Leveraging Dark Knowledge for Intrinsic Multimodal Out-of-Distribution Detection

Nimeshika Udayangani[1], Sarah Erfani[1], and Christopher Leckie[1]

School of Computing and Information Systems, The University of Melbourne, Parkville, VIC, Australia
nhewadehigah@student.unimelb.edu.au

Abstract. Out-of-distribution (OOD) detection is crucial for the safe deployment of deep neural models in applications such as autonomous driving. With the emerging multimodal nature of modern applications, recent attention has shifted toward OOD detection in multimodal settings. However, current multimodal OOD detection methods fail to fully exploit the synergy among modalities: they treat all modalities equally, disregarding their varying detection performance, and they are unable to capture the diverse uncertainty information encoded at the logit level. In this paper, we propose to exploit the *dark knowledge* within unimodal experts as the key to revealing their synergy. To this end, we introduce a self multimodal OOD distillation framework, which leverages logits as uncertainty-aware soft targets to train a holistic model that operates in the joint embedding space of all modalities. Specifically, the proposed framework accounts for the negative effects of underperforming modalities and effectively fuses both the rich feature-level knowledge and the logit-level knowledge of modalities. As a result, our method improves the performance of current state-of-the-art multimodal OOD detection methods, achieving gains of up to 30% across diverse OOD detection benchmarks, spanning two tasks and five multimodal OOD datasets.

Keywords: Out-of-distribution detection · Multimodal learning

1 Introduction

Out-of-distribution (OOD) detection aims to reliably identify samples that differ from the training distribution, which is a crucial requirement for deploying deep neural networks in safety-critical domains such as autonomous driving [22] and healthcare [7]. Consequently, a significant body of research has been dedicated to OOD detection [37], where the most standard and fundamental approach is to derive a confidence score from a deep neural classifier [8, 9, 20]. Many other approaches have also been proposed, ranging from distance-based methods [17, 30, 33] to retraining strategies that enhance ID–OOD separability [5, 10, 32]. However, they have predominantly focused on single-modality inputs, such as images or videos, despite the inherently multimodal nature of real-world applications [18].

More recently, Dong *et al*. [4] introduced the first multimodal OOD benchmark, enabling methods to leverage the complementary nature of various modalities for OOD detection. They observed that distributional shift manifests as increased diversity in predictions across modalities and proposed an algorithm to amplify the OOD discrepancy during training. Since then, several works have been proposed [18, 19] for multimodal OOD detection. However, these methods are unable to capture the full synergy between modalities due to several limiting factors. For example, they employ complex retraining strategies to enhance the embedding space for better OOD separation, including contrastive learning [18], maximizing disagreement between models on OOD samples [4], or improving supervision from unknown data via outlier synthesis [19], while paying limited attention to the intrinsic OOD detection capability of unimodal models. A final classifier is then employed by fusing the enhanced embeddings from unimodal models, incorporating one-hot targets. However, they overlook the diverse uncertainty knowledge contained at the logit level of each unimodal expert, which can be incorporated to strengthen OOD detection. In particular, a key challenge underlying OOD overconfidence is that *ground-truth* uncertainty labels for training samples are unavailable [16]. Instead, only one-hot class labels are available, which cause the model to treat all class samples uniformly, despite their varying degrees of uncertainty, potentially leading to overconfident predictions [38]. On the other hand, we found that, as illustrated in Fig. 1, some modalities (e.g., video) can achieve stronger detection performance than others (e.g., optical flow)—a fact largely overlooked in previous research, where all modalities are treated equally without accounting for their varying detection capabilities.

To address these challenges and fully exploit the synergy across modalities for OOD detection, we propose a self multimodal OOD distillation framework, which trains a holistic model whose guidance is self-provided by unimodal experts serving as teachers. Traditionally, the hidden information within the teacher has been referred to as *dark knowledge* [12]. Accordingly, our framework leverages the uncertainty-aware dark knowledge encoded at the logit level of unimodal experts to strengthen multimodal OOD detection performance. Given that ground-truth uncertainty is unavailable, we first estimate predictive uncertainty by approximating it through the collective behavior of all models, while accounting for underperforming modalities. We then leverage this estimated uncertainty as soft targets to train a joint classifier in the embedding space shared across all modalities, thereby reducing OOD overconfi-

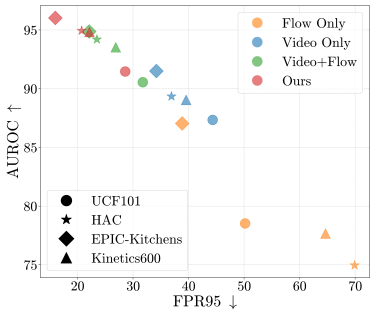

Fig. 1: OOD detection performance on the HMDB51 dataset across different modalities. The detection performance of *Flow* is substantially lower than that of *Video*. Multimodal OOD detection (*Video+Flow*, obtained using one of the SOTA [4]) improves over unimodal OOD, while our method more effectively harnesses the synergy among modalities.

dence and accounting for disparities among modalities. This joint classifier can be conceptually regarded as the student, guided by the same ensemble of unimodal teachers, who also collectively share the embedding space. In this way, our method exploits both the diverse uncertainty knowledge encoded at the logit level and the representational knowledge at the feature level of unimodal experts, thereby revealing their ultimate synergy. As part of this work, we also provide a comprehensive analysis of how the uncertainty-aware dark knowledge in our framework contributes to improving multimodal OOD detection performance, which cannot be achieved by naively combining ensembles or merely fusing features across modalities.

This principled and effective framework of ours allows us to integrate it as a plug-in to mine the intrinsic OOD detection capabilities of a given multimodal model set, thereby alleviating the burden of retraining large modality-specific models. A simple integration of our framework into a vanilla baseline yields substantial improvements in OOD performance (Fig. 1), achieving comparable or even superior results to current retraining-based state-of-the-art multimodal OOD detection methods. Moreover, being agnostic to these methods, it further improves their performance with gains up to 30% across diverse OOD detection benchmarks, spanning two tasks and five OOD datasets.

2 Related Work

2.1 Out-of-Distribution (OOD) Detection

A common approach to OOD detection is to leverage confidence scores derived from deep neural classifiers, such as the maximum softmax probability (MSP) [9], Energy [20]—which also mirrors the class-conditional probability— Generalized Entropy (GEN) [21], and the maximum of logit (MaxLogit) [8]. In contrast to these classifier-based scores, kNN [30] and Mahalanobis [17] rely on distance metrics in feature space for OOD detection, while Virtual Logit Matching (VIM) [35] and SupLID [33] integrates information from both feature and logit spaces to define the OOD score. Other approaches, such as ReAct [29] and ASH [2], enhance existing scoring functions by modifying input activations. In contrast to the above post-hoc methods, retraining-based approaches [5, 10, 32] aim to enhance representations for better ID–OOD separability. However, all these methods predominantly focus on single-modality inputs, without accounting for the multimodal nature of recent applications.

