A Unified Approach Towards Active Learning and Out-of-Distribution Detection

Anonymous authors
Paper under double-blind review

Abstract

In real-world applications of deep learning models, active learning (AL) strategies are essential for identifying label candidates from vast amounts of unlabeled data. In this context, robust out-of-distribution (OOD) detection mechanisms are crucial for handling data outside the target distribution during the application's operation. Usually, these problems have been addressed separately. In this work, we introduce SISOM as a unified solution designed explicitly for AL and OOD detection. By combining feature space-based and uncertainty-based metrics, SISOM leverages the strengths of the currently independent tasks to solve both effectively without requiring specific training schemes. We conducted extensive experiments showing the problems arising when migrating between both tasks. In our experiments SISOM underlined its effectiveness by achieving first place in one of the commonly used OpenOOD benchmark settings and top-3 places in the remaining two for near-OOD data. In AL, SISOM [1] delivers top performance in common image benchmarks.

1 Introduction

Large-scale deep learning models encounter several data-centric challenges during training and operation, particularly in real-world problems such as mobile robotic perception (Cai & Koutsoukos, 2020) and autonomous driving (Nitsch et al., 2021). On the one hand, these models require vast amounts of data and labels for training, driven by the uncontrolled nature of real-world tasks. On the other hand, even when trained with extensive data, these models can behave unpredictably when encountering samples that deviate significantly from the training data, known as out-of-distribution (OOD) data.

Active learning (AL) addresses the first limitation by guiding the selection of label candidates. In the traditional pool-based AL scenario (Settles, 2010), models start with a small labeled training set and can iteratively query data and its labels from an unlabeled data pool. The selection is based on model metrics such as uncertainty, diversity, or latent space encoding. One AL cycle concludes with the model being trained on the labeled subset, including the newly added samples.

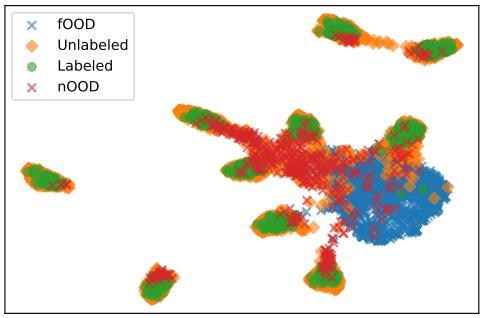


Figure 1: CIFAR-10 UMAP plot of unlabeled, near and far OOD data compared to labeled data. For details see Appendix B.4.

The second challenge, dealing with unknown data during operation, is typically addressed by OOD detection. OOD detection distinguishes between in-distribution (InD) data used for training the model and OOD samples, which differ from the training distribution. Literature differentiates between near-ODD and far-OOD, which can be categorized by the type of distribution shifts occurring. Yang et al. (2022); Zhang et al. (2023) defines near-OOD as a pure covariate shift, while far-OOD often contains a semantic shift.

Recent works in AL combined AL and OOD detection into the Open-set AL (OSAL) scenario, assuming the existence of OOD data in the unlabeled pool, enabling a selection from the data pool without prior

[1] Code will be published upon acceptance - for review in supplementary

knowledge. However, existing methods Ning et al. (2022); Yang et al. (2023); Du et al. (2021) usually rely on separate components for OOD separation and data selection, respectively.

As a result, a consideration of the connection between both tasks has rarely been done. From a method perspective, both methodologies utilize common metrics, such as uncertainty, latent space distances, and energy. In addition, a sample detected by such metrics can be, on the one hand, a novel AL sample that is insufficiently represented by the current training distribution. On the other hand, the sample can pose a covariant shift in an OOD setting. Considering both cases as depicted in Fig. 1 show an ambiguity and overlap of both sample categories. This raises the question whether an examination of the ambiguity of and relations between the respective samples can provide valuable insights for designing approaches for both tasks and OSAL[2], where current methods mostly consider both subtasks as independent.

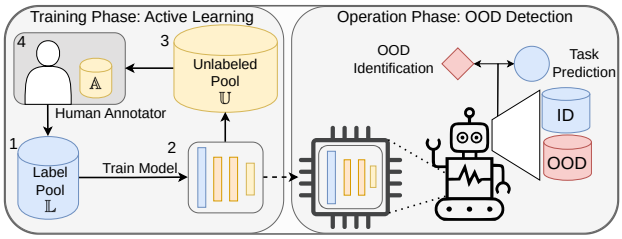


Figure 2: Real-world application life cycle comprising active learning in the training phase (left) and out-of-distribution detection in the operation phase (right).

Besides the methodology perspective, mobile robotic applications often train neural networks on recordings and afterward deploy them for operation as shown in Fig. 2 as an integrated life cycle. Given the amount of collected data, AL is applied for label-efficient training. During operation, OOD detection is employed to observe if the current operation is within the trained domain, which is necessary for real-world operation domains. Existing works address both challenges, which are separate methods that can lead to diverging goals like specific training schemes. Addressing these tasks separately introduces significant overhead, especially for deployment and development like hyperparameter optimization or the training of auxiliary models.

Our work explores the connection between AL and OOD detection, introducing a unified approach for both tasks that leverages their mutual strengths. Specifically, we propose **S**imultaneous **I**nformative **S**ampling and **O**utlier **M**ining (**SISOM**), which uses enriched feature space distances based on coverage creating a symbiosis between AL and OOD detection. By exploiting the ambiguity of both tasks, SISOM achieves top performance across most near-OOD and AL benchmarks. By uniting both tasks, SISOM simplifies the application life cycle by *eliminating the need for a separate OOD design phase* and resolves conflicting design goals. This perspective contrasts with open-set AL, where OOD data is incorporated into the unlabeled pool, forming a combined task. Additionally, SISOM provides a *novel latent space analysis* for *post-training latent space refinement* and a first-of-its-kind *self-balancing of uncertainty and diversity metrics*.

In summary, *our contributions* are as follows:

- We propose **S**imultaneous **I**nformative **S**ampling and **O**utlier **M**ining (**SISOM**), a novel method designed for both OOD detection *and* AL.
- We introduce a latent space analysis enabling an *optimization loop* for further *post-training latent space refinement* and a *self-balanced uncertainty diversity fusion*.
- In extensive experiments, we demonstrate SISOM effectiveness in common image AL *and* OOD benchmarks against highly specialized state-of-the-art methods.

2 Preliminaries

Active Learning: AL is a subfield of machine learning designed to reduce the number of required labels by querying a set of new samples $\mathbb{A}$ of a query size q in a cyclic process. Let $\mathcal{X}$ represent a set of samples and $\mathcal{Y}$ a set of labels. AL starts with an initially labeled pool $\mathbb{L}$, containing data samples with features $\mathbf{x}$ and corresponding label y, and an unlabeled pool $\mathbb{U}$ where only $\mathbf{x}$ is known. However, y can be queried from a human oracle. We further assume that $\mathbb{L}$ and $\mathbb{U}$ are samples from a distribution Ω. In each cycle, a model

[2]In fact, Schmidt et al. (2025) employed SISOMe for OSAL based on a preprint version of this work.

f is trained such that $f : \mathcal{X}_{\mathbb{L}} \rightarrow \mathcal{Y}_{\mathbb{L}}$. This model then selects new samples from $\mathbb{U}$ based on a query strategy $Q(\mathbf{x}, f)$, which utilizes (intermediate) model outputs. As a result, the newly annotated set $\mathbb{A}$ is added to the labeled pool $\mathbb{L}^{i+1}$ and removed from the unlabeled pool $\mathbb{U}^{i+1}$.

Out of Distribution Detection: Ancillary, OOD detection assumes a model $f : \mathcal{X}_{\mathbb{L}} \rightarrow \mathcal{Y}_{\mathbb{L}}$ trained on our training data $\{\mathbf{x}, y\} \in \mathbb{L}$ which have been sampled from the distribution Ω. During evaluation or inference, a model f encounters data samples $\tilde{\mathbf{x}}$ from a distribution Θ and Ω, where $\Omega \cap \Theta = \emptyset$ and $\tilde{\mathbf{x}} \notin \mathbb{L}$. Data sampled from Ω are referred to as InD data, while samples from Θ are referred to as OOD data. Based on the trained model f, a metric S is used to determine whether a sample x is sampled from Ω or Θ.

$$G(\mathbf{x}, f) = \begin{cases} \text{InD} & \text{if } S(\mathbf{x}; f) \geq \lambda \\ \text{OOD} & \text{if } S(\mathbf{x}; f) < \lambda \end{cases} \tag{1}$$

OOD detection is further categorized into near- and far-OOD (Zhang et al., 2023). Far-OOD refers to completely unrelated data, such as comparing MNIST (LeCun et al., 1998) to CIFAR-100 (Krizhevsky et al., 2009), while CIFAR-10 (Krizhevsky et al., 2009) to CIFAR-100 would be considered as near-OOD. OpenODD (Yang et al., 2022) ranks near-OOD detection as more challenging.

