PViT: Prior-Augmented Vision Transformer for Out-of-Distribution Detection

Anonymous authors
Paper under double-blind review

Abstract

Vision Transformers (ViTs) have achieved remarkable success over various vision tasks, yet their robustness against data distribution shifts and inherent inductive biases remain underexplored. To enhance the robustness of ViT models for image Out-of-Distribution (OOD) detection, we introduce a novel and generic framework named Prior-augmented Vision Transformer (PViT). We train PViT to predict class labels while taking as input both image tokens and the prior class logits from a pretrained model. During inference, PViT identifies OOD samples by quantifying the divergence between the predicted class logits and the prior logits obtained from pre-trained models. Unlike existing state-of-the-art(SOTA) OOD detection methods, PViT shapes the decision boundary between ID and OOD by utilizing the proposed prior guided confidence, without requiring additional data modeling, generation methods, or structural modifications. Extensive experiments on the large-scale ImageNet benchmark, evaluated against over seven OOD datasets, demonstrate that PViT significantly outperforms existing SOTA OOD detection methods in terms of FPR95 and AUROC.

1 Introduction

In recent years, the Transformer, characterized by its innovative attention mechanism, has achieved a significant success in various domains, extending its success from natural language processing to various vision tasks. The inception of the Vision Transformer (ViT) represents a pivotal moment in the adaptation of Transformer architectures for vision applications, setting the stage for subsequent models that exhibit remarkable performance enhancements through increased depth and scale, albeit at the cost of heightened computational demands Dosovitskiy et al. (2021). However, the exploration into enhancing Out-of-Distribution (OOD) detection within these architectures has lagged, especially when compared to the extensive research conducted for Convolutional Neural Networks (CNNs)-based models Xie et al. (2024).

OOD detection is a crucial machine learning technique that aims to identify test samples from distributions divergent from the training data distribution. This technique is essential for differentiating between inputs that are part of the training distribution and those that are not. The importance of proficient OOD detection is underscored in safety-critical real-world deployments, where encountering novel classes is inevitable. Addressing the generalization and OOD detection capabilities of ViTs becomes imperative.

We explore the strategic incorporation of prior knowledge in vision models to enhance their safety-related capabilities. Just as humans leverage contextual cues to identify unfamiliar objects, AI models can similarly benefit from prior information to improve their ability to accurately classify data and detect distribution shifts. This insight motivates us to investigate whether integrating prior knowledge from pre-trained models can strengthen OOD detection in ViTs.

In this work, we leverage the advantages of ViTs to design a scalable solution that improves their robustness and performance for OOD detection. Compared to traditional CNNs, ViTs directly operate on sequences of image patches, producing results through the *attention* mechanism. ViTs are effective at capturing long-range dependencies between patches but often neglect local feature extraction, as 2D patches are projected into vectors using a simple linear layer. Some recent studies have begun to focus on enhancing the modeling

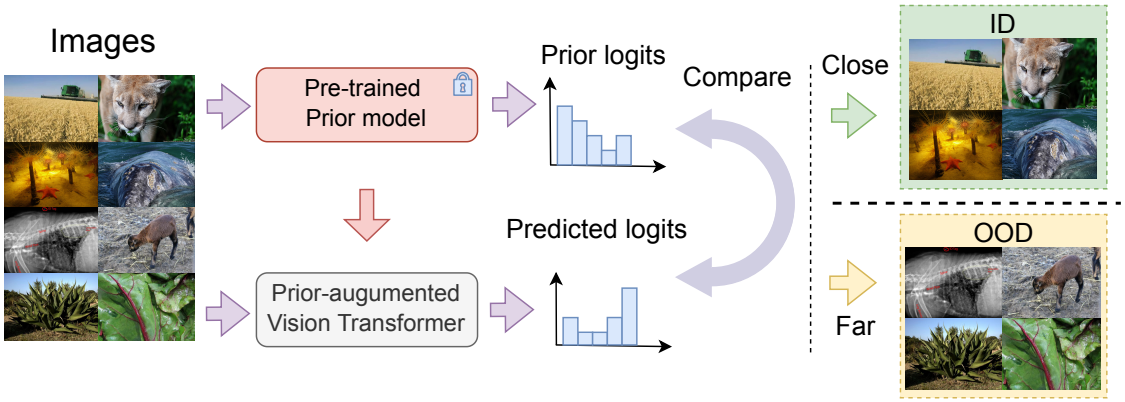


Figure 1. A brief overview of the proposed method. The ID images are taken from IMAGENET-1K, and the OOD images are sourced from OPENIMAGE-O. The distinction between ID and OOD images is made by measuring the difference between the prior logits and the predicted logits.

capacity for local information Liu et al. (2021). Enabling ViTs to incorporate additional contextual data broadens their analytical scope beyond the immediate visual input, creating a strategic opportunity to integrate prior knowledge derived from high-performing pre-trained vision models into the learning process Han et al. (2020).

Following the idea of integrating the prior knowledeg into a ViT, we propose the novel Prior-augmented Vision Transformer (PViT) for OOD detection. As illustrated in Fig. 1, PViT is designed to generate predictions that closely align with the prior logits for In-Distribution (ID) data while exhibiting significant divergence for OOD data. The prior knowledge is derived from a pre-trained model on the ID dataset, referred to as the *prior model* in this paper. PViT is trained using the prior predictions generated by the prior model. During inference, PViT employs the proposed *Prior Guide Energy* (PGE) score to effectively distinguish OOD instances by quantifying the divergence between the prior logits and the predicted logits.

We demonstrate that our proposed framework, PViT, is highly effective for OOD detection, particularly on large-scale datasets such as IMAGENET. Compared to state-of-the-art OOD detection methods, PViT achieves remarkable performance improvements, reducing the FPR95 by up to 20% and increasing the AUROC by up to 7% compared to the best baseline. Additionally, PViT eliminates the need for generating synthetic outlier data while maintaining high accuracy on ID datasets.

The key contributions of this paper are summarized as follows:

1. We introduce PViT, a novel and generic framework that integrates prior knowledge into ViT, thereby enhancing model robustness and OOD detection capabilities.
2. We introduce the Prior Guide Energy as an effective scoring method for OOD detection by measuring the similarity between the prior class logits and the predicted class.
3. We conduct comprehensive experiments on various benchmarks across a diverse set of ID and OOD datasets, providing qualitative analyses of PViT and offering insightful discussions on the impact of incorporating prior knowledge into ViT models.

The remainder of this paper is structured as follows. Section 2 reviews related work and provides the necessary background for our study. Section 3 presents a detailed description of our proposed PViT framework, including its architecture and the novel prior-guided OOD scoring mechanism. In Section 4, we evaluate the performance of PViT through extensive experiments on large-scale OOD benchmarks, including IMAGENET-1K and CIFAR. Section 4.2 presents comprehensive ablation studies with both quantitative and qualitative analyses. Section 5 discusses the implications of our approach, highlights its current limitations, and outlines potential directions for future research. Finally, Section 6 summarizes the main results

and contributions. The *Appendices* provide additional details on the datasets and the models used in this paper to ensure reproducibility.

2 Related Works

2.1 Out-of-Distribution (OOD) Detection

Deep learning models are often overconfident when classifying samples from different semantic distributions, leading to inappropriate predictions in tasks such as image classification and text categorization. This issue has prompted the emergence of the field of OOD detection, which requires models to reject inputs that are semantically different from the training distribution and should not be predicted by the model Hendrycks & Gimpel (2017). OOD detection is a critical area of research aimed at ensuring the safe deployment of AI systems. Over the years, various methods for OOD detection have emerged, broadly categorized into techniques focused on network modifications and score-based approaches to distinguish between ID and OOD samples in the embeddings or latent feature spaces Liang et al. (2018). Methods modifying network behavior often employ techniques like truncation. For example, ODIN Liang et al. (2018) perturbs the input using gradient vectors to amplify detection scores, while ReAct Sun et al. (2021) applies thresholding to clip hidden layer activations. These approaches enhance the network's ability to separate ID and OOD samples.

Score-based methods involve developing scalar metrics to quantify the likelihood of a sample being OOD. Classifier-based approaches, often referred to as *confidence scoring*, leverage the neural network's classification layer. A seminal work in this domain is the Maximal Softmax Probability (MSP) method, which serves as a baseline for OOD detection Hendrycks & Gimpel (2017). Subsequent advancements include the energy function Liu et al. (2020), which provides a bias-free estimation of class-conditional probabilities, and the maximum-of-logit technique Hendrycks et al. (2022), which combines class likelihood with feature magnitude for improved performance.

Distance-based methods form another key category of OOD detection, identifying samples based on their spatial relationship to ID data in the feature space. The Mahalanobis detector Lee et al. (2018) computes distances to class-wise means with shared feature covariance, while SSD assumes a single Gaussian distribution for ID samples Sehwag et al. (2021). Non-parametric methods like k-Nearest Neighbors (k-NN) offer precise boundary delineation and have been improved through NNGuide, which enhances differentiation in distant datasets Sun et al. (2022); Park et al. (2023). Apart from calculating the distance between samples and class centroids, feature norm that in the orthogonal complement space of the principal space is shown effective on OOD detection Wang et al. (2022b).

Other popular OOD detection methods include enhancing the model robustness by creating *outliers*, also referred to as outlier exposure approaches. These methods impose a strong assumption on the availability of OOD training data, which can be infeasible in practice. When no OOD samples are available, some methods attempt to synthesize OOD samples to enable ID/OOD separability. Existing works leverage GANs to generate OOD training samples and force the model predictions to be uniform, generate boundary samples in the low-density region, or produce high-confidence OOD samples. However, synthesizing images in the high-dimensional pixel space can be difficult to optimize. Recent work, VOS Du et al. (2022c), proposed synthesizing virtual outliers from the low-likelihood region in the feature space, which is more tractable given the lower dimensionality. In object detection, similar algorithm has been applied proposes synthesizing unknown objects from videos in the wild using spatial-temporal unknown distillation Du et al. (2022b). Recent advances focus on localizing OOD regions in complex visual environments, such as urban driving scenarios Du et al. (2022a). Such outlier exposure methods often require additional training and generating the synthesized data, which reduces scalability and adaptability. Compared to existing state-of-the-art methods, our proposed PViT distinguishes itself by eschewing reliance on synthesized data or external outliers for training, thereby enhancing its scalability and adaptability across diverse frameworks Du et al. (2023).

More recently, Vision-Language Model (VLM) based approaches have emerged as a promising direction for OOD detection Ming et al. (2022); Wang et al. (2023); Jiang et al. (2024); Miyai et al. (2025). These methods leverage the semantic understanding capabilities of pre-trained vision-language models such as

CLIP Radford et al. (2021) to distinguish ID from OOD samples without requiring any training on the target ID dataset. MCM Ming et al. (2022) computes maximum concept matching scores using CLIP's text encoder, while GL-MCM Miyai et al. (2025) extends this approach by incorporating both global and local concept matching. CLIPN Wang et al. (2023) trains additional "no" prompts to explicitly recognize OOD samples, and NegLabel Jiang et al. (2024) introduces negative labels for improved detection. While these VLM-based methods achieve strong performance, they require access to large pre-trained vision-language models, which may not always be available or computationally feasible. In contrast, PViT operates with standard vision-only models, offering a complementary approach for scenarios where VLM resources are limited.