2.2 Multimodal OOD Detection

With the emergence of large vision–language models [27], the task of *multimodal OOD detection* was initially associated with approaches that exploited semantic information from text labels [23, 36], although their scope remained restricted to image-only benchmarks. To address this, Dong *et al.* [4] introduced the first multimodal OOD benchmark incorporating multiple modality combinations (i.e., video, audio, and optical flow), and observed that OOD samples tend

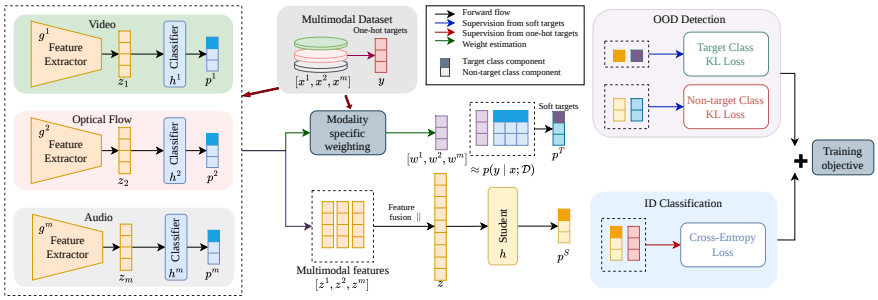

Fig. 2: An overview of the proposed self multimodal OOD distillation framework.

to exhibit higher discrepancy among unimodal prediction distributions. They proposed the Agree-to-Disagree algorithm to amplify this discrepancy during training and further introduced multimodal outlier synthesis (NP-Mix) to obtain a more discriminative embedding space for OOD detection. More recently, DPU [18] employed a contrastive learning strategy to account for intra-class variability in multimodal OOD detection, while Feature Mixing [19] introduced yet another synthetic outlier–based approach for training. However, these methods are unable to reveal the full synergy between modalities, in that they treat all modalities equally, without considering their varying detection performance. In addition, they are unable to capture holistic uncertainty across modalities, leaving the rich uncertainty-aware information within the logit layers of unimodal experts untapped. Moreover, they rely on complex retraining approaches, especially with a focus on providing a more discriminative embedding space for OOD detection, while little attention has been given to harnessing the intrinsic OOD detection capability of the ensemble itself.

2.3 Logit Distillation

Logit distillation, which transfers the *dark knowledge* within logits of a large model to a smaller one, was initially proposed for model compression [12], but has since been shown to also improve model accuracy [11, 31, 40]. For example, retraining with self-distillation, has been found to further improve accuracy [40]. Zhao *et al*. [41] sought to shed light on dark knowledge by decoupling it into two components: binary knowledge related to the target class and relational knowledge pertaining to the non-target classes. More closely related to our work, Yang *et al*. [38] employed self-knowledge distillation for OOD detection to mitigate the negative effect from atypical samples during training. However, their method is limited to single modalities, whereas we leverage it to harness the synergy among multiple modalities. Logit distillation has also been explored in multimodal applications for cross-modal adaptation [26], but not to OOD detection. To the best of our knowledge, this is the first work that methodically leverages and analyzes the role of dark knowledge in improving multimodal OOD detection.

3 Methodology

In this section, we present our self multimodal OOD distillation framework. We begin with a formal introduction to the multimodal OOD detection problem, followed by an overview of the framework and a detailed description of the method. We also provide further insights into uncertainty-aware dark knowledge within our framework at the end of this section.

3.1 Problem Definition

Let $\mathcal{D}_{\text{in}} = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^{n}$ be our training set, where each $\boldsymbol{x}_i \in \mathcal{X}$ is an input sample and $y_i \in \mathcal{Y} = \{1, 2, \ldots, C\}$ is the corresponding class label. In a multimodal setting each training sample $\boldsymbol{x}_i$ is comprised of M modalities, denoted as $\boldsymbol{x}_i = \{\, \boldsymbol{x}_i^m \mid m = 1, \ldots, M \,\}$. Let $\{f^m : \mathcal{X}^m \to \mathbb{R}^C\}_{m=1}^{M}$ be a set of neural networks each trained on a single modality (e.g., video). The goal of multimodal OOD detection is to effectively combine information from all modalities in order to correctly identify samples with semantic shifts. Each model f^m comprises a feature extractor $g^m(\cdot)$, which extracts an embedding $\boldsymbol{z}^m$ for its corresponding modality m and the usual practice is to employ a final classifier $h(\cdot)$, which takes the combined embeddings from all modalities as input and outputs a prediction probability $\boldsymbol{p}^{\text{S}} = \delta(h([\, g^1(\boldsymbol{x}^1) \,||\, \ldots \,||\, g^M(\boldsymbol{x}^M) \,])) = \delta(h([\, \boldsymbol{z}^1 \,||\, \ldots \,||\, \boldsymbol{z}^M \,])) = \delta(h(\boldsymbol{z}))$, where $||$ denotes concatenation and $\delta(\cdot)$ denotes the softmax function.

When deploying in the real world, h should not only accurately classify known samples from the in-distribution (ID), but also detect samples that exhibit semantic shifts compared to ID samples and do not belong to any class in $\mathcal{Y}$, i.e., out-of-distribution (OOD) samples. This aims to define a separate scoring function $S : \mathcal{X} \to \mathbb{R}$, which assigns a high score $S(\boldsymbol{x})$ for samples $\boldsymbol{x} \sim \mathcal{D}_{\text{in}}$ and a low score otherwise. A threshold η is then used to classify each sample $\boldsymbol{x} \in \mathcal{X}$ as ID or OOD:

$$\text{Decision}(\boldsymbol{x}) = \begin{cases} \text{ID}, & \text{if } S(\boldsymbol{x}) \geq \eta, \\ \text{OOD}, & \text{if } S(\boldsymbol{x}) < \eta. \end{cases}$$

