Advances in Multimodal Adaptation and Generalization: From Traditional Approaches to Foundation Models

Hao Dong, Moru Liu, Kaiyang Zhou, Eleni Chatzi, Juho Kannala, Cyrill Stachniss and Olga Fink

Abstract—In real-world scenarios, achieving domain adaptation and generalization poses significant challenges, as models must adapt to or generalize across unknown target distributions. Extending these capabilities to unseen multimodal distributions, i.e., multimodal domain adaptation and generalization, is even more challenging due to the distinct characteristics of different modalities. Significant progress has been made over the years, with applications ranging from action recognition to semantic segmentation. Besides, the recent advent of large-scale pre-trained multimodal foundation models, such as CLIP, has inspired works leveraging these models to enhance adaptation and generalization performances or adapting them to downstream tasks. This survey provides the first comprehensive review of recent advances from traditional approaches to foundation models, covering: (1) Multimodal domain adaptation; (2) Multimodal test-time adaptation; (3) Multimodal domain generalization; (4) Domain adaptation and generalization with the help of multimodal foundation models; and (5) Adaptation of multimodal foundation models. For each topic, we formally define the problem and thoroughly review existing methods. Additionally, we analyze relevant datasets and applications, highlighting open challenges and potential future research directions. We maintain an active repository that contains up-to-date literature and supports research activities in these fields at https://github.com/donghao51/Awesome-Multimodal-Adaptation.

Index Terms—Domain generalization, Domain adaptation, Multimodal learning, Foundation models, Test-time adaptation

1 Introduction

Domain adaptation (DA) and domain generalization (DG) have garnered significant attention in the research community [1], [2]. In real-world applications such as robotics [3], [4], action recognition [5], and anomaly detection [6], [7], it is essential for models trained on limited source domains to perform well on novel target domains. To address distribution shift challenges, numerous DA and DG algorithms have been proposed, including distribution alignment [8], domain-invariant feature learning [9], feature disentanglement [10], data augmentation [11], and meta-learning [12]. However, most of these algorithms are designed for unimodal data, such as images or time series data. The emergence of large-scale multimodal datasets has underscored the need to address multimodal domain adaptation (MMDA) and generalization (MMDG) across multiple modalities, including audio-video [13], image-language [14], and LiDAR-camera [15]. Fig. 1 illustrates the distinction between unimodal and multimodal DA/DG, where MMDA and MMDG integrate information from multiple modalities to enhance generalization ability.

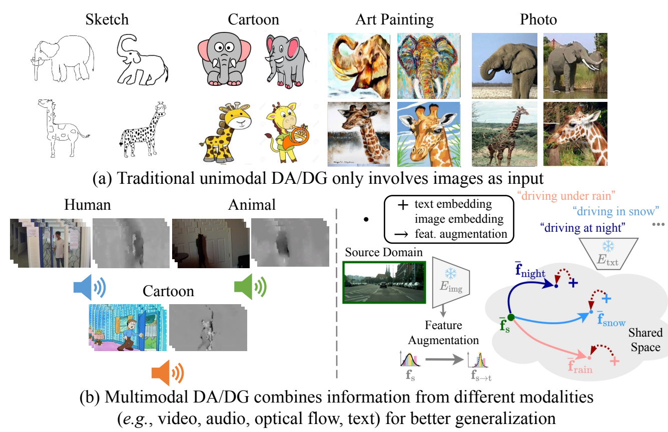


Fig. 1. The difference between unimodal and multimodal DA/DG.

In recent years, MMDA and MMDG have achieved significant progress in areas such as action recognition [16] and semantic segmentation [17]. A central challenge in MMDA and MMDG is effectively leveraging complementary information from diverse modalities to enhance generalization performance—an area where unimodal DA and DG approaches often fall short. For instance, the approach by Munro and Damen [16] incorporates the within-modal adversarial alignment alongside multimodal self-supervised alignment for MMDA. Multimodal test-time adaptation (MMTTA) [18] is a special form of MMDA that focuses on adapting a pre-trained source multimodal model to a target domain online, without accessing source domain data.

The advent of large-scale multimodal foundation models (MFMs), such as contrastive language–image pretraining (CLIP) [14] and stable diffusion [19], has opened new research directions for DA and DG. These efforts aim to enhance generalization capabilities using MFMs or adapt MFMs to downstream tasks. For example, Dunlap et al. [20]

- Hao Dong and Eleni Chatzi are with ETH Zürich, Switzerland.
- Moru Liu is with the Technical University of Munich, Germany.
- Kaiyang Zhou is with Hong Kong Baptist University, China.
- Juho Kannala is with Aalto University, Finland.
- Cyrill Stachniss is with the University of Bonn, Germany, the University of Oxford, UK, and the Lamarr Institute for Machine Learning and Artificial Intelligence, Germany.
- Olga Fink is with EPFL, Switzerland.