3 Related Work

Given the disentanglement of fields, we review the related work individually.

Active Learning: AL mainly considers the pool-based and stream-based scenario (Settles, 2010), where data is either queried from a pool in a data center or a stream on the fly. For deep learning, the majority of current research deals with pool-based AL (Ren et al., 2021). However, further scenarios have been evaluated by Schmidt & Günnemann (2023) and Schmidt et al. (2024). Independent of the scenarios, samples are selected either by prediction uncertainty, latent space diversity, or auxiliary models. A majority of the uncertainty-based methods rely on sampling - like Monte Carlo Dropout (Gal & Ghahramani, 2016) - or employ ensembles (Beluch et al., 2018; Lakshminarayanan et al., 2017). To additionally ensure batch diversity Kirsch et al. (2019) used the joint mutual information. The uncertainty concepts have been employed and further developed for major computer vision tasks, including object detection (Feng et al., 2019; Schmidt et al., 2020), 3D object detection (Hekimoglu et al., 2022; Park et al., 2023b), and semantic segmentation (Huang et al., 2018). One of the few works breaking the gap between both tasks (Shukla et al., 2022) modified an OOD detection method for pose estimation. Dutta et al. (2025), proposed to use imprecise probability by Imprecise Neural Networks for statistical verification of autonomous system to enable save operation and AL for in reinforcement learning. Mukhoti et al. (2023) proposed DDU, an uncertainty baseline based on spectral convolutions and Gaussian mixture models, which show improvements against other general uncertainty approaches on AL and OOD detection compared to other uncertainty approaches. Given the requirements of spectral convolutions, DDU is not flexible applicable for all use cases. In contrast, diversity-based approaches aim to select key samples to cover the whole dataset. Sener & Savarese (2018) proposed to choose a CoreSet of the latent space using a greedy optimization. Yehuda et al. (2022b) selected samples having high coverage in a fixed radius for low data regimes. Mishal & Weinshall (2024) extends the approach for more data regimes dynamic strategy mixing. Ash et al. (2020) enriched the latent space dimensions to the dimensions of the gradients and included uncertainty in this way. The concept of combining uncertainty with diversity has been further refined for 3D object detection (Yang et al., 2024; Luo et al., 2023). Liang et al. (2024) combined different diversity metrics for the same task. In semantic segmentation, Surprise Adequacy (Kim et al., 2020) has been employed to measure how surprising a model finds a new instance. Yi et al. (2022) used auxiliary tasks for unsupervised model training to select diverse samples. Besides the metric-based approach, the selection can also be made by auxiliary models mimicking diversity and uncertainty. These approaches range from loss estimation (Yoo & Kweon, 2019), autoencoder-based approaches (Sinha et al., 2019; Zhang et al., 2020; Kim et al., 2021) and graph models (Caramalau et al., 2021), to teacher-student approaches (Peng et al., 2021; Hekimoglu et al., 2024).

Out-of-Distribution Detection: To facilitate a fair comparison and evaluation of OOD methods, benchmarking frameworks like OpenOOD (Yang et al., 2022; Zhang et al., 2023) have been introduced, which

categorizes the methods into preprocessing methods altering the training process and postprocessing methods being applied after training. Preprocessing techniques include augmenting training data like mixing (Zhang et al., 2018; Tokozume et al., 2018) different samples or applying fractals to images (Hendrycks et al., 2022). Postprocessing approaches include techniques of manipulations on neurons and weights of the trained network, such as filtering for important neurons (Ahn et al., 2023; Djurisic et al., 2022), or weights (Sun & Li, 2022), or clipping neuron values to reduce OOD-induced noise (Sun et al., 2021). Recent advancements in post-hoc network enhancement include SCALE (Xu et al., 2024), which enhances OOD detection by activation scaling. Logit-based approaches encompass the model output to estimate uncertainties using temperature-scaling (Liang et al., 2018), modified entropy scores (Liu et al., 2023), energy scores (Liu et al., 2020; Elflein et al., 2021) or ensembles (Arpit et al., 2022). Other methods use distances in the feature space, such as the Mahalanobis distance between InD and OOD samples (Lee et al., 2018), rely on gradients (Liang et al., 2018; Hsu et al., 2020; Huang et al., 2021; Schwinn et al., 2021), estimate densities (Charpentier et al., 2020; 2022) or apply k-nearest-neighbor on latent space distances (Sun et al., 2022). Nearest Neighbor Guidance (NNGuide) (Park et al., 2023a) refines classifier-based scores by respecting the data manifold's boundary geometry. A different branch operates on the features directly and evaluates properties like the Norm (Yu et al., 2023) or performs rank reductions via SVD (Song et al., 2022). NAC (Liu et al., 2024) combined gradient information with a density approach, where a probability density function over InD samples is estimated. CombOOD (Rajasekaran et al., 2024) is a semi-parametric framework that combines nearest-neighbor and Mahalanobis distances to improve OOD detection accuracy.

OpenSet Active Learning: The emerging field of OpenSet AL considers both tasks in one cycle, assuming the AL pool is polluted by OOD samples. Existing approaches (Ning et al., 2022; Park et al., 2022; Yang et al., 2023; Safaei et al., 2024) present strong results by primarily tackling both tasks by *separate* modules containing auxiliary models. As both unlabeled and OOD samples are considered decoupled with uncorrelated modules, this field is orthogonal to our examination of correlation and entanglement. Alternative approaches such as SIMILAR (Kothawade et al., 2021) are being explored, which proposes a unified active learning framework using submodular information measures to handle OOD data within the unlabeled set. However, the computation of these measures introduces additional computational overhead and requires access to initially known OOD data. Furthermore, Stojnić et al. (2024) built on top of the open-set classifier concept and created an ensemble of models with an additional OOD class for semi-supervised AL, which comes with the increased computational demands. In contrast to works in this field, we investigate the correlation between unlabeled and OOD samples to provide a unified metric for both. We believe that this field benefits from the joint consideration of AL and OOD samples as well as an examination of their ambiguity.

While various works exist in OOD and AL, both tasks are considered independent. Even in OSAL, the tasks are **mostly** considered by independent method components. Some uncertainty methods are evaluated on both tasks but limit their evaluation to the uncertainty domain. Current state-of-the-art approaches are often specified for one task. In addition, the application life cycle consideration is unexplored.

4 Methodology

To address both AL and OOD detection tasks in a unified method to simplify real-world applications, we need to first understand the goals of these two tasks. AL aims to identify and select samples that are beneficial for training and increase the model's performance. These samples typically position themselves between the existing clusters in the latent space or near the decision boundaries. OOD detection targets the identification of data outside the training data and, therefore, outside the known clusters in latent space. Given the definition of far- and near-OOD, near-ODD is closer to InD data and located close to the decision boundaries and in between the existing clusters. Liu & Qin (2024) recently showed that OOD is generally closer to the decision boundary than InD, confirming this hypothesis. Fig. 1 investigates this hypothesis and shows the overlap n distribution of interesting unlabeled data and (near-)OOD data compared to distance to far-OOD samples.

To target these overlapping regions, we designed a method that focused on the latent space regions between the clusters. To do so, SISOM employs an enlarged feature space **Coverage (1)** and increases expressiveness by weighting important neurons in a **Feature Enhancement (2)**. Based on this feature representation,

we refine the AL selection and the InD and OOD border by using an inner-to-outer class **Distance Ratio (3)**, guiding it to unexplored and decision boundary regions. As feature space distances are prone to poorly defined latent space representations, we introduce **Feature Space Analysis (4)** providing a self-deciding fusion of our distance metric with an uncertainty-based energy score. Optionally, our previous analysis enables us to optimize the **Sigmoid Steepness (5)**, providing a further refinement of the feature space representations from **(2)**. An overview is depicted in Fig. 3. With this setup, SISOM specifically addresses both tasks without any requirements on the training scheme.

(1) Coverage: We aim to identify the regions of the samples that are interesting and unexplored for AL as well as OOD samples in latent space. To do so, we rely on an informative latent space covering as much information as possible.

To increase the information gain, we cover the full network and define the feature space representation of an input sample $\mathbf{x}$ as a concatenation of the latent space of multiple layers h_j in a set of selected layers H in Eq. (2). This approach follows the procedures of neural coverage (Kim et al., 2019; Liu et al., 2024) and is contrasting to most diversity-based AL approaches (Sener & Savarese, 2018; Ash et al., 2020), which use a single layer.

$$\mathbf{z} = h_1(\mathbf{x}) \oplus \cdots \oplus h_j(\mathbf{x}) \oplus \cdots \oplus h_n(\mathbf{x}) \tag{2}$$

Given the feature space $\mathbf{z}$, we further denote $\mathbb{Z}_U$ as a set of feature space representations of unlabeled samples from $\mathbb{U}$, while $\mathbb{Z}_L$ denotes the set of representations of all labeled samples $\mathbb{L}$.