3.2 Motivation and Method Overview

In this section, we present the motivation behind our approach and provide a brief overview of our framework illustrated in Fig. 2. A common approach to combining complementary knowledge from different modalities is to train a joint classifier (h) in the combined embedding space of modalities, which yields improvements in OOD detection over unimodal models (see *Video+Flow* in Fig. 1). However, this alone is unable to fully leverage the synergy between modalities, as it does not properly account for underperforming modalities, and it fails to fully exploit the uncertainty-related knowledge encoded in modality-specific models. As shown in Fig. 1, some modalities (e.g., optical flow) exhibit weaker OOD detection performance than others (e.g., video). Specifically, this disparity across modalities–whose weakness partly arises from OOD overconfidence– is difficult

to address solely at the feature level; therefore, we turn to the logit level for complementary guidance. In particular, a key challenge underlying OOD overconfidence is that *ground-truth* uncertainty labels for training samples are not available. Therefore, we first model the predictive uncertainty by treating the collective behavior of all models as an approximation to the true uncertainty. We then employ logit distillation to leverage this estimated uncertainty as soft targets to train the joint classifier (h in Fig. 2). In this framework, h acts as the student, and its guidance is self-provided by the set of unimodal teachers that collectively share the embedding space. We also employ decoupled KL divergence [41] for logit distillation, separating it into target and non-target class components (the purple block in Fig. 2), which allows better weight adjustment across tasks of varying difficulty (e.g., near-OOD vs. far-OOD). Ultimately, our framework not only fuses rich feature-space knowledge but also leverages the uncertainty-aware knowledge residing in the logit space of unimodal teachers in a weighted manner, thus revealing the synergy among them. Building on these insights, we present a detailed and formal description of our framework in the next section.

3.3 Self Multimodal OOD Distillation Framework

To improve the multimodal OOD detetcion we focus on both feature-space and logit-space knowledge of multimodal models. We begin with the joint classifier h that operates on the combined embeddings space of unimodal models $\{f^m\}_{m=1}^M$. In order to model the holistic predictive capability of all modalities, we further include a separate classifier $h^m(\cdot)$ for each modality m in order to obtain predictions from each modality individually. The conditional distribution over classes from the m-th modality is then given by $p(y \mid \boldsymbol{x}^m, f^m) := \delta(h^m(g^m(\boldsymbol{x}^m)))$. The expected predictive distribution for a sample $\boldsymbol{x} \in \mathcal{X}$ can be estimated from the predictive posterior $p(y \mid \boldsymbol{x}; \mathcal{D})$, which can be expressed from a Bayesian perspective as:

$$p(y \mid \boldsymbol{x}, \mathcal{D}) = \int_{f \in \mathcal{F}} p(y \mid \boldsymbol{x}, f)\, p(f \mid \mathcal{D})\, df, \tag{1}$$

where $\mathcal{F}$ denotes the space of predictors. For a given finite set of M diverse models, this integral is typically approximated by a sum over the individual models [16]:

$$p(y \mid \boldsymbol{x}, \mathcal{D}) \approx \sum_{m=1}^{M} p(y \mid \boldsymbol{x}^m, f^m)\, p(f^m \mid \mathcal{D}). \tag{2}$$

By Bayes' rule, $p(f^m \mid \mathcal{D}) = \frac{p(\mathcal{D}|f^m)\, p(f^m)}{p(\mathcal{D})}$. Under a uniform prior over the M models (i.e., not favoring any modality in advance), i.e., $p(f^m) = \frac{1}{M}$, this reduces to $p(f^m \mid \mathcal{D}) = \frac{p(\mathcal{D}|f^m)}{\sum_{j=1}^{M} p(\mathcal{D}|f^j)}$. Substituting this into Eq. (2) yields a likelihood-weighted sum across modalities:

$$p(y \mid \boldsymbol{x}, \mathcal{D}) \approx \sum_{m=1}^{M} w_m\, p(y \mid \boldsymbol{x}^m, f^m), \tag{3}$$

where the weights satisfy $w_m \propto p(\mathcal{D} \mid f^m)$ and $\sum_{m=1}^{M} w_m = 1$. In particular, each w_m can be expressed in terms of the cross-entropy loss (equivalently, the negative log-likelihood) that model f^m exhibits on the dataset $\mathcal{D}$, i.e.,

$$w_m = \frac{\exp\big(-\mathrm{CE}(f^m, \mathcal{D})\big)}{\sum_{j=1}^{M} \exp\big(-\mathrm{CE}(f^j, \mathcal{D})\big)}, \tag{4}$$

where $\mathrm{CE}(f^m, \mathcal{D})$ denotes the average cross-entropy loss of model f^m on $\mathcal{D}$.

Our goal is to capture the full predictive capability and uncertainty expressed in Eq. (3) to be reflected in the final classifier h, such that it inherits both the diverse uncertainty knowledge present at the logit level of each unimodal model and the strong discriminative power from the feature level. For convenience, we denote $p(y \mid \boldsymbol{x}, \mathcal{D})$ as $\boldsymbol{p}^{\mathrm{T}}$ and $p(y \mid \boldsymbol{x}^m, f^m)$ as $\boldsymbol{p}^m$. Accordingly, h is trained by minimizing the cross-entropy with soft targets given by $\boldsymbol{p}^{\mathrm{T}}$. From the perspective of knowledge distillation (KD), the output $\boldsymbol{p}^{\mathrm{S}}$ of h can be conceptually regarded as the student distribution while, $\boldsymbol{p}^{\mathrm{T}}$ serves as the teacher distribution. Consequently, this corresponds to minimizing the KL divergence between teacher and student distributions[1]:

$$\mathcal{L}_{\mathrm{KL}} = \mathrm{KL}(\boldsymbol{p}^{\mathrm{T}} \,\|\, \boldsymbol{p}^{\mathrm{S}}). \tag{5}$$

Notably, this loss formulation given by Eq. (5) and Eq. (3), accounts for the effect of underperforming unimodal models when training h by ranking the teachers rather than treating them uniformly. To practically implement w_m, the weights are estimated on each mini-batch $\mathcal{B}$, as a stochastic approximation to $\mathcal{D}$.

We further decompose the KL divergence, following [41], into target and non-target class components, to gain deeper insights and better leverage the dark knowledge contained in the multimodal teacher logits for OOD detection. We therefore reformulate Eq. (5) as a weighted sum of these two components (the derivation is deferred to Sec. B):

$$\mathcal{L}_{\mathrm{KL}} = \mathrm{KL}\Big(\boldsymbol{b}^{\mathrm{T}} \,\|\, \boldsymbol{b}^{\mathrm{S}}\Big) + \big(1 - p_t^{\mathrm{T}}\big)\, \mathrm{KL}\Big(\hat{\boldsymbol{p}}^{\mathrm{T}} \,\|\, \hat{\boldsymbol{p}}^{\mathrm{S}}\Big), \tag{6}$$

where the notation $\boldsymbol{b} = [\, p_t,\, p_{\backslash t}\,] \in \mathbb{R}^2$ denotes the binary probabilities of the target class (p_t) and all non-target classes ($p_{\backslash t} = 1 - p_t$). The notation $\hat{\boldsymbol{p}} = [\, \hat{p}_1, \ldots, \hat{p}_{t-1}, \hat{p}_{t+1}, \ldots, \hat{p}_C\,] \in \mathbb{R}^{(C-1)}$ is used to independently model probabilities among non-target classes, i.e., the probabilities normalized excluding the t-th class.

