

Open-set Cross Modal Generalization via Multimodal Unified Representation

Hai Huang^{1,2} Yan Xia¹ Shulei Wang¹ Hanting Wang¹
Minghui Fang¹ Shengpeng Ji¹ Sashuai Zhou¹ Tao Jin¹ Zhou Zhao^{1,2†}
¹ Zhejiang University ² Shanghai Artificial Intelligence Laboratory
haihuangcode@outlook.com zhaozhou@zju.edu.cn

Abstract

This paper extends Cross Modal Generalization (CMG) to open-set environments by proposing the more challenging Open-set Cross Modal Generalization (OSCMG) task. This task evaluates multimodal unified representations in open-set conditions, addressing the limitations of prior closed-set cross-modal evaluations. OSCMG requires not only cross-modal knowledge transfer but also robust generalization to unseen classes within new modalities, a scenario frequently encountered in real-world applications. Existing multimodal unified representation work lacks consideration for open-set environments. To tackle this, we propose *MICU*, comprising two key components: *Fine-Coarse Masked multimodal InfoNCE (FCMI)* and *Cross modal Unified Jigsaw Puzzles (CUJP)*. *FCMI* enhances multimodal alignment by applying contrastive learning at both holistic semantic and temporal levels, incorporating masking to enhance generalization. *CUJP* enhances feature diversity and model uncertainty by integrating modality-agnostic feature selection with self-supervised learning, thereby strengthening the model's ability to handle unknown categories in open-set tasks. Extensive experiments on CMG and the newly proposed OSCMG validate the effectiveness of our approach. The code is available at <https://github.com/haihuangcode/CMG>.

1. Introduction

To address the challenge of scarce annotated data in downstream tasks involving rare modalities (e.g., point clouds, EEG signals), Cross Modal Generalization (CMG) [42] has been introduced as a novel task. This paradigm aims to establish unified representations through fine-grained pre-training on large-scale paired multimodal datasets, mapping semantically equivalent information across different modalities into a shared discrete dictionary. This framework enables zero-shot transfer of knowledge and capabil-

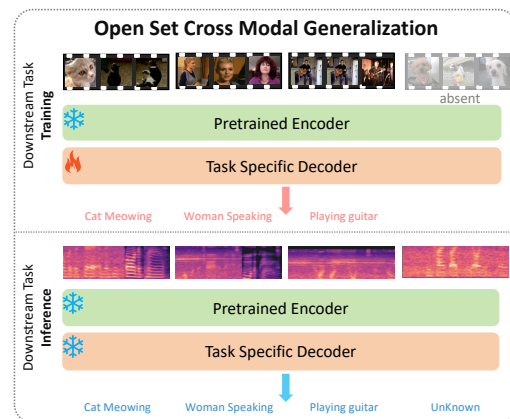


Figure 1. After unsupervised pretraining, the model is directly transferred to unseen modalities and unseen categories in downstream tasks.

ities learned from common modalities (such as images and text) to rare modalities in downstream applications, without requiring additional modality-specific annotations.

The method proposed by Xia *et al.* [42] has achieved promising fine-grained semantic alignment results through feature disentangling and cross-modal contrastive prediction. However, their work relies on a closed-set assumption, where training and test classes remain consistent across tasks. In practical applications, the target modality for transfer often includes categories that do not exactly match those in the source domain. Directly applying previous method [12, 22, 25, 33, 42, 49] for cross-modal generalization would lead to misclassification of these unknown categories, limiting its applicability in real-world scenarios.

Therefore, we introduce the **Open-Set Cross-Modal Generalization (OSCMG)** task, designed to enhance models' cross-modal generalization capabilities in open-set environments. The OSCMG task requires models not only to achieve unified representations across different modalities but also to ensure that these representations are highly generalizable, enabling effective distinction between known and unknown classes. Specifically, this approach pretrains the model in an unsupervised setting, then fine-tunes it

[†]Corresponding author

$C_s \neq C_t$	Multimodal	$M_s \neq M_t$	Task Setting
-	-	-	DG [3]
✓	-	-	OSDG [34]
-	✓	-	MMDG [10]
✓	✓	-	MM-OSDG [9]
-	✓	✓	CMG [42]
✓	✓	✓	OSCMG

Table 1. The differences between OSCMG and other related tasks. M_s and M_t represent the source and target modalities, while C_s and C_t represent the labels of the source and target modalities.

on downstream tasks with modality a , containing only the class set V , and enables it to generalize to modality b , which includes a broader class set U , where $V \subset U$; a graphical depiction can be seen in Figure 1. Similar open-set tasks include Open-Set Domain Generalization (OSDG)[34] and Multimodal Open-Set Domain Generalization (MM-OSDG)[9], which extend the challenges of Domain Generalization (DG)[3] and Multimodal Domain Generalization (MMDG)[10] to open-domain scenarios, with specific differences outlined in Tab. 1.

As a novel task, OSCMG primarily encompasses two challenging aspects. (1) To achieve cross-modal generalization, it is crucial to establish effective multimodal unified representations. However, previous works have predominantly focused on alignment at a singular level. For instance, methods like CLIP [31] and ImageBind [15] perform coarse-grained alignment by average pooling features from different modalities, which can easily overlook fine-grained cross-modal alignment relationships. Xia [42] addresses the challenge of fine-grained multimodal alignment through cross-modal contrastive predictive coding. Nonetheless, it tends to overlook the holistic semantic associations between different modalities. (2) Since large-scale labeled multimodal data is difficult to obtain, the construction of a unified representation primarily relies on learning from vast amounts of unlabeled multimodal data. We propose OSCMG to evaluate the performance of unified representation under more challenging conditions, thus adopting an unsupervised setting. This setting renders most existing label-dependent OSDG methods, such as DAML [34] and MEDIC [41], inapplicable to OSCMG. In contrast, MMJP [9], designed as a self-supervised learning approach that does not require label information, was proposed to tackle the MM-OSDG challenge and has demonstrated strong generalization capabilities in open-domain multimodal scenarios. However, MMJP is not suitable for the OSCMG task. Its core mechanism relies on utilizing all modalities to perform the jigsaw puzzles, leveraging cross-modal complementarity to enhance performance in MM-OSDG. This design makes MMJP highly sensitive to modality-specific semantics, as it depends on information from all modalities during training. However, such sensitiv-

ity can be detrimental to the learning of a unified representation, as it emphasizes modality-specific features that may negatively impact the representation’s generalization [42].

