Modality-Aware Out-of-Distribution Detection for Multi-Modal Action Recognition

Lars Doorenbos[1,2], Duc Manh Vu[1], Serdar Ozsoy[1], and Juergen Gall[1,2]

[1] University of Bonn, Germany
[2] Lamarr Institute for Machine Learning and Artificial Intelligence, Germany
{doorenbos,soezsoy,gall}@iai.uni-bonn.de

Abstract. The incorporation of additional modalities into action recognition models increases their performance across a wide range of settings. However, how this additional information can contribute to making the models more robust remains underexplored, particularly for the case of multi-modal out-of-distribution (OOD) detection. While methods exist that regularize the multi-modal training process with OOD detection in mind, they still apply off-the-shelf OOD detectors designed for the uni-modal case during inference, discarding important information. Based on an interesting relationship we find between the multi-modal and uni-modal predictions, we propose to use this signal to build a post-hoc detector explicitly designed for the multi-modal scenario. We combine this new source of information with a feature-space score, which detects off-manifold samples in the multi-modal space, and normalize them by the multi-modal logits. In doing so, the proposed hybrid detector is compatible with existing training-time approaches and consistently improves performance. Experiments on a wide range of established datasets from the MultiOOD benchmark show that, on average, our approach outperforms the state of the art. Our results show the importance of explicitly considering the different modalities at inference time for multi-modal OOD detection.

1 Introduction

Integrating additional modalities into action recognition systems offers a powerful means to enhance their capabilities. The complementary information present in, for example, audio and optical flow enables these systems to consider aspects impossible to obtain from RGB frames alone, greatly increasing their success [4, 5, 34, 38]. Yet, despite the large interest in improving their performance, little attention has been given to how robust these systems remain when operating outside of controlled settings.

One of the key challenges in designing robust action recognition models, as well as machine learning systems more broadly, is identifying when test samples differ from those observed during training. These *out-of-distribution* (OOD) samples will be confidently handled as belonging to one of the training classes, resulting in silent failures that negatively impact their reliability. To combat this,

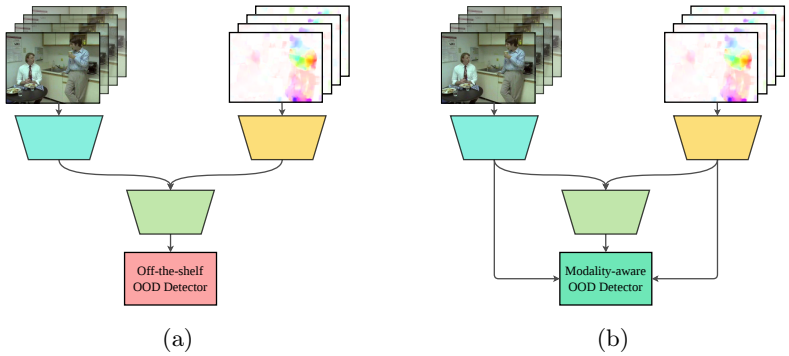


Fig. 1: Motivation. a) Current methods for multi-modal OOD detection apply off-the-shelf detectors designed for the uni-modal case at inference time. b) In contrast, we propose a modality-aware OOD detector that improves multi-modal OOD detection.

OOD detection [16, 42] methods learn scoring functions that measure the level of *out-of-distributionness* of test samples, such that they can be filtered out. However, most current methods are designed for the uni-modal case, leaving multi-modal classifiers without tailored OOD detectors.

Only recently, a handful of works have started to consider the multi-modal video OOD detection problem [9, 26, 30], following the introduction of the first multi-modal OOD detection benchmark [9]. They do so by introducing training-time regularization methods that try to enhance the OOD detection performance of post-hoc detectors at inference time. However, they do not propose new OOD detectors that leverage multiple modalities at inference time. Instead, they apply off-the-shelf *uni-modal* detectors to the joint representation of the modalities, as illustrated in Fig. 1. In this way, the multi-modality is not fully exploited for OOD detection at inference time.

In this work, we address this gap and propose a post-hoc *modality-aware* OOD detector that is motivated by an interesting relationship we identify between the uni-modal and multi-modal predictions for action recognition: there is a high correlation between the multi-modal prediction and an aggregation of the uni-modal predictions for ID samples. In contrast, when presented with OOD samples, this relation between uni- and multi-modal predictions will be altered. As such, we find that the relations between the single- and multi-modal decisions are a strong indication of normality. Importantly, this property separates in-distribution (ID) and OOD samples from a different perspective than the standard feature- or prediction-based methods by modeling the impact of multi-modal integration. Indeed, we show that the relation between uni-modal and multi-modal predictions provides a strong cue for OOD samples ignored by previous methods.

Building on these findings, we propose a modality-aware detector that improves OOD detection performance in multi-modal action recognition scenarios. We do so by integrating this information missed by modality-agnostic scoring

functions into a hybrid detector through combining the multi-modal score with a feature-based one and normalizing them using the multi-modal logits. By considering these complementary aspects together, our method is able to detect a wide variety of OOD samples. We evaluate our approach on the Multi-OOD benchmark [9], which comprises five datasets, up to three modalities, and a total of 14 experiments. Our approach consistently outperforms previous baselines. On the challenging Kinetics-600 Far-OOD experiments, which have the largest ID variety, we outperform the state of the art by 6.2 FPR, an improvement of 13.5%. In summary, our main contributions are:

- We identify an interesting relationship between uni-modal and multi-modal predictions that can be used to separate ID from OOD samples.
- We propose a modality-aware OOD detector, which incorporates both uni- and multi-modal information into a hybrid detector to make its predictions during inference.
- We demonstrate that our method outperforms previous methods across a wide range of experiments, establishing a new state of the art on the Multi-OOD benchmark.

2 Related Work

Uni-modal OOD Detection. Research into uni-modal OOD detection can be broadly categorized into two types. One direction of research involves applying regularization at training time to boost downstream OOD detection performance. This can be done, for instance, by modifying the loss function to encourage calibrated predictions [7, 24], designing auxiliary self-supervised objectives [1, 41], or knowledge distillation [39, 43]. A more recent trend is to explicitly use outliers during training. These outliers can come from large auxiliary datasets [17] or be derived from the training data itself [11–13]. During inference, these methods typically rely on scoring functions related to the regularization approach to detect OOD samples, such as the prediction confidence or entropy. While these methods can be successful, the requirement to train models from scratch limits their general applicability.

Instead, most detectors opt for a post-hoc approach that can be used with an already trained model, as these models natively contain a variety of signals that can detect OOD samples. For instance, the scores can be derived from the predicted probabilities [16, 28] or logits [15, 27], as the predictions for OOD samples will be lower than those for ID samples. Similarly, back-propagating the gradients for specifically designed loss functions can highlight OOD samples [2, 18, 35], and they will lie outside of the distribution of the training features in the feature space of the trained network [20, 25, 33, 37].