2.2 Vision Transformers

Originally proposed for machine translation, Transformers have ascended to the state-of-the-art in numerous Natural Language Processing (NLP) tasks Vaswani et al. (2017). The vanilla ViT Dosovitskiy et al. (2021), representing the first adaptation of a purely Transformer-based model for image classification, has shown competitive performance with state-of-the-art CNNs. Like their NLP counterparts, ViTs lack the local receptive fields and weight-sharing properties of CNNs. Instead, they use positional encodings and self-attention to capture positional relationships.

Following the paradigm of ViT showing superior results on remarkable performance, a series of variants of ViT have been proposed to improve the performance on a variety of visual tasks, such as image classificationZhang et al. (2024), image segmentation Guo et al. (2024), and object detection Hua et al. (2025). DeiT Touvron et al. (2021), also known as Data-efficient image transformer, is later proposed as a competitive convolution-free transformer by training on only the ImageNet database. Swin Transformers Liu et al. (2021) performs local attention within a window and introduces a shifted window partitioning approach for cross-window connections.

Other than pure vision tasks, Transformers has also been used for Bayesian inference. A recent study exploring the use of Transformers for Bayesian Inference Müller et al. (2022) has broadened the scope of their applicability. This research demonstrates that when trained on prior samples, Transformers are capable of effectively approximating the posterior predictive distribution (PPD), even in scenarios involving small tabular datasets Hollmann et al. (2022). In contrast, our method is designed to work with large-scale image data, showcasing the versatility of taking advatages of prior information in handling diverse data scales and types.

While the introduction of specialized tokens in ViTs is a relatively unexplored area, our work pioneers the use of a *prior token* for OOD detection. The concept of a *prior token* is not first introduced by us, as seen in applications like MatteFormer Park et al. (2022) for image matting, which integrates trimap information via a Prior-Attentive Swin Transformer block. Our approach, however, diverges significantly as it repurposes this concept for enhancing OOD detection in ViTs.

3 Methodology

This section lays the foundation of our approach, starting with the problem setup (Section. 3.1) to establish the necessary background. We then offer a comprehensive overview of our PViT presented in Section. 3.2, detailing its architecture and functionality with Figure. 2. Finally, we explore our OOD scoring mechanism, particularly emphasizing the role of the PGE score in differentiating OOD instances in Section. 3.3.

3.1 Preliminaries

In the context of image classification, let $\mathcal{X} = \mathbb{R}^d$ represent the input space, and $\mathcal{Y} = \{1, \ldots, K\}$ denote the finite set of labels for K classes. The training dataset $\mathcal{D}_{\text{in}}^{\text{train}} = \{(\mathbf{x}_i, y_i)\}_{i=1}^{n}$ consists of pairs $(\mathbf{x}_i, y_i)$, where a classification function $f_\theta : \mathcal{X} \to \mathbb{R}^K$ predicts class scores. The predicted label $\widehat{y}$ is obtained as $\widehat{y} = \arg\max_k f_\theta^{(k)}(\mathbf{x})$, corresponding to the class with the highest score.

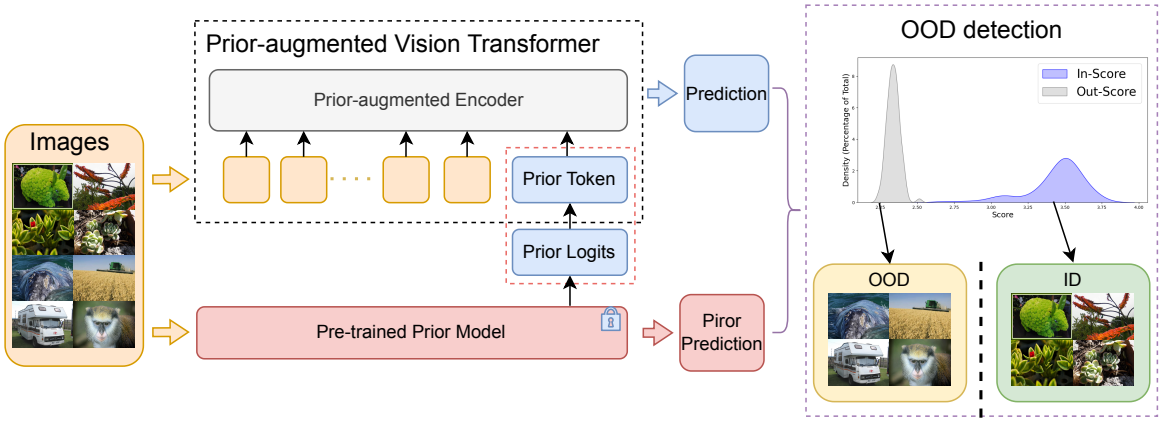


Figure 2. Framework of our proposed PViT. During the training stage, PViT processes the ID image patches $\mathcal{D}_{\text{in}}^{\text{train}}$ alongside the prior token $\mathbf{T}_{\text{prior}}$, which embeds prior knowledge from the pre-trained prior model. During testing, the prior model θ_{prior} continues to provide the prior logits for the OOD data $\mathcal{D}_{\text{out}}^{\text{test}}$ to PViT. The predicted class logits are then used to calculate the prior-guided OOD score, enabling the differentiation between ID and OOD data. Other components, including position embeddings, the classification (cls) token, and the flattening of image patches.

For testing on unseen data, the objective is to train a model capable of distinguishing OOD inputs $\mathcal{D}_{\text{out}}^{\text{test}} = \{(\mathbf{x}_j, y_j)\}_{j=1}^{m}$, where labels y_j do not belong to $\mathcal{Y}$. To achieve this, a binary classification approach, also referred to as the decision rule for OOD detection, is employed:

$$\mathbf{x} = \begin{cases} \text{ID} & \text{if } S(\mathbf{x};\theta) \geq \gamma, \\ \text{OOD} & \text{if } S(\mathbf{x};\theta) < \gamma, \end{cases} \tag{1}$$

where the threshold γ is selected to ensure high classification accuracy for ID data, typically set at 95%. The score $S(\mathbf{x};\theta)$, also known as 'confidence', represents the classifier-based detection score.

3.2 Prior-augmented Vision Transformer (PViT)

The architecture of the Prior-augmented Vision Transformer (PViT) is depicted in Fig. 2. The implementation of PViT follows the foundational structure of the vanilla Vision Transformer (ViT) Dosovitskiy et al. (2021). In the conventional ViT, an input image $\mathbf{x} \in \mathbb{R}^{H\times W\times C}$ is transformed into a sequence of flattened 2D patches $\mathbf{x}_p \in \mathbb{R}^{N\times(P^2\cdot C)}$. Here, (H, W) represents the resolution of the original image, C denotes the number of channels, (P, P) defines the resolution of each patch, and $N = \frac{HW}{P^2}$ indicates the total number of patches, which effectively determines the sequence length for the Transformer.

Similar to the vanilla ViT, a learnable embedding $\mathbf{z}_0^0$, initially set to $\mathbf{x}_{\text{class}}$, serves as the class embedding. This embedding is designed to capture the global image representation and is iteratively updated throughout the Transformer layers. After L layers of the Transformer encoder, its final state, $\mathbf{z}_L^0 \in \mathbb{R}^{1\times D}$, serves as the aggregated image representation, denoted by $\mathbf{y} \in \mathbb{R}^D$, where D is the dimensionality of the embedding space. Given an input image, the model first computes patch embeddings $\mathbf{E}_{\text{patches}} \in \mathbb{R}^{N\times D}$, where N represents the number of patches. A class token $\mathbf{t}_{\text{cls}} \in \mathbb{R}^{1\times D}$ is then prepended to this sequence of embeddings. Positional encodings $\mathbf{P}_{\text{pos}} \in \mathbb{R}^{(N+1)\times D}$ are added to provide the sequence with spatial information, resulting in the final sequence of embeddings $\mathbf{E}_{\text{pos}} = [\mathbf{t}_{\text{cls}}; \mathbf{E}_{\text{patches}}] + \mathbf{P}_{\text{pos}}$.

Prior Token Integration. Given a pre-trained *prior model* parameterized by θ_{prior}, the prior logits vector $\mathbf{p} \in \mathbb{R}^K$ represents the classification output logits, serving as prior knowledge for PViT. To incorporate this prior knowledge into the ViT architecture, we introduce a special token, termed the *prior token*. This token, $\mathbf{t}_{\text{prior}} \in \mathbb{R}^{1\times D}$, encapsulates the prior knowledge and is input alongside the patch tokens and the class token into the prior-augmented encoder, where it is processed by the attention mechanism.

To create the prior token, the logits vector $\mathbf{p} \in \mathbb{R}^K$ from the pre-trained classifier is first normalized using the softmax function. These normalized logits are then projected into the embedding dimension D to form the prior token $\mathbf{t}_{\text{prior}}$:

$$\mathbf{t}_{\text{prior}} = \mathbf{W}_{\text{proj}} \cdot \text{softmax}(\mathbf{p}), \tag{2}$$

where $\mathbf{W}_{\text{proj}} \in \mathbb{R}^{D \times K}$ is a learnable projection matrix designed to transform the class-wise priors into the embedding space, aligning them dimensionally with the patch embeddings.

The prior token is then scaled by a factor $\alpha \in \mathbb{R}$, a hyperparameter that modulates the influence of prior knowledge. This scaling balances the model's attention between the prior token and the image-derived embeddings, optimizing overall performance. The scaled prior token $\mathbf{t}_{\text{prior}}$ is then replicated across the batch, resulting in:

$$\mathbf{T}_{\text{prior}} = \alpha \cdot \mathbf{t}_{\text{prior}} \otimes \mathbf{1}_{B \times 1 \times D}, \tag{3}$$

where B denotes the batch size, and $\otimes$ represents the outer product with a vector of ones, effectively broadcasting $\mathbf{t}_{\text{prior}}$ across the batch. The batch-level prior token $\mathbf{T}_{\text{prior}}$ is appended to the positionally encoded sequence, forming the complete input $\mathbf{T} = [\mathbf{E}_{\text{pos}}; \mathbf{T}_{\text{prior}}]$ for the encoder.

The concatenated sequence $\mathbf{T}$ is processed through the Transformer encoder layers to yield the final representations. Our model follows the architecture of the vanilla ViT, employing multi-headed self-attention (MSA) (Eq. (5)) and multi-layer perceptron (MLP) blocks (Eq. (6)). Layer normalization (LN) is applied before each block (Eq. (7)), as described in the following equations:

$$\mathbf{Z}_0 = \mathbf{T}, \tag{4}$$

$$\mathbf{Z}'_\ell = \text{MSA}(\text{LN}(\mathbf{Z}_{\ell-1})) + \mathbf{Z}_{\ell-1}, \quad \ell = 1 \dots L, \tag{5}$$

$$\mathbf{Z}_\ell = \text{MLP}(\text{LN}(\mathbf{Z}'_\ell)) + \mathbf{Z}'_\ell, \quad \ell = 1 \dots L, \tag{6}$$

$$\mathbf{Y} = \text{LN}(\mathbf{Z}_L[0]), \tag{7}$$

where $\mathbf{Z}_L[0]$ represents the final layer's class token representation. The output $\mathbf{Y} \in \mathbb{R}^D$ serves as the input to a classifier head for the task at hand.