To address these challenges, we propose MICU, a novel approach that combines strong generalization with enhanced multimodal alignment through two key components: Fine-Coarse Masked Multimodal InfoNCE (FCMI) and Cross-modal Unified Jigsaw Puzzles (CUJP). (1) FCMI refines and strengthens multimodal alignment by applying masked contrastive learning at both inter-sample (holistic semantic) and intra-sample (temporal) levels, thereby capturing broad semantic consistency and fine-grained alignment to construct a more effective multimodal unified representation space. (2) Considering the unsupervised setting of OSCMG pre-training, we adopt a self-supervised learning approach that does not require label information. However, as previously mentioned, MMJP [9] is highly sensitive to modality-specific semantic features, whereas OSCMG aims to learn a unified representation by minimizing the influence of modality-specific information. To address this, we propose CUJP, which disregards modality distinctions and treats all modalities as a single unified modality. During the jigsaw puzzle process, CUJP randomly selects feature split blocks from any modality, enabling modality-agnostic learning. Furthermore, benefiting from the partitioning mechanism of the jigsaw puzzle, CUJP achieves finer-grained alignment compared to previous unified representation approaches that primarily focus on aligning entire samples [22, 42, 49], ensuring consistency at the block level. Additionally, CUJP significantly reduces computational complexity compared to MMJP, as it does not require using all feature blocks from every modality. For instance, in a three-modality setting where each modality’s features are split into four parts, MMJP requires $12! = 479001600$ sorting computations, whereas CUJP only requires $4! = 24$, leading to a substantial improvement in computational efficiency. Our contributions can be summarized as follows:

- We propose **OSCMG**, which enables the evaluation of multimodal unified representations under more realistic and complex challenges. This approach evaluates the model’s ability not only to generalize across modalities but also to transfer knowledge to unseen categories.
- We propose **MICU**, which comprises **FCMI** and **CUJP**. FCMI achieves multimodal alignment through fine- and coarse-grained contrastive learning across temporal and holistic semantic levels, enhanced by a masking mechanism. CUJP enhances modality-agnostic performance by integrating discrete unified representations with a jigsaw puzzle approach, splitting and randomly rearranging the representations.
- Our model achieves state-of-the-art performance on both CMG and OSCMG tasks, demonstrating the effectiveness of the proposed methods.

2. Related Work

Multimodal Unified Representation. Recent efforts in multimodal unified representation focus on aligning different modalities in a shared latent space [1, 29, 33], training modal-general encoders for cross-modal feature extraction [8, 40], and using cross-modal knowledge distillation to facilitate information transfer between modalities [28, 33]. Bridging techniques have also been proposed to connect continuous representation spaces to leverage complementary strengths [48]. To improve interpretability, codebooks or prototypes are used for unified representations, mapping multimodal features into discrete forms [12, 16–18, 22, 25, 42, 49]. For instance, Duan *et al.* [12] uses Optimal Transport to align features with prototypes, while Zhao *et al.* [49] enhances mutual information via self-cross-reconstruction. Xia *et al.* [42] addresses imperfect alignment by mapping sequences into a common discrete space. We retained the consideration that paired multimodal data may not be perfectly aligned and proposed FCMI, which is easier to train compared to decoupling-based methods. Furthermore, we combined the highly effective Jigsaw Puzzle approach from the self-supervised learning domain with discrete representations, introducing CUJP, which achieves better unified representation performance.

Domain and Cross-Modal Generalization. DG has been instrumental in enabling models to generalize to unseen target domains without direct access to target domain data, and has found applications in diverse fields such as medical imaging [20, 23] and action recognition [30]. Common DG methods include feature representation learning [14, 27, 39], data augmentation [38, 47], and domain-agnostic learning strategies such as domain adversarial learning [14, 43, 45] and meta-learning [19] to handle domain shifts. As multimodal research has advanced, MMDG [10, 30] emerged to address the additional complexity of generalizing across different modalities.

In scenarios where the target domain may include categories unseen during training, OSDG[34] addresses both domain generalization and unknown class detection. This concept has further developed into MM-OSDG[9], with tasks such as MOOSA leveraging multimodal self-supervised learning to enhance generalization and open-set recognition in multimodal contexts. Similarly, CMG, like MMDG, faces challenges in open-set environments. To bridge this gap in evaluating multimodal unified representations, we propose the Open-set Cross-Modal Generalization (OSCMG) task, which requires models to transfer knowledge across modalities and adapt to unseen classes within new modalities.

3. Method

In this section, we first provide a detailed definition of the proposed OSCMG task, followed by an introduction to our new architecture, MICU, designed to address this challenge. MICU primarily integrates the concepts of masked contrastive learning and self-supervised learning. We will introduce its two constituent modules separately, whereas Figure 2 illustrates the overall model architecture.

3.1. Open-set Cross Modal Generalization

OSCMG shares the same pre-training setup as CMG, where multimodal data is learned in an unsupervised manner to obtain a unified multimodal representation. The key difference lies in the evaluation of downstream tasks, OSCMG is designed to assess a model’s cross-modal generalization ability under open-set conditions. Specifically, it evaluates the model’s capacity to transfer knowledge from a source modality to a target modality while handling unseen classes absent in the source modality. During training, the model is trained on a source modality M_s and tested on a target modality M_t , where the class set of the source modality V is a subset of the class set U in the target modality, i.e., $V \subset U$. This setup challenges the model to generalize across modalities while also adapting to novel categories not encountered during training, providing a more comprehensive evaluation of cross-modal learning capabilities.

During training, the model learns representations for inputs from a source modality using the encoder Φ^{M_s} and the downstream decoder \mathbf{D} . The process is formulated as follows:

$$\mathbf{E}(\mathbf{D}(\Phi^{M_s}(\mathbf{x}_i^{M_s})), \mathbf{y}_i^{M_s}). \quad (1)$$

where $\mathbf{x}_i^{M_s}$ is the input, $\mathbf{y}_i^{M_s}$ is the corresponding label, and \mathbf{E} denotes the evaluation function. In the testing phase, the model is evaluated on a different target modality M_t , assessing its generalization capability:

$$\mathbf{E}(\mathbf{D}(\Phi^{M_t}(\mathbf{x}_i^{M_t})), \mathbf{y}_i^{M_t}). \quad (2)$$

The parameters of the encoders Φ^{M_s} and Φ^{M_t} remain frozen during both downstream training and testing, as they are fully determined during the pre-training process. With only the parameters of the decoder \mathbf{D} being updated during downstream training. Additionally, the encoders are derived from a multimodal model pretrained in an unsupervised manner, while the decoder varies according to the downstream task, typically implemented as a linear probe.