More recently, there has been a shift toward post-hoc methods that adopt a hybrid approach: rather than considering these signals independently, they combine information from multiple sources. Due to the complementary nature of these different aspects, this leads to improved performance. Examples are ViM [40] and CADRef [29], which combine information from both feature space

and the logits, and the hybrid versions of GEN [32], which combine features with probabilities. In this work, we also adopt a hybrid, post-hoc approach, but introduce and integrate a new source of information unique to the multi-modal case: the relationship between uni-modal and multi-modal predictions.

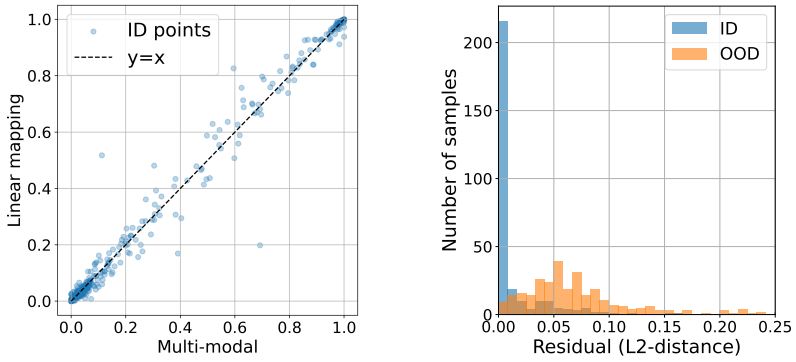
Multi-modal Video OOD Detection. In contrast to the uni-modal case, research into multi-modal OOD detection has only looked at training time regularization until now, without developing new methods to detect OOD samples during inference. Specifically, the emerging area of multi-modal OOD detection for action recognition started with the introduction of the Multi-OOD benchmark [9], along with the Agree-to-Disagree (A2D) algorithm, which serves as the foundation for all subsequent work. A2D is based on the observation that uni-modal predictions tend to agree more for ID samples than for OOD samples. To further encourage this during training where no OOD data are available, A2D introduces a regularization term to increase the discrepancy between uni-modal predictions after excluding the ground-truth class. Follow-up work improved upon A2D by dynamic weighting [26] and refined outlier synthesis [30]. At inference time, all these methods apply existing OOD detectors designed for the *uni-modal* case to the trained networks.

Our work differs in an important aspect: we do not introduce a training-time regularization method; instead, we present the first OOD detector that explicitly considers the different modalities at inference time. As a post-hoc method, our approach is training-free and can be used together with any of the previous training-time methods to boost performance.

3 Modality-Aware OOD detection

The goal of OOD detection is to determine whether input samples belong to one of the classes seen at training time, in which case they are considered ID, and OOD otherwise. In the case of multi-modal action recognition, these samples consist of observations of the same action in multiple modalities, such as video, audio, and optical flow. Formally, for a multi-modal training dataset $\{\mathbf{x}_i, y_i\}_{i=1}^{N}$ with labels $y \in \mathcal{Y}$, each sample $\mathbf{x}_i \in \mathcal{X}$ consists of M modalities, $\mathbf{x}_i = \{\mathbf{x}_i^{(m)} | m = 1, ..., M\}$. Based on the training dataset, multi-modal OOD detection aims to find a function $g(\mathbf{x}) : \mathcal{X} \rightarrow \mathbb{R}$ that assigns the degree of *out-of-distributionness* to a multi-modal test sample $\mathbf{x}$ at inference time.

All previous works in multi-modal OOD detection rely on the A2D method [9, 26, 30]. While these methods show that the relations between modalities can contain important information for detecting anomalies, they all rely on detectors designed for the uni-modal case to make their decisions at inference time. This observation raises the question of why these relations are only used during training. While we show the benefits of a scoring function that explicitly uses multi-modality to detect OOD samples at inference time compared to previous approaches in Sec. 4, we first provide an analysis of uni-modal and multi-modal predictions, which motivates our proposed approach.



(a) (b)

Fig. 2: Example of our observed relation between uni- and multi-modal predictions. (a) The multi-modal predictions can be closely approximated by a linear mapping from the single-modal predictions (Pearson correlation = 0.998) for ID samples, shown here for HMDB. (b) The distance between the probabilities obtained by the linear mapping from the uni-modal heads versus those of the multi-modal head provides a strong signal for OOD detection.

To this end, we formalize the typical architecture used in multi-modal OOD detection for action recognition. We assume access to a multi-modal classification head that combines all modalities, and uni-modal classification heads for each individual modality. This setup is implicitly present in all current multi-modal OOD detection methods, including A2D [9] and DPU [26], and it is not a limitation in practice since heads can be easily added to existing multi-modal architectures, as we show in the supplementary material. At its core, the network is trained with $M+1$ cross-entropy losses. Each modality encoder e_m has a corresponding uni-modal classification head h_m, and the multi-modal head h_f operates on all features combined. The loss is then given by:

$$\mathcal{L}_{cls}(\mathbf{x}, y) = \mathrm{CE}(\mathbf{z}^{(f)}, y) + \sum_{m=1}^{M} \mathrm{CE}(\mathbf{z}^{(m)}, y), \tag{1}$$

where $\mathbf{z}^{(m)} = h_m(e_m(\mathbf{x}^{(m)}))$ are the predicted logits for modality m, and $\mathbf{z}^{(f)} = h_f(e_1(\mathbf{x}^{(1)}), \cdots, e_M(\mathbf{x}^{(M)}))$ operates on the combined embeddings.

Given this structure, we next examine how the uni-modal and multi-modal predictions relate to each other. Fig. 2 shows an interesting relation between the predictions of the uni-modal heads and the prediction of the multi-modal head that we will exploit. For ID samples, there is a high correlation between the predictions of the multi-modal classification head and a linear mapping from the predictions of the uni-modal heads, whereas this relation is altered for OOD samples, as shown in Fig. 2(b). Building on this intuition, we find that the relations between the single- and multi-modal predictions are a strong indication of normality.

We now describe how we design a novel OOD scoring function that uses this property to detect OOD samples in Section 3.1. Then, we describe in Sections 3.2 and 3.3 how we integrate this into a new post-hoc multi-modal OOD detector.

3.1 Modality-Aware OOD Scoring Function

To find the linear mapping between the logits predicted by the single-modal heads and the multi-modal head of the trained model, we solve a linear model with intercept on a validation set of N ID samples for each class c separately,

$$\tilde{\mathbf{w}}_c = \arg\min_{\tilde{\mathbf{w}}_c} ||\tilde{\mathbf{A}}_c \tilde{\mathbf{w}}_c - \mathbf{b}_c||_2, \tag{2}$$

where $\tilde{\mathbf{A}}_c = [\mathbf{A}_c, \mathbf{1}] \in \mathbb{R}^{N \times (M+1)}$ contains the uni-modal logits for class c and $\mathbf{b}_c \in \mathbb{R}^N$ are the multi-modal logits. Solving this system for $\tilde{\mathbf{w}}_c = [\mathbf{w}_c, w_{0,c}] \in \mathbb{R}^{M+1}$ gives the relative contributions of the modalities and intercept $w_{0,c}$ when approximating the multi-modal logits for that class.

