Towards Multimodal Open-Set Domain Generalization and Adaptation through Self-supervision

Hao Dong[1] , Eleni Chatzi[1] , and Olga Fink[2]

[1] ETH Zürich
{hao.dong, chatzi}@ibk.baug.ethz.ch
[2] EPFL
olga.fink@epfl.ch

Abstract. The task of open-set domain generalization (OSDG) involves recognizing novel classes within unseen domains, which becomes more challenging with multiple modalities as input. Existing works have only addressed unimodal OSDG within the meta-learning framework, without considering multimodal scenarios. In this work, we introduce a novel approach to address Multimodal Open-Set Domain Generalization (MM-OSDG) for the first time, utilizing self-supervision. To this end, we introduce two innovative multimodal self-supervised pretext tasks: Masked Cross-modal Translation and Multimodal Jigsaw Puzzles. These tasks facilitate the learning of multimodal representative features, thereby enhancing generalization and open-class detection capabilities. Additionally, we propose a novel entropy weighting mechanism to balance the loss across different modalities. Furthermore, we extend our approach to tackle also the Multimodal Open-Set Domain Adaptation (MM-OSDA) problem, especially in scenarios where unlabeled data from the target domain is available. Extensive experiments conducted under MM-OSDG, MM-OSDA, and Multimodal Closed-Set DG settings on the EPIC-Kitchens and HAC datasets demonstrate the efficacy and versatility of the proposed approach. Our source code is publicly available[1].

Keywords: Multimodal Learning · Open-Set Domain Generalization · Open-Set Domain Adaptation

1 Introduction

Domain generalization (DG) and adaptation (DA) significantly improve the robustness and adaptability of machine learning models, enabling them to perform effectively across diverse and previously unseen environments [51]. This enhancement ensures that the models are more readily transferable to real-world applications, such as autonomous driving [11,13] and action recognition [8]. To address the challenges posed by distribution shifts, a wide range of DG and DA

---

[1] https://github.com/donghao51/MOOSA

Task	Only one modality?	Need $\mathcal{C}_s = \mathcal{C}_t$?	Access to $\mathcal{D}_t$?
Domain Adaptation [15]	✓	✓	✓
Domain Generalization [6]	✓	✓	✗
Open-Set Domain Adaptation [32]	✓	✗	✓
Open-Set Domain Generalization [47]	✓	✗	✗
Multimodal Domain Adaptation [37]	✗	✓	✓
Multimodal Domain Generalization [12]	✗	✓	✗
Multimodal Open-Set Domain Adaptation (proposed)	✗	✗	✓
Multimodal Open-Set Domain Generalization (proposed)	✗	✗	✗

Table 1: Illustration of the difference between the proposed Multimodal Open-Set DG and DA and other related tasks. $\mathcal{D}_t$ denote the target domains. $\mathcal{C}_s$ and $\mathcal{C}_t$ denote the label space of source and target domains, respectively.

algorithms have been introduced. These algorithms include domain-invariant feature learning [36], feature disentanglement [42], data augmentation [54], and meta-learning [28]. However, the majority of these algorithms are designed for unimodal data, such as images [27] or time series data [34]. With the increasing need to process multimodal data in real-world applications [5,8], it has become imperative to extend these methods to support multimodal DG across a variety of modalities, including audio-video [24,55] and vision-language [23,44]. In response to this challenge, several approaches, such as RNA-Net [43] and SimM-MDG [12], have been proposed to address the complexities of multimodal DG.

An inherent assumption in both DG and multimodal DG is the alignment of label spaces between source and target domains. However, real-world applications, such as autonomous driving [3], often feature target domains with novel categories not present in the source label space. As a result, the learned model may struggle with samples from these novel categories, significantly degrading the robustness of existing DG and multimodal DG methods. This setup, where the target domain contains unknown or open classes not seen in source domains, is referred to as open-set DG. Several unimodal open-set DG approaches, including DAML [47] and MEDIC [52], have been developed within the meta-learning framework to tackle this issue. CrossMatch [57] utilizes an adversarial data augmentation strategy to generate auxiliary samples beyond the source label space. Yet, none specifically address the challenge of Multimodal Open-Set DG (MM-OSDG), which is the primary focus of this paper. The goal of MM-OSDG is to train a model using data from several source domains across two or more modalities, enabling it to effectively generalize to previously unseen target domains with the same modalities and including samples from unknown classes. The key challenge of MM-OSDG is to efficiently leverage complementary information from diverse modalities to improve generalization and open-class detection performance, areas where current unimodal open-set DG approaches fall short. We summarize the distinctions between our proposed MM-OSDG problem and various related problems in Tab. 1.

To address the challenge of domain shift in unimodal setups, several studies have explored the integration of self-supervision techniques for DG and DA [4,6]. This approach involves addressing a self-supervised pretext task concurrently with the primary supervised problem, which leads to the learning of resilient cross-domain features conducive to robust generalization. Additionally, recent research has demonstrated the utility of leveraging outputs from self-supervised models for anomaly detection, facilitating the discrimination between normal

and anomalous samples [2, 18] – a scenario similar to open-class detection. In particular, Rotation-based Open Set (ROS) [4] explores rotation recognition as a self-supervised task for both unknown class detection and domain adaptation. Meanwhile, results from MOOD [30] illustrate that a reconstruction-based pretext task forces the network to learn the real data distribution of the in-distribution (ID) samples, thereby enlarging the divergence between the out-of-distribution (OOD) and ID samples.

Inspired by the success of the self-supervised pretext tasks in robust feature learning and OOD sample detection in unimodal setups, we propose a novel approach referred to as ***MOOSA*** to address the **M**ultim**o**dal **O**pen-**S**et Domain Generalization and **A**daptation problems using multimodal self-supervised tasks. Specifically, we propose a generative task, termed Masked Cross-modal Translation, and a contrastive task, denoted as Multimodal Jigsaw Puzzles. These tasks are complementary to each other to effectively learn multimodal representative features conducive to both generalization and open-class detection. An entropy weighting mechanism is further introduced to balance the loss across different modalities. Additionally, our methodology is extended to handle the Multimodal Open-Set Domain Adaptation (MM-OSDA) problem in scenarios where unlabeled target data is available. The efficacy of the proposed approach is thoroughly validated through extensive experiments conducted on two multimodal DG benchmark datasets: EPIC-Kitchens [8] and Human-Animal-Cartoon (HAC) [12]. Our contributions can be summarized as follows:

– We delve into the unexplored area of Multimodal Open-Set Domain Generalization, a concept with significant implications for real-world applications. It requires a model trained on diverse source domains to generalize effectively to new, unseen target domains, which share the same modalities but include samples from previously unknown classes.

– To tackle MM-OSDG, we introduce two complementary multimodal self-supervised tasks – Masked Cross-modal Translation and Multimodal Jigsaw Puzzles – along with an entropy weighting mechanism. We also extend our method to the novel Multimodal Open-Set Domain Adaptation setup.

– The efficacy and versatility of our proposed approach are validated through extensive experiments conducted across MM-OSDG, MM-OSDA, and Multimodal Closed-Set DG settings.