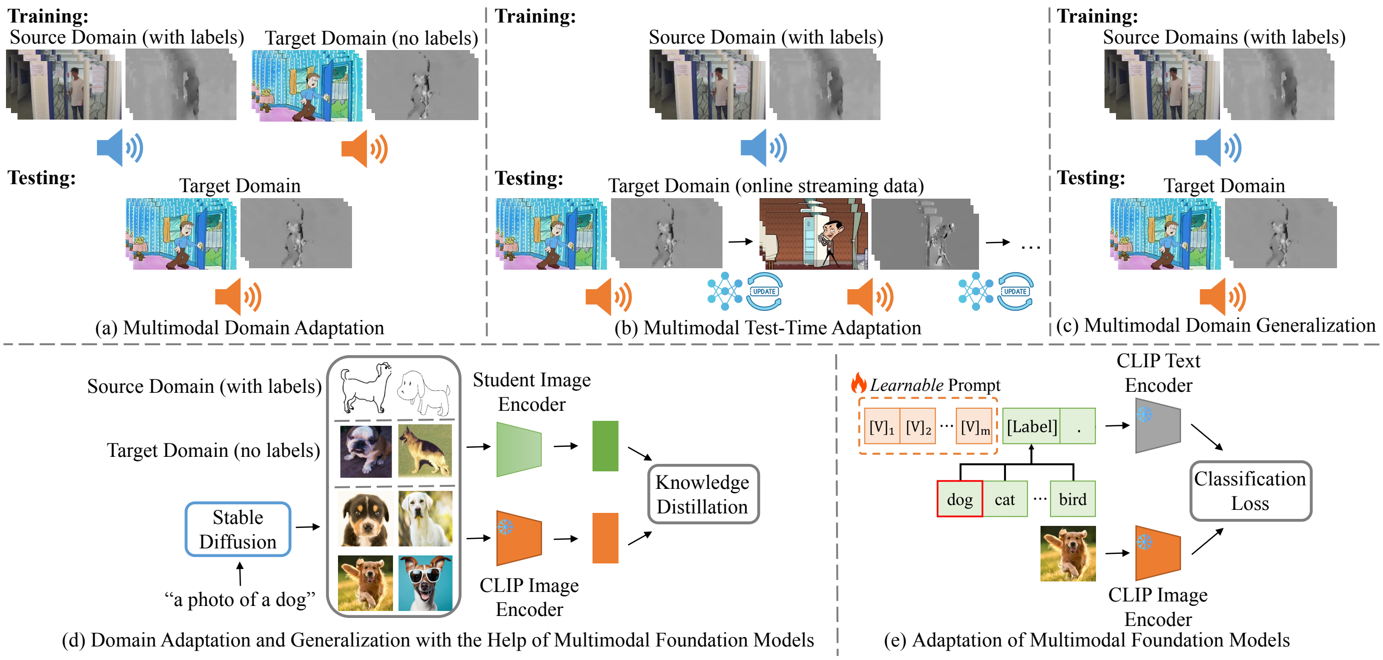


Fig. 2. Five multimodal adaptation scenarios are discussed in this survey. (a) Multimodal domain adaptation, (b) Multimodal test-time adaptation, and (c) Multimodal domain generalization, which represent traditional multimodal settings with varying access to source and target domain data. Additionally, we examine more recent foundation models including (d) unimodal domain adaptation and generalization assisted by multimodal foundation models, and (e) the adaptation of multimodal foundation models to downstream tasks.

extend image embeddings to unseen domains using language, while Huang et al. [21] distill CLIP's knowledge into a smaller student model for domain generalization. Additionally, Zhou et al. [22] adapt CLIP-like vision-language models (VLMs) for downstream image recognition by modeling a prompt's context words with learnable vectors.