When applied to a test sample, we use the learned mappings to transform the uni-modal logits to an estimator of the logits of the multi-modal head,

$$\bar{\mathbf{z}}^{(f)} = \left[\tilde{\mathbf{z}}_1^T \tilde{\mathbf{w}}_1, \cdots, \tilde{\mathbf{z}}_C^T \tilde{\mathbf{w}}_C\right], \tag{3}$$

where $\tilde{\mathbf{z}}_c = [\mathbf{z}_c, 1]$ are the logits per modality for class c.

After taking the softmax, the error between the predicted and observed multi-modal distribution will be larger for OOD samples, which can be used as a scoring function:

$$s(\mathbf{x}) = d(\text{softmax}(\bar{\mathbf{z}}^{(f)}), \text{softmax}(\mathbf{z}^{(f)})), \tag{4}$$

with $d(\cdot, \cdot)$ any distance function defined over probability distributions, for which we use the L_2 norm in our experiments. Conceptually, this implies that the linear mapping between uni-modal and multi-modal predictions will be altered when dealing with abnormal samples, which is shown in Fig. 2(b). In particular, our approach is motivated by the fact that uni-modal representations converge to a shared representation space [19]. It is thus intuitive that multi-modal predictions of ID samples can be estimated from uni-modal predictions, as they are in the same vector space. In this work, we find that this shared space does not hold for OOD samples and exploit this for OOD detection.

3.2 Feature Space Integration

The most successful OOD detection algorithms incorporate information from multiple sources, typically combining both features and predicted probabilities (e.g., [40]). The relation between uni- and multi-modal predictions, described in the previous section, can add another dimension to these hybrid methods. We now describe how we obtain information from the multi-modal feature space and integrate the predictions to obtain our final detector.

In feature space, cross-modal connections can be modeled by considering the manifold of the concatenated features. Then, off-manifold directions are strong

indicators of OOD samples. We represent these directions by the principal components with the lowest variance, i.e., the linear invariants [10]. Any deviations in these directions constitute off-manifold samples that are OOD.

To obtain the invariants, we perform an Eigen decomposition of the sample covariance matrix $\mathbf{\Sigma}$ of the normalized concatenated features of dimensionality D,

$$\mathbf{\Sigma} = \mathbf{Q}\mathbf{\Lambda}\mathbf{Q}^\intercal, \tag{5}$$

where $\mathbf{Q} \in \mathbb{R}^{D\times D}$ contains the Eigenvectors of $\mathbf{\Sigma}$, and $\mathbf{\Lambda} \in \mathbb{R}^{D\times D}$ its Eigenvalues. To address high-dimensional and low-data regimes, we employ the hyperparameter-free Ledoit-Wolf shrinkage method [23] to compute $\mathbf{\Sigma}$. Then, in order to focus on the common features of the entire dataset, we use the k Eigenvectors with the smallest Eigenvalues, denoted as $\mathbf{Q}_k \in \mathbb{R}^{k\times D}$ and $\mathbf{\Lambda}_k \in \mathbb{R}^{k\times k}$. We can set k adaptively by tying it to the number of principal components of the data that jointly explain less than $p\%$ of the variance, where p is a hyperparameter of our method. This way, if the training data lies on a lower-dimensional manifold, more dimensions will be used to detect OOD.

Any deviation in these tight dimensions is a strong signal that a sample lies off-manifold and is OOD. Furthermore, deviations in dimensions with smaller variance should have a larger impact than similar-sized deviations in dimensions with larger variance. Therefore, we use the Mahalanobis distance to compute our feature-level score,

$$r(\mathbf{x}) = \sqrt{(\mathbf{x}-\mu)^\intercal \mathbf{Q}_k \mathbf{\Lambda}_k^{-1} \mathbf{Q}_k^\intercal (\mathbf{x}-\mu)}, \tag{6}$$

where μ is the mean of the training features.

Applying $r(\mathbf{x})$ to the concatenated features $(e_1(\mathbf{x}^{(1)}), \cdots, e_M(\mathbf{x}^{(M)}))$ of all modalities will find the off-manifold directions in the training set and score test samples accordingly. We show this for a toy example in Fig. 3.

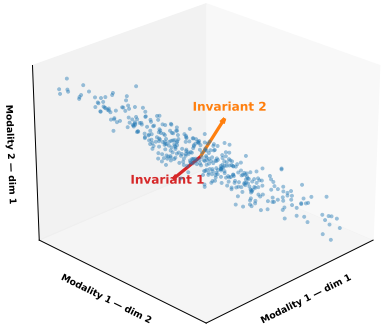


Fig. 3: Finding the invariants on a toy example. The principal components with the lowest variance describe off-manifold directions in the multi-modal feature space. We use these to detect feature-level OOD samples.

3.3 Final Score

Beyond the relation between uni- and multi-modal predictions and feature space, the final predictions of the classifier are an indicator of OOD from a third point of view, and we find that performance can be further improved by taking a holistic approach and combining the three different aspects. We incorporate the previous scores with the predicted distribution as virtual logits [40] with an additional scaling factor. We first compute the mean and standard deviation of the highest

Table 1: Kinetics-600 Far-OOD Detection results with video and optical flow. A2D results taken from [9], FM results taken from [30]. We report the mean and standard deviation over three runs. Our method achieves the best overall results.

Method	OOD Datasets										Acc
	HMDB51		UCF101		HAC		EPIC-Kitchen		Average		
	FPR ↓	AUC ↑	FPR ↓	AUC ↑	FPR ↓	AUC ↑	FPR ↓	AUC ↑	FPR ↓	AUC ↑	
$A2D_p$	61.7	80.6	57.0	78.0	**42.0**	**87.0**	40.1	85.1	50.2	82.7	73.7
$A2D_f$	63.7	79.1	62.2	75.3	46.5	85.2	36.1	88.7	52.1	82.1	72.5
$A2D_h$	63.3	74.0	66.3	74.1	53.8	81.1	34.5	87.7	54.5	79.2	73.7
FM	62.9	74.3	67.7	74.4	54.9	80.6	33.5	87.6	54.8	79.2	73.7
DPU_p	$54.0_{\pm 2.4}$	$83.2_{\pm 0.4}$	$52.6_{\pm 1.5}$	$81.8_{\pm 0.4}$	$46.6_{\pm 2.4}$	$84.8_{\pm 1.2}$	$30.5_{\pm 0.5}$	$89.7_{\pm 0.9}$	45.9	84.9	75.2
DPU_f	$58.5_{\pm 6.5}$	$84.9_{\pm 0.7}$	$57.7_{\pm 3.1}$	$82.9_{\pm 1.1}$	$52.3_{\pm 3.5}$	$84.6_{\pm 0.8}$	$\mathbf{24.5}_{\pm 0.5}$	$\mathbf{91.7}_{\pm 0.9}$	48.3	86.0	75.2
DPU_h	$59.7_{\pm 6.1}$	$79.5_{\pm 3.2}$	$59.2_{\pm 3.6}$	$76.7_{\pm 3.5}$	$48.6_{\pm 2.3}$	$82.6_{\pm 1.3}$	$25.6_{\pm 0.5}$	$90.8_{\pm 0.8}$	48.3	82.4	75.2
Ours	$\mathbf{49.9}_{\pm 4.2}$	$\mathbf{88.5}_{\pm 0.6}$	$\mathbf{37.6}_{\pm 2.3}$	$\mathbf{90.8}_{\pm 0.8}$	$44.2_{\pm 2.5}$	$86.2_{\pm 1.0}$	$27.1_{\pm 1.2}$	$90.9_{\pm 1.0}$	**39.7**	**89.1**	75.2

