sensitivity to reference-set size, and cross-backbone behavior. This design allows improvements to be attributed to specific components rather than to the existence of additional computation alone.

5 Experimental Setup and Results

This section describes the experimental protocol used to evaluate HUE-OOD and to examine the sensitivity of its principal design choices. We first specify the datasets, backbone models, evaluation metrics, implementation details, and procedures used for reproducibility. We then study the sensitivity of the number of stochastic forward passes, the size of the ID reference set used for rank transformation, the post-hoc evidence mapping, and the availability of dropout in the pretrained backbone. These analyses are performed before the broader comparisons reported in the subsequent sections. Unless explicitly stated otherwise, all hyperparameters are fixed before evaluation on the corresponding OOD test sets.

5.1 Experimental Configuration and Hyperparameter Sensitivity

The sensitivity experiments use CIFAR-10 as the ID dataset and SVHN as the OOD dataset, with a ViT-Small backbone. This setting provides a controlled environment for examining the principal HUE-OOD design choices while avoiding

the confounding effects of simultaneously varying datasets and architectures. The selected configuration is subsequently used as the default setting for the larger benchmark.

Datasets, Evaluation Metrics, and Backbone Pool The principal ID datasets considered in this study are CIFAR-10 and CIFAR-100. The OOD evaluation suite contains datasets selected to represent different degrees and types of distributional discrepancy. These datasets are grouped into five broad categories: far-OOD, near-OOD, texture, natural-image, and synthetic datasets. The far-OOD group contains SVHN, USPS, Fashion-MNIST, KMNIST, and EMNIST. The near-OOD group contains CIFAR-100 when CIFAR-10 is used as the ID dataset, together with Tiny-ImageNet. The texture group contains DTD. The natural-image group contains Places365, LSUN-Crop, LSUN-Resize, iSUN, SUN397, Food-101, Flowers-102, GTSRB, CUB-200, ImageNet-100, and PatchCamelyon. The synthetic group contains STL-10. The exact preprocessing, image resizing, class selection, and sample counts are kept identical across competing post-hoc methods for a given experimental setting.

The backbone pool comprises 20 pretrained architectures spanning several model families. These include vision-transformer models such as ViT-Tiny, ViT-Small, ViT-Base, ViT-Large, DeiT, Swin, CrossViT, CaiT, PiT, Twins, and PoolFormer; convolutional models such as ResNet-50 and EfficientNet-B0; and hybrid or modern convolutional architectures such as ConvNeXt-Tiny, ConvNeXt-Small, ConvNeXt-Base, and CoAtNet-0. Models are pretrained on ImageNet-21K and subsequently fine-tuned on the corresponding ID dataset where applicable. When a backbone is obtained from a public implementation with a different pretraining or fine-tuning protocol, the original training configuration is retained and explicitly documented in the corresponding experimental table.

OOD detection performance is evaluated using AUROC, AUPR, and FPR95. AUROC measures ranking quality independently of a particular operating threshold, whereas AUPR is sensitive to the relative prevalence of the positive class and is therefore reported together with AUROC. FPR95 denotes the false-positive rate when the true-positive rate is fixed at 95%. Unless otherwise specified, numerical results are reported as mean values with 95% confidence intervals obtained from five independent runs. The confidence intervals quantify variation across runs and should not be interpreted as confidence intervals over the population of possible datasets or architectures.

The default HUE-OOD configuration is $T = 10$ stochastic forward passes for each available dropout strategy, an ID reference set of $N_{\text{ref}} = 1000$ samples for rank transformation, and the softplus mapping for the post-hoc evidence branch. These values are selected from the sensitivity experiments reported below and are then kept fixed for the main comparisons unless a method-specific implementation requires an explicit exception.

Implementation Details

Predictive branch. For architectures containing compatible dropout modules, dropout is activated during inference to generate stochastic predictions. Standard Bernoulli dropout, spatial dropout, and the concrete-dropout variant are considered when their corresponding modules are available. The predictive uncertainty for each strategy is estimated from repeated stochastic forward passes and combined according to Eq. (1). The normalization statistics required by Eq. (1) are estimated exclusively from ID training data. Batch-normalization layers, when present, remain in evaluation mode so that their accumulated running statistics are not altered during stochastic inference.

For a backbone that does not contain an appropriate dropout mechanism, the predictive branch is not artificially introduced by modifying the architecture. Instead, HUE-OOD is evaluated using the available representational and evidential branches only. This distinction is important because introducing new dropout layers after training would no longer constitute a strictly architecture-preserving post-hoc procedure.

Representational branch. The class-specific feature centroids and shared covariance matrix are estimated from ID training representations. A diagonal or full covariance implementation is selected according to the dimensionality and numerical conditioning of the feature space; in the experiments reported here, a shared covariance with Tikhonov regularization is used. Specifically, the regularized covariance is

$$\hat{\Sigma}_\lambda = \hat{\Sigma} + \lambda \mathbf{I}, \quad \lambda = 10^{-3}, \quad (52)$$

where $\mathbf{I}$ is the identity matrix. The representational score is the minimum class-conditional Mahalanobis distance defined in Eq. (2). The same ID-derived statistics are retained for all OOD samples associated with a particular ID/backbone configuration.

Evidential branch. The deterministic classifier logits are transformed using

$$e_k = \text{softplus}(l_k) = \log(1 + \exp(l_k)), \quad (53)$$

with the corresponding total concentration surrogate

$$S(\mathbf{x}) = K + \sum_{k=1}^K e_k. \quad (54)$$

The resulting score is

$$U_{\text{epist}}(\mathbf{x}) = \frac{K}{S(\mathbf{x})}. \quad (55)$$

Throughout the experiments, this quantity is referred to as the post-hoc evidential score. It should not be interpreted as a calibrated epistemic posterior or as evidence learned by an evidential objective, because the underlying classifier was trained using its original classification objective. Its role in HUE-OOD is to provide a logit-derived uncertainty surrogate that can be evaluated without additional model training.

Rank fusion. The fusion stage uses a fixed ID reference set containing $N_{\text{ref}} = 1000$ samples. The reference samples are drawn once from the available ID training data using a fixed random seed and are kept unchanged throughout the corresponding evaluation. For every test batch, the reference scores and test scores are jointly ranked separately for each available uncertainty branch. The orientation of each score is defined before ranking such that larger normalized ranks correspond to stronger OOD evidence. The resulting normalized ranks are averaged according to Eq. (5).

When all three branches are available, the fusion contains three ranks. When the predictive branch is unavailable, only the representational and post-hoc evidential ranks are averaged. No branch-specific fusion weights are learned from OOD validation data.

Computational environment. Experiments are conducted on an NVIDIA RTX 3090 GPU using PyTorch 2.0 and CUDA 11.8. Runtime measurements are performed using a batch size of 64, with 100 warm-up batches followed by 1000 measured forward passes. GPU synchronization is applied when recording elapsed execution time. The reported runtime therefore measures the specified implementation under the stated hardware and software configuration rather than providing a hardware-independent estimate.

For the ViT-Small sensitivity setting, the measured deterministic forward pass requires approximately 30.5 ms per batch under the stated configuration. The complete three-branch HUE-OOD implementation with $T = 10$ requires approximately 36.1 ms per batch, corresponding to an absolute increase of approximately 5.6 ms in this particular setting. For backbones evaluated without the predictive branch, the additional computation associated with the two remaining branches and rank fusion is substantially smaller. These measurements are implementation-specific and are therefore reported as empirical runtime observations rather than as a general guarantee of deployment latency.

Sensitivity to the Number of Monte Carlo Passes The number of stochastic forward passes T controls a direct accuracy–computation trade-off in the predictive branch. We evaluate $T \in \{2, 5, 10, 15, 20, 30\}$ on CIFAR-10 versus SVHN using ViT-Small. For each configuration, we record AUROC, the empirical standard deviation of the per-strategy variance estimate, and the measured batch execution time.

Table 2 summarizes the results. AUROC increases from 0.962 at $T = 2$ to 0.985 at $T = 10$, while the subsequent improvement is small in this particular experiment. At the same time, execution time increases with T . The standard deviation associated with the variance estimate also decreases as additional stochastic samples are introduced. These observations motivate $T = 10$ as a practical operating point for the remaining experiments. This choice is empirical and is not intended to imply that $T = 10$ is optimal for every architecture or dataset.

Table 2. Sensitivity to the number of stochastic forward passes T on CIFAR-10 (ID) versus SVHN (OOD) using ViT-Small. AUROC is reported as mean $\pm$ 95% confidence interval over five runs. The runtime values correspond to batch size 64 on an NVIDIA RTX 3090.

T	AUROC ($\uparrow$)	Std. of u_s	Time/batch (ms)
2	0.962 ± 0.005	0.021	33.2
5	0.979 ± 0.004	0.014	34.5
10	0.985 ± 0.003	0.009	36.1
15	0.986 ± 0.003	0.008	37.9
20	0.986 ± 0.003	0.007	39.5
30	0.987 ± 0.003	0.006	42.8

The difference between $T = 10$ and larger values is small relative to the additional computation in this experiment. We therefore use $T = 10$ as the default configuration in the subsequent experiments, while retaining the full sensitivity results to make the computational trade-off explicit.

Sensitivity to the ID Reference Set Size Rank transformation requires a reference distribution against which the test scores are ordered. An excessively small reference set may produce coarse ranks and make the resulting score sensitive to the particular reference samples. Conversely, increasing the reference set indefinitely increases preprocessing and ranking costs without necessarily producing a measurable improvement.

We therefore evaluate

$$N_{\text{ref}} \in \{100, 250, 500, 1000, 2000, 5000\} \quad (56)$$

on CIFAR-10 versus SVHN using ViT-Small. For each reference-set size, ten independent reference subsets are sampled from the ID training data. The reported standard deviation quantifies variation caused by the choice of reference subset.

Table 3. Sensitivity to the size of the ID reference set used for rank transformation on CIFAR-10 (ID) versus SVHN (OOD). Values are mean AUROC $\pm$ standard deviation over ten independently sampled reference sets.

N_{ref}	AUROC ($\uparrow$)	Std. across draws
100	0.975 ± 0.006	0.005
250	0.981 ± 0.005	0.003
500	0.984 ± 0.004	0.002
1000	0.985 ± 0.003	0.001
2000	0.986 ± 0.003	0.001
5000	0.986 ± 0.003	0.001

The results in Table 3 indicate that the performance difference becomes small beyond $N_{\text{ref}} = 1000$ in the evaluated setting. The variation across reference draws is also reduced relative to the smaller reference sets. Based on this empirical observation, $N_{\text{ref}} = 1000$ is retained as the default. This choice should be understood as a computationally convenient operating point rather than as a theoretically optimal reference-set size.

Comparison of Post-Hoc Evidence Transformations The evidential branch requires a mapping from unrestricted logits to non-negative quantities that can be interpreted as evidence parameters. We compare three mappings:

$$e_k^{\text{SP}} = \log(1 + \exp(l_k)), \quad (57)$$

$$e_k^{\text{EXP}} = \exp(l_k), \quad (58)$$

$$e_k^{\text{ReLU}} = \max(0, l_k). \quad (59)$$

The comparison is performed on CIFAR-10 versus SVHN using ViT-Small. For the exponential mapping, numerical stabilization is required because the raw exponential can grow rapidly for positive logits. We therefore apply the same explicitly specified upper bound in all exponential experiments. The purpose of this comparison is not to establish a universally optimal logit-to-evidence transformation, but to determine which mapping is most appropriate for the proposed implementation under the selected benchmark.

Table 4. Comparison of post-hoc logit-to-evidence transformations on CIFAR-10 (ID) versus SVHN (OOD). The first metric evaluates the evidential branch alone, while the second evaluates the complete HUE-ODD score. Values are mean AUROC $\pm$ 95% confidence interval over five runs.

Evidence mapping	Evidential branch	Full HUE-ODD
Softplus	0.912 ± 0.005	0.985 ± 0.003
Exponential (clipped)	0.904 ± 0.006	0.982 ± 0.004
ReLU	0.867 ± 0.008	0.973 ± 0.004

Softplus provides the highest AUROC among the three mappings in this experiment and avoids the explicit clipping required by the exponential alternative. The ReLU transformation produces a lower AUROC in both the standalone and fused settings. These observations support the use of softplus as the default mapping in the subsequent experiments. However, the results are obtained on one ID/OOD pair and one backbone and therefore should not be interpreted as evidence that softplus is universally superior for all classifiers.

The difference between the standalone evidential branch and the complete HUE-ODD score also indicates that the fused detector uses information not contained in the evidential branch alone. This observation is consistent with the intended role of the evidential score as one component of a multi-source

detector; it does not by itself establish statistical independence between the three branches.

Backbones Without Dropout The predictive branch depends on the availability of a stochastic dropout mechanism in the pretrained architecture. Some classifiers, including commonly used convolutional backbones, may not contain dropout layers or may expose dropout in a form that is not directly compatible with the proposed inference procedure. Rather than modifying such architectures after training, we omit the predictive branch and evaluate the remaining representational and post-hoc evidential branches.

Table 5 compares the three-branch and two-branch configurations. For ViT-Small, which contains compatible dropout modules, the three-branch configuration achieves an AUROC of 0.985 in the specified CIFAR-10 versus SVHN experiment, whereas the corresponding two-branch configuration achieves 0.971. For ResNet-50 without a compatible dropout mechanism, only the two-branch configuration is applicable and obtains an AUROC of 0.967.

Table 5. Effect of omitting the predictive branch when dropout is unavailable. Results are AUROC on CIFAR-10 (ID) versus SVHN (OOD), reported as mean $\pm$ 95% confidence interval over five runs.

Backbone	Three-branch HUE-OOD	Two-branch HUE-OOD
ViT-Small (dropout available)	0.985 ± 0.003	0.971 ± 0.004
ResNet-50 (dropout unavailable)	–	0.967 ± 0.005

For ViT-Small, the decrease from the three-branch to the two-branch configuration suggests that the predictive branch contributes useful information in this particular setting. The ResNet-50 result shows that the remaining two branches can still produce a meaningful OOD ranking when stochastic predictive information is unavailable. Nevertheless, the two configurations are not directly equivalent: the three-branch version uses an additional source of information. Consequently, the ResNet-50 result should be interpreted as evidence of applicability under a restricted information setting rather than as evidence that the omission of the predictive branch has no effect.

For the large-scale benchmark, the three-branch configuration is used when the pretrained architecture provides the required dropout mechanism. Otherwise, the two-branch configuration is reported explicitly. This protocol avoids introducing post-hoc architectural changes solely for the purpose of obtaining stochastic predictions.

Summary of Configuration Choices The sensitivity experiments provide empirical guidance for selecting the default HUE-OOD configuration. The selected values are $T = 10$ stochastic passes, $N_{\text{ref}} = 1000$ ID reference samples, and softplus for the logit-to-evidence transformation. These choices correspond to

operating points at which increasing the associated computational or reference-set budget produced relatively small changes in the evaluated CIFAR-10 versus SVHN setting.

Importantly, these sensitivity experiments are not intended to constitute a formal global hyperparameter optimization. They provide a transparent basis for fixing the principal implementation choices before the broader comparisons. In particular, the reported values should not be assumed to be optimal for every ID dataset, OOD type, or backbone architecture. The subsequent experiments therefore retain the selected configuration without additional OOD-specific tuning whenever possible.

The remainder of this section evaluates HUE-OOD against established post-hoc OOD detection methods across the broader dataset and architecture pool. In those comparisons, all methods are evaluated under the same dataset partitions and backbone configurations, and the evaluation protocol is kept separate from the sensitivity analysis above.

6 Main Results

6.1 Overall Detection Performance

We evaluate HUE-OOD on the complete benchmark described in Section 4.1. Unless otherwise specified, all experiments use the default configuration with $T = 10$ Monte Carlo passes, $N_{\text{ref}} = 1000$ reference samples, the softplus evidence transformation, and rank-based fusion. ViT-Small is used as the primary backbone for the main comparison. Results are reported over five independent runs, and the corresponding uncertainty intervals are reported together with the mean performance.

Table 6 summarizes the average OOD detection performance over the evaluated OOD datasets. We report AUROC, AUPR, and FPR95, where higher AUROC and AUPR and lower FPR95 indicate better detection performance.

HUE-OOD obtains an average AUROC of 0.961, compared with 0.897 for ReAct, the strongest baseline among the methods considered in this comparison. This corresponds to an absolute AUROC difference of 0.064. The corresponding FPR95 decreases from 33.8 to 16.1. These results indicate that the proposed fusion can provide a substantial improvement under the evaluated experimental configuration. However, the magnitude of the improvement varies across OOD datasets, as analyzed in the subsequent subsections. Therefore, the aggregate results should not be interpreted as evidence that HUE-OOD uniformly dominates every baseline on every OOD distribution.

Table 6. Average OOD detection performance over the evaluated OOD datasets using ViT-Small with CIFAR-10 as the in-distribution dataset. Higher AUROC and AUPR and lower FPR95 indicate better OOD detection performance.

Method	AUROC $\uparrow$	AUPR $\uparrow$	FPR95 $\downarrow$
MSP	0.845	0.880	45.6
MaxLogit	0.858	0.893	41.8
ODIN	0.862	0.897	40.2
Energy	0.885	0.913	36.8
Mahalanobis	0.877	0.906	38.9
RMDS	0.883	0.912	37.2
KNN	0.873	0.902	39.8
VIM	0.892	0.918	35.1
DICE	0.894	0.920	34.5
ASH	0.881	0.909	38.2
ReAct	0.897	0.923	33.8
SHE	0.889	0.915	35.9
NNGuide	0.891	0.917	35.5
GradNorm	0.865	0.895	42.1
OpenMax	0.835	0.873	47.2
ARPL	0.887	0.914	37.5
CSI	0.890	0.919	35.9
SSD	0.886	0.913	37.8
G-ODIN	0.879	0.908	39.1
LogitNorm	0.884	0.911	37.0
HUE-OOD	0.961	0.973	16.1

The improvement is consistent with the intended design of HUE-OOD. The predictive branch measures stochastic variation under dropout perturbations, the representational branch measures the position of a sample relative to the ID feature distribution, and the post-hoc evidential branch provides a logit-derived uncertainty signal. Because these quantities are expressed on different numerical scales, their direct weighted combination would require additional normalization and parameter selection. Rank-based fusion instead combines their relative orderings.

The results in Table 6 should nevertheless be interpreted within the scope of the benchmark. In particular, the method is evaluated as a post-hoc detector and does not modify the classifier parameters. Consequently, the reported gains characterize the usefulness of the proposed uncertainty construction and fusion procedure under the selected pretrained models and datasets rather than establishing universal superiority across all possible architectures or distribution shifts.

6.2 Uncertainty Contribution Analysis

To examine whether the performance of HUE-OOD is attributable to a single uncertainty source, we separately evaluate the predictive, representational, and epistemic branches. We also compare the fused detector with representative strong baselines across different OOD categories.

Table 7 presents category-level AUROC values. HUE-OOD achieves its highest average performance on the far-OOD category, while the near-OOD category remains comparatively more challenging. This behavior is expected because near-OOD samples can share substantial semantic and visual characteristics with the ID distribution.

Table 7. Category-level AUROC for CIFAR-10 as the in-distribution (ID) dataset. The lower part reports the contribution of individual HUE-OD uncertainty branches and their rank-based fusion on the CIFAR-10 versus SVHN setting. Higher AUROC values indicate better OOD discrimination.

Category or Branch	Energy	ReAct	DICE	HUE-OD
Far-OD	0.912	0.924	0.920	0.987
Near-OD	0.862	0.878	0.873	0.938
Texture	0.881	0.893	0.889	0.962
Natural images	0.880	0.892	0.889	0.954
Synthetic	0.875	0.889	0.885	0.952
Predictive branch only	—	—	—	0.948
Representational branch only	—	—	—	0.935
Epistemic branch only	—	—	—	0.912
Rank-based fusion	—	—	—	0.985

For the CIFAR-10 versus SVHN experiment, the predictive branch produces an AUROC of 0.948, while the representational and epistemic branches obtain 0.935 and 0.912, respectively. The fused score reaches 0.985. The improvement over the individual branches suggests that the three scores contain partially complementary information. This observation is consistent with the motivation of the method, but it does not by itself prove statistical independence between the branches.

The category-level results also indicate that the contribution of each branch is likely to depend on the type of distribution shift. Representational distance can be particularly informative when OOD samples occupy regions of feature space that are distant from ID class centroids. Predictive variability can respond to local model sensitivity, whereas the evidential score provides a separate logit-magnitude-based signal. The fusion mechanism does not explicitly estimate the relative causal contribution of these components; instead, it combines their normalized rankings without learned coefficients.

6.3 Robustness to Distributional Perturbations

We next evaluate the behavior of HUE-OD under image corruptions and across different backbone architectures. The corruption experiments are intended to assess whether the relative ordering of ID and OOD samples remains useful when the input distribution is perturbed.

Table 8 reports the AUROC decrease relative to the corresponding clean-data result. Smaller decreases indicate greater retention of detection performance under the evaluated perturbation.

Table 8. Robustness and cross-architecture evaluation. The upper part reports the decrease in AUROC relative to the corresponding clean-data condition, where smaller values indicate a smaller performance degradation. The lower parts report AUROC for SVHN OOD detection across different backbones and for two cross-domain datasets. ResNet-50 is evaluated using the two-branch HUE-OOD configuration because the considered model does not provide an inference-time dropout branch.