Specifically, in Eq. (6), the first term transfers knowledge via binary logit distillation for the target class, referred to as target class KD (TCKD). The second term, referred to as non-target class KD (NCKD), considers the knowledge among the non-target logits. However, these two KD terms are coupled with $1 - p_t^{\mathrm{T}}$ in Eq. (6). To better adjust the importance of each KD term, the hyperparameters α and β are introduced in Eq. (7), following [41]. We further provide

[1] We omit the temperature τ in [12] without loss of generality, to allow the student to better imitate the predictive posterior and to keep it in its natural form ($\tau = 1$), thereby preserving its diversity [34, 40].

insights into each of these KD losses in the next section. Finally, we adopt the standard cross-entropy loss $\mathcal{L}_{\text{CE}}$ as the task loss, and compute the final objective for training the classifier h as:

$$\mathcal{L} = \mathcal{L}_{\text{CE}} + \alpha \underbrace{\text{KL}\left(\boldsymbol{b}^{\text{T}} \,\|\, \boldsymbol{b}^{\text{S}}\right)}_{\text{TCKD}} + \beta \underbrace{\text{KL}\left(\hat{\boldsymbol{p}}^{\text{T}} \,\|\, \hat{\boldsymbol{p}}^{\text{S}}\right)}_{\text{NCKD}}. \tag{7}$$

In particular, the teachers' guidance itself is integrated into the joint classification head that operates on the combined embedding space of modalities, rather than relying on a separate student model that would otherwise need to learn low-level perception of modalities. Overall, it leverages self-generated, uncertainty-aware targets—extracted from the logit space of multimodal models—to effectively learn holistic uncertainty knowledge across modalities. Therefore, the proposed framework harnesses the intrinsic OOD detection capability of a given set of multimodal models by exploiting both feature-level and logit-level information across the modalities.

3.4 Further Insights into Uncertainty-aware Dark Knowledge

To gain deeper insights into how the uncertainty-aware dark knowledge contained in multimodal teacher logits contributes to OOD detection, we reformulate Eq. (5) as the combination of TCKD and NCKD. In particular, Zhao *et al*. [41] study their effect in the task of image classification. In contrast, below we investigate the role of these two forms of dark knowledge for multimodal OOD detection.

Multimodal near-OOD detection mostly benefits from TCKD. TCKD transfers knowledge about the relative "difficulty" of training samples, i.e., the knowledge describes how difficult it is to recognize each sample [41]. In our task, TCKD can be interpreted as binary uncertainty-aware supervision of the target class, which guides the student to identify difficult ID samples that lie near the decision boundary (i.e., with high uncertainty in the target class). This is particularly relevant to near-OOD detection, where the fundamental challenge is that near-OOD samples also lie close to ID decision boundaries (e.g, given 'cat' and 'dog' as ID classes, images of 'fox' will be near-OOD, see Fig. 3) and often exhibit ambiguity with those ID samples. Therefore, we argue that the near-OOD task mostly benefits from TCKD rather than NCKD (which focuses on the uncertainty knowledge among non-target classes) in the proposed self multimodal

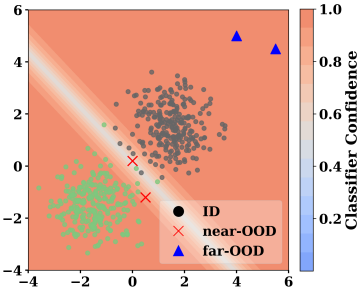

Fig. 3: Classifier-based confidence for OOD detection. It assigns low scores in the small in-between area but may still suffer from overconfidence issues.

Table 1: The effect of TCKD and NCKD for near-OOD detection. Values in brackets represent the performance improvement over the baseline.

TCKD	NCKD	HMDB51 25/26			UCF101 50/51		
		FPR95	AUROC	ID ACC	FPR95	AUROC	ID ACC
–	–	38.78	88.83	89.76	10.10	98.06	99.71
✓	✓	34.81 (-3.97)	90.07 (+1.24)	90.37 (+0.61)	5.98 (-4.12)	98.57 (+0.51)	99.79 (+0.08)
–	✓	35.82 (-2.96)	89.69 (+0.86)	89.89 (+0.13)	6.76 (-3.34)	98.49 (+0.43)	99.63 (-0.08)
✓	–	34.99 (-3.79)	90.12 (+1.29)	90.50 (+0.74)	5.92 (-4.18)	98.57 (+0.51)	99.81 (+0.10)

OOD distillation framework. This claim is further validated by the empirical results presented in Tab. 1, where we individually study the effects of TCKD and NCKD on two near-OOD benchmarks. The results imply that applying TCKD alone is comparable to, and even better than, classical KD (which combines both TCKD and NCKD, with results shown in the second row) for the near-OOD task.

Multimodal Far-OOD detection mostly benefits from NCKD. In contrast to near-OOD samples that reside close to decision boundaries, far-OOD samples lie far from the ID clusters or centers and therefore are less influenced by TCKD. However, a common challenge in the far-OOD case is that classifiers can still be overconfident in these sparse far-OOD regions (Fig. 3)—while they assign low confidence primarily near decision boundaries, they often misclassify far-OOD samples as belonging to one of the ID classes [24]. This overconfidence can be attributed to the fact that the model is trained with one-hot labels, which compel it to treat even atypical ID samples as strong representatives of their target classes, under the provided high-confidence supervision [38]. Therefore, rather than relying on the binary uncertainty knowledge provided by TCKD, which is still related to the target class, the student benefits from NCKD on these ID samples, which emphasizes uncertainty knowledge among non-target classes. To further validate this claim, we evaluate the individual effects of TCKD and NCKD on four far-OOD datasets in Tab. 2. The results show that applying NCKD alone, or assigning more weight to NCKD, can yield better detection performance (with gains of up to 15% on EPIC-Kitchens) on the multimodal far-OOD task.

Our method improves OOD detection, which cannot be achieved by the joint probability distribution or by naively combining it with predictions from individual modalities. Our method significantly improves both near-OOD and far-OOD detection by reducing overconfidence and false positive rates through learning with self uncertainty-aware dark knowledge distilled from the teachers. Notably, this performance gain (see Sec. 4.3 for empirical validation) cannot be achieved by relying on inference from the joint probability distribution across all modalities, $\boldsymbol{p}^{\mathrm{S}}$ (trained solely with the task loss), nor by averaging the predicted probabilities of the individual modalities and the joint classifier: $\bar{\boldsymbol{p}}_{\mathrm{all}} = \frac{1}{M+1}\left(\sum_{m=1}^{M} \boldsymbol{p}^{m} + \boldsymbol{p}^{\mathrm{S}}\right)$.

