OUT-OF-DISTRIBUTION DETECTION WITH ANGULAR-MAGNITUDE LIKELIHOOD AND TARGETED FEATURE REFINEMENT

Atik Garg *Yu-Shuen Wang*

National Yang Ming Chiao Tung University

ABSTRACT

Reliable out-of-distribution detection is pivotal for deploying deep learning in open-world scenarios. However, existing methods often rely on a single feature modality, either semantic direction (confidence) or activation intensity (norm), limiting their robustness against diverse shifts. To bridge this gap, we propose a unified framework that harmonizes representation quality with decision certainty. We introduce the angular-magnitude likelihood, which jointly evaluates semantic alignment and signal typicality. To ensure the reliability of the magnitude component, we integrate convolutional block attention modules into discriminative layers. This approach purifies feature norms without diluting low-level texture information. Experiments demonstrate that our framework achieves state-of-the-art performance.

***Index Terms*—** Out-of-Distribution Detection, Score Fusion, Machine Learning

1. INTRODUCTION

Deep neural networks have demonstrated remarkable success in closed-set recognition tasks. However, their deployment in safety-critical applications, such as autonomous driving and medical diagnosis, is hindered by their inability to handle unknown data. A reliable system must not only classify known inputs accurately but also reject samples that deviate from the training distribution.

Despite recent progress, a fundamental limitation persists in how current methods interpret latent features. Confidence-based methods (e.g., MSP [1], ODIN [2]) primarily utilize the angular information of the feature vector relative to class boundaries. While effective for semantic classification, they are prone to the *scaling issue*, where out-of-distribution (OOD) samples with correct directions but weak signals trigger high-confidence false positives. Conversely, feature-norm-based methods [3, 4] leverage the magnitude of activations to gauge signal familiarity. However, these methods often lack probabilistic interpretation and ignore the semantic direction, making them susceptible to high-energy background noise.

To bridge this gap, we propose a framework grounded in the joint likelihood of the feature space. We argue that a robust OOD detector must verify both what the object is (angular alignment) and how typical the signal is (magnitude intensity). To this end, we introduce the angular-magnitude likelihood (AML), a novel scoring function that mathematically fuses calibrated softmax probability and feature norm to reconstruct the full feature likelihood via polar decomposition. Complementing this theoretical model, we implement an attention refinement strategy. Unlike universal attention application, we selectively enhance only the most discriminative convolutional blocks. This prevents the dilution of discriminative focus and ensures that feature norms serve as a robust indicator for the AML score.

Experiments verify that synergizing attention-refined features with dual-feature likelihood estimation achieves state-of-the-art performance. Our evaluation spans standard Far-OOD tasks, challenging Near-OOD scenarios where semantic gaps are minimal, and complex real-world applications in medical pathology and remote sensing. These results demonstrate the framework's robustness and versatility across varying degrees of distribution shifts.

2. RELATED WORK

Existing OOD detection methods typically derive scoring functions from the latent representations of a pre-trained model. These approaches can be broadly categorized based on the specific signals they exploit: softmax confidence, feature magnitude, and external semantic priors.

Confidence and Distribution-based Methods. A foundational line of work relies on the output probabilities of the classifier. [1] proposed the maximum softmax probability (MSP) baseline, identifying OOD samples based on low prediction confidence. Subsequent works improved this by calibrating the distribution; for instance, ODIN [2] introduced temperature scaling and input perturbations to amplify the separability between in-distribution (ID) and out-of-distribution (OOD). Similarly, [5] proposed using the energy score, which aligns theoretically with the probability

This work is partially supported by the National Science and Technology Council (NSTC), Taiwan, under Contract:113-2221-E-A49 -149 -MY3, 114-2221-E-A49 -129 -MY3, and 115-2425-H-A49 -001 -, and by the Higher Education Sprout Project of the National Yang Ming Chiao Tung University and Ministry of Education (MOE), Taiwan

density of the input. Beyond output scores, other methods focus on rectifying the activation space. Techniques like ReAct [6] and DICE [7] truncate extreme activations in the penultimate layer to mitigate the noise caused by OOD inputs. While effective, these methods primarily rely on the angular alignment of features relative to class boundaries, often overlooking the intensity of the signal.