In the context of image classification, the primary training objective for PViT is to minimize the divergence between the model's predicted distribution and the true label distribution. The overall training objective is achieved through the minimization of the cross-entropy loss function $\mathcal{L}_{\text{CE}}$, which is formulated as:

$$\mathcal{L}_{\text{CE}} = -\sum_{i=1}^{K} \log P_{\text{PViT}}(y_i | x_i, \mathcal{D}, \pi; \theta), \tag{8}$$

where θ represents the parameters of PViT, y_i the true labels, x_i the input data, $\mathcal{D}$ the dataset, and π the prior information.

3.3 Prior Guide Energy for OOD Detection

Given a base confidence score function $S_{\text{base}}(\mathbf{x}; \theta)$, we propose a Prior Guide Energy (PGE) method to effectively differentiate between ID and OOD data by incorporating prior knowledge:

$$S_{\text{PGE}}(\mathbf{x}; \theta) = S_{\text{base}}(\mathbf{x}; \theta) \cdot G(\mathbf{x}; \theta, \theta_{\text{prior}}), \tag{9}$$

where $G(\mathbf{x}; \theta, \theta_{\text{prior}})$ is the *guidance term*, designed to measure the similarity between the prior embeddings and the outputs of PViT.

Energy as a Base Confidence Score. The base confidence score $S_{\text{base}}(\mathbf{x};\theta)$ is derived from the *Energy* score Liu et al. (2020), defined as:

$$E(\mathbf{x};\theta) = -\log \sum_{i=1}^{K} e^{f_i(\mathbf{x};\theta)}, \tag{10}$$

where $f_i(\mathbf{x};\theta)$ denotes the logit corresponding to class i output by the model. The effectiveness of the energy score for OOD detection arises from its relationship with the model's learned representation. The energy score is particularly effective for OOD detection due to the following properties:

- **Push-Pull Dynamics**: During training with the negative log-likelihood (NLL) loss, the energy of the correct label is minimized while the energies of incorrect labels are increased, creating a sharp confidence boundary between ID and OOD examples.
- **Free Energy Interpretation**: The energy score implicitly incorporates the *Free Energ* of the system (log partition function), allowing it to model the overall uncertainty of predictions across all classes.
- **Non-probabilistic Efficiency**: The score is efficiently computed via the `logsumexp` operator, making it computationally advantageous compared to probabilistic density estimates.

For ID data, the logits $f_i(\mathbf{x};\theta)$ form a concentrated distribution, resulting in low energy scores. Conversely, for OOD data, the logits tend to distribute more uniformly across all classes, leading to higher energy scores. This separation is a natural consequence of the NLL training objective, which explicitly pushes down the energy for ID data while increasing the energy for irrelevant classes. To align with the conventional OOD detection definition, we using the negative energy score, $-E(\mathbf{x};\theta)$, as $S_{\text{base}}(\mathbf{x};\theta)$.

Prior Guidance Term. While the energy score is a powerful standalone confidence measure, it can benefit from additional prior information to enhance its discriminative ability. To this end, we introduce the guidance term $G(\mathbf{x};\theta,\theta_{\text{prior}})$, which evaluates the similarity between prior knowledge and the predictions of the current model. Here we introduce one of the possible options: Cross Entropy (CE). CE is a widely used metric for quantifying the cost of matching multi-class predictions, and it is recognized as an effective and direct method for defining training targets in classification tasks. As demonstrated by the results presented in Section 4, CE emerges as the optimal guidance term by utilizing the prior logits and the predicted class as inputs:

$$G(\mathbf{x};\theta,\theta_{\text{prior}}) = -\sum_{i=1}^{K} y_i \log\left(q_i(\mathbf{x};\theta_{\text{prior}})\right), \tag{11}$$

where $q_i(\mathbf{x};\theta_{\text{prior}})$ is the probability of class i based on the prior logits, and y_i is the predicted class by PViT. This guidance term measures the dissimilarity between the model's predictions and the prior distribution. A higher cross entropy score indicates greater alignment with the prior distribution, suggesting that the data are likely ID, while a lower score suggests potential OOD data.

Alternative Guidance Terms. Beyond CE, the guidance term $G(\mathbf{x};\theta,\theta_{\text{prior}})$ can also be instantiated using other divergence measures between the prior logits and PViT's predictions:

- **Euclidean Distance (ED):** $G_{\text{ED}}(\mathbf{x}) = \sqrt{\sum_{i=1}^{K}(p_i - q_i)^2}$, where p_i and q_i are the corresponding elements from the prior and predicted logits vectors.
- **KL Divergence (KL):** $G_{\text{KL}}(\mathbf{x}) = \sum_{i=1}^{K} P(x_i) \log \frac{P(x_i)}{Q(x_i)}$, where P denotes the prior distribution and Q the predicted distribution.

These alternative guidance terms replace $G(\mathbf{x};\theta,\theta_{\text{prior}})$ in Eq. (9). As demonstrated in the ablation study (Section 4.2), CE yields the best overall performance and is therefore used as the default guidance term

throughout our experiments. Unless otherwise specified, all PViT results reported in this paper use CE as the guidance term.

Overall PGE Score. By combining the base confidence score and the prior guidance term, the PGE score is defined as:

$$S_{\text{PGE}}(\mathbf{x};\theta) = \underbrace{S_{\text{energy}}(\mathbf{x};\theta)}_{\uparrow \text{ for in-distribution } \mathbf{x}} \cdot \underbrace{G(\mathbf{x};\theta,\theta_{\text{prior}})}_{\uparrow \text{ for in-distribution } \mathbf{x}} . \tag{12}$$

The guidance term amplifies the base confidence score, resulting in a higher overall PGE score for ID data. Conversely, OOD data are characterized by lower PGE scores. By setting an appropriate threshold γ, the PGE score can effectively separate ID and OOD data:

$$\begin{cases} S_{\text{PGE}}(\mathbf{x};\theta) > \gamma, & \text{if } \mathbf{x} \in \mathcal{D}_{\text{in}}, \\ S_{\text{PGE}}(\mathbf{x};\theta) \leq \gamma, & \text{if } \mathbf{x} \in \mathcal{D}_{\text{out}}. \end{cases} \tag{13}$$

The next section presents a detailed description of the datasets, evaluation metrics and the overall evaluation of the proposed OOD detection approach.

4 Experiments

Datasets. To assess the model performance, we use the large-scale ImageNet-1K Deng et al. (2009) and the small-scale Cifar Krizhevsky (2009) datasets as our ID training datasets. We use the standard train/validation/test splits for training and testing. In the main results reported in Tab. 2 and Tab. 1 where ImageNet-1K is used as ID data, we employ a range of natural image datasets as OOD benchmarks, including iNaturalist Van Horn et al. (2018), Sun Xiao et al. (2010), Textures Cimpoi et al. (2014), Places Zhou et al. (2017), NINCO Bitterwolf et al. (2023), OpenImage-O Wang et al. (2022a), and SSB-hard Vaze et al. (2022). In Tab. 4 where Cifar-100 is used as the ID dataset, Cifar-10 and Cifar-100 are used interchangeably as ID and OOD datasets due to their similarities yet distinct characteristics. The following OOD test datasets are used to evaluate PViT: Cifar-10 Krizhevsky (2009), Textures Cimpoi et al. (2014), Places Zhou et al. (2017), Sun Xiao et al. (2010), and Svhn Netzer et al. (2011).

Training Details. The PViT is trained with a configuration that includes a hidden dimension of 384, a depth of 12 layers, 6 MSA heads, and a MLP dimension of 768. The Adam optimizer is used with hyperparameters $\beta_1 = 0.9$ and $\beta_2 = 0.999$ over 20 epochs. Training begins with an initial learning rate of 0.1 and employs a batch size of 256, a momentum of 0.9, and a weight decay of 1×10^{-3}. A linear learning rate decay schedule is applied after 5 warm-up epochs. For different ID datasets, we utilize various prior models, including ResNet He et al. (2016) and ViT models Dosovitskiy et al. (2021) along with their variants Touvron et al. (2021); Singh et al. (2022). For ImageNet-1K as the ID dataset, the ViT models are pre-trained on ImageNet-21K and subsequently fine-tuned on ImageNet-1K. All pre-trained models used as prior models are publicly available.

Evaluation Metrics. For assessing the performance of our proposed models in OOD detection, we employ two evaluation metrics: (1) **FPR95**, which measures the false positive rate of OOD samples when the true positive rate for ID samples is at 95%; (2) **AUROC**, which computes the Area Under the Receiver Operating Characteristic Curve.

4.1 Evaluation on OOD Detection

We evaluate PViT for OOD detection against competitive baselines, including MSP Hendrycks & Gimpel (2017), MaxLogit score Hendrycks et al. (2022), Mahalanobis score Lee et al. (2018), Energy score Liu et al.

OOD Datasets	iNaturalist		SUN		Places		Textures		NINCO		OpenImage_O		SSB-hard		Mean	
Metrics	FPR95	AUROC	FPR95	AUROC	FPR95	AUROC	FPR95	AUROC	FPR95	AUROC	FPR95	AUROC	FPR95	AUROC	FPR95	AUROC
Methods	↓	↑	↓	↑	↓	↑	↓	↑	↓	↑	↓	↑	↓	↑	↓	↑
MSP	51.52	88.16	66.52	80.94	68.67	80.38	60.23	82.99	73.26	78.08	59.93	84.81	69.02	84.73	64.16	82.87
MaxLogit	52.26	85.24	66.88	76.39	69.14	75.06	56.68	81.69	74.02	72.38	58.75	81.55	64.43	85.64	63.17	79.71
Mahalanobis	21.28	95.59	61.55	85.12	61.48	83.92	53.94	87.35	58.74	85.98	44.03	91.73	70.85	76.58	53.12	86.61
Energy	64.08	79.24	72.77	70.28	74.31	68.44	58.48	79.30	78.46	66.03	64.93	76.46	59.13	88.16	67.45	75.42
SSD	25.58	94.21	65.75	81.04	66.58	79.90	53.35	85.28	61.93	82.89	48.23	90.14	66.57	80.58	55.43	84.86
ViM	17.12	96.45	58.87	83.14	60.52	81.24	48.09	88.17	56.09	84.98	41.87	91.93	68.92	76.31	50.21	86.03
KNN	71.35	85.55	79.51	79.48	78.61	78.27	68.58	82.76	78.94	76.73	66.11	85.86	59.56	**90.23**	71.81	82.70
NNGuide	66.93	88.57	76.98	81.94	76.33	80.61	64.84	85.68	77.56	80.14	65.42	87.93	63.76	89.19	70.26	84.87
PViT	**2.26**	**99.39**	30.98	**93.38**	40.35	90.99	**32.96**	**92.05**	41.63	90.74	**16.59**	**96.58**	58.32	84.49	31.87	**92.52**
PViT + ED	3.27	99.08	**30.85**	93.00	**38.24**	**91.28**	33.30	90.63	**38.23**	**90.93**	18.78	95.81	**56.78**	84.72	**31.35**	92.21

Table 1. OOD detection results for PViT using the pretrained DeiT Touvron et al. (2021) as the prior model. All benchmarks are evaluated on the same prior model with an ID accuracy of **81.07%**, while PViT achieves an ID accuracy of **81.33%**. OOD detection results are reported for IMAGENET-1K as the ID data. **Bold** numbers indicate superior results.