3.2. Fine-Coarse Masked Multimodal InfoNCE

In the field of multimodal unified representation, contrastive learning is a widely used alignment method. Liu *et al.* [22] enhanced discrete representations through contrastive learning, significantly improving unified representation performance, while Xia *et al.* [42] incorporated cross-modal

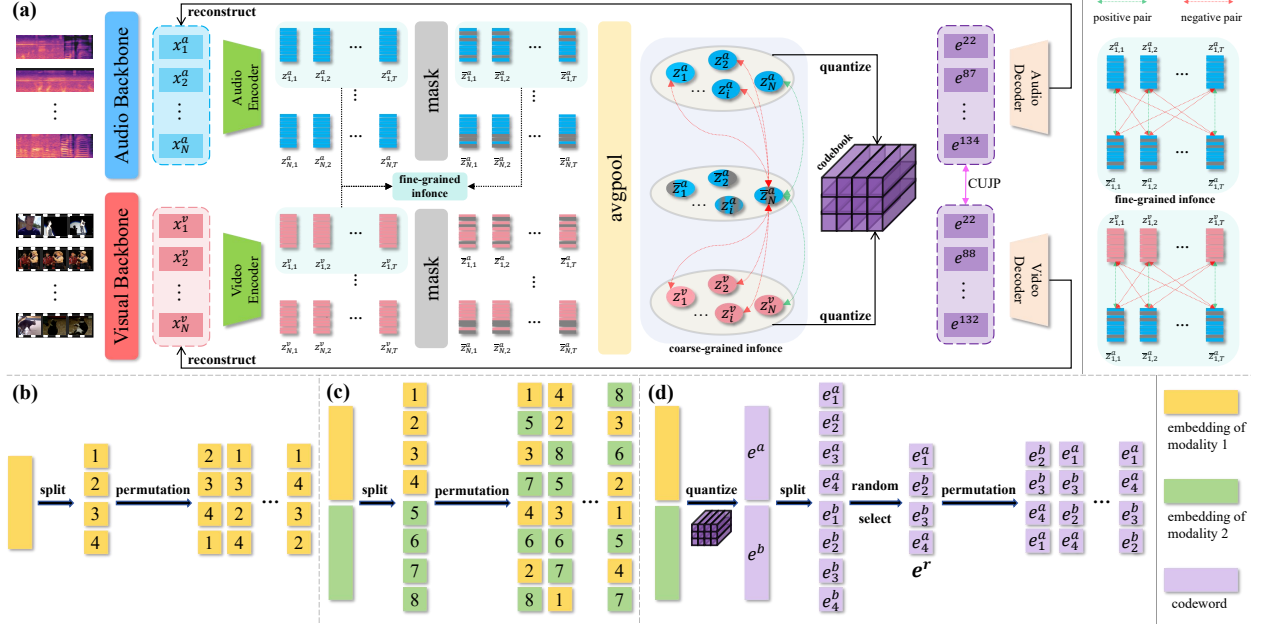


Figure 2. (a) The architecture of MICU, illustrated with an example of fine and coarse InfoNCE with masked audio and video, as well as with masked audio and audio. (b) Single-modal Jigsaw Puzzles. (c) Multimodal Jigsaw Puzzles. (d) Our proposed Cross modal Unified Jigsaw Puzzles.

contrastive learning into their disentanglement framework. Building on these foundations, we introduce **FCMI**, an improved InfoNCE approach designed for multimodal unified representation. FCMI strengthens alignment by applying contrastive learning at both inter-sample (holistic semantic) and intra-sample (temporal) levels, ensuring both broad semantic consistency and fine-grained alignment. To enhance model generalization, we introduce masking within contrastive learning, inspired by SemSeg [44], which builds class embeddings to recognize unknown categories, and Mask2Anomaly [32], which uses masked contrastive learning to sharpen the boundary between known and anomalous classes.

As shown in Figure 2(a), FCMI is divided into two parts: fine-grained and coarse-grained masked contrastive learning. The paired features extracted by the backbone from each modality are denoted as $\{(\mathbf{x}_i^a, \mathbf{x}_i^b)\}$, where $\{a, b\}$ representing paired modals. For each modality, an encoder Φ^m , where $m \in \{a, b\}$, is introduced to map the features to a uniform feature size $\mathbf{z}_i^m \in \mathbb{R}^{T \times D}$, where T and D represent the audio-video time dimension and the latent feature dimension, respectively:

$$\mathbf{z}_i^m = \Phi^m(\mathbf{x}_i^m), m \in \{a, b\}. \quad (3)$$

We then apply a mask to the features, resulting in $\bar{\mathbf{z}}_i^m = \text{Mask}(\mathbf{z}_i^m)$. This masking is sample-specific, meaning the mask is consistent across different timesteps for the same sample. To ensure effective cross-modal masked contrastive

learning, the masked positions are aligned across corresponding samples' different modalities, which will be discussed further in Figure 4.

The fine-grained masked contrastive learning is applied to different timesteps of a single sample pair. The masked features at a specific timestep are contrasted with the unmasked features of the corresponding timestep from other modalities as positive pairs, while the remaining timesteps serve as negative pairs.

$$L_{\text{fine}} = -\frac{1}{N} \frac{1}{T} \sum_{i=1}^N \sum_{j=1}^T \log \left[\frac{\exp(\bar{\mathbf{z}}_{i,j}^m \cdot (\mathbf{z}_{i,j}^n)^\top / \tau)}{\sum_{k=1}^T \exp(\bar{\mathbf{z}}_{i,j}^m \cdot (\mathbf{z}_{i,k}^n)^\top / \tau)} \right], \quad m, n \in \{a, b\}, \quad (4)$$

where N represents the number of samples, \top denotes transpose, and τ is the temperature parameter. Both m and n can represent the same modality, allowing for cross-modal as well as intra-modal alignment. This loss enables the model to learn fine-grained cross-modal alignment. Adjacent modalities time steps can serve as hard negatives, a strategy that effectively enhances contrastive learning by enforcing finer temporal discrimination and improving robustness.

Simultaneously, coarse-grained masked contrastive learning is applied across samples, where the masked features of a single sample are contrasted with the corresponding complete features from other modalities as positive pairs, and other samples as negative pairs.

$$L_{\text{coarse}} = -\frac{1}{N} \sum_{i=1}^N \log \left[\frac{\exp(\bar{\mathbf{z}}_i^m \cdot (\mathbf{z}_i^n)^\top / \tau)}{\sum_{j=1}^N \exp(\bar{\mathbf{z}}_i^m \cdot (\mathbf{z}_j^n)^\top / \tau)} \right], \quad (5)$$

$$m, n \in \{a, b\},$$

this loss facilitates the learning of multimodal alignment at the holistic semantic level.

The cross-modal InfoNCE between unmasked features is not applied, as indirect alignment has already been achieved through modality masking. Adding this extra computation would not significantly improve the results.

3.3. Cross Modal Unified Jigsaw Puzzle

Previous studies [3, 26] have used Jigsaw puzzles to learn visual representations, where the task is to reconstruct an original image from shuffled parts. MMJP [9] extended this idea to MM-OSDG. While CUJP shares the use of Jigsaw puzzles with MMJP, it differs by operating on unified discrete representations rather than shuffling all modality parts. Specifically, CUJP utilizes quantized features $\hat{\mathbf{z}}_{i,t}^m$ from the codebook, where each segment is a randomly selected code-word e from any modality. This design significantly enhances modality-agnostic feature diversity and uncertainty, making CUJP particularly well-suited for OSCMG. It effectively integrates the advantages of MMJP in open-domain multimodal learning while preserving the unified representation property, which does not require modality-specific information. The illustrations of the three different Jigsaw puzzles are shown in subfigures (b), (c), and (d) of Figure 2.