Feature Norm-based Methods. Recently, the magnitude (ℓ_2 norm) of feature vectors has emerged as a distinct and powerful OOD signal. [8] established the theoretical foundation, demonstrating that ID samples generally trigger higher activation norms than OOD samples due to the concentration of measure phenomenon. Building on this, NAN [3] introduced a negative-aware norm to capture both activation and deactivation patterns. Block Selection [4] revealed that intermediate layers may contain more discriminative norm information than the final layer. However, relying solely on feature norms can be risky, as high-contrast background noise in OOD images can occasionally trigger high norms without semantic meaning. Our work aims to bridge this gap by fusing norm-based typicality with confidence-based semantics.

Methods Utilizing External Information. A stream of research seeks to enhance detection by incorporating external resources. Outlier exposure (OE) methods [9, 10] improve robustness by explicitly training the model with auxiliary datasets effectively teaching the network what unknown looks like. More recently, the focus has shifted towards leveraging the rich semantic priors of vision-language models (VLMs). Methods like CATEX [11] utilize the joint embedding space of models like CLIP [12] to perform zero-shot OOD detection by matching images against text descriptions. While powerful, these approaches come with limitations: OE requires careful curation of auxiliary data, while VLMs often suffer from domain shifts in specialized tasks (e.g., medical imaging) and incur high computational costs. In contrast, our framework focuses on maximizing the utility of the intrinsic features within a standard backbone, free from dependencies on external datasets or large-scale pre-trained priors.

3. METHOD

We jointly assess the semantic direction and activation intensity of the latent representation to distinguish ID samples from OOD samples. To this end, we introduce an attention refinement strategy that enhances the discriminative capacity of the learned representations. The overall training and inference pipeline is summarized in Algorithm 1.

Algorithm 1 Training and Inference with CBAM and Angular-Magnitude Likelihood Score

Require: Training data $\mathcal{D} = \{(x_i, y_i)\}$, Jigsaw training data $\mathcal{D}' = \{(x'_i, y'_i)\}$, model M, total epochs N, cross-entropy loss $\mathcal{L}_{\text{CE}}$, convocational blocks$\{b^1, b^2, \ldots, B\}$
Ensure: OOD score $s_{\text{OOD}}(x)$ for a test sample x
procedure TRAINING
  Train M on $\mathcal{D}$ for 10 epochs using $\mathcal{L}_{\text{CE}}$
  Identify convolutional block $b^\star$ with highest feature-norm ratio between $\mathcal{D}$ and $\mathcal{D}'$
  Insert CBAM modules into blocks $\{b^\star, \ldots, B\}$ to form M_{CBAM}
  Train M_{CBAM} for N epochs using $\mathcal{L}_{\text{CE}}$
end procedure
procedure INFERENCE(Sample x)
  Compute Norm(x) at block $b^\star$
  Compute MSP(x).
  Compute OOD score $s_{\text{OOD}}(x)$ using Eq 1.
  Classify x as OOD if $s_{\text{OOD}}(x) < threshold$.
end procedure

3.1. Angular-Magnitude Likelihood

Current OOD detection methods often rely on a single modality of the latent representation: feature-norm methods capture the *intensity* of activations but ignore class-relative directions, while confidence-based methods capture *directional* alignment but overlook the magnitude of the signal. To bridge this gap, we propose the angular-magnitude likelihood (AML).

3.1.1. A Joint Likelihood Perspective

To theoretically ground our approach, we formulate OOD detection as a density estimation problem within the deep feature space. Let $z \in \mathbb{R}^d$ denote the feature vector from the penultimate layer. Our objective is to design a scoring function $S(x)$ that approximates the log-likelihood of a sample belonging to the ID manifold, i.e., $\log P(z|\text{ID})$.

By applying polar decomposition, we express the feature vector z in terms of its magnitude $r = \|z\|_2$ and direction $\hat{z} = z/\|z\|_2$. Assuming that the magnitude and angular components are asymptotically independent in high-dimensional spaces, the joint probability density can be factorized as:

$$P(z|\text{ID}) \approx P(\hat{z}|\text{ID}) \cdot P(r|\text{ID})$$

Taking the logarithm of both sides, we derive that our target OOD score $S(x)$ should naturally consist of two additive components:

$$S(x) \propto \log P(z|\text{ID}) \approx \underbrace{\log P(\hat{z}|\text{ID})}_{\text{Angular Component}} + \underbrace{\log P(r|\text{ID})}_{\text{Magnitude Component}}$$

This decomposition reveals that a robust OOD detector must jointly evaluate whether a sample is aligned with ID semantic directions ($\hat{z}$) and whether it possesses a typical activation magnitude (r).

