This CVPR paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

DPU: Dynamic Prototype Updating for Multimodal Out-of-Distribution Detection

Shawn Li[1], Huixian Gong[1], Hao Dong[2], Tiankai Yang[1], Zhengzhong Tu[3], Yue Zhao[1]
[1]University of Southern California [2]ETH Zürich [3]Texas A&M University
{li.li02, huixian, tiankaiy, yzhao010}@usc.edu, hao.dong@ibk.baug.ethz.ch, tzz@tamu.edu

Abstract

Out-of-distribution (OOD) detection is crucial for ensuring the robustness of machine learning models by identifying samples that deviate from the training distribution. While traditional OOD detection has predominantly focused on single-modality inputs, such as images, recent advancements in multimodal models have shown the potential of utilizing multiple modalities (e.g., video, optical flow, audio) to improve detection performance. However, existing approaches often neglect intra-class variability within in-distribution (ID) data, assuming that samples of the same class are perfectly cohesive and consistent. This assumption can lead to performance degradation, especially when prediction discrepancies are indiscriminately amplified across all samples. To address this issue, we propose ***Dynamic Prototype Updating*** *(DPU), a novel plug-and-play framework for multimodal OOD detection that accounts for intra-class variations. Our method dynamically updates class center representations for each class by measuring the variance of similar samples within each batch, enabling tailored adjustments. This approach allows us to intensify prediction discrepancies based on the updated class centers, thereby enhancing the model's robustness and generalization across different modalities. Extensive experiments on two tasks, five datasets, and nine base OOD algorithms demonstrate that DPU significantly improves OOD detection performances, setting a new state-of-the-art in multimodal OOD detection, including improvements up to* 80% *in Far-OOD detection. To improve accessibility and reproducibility, our code is released at* https://github.com/lili0415/DPU-OOD-Detection.

1. Introduction

Out-of-distribution (OOD) detection aims to identify samples that differ from the in-distribution (ID) data in ways that challenge the model's ability to generalize [2, 14, 16, 25, 36]. It is crucial for enhancing the safety and robustness of machine learning models across various domains, such as autonomous driving [8, 26], medical imaging [19], robotics [3, 9], and other applications [4, 12, 15, 21, 28, 42, 43, 48, 50]. In recent years, numerous OOD detection algorithms have been developed, spanning classification-based and distance-based methods [7, 17, 30, 31, 40, 49]. Traditionally, OOD detection has focused on *single-modality* inputs, such as images or videos. With the emergence of large Vision-Language Models [37, 52], researchers are exploring OOD detection with the assistance of language modalities [29, 33, 46]. However, their evaluations remain limited to benchmarks containing *only images*. Effectively leveraging multimodal features (*e.g.*, video, optical flow, and audio) remains an open challenge, requiring further research.

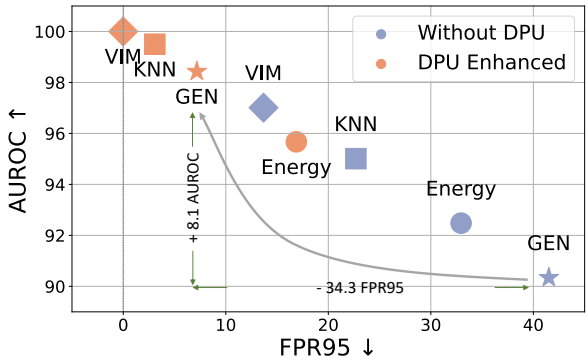


Figure 1. Performance of our DPU applied to four base OOD methods in the Multimodal Far-OOD Detection task (§4.2), using HMDB51 as the ID dataset and Kinetics600 as the OOD dataset. Red symbols denote the OOD methods enhanced by DPU, demonstrating that DPU significantly improves their performances.

Current Work. Dong et al. [11] introduced the first multimodal OOD benchmark and framework, identifying the phenomenon of modality prediction discrepancy. This phenomenon reveals that softmax prediction discrepancies across different modalities are negligible for ID data but significant for OOD data. By amplifying this prediction discrepancy during training, they observed improvements in OOD detection performance. However, a key assumption in previous multimodal OOD detection studies is that all samples within a given class are entirely in-distribution [11, 49], implying perfect cohesion among samples within

the same class. This assumption rarely holds true in real-world applications, where intra-class variability is common. As a result, applying uniform discrepancy intensification across all training samples can degrade the model's ID prediction accuracy [11]. When the discrepancy is intensified on class-center samples—typically exhibit consistent predictions across all modalities—this consistency is disrupted, causing significant model confusion and performance degradation, as shown in Appx. C.4 case study.

Our Proposal. To tackle the challenge of intra-class variations in existing multimodal OOD detection methods, we introduce a novel approach called **D**ynamic **P**rototype **U**pdating (DPU). DPU dynamically adjusts the multimodal prediction discrepancy to ensure high intra-class cohesion and clear inter-class separation, leveraging instance-level training invariance. The core idea of DPU is to update the prototype representations [24, 27] of each class at a dynamic, sample-specific rate, resulting in more precise and robust model performance. These prototype representations act as central reference points, capturing the key features of each class (see Fig. 2 for an overview). To establish a reliable representation space, we first introduce the Cohesive-Separate Contrastive Training procedure (§3.3), which applies marginal contrastive learning to strengthen intra-class cohesion while preserving distinctions between individual samples. Building on this, we design the Dynamic Prototype Approximation mechanism (§3.4), which adaptively refines prototype representations based on observed sample variances. This adaptive updating helps mitigate the negative impact of outliers on prototype evolution, stabilizing the learning process. Using these refined prototypes, we further adjust the multimodal prediction discrepancy for each sample according to its similarity to its class prototype (§3.5). Finally, OOD models make predictions by leveraging both the joint probability distribution across all modalities and the distinct information from each modality. In summary, we make the following contributions:

- **New Observations in Multimodal OOD Detection**. We are the first to identify and explore the negative impact of intra-class variations within ID data for OOD detection.
- **Novel, Model-Agnostic Framework**. We propose DPU, a flexible, plug-and-play method that effectively handles intra-class variations and is compatible with various existing OOD detection models. As shown in Fig. 1, DPU enhances performance across various OOD methods.
- **Effectiveness**. Comprehensive experiments demonstrate the effectiveness of the proposed method across two tasks, five datasets, and nine base OOD methods. DPU significantly improves the performance of all benchmark models, achieving new state-of-the-art results, including improvements of around **10%** across all metrics for Near-OOD detection and up to **80%** for Far-OOD detection.

2. Related Work (Extended Ver. in Appx. A)

OOD Detection. OOD detection identifies samples that deviate from the training distribution while preserving ID accuracy. Key methods include post hoc techniques such as Maximum Softmax Probability (MSP) [16] and enhancements like temperature scaling [30], along with activation-modifying techniques (e.g., ReAct [40]) and distance-based approaches (e.g., Mahalanobis [23]). Recent hybrid methods, including VIM [45], integrate features and logits, while our method uniquely addresses intra-class variability in ID data to enhance existing base OOD detection methods.