predicted logit over the validation samples and use it to scale $s(\mathbf{x})$ and $r(\mathbf{x})$ to have the same statistics, i.e.,

$$\alpha = \frac{\sigma_l}{\sigma_s}, \quad \beta = \mu_l - \alpha\mu_s \tag{7}$$

$$s'(\mathbf{x}) = \gamma_s(\alpha s(\mathbf{x}) + \beta), \tag{8}$$

where μ_l, σ_l, μ_s, and σ_s are the means and standard deviations of the highest logit and multi-OOD score, respectively, and analogously construct $r'(\mathbf{x})$ from $r(\mathbf{x})$. The hyperparameters γ_s and γ_r control the relative strength of the components. Then, the scaled scores are integrated with the predicted logits to obtain the final OOD score,

$$g(\mathbf{x}) = \frac{e^{s'(\mathbf{x})} + e^{r'(\mathbf{x})}}{\sum_{i=1}^{C} e^{\mathbf{z}_i^{(f)}} + e^{s'(\mathbf{x})} + e^{r'(\mathbf{x})}}. \tag{9}$$

Intuitively, this normalized score acts as a logit for the OOD class: confident predictions into one of the training classes will decrease the OOD probability, whereas abnormal multi-modal representations will increase it.

4 Experiments

Evaluation protocol. We follow the Multi-OOD benchmark [9] with Near- and Far-OOD experiments using Kinetics-600 [3], HMDB51 [21], UCF101 [36], EPIC-Kitchen [6], and HAC [8]. We evaluate our method against the strongest reported versions of previous multi-modal OOD methods: A2D with NP-Mix (A2D) [9], DPU [26], and FM [30]. As these methods are evaluated using a variety of off-the-shelf detectors, we compare against the best-performing variant of each baseline method across three detector categories: probabilities/logits-based, feature-based, and hybrid versions. We denote these with subscripts $_{\text{p}}$, $_{\text{f}}$, and $_{\text{h}}$, respectively. Thus, $A2D_f$ denotes the best results of A2D with a feature-based scoring function. We give the exact method used for every number and full

Table 2: HMDB51 Far-OOD Detection results with video and optical flow. FM results taken from [30]. We report the mean and standard deviation over three runs. Our method achieves the best overall results.

Method	OOD Datasets										Acc
	Kinetics-600		UCF101		HAC		EPIC-Kitchen		Average		
	FPR ↓	AUC ↑	FPR ↓	AUC ↑	FPR ↓	AUC ↑	FPR ↓	AUC ↑	FPR ↓	AUC ↑	
$A2D_p$	$24.1_{\pm 5.7}$	$94.3_{\pm 1.3}$	$35.5_{\pm 2.7}$	$90.2_{\pm 0.1}$	$22.7_{\pm 4.4}$	$94.6_{\pm 0.9}$	$9.5_{\pm 5.1}$	$97.4_{\pm 1.1}$	23.0	94.1	86.9
$A2D_f$	$11.7_{\pm 0.4}$	$97.1_{\pm 0.2}$	$31.6_{\pm 2.0}$	$89.9_{\pm 1.0}$	$9.4_{\pm 1.4}$	$97.7_{\pm 0.4}$	$10.9_{\pm 1.5}$	$96.5_{\pm 0.3}$	15.9	95.3	86.9
$A2D_h$	$10.5_{\pm 1.8}$	$97.8_{\pm 0.3}$	$28.6_{\pm 2.1}$	$91.6_{\pm 0.7}$	$8.0_{\pm 1.7}$	$98.3_{\pm 0.3}$	$5.7_{\pm 1.4}$	$98.4_{\pm 0.3}$	13.2	96.5	86.9
FM	19.6	94.7	34.3	90.1	16.0	95.5	10.2	96.3	20.0	94.2	87.0
DPU_p	$19.3_{\pm 0.2}$	$96.0_{\pm 0.4}$	$27.2_{\pm 1.9}$	$93.7_{\pm 0.5}$	$14.8_{\pm 1.3}$	$96.9_{\pm 0.4}$	$6.3_{\pm 7.0}$	$\mathbf{98.7}_{\pm 1.2}$	16.9	96.3	87.4
DPU_f	$10.4_{\pm 1.7}$	$97.7_{\pm 0.3}$	$24.5_{\pm 1.3}$	$92.4_{\pm 0.7}$	$9.2_{\pm 1.3}$	$97.9_{\pm 0.7}$	$10.4_{\pm 2.0}$	$96.8_{\pm 0.8}$	13.6	96.2	87.4
DPU_h	$8.5_{\pm 1.0}$	$\mathbf{98.1}_{\pm 0.0}$	$23.2_{\pm 3.6}$	$93.0_{\pm 0.8}$	$\mathbf{6.8}_{\pm 0.8}$	$\mathbf{98.6}_{\pm 0.3}$	$6.0_{\pm 1.2}$	$98.4_{\pm 0.5}$	11.1	**97.0**	87.4
Ours	$\mathbf{8.3}_{\pm 0.5}$	$97.6_{\pm 0.5}$	$\mathbf{20.4}_{\pm 1.3}$	$\mathbf{94.8}_{\pm 0.8}$	$8.2_{\pm 1.4}$	$97.1_{\pm 0.6}$	$\mathbf{3.6}_{\pm 1.9}$	$98.3_{\pm 0.8}$	**10.1**	**97.0**	87.4

Table 3: Multi-modal Near-OOD Detection results using video and optical flow. We report the mean and standard deviation over three runs. ID accuracies are shown in the supplementary material. Our method reaches the highest average performance.