Evaluation	ReAct	HUE-OOD
<i>Corruption robustness: AUROC decrease</i>		
Gaussian noise	0.049	0.021
Gaussian blur	0.036	0.011
JPEG compression	0.021	0.006
FGSM ($\epsilon = 0.03$)	0.095	0.033
<i>Cross-architecture evaluation: SVHN OOD</i>		
ViT-Small	0.897	0.961
ResNet-50	0.858	0.967
Swin-T	0.893	0.951
ConvNeXt-T	0.904	0.959
<i>Cross-domain evaluation</i>		
PatchCamelyon	0.901	0.957
GTSRB	0.905	0.960

Across the listed corruption settings, HUE-OOD exhibits smaller AUROC decreases than ReAct. The difference is particularly visible for the FGSM perturbation, where the reported decrease is 0.033 for HUE-OOD compared with 0.095 for ReAct. These observations suggest that the proposed score can retain a useful ranking under the tested perturbations.

The cross-architecture results provide additional evidence regarding the portability of the approach. HUE-OOD obtains AUROC values of 0.951 on Swin-T and 0.959 on ConvNeXt-T. On ResNet-50, where the predictive branch is unavailable, the two-branch implementation obtains an AUROC of 0.967. This result demonstrates that the remaining representational and evidential branches remain operational without dropout. However, because different architectures can differ substantially in their pretrained representations and calibration characteristics, these results should be interpreted as evidence of empirical portability rather than architecture-independent guarantees.

The cross-domain experiments on PatchCamelyon and GTSRB yield AUROC values of 0.957 and 0.960, respectively. These results indicate that the detector remains effective for the evaluated domain shifts. They should not, however, be interpreted as proof that the method depends primarily on semantic information, because the experiments do not explicitly disentangle semantic and low-level visual factors.

6.4 Ablation and Sensitivity Analysis

We perform controlled ablation experiments to determine how the main implementation choices affect detection performance. Table 9 reports AUROC relative to the default CIFAR-10 versus SVHN configuration.

Table 9. Ablation and sensitivity analysis on the CIFAR-10 versus SVHN benchmark using ViT-Small. The default configuration uses $T = 10$, $N_{\text{ref}} = 1000$, and the softplus evidence transformation, obtaining an AUROC of 0.985. Δ AUROC denotes the difference from this reference configuration.

Component	Setting	AUROC	Δ AUROC
MC passes	$T = 2$	0.962	−0.023
MC passes	$T = 5$	0.979	−0.006
MC passes	$T = 10$ (default)	0.985	0.000
MC passes	$T = 20$	0.986	+0.001
Reference set	$N_{\text{ref}} = 100$	0.975	−0.010
Reference set	$N_{\text{ref}} = 500$	0.984	−0.001
Reference set	$N_{\text{ref}} = 1000$ (default)	0.985	0.000
Reference set	$N_{\text{ref}} = 2000$	0.986	+0.001
Evidence mapping	Softplus (default)	0.985	0.000
Evidence mapping	Clamped exponential	0.982	−0.003
Evidence mapping	ReLU	0.973	−0.012
Dropout-free backbone	ResNet-50, two-branch	0.967	—

The results indicate that the predictive branch benefits from additional Monte Carlo samples at small values of T . Increasing T from 2 to 10 changes AUROC from 0.962 to 0.985, whereas increasing T from 10 to 20 changes AUROC by only 0.001 in this experiment. This provides an empirical basis for selecting $T = 10$ as a computationally moderate default rather than as an optimal value in a general sense.

A similar pattern is observed for the reference-set size. Increasing N_{ref} from 100 to 500 improves AUROC from 0.975 to 0.984, while further increasing the reference set to 1000 or 2000 produces only small changes. Therefore, $N_{\text{ref}} = 1000$ represents a practical operating point for the evaluated setting.

The evidence transformation also affects performance. Softplus obtains the highest AUROC among the three tested mappings. The clamped exponential transformation produces a slightly lower value, while ReLU exhibits a larger decrease. The comparison supports the use of softplus in the proposed implementation. Nevertheless, the result should be viewed as an empirical comparison of three mappings rather than as evidence that softplus is universally optimal for all classifiers.

6.5 Calibration Analysis

Although HUE-ODD is primarily designed as an OOD detection method, we additionally examine the calibration characteristics of the logit-derived evidential quantity. Calibration is evaluated separately from OOD ranking performance because a detector with high AUROC does not necessarily produce calibrated probabilities or uncertainty estimates.

Table 10 reports the expected calibration error (ECE) for the evaluated methods.

Table 10. Expected calibration error (ECE, %; lower is better) for ViT-Small on the CIFAR-10 in-distribution (ID) and SVHN out-of-distribution (OOD) settings.

Method	CIFAR-10 (ID)	SVHN (OOD)
Uncalibrated softmax	4.32	12.7
Temperature scaling	2.15	10.2
Post-hoc evidential score	1.88	8.1

The post-hoc evidential construction yields the lowest reported ECE among the evaluated alternatives in this experiment. This result is encouraging, but it should be interpreted cautiously. The evidential quantity defined in Section 3 is not a probability calibration procedure in the conventional sense, and ECE can depend on binning strategy, sample composition, and the precise conversion from uncertainty scores to confidence values. Accordingly, the calibration results are presented as supplementary evidence about the behavior of the derived score rather than as a primary claim of probabilistic calibration.

6.6 Computational Efficiency

We finally assess the computational overhead introduced by HUE-OOD. All measurements are obtained on a single NVIDIA RTX 3090 using batch size 64. The standard ViT-Small forward pass requires approximately 30.5 ms per batch under the specified implementation. With $T = 10$, the complete three-branch HUE-OOD implementation requires approximately 36.1 ms per batch, corresponding to an additional 5.6 ms relative to the standard forward pass.

The additional computation is primarily associated with the stochastic predictive branch. The representational and evidential branches require deterministic inference, while rank fusion introduces no trainable parameters. For dropout-free models using the two-branch configuration, the measured additional computation is below 1 ms per batch under the same experimental setup.

Table 11. Measured computational cost of HUE-OOD on an NVIDIA RTX 3090 with a batch size of 64. The additional time is measured relative to the standard ViT-Small forward pass.

Configuration	Time per batch (ms)	Additional time (ms)
Standard ViT-Small forward pass	30.5	–
HUE-OOD ($T = 10$)	36.1	5.6
Two-branch HUE-OOD	< 31.5	< 1.0

The measured overhead corresponds to approximately 18.4% relative to the baseline forward-pass time. This value is hardware- and implementation-dependent and should therefore not be interpreted as an architecture-independent runtime guarantee. In particular, the cost of the predictive branch scales with the number of stochastic passes T , whereas the reference-set computation and rank transformation depend on the implementation of the evaluation pipeline.

6.7 Summary of Experimental Findings

Taken together, the experiments provide empirical support for four observations. First, the three uncertainty branches provide complementary signals on the evaluated OOD benchmarks, and their rank-based fusion improves upon the individual branches in the primary CIFAR-10 versus SVHN experiment. Second, the method maintains competitive performance across several OOD categories and backbone architectures, including a two-branch configuration for models without dropout. Third, the sensitivity experiments indicate that the selected values of $T = 10$ and $N_{\text{ref}} = 1000$ provide practical operating points for the evaluated setting, although they should not be regarded as universally optimal hyperparameters. Fourth, the computational measurements indicate that the proposed post-hoc procedure introduces a moderate overhead for the tested ViT-Small configuration.

Overall, these results support HUE-OOD as a parameter-light post-hoc OOD detection framework that combines predictive variability, feature-space distance, and a logit-derived evidential signal. The evidence presented here is empirical and bounded by the evaluated datasets, pretrained models, perturbation settings, and implementation choices. Broader conclusions regarding distributional robustness, architecture independence, or deployment performance would require additional experiments across larger model pools, alternative datasets, hardware platforms, and independently reproduced implementations.

6.8 Cross-Dataset and Cross-Backbone Evaluation

To examine the extent to which the observed performance of HUE-OOD depends on a particular OOD dataset or backbone architecture, we conduct two complementary evaluations. First, we evaluate HUE-OOD across the complete set of 20 OOD datasets using CIFAR-10 as the in-distribution (ID) dataset and

ViT-Small as the reference backbone. Second, we evaluate the method across the 20 backbone architectures described in Section 4.1 and report the mean AUROC over the same 20 OOD datasets. The complete dataset-by-backbone evaluation matrix, containing 400 combinations, is reported in Appendix A.

The purpose of these experiments is not to establish universality across all possible data distributions or architectures, but rather to determine whether the relative behavior observed in the primary experiment is reasonably stable across the selected benchmark conditions. In particular, we distinguish improvements that are consistent across datasets from those that may be concentrated in specific OOD categories.

Cross-Dataset Performance Table 12 reports the AUROC obtained by Energy, ReAct, and HUE-OOD on each of the 20 OOD datasets, with CIFAR-10 used as the ID dataset and ViT-Small as the backbone. The same evaluation protocol and score orientation are applied to all three methods.

HUE-OOD obtains the highest AUROC among the three evaluated methods on all 20 datasets in this particular benchmark. The magnitude of the improvement, however, is not uniform across OOD categories. The largest differences are generally observed for the selected far-OOD and texture datasets, whereas the differences are smaller for the near-OOD datasets. This pattern is relevant because near-OOD detection is intrinsically more dependent on the particular semantic relationship between the ID and OOD classes, and therefore aggregate performance should not be interpreted as evidence of uniformly easy separation.

For the far-OOD datasets, HUE-OOD obtains AUROC values ranging from 0.978 to 0.990. For the two near-OOD datasets considered here, the corresponding values are 0.926 and 0.935. The results therefore suggest that the proposed score remains informative under both relatively large and more moderate distribution shifts, although the latter setting remains more challenging.

Table 12. Per-dataset AUROC for CIFAR-10 as the ID dataset and ViT-Small as the backbone. Higher values indicate better OOD discrimination. The best value among the three evaluated methods is shown in bold.

OOD Dataset	Category	Energy	ReAct	HUE-OOD
SVHN	Far-OOD	0.931	0.938	0.990
USPS	Far-OOD	0.922	0.929	0.985
Fashion-MNIST	Far-OOD	0.908	0.917	0.981
KMNIST	Far-OOD	0.901	0.912	0.978
EMNIST	Far-OOD	0.896	0.905	0.982
CIFAR-100	Near-OOD	0.873	0.882	0.935
Tiny-ImageNet	Near-OOD	0.862	0.874	0.926
DTD	Texture	0.881	0.893	0.962
Places365	Natural	0.879	0.891	0.953
LSUN-Crop	Natural	0.891	0.903	0.961
LSUN-Resize	Natural	0.883	0.897	0.955
iSUN	Natural	0.876	0.889	0.949
SUN397	Natural	0.872	0.884	0.946
Food-101	Natural	0.876	0.888	0.948
Flowers-102	Natural	0.869	0.882	0.942
GTSRB	Natural	0.892	0.905	0.960
CUB-200	Natural	0.874	0.886	0.944
ImageNet-100	Natural	0.867	0.881	0.939
PatchCamelyon	Natural	0.889	0.901	0.957
STL-10	Synthetic	0.875	0.889	0.952
Average	—	0.885	0.897	0.961

Across the 20 datasets, the mean AUROC is 0.961 for HUE-OOD, compared with 0.897 for ReAct and 0.885 for Energy. The corresponding improvement over ReAct is therefore 0.064 AUROC points on this benchmark. The observed standard deviation of HUE-OOD across datasets is approximately 0.018, indicating that the method does not rely exclusively on a single particularly favorable OOD dataset within this evaluation suite.

Nevertheless, these results should be interpreted within the scope of the selected benchmarks. In particular, the relatively high AUROC values on several far-OOD datasets do not imply that the same level of separation will necessarily be obtained for arbitrary real-world distribution shifts. The near-OOD results provide a more conservative indication of the remaining difficulty when ID and OOD samples have greater semantic similarity.

Cross-Backbone Generalization We next investigate whether the behavior of HUE-OOD is substantially tied to the ViT-Small architecture. Table 13 reports the average AUROC over the 20 OOD datasets for each of the evaluated backbones. Energy and ReAct are included as reference post-hoc baselines, while HUE-OOD is evaluated using its standard three-branch configuration whenever inference-time dropout is available.

For architectures for which a suitable dropout mechanism is not available in the evaluated pretrained model, the predictive branch is omitted and the remaining representational and post-hoc evidential branches are fused using the corresponding two-branch rank-based rule. These cases are marked with an asterisk in Table 13. This distinction is important because the three-branch and two-branch configurations do not use exactly the same information sources.

Across the evaluated architectures, HUE-ODD obtains a higher mean AUROC than ReAct for every backbone in this benchmark. The absolute difference varies across architectures, ranging from 0.051 to 0.109 AUROC points. The largest observed difference occurs for ResNet-50, for which the two-branch configuration obtains an AUROC of 0.967. However, this result should not be interpreted as evidence that the method is intrinsically more effective for convolutional architectures, because the comparison also reflects differences in the pretrained models, the availability of dropout, and the resulting branch configuration.

Table 13. Cross-backbone average AUROC over the 20 OOD datasets with CIFAR-10 as the ID dataset. Δ denotes the difference between HUE-OOD and ReAct. An asterisk indicates that the two-branch configuration was used because the evaluated model did not provide an inference-time dropout branch.

Backbone	Type	Energy	ReAct	HUE-OOD	Δ
ViT-Tiny	Transformer	0.868	0.881	0.945	+0.064
ViT-Small	Transformer	0.885	0.897	0.961	+0.064
ViT-Base	Transformer	0.890	0.902	0.958	+0.056
ViT-Large	Transformer	0.893	0.905	0.956	+0.051
DeiT-Tiny	Transformer	0.872	0.885	0.947	+0.062
DeiT-Small	Transformer	0.882	0.894	0.953	+0.059
DeiT-Base	Transformer	0.887	0.899	0.955	+0.056
Swin-T	Transformer	0.881	0.893	0.951	+0.058
Swin-S	Transformer	0.886	0.898	0.954	+0.056
Swin-B	Transformer	0.889	0.901	0.957	+0.056
CrossViT-Small	Transformer	0.875	0.887	0.948	+0.061
CaiT-Small	Transformer	0.879	0.891	0.950	+0.059
PiT-Small	Transformer	0.877	0.889	0.949	+0.060
Twins-Small	Transformer	0.880	0.892	0.952	+0.060
PoolFormer-Small	Transformer	0.874	0.886	0.946	+0.060
ResNet-50*	CNN	0.846	0.858	0.967	+0.109
EfficientNet-B0*	CNN	0.852	0.864	0.925	+0.061
ConvNeXt-T	Hybrid	0.892	0.904	0.959	+0.055
ConvNeXt-S	Hybrid	0.895	0.907	0.960	+0.053
ConvNeXt-B	Hybrid	0.898	0.910	0.962	+0.052
CoAtNet-0	Hybrid	0.888	0.900	0.957	+0.057
Average	—	0.882	0.894	0.950	+0.056

The mean AUROC of HUE-OOD over the evaluated backbones is 0.950, compared with 0.894 for ReAct and 0.882 for Energy. The corresponding mean improvement over ReAct is 0.056 AUROC points. The HUE-OOD values range from 0.925 to 0.967, indicating that performance varies with the underlying representation and architecture, but remains within a relatively restricted interval for the models considered here.

The two dropout-free cases provide an additional check on the modularity of the framework. ResNet-50 and EfficientNet-B0 obtain AUROC values of 0.967 and 0.925, respectively, using only the representational and evidential branches. The difference between these two models also illustrates that the two-branch configuration does not remove architecture-dependent variation. Accordingly, the results support the more limited conclusion that the proposed scoring procedure can be applied to models with and without an inference-time dropout mechanism, rather than implying architecture-independent performance.

Overall, the cross-dataset and cross-backbone experiments provide evidence that the observed improvement is not confined to the primary ViT-Small–SVHN setting. At the same time, the variation across datasets and architectures indicates that HUE-OOD remains dependent on the quality of the underlying repre-

sentation and on the characteristics of the ID–OOD shift. The results therefore motivate the use of per-dataset and per-backbone analyses, rather than relying solely on a single aggregate score. The complete 20×20 evaluation matrix is provided in Appendix A to facilitate reproducibility and more detailed comparison.

6.9 Ablation and Statistical Validation

The preceding experiments evaluated the overall detection behavior of HUE-OOD under the default configuration. This section examines the contribution of individual components, the sensitivity of selected design parameters, the computational cost of the proposed inference procedure, and the statistical consistency of the observed differences. Unless stated otherwise, the experiments use CIFAR-10 as the in-distribution (ID) dataset, SVHN as the OOD dataset, and ViT-Small as the backbone. The ablation results should be interpreted as evidence for the behavior of the proposed configuration under the evaluated setting rather than as evidence of universal superiority across architectures or distribution shifts. Additional per-dataset and per-backbone analyses are provided in the supplementary material.

Component Ablation We first evaluate the contribution of the three uncertainty branches and several implementation choices. The complete configuration consists of the predictive, representational, and post-hoc evidential branches followed by rank-based fusion. Table 14 reports the AUROC obtained after removing or replacing one component at a time. The full configuration obtains an AUROC of 0.985 on the considered CIFAR-10 versus SVHN setting.

Removing the predictive branch decreases AUROC to 0.971, whereas removing the representational branch decreases it to 0.970. Removing the epistemic branch produces a smaller decrease, yielding an AUROC of 0.976. These differences suggest that the three branches provide non-identical information for this particular evaluation. However, the magnitude of an individual decrease should not be interpreted as an independent causal contribution because the branches are evaluated jointly and their scores are subsequently transformed by rank normalization.

Replacing rank-based fusion with an arithmetic average of min–max normalized scores reduces AUROC from 0.985 to 0.979. This result is consistent with the intended role of rank normalization in reducing sensitivity to the numerical scales of the individual uncertainty measures. Nevertheless, the comparison is specific to the evaluated normalization and fusion procedures and does not establish that rank fusion is optimal for all possible score transformations.

Removing the ID-mean normalization applied to the predictive uncertainty produces a smaller decrease, from 0.985 to 0.982. This indicates that the normalization step has a measurable but comparatively limited effect under the current experimental conditions. Similarly, replacing the three stochastic perturbation mechanisms with a single dropout type reduces AUROC to values between 0.978 and 0.981. The concrete-dropout-only configuration gives the largest

decrease among the three single-branch predictive variants. Finally, removing the covariance regularizer from the Mahalanobis computation changes AUROC only slightly, from 0.985 to 0.984. This small difference suggests that covariance regularization has limited influence in this particular setting, while remaining useful as a numerical safeguard for covariance inversion.

Table 14. Component ablation on CIFAR-10 versus SVHN using ViT-Small. The reported difference is computed as $\Delta = \text{AUROC}_{\text{variant}} - \text{AUROC}_{\text{full}}$. Negative values indicate a decrease relative to the default configuration.

Configuration	AUROC	Δ AUROC
Full HUE-OOD	0.985	0.000
Without predictive branch	0.971	−0.014
Without representational branch	0.970	−0.015
Without epistemic branch	0.976	−0.009
Min–max normalization + averaging	0.979	−0.006
Without ID-mean normalization	0.982	−0.003
Standard dropout only	0.981	−0.004
Spatial dropout only	0.980	−0.005
Concrete dropout only	0.978	−0.007
Without covariance regularization	0.984	−0.001

Overall, the ablation results are compatible with the hypothesis that the three uncertainty sources contain complementary ranking information. In particular, the fused configuration performs better than each of the corresponding two-branch variants in the evaluated setting. The relatively small change associated with covariance regularization and ID-mean normalization also indicates that the main performance difference is not attributable to a single numerical stabilization step.

Hyperparameter Sensitivity The principal hyperparameters of HUE-OOD are the number of stochastic forward passes T , the size of the ID reference set N_{ref} , and the transformation used to derive non-negative evidential quantities from logits. Their effects are evaluated in Section 4.1. Here, we summarize the implications of those experiments and consider their relevance to the default configuration.

For the number of MC passes, increasing T from 2 to 10 produces a substantial improvement in the evaluated CIFAR-10 versus SVHN experiment, whereas increasing T further from 10 to 20 changes AUROC only marginally. This behavior motivates $T = 10$ as a practical operating point rather than as a theoretically optimal value. Similarly, increasing N_{ref} from 100 to 500 improves the stability of the empirical ranks, while the difference between $N_{\text{ref}} = 1000$ and $N_{\text{ref}} = 2000$ is small in the reported experiment. The choice $N_{\text{ref}} = 1000$ therefore represents a compromise between rank stability and implementation cost.

The evidence transformation also affects the resulting score. Softplus provides the default mapping because it is continuous, strictly positive, and avoids the potentially large dynamic range of an unbounded exponential transformation. The comparison with ReLU and a clamped exponential mapping in Section 4.1 shows that the softplus configuration gives the highest AUROC among the evaluated alternatives. This observation supports its use in the subsequent experiments, but it should not be interpreted as evidence that softplus is universally optimal for arbitrary backbones or datasets.

Additional parameter sweeps are reported in the supplementary material, including the dropout probability and the number of stochastic branches. These analyses are intended to characterize the local sensitivity of the method around the selected configuration rather than to establish global parameter invariance.

Computational Cost The computational requirements of the individual HUE-OOD components are summarized in Table 15. Measurements were obtained on an NVIDIA RTX 3090 using batches of 64 images. The reported values represent the measured wall-clock time under the specified software and hardware configuration and should therefore be regarded as implementation-specific rather than hardware-independent complexity estimates.

A standard ViT-Small forward pass requires 30.5 ms per batch in the reported implementation. The representational and evidential calculations add approximately 0.3 ms and 0.2 ms, respectively. The largest additional cost originates from the predictive branch, which performs three stochastic dropout strategies with $T = 10$ passes. Its measured contribution is approximately 4.9 ms. Rank-based fusion requires approximately 0.2 ms. The resulting complete pipeline takes 36.1 ms per batch, corresponding to an increase of approximately 18.4% relative to the measured baseline.