Multimodal OOD Detection. Recent efforts extend OOD detection to multimodal contexts, particularly in vision-language systems [33, 46]. Approaches like Maximum Concept Matching (MCM) [33] focus on aligning features to determine OOD scores. Additionally, Dong et al. [11] introduced a new multimodal OOD benchmark that includes video, optical flow, and audio, and revealed prediction discrepancies across different modalities. However, these methods (used as baselines for comparisons in §4) often overlook the rich contextual information provided by modalities, which our dynamic DPU framework leverages.

3. Proposed DPU Framework

3.1. Problem Statement and Preliminaries

Given a training set $\mathbf{D} = \{(x_i, y_i)\}_{i=1}^{n}$, where each $x_i \in \mathcal{X}$ is an input sample and $y_i \in \mathcal{Y} = \{1, 2, \ldots, C\}$ is a class label, OOD detection aims to distinguish between ID and OOD samples. OOD samples exhibit *semantic shifts* compared to ID samples and do not belong to any class in $\mathcal{Y}$. The model employs a feature extractor $g(\cdot)$ to obtain features and a classifier $h(\cdot)$ to generate predictions, yielding probability output $\hat{p}$. When the input data includes multiple modalities, we extend this framework to *multimodal OOD detection*, which we define as follows.

Problem 1 (Multimodal OOD Detection) *Each training sample x_i consists of M modalities, denoted as $x_i = \{x_i^k \mid k = 1, \ldots, M\}$. Multimodal OOD detection combines information from all modalities to make predictions.*

The marginal distribution of ID data is denoted as P_{in}, while OOD samples encountered during testing are drawn from a different marginal distribution P_{out}. The objective of OOD detection is to construct a decision function G that classifies each test sample $x \in \mathcal{X}$ as ID or OOD:

$$G(x; g, h) = \begin{cases} 0 & \text{if } x \sim \mathcal{D}_{\text{out}}, \\ 1 & \text{if } x \sim \mathcal{D}_{\text{in}}. \end{cases} \tag{1}$$

Key Challenges in Multimodal OOD Detection. In multimodal OOD detection, previous research has primarily focused on post-processing techniques applied to logit probabilities [16, 17, 40], with limited emphasis on optimizing the embedding space where learning occurs. While multimodal data offers potential synergy by allowing different

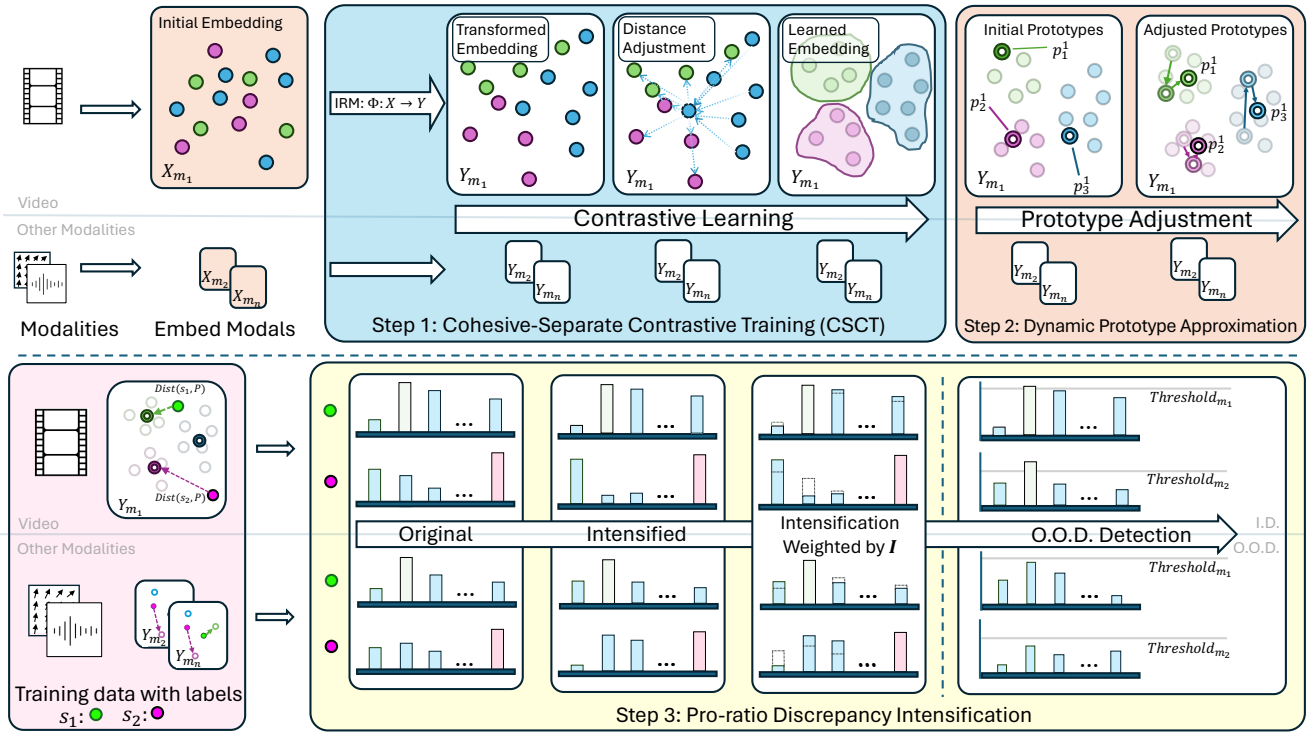


Figure 2. Overview of our DPU. It dynamically adjusts multimodal discrepancy intensification based on each sample's distance to its class prototype. Key components include: (Step 1 ■, §3.3) Cohesive-Separate Contrastive Training (CSCT), which aims to preserve intra-class cohesion and inter-class distinctions while capturing within-class variances; (Step 2 ■, §3.4) Dynamic Prototype Approximation (DPA), which refines prototypes to ensure they remain representative despite class outliers; (Step 3 ■, §3.5) Pro-ratio Discrepancy Intensification (PDI), which adjusts discrepancy based on sample-prototype similarity, boosting ID accuracy and robustness. Finally, OOD models leverage both joint and modality-specific features for robust OOD detection.

modalities to complement each other, how best to leverage this synergy remains an open question.

An essential yet often overlooked challenge is *managing the natural intra-class variations within multimodal ID data*. In real-world applications, samples within the same class may exhibit diverse patterns across different modalities, introducing significant variability. The current leading approach, which uniformly amplifies prediction discrepancies across all samples [11], risks confusing those samples close to the class center, potentially degrading model performance by weakening class cohesion.