The Angular Component via MSP. The first term, $\log P(\hat{z})$, represents the semantic alignment of the feature with known class concepts. Recent theoretical findings on

the neural collapse phenomenon [13] demonstrate that during the terminal phase of training, classifier weights w_k converge to an equiangular tight frame structure, where their norms become approximately equal ($\|w_k\| \approx C$). Under this condition, the logit $f_k(x) = w_k^T z$ becomes proportional to the cosine similarity:

$$f_k(x) = w_k^T z = \|w_k\|\|z\| \cos(\theta_k) \approx C \cdot r \cdot \cos(\theta_k),$$

where θ_k is the angle between w_k and z. Since the softmax operation normalizes the logits, namely,

$$MSP(x) = \max_k \frac{e^{C \cdot r \cdot \cos\theta_k}}{\sum_j e^{C \cdot r \cdot \cos\theta_j}}$$

the resulting MSP effectively captures the relative directional probability $P(\hat{z}|\text{Classes})$, invariant to the absolute scale of r. Therefore, we posit that $\log MSP(x) \propto \log P(\hat{z}|\text{ID})$.

It is worth noting that standard networks are often overconfident, producing sharp softmax distributions for both ID and OOD samples, which renders the angular component (i.e., MSP) indistinguishable. To address this, we adopt the strategy proposed in ODIN [2] to apply temperature scaling (i.e., $\tau = 5$) during the *inference phase*. This calibration disproportionately affects OOD samples (pushing them towards a uniform distribution) while preserving the relative confidence of ID samples, thereby restoring the discriminative power of the angular component in our fusion score.

The Magnitude Component via Feature Norm. Relying solely on the angular component leads to overconfidence, as OOD samples may share a direction with ID classes but lack the requisite signal strength. Following the insight discussed in [3], ID feature representations are optimized to elicit strong activations, whereas OOD samples typically induce smaller norms. Thus, we obtain $\log FN(x) \propto \log P(r|\text{ID})$. In our implementation, $FN(x)$ is the l_2 norm of the feature vector extracted from the most discriminative block b^* [4].

3.1.2. Score Fusion

Based on the derivation above, a robust OOD detector must jointly constrain both class direction and feature magnitude. However, simply summing the two terms is suboptimal due to the discrepancy in dynamic ranges between the bounded probability $MSP \in [0, 1]$ and the unbounded feature norm $FN \in [0, \infty)$. To fuse these modalities effectively, we define the fused score $S(x)$ as:

$$S(x) = \alpha \cdot \log\big(MSP(x) + \epsilon\big) + (1 - \alpha) \cdot \log\big(FN(x) + \epsilon\big), \tag{1}$$

where $\alpha \in [0, 1]$ is a balancing coefficient that controls the contribution of the angular and magnitude components, ϵ is a small scalar (e.g., 10^{-6}) to prevent numerical instability. By combining the feature norm with this calibrated probability, AML reconstructs the full feature likelihood, ensuring that samples are identified as ID only if they possess both the correct semantic direction and the typical activation magnitude.

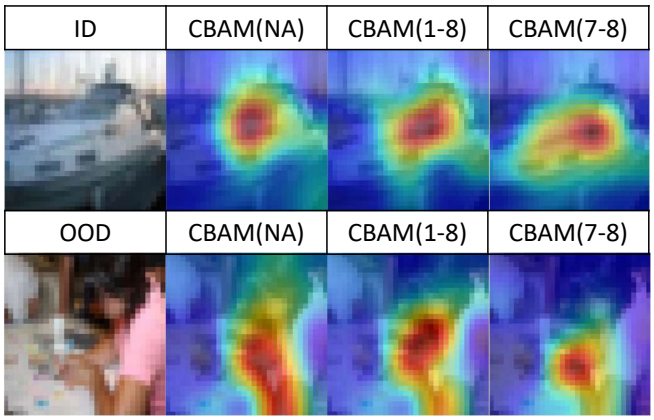

Fig. 1. Visualization of the activation maps demonstrates that CBAM is most effective when applied on blocks 7 and 8.