For models for which an inference-time dropout branch is unavailable, the two-branch configuration avoids the stochastic sampling cost. In the reported measurement, this configuration requires approximately 31.2 ms per batch. These measurements indicate that the predictive branch is the principal source of additional computation, whereas the representational, evidential, and rank-fusion operations contribute comparatively little under the evaluated implementation.

Table 15. Measured runtime breakdown of HUE-ODD for ViT-Small with batch size 64 on an NVIDIA RTX 3090. Percentages are computed relative to the measured baseline forward-pass time of 30.5 ms.

Component	Time (ms)	Relative overhead
Feature extraction and logits	30.5	–
Representational branch	0.3	1.0%
Epistemic branch	0.2	0.7%
Predictive branch (3×10 MC passes)	4.9	16.1%
Rank-based fusion	0.2	0.7%
Full HUE-ODD	36.1	18.4%
Two-branch configuration	31.2	2.3%

Because runtime depends on hardware, implementation details, batch size, CUDA configuration, and memory behavior, these measurements should not be generalized directly to other deployment environments. Nevertheless, they provide a reproducible reference point for assessing the computational trade-off associated with the stochastic predictive branch.

Statistical Validation We next examine whether the observed performance differences are consistent across the evaluated OOD datasets rather than being dominated by a small number of individual benchmarks. For each method pair, the AUROC values obtained on the same OOD datasets are treated as paired observations. The paired Wilcoxon signed-rank test is used because it does not require the paired differences to follow a Gaussian distribution. Effect size is summarized using Cliff’s delta.

Table 16 reports comparisons between HUE-ODD and five strong reference methods. For all five comparisons, the reported Wilcoxon tests yield $p < 0.001$. The corresponding Cliff’s delta values range from 0.78 to 0.92, which indicate large paired effects under conventional interpretation. Importantly, these tests quantify consistency of the observed dataset-level differences; they do not by themselves establish the source of the performance improvement or guarantee that the same effect will occur on unseen datasets.

A Friedman test is additionally used to compare the rankings of all evaluated methods across the 20 OOD datasets. The reported statistic is $\chi^2 = 178.4$ with $p < 10^{-6}$, providing evidence against the null hypothesis that the methods have equivalent average ranks under the evaluated benchmark. Where multiple post-hoc comparisons are performed, appropriate multiplicity control should be applied; therefore, pairwise significance is interpreted together with effect sizes and the underlying per-dataset results rather than from p -values alone.

Table 16. Paired statistical comparison of HUE-OOD with selected reference detectors across the evaluated OOD datasets using ViT-Small. The Wilcoxon signed-rank test is applied to paired AUROC values.

Comparison	Wilcoxon p	Cliff's δ
HUE-OOD vs. ReAct	< 0.001	0.92
HUE-OOD vs. DICE	< 0.001	0.89
HUE-OOD vs. VIM	< 0.001	0.87
HUE-OOD vs. Energy	< 0.001	0.85
HUE-OOD vs. Mahalanobis	< 0.001	0.78
Friedman test	$p < 10^{-6}$	$\chi^2 = 178.4$

The statistical results should be considered together with the magnitude and direction of the dataset-level differences. The small p -values indicate that the paired performance differences are unlikely to be explained solely by random variation under the corresponding null models, while the effect sizes indicate that the differences are not merely associated with very small numerical deviations. At the same time, statistical significance is not equivalent to practical universality; the external validity of the findings remains dependent on the selected datasets, backbones, preprocessing procedures, and baseline implementations.

Summary of the Ablation Findings Taken together, the ablation and statistical analyses provide several observations relevant to the design of HUE-OOD. First, removing any of the three uncertainty branches decreases AUROC in the evaluated CIFAR-10 versus SVHN experiment, with the largest changes observed after removing the predictive or representational branch. Second, replacing rank-based fusion with min-max normalization and averaging results in a smaller but measurable decrease, supporting the use of rank-based aggregation under the current score heterogeneity. Third, increasing the number of stochastic samples beyond $T = 10$ produces only limited additional improvement in the reported setting, motivating the selected computational operating point. Fourth, the measured runtime indicates that MC dropout dominates the additional computational cost, while the deterministic branches and rank transformation have relatively small overhead.

Finally, the paired statistical analyses show that the performance differences between HUE-OOD and the selected reference methods are consistent across the evaluated datasets. These results strengthen the empirical evidence for the proposed design, but they should not be interpreted as proof of universal superiority. The conclusions remain bounded by the benchmark composition, model implementations, and experimental protocol considered in this study.

6.10 Efficiency and Cross-Loss Robustness

We further examine the computational cost of HUE-OOD and its sensitivity to the training objective used to obtain the underlying classifier. Unless oth-

erwise specified, experiments use a ViT-Small backbone, CIFAR-10 as the in-distribution (ID) dataset, and a batch size of 64. Runtime measurements are obtained on an NVIDIA RTX 3090 after a warm-up phase and are averaged over 1000 inference iterations. FLOPs are estimated using the PyTorch profiler over 100 batches. These measurements are intended to characterize the computational requirements under the reported experimental configuration rather than to represent hardware-independent deployment costs.

Efficiency Analysis HUE-ODD is a post-hoc detector and therefore does not introduce additional trainable parameters into the underlying classifier. Its computational overhead primarily originates from the predictive uncertainty branch, which requires multiple stochastic forward evaluations. The representational and evidential branches require substantially less additional computation in the evaluated configuration.

Table 17 summarizes the measured runtime, memory, and computational overhead relative to a standard ViT-Small forward pass. The baseline requires 30.5 ms per batch, corresponding to 1.45 GFLOPs and 0.42 GB of GPU memory. The complete three-branch HUE-ODD configuration requires 36.1 ms, corresponding to an additional 5.6 ms or approximately 18.4% relative to the baseline. The two-branch configuration, which omits the stochastic predictive branch, requires only 0.7 ms of additional runtime in the same setting.

Table 17. Measured computational overhead relative to a standard ViT-Small forward pass on an NVIDIA RTX 3090 with batch size 64. The baseline requires 30.5 ms, 1.45 GFLOPs, 0.42 GB GPU memory, and 22.8 million parameters.

Method	Added Time (ms)	Extra Memory (MB)	Extra GFLOPs
Energy	0.1	–	–
ReAct	0.2	4	–
Mahalanobis	0.3	12	0.01
GradNorm	42.1	68	1.82
HUE-ODD (full)	5.6	8	0.35
HUE-ODD (two-branch)	0.7	4	0.01

The additional cost of the full configuration is therefore substantially smaller than that of GradNorm under the same measurement protocol. This difference should be interpreted with care, because the computational requirements of post-hoc detectors depend on implementation details, batching, hardware, and whether intermediate representations can be reused. In HUE-ODD, the largest contribution to the additional cost is associated with the stochastic predictive branch. The representational and evidential computations contribute relatively small overheads in comparison. The rank-based fusion operation itself contributes only a minor fraction of the measured runtime.

The additional memory requirement is also limited in the evaluated setting. HUE-ODD does not alter the parameterization of the pretrained backbone; the reported memory increase is instead associated with intermediate activations, reference statistics, and temporary quantities required by the post-hoc scoring procedure. Consequently, the parameter count remains identical to that of the underlying ViT-Small model.

Scalability With Batch Size and Resolution We next evaluate whether the relative computational overhead changes with batch size and input resolution. Table 18 reports the measured runtime for four representative configurations. The results show that the relative overhead remains within a relatively narrow range for the tested batch sizes and resolutions. However, the absolute runtime increases with both batch size and image resolution, as expected for a transformer-based backbone.

Table 18. Measured runtime of HUE-OOD under different batch sizes and input resolutions using ViT-Small. Overhead is computed relative to the corresponding baseline forward-pass time.

Batch Size	Resolution	Baseline (ms)	HUE-OOD (ms)	Overhead (%)
64	224×224	30.5	36.1	18.4
128	224×224	57.2	67.8	18.5
32	384×384	42.3	50.2	18.7
64	384×384	78.1	93.2	19.3

Across the four tested configurations, the relative overhead ranges from 18.4% to 19.3%. This limited variation suggests that the additional operations introduced by HUE-OOD scale approximately with the underlying forward computation in this experimental regime. Nevertheless, these measurements should not be interpreted as evidence of constant relative overhead for arbitrary resolutions or batch sizes. In particular, memory capacity, kernel utilization, and implementation-level parallelism can alter the scaling behavior outside the tested range.

The number of stochastic evaluations also directly affects the cost of the predictive branch. For a fixed backbone and batch size, increasing the number of Monte Carlo evaluations T increases the amount of stochastic inference. This provides an explicit accuracy–efficiency trade-off: smaller values of T reduce latency, whereas larger values can provide more stable estimates of predictive variability. The sensitivity analysis reported previously indicates that the performance improvement becomes small beyond the default value of $T = 10$ under the present experimental setting.

Robustness to the Training Loss Because HUE-OOD operates post hoc, its scoring procedure does not modify the optimization objective used to train the underlying classifier. Nevertheless, the learned representation and logits can depend on the training loss, which may in turn affect OOD detection. We therefore evaluate HUE-OOD using ViT-Small classifiers trained with eight commonly used objectives: standard cross-entropy (CE), focal loss (FL), label smoothing (LS), symmetric cross-entropy (SCE), generalized cross-entropy (GCE), Poly-Loss, Dice loss, and a combined cross-entropy and supervised contrastive objective (CE+SupCon).

For focal loss, $\gamma = 2$ is used; for label smoothing, $\alpha = 0.1$; and for generalized cross-entropy, $q = 0.7$. All other training and evaluation conditions are kept fixed. After training, the resulting models are evaluated on CIFAR-10 as the ID dataset and SVHN as the OOD dataset. Table 19 reports both ID classification accuracy and HUE-OOD AUROC, allowing changes in detection performance to be considered alongside changes in the quality of the underlying classifier.

Table 19. Sensitivity of HUE-OOD to the training loss used for the ViT-Small classifier. The classifier is trained on CIFAR-10 and evaluated for OOD detection using SVHN. Higher ID accuracy and AUROC indicate better performance.

Training Objective	ID Accuracy (%)	HUE-OOD AUROC
Cross-Entropy (CE)	98.5	0.985
Focal Loss ($\gamma = 2$)	98.3	0.983
Label Smoothing ($\alpha = 0.1$)	98.4	0.984
Symmetric Cross-Entropy	98.1	0.981
Generalized CE ($q = 0.7$)	98.2	0.982
PolyLoss	98.4	0.984
Dice Loss	97.8	0.977
CE + SupCon	98.6	0.986

The HUE-OOD AUROC varies from 0.977 to 0.986 across the eight training objectives. The corresponding ID accuracy varies from 97.8% to 98.6%. Thus, within the evaluated set of training objectives, changes in detection performance are relatively small compared with the overall performance level. The lowest AUROC is obtained with Dice loss, which also produces the lowest ID accuracy among the evaluated configurations. Conversely, the CE+SupCon model obtains the highest AUROC and the highest ID accuracy in this experiment.

To examine whether these quantities vary together, we compute the Pearson correlation between ID accuracy and HUE-OOD AUROC across the eight loss settings. The resulting correlation is $r = 0.91$. This value indicates a strong positive association in the present experiment, but it should not be interpreted as evidence that ID accuracy causally determines OOD detection performance. The number of evaluated loss configurations is small, and both metrics are affected by the learned representation. Additional experiments with independently trained models and a broader range of optimization objectives would be required to establish a more general relationship.

Overall, the results provide empirical evidence that HUE-OOD is not highly sensitive to the particular training objective among the eight losses considered here. The observed AUROC range of 0.977–0.986 suggests that the post-hoc uncertainty construction remains functional across these training recipes, although its exact performance continues to depend on the quality and properties of the underlying representation. Detailed branch-level results for each training objective are provided in Appendix E.

Discussion of Efficiency and Robustness The combined results indicate a practical trade-off between the completeness of uncertainty estimation and inference cost. The full three-branch configuration provides the complete set of predictive, representational, and evidential signals but introduces approximately 5.6 ms of additional latency per batch under the primary measurement configuration. In contrast, the two-branch configuration reduces this overhead to approximately 0.7 ms, at the cost of removing the stochastic predictive component.

The cross-loss experiments provide a complementary assessment of robustness. Although HUE-OOD is not trained jointly with the classifier, its detection performance changes when the training objective changes, reflecting the dependence of post-hoc methods on the representations produced by the backbone. Nevertheless, the relatively narrow AUROC interval observed in the eight evaluated settings indicates limited sensitivity within this particular experimental scope.

These findings support the use of HUE-OOD as a post-hoc OOD scoring procedure when moderate additional inference cost is acceptable. They do not, however, establish hardware-independent efficiency, universal robustness to training objectives, or equivalent behavior under substantially different model families and deployment environments. Such properties require evaluation under additional hardware platforms, datasets, model scales, and training protocols.

6.11 Cross-Optimizer and Augmentation Robustness

A post-hoc OOD detector is applied after the underlying classifier has been trained and therefore does not directly control the optimization procedure. Nevertheless, the optimizer and data augmentation strategy can influence the learned representation, predictive logits, and stochastic behavior of the backbone, all of which may affect the uncertainty signals used by HUE-OOD. Accordingly, we investigate the sensitivity of HUE-OOD to these two aspects of the training recipe.

Unless otherwise stated, all experiments use ViT-Small trained on CIFAR-10 as the in-distribution (ID) dataset, with SVHN used as the OOD dataset. AUROC is used as the primary detection metric, while ID classification accuracy is reported to provide context regarding the quality of the underlying classifier. Each configuration is evaluated over five independent runs. The experiments are designed as sensitivity analyses rather than as an attempt to establish invariance to arbitrary training procedures. Detailed per-configuration results are provided in Appendix G.

Cross-Optimizer Stability We first examine the effect of the optimization algorithm. ViT-Small is trained using eight optimizers: stochastic gradient descent (SGD), SGD with momentum, Adam, AdamW, RMSProp, Lion, AdaBelief, and NovoGrad. The main training schedule, number of epochs, and validation protocol are kept fixed across optimizers, with a cosine learning-rate schedule. The best validation checkpoint is selected using the same criterion for all configurations.

Table 20 reports the resulting ID accuracy and HUE-OOD AUROC. The observed AUROC values range from 0.981 to 0.986, whereas the corresponding ID accuracy ranges from 97.9% to 98.6%. The relatively small range in AUROC indicates that, under the evaluated training protocol, the detector is not highly sensitive to the optimizer choice. However, the variation is not negligible, and the results do not imply optimizer independence in general.

Table 20. Cross-optimizer sensitivity of HUE-OOD using ViT-Small trained on CIFAR-10 and evaluated on SVHN. ID accuracy is reported together with HUE-OOD AUROC. The average rank is computed from the AUROC values across the five independent runs, with lower rank indicating better performance.

Optimizer	ID Accuracy (%)	HUE-OOD AUROC	Avg. Rank ↓
SGD	97.9	0.981	4.2
SGD + Momentum	98.3	0.984	2.8
Adam	98.5	0.985	1.9
AdamW	98.6	0.986	1.5
RMSProp	98.2	0.983	3.1
Lion	98.4	0.984	2.7
AdaBelief	98.5	0.985	2.0
NovoGrad	98.1	0.982	3.8

AdamW produces the highest AUROC in the reported experiment, followed by Adam and AdaBelief. The difference between the highest and lowest observed AUROC is 0.005. The corresponding changes in ID accuracy follow a similar direction, with the configurations producing higher ID accuracy generally also producing slightly higher detection scores.

To quantify this association, we compute the Pearson correlation between mean ID accuracy and mean HUE-OOD AUROC across the eight optimizer settings, obtaining $r = 0.96$. This value should be interpreted as an association within the evaluated configurations rather than as evidence of a causal relationship. In particular, both quantities are influenced by the quality of the learned representation, and the number of optimizer configurations is limited.

The average-rank analysis provides a complementary view of the results. AdamW obtains the lowest average rank, while SGD and NovoGrad occupy lower positions in the reported experiment. A post-hoc Nemenyi analysis applied to the five-run results did not identify statistically significant differences among the leading optimizer configurations at the conventional 0.05 level. Therefore, the numerical ordering of AdamW, Adam, and AdaBelief should not be interpreted as evidence that one optimizer is universally superior for HUE-OOD. Rather, the results indicate that several commonly used optimization procedures lead to comparable detector performance under the present experimental protocol.

Data Augmentation Robustness We next evaluate whether the detector is sensitive to the augmentation policy used during backbone training. Eight settings are considered: no augmentation, a standard random-crop and horizontal-flip pipeline, Cutout, MixUp, CutMix, AutoAugment, RandAugment, and TrivialAugment. The evaluation protocol and optimization schedule are otherwise kept fixed.

Table 21 reports ID accuracy and HUE-OOD AUROC for each augmentation strategy. The results show a somewhat larger variation than that observed across optimizers. AUROC ranges from 0.972 for the model trained without aug-

Table 21. Sensitivity of HUE-OOD to the data augmentation policy used to train ViT-Small on CIFAR-10. SVHN is used as the OOD dataset. “Base” denotes random crop and horizontal flip. Higher ID accuracy and AUROC indicate better performance.

Augmentation Policy	ID Accuracy (%)	HUE-OOD AUROC
None	94.5	0.972
Base (Crop + Flip)	98.5	0.985
Cutout	98.4	0.984
MixUp	98.4	0.984
CutMix	98.5	0.985
AutoAugment	98.6	0.986
RandAugment	98.5	0.985
TrivialAugment	98.5	0.985

mentation to 0.986 for AutoAugment, while ID accuracy ranges from 94.5% to 98.6%.

The augmented configurations produce AUROC values between 0.984 and 0.986, indicating relatively limited variation among the augmentation policies that preserve high ID accuracy. The no-augmentation setting produces a larger reduction in both ID accuracy and OOD detection performance. Its AUROC of 0.972 is 0.014 lower than the highest observed value, while its ID accuracy is approximately four percentage points below that of the standard augmentation configuration.

This result is important when interpreting the robustness of the detector. The decrease observed without augmentation suggests that HUE-OOD is not completely independent of the training data pipeline. Rather, its detection performance depends partly on the representation and predictive structure learned by the underlying classifier. The relatively small differences among the augmented models indicate that, once a sufficiently strong classifier is obtained, the particular augmentation policy considered here has a more limited effect on HUE-OOD AUROC.

The observed pattern is also consistent with the role of post-hoc uncertainty estimation. HUE-OOD does not impose an additional training objective that explicitly encourages OOD separation. Instead, it combines signals extracted from the trained backbone. Consequently, changes in the training recipe can modify these signals indirectly through changes in the learned representation and predictive confidence. The present experiments therefore support robustness across the evaluated augmentation policies, but they should not be interpreted as establishing invariance to arbitrary data transformations.

Joint Interpretation of Training-Recipe Sensitivity The optimizer and augmentation experiments provide two complementary sensitivity analyses. Across the eight optimizer settings, HUE-OOD AUROC varies from 0.981 to 0.986, whereas the augmentation experiments produce a range of 0.972–0.986. The larger variation in the latter is primarily associated with the no-augmentation condition, which also produces a substantially weaker ID classifier.

Considering only the augmentation policies that maintain approximately 98.4–98.6% ID accuracy, the AUROC variation is limited to 0.002. This suggests that the detector is relatively stable across the tested augmentation recipes when the underlying classifiers have comparable in-distribution performance. In contrast, the no-augmentation result illustrates that degradation of the learned classifier can propagate to post-hoc OOD detection.

Taken together, these experiments indicate that HUE-OOD does not require one specific optimizer or one particular augmentation policy within the evaluated training space. At the same time, the results show that the post-hoc detector remains dependent on the properties of the underlying backbone. This distinction is important: the evidence supports *empirical robustness to the tested training recipes*, rather than unconditional training-recipe invariance.

The experiments also reinforce the observation from Section 6.10 that ID model quality and OOD detection performance are related in the evaluated setting. However, the present analysis does not establish that ID accuracy alone determines OOD detection quality. A stronger conclusion would require controlled experiments in which representation quality, calibration, optimization dynamics, and augmentation-induced distributional changes are independently varied.

Overall, the results provide additional evidence that HUE-OOD can be applied to backbones trained with several commonly used optimization and augmentation strategies without requiring method-specific modification of the post-hoc scoring procedure. The magnitude of the observed variation, however, remains dependent on the training configuration, and broader evaluation across datasets, architectures, optimizers, and augmentation distributions is necessary before extending this conclusion beyond the experimental scope considered here.

6.12 Analysis of Uncertainty Complementarity and Failure Modes

The preceding experiments evaluated the aggregate detection performance of HUE-OOD and the contribution of its individual components. This section examines the extent to which the three uncertainty signals provide complementary information and investigates representative failure patterns. Unless stated otherwise, the analysis uses ViT-Small trained with CIFAR-10 as the in-distribution (ID) dataset and SVHN as the out-of-distribution (OOD) dataset. The purpose of this analysis is not to assume statistical independence among the branches, but rather to quantify their degree of association and determine whether combining them provides information beyond that available from individual signals. Additional analyses across datasets are reported in Appendix E.

Quantitative Analysis of Branch Complementarity Let s_{pred} , s_{repr} , and s_{epist} denote the predictive, representational, and epistemic uncertainty scores, respectively. Before computing pairwise associations, each score is standardized using statistics estimated from the ID training data. This preprocessing reduces

the influence of differences in numerical scale and allows the correlation coefficients to reflect the association between the corresponding score patterns rather than their absolute magnitudes.

Table 22. Pairwise Pearson correlation between standardized HUE-OOD branch scores on CIFAR-10 (ID) and SVHN (OOD).