(2) Feature Enhancement: To enhance the expressiveness of our defined latent space, improving class separation, we introduce a weighting of individual layers. Prior research (Huang et al., 2021; Liu et al., 2024) have demonstrated that the gradients of neurons with respect to the KL divergence of the model's output and a uniform distribution encapsulate valuable information for OOD detection.

We apply the technique to improve the features further and enrich these by representing the individual contribution of each neuron i, denoted as g_i. This gradient describes each neuron's contribution to the actual output being different from the uniform distribution. A low value suggests that the neuron has little influence on the prediction of a given input sample. Conversely, if the value is high, the respective neuron is crucial for the decision process.

Thus, the gradient vector can be interpreted as a saliency weighting for the activation values in the feature space to support separability. In detail, we compute the gradient of the Kullback-Leibler (KL) divergence between an uniform distribution u and the softmax output distribution $f(\mathbf{x})$ for an input sample $\mathbf{x}$ with respect to the selected features $\mathbf{z}$:

$$\mathbf{g} = \frac{\partial D_{KL}(u || f(\mathbf{x}))}{\partial \mathbf{z}}. \tag{3}$$

We create a weighted feature representation $\tilde{\mathbf{z}}$ by multiplying the calculated saliency element-wise with the feature space representation $\mathbf{z}$ using the hadamard product $\odot$ and restrict it to $[-1, 1]$ with the sigmoid function σ:

$$\tilde{\mathbf{z}} = \sigma(\mathbf{z} \odot \mathbf{g}). \tag{4}$$

The resulting gradient-weighted feature representation effectively prioritizes the most influential neurons for each input. This facilitates the identification of inputs activating atypical influence patterns, which is significant for AL as well as OOD detection. A qualitative analysis demonstrating the effect of the feature enrichment is given in Appendix A.4.

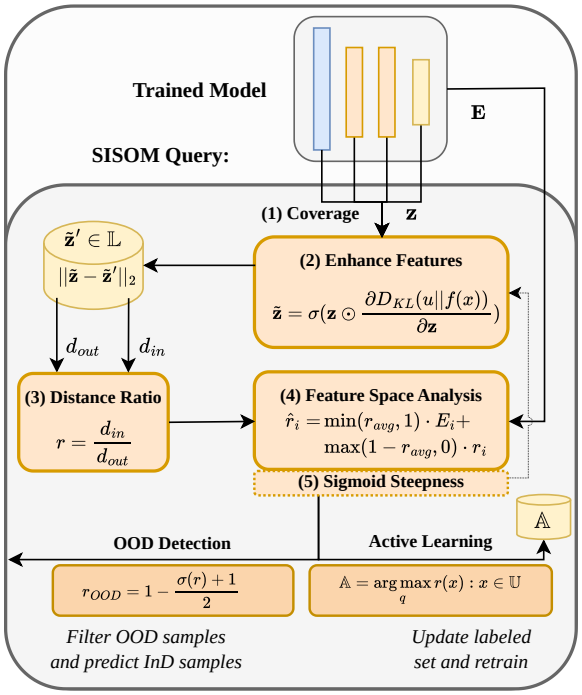
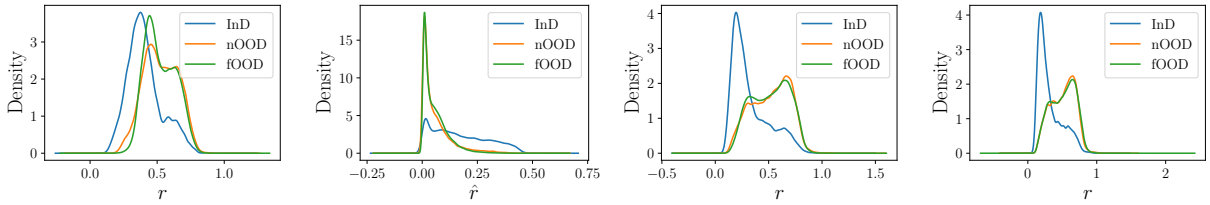


Figure 3: SISOM framework for OOD detection and AL combined.



(a) SISOM (b) SISOMe (c) SISOM +OS (d) SISOM +OS+RS

Figure 4: Density plots for SISOM with energy, Optimal Sigmoid Steepness (OS) and Reduced Subset Selection (RS) on CIFAR-100 with near-OOD (nOOD) and far-OOD (fOOD) as defined in OpenOOD.

(3) Distance Ratio: After we defined and enhanced our latent space, we design our metric to identify the respective samples. Contrasting to other works in the latent space domain for AL and OOD detection (Sener & Savarese, 2018) which, rely on simple distance metrics, we take inspiration from complex distance metrics (Kim et al., 2019) for detecting adversarial examples.

We assume the location of important samples in between the existing clusters in latent space. While far-OOD and curious AL samples are easier to detect due to larger latent space distances, near-OOD or AL samples close to the decision boundary, are more difficult to detect and often more important. To identify samples in these regions, we rely on a distance quotient between inner-class and outer-class distances, boosting the detection of samples close to the decision boundary.

The inner-class distance d_{in} is defined as the minimal feature space distance to a known sample of the same class c as the predicted pseudo-class of the given sample. The outer-class distance d_{out} represents the minimal feature space distance to a known sample of a different class than the sample's pseudo-class.

$$d_{in}(\tilde{\mathbf{z}}) = \min_{\mathbf{z}' \in \mathbb{Z}_L(c'=c)} ||\tilde{\mathbf{z}} - \tilde{\mathbf{z}}'||_2 \qquad (5) \qquad\qquad d_{out}(\tilde{\mathbf{z}}) = \min_{\mathbf{z}' \in \mathbb{Z}_L(c' \neq c)} ||\tilde{\mathbf{z}} - \tilde{\mathbf{z}}'||_2 \qquad (6)$$

The distance is computed on the gradient-enhanced feature space $\tilde{\mathbf{z}}$ defined in Eq. (4) with $\mathbf{z}'$ describing the nearest sample from the set of known samples $\mathbb{Z}_L$.

In many state-of-the-art works on AL, computationally expensive distance calculations are often present (Sener & Savarese, 2018; Ash et al., 2020; Caramalau et al., 2021). To make our approach more efficient for AL and feasible for large-scale OOD detection tasks, we select a representative subset $\mathbb{T} \subset \mathbb{Z}_L$ as a comparison set, thereby significantly reducing computational overhead. We utilize a fixed radius neighborhood to select samples, maximizing the coverage of the dataset within this radius in the feature space for each class. The effect of this subset selection is further investigated in Section 5.

Our SISOM score r reflects the distance between each neuron's weighted feature representation in the latent space and the nearest sample of the predicted class relative to the closest distance to a sample from a different class for a sample $\mathbf{x}$ and the respective representation $\tilde{\mathbf{z}}$:

$$r(\mathbf{x}) = \frac{d_{in}}{d_{out}}. \qquad (7)$$

An extended comparison of the different distance metrics and their ability to separate InD and OOD is shown in Appendix A.4, while a SISOM is depicted in Fig. 4a. We further omit $\mathbf{x}$ for r after Eq. (8).

Since we calculated a scale-value for each sample, we follow the most commonly used top-k selection (Yoo & Kweon, 2019; Kim et al., 2021; Gal & Ghahramani, 2016) to select q samples with the highest distance ratio r, with q being the AL query size:

$$\mathbb{A} = \text{argmax}_q r(\mathbf{x}) : \mathbf{x} \in \mathbb{U}. \qquad (8)$$

For OOD Detection, we map the distance ratios r to an interval $[0; 1]$ with the strictly monotonically decreasing function:

$$r_{OOD} = 1 - \frac{\sigma(r) + 1}{2}. \qquad (9)$$

(4) Feature Space Analysis: Having a well-defined latent space is crucial for SISOM to attain optimal performance. Furthermore, we hypothesize that techniques relying on feature space metrics are more dependent on feature space separation than uncertainty-based methods. This dependency is important for SISOM as it utilizes a quotient of feature space metrics. Nevertheless, obtaining a well-defined and separable latent space may pose challenges in specific contexts and tasks.

To estimate the separability of feature space, we compute the average distance ratio r_{avg} using Eq. (4) and Eq. (7) for the known set as:

$$r_{avg} = \frac{1}{|\mathbb{L}|} \sum_{\tilde{\mathbf{z}} \in \mathbb{L}} \frac{d_{in}(\tilde{\mathbf{z}})}{d_{out}(\tilde{\mathbf{z}})} = \frac{1}{|\mathbb{L}|} \sum_{\mathbf{z} \in \mathbb{L}} \frac{d_{in}(\sigma(\mathbf{z} \odot \mathbf{g}))}{d_{out}(\sigma(\mathbf{z} \odot \mathbf{g}))}. \tag{10}$$

A lower r_{avg} value indicates better separation of the samples in the enhanced feature space, implying that samples of the same class are relatively closer together than samples of different classes.