Method	Datasets									
	HMDB51		UCF101		Kinetics-600		EPIC-Kitchen		Average	
	FPR ↓	AUC ↑	FPR ↓	AUC ↑	FPR ↓	AUC ↑	FPR ↓	AUC ↑	FPR ↓	AUC ↑
$A2D_p$	$38.4_{\pm 3.4}$	$89.1_{\pm 0.7}$	$8.0_{\pm 1.8}$	$98.3_{\pm 0.2}$	$\mathbf{61.2}_{\pm 2.9}$	$\mathbf{77.4}_{\pm 0.7}$	$70.0_{\pm 4.9}$	$71.7_{\pm 3.4}$	44.4	84.1
$A2D_f$	$36.5_{\pm 2.2}$	$89.8_{\pm 0.2}$	$12.7_{\pm 0.4}$	$97.4_{\pm 0.2}$	$64.0_{\pm 0.2}$	$77.1_{\pm 0.8}$	$75.8_{\pm 2.3}$	$68.1_{\pm 1.9}$	47.3	83.1
$A2D_h$	$36.2_{\pm 0.8}$	$88.2_{\pm 0.5}$	$7.0_{\pm 0.9}$	$\mathbf{98.4}_{\pm 0.2}$	$63.2_{\pm 3.3}$	$77.3_{\pm 1.3}$	$71.5_{\pm 2.8}$	$67.1_{\pm 2.2}$	44.5	82.8
DPU_p	$35.9_{\pm 2.4}$	$89.0_{\pm 0.4}$	$10.0_{\pm 4.6}$	$97.7_{\pm 0.8}$	$63.4_{\pm 0.2}$	$76.9_{\pm 0.7}$	$70.7_{\pm 2.1}$	$68.5_{\pm 0.7}$	45.0	83.0
DPU_f	$35.4_{\pm 3.3}$	$89.6_{\pm 0.5}$	$10.6_{\pm 1.8}$	$97.6_{\pm 0.2}$	$66.6_{\pm 1.1}$	$76.6_{\pm 0.8}$	$76.0_{\pm 3.7}$	$67.0_{\pm 1.8}$	47.2	82.7
DPU_h	$34.9_{\pm 2.2}$	$89.1_{\pm 0.8}$	$7.8_{\pm 0.5}$	$98.2_{\pm 0.2}$	$63.8_{\pm 0.3}$	$77.1_{\pm 0.7}$	$72.8_{\pm 1.8}$	$66.5_{\pm 0.2}$	44.8	82.7
Ours	$\mathbf{33.8}_{\pm 3.0}$	$\mathbf{90.3}_{\pm 0.3}$	$\mathbf{6.5}_{\pm 1.6}$	$\mathbf{98.4}_{\pm 0.3}$	$63.0_{\pm 1.1}$	$77.0_{\pm 0.4}$	$\mathbf{69.0}_{\pm 1.3}$	$\mathbf{72.1}_{\pm 0.8}$	**43.1**	**84.5**

dataset details in the supplementary material. Following standard procedure, we report the False Positive Rate at 95%, True Positive Rate (FPR), and the Area Under the ROC Curve (AUC). We also report the in-distribution accuracy of the classifiers (Acc).

Implementation details. All methods are implemented with the same architectures. Similar to previous work [9, 26, 30], we use the SlowFast [14] model pre-trained on Kinetics-400 to encode videos. For optical flow, we also use SlowFast pre-trained on Kinetics-400, with a slow-only pathway. We train the baselines using the official settings provided in their repositories across three seeds to assess the variability in results unless specified otherwise. We apply our method to networks trained with DPU, use the L2-distance for $d(\cdot, \cdot)$, and fix γ_s to 0.4, γ_r to 1.25, and k to $p = 0.5\%$ in all experiments.[3]

4.1 Main Results

Far-OOD. We show the results for Far-OOD detection with Kinetics-600 as the in-distribution in Tab. 1 and with HMDB51 as the in-distribution in Tab. 2. On

[3] The code is available at `https://github.com/LarsDoorenbos/modality-aware-ood`

Table 4: Triple-modal OOD detection. *Left:* Results with audio, video, and optical flow over three seeds for EPIC-Kitchens Near-OOD and EPIC-Kitchens:Kinetics Far-OOD. *Right:* Ablating the benefits of multiple modalities for OOD detection on EPIC-Kitchens:Kinetics. Integrating all modalities gives the best results.

	Near AUC	Far AUC	Average FPR	Average AUC
$A2D_p$	**73.3**$_{\pm 2.6}$	78.8$_{\pm 3.3}$	64.2	76.0
$A2D_f$	65.8$_{\pm 2.4}$	87.5$_{\pm 4.0}$	58.7	76.6
$A2D_h$	52.0$_{\pm 5.8}$	**98.0**$_{\pm 0.8}$	45.9	75.0
Ours	65.5$_{\pm 2.4}$	**98.0**$_{\pm 0.3}$	**41.6**	**81.8**

Video	Optical Flow	Audio	Far FPR	Far AUC
✓	✗	✗	15.4	97.0
✗	✓	✗	28.9	93.4
✗	✗	✓	48.5	86.7
✓	✓	✗	12.8	97.2
✓	✓	✓	**10.8**	**98.0**

the Kinetics-600 experiments, we set a new state of the art with HMDB51 and UCF101 as the OOD dataset, and reach the second-best AUC on the remaining two. On average, this leads to an improvement of 6.2 FPR over the next best method. Importantly, this baseline uses the same trained network, highlighting the benefit of our scoring function. We find that the same pattern holds when HMDB51 is used as the in-distribution: our method achieves the overall best performance with an average FPR of 10.1 and AUC of 97.0.

Near-OOD. The Near-OOD results in Tab. 3 show that our approach also works well in this setting. We achieve the best results in all cases except for Kinetics, leading to an overall improvement of 1.3 in average FPR over the next-best method, $A2D_p$. Interestingly, $A2D_p$ scored far worse on Far-OOD: for instance, its average FPR is 12.9 points worse with HMDB as the ID (Tab. 2). For methods using the same network weights, we improve by 1.7 average FPR over DPU_h. In general, these baselines are less consistent, with DPU_h being the best DPU version for Near-OOD, but the worst on Kinetics Far-OOD.

Three modalities. We show triple-modal results in Tab. 4(left) where we use A2D to train the backbone, as DPU does not provide training code for the triple-modal case. In this setting, the probability-based method $A2D_p$ scores very well in the Near-OOD setting, but 19.2 AUC points behind our method for Far-OOD. For the hybrid method $A2D_h$ this is reversed, scoring high for Far-OOD but 13.5 points below our method on Near-OOD. Overall, our method achieves the best average performance. Furthermore, these results also confirm that our method works when A2D is used to train the network.

4.2 The benefits of multiple modalities

The complementary information present in different modalities increases the capabilities of action recognition systems. Here, we show that the same holds for our proposed method. We conduct an experiment using different combinations of the three modalities for the EPIC-Kitchens:Kinetics task and present the results in Tab. 4(right). We find that video is the strongest modality of the

Table 5: Results with early- and mid-fusion networks for Kinetics-600 as the ID dataset. Our method is also successful with attention-based early- and mid-fusion.