2 Related Work

Self-supervised Learning generates supervisory signals based on the input data, alleviating the need for expensive manual labeling and improving performance across almost all types of downstream tasks, including domain generalization [6] and anomaly detection [18]. The majority of self-supervised tasks have so far focused on a single modality. Examples of such self-supervised tasks include predicting the relative position of image patches [10], colorizing grayscale images [26], recognizing the rotation of an image [17], or inpainting removed

patches [41]. Recently, some approaches have been proposed to tackle the specific challenges of multimodal data, such as multimodal alignment [37], relative norm alignment [43], and cross-modal translation [12].

Domain Generalization entails training a model on multiple source domains to achieve robust generalization to previously unseen target domains. This is different from domain adaptation, where unlabeled target domain data is available during training. Prior research [51] has identified three main categories of domain generalization methods: data manipulation, representation learning, and learning strategies. Strategies involving data manipulation aim to enhance generalization performance by augmenting the diversity of training data [48,54,56]. Representation learning methods strive to acquire domain-invariant representations employing techniques such as domain-adversarial neural networks [16,29], explicit feature distribution alignment [49], and instance normalization [39]. Additionally, various approaches leverage learning strategies to improve generalization performance, including meta-learning [28], gradient operation [22], and self-supervised learning [6].

Open-Set DA and DG address scenarios where the target domain, with domain shifts, encompasses classes not present in the source domain [40]. In the context of open-set DA, several approaches have been proposed to tackle this challenge, such as assigning target domain images to source categories while discarding unrelated target domain images [40], and employing adversarial training to separate unknown target samples [46]. Separate To Adapt (STA) [32] introduces a progressive separation mechanism for unknown and known class samples. ROS [4] investigates the utility of rotation recognition for unknown class detection. Adjustment and Alignment (ANNA) [31] addresses semantic-level bias by incorporating front-door adjustment and decoupled causal alignment modules. For open-set DG, Domain-Augmented Meta-Learning (DAML) [47] adopts a meta-learning framework to learn open-domain generalizable representations and augment domains at both feature and label levels. CrossMatch [57] employs an adversarial data augmentation strategy to generate auxiliary samples beyond the source label space. MEDIC [52] considers gradient matching towards inter-domain and inter-class splits simultaneously to find a generalizable boundary balanced for all tasks. However, none of these approaches tackle the challenge posed by multiple modalities.

Multimodal DA and DG tackle the challenges of DA and DG involving multiple modalities. For instance, MM-SADA [37] proposes a self-supervised alignment approach coupled with adversarial alignment for multimodal DA, while Kim *et al.* [25] leverage cross-modal contrastive learning to align cross-modal and cross-domain representations. Zhang *et al.* [55] introduce an audio-adaptive encoder and an audio-infused recognizer to address domain shifts. RNA-Net [43] addresses the multimodal DG problem by introducing a relative norm alignment loss to balance audio and video feature norms. SimMMDG [12] presents a universal framework for multimodal domain generalization, involving the separation of features within each modality into modality-specific and modality-shared components, along with constraints to facilitate meaningful representation learning.
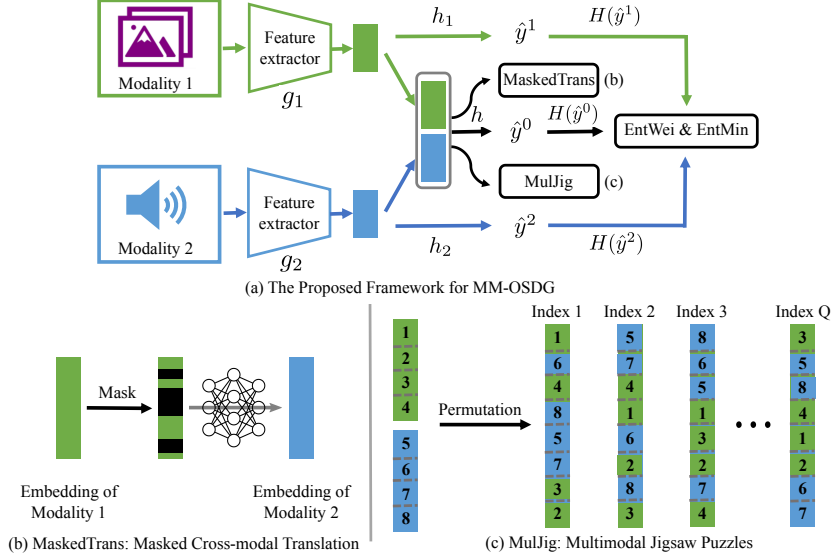
(a) The Proposed Framework for MM-OSDG

(b) MaskedTrans: Masked Cross-modal Translation

(c) MulJig: Multimodal Jigsaw Puzzles

Fig. 1: Our proposed *MOOSA* framework for MM-OSDG. EntWei & EntMin: Entropy Weighting and Minimization.

Different from prior efforts, our work represents the first attempt to address the Multimodal Open-Set DG and DA, scenarios that are both more practical and significantly more challenging.

3 Methodology

In this section, we first introduce the formulation of the Multimodal Open-Set Domain Generalization problem. Subsequently, we present our novel framework designed to address this challenge. This framework incorporates the use of self-supervised pretext tasks, specifically the proposed Masked Cross-modal Translation and Multimodal Jigsaw Puzzles, complemented by an entropy weighting mechanism. Finally, we extend our framework to tackle Multimodal Open-Set Domain Adaptation. A visual representation of our approach is depicted in Fig. 1.

3.1 Multimodal Open-Set Domain Generalization

In the context of MM-OSDG, we work with several source domains, denoted as $S = \{\mathcal{D}_1, \mathcal{D}_2, ..., \mathcal{D}_S\}$, each sharing a common label space $\mathcal{C}$. We also consider an unseen target domain $\mathcal{T} = \{\mathcal{D}_T\}$ that introduces an extended label space $\mathcal{C} \cup \mathcal{U}$, with the stipulation that $\mathcal{C}$ and $\mathcal{U}$ do not overlap $\mathcal{C} \cap \mathcal{U} = \varnothing$. Each source domain, indexed by s and comprising N_s samples, is represented as $\mathcal{D}_s = \{(\mathbf{x}_i^s, y_i^s)\}_{i=1}^{N_s}$ following the distribution P_{XY}^s. Here, $\mathbf{x}_i^s$ denotes the i-th sample from the sample space $\mathcal{X}$ and y_i^s is the corresponding label, which belongs to the shared label space $\mathcal{C}$. Each sample instance $\mathbf{x}_i^s$ is comprised of M modalities,

denoted as $\mathbf{x}_i^s = \{(x_i^s)_k \mid k = 1, \cdots, M\}$. The joint distributions between each pair of domains differ: $P_{XY}^i \neq P_{XY}^j, 1 \leq i \neq j \leq S$. The objective of MM-OSDG is to develop a model $f : \mathcal{X} \rightarrow \mathcal{Y}$ that leverages data from multiple source domains S with a label space $\mathcal{C}$ and M modalities. This model aims to seamlessly adapt to unseen target domains $\mathcal{D}_T$ that include both known classes from $\mathcal{C}$ and unknown classes from $\mathcal{U}$, with the combined label space being $\mathcal{C} \cup \mathcal{U}$. Consequently, the model is expected to accurately classify each target sample into the appropriate class from $\mathcal{C}$, or designate it as "unknown" if it falls within $\mathcal{U}$. Importantly, during training, data from the target domain $\mathcal{D}_T$ is inaccessible, and there is a distribution shift between the source and the target domains, $P_{XY}^T \neq P_{XY}^s$ for all $s \in \{1, \cdots, S\}$.