3.2. Overview of DPU

To address the challenge above, it is key to balance *intra-class cohesion* with *inter-class separation*. We propose a novel approach that dynamically scales the intensification of prediction discrepancies based on each sample's similarity to its class prototype. This adaptive strategy allows samples near the class center to maintain low discrepancy, preserving cohesive predictions, while more distant samples experience higher discrepancy to increase the model's sensitivity to outlying patterns. This forms the foundation of our proposed DPU. As shown in Fig. 2, DPU integrates three key components to effectively handle the unique challenges of multimodal OOD detection. The primary objective is to *create a stable representation space with strong intra-class cohesion and distinct inter-class separation*, enabling adaptive discrepancy intensification and enhancing the model's predictive accuracy across modalities.

As shown in the flowchart, Fig. 2, the first component, *Cohesive-Separate Contrastive Training (CSCT)* (§3.3, Step 1 ■ in the figure), enhances intra-class consistency while maintaining clear distinctions between classes, improving the model's generalization across data distributions. However, class outliers and noise can interfere with prototype representations, reducing reliability. To mitigate this, we introduce *Dynamic Prototype Approximation (DPA)* (§3.4, Step 2 ■ in the figure), which adaptively updates class prototypes by weighting samples according to their similarity to the prototype, ensuring prototypes remain representative of each class. Additionally, we incorporate *Pro-ratio Discrepancy Intensification (PDI)* and adaptive outlier synthesis (§3.5, Step 3 ■ in the figure) to enhance the model's ability to distinguish ID from OOD samples. These components adjust prediction discrepancies across modalities, strengthening the model's detection performance. Finally, DPU leverages both the joint probability distribution across all modalities and the unique information from each modality to make robust ID/OOD predictions.

3.3. Step 1: Cohesive-Separate Contrastive Training (CSCT) for Intra-class Cohesion

Motivation. In multimodal OOD detection, it is important to maintain strong intra-class cohesion—ensuring that samples from the same class are represented consistently—while still capturing subtle variations within each class [1, 5]. Meanwhile, clear separation between different classes is essential for effective differentiation in the learned representation space. Balancing these objectives is key to achieving robust performance across different modalities.

Inspired by Arjovsky et al. [1], we employ the invariant risk minimization paradigm to (*i*) construct a representation space that is both intra-class cohesive and inter-class separated, and (*ii*) minimize and measure class-wise variances within batches. Specifically, we define a variant representation function $\Phi : X \to H$ that elicits an invariant predictor $\omega \circ \Phi : H \to Y$ across a set of positive environments ε. This predictor is invariant if the classifier ω is optimal for all samples within the set. The learning objective is:

$$\min_{\omega,\Phi} \sum_{e} R^e(\omega \circ \Phi),$$
$$\text{s.t. } \omega \in \arg\min_{\bar{\omega}:H\to Y} R^e(\bar{\omega} \circ \Phi), \quad \forall e \in \varepsilon.$$

Following [1], this objective can be instantiated as:

$$\mathcal{L}_{csct}(\omega, \Phi) = \sum_{e\in\varepsilon} R^e(\Phi) + \lambda \mathbf{D}(\omega, \Phi, e), \tag{2}$$

where $R^e(\cdot)$ is an empirical-based loss function, $\mathbf{D}(\cdot)$ is a parameterization of invariant risks, and λ is a hyperparameter (HP) controlling the balance between objectives. Eq. (2) aims to learn a feature representation $\Phi(\cdot)$ that induces a classifier $\omega(\cdot)$, which remains optimal across all samples in the positive environments $e \in \varepsilon$.

To implement Eq. (2) in our scenario, we use a two-part strategy. First, we apply robust marginal contrastive learning to construct a representation space that is both cohesive within each class and well-separated between classes. This approach aligns naturally with the goals of contrastive learning, which is designed to bring similar samples closer together while pushing dissimilar samples further apart [35]. Second, we measure and minimize class-wise variances within batches to capture subtle intra-class variations and ensure consistency in the learned representations. Together, these steps enable the model to maintain meaningful inter-class distinctions while stabilizing intra-class representations, creating a robust feature space.

Robust Marginal Contrastive Learning. Consider a batch of video samples $V_1, V_2, \ldots, V_n$ and their corresponding data from other modalities, $M_1^k, M_2^k, \ldots, M_n^k$, where $k \in \{1, \ldots, M\}$ represents the modality type, and n is the batch size. For each video sample V_j, we define a positive set PS_j consisting of samples from the same class and a negative set NG_j composed of samples from different classes.

To enhance robustness and sensitivity to inter-class distinctions, we start with the standard InfoNCE loss, a popular contrastive learning objective [35], which is defined as:

$$\mathcal{L}_{rmcl} = \sum_{j=1}^{n} -\log \frac{f_{\text{pos}}}{f_{\text{pos}} + f_{\text{neg}}}, \tag{3}$$

where f_{pos} denotes the similarity scores for samples in the positive set PS_j, and f_{neg} represents the similarity scores for samples in the negative set NG_j.

We compute the similarity scores for positive samples using the arc-cosine function, which allows us to effectively measure the angular distance between representations. We then add an angular margin m to further enhance robustness by emphasizing inter-class distinctions. This adjustment helps to accentuate the differences among classes, improving the model's sensitivity to subtle variations within each class. We defer the details of the modifications to the contrastive loss to Appx. B.1. While contrastive learning can effectively separate different classes, capturing nuanced variations remains challenging, highlighting the need for our refined approach.

Variance Representation Control. To ensure stable intra-class representations, we introduce an invariant representation regularization term. This regularization encourages invariant representations $\Phi(\cdot)$ to yield consistent prediction distributions across all environments $e \in \varepsilon$. We define it as:

$$\mathcal{L}_{irm} = \sum_{j=1}^{n} \text{Var}\left(\mathcal{L}^j\right), \tag{4}$$

where $\mathcal{L}^j = \{\mathcal{L}_{rmcl}(j) \mid j \in PS_j\}$ is the set of loss values for samples within the same positive set. By minimizing the variance of these loss values, the model achieves stable and consistent representation learning for each class [1].

Combining the objectives from robust marginal contrastive learning and variance representation control, we implement CSCT loss in Eq. (2) as:

$$\mathcal{L}_{csct} = \underbrace{\mathcal{L}_{rmcl}}_{\text{Eq. (3)}} + \lambda \cdot \underbrace{\mathcal{L}_{irm}}_{\text{Eq. (4)}}, \tag{5}$$

where λ balances the influence of variance minimization.

CSCT constructs a representation space that is cohesive within classes and distinct across classes, leveraging robust marginal contrastive learning and variance control. In the next step, we build on this representation space to dynamically update prototype representations for each class.

3.4. Step 2: Dynamic Prototype Approximation

Motivation. Building on the representation space created through CSCT, a dynamic approach to prototype learning

is essential for capturing subtle intra-class variations and enhancing class representations. By updating prototypes based on sample-wise similarity, we reduce the influence of outliers, allowing each prototype to more accurately represent its class's central features.