Method	Near-OOD			Far-OOD Datasets										
	Kinetics			HMDB51		UCF101		HAC		EPIC-Kitchen		Average		
	FPR ↓	AUC ↑	Acc	FPR ↓	AUC ↑	FPR ↓	AUC ↑	FPR ↓	AUC ↑	FPR ↓	AUC ↑	FPR ↓	AUC ↑	Acc
Early fusion														
DPU_p	63.7	78.6	81.2	70.0	78.4	71.1	73.9	56.2	81.8	38.0	86.2	59.8	79.8	72.7
DPU_f	61.2	**79.3**	81.2	58.8	81.7	55.9	82.0	49.6	83.8	35.9	87.4	52.3	82.8	72.7
DPU_h	61.6	79.2	81.2	58.8	81.8	54.8	82.3	47.7	83.8	35.9	87.4	51.8	82.9	72.7
Ours	**59.9**	79.0	81.2	**53.7**	**88.1**	**40.0**	**91.0**	**46.3**	**86.0**	**33.2**	**87.9**	**46.6**	**86.4**	72.7
Mid fusion														
DPU_p	63.9	78.5	81.0	69.2	79.6	70.0	74.5	53.8	83.3	33.1	88.8	58.0	80.9	74.4
DPU_f	62.1	**79.0**	81.0	56.3	83.4	57.5	81.5	50.9	84.1	**28.3**	**90.1**	51.0	83.6	74.4
DPU_h	63.1	79.0	81.0	56.3	83.4	57.1	81.5	50.6	84.0	28.8	**90.1**	51.2	83.6	74.4
Ours	**60.0**	78.7	81.0	**48.3**	**89.0**	**40.2**	**90.2**	**47.6**	**85.8**	29.9	89.5	**45.2**	**86.4**	74.4

three, followed by optical flow. By itself, audio performs 33.1 FPR worse than video. However, when integrated with the other modalities in our method, audio still brings an improvement of 2.0 FPR over video and flow alone.

4.3 Effectiveness under different fusion mechanisms

In the previous sections, we followed the late fusion architecture design used in all previous multi-modal OOD studies. Here, we additionally validate our approach on attention-based early- and mid-fusion networks using the same settings as before. To do so, we train two DPU networks replacing the multi-modal head h_f with a transformer: one with cross-attention layers throughout (early-fusion) and one with cross-attention starting halfway through the network (mid-fusion).

We show the results for both Near- and Far-OOD in Tab. 5, where we find similar trends as the main results: our method outperforms the baselines in this regime by 5.2-5.8 average FPR. Note that the performance of all methods can be generally higher or lower than previous experiments, depending on how well the architecture fits the action recognition task, as reflected in the ID accuracy. Nonetheless, all methods use the same network, so this does not affect the OOD detection performance comparisons in the table.

4.4 The importance of a hybrid detector

The motivation behind adopting a hybrid approach is that different sources of information provide complementary signals. As the space of OOD samples is large, unifying these sources allows the detection of a broader spectrum of OOD samples. To validate this, we conduct an experiment to assess the benefits of each of the three information sources we use in our detector, i.e., the multi-modal $s(\mathbf{x})$, feature $r(\mathbf{x})$, and probability $g(\mathbf{x})$ signals, in Tab. 6.

The results reveal several notable trends. First, while non-hybrid approaches can achieve good results in isolated cases, they do not perform well for both types of OOD. For example, $r(\mathbf{x})$ trails behind the FPR achieved by our full method on Near-OOD by 7.6 points. Second, hybrid approaches that combine two modalities

Table 6: Ablating hybrid components on Kinetics-600 Near-OOD and Kinetics-600:HAC Far-OOD. We find that each component of our method is important. (†) we use the energy score [31] for only $g(\mathbf{x})$, corresponding to Eq. (9).

$s(\mathbf{x})$	$r(\mathbf{x})$	$g(\mathbf{x})$	Near FPR	Far FPR
✓	✗	✗	64.1	57.1
✗	✓	✗	71.6	52.7
✗	✗	✓†	64.8	46.8
✓	✓	✗	68.4	45.9
✗	✓	✓	64.8	41.7
✓	✗	✓	64.6	47.5
✓	✓	✓	**64.0**	**41.6**

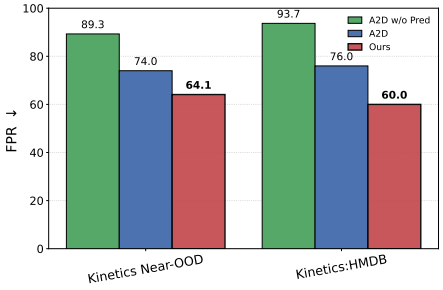


Fig. 4: Despite being trained with A2D, the difference between uni-modal predictions does not provide a strong indicator for OOD. Instead, we find that the relation between the predictions of the uni-modal heads and the prediction of the multi-modal head is a strong signal of normality.

can alleviate some of these shortcomings. The combination of $r(\mathbf{x})$ and $g(\mathbf{x})$, for example, reaches the second-best FPR for the Far-OOD setting. Nonetheless, the best results are achieved by combining all three sources, highlighting the importance of a holistic approach to OOD detection for multi-modal action recognition.

4.5 Limitations of Modality Disagreement

The A2D method [9] is motivated by the observation that the distance between the predictions on individual modalities is larger for OOD samples, and their regularization term aims to increase this disagreement after excluding the ground-truth class. Despite this, A2D and related methods ignore this property at inference time, instead relying on off-the-shelf OOD detectors. Nonetheless, A2D encodes modality disagreement as a signal for OOD in two ways: the distance between the full uni-modal probability distributions, which we refer to as *A2D*, and the distance between uni-modal predictions excluding the predicted class, denoted by *A2D w/o Pred*. To test their efficacy, we evaluate both variants as OOD detectors and compare them to our proposed score $s(\mathbf{x})$. As shown in Fig. 4, the discrepancy between uni-modal predictions alone is a weak indicator of OOD. In contrast, our score based on the relation between uni- and multi-modal predictions provides a stronger signal, improving the FPR on Kinetics Near-OOD by 9.9 FPR and on Kinetics:HMDB Far-OOD by 16.0 FPR.

4.6 Ablations

Complexity of w. We ablate the choice for our linear system ("class-conditional") in Eq. (2) with respect to other possible versions. Specifically, we compare to a

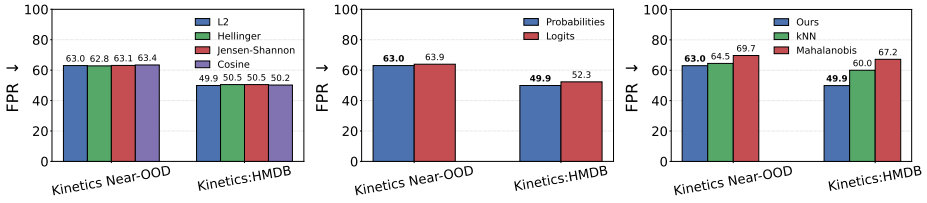


Fig. 5: Ablation studies for our framework. We evaluate the impact of (a) the distance function, (b) the choice for probabilities over logits in Eq. (4), and (c) variants of feature-based scores for $r(\mathbf{x})$.