Despite significant progress made in the field recently, there is no comprehensive survey that summarizes the main ideas of multimodal adaptation and generalization. This survey paper aims to provide a detailed literature review of algorithms developed over the last decade and to offer insights into potential future research directions. This paper covers five adaptation scenarios (Fig. 2 and Fig. 3) and is organized as follows. Sec. 2 discusses related research areas. Sec. 3 introduces the multimodal domain adaptation problem and highlights major solutions for action recognition and semantic segmentation. Sec. 4 and Sec. 5 present representative methods for multimodal test-time adaptation and domain generalization, respectively. Sec. 6 explores how multimodal foundation models can help improve DA and DG. Sec. 7 reviews popular methods for adapting MFMs to downstream tasks. Sec. 8 summarizes major applications and datasets. Finally, we outline potential future directions in Sec. 9 and conclude the paper in Sec. 10.

Comparison with previous surveys. While our survey contributes to the broader areas of DA and DG, which have been reviewed in prior works [1], [2], our specific focus is on multimodal adaptation and generalization, i.e. methods that involve multiple modalities. The survey by Zhang et al. [23] only covers an overview of the adaptation of VLMs before 2023. In contrast, we unify the discussion of traditional approaches for MMDA, MMTTA, and MMDG, the role of advanced MFMs in enhancing DA and DG, as well as the recent methods for adapting MFMs to downstream tasks.

2 Related Research Topics

2.1 Domain Adaptation

Domain adaptation seeks to enhance model performance in the target domain by leveraging labeled source data and unlabeled target data [1]. Traditional DA methods focus on unimodal scenarios with images as the major input. Common approaches include aligning feature distributions using discrepancy metrics [8], employing adversarial learning in input or feature spaces [131], [132], and utilizing reconstruction-based methods [133]. Additionally, techniques such as data augmentation [11] and self-training [134], have also been extensively explored. Depending on assumptions about label set relationships between source and target domains, DA is further categorized into partial-set [135], open-set [136], and universal DA [137].

2.2 Domain Generalization

Domain generalization aims to generalize models to unseen target domains without accessing target data during training. DG methods can be broadly grouped into data manipulation, representation learning, and learning strategies [2]. Data manipulation methods, such as [138], enhance data diversity, while representation learning approaches [139] focus on extracting domain-invariant features. Additionally, learning strategies like meta-learning [12] and self-supervised learning [140] have demonstrated improved generalization performance across domains. Shu et al. [141] also address the open-set DG problem where the target domain has private classes.

2.3 Test-time Adaptation

Test-time adaptation (TTA) seeks to adapt a pre-trained model on the source domain online, addressing distribution