To explicitly represent the unified representation of different modalities, we utilize a shared latent codebook $\mathbf{E} \in \mathbb{R}^{H \times D}$ across multi modalities. We apply a vector quantization VQ operation to map the multimodal features \mathbf{z}_i^a and \mathbf{z}_i^b into discrete latent codes. Here, $t \in [0, T)$, and T , H , and D represent the time steps, the size of the discrete latent space, and the hidden dimension, respectively.

$$\hat{\mathbf{z}}_{i,t}^m = VQ(\Phi^m(\mathbf{x}_{i,t}^m)) = VQ(\mathbf{z}_{i,t}^m) = e_l, \quad (6)$$

where $l = \operatorname{argmin}_j \|\Phi^m(x) - e_j\|_2$, $m \in \{a, b\}$.

Not all $\hat{\mathbf{z}}_{i,t}^m$ are utilized in the process, and each segment is treated as modality-agnostic, enhancing uncertainty to aid open-set detection. This contrasts with MMJP, which explicitly differentiates between modalities.

The modality codes are divided into O segments of equal length: $e^a = [e_1^a, e_2^a, \dots, e_O^a]$ and $e^b = [e_1^b, e_2^b, \dots, e_O^b]$. These segments are randomly selected across modalities to form $e^r = [e_1^m, e_2^m, \dots, e_O^m]$, where $m \in \{a, b\}$. One possible permutation is $\tilde{e}^o = [e_2^{m_2}, e_O^{m_n}, \dots, e_1^{m_1}]$. The O segments are subsequently shuffled to produce different permutations, yielding a total of $O!$ possible combinations. Among these, we randomly sample P permutations and assign each a unique index to serve as its label.

An auxiliary classification task is introduced for each sample instance, formulated as $\{(\tilde{e} \in \tilde{e}^o, o)\}_{o=1}^P$, where $\tilde{e} \in \tilde{e}^o$ denotes the recomposed embeddings, and $o \in \{1, \dots, P\}$ indicates the associated permutation index. The goal is to optimize the cross-modal jigsaw loss $L_{\text{cujp}}(\mathcal{H}(\tilde{e}), o)$, with \mathcal{H} being the classifier used for recognizing the permutation, and L_{cujp} denoting the conventional cross-entropy loss. Furthermore, as the combined feature dimension in CUJP matches that of a single modality, the number of required permutations is reduced, enhancing computational efficiency.

3.4. Final Loss

In addition to the previously mentioned losses, the following losses are also required:

$$\underbrace{\|\mathbf{x}_i^m - D(\hat{\mathbf{e}}_i^m)\|_2^2}_{L_{\text{recon}}} + \underbrace{\|\Phi^m(\mathbf{x}_i^m) - \text{sg}[\mathbf{e}]\|_2^2}_{L_{\text{commit}}} \quad (7)$$

Here, sg denotes the stop-gradient operation. The reconstruction loss, L_{recon} , measures the difference between the outputs of each modality projector Φ^m and the original inputs using Mean Squared Error (MSE). The commitment loss, L_{commit} , computes the MSE between the encoder results and their quantized codes. In this work, we replace the traditional VQ loss with Exponential Moving Average (EMA), as EMA offers greater robustness. The final loss is as follows, $\lambda_1, \lambda_2, \lambda_3, \lambda_4$ are hyperparameters:

$$L = \lambda_1(L_{\text{fine}} + L_{\text{coarse}}) + \lambda_2 L_{\text{cujp}} + \lambda_3 L_{\text{recon}} + \lambda_4 L_{\text{commit}} \quad (8)$$

4. Experiment

4.1. Experimental Setting

Pretrain: We use VGGsound-AVEL40K [5, 51] with text provided by [42] to train unified representation.

Downstream: We propose the OSCMG problem, which includes three tasks: classification on the AVE [36] and UCF [35] datasets, and a cross-dataset classification task between UCF and VGG [5] (UCF \leftrightarrow VGG). The AVE dataset originally contains 28 classes. We split the data based on the original labels into a 1:1 and 3:1 ratio, resulting in 14-class or 21-class training sets, which are then tested on the full 28 classes. For UCF, after filtering out classes without audio data from the original 101 classes, we obtained 51 classes. The data was split into training sets with either 17 or 34 classes in a 1:2 and 2:1 ratio, while testing was performed on the complete 51 classes. For UCF \leftrightarrow VGG, we filtered the labels to retain 16 common classes between UCF and VGG, splitting them into 1:1 and 3:1 ratios. This resulted in training sets with 8 or 12 classes, and testing was conducted on all 16 classes. It is important to note that some UCF classes do not have audio data, so in UCF \leftrightarrow VGG,

Dataset	Method	Split1						Split2					
		V→A			A→V			V→A			A→V		
		OS*	UNK	HOS	OS*	UNK	HOS	OS*	UNK	HOS	OS*	UNK	HOS
AVE	CODIS [12]	36.41	47.33	41.16	26.31	37.29	30.85	34.51	55.76	42.63	27.71	52.41	36.25
	TURN [49]	35.37	49.26	41.18	27.13	39.41	32.14	31.73	58.13	41.05	25.89	56.26	35.46
	CMCM [22]	39.09	53.48	45.17	30.21	45.93	36.45	34.51	62.86	44.56	30.78	61.31	40.98
	DCID [42]	45.29	59.78	51.54	34.98	42.46	38.36	41.14	68.60	51.44	34.18	67.44	45.37
	MICU	51.57	57.54	54.39	34.98	64.80	45.43	47.15	79.07	59.08	35.44	80.23	49.17
UCF	CODIS [12]	17.51	43.17	24.91	23.66	49.04	31.92	16.33	45.32	24.01	17.80	43.78	25.31
	TURN [49]	15.43	43.39	22.76	22.05	53.75	31.27	17.41	44.76	25.07	18.43	44.96	26.14
	CMCM [22]	21.41	50.09	30.00	25.38	51.63	34.03	18.78	46.72	26.79	21.67	47.87	29.83
	DCID [42]	25.08	55.06	34.46	29.62	53.35	38.09	18.52	58.97	28.18	25.83	48.28	33.65
	MICU	29.40	61.69	39.82	27.48	72.96	39.92	24.33	60.05	34.64	23.90	68.25	35.41
UCF(v)↔VGG(a)	CODIS [12]	62.75	75.35	68.48	43.61	63.71	51.78	47.71	79.16	59.54	41.61	72.14	52.78
	TURN [49]	59.73	78.52	67.85	41.52	64.40	50.49	51.31	75.53	61.11	40.73	75.62	52.94
	CMCM [22]	68.44	77.17	72.54	43.67	68.89	53.45	50.17	84.62	62.99	44.61	78.43	56.87
	DCID [42]	79.16	88.53	83.58	56.47	77.34	65.28	54.97	95.83	69.87	50.00	83.22	62.49
	MICU	81.72	93.23	87.09	68.71	70.70	69.69	66.77	87.18	75.62	47.43	86.13	61.17

Table 2. Comparison of our model with previous SOTA models on OSCMG. Split1 and Split2 represent different class partitioning schemes of the training set for each dataset, where Split1 corresponds to the scheme with fewer classes in the training set.

we only use UCF’s video modality (v) paired with VG-
GSound’s audio modality (a).