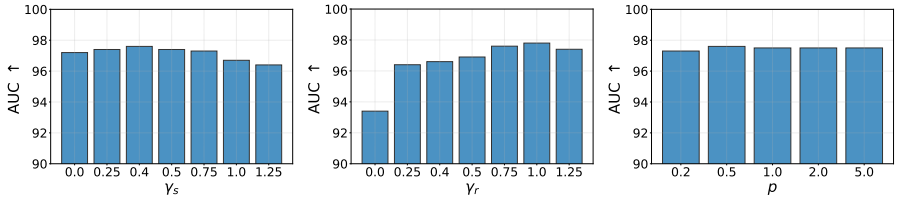


Fig. 6: Sensitivity to hyperparameters. We show the AUC for HMDB51:Kinetics Far-OOD for different settings of the hyperparameters a) γ_s, b) γ_r, and c) p. Our method is robust to their settings.

fully connected $\mathbf{W} \in \mathbb{R}^{(MC)\times C}$ ("fully connected"), where every uni-modal logit influences every multi-modal logit, and a version with only a single $\tilde{\mathbf{w}}$ shared between all classes ("single linear"). We show the results in Tab. 7(left), where we find that the overall performance is stable, with the class-conditional model slightly outperforming the other two variants.

Other choices. We ablate the choice of distance function (Eq. (4)) in Fig. 5(a). The differences between the distance functions are marginal: out of the four distances considered, the maximum difference in FPR on the two experiments is 0.6. In Fig. 5(b), we justify our choice to use the distance between probability distributions rather than logits. In general, the logits for ID samples have a higher norm, which makes the distance-based comparison imbalanced. By considering probability distributions instead, the resulting values for ID and OOD samples have the same range, leading to better performance. Finally, we compare our feature-based score $r(\mathbf{x})$ with the kNN distance [37] and the class-wise Mahalanobis method [25]. The results in Fig. 5(c) confirm that the choice for invariants results in the best performance.

Sensitivity to validation set size. We test the sensitivity of our method to the validation set size and show the results in Tab. 7(right). We do not find a strong degradation in performance; even when only using 1% of the validation set, the performance of our overall model remains within 0.2 points of the

Table 7: Ablating w complexity and validation set size. *Left:* We report the results for Kinetics Near-OOD with early and late fusion architectures. Our class-conditional approach gives the best balance. *Right:* Robustness of results with respect to subsampling the validation set for Kinetics Near-OOD and Kinetics:HMDB. Our results remain stable even when only using a fraction of the validation samples.

	Early		Late	
	FPR	AUC	FPR	AUC
Single linear	59.9	79.0	63.3	77.1
Class-conditional	59.9	79.0	63.0	77.0
Fully connected	59.9	79.0	63.4	77.0

	Near		Far	
	FPR	AUC	FPR	AUC
0.5%	63.3	76.5	50.5	88.0
1%	63.2	77.2	50.8	88.3
5%	63.0	77.0	50.5	88.4
10%	63.0	77.1	50.3	88.5
Full	63.0	77.0	49.9	88.5

AUC obtained with the full validation set. This is a direct consequence of only estimating the intercept plus one parameter per modality for each class.

Sensitivity to hyperparameters. Finally, we probe the effect of the hyperparameters of our method: γ_s, γ_r, and p. We present the results in Fig. 6, which demonstrate that our method is generally robust to their settings.

Limitations. Our method inherits a limitation of the broader multi-modal OOD field: a reliance on architectures with exposed uni-modal classifiers. However, we go beyond previous works in this domain by showing that our method also works under various fusion mechanisms. Moreover, we show in the supplementary material that our method still performs well when adding uni-modal heads post hoc to an already trained classifier. Qualitative examples including failure cases are provided in the supplementary material.

5 Conclusion

We presented the first modality-aware OOD detector for multi-modal action recognition. Our approach is based on the observation that the probability distribution predicted by the multi-modal classifier can be modeled as an aggregation of the uni-modal predictions. The residual error between the observed and predicted distributions is higher for OOD samples, leading to a novel modality-aware OOD scoring function. We integrate this score with information from the feature space by modeling the invariants in the multi-modal manifold and the final predictions as a virtual logit. Through combining these complementary signals, we outperform the state of the art on average on the challenging MultiOOD benchmark. Overall, our findings demonstrate that our approach can enhance the robustness of multi-modal action recognition systems. As our approach does not make any modality-specific assumptions, it generalizes well to other domains, as shown by an initial exploration in the supplementary material.

Acknowledgements

The work has been supported by the Federal Ministry of Research, Technology and Space (BMFTR) under grant no. 16IS22094A WEST-AI and the ERC Consolidator Grant FORHUE (101044724). For the computations involved in this research, we acknowledge EuroHPC Joint Undertaking for awarding us access to Leonardo at CINECA, Italy, through EuroHPC Regular Access Call - proposal No. EHPC-REG-2025R01-218.

References

1. Ahmed, F., Courville, A.: Detecting semantic anomalies. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 34, pp. 3154–3162 (2020)
2. Behpour, S., Doan, T.L., Li, X., He, W., Gou, L., Ren, L.: Gradorth: A simple yet efficient out-of-distribution detection with orthogonal projection of gradients. Advances in Neural Information Processing Systems **36**, 38206–38230 (2023)
3. Carreira, J., Noland, E., Banki-Horvath, A., Hillier, C., Zisserman, A.: A short note about kinetics-600. arXiv preprint arXiv:1808.01340 (2018)
4. Carreira, J., Zisserman, A.: Quo vadis, action recognition? a new model and the kinetics dataset. In: proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 6299–6308 (2017)
5. Chéron, G., Laptev, I., Schmid, C.: P-cnn: Pose-based cnn features for action recognition. In: Proceedings of the IEEE international conference on computer vision. pp. 3218–3226 (2015)
6. Damen, D., Doughty, H., Farinella, G.M., Fidler, S., Furnari, A., Kazakos, E., Moltisanti, D., Munro, J., Perrett, T., Price, W., et al.: Scaling egocentric vision: The epic-kitchens dataset. In: Proceedings of the European conference on computer vision (ECCV). pp. 720–736 (2018)
7. DeVries, T., Taylor, G.W.: Learning confidence for out-of-distribution detection in neural networks. arXiv preprint arXiv:1802.04865 (2018)
8. Dong, H., Nejjar, I., Sun, H., Chatzi, E., Fink, O.: Simmmdg: A simple and effective framework for multi-modal domain generalization. Advances in Neural Information Processing Systems **36**, 78674–78695 (2023)
9. Dong, H., Zhao, Y., Chatzi, E., Fink, O.: Multiood: Scaling out-of-distribution detection for multiple modalities. Advances in Neural Information Processing Systems **37**, 129250–129278 (2024)
10. Doorenbos, L., Sznitman, R., Márquez-Neila, P.: Data invariants to understand unsupervised out-of-distribution detection. In: European Conference on Computer Vision. pp. 133–150. Springer (2022)
11. Doorenbos, L., Sznitman, R., Márquez-Neila, P.: Non-linear outlier synthesis for out-of-distribution detection. arXiv preprint arXiv:2411.13619 (2024)
12. Du, X., Sun, Y., Zhu, J., Li, Y.: Dream the impossible: Outlier imagination with diffusion models. Advances in Neural Information Processing Systems **36** (2024)
13. Du, X., Wang, Z., Cai, M., Li, Y.: Vos: Learning what you don't know by virtual outlier synthesis. In: Proceedings of the International Conference on Learning Representations (2022)
14. Feichtenhofer, C., Fan, H., Malik, J., He, K.: Slowfast networks for video recognition. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 6202–6211 (2019)