(2020), SSD Sehwag et al. (2021), ViM Wang et al. (2022b), KNN Sun et al. (2022), and NNGuide Park et al. (2023). For fairness, we exclude synthesis-based OOD methods (e.g., VOS Du et al. (2022c) and Dream-OOD Du et al. (2023)), since PViT can be combined with synthesis-based training. Unless otherwise specified, we set the prior-token scaling factor α to 0.1 and use Cross Entropy (CE) as the default guidance term in Eq. (9). Results denoted as "PViT" in the tables use CE guidance; where Euclidean Distance (ED) is used instead, it is explicitly indicated as "PViT + ED".

OOD Datasets	iNaturalist		SUN		Places		Textures		NINCO		OpenImage_O		SSB-hard		Mean	
Metrics	FPR95	AUROC	FPR95	AUROC	FPR95	AUROC	FPR95	AUROC	FPR95	AUROC	FPR95	AUROC	FPR95	AUROC	FPR95	AUROC
Methods	↓	↑	↓	↑	↓	↑	↓	↑	↓	↑	↓	↑	↓	↑	↓	↑
MSP Hendrycks & Gimpel (2017)	15.33	97.19	46.48	90.12	53.34	87.89	42.82	89.61	51.58	88.31	27.85	94.98	67.41	80.98	43.55	89.87
MaxLogit Hendrycks et al. (2022)	36.96	94.00	77.03	81.56	80.82	78.65	65.20	84.01	81.15	75.57	68.05	86.29	92.28	63.62	71.64	80.53
Energy Liu et al. (2020)	31.18	92.64	61.58	82.66	64.99	81.83	59.61	81.07	63.80	82.82	44.86	89.03	79.21	73.03	57.89	83.30
SSD Sehwag et al. (2021)	18.17	96.65	50.50	89.36	55.97	87.35	47.07	88.69	54.12	87.88	31.41	94.31	71.02	80.27	46.89	89.22
ViM Wang et al. (2022b)	70.91	80.81	93.42	60.38	95.42	56.20	85.46	69.19	92.85	53.42	87.02	64.36	92.98	50.81	88.30	62.17
KNN Sun et al. (2022)	78.14	74.68	91.20	59.32	94.04	55.43	89.34	63.16	93.33	52.17	91.21	59.42	94.61	48.49	90.27	58.95
NNGuide Park et al. (2023)	81.44	76.72	94.97	51.18	95.88	49.40	85.20	68.99	94.85	55.80	92.75	60.46	96.71	43.87	91.68	58.06
PViT	**13.08**	**97.56**	41.56	90.99	49.97	88.40	**39.56**	**90.43**	**49.20**	**88.33**	**25.62**	**95.26**	61.31	**82.19**	**40.04**	**90.45**
PViT + ED	18.03	96.18	**28.75**	**91.85**	**37.01**	**90.24**	78.01	79.42	51.02	87.08	49.12	88.92	**60.23**	80.86	46.02	87.79

Table 2. OOD detection results for PViT using the original ViT-B/16 as the prior model. All benchmarks are evaluated on the same prior model with an ID accuracy of **75.73%**, while PViT achieves an ID accuracy of **76.61%**. OOD detection results are reported for IMAGENET-1K as the ID dataset. Arrows (↑ and ↓) indicate that larger or smaller values are better, respectively. All values are percentages. **Bold** numbers indicate superior results. **CE** refers to using cross entropy as the prior guidance, and **ED** refers to using Euclidean Distance as the prior guidance.

As shown in Tab. 2 and Tab. 1, PViT is evaluated across a wide range of seven OOD datasets to demonstrate its superior performance. Our experiments show that PViT exhibits remarkable performance on large-scale ID datasets ImageNet-1k, particularly when utilizing ViT-based prior models. Notably, even when compared to the Mahalanobis and ViM detectors, both of which are known for their strong performance in ViT-based architectures due to the Gaussian nature of the vision transformer embedding space, PViT significantly outperforms these benchmarks in both the prior models of vanilla ViT variants.

Comparison with Vision-Language Model (VLM) Based Methods. Recent advances in VLM-based OOD detection methods Ming et al. (2022); Wang et al. (2023); Jiang et al. (2024); Miyai et al. (2025) have shown promising results by leveraging the zero-shot capabilities of CLIP Radford et al. (2021). To provide a comprehensive comparison, we evaluate PViT against these state-of-the-art VLM-based methods on a standardized benchmark of four OOD datasets (iNaturalist, SUN, Places, and Textures), as shown in Tab. 3.

It is important to note the fundamental difference in methodology: VLM-based methods rely on pre-trained vision-language models (e.g., CLIP ViT-B/16) that have been trained on large-scale image-text pairs, providing inherent semantic understanding for distinguishing ID from OOD samples. In contrast, PViT operates without requiring such external knowledge, using only the prior predictions from a standard ViT model trained on the ID dataset.

Method	Venue	iNaturalist		SUN		Places		Textures		Mean	
		FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑
Vision-Language Model (VLM) based methods (CLIP ViT-B/16)											
MCM Ming et al. (2022)	NeurIPS'22	30.91	94.61	37.59	92.57	44.69	89.77	57.77	86.11	42.74	90.77
GL-MCM Miyai et al. (2025)	IJCV'25	17.42	96.44	30.75	93.44	37.62	90.63	55.20	85.54	35.25	91.51
CLIPN-A Wang et al. (2023)	ICCV'23	23.94	95.27	26.17	93.93	33.45	92.28	40.83	90.93	31.10	93.10
NegLabel Jiang et al. (2024)	ICLR'24	**1.91**	**99.49**	**20.53**	**95.49**	**35.59**	91.64	43.56	90.22	**25.40**	**94.21**
Mysteries Jara-Rodriguez et al. (2025)	NeurIPS'25	-	-	-	-	-	-	-	-	30.70	93.54
PViT (Ours) - Non-VLM method											
PViT (ViT-B/16)	-	13.08	97.56	41.56	90.99	49.97	88.40	**39.56**	**90.43**	36.04	91.85
PViT (DeiT)	-	2.26	99.39	30.98	93.38	40.35	**90.99**	32.96	92.05	26.64	93.95

Table 3. Comparison with VLM-based OOD detection methods on four standardized OOD datasets. VLM-based methods use CLIP ViT-B/16, while PViT uses standard ViT models without vision-language pre-training. **Bold** numbers indicate the best results in each column. Results for VLM methods are from their respective papers.

Methods	CIFAR10		Textures		Places		SUN		SVHN		Mean	
	FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑
MSP	79.98	78.68	80.02	78.14	82.52	73.34	74.15	82.06	79.46	80.20	78.44	79.17
MaxLogit	**79.36**	**79.35**	78.76	79.33	82.17	73.53	70.45	84.72	77.69	83.20	76.88	80.90
Energy	79.62	79.17	77.75	79.30	82.62	73.36	67.35	85.12	76.82	83.52	76.06	80.98
SSD	92.94	55.31	79.04	71.07	92.41	58.98	77.66	71.67	90.39	68.12	87.77	63.47
ViM	93.57	54.50	82.73	66.53	92.61	56.53	77.59	68.25	85.64	69.28	87.69	61.45
KNN	99.96	31.84	99.50	49.82	99.15	42.16	99.80	46.62	99.92	48.26	99.72	42.87
NNGuide	80.02	78.46	**75.85**	**80.23**	**81.52**	**74.46**	**62.36**	**86.44**	80.64	82.94	**75.20**	**81.39**
PViT	81.68	78.99	78.72	79.10	87.12	71.17	65.02	85.09	**73.87**	**83.57**	79.77	79.25
PViT + KL	84.74	77.29	86.79	76.04	90.68	70.64	75.80	80.59	84.17	79.08	85.20	76.96
PViT + ED	86.88	76.89	81.28	77.90	83.09	71.32	71.73	83.63	79.68	81.04	81.90	77.94

Table 4. OOD detection results for PViT with ResNet18 as the prior model, where CIFAR-100 as the ID dataset. The ID accuracy of the prior model ResNet18 on CIFAR-100 is **77.27%**. PViT achieves ID accuracies of **78.84%**. **Bold** numbers indicate superior results.

Despite this methodological disadvantage, PViT with DeiT as the prior model achieves competitive performance (FPR95 = 26.64%, AUROC = 93.95%), approaching the state-of-the-art NegLabel Jiang et al. (2024) (FPR95 = 25.40%, AUROC = 94.21%) while significantly outperforming MCM Ming et al. (2022) (FPR95 = 42.74%, AUROC = 90.77%) and GL-MCM Miyai et al. (2025) (FPR95 = 35.25%, AUROC = 91.51%). This demonstrates that PViT provides a strong alternative for scenarios where vision-language pre-training is not available or when computational resources for VLM inference are limited.

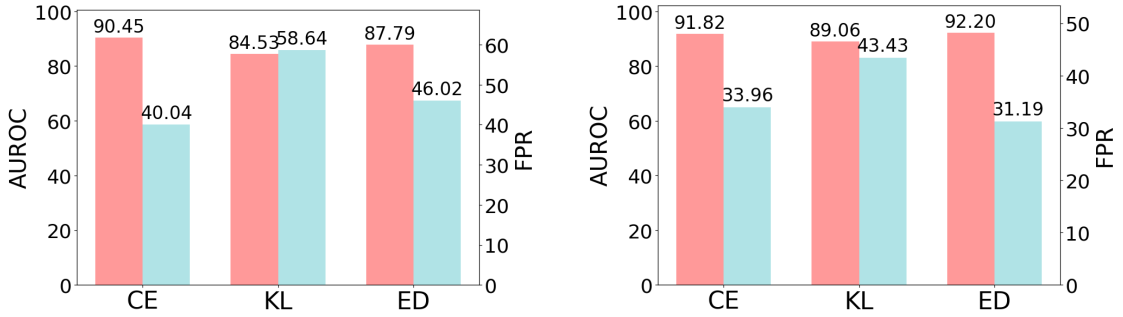
To further assess the robustness of PViT, we evaluate its performance on the small-scale CIFAR-100 dataset as the ID data, as shown in Table 4. Although PViT does not consistently outperform all baseline methods across every metric, it remains highly competitive, ranking among the top performers on average over five OOD datasets. The slightly lower performance relative to some baselines is likely attributable to two factors. First, the limited number of training samples in the small-scale CIFAR-100 dataset may hinder optimal model learning. Second, the necessity to upscale images from 32×32 to 224×224 to maintain consistency with PViT's configuration may introduce artifacts that affect detection performance.

4.2 Ablation Studies

Ablation Study on OOD Prior Guidance.

Although we have introduced using the CE as the guidance term in Eq. (11) for detecting OOD instances, we also considered other metrics to measure the difference between the priors and the predicted logits: 1) Euclidean Distance (ED). Euclidean Distance is a geometric measure calculating the "straight-line" distance between two points in Euclidean space. For vectors of predicted logits and prior probabilities, it is computed as $\sqrt{\sum(p_i - q_i)^2}$, with p_i and q_i being the corresponding elements from the prior and predicted logits vectors, respectively. 2) Kullback–Leibler Divergence (KL-Divergence), used for measuring the distance between two probability distributions, is defined as $\mathrm{KL}(P \parallel Q) = \sum P(x) \log \frac{P(x)}{Q(x)}$, where P represents the true distribution of data (priors in our context) and Q denotes the distribution inferred by the model (predicted logits).

As visualized in Fig. 4, the score distributions reveal that our prior-guided energy score better distinguishes between ID and OOD data compared to the original energy score. Both CE and ED guidance terms produce



(a) Results of ViT-B/16 as the prior model.

(b) Results of using ViT-LP as the prior model.