The predictive function f comprises M feature extractors $g_k(\cdot)$, a classifier $h(\cdot)$, and the Softmax function $\delta(\cdot)$. Each feature extractor $g_k(\cdot)$ is tasked with extracting a feature embedding $\mathbf{E}^k$ for its corresponding modality k, and the classifier $h(\cdot)$ takes the combined embeddings from all modalities as input to produce the final output prediction $\hat{y}^0$, which can be written as:

$$\begin{aligned}
\hat{y}^0 = \delta(f(\mathbf{x})) &= \delta(h([g_1((x)_1), ..., g_k((x)_k), ..., g_M((x)_M)])) \\
&= \delta(h([\mathbf{E}^1, ..., \mathbf{E}^k, ..., \mathbf{E}^M])).
\end{aligned} \tag{1}$$

We further include a separate classifier $h_k(\cdot)$ for each modality k, with the prediction from the k-th modality as $\hat{y}^k = \delta(h_k(g_k((x)_k)))$.

3.2 Motivation: Self-supervised Pretext Tasks

Self-supervised pretext tasks are commonly classified into generative and contrastive methods [33]. Contrastive approaches, such as context spatial relations [17, 35] and instance discrimination [7, 21], focus on encouraging class-invariance between different instances [33], thereby contributing to generalization. However, recent findings [45] indicate that contrastive methods may tend to learn specific patterns of categories, which are beneficial for classification, but less conducive to understanding intrinsic data distributions of in-distribution samples. In contrast, generative methods [9, 20] force the network to learn the real data distribution of in-distribution samples during training, enhancing the divergence between in-distribution and out-of-distribution samples [30], making them particularly valuable for open-class detection. Motivated by these observations, we propose using both generative and contrastive tasks to achieve robust generalization and reliable open-class detection. Accordingly, we propose two novel multimodal self-supervised pretext tasks aligned with these two objectives. Multimodal self-supervised tasks are challenging, requiring the joint consideration of all modality properties to ensure meaningful outcomes.

3.3 Generative Task: Masked Cross-modal Translation

The cross-modal translation module introduced in SimMMDG [12] can be regarded as an auto-encoding generative approach. It aims to ensure the meaningfulness of learned features by exploiting the implicit relationships inherent

between the M modalities within the same data instance. A multi-layer perceptron (MLP) is used to translate the feature embedding $\mathbf{E}$ across modalities. For instance, translating the embedding $\mathbf{E}^i$ of the i-th modality to the j-th modality using $MLP_{\mathbf{E}^i \to \mathbf{E}^j}$ results in the translated embedding $\mathbf{E}^j_t$. The aim is to minimize the ℓ_2 distance between the real embedding of the j-th modality, $\mathbf{E}^j$, and the translated embedding, $\mathbf{E}^j_t$, defining the cross-modal translation loss as:

$$\mathcal{L}_{trans} = \frac{1}{M(M-1)} \sum_{i=1}^{M} \sum_{j \neq i} ||MLP_{\mathbf{E}^i \to \mathbf{E}^j}(\mathbf{E}^i) - \mathbf{E}^j||_2^2. \tag{2}$$

Although cross-modal translation has demonstrated effectiveness in various multimodal DG tasks [12], inspired by the masked image modeling pretext task [1] and its application in OOD detection [30], we propose masked cross-modal translation tailored to better align with the MM-OSDG task. Instead of directly translating $\mathbf{E}^i$ to $\mathbf{E}^j$, we randomly mask a percentage of $\mathbf{E}^i$ and subsequently translate this masked embedding, denoted as $Mask(\mathbf{E}^i)$, to the target embedding, as depicted in Fig. 1 (b). The masking operation, $Mask(\mathbf{E}^i)$, randomly sets a β percentage of elements in $\mathbf{E}^i$ to 0. The masked cross-modal translation loss is then defined based on this operation as follows:

$$\mathcal{L}_{MaskedTrans} = \frac{1}{M(M-1)} \sum_{i=1}^{M} \sum_{j \neq i} ||MLP_{\mathbf{E}^i \to \mathbf{E}^j}(Mask(\mathbf{E}^i)) - \mathbf{E}^j||_2^2. \tag{3}$$

With this masking procedure, the model learns to infer the complete embeddings of the target modality from the remaining embeddings of the source modality. This procedure enables the model to leverage intrinsic data distributions across diverse modalities as priors, rather than solely focusing on discerning different patterns among categories during the classification process. By benefiting from these intrinsic priors, the model achieves a more representative feature representation of the known classes, thereby increasing the divergence between the known and unknown samples. This observation is further supported by the visualization results in Fig. 3.

3.4 Contrastive Task: Multimodal Jigsaw Puzzles

Context-based spatial relation prediction [10, 17] falls within the category of contrastive objectives [33]. Previous studies [6, 38] have explored learning visual representations through solving Jigsaw puzzles, a task that involves recovering an original image from its shuffled parts. In this work, we extend the Jigsaw puzzle-solving task to multimodal scenarios. Given the heterogeneity of different modalities, shuffling them in the input space and then randomly combining them poses a challenge. Therefore, our approach involves shuffling in the embedding space for all modalities. Given the embedding $\mathbf{E}^i$ of the i-th modality, we split it into P parts as $\mathbf{E}^i = [\mathbf{E}^{i,1}, \mathbf{E}^{i,2}, ..., \mathbf{E}^{i,P}]$ with equal length. Similarly, for the j-th modality, we have $\mathbf{E}^j = [\mathbf{E}^{j,1}, \mathbf{E}^{j,2}, ..., \mathbf{E}^{j,P}]$. Combining all $M \cdot P$ split parts

from M modalities, we randomly shuffle them to generate different permutations. For example, with two modalities i and j and by setting P to 2, one possible permutation is $\widetilde{\mathbf{E}}^p = [\mathbf{E}^{j,1}, \mathbf{E}^{i,2}, \mathbf{E}^{j,2}, \mathbf{E}^{i,1}]$. The shuffling is performed across modalities, which makes the task particularly challenging. Out of the $(M \cdot P)!$ possible permutations, we randomly select a set of Q elements and assign an index to each entry as the label. Then, we can define an auxiliary classification task on each sample instance as $\{(\widetilde{\mathbf{E}}^p, p)\}_{p=1}^{Q}$, where $\widetilde{\mathbf{E}}^p$ indicates the recomposed embeddings and $p \in \{1, \ldots, Q\}$ denotes the corresponding permutation index. To achieve this, we aim to minimize the multimodal jigsaw loss $\mathcal{L}_{MulJig}(h_p(\widetilde{\mathbf{E}}^p), p)$, were h_p is the classifier for permutation recognition, and $\mathcal{L}_{MulJig}$ is the standard cross-entropy loss. Fig. 1 (c) provides an illustration of this process.

By solving the multimodal Jigsaw puzzles problem, the model must not only to recover the correct order of the shuffled embeddings for each modality but also discern the source modality of each shuffled part. This process enables the model to learn multimodal representative features conducive to robust generalization.