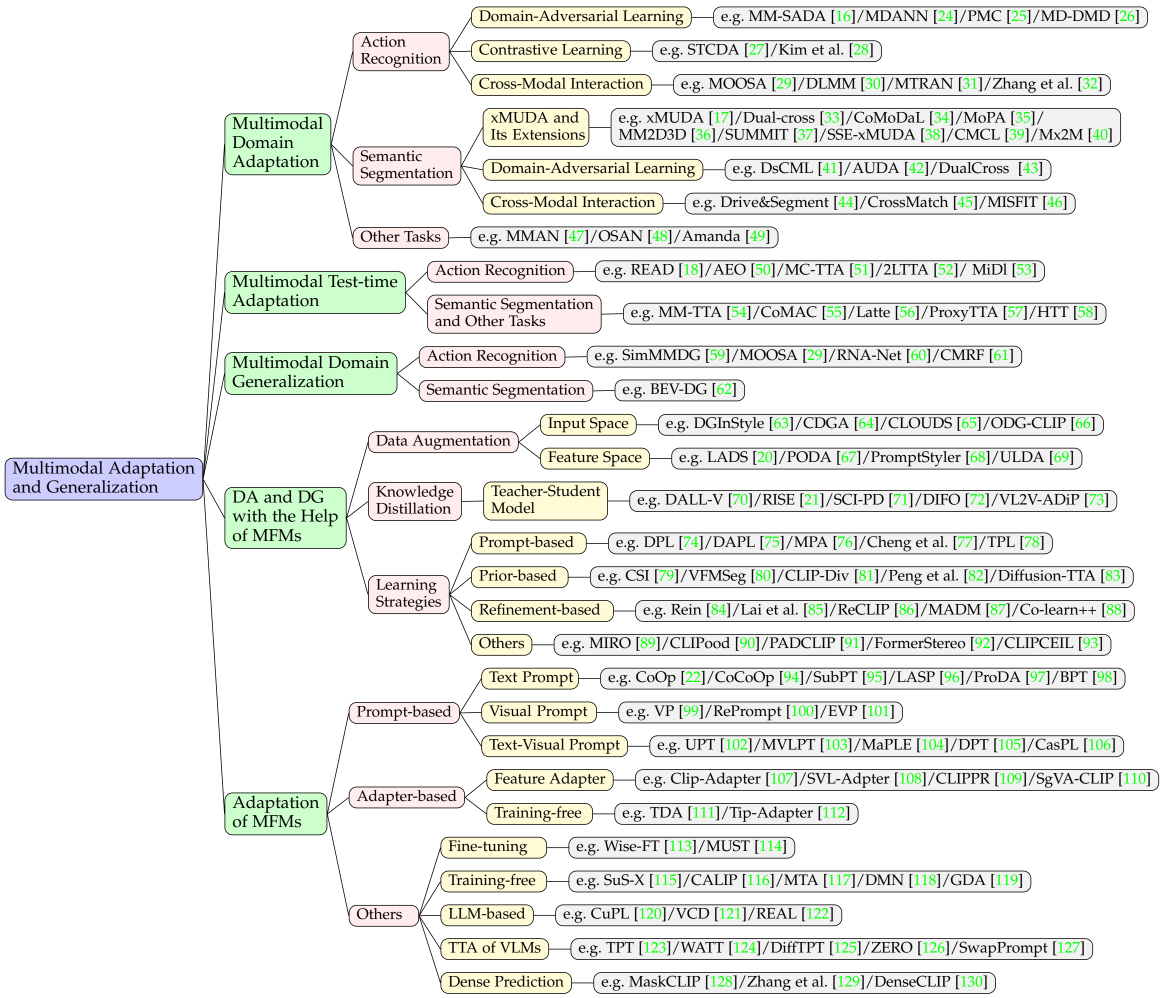


Fig. 3. Taxonomy of multimodal adaptation and generalization methods.

shifts without requiring access to either source data or target labels. Online TTA methods [142], [143] update specific model parameters using incoming test samples based on unsupervised objectives such as entropy minimization and pseudo-labels. Robust TTA methods [144], [145] address more complex and practical scenarios, including label shifts, single-sample adaptation, and mixed domain shifts. Continual TTA approaches [146], [147] target the continual and evolving distribution shifts encountered over test time. For further information, we refer the reader to [148], [149].

2.4 Multimodal Learning

Multimodal learning leverages the complementary strengths of diverse modalities to enhance representation learning and contextual understanding. Prominent multimodal learning directions can be categorized into multimodal representation learning [150], [151], fusion methods [152], [153], alignment [154], [155], etc. For further information, we refer the reader to [156], [157].

2.5 Self-supervised Learning

Self-supervised learning (SSL) aims to learn from unlabeled data by obtaining supervision signals from pretext tasks, such as predicting transformations [158], [159], reconstructing missing components [160], [161], or optimizing contrastive objectives [162], [163]. By capturing intrinsic data structures, SSL enables learning robust and domain-invariant representations, making it an essential component for DA and DG. In the multimodal context, SSL is also exploited through tasks such as multimodal alignment [164], cross-modal translation [165], and relative norm alignment [166]. These pretext tasks have been effectively integrated into MMDA and MMDG frameworks, including recent methods such as [16], [29]. For further information on SSL, we refer the reader to existing literature [167], [168].

2.6 Foundation Models

Foundation models are large-scale models pre-trained on vast amounts of datasets to serve as versatile starting points for a wide range of downstream tasks. These models exhibit strong generalization capabilities, allowing them to adapt to various applications with minimal fine-tuning. Prominent examples include language models like GPT [169], vision models like SAM [170] and DINO [171], vision-language models like CLIP [14] and Flamingo [172], and generative models like stable diffusion [19]. For further information on foundation models, we refer the reader to [173].

3 Multimodal Domain Adaptation

Multimodal domain adaptation (MMDA) aims at adapting a model trained on a source domain to perform effectively on a target domain while leveraging multiple modalities of data (e.g., video, audio, and optical flow). MMDA utilizes both labeled data from the source domain and unlabeled data from the target domain during adaptation.