Unlike conventional methods that equally weight all samples within a class, we adjust prototypes according to the variance within each batch, enabling more representative and adaptive prototypes. Thus, We define the total prototype space as $P_{\text{ty}_k} \in \mathbf{D}^{L\times Q}$, where L is the feature dimension, Q is the number of classes, and $k \in \{1, \ldots, M\}$ denotes the modality. The prototype space is updated dynamically by variance observed during the CSCT process.

For a given batch S, we construct positive sets $PS = \{ps_1, ps_2, \ldots, ps_P\}$, where each set corresponds to a unique class label, and P is the number of distinct class labels in S. For each positive set in PS with class label y, we calculate the average representation embedding:

$$H^y_{\text{av}_k} = \frac{1}{N^y} \sum_{i=1}^{N^y} F^k_i, \tag{6}$$

where N^y is the number of samples with label y in batch S, and F^k_i represents the embedding (in modality k) for sample i. This average embedding serves as a central representation for all samples with label y in the batch.

Using this average embedding, we update the prototype for each class y through a moving average approach:

$$P^y_{\text{ty}_k} = \beta P^y_{\text{ty}_k} + (1-\beta) \underbrace{\frac{1}{\gamma + \text{Var}(\mathcal{L}^j) N^y}}_{\text{Update Rate}} \left(H^y_{\text{av}_k} - P^y_{\text{ty}_k}\right), \tag{7}$$

where β and γ are HPs that control the update rate and influence of variance, allowing prototypes to adapt to each batch. The update rate is higher when the variance $\text{Var}(L^j)$ is lower or when the batch size N^y is smaller, enabling faster adaptation when the class representations are more consistent or when fewer samples are available.

3.5. Step 3: Pro-ratio Discrepancy Intensification

Motivation. In multimodal OOD detection, differences in predictions across various modalities are a strong signal for distinguishing ID samples from OOD samples [11]. The dynamic prototypes we have learned provide a basis for selectively amplifying these prediction discrepancies, enabling the model to more effectively detect OOD samples by emphasizing the separation between ID and OOD data.

To achieve this adaptive intensification, we define an intensification rate for each sample i with class label y, which scales the discrepancy by the similarity between the sample's feature representation and its class prototype:

$$\mathcal{L}_{pdi} = -\underbrace{\mu \cdot \left(1 - \text{Sigmoid}(F^v_i \cdot (P^y_{ty_v})^T)\right)}_{\text{Intensification Rate}} \cdot \text{Discr}(\hat{p}^{k_1}_i, \hat{p}^{k_2}_i), \tag{8}$$

where μ is an HP that controls the intensity, F^v_i is sample's video feature embedding, $\text{Sigmoid}(\cdot)$ modulates its similarity to the class prototype, $\hat{p}^{k_1}_i$ and $\hat{p}^{k_2}_i$ denote prediction probabilities of different modalities of sample i, and $\text{Discr}(\cdot)$ represents a distance metric (we use the Hellinger distance following [11]) that quantifies the discrepancy between probability distributions across modalities.

Adaptive Outlier Synthesis. To further improve DPU's OOD detection capability, we apply an adaptive outlier synthesis technique. The goal of this technique is to generate synthetic samples that help the model learn to distinguish between ID and OOD data with greater accuracy. Specifically, we create synthetic outliers by fusing prototypes from different modalities and classes.

For each class, we concatenate its prototypes across different modalities, denoted as $\bar{P} = P_{\text{ty}_{i_1}} \oplus P_{\text{ty}_{i_2}}$, where $i_1, i_2 \in \{\text{video, flow, audio}\}$, and $\oplus$ represents concatenation. To form a synthetic outlier, we select a class prototype $\bar{P}_{y_1}$ and choose another prototype $\bar{P}_{y_2}$ from the top K nearest prototypes of $\bar{P}_{y_1}$. We fuse these prototypes as follows:

$$\bar{P}_{\text{fuse}} = \eta \cdot \bar{P}_{y_1} + (1-\eta) \cdot \bar{P}_{y_2}, \tag{9}$$

where η is an HP controlling the fusion balance.

We then divide $\bar{P}_{\text{fuse}}$ into two separate fused representations, $\bar{P}^{\text{fuse}}_{y_1}$ and $\bar{P}^{\text{fuse}}_{y_2}$, and optimize the model using the following objective function:

$$\mathcal{L}_{aos} = -\left(\text{Discr}(\bar{P}^{\text{fuse}}_{y_1}, \bar{P}^{\text{fuse}}_{y_2}) + E(\bar{P}^{\text{fuse}}_{y_1}) + E(\bar{P}^{\text{fuse}}_{y_2})\right), \tag{10}$$

where $E(\cdot)$ is an entropy function that measures prediction uncertainty. By optimizing this objective, the model learns to generalize and distinguish between ID and OOD samples more effectively, leveraging both adaptive intensification and synthetic outlier generation.

Following [11], our multimodal OOD model leverages the joint probability distribution across all modalities and individual predictions from each modality. The model uses M modality-specific feature extractors and a shared classifier, with each modality k also having its own classifier for generating individual prediction probabilities.

The base loss $\mathcal{L}_{base}$ combines the cross-entropy losses for both joint and individual modality outputs to encourage accurate predictions across modalities. Finally, the complete objective function for OOD detection integrates additional regularization components from the previous steps:

$$\mathcal{L} = \mathcal{L}_{base} + \delta \cdot \mathcal{L}_{csct} + \mathcal{L}_{pdi} + \kappa \cdot \mathcal{L}_{aos}, \tag{11}$$

where δ and κ are HPs that balance the influence of the CSCT, discrepancy intensification, and adaptive outlier synthesis losses, thereby optimizing the model's performance on OOD detection. See details of HP settings in Appx. B.2.

Table 1. Multimodal Near-OOD Detection results using video and optical flow (↑ the higher the better; ↓ the lower the better). 'A2D' refers to the Agree-to-Disagree baseline [11], while 'AN' combines Agree-to-Disagree with NP-Mix. DPU consistently enhances all metrics across various datasets and OOD detection models, showcasing its effectiveness and adaptability in multimodal OOD detection.