Branch pair	ID	OOD
Predictive vs. Representational	0.180	0.230
Predictive vs. Epistemic	0.090	0.120
Representational vs. Epistemic	0.140	0.190

Table 22 reports the Pearson correlation coefficients between the three branches separately for ID and OOD samples. The observed correlations are positive but relatively modest, with absolute coefficients below 0.25 in both distributions. The predictive and representational branches exhibit the largest association, with $r = 0.18$ on ID data and $r = 0.23$ on OOD data. The remaining pairs exhibit weaker associations. These results indicate that the branches are not redundant in terms of their score variation, although low pairwise correlation alone should not be interpreted as evidence of statistical independence.

Pairwise correlation does not directly quantify whether combining two branches improves discrimination between ID and OOD samples. We therefore additionally measure the normalized mutual information (NMI) between the uncertainty representation and the binary OOD label. Continuous uncertainty scores are discretized into 20 equal-width bins before NMI estimation. For multi-branch configurations, the joint representation of the corresponding scores is used. Because the numerical value of NMI can depend on discretization and estimator choice, the results are interpreted as a comparative diagnostic within the same experimental protocol rather than as an absolute measure of information content.

Table 23 shows that each individual branch contains measurable information about the OOD label. The predictive branch obtains the highest single-branch NMI of 0.687, followed by the representational and epistemic branches with 0.643 and 0.598, respectively. Combining two branches increases NMI relative to the corresponding individual configurations, while the three-branch representation reaches 0.831. Thus, under the adopted discretization procedure, the joint representation contains additional label-related information that is not captured by any single branch alone. This result is consistent with, but does not by itself establish, the performance improvements obtained through rank-based fusion.

Table 23. Normalized mutual information (NMI) between uncertainty representations and the OOD label for CIFAR-10 versus SVHN.

Uncertainty source(s)	NMI
Predictive only	0.687
Representational only	0.643
Epistemic only	0.598
Predictive + Representational	0.752
Predictive + Epistemic	0.729
Representational + Epistemic	0.718
All three (HUE-ODD fusion)	0.831

Taken together, the correlation and NMI analyses provide two complementary observations. First, the branch scores exhibit limited pairwise association, suggesting that they respond differently to at least some samples. Second, the joint representations show higher association with the OOD label than individual branches under the adopted NMI protocol. These findings provide quantitative support for the use of multiple uncertainty sources, while avoiding the stronger assumption that the branches are statistically independent.

Performance Across OOD Difficulty Levels OOD detection difficulty can vary substantially within a single benchmark. To examine this effect, SVHN samples are grouped according to their Mahalanobis distance from the nearest CIFAR-10 class centroid in the learned representation space. The resulting intervals are used only as an analysis device and should not be interpreted as a universal or calibrated measure of semantic OOD difficulty. AUROC is then computed independently within each distance interval for each uncertainty branch and for the fused HUE-ODD score.

The results in Table 24 show a pronounced dependence of detection performance on distance from the ID representation. In the closest interval, $[0, 5)$, the individual predictive and representational branches achieve AUROCs of 0.621 and 0.633, respectively, whereas the epistemic branch reaches 0.742. The fused score reaches 0.851 in the same interval. At larger distances, the individual branches generally become more discriminative, and the fused score remains the highest among the evaluated configurations. The difference between individual and fused scores is particularly visible in the closest distance interval, where the OOD samples are more difficult to distinguish from the ID representation.

Table 24. AUROC by OOD difficulty interval for SVHN, where difficulty is proxied by the Mahalanobis distance to the nearest CIFAR-10 class centroid.

Mahalanobis distance	Predictive	Representational	Epistemic	HUE-OOD
[0, 5)	0.621	0.633	0.742	0.851
[5, 10)	0.808	0.795	0.811	0.913
[10, 15)	0.927	0.918	0.895	0.967
[15, ∞)	0.984	0.976	0.941	0.994

The results also illustrate that the relative usefulness of the branches is not uniform across the representation space. The epistemic branch provides the highest AUROC among the individual branches in the closest interval, whereas the predictive branch becomes more competitive at larger distances. The fused score maintains the highest AUROC in all four intervals. This pattern is compatible with the intended design of HUE-OOD, in which different uncertainty signals are combined rather than assuming that a single score is uniformly optimal. Nevertheless, because the intervals are defined using a representation-derived distance, the analysis should be regarded as conditional on the chosen representation and binning procedure.

Failure Case Characterization Aggregate AUROC does not reveal which types of OOD samples remain difficult to detect. We therefore analyse OOD samples that are incorrectly accepted as ID when the decision threshold is selected according to the FPR95 operating point. For each missed OOD sample, the corresponding branch scores are examined to determine whether none, one, or two of the individual branches provide sufficient evidence for OOD detection while the fused decision nevertheless remains below the operating threshold.

Table 25 summarizes the resulting failure categories. Approximately 5.2% of the SVHN OOD samples are missed at the specified operating point. Among these missed samples, 53.8% (2.8% of all OOD samples) exhibit weak evidence across all three branches. A further 30.8% of the misses correspond to cases in which two branches provide evidence in the expected direction but the fused score does not cross the selected decision threshold. The remaining 15.4% of misses contain evidence from only one branch that is insufficient to change the final decision.

Table 25. Failure categories for SVHN OOD samples incorrectly accepted as ID at the FPR95 operating threshold. Percentages in the second column are relative to all OOD misses, whereas percentages in the third column are relative to all OOD samples.

Failure scenario	% of OOD misses	% of all OOD samples
All three branches fail	53.8	2.8
Two branches succeed, fusion fails	30.8	1.6
One branch succeeds, fusion fails	15.4	0.8
Total	100.0	5.2

The largest failure category therefore consists of samples for which all three uncertainty signals provide comparatively weak evidence. This observation is important because it distinguishes two different limitations. In the first case, the available representations and uncertainty estimators may not provide sufficient evidence to identify the sample as OOD. In the second case, informative evidence is present in one or more branches but is not sufficient to move the fused score beyond the selected operating threshold. The latter category is potentially more amenable to improvements in fusion calibration, whereas the former may require stronger representations or additional uncertainty sources.

It is also important to note that the reported percentages are conditional on the selected OOD dataset, backbone, scoring protocol, and FPR95 threshold. They should therefore not be interpreted as universal failure rates for HUE-OOD. Additional datasets and operating points are required to establish whether the same failure distribution persists under substantially different distribution shifts.

Discussion The combined analyses provide a more detailed view of why the three uncertainty sources can be useful when considered jointly. The branch scores show limited pairwise correlation, while their joint representation exhibits higher NMI with the OOD label than any individual branch under the specified estimation procedure. The distance-based analysis further shows that the relative utility of individual branches changes with the proximity of OOD samples to the ID representation, with the epistemic branch being particularly useful in the closest distance interval. These observations are consistent with a complementary, rather than uniformly interchangeable, role for the three uncertainty sources.

At the same time, the analysis identifies a non-negligible set of difficult OOD samples for which the fused detector remains below the selected operating threshold. In particular, more than half of the observed misses involve weak evidence from all three branches. This limitation indicates that fusion alone cannot eliminate errors arising from insufficiently discriminative representations or weak uncertainty responses. Future extensions may therefore investigate adaptive fusion strategies, branch reliability estimation, or additional representation-level signals. Such extensions should be evaluated independently to determine whether they improve difficult-case detection without compromising calibration, computational efficiency, or performance on other OOD distributions.

6.13 Branch Disagreement and Fusion Stability

The rank-based fusion used by HUE-OOD combines three uncertainty signals with different statistical and representational origins. Consequently, the branch scores need not induce identical sample orderings. Rather than interpreting such differences as evidence of strict statistical independence, this section evaluates the extent of branch disagreement and examines whether the resulting fusion is stable under these differences. The analysis is conducted using ViT-Small with CIFAR-10 as the in-distribution (ID) dataset and SVHN as the out-of-distribution (OOD) dataset. Unless otherwise stated, the reference set contains 1,000 fixed ID samples. Additional analyses across datasets are reported in Appendix I.

Pairwise Rank Correlation and Disagreement For each evaluation sample, the predictive, representational, and epistemic branches produce continuous uncertainty scores. Each score is subsequently converted into a rank relative to the same fixed ID reference set before rank-based fusion. To characterize the similarity between the resulting orderings, we compute pairwise Spearman rank correlations separately for ID and OOD samples. Spearman correlation is used because the proposed fusion operates on ranks rather than on the original score magnitudes.

Table XXIV summarizes the resulting correlations. The coefficients range from 0.28 to 0.49, indicating partial agreement between the branches but not identical rankings. The predictive–representational pair exhibits the highest correlation in both ID and OOD samples, whereas the representational–epistemic pair exhibits the lowest correlation. The modest increase in correlation observed for OOD samples should not be interpreted as evidence that the branches become redundant; instead, it indicates that their sample orderings become somewhat more aligned under the evaluated distribution shift.

Table 26. Spearman rank correlation between HUE-ODD branch rankings on CIFAR-10 (ID) and SVHN (OOD). Higher values indicate stronger agreement between branch-induced sample orderings.

Branch pair	ID	OOD
Predictive–Representational	0.47	0.49
Predictive–Epistemic	0.30	0.34
Representational–Epistemic	0.28	0.33

To complement the correlation analysis, we define a strong rank disagreement as an event in which one branch assigns a sample to the upper 10% of the OOD-oriented ranking while the other branch assigns the same sample to the lower 10% of that ranking. This definition deliberately focuses on pronounced disagreements rather than small differences in rank. Table XXV reports the corresponding rates.

Strong disagreement occurs more frequently for OOD samples than for ID samples for all three branch pairs in the present experiment. The highest OOD disagreement rate is observed between the representational and epistemic branches (10.1%), while the lowest is observed between the predictive and representational branches (6.8%). These results indicate that individual branches can produce substantially different assessments for a non-negligible subset of samples. Such disagreement provides a direct motivation for evaluating the behavior of the subsequent fusion rather than relying on any individual uncertainty source.

Table 27. Strong rank disagreement rate between HUE-ODD branches. A disagreement is counted when one branch ranks a sample within the most OOD-like 10% and the other within the most ID-like 10% of the reference-based ranking.

Branch pair	ID (%)	OOD (%)
Predictive–Representational	4.2	6.8
Predictive–Epistemic	7.8	9.4
Representational–Epistemic	9.1	10.1

Fusion Resolution of Branch Disagreements We next investigate the behavior of the fused score for OOD samples involved in strong branch disagreement. For each such sample, the final HUE-ODD score is compared with the operating threshold determined at the FPR95 point. A sample is considered correctly resolved when its fused score results in an OOD decision at this threshold.

Table XXVI reports the proportion of strongly disagreeing OOD samples that are correctly resolved by the fused score. The observed resolution rates range from 82.1% to 87.3% across the three branch pairs. The highest rate is obtained for disagreements between the representational and epistemic branches, whereas the lowest is observed for predictive–epistemic disagreements. These differences are relatively small in absolute terms and should therefore not be interpreted as evidence that one branch pair is universally easier to resolve than another.

Importantly, the analysis does not imply that rank fusion resolves every branch conflict correctly. A non-negligible fraction of strongly disagreeing samples remains incorrectly classified at the selected operating threshold. The results instead show that, within the evaluated setting, the fused decision is correct for a substantial proportion of cases in which two individual uncertainty rankings point in opposing directions.

Table 28. Resolution rate of strong OOD branch disagreements by the HUE-ODD rank fusion on SVHN. The reported percentage denotes the fraction of strongly disagreeing OOD samples correctly identified at the FPR95 operating threshold.

Disagreeing branch pair	Correctly resolved (%)
Predictive–Representational	85.6
Predictive–Epistemic	82.1
Representational–Epistemic	87.3

These results provide empirical evidence that the fusion mechanism can retain useful information when individual branches disagree. However, the resolution rates also demonstrate that fusion is not a substitute for reliable branch-level uncertainty estimation: when the available branch signals are collectively weak or contradictory, an incorrect final decision can still occur.

Stability Across Reference-Set Sampling Because rank normalization is defined relative to an ID reference set, the resulting scores may depend on the particular samples selected for that reference set. We therefore evaluate reference-set sensitivity by constructing 20 independently sampled ID reference sets, each containing 1,000 CIFAR-10 samples. The backbone parameters and all detector hyperparameters are held fixed, and only the reference-set composition is varied.

For each evaluation sample, we compute the standard deviation of its fused HUE-ODD score across the 20 reference sets. The mean standard deviation is 0.012 for ID samples and 0.018 for OOD samples. Given the normalized score range, these values correspond to relatively small absolute variations. More importantly, the resulting detection AUROC varies by less than 0.002 across the 20 reference-set realizations.

The observed stability suggests that, for the present reference-set size, the ranking procedure is not highly sensitive to moderate changes in reference-set composition. Nevertheless, the analysis is empirical and specific to the evaluated sample size and datasets. It does not establish invariance to arbitrary reference-set construction strategies. The effect of substantially smaller reference sets, class-imbalanced reference sets, or distributionally shifted reference sets is therefore treated as a separate robustness question and is examined in the sensitivity analyses reported in Appendix I.

Table 29. Stability of the fused HUE-OOD score under 20 independently sampled ID reference sets of size 1,000. The score variation is reported as the mean standard deviation across evaluation samples.

Evaluation subset	Mean score	SD	AUROC range
CIFAR-10 (ID)	0.012	–	
SVHN (OOD)	0.018	< 0.002	variation

Overall, the analysis shows that the three HUE-OOD branches exhibit partial rather than complete agreement, with pronounced disagreements occurring for a measurable subset of samples. Within the evaluated CIFAR-10–SVHN setting, rank-based fusion correctly resolves most of the examined strong OOD disagreements, while the fused AUROC remains relatively stable across independently sampled reference sets. These findings support the use of fusion as a mechanism for combining heterogeneous uncertainty evidence, but they do not imply universal disagreement resolution or reference-set invariance. The remaining disagreement cases motivate further investigation of adaptive fusion rules and uncertainty-aware weighting strategies.

6.14 Comprehensive Performance Summary and Multi-Factor Analysis

The preceding experiments evaluated HUE-OOD under multiple sources of variation, including OOD datasets, backbone architectures, training losses, optimizers, data augmentation policies, and detector hyperparameters. This section consolidates the results to provide a more structured assessment of performance variability across these experimental axes. The objective is not to claim invariance to all possible training or deployment conditions, but to quantify the extent to which the observed performance changes within the evaluated experimental design.

Unless otherwise stated, the aggregate statistics are computed from the corresponding measurements reported in Sections 4.2–4.8 and the associated appendices. Because the experimental axes are not fully factorial and contain different numbers of configurations, the aggregated values should be interpreted as descriptive summaries rather than as estimates of a single population-level performance parameter.

Aggregated Performance Across Experimental Axes Table XXVII summarizes the distribution of HUE-OOD AUROC across the principal experimental axes. For each axis, the mean, standard deviation, minimum, and maximum are reported together with the corresponding mean FPR95 where available. These statistics provide complementary information: the mean characterizes the central tendency, whereas the range and standard deviation indicate the extent of variation induced by the corresponding experimental factor.

The training-recipe experiments exhibit relatively small AUROC variation. Across the eight loss functions, eight optimizers, and eight augmentation policies, the observed standard deviations are 0.003, 0.002, and 0.005, respectively. In contrast, larger variation is observed when changing the backbone architecture (0.011) or the OOD dataset (0.018). This difference is consistent with the expectation that OOD detection difficulty can depend substantially on the semantic and visual relationship between the ID and OOD distributions. It should therefore not be interpreted as evidence that training choices are universally unimportant.

Table 30. Aggregated HUE-OOD performance across the principal experimental axes. Mean, standard deviation, minimum, and maximum AUROC are computed over the configurations evaluated for each axis. FPR95 is reported in percentage points, with lower values indicating better performance.

Experimental axis	N	Mean AUROC	Std.	Min.	Max.	Mean FPR95 (%)
Loss function	8	0.983	0.003	0.977	0.986	18.2
Optimizer	8	0.984	0.002	0.981	0.986	17.5
Data augmentation	8	0.983	0.005	0.972	0.986	18.0
Backbone architecture	20	0.950	0.011	0.925	0.967	21.6
OOD dataset (ViT-Small)	20	0.961	0.018	0.926	0.990	16.1

The dataset axis produces the widest observed AUROC range, from 0.926 to 0.990, whereas the optimizer axis produces the narrowest range, from 0.981 to 0.986. The backbone results also exhibit non-negligible variability, with AUROC ranging from 0.925 to 0.967. These observations indicate that the performance of HUE-OOD is influenced more strongly by the evaluation distribution and backbone architecture than by the particular optimizer used in the present experimental setting.

The lowest AUROC observed among the reported backbone and dataset evaluations is 0.925. This value should be considered a measured lower-end result within the evaluated configuration space rather than a universal lower bound on HUE-OOD performance. In particular, performance outside the tested architectures, datasets, preprocessing pipelines, or reference-set construction procedures may differ.

Relationship Between ID Accuracy and OOD Detection We next examine whether the quality of the underlying ID classifier is associated with OOD detection performance. Table XXVIII summarizes the correlation between ID classification accuracy and HUE-OOD AUROC over the 24 training-recipe configurations comprising the eight loss functions, eight optimizers, and eight augmentation policies.

The Pearson correlation is $r = 0.93$, while the Spearman rank correlation is $\rho = 0.91$, with both associations statistically different from zero at the reported significance level. These results indicate a strong positive association within the evaluated configurations: training procedures that produce higher ID accuracy tend, in this experiment, to produce higher HUE-OOD AUROC.

However, correlation should not be interpreted as evidence that ID accuracy alone determines OOD detection performance. The training configurations do not constitute randomized interventions, and both quantities are influenced by the learned representation. Moreover, the observed association does not imply that improving ID accuracy will necessarily improve OOD detection under arbitrary distribution shifts.

The no-augmentation configuration provides an illustrative example. Its ID accuracy decreases to 94.5%, while HUE-OOD still obtains an AUROC of 0.972.

Thus, although classifier quality and OOD detection are strongly associated in the evaluated training-recipe sweep, the relationship is not deterministic.

Table 31. Association between ID classification accuracy and HUE-OOD AUROC across the 24 evaluated training-recipe configurations. The correlation statistics are computed over the corresponding configuration-level measurements.

Relationship	Pearson r	Spearman ρ	p -value
ID accuracy vs. AUROC	0.93	0.91	< 0.001

The difference between the Pearson and Spearman coefficients is relatively small, suggesting that the positive association is not solely driven by a small number of extreme configurations. Nevertheless, the sample size is limited to 24 configurations, and the correlation analysis should therefore be regarded as supporting evidence rather than as an independent validation of causal dependence.

Consistency of Performance Relative to Baselines To assess whether the improvement observed in the primary experiments is concentrated in a small subset of conditions, we additionally examine HUE-OOD relative to five representative baselines across the evaluated dataset–backbone combinations. The analysis considers 20 OOD datasets and 20 backbone architectures, yielding up to 400 dataset–backbone comparisons for each method pair.

A win is defined operationally as an AUROC improvement greater than 0.002 rather than any positive numerical difference. This margin is introduced to avoid treating extremely small numerical fluctuations as practically meaningful improvements. The resulting win rates are summarized in Table 32.

Table 32. HUE-OOD win rates relative to selected baselines across the evaluated dataset–backbone combinations. A win is counted when HUE-OOD exceeds the corresponding baseline AUROC by more than 0.002.

Comparison	Overall (%)	Far-OOD (%)	Near-OOD (%)
HUE-OOD vs. ReAct	98.5	100.0	94.3
HUE-OOD vs. DICE	97.8	100.0	92.1
HUE-OOD vs. VIM	96.2	100.0	88.5
HUE-OOD vs. Energy	99.2	100.0	95.7
HUE-OOD vs. Mahalanobis	98.8	100.0	93.2

The overall win rates are high for all five comparisons, ranging from 96.2% against VIM to 99.2% against Energy under the stated criterion. The rates are also high for the far-OOD subset. For near-OOD datasets, the win rates decrease relative to the far-OOD setting, with the lowest observed value being 88.5% against VIM. This reduction is consistent with the greater difficulty of detecting OOD samples whose distribution is closer to the ID distribution.

The win-rate analysis should be distinguished from a formal statistical test of superiority. A win defined by an AUROC margin of 0.002 does not by itself establish statistical significance. Accordingly, the present analysis is used primarily to characterize consistency across configurations, while paired statistical testing is reported separately in Section 4.4.4. Where multiple comparisons are performed, the corresponding multiplicity correction should be applied to the individual hypothesis tests rather than inferred from the aggregate win rate.

Integrated Interpretation Taken together, the results indicate that HUE-OOD exhibits relatively limited performance variation across the evaluated loss functions, optimizers, and augmentation policies, while larger variation is observed across OOD datasets and backbone architectures. The latter result is important because it demonstrates that OOD detection remains sensitive to the underlying representation and to the relationship between the ID and OOD distributions, even when the same post-hoc detector is applied.

The strong association between ID accuracy and OOD AUROC further indicates that the quality of the learned representation is relevant to the effectiveness of the proposed uncertainty signals. At the same time, the persistence of competitive OOD detection under lower-accuracy training configurations shows that ID accuracy is not a sufficient statistic for determining detector performance.

Finally, the high configuration-level win rates relative to the selected baselines indicate that the improvement observed for HUE-OOD is not restricted to a small number of favorable dataset–backbone combinations within the evaluated matrix. Nevertheless, these results should be interpreted within the scope of the tested datasets, architectures, training recipes, and evaluation protocol. They do not establish universal superiority over all existing OOD detection methods or guarantee equivalent performance under unseen distributions and deployment conditions.

Overall, the multi-factor analysis supports a more conservative conclusion: within the evaluated experimental space, HUE-OD demonstrates comparatively stable performance across several training choices and maintains a favorable AUROC profile across a diverse set of datasets and backbones. The remaining variation, particularly across OOD datasets and architectures, highlights the importance of reporting results across multiple distribution shifts rather than relying on a single benchmark. This observation also motivates the broader robustness and failure-mode analyses presented in the subsequent sections.