Table 2: The effect of TCKD and NCKD for far-OOD detection, on HMDB51 [15] ID dataset. Values in brackets represent the performance improvement over the baseline.

TCKD	NCKD	UCF101		HAC		EPIC-Kitchens		Kinetics600	
		FPR95	AUROC	FPR95	AUROC	FPR95	AUROC	FPR95	AUROC
–	–	31.77	90.55	23.49	94.20	22.12	94.87	26.91	93.53
✓	✓	29.49 (-2.28)	90.80 (+0.25)	21.19 (-2.30)	94.77 (+0.57)	18.91 (-3.21)	95.13 (+0.26)	22.49 (-4.42)	94.70 (+1.17)
✓	–	29.83 (-1.94)	90.70 (+0.15)	21.48 (-2.01)	94.71 (+0.51)	19.59 (-2.53)	94.96 (+0.09)	22.81 (-4.10)	94.64 (+1.11)
–	✓	28.60 (-3.17)	91.46 (+0.91)	20.75 (-2.74)	94.94 (+0.74)	16.01 (-6.11)	96.03 (+1.16)	22.17 (-4.74)	94.83 (+1.30)

4 Evaluation

In this section, we first describe our experimental setup, then present the main results on diverse multimodal OOD detection benchmarks, followed by ablation studies.

4.1 Experimental Settings

Datasets and Tasks. Following [4, 18, 19], we evaluate the proposed method on a diverse set of multimodal OOD benchmarks, comprising five action recognition datasets (HMDB51 [15], UCF101 [28], Kinetics-600 [13], HAC [3], and EPIC-Kitchens [1]) and two tasks, namely multimodal near-OOD detection and multimodal far-OOD detection. HMDB51 and UCF101 provide video and optical flow modalities, whereas the others additionally include audio modality. For the near-OOD detection task, we evaluate on four datasets: EPIC-Kitchens 4/4, a subset of EPIC-Kitchens divided into four classes for training (ID) and four classes for testing (OOD); HMDB51 25/26, UCF101 50/51, and Kinetics-600 129/100, which are similarly derived from HMDB51, UCF101, and Kinetics-600, respectively. For the far-OOD detection task, either HMDB51 or Kinetics-600 is used as the ID dataset, with the remaining datasets serving as OOD.

Evaluation Metrics. We use widely adopted metrics for OOD detection [4,18], including the area under the receiver operating characteristic curve (AUROC), the false positive rate at 95% true positive rate (FPR95), and the ID classification accuracy (ID ACC).

Baselines. As the vanilla baseline model (Base), we train all classifiers, including each unimodal model and a combined classifier (with the same architecture as ours), using only the task loss, i.e., cross-entropy loss. Following [4, 18, 19], we adopt the SlowFast network [6], initialized with pre-trained weights from Kinetics-600. As an easy plug-and-play approach, we further evaluate our method on recent SOTA models as backbones that incorporate retraining approaches, including the Agree-to-Disagree algorithm (A2D) [4], the combination of A2D and NP-Mix (AN) [4], DPU [18], and Feature Mixing (FM) [19]. For a fair comparison, we use the same classifier architecture as theirs, but retrain them using our framework while keeping the backbone models fixed.

Table 3: Multimodal near-OOD detection results using video and optical flow. The best results are bolded. Results averaged over six random runs.

Method	HMDB51 25/26			UCF101 50/51			EPIC-Kitchens 4/4			Kinetics600 129/100		
	FPR95	AUROC	ID ACC	FPR95	AUROC	ID ACC	FPR95	AUROC	ID ACC	FPR95	AUROC	ID ACC
Base	38.78	88.83	89.76	10.10	98.06	99.71	75.00	66.70	71.27	64.61	76.59	80.31
+Ours	**34.99**	**90.12**	**90.50**	**5.92**	**98.57**	**99.81**	**73.84**	**68.36**	**72.65**	**61.37**	**78.03**	**81.60**
A2D	38.34	88.22	90.63	7.09	98.19	99.61	**66.23**	**71.04**	71.46	63.04	76.47	79.52
+Ours	**37.86**	**88.78**	**90.76**	**5.32**	**98.51**	**99.69**	69.40	70.24	**72.46**	**61.64**	**77.80**	**81.45**
AN	**33.77**	88.80	90.20	7.96	98.24	99.71	67.16	71.53	71.64	62.91	76.93	80.54
+Ours	34.25	**90.07**	**90.54**	**5.51**	**98.56**	**99.77**	**65.19**	**73.15**	**72.54**	**61.80**	**77.93**	**81.89**
DPU	34.42	89.15	92.16	**7.57**	98.17	**99.81**	**63.81**	**71.46**	72.39	61.59	77.50	81.07
+Ours	**33.25**	**89.65**	**92.77**	7.84	**98.31**	99.77	64.70	71.11	**73.02**	**59.46**	**78.33**	**81.96**
FM	45.10	87.29	89.11	8.06	97.92	99.61	**71.83**	68.49	72.76	64.10	76.16	80.11
+Ours	**37.73**	**88.99**	**90.94**	**6.06**	**98.49**	**99.71**	72.16	**68.93**	**73.06**	**62.75**	**78.13**	**81.80**

Configuration. As mentioned, our method requires no retraining of the original unimodal models and involves only retraining a combined classifier. We use the same model architectures as in the baselines above and retrain their final classifiers using the proposed framework for 10 epochs. Training is performed with the Adam optimizer [14], an initial learning rate of 0.0001, cosine annealing learning-rate scheduler and a batch size of 16 on an NVIDIA-A100 GPU. Following [41] we set the loss terms of KD and CE to 1.0 each. The hyperparameters are chosen as $\alpha = 0.1$, $\beta = 0.9$ for the far-OOD task, and $\alpha = 0.8$, $\beta = 0.2$ for the near-OOD task.

4.2 Main Results

Results for the near-OOD detection task with all five datasets are presented in Tab. 3. Our method consistently achieves the best values across key metrics, including FPR95, AUROC, and ID accuracy on all baselines. Notably, simply incorporating our framework into a vanilla baseline (Base+Ours) yields substantial improvements in OOD performance, achieving comparable or even superior results to SOTA multimodal OOD detection methods (*e.g.* A2D, AN, DPU, FM in Tab. 3) that rely on complex retraining approaches.

Our method is model-agnostic for diverse multimodal OOD detection methods. As seen, across different models using diverse training strategies, including A2D, AN, DPU, and FM, our method consistently improves performance on nearly every evaluation metric despite their diversity. For example, on the UCF101 dataset with AN, our method reduces FPR95 by up to 31%, while on HMDB51 with FM it achieves a reduction of 17%.