Method	HMDB51 25/26			UCF101 50/51			Kinetics600 129/110			EPIC-Kitchen		
	FPR95 ↓	AUROC ↑	ID ACC ↑	FPR95 ↓	AUROC ↑	ID ACC ↑	FPR95 ↓	AUROC ↑	ID ACC ↑	FPR95 ↓	AUROC ↑	ID ACC ↑
MSP [16]	44.66	87.74	89.32	22.14	95.73	99.22	64.08	76.06	80.11	76.31	67.59	71.46
+A2D	38.78	88.37	90.64	7.09	98.19	99.61	63.04	76.47	79.50	66.23	67.91	71.83
+AN	33.99	88.79	89.98	7.96	98.24	99.71	62.91	76.92	80.52	67.91	71.52	71.64
+Ours	34.20	**89.15**	**92.16**	**7.57**	98.17	**99.81**	**61.59**	**77.50**	**81.07**	**63.81**	**71.47**	**72.39**
Energy [31]	43.36	87.46	89.32	22.52	96.06	99.22	68.75	75.49	80.11	76.68	68.29	71.46
+A2D	39.22	88.84	90.63	9.81	98.16	99.61	64.59	76.45	79.50	66.98	72.45	71.46
+AN	36.38	88.91	89.98	6.50	98.48	99.71	63.69	77.11	80.52	67.91	73.79	71.64
+Ours	**35.07**	**89.52**	**92.16**	7.48	**98.43**	**99.81**	**62.56**	**77.40**	**81.07**	**63.62**	**74.13**	**72.39**
Maxlogit [17]	43.36	87.75	89.32	22.52	96.02	99.22	68.73	75.98	80.11	76.68	68.29	71.46
+A2D	39.22	88.93	90.63	9.81	98.15	99.61	64.57	76.92	79.50	66.98	72.23	71.46
+AN	36.38	89.06	89.98	6.50	98.49	99.71	63.65	77.55	80.52	66.98	73.48	71.64
+Ours	**35.08**	**89.60**	**92.16**	7.57	98.39	**99.81**	**62.46**	**77.76**	**81.07**	**63.62**	**73.50**	**72.39**
Mahalanobis [23]	40.31	85.28	89.32	12.14	97.14	99.22	93.51	35.83	80.11	98.69	42.99	71.46
+A2D	44.88	86.99	90.63	8.74	98.07	99.61	92.86	50.65	79.50	95.52	44.43	71.46
+AN	41.61	87.69	89.98	7.86	97.99	99.71	92.49	50.05	80.52	94.78	44.37	71.64
+Ours	**36.17**	**89.53**	**92.16**	**6.60**	**98.68**	**99.81**	**89.27**	**54.17**	**81.07**	95.34	**45.62**	**72.39**
ReAct [40]	42.05	87.79	89.32	25.63	95.85	99.32	72.40	73.80	80.35	83.96	65.89	71.08
+A2D	37.91	89.09	90.63	10.39	98.12	99.61	71.26	74.29	79.64	68.66	72.03	70.52
+AN	37.47	88.63	90.20	12.43	97.33	99.90	67.95	75.55	80.70	66.60	73.11	72.39
+Ours	**35.08**	**89.51**	**92.16**	**7.48**	**98.43**	**99.81**	**62.57**	**77.41**	**81.07**	71.46	71.24	69.78
ASH [7]	53.59	87.16	89.54	32.14	94.02	99.22	69.24	76.16	79.62	76.87	67.92	70.15
+A2D	42.05	87.72	90.41	12.52	97.43	99.42	64.28	77.28	79.15	63.25	74.73	67.72
+AN	36.17	89.30	89.32	10.68	97.80	99.81	63.77	77.44	79.44	66.23	73.06	66.23
+Ous	**35.08**	**89.51**	**92.16**	**7.48**	**98.43**	**99.81**	**62.57**	**77.41**	**81.07**	71.46	71.24	69.78
GEN [32]	43.79	87.49	89.32	23.79	95.54	99.22	69.03	75.33	80.11	76.87	68.52	71.46
+A2D	38.56	88.61	90.63	8.83	98.12	99.61	62.28	77.08	79.50	65.86	73.05	71.46
+AN	35.95	89.78	89.32	7.67	92.28	99.71	62.95	76.95	80.52	65.30	75.17	71.64
+Ours	**34.64**	89.71	**92.16**	**6.41**	**98.41**	**99.81**	**62.10**	**78.05**	**81.07**	**63.06**	74.22	**72.39**
KNN [41]	42.92	88.46	89.32	15.63	96.93	99.22	68.67	74.64	80.11	75.93	63.60	71.46
+A2D	33.33	89.59	90.63	7.48	98.39	99.61	65.71	76.23	79.50	72.39	67.83	71.46
+AN	33.77	90.05	89.98	12.04	97.65	99.71	66.81	74.19	80.52	71.27	69.94	71.64
+Ours	**32.90**	89.36	**92.16**	**11.84**	**97.74**	**99.81**	**66.71**	**75.34**	**81.07**	71.64	68.03	**72.39**
VIM [47]	36.82	88.06	89.32	12.52	97.66	99.22	68.77	75.47	80.11	77.05	65.60	71.46
+A2D	33.77	89.37	90.63	6.80	98.62	99.61	64.59	76.45	79.50	66.42	68.43	71.46
+AN	34.64	88.80	89.98	4.56	98.73	99.71	63.67	77.12	80.52	67.72	69.09	71.64
+Ours	**32.46**	**89.93**	**92.16**	6.79	98.56	**99.81**	**62.32**	**77.51**	**81.07**	**66.60**	**70.50**	**72.39**

4. Experimental Results

4.1. Datasets and Evaluation Metrics

Datasets. Following [11], we evaluate our method on 5 datasets: HMDB51 [22], UCF101 [39], Kinetics-600 [20], HAC [10], and EPIC-Kitchens [6]. See details in Appx. C.1.

Evaluation Metrics. We evaluate the performance via the use of the following metrics: (1) the false positive rate (FPR95, the lower, the better) of OOD samples when the true positive rate of ID samples is at 95%, (2) the area under the receiver operating characteristic curve (AUROC), and (3) ID classification accuracy (ID ACC).

4.2. Tasks and Implementation Details

Tasks. We evaluate DPU on two tasks: Near-OOD detection and Far-OOD detection [11]. We present the results on {video, optical flow} in the main paper, and defer the results on full modality including {video, optical flow, audio} to Appendix. See details in Appx. C.2.

Base OOD Methods and Baselines. We include MSP [16], Energy [31], Maxlogit [17], Mahalanobis [23], ReAct [40], ASH [7], GEN [32], KNN [41], and VIM [47] as our base OOD methods to be extended to the multimodal setting. The training algorithms range from probability space (MSP, GEN), logit space (Energy, Maxlogit), feature space (Mahalanobis, KNN), penultimate activation manipulations (ReAct, ASH), to a combination of logit and feature space (VIM). We employ the SOTA Agree-to-Disagree algorithm (A2D) [11] and the combination of A2D and NP-Mix algorithm (AN) [11] as the baseline methods.

Implementation and Hardware. We use the SlowFast

Table 2. Multimodal Far-OOD Detection results using video and optical flow, with HMDB51 as the ID dataset (↑ the higher the better; ↓ the lower the better). 'AN' represents the combined use of Agree-to-Disagree and NP-Mix algorithms.