Figure 3. Ablation study on different scoring rules of PViT. The ID data is IMAGENET-1K. The results are average results over seven OOD datasets in Table. 2.

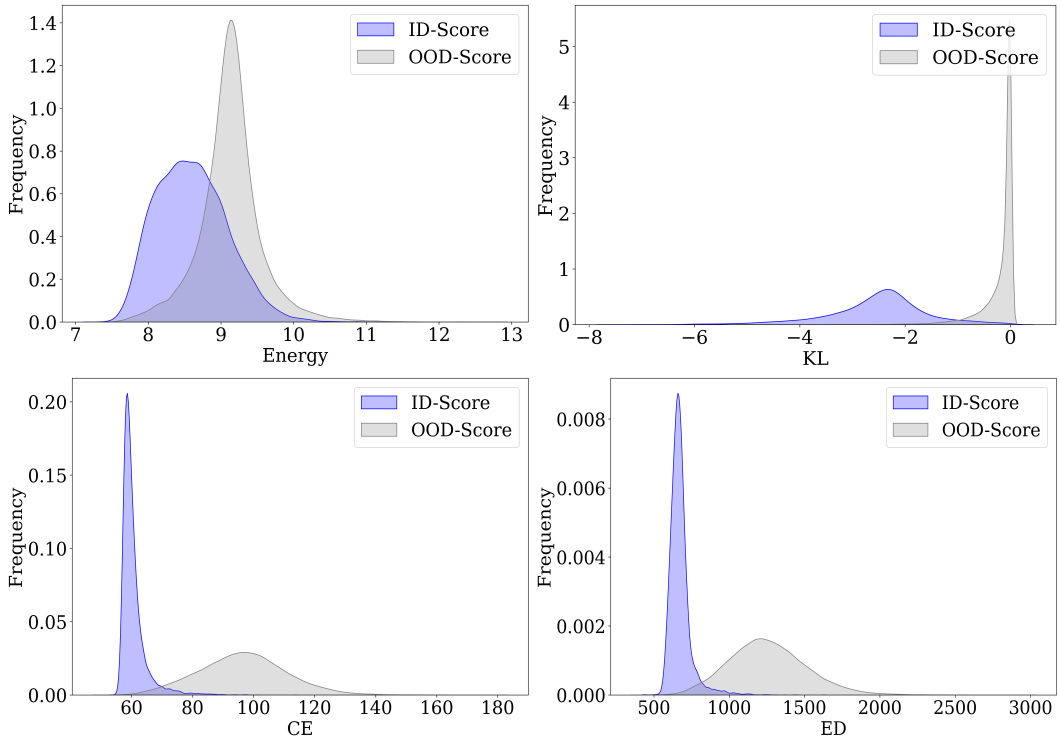
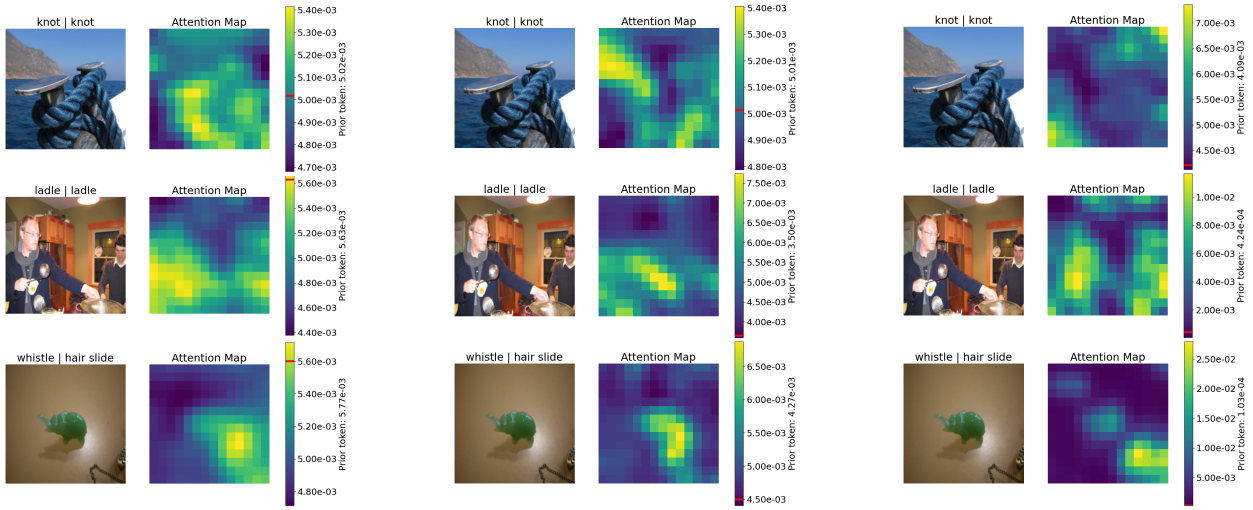


Figure 4. Score distributions with IMAGENET-1K as ID data and iNATURALIST as OOD data. The scores are calculated by PViT with ViT-LP as the prior model.

similar results, albeit with different score values. The performance comparison in Fig. 3 further demonstrates that ED achieves performance comparable to that of the CE as the guidance.

Ablation Study on Effect of Priors.

To demonstrate the efficacy of our integrated prior token in guiding PViT to effectively differentiate between ID and OOD data, we conduct an ablation study comparing PViT with a vanilla ViT model. Specifically, we evaluate the OOD detection performance both without and with prior knowledge integration. In the "w/o Prior" setting, we directly calculate the divergence between a vanilla ViT model and the prior models without any prior token integration. In the "w/ Prior" setting, PViT is trained with integrated prior tokens as described in Sec. 3. The results are presented in Tab. 5.



(a) $\alpha = 10$. (b) $\alpha = 1$. (c) $\alpha = 0.1$.

Figure 5. Visualization of attention maps with varying scaling factors α for prior token embedding, generated from the last layer and the first MSA head. The attention weight for the prior token is highlighted with a red line on the color bar. The first two rows of figures are taken from IMAGENET-1K, representing ID data. The third row, representing the OOD data, which is taken from OPENIMAGE_O, illustrates the differential responses of PViT to both ID and OOD data. Labels above the original figure on the **left** indicate predictions made by the prior model, while the labels on the **right** correspond to the predictions made by PViT.

As shown in Tab. 5, the performance of the vanilla ViT without prior integration is significantly inferior across all three types of prior guidance. Without prior tokens, OOD detection using CE as prior guidance achieves nearly 100% FPR95 on the mean of seven OOD datasets. Similarly, the results obtained with KL and ED as prior guidance are also suboptimal. In stark contrast, PViT with integrated prior tokens (highlighted rows) achieves dramatically improved performance. For instance, with vit-b-16 as the prior model, PViT reduces the mean FPR95 from 99.81% (w/o Prior, CE) to 31.87% (w/ Prior, CE), while improving AUROC from 18.32% to 92.52%. Similar improvements are observed with vit-lp as the prior, where PViT reduces the mean FPR95 from 99.81% to 40.04%. These findings underscore that the ViT architecture, in its vanilla form, lacks the inherent capacity to effectively differentiate between OOD and ID data, and that the integration of prior tokens is crucial for PViT to achieve robust OOD detection.

Ablation Study on Scaling the Priors. Fig. 5 illustrates the impact of scaling prior weights α on PViT's output. Our analysis aims to mitigate the excessive influence of priors. It is observed that a higher α leads to more attention on the prior token, potentially reducing the focus on image patches. Conversely, a lower α may enhance attention on image patches. Optimal performance is achieved when PViT balances its focus between the prior token and image tokens. This balance results in a clearer distinction between priors and predictions, underpinning the effectiveness of PViT in OOD detection.

To further investigate the trade-offs in attention weights among patch tokens, we visualize the attention map in Fig. 5, using image examples with corresponding prior and predicted labels. The color bar in these maps ranges from highest to lowest attention, highlighting the impact of the scaling factor α on the model's attention mechanism. Notably, the attention weight for the prior token, indicative of PViT's focus on the prior token, is demarcated with a red line on the color bar. As a result of OOD detection, we can see that in the third row of figures, PViT often produces predictions that notably differ from the prior model's.

5 Theoretical Insights

5.1 Bayesian Explanation

Bayesian Neural Networks (BNNs) have been explored for OOD detection in various studies. BNNs utilize probability distributions over model parameters to represent uncertainties in predictions. In the context of OOD detection, BNNs can be employed by comparing uncertainties between the model's predictions on given inputs and known ID data. However, the suitability of BNNs for OOD detection has been a subject of debate in recent works Henning et al. (2021).

From a Bayesian perspective, our approach can be interpreted as utilizing prior knowledge from an ID dataset to establish the Predictive Posterior Distribution (PPD) within a Bayesian framework. This aligns with the concept of Transformers facilitating Bayesian inference Müller et al. (2022). In our model, the priors can be treated akin to mean values of sampled priors from a Bayesian model, positioning PViT to approximate the posterior distribution $P(y|x, \mathcal{D}, \pi)$. This approximation follows the equation $p(y|x, \mathcal{D}) = \int_t p(y|x, t)p(t|\mathcal{D})$, where $P(x|y, \mathcal{D}, \pi)$ denotes the likelihood of observing the data given the label and priors, $P(y|\mathcal{D}, \pi)$ represents the prior probability informed by the dataset and the prior model, and $P(x|\mathcal{D}, \pi)$ signifies the evidence, usually computed by marginalizing over the label space. Notably, the Transformer-based Bayesian inference approach Müller et al. (2022) is specifically designed for single-sequence data, aimed at providing ultra fast Bayesian inference in a single forward pass Hollmann et al. (2022). This approach, while not directly applicable to image data due to ViTs' processing limitations, sheds light on the efficacy of our PViT in OOD detection. By capturing uncertainties through Bayesian inference, it provides a compelling explanation for PViT's robust performance in identifying OOD samples.

5.2 Inductive Bias

Gernarally, ViTs and CNNs are believed that they exhibit fundamentally different inductive biases. ViTs inherently focus on global image patterns by treating image patches as analogous to tokens. This global perspective contrasts sharply with the local feature emphasis of CNNs, which inherently encode a bias towards local spatial hierarchies and proximities. While this enables ViTs to excel in tasks requiring holistic image comprehension, their lack of built-in locality bias may limit effectiveness in tasks where detailed local feature analysis is crucial Xu et al. (2021).

ViTs can adopt inductive biases through data augmentation or hybrid architectures, improving their local feature processing, traditionally a strength of CNNs. Our study introduces embedding additional prior information into ViTs, enhancing their robustness and serving as a method to incorporate inductive bias. This strategy marks a new path for embedding manually designed inductive biases into ViTs, potentially boosting their robustness and explainability. By introducing additional prior information, we hope our PViT can compensate for the less pronounced traditional inductive biases in ViTs, where such augmentation can significantly refine their performance, particularly in challenging scenarios where inductive biases play a crucial role.

6 Conclusions and Future Work

In this work, we present Prior-augmented Vision Transformer, a novel and generic framework for OOD detection. PViT uniquely integrates prior knowledge as a prior token to be trained to approximate the true label, allowing for effective differentiation of OOD data by examining the relative distances between model predictions and the prior logits. Our empirical results demonstrate that PViT achieves outstanding performance in OOD detection benchmarks. Moreover, the innovative integration of prior knowledge by PViT not only enhances OOD detection capabilities but also suggests a versatile approach for the strategic planning and control of large vision models tailored to specific practical applications.