3.1 Problem Definition

In MMDA, we have a labeled source domain $\mathcal{D}_{src}$ and an unlabeled target domain $\mathcal{D}_{target}$, where $\mathcal{D}_{src} = \{(\mathbf{x}_j^s, y_j)\}_{j=1}^{n_s}$ represents the source domain with n_s labeled data instances, and $\mathcal{D}_{target} = \{\mathbf{x}_j^t\}_{j=1}^{n_t}$ denotes the target domain with n_t data instances. Each data instance $\mathbf{x}_j$ in source and target domain is composed of M different modalities, expressed as $\mathbf{x}_j = \{(\mathbf{x}_j)_m \mid m = 1, \ldots, M\}$. Labels for both domains are given as $y_j \in \mathcal{Y} \subset \mathbb{R}$, while labels for the target domain are unavailable during training. The joint distributions of inputs and labels differ across source and target domains, i.e., $P_{XY}^{target} \neq P_{XY}^{src}$. The goal of MMDA is to learn a robust predictive function $f : \mathbf{X} \to \mathcal{Y}$ on $\mathcal{D}_{src}$ and $\mathcal{D}_{target}$ that minimizes the prediction error on the unlabeled target domain $\mathcal{D}_{target}$ under domain shift scenarios:

$$f = \arg\min_{f} \mathbb{E}_{(\mathbf{x},y)\in\mathcal{D}_{target}}[\ell(f(\mathbf{x}), y)], \tag{1}$$

where $\mathbb{E}$ denotes the expectation, and $\ell(\cdot,\cdot)$ is the loss function. Existing research on MMDA primarily focuses on two tasks—the action recognition task with video, audio, and optical flow modalities, and the semantic segmentation task with LiDAR point cloud and RGB images. We discuss them separately in the following sections.

3.2 MMDA for Action Recognition

In this section, we introduce existing MMDA methods for action recognition in detail and categorize them into domain-adversarial learning, contrastive learning, and cross-modal interaction.

3.2.1 Domain-Adversarial Learning

Adversarial learning-based approaches effectively align multimodal features across domains by leveraging adversarial objectives [174] to learn domain-invariant representations. For example, Qi et al. [24] leverage an adversarial objective to jointly attend and fuse multimodal representation to learn domain-invariant features across modalities. MM-SADA [16] incorporates within-modal adversarial alignment alongside multimodal self-supervised alignment for MMDA, as shown in Fig. 4. Given a binary domain label, d, indicating if an example $x \in \mathcal{D}_{src}$ or $x \in \mathcal{D}_{target}$, the domain discriminator for modality m is defined as:

$$\mathcal{L}_d^m = \sum_{x\in\{\mathcal{D}_{src},\mathcal{D}_{target}\}} -d\log(D^m(F^m(x))) - (1-d)\log(1-D^m(F^m(x))), \tag{2}$$

where D^m is the domain discriminator for modality m and F^m is the feature extractor. The multimodal self-supervised alignment loss aims to learn the temporal correspondence between modalities and is defined as:

$$\mathcal{L}_c = \sum_{x\in\{\mathcal{D}_{src},\mathcal{D}_{target}\}} -c\log C(F^1(x), ..., F^M(x)), \tag{3}$$

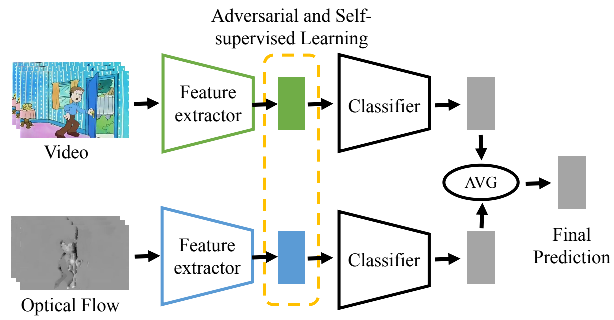


Fig. 4. The MM-SADA [16] architecture for MMDA, where adversarial and self-supervised learning are used to align multimodal features.

where C is the self-supervised correspondence classifier head and c is a binary label defining if modalities correspond. Moving beyond simple alignment, Zhang et al. [25] enhance cross-modal collaboration by selecting reliable pseudo-labeled target samples while also addressing missing modality scenarios—where adversarial learning is leveraged to generate absent modalities while preserving semantic integrity. Yin et al. [26] further extend adversarial learning to temporal sequences, using mix-sample adversarial learning to capture domain-invariant temporal dependencies while dynamically distilling knowledge across modalities to boost adaptability.

3.2.2 Contrastive Learning

Contrastive learning [175] is a powerful technique for learning transferable representations by pulling positive pairs closer in feature space while pushing negative pairs apart. In MMDA, it helps align features across both domains and modalities. For instance, Song et al. [27] jointly align clip- and video-level features using self-supervised contrastive learning while minimizing video-level domain discrepancy, enhancing category-aware alignment and cross-domain generalization. Kim et al. [28] leverages contrastive learning with modality- and domain-specific sampling strategies to jointly regularize cross-modal and cross-domain feature representations.