3.2. Feature Refinement

To maximize the efficacy of the proposed AML score, it is crucial that the magnitude component (feature norm) accurately reflects the semantic typicality of ID samples. Standard norm-based methods rely on identifying discriminative blocks by computing the *norm ratio* between features from the ID dataset $\mathcal{D}$ and a pseudo-OOD dataset $\mathcal{D}'$, generated via patch-wise jigsaw transformations [4]. However, as visualized in Figure 1, the baseline model suffers from *diffused focus*. It fails to concentrate high-norm activations on the semantic objects of ID samples and frequently activates irrelevant background regions in OOD samples.

To mitigate this, we propose a *attention refinement* strategy by integrating the convolutional block attention module (CBAM) [14]. While CBAM, comprising channel and spatial attention sub-modules, is traditionally inserted after every convolutional block, our empirical analysis reveals that such *universal application yields diminishing returns* (see CBAM(1-8) in Figure 1). We posit that applying attention mechanisms to early layers disrupts the extraction of semantic features from low-level textures. Instead, we adopt a two-stage training protocol to precisely locate and refine semantic features. (1) *Block Identification:* We first train the backbone for a warm-up period (e.g., 10 epochs) using the cross-entropy loss $\mathcal{L}_{\text{CE}}$. We then compute the norm ratio for all blocks to identify the most discriminative block $b^\star$. (2) *Targeted Integration:* We insert CBAM modules exclusively after $b^\star$ and its subsequent blocks. The model is then trained for the remaining epochs.

4. EXPERIMENTS

Following the standard protocol established in [4], we evaluate performance by mixing the ID test set with samples from specific OOD datasets. We adopt a *parameter-free evaluation* strategy, quantifying performance using the area under

Method	Venue	Far OOD														Near OOD	
		SVHN		Textures		LSUN-C		LSUN-R		iSUN		Places365		Average		CIFAR-100	
		FPR	AUC	FPR	AUC	FPR	AUC	FPR	AUC	FPR	AUC	FPR	AUC	FPR	AUC	FPR	AUC
MSP [1]	ICLR'17	52.12	92.20	59.47	89.56	32.83	95.62	48.35	93.07	50.30	92.58	60.70	88.42	50.63	91.91	54.83	85.49
ODIN [2]	ICLR'18	33.83	93.03	45.49	90.01	7.29	98.62	20.05	96.56	23.09	96.01	45.06	89.86	29.14	94.02	88.27	72.06
EN [5]	NeurIPS'20	30.47	94.05	45.83	90.37	7.21	98.63	23.62	95.93	27.14	95.34	43.67	90.29	29.66	94.10	43.58	89.18
FN [4]	CVPR'23	7.13	98.65	31.18	92.31	**0.07**	**99.96**	27.08	95.25	26.02	95.38	62.54	84.62	25.67	94.36	48.94	87.62
RFW [15]	TPAMI'24	54.81	87.67	48.81	87.50	15.89	96.75	36.58	92.59	36.31	92.70	47.19	87.75	39.93	90.83	45.94	90.22
OTOD [16]	ICASSP'25	22.87	91.25	41.59	88.92	5.89	97.48	30.72	94.64	32.74	94.30	41.34	89.51	29.19	92.68	45.66	86.83
PRO [17]	CVPR'25	19.82	94.28	35.48	90.65	0.56	99.85	29.97	95.78	14.55	97.56	**32.49**	90.43	22.15	94.76	43.44	90.88
Ours (CBAM)		2.95	99.40	18.85	96.28	0.21	99.92	15.35	97.23	11.57	97.87	56.77	88.31	17.62	96.50	43.18	91.25
Ours (CBAM+AML)		**2.22**	**99.54**	**12.57**	**97.82**	0.15	99.94	**5.99**	**98.66**	**4.49**	**98.92**	38.22	**92.80**	**10.61**	**97.95**	**41.73**	**92.49**

Table 1. OOD detection results on ResNet-18 with CIFAR-10 as the ID dataset. The best and second-best results are highlighted in bold and underlined, respectively.

	CIFAR-100		CIFAR-10			
Method	ResNet18		WRN-28-10		VGG11	
	FPR	AUC	FPR	AUC	FPR	AUC
MSP [1]	73.02	82.43	41.49	91.84	64.77	88.73
ODIN [2]	63.87	85.94	29.22	90.95	47.52	91.56
EN [5]	71.45	84.96	28.65	91.99	46.46	91.67
FN [4]	60.27	84.09	13.53	97.33	39.34	91.18
RFW [15]	77.55	77.18	23.86	93.82	48.31	89.45
OTOD [16]	60.44	78.69	27.17	94.93	48.21	87.08
PRO [17]	60.03	85.43	17.64	96.28	46.40	89.81
Ours (CBAM + AML)	**39.72**	**89.79**	**7.56**	**98.43**	**33.16**	**93.25**

Table 2. Our method is evaluated on multiple network architectures using CIFAR-10 and CIFAR-100 as ID datasets.