To mitigate possible performance disparities of SISOM in difficult separable domains, we introduce a novel self-deciding process combining SISOM with the uncertainty-based energy score $E(\mathbf{x}) = -\log \sum_{i=1}^{c} \exp(f(\mathbf{x})_i)$ estimated from the model's output logits $f(\mathbf{x})$. We utilize r_{avg} to weights the combination of uncertainty with our SISOM diversity metric as follows for an instance i:

$$\hat{r}_i = \min(r_{avg}, 1) \cdot E_i + \max(1 - r_{avg}, 0) \cdot r_i. \tag{11}$$

The created SISOMe score $\hat{r}$ balances between uncertainty and diversity and relies either more on the energy score E or the distance ratio r_i, based on the feature space separability. If $r_{avg} \rightarrow 1$, indicating poorly separated classes, $\hat{r}_i$ relies more on the energy score. Conversely, if $r_{avg} \rightarrow 0$, suggesting a well-separated feature space, $\hat{r}_i$ relies more on the distance ratio. A density outline of our combined approach SISOMe is given in Fig. 4b. Alternatively, one can replace r_{avg} with a tuneable hyperparameter in Eq. (11).

(5) Sigmoid Steepness: Since Eq. (10) depends on the sigmoid function defined in Eq. (4), the sigmoid function has a large influence on the enhanced feature space $\tilde{\mathbf{z}}$. An additional hyperparameter α can influence the sigmoid function's steepness. As $\mathbf{z}$ is concatenated from different layers in Eq. (2), the sigmoid can be applied to each layer j individually. This allows for a more nuanced control over the influence of each neuron's contribution to the final decision, and so influences the separability of the feature space. We define the sigmoid using the steepness parameter α as:

$$\sigma_j(\mathbf{x}) = \frac{1}{1 + e^{-\alpha_j \mathbf{x}}}; \quad \{\alpha_j : h_j \in \mathbf{z} \;\; \forall j\}. \tag{12}$$

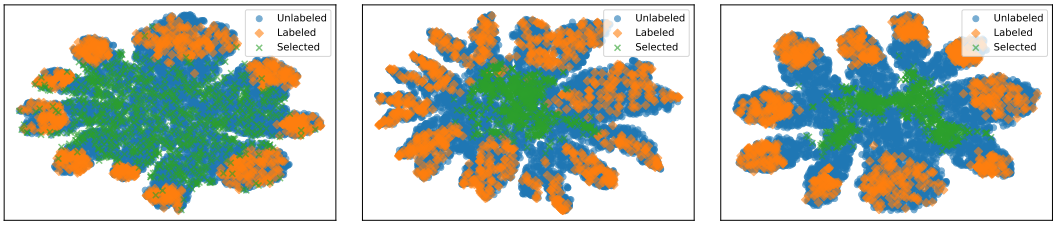
Relating to Eq. (4), the set α of steepness parameters of the sigmoid function for each layer h_j, determines the degree of continuity or discreteness of the features within that layer. By applying a layerwise sigmoid, Eq. (4) is formulated as follows:

$$\begin{aligned} \tilde{\mathbf{z}} &= \sigma_1(h_1(\mathbf{x}) \odot g_{i,1}) \oplus \cdots \oplus \sigma_j(h_j(\mathbf{x}) \odot \mathbf{g}_{i,j}) \oplus \cdots \oplus \sigma_n(h_n(\mathbf{x}) \odot \mathbf{g}_{i,n}), \\ \text{with} \quad \mathbf{g}_{i,j} &= \frac{\partial D_{KL}(u || f(\mathbf{x}))}{\partial h_{j,i}}; \quad \forall j. \end{aligned} \tag{13}$$

Following this consideration we can select α values which optimize the feature space separability metric r_{avg} from Eq. (10) by minimizing $\alpha_{opt} = \arg\min_{\alpha} r_{avg}(\alpha)$. Besides the quantitative assessment of our Feature Space Analysis and Sigmoid Steepness in Section 5, the influence of the Sigmoid Steepness is shown in Fig. 4c.

5 Experiments

To evaluate the abilities of SISOM for the real-world application life cycle (Fig. 2), we conducted comprehensive experiments on **AL** and **OOD detection** individually. We utilize the commonly used closed-set pool-based AL scenario (Settles, 2010) for AL. For OOD detection, we employ the extensive OpenOOD benchmark (Yang et al., 2022; Zhang et al., 2023). In addition, we conduct OOD detection using models trained by AL to evaluate the whole application life cycle steps from Fig. 2.



(a) CoreSet (b) Loss Learning (c) SISOM (ours)

Figure 5: T-SNE feature space comparison of Loss Learning, CoreSet and SISOM for SVHN on cycle 1. SISOM effectively targets the areas in-between the clusters.

An evaluation of the **OSAL** scenario, where OOD data is present in the unlabeled AL pool, is out of scope in this work. Existing OSAL (Ning et al., 2022; Park et al., 2022; Yang et al., 2023; Safaei et al., 2024) approaches employ **two** components one for each task, while this work focuses on a single component which solves both based on an ambiguity. This ambiguity would confuse the proposed method when applied independently to the compound task. In addition, the OOD filtering components often require access to OOD samples, which is not permitted in the classic OOD detection setup. However, a recent work (Schmidt et al., 2025) effectively integrated SISOM with additional components into a framework addressing OSAL.

Latent Space Assessment: In Fig. 1, we compare near- and far-OOD data as well as unlabeled data with the labeled data. It can be seen that the near-OOD data and the unlabeled data are positioned close to the individual clusters of the InD labeled data. This confirms our hypothesis that near-OOD can be ambiguous to unlabeled data. Moreover, in real-world applications, the model is more likely to encounter semantic shifts along with contextually similar data, like TinyImagenet (Le & Yang, 2015) to CIFAR-10 Krizhevsky et al. (2009) instead of contextual shifted far-OOD data like MNIST (LeCun et al., 1998) to CIFAR-10.

To validate the assumptions made in Section 4 for AL, we examine the configuration of the latent space of our selection in the AL experiments. The objective of our method is to select samples in the decision boundary region for the AL case. In Fig. 5, we compare CoreSet and Loss Learning with SISOM. It can be observed that CoreSet, as intended, exhibits high diversity in unseparated regions. The pseudo-uncertainty-based Loss Learning method is more concentrated in its selection but fails to diversify the selection across all decision boundaries. In contrast, SISOM, as shown in Fig. 5c, focuses on the decision boundary while successfully covering the entire area between the unseparated samples, demonstrating the effectiveness of our SISOM in targeting the region for AL and near-OOD.

Active Learning: AL exists in numerous variants; given the focus on showing the relation between OOD detection and AL, we adhered to common close-set AL benchmark settings (Yoo & Kweon, 2019; Ash et al., 2020) with the in this context widely evaluated datasets, namely CIFAR-10 and CIFAR-100 (Krizhevsky et al., 2009), as well as SVHN (Netzer et al., 2011), in conjunction with different ResNet models (He et al., 2016). Since we assume a complete learning life-cycle, we mostly focus on the high and mid data regions (Lüth et al., 2023), as low data regions would not allow reasonable performance for deployment. We selected our query sizes accordingly and followed the commonly suggested sizes of $q = 1000$ for CIFAR-10 (Yoo & Kweon, 2019; Lüth et al., 2023) and $q = 2000$ for CIFAR-100 (Caramalau et al., 2021). We employ several baselines, including **CoreSet** (Sener & Savarese, 2018), **CoreGCN** (Caramalau et al., 2021), **Random**, **Badge** (Ash et al., 2020), and **Loss Learning (LLoss)** (Yoo & Kweon, 2019). Additionally, we adapted **NAC** (Liu et al., 2024) from OOD detection to AL to assess the transferability from OOD to AL.

Since more recent AL methods use pre-text tasks or unsupervised weights to select samples, we additionally consider an additional Semi-Supervised AL setup for a fair comparison, including more recent approaches. In this setting we compare SISOM with **TypiClust** Yehuda et al. (2022a), **ProbCover** (Yehuda et al., 2022b) and **PT4AL** Yi et al. (2022). Additionally, in this setup, SISOM can profit from the well-defined feature space. All experiment details, including parameters and settings, are in Appendix B.1.