The CMG problem includes four tasks: cross-modal classification on AVE [36] and UCF↔VGG [5, 35], and cross-modal localization tasks on AVVP [37] and AVE→AVVP. Additionally, we conducted experiments on cross-modal zero-shot retrieval.

Evaluation Metrics: The evaluation metrics used in OSCMG are OS, UNK, and HOS, which have been widely adopted in prior open-set recognition works [2, 9, 21]. The HOS metric is calculated as $HOS = \frac{2 \times OS^* \times UNK}{OS^* + UNK}$, where OS* refers to the accuracy for known categories, and UNK corresponds to the accuracy for unknown categories. Unlike OS, HOS offers a more comprehensive performance measure by balancing results across known and unknown classes, which is crucial when accuracy for unknown classes is notably lower, underscoring the need for effective detection of unknown categories. For CMG, we employ different evaluation metrics depending on the task. Precision is used for classification tasks on AVE [36], VGG [50, 51], and UCF [35], while the F1-score is utilized for localization tasks on AVVP [37] and AVE→AVVP. For cross-modal zero-shot retrieval [4, 11], recall is the primary evaluation metric.

Implementation Details: We compare our model against several state-of-the-art methods in multimodal unified discrete representations and multimodal domain generalization, including CODIS [12], TURN [49], CMCM [22], and DCID [42]. These models are evaluated across our tasks and various downstream scenarios. For both L_{fine} and L_{coarse} , the temperature parameter τ is set to 1.0, the mask ratio of FCMI is set to 30%. All experiments, as shown in Tables 2, 3, 4, 7, 8, and Figures 6, 3, 4, use a codebook

size of 400 with an embedding dimension of 256. To ensure a fair comparison, all experiments, except those in Tables 5, 6, follow the same backbone settings as DCID [42]. However, since DCID employs relatively outdated backbones for video and audio, we introduce Swin-V2-L [24] and HTS-AT [6] as enhanced alternatives in Table 5 for video and audio, respectively. Additionally, in Table 6, we conduct new modality pairing experiments involving video, audio, and optical flow, where the backbones used are Swin-V2-L, HTS-AT, and SlowOnly [13], respectively. As the source dataset for the optical flow modality is not provided for both pretraining and downstream tasks, we use the TV-L1 [46] algorithm for optical flow extraction to ensure data consistency. $\lambda_1, \lambda_2, \lambda_3, \lambda_4$ set to 1, 2, 1, and 1, respectively.

4.2. Performance Analysis

In the tables below, **bold** numbers indicate the best results, V, A, T and F represent Video, Audio, Text, and Optical Flow, respectively.

Open-set Cross Modal Generalization: As shown in Table 2, we compared our proposed MICU model with the previous SOTA multimodal unified representation models on the newly introduced OSCMG task. It can be observed that MICU significantly outperforms the previous SOTA models in 11 of the most important HOS metrics. The only exception is the Split2 HOS metric for VGG(a)→UCF(v), where it ranks second with a value close to first place. This demonstrates the effectiveness of our proposed method on OSCMG, regardless of the dataset, its splits, or the cross-modal direction.

Cross Modal Generalization: To prove that our model excels not only on the newly proposed OSCMG task, but also on the well-established CMG task, we conducted a detailed

comparison with previous SOTA models. As shown in Table 3, MICU outperforms the previous models by a significant margin, with all 8 evaluation metrics showing clear and consistent improvements. The smallest observed improvement is as high as 2.0%, further underscoring the robustness and superior generalizability of our approach across a wide range of tasks.

Method	AVE		AVVP		AVE→AVVP		UCF(v)↔VGG(a)	
	V→A	A→V	V→A	A→V	V→A	A→V	V→A	A→V
CODIS [12]	36.8	39.7	32.7	32.6	40.8	40.6	50.8	45.2
TURN [49]	37.6	39.2	32.4	32.2	40.6	41.4	50.4	46.1
CMCM [22]	46.3	45.8	36.1	35.2	47.1	48.2	51.2	48.3
DCID [42]	54.1	55.0	40.4	40.8	53.0	52.4	67.1	60.6
MICU	56.1	57.1	45.2	48.2	56.3	54.9	75.3	64.5

Table 3. Comparison of our model with previous SOTA models on CMG.

Cross Modal Zero-shot Retrieval: As shown in Table 4, we also conducted Zero-shot Retrieval on two tasks, $V \leftrightarrow T$ and $A \leftrightarrow T$, to demonstrate that our model still maintains an advantage in the unified representation of other modalities.

Method	MSCOCO($V \leftrightarrow T$)			Clotho($A \leftrightarrow T$)		
	R@1	R@5	R@10	R@1	R@5	R@10
CMCM [22]	0.50	4.20	7.20	1.62	8.04	14.87
DCID [42]	0.80	5.00	8.30	2.06	9.00	16.70
MICU	1.30	5.00	8.80	2.44	10.96	18.95

Table 4. Comparison of our model with previous SOTA models on Zero-shot Retrieval.

Experiments with stronger backbones: As shown in Table 5, all models exhibit significant performance improvements with enhanced backbones. However, under the same backbone settings, our proposed MICU consistently maintains a clear advantage, further demonstrating the effectiveness of our approach.

Experiments with more modality combinations: As shown in Table 6, all experiments are conducted using enhanced backbones, where each value represents the average result of two generalization directions. For example,

Dataset	Method	Split1				Split2			
		Original		Enhanced		Original		Enhanced	
		V→A	A→V	V→A	A→V	V→A	A→V	V→A	A→V
AVE	CMCM	45.17	36.45	48.09	40.63	44.56	40.98	48.61	45.79
	DCID	51.54	38.36	54.03	41.65	51.44	45.37	54.97	50.27
	MICU	54.39	45.43	57.76	48.42	59.08	49.17	61.57	53.13
UCF	CMCM	30.00	34.03	32.72	35.62	26.79	29.83	30.25	32.13
	DCID	34.46	38.09	35.81	40.15	28.18	33.65	30.54	37.18
	MICU	39.82	39.92	41.27	42.31	34.64	35.41	36.83	39.01
UCF(vf)-VGG(a)	CMCM	72.54	53.45	76.42	60.20	62.99	56.87	66.36	59.98
	DCID	83.58	65.28	87.69	68.61	69.87	62.49	74.23	65.01
	MICU	87.09	69.69	90.62	74.64	75.62	61.17	77.10	65.74

Table 5. Comparison with previous SOTA methods on OSCMG, evaluated using HOS. Original and Enhanced refer to respective backbones in Implementation Details.