Method	OOD Datasets								ID ACC
	Kinetics600		UCF101		HAC		EPIC-Kitchen		
	FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑	
Energy	32.95	92.48	44.93	87.95	32.95	92.28	8.10	97.70	87.23
+AN	24.52	93.96	36.49	89.67	22.92	94.41	6.96	97.53	86.89
+Ours	**21.89**	**95.81**	**33.30**	**92.76**	**20.64**	**95.39**	**4.33**	**98.46**	**87.34**
ASH	51.20	87.81	53.93	84.22	42.99	90.23	19.95	95.21	86.20
+AN	27.82	93.17	38.43	89.52	23.03	94.45	6.84	98.23	86.20
+Ours	**26.45**	86.62	**37.04**	**89.95**	**20.64**	**95.39**	**4.33**	**98.46**	**87.34**
GEN	41.51	90.34	46.18	87.91	38.31	91.28	8.21	98.26	87.23
+AN	25.66	93.50	37.40	91.19	24.63	94.28	5.25	98.98	86.89
+Ours	**7.18**	**98.43**	**36.37**	90.93	**21.43**	**95.66**	**3.19**	**99.30**	**87.34**
KNN	22.69	95.01	39.34	89.28	20.75	96.02	9.97	97.92	87.23
+AN	15.05	96.96	33.06	91.92	13.45	97.25	5.47	98.97	86.89
+Ours	**3.08**	**99.50**	**31.81**	**92.52**	**14.48**	**97.26**	5.87	**98.97**	**87.34**
VIM	13.68	97.01	33.87	91.45	13.45	97.12	5.93	98.15	87.23
+AN	9.24	98.04	26.45	92.34	6.04	98.56	5.36	98.09	86.89
+Ours	**0.01**	**99.99**	27.59	90.86	**3.87**	**99.16**	**1.71**	**99.62**	**87.34**

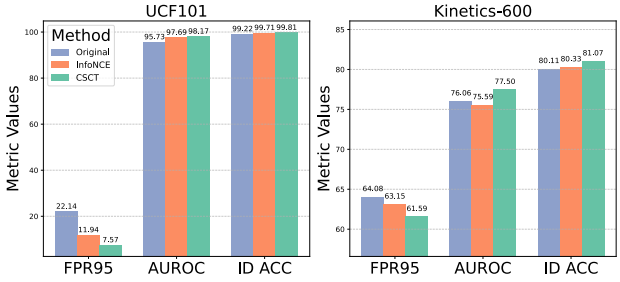


Figure 3. Ablation study on contrastive learning methods. The experiment is conducted using MSP on Near-OOD detection with UCF101, and Kinetics-600 dataset. Original: plain MSP. InfoNCE: the classic contrastive learning method [35]. CSCT: proposed Cohesive-Separate Contrastive Training.

model [13] as our feature extractor, using pre-trained weights from Kinetics-600 [20]. All parameters of the SlowFast model are frozen throughout the training process to maintain consistency. All models are implemented on the MultiOOD codebase [11] and run on a multi-NVIDIA RTX 6000 Ada workstation. See full details in Appx. C.3.

4.3. Comparison with State-of-the-Art (SOTA)

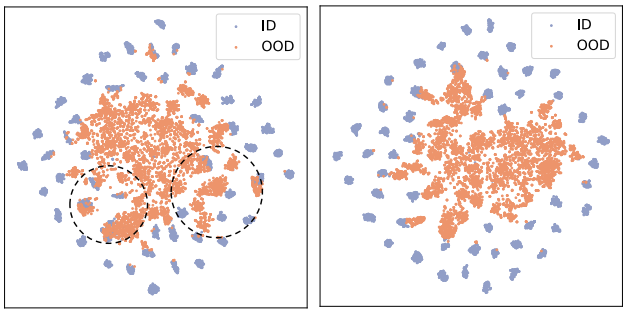
Tables 1 and 2 and Appx. Tables A and B show the effectiveness of DPU in comparison to SOTA methods on 2 tasks and 9 OOD detection methods with all 5 datasets. DPU consistently achieves the best or near-best values for key metrics, including FPR95, AUROC, and ID Accuracy.

DPU is model-agnostic for diverse OOD detection methods. We evaluate DPU across nine base OOD methods with different training strategies, including probability space, logit space, feature space, penultimate activation manipulations, and combinations of logit and feature space techniques. Despite this diversity, DPU consistently improves all these OOD methods on nearly every evaluation metric. For example, on the HMDB51 dataset with the Energy OOD detection, DPU significantly reduces FPR95 from 43.36 to 35.07, while increasing AUROC from 87.46 to 89.52 and ID ACC from 89.32 to 92.16. These results highlight DPU's ability to markedly lower false positive rates and boost identification accuracy for various OOD models.

DPU is effective on multiple OOD tasks. We conduct experiments on both Near (Tab. 1) and Far (Tab. 2 and Tab. A) OOD detection tasks, and DPU consistently demonstrates superior performance across both. We observe that DPU consistently outperforms the baseline methods on both Near-OOD and Far-OOD detection tasks, showing its applicability and effectiveness. As shown in Fig. 1 and Tab. 1, DPU yields substantial performance improvements, with gains of around 10% across all metrics for Near-OOD and up to 80% in certain metrics for Far-OOD detection.

DPU performs efficiently across diverse datasets. We evaluate DPU on a total of five datasets, each representing different video styles, ranging from digitized movies and daily YouTube videos to cartoon figures and kitchen environments. Across every OOD method and dataset combination, DPU consistently lowers FPR95 values and enhances AUROC and ID ACC. This demonstrates its robust capacity to identify Near-OOD samples with higher accuracy and reliability. In datasets like HMDB51 and Kinetics600, DPU achieves particularly notable improvements, especially in metrics such as AUROC, where it significantly surpasses other methods. This indicates DPU's adaptability and stable performance across diverse datasets and OOD detection models, reinforcing its broad applicability and alignment with improved OOD detection outcomes.

More experimental results including Far-OOD detection with Kinetics-600 as the ID, and Near-OOD detection with



(a) Embeddings without DPU. (b) DPU-enhanced embeddings.

Figure 4. Visualization of the learned embeddings on ID and OOD data using t-SNE on the UCF101 50/51 dataset before and after training with DPU. We observe better separation after using DPU.

full modality training are in Appx. C.5.

Qualitative Analysis. As shown in Fig. 4, we provide visualizations of the learned embeddings for ID and OOD data using t-SNE on the UCF101 50/51 dataset before and after training with DPU. Embeddings learned with DPU exhibit greater robustness, with minimal overlap between ID and OOD embeddings, further showing its effectiveness in enhancing class separability and overall detection reliability.

4.4. Ablation Studies

We assess our method by employing different contrastive learning methods (§3.3), and using fixed discrepancy intensification ratios instead of our adaptive one (§3.5).

Ablation Study on CSCT (Step 1). We assess the impact of various contrastive learning methods on near-OOD detection performance using MSP (see Fig. 3). The methods compared include the original MSP (no contrastive learning), InfoNCE (a widely-used contrastive method) [35], and our DPU using CSCT. Results show that CSCT consistently outperforms other approaches on UCF101 and Kinetics-600 datasets. Specifically, CSCT achieves the lowest FPR95 and highest AUROC, showing improved detection accuracy. Also, CSCT demonstrates the highest ID ACC across datasets, highlighting its strength in maintaining intra-class cohesion and enhancing inter-class separation.