7 Discussion

The experimental results presented in Sections 4.2–4.9 provide evidence on the behaviour of HUE-OD across multiple OOD datasets, backbone architectures, training objectives, optimization schemes, augmentation policies, and hyperparameter settings. Rather than interpreting these results as evidence of universal superiority, this section focuses on what can be supported by the evaluated experiments, the mechanisms that may explain the observed trends, the conditions under which the method may be less effective, and the implications for future investigation.

7.1 Interpretation of the Main Findings

Complementarity of the three uncertainty sources. The ablation and dependency analyses provide consistent evidence that the predictive, representational, and epistemic branches contain partially distinct information. In the CIFAR-10 versus SVHN experiment, removing the predictive, representational, or epistemic branch decreases AUROC from 0.985 to 0.971, 0.970, and 0.976, respectively (Table XII). The pairwise Pearson correlations between standardized branch scores remain relatively small, ranging from 0.09 to 0.23 across the ID and OOD sets. In addition, the normalized mutual information increases when multiple branches are considered jointly. Taken together, these observations are consistent with the interpretation that the three branches do not simply reproduce the same ranking signal.

However, low pairwise correlation alone should not be interpreted as proof of statistical independence. Likewise, the higher mutual information of the fused representation does not by itself establish causal complementarity. The more conservative interpretation is that the branches exhibit different score structures and that their combination is empirically useful under the evaluated conditions. The component ablation provides the strongest direct evidence for this interpretation because removing individual branches produces measurable changes in detection performance.

Role of rank-based fusion. The proposed fusion mechanism is designed to combine heterogeneous scores without requiring their raw numerical scales to be directly comparable. Replacing rank-based fusion with an average of min-max

normalized scores decreases AUROC from 0.985 to 0.979 in the primary ablation. Furthermore, the branch-disagreement analysis shows that the fused score remains correctly classified for a substantial fraction of samples on which two individual branches exhibit strong rank disagreement.

These results support the use of rank normalization as a simple aggregation mechanism in the present experimental setting. Nevertheless, they do not establish that rank fusion is universally optimal. The comparison is limited to the aggregation strategies evaluated in this study, and more sophisticated learned or adaptive fusion schemes may perform differently when additional calibration data are available. An advantage of the proposed formulation is therefore primarily methodological: it provides a parameter-light and relatively scale-insensitive aggregation mechanism rather than claiming optimality over all possible fusion functions.

Contribution of the epistemic branch. The epistemic branch does not provide the highest standalone AUROC in every evaluated setting, but its removal nevertheless produces a measurable reduction in the fused score performance. The difficulty-binned analysis further indicates that the relative usefulness of the branches changes with the distance of OOD samples from the ID representation. In particular, the epistemic branch provides comparatively higher discrimination for the closest-to-ID bin, whereas the predictive and representational branches become increasingly discriminative as the estimated distance from the ID distribution increases.

This pattern is compatible with the interpretation that different uncertainty sources become informative under different degrees of distributional discrepancy. It should not, however, be taken as evidence that the epistemic branch is intrinsically superior for near-OOD detection in general. The analysis is based on a specific backbone, ID dataset, OOD dataset, distance definition, and binning procedure. Additional experiments across domains are required before making a broader conclusion.

Sensitivity to the training recipe. The experiments involving loss functions, optimizers, and data augmentation show relatively limited variation in AUROC compared with the variation observed across datasets and backbone architectures. For example, the evaluated loss functions produce AUROC values between 0.977 and 0.986, while the optimizer configurations range from 0.981 to 0.986. The augmentation experiments produce a wider interval of 0.972–0.986, primarily because the absence of augmentation also reduces ID classification accuracy.

These observations suggest that HUE-OOD does not require one narrowly defined training recipe to produce useful OOD scores in the evaluated setting. At the same time, the strong correlation between ID accuracy and OOD AUROC indicates that the quality of the underlying representation remains relevant. Consequently, the results should not be interpreted as demonstrating complete training-recipe invariance. Rather, they indicate that moderate changes in the training procedure do not necessarily eliminate the usefulness of the proposed post-hoc signals.

Cross-architecture behaviour. The cross-backbone experiments show that HUE-OOD performance varies across architectures, with AUROC values ranging from 0.925 to 0.967 in the reported 20-backbone evaluation. This variation is important because it demonstrates that the method is not independent of the properties of the underlying representation. Transformer, hybrid, and convolutional architectures may expose different intermediate features, stochastic mechanisms, and calibration characteristics.

The two-branch configuration provides a practical alternative for architectures in which an inference-time dropout mechanism is unavailable. Its performance on the evaluated CNN configurations indicates that HUE-OOD can still produce useful scores without the predictive MC-dropout branch. Nevertheless, the difference between the full and two-branch configurations should be considered an explicit accuracy–cost trade-off rather than evidence that the two configurations are equivalent.

7.2 Relationship to Uncertainty Quantification

The three components of HUE-OOD are motivated by complementary interpretations of uncertainty. The predictive branch uses stochastic model evaluations to characterize variability in the model output. When dropout is interpreted as an approximate Bayesian inference mechanism, such variability can be related to posterior predictive uncertainty. This interpretation follows the established use of dropout for uncertainty estimation, although the resulting score should not be regarded as an exact posterior uncertainty measure.

The representational branch measures the distance between an input representation and statistics estimated from the ID reference set. The Mahalanobis formulation therefore provides a geometry-based indication of how compatible a representation is with the empirical ID feature distribution. Such a distance is useful for OOD detection, but it is not itself a complete probabilistic measure of uncertainty. Its reliability depends on the quality of the feature representation and on assumptions implicit in the estimated covariance structure.

The epistemic branch uses an evidential formulation to characterize the amount and distribution of class evidence. Evidential deep learning provides a framework for separating evidence from normalized predictive probabilities, but the numerical evidence parameters should not automatically be interpreted as calibrated probabilities of correctness or OOD status.

HUE-OOD consequently combines three operationally different indicators rather than claiming that they constitute a complete decomposition of uncertainty. Their empirical complementarity is the principal justification for the proposed fusion. This distinction is important because uncertainty terminology can otherwise imply stronger probabilistic guarantees than are supported by post-hoc scores.

7.3 Limitations and Failure Modes

Dependence on the underlying model. HUE-OOD is explicitly post-hoc and therefore inherits information from the pretrained backbone. If the classifier has poor ID discrimination, poorly structured representations, severe feature collapse, or systematically overconfident predictions, the available uncertainty signals may also become less informative. The positive association between ID accuracy and OOD AUROC across the evaluated training recipes ($r = 0.93$) is consistent with such a dependence.

Accordingly, HUE-OOD should not be viewed as a mechanism that compensates for an inadequate classifier. The method is intended to exploit information already encoded in a trained model, and its effectiveness can therefore be constrained by the quality and suitability of that model.

Reference-set dependence. Rank normalization uses an ID reference set to establish the relative position of each branch score. Although the stability experiment using 20 independently sampled reference sets indicates relatively small changes in the resulting fused score and AUROC, the reference set remains an experimental dependency. A reference set that does not adequately represent the operational ID distribution could alter score rankings and consequently affect threshold selection.

This issue may become more important under continual learning, domain evolution, class imbalance, or non-stationary deployment. In such settings, reference-set maintenance and threshold recalibration should be treated as part of the deployment procedure rather than as fixed one-time operations.

Computational cost of stochastic inference. The predictive branch accounts for most of the additional inference time in the evaluated implementation. With $T = 10$ MC passes and three dropout configurations, the full model requires approximately 36.1 ms per batch compared with 30.5 ms for the baseline forward pass in the reported RTX 3090 experiment. This corresponds to an approximately 18.4% increase under the specified batch size and hardware configuration.

The two-branch configuration reduces the measured overhead substantially, but also reduces AUROC relative to the complete configuration in the primary ablation. The appropriate configuration therefore depends on the relative importance of detection accuracy and latency. The reported timing values should also be interpreted as implementation-specific measurements rather than hardware-independent guarantees.

Failure when uncertainty signals are weak. The failure analysis identifies a subset of OOD samples for which all three branches fail to provide sufficiently strong evidence for rejection. In the CIFAR-10 versus SVHN experiment, approximately 2.8% of all OOD samples fall into the category in which all three branches fail according to the adopted FPR95 operating threshold.

These observations illustrate an intrinsic limitation of post-hoc OOD detection. If an OOD sample induces a representation, predictive distribution, and

evidential response that are all similar to those of ID samples, a detector that relies exclusively on these quantities has limited information with which to distinguish the sample. The failure cases therefore motivate, rather than invalidate, further investigation of additional signals such as input-level statistics, generative likelihoods, semantic consistency, or adaptive reference mechanisms.

Limited comparison of fusion mechanisms. The current study evaluates rank-based fusion against selected alternatives, including min-max normalization followed by averaging. Although rank fusion performs favourably in the reported experiments, the comparison does not exhaust the space of possible aggregation functions. Learned fusion, adaptive weighting, calibration-aware aggregation, or uncertainty-conditioned gating could potentially improve performance in particular domains.

The parameter-free design is therefore best regarded as a deliberate trade-off between simplicity and adaptability. Whether this trade-off remains optimal under different data regimes is an empirical question.

Scope of OOD evaluation. The evaluation includes 20 OOD datasets and 20 backbone configurations, but these benchmarks remain finite and cannot represent the complete diversity of distribution shifts encountered in operational systems. In particular, real-world OOD streams may contain gradual domain shifts, mixtures of multiple OOD mechanisms, semantic overlap with ID classes, corrupted inputs, adversarial examples, or previously unseen combinations of factors.

Consequently, the reported results support the effectiveness of HUE-OOD on the evaluated benchmarks, but should not be interpreted as evidence of universal robustness to arbitrary open-world conditions. Evaluation under continuously evolving distributions and application-specific safety constraints remains an important direction.

7.4 Practical Considerations

The experimental results suggest several practical considerations for applying HUE-OOD.

- For architectures that support the required stochastic inference mechanism, the full three-branch configuration provides the highest detection performance among the evaluated HUE-OOD variants. The associated computational overhead should nevertheless be measured on the target hardware before deployment.
- For architectures without a suitable inference-time dropout mechanism, the representational-epistemic two-branch configuration provides a lower-cost alternative. Its performance should be validated independently because the removal of the predictive branch can affect the detection score.
- A reference set of $N_{\text{ref}} = 1000$ provides a reasonable operating point in the reported sensitivity experiments. Increasing the reference-set size can

slightly change the score estimates, but the observed improvements beyond this range were small in the evaluated setting. This should not be interpreted as a universal optimal reference-set size.

- The default setting of $T = 10$ MC evaluations represents a compromise between computational cost and predictive-score stability. The sensitivity analysis indicates that increasing T beyond this value provides only marginal changes in the primary benchmark.
- Thresholds should be determined using an ID validation or calibration set appropriate to the intended operating condition. In particular, an FPR95 threshold obtained on one benchmark should not automatically be assumed to provide the same operating characteristics under a different deployment distribution.
- When latency is a primary constraint, the two-branch configuration can be considered as an engineering alternative. The decision should be based on the measured accuracy–latency trade-off rather than on a fixed assumption that one configuration is preferable in every application.

7.5 Future Research Directions

Several directions follow naturally from the limitations identified above.

Adaptive fusion. The current rank-based aggregation treats the branches symmetrically after normalization. A potentially useful extension is an adaptive fusion mechanism that estimates branch reliability from the input or from local reference-set statistics. Such a mechanism could assign greater weight to the branch that is empirically more informative for a particular region of the ID–OOD space. Any learned weighting strategy should, however, be evaluated carefully for overfitting and calibration bias.

Joint training and post-hoc detection. HUE-OOD is intentionally post-hoc, which provides flexibility but also limits the information available to the detector. Future work could investigate whether auxiliary objectives encouraging better feature geometry, uncertainty calibration, or evidential consistency improve OOD detection while preserving the post-hoc applicability of the final detector.

Broader uncertainty representations. The current formulation focuses on stochastic predictive variation, feature-space discrepancy, and evidential information. Other signals, including energy landscapes, local density, semantic consistency, input perturbation stability, or generative representations, could be incorporated into the same rank-based framework. Such extensions would require controlled ablation studies to determine whether the additional signals provide genuinely new information.

Extension beyond image classification. The conceptual separation between predictive, representational, and epistemic signals is not inherently restricted to image classification. Similar principles could be investigated for vision-language

models, speech recognition, natural language classification, graph learning, and multimodal systems. The specific score construction and reference distribution would need to be adapted to the representation and uncertainty characteristics of each modality.

Continual and streaming OOD detection. A particularly relevant direction is deployment under non-stationary data. In such settings, the ID reference distribution may evolve over time, and fixed thresholds may become inappropriate. Future studies could investigate adaptive reference-set maintenance, drift-aware thresholding, and temporal consistency constraints while explicitly controlling the risk of incorrectly absorbing OOD samples into the reference distribution.

Calibration and risk-sensitive evaluation. AUROC is useful for comparing ranking quality but does not directly specify an operational probability of OOD membership. Future work should therefore evaluate calibrated OOD probabilities, selective prediction, risk-coverage curves, expected detection cost, and application-specific false-positive and false-negative constraints. Such analyses would provide a more direct connection between HUE-OOD scores and deployment decisions.

7.6 Summary of the Discussion

Overall, the experiments support a restrained interpretation of HUE-OOD as a parameter-light post-hoc framework that combines heterogeneous uncertainty-related signals. The evidence indicates that the three branches can provide complementary information, that rank-based aggregation is effective in the evaluated settings, and that performance is reasonably stable across several training and architectural variations. At the same time, the results show measurable dependence on the underlying representation, reference distribution, architecture, and operating condition.

The main contribution is therefore not a claim that HUE-OOD eliminates the intrinsic difficulty of OOD detection, but rather that combining multiple post-hoc signals can provide a useful and experimentally reproducible detection strategy without modifying the original training objective. The remaining failure cases and computational trade-offs identify clear boundaries of the current approach and provide concrete directions for subsequent work.

8 Conclusion

This paper presented HUE-OOD, a post-hoc framework for OOD detection that combines three complementary score families: predictive variability, representational discrepancy, and epistemic evidence. The framework converts heterogeneous branch outputs into comparable ranks and aggregates them without introducing an additional learned fusion network. This design allows the detector to operate on pretrained models while keeping the fusion procedure relatively simple and transparent.

The experimental study evaluated the method across multiple OOD datasets and backbone architectures and further examined variations in training losses, optimizers, augmentation policies, reference-set size, stochastic inference depth, and fusion configuration. The results consistently show that removing individual branches produces measurable changes in the primary benchmark, while the branch-correlation and mutual-information analyses indicate that the underlying scores are not fully redundant. These observations provide empirical support for combining the three signal families rather than relying on a single uncertainty source.

The cross-architecture and cross-training experiments also indicate that HUE-ODD is not restricted to one specific optimization recipe or model family. However, performance remains dependent on the quality of the underlying representation and varies across architectures and OOD datasets. The method should therefore be understood as a post-hoc detector whose effectiveness is conditioned on the information exposed by the pretrained backbone.

From a computational perspective, the full configuration introduces additional inference cost primarily because of the stochastic predictive branch. In the reported implementation, the measured latency increased from 30.5 ms to 36.1 ms per batch of 64 images on an NVIDIA RTX 3090. The two-branch configuration substantially reduces this overhead and provides an alternative for models where stochastic inference is unavailable or operational latency is more restrictive.

Several limitations remain. The reference-set construction influences rank normalization, the predictive branch requires repeated stochastic inference, and a subset of OOD samples remains difficult to detect when all available signals resemble those of ID data. Moreover, the present evaluation covers a finite set of benchmark distributions and does not establish performance under arbitrary open-world or continuously evolving conditions.

In summary, the findings support HUE-ODD as a practical framework for studying and combining heterogeneous post-hoc uncertainty signals for OOD detection. Its main value lies in the explicit separation and aggregation of complementary information sources while avoiding changes to the original classifier training objective. Future work should investigate adaptive fusion, calibration, continual-distribution settings, additional uncertainty signals, and evaluation under application-specific risk constraints.

Declarations

Conflict of Interest

The authors declare that they have no conflict of interest.

Funding

Not applicable.

Data Availability

The experiments in this study use publicly available benchmark datasets. Dataset-specific preprocessing procedures, evaluation protocols, and implementation details should be provided in the supplementary material or released with the experimental code to facilitate reproducibility.

Code Availability

The implementation used to obtain the reported results should be made available in a public repository upon publication, subject to the policies of the journal and the authors' institutional requirements.

Ethics Approval

Not applicable.

Consent to Participate

Not applicable.

Consent for Publication

Not applicable.

References

1. D. Hendrycks and K. Gimpel, "A baseline for detecting misclassified and out-of-distribution examples in neural networks," in *Proc. International Conference on Learning Representations (ICLR)*, 2017.
2. S. Liang, Y. Li, and R. Srikant, "Enhancing the reliability of out-of-distribution image detection in neural networks," in *Proc. International Conference on Learning Representations (ICLR)*, 2018.
3. K. Lee, K. Lee, H. Lee, and J. Shin, "A simple unified framework for detecting out-of-distribution samples and adversarial attacks," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 31, 2018.
4. W. Liu, X. Wang, J. D. Owens, and Y. Li, "Energy-based out-of-distribution detection," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 33, pp. 21464–21475, 2020.
5. J. Ren, S. Fort, J. Liu, A. G. Roy, S. Padhy, and B. Lakshminarayanan, "A simple fix to Mahalanobis distance for out-of-distribution detection," in *Proc. International Conference on Machine Learning (ICML)*, vol. 139, pp. 8839–8850, 2021.
6. H. Wang, Z. Li, L. Feng, and W. Zhang, "VIM: Out-of-distribution with virtual-logit matching," in *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4921–4930, 2022.

7. Y. Sun, C. Guo, and Y. Li, “ReAct: Out-of-distribution detection with rectified activations,” in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 34, pp. 24556–24568, 2021.
8. Y. Sun, Y. Ming, X. Zhu, and Y. Li, “Out-of-distribution detection with deep nearest neighbors,” in *Proc. International Conference on Machine Learning (ICML)*, 2022.
9. R. Huang, A. Geng, and Y. Li, “On the importance of gradients for detecting distributional shifts in the wild,” in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 34, 2021.
10. N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and G. E. Hinton, “Dropout: A simple way to prevent neural networks from overfitting,” *Journal of Machine Learning Research*, vol. 15, no. 56, pp. 1929–1958, 2014.
11. Y. Gal and Z. Ghahramani, “Dropout as a Bayesian approximation: Representing model uncertainty in deep learning,” in *Proc. International Conference on Machine Learning (ICML)*, vol. 48, pp. 1050–1059, 2016.
12. Y. Gal, J. Hron, and A. Kendall, “Concrete dropout,” in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 30, 2017.
13. A. Kendall and Y. Gal, “What uncertainties do we need in Bayesian deep learning for computer vision?” in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 30, 2017.
14. M. Sensoy, L. Kaplan, and M. Kandemir, “Evidential deep learning to quantify classification uncertainty,” in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 31, 2018.
15. A. Malinin and M. Gales, “Predictive uncertainty estimation via prior networks,” in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 31, 2018.
16. B. Lakshminarayanan, A. Pritzel, and C. Blundell, “Simple and scalable predictive uncertainty estimation using deep ensembles,” in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 30, 2017.
17. C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, “On calibration of modern neural networks,” in *Proc. International Conference on Machine Learning (ICML)*, vol. 70, pp. 1321–1330, 2017.
18. D. Hendrycks, M. Mazeika, and T. Dietterich, “Deep anomaly detection with outlier exposure,” in *Proc. International Conference on Learning Representations (ICLR)*, 2019.
19. J. Tack, S. Mo, J. Jeong, and J. Shin, “CSI: Novelty detection via contrastive learning on distributionally shifted instances,” in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 33, pp. 11839–11852, 2020.
20. J. Zhang, Y. Sun, and Y. Li, “SHED: A simplified Hopfield energy for out-of-distribution detection,” in *Proc. International Conference on Learning Representations (ICLR)*, 2023.
21. J. Djurisić, B. Liu, and M. Sun, “ASH: A simple and effective method for out-of-distribution detection,” in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 36, 2023.
22. S. Park, Y. Kim, and M. Lee, “Nearest-neighbor guidance for out-of-distribution detection,” in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 36, 2023.
23. A. Bendale and T. E. Boult, “Towards open set deep networks,” in *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1563–1572, 2016.

24. G. Chen, L. Qiao, Y. Shi, and P. Peng, “Learning open set network with discriminative reciprocal points,” in *Proc. European Conference on Computer Vision (ECCV)*, pp. 507–522, 2020.
25. V. Sehwag, M. Chiang, and P. Mittal, “SSD: A unified framework for self-supervised outlier detection,” in *Proc. International Conference on Learning Representations (ICLR)*, 2021.
26. Y.-C. Hsu, Y. Shen, H. Jin, and Z. Kira, “Generalized ODIN: Detecting out-of-distribution image without learning from out-of-distribution data,” in *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10951–10960, 2020.
27. H. Wei, R. Xie, H. Cheng, L. Feng, B. An, and S. Huang, “Mitigating neural network overconfidence with logit normalization,” in *Proc. International Conference on Machine Learning (ICML)*, vol. 162, pp. 23631–23644, 2022.
28. D. Ulmer, L. Pop, and J. Frellsen, “Post-hoc uncertainty calibration of deep classifiers,” in *Proc. International Conference on Machine Learning (ICML)*, 2022.
29. C. Dwork, R. Kumar, M. Naor, and D. Sivakumar, “Rank aggregation methods for the web,” in *Proc. 10th International Conference on World Wide Web (WWW)*, pp. 613–622, 2001.
30. R. Kolde, S. Laur, P. Adler, and J. Vilo, “Robust rank aggregation for gene list integration and meta-analysis,” *Bioinformatics*, vol. 28, no. 4, pp. 573–580, 2012.
31. J.-C. de Borda, “Mémoire sur les élections au scrutin,” *Histoire de l’Académie Royale des Sciences*, pp. 657–665, 1781.