3.2.3 Cross-Modal Interaction

Cross-modal interaction methods enhance multimodal feature learning by fostering information exchange between modalities during adaptation, enabling models to capture complementary and interdependent relationships across modalities. For instance, Lv et al. [30] model modality-specific classifiers as teacher-student sub-models, using prototype-based reliability measurement for adaptive teaching and asynchronous curriculum learning, and employing reliability-aware fusion for robust final decisions. Huang et al. [31] address source-free MMDA by leveraging self-entropy-guided Mixup [11] to generate synthetic samples and aligning them with hypothetical source-like samples using multimodal and temporal relative alignment. Zhang et al. [32] take a different perspective by proposing an audio-adaptive encoder and an audio-infused recognizer to tackle domain shifts in action recognition across scenes, viewpoints, and actors. By leveraging domain-invariant audio activity information, they refine visual representations through absent activity learning and enhance silent task

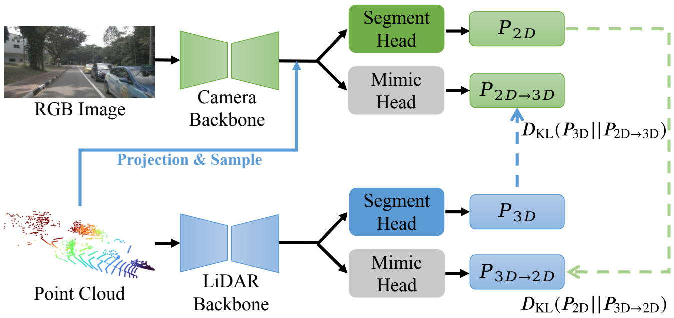


Fig. 5. The xMUDA [17] architecture for MMDA, which promotes cross-modal prediction consistency through multi-head mutual mimicking.

recognition with visual cues. Yang et al. [176] demonstrate that enhancing the transferability of each modality through cross-modal interaction before performing cross-domain alignment is more effective than directly aligning the multimodal inputs. Recently, Dong et al. [29] tackle MMDA under an open-set setup by designing two self-supervised tasks—masked cross-modal translation and multimodal Jigsaw puzzles—to learn robust multimodal features for generalization and open-class detection, along with an entropy weighting mechanism to balance modality-specific losses.

3.3 MMDA for Semantic Segmentation

In this section, we introduce existing MMDA methods for semantic segmentation in detail and categorize them into xMUDA and its extensions, domain-adversarial learning, and cross-modal interaction.

3.3.1 xMUDA and Its Extensions

Jaritz et al. [17] introduces the first MMDA framework named xMUDA for 3D semantic segmentation (3DSS), promoting cross-modal prediction consistency through multi-head mutual mimicking (Fig. 5). An unsupervised cross-modal divergence loss is applied to both the source and target domains to ensure effective cross-modal alignment:

$$\begin{aligned}\mathcal{L}_{xM} &= D_{KL}(P_x^{(n,c)} \| Q_x^{(n,c)}) \\ &= -\frac{1}{N}\sum_{n=1}^{N}\sum_{c=1}^{C} P_x^{(n,c)} \log \frac{P_x^{(n,c)}}{Q_x^{(n,c)}},\end{aligned} \tag{4}$$

where $(P, Q) \in \{(P_{2D}, P_{3D\to2D}), (P_{3D}, P_{2D\to3D})\}$, N the number of 3D points and C is the number of classes. Here, P denotes the target distribution from the main prediction, while Q represents the mimicking prediction used to approximate P. xMUDA also has its variant xMUDA$_{\mathrm{PL}}$, which incorporates pseudo-labels for self-training and serves as a strong baseline for MMDA. As a pioneering work, xMUDA also establishes a new benchmark using nuScenes [177], A2D2 [178], and SemanticKITTI [179] datasets with three adaptation scenarios: day-to-night, country-to-country, and dataset-to-dataset. Many follow-up works extend xMUDA from different perspectives.

Extension via Data Augmentation. Data augmentation techniques have been explored to enhance cross-modal alignment in xMUDA. For example, Li et al. [33] propose a multimodal style transfer strategy and a target-aware teacher framework to perform cross-domain and cross-modal knowledge distillation on source and synthesized target-style data. Complementing this, Chen et al. [34] employ CutMix [180] and Mix3D [181] to augment 2D and 3D training data, facilitating 2D-3D interaction and intra-domain cross-modal learning. Recently, Cao et al. [35] integrate xMUDA's pipeline with 3D rare objects collected from real-world scenarios and pixel-wise supervision from the SAM [170] model, addressing imbalanced supervision and significantly improving rare object segmentation.