Our method is effective in multiple OOD tasks. Results for the far-OOD task with HMDB as the ID dataset are presented in Tab. 4, and with Kinetics-600 as the ID dataset in Tab. 7 in the supplementary material. Similar to the near-OOD task, our method yields considerable improvements (with gains of up to 30%) in OOD detection performance, particularly in terms of

Table 4: Multimodal far-OOD detection results using video and optical flow, with HMDB51 as the ID dataset. The best results are bolded. Results averaged over six random runs.

Method	Kinetics-600		UCF101		EPIC-Kitchens		HAC		Average		ID ACC
	FPR95	AUROC	FPR95	AUROC	FPR95	AUROC	FPR95	AUROC	FPR95	AUROC	
Base	26.91	93.53	31.47	90.10	22.12	94.87	23.49	94.20	26.00	93.18	87.46
+Ours	**22.17**	**94.83**	**28.60**	**91.46**	**16.01**	**96.03**	**20.75**	**94.94**	**21.88**	**94.31**	**88.12**
A2D	20.18	95.12	33.87	90.29	12.43	96.53	**15.85**	95.82	20.58	94.44	87.34
+Ours	**17.40**	**95.73**	**27.55**	**91.03**	**8.73**	**97.33**	16.47	**96.07**	**17.54**	**95.04**	**87.69**
AN	24.29	93.99	36.94	89.71	**7.18**	**97.60**	23.15	94.45	22.89	93.94	86.66
+Ours	**19.98**	**94.96**	**29.35**	**91.26**	10.40	96.05	**19.38**	**94.75**	**19.78**	**94.49**	**86.96**
DPU	20.75	**95.35**	28.39	**92.41**	**4.33**	**98.46**	20.64	**95.40**	18.53	**95.41**	87.34
+Ours	**19.11**	95.32	**25.25**	92.26	6.50	97.66	**18.15**	95.36	**17.25**	95.16	**87.94**
FM	20.30	**94.85**	34.89	89.97	**9.01**	**96.42**	19.27	**95.25**	20.87	**94.12**	**86.32**
+Ours	**16.88**	94.79	**27.99**	**90.64**	12.03	93.32	**16.88**	94.94	**18.95**	93.92	86.26

FPR95 and AUROC, while maintaining comparable or even better ID accuracy across all baseline algorithms. This demonstrates the effectiveness of leveraging uncertainty-aware dark knowledge for multimodal OOD detection.

Performs robustly across diverse datasets. Across five diverse datasets covering a wide range of video styles, including digitized movies, cartoon figures, everyday YouTube videos and kitchen environments, our method consistently reduces FPR95 and improves both AUROC and ID accuracy.

Our method is adaptable and effective across various combinations of modalities, not limited to video and optical flow. Results in Tab. 8 in the supplementary material and Tab. 4 show that performance is consistently improved regardless of the modality combinations used.

4.3 Ablation Studies

We conduct ablation studies with four main objectives: (1) to evaluate the compatibility of our method with various post-hoc OOD scores; (2) to analyze the contribution of each individual component of our framework; (3) to analyze the sensitivity to its key hyperparameters; and (4) to investigate the robustness of our framework under corrupted modality conditions.

Compatibility with various post-hoc OOD scores. We evaluate our method across seven post-hoc OOD scoring methods with different strategies: probability space (MSP), logit space (MaxLogit, Energy, GEN), penultimate activation manipulations (ReAct, ASH), and combinations of logit- and feature-space techniques (VIM), beyond the default model scores used in our main experiments. Results in Fig. 4 show that, despite this diversity, our method consistently improves the performance of all OOD scoring methods (with detailed results provided in Tab. 9 in the supplementary material).

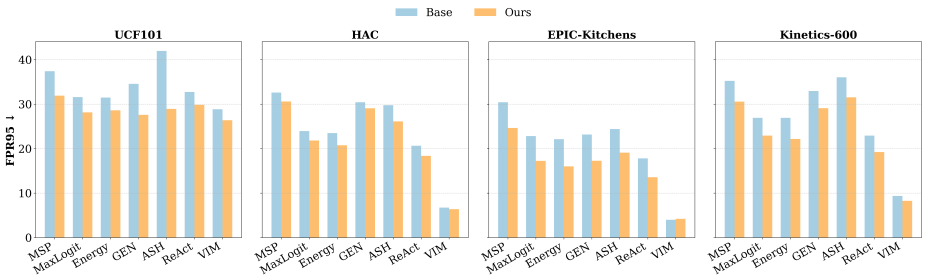

Fig. 4: Compatibility of our method with various post-hoc OOD scoring methods.

Importance of the components. Our self-multimodal OOD distillation framework improves OOD detection by leveraging uncertainty-aware dark knowledge from unimodal experts through logit distillation while explicitly accounting for modality imbalance. To validate the contribution of each component, we conduct a comprehensive ablation study across four OOD datasets (Tab. 5). First, we compare against unimodal models trained using only the task loss. As also shown in Fig. 1, the optical flow modality performs substantially worse than the video modality, highlighting the modality imbalance commonly encountered in multimodal settings. Next, we include a naive ensemble baseline that averages the predictions of the unimodal experts without feature fusion (*Ensemble w/o FF*) to quantify the improvement obtained solely from ensembling unimodal predictions. Then, we compare the performance of our method against a joint classifier h trained on fused multimodal features using only the task loss (i.e., equivalent to setting both α and β to 0 in Eq. (7)), denoted as *Vanilla*. We also compare against an ensemble baseline that averages the predicted probabilities of the individual modalities and the joint classifier ($\bar{\boldsymbol{p}}_{\text{all}}$, cf. Sec. 3.4), denoted as *Ensemble w/ FF*. Together, these baselines quantify the improvements achievable through multimodal feature fusion and prediction aggregation alone, without logit distillation.

Our method consistently surpasses all these approaches, highlighting the importance of uncertainty-aware dark knowledge for improving OOD detection performance beyond what can be achieved through naive ensemble averaging or feature fusion alone. We further evaluate a variant without modality-specific weights (in Eq. (3)), reported as *Uniform KD*, demonstrating that the primary performance gains arise from leveraging uncertainty-aware dark knowledge, while modality-specific weighting provides additional improvements by accounting for underperforming modalities. We also compare another knowledge transfer framework, in which $\boldsymbol{p}^T$ is replaced by the predictive distribution of the best-performing modality while still employing multimodal feature fusion, denoted as *Best KD*. Our method still outperforms this variant, suggesting that complementary uncertainty information from multiple modalities provides more informative supervision than relying solely on the strongest individual modality.