Dataset	Method	Split1			Split2		
		V↔A	V↔F	A↔F	V↔A	V↔F	A↔F
AVE	CMCM	43.73	45.16	40.98	47.61	47.97	34.62
	DCID	45.90	48.49	44.21	51.87	52.94	50.78
	MICU	52.15	54.28	52.09	56.96	58.63	55.43
UCF	CMCM	33.09	34.64	32.68	29.63	31.45	27.35
	DCID	37.16	38.47	35.26	33.15	36.34	33.81
	MICU	40.58	41.92	39.81	36.48	38.17	35.59
UCF(vf)-VGG(a)	CMCM	66.37	68.03	64.70	60.56	62.89	60.17
	DCID	78.41	79.24	78.05	66.40	69.45	65.89
	MICU	81.59	82.25	80.26	70.24	71.57	70.44

Table 6. Comparison with previous SOTA methods on OSCMG, evaluated using HOS. The experimental modalities include Video (V), Audio (A), and Optical Flow (F).

$V \leftrightarrow A$ denotes the mean of $V \rightarrow A$ and $A \rightarrow V$. Our method consistently maintains a clear advantage in tasks involving optical flow, demonstrating its adaptability beyond specific modality settings. Additionally, $V \leftrightarrow F$ achieves the best overall performance, likely due to the inherent similarity between video (V) and optical flow (F) modalities.

Jigsaw Puzzles: We conducted additional discussions on Jigsaw Puzzles, focusing on experiments with "without Jigsaw Puzzles," MMJP [9] using a 6-segment split, and our proposed CUJP with 2, 4, and 8-segment splits. The limitation on the number of segments is due to the unified representation features being 256-dimensional, so the number of splits must evenly divide 256, which leads CUJP to use 2, 4, and 8 splits. In contrast, MMJP requires the features of all three modalities to be split simultaneously, which results in a multiplication factor of 3. For instance, if each modality has 2 splits, MMJP will use 6 segments. However, if each modality has 4 splits, MMJP would require 12! factorial permutations, which our experiments showed resulted in excessively long computation times. Therefore, MMJP is limited to 6 segments in this study.

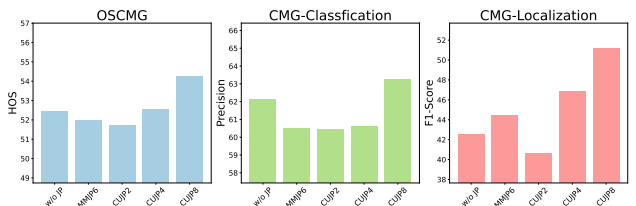


Figure 3. Experimental results of different Jigsaw Puzzles.

The specific experimental results are shown in Figure 3, where we separate the classification and localization tasks of CMG into two charts to display the model differences more clearly. It can be observed that MMJP6 performs worse than w/o jp in both OSCMG and CMG-Classification, showing improvements only in CMG-Localization. In contrast, CUJP’s performance improves as the number of splits increases, showing a clear upward

Dataset	L_{fine}	L_{coarse}	L_{cujp}	Split1						Split2					
				V→A			A→V			V→A			A→V		
				OS*	UNK	HOS	OS*	UNK	HOS	OS*	UNK	HOS	OS*	UNK	HOS
AVE	✓	-	-	8.07	10.61	9.17	7.17	13.97	9.48	4.43	26.74	7.60	5.06	16.28	7.72
	-	✓	-	45.29	54.75	49.57	29.15	65.92	40.42	38.61	83.72	52.85	26.90	81.40	40.43
	-	-	✓	7.17	28.49	11.46	7.62	35.75	12.57	5.38	19.77	8.46	5.38	36.05	9.36
	✓	✓	-	49.33	59.78	54.05	40.39	51.19	45.15	43.04	77.91	55.45	36.73	72.09	48.67
	✓	-	✓	7.17	12.23	41.34	0.90	99.44	1.78	5.38	17.44	8.22	5.70	1.16	1.93
	-	✓	✓	43.50	71.51	54.09	35.87	47.58	40.90	45.57	63.47	53.05	30.38	48.84	37.46
	✓	✓	✓	51.57	57.54	54.39	34.98	64.80	45.43	47.15	79.07	59.08	35.44	80.23	49.17
UCF	✓	-	-	4.28	23.54	7.24	7.08	4.12	5.21	2.16	20.74	3.92	3.53	0.82	1.32
	-	✓	-	24.86	62.11	35.51	30.76	48.86	37.75	20.25	56.75	29.85	23.86	56.57	33.56
	-	-	✓	5.85	20.96	9.15	6.12	24.59	9.80	2.90	18.43	5.00	3.53	18.98	5.95
	✓	✓	-	29.06	60.30	39.22	27.10	68.69	38.87	24.25	56.07	33.86	27.93	52.63	36.49
	✓	-	✓	8.39	20.08	11.83	6.95	8.98	7.83	3.96	20.79	6.65	2.66	18.93	4.67
	-	✓	✓	28.53	58.89	38.44	30.23	58.56	39.88	22.45	66.08	33.52	24.96	63.90	35.90
	✓	✓	✓	29.40	61.69	39.82	27.48	72.96	39.92	24.33	60.05	34.64	23.90	68.25	35.41
UCF(v)↔VGG(a)	✓	-	-	13.47	30.05	18.60	0.17	96.80	0.34	9.04	25.71	13.38	1.61	98.43	3.16
	-	✓	-	70.80	91.79	79.94	60.95	70.23	65.26	60.40	84.94	70.60	45.24	72.93	55.84
	-	-	✓	12.26	67.60	20.76	12.07	43.65	18.91	6.08	67.80	11.16	9.00	53.92	15.42
	✓	✓	-	79.16	82.05	80.58	65.26	70.70	67.87	58.39	68.34	62.97	51.03	63.76	56.69
	✓	-	✓	2.76	86.50	5.35	9.57	51.96	16.16	13.19	9.03	10.72	5.98	61.75	10.90
	-	✓	✓	75.86	86.83	80.98	65.26	73.12	68.97	63.59	81.16	71.31	54.88	69.13	61.19
	✓	✓	✓	81.72	93.23	87.09	68.71	70.70	69.69	66.77	87.18	75.62	47.43	86.13	61.17

Table 7. Ablation study of the three losses proposed by our model on OSCMG.

trend, with CUJP8 significantly outperforming all other configurations. Additionally, CUJP4 already consistently outperforms MMJP6, which demonstrates that for tasks related to multimodal unified representations, the CUJP setup is more suitable.