Extension via Fusion. Beyond augmentation, fusion-based strategies refine xMUDA by improving information exchange between modalities. For instance, Wu et al. [182] perform cross-modal and cross-domain alignments through knowledge distillation using fused cross-modal representations, maximizing correlation and complementarity between heterogeneous modalities to mitigate domain shift. Cardace et al. [36] further strengthen fusion by feeding depth features into the 2D branch and dynamically enriching the 3D network with RGB features. Through middle fusion across both branches, intrinsic cross-modal complementarity is effectively exploited. Simons et al. [37] take a different approach, incorporating dynamic selection of fused and unfused rectified pseudo-labels for self-training to address source-free MMDA for 3DSS.

Extension via Cross-modal Interaction. Zhang et al. [38] present plane-to-spatial and discrete-to-textured self-supervised tasks to train the model under a mixed domain setting to enhance modality-specific learning and mitigate domain shift. Xing et al. [39] enhance xMUDA with cross-modal contrastive learning and a neighborhood feature aggregation module, strengthening 2D-3D consistency across domains while capturing richer context information. Zhang et al. [40] take this further by incorporating masked cross-modal modeling to mitigate large domain gaps and introducing dynamic cross-modal filters for feature matching, enabling the method to dynamically leverage more suitable 2D-3D complementarity and improve overall adaptability.

3.3.2 Domain-Adversarial Learning

Similar to the action recognition task, domain-adversarial learning methods for 3DSS leverage adversarial objectives to learn domain-invariant representations. For example, Peng et al. [41] introduce sparse-to-dense feature alignment for intra-domain point-pixel correspondence and adversarial learning across domains and modalities for inter-domain alignment. In contrast, Liu et al. [42] focus adversarial learning on the image modality and propose a threshold-moving strategy to mitigate data imbalance during inference. Other than pure adversarial alignment, Man et al. [43] introduce a distillation framework that transfers knowledge from a LiDAR teacher model to a camera student model via feature supervision on depth estimation and BEV embeddings. Multi-stage adversarial learning further aligns feature spaces across domains.

3.3.3 Cross-Modal Interaction

To foster cross-modal interaction, Vobecky et al. [44] introduce a cross-modal unsupervised approach for 2D semantic segmentation (2DSS) using unannotated paired LiDAR and camera data. It first extracts 3D-consistent object segments based on their geometrical properties and applies projection and clustering to generate 2D pseudo-ground truth, enabling knowledge distillation with cross-modal spatial constraints. Yin et al. [45] address source-free MMDA for 2DSS by integrating a multimodal auxiliary network during training. They employ middle fusion and enforce prediction consistency between augmented depth-RGB pairs to enable cross-modal learning. Rizzoli et al. [46] further enhance multimodal learning by integrating depth data into a vision transformer at the input, feature, and output stages. Color and depth style transfer enable early domain alignment, while cross-modal self-attention generates mixed features for better semantic extraction. Bultmann et al. [183] enable real-time semantic inference and fusion of LiDAR, RGB, and thermal sensor modalities for semantic segmentation and object detection, using a late fusion approach and label propagation to adapt across sensors and domains.

3.4 MMDA for Other Tasks

Beyond action recognition and semantic segmentation, MMDA has been explored in diverse tasks. Ma et al. [47] address MMDA for cross-domain object and event recognition tasks by using stacked attention to learn semantic representations and applying multi-channel constraints to enhance category discrimination. Liu et al. [48] introduce a tensor-based alignment module to explore the relationship between domains and modalities and a dynamic domain generator to create transitional samples, achieving superior performance in multimodal sentiment analysis and video text classification tasks. More recently, Zhang et al. [49] addressed MMDA for emotion recognition by independently learning optimal representations for each modality and adaptively balancing domain alignment across modalities through dynamic weighting.

4 Multimodal Test-time Adaptation

In contrast to multimodal domain adaptation where both source and target domain data are available during adaptation, multimodal test-time adaptation (MMTTA) aims to adapt a pre-trained source model online to a target domain without accessing data from the source domain.