15. Hendrycks, D., Basart, S., Mazeika, M., Mostajabi, M., Steinhardt, J., Song, D.: Scaling out-of-distribution detection for real-world settings. International Conference on Machine Learning (2022)
16. Hendrycks, D., Gimpel, K.: A baseline for detecting misclassified and out-of-distribution examples in neural networks. Proceedings of International Conference on Learning Representations (2017)
17. Hendrycks, D., Mazeika, M., Dietterich, T.: Deep anomaly detection with outlier exposure. International Conference on Learning Representations (2019)
18. Huang, R., Geng, A., Li, Y.: On the importance of gradients for detecting distributional shifts in the wild. Advances in Neural Information Processing Systems **34**, 677–689 (2021)
19. Huh, M., Cheung, B., Wang, T., Isola, P.: Position: The platonic representation hypothesis. In: Forty-first International Conference on Machine Learning (2024)
20. Kamoi, R., Kobayashi, K.: Why is the mahalanobis distance effective for anomaly detection? arXiv preprint arXiv:2003.00402 (2020)
21. Kuehne, H., Jhuang, H., Garrote, E., Poggio, T., Serre, T.: Hmdb: a large video database for human motion recognition. In: 2011 International conference on computer vision. pp. 2556–2563. IEEE (2011)
22. Lai, K., Bo, L., Ren, X., Fox, D.: A large-scale hierarchical multi-view rgb-d object dataset. In: 2011 IEEE international conference on robotics and automation. pp. 1817–1824. IEEE (2011)
23. Ledoit, O., Wolf, M.: A well-conditioned estimator for large-dimensional covariance matrices. Journal of multivariate analysis **88**(2), 365–411 (2004)
24. Lee, K., Lee, H., Lee, K., Shin, J.: Training confidence-calibrated classifiers for detecting out-of-distribution samples. International Conference on Learning Representations (2018)
25. Lee, K., Lee, K., Lee, H., Shin, J.: A simple unified framework for detecting out-of-distribution samples and adversarial attacks. Advances in neural information processing systems **31** (2018)
26. Li, S., Gong, H., Dong, H., Yang, T., Tu, Z., Zhao, Y.: Dpu: Dynamic prototype updating for multimodal out-of-distribution detection. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 10193–10202 (2025)
27. Liang, J., Hou, R., Hu, M., Chang, H., Shan, S., Chen, X.: Revisiting logit distributions for reliable out-of-distribution detection. Advances in Neural Information Processing Systems (2025)
28. Liang, S., Li, Y., Srikant, R.: Enhancing the reliability of out-of-distribution image detection in neural networks. In: Proceedings of the International Conference on Learning Representations (2018)
29. Ling, Z., Chang, Y., Zhao, H., Zhao, X., Chow, K., Deng, S.: Cadref: Robust out-of-distribution detection via class-aware decoupled relative feature leveraging. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 4968–4977 (2025)
30. Liu, M., Dong, H., Kelly, J., Fink, O., Trapp, M.: Extremely simple multimodal outlier synthesis for out-of-distribution detection and segmentation. Advances in Neural Information Processing Systems (2025)
31. Liu, W., Wang, X., Owens, J., Li, Y.: Energy-based out-of-distribution detection. Advances in Neural Information Processing Systems (2020)
32. Liu, X., Lochman, Y., Zach, C.: Gen: Pushing the limits of softmax-based out-of-distribution detection. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 23946–23955 (2023)

33. Mueller, M., Hein, M.: Mahalanobis++: Improving ood detection via feature normalization. Internation Conference on Machine Learning (2025)
34. Shaikh, M.B., Chai, D., Islam, S.M.S., Akhtar, N.: Multimodal fusion for audio-image and video action recognition. Neural Computing and Applications **36**(10), 5499–5513 (2024)
35. Sharifi, S., Entesari, T., Safaei, B., Patel, V.M., Fazlyab, M.: Gradient-regularized out-of-distribution detection. In: European Conference on Computer Vision. pp. 459–478. Springer (2024)
36. Soomro, K., Zamir, A.R., Shah, M.: Ucf101: A dataset of 101 human actions classes from videos in the wild. arXiv preprint arXiv:1212.0402 (2012)
37. Sun, Y., Ming, Y., Zhu, X., Li, Y.: Out-of-distribution detection with deep nearest neighbors. In: Proceedings of the International Conference on Machine Learning. pp. 20827–20840 (2022)
38. Sun, Z., Ke, Q., Rahmani, H., Bennamoun, M., Wang, G., Liu, J.: Human action recognition from various data modalities: A review. IEEE transactions on pattern analysis and machine intelligence **45**(3), 3200–3225 (2022)
39. Tang, K., Hou, C., Peng, W., Fang, X., Wu, Z., Nie, Y., Wang, W., Tian, Z.: Simplification is all you need against out-of-distribution overconfidence. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 5030–5040 (2025)
40. Wang, H., Li, Z., Feng, L., Zhang, W.: Vim: Out-of-distribution with virtual-logit matching. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 4921–4930 (2022)
41. Winkens, J., Bunel, R., Roy, A.G., Stanforth, R., Natarajan, V., Ledsam, J.R., MacWilliams, P., Kohli, P., Karthikesalingam, A., Kohl, S., et al.: Contrastive training for improved out-of-distribution detection. arXiv preprint arXiv:2007.05566 (2020)
42. Yang, J., Zhou, K., Li, Y., Liu, Z.: Generalized out-of-distribution detection: A survey. International Journal of Computer Vision **132**(12), 5635–5662 (2024)
43. Yang, Y., Xu, H.: Strengthen out-of-distribution detection capability with progressive self-knowledge distillation. In: Forty-second International Conference on Machine Learning (2025)