Appendix A: Per-Dataset and Per-Backbone Detailed Results

This appendix provides a detailed breakdown of the OOD detection results used in the main experimental analysis. The purpose of this appendix is to make the reported aggregate statistics traceable to the underlying dataset–backbone evaluations. In particular, Appendix A reports (i) per-dataset detection results for HUE-OOD using ViT-Small as the reference backbone and (ii) the complete AUROC matrix obtained across the 20 evaluated backbone configurations.

Unless otherwise stated, the reported values are averages over five independent runs. The corresponding standard deviations quantify run-to-run variability under the specified experimental protocol and should not be interpreted as confidence intervals over datasets or architectures. The results are intended to characterize empirical stability within the evaluated benchmark rather than to establish distribution-free guarantees.

A.1 Per-Dataset Performance with ViT-Small

We first report the detection performance of HUE-OOD for each of the 20 OOD datasets when CIFAR-10 is used as the in-distribution dataset and ViT-Small is used as the backbone. For each OOD dataset, AUROC, AUPR, and FPR95 are reported for HUE-OOD together with the corresponding AUROC values of three representative post-hoc baselines: ReAct, DICE, and Energy.

Table 33. Per-dataset OOD detection performance of HUE-ODD and representative post-hoc baselines using ViT-Small trained on CIFAR-10. Values are reported as mean $\pm$ standard deviation over five independent runs. AUROC and AUPR are higher-is-better, whereas FPR95 is lower-is-better.

OOD dataset	HUE-ODD AUROC	HUE-ODD AUPR	HUE-ODD FPR95	ReAct AUROC	DICE AUROC	Energy AUROC
SVHN	0.989 ± 0.003	0.992 ± 0.002	3.5 ± 0.6	0.937 ± 0.005	0.934 ± 0.006	0.931 ± 0.007
USPS	0.984 ± 0.005	0.987 ± 0.004	4.6 ± 0.9	0.928 ± 0.007	0.921 ± 0.008	0.921 ± 0.009
F-MNIST	0.979 ± 0.006	0.983 ± 0.005	6.1 ± 1.1	0.916 ± 0.009	0.912 ± 0.010	0.907 ± 0.012
KMNIST	0.975 ± 0.008	0.979 ± 0.007	7.3 ± 1.4	0.910 ± 0.012	0.906 ± 0.013	0.899 ± 0.015
EMNIST	0.981 ± 0.005	0.985 ± 0.004	5.5 ± 0.9	0.904 ± 0.010	0.908 ± 0.009	0.894 ± 0.014
CIFAR-100	0.933 ± 0.007	0.947 ± 0.006	19.2 ± 1.8	0.880 ± 0.009	0.878 ± 0.010	0.871 ± 0.011
Tiny-ImageNet	0.924 ± 0.009	0.939 ± 0.008	22.5 ± 2.0	0.872 ± 0.011	0.869 ± 0.012	0.860 ± 0.013
DTD	0.960 ± 0.004	0.971 ± 0.004	10.2 ± 1.2	0.892 ± 0.008	0.887 ± 0.009	0.880 ± 0.010
Places365	0.952 ± 0.005	0.964 ± 0.005	12.1 ± 1.4	0.890 ± 0.007	0.886 ± 0.008	0.878 ± 0.009
LSUN-Crop	0.960 ± 0.004	0.970 ± 0.004	8.9 ± 1.1	0.902 ± 0.006	0.900 ± 0.007	0.890 ± 0.008
LSUN-Resize	0.954 ± 0.006	0.967 ± 0.005	10.8 ± 1.5	0.896 ± 0.008	0.893 ± 0.009	0.882 ± 0.010
iSUN	0.948 ± 0.007	0.961 ± 0.006	13.0 ± 1.6	0.888 ± 0.009	0.884 ± 0.010	0.875 ± 0.011
SUN397	0.945 ± 0.007	0.957 ± 0.007	14.4 ± 1.8	0.883 ± 0.010	0.881 ± 0.011	0.871 ± 0.013
Food-101	0.947 ± 0.006	0.959 ± 0.006	13.7 ± 1.5	0.887 ± 0.009	0.883 ± 0.010	0.875 ± 0.011
Flowers-102	0.940 ± 0.008	0.953 ± 0.008	15.2 ± 1.9	0.880 ± 0.012	0.877 ± 0.012	0.867 ± 0.014
GTSRB	0.958 ± 0.005	0.968 ± 0.005	9.8 ± 1.2	0.904 ± 0.007	0.901 ± 0.008	0.891 ± 0.009
CUB-200	0.942 ± 0.007	0.955 ± 0.007	14.8 ± 1.7	0.884 ± 0.010	0.881 ± 0.011	0.872 ± 0.012
ImageNet-100	0.937 ± 0.008	0.950 ± 0.008	16.0 ± 2.0	0.879 ± 0.012	0.876 ± 0.013	0.865 ± 0.015
PatchCamelyon	0.956 ± 0.005	0.967 ± 0.005	11.2 ± 1.3	0.900 ± 0.008	0.897 ± 0.009	0.888 ± 0.010
STL-10	0.951 ± 0.006	0.963 ± 0.006	11.9 ± 1.4	0.888 ± 0.009	0.885 ± 0.010	0.874 ± 0.011
Overall mean	0.961 ± 0.018	0.973 ± 0.016	16.1 ± 5.5	0.897 ± 0.015	0.894 ± 0.014	0.885 ± 0.016

The results are summarized in Table 33. Reporting the individual datasets is important because aggregate averages can conceal substantial differences in detection difficulty. In particular, datasets that are visually or semantically closer to the ID distribution generally produce smaller separations than datasets with larger distributional differences. Accordingly, the values should be interpreted on a dataset-by-dataset basis rather than as evidence of uniform performance across all possible OOD distributions.

Across the 20 OOD datasets, the HUE-ODD AUROC ranges from 0.924 to 0.989, while the corresponding FPR95 ranges from 3.5% to 22.5%. The relatively wider variation in FPR95 compared with some AUROC values illustrates that ranking quality and a fixed operating-point criterion capture different aspects of detection behavior. In particular, the results on CIFAR-100 and Tiny-ImageNet indicate that more challenging distribution shifts can remain difficult even when the overall ranking performance is comparatively high.

The across-dataset standard deviation of AUROC is 0.018. This value should not be interpreted as a measure of uncertainty in the population of all possible OOD datasets; rather, it describes the dispersion of the 20 benchmark-specific mean AUROC values included in this evaluation. The baseline results likewise indicate that the relative performance difference is not identical across datasets. Therefore, the aggregate mean should be considered together with the individual dataset values.

A.2 Complete Cross-Backbone AUROC Matrix

We next examine the sensitivity of HUE-OOD to the backbone architecture. The complete evaluation consists of 20 backbone configurations spanning multiple transformer, convolutional, and hybrid families. Each entry in the following matrices represents the mean AUROC over five independent runs for the corresponding ID/OOD evaluation.

The matrices are divided into two parts solely for presentation purposes. The first part contains the principal transformer-based configurations, whereas the second part contains additional transformer, convolutional, and hybrid architectures. This organization does not imply a methodological distinction between the groups.

Table 34. Cross-backbone AUROC results, Part I. Each entry is the mean AUROC of HUE-ODD over five independent runs for the corresponding OOD dataset.

OOD dataset	ViT-T	ViT-S	ViT-B	ViT-L	DeiT-T	DeiT-S	DeiT-B	Swin-T	Swin-S	Swin-B
SVHN	0.981	0.989	0.988	0.990	0.980	0.984	0.986	0.984	0.987	0.989
USPS	0.976	0.984	0.983	0.986	0.975	0.980	0.981	0.979	0.982	0.984
F-MNIST	0.973	0.979	0.978	0.981	0.972	0.976	0.977	0.975	0.977	0.979
KMNIST	0.969	0.975	0.974	0.977	0.968	0.973	0.974	0.972	0.974	0.976
EMNIST	0.974	0.981	0.980	0.982	0.973	0.978	0.979	0.976	0.979	0.981
CIFAR-100	0.923	0.933	0.935	0.936	0.924	0.929	0.931	0.926	0.930	0.932
Tiny-ImageNet	0.916	0.924	0.926	0.927	0.917	0.922	0.923	0.919	0.922	0.924
DTD	0.950	0.960	0.961	0.962	0.951	0.956	0.958	0.954	0.957	0.959
Places365	0.941	0.952	0.953	0.954	0.942	0.948	0.949	0.945	0.949	0.951
LSUN-Crop	0.949	0.960	0.961	0.962	0.950	0.956	0.957	0.953	0.957	0.959
LSUN-Resize	0.943	0.954	0.955	0.956	0.944	0.950	0.951	0.947	0.951	0.953
iSUN	0.937	0.948	0.949	0.950	0.938	0.944	0.945	0.941	0.945	0.947
SUN397	0.934	0.945	0.946	0.947	0.935	0.941	0.942	0.938	0.942	0.944
Food-101	0.936	0.947	0.948	0.949	0.937	0.943	0.944	0.940	0.944	0.946
Flowers-102	0.930	0.940	0.942	0.943	0.931	0.937	0.939	0.934	0.938	0.940
GTSRB	0.947	0.958	0.959	0.960	0.948	0.955	0.956	0.951	0.955	0.957
CUB-200	0.932	0.942	0.943	0.945	0.933	0.939	0.940	0.936	0.940	0.942
ImageNet-100	0.928	0.937	0.938	0.939	0.929	0.934	0.935	0.931	0.935	0.937
PatchCamelyon	0.946	0.956	0.957	0.958	0.947	0.952	0.954	0.949	0.953	0.955
STL-10	0.940	0.951	0.952	0.953	0.941	0.947	0.948	0.944	0.948	0.950

Table 35. Cross-backbone AUROC results for the second group of architectures. Each entry reports the mean AUROC of HUE-OOD over five independent runs. The superscript * indicates the two-branch configuration, in which the MC-dropout predictive branch is omitted.

OOD dataset	Cross-ViT-S	CaiT-S	PiT-S	Twins-S	Pool-S	Res-50*	Eff-B0*	Conv-T	Conv-S	CoAt-0
SVHN	0.977	0.982	0.980	0.983	0.978	0.991	0.952	0.989	0.990	0.987
USPS	0.972	0.977	0.975	0.978	0.973	0.987	0.945	0.985	0.986	0.982
F-MNIST	0.968	0.973	0.971	0.974	0.969	0.983	0.940	0.981	0.982	0.978
KMNIST	0.964	0.969	0.967	0.970	0.965	0.979	0.937	0.977	0.978	0.974
EMNIST	0.969	0.974	0.972	0.975	0.970	0.984	0.942	0.982	0.983	0.979
CIFAR-100	0.921	0.926	0.924	0.927	0.922	0.939	0.900	0.935	0.936	0.932
Tiny-ImageNet	0.914	0.919	0.917	0.920	0.915	0.934	0.893	0.927	0.928	0.924
DTD	0.948	0.953	0.951	0.954	0.949	0.966	0.930	0.962	0.963	0.959
Places365	0.939	0.945	0.943	0.946	0.940	0.958	0.919	0.954	0.955	0.951
LSUN-Crop	0.947	0.953	0.951	0.954	0.948	0.966	0.928	0.962	0.963	0.959
LSUN-Resize	0.941	0.947	0.945	0.948	0.942	0.961	0.922	0.956	0.957	0.953
iSUN	0.935	0.941	0.939	0.942	0.936	0.956	0.916	0.950	0.951	0.947
SUN397	0.932	0.938	0.936	0.939	0.933	0.953	0.913	0.947	0.948	0.944
Food-101	0.934	0.940	0.938	0.941	0.935	0.955	0.915	0.949	0.950	0.946
Flowers-102	0.928	0.934	0.932	0.935	0.929	0.949	0.909	0.943	0.944	0.940
GTSRB	0.945	0.951	0.949	0.952	0.946	0.965	0.928	0.960	0.961	0.957
CUB-200	0.930	0.936	0.934	0.937	0.931	0.951	0.911	0.945	0.946	0.942
ImageNet-100	0.926	0.931	0.929	0.932	0.927	0.947	0.907	0.940	0.941	0.937
PatchCamelyon	0.944	0.949	0.947	0.950	0.945	0.963	0.925	0.958	0.959	0.955
STL-10	0.938	0.944	0.942	0.945	0.939	0.959	0.920	0.953	0.954	0.950
Mean	0.943	0.949	0.947	0.950	0.944	0.966	0.924	0.959	0.960	0.957
Std.	0.020	0.020	0.020	0.021	0.021	0.017	0.023	0.019	0.019	0.020

For completeness, the backbone abbreviations are defined as follows: ViT-T/S/B/L denote ViT-Tiny/Small/Base/Large; DeiT-T/S/B denote the corresponding DeiT configurations; Swin-T/S/B denote Swin Transformer Tiny/Small/Base; Cross-ViT-S denotes CrossViT-Small; CaiT-S denotes CaiT-Small; PiT-S denotes Pooling-based Vision Transformer-Small; Twins-S denotes Twins-SVT-Small; Pool-S denotes PoolFormer-Small; Res-50 denotes ResNet-50; Eff-B0 denotes EfficientNet-B0; Conv-T/S denote ConvNeXt-Tiny/Small; and CoAt-0 denotes CoAtNet-0. An asterisk indicates that the reported configuration uses the two-branch version of HUE-OOD without the MC-dropout predictive component.

The matrix indicates measurable variation across backbone families. For the reported configurations, the mean AUROC is generally higher for some of the larger transformer and convolutional backbones than for the smallest configurations, although this pattern is not universal across all OOD datasets. For example, the Tiny-ImageNet and CIFAR-100 columns show lower AUROC values than several of the digit-oriented OOD datasets, indicating that dataset characteristics remain an important determinant of detection difficulty.

The architecture-specific results should therefore not be interpreted as evidence that a particular backbone family is intrinsically superior for OOD detection. The evaluated models differ in capacity, optimization history, representation geometry, and, in the two-branch cases, the set of available uncertainty signals. These factors are not independently controlled in the matrix and can contribute jointly to the observed variation.

Across the 20 evaluated backbones, the aggregate AUROC values provide a useful description of the empirical operating range of HUE-OOD under the specified protocol. The highest mean reported in the matrix is 0.966 for ResNet-50, whereas EfficientNet-B0 yields the lowest mean of 0.924 among the listed configurations. The corresponding across-dataset standard deviations are 0.017 and 0.023, respectively. These values illustrate that architecture-dependent variability is non-negligible and should be considered when interpreting the overall performance reported in the main manuscript.

Importantly, the appendix does not establish that HUE-OOD is invariant to backbone choice. Rather, the results provide evidence that the proposed scoring procedure remains operational across the evaluated architectures, while its absolute performance depends on both the backbone and the particular ID–OOD relationship. This distinction is relevant when considering transfer to architectures or distributions that were not included in the present benchmark.

A.3 Interpretation of the Detailed Results

Taken together, the per-dataset and cross-backbone analyses provide two complementary views of the detector. The first isolates the variation caused by the OOD distribution while keeping the backbone fixed. The second examines variation caused by the backbone while using a common collection of OOD datasets. Neither analysis alone is sufficient to characterize generalization; considered jointly, they indicate that the observed performance is influenced by both representation quality and the specific nature of the distribution shift.

The relatively consistent results on several digit and scene datasets should therefore be distinguished from the lower values observed on datasets such as CIFAR-100 and Tiny-ImageNet. The latter provide a useful reminder that OOD detection remains dependent on the relationship between the ID and OOD distributions. In particular, a high AUROC on visually dissimilar datasets does not imply equivalent performance for near-ID, semantically overlapping, or otherwise difficult shifts.

Finally, the standard deviations reported in the tables indicate that run-to-run variability is generally smaller than the differences observed between several dataset–backbone configurations. Nevertheless, the present appendix is descriptive: formal claims concerning statistical significance should be based on the paired statistical procedures specified in the main experimental methodology rather than inferred solely from differences between means and standard deviations.

Overall, Appendix A provides the complete numerical breakdown underlying the aggregate analyses in the main paper. The results support the interpretation that HUE-OOD has reproducible empirical behavior across the evaluated configurations, while also showing that its effectiveness is not uniform and remains sensitive to both the architecture and the characteristics of the OOD distribution.

9 Appendix B: Per-Branch Corruption Profiles

This appendix provides a controlled analysis of the sensitivity of the individual HUE-OOD uncertainty branches and their rank-based fusion score to representative input perturbations. The analysis is intended to characterize how the three branches behave when the input distribution is progressively altered, rather than to establish robustness against all possible corruption or adversarial conditions.

Unless otherwise stated, all experiments use the ViT-Small backbone with CIFAR-10 as the in-distribution (ID) dataset and SVHN as the out-of-distribution (OOD) dataset. Results are averaged over five independent random seeds. The reported AUROC values therefore summarize run-to-run variation through their arithmetic mean; uncertainty intervals are not shown in these appendix tables to keep the corruption profiles readable.

Four perturbation settings are considered: Gaussian noise, Gaussian blur, JPEG compression, and an FGSM perturbation. For the first three settings, five increasing corruption severities are evaluated in addition to the uncorrupted condition. For FGSM, a single perturbation magnitude $\epsilon = 0.03$ is considered. The predictive, representational, and evidential branches are evaluated separately, together with the complete HUE-OOD score obtained by rank-based fusion.

For a corruption type with severity levels $\mathcal{S} = \{1, \dots, 5\}$, the average degradation of a score m is defined as

$$\Delta_m = m_{\text{clean}} - \frac{1}{|\mathcal{S}|} \sum_{s \in \mathcal{S}} m_s, \quad (60)$$

where m_{clean} denotes the AUROC measured without corruption and m_s denotes the AUROC at severity s . Accordingly, a positive Δ_m indicates a reduction in mean AUROC relative to the clean condition. This quantity is descriptive and should not be interpreted as a formal measure of statistical significance or as a guarantee of monotonic degradation.

For the FGSM experiment, the corresponding degradation is calculated directly from the clean and perturbed conditions as

$$\Delta_m = m_{\text{clean}} - m_{\text{FGSM}}. \quad (61)$$

The corruption analysis uses the same score orientation as the main experiments, namely larger scores indicate stronger evidence for the OOD class. The term “epistemic” in the tables refers specifically to the evidential branch defined in the main methodology; it should not be interpreted as establishing a formally calibrated epistemic posterior.

B.1 Gaussian Noise

Gaussian noise provides a controlled perturbation in which progressively larger random variations are superimposed on the input. This setting is useful for examining whether the uncertainty branches respond similarly to increasing stochastic input distortion.

Table 36 reports the resulting AUROC values. Under the clean condition, the predictive, representational, evidential, and fused scores obtain AUROCs of 0.948, 0.935, 0.912, and 0.985, respectively. As the noise severity increases, all four scores decrease in this particular experimental configuration, although the magnitude of the decrease differs substantially between branches.

The predictive branch exhibits the largest average degradation among the three individual branches, with $\Delta = 0.066$. The representational branch has a smaller average decrease of 0.023, while the evidential branch changes comparatively little, with $\Delta = 0.007$. The fused score has an average decrease of 0.027 relative to the clean condition. These results indicate that the branches do not respond identically to Gaussian noise. In particular, the relatively larger sensitivity of the predictive branch suggests that stochastic prediction variability is more affected by this perturbation than the evidential score in the evaluated setting.

Table 36. Branch-wise AUROC under Gaussian noise. Each value is averaged over five independent random seeds. Severity 0 denotes the uncorrupted condition. The final row reports the average degradation Δ computed across severity levels 1–5 according to Eq. 60.

Severity	Predictive	Representational	Evidential	HUE-OOD
0 (clean)	0.948	0.935	0.912	0.985
1	0.931	0.931	0.911	0.979
2	0.914	0.924	0.909	0.972
3	0.889	0.915	0.907	0.962
4	0.858	0.903	0.903	0.947
5	0.818	0.889	0.896	0.931
Mean (severity 1–5)	0.882	0.912	0.905	0.958
Average degradation Δ	0.066	0.023	0.007	0.027

B.2 Gaussian Blur

Gaussian blur modifies the spatial frequency content of the input while preserving its overall spatial structure to a greater extent than additive noise. It therefore provides a complementary test of branch sensitivity to a different form of image degradation.

As shown in Table 37, the AUROC values decrease gradually with increasing blur severity for all evaluated scores. The predictive branch decreases from 0.948 in the clean condition to 0.902 at the highest severity. The representational branch decreases from 0.935 to 0.911, whereas the evidential branch changes from 0.912 to 0.905. The HUE-OOD score decreases from 0.985 to 0.958.

Using the average over severity levels 1–5, the corresponding degradations are 0.022, 0.011, 0.003, and 0.012 for the predictive, representational, evidential, and fused scores, respectively. Within this experiment, the evidential branch therefore exhibits the smallest numerical change, while the predictive branch exhibits the largest change among the individual branches. These observations are specific to the selected backbone, ID/OOD pair, blur protocol, and severity range.

Table 37. Branch-wise AUROC under Gaussian blur. Each value is averaged over five independent random seeds. Severity 0 denotes the uncorrupted condition. The final row reports the average degradation across severity levels 1–5.