Limitations. Despite its promising performance, PViT does have certain limitations. Its accuracy is closely tied to the quality and structure of the prior model—particularly the ID prior accuracy. Additionally, while training on ID data enables rapid convergence and leverages prior knowledge, it also adds complexity to the

training process. Moreover, during inference, the requirement to process both the prior model and PViT increases computational costs.

Future Work. In future work, we plan to further explore and extend the capabilities of PViT. Our study demonstrates that PViT naturally embeds a beneficial inductive bias into large vision models, which representing the class of AI models that is both rapidly expanding and evolving. As the field of computer vision increasingly mirrors the trajectory of Large Language Models (LLMs), these vision models are now tasked with the challenge of "planning", meaning they must direct their capabilities towards specific, controlled outcomes. Looking ahead, we aim to incorporate varying levels of prior knowledge into PViT to enable large vision models to be fine-tuned for customized objectives and to enhance OOD detection across diverse scenarios. Additionally, PViT shows promise for improving state-of-the-art architectures like Swin Transformers, potentially unlocking further innovations in computer vision.

References

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.

Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3606–3613, 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 IEEE Conference on Computer Vision and Pattern Recognition*, 2009.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *Proceedings of International Conference on Learning Representations*, 2021.

Xuefeng Du, Gabriel Gozum, Yifei Ming, and Yixuan Li. Siren: Shaping representations for detecting out-of-distribution objects. In *Proceedings of Advances in Neural Information Processing Systems*, 2022a.

Xuefeng Du, Xin Wang, Gabriel Gozum, and Yixuan Li. Unknown-aware object detection: Learning what you don't know from videos in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022b.

Xuefeng Du, Zhaoning Wang, Mu Cai, and Yixuan Li. Vos: Learning what you don't know by virtual outlier synthesis. In *Proceedings of the International Conference on Learning Representations*, 2022c.

Xuefeng Du, Yiyou Sun, Xiaojin Zhu, and Yixuan Li. Dream the impossible: Outlier imagination with diffusion models. In *Proceedings of Advances in Neural Information Processing Systems*, 2023.

Xiayu Guo, Xian Lin, Xin Yang, Li Yu, Kwang-Ting Cheng, and Zengqiang Yan. Uctnet: Uncertainty-guided cnn-transformer hybrid networks for medical image segmentation. *Pattern Recognition*, 152:110491, 2024. ISSN 0031-3203.

Kai Han, Yunhe Wang, Hanting Chen, Xinghao Chen, Jianyuan Guo, Zhenhua Liu, Yehui Tang, An Xiao, Chunjing Xu, Yixing Xu, et al. A survey on visual transformer. *arXiv preprint arXiv:2012.12556*, 2020.

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

Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. In *Proceedings of International Conference on Learning Representations, ICLR 2017*, 2017.

Dan Hendrycks, Steven Basart, Mantas Mazeika, Andy Zou, Joe Kwon, Mohammadreza Mostajabi, Jacob Steinhardt, and Dawn Song. Scaling out-of-distribution detection for real-world settings. In *Proceedings of the International Conference on Machine Learning (ICML)*, pp. 8759–8773, 2022.

Christian Henning, Francesco D'Angelo, and Benjamin F. Grewe. Are bayesian neural networks intrinsically good at out-of-distribution detection? *Computing Research Repository (CoRR)*, 2021.

Noah Hollmann, Samuel Müller, Katharina Eggensperger, and Frank Hutter. Tabpfn: A transformer that solves small tabular classification problems in a second. *Computing Research Repository (CoRR)*, 2022.

Xia Hua, Xiaopeng Cui, Xinghua Xu, Shaohua Qiu, and Zhong Li. Weakly supervised underwater object real-time detection based on high-resolution attention class activation mapping and category hierarchy. *Pattern Recognition*, 159:111111, 2025. ISSN 0031-3203.

Pedro Jara-Rodriguez, Cristobal Opazo, et al. Mysteries of the deep: Role of intermediate representations in out of distribution detection. In *Advances in Neural Information Processing Systems*, 2025.

Xue Jiang, Feng Liu, Zhen Fang, Hong Chen, Tongliang Liu, Feng Zheng, and Bo Han. Negative label guided OOD detection with pretrained vision-language models. In *Proceedings of International Conference on Learning Representations*, 2024.

Alex Krizhevsky. *Learning Multiple Layers of Features from Tiny Images*. PhD thesis, University of Toronto, 2009.

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 Advances in Neural Information Processing Systems*, pp. 7167–7177, 2018.

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

Weitang Liu, Xiaoyun Wang, John Owens, and Yixuan Li. Energy-based out-of-distribution detection. In *Proceedings of Advances in Neural Information Processing Systems*, 2020.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of ICCV 2021*, October 2021.

Yifei Ming, Ziyang Cai, Jiuxiang Gu, Yiyou Sun, Wei Li, and Yixuan Li. Delving into out-of-distribution detection with vision-language representations. In *Proceedings of Advances in Neural Information Processing Systems*, 2022.

Atsuyuki Miyai, Qing Yu, Go Irie, and Kiyoharu Aizawa. GL-MCM: Global and local maximum concept matching for zero-shot out-of-distribution detection. *International Journal of Computer Vision*, 2025.

S. Müller, N. Hollmann, S. Arango, J. Grabocka, and F. Hutter. Transformers can do bayesian inference. In *Proceedings of International Conference on Learning Representations*, 2022.

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 NIPS Workshop on Deep Learning and Unsupervised Feature Learning, 2011*, 2011.

GyuTae Park, SungJoon Son, JaeYoung Yoo, SeHo Kim, and Nojun Kwak. Matteformer: Transformer-based image matting via prior-tokens. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 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, 2023.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *Proceedings of the International Conference on Machine Learning*, pp. 8748–8763, 2021.

Vikash Sehwag, Mung Chiang, and Prateek Mittal. SSD: A unified framework for self-supervised outlier detection. In *Proceedings of International Conference on Learning Representations*, 2021.

Mannat Singh, Laura Gustafson, Aaron Adcock, Vinicius de Freitas Reis, Bugra Gedik, Raj Prateek Kosaraju, Dhruv Mahajan, Ross Girshick, Piotr Dollár, and Laurens Van Der Maaten. Revisiting weakly supervised pre-training of visual perception models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 804–814, 2022.

Yiyou Sun, Chuan Guo, and Yixuan Li. React: Out-of-distribution detection with rectified activations. In *Proceedings of Advances in Neural Information Processing Systems*, 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*, pp. 20827–20840, 2022.

Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *Proceedings of International conference on machine learning*, pp. 10347–10357. PMLR, 2021.

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 Conference on Computer Cision and Pattern Recognition*, pp. 8769–8778, 2018.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Proceedings of Advances in Neural Information Processing Systems*, volume 30, 2017.

Sagar Vaze, Kai Han, Andrea Vedaldi, and Andrew Zisserman. Open-set recognition: A good closed-set classifier is all you need. In *Proceedings of International Conference on Learning Representations*, 2022.

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*, 2022a.

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*, pp. 4921–4930, 2022b.

Hualiang Wang, Yi Li, Huifeng Yao, and Xiaomeng Li. CLIPN for zero-shot OOD detection: Teaching CLIP to say no. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1802–1812, 2023.

Jianxiong Xiao, James Hays, Krista A Ehinger, Aude Oliva, and Antonio Torralba. Sun database: Large-scale scene recognition from abbey to zoo. In *Proceedings of 2010 IEEE computer society conference on computer vision and pattern recognition*, pp. 3485–3492. IEEE, 2010.

Xin Xie, Dengquan Wu, Mingye Xie, and Zixi Li. Ghostformer: Efficiently amalgamated cnn-transformer architecture for object detection. *Pattern Recognition*, 148:110172, 2024. ISSN 0031-3203.

Yufei Xu, Qiming Zhang, Jing Zhang, and Dacheng Tao. Vitae: Vision transformer advanced by exploring intrinsic inductive bias. *Computing Research Repository (CoRR)*, 34, 2021.

Zi-Chao Zhang, Zhen-Duo Chen, Yongxin Wang, Xin Luo, and Xin-Shun Xu. A vision transformer for fine-grained classification by reducing noise and enhancing discriminative information. *Pattern Recognition*, 145:109979, 2024. ISSN 0031-3203.

Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(6): 1452–1464, 2017.

OOD			iNaturalist		SUN		Places		Textures		NINCO		OpenImage_O		SSB-hard		Mean	
Prior model	Setting	$G(\mathbf{x})$	FPR95	AUROC	FPR95	AUROC	FPR95	AUROC	FPR95	AUROC	FPR95	AUROC	FPR95	AUROC	FPR95	AUROC	FPR95	AUROC
ResNet50	w/o Prior	CE	100.00	12.32	100.00	19.81	100.00	20.13	100.00	16.93	100.00	18.72	100.00	14.44	100.00	29.51	100.00	18.84
		KL	94.45	79.78	94.08	76.27	94.09	76.04	84.63	80.64	90.23	79.10	90.82	81.42	94.06	69.37	91.77	77.52
		ED	74.32	81.70	70.20	81.95	69.20	81.63	84.59	73.63	83.48	72.86	82.58	74.65	85.25	63.70	78.52	75.73
vit-b-16	w/o Prior	CE	100.00	10.54	99.68	19.92	99.65	20.10	99.73	16.89	99.97	18.21	100.00	13.57	99.61	28.99	99.81	18.32
		KL	93.23	86.42	92.59	78.70	91.64	78.91	81.95	82.82	89.11	81.61	88.90	85.05	96.17	69.37	90.51	80.41
		ED	98.96	63.86	84.09	66.41	86.48	64.07	64.20	67.84	92.04	60.45	83.42	64.65	90.24	60.90	85.63	64.03
	w/ Prior	CE	**2.26**	**99.39**	**30.98**	**93.38**	**40.35**	**90.99**	**32.96**	**92.05**	**41.63**	**90.74**	**16.59**	**96.58**	**58.32**	**84.49**	**31.87**	**92.52**
		ED	3.27	99.08	30.85	93.00	38.24	91.28	33.30	90.63	38.23	90.93	18.78	95.81	56.78	84.72	31.35	92.21
vit-lp	w/o Prior	CE	100.00	10.54	99.68	19.92	99.65	20.10	99.73	16.89	99.97	18.21	100.00	13.57	99.61	28.99	99.81	18.32
		KL	93.23	86.42	92.59	78.70	91.64	78.91	81.95	82.82	89.11	81.61	88.90	85.05	96.17	69.37	90.51	80.41
		ED	98.96	63.86	84.09	66.41	86.48	64.07	64.20	67.84	92.04	60.45	83.42	64.65	90.24	60.90	85.63	64.03
	w/ Prior	CE	**13.08**	**97.56**	**41.56**	**90.99**	**49.97**	**88.40**	**39.56**	**90.43**	**49.20**	**88.33**	**25.62**	**95.26**	**61.31**	**82.19**	**40.04**	**90.45**
		ED	18.03	96.18	28.75	91.85	37.01	90.24	78.01	79.42	51.02	87.08	49.12	88.92	60.23	80.86	46.02	87.79

Table 5. Ablation study on the effect of priors. "w/o Prior" denotes OOD detection by directly computing the divergence between a vanilla ViT and the prior model without prior token integration. "w/ Prior" denotes PViT with integrated prior tokens (highlighted rows). ID dataset is IMAGENET-1K. For vit-b-16, PViT results use DeiT as the prior model; for vit-lp, PViT results use ViT-B/16 as the prior model.