CIFAR-10: In the CIFAR-10 benchmark depicted in Fig. 6a, SISOM and SISOMe demonstrate rapid progress and consistent performance, surpassing other methods in all selection cycles. Furthermore, as the sample size

increase, our method maintains its superiority over Learning Loss and CoreSet. NAC does not demonstrate superior performance compared to Random. To underline the versatility of SISOM, we analyze additional mid-data regimens and different query sizes as proposed by Lüth et al. (2023) in Appendix A.3. In the mid data regime in Fig. 10a. In this data regime, SISOM and SISOMe can maintain their performance shown in the high data region. As SISOM does not involve continuous distance updates and applies a top-k selection, it could suffer from less diversity in the selected batch. To investigate this problem, we perform additional experiments with reduced and increased query sizes in Fig. 9a and Fig. 9b (Appendix A.2), respectively. SISOM's high performance in both settings strengthens our hypothesis that targeting the areas between clusters leads to higher batch diversification than purely uncertainty-based approaches.

CIFAR-100: After examining SISOM in datasets with a limited number of classes, we examine the AL setup on the larger CIFAR-100 dataset and report the results in Fig. 6b. In this setting, all methods are less stable in their ranking compared to the other dataset, reflecting the increased difficulty of the dataset. The complexity of the dataset requires more data for the model to perform effectively. While in the early stages, pure diversity-based methods are in the lead, SISOM gains velocity in the last selection steps and achieves the highest performance difference only in the last step SISOMe is more effective.

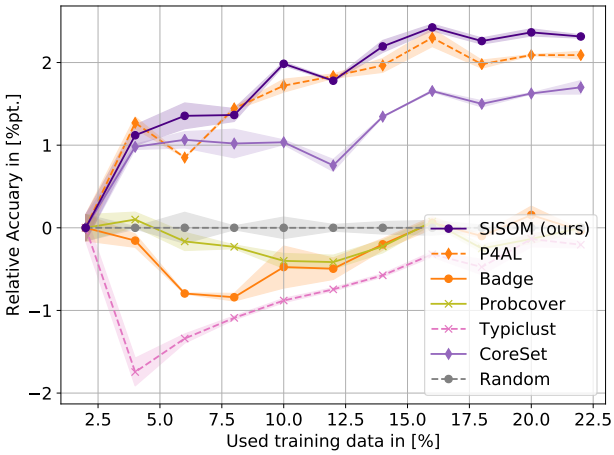
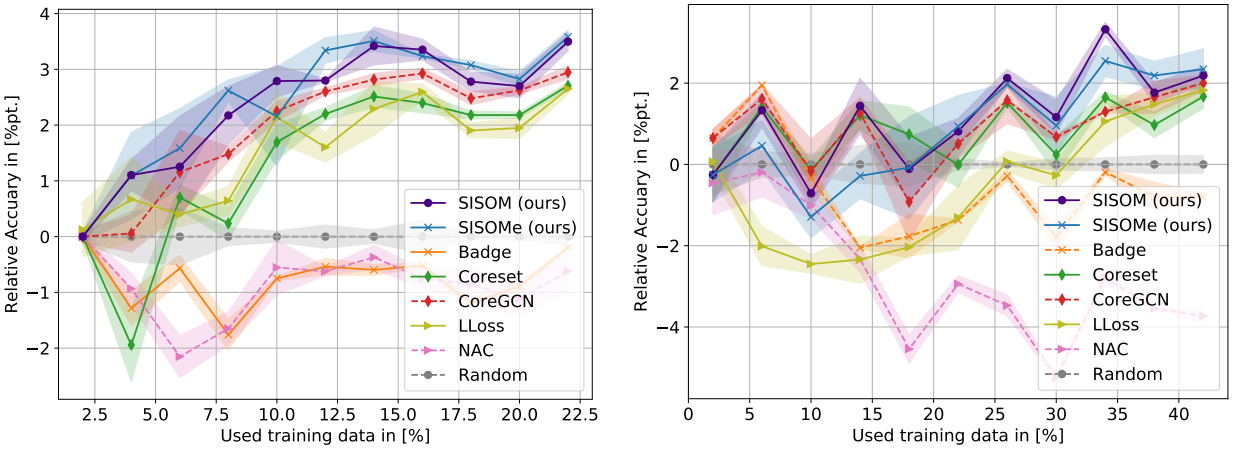


Figure 7: Comparison of accuracy relative to random selection of active learning methods on a semi-supervised CIFAR-10 setup with indicated standard errors for ResNet 18.

Semi-Supervised: In Fig. 7, we compare different approaches designed for a semi-supervised AL setup. For this experimental setup, every approach builds upon the unsupervised pre-train weights, which are required for TypiClust and ProbCover. While both approaches do not perform strongly in the classic data regime, PT4AL shows a good performance by leveraging training on auxiliary pre-tasks. However, SISOM strongly profits from the pre-training and outperforms PT4AL *without* additional pre-task training.

SVHN: Following the experiments on CIFAR-10 and CIFAR-100 we conducts experiments on SVHN and report them in Appendix A.1.



(a) CIFAR-10

(b) CIFAR-100

Figure 6: Comparison of different active learning methods on CIFAR-10 and CIFAR-100 with ResNet18 with indicated standard errors. Relative accuracy is the absolute performance difference between a method and random selection.

Additionally Active Learning Experiments: To underline the flexibility of SISOM, we conduct additional experiments on different models. Appendix A.3 and Fig. 10b show experiments with a larger ResNet version. The larger model leads to lower performance differences between all methods. SISOMe struggles in the early cycles with latent space assessment but eclipses others and SISOM after a few cycles. Finally, SISOM and SISOMe achieve the highest accuracy in the final cycles.

In all AL experiments, SISOM achieved state-of-the-art performance across all three datasets, confirming its viability for AL. Although SISOM initially lags behind on CIFAR-100, it surpasses other methods in later selection cycles as more training data improves feature space separation.

Out-of-Distribution Detection: Following our evaluation of SISOM on classic AL benchmarks, we utilize the OpenOOD framework Yang et al. (2022); Zhang et al. (2023) to assess its performance on the OOD detection task. We employ to the recommended benchmarks on CIFAR-10 (Krizhevsky et al., 2009), CIFAR-100 (Krizhevsky et al., 2009), and ImageNet 1k (Deng et al., 2009). For a fair evaluation, we use the official leader board Cross-Entropy training checkpoints and report near- and far-OOD results. The results are reported as aligned to the framework without standard deviation. In addition to the official benchmark, we evaluate the life cycle setting. The assignment of OOD datasets follows the framework's categorization Yang et al. (2022); Zhang et al. (2023) and is reported with additional settings in Appendix B.2. For all experiments, we employ the implementation provided by the OpenOOD framework when it is available. The baselines used for validation include **NAC** (Liu et al., 2024), **Ash** (Djurisic et al., 2022), **KNN** (Sun et al., 2022), **Odin** (Hsu et al., 2020), **ReAct** (Sun et al., 2021), **MSP** (Hendrycks & Gimpel, 2016), **Energy** (Liu et al., 2020), **Dice** (Sun & Li, 2022), **SCALE** (Xu et al., 2024), **CombOOD** (Rajasekaran et al., 2024), **NNGuide** (Park et al., 2023a), **RankFeat** (Yu et al., 2023), **FeatureNorm** (Song et al., 2022) and **GEN** (Liu et al., 2023) and were selected based on their near-OOD performance. Moreover, we ported the **CoreSet** (Sener & Savarese, 2018) AL method to verify the transferability from AL to OOD.

In Table 1, we examine the performance of SISOM and SISOMe for the three benchmarks. As stated before, the near-OOD data is more relevant for this task. We focus on the near-OOD evaluation and report the far-OOD results in Appendix A.5. For the *CIFAR-10* benchmark SISOMe and SISOM achieve the highest AUROC score for near-OOD data, respectively. In the *CIFAR-100* evaluation SISOMe ranks as the third-best method for near-OOD and standing apart from the individual metrics, SISOM and Energy. GEN, which shows weaker performance in the other setting, achieves the first place. This is an interesting finding since, in contrast to CIFAR-10, Energy achieves better performance than SISOM among the individual metrics on CIFAR-100. This supports our hypothesis that we obtain stronger performances in well-separated and poorly-separated feature spaces by considering the average ratio r_{avg} as a proxy for feature separation. The *ImageNet 1k* benchmark suggested by OpenOOD contains more classes and is a much larger dataset than the previous ones. In this setup SISOMe achieve the third-best scores on near-OOD, making it the only method with three top-three rankings. In addition, when aggregating the individual ranks across the three benchmarks, SISOMe ranks first. Interestingly, the NAC method, which showed high performance in CIFAR-

Table 1: Near-OOD AUROC benchmark for CIFAR-10, CIFAR-100, and ImageNet1k with Cross-Entropy training. Dataset and official checkpoints according to OpenOOD. It can be seen that most methods focus on one particular setting. Given the number of baselines for each dataset, we estimated a rank for each benchmark and aggregated them for an overall ranking reported in the first row. The top three ranks are marked in gray from dark to light, with the top method name also in bold. For individual datasets, the top three ranks are indicated from first to third using **bold**, double underlined or single underlined formatting.