3.5   Entropy Weighting and Minimization

Distinct modalities may contribute differently to the final predictions under various circumstances. For instance, in recognizing the "running" action within a noisy gym, video data may be more reliable than audio data for the final prediction. Conversely, in scenarios characterized by low environmental light or partial shading of the camera, audio input may prove more dependable. To address the need for balancing the contributions of different modalities during training, we propose employing the prediction entropy of each modality to weigh their respective losses. The entropy $H(\hat{y})$ of model predictions $\hat{y}$ is calculated as $H(\hat{y}) = -\sum_c p(\hat{y}_c) \log p(\hat{y}_c)$, where $\hat{y}_c$ represents the probability of class c. Entropy serves as a metric for prediction confidence, with higher confidence yielding lower entropy values. For each prediction $\hat{y}^k$ from the k-th modality, we calculate an entropy $H(\hat{y}^k)$ and a cross-entropy loss $\mathcal{L}_{cls}^k = CE(\hat{y}^k, y)$. The objective is to assign less weight to the modality with higher entropy in contributing to the final prediction loss. Given $M+1$ entropy values $H(\hat{y}^0)$, $H(\hat{y}^1)$, ..., $H(\hat{y}^M)$, the weight of the k-th prediction is calculated by:

$$w_k = \frac{e^{-H(\hat{y}^k)/T}}{\sum_{i=0}^{M} e^{-H(\hat{y}^i)/T}}, \tag{4}$$

where T is the temperature. Then, the final prediction loss is defined as:

$$\mathcal{L}_{cls} = \sum_k w_k \mathcal{L}_{cls}^k. \tag{5}$$

Similar to previous work [50], we also add an entropy minimization loss as:

$$\mathcal{L}_{EntMin} = \sum_k H(\hat{y}^k). \tag{6}$$

3.6   Final Loss and Inference

The final loss is obtained as the weighted sum of the previously defined losses:

$$\mathcal{L} = \mathcal{L}_{cls} + \alpha_1 \mathcal{L}_{MaskedTrans} + \alpha_2 \mathcal{L}_{MulJig} + \alpha_3 \mathcal{L}_{EntMin}, \tag{7}$$

where $\mathcal{L}_{cls}$ is the cross-entropy loss for classification. The hyperparameters α_1, α_2, and α_3 control the relative importance of the Masked Cross-modal Translation, Multimodal Jigsaw Puzzles, and Entropy Minimization, respectively.

During inference, we use $\hat{y}^0$ as the final prediction. Following the approach outlined in previous research [53], we employ a threshold on the prediction confidence. Samples with a confidence level lower than the specified threshold are subsequently labeled as belonging to an open class.

3.7   Extension to Multimodal Open-Set Domain Adaptation

Different from MM-OSDG, in MM-OSDA, we have access to sample instances $\mathcal{D}_T = \{(\mathbf{x}_i^T)\}_{i=1}^{N_T}$ from the target domain during training. However, the labels y_i^T for each $\mathbf{x}_i^T$ are not available, where $y_i^T \in \mathcal{C} \cup \mathcal{U}$, encompassing both known and unknown classes. The first step is to distinguish the known samples ($y_i^T \in \mathcal{C}$) and unknown samples ($y_i^T \in \mathcal{U}$) from $\mathcal{D}_T$ to mitigate the risk of negative transfer. Given a sample $\mathbf{x}^T$ from the target domain, we calculate a prediction $\hat{y}^0 = \delta(f(\mathbf{x^T}))$ and use $max(\hat{y}^0)$ as the prediction confidence to identify unknown classes. The assumption is that the model exhibits less confidence in predicting unknown samples compared to known ones. A target sample is classified as unknown if $max(\hat{y}^0)$ falls below a threshold τ, where we typically set $\tau = 0.5$.

After filtering out all unknown samples from the target domain, we apply our self-supervised pretext tasks to the remaining known samples $\widetilde{\mathcal{D}}_T$ to align the source and target domains. For a given target sample $\widetilde{\mathbf{x}}^T$ from $\widetilde{\mathcal{D}}_T$, we first calculate the embedding for each modality as $[\widehat{\mathbf{E}}^1, ..., \widehat{\mathbf{E}}^k, ..., \widehat{\mathbf{E}}^M]$ and then use $\mathcal{L}_{MaskedTrans}$ and $\mathcal{L}_{MulJig}$ to calculate the losses associated with different self-supervised pretext tasks. Additionally, we incorporate $\mathcal{L}_{EntMin}$ into the prediction of target samples. Finally, we can redefine our loss under the MM-OSDA setting as:

$$\mathcal{L} = \mathcal{L}_{cls} + \alpha_1(\mathcal{L}_{MaskedTrans}^s + \mathcal{L}_{MaskedTrans}^t) + \alpha_2(\mathcal{L}_{MulJig}^s + \mathcal{L}_{MulJig}^t) \\ + \alpha_3(\mathcal{L}_{EntMin}^s + \mathcal{L}_{EntMin}^t), \tag{8}$$

where $\mathcal{L}^s$ and $\mathcal{L}^t$ correspond to different losses applied to source and target domain data respectively.

4   Experiments

4.1   Experimental Setting

Dataset. We use the EPIC-Kitchens [8] and HAC [12] datasets for experiments and adjust these to the open-set setups. We follow the experimental protocol

used in previous research [12]. The EPIC-Kitchens dataset includes eight actions ('put', 'take', 'open', 'close', 'wash', 'cut', 'mix', and 'pour') recorded across three distinct kitchens, constituting three separate domains D1, D2, and D3. The HAC dataset comprises seven actions ('sleeping', 'watching TV', 'eating', 'drinking', 'swimming', 'running', and 'opening door') performed by humans, animals, and cartoon figures, forming three distinct domains H, A, and C. Both datasets feature three modalities, including video, audio, and optical flow. For both datasets, we organized the action names in alphabetic order, designating the first class as unknown and the remaining classes as known. Additionally, we explored the impact of different levels of openness in the ablation study.

Evaluation Metrics. For performance evaluation, we employ three widely recognized evaluation metrics commonly used in OSDA literature [4, 31, 32]. In OSDA, rather than using average per-class accuracy OS, we present results using the harmonic mean HOS. The HOS is calculated as $\frac{2 \times OS^* \times UNK}{OS^* + UNK}$, where OS^* represents the accuracy in known categories, and UNK represents the accuracy in unknown classes. The HOS metric offers a balance between accuracy in known and unknown categories, providing a more nuanced evaluation than OS alone. This distinction is particularly important in scenarios where the accuracy of unknown classes is notably low, emphasizing the significance of effective detection of unknown categories. The HOS is alternatively denoted as H-score in some paper [47, 52].

Baselines. As the first to address the Multimodal Open-Set DG and DA problems, we benchmark our method against several state-of-the-art (SOTA) Multimodal DG and DA approaches, including MM-SADA [37], RNA-Net [43], and SimMMDG [12]. Additionally, MEDIC [52], the SOTA unimodal Open-Set DG method based on the meta-learning framework, is extended to multimodal scenarios for a comprehensive comparison. The DeepAll approach involves feeding all data from source domains to the network without employing any domain generalization strategies. For all baselines, experiments are conducted using the official code provided by the respective authors, and we apply the same prediction confidence-based method as in this paper to detect unknown classes. More implementation details are provided in the Appendix.