Ablation Study on Pro-ratio Discrepancy Intensification. To investigate the impact of varying levels of discrepancy intensification, we substitute the adaptive $\mathbf{I}$ with several fixed rates. As presented in Table 3, we replace $\mathbf{I}$ with $\{0.1, 0.3, 0.5, 0.7\}$ and report the results on the UCF101 and Kinetics-600 Near-OOD datasets. From the table, it is evident that any fixed, pre-defined rate fails to outperform DPU. For example, while lower fixed rates (e.g., 0.1 or 0.3) demonstrate improved performance on UCF101, they perform poorly on Kinetics-600. Conversely, higher fixed rates lead to more balanced results between the two datasets but still cannot match the overall effectiveness of our adaptive method. This limitation arises because fixed rates do not accommodate the intra-class variances of ID samples, causing either under-emphasis or over-emphasis in intensifying discrepancies. By dynamically adjusting the discrepancy based on the data's characteristics, our approach better balances intra-class cohesion and inter-class separability, enhancing the detection performance across datasets.

Table 3. Ablation study on the discrepancy intensification rate for multimodal Near-OOD using the Mahalanobis with UCF101 and Kinetics-600 datasets. It highlights that DPU achieves the best performance across all metrics, showing the advantages of DPU's adaptive intensification strategy over fixed intensification ratios.

Intensification Ratio	UCF101			Kinetics-600		
	FPR95	AUROC	ID ACC	FPR95	AUROC	ID ACC
0.1	10.19	97.88	99.51	92.49	50.05	80.52
0.3	6.89	98.44	99.81	93.43	50.58	79.91
0.5	7.86	97.99	99.71	93.59	50.10	79.93
0.7	8.25	98.28	99.81	92.68	50.81	80.42
DPU	**6.60**	**98.68**	**99.81**	**89.27**	**54.17**	**81.07**

5. Conclusion, Limitations, and Future Work

The DPU framework addresses the critical challenge of intra-class variability in multimodal OOD detection through adaptive, sample-specific prototype updates. Extensive experiments on diverse datasets and tasks demonstrate DPU 's leading performance, achieving substantial improvements in OOD detection and, in some cases, perfect ID/OOD classification accuracy. These results highlight DPU's potential for robust deployment in high-stakes applications.

Limitations and Future Work. While DPU performs well across a range of datasets, its effectiveness may vary with significantly larger or more complex multimodal datasets, which could introduce scalability challenges. Future work could focus on enhancing DPU 's efficiency for these larger datasets and optimizing it for real-time applications, such as autonomous driving and healthcare, where prompt responses are essential. Additionally, exploring cross-domain OOD detection would broaden DPU 's applicability, supporting its use in diverse real-world environments.

Broader Impact and Ethics Statement

Broader Impact Statement: DPU significantly advances multimodal OOD detection by addressing intra-class variability, a crucial factor for reliable model performance in fields like healthcare, autonomous systems, and security. By enhancing model robustness against unknown data distributions, DPU empowers applications in dynamic, high-stakes environments where accuracy and reliability are paramount. This adaptability ensures systems remain effective when confronted with new or evolving data patterns.

Ethics Statement: Our research complies with ethical standards, emphasizing privacy, fairness, and transparency. DPU is designed to mitigate risks associated with biased predictions and potential privacy violations, particularly relevant in sensitive applications like surveillance and medical diagnosis. Promoting more consistent and fair OOD detection, DPU helps reduce ethical concerns tied to unexpected model behaviors, especially meaningful for emerging ones.

Acknowledgments

This work was partially supported by the National Science Foundation under Award No. 2346158. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

References

[1] Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization. *arXiv preprint*, 2019. 4

[2] Yichen Bai, Zongbo Han, Bing Cao, Xiaoheng Jiang, Qinghua Hu, and Changqing Zhang. Id-like prompt learning for few-shot out-of-distribution detection. In *CVPR*, pages 17480–17489, 2024. 1

[3] Hermann Blum, Paul-Edouard Sarlin, Juan Nieto, Roland Siegwart, and Cesar Cadena. Fishyscapes: A benchmark for safe semantic segmentation in autonomous driving. In *ICCVW*, pages 0–0, 2019. 1

[4] Wonwoo Cho, Jeonghoon Park, and Jaegul Choo. Training auxiliary prototypical classifiers for explainable anomaly detection in medical image segmentation. In *WACV*, pages 2624–2633, 2023. 1

[5] Yo Joong Choe, Jiyeon Ham, and Kyubyong Park. An empirical study of invariant risk minimization. *arXiv preprint arXiv:2004.05007*, 2020. 4

[6] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Sanja Fidler, Antonino Furnari, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, and Michael Wray. Scaling egocentric vision: The epic-kitchens dataset. In *ECCV*, 2018. 6, 2

[7] Andrija Djurisic, Nebojsa Bozanic, Arjun Ashok, and Rosanne Liu. Extremely simple activation shaping for out-of-distribution detection. *arXiv preprint arXiv:2209.09858*, 2022. 1, 6

[8] Hao Dong, Xianjing Zhang, Jintao Xu, Rui Ai, Weihao Gu, Huimin Lu, Juho Kannala, and Xieyuanli Chen. Superfusion: Multilevel lidar-camera fusion for long-range hd map generation. *arXiv preprint arXiv:2211.15656*, 2022. 1

[9] Hao Dong, Xieyuanli Chen, Simo Särkkä, and Cyrill Stachniss. Online pole segmentation on range images for long-term lidar localization in urban environments. *Robotics and Autonomous Systems*, 159:104283, 2023. 1

[10] Hao Dong, Ismail Nejjar, Han Sun, Eleni Chatzi, and Olga Fink. SimMMDG: A simple and effective framework for multi-modal domain generalization. In *NeurIPS*, 2023. 6, 2

[11] Hao Dong, Yue Zhao, Eleni Chatzi, and Olga Fink. Multiood: Scaling out-of-distribution detection for multiple modalities. In *NeurIPS*, 2024. 1, 2, 3, 5, 6, 7, 4

[12] 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 *CVPR*, pages 13678–13688, 2022. 1

[13] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition. In *ICCV*, pages 6202–6211, 2019. 7, 3

[14] Mark S. Graham, Walter H. L. Pinaya, Petru-Daniel Tudosiu, Parashkev Nachev, Sebastien Ourselin, and M. Jorge Cardoso. Denoising diffusion models for out-of-distribution detection. In *CVPRW*, pages 2948–2957, 2023. 1

[15] Nan Hao, Yuangang Li, Kecheng Liu, Songtao Liu, Yingzhou Lu, Bohao Xu, Chenhao Li, Jintai Chen, Ling Yue, Tianfan Fu, et al. Artificial intelligence-aided digital twin design: A systematic review. 2024. 1

[16] Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. In *ICLR*, 2017. 1, 2, 6

[17] 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. *ICML*, 2022. 1, 2, 6