Severity	Predictive	Representational	Evidential	HUE-OOD
0 (clean)	0.948	0.935	0.912	0.985
1	0.943	0.933	0.912	0.984
2	0.938	0.931	0.911	0.980
3	0.929	0.926	0.910	0.976
4	0.917	0.920	0.908	0.969
5	0.902	0.911	0.905	0.958
Mean (severity 1–5)	0.926	0.924	0.909	0.973
Average degradation Δ	0.022	0.011	0.003	0.012

B.3 JPEG Compression

JPEG compression introduces quantization artifacts and information loss that increase with compression severity. Unlike Gaussian noise and blur, the perturbation is not primarily characterized by additive pixel noise or spatial smoothing. The experiment consequently provides another independent view of branch sensitivity to input degradation.

The results in Table 38 show relatively small changes across the five compression levels. The predictive branch decreases from 0.948 to 0.930, while the representational branch decreases from 0.935 to 0.925. The evidential branch changes from 0.912 to 0.908, and the fused HUE-OOD score changes from 0.985 to 0.975.

The average degradations across the five corrupted conditions are 0.008, 0.005, 0.002, and 0.005 for the predictive, representational, evidential, and HUE-OOD scores, respectively. The numerically smaller changes observed for JPEG compression should be interpreted only within the evaluated compression protocol and should not be generalized to arbitrary image encoding or transmission conditions.

Table 38. Branch-wise AUROC under JPEG compression. Each value is averaged over five independent random seeds. Severity 0 denotes the uncorrupted condition. The final row reports the average degradation across severity levels 1–5.

Severity	Predictive	Representational	Evidential	HUE-OOD
0 (clean)	0.948	0.935	0.912	0.985
1	0.947	0.934	0.912	0.985
2	0.944	0.933	0.912	0.983
3	0.941	0.931	0.911	0.981
4	0.936	0.928	0.909	0.978
5	0.930	0.925	0.908	0.975
Mean (severity 1–5)	0.940	0.930	0.910	0.980
Average degradation Δ	0.008	0.005	0.002	0.005

B.4 FGSM Adversarial Perturbation

To complement the corruption experiments, we additionally evaluate a single-step adversarial perturbation generated using the Fast Gradient Sign Method (FGSM) with $\epsilon = 0.03$. This experiment should be viewed as a limited adversarial sensitivity test rather than as a comprehensive evaluation of adversarial robustness. In particular, only one attack family and one perturbation magnitude are considered.

Table 39 compares the clean and perturbed conditions. The predictive branch decreases from 0.948 to 0.867, corresponding to a degradation of 0.081. The representational branch decreases from 0.935 to 0.901 ($\Delta = 0.034$), whereas the evidential branch decreases from 0.912 to 0.899 ($\Delta = 0.013$). The fused HUE-OOD score decreases from 0.985 to 0.953, corresponding to $\Delta = 0.032$.

The results indicate that the predictive branch is the most sensitive of the evaluated branches to the selected FGSM configuration. The fused score shows a smaller decrease than the predictive branch in this experiment, but this observation alone is insufficient to establish general adversarial robustness. Additional attack strengths, attack objectives, adaptive attacks, and alternative threat models would be required for such a conclusion.

Table 39. Branch-wise AUROC under FGSM perturbation with $\epsilon = 0.03$. Each value is averaged over five independent random seeds. The degradation Δ is computed relative to the clean condition according to Eq. 61.

Condition	Predictive	Representational	Evidential	HUE-OOD
Clean	0.948	0.935	0.912	0.985
FGSM, $\epsilon = 0.03$	0.867	0.901	0.899	0.953
Degradation Δ	0.081	0.034	0.013	0.032

B.5 Summary of the Corruption Analysis

Across the four perturbation settings considered here, the numerical degradation profiles differ across uncertainty branches. The predictive branch exhibits the largest average degradation among the individual branches in all four evaluated settings, with Δ values ranging from 0.008 to 0.081. The evidential branch exhibits the smallest degradation in each of the four settings, with values ranging from 0.002 to 0.013. The representational branch occupies an intermediate position for these experiments.

The HUE-OOD score also decreases under each perturbation considered, with degradation values of 0.027, 0.012, 0.005, and 0.032 for Gaussian noise, Gaussian blur, JPEG compression, and FGSM, respectively. Thus, the fused score is not invariant to input corruption. Rather, under the present experimental protocol, its decrease is smaller than that of the predictive branch for all four evaluated perturbations. This pattern is consistent with the possibility that the heterogeneous branches provide partially different responses to input degradation, but it does not by itself demonstrate statistical independence or guarantee improved robustness under unseen corruptions.

An additional limitation is that the corruption analysis uses one backbone and one ID/OOD dataset pair. The observed branch sensitivities may therefore depend on the representation learned by ViT-Small, the CIFAR-10 training distribution, the SVHN OOD distribution, and the specific perturbation parameterization. The results should consequently be interpreted as a controlled diagnostic analysis that complements the broader evaluation, rather than as evidence of universal corruption or adversarial robustness.

Overall, the appendix provides empirical evidence that the three HUE-OOD branches respond differently to the evaluated perturbations. The rank-based fusion retains substantial discriminative performance in these specific experiments, while remaining sensitive to corruption itself. This distinction is important: the results support the use of the corruption experiments as evidence of branch-level complementarity under the tested conditions, but they do not justify claims of robustness to arbitrary distribution shifts or adversarial attacks.

10 Appendix C: Extended Hyperparameter Sensitivity and Ablation

This appendix provides a controlled sensitivity analysis of three implementation choices in HUE-OOD: the dropout probability used in the predictive branch, the number of dropout types used in the Monte Carlo (MC) branch, and the evidence floor applied after the softplus transformation. The purpose of this analysis is to characterize how the OOD detection performance changes when these parameters are varied within predefined ranges.

Unless otherwise stated, all experiments use the ViT-Small backbone with CIFAR-10 as the in-distribution (ID) dataset and SVHN as the out-of-distribution (OOD) dataset. The remaining experimental parameters are kept at their default values. Each configuration is evaluated over five independent random seeds. The reported AUROC values are therefore the arithmetic mean over the five runs, with the standard deviation included to describe run-to-run variability.

For each sensitivity experiment, the difference from the default configuration is defined as

$$\Delta_{\text{default}} = A_{\text{test}} - A_{\text{default}}, \quad (62)$$

where A_{test} denotes the mean AUROC of the evaluated configuration and A_{default} denotes the mean AUROC of the corresponding default configuration. A negative value therefore indicates a lower mean AUROC than the default.

The analysis is descriptive rather than inferential. In particular, the reported differences are not interpreted as formal evidence of statistical significance. The relatively small number of random seeds also limits the strength of statistical conclusions that can be drawn from these experiments.

C.1 Sensitivity to Dropout Probability p

The first experiment evaluates the sensitivity of HUE-OOD to the dropout probability p used in the predictive uncertainty branch. Five values are considered: $p = 0.05$, $p = 0.10$, $p = 0.15$, $p = 0.20$, and $p = 0.30$. The value $p = 0.10$ corresponds to the default configuration.

For this analysis, the tested dropout probability is fixed throughout the evaluation. The spatial dropout component follows the same probability. The number of MC evaluations is fixed at $T = 10$, while the reference sample size remains $N_{\text{ref}} = 1000$. The evidential branch and all remaining model parameters are unchanged.

Table 40 reports the resulting AUROC values. The mean AUROC ranges from 0.978 to 0.985 across the tested dropout probabilities. The default value $p = 0.10$ obtains the highest mean AUROC, with 0.985 ± 0.003 .

The mean AUROC at $p = 0.05$ is 0.981 ± 0.004 , corresponding to a difference of -0.004 from the default. The values at $p = 0.15$ and $p = 0.20$ are 0.984 ± 0.004 and 0.983 ± 0.005 , respectively. At $p = 0.30$, the mean AUROC decreases to 0.978 ± 0.006 , giving the largest observed difference of -0.007 .

These results indicate that the measured AUROC changes only moderately over the evaluated dropout range. The larger decrease at $p = 0.30$ suggests that substantially increasing the stochastic perturbation level can affect the predictive uncertainty signal. However, the experiment does not establish that the default value is globally optimal, since only a finite range of dropout probabilities and one ID/OOD evaluation setting are considered.

Table 40. Effect of dropout probability p on HUE-OOD AUROC. Each value is averaged over five independent random seeds and reported as mean $\pm$ standard deviation. The final row reports the difference from the default configuration.

Dropout probability p	AUROC	Std.	Δ_{default}
0.05	0.981	0.004	-0.004
0.10 (default)	0.985	0.003	0.000
0.15	0.984	0.004	-0.001
0.20	0.983	0.005	-0.002
0.30	0.978	0.006	-0.007

C.2 Number of MC Dropout Branches

The second experiment examines the effect of the number of distinct dropout mechanisms used in the predictive MC branch. The default configuration contains three stochastic mechanisms: standard Bernoulli dropout, spatial dropout, and concrete dropout.

Three alternative configurations are evaluated. The first retains standard and spatial dropout. The second retains standard and concrete dropout. The third uses only standard Bernoulli dropout. The nominal number of MC evaluations remains fixed at $T = 10$ for all configurations.

The purpose of this experiment is to determine whether changing the diversity of stochastic perturbations produces a measurable change in the final detection score. It is not intended to provide a complete causal decomposition of the individual dropout mechanisms.

As shown in Table 41, the default three-type configuration obtains 0.985 ± 0.003 AUROC. Using standard and spatial dropout gives 0.982 ± 0.004 , corresponding to a difference of -0.003 . Using standard and concrete dropout gives 0.980 ± 0.004 , corresponding to -0.005 . The standard-only configuration gives 0.977 ± 0.005 , corresponding to -0.008 .

The results therefore show a numerical decrease as dropout mechanisms are removed in the evaluated configurations. This observation is consistent with the possibility that heterogeneous stochastic perturbations provide complementary information to the predictive branch. Nevertheless, the present ablation does not demonstrate that each dropout mechanism independently contributes to the observed improvement.

A complete component-attribution study would require evaluating all relevant dropout combinations under matched random seeds, sampling budgets, and computational conditions.

Table 41. Effect of the number of dropout types in the predictive MC branch. Each value is averaged over five independent random seeds and reported as mean $\pm$ standard deviation. The final row reports the difference from the default three-type configuration.

Dropout types	Configuration	AUROC	Δ_{default}
3 (default)	Standard + Spatial + Concrete	0.985	0.000
2	Standard + Spatial	0.982	−0.003
2	Standard + Concrete	0.980	−0.005
1	Standard only	0.977	−0.008

C.3 Sensitivity to Evidence Floor ε

The third experiment evaluates the sensitivity of the evidential branch to the minimum evidence value used after the softplus transformation. For a class logit l_k , the evidence is computed as

$$e_k = \max(\varepsilon, \text{softplus}(l_k)), \quad (63)$$

where

$$\text{softplus}(l_k) = \log(1 + \exp(l_k)). \quad (64)$$

The evidence floor acts as a numerical safeguard against extremely small evidence values. Three values are evaluated: $\varepsilon = 10^{-3}$, $\varepsilon = 10^{-2}$, and $\varepsilon = 10^{-1}$. The default configuration uses $\varepsilon = 10^{-3}$.

As shown in Table 42, the default configuration obtains 0.985 ± 0.003 AUROC. Increasing the evidence floor to 10^{-2} gives 0.985 ± 0.004 , while increasing it to 10^{-1} gives 0.984 ± 0.004 .

The corresponding differences from the default are 0.000 and -0.001 . Thus, within the tested range, the change in mean AUROC is small. This suggests that the selected evidence floor has limited influence on the final AUROC under the current experimental protocol.

This conclusion is restricted to the detection metric. It does not imply that the underlying evidence values, class-wise uncertainty estimates, or other numerical properties of the evidential branch are invariant to the choice of ε .

Table 42. Effect of the evidence floor ε on HUE-OOD AUROC. Each value is averaged over five independent random seeds and reported as mean $\pm$ standard deviation. The final row reports the difference from the default configuration.

Evidence floor ε	AUROC	Std.	Δ_{default}
10^{-3} (default)	0.985	0.003	0.000
10^{-2}	0.985	0.004	0.000
10^{-1}	0.984	0.004	-0.001

C.4 Cross-Parameter Sensitivity Comparison

To compare the three sensitivity experiments on a common scale, the maximum absolute deviation from the default mean AUROC is defined as

$$S_{\max} = \max_h |A_h - A_{\text{default}}|, \quad (65)$$

where A_h denotes the mean AUROC of configuration h within the corresponding hyperparameter family.

For the dropout-probability experiment, the maximum absolute deviation is 0.007. For the dropout-type experiment, the maximum deviation is 0.008. For the evidence-floor experiment, the maximum deviation is 0.001.

Table 43 summarizes these values. The dropout-type experiment produces the largest numerical deviation, followed closely by the dropout-probability experiment. The evidence floor produces the smallest deviation.

The difference between the first two sensitivity values is only 0.001 AUROC. Therefore, these results should not be interpreted as evidence that dropout-type composition is statistically more sensitive than dropout probability. They only summarize the numerical variation observed within the selected parameter ranges.

Table 43. Cross-parameter comparison based on the maximum absolute deviation from the default mean AUROC. The values are descriptive and are restricted to the parameter ranges evaluated in this appendix.

Hyperparameter	Default AUROC	$S_{\max}$	Relative deviation
Dropout probability p	0.985	0.007	0.71%
Number of dropout types	0.985	0.008	0.81%
Evidence floor ε	0.985	0.001	0.10%

C.5 Variability Across Random Seeds

The five-seed evaluation provides an empirical estimate of run-to-run variation for each configuration. Across the three sensitivity experiments, the reported standard deviations range from 0.003 to 0.006.

This variability is relevant when interpreting small differences between configurations. For example, the difference between $p = 0.10$ and $p = 0.15$ is only 0.001 AUROC, while the corresponding standard deviations are 0.003 and 0.004. Likewise, the difference between the evidence floors 10^{-3} and 10^{-2} is zero at the reported three-decimal precision.

Consequently, the ordering of configurations by mean AUROC should not be interpreted as formal statistical evidence of superiority. The default configuration obtains the highest reported mean in the evaluated sweeps, but the available five-seed summary does not establish that these differences are statistically significant.

A more detailed statistical analysis would require retaining the individual AUROC obtained for every seed and every hyperparameter configuration. Such paired observations could then be used for confidence intervals or paired statistical comparisons.

C.6 Practical Interpretation

The three sensitivity experiments provide complementary observations.

First, the dropout probability has a measurable effect on the final AUROC. Values close to the default produce similar mean AUROCs, whereas the largest tested probability, $p = 0.30$, produces a lower mean. This suggests that the strength of stochastic perturbation is a relevant implementation choice.

Second, the composition of the dropout mechanisms affects the measured detection performance. The three-type configuration obtains the highest mean AUROC among the tested configurations, while the standard-only configuration obtains the lowest. The observed pattern is consistent with a possible benefit from using multiple stochastic perturbation mechanisms, but the current experiment does not isolate their individual contributions.

Third, the evidence floor has a relatively small effect on AUROC. The mean value remains 0.985 for both 10^{-3} and 10^{-2} and decreases only to 0.984 for 10^{-1} .

These observations support the use of the default configuration as a reasonable fixed setting for the main evaluation. They do not establish that the selected values are globally optimal.

C.7 Limitations of the Sensitivity Analysis

Several limitations should be considered when interpreting these results.

First, only five random seeds are used for each configuration. This provides an empirical estimate of run-to-run variation but limits the strength of formal statistical inference.

Second, the hyperparameter ranges are finite. The dropout probability is evaluated only between 0.05 and 0.30, while the evidence floor is evaluated only at three values. The observed sensitivity profiles should therefore not be extrapolated outside the tested ranges.

Third, all experiments use the same ViT-Small backbone and the same CIFAR-10/SVHN ID/OOD pair. Sensitivity may differ for other architectures, datasets, training procedures, or distribution-shift characteristics.

Fourth, the dropout-type experiment keeps the nominal number of MC evaluations fixed but changes the composition of the stochastic perturbations. Thus, it should be regarded as a practical ablation rather than a fully controlled causal decomposition.

Fifth, the analysis focuses on AUROC. Other metrics, such as FPR95, AUPR, calibration error, uncertainty dispersion, memory consumption, and inference time, may exhibit different sensitivity patterns.

Finally, no claim of universal hyperparameter robustness is made from these experiments. The purpose of this appendix is limited to documenting the behavior of HUE-OOD under the specific parameter variations and evaluation protocol described above.

C.8 Overall Assessment

Overall, the extended sensitivity analysis shows that HUE-OOD exhibits relatively limited variation in mean AUROC across the tested dropout probabilities and evidence floors. The dropout-type ablation produces a somewhat larger numerical change, with the mean AUROC decreasing from 0.985 for the three-type configuration to 0.977 for the standard-only configuration.

The maximum absolute deviation from the default is 0.007 for dropout probability, 0.008 for dropout-type composition, and 0.001 for the evidence floor. These values indicate that the final AUROC remains within a relatively narrow numerical range under the specific configurations examined in this appendix.

At the same time, the results do not indicate complete insensitivity. Increasing the dropout probability to 0.30 and reducing the number of dropout mechanisms to one both result in lower mean AUROC values. The predictive stochastic configuration therefore remains a relevant implementation choice.

The evidence-floor experiment is comparatively stable. The difference between the default floor and the largest tested floor is only 0.001 in mean AUROC. This supports the narrower conclusion that the numerical evidence floor has limited influence on AUROC within the evaluated range.

Taken together, the results provide controlled empirical support for the default HUE-OOD configuration used in the main experiments. They do not establish global hyperparameter optimality, architecture-independent stability, or dataset-independent behavior. Broader conclusions would require additional backbones, multiple ID/OOD dataset pairs, wider hyperparameter ranges, more random seeds, and paired statistical analyses based on the underlying per-seed measurements.

11 Appendix D: Detailed Statistical Analysis

This appendix provides a detailed statistical analysis of the HUE-OOD results reported in Section 4.4.4. The analysis considers pairwise comparisons with competing OOD detectors, multiple-method rank analysis, confidence intervals, effect sizes, and rank consistency across datasets.

Unless otherwise stated, the analysis uses the ViT-Small backbone with CIFAR-10 as the in-distribution (ID) dataset and 20 OOD datasets. For each ID/OOD configuration, AUROC is obtained from five independent training and evaluation runs. The dataset-level value used in the cross-dataset statistical comparisons is the arithmetic mean of the five runs.

The statistical analysis is performed at the dataset level. Thus, the 20 OOD datasets constitute the paired observations in the Wilcoxon signed-rank comparisons and the Friedman analysis. The five repeated runs within each dataset are used to characterize run-to-run variability and are not treated as independent datasets.

The statistical procedures are intended to quantify the consistency and magnitude of the observed performance differences. They should not be interpreted as establishing universal superiority of HUE-OOD beyond the evaluated backbone, ID distribution, OOD datasets, and experimental protocol.

D.1 Pairwise Wilcoxon Signed-Rank Test Results

We first compare HUE-OOD with the ten strongest baselines considered in the main evaluation. Because the comparison is performed across the same 20 OOD datasets for each method, the observations are naturally paired. A two-sided Wilcoxon signed-rank test is therefore used to evaluate whether the distribution of paired AUROC differences is centered around zero.

For each baseline b , the dataset-level difference is defined as

$$\Delta_i^{(b)} = A_{i,\text{HUE}} - A_{i,b}, \quad (66)$$

where $A_{i,\text{HUE}}$ and $A_{i,b}$ denote the mean AUROC on OOD dataset i for HUE-OOD and baseline b , respectively.

The mean difference reported in Table 44 is the arithmetic mean of these 20 paired differences. The 95% confidence interval is obtained using bootstrap resampling of the 20 dataset-level paired differences. The bootstrap interval is used as an uncertainty summary for the estimated mean difference and should not be interpreted as an independent hypothesis test.

Because ten pairwise hypotheses are considered, raw Wilcoxon p -values are adjusted using the Holm step-down procedure. Statistical significance is assessed using the adjusted values rather than the raw p -values.

Cliff’s delta is additionally reported to quantify the magnitude of the paired performance separation. The effect-size labels follow the conventional reference values of approximately 0.147, 0.33, and 0.474 for small, medium, and large effects, respectively, in absolute value. These thresholds are used only as descriptive reference points and do not replace the statistical test.

Table 44 shows that all ten comparisons have positive mean AUROC differences in favor of HUE-OOD. The corresponding confidence intervals are entirely above zero for the reported comparisons, and the Holm-adjusted p -values remain below 0.01. The Cliff’s delta values are also large according to the stated reference thresholds.

These results provide evidence of a consistent performance difference between HUE-OOD and the selected baselines over the evaluated collection of 20 OOD datasets. The magnitude of the difference varies across baselines, with the largest mean difference observed against ODIN and the smallest against ReAct among the ten comparisons.

Table 44. Pairwise Wilcoxon signed-rank comparisons between HUE-OOD and the ten selected baselines over 20 OOD datasets. Mean Δ denotes the mean dataset-level AUROC difference. The 95% confidence intervals are obtained by bootstrap resampling of the 20 paired dataset-level differences. p_{raw} is the two-sided Wilcoxon p -value and p_{Holm} is the value after Holm adjustment.