4.2   Results

Multimodal Open-set DG.   Tab. 2 and Tab. 3 present the results under the multimodal open-set DG setting, where the model is trained on several source domains with multiple modalities and tested on a single target domain with open classes. To verify the adaptability of our framework to various modalities, we conduct experiments by combining any two modalities, as well as all three modalities. The results on the EPIC-Kitchens dataset demonstrate that *MOOSA* consistently outperforms all baselines by a substantial margin across all cases, yielding average improvements in HOS of up to 5.63%. Notably, our performance further improves when combining all three modalities, surpassing the results achieved with any two modalities. Conversely, some baseline methods fail to exhibit improved results with an increased number of modalities,

	Modality			D2, D3 → D1			D1, D3 → D2			D1, D2 → D3			*Mean*		
	Video	Audio	Flow	OS*	UNK	**HOS**	OS*	UNK	**HOS**	OS*	UNK	**HOS**	OS*	UNK	**HOS**
DeepAll	✓	✓		38.90	50.00	43.76	47.69	44.64	46.12	52.16	42.25	46.69	46.25	45.63	45.52
MM-SADA [37]	✓	✓		35.41	50.00	41.46	47.12	50.00	48.52	51.16	46.48	48.71	44.56	48.83	46.23
RNA-Net [43]	✓	✓		39.90	38.24	39.05	43.80	53.57	48.20	46.29	59.15	51.94	43.33	50.32	46.40
MEDIC [52]	✓	✓		38.40	55.88	45.52	34.44	69.64	46.09	39.98	59.15	47.71	37.61	61.56	46.44
SimMMDG [12]	✓	✓		34.41	70.59	46.27	46.83	51.79	49.18	43.19	63.38	51.37	41.48	61.92	48.94
MOOSA (ours)	✓	✓		41.90	82.35	**55.54**	45.53	64.29	**53.31**	45.52	69.01	**54.85**	44.32	71.88	**54.57**
DeepAll	✓		✓	47.88	47.06	47.47	48.42	44.64	46.45	50.28	36.62	42.38	48.86	42.77	45.43
MM-SADA [37]	✓		✓	44.39	47.06	45.69	40.20	48.21	43.84	45.96	47.89	46.90	43.52	47.72	45.48
RNA-Net [43]	✓		✓	48.13	50.00	49.05	50.72	48.21	49.44	51.61	39.44	44.71	50.15	45.88	47.73
MEDIC [52]	✓		✓	44.14	61.76	51.49	43.37	60.71	50.60	40.42	56.34	47.07	42.64	59.60	49.72
SimMMDG [12]	✓		✓	42.14	67.65	51.93	41.93	78.57	54.68	41.75	64.79	50.78	41.94	70.34	52.46
MOOSA (ours)	✓		✓	45.14	61.76	**52.16**	50.00	62.50	**55.56**	47.84	54.93	**51.14**	47.66	59.73	**52.95**
DeepAll		✓	✓	40.90	44.12	42.45	46.25	37.50	41.42	38.32	46.48	42.01	41.82	42.70	41.96
MM-SADA [37]		✓	✓	38.90	47.06	42.59	45.97	42.86	44.36	45.52	40.85	43.05	43.46	43.59	43.33
RNA-Net [43]		✓	✓	37.16	52.94	43.67	46.69	46.43	46.56	47.51	40.85	43.93	43.79	46.74	44.72
MEDIC [52]		✓	✓	44.39	47.06	45.69	38.47	60.71	47.10	40.53	49.30	44.49	41.13	52.36	45.76
SimMMDG [12]		✓	✓	43.64	52.94	47.84	46.54	67.86	55.21	42.35	60.45	49.63	43.84	62.88	51.56
MOOSA (ours)		✓	✓	42.64	55.88	**48.37**	46.69	75.00	**57.55**	42.19	57.75	**48.76**	43.84	62.88	**51.56**
DeepAll	✓	✓	✓	39.15	50.00	43.92	52.45	35.71	42.49	53.82	39.44	45.52	48.47	41.72	43.98
MM-SADA [37]	✓	✓	✓	38.90	41.18	40.01	45.68	51.79	48.54	48.39	49.30	48.84	43.00	49.30	45.89
RNA-Net [43]	✓	✓	✓	40.15	50.00	44.54	48.85	44.64	46.65	48.39	49.30	48.84	45.80	47.98	46.68
MEDIC [52]	✓	✓	✓	43.64	64.71	52.13	47.84	64.29	54.86	45.29	59.15	51.30	45.59	62.72	52.76
SimMMDG [12]	✓	✓	✓	40.15	73.53	51.94	45.68	78.57	57.77	39.09	78.87	52.27	41.64	76.99	53.99
MOOSA (ours)	✓	✓	✓	46.13	73.52	**56.70**	53.03	67.86	**59.53**	48.84	60.56	**54.07**	49.33	67.31	**56.77**

Table 2: Multimodal **Open-set DG** with different combinations of modalities on EPIC-Kitchens dataset.

	Modality			A, C → H			H, C → A			H, A → C			*Mean*		
	Video	Audio	Flow	OS*	UNK	**HOS**	OS*	UNK	**HOS**	OS*	UNK	**HOS**	OS*	UNK	**HOS**
DeepAll	✓	✓	✓	66.02	71.72	68.75	74.51	54.14	62.71	44.59	55.56	49.48	61.71	60.47	60.31
MM-SADA [37]	✓	✓	✓	60.13	72.22	65.63	76.46	52.63	62.35	40.64	56.35	47.23	59.08	60.40	58.40
RNA-Net [43]	✓	✓	✓	70.31	67.68	68.97	72.06	51.88	60.33	46.99	55.56	50.91	63.12	58.37	60.07
MEDIC [52]	✓	✓	✓	56.18	77.78	65.24	58.47	69.17	63.37	42.93	61.11	50.43	52.53	69.35	59.68
SimMMDG [12]	✓	✓	✓	62.83	77.27	69.30	63.78	75.19	69.01	42.41	64.29	51.11	56.34	72.25	63.14
MOOSA (ours)	✓	✓	✓	66.69	78.79	**72.24**	66.36	75.94	**70.83**	45.01	62.70	**52.40**	59.35	72.48	**65.16**

Table 3: Multimodal **Open-set DG** with all modalities on HAC dataset. The results for different combinations of any two modalities are in the Appendix.

suggesting an inability to effectively leverage the complementary information between modalities. To validate the versatility of *MOOSA*, we extend the experiments to the HAC dataset, with results consistent with those obtained on the EPIC-Kitchens dataset, as detailed in Tab. 3 and in the Appendix. Our method consistently outperforms all baselines by a significant margin in most cases, yielding average improvements in terms of *HOS* of up to 3.37%.