[18] Wei Ji, Li Li, Hao Fei, Xiangyan Liu, Xun Yang, Juncheng Li, and Roger Zimmermann. Towards complex-query referring image segmentation: A novel benchmark. *ACM Trans. Multimedia Comput. Commun. Appl.*, 2024. 1

[19] Davood Karimi and Ali Gholipour. Improving calibration and out-of-distribution detection in deep models for medical image segmentation. *IEEE transactions on artificial intelligence*, 4(2):383–397, 2022. 1

[20] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*, 2017. 6, 7, 2, 3

[21] Gyuhak Kim, Sepideh Esmaeilpour, Changnan Xiao, and Bing Liu. Continual learning based on ood detection and task masking. In *CVPRW*, pages 3856–3866, 2022. 1

[22] Hildegard Kuehne, Hueihan Jhuang, Estíbaliz Garrote, Tomaso Poggio, and Thomas Serre. Hmdb: a large video database for human motion recognition. In *ICCV*, 2011. 6, 2

[23] Kimin Lee, Kibok Lee, Honglak Lee, and Jinwoo Shin. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. In *NeurIPS*, 2018. 2, 6, 1

[24] Gen Li, Varun Jampani, Laura Sevilla-Lara, Deqing Sun, Jonghyun Kim, and Joongkyu Kim. Adaptive prototype learning and allocation for few-shot segmentation. In *CVPR*, pages 8334–8343, 2021. 2

[25] Jingyao Li, Pengguang Chen, Zexin He, Shaozuo Yu, Shu Liu, and Jiaya Jia. Rethinking out-of-distribution (ood) detection: Masked image modeling is all you need. In *CVPR*, pages 11578–11589, 2023. 1

[26] Jinlong Li, Baolu Li, Zhengzhong Tu, Xinyu Liu, Qing Guo, Felix Juefei-Xu, Runsheng Xu, and Hongkai Yu. Light the night: A multi-condition diffusion framework for unpaired low-light enhancement in autonomous driving. In *CVPR*, pages 15205–15215, 2024. 1

[27] Li Li, Chenwei Wang, You Qin, Wei Ji, and Renjie Liang. Biased-predicate annotation identification via unbiased visual predicate representation. In *ACM MM*, pages 4410–4420, 2023. 2

[28] Li Li, Wei Ji, Yiming Wu, Mengze Li, You Qin, Lina Wei, and Roger Zimmermann. Panoptic scene graph generation

with semantics-prototype learning. In *AAAI*, pages 3145–3153, 2024. 1

[29] Li Li, You Qin, Wei Ji, Yuxiao Zhou, and Roger Zimmermann. Domain-wise invariant learning for panoptic scene graph generation. In *ICASSP*, pages 3165–3169, 2024. 1

[30] Ming Liang, Bin Yang, Shenlong Wang, and Raquel Urtasun. Deep continuous fusion for multi-sensor 3d object detection. In *ECCV*, 2018. 1, 2

[31] Weitang Liu, Xiaoyun Wang, John D Owens, and Yixuan Li. Energy-based out-of-distribution detection. In *NeurIPS*, 2020. 1, 6

[32] Xixi Liu, Yaroslava Lochman, and Christopher Zach. Gen: Pushing the limits of softmax-based out-of-distribution detection. In *CVPR*, 2023. 6, 1

[33] Yifei Ming, Ziyang Cai, Jiuxiang Gu, Yiyou Sun, Wei Li, and Yixuan Li. Delving into out-of-distribution detection with vision-language representations. In *NeurIPS*, 2022. 1, 2

[34] Jonathan Munro and Dima Damen. Multi-modal Domain Adaptation for Fine-grained Action Recognition. In *CVPR*, 2020. 2

[35] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018. 4, 7, 8, 1

[36] Yuehan Qin, Yichi Zhang, Yi Nian, Xueying Ding, and Yue Zhao. Metaood: Automatic selection of ood detection models. *arXiv preprint arXiv:2410.03074*, 2024. 1

[37] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *ICML*, 2021. 1

[38] Karen Simonyan and Andrew Zisserman. Two-stream convolutional networks for action recognition in videos. In *NeurIPS*, 2014. 1

[39] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012. 6, 2, 4

[40] Yiyou Sun, Chuan Guo, and Yixuan Li. React: Out-of-distribution detection with rectified activations. In *NeurIPS*, 2021. 1, 2, 6

[41] Yiyou Sun, Yifei Ming, Xiaojin Zhu, and Yixuan Li. Out-of-distribution detection with deep nearest neighbors. *ICML*, 2022. 6, 1

[42] Zhensu Sun, Li Li, Yan Liu, Xiaoning Du, and Li Li. On the importance of building high-quality training datasets for neural code search. In *ICSE*, page 1609–1620, 2022. 1

[43] Xinyu Tian, Jing Zhang, Mochu Xiang, and Yuchao Dai. Modeling the distributional uncertainty for salient object detection models. In *CVPR*, pages 19660–19670, 2023. 1

[44] Angelina Wang and Olga Russakovsky. Overwriting pretrained bias with finetuning data. In *ICCV*, pages 3957–3968, 2023. 1

[45] Haoqi Wang, Zhizhong Li, Litong Feng, and Wayne Zhang. Vim: Out-of-distribution with virtual-logit matching. In *CVPR*, 2022. 2, 1

[46] Hualiang Wang, Yi Li, Huifeng Yao, and Xiaomeng Li. Clipn for zero-shot ood detection: Teaching clip to say no. In *ICCV*, 2023. 1, 2

[47] Peng Wang, An Yang, Rui Men, Junyang Lin, Shuai Bai, Zhikang Li, Jianxin Ma, Chang Zhou, Jingren Zhou, and Hongxia Yang. Ofa: Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. In *ICML*, 2022. 6, 4

[48] Haoyan Xu, Kay Liu, Zhengtao Yao, Philip S Yu, Kaize Ding, and Yue Zhao. Lego-learn: Label-efficient graph open-set learning. *arXiv preprint arXiv:2410.16386*, 2024. 1

[49] Jingkang Yang, Pengyun Wang, Dejian Zou, Zitang Zhou, Kunyuan Ding, Wenxuan Peng, Haoqi Wang, Guangyao Chen, Bo Li, Yiyou Sun, et al. Openood: Benchmarking generalized out-of-distribution detection. In *NeurIPS*, 2022. 1

[50] Wenting Yi, Wai Kit Chan, Hiu Hung Lee, Steven T Boles, and Xiaoge Zhang. An uncertainty-aware deep learning model for reliable detection of steel wire rope defects. *IEEE Transactions on Reliability*, 2023. 1

[51] Christopher Zach, Thomas Pock, and Horst Bischof. A duality based approach for realtime tv-l 1 optical flow. In *Pattern Recognition*, 2007. 2

[52] Ao Zhang, Hao Fei, Yuan Yao, Wei Ji, Li Li, Zhiyuan Liu, and Tat-Seng Chua. Vpgtrans: Transfer visual prompt generator across llms. In *NeurIPS*, pages 20299–20319. Curran Associates, Inc., 2023. 1