Dataset	SISOMe*	SISOM	NAC	KNN	CoreSet	GEN	MSP	Energy	ReAct	Feat.Norm	ODIN	RankFeat	KLM	DICE	ASH	SCALE	CombOOD	NNGuide
Overall Rank	**1**	4	10	9	14	2	8	7	6	16	11	17	13	15	12	5	3	8
CIFAR-10	**91.76**	91.40	90.93	90.64	90.34	88.20	88.03	87.58	87.11	85.52	82.87	79.46	79.19	78.34	75.27	82.55	91.13	87.56
CIFAR-100	81.10	79.42	75.90	80.18	75.69	**81.31**	80.27	80.91	80.77	47.87	79.90	61.88	76.56	79.38	78.20	80.99	78.77	81.25
ImageNet 1k	78.59	77.33	74.43	71.10	-	76.85	76.02	76.03	77.38	67.57	74.75	50.99	76.64	73.07	78.17	81.36	**95.22**	73.57

Table 2: OOD benchmark for CIFAR-10 using the AL checkpoints of SISOM.

AUROC	SISOMe	ReAct	GEN	MSP	ASH	NAC	RankFeat
near-OOD	**86.84**	**86.84**	85.43	84.37	83.39	82.26	60.20
far-OOD	**88.39**	<u>87.72</u>	86.04	84.85	87.33	85.06	56.73

10, ranks much lower, and KNN, the third-best method in CIFAR-10, ranks last. Meanwhile, ASH, which ranks forth in this benchmark, ranked last in the CIFAR-10 setting. This underlines the difficulty of the different datasets and their benchmarked OOD pairs. In the real-world applications less relevant *far-OOD* evaluation in Appendix A.5, the dependence on the performance between datasets and methods persists. Despite this, SISOMe and SISOM achieve high scores, with SISOMe remaining the only method with a top-3 ranking in all benchmarks, leading to an overall top-one rank.

Full Life Cycle: To evaluate the effectiveness of SISOM in a life cycle setting, we conduct OOD benchmark experiments with models trained in an AL cycle. We used the same setting as for the benchmark CIFAR-10 experiments with similar near- and far-OOD. The exact setup and training scheme of the checkpoints is described in Appendix B. In Table 2, all methods suffer from less training data, however SISOMe archived the top performance for both OOD categories, making it suitable for the full application life cycle. A more detailed analysis including different intermediate cycles is shown in Appendix A.5.

Overall benchmarks, SISOMe is the only approach, being consistently under the top three ranks and even secured first place in two of them. Excluding SISOMe, SISOM achieved one top-three ranking and one top-one ranking. As intended, our method performs relatively better on near-OOD data than on far-OOD data. This is understandable, as the ratio between inner and outer class distance is higher for data close to the training data distribution, while the quotient is lower for far-OOD. Additionally, near-OOD is closer to the data of interest for AL selection. According to (Yang et al., 2022), near-OOD is considered the more challenging task and is more likely to occur in real-world applications. Thus, a higher score on near-OOD may be preferred in practice.

Ablations Studies: In an ablation study, we examine the effect of unsupervised feature space analysis and reduce labeled set $\mathbb{T}$. A study of the individual components of SISOM is given in Appendix A.4.

<u>Optimal Sigmoid Steepness:</u> In our feature space analysis in Section 4, we derived r_{avg} in Eq. (10) as a proxy for the feature space separability. Due to the distance concept of SISOM, we hypothesize that it works better in well-separated feature spaces. To examine this, we conduct a random search for different α sets and record the different r_{avg} values. To reduce the search space, we follow the premise postulated in Section 4 that generally, deeper layers require a steeper sigmoid curve, i.e., a higher α_j value due to the nature of the features captured within these layers. After computing every r_{avg} value for each combination of α, we select the α_{opt} set that minimizes r_{avg}. Formally, this can be written as:

$$\alpha_{\text{opt}} = \arg\min_{\alpha} r_{\text{avg}}(\alpha)$$

In Table 3, an optimized set α_{opt} is marked with OS. As it can be seen, a set with better feature space separation leads to increased performance for CIFAR-100 and ImageNet, partly confirming our hypothesis. In CIFAR-10, however, the original set of parameters yields the best results. One explanation might be that, in CIFAR-10, the different classes are already well separated, such that optimization on this separation yields no improvement and leads to an overfitting behavior.

<u>Reduced Subset Selection:</u> For larger datasets, distance-based approaches like CoreSet (Sener & Savarese, 2018) or Badge (Ash et al., 2020) suffer from huge computational efforts, which is problematic for OOD detection, too. In Section 4, we suggested to use a reduced subset $\mathbb{T}$ of the comparison set $\mathbb{Z}_L$, selecting class-wise samples with the most neighbors in a given radius. For each dataset, we select a total of 10% of the samples for each class, drastically increasing inference speed. We compare the effect of our reduced subset selection (RS) in Table 3 and highlight it qualitatively in Fig. 4d. A comparison of the preprocessing steps for SISOM in Table 3 indicates that the AUROC near-OOD score has improved for all datasets. It can be observed that pre-selection enhances feature space separability based on the r_{avg} column. This also supports our hypothesis from the previous subsection. For both ImageNet and CIFAR-100, the combination

Table 3: Ablation Study on Optimal Sigmoid Steepness (OS) and Reduced Subset Selection (RS) on Near OOD Benchmarks.

Method	ImageNet		CIFAR 100		CIFAR 10	
	AUROC_n	r_{avg}	AUROC_n	r_{avg}	AUROC_n	r_{avg}
SISOM	77.21	0.270	75.93	0.33	<u>91.33</u>	0.26
SISOM , OS	**77.4**	0.266	<u>79.56</u>	0.19	90.37	0.099
SISOM , RS	77.33	0.249	76.07	0.31	**91.40**	0.24
SISOM , OS, RS	<u>77.37</u>	0.245	**79.69**	0.18	90.54	0.086

Table 4: Runtime comparison of different subset selection sizes for SISOMe in seconds. The setup time is the time to build up the set for comparison, while eval time measures one run over all OpenOOD evaluation datasets for CIFAR-100.

Subset Size $\mathbb{T}$	100%	60%	20%	5%
Setup Time [s]	291.17 ± 1.14	60.87 ± 0.93	40.63 ± 0.79	14.53 ± 0.92
Eval Time [s]	885.67 ± 3.31	688.34 ± 6.27	376.57 ± 1.60	251.67 ± 0.74
AUROC near OOD	80.97	81.01	81.03	81.10

of feature analysis and pre-selection yield increased performance. However, for CIFAR-10, the additional analysis of the feature space did not lead to improved performance. Considering the low average relevance score, the selected values may have overly constrained the feature space, resulting in an overfitting behavior. All parameters are detailed in Appendix B.3.

<u>Runtime Considerations:</u> In Table 4, we compare the runtime of SISOMe for different subset sizes for a full OpenOOD CIFAR-100 evaluation. It can be seen that with decreasing subset size, the computational efficiency is increasing for setup and evaluation time. Perhaps surprisingly, even the performance is slightly improving with decreasing subset size. One reason might be, that a smaller subset is a more compact, outlier-free representation of the data distribution. We also compare SISOM 's and SISOMe 's runtime to other OOD detection and AL methods in Appendix A.5. In general, approaches without distance calculations are faster than diversity-based methods such as SISOM .

6 Conclusion

We proposed SISOM , the first approach designed explicitly to solve OOD detection and AL jointly, providing an effective simplification in real-world application life cycles by eliminating an additional OOD detection design phase and avoiding conflicting goals of AL and OOD detection. By weighting latent space features with KL divergence of the neuron activations and relating them to the latent space clusters of the different classes SISOM achieves state-of-the-art performance in both tasks, *without* requiring a specific training scheme. In addition, SISOM provides a novel feature space analysis scheme enabling a post-training feature space refinement as well as a self-balancing uncertainty and diversity fusion introduced as SISOMe . Our experiments show that unlabeled samples and near-OOD data can be ambiguous, which SISOM can leverage. In the common OpenOOD benchmarks, SISOM archives the *top-1* performance in *one of the three benchmarks* and the is the only method with top-three places in all benchmarks for challenging near-OOD scenario. For AL, SISOM surpasses state-of-the-art approaches in eight different benchmark settings. While current state-of-the-art approaches are highly specialized for either AL or OOD detection, SISOM solves both tasks with the same approach. This versatility makes SISOM well-suited for real-world applications, like environment sensing, which face high label costs, abundant unlabeled data, and OOD samples at inference. In addition, SISOM provides important insights into the ambiguity of unlabeled and near-OOD samples.