Multimodal Single-source Open-set DG. The inclusion of domain labels is not required within our framework. This characteristic makes our method immediately applicable to multimodal single-source open-set DG without necessitating modifications, where the model is trained exclusively on a single source domain and subsequently tested across multiple target domains. The *HOS* results (Tab. 4) within the Multimodal

Source:	D1		D2		D3		
Method Target:	D2	D3	D1	D3	D1	D2	*Mean*
DeepAll	40.63	39.92	37.34	38.17	33.35	41.00	38.40
MM-SADA [37]	40.78	38.05	33.23	36.14	35.48	41.55	37.54
RNA-Net [43]	40.95	40.86	32.61	38.39	32.01	44.62	38.24
SimMMDG [12]	43.28	43.67	43.91	48.68	45.48	**51.18**	46.03
MOOSA (ours)	**49.04**	**46.55**	**44.01**	**51.89**	**50.39**	49.00	**48.48**

Table 4: Multimodal **Single-source Open-set DG** with video and audio modalities on EPIC-Kitchens dataset. The *HOS* value is reported.

Single-source Open-set DG paradigm demonstrate robust generalization to unseen domains of *MOOSA*, with an average *HOS* improvement of up to 2.45%.

	D1 → D2			D1 → D3			D2 → D1			D2 → D3			D3 → D1			D3 → D2			*Mean*		
	OS*	UNK	**HOS**	OS*	UNK	**HOS**	OS*	UNK	**HOS**	OS*	UNK	**HOS**	OS*	UNK	**HOS**	OS*	UNK	**HOS**	OS*	UNK	**HOS**
DeepAll	37.61	42.86	40.06	37.54	35.21	36.34	35.66	44.12	39.44	45.85	49.30	47.51	40.40	44.12	42.18	37.61	32.14	34.66	39.11	41.29	40.03
MM-SADA [37]	36.89	44.64	40.40	36.99	52.11	43.27	32.17	44.12	37.21	45.29	50.70	47.85	34.16	41.18	37.34	35.59	48.21	40.95	36.85	46.83	41.17
RNA-Net [43]	42.51	37.50	39.85	30.56	49.30	37.73	32.17	41.18	36.12	46.95	33.80	39.31	41.65	35.29	38.21	48.56	35.71	41.16	40.40	38.80	38.73
SimMMDG [12]	33.86	41.07	37.12	39.31	56.34	46.31	37.16	58.82	45.54	40.31	71.83	51.64	31.17	52.94	39.24	42.51	57.14	48.75	37.39	56.36	44.77
MOOSA (ours)	35.73	67.86	**46.82**	45.07	50.70	**47.72**	40.15	64.71	**49.55**	46.62	59.15	**52.15**	38.15	55.88	**45.35**	45.39	60.71	**51.94**	41.85	59.84	**48.92**

Table 5: Multimodal **Open-set DA** with video and audio modalities on EPIC-Kitchens dataset.

	Modality			EPIC-Kitchens dataset				HAC dataset			
Method	Video	Audio	Flow	D2, D3 → D1	D1, D3 → D2	D1, D2 → D3	*Mean*	A, C → H	H, C → A	H, A → C	*Mean*
DeepAll	✓	✓	✓	55.63	59.20	58.01	57.61	69.07	71.30	51.47	63.95
MM-SADA [37]	✓	✓	✓	51.72	58.40	59.34	56.49	72.53	72.19	55.51	66.74
RNA-Net [43]	✓	✓	✓	52.41	57.20	60.16	56.59	69.00	73.40	51.65	64.68
SimMMDG [12]	✓	✓	✓	**63.68**	70.13	67.76	67.19	77.65	**79.03**	56.62	71.10
MOOSA (ours)	✓	✓	✓	63.22	**70.27**	**68.38**	**67.29**	**78.01**	77.15	**61.40**	**72.19**

Table 6: Multimodal **Closed-set DG** with all three modalities on EPIC-Kitchens and HAC datasets. The results for different combinations of any two modalities are in the Appendix.

Multimodal Open-set DA. In this setting, unlabeled data from the target domain is available during training. As we are the first to address the MM-OSDA problem, we extend several methods to this setup and consider them as baselines. For MM-SADA [37], we employ the within-modal adversarial alignment loss to align the distribution of each modality between source and target domains. In the case of RNA-Net [43], we apply the relative norm alignment loss on the target domain data. Additionally, for SimMMDG [12], we incorporate their contrastive loss, distance loss, and cross-modal translation loss to the target data to align the distribution across domains. As depicted in Tab. 5, our method achieves competitive performance compared to the best baseline, yielding an improvement of 4.15% on *HOS*.

Multimodal Closed-set DG. We also evaluate *MOOSA* under the multimodal closed-set DG setup [12], where the label spaces of the source and target domains are identical, as presented in Tab. 6 and detailed further in the Appendix. In comparison with the current SOTA multimodal DG baseline, SimMMDG [12], our method achieves further improvements of up to 1.30%. This result suggests that the proposed multimodal self-supervised pretext tasks and entropy weighting mechanism contribute significantly to the model's ability to learn more discriminative features for enhanced generalization.

4.3   Ablation Studies and Analysis

Disparate label sets across domains. In previous MM-OSDG experiments, we assumed that the label sets of all source domains were identical, and the target domain included all label sets from source domains along with additional unknown classes. However, in practical scenarios, different source domains may comprise distinct label sets, while the target domain might not include all label sets present in the source domains. Tab. 7 illustrates the results under this more challenging setup, where our method consistently demonstrates improvements in comparison to other baselines. The details of the label split under this setup are provided in the Appendix.

	D2, D3 → D1			D1, D3 → D2			D1, D2 → D3			*Mean*		
	OS*	UNK	**HOS**	OS*	UNK	**HOS**	OS*	UNK	**HOS**	OS*	UNK	**HOS**
DeepAll	49.04	35.29	41.05	52.47	30.36	38.46	50.45	39.44	44.27	50.65	35.03	41.26
MM-SADA [37]	41.76	47.06	44.25	52.47	39.29	44.93	50.27	38.03	43.30	48.17	41.46	44.16
RNA-Net [43]	48.28	38.24	42.67	51.61	32.14	39.61	48.28	35.21	40.72	49.39	35.20	41.00
SimMMDG [12]	48.28	47.06	47.66	39.35	62.50	48.30	43.92	54.93	48.81	43.85	54.83	48.26
MOOSA (ours)	49.04	61.76	**54.67**	46.24	57.14	**51.11**	45.92	59.15	**51.70**	47.07	59.35	**52.49**

Table 7: Abation on disparate label sets across domains for MM-OSDG with video and audio modalities on EPIC-Kitchens dataset.

	D2, D3 → D1			D1, D3 → D2			D1, D2 → D3			*Mean*		
	OS*	UNK	**HOS**	OS*	UNK	**HOS**	OS*	UNK	**HOS**	OS*	UNK	**HOS**
MEDIC [52] (V)	36.41	58.82	44.98	34.73	66.07	45.53	31.34	59.15	40.97	34.16	61.35	43.83
MEDIC [52] (A)	22.44	47.06	30.39	27.67	62.50	38.35	35.44	46.48	40.21	28.52	52.01	36.32
MEDIC [52] (F)	36.91	64.71	47.00	45.97	58.93	51.65	42.30	56.34	48.32	41.73	59.99	48.99
MOOSA (ours) (V+A+F)	46.13	73.52	**56.70**	50.72	76.79	**61.09**	46.84	77.46	**58.38**	47.90	75.92	**58.72**

Table 8: Compared with unimodal open-set DG. V: video, A: audio, F: optical flow.