A Implementation Details for Reproducibility

A.1 ID datasets

CIFAR10 consists of 60,000 32x32 color images in 10 classes, with 6,000 images per class Krizhevsky (2009). The dataset is divided into 50,000 training images and 10,000 test images. The classes are mutually exclusive and include objects such as cats, dogs, trucks, and ships.

CIFAR100 is similar to CIFAR10 but contains 100 classes each consisting of 600 images Krizhevsky (2009). Each class is divided into 500 training images and 100 testing images. The 100 classes in CIFAR100 are grouped into 20 super-classes, each comprising 5 sub-classes.

ImageNet-1K, a subset of the larger ImageNet dataset, contains over 1.2 million images spanning 1,000 classes Deng et al. (2009). ImageNet-1K is a standard benchmark in computer vision research. Its classes encompass a wide range of objects, including various species of animals, plants, and everyday objects.

A.2 OOD datasets

Textures comprises diverse images of textures categorized into several classes, providing a unique challenge for texture recognition and classification models Cimpoi et al. (2014). It is widely used for evaluating model robustness against textural variations.

SVHN (Street View House Numbers) contains digit images obtained from house numbers in Google Street View images. It includes over 600,000 digit images, making it a comprehensive dataset for digit classification tasks Netzer et al. (2011).

Places is a large-scale dataset of scene-centric images. With 365 scene categories and over 1.8 million images, it is extensively used for scene recognition and contextual understanding in images Zhou et al. (2017).

Sun is a dataset of natural scene images under varying illumination and weather conditions, often used for assessing model performance in diverse environmental settings Xiao et al. (2010).

SSB-hard and **NINCO** are datasets specifically designed for evaluating OOD detection in neural networks. SSB-hard focuses on subtly different classes, while NINCO provides near-in-context OOD examples, presenting unique challenges for OOD detection Vaze et al. (2022); Bitterwolf et al. (2023).

OpenImage-O is a subset of the OpenImages dataset, tailored for OOD detection. It includes a wide range of object categories not present in standard datasets like ImageNet, making it ideal for testing the generalizability of models Wang et al. (2022a).

iNaturalist contains images of natural world Van Horn et al. (2018). It has 13 super-categories and 5,089 sub-categories covering plants, insects, birds, mammals, and so on. We use the subset that contains 110 plant classes which do not overlap with ImageNet-1K.

A.3 Prior Models

Pretrained ResNet-18 on CIFAR10: Achieving a test accuracy of 0.9498, this model was trained with a batch size of 128 over 300 epochs, and validation every 5 epochs. The SGD optimizer was used with a learning rate of 0.1, momentum of 0.9, and weight decay of 0.0005, paired with a ReduceLROnPlateau scheduler. This pretrained ResNet-18 model is available on Hugging Face.[1]

[1] `https://huggingface.co/edadaltocg/resnet18_cifar10`

Methods	CIFAR10		TEXTURES		PLACES		SUN		SVHN		Mean	
	FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑
MSP	79.98	78.68	80.02	78.14	82.52	73.34	74.15	82.06	79.46	80.20	78.44	79.17
MaxLogit	**79.36**	**79.35**	78.76	79.33	82.17	73.53	70.45	84.72	77.69	83.20	76.88	80.90
Energy	79.62	79.17	77.75	79.30	82.62	73.36	67.35	85.12	76.82	83.52	76.06	80.98
SSD	92.94	55.31	79.04	71.07	92.41	58.98	77.66	71.67	90.39	68.12	87.77	63.47
ViM	93.57	54.50	82.73	66.53	92.61	56.53	77.59	68.25	85.64	69.28	87.69	61.45
KNN	99.96	31.84	99.50	49.82	99.15	42.16	99.80	46.62	99.92	48.26	99.72	42.87
NNGuide	80.02	78.46	**75.85**	**80.23**	**81.52**	**74.46**	**62.36**	**86.44**	80.64	82.94	**75.20**	**81.39**
PViT	81.68	78.99	78.72	79.10	87.12	71.17	65.02	85.09	**73.87**	**83.57**	79.77	79.25
PViT + KL	84.74	77.29	86.79	76.04	90.68	70.64	75.80	80.59	84.17	79.08	85.20	76.96
PViT + ED	86.88	76.89	81.28	77.90	83.09	71.32	71.73	83.63	79.68	81.04	81.90	77.94

Table B1. Full OOD detection results for CIFAR100 as the ID data.

Pretrained ViT on CIFAR10: The model, with a test accuracy of 0.9788 and a loss of 0.2564, was trained using an Adam optimizer with a learning rate of 5e-05, and a linear learning rate scheduler. This model can be found on Hugging Face.[2]

Pretrained ResNet-18 on CIFAR100: This model recorded a test accuracy of 0.7926. Using SGD as the optimizer with a learning rate of 0.1, momentum of 0.9, and weight decay of 0.0005, along with a CosineAnnealingLR scheduler, the model is available on Hugging Face.[3]

Pretrained ViT on CIFAR100: The fine-tuned version of google/vit-base-patch16-224-in21k on CIFAR100 achieved an accuracy of 0.8985 and a loss of 0.4420. It used a learning rate of 0.0002, with train and eval batch sizes of 16 and 8. The pretrained ViT is available at Hugging Face.[4]

Pretrained ResNet-50 on IMAGENET-1K: Provided by Microsoft, this model achieved an accuracy of approximately 67.35%. The model can be accessed at Hugging Face. [5]

Pretrained Google ViT on IMAGENET-1K: The pretrained orginal ViT, fine-tuned on ImageNet-1K, was initially trained on ImageNet-21k. This model is available at Hugging Face. [6].

Pretrained ViT Variants on ImageNet-1K: We explore various ViT configurations as the prior models, including models trained with different approaches. These configurations include weights trained using the DeIT training recipe, as well as models with the original frozen SWAG trunk weights combined with a linear classifier. The models are avaible in PyTorch. [7].

B Additional Evaluation Results

B.1 Full Evaluation Results on CIFAR100

Full evaluation results on CIFAR100 are provided in Tab. B1.

B.2 Full Evaluation Results on IMAGENET-1k

Here we provide addition evaluation results on IMAGENET-1k with different prior models including other ViTs including DeiT and ViT-swag, and the CNN models including ResNet and RegNet in Fig B2.

[2] https://huggingface.co/aaraki/vit-base-patch16-224-in21k-finetuned-cifar10

[3] https://huggingface.co/edadaltocg/resnet18_cifar100

[4] https://huggingface.co/Ahmed9275/Vit-Cifar100

[5] https://huggingface.co/microsoft/resnet-18

[6] https://huggingface.co/google/vit-base-patch16-224

[7] https://pytorch.org/vision/main/models/generated/torchvision.models.vit_b_16.html#torchvision.models.vit_b_16

OOD Datasets		iNaturalist		SUN		Places		Textures		NINCO		OpenImage_O		SSB-hard		Mean		ID
Methods	Model	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	ACC
MSP	DeiT	51.52	88.16	66.52	80.94	68.67	80.38	60.23	82.99	73.26	78.08	59.93	84.81	69.02	84.73	64.16	82.87	81.07
MaxLogit	DeiT	52.26	85.24	66.88	76.39	69.14	75.06	56.68	81.69	74.02	72.38	58.75	81.55	64.43	85.64	63.17	79.71	81.07
Mahalanobis	DeiT	21.28	95.59	61.55	85.12	61.48	83.92	53.94	87.35	58.74	85.98	44.03	91.73	70.85	76.58	53.12	86.61	81.07
Energy	DeiT	64.08	79.24	72.77	70.28	74.31	68.44	58.48	79.30	78.46	66.03	64.93	76.46	59.13	88.16	67.45	75.42	81.07
SSD	DeiT	25.58	94.21	65.75	81.04	66.58	79.90	53.35	85.28	61.93	82.89	48.23	90.14	66.57	80.58	55.43	84.86	81.07
ViM	DeiT	17.12	96.45	58.87	83.14	60.52	81.24	48.09	88.17	56.09	84.98	41.87	91.93	68.92	76.31	50.21	86.03	81.07
KNN	DeiT	71.35	85.55	79.51	79.48	78.61	78.27	68.58	82.76	78.94	76.73	66.11	85.86	59.56	90.23	71.81	82.70	81.07
NNGuide	DeiT	66.93	88.57	76.98	81.94	76.33	80.61	64.84	85.68	77.56	80.14	65.42	87.93	63.76	89.19	70.26	84.87	81.07
PViT	DeiT	2.26	99.39	30.98	93.38	40.35	90.99	32.96	92.05	41.63	90.74	16.59	96.58	58.32	84.49	31.87	92.52	81.33
PViT + KL	DeiT	10.47	97.35	55.52	87.66	60.90	85.76	58.23	87.95	59.87	87.75	39.66	93.57	76.67	77.35	51.62	88.20	81.33
PViT + ED	DeiT	3.27	99.08	30.85	93.00	38.24	91.28	33.30	90.63	38.23	90.93	18.78	95.81	56.78	84.72	31.35	92.21	81.33
MSP	ViT-swag	27.58	93.96	57.43	85.18	61.13	84.34	53.21	84.96	61.16	84.37	44.32	89.89	75.10	76.41	54.27	79.08	85.29
MaxLogit	ViT-swag	13.19	97.19	47.45	86.80	54.36	83.13	44.70	86.11	50.97	86.46	28.41	93.23	70.13	77.11	44.17	77.89	85.29
Mahalanobis	ViT-swag	4.94	98.85	58.77	88.15	65.62	85.45	43.49	90.30	41.54	91.17	23.87	95.92	68.17	76.97	43.77	79.25	85.29
Energy	ViT-swag	12.64	97.34	48.05	86.48	56.41	82.22	46.79	85.72	52.14	86.12	28.33	93.31	71.69	76.59	45.15	86.83	85.29
SSD	ViT-swag	11.19	97.50	84.91	70.87	88.23	65.10	69.38	79.81	67.00	78.96	45.01	88.62	89.96	59.10	65.10	70.91	85.29
ViM	ViT-swag	3.42	99.20	49.67	88.05	59.69	83.68	42.55	88.46	44.20	89.61	20.50	95.79	74.57	74.74	42.09	77.75	85.29
KNN	ViT-swag	29.46	94.07	72.15	83.88	74.17	81.47	51.21	87.17	68.80	82.50	45.25	91.49	86.43	66.51	61.07	77.27	85.29
NNGuide	ViT-swag	9.17	97.96	45.64	90.03	53.82	87.25	39.26	90.01	49.80	89.63	23.11	95.47	73.65	77.23	42.06	89.65	85.29
PViT	ViT-swag	17.73	96.37	55.03	84.24	63.63	78.82	54.41	83.35	59.00	83.70	34.99	91.95	76.19	74.00	51.57	84.63	85.07
PViT + KL	ViT-swag	44.15	94.11	66.00	86.22	70.09	83.98	60.99	87.29	65.02	86.26	49.70	92.23	75.92	77.99	61.70	86.87	85.07
PViT + ED	ViT-swag	15.14	95.14	36.52	89.48	43.19	88.13	48.62	83.70	48.21	85.14	32.93	89.62	68.56	76.24	41.88	86.78	85.07
MSP	ResNet50	29.73	93.78	59.62	84.56	60.91	84.28	49.98	84.90	60.97	84.79	47.49	89.68	80.27	73.25	55.57	85.03	78.73
MaxLogit	ResNet50	22.12	95.99	50.94	88.43	53.79	87.37	42.23	88.42	57.86	87.01	41.68	92.23	78.91	74.09	49.65	87.65	78.73
Mahalanobis	ResNet50	34.97	94.79	64.98	86.55	70.30	83.92	15.02	95.52	63.75	87.13	37.51	93.89	85.41	71.71	53.13	87.64	78.73
Energy	ResNet50	20.99	96.17	47.10	88.91	51.18	87.70	39.34	88.89	58.08	86.96	41.63	92.32	78.49	74.01	48.12	87.85	78.73
SSD	ResNet50	33.68	94.64	56.00	88.34	65.09	84.51	11.79	96.54	69.00	83.63	37.93	93.29	88.30	66.49	51.68	86.78	78.73
ViM	ResNet50	16.68	96.87	39.37	90.86	49.24	88.48	15.87	94.20	61.84	86.95	28.20	94.59	85.59	72.17	42.40	89.16	78.73
KNN	ResNet50	37.60	93.87	54.53	86.98	63.32	83.54	17.43	94.13	69.45	82.02	40.75	92.46	90.07	61.65	53.31	84.95	78.73
NNGuide	ResNet50	11.98	97.47	31.67	91.66	38.91	90.12	24.96	91.51	60.99	86.56	31.62	93.66	82.44	73.06	40.37	89.15	78.73
PViT	ResNet50	23.25	95.97	46.11	89.04	50.99	87.68	38.32	89.05	59.24	86.53	43.55	92.01	78.93	73.62	48.63	87.70	78.56
PViT + KL	ResNet50	37.11	94.12	69.32	84.66	71.61	83.74	45.50	87.27	64.09	85.90	51.72	91.03	81.12	73.06	60.07	85.68	78.56
PViT + ED	ResNet50	16.89	96.09	36.10	90.21	40.97	89.21	56.24	85.94	54.71	87.19	42.46	91.27	74.99	74.39	46.05	88.73	78.56
MSP	RegNet	23.62	94.64	52.50	86.58	56.82	85.13	49.22	86.47	53.23	86.77	34.64	91.94	71.97	77.96	48.86	87.07	86.01
MaxLogit	RegNet	7.79	98.04	31.68	91.55	41.05	88.08	32.68	91.19	34.31	91.44	16.74	95.68	54.33	84.39	31.23	91.48	86.01
Mahalanobis	RegNet	2.22	99.36	49.33	89.85	61.86	85.77	27.91	93.90	38.23	92.06	19.51	96.48	71.89	77.71	38.71	90.73	86.01
Energy	RegNet	6.68	98.28	29.41	91.88	40.51	87.97	30.85	91.48	34.03	91.53	16.19	95.81	52.44	84.82	30.02	91.68	86.01
SSD	RegNet	5.11	98.82	60.35	85.88	70.88	80.27	38.14	92.58	54.30	86.31	28.77	93.54	83.70	69.72	48.75	86.73	86.01
ViM	RegNet	1.97	99.52	28.21	93.15	42.74	89.05	20.55	95.58	34.04	92.71	13.57	97.15	65.66	81.27	29.53	92.63	86.01
KNN	RegNet	4.32	98.76	46.15	88.45	56.30	85.15	28.33	91.93	56.99	85.57	21.30	95.51	86.19	67.01	42.80	87.48	86.01
NNGuide	RegNet	1.83	99.57	21.58	94.44	31.47	91.87	17.00	95.82	29.40	93.58	10.79	97.73	55.86	84.09	23.99	93.87	86.01
PViT	RegNet	14.21	96.64	51.93	83.34	60.53	78.38	48.10	86.06	54.13	83.87	30.85	91.65	72.58	71.77	47.48	84.53	83.25
PViT + KL	RegNet	23.71	94.96	69.04	82.56	69.99	81.10	61.68	86.14	61.04	86.12	38.60	92.00	81.28	73.54	57.91	85.20	83.25
PViT + ED	RegNet	64.39	90.09	68.89	80.76	75.88	78.90	77.38	79.71	62.08	82.84	64.64	83.94	78.24	71.57	70.21	81.12	83.25