Table 5: Ablation study of components for far-OOD detection with HMDB51 as the ID dataset. The best results are bolded. Results averaged over six random runs.

Method	UCF101		HAC		EPIC-Kitchens		Kinetics-600		Average		ID ACC
	FPR95	AUROC	FPR95	AUROC	FPR95	AUROC	FPR95	AUROC	FPR95	AUROC	
Flow only	50.20	78.52	69.92	74.96	38.86	87.04	64.68	77.66	55.92	79.55	75.14
Video only	44.36	87.34	36.94	89.35	34.21	91.50	39.57	89.05	38.77	89.31	85.63
Ensemble w/o FF	32.84	90.10	28.85	93.06	20.64	95.18	30.33	92.84	28.16	92.80	87.91
Vanilla	31.47	90.10	23.49	94.20	22.12	94.87	26.91	93.53	26.00	93.18	87.46
Ensemble w/ FF	32.50	90.21	25.88	93.70	21.44	95.14	28.05	93.25	26.97	93.08	88.03
Uniform KD	29.44	91.26	22.37	94.62	16.62	95.87	23.40	94.58	22.96	94.08	**88.21**
Best KD	30.13	91.29	20.84	94.96	16.90	95.68	22.19	94.75	22.52	94.17	87.89
Ours	**28.60**	**91.46**	**20.75**	**94.94**	**16.01**	**96.03**	**22.17**	**94.83**	**21.88**	**94.31**	88.12

Hyperparameter Analysis In Fig. 5, we investigate the impact of the hyperparameters α and β in our method across three far-OOD datasets. Following the basic settings in [25, 41], we set the loss weights of the KD and CE terms to 1.0. As discussed in Sec. 3.4, assigning a higher value to β, which emphasizes the NCKD component, leads to improved detection performance in multimodal far-OOD detection. Furthermore, the detection performance consistently remains above the baseline across all tested combinations of α and β, demonstrating the stability and robustness of our framework under varying hyperparameter settings.

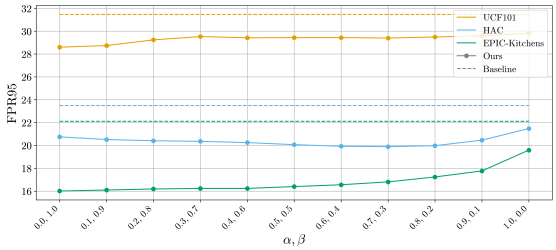

Fig. 5: OOD detection performance with varying α and β for far-OOD detection using video and optical flow, with HMDB51 as the ID dataset.

For real-world deployment where the type of OOD shift is unknown, we recommend a balanced and task-agnostic configuration such as $\alpha = 0.5$ and $\beta = 0.5$, which performs robustly across all benchmarks and OOD scenarios. Alternatively, practitioners may adopt a proxy-OOD tuning strategy, which is conventional in prior OOD detection methods [4, 10, 39]. In practice, many real-world applications naturally prioritize one type of OOD. Some systems primarily focus on near-OOD detection (e.g., fine-grained classification, medical imaging, product or defect inspection) [39], where far-OOD samples are trivial and typically filtered earlier in the pipeline. Conversely, some applications prioritize far-OOD (e.g., OCR scanners detecting non-text inputs need to be rejected early

in pipeline). Our framework allows practitioners to adjust α and β according to such priorities, but importantly, no prior knowledge of the exact OOD type is required to benefit from the method, as it consistently improves both near- and far-OOD detection under a wide range of settings.

Weak-Modality Analysis. To investigate the robustness of our framework under corrupted modality conditions and validate its ability to effectively handle weaker modalities, we conduct a weak-modality experiment by corrupting the Flow model through random permutation of its class-specific weights while keeping the Video modality unchanged. Since teacher contributions are weighted according to the modality-specific classification uncertainty in Eq. (4), which naturally penalizes unreliable or miscalibrated predictions, the corrupted modality is automatically assigned a lower influence during distillation. The results in Tab. 6 show that our method degrades more gracefully under weak-modality conditions, whereas both ensemble-based baselines and discrepancy-based methods deteriorate substantially (see supplementary Sec. C for further analysis).

Table 6: Comparison under corrupted *Flow* modality settings on near-OOD detection benchmarks. Results are averaged over six random runs. The best results are bolded.

Method	HMDB51 25/26				UCF101 50/51			
	FPR95	AUROC	ID ACC	RMSCE	FPR95	AUROC	ID ACC	RMSCE
Ensemble	55.12	85.98	88.45	7.11	33.50	94.90	99.42	9.08
Ensemble w/o FF	55.99	85.55	87.80	12.74	57.38	89.53	79.71	5.34
A2D	51.63	85.96	88.89	11.51	21.26	96.56	99.42	**4.57**
Ours	**50.24**	**87.12**	**89.41**	**7.02**	**7.38**	**98.37**	**99.86**	4.59
Energy weights	51.46	86.10	89.37	7.60	19.24	96.78	99.67	15.59

5 Conclusion

In this work, we explore the potential of dark knowledge within multimodal models to strengthen OOD detection. Building on this, we propose a self multimodal distillation framework that leverages both logit-space uncertainty knowledge and feature-space knowledge from a given set of multimodal models to harness their intrinsic OOD detection capability, while effectively accounting for underperforming modalities. Extensive experiments demonstrate that our method consistently improves multimodal OOD detection, further enhances existing approaches, and reveals the full synergy among modalities.

Acknowledgements

This research used the LIEF HPC-GPGPU Facility at the University of Melbourne, established with LIEF Grant LE170100200. Sarah Erfani is in part supported by Australian Research Council (ARC) Discovery Early Career Researcher Award (DECRA) DE220100680.