Ablation Study: Since L_{recon} and L_{commit} are standard losses for discrete representations and not the novelty of this paper, their effectiveness has been established in prior work. Therefore, our ablation study focuses on the newly proposed loss.

As shown in Table 7, we conducted a detailed ablation study on the three newly proposed losses in the MICU architecture, namely L_{fine} , L_{coarse} , and L_{cujp} . First, by observing the first three rows for each dataset, it is evident that L_{coarse} is the foundation of the model, as without it, a unified representation cannot be constructed. This is apparent because L_{coarse} represents contrastive learning of overall semantics, and without overall semantics, a representation space cannot be built. Next, comparing the 2nd and 4th rows, it can be observed that the combination of L_{coarse} and L_{fine} further improves the model’s performance, with noticeable gains in 11 HOS metrics. This indicates that L_{fine} provides fine-grained temporal knowledge that L_{coarse} alone cannot learn, helping the model construct a more refined representation space. Similarly, the comparison between the 2nd and 6th rows also shows improvements in 11 HOS metrics, indicating that L_{cujp} also helps build a better representation space, with the modality-agnostic Jigsaw Puzzles proving to be highly effective. The 5th row shows the same effect as the 2nd row, confirming that with-

out contrastive learning of overall semantics, a representation space cannot be constructed. The 7th row demonstrates that the combination of all three components achieves the optimal result.

Additional Experiments: Further experiments, including the mask setting of FCMI (Sec 6), codebook size hyperparameter selection (Sec 7), ablation study on CMG (Sec 8), computational efficiency analysis (Sec 9), and visualization of the discrete representation space (Sec 10), are provided in the supplementary material.

5. Conclusion

To advance the evaluation of multimodal unified representations in complex scenarios, we introduce the Open-set Cross-Modal Generalization (OSCMG) task, which specifically addresses the challenges of open-set detection and multimodal alignment. To tackle these challenges, we propose the MICU method, which integrates two key components: Fine-Coarse Masked Multimodal InfoNCE and Cross-Modal Unified Jigsaw Puzzle. These components offer complementary strategies, combining fine-grained masked contrastive learning with modality-agnostic self-supervised learning to enhance generalization and alignment across diverse modalities. Our approach achieves state-of-the-art performance on the OSCMG task and demonstrates significant improvements over previous models on the CMG task. Overall, we introduce a novel task to evaluate the performance of multimodal unified representations in open-set domains, and propose a new method to effectively address the challenges posed by this task.

Acknowledgments

This work was supported by National Key R&D Program of China (2022ZD0162000) and National Natural Science Foundation of China (62222211).

References

- [1] Alex Andonian, Shixing Chen, and Raffay Hamid. Robust cross-modal representation learning with progressive self-distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16430–16441, 2022. 3
- [2] Silvia Bucci, Mohammad Reza Loghmani, and Tatiana Tommasi. On the effectiveness of image rotation for open set domain adaptation. In *European conference on computer vision*, pages 422–438. Springer, 2020. 6
- [3] Fabio M Carlucci, Antonio D’Innocente, Silvia Bucci, Barbara Caputo, and Tatiana Tommasi. Domain generalization by solving jigsaw puzzles. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2229–2238, 2019. 2, 5
- [4] David Chen and William B Dolan. Collecting highly parallel data for paraphrase evaluation. In *Proceedings of the 49th annual meeting of the association for computational linguistics: human language technologies*, pages 190–200, 2011. 6
- [5] Honglie Chen, Weidi Xie, Andrea Vedaldi, and Andrew Zisserman. Vggsound: A large-scale audio-visual dataset. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 721–725. IEEE, 2020. 5, 6
- [6] Ke Chen, Xingjian Du, Bilei Zhu, Zejun Ma, Taylor Berg-Kirkpatrick, and Shlomo Dubnov. Hts-at: A hierarchical token-semantic audio transformer for sound classification and detection. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 646–650. IEEE, 2022. 6
- [7] Sihan Chen, Xingjian He, Longteng Guo, Xinxin Zhu, Weining Wang, Jinhui Tang, and Jing Liu. Valor: Vision-audio-language omni-perception pretraining model and dataset. *arXiv preprint arXiv:2304.08345*, 2023. 1
- [8] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. Uniter: Universal image-text representation learning. In *European conference on computer vision*, pages 104–120. Springer, 2020. 3
- [9] Hao Dong, Eleni Chatzi, and Olga Fink. Towards multimodal open-set domain generalization and adaptation through self-supervision. *arXiv preprint arXiv:2407.01518*, 2024. 2, 3, 5, 6, 7, 1
- [10] Hao Dong, Ismail Nejjar, Han Sun, Eleni Chatzi, and Olga Fink. Simmmdg: A simple and effective framework for multi-modal domain generalization. *Advances in Neural Information Processing Systems*, 36, 2024. 2, 3
- [11] Konstantinos Drossos, Samuel Lipping, and Tuomas Virtanen. Clotho: An audio captioning dataset. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 736–740. IEEE, 2020. 6
- [12] Jiali Duan, Liqun Chen, Son Tran, Jinyu Yang, Yi Xu, Belinda Zeng, and Trishul Chilimbi. Multi-modal alignment using representation codebook. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15651–15660, 2022. 1, 3, 6, 7
- [13] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6202–6211, 2019. 6
- [14] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario March, and Victor Lempitsky. Domain-adversarial training of neural networks. *Journal of machine learning research*, 17(59):1–35, 2016. 3
- [15] Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. Imagebind: One embedding space to bind them all. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15180–15190, 2023. 2
- [16] Hai Huang, Yan Xia, Shengpeng Ji, Shulei Wang, Hanting Wang, Minghui Fang, Jieming Zhu, Zhenhua Dong, Sashuai Zhou, and Zhou Zhao. Enhancing multimodal unified representations for cross modal generalization. *arXiv preprint arXiv:2403.05168*, 2024. 3
- [17] Hai Huang, Shulei Wang, and Yan Xia. Semantic residual for multimodal unified discrete representation. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2025.
- [18] Hai Huang, Yan Xia, Sashuai Zhou, Hanting Wang, Shulei Wang, and Zhou Zhao. Bridging domain generalization to multimodal domain generalization via unified representations. *arXiv preprint arXiv:2507.03304*, 2025. 3
- [19] Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy Hospedales. Learning to generalize: Meta-learning for domain generalization. In *Proceedings of the AAAI conference on artificial intelligence*, 2018. 3
- [20] Haoliang Li, YuFei Wang, Renjie Wan, Shiqi Wang, Tie-Qiang Li, and Alex Kot. Domain generalization for medical imaging classification with linear-dependency regularization. *Advances in neural information processing systems*, 33:3118–3129, 2020. 3
- [21] Wuyang Li, Jie Liu, Bo Han, and Yixuan Yuan. Adjustment and alignment for unbiased open set domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24110–24119, 2023. 6
- [22] Alexander H Liu, SouYoung Jin, Cheng-I Jeff Lai, Andrew Rouditchenko, Aude Oliva, and James Glass. Cross-modal discrete representation learning. *arXiv preprint arXiv:2106.05438*, 2021. 1, 2, 3, 6, 7
- [23] Quande Liu, Cheng Chen, Jing Qin, Qi Dou, and Pheng-Ann Heng. Feddg: Federated domain generalization on medical image segmentation via episodic learning in continuous frequency space. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1013–1023, 2021. 3