MulJig	MaskedTrans	EntWei	EntMin	*Mean* **HOS**
				50.62
✓				51.88
	✓			51.60
✓	✓			53.91
✓	✓	✓		54.27
✓	✓	✓	✓	**54.57**

Table 9: Ablations of each proposed module in our *MOOSA*.

Self-supervised Tasks	*Mean* **HOS**
Multimodal Alignment [37]	51.17
Cross-modal Translation [12]	48.94
Relative Norm Alignment [43]	50.55
Cross-modal Distillation [19]	50.88
Masked Cross-modal Translation	51.60
Multimodal Jigsaw Puzzles	51.88
MaskedTrans+MulJig	**53.91**

Table 10: Abation on different self-supervised tasks.

Compared with unimodal open-set DG: Tab. 8 offers a comparison against unimodal open-set DG algorithms that exclusively utilize either video, audio, or optical flow as inputs. We selected the SOTA unimodal open-set DG method, MEDIC, as our baseline. By leveraging information from multiple modalities, our *MOOSA* demonstrates substantial improvements up to 7.78% in *HOS* compared to unimodal open-set DG methods.

Ablation on each proposed module. We conducted comprehensive ablation studies to investigate the role of each module of our method on the EPIC-Kitchens dataset, as illustrated in Tab. 9. The inclusion of either masked cross-modal translation or multimodal Jigsaw puzzles as pretext tasks demonstrates performance benefits. The combination of these two tasks further enhances the results, verifying their complementarity. The best performance is achieved when incorporating entropy weighting and minimization to balance the loss from different modalities.

On the effectiveness of other self-supervised pretext tasks. We also explore the feasibility of replacing our multimodal self-supervised tasks with alternative pretext tasks. As depicted in Tab. 10, multimodal alignment [37] and cross-modal distillation [19] prove to be effective as pretext tasks for MM-OSDG. However, our proposed masked cross-modal translation and multimodal Jigsaw puzzles emerge as superior self-supervised tasks in this context.

Varying the ratio of known to unknown classes. The term "openness" denotes the proportion of unknown target classes during the test phase. To assess the adaptability of our model across different levels of openness, we perform experiments on the EPIC-Kitchens dataset. The results presented in Fig. 2 demonstrate that increasing the number of open classes presents more challenges for
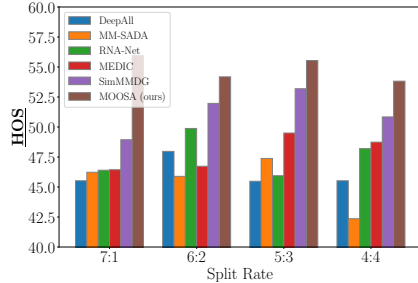
Fig. 2: The average *HOS* with the varying known-unknown split rates on EPIC-Kitchens dataset.
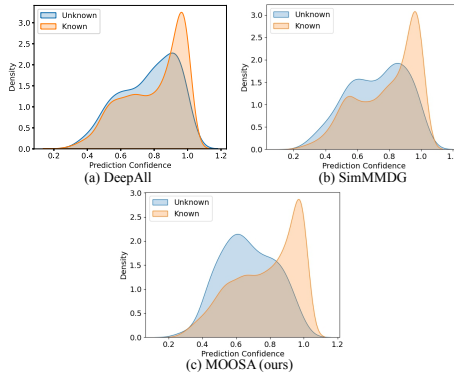
Fig. 3: The distribution curves of prediction confidence for unknown and known classes obtained using various methods.

the classification task, leading to reduced model accuracy. However, our method consistently outperforms alternative approaches, showcasing significant improvements and validating the robustness of our model across diverse scenarios.

Visualization of confidence score distributions for known and unknown classes. In Fig. 3, we illustrate the distribution curves of prediction confidence for unknown and known classes across different methods. A smaller overlap between unknown and known data indicates superior open-class detection performance, whereas a larger overlap suggests weaker detection results. The unknown curve (depicted in blue) for our method exhibits a distinctive peak at a lower position, resulting in minimal overlap with other known data. This distinct characteristic underscores the superior open-class detection capability achieved by our method. This success can be attributed to the high-quality in-distribution feature representation.

5   Conclusion

In this paper, we introduce the Multimodal Open-Set Domain Generalization problem and propose a novel solution *MOOSA* employing self-supervised pretext tasks. Our approach involves the development of two complementary multimodal self-supervised tasks, namely Masked Cross-modal Translation and Multimodal Jigsaw Puzzles, aimed at learning multimodal representative features for robust generalization and effective open-class detection. To maintain a balance in the loss across different modalities, we introduce an Entropy Weighting mechanism during training. Furthermore, we extend our framework to address Multimodal Open-Set Domain Adaptation. Through extensive evaluations on the EPIC-Kitchens and HAC datasets, our method demonstrates superior performance compared to competing baselines. Notably, our approach exhibits robustness across variations in label sets, openness levels, and parameter choices. In future work, methods proposed in Multimodal OOD Detection [14] could be good alternatives for open-class detection.

Acknowledgments

References

1. Bao, H., Dong, L., Piao, S., Wei, F.: Beit: Bert pre-training of image transformers. In: ICLR (2022)
2. Bergman, L., Hoshen, Y.: Classification-based anomaly detection for general data. arXiv preprint arXiv:2005.02359 (2020)
3. Blum, H., Sarlin, P.E., Nieto, J., Siegwart, R., Cadena, C.: Fishyscapes: A benchmark for safe semantic segmentation in autonomous driving. In: ICCVW. pp. 0–0 (2019)
4. Bucci, S., Loghmani, M.R., Tommasi, T.: On the effectiveness of image rotation for open set domain adaptation. In: ECCV (2020)
5. Caesar, H., Bankiti, V., Lang, A.H., Vora, S., Liong, V.E., Xu, Q., Krishnan, A., Pan, Y., Baldan, G., Beijbom, O.: nuscenes: A multimodal dataset for autonomous driving. In: CVPR (2020)
6. Carlucci, F.M., D'Innocente, A., Bucci, S., Caputo, B., Tommasi, T.: Domain generalization by solving jigsaw puzzles. In: CVPR (2019)
7. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A simple framework for contrastive learning of visual representations. In: ICML. pp. 1597–1607 (2020)
8. Damen, D., Doughty, H., Farinella, G.M., Fidler, S., Furnari, A., Kazakos, E., Moltisanti, D., Munro, J., Perrett, T., Price, W., Wray, M.: Scaling egocentric vision: The epic-kitchens dataset. In: ECCV (2018)
9. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018)
10. Doersch, C., Gupta, A., Efros, A.A.: Unsupervised visual representation learning by context prediction. In: ICCV (2015)
11. Dong, H., Chen, X., Särkkä, S., Stachniss, C.: Online pole segmentation on range images for long-term lidar localization in urban environments. Robotics and Autonomous Systems 159, 104283 (2023)
12. Dong, H., Nejjar, I., Sun, H., Chatzi, E., Fink, O.: SimMMDG: A simple and effective framework for multi-modal domain generalization. In: NeurIPS (2023)
13. Dong, H., Zhang, X., Xu, J., Ai, R., Gu, W., Lu, H., Kannala, J., Chen, X.: Superfusion: Multilevel lidar-camera fusion for long-range hd map generation. arXiv preprint arXiv:2211.15656 (2022)
14. Dong, H., Zhao, Y., Chatzi, E., Fink, O.: MultiOOD: Scaling Out-of-Distribution Detection for Multiple Modalities. arXiv preprint arXiv:2405.17419 (2024)
15. Ganin, Y., Lempitsky, V.: Unsupervised domain adaptation by backpropagation. In: ICML (2015)
16. Ganin, Y., Ustinova, E., Ajakan, H., Germain, P., Larochelle, H., Laviolette, F., Marchand, M., Lempitsky, V.: Domain-adversarial training of neural networks. The journal of machine learning research 17(1), 2096–2030 (2016)
17. Gidaris, S., Singh, P., Komodakis, N.: Unsupervised representation learning by predicting image rotations. arXiv preprint arXiv:1803.07728 (2018)