References

1. 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: ECCV. pp. 720–736 (2018)
2. Djurisic, A., Bozanic, N., Ashok, A., Liu, R.: Extremely simple activation shaping for out-of-distribution detection. arXiv preprint arXiv:2209.09858 (2022)
3. Dong, H., Nejjar, I., Sun, H., Chatzi, E., Fink, O.: Simmmdg: A simple and effective framework for multi-modal domain generalization. NeurIPS **36**, 78674–78695 (2023)
4. Dong, H., Zhao, Y., Chatzi, E., Fink, O.: Multiood: Scaling out-of-distribution detection for multiple modalities. NeurIPS **37**, 129250–129278 (2024)
5. Du, X., Wang, Z., Cai, M., Li, Y.: Vos: Learning what you don't know by virtual outlier synthesis. arXiv preprint arXiv:2202.01197 (2022)
6. Feichtenhofer, C., Fan, H., Malik, J., He, K.: Slowfast networks for video recognition. In: ICCV. pp. 6202–6211 (2019)
7. Fink, O., Wang, Q., Svensen, M., Dersin, P., Lee, W.J., Ducoffe, M.: Potential, challenges and future directions for deep learning in prognostics and health management applications. Engineering Applications of Artificial Intelligence **92**, 103678 (2020)
8. Hendrycks, D., Basart, S., Mazeika, M., Zou, A., Kwon, J., Mostajabi, M., Steinhardt, J., Song, D.: Scaling out-of-distribution detection for real-world settings. In: ICML. pp. 8759–8773 (2022)
9. Hendrycks, D., Gimpel, K.: A baseline for detecting misclassified and out-of-distribution examples in neural networks. In: ICLR (2017)
10. Hendrycks, D., Mazeika, M., Dietterich, T.G.: Deep anomaly detection with outlier exposure. In: ICLR (2019)
11. Hewa Dehigahawattage, N.U., Nandakishor, K., Palaniswami, M.: Learning to reason: Temporal saliency distillation for interpretable knowledge transfer. In: Proceedings of the 28th European Conference on Artificial Intelligence (ECAI 2025). Frontiers in Artificial Intelligence and Applications (FAIA), IOS Press, Bologna, Italy (2025). https://doi.org/10.3233/FAIA251144
12. Hinton, G., Vinyals, O., Dean, J.: Distilling the knowledge in a neural network. arXiv preprint arXiv:1503.02531 (2015)
13. Kay, W., Carreira, J., Simonyan, K., Zhang, B., Hillier, C., Vijayanarasimhan, S., Viola, F., Green, T., Back, T., Natsev, P., et al.: The kinetics human action video dataset. arXiv preprint arXiv:1705.06950 (2017)
14. Kingma, D.P.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
15. Kuehne, H., Jhuang, H., Garrote, E., Poggio, T., Serre, T.: Hmdb: a large video database for human motion recognition. In: ICCV. pp. 2556–2563. IEEE (2011)
16. Lakshminarayanan, B., Pritzel, A., Blundell, C.: Simple and scalable predictive uncertainty estimation using deep ensembles. NeurIPS **30** (2017)
17. Lee, K., Lee, K., Lee, H., Shin, J.: A simple unified framework for detecting out-of-distribution samples and adversarial attacks. In: NeurIPS (2018)
18. Li, S., Gong, H., Dong, H., Yang, T., Tu, Z., Zhao, Y.: Dpu: Dynamic prototype updating for multimodal out-of-distribution detection. In: CVPR. pp. 10193–10202 (2025)
19. Liu, M., Dong, H., Kelly, J., Fink, O., Trapp, M.: Extremely simple multimodal outlier synthesis for out-of-distribution detection and segmentation. arXiv preprint arXiv:2505.16985 (2025)

20. Liu, W., Wang, X., Owens, J., Li, Y.: Energy-based out-of-distribution detection. In: NeurIPS. pp. 21464–21475 (2020)
21. Liu, X., Lochman, Y., Zach, C.: Gen: Pushing the limits of softmax-based out-of-distribution detection. In: CVPR. pp. 23946–23955 (2023)
22. Liu, Y., Ding, C., Tian, Y., Pang, G., Belagiannis, V., Reid, I., Carneiro, G.: Residual pattern learning for pixel-wise out-of-distribution detection in semantic segmentation. In: ICCV. pp. 1151–1161 (2023)
23. Ming, Y., Cai, Z., Gu, J., Sun, Y., Li, W., Li, Y.: Delving into out-of-distribution detection with vision-language representations. NeurIPS **35**, 35087–35102 (2022)
24. Park, J., Jung, Y.G., Teoh, A.B.J.: Nearest neighbor guidance for out-of-distribution detection. In: ICCV. pp. 1686–1695 (2023)
25. Park, W., Kim, D., Lu, Y., Cho, M.: Relational knowledge distillation. In: CVPR. pp. 3967–3976 (2019)
26. Radevski, G., Grujicic, D., Blaschko, M., Moens, M.F., Tuytelaars, T.: Multimodal distillation for egocentric action recognition. In: ICCV. pp. 5213–5224 (2023)
27. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: ICML. pp. 8748–8763. PmLR (2021)
28. 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)
29. Sun, Y., Guo, C., Li, Y.: React: Out-of-distribution detection with rectified activations. NeurIPS **34**, 144–157 (2021)
30. Sun, Y., Ming, Y., Zhu, X., Li, Y.: Out-of-distribution detection with deep nearest neighbors. In: ICML. pp. 20827–20840. PMLR (2022)
31. Udayangani, H.D.N., Desai, N., Palaniswami, M.: Generative diffusion prior distillation for long-context knowledge transfer. In: The Fourteenth International Conference on Learning Representations (2026)
32. Udayangani, N., Dolatabadi, H.M., Erfani, S., Leckie, C.: Exploiting inter-sample information for long-tailed out-of-distribution detection. In: Proceedings of the Winter Conference on Applications of Computer Vision. pp. 8535–8544 (2025)
33. Udayangani, N., Erfani, S., Leckie, C.: Suplid: Geometrical guidance for out-of-distribution detection in semantic segmentation. In: Proceedings of the 34th ACM International Conference on Information and Knowledge Management. pp. 2905–2914 (2025)
34. Wang, D., Xu, J., Lu, S., Wei, X., Wang, L.: Ensemble distillation for out-of-distribution detection. In: 2023 IEEE 29th International Conference on Parallel and Distributed Systems (ICPADS). pp. 248–253. IEEE (2023)
35. Wang, H., Li, Z., Feng, L., Zhang, W.: Vim: Out-of-distribution with virtual-logit matching. In: CVPR. pp. 4921–4930 (2022)
36. Wang, H., Li, Y., Yao, H., Li, X.: Clipn for zero-shot ood detection: Teaching clip to say no. In: CVPR. pp. 1802–1812 (2023)
37. Yang, J., Zhou, K., Li, Y., Liu, Z.: Generalized out-of-distribution detection: A survey. IJCV **132**(12), 5635–5662 (2024)
38. Yang, Y., Xu, H.: Strengthen out-of-distribution detection capability with progressive self-knowledge distillation. In: ICML (2025)
39. Zhang, J., Inkawhich, N., Linderman, R., Chen, Y., Li, H.: Mixture outlier exposure: Towards out-of-distribution detection in fine-grained environments. In: WACV. pp. 5531–5540 (2023)
40. Zhang, L., Song, J., Gao, A., Chen, J., Bao, C., Ma, K.: Be your own teacher: Improve the performance of convolutional neural networks via self distillation. In: ICCV. pp. 3713–3722 (2019)

41. Zhao, B., Cui, Q., Song, R., Qiu, Y., Liang, J.: Decoupled knowledge distillation. In: CVPR. pp. 11953–11962 (2022)