Method	iNaturalist		SUN		PLACES		Open-O		Avg	
	FPR	AUC	FPR	AUC	FPR	AUC	FPR	AUC	FPR	AUC
MSP [1]	55.0	87.7	70.8	80.9	74.0	79.8	95.3	50.0	73.8	74.6
EN [5]	55.7	90.0	59.3	85.9	64.9	82.9	95.6	50.3	68.9	77.2
RFW [15]	69.5	85.8	73.5	80.7	76.6	78.5	88.0	59.5	76.9	76.1
ODIN [2]	47.7	89.7	60.2	84.6	67.9	81.8	94.9	49.3	67.7	76.3
FN [4]	22.0	95.8	42.9	90.2	56.8	85.0	88.1	58.1	52.5	82.3
PRO [17]	45.3	90.0	61.7	85.2	62.8	83.3	94.2	50.7	66.0	77.3
Ours (CBAM)	17.7	96.4	35.4	91.7	47.3	87.3	87.9	56.4	47.0	82.9
Ours (CBAM + AML)	**12.4**	**97.6**	**29.4**	**93.5**	**39.6**	**90.6**	**85.6**	**62.0**	**41.7**	**85.9**

Table 3. Comparison of OOD detection performance on the ImageNet dataset.

the receiver operating characteristic curve (AUROC) and the false positive rate at 95% true positive rate (FPR95). To dissect the contribution of each component, we denote the model trained with our attention strategy alone as CBAM, while the dual-feature scoring is referred to as AML. For comprehensive training details, including backbone specifications, optimization settings, and hyperparameter configurations, please refer to the Appendix.

4.1. Evaluation on Benchmarks

Far-OOD Performance. We initially follow the standard protocol, using the CIFAR-10 and CIFAR-100 datasets as the ID set and diverse datasets (e.g., SVHN, LSUN, Textures) as the OOD set. As detailed in Table 1, employing CBAM alone improves detection accuracy over the baseline. This confirms that refining features in discriminative blocks effectively suppresses spurious background activations. While improvements on LSUN-Crop are moderate due to the lack of background context, our method demonstrates significant gains across the majority of datasets. Crucially, integrating the AML scoring function further boosts performance by fusing classification confidence with feature typicality.

Near-OOD Robustness. To further stress-test our model, we evaluate it on the *Near-OOD* scenario, where the semantic gap between CIFAR-10 and CIFAR-100 is minimal. This setting requires the model to distinguish fine-grained semantic differences rather than relying on low-level statistical shifts. As shown in Table 1, our approach achieves the highest performance across both FPR95 and AUROC metrics compared to baselines. This indicates the robustness of our method even when the OOD samples share similar visual structures with the ID classes.

Generalization across Backbones. We also evaluate our method across varying architectures, including VGG-11, WRN-28-10, and ResNet-18. Table 2 summarizes the performance. As indicated, our approach consistently outperforms existing methods across all backbones, demonstrating its versatility and robustness independent of the underlying network depth or width.

Large-Scale Scalability. Finally, to validate the scalability of our approach, we conduct experiments on the more challenging ImageNet benchmark using ResNet-50. As presented in Table 3, the proposed framework exhibits superior performance in high-dimensional semantic spaces. This substantial margin underscores the effectiveness of checking both angular alignment and magnitude typicality when dealing with large-scale, semantically complex data.

Method	Venue	MIDOG		SpaceNet-8	
		Avg. FPR	Avg. AUC	FPR	AUC
CATEX (CLIP) [11]	NeurIPS'23	64.57	75.90	29.33	82.19
SCALE (RNet50) [18]	ICLR'24	58.48	76.44	25.29	86.68
NAC-UE (RNet50) [19]	ICLR'24	54.84	77.26	25.74	86.25
OTOD (RNet50) [16]	ICASSP'25	56.51	73.90	30.57	83.81
Ours (CBAM+AML) (RNet50)		**47.76**	**81.97**	**23.15**	**89.58**

Table 4. OOD detection results on the MIDOG and SpaceNet-8 datasets. For MIDOG, we report the average performance across OOD subsets.