Table B2. Additional OOD detection results for iNaturalist-1k as the ID data.

OOD Dataset	Energy		KL		Cross Entropy		Euclidean Distance	
	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑
Texture	0.67	99.74	0.78	99.18	0.84	99.55	0.78	99.29
SVHN	25.61	94.43	24.05	95.00	25.06	95.64	24.83	93.31
Places365	9.43	95.96	10.88	96.02	8.31	97.89	8.48	96.67
LSUN_C	11.16	96.34	15.07	95.24	13.45	97.34	11.50	97.25
LSUN_Resize	39.62	88.06	45.48	87.55	39.45	93.62	38.73	89.92
iSUN	37.22	87.95	43.75	87.99	37.00	93.85	34.60	90.55
CIFAR100	34.21	89.54	36.27	87.72	31.36	93.01	34.65	88.79
Mean Test Results	24.56	93.15	22.54	92.67	22.21	95.84	23.38	93.33

Table B3. OOD detection results for CIFAR10 as the ID data. Scale factor $\alpha = 1$. ↑ indicates larger values are better and ↓ indicates smaller values are better. All values are percentages.

B.3 Full Evaluation Results on CIFAR10

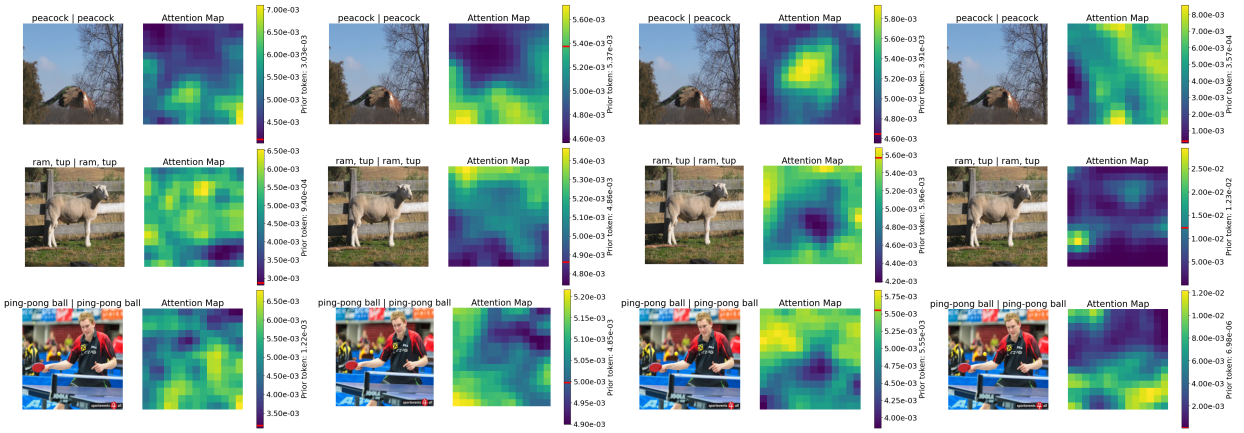
Comprehensive results of our evaluation on CIFAR10 are provided in Tab. B3 and Tab. B4. These results represent the average performance across seven OOD datasets: Texture, SVHN, Places365, LSUN_C, LSUN_Resize, iSUN, and CIFAR100. For a scale factor of $\alpha = 1$, the detailed results are provided in Table B3. In contrast, Table B4 presents the results for a smaller scale factor of $\alpha = 1e - 3$.

B.4 Additional Visualization of the Attention Maps

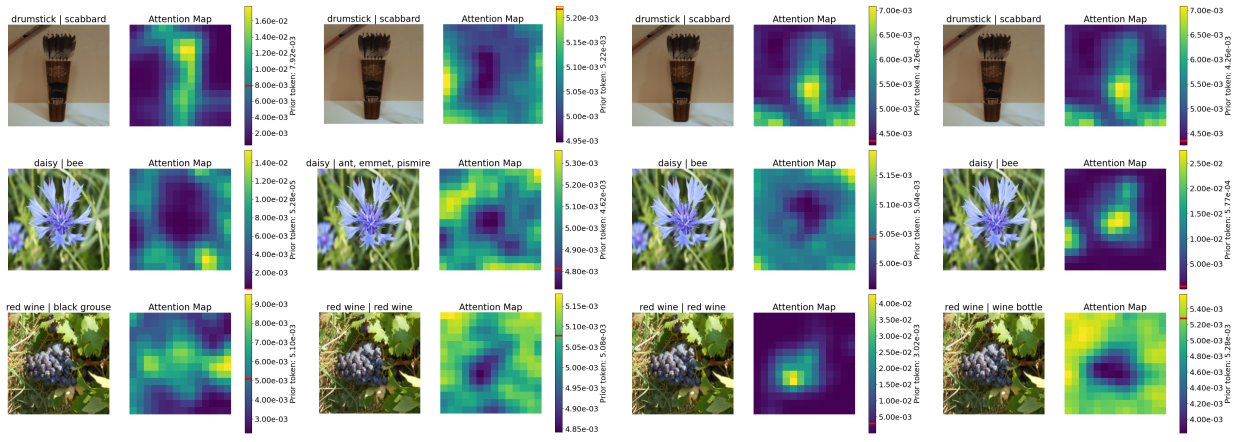
In addition to visualization of the attention maps presented in the paper, we provide further visualizations to compare different scaling factors α as shown in Fig. B1a and Fig. B1b.

OOD Dataset	Energy		KL		Cross Entropy		Euclidean Distance	
	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑
Texture	0.95	99.65	48.10	90.37	1.06	99.18	11.89	96.63
SVHN	30.25	93.18	71.21	69.92	22.54	94.50	45.93	77.21
Places365	7.59	97.71	46.99	83.43	10.60	97.07	22.43	89.52
LSUN_C	11.50	97.57	51.62	82.03	13.90	96.54	24.55	89.83
LSUN_Resize	43.53	89.36	79.69	51.58	42.02	91.10	54.41	68.40
iSUN	37.44	90.19	77.18	55.88	36.66	91.43	54.07	70.98
CIFAR100	33.71	89.95	74.89	66.06	33.20	90.41	50.33	74.63
Mean Test Results	23.57	93.94	64.24	71.32	22.86	94.32	37.66	81.03

Table B4. OOD detection results for CIFAR10 as the ID data. Scale factor $\alpha = 1e-3$. ↑ indicates larger values are better and ↓ indicates smaller values are better. All values are percentages.



(a) Additional visualization of attention maps for ID images with varying scaling factors α for prior token embedding, generated from the last layer and the first MSA head. From left to right, each column represents the scale factor value of $\alpha = 100$, $\alpha = 5$, $\alpha = 1$, $\alpha = 0.1$. The figures are sourced from ImageNet-1K, representing ID data.



(b) Additional visualization of attention maps for OOD images with varying scaling factors α for prior token embedding, generated from the last layer and the first MSA head. From left to right, each column represents the scale factor value of $\alpha = 100$, $\alpha = 5$, $\alpha = 1$, $\alpha = 0.1$. The figures are sourced from OpenImage_O, representing OOD data.

Figure B1. Additional visualizations of attention maps for ID and OOD data under different scaling factors of α. Each subfigure illustrates how varying scaling factors affect attention maps for ID and OOD data respectively.