In future work, we aim to leverage the relation of near-OOD data and unlabeled samples to explore tasks setup like open-set AL where currently two components are required. In addition, we aim extend SISOM to more complex tasks and investigate the effect batch diversification techniques on our two task approach.

References

Yong Hyun Ahn, Gyeong-Moon Park, and Seong Tae Kim. LINe: Out-of-Distribution Detection by Leveraging Important Neurons. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

Devansh Arpit, Huan Wang, Yingbo Zhou, and Caiming Xiong. Ensemble of averages: Improving model selection and boosting performance in domain generalization. In *Proceedings of the International Conference on Neural Information Processing Systems (NeurIPS)*, 2022.

Jordan T Ash, Chicheng Zhang, Akshay Krishnamurthy, John Langford, and Alekh Agarwal. Deep batch active learning by diverse, uncertain gradient lower bounds. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2020.

William H Beluch, Tim Genewein, Andreas Nürnberger, and Jan M Köhler. The power of ensembles for active learning in image classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

Julian Bitterwolf, Maximilian Mueller, and Matthias Hein. In or out? fixing imagenet out-of-distribution detection evaluation. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2023.

Feiyang Cai and Xenofon Koutsoukos. Real-time out-of-distribution detection in learning-enabled cyber-physical systems. In *2020 ACM/IEEE 11th International Conference on Cyber-Physical Systems (ICCPS)*, pp. 174–183, 2020. doi: 10.1109/ICCPS48487.2020.00024.

Razvan Caramalau, Binod Bhattarai, and Tae-Kyun Kim. Sequential graph convolutional network for active learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

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

Bertrand Charpentier, Oliver Borchert, Daniel Zügner, Simon Geisler, and Stephan Günnemann. Natural posterior network: Deep bayesian uncertainty for exponential family distributions. In *International Conference on Machine Learning (ICML)*, 2022.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *Proceedings of the 37th International Conference on Machine Learning*, pp. 1597–1607. PMLR, 2020.

M. Cimpoi, S. Maji, I. Kokkinos, S. Mohamed, and and A. Vedaldi. Describing Textures in the Wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009.

Andrija Djurisic, Nebojsa Bozanic, Arjun Ashok, and Rosanne Liu. Extremely simple activation shaping for out-of-distribution detection. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2022.

Pan Du, Suyun Zhao, Hui Chen, Shuwen Chai, Hong Chen, and Cuiping Li. Contrastive coding for active learning under class distribution mismatch. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021.

Souradeep Dutta, Michele Caprio, Vivian Lin, Matthew Cleaveland, Kuk Jin Jang, Ivan Ruchkin, Oleg Sokolsky, and Insup Lee. Distributionally robust statistical verification with imprecise neural networks. In *ACM International Conference on Hybrid Systems: Computation and Control*, 8 2025.

Sven Elflein, Bertrand Charpentier, Daniel Zügner, and Stephan Günnemann. On out-of-distribution detection with energy-based models. *arXiv*, 2107.08785, 2021.

Di Feng, Xiao Wei, Lars Rosenbaum, Atsuto Maki, and Klaus Dietmayer. Deep active learning for efficient training of a lidar 3d object detector. In *Proceedings of the IEEE Intelligent Vehicles Symposium (IV)*, 2019.

Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2016.

Wouter Van Gansbeke, Simon Vandenhende, Stamatios Georgoulis, Marc Proesmans, and Luc Van Gool. Scan: Learning to classify images without labels. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 5 2020.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

Aral Hekimoglu, Michael Schmidt, Alvaro Marcos-Ramiro, and Gerhard Rigoll. Efficient active learning strategies for monocular 3d object detection. In *Proceedings of the IEEE Intelligent Vehicles Symposium (IV)*, 2022.

Aral Hekimoglu, Michael Schmidt, and Alvaro Marcos-Ramiro. Monocular 3d object detection with lidar guided semi supervised active learning. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2024.

Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2016.

Dan Hendrycks, Andy Zou, Mantas Mazeika, Leonard Tang, Bo Li, Dawn Song, and Jacob Steinhardt. Pixmix: Dreamlike pictures comprehensively improve safety measures. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

Yen-Chang Hsu, Yilin Shen, Hongxia Jin, and Zsolt Kira. Generalized odin: Detecting out-of-distribution image without learning from out-of-distribution data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

Po Yu Huang, Wan Ting Hsu, Chun Yueh Chiu, Ting Fan Wu, and Min Sun. Efficient uncertainty estimation for semantic segmentation in videos. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018.

Rui Huang, Andrew Geng, and Yixuan Li. On the importance of gradients for detecting distributional shifts in the wild. In *Proceedings of the International Conference on Neural Information Processing Systems (NeurIPS)*, 2021.

Jinhan Kim, Robert Feldt, and Shin Yoo. Guiding Deep Learning System Testing Using Surprise Adequacy. In *Proceedings of the International Conference on Software Engineering (ICSE)*, 2019.

Jinhan Kim, Jeongil Ju, Robert Feldt, and Shin Yoo. Reducing DNN labelling cost using surprise adequacy: an industrial case study for autonomous driving. In *Proceedings of the Joint European Software Engineering Conference and Symposium on the Foundations of Software Engineering (ESEC/FSE)*, 2020.

Kwanyoung Kim, Dongwon Park, Kwang In Kim, and Se Young Chun. Task-aware variational adversarial active learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

Andreas Kirsch, Joost Van Amersfoort, and Yarin Gal. Batchbald: Efficient and diverse batch acquisition for deep bayesian active learning. In *Proceedings of the International Conference on Neural Information Processing Systems (NeurIPS)*, 2019.

Suraj Kothawade, Nathan Beck, Krishnateja Killamsetty, and Rishabh Iyer. Similar: Submodular information measures based active learning in realistic scenarios. *Advances in Neural Information Processing Systems*, 34:18685–18697, 2021.

Alex Krizhevsky, Vinod Nair, and Geoffrey Hinton. Learning multiple layers of features from tiny images. Technical report, Canadian Institute for Advanced Research, 2009. URL `http://www.cs.toronto.edu/~kriz/cifar.html`.

kuangliu. pytorch-cifar, 2021. URL `https://github.com/kuangliu/pytorch-cifar`. GitHub repository.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Proceedings of the International Conference on Neural Information Processing Systems (NeurIPS)*, 2017.

Ya Le and Xuan Yang. Tiny imagenet visual recognition challenge. Technical Report 7, Stanford Computer Vision Lab, 2015.

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. In *Proceedings of the IEEE*, 1998.

Kimin Lee, Kibok Lee, Honglak Lee, and Jinwoo Shin. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. In *Proceedings of the International Conference on Neural Information Processing Systems (NeurIPS)*, 2018.

Shiyu Liang, Yixuan Li, and R Srikant. Enhancing the reliability of out-of-distribution image detection in neural networks. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2018.

Zhihao Liang, Xun Xu, Shengheng Deng, Lile Cai, Tao Jiang, and Kui Jia. Exploring diversity-based active learning for 3d object detection in autonomous driving. *IEEE Transactions on Intelligent Transportation Systems*, 25(11):15454–15466, 2024.

Litian Liu and Yao Qin. Fast decision boundary based out-of-distribution detector. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2024.

Weitang Liu, Xiaoyun Wang, John D. Owens, and Yixuan Li. Energy-based Out-of-distribution Detection. In *Proceedings of the International Conference on Neural Information Processing Systems (NeurIPS)*, October 2020.

Xixi Liu, Yaroslava Lochman, and Christopher Zach. GEN: Pushing the Limits of Softmax-Based Out-of-Distribution Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

Yibing Liu, Chris Xing Tian, Haoliang Li, Lei Ma, and Shiqi Wang. Neuron Activation Coverage: Rethinking Out-of-distribution Detection and Generalization. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2024.

Alejandro López-Cifuentes, Marcos Escudero-Vinolo, Jesús Bescós, and Álvaro García-Martín. Semantic-aware scene recognition. *Pattern Recognition*, 102, 2020. Publisher: Elsevier.

Yadan Luo, Zhuoxiao Chen, Zijian Wang, Xin Yu, Zi Huang, and Mahsa Baktashmotlagh. Exploring active 3d object detection from a generalization perspective. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2023.

Carsten T Lüth, Till J Bungert, Lukas Klein, and Paul F Jaeger. Navigating the pitfalls of active learning evaluation: A systematic framework for meaningful performance assessment. In *Proceedings of the International Conference on Neural Information Processing Systems (NeurIPS)*, 2023.

Inbal Mishal and Daphna Weinshall. Dcom: Active learning for all learners. *Arvix*, 7 2024.