4.2. Evaluation on Real-World Domains

Standard OOD benchmarks often rely on semantic shifts between natural images (e.g., Dog vs. Truck). To assess applicability in specialized, safety-critical domains, we evaluate our method on two real-world datasets: MIDOG [20] and SpaceNet-8 [21]. *MIDOG* is a pathology dataset for mitosis detection. We treat images from a specific acquisition center as ID, while samples from different scanners, staining protocols, or species are treated as OOD. *SpaceNet-8* is a remote sensing dataset. We designate non-flooded regions as ID and flooded regions as OOD, testing the model's ability to detect environmental anomalies.

In these experiments, distinct from the CIFAR/ImageNet evaluations, we extend our comparison to include **CATEX** [11], a recent OOD detection method based on the CLIP vision-language model. To adapt CATEX to these tasks, we followed its original protocol by constructing domain-specific textual prompts and finetuning the model to explicitly detect visual patterns that deviate from the expected data distribution, such as unfamiliar cellular textures in MIDOG or unseen terrain domains in SpaceNet-8. However, as reported in Table 4, even with task-specific finetuning, this VLM-based approach performs sub-optimally compared to ResNet-50 baselines. We attribute this degradation to the persistent *semantic gap*: CLIP's representations, pre-trained on internet-scale natural images, are optimized for object-level semantics rather than the fine-grained, repetitive textural cues required for medical histology or aerial topography. In contrast, by learning the intrinsic feature norms and angular distributions of the specific ID domain (e.g., specific tissue textures), our framework effectively identifies domain shifts, outperforming baselines in these challenging real-world scenarios.

5. ABLATION STUDY

5.1. Sensitivity Analysis of Hyperparameters

We investigate the impact of two critical hyperparameters in our framework: the balancing coefficient α in the AML scoring function and the temperature scaling parameter τ. Table 5 summarizes the results. As indicated, the detection performance remains consistently high across a broad range of α and τ, indicating that the proposed AML score is insensitive to minor fluctuations in these hyperparameters.

Effect of α			**Effect of τ**		
α	FPR	AUC	τ	FPR	AUC
0.0	17.50	96.55	1	14.41	97.20
0.1	15.51	96.70	2	14.29	97.32
0.2	14.41	97.20	5	10.60	97.94
0.3	14.43	97.16	6	10.68	97.16
0.4	14.90	97.06	10	11.66	97.71

Table 5. The effect of varying trade-off α and temperature τ on OOD detection performance.

CBAM Convolution Blocks					
Block	FPR	AUC	*Block*	FPR	AUC
NA	25.67	94.36	1 - 8	18.32	96.49
7	21.94	95.99	8	18.80	96.44
7,8	17.50	96.55	6,7,8	20.11	96.27

Table 6. Ablation study of CBAM applied to convolutional blocks (1–8). "NA" indicates CBAM not applied.

5.2. Impact of Attention Placement Strategy

Table 6 compares the performance of integrating CBAM into different stages of the backbone network. Our empirical results align with the hypothesis that attention should be applied selectively. Inserting CBAM exclusively after the most discriminative block ($b^\star$, Block 7) and its successors yields the best OOD detection performance. This configuration refines high-level semantic features where the separation between ID and OOD is most distinct. In contrast, applying CBAM universally (Blocks 1-8) or to earlier layers leads to a notable performance drop. We attribute this to *feature dispersion*: applying strong attention mechanisms to early layers interferes with the extraction of low-level textures and edges, which disrupts the foundational representations required for the final feature norm calculation.

6. CONCLUSION

In this study, we addressed the fundamental limitations of existing OOD detection methods, which typically rely on a single modality of the latent representation. We proposed a unified framework grounded in the polar decomposition of feature space, introducing the angular-magnitude likelihood to jointly evaluate semantic direction and activation intensity. Complementing this scoring function, we developed an attention refinement strategy that selectively enhances discriminative blocks to purify feature norms. Extensive experiments on standard benchmarks and real-world datasets demonstrate the effectiveness of our method.

7. REFERENCES

[1] Dan Hendrycks and Kevin Gimpel, "A baseline for detecting misclassified and out-of-distribution examples in neural networks," in *ICLR*, 2017.