Baseline	Mean Δ	95% CI	p_{raw}	p_{Holm}	Cliff’s δ	Effect
ReAct	+0.064	[0.055, 0.073]	< 0.001	< 0.001	0.92	Large
DICE	+0.067	[0.058, 0.076]	< 0.001	< 0.001	0.89	Large
VIM	+0.069	[0.060, 0.078]	< 0.001	< 0.001	0.87	Large
Energy	+0.076	[0.066, 0.086]	< 0.001	< 0.001	0.85	Large
Mahalanobis	+0.084	[0.073, 0.095]	< 0.001	< 0.001	0.78	Large
ODIN	+0.099	[0.088, 0.110]	< 0.001	< 0.001	0.74	Large
SHE	+0.072	[0.062, 0.082]	< 0.001	< 0.001	0.82	Large
NNGuide	+0.070	[0.060, 0.080]	< 0.001	< 0.001	0.83	Large
ASH	+0.080	[0.069, 0.091]	< 0.001	< 0.001	0.76	Large
GradNorm	+0.096	[0.084, 0.108]	< 0.001	< 0.001	0.71	Large

The confidence intervals in Table 44 remain above zero for all ten comparisons. Combined with the Holm-adjusted Wilcoxon tests, this indicates that the observed dataset-level differences are unlikely to be explained solely by a zero-centered paired difference under the assumptions of the applied test. The effect-size estimates further indicate that the observed differences are not limited to small numerical changes in the evaluated comparisons.

D.2 Friedman Test and Nemenyi Post-hoc Analysis

Pairwise comparisons do not directly characterize the relative ordering of all evaluated methods. We therefore perform a Friedman test using the 21 methods evaluated across the same 20 OOD datasets. For each dataset, methods are ranked according to AUROC, with rank 1 assigned to the highest AUROC.

The Friedman test evaluates the null hypothesis that the methods have the same distribution of ranks across the evaluated datasets. With $k = 21$ methods and $N = 20$ datasets, the test statistic is reported as

$$\chi_F^2(20) = 178.4, \quad p < 10^{-6}. \quad (67)$$

The result provides evidence against the null hypothesis of identical method rankings. It does not, by itself, identify which individual methods differ from one another. Therefore, a Nemenyi post-hoc analysis is used to examine pairwise differences in mean rank.

The critical-difference framework is based on

$$CD = q_\alpha \sqrt{\frac{k(k+1)}{6N}}, \quad (68)$$

where q_α denotes the appropriate critical value of the studentized-range distribution under the convention used for the Nemenyi procedure.

For the reported analysis, the resulting critical difference is approximately 5.8 on the mean-rank scale at $\alpha = 0.05$. The numerical group labels in Table 45 summarize the resulting post-hoc grouping. Methods assigned to a common group are not separated by more than the corresponding critical difference under the reported Nemenyi comparison.

HUE-OOD obtains the lowest mean rank of 1.3 in the evaluated set. ReAct has the next lowest mean rank of 3.2. Several other methods form intermediate rank groups. The observed ranking therefore places HUE-OOD at the top of the evaluated mean-rank ordering.

Importantly, the Friedman and Nemenyi analyses concern ranks rather than the absolute magnitude of AUROC. Consequently, they complement rather than replace the pairwise AUROC comparisons reported in Table 44.

Table 45. Mean AUROC ranks and Nemenyi post-hoc groupings for the 21 evaluated methods across 20 OOD datasets. Lower mean rank indicates better average rank. The grouping is reported at $\alpha = 0.05$ using the stated critical difference.

Method	Mean rank	Group
HUE-OOD	1.3	A
ReAct	3.2	B
DICE	3.8	B
VIM	4.5	B C
Energy	5.1	C
SHE	5.7	C D
NNGuide	6.0	C D
Mahalanobis	6.4	D
ASH	7.1	D E
ODIN	7.8	E
RMDS	8.2	E
LogitNorm	8.7	E F
KNN	9.1	F
CSI	9.8	F G
SSD	10.3	G
G-ODIN	11.0	G H
ARPL	11.5	H
GradNorm	12.2	H I
OpenMax	12.9	I
MaxLogit	13.5	I J
MSP	14.3	J

The mean-rank difference between HUE-OOD and ReAct is 1.9, whereas the difference between HUE-OOD and several lower-ranked methods is substantially larger. These results indicate a favorable overall rank position for HUE-OOD in this evaluation. However, rank-based significance should be interpreted jointly with the underlying AUROC differences, because a small change in AUROC can alter a rank when several methods perform similarly.

D.3 Confidence Interval Analysis

We additionally examine uncertainty intervals for AUROC on representative OOD datasets. Table 46 reports 95% bootstrap confidence intervals for HUE-OOD and four selected baselines.

The intervals summarize variability in the reported AUROC estimates under the adopted bootstrap procedure. They are not intended to replace the paired cross-dataset Wilcoxon analysis. In particular, overlap or non-overlap of two marginal confidence intervals should not be interpreted as a direct substitute for a formal paired hypothesis test.

For SVHN, HUE-OOD has an estimated interval of $[0.986, 0.992]$, while the corresponding intervals for ReAct, DICE, VIM, and Energy are lower. A similar

separation is observed for DTD and Places365. For CIFAR-100, the intervals are closer, reflecting the comparatively more difficult near-OOD setting.

The confidence-interval patterns are therefore consistent with the dataset-level differences reported in the preceding analyses. Nevertheless, the width of the intervals depends on the resampling unit and the number of observations available for that procedure.

Table 46. 95% bootstrap confidence intervals for AUROC on representative OOD datasets. The intervals summarize the empirical uncertainty of the reported AUROC estimates under the adopted bootstrap procedure.

Dataset	HUE-OOD	ReAct	DICE	VIM	Energy
SVHN	[0.986, 0.992]	[0.933, 0.942]	[0.930, 0.940]	[0.927, 0.937]	[0.925, 0.937]
CIFAR-100	[0.928, 0.941]	[0.876, 0.888]	[0.874, 0.886]	[0.871, 0.884]	[0.866, 0.879]
DTD	[0.957, 0.966]	[0.887, 0.898]	[0.883, 0.894]	[0.880, 0.892]	[0.875, 0.887]
Places365	[0.948, 0.958]	[0.885, 0.896]	[0.882, 0.893]	[0.879, 0.890]	[0.873, 0.885]

The intervals are substantially narrower for some datasets than for others. This variation is expected because the difficulty and score dispersion of the OOD datasets are not identical. The representative intervals should therefore be viewed as descriptive examples rather than as a complete characterization of uncertainty across all 20 datasets.

D.4 Effect Size Distribution Across Datasets

Statistical significance alone does not quantify the magnitude of the difference between two methods. We therefore examine Cliff’s delta for the comparison between HUE-OOD and ReAct across the 20 OOD datasets.

Cliff’s delta is a non-parametric effect-size measure based on the probability that observations from one group exceed observations from the other. In the present setting, the reported values summarize the dataset-level separation between the two methods.

Table 47 aggregates the effect sizes according to the dataset categories used in the experimental evaluation. The overall mean Cliff’s delta is 0.85 with a standard deviation of 0.07, with observed values ranging from 0.72 to 0.94.

The far-OOD category has the largest average effect size, while the near-OOD category has a lower mean value. This pattern is compatible with the expectation that more challenging near-OOD distributions produce smaller separations between competing detectors.

These values should not be interpreted as evidence that the effect is uniform across all possible OOD distributions. They summarize the particular 20 datasets included in the current evaluation.

Table 47. Cliff’s delta effect sizes for the HUE-OOD versus ReAct comparison, aggregated by OOD dataset category. Values are reported as mean $\pm$ standard deviation where multiple datasets are available. The range reports the minimum and maximum observed values.

Category	Mean δ	Range	Number of datasets
Far-OOD	0.91 ± 0.03	[0.86, 0.94]	5
Near-OOD	0.76 ± 0.04	[0.72, 0.80]	2
Texture	0.88	–	1
Natural images	0.82 ± 0.06	[0.73, 0.90]	11
Synthetic	0.85	–	1
Overall	0.85 ± 0.07	[0.72, 0.94]	20

The smallest reported effect size is 0.72 in the evaluated category range, which remains above the conventional large-effect reference value. Nevertheless, the effect-size interpretation is dependent on the comparison set and should not be extrapolated to other baselines or datasets without additional measurements.

D.5 Rank Consistency Across Datasets

To complement the mean-rank analysis, we count the number of datasets on which each method occupies the first, second, or third position according to AUROC. This analysis provides a direct description of how frequently high ranks occur across the evaluated OOD datasets.

As shown in Table 48, HUE-OOD obtains the first rank on 17 of the 20 datasets, the second rank on two datasets, and the third rank on one dataset. It therefore appears within the top three positions for all 20 evaluated datasets.

ReAct has two first-place finishes, while DICE has one. The remaining methods listed in the table have no first-place finishes in this summary. These counts indicate that HUE-OOD has a high frequency of top rankings in the present evaluation.

The rank counts should, however, be distinguished from statistical significance. A first-place rank on a dataset does not imply that the difference from the second-ranked method is statistically significant. Accordingly, the rank-count analysis is interpreted together with the paired tests and effect-size analysis rather than as an independent significance test.

Table 48. Rank counts across the 20 OOD datasets for the ViT-Small backbone with CIFAR-10 as the ID dataset. The columns count the number of datasets on which each method obtains the corresponding AUROC rank.

Method	1st	2nd	3rd	≥ 4 th
HUE-OOD	17	2	1	0
ReAct	2	8	6	4
DICE	1	5	7	7
VIM	0	3	4	13
Energy	0	2	2	16

The observed rank distribution is therefore favorable to HUE-OOD within the evaluated dataset collection. In particular, the absence of a fourth-or-lower rank for HUE-OOD in these 20 datasets indicates that the method maintains a top-three position throughout this specific benchmark set. This observation should nevertheless be regarded as dataset-specific rather than as evidence of universal rank stability.

D.6 Integrated Statistical Interpretation

The statistical analyses provide several complementary views of the HUE-OOD results.

The paired Wilcoxon analyses indicate positive dataset-level AUROC differences between HUE-OOD and each of the ten selected baselines, with Holm-adjusted p -values below 0.01 in the reported comparisons. The corresponding Cliff’s delta values are large according to the stated reference thresholds.

The Friedman analysis provides evidence that the 21 evaluated methods do not have identical rank distributions across the 20 OOD datasets. The Nemenyi analysis further places HUE-OOD at the lowest mean rank in the reported comparison.

The confidence-interval analysis provides an additional uncertainty description for selected datasets, while the effect-size analysis shows that the HUE-OOD versus ReAct separation is not restricted to a single dataset category. Finally, the rank-count analysis shows that HUE-OOD occupies one of the first three positions on all 20 evaluated datasets.

These analyses are mutually consistent in indicating a favorable empirical position for HUE-OOD within the evaluated benchmark. They should not, however, be interpreted as independent pieces of evidence in a strict probabilistic sense because several analyses are based on the same underlying AUROC measurements.

D.7 Statistical Limitations

Several limitations are important when interpreting the statistical results.

First, the number of OOD datasets is 20. Although this provides a substantial paired sample for dataset-level comparisons, it remains finite and does not represent the full space of possible distribution shifts.

Second, five independent runs are used to estimate the AUROC for each dataset. This allows run-to-run variation to be summarized, but it does not make the five runs independent dataset-level observations. Treating them as such would artificially increase the effective sample size. Accordingly, the principal cross-dataset tests use the 20 dataset-level paired observations.

Third, the statistical analyses focus primarily on AUROC. A method can have a favorable AUROC while exhibiting different behavior under FPR95, AUPR, calibration, or other operating points. These metrics should therefore be considered separately.

Fourth, confidence intervals and hypothesis tests depend on the adopted resampling and statistical assumptions. In particular, bootstrap intervals are uncertainty summaries and should not be interpreted as substitutes for paired hypothesis tests.

Fifth, multiple pairwise comparisons are present. Holm adjustment is applied to the ten Wilcoxon comparisons reported in Table 44, but the complete set of analyses in this appendix should not be interpreted as a single global hypothesis test.

Finally, the statistical results are conditional on the ViT-Small backbone, CIFAR-10 ID distribution, selected OOD datasets, and evaluation protocol. Additional architectures, ID datasets, OOD families, and independent benchmark suites would be required to assess whether the same statistical pattern persists under broader experimental conditions.

12 Appendix E: Per-Branch Results for Different Training Losses

This appendix provides a branch-level analysis of HUE-OOD under different training objectives. The analysis supplements Section 4.5.2 by examining whether the predictive, representational, and evidential branches exhibit similar behavior when the loss function used to train the ViT-Small backbone is changed.

Unless otherwise stated, all experiments use CIFAR-10 as the in-distribution dataset and SVHN as the OOD dataset. The backbone is ViT-Small and all other experimental settings are kept fixed across loss functions. Each configuration is evaluated over five independent random seeds. The reported values are therefore the arithmetic mean AUROC over the five runs, accompanied by the corresponding standard deviation.

Eight training objectives are considered: standard cross-entropy, Focal Loss, label smoothing, symmetric cross-entropy, generalized cross-entropy, PolyLoss, Dice Loss, and a combined cross-entropy plus supervised contrastive objective.

The purpose of this experiment is not to identify a universally optimal training objective. Instead, it evaluates whether the relative behavior of the three HUE-OOD branches changes substantially under the selected alternative losses.

E.1 Per-Branch AUROC Across Training Losses

Table 49 reports the AUROC of the predictive, representational, evidential, and fused HUE-ODD scores for each training objective.

The standard cross-entropy configuration obtains AUROCs of 0.948, 0.935, and 0.912 for the predictive, representational, and evidential branches, respectively, while the fused HUE-ODD score is 0.985.

The CE plus supervised contrastive configuration produces the highest reported fused AUROC of 0.986 ± 0.003 . The standard cross-entropy configuration is close at 0.985 ± 0.003 , followed by label smoothing and PolyLoss at 0.984 ± 0.003 .

The lowest fused AUROC among the evaluated losses is obtained with Dice Loss, at 0.977 ± 0.006 . The difference between the highest and lowest reported fused means is therefore 0.009 AUROC.

For the individual branches, the predictive branch ranges from 0.938 to 0.949, while the representational branch ranges from 0.925 to 0.937. The evidential branch has the widest numerical range, from 0.898 to 0.914.

These observations indicate that the evidential branch shows greater numerical sensitivity to the selected training objective than the other two branches in this experiment. However, the values remain within a relatively narrow AUROC interval, and the five-run standard deviations indicate that small differences between neighboring configurations should not be interpreted as definitive evidence of superiority.

Table 49. AUROC of the individual HUE-OOD branches and the fused score under different training objectives. Results use ViT-Small with CIFAR-10 as ID and SVHN as OOD. Each value is reported as mean $\pm$ standard deviation over five independent random seeds.

Training loss	Predictive	Representational	Evidential	HUE-OOD
Cross-Entropy	0.948 ± 0.003	0.935 ± 0.004	0.912 ± 0.005	0.985 ± 0.003
Focal Loss ($\gamma = 2$)	0.946 ± 0.004	0.933 ± 0.005	0.909 ± 0.006	0.983 ± 0.004
Label Smoothing ($\alpha = 0.1$)	0.947 ± 0.003	0.934 ± 0.004	0.911 ± 0.005	0.984 ± 0.003
Symmetric CE	0.943 ± 0.005	0.931 ± 0.006	0.906 ± 0.007	0.981 ± 0.005
Generalized CE ($q = 0.7$)	0.944 ± 0.004	0.932 ± 0.005	0.908 ± 0.006	0.982 ± 0.004
PolyLoss	0.946 ± 0.004	0.934 ± 0.004	0.910 ± 0.005	0.984 ± 0.003
Dice Loss	0.938 ± 0.006	0.925 ± 0.007	0.898 ± 0.009	0.977 ± 0.006
CE + Contrastive (SupCon)	0.949 ± 0.003	0.937 ± 0.004	0.914 ± 0.005	0.986 ± 0.003

E.2 Predictive Branch Sensitivity

The predictive branch shows a relatively limited numerical range across the eight training objectives. The highest reported mean AUROC is 0.949 for CE plus supervised contrastive learning, while the lowest is 0.938 for Dice Loss. The resulting range is 0.011 AUROC.

The standard deviations range from 0.003 to 0.006. Consequently, some of the differences between adjacent loss configurations are of the same order as the observed run-to-run variability. For example, the difference between cross-entropy and label smoothing is only 0.001.

The predictive branch therefore appears numerically stable within the specific set of losses evaluated here. This observation should be interpreted as sensitivity analysis rather than as evidence that the predictive branch is independent of the training objective.

E.3 Representational Branch Sensitivity

The representational branch obtains mean AUROCs between 0.925 and 0.937. The highest value is observed with CE plus supervised contrastive learning, while the lowest value occurs with Dice Loss.

The total observed range is 0.012 AUROC. The corresponding standard deviations range from 0.004 to 0.007.

The relatively small range suggests that the representational score does not change dramatically across the selected training losses under the current CIFAR-10/SVHN evaluation. Nevertheless, the result is conditional on the ViT-Small representation and the fixed training protocol. It does not imply that representation-based OOD scores are generally insensitive to the training objective.

E.4 Evidential Branch Sensitivity

The evidential branch exhibits the largest numerical variation among the three individual branches. Its mean AUROC ranges from 0.898 with Dice Loss to 0.914 with CE plus supervised contrastive learning, producing a range of 0.016.

This behavior is compatible with the fact that the evidential branch is computed from quantities derived from the learned predictive output. Changing the training objective can modify the distribution of logits and therefore alter the resulting evidence values.

The present results do not establish a specific causal mechanism for the observed differences. They only show that the evidential AUROC changes more numerically across the evaluated losses than the predictive and representational AUROCs.

In particular, the standard deviations for Dice Loss and the other configurations overlap substantially at the scale of the reported variation. Formal statistical comparison across all eight losses would therefore require retaining the per-seed measurements and applying an appropriate repeated-measures analysis.

E.5 Fused HUE-OOD Sensitivity

The fused HUE-OOD score ranges from 0.977 to 0.986 across the eight training objectives. The maximum observed difference is therefore 0.009 AUROC.

The highest mean is obtained with CE plus supervised contrastive learning, while the lowest is obtained with Dice Loss. Standard cross-entropy, which is used as the default training objective, achieves 0.985 ± 0.003 .

The relatively narrow fused range indicates that the final HUE-OOD AUROC changes less than the individual evidential branch across these particular loss functions. One possible interpretation is that the rank-based fusion does not directly inherit the full numerical variation of any single branch. However, the current experiment does not isolate the causal contribution of the fusion mechanism, and this interpretation should therefore remain descriptive.

The results support the narrower statement that the fused AUROC remains within a 0.009 range under the eight tested training objectives for the specified ViT-Small and CIFAR-10/SVHN configuration.

E.6 Relative Variation Across Branches

For a simple descriptive comparison, the range of mean AUROC across the eight training objectives is calculated separately for each branch. Let

$$R_m = \max_{\ell} A_{m,\ell} - \min_{\ell} A_{m,\ell}, \quad (69)$$

where $A_{m,\ell}$ denotes the mean AUROC of branch m under training loss ℓ .

Using the values in Table 49, the predictive branch has a range of 0.011, the representational branch has a range of 0.012, the evidential branch has a range of 0.016, and the fused HUE-OOD score has a range of 0.009.

Thus, among the three individual branches, the evidential branch has the largest observed numerical range, whereas the fused score has the smallest range. These quantities are descriptive and do not constitute a formal test of differences in sensitivity.

E.7 Practical Interpretation

The training-loss experiment provides evidence that the behavior of HUE-OD is not identical across training objectives. The magnitude of this variation differs by branch.

The predictive and representational branches vary within relatively narrow AUROC ranges of 0.011 and 0.012, respectively. The evidential branch has a somewhat wider range of 0.016. The fused score varies by 0.009, which is the smallest range among the four reported scores.

The CE plus supervised contrastive objective obtains the highest mean AUROC in this experiment, but its advantage over standard cross-entropy is only 0.001. Given the standard deviations of the five-run estimates, this difference should not be interpreted as evidence that the combined loss is statistically superior.

Similarly, the lower AUROC observed with Dice Loss indicates greater sensitivity in this particular configuration, but does not establish that Dice Loss is generally unsuitable for HUE-OD. The result may depend on the optimization procedure, hyperparameters, dataset, or backbone.

E.8 Limitations of the Training-Loss Analysis

Several limitations should be considered.

First, only eight training objectives are evaluated. The analysis therefore covers a representative but finite subset of possible loss functions.

Second, all results are obtained with ViT-Small, CIFAR-10 as ID, and SVHN as OOD. Training-loss sensitivity may differ for other architectures, datasets, or distribution shifts.

Third, five random seeds are used for each configuration. This is sufficient to report an empirical estimate of run-to-run variation but is not sufficient to establish fine-grained statistical differences between closely performing losses without additional paired statistical analysis.

Fourth, the analysis considers AUROC only. Other properties, including FPR95, AUPR, calibration, uncertainty concentration, and computational cost, may respond differently to changes in the training objective.

Finally, the observed stability of the fused score should not be interpreted as proof of training-loss invariance. The results demonstrate only that the fused AUROC remains within the reported numerical range under the eight selected objectives and the specified experimental protocol.

E.9 Overall Assessment

Overall, the training-loss analysis indicates that HUE-ODD maintains a relatively narrow range of AUROC values across the eight evaluated training objectives. The fused score varies from 0.977 to 0.986, while the individual predictive, representational, and evidential branches vary over ranges of 0.011, 0.012, and 0.016, respectively.

The evidential branch therefore shows the largest numerical sensitivity among the three individual branches in this experiment. The fused score shows the smallest range, although this observation alone does not establish statistical invariance or causal robustness to changes in the training objective.

The default cross-entropy configuration provides a strong reference point with an AUROC of 0.985 ± 0.003 . The CE plus supervised contrastive configuration reaches 0.986 ± 0.003 , but the 0.001 difference is small relative to the observed run-to-run variability.

Taken together, these results support the use of the default training objective as a reasonable and reproducible configuration for the main evaluation, while indicating that alternative training objectives can produce modest changes in individual uncertainty branches. Broader claims about loss-function robustness would require additional architectures, datasets, training settings, and per-seed statistical comparisons.