- [24] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021. 6
- [25] Jiaseen Lu, Christopher Clark, Rowan Zellers, Roozbeh Motlaghi, and Aniruddha Kembhavi. Unified-io: A unified model for vision, language, and multi-modal tasks. *arXiv preprint arXiv:2206.08916*, 2022. 1, 3
- [26] Mehdi Noroozi and Paolo Favaro. Unsupervised learning of visual representations by solving jigsaw puzzles. In *European conference on computer vision*, pages 69–84. Springer, 2016. 5
- [27] Xingang Pan, Ping Luo, Jianping Shi, and Xiaoou Tang. Two at once: Enhancing learning and generalization capacities via ibn-net. In *Proceedings of the european conference on computer vision (ECCV)*, pages 464–479, 2018. 3
- [28] Fabrizio Pedersoli, Dryden Wiebe, Amin Banitalebi, Yong Zhang, George Tzanetakis, and Kwang Moo Yi. Estimating visual information from audio through manifold learning. *arXiv preprint arXiv:2208.02337*, 2022. 3
- [29] Stavros Petridis, Themos Stafylakis, Pingchuan Ma, Georgios Tzimiropoulos, and Maja Pantic. Audio-visual speech recognition with a hybrid ctc/attention architecture. In *2018 IEEE Spoken Language Technology Workshop (SLT)*, pages 513–520. IEEE, 2018. 3
- [30] Mirco Planamente, Chiara Plizzari, Emanuele Alberti, and Barbara Caputo. Domain generalization through audio-visual relative norm alignment in first person action recognition. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 1807–1818, 2022. 3
- [31] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 2
- [32] Shyam Nandan Rai, Fabio Cermelli, Barbara Caputo, and Carlo Masone. Mask2anomaly: Mask transformer for universal open-set segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024. 4
- [33] Pritam Sarkar and Ali Etemad. Xkd: Cross-modal knowledge distillation with domain alignment for video representation learning. *arXiv preprint arXiv:2211.13929*, 2022. 1, 3
- [34] Yang Shu, Zhangjie Cao, Chenyu Wang, Jianmin Wang, and Mingsheng Long. Open domain generalization with domain-augmented meta-learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9624–9633, 2021. 2, 3
- [35] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012. 5, 6
- [36] Yapeng Tian, Jing Shi, Bochen Li, Zhiyao Duan, and Chenliang Xu. Audio-visual event localization in unconstrained videos. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 247–263, 2018. 5, 6
- [37] Yapeng Tian, Dingzeyu Li, and Chenliang Xu. Unified multisensory perception: Weakly-supervised audio-visual video parsing. In *Computer Vision—ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part III 16*, pages 436–454. Springer, 2020. 6
- [38] Josh Tobin, Rachel Fong, Alex Ray, Jonas Schneider, Wojciech Zaremba, and Pieter Abbeel. Domain randomization for transferring deep neural networks from simulation to the real world. In *2017 IEEE/RSJ international conference on intelligent robots and systems (IROS)*, pages 23–30. IEEE, 2017. 3
- [39] Eric Tzeng, Judy Hoffman, Ning Zhang, Kate Saenko, and Trevor Darrell. Deep domain confusion: Maximizing for domain invariance. *arXiv preprint arXiv:1412.3474*, 2014. 3
- [40] Teng Wang, Wenhao Jiang, Zhichao Lu, Feng Zheng, Ran Cheng, Chengguo Yin, and Ping Luo. Vlmixer: Unpaired vision-language pre-training via cross-modal cutmix. In *International Conference on Machine Learning*, pages 22680–22690. PMLR, 2022. 3
- [41] Xiran Wang, Jian Zhang, Lei Qi, and Yinghuan Shi. Generalizable decision boundaries: Dualistic meta-learning for open set domain generalization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11564–11573, 2023. 2
- [42] Yan Xia, Hai Huang, Jieming Zhu, and Zhou Zhao. Achieving cross modal generalization with multimodal unified representation. *Advances in Neural Information Processing Systems*, 36, 2024. 1, 2, 3, 5, 6, 7
- [43] WANG Xuesong, LI Yiran, and CHENG Yuhu. Hyper-spectral image classification based on unsupervised heterogeneous domain adaptation cyclegan. *Chinese Journal of Electronics*, 29(4):608–614, 2020. 3
- [44] Yifei Yang, ZhongXiang Zhou, Jun Wu, Yue Wang, and Rong Xiong. Class semantics modulation for open-set instance segmentation. *IEEE Robotics and Automation Letters*, 2024. 4
- [45] ZHANG Yun, WANG Nianbin, and CAI Shaobin. Learning domain-invariant and discriminative features for homogeneous unsupervised domain adaptation. *Chinese Journal of Electronics*, 29(6):1119–1125, 2020. 3
- [46] Christopher Zach, Thomas Pock, and Horst Bischof. A duality based approach for realtime tv-l1 optical flow. In *Pattern Recognition: 29th DAGM Symposium, Heidelberg, Germany, September 12-14, 2007. Proceedings 29*, pages 214–223. Springer, 2007. 6
- [47] H Zhang, M Cisse, Y Dauphin, and D Lopez-Paz. mixup: Beyond empirical risk management. In *6th Int. Conf. Learning Representations (ICLR)*, pages 1–13, 2018. 3
- [48] Ziang Zhang, Zehan Wang, Luping Liu, Rongjie Huang, Xize Cheng, Zhenhui Ye, Huadai Liu, Haifeng Huang, Yang Zhao, Tao Jin, et al. Extending multi-modal contrastive representations. *Advances in Neural Information Processing Systems*, 37:91880–91903, 2024. 3
- [49] Yang Zhao, Chen Zhang, Haifeng Huang, Haoyuan Li, and Zhou Zhao. Towards effective multi-modal interchanges in zero-resource sounding object localization. *Advances in Neural Information Processing Systems*, 35:38089–38102, 2022. 1, 2, 3, 6, 7

- [50] Jinxing Zhou, Liang Zheng, Yiran Zhong, Shijie Hao, and Meng Wang. Positive sample propagation along the audio-visual event line. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8436–8444, 2021. [6](#)
- [51] Jinxing Zhou, Dan Guo, and Meng Wang. Contrastive positive sample propagation along the audio-visual event line. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022. [5](#), [6](#)