[2] Shiyu Liang, Yixuan Li, and Rayadurgam Srikant, "Enhancing the reliability of out-of-distribution image detection in neural networks," in *ICLR*, 2018.

[3] Jaewoo Park, Jacky Chen Long Chai, Jaeho Yoon, and Andrew Beng Jin Teoh, "Understanding the feature norm for out-of-distribution detection," in *ICCV*, October 2023, pp. 1557–1567.

[4] Yeonguk Yu, Sungho Shin, Seongju Lee, Changhyun Jun, and Kyoobin Lee, "Block selection method for using feature norm in out-of-distribution detection," in *CVPR*, 2023, pp. 15701–15711.

[5] Weitang Liu, Xiaoyun Wang, John D. Owens, and Yixuan Li, "Energy-based out-of-distribution detection," *NeurIPS*, vol. 33, pp. 21464–21475, 2020.

[6] Yiyou Sun, Chuan Guo, and Yixuan Li, "React: Out-of-distribution detection with rectified activations," *NeurIPS*, vol. 34, pp. 144–157, 2021.

[7] Yiyou Sun and Yixuan Li, "Dice: Leveraging sparsification for out-of-distribution detection," in *ECCV*. Springer, 2022, pp. 691–708.

[8] Zhen Fang, Yixuan Li, Jie Lu, Jiahua Dong, Bo Han, and Feng Liu, "Is out-of-distribution detection learnable?," *NeurIPS*, vol. 35, pp. 37199–37213, 2022.

[9] Dan Hendrycks, Mantas Mazeika, and Thomas Dietterich, "Deep anomaly detection with outlier exposure," in *ICLR*, 2019.

[10] Jiefeng Chen, Yixuan Li, Xi Wu, Yingyu Liang, and Somesh Jha, "Atom: Robustifying out-of-distribution detection using outlier mining," in *ECML*. Springer, 2021, pp. 430–445.

[11] Kai Liu, Zhihang Fu, Chao Chen, Sheng Jin, Ze Chen, Mingyuan Tao, Rongxin Jiang, and Jieping Ye, "Category-extensible out-of-distribution detection via hierarchical context descriptions," *NeurIPS*, vol. 36, pp. 33241–33261, 2023.

[12] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Gretchen Sutskeverand, "Learning transferable visual models from natural language supervision," in *International conference on machine learning*. PMLR, 2021, pp. 8748–8763.

[13] Vardan Papyan, XY Han, and David L Donoho, "Prevalence of neural collapse during the terminal phase of deep learning training," *National Academy of Sciences*, vol. 117, no. 40, pp. 24652–24663, 2020.

[14] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon, "Cbam: Convolutional block attention module," in *ECCV*, 2018, pp. 3–19.

[15] Yue Song, Wei Wang, and Nicu Sebe, "Rankfeat&rankweight: Rank-1 feature/weight removal for out-of-distribution detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

[16] Heng Gao, Zhuolin He, and Jian Pu, "Detecting ood samples via optimal transport scoring function," in *IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2025, pp. 1–5.

[17] Wenxi Chen, Raymond A Yeh, Shaoshuai Mou, and Yan Gu, "Leveraging perturbation robustness to enhance out-of-distribution detection," in *CVPR*, 2025, pp. 4724–4733.

[18] Kai Xu, Rongyu Chen, Gianni Franchi, and Angela Yao, "Scaling for training time and post-hoc out-of-distribution detection enhancement," in *ICLR*, 2024.

[19] Yibing Liu, Chris Xing Tian, Haoliang Li, Lei Ma, and Shiqi Wang, "Neuron activation coverage: Rethinking out-of-distribution detection and generalization," in *ICLR*, 2024.

[20] Marc Aubreville, Frauke Wilm, Nikolas Stathonikos, Katharina Breininger, Taryn A Donovan, Samir Jabari, Mitko Veta, Jonathan Ganz, Jonas Ammeling, Robert Klopfleisch van Diest, Paul J, and Christof A. Bertram, "A comprehensive multi-domain dataset for mitotic figure detection," *Scientific data*, vol. 10, no. 1, pp. 484, 2023.

[21] Ronny Hänsch, Jacob Arndt, Dalton Lunga, Matthew Gibb, Tyler Pedelose, Arnold Boedihardjo, Desiree Petrie, and Todd M Bacastow, "Spacenet 8-the detection of flooded roads and buildings," in *CVPR*, 2022, pp. 1472–1480.