Jishnu Mukhoti, Andreas Kirsch, Joost Van Amersfoort, Philip H S Torr, and Yarin Gal. Deep deterministic uncertainty: A new simple baseline. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, and Andrew Y. Ng. Reading digits in natural images with unsupervised feature learning. In *Proceedings of the International Conference on Neural Information Processing Systems (NeurIPS) Workshop on Deep Learning and Unsupervised Feature Learning*, 2011.

Kun-Peng Ning, Xun Zhao, Yu Li, and Sheng-Jun Huang. Active learning for open-set annotation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 1 2022.

Julia Nitsch, Masha Itkina, Ransalu Senanayake, Juan Nieto, Max Schmidt, Roland Siegwart, Mykel J. Kochenderfer, and Cesar Cadena. Out-of-distribution detection for automotive perception. In *2021 IEEE International Intelligent Transportation Systems Conference (ITSC)*, pp. 2938–2943, 2021. doi: 10.1109/ITSC48978.2021.9564545.

Dongmin Park, Yooju Shin, Jihwan Bang, Youngjun Lee, Hwanjun Song, and Jae-Gil Lee. Meta-query-net: Resolving purity-informativeness dilemma in open-set active learning. In *Proceedings of the International Conference on Neural Information Processing Systems (NeurIPS)*, 10 2022.

Jaewoo Park, Yoon Gyo Jung, and Andrew Beng Jin Teoh. Nearest neighbor guidance for out-of-distribution detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 1686–1695, 2023a.

Younghyun Park, Wonjeong Choi, Soyeong Kim, Dong-Jun Han, and Jaekyun Moon. Active learning for object detection with evidential deep learning and hierarchical uncertainty aggregation. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2023b.

Fengchao Peng, Chao Wang, Jianzhuang Liu, Zhen Yang Noah, and Ark Lab. Active learning for lane detection: A knowledge distillation approach. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021.

Magesh Rajasekaran, Md Saiful Islam Sajol, Frej Berglind, Supratik Mukhopadhyay, and Kamalika Das. Combood: A semiparametric approach for detecting out-of-distribution data for image classification. In *Proceedings of the 2024 SIAM International Conference on Data Mining (SDM)*, pp. 643–651. SIAM, 2024.

Pengzhen Ren, Yun Xiao, Xiaojun Chang, Po-Yao Huang, Zhihui Li, Brij B Gupta, Xiaojiang Chen, and Xin Wang. A survey of deep active learning. *ACM computing surveys (CSUR)*, 54(9):1–40, 2021.

Bardia Safaei, Vibashan VS, Celso M. de Melo, and Vishal M. Patel. Entropic open-set active learning. In *AAAI*, 12 2024.

Sebastian Schmidt and Stephan Günnemann. Stream-based active learning by exploiting temporal properties in perception with temporal predicted loss. In *Proceedings of the British Machine Vision Conference (BMVC)*, 2023.

Sebastian Schmidt, Qing Rao, Julian Tatsch, and Alois Knoll. Advanced active learning strategies for object detection. In *Proceedings of the IEEE Intelligent Vehicles Symposium (IV)*, 2020.

Sebastian Schmidt, Lukas Stappen, Leo Schwinn, and Stephan Günnemann. Generalized synchronized active learning for multi-agent-based data selection on mobile robotic systems. *IEEE Robotics and Automation Letters*, 9(10):8659–8666, 2024. doi: 10.1109/LRA.2024.3444670.

Sebastian Schmidt, Leonard Schenk, Leo Schwinn, and Stephan Günnemann. Joint out-of-distribution filtering and data discovery active learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

Leo Schwinn, An Nguyen, René Raab, Leon Bungert, Daniel Tenbrinck, Dario Zanca, Martin Burger, and Bjoern Eskofier. Identifying untrustworthy predictions in neural networks by geometric gradient analysis. In *Conference on Uncertainty in Artificial Intelligence (UAI)*, 2021.

Ozan Sener and Silvio Savarese. Active learning for convolutional neural networks: A core-set approach. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 8 2018.

Burr Settles. Active learning literature survey. Technical report, University of Wisconsin–Madison, 2010.

Megh Shukla, Roshan Roy, Pankaj Singh, Shuaib Ahmed, and Alexandre Alahi. VL4Pose: Active Learning Through Out-Of-Distribution Detection For Pose Estimation. In *Proceedings of the British Machine Vision Conference (BMVC)*, 2022.

Samarth Sinha, Sayna Ebrahimi, and Trevor Darrell. Variational adversarial active learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019.

Yue Song, Nicu Sebe, and Wei Wang. Rankfeat: Rank-1 feature removal for out-of-distribution detection. In *Proceedings of the International Conference on Neural Information Processing Systems (NeurIPS)*, 2022.

Vladan Stojnić, Zakaria Laskar, and Giorgos Tolias. Training ensembles with inliers and outliers for semi-supervised active learning. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2024.

Yiyou Sun and Yixuan Li. DICE: Leveraging Sparsification for Out-of-Distribution Detection. In *Proceedings of the European Conference on Computer Vision (ECCV)*, November 2022.

Yiyou Sun, Chuan Guo, and Yixuan Li. React: Out-of-distribution detection with rectified activations. In *Proceedings of the International Conference on Neural Information Processing Systems (NeurIPS)*, 2021.

Yiyou Sun, Yifei Ming, Xiaojin Zhu, and Yixuan Li. Out-of-distribution detection with deep nearest neighbors. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2022.

Yuji Tokozume, Yoshitaka Ushiku, and Tatsuya Harada. Between-class learning for image classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

Grant Van Horn, Oisin Mac Aodha, Yang Song, Yin Cui, Chen Sun, Alex Shepard, Hartwig Adam, Pietro Perona, and Serge Belongie. The inaturalist species classification and detection dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

Sagar Vaze, Kai Han, Andrea Vedaldi, and Andrew Zisserman. Open-set recognition: A good closed-set classifier is all you need. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2021.

Haoqi Wang, Zhizhong Li, Litong Feng, and Wayne Zhang. Vim: Out-of-distribution with virtual-logit matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

weiaicunzai. pytorch-cifar100, 2022. URL `https://github.com/weiaicunzai/pytorch-cifar100`. GitHub repository.

Kai Xu, Rongyu Chen, Gianni Franchi, and Angela Yao. Scaling for training time and post-hoc out-of-distribution detection enhancement. In *The Twelfth International Conference on Learning Representations*, 2024. URL `https://openreview.net/forum?id=RDSTjtnqCg`.

Chenhongyi Yang, Lichao Huang, and Elliot J. Crowley. Plug and play active learning for object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 11 2024.

Jingkang Yang, Pengyun Wang, Dejian Zou, Zitang Zhou, Kunyuan Ding, WenXuan Peng, Haoqi Wang, Guangyao Chen, Bo Li, Yiyou Sun, Xuefeng Du, Kaiyang Zhou, Wayne Zhang, Dan Hendrycks, Yixuan Li, and Ziwei Liu. OpenOOD: Benchmarking generalized out-of-distribution detection. In *Proceedings of the International Conference on Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks Track*, 2022.

Yang Yang, Yuxuan Zhang, Xin Song, and Yi Xu. Not all out-of-distribution data are harmful to open-set active learning. In *Proceedings of the International Conference on Neural Information Processing Systems (NeurIPS)*, 2023.

Ofer Yehuda, Avihu Dekel, Guy Hacohen, and Daphna Weinshall. Active learning through a covering lens. In *Advances in Neural Information Processing Systems*, 2022a.

Ofer Yehuda, Avihu Dekel, Guy Hacohen, and Daphna Weinshall. Active learning through a covering lens. In *Proceedings of the International Conference on Neural Information Processing Systems (NeurIPS)*, 2022b.

John Seon Keun Yi, Minseok Seo, Jongchan Park, and Dong-Geol Choi. Pt4al: Using self-supervised pretext tasks for active learning. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2022.

Donggeun Yoo and In So Kweon. Learning loss for active learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

Yeonguk Yu, Sungho Shin, Seongju Lee, Changhyun Jun, and Kyoobin Lee. Block selection method for using feature norm in out-of-distribution detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

Beichen Zhang, Liang Li, Shijie Yang, Shuhui Wang, Zheng-Jun Zha, and Qingming Huang. State-relabeling adversarial active learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2018.

Jingyang Zhang, Jingkang Yang, Pengyun Wang, Haoqi Wang, Yueqian Lin, Haoran Zhang, Yiyou Sun, Xuefeng Du, Kaiyang Zhou, Wayne Zhang, Yixuan Li, Ziwei Liu, Yiran Chen, and Hai Li. OpenOOD v1.5: Enhanced benchmark for out-of-distribution detection. In *Proceedings of the International Conference on Neural Information Processing Systems (NeurIPS) Workshop on Distribution Shifts: New Frontiers with Foundation Models*, 2023.