18. Golan, I., El-Yaniv, R.: Deep anomaly detection using geometric transformations. In: NeurIPS (2018)
19. Gupta, S., Hoffman, J., Malik, J.: Cross modal distillation for supervision transfer. In: CVPR (2016)
20. He, K., Chen, X., Xie, S., Li, Y., Dollár, P., Girshick, R.: Masked autoencoders are scalable vision learners. In: CVPR (2022)
21. He, K., Fan, H., Wu, Y., Xie, S., Girshick, R.: Momentum contrast for unsupervised visual representation learning. In: CVPR (2020)
22. Huang, Z., Wang, H., Xing, E.P., Huang, D.: Self-challenging improves cross-domain generalization. In: ECCV (2020)
23. Jia, C., Yang, Y., Xia, Y., Chen, Y.T., Parekh, Z., Pham, H., Le, Q., Sung, Y.H., Li, Z., Duerig, T.: Scaling up visual and vision-language representation learning with noisy text supervision. In: ICML (2021)
24. Kazakos, E., Nagrani, A., Zisserman, A., Damen, D.: Epic-fusion: Audio-visual temporal binding for egocentric action recognition. In: ICCV (2019)
25. Kim, D., Tsai, Y.H., Zhuang, B., Yu, X., Sclaroff, S., Saenko, K., Chandraker, M.: Learning cross-modal contrastive features for video domain adaptation. In: ICCV (2021)
26. Larsson, G., Maire, M., Shakhnarovich, G.: Colorization as a proxy task for visual understanding. In: CVPR (2017)
27. Li, D., Yang, Y., Song, Y.Z., Hospedales, T.M.: Deeper, broader and artier domain generalization. In: ICCV (2017)
28. Li, D., Yang, Y., Song, Y.Z., Hospedales, T.M.: Learning to generalize: Meta-learning for domain generalization. In: AAAI (2018)
29. Li, H., Pan, S.J., Wang, S., Kot, A.C.: Domain generalization with adversarial feature learning. In: CVPR (2018)
30. Li, J., Chen, P., He, Z., Yu, S., Liu, S., Jia, J.: Rethinking out-of-distribution (ood) detection: Masked image modeling is all you need. In: CVPR (2023)
31. Li, W., Liu, J., Han, B., Yuan, Y.: Adjustment and alignment for unbiased open set domain adaptation. In: CVPR (2023)
32. Liu, H., Cao, Z., Long, M., Wang, J., Yang, Q.: Separate to adapt: Open set domain adaptation via progressive separation. In: CVPR (2019)
33. Liu, X., Zhang, F., Hou, Z., Mian, L., Wang, Z., Zhang, J., Tang, J.: Self-supervised learning: Generative or contrastive. IEEE transactions on knowledge and data engineering **35**(1), 857–876 (2021)
34. Lu, W., Wang, J., Sun, X., Chen, Y., Xie, X.: Out-of-distribution representation learning for time series classification. In: ICLR (2023)
35. Misra, I., Maaten, L.v.d.: Self-supervised learning of pretext-invariant representations. In: CVPR (2020)
36. Muandet, K., Balduzzi, D., Schölkopf, B.: Domain generalization via invariant feature representation. In: ICML (2013)
37. Munro, J., Damen, D.: Multi-modal domain adaptation for fine-grained action recognition. In: CVPR (2020)
38. Noroozi, M., Favaro, P.: Unsupervised learning of visual representations by solving jigsaw puzzles. In: ECCV (2016)
39. Pan, X., Luo, P., Shi, J., Tang, X.: Two at once: Enhancing learning and generalization capacities via ibn-net. In: ECCV (2018)
40. Panareda Busto, P., Gall, J.: Open set domain adaptation. In: ICCV (2017)
41. Pathak, D., Krahenbuhl, P., Donahue, J., Darrell, T., Efros, A.A.: Context encoders: Feature learning by inpainting. In: CVPR (2016)

42. Piratla, V., Netrapalli, P., Sarawagi, S.: Efficient domain generalization via common-specific low-rank decomposition. In: ICML (2020)
43. Planamente, M., Plizzari, C., Alberti, E., Caputo, B.: Domain generalization through audio-visual relative norm alignment in first person action recognition. In: WACV (2022)
44. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I.: Learning transferable visual models from natural language supervision. In: ICML (2021)
45. Saha, A., Subramanya, A., Pirsiavash, H.: Hidden trigger backdoor attacks. In: AAAI (2020)
46. Saito, K., Yamamoto, S., Ushiku, Y., Harada, T.: Open set domain adaptation by backpropagation. In: ECCV (2018)
47. Shu, Y., Cao, Z., Wang, C., Wang, J., Long, M.: Open domain generalization with domain-augmented meta-learning. In: CVPR (2021)
48. Tobin, J., Fong, R., Ray, A., Schneider, J., Zaremba, W., Abbeel, P.: Domain randomization for transferring deep neural networks from simulation to the real world. In: IROS (2017)
49. Tzeng, E., Hoffman, J., Zhang, N., Saenko, K., Darrell, T.: Deep domain confusion: Maximizing for domain invariance. arXiv preprint arXiv:1412.3474 (2014)
50. Wang, D., Shelhamer, E., Liu, S., Olshausen, B., Darrell, T.: Tent: Fully test-time adaptation by entropy minimization. arXiv preprint arXiv:2006.10726 (2020)
51. Wang, J., Lan, C., Liu, C., Ouyang, Y., Qin, T., Lu, W., Chen, Y., Zeng, W., Yu, P.: Generalizing to unseen domains: A survey on domain generalization. IEEE Transactions on Knowledge and Data Engineering (2022)
52. Wang, X., Zhang, J., Qi, L., Shi, Y.: Generalizable decision boundaries: Dualistic meta-learning for open set domain generalization. In: ICCV (2023)
53. You, K., Long, M., Cao, Z., Wang, J., Jordan, M.I.: Universal domain adaptation. In: CVPR (2019)
54. Zhang, H., Cisse, M., Dauphin, Y.N., Lopez-Paz, D.: mixup: Beyond empirical risk minimization. In: ICLR (2018)
55. Zhang, Y., Doughty, H., Shao, L., Snoek, C.G.: Audio-adaptive activity recognition across video domains. In: CVPR (2022)
56. Zhou, K., Yang, Y., Hospedales, T., Xiang, T.: Deep domain-adversarial image generation for domain generalisation. In: AAAI (2020)
57. Zhu, R., Li, S.: Crossmatch: Cross-classifier consistency regularization for open-set single domain generalization. In: ICLR (2021)