4.1 Problem Definition

Let $\mathcal{D}_{src} = \{(\mathbf{x}_j^s, y_j)\}_{j=1}^{n_s}$ represent the source domain dataset which follows the distribution P_{XY}^{src} and each sample $\mathbf{x}_j$ consists of M modalities, denoted as $\mathbf{x}_j = \{x_j^m \mid m = 1, \cdots, M\}$. Similarly, let $\mathcal{D}_{target} = \{\mathbf{x}_j^t\}_{j=1}^{n_t}$ represent the target domain dataset with distribution P_{XY}^{target}. The label spaces for both domains are given as $y_j \in \mathcal{Y} \subset \mathbb{R}$. Let $f : \mathbf{X} \rightarrow \mathcal{Y}$ denote a neural network trained on the source distribution P_{XY}^{src}. In MMTTA, f consists of M feature extractors $g_m(\cdot)$ and a classifier $h(\cdot)$. Each feature extractor $g_m(\cdot)$ processes modality m to produce an embedding $\mathbf{E}^m$, and the classifier $h(\cdot)$ combines these embeddings to generate a prediction probability $\hat{p}$:

$$\begin{aligned} \hat{p} = \delta(f(\mathbf{x})) &= \delta(h([g_1(x^1), ..., g_M(x^M)])) \\ &= \delta(h([\mathbf{E}^1, ..., \mathbf{E}^M])). \end{aligned} \tag{5}$$

where $\delta(\cdot)$ denotes the softmax function. Given a well-trained multimodal source model $f(\mathbf{x})$ on $\mathcal{D}_{src}$, MMTTA aims to adapt this model online to the unlabeled target domain $\mathcal{D}_{target}$, where $P_{XY}^{target} \neq P_{XY}^{src}$.

4.2 Methods for Multimodal Test-time Adaptation

MMTTA is a relatively new research direction and few works have been proposed to address this challenging problem. Existing research on MMTTA primarily focuses on action recognition, semantic segmentation, and other tasks.

4.2.1 MMTTA for Action Recognition

READ by Yang et al. [18] addresses MMTTA under reliability bias, where the information discrepancies across different modalities are derived from intra-modal distribution shifts. Instead of updating the normalization statistics and transformation parameters in the batch normalization layer as in previous TTA methods [142], [144], READ proposes to modulate the attention between modalities in a self-adaptive way for reliable fusion. Besides, READ also introduces a novel confidence-aware loss function $\mathcal{L}_{ra}$ for robust multimodal adaptation:

$$\mathcal{L}_{ra} = \frac{1}{B}\sum_{i=1}^{B} p_i \log \frac{e\gamma}{p_i}, \tag{6}$$

where B is the batch size, p_i is the confidence of the prediction $\hat{p}_i$, i.e., $p_i = \max(\hat{p}_i)$, and γ is a threshold for confident prediction. $\mathcal{L}_{ra}$ helps the model focus more on the high-confident prediction while preventing the noise from the low-confident predictions. Xiong et al. [51] use teacher-student memory banks and self-assembled source-friendly feature reconstruction to align multimodal prototypes and reduce domain shift during MMTTA. Lei et al. [52] adopt a two-level objective function including Shannon entropy loss and a diversity-promoting loss to consider both intra-modal distribution shift and cross-modal reliability bias in the modality fusion block. Beyond domain shifts, Dong et al. [50] extend MMTTA to the open-set setting, where unknown categories appear during test-time adaptation. They propose adaptive entropy-aware optimization (AEO), which effectively amplifies the entropy difference between known and unknown samples during online adaptation. AEO consists of two primary components: unknown-aware adaptive entropy optimization and adaptive modality prediction discrepancy optimization. The unknown-aware adaptive entropy optimization module adaptively weights and optimizes each sample based on its prediction uncertainty and is defined as:

$$W_{ada} = \tanh(\beta \cdot (H(\hat{p}) - \alpha)), \tag{7}$$

$$\mathcal{L}_{AdaEnt} = -H(\hat{p})W_{ada}, \tag{8}$$

where $\tanh$ is the hyperbolic tangent function, W_{ada} is the adaptive weight assigned to each sample, $H(\hat{p})$ is the normalized entropy of prediction $\hat{p}$, computed as $H(\hat{p}) = -(\sum_c \hat{p}_c \log \hat{p}_c)/\log(C)$, with C being the number of classes. The adaptive modality prediction discrepancy optimization module optimizes the prediction discrepancy across modalities and is defined as:

$$\mathcal{L}_{AdaDis} = -(Dis(\hat{p}^1, \hat{p}^2))W_{ada}, \tag{9}$$

where W_{ada} is the adaptive weight calculated in Eq. (7). The effectiveness and versatility of AEO are also validated through extensive experiments in challenging long-term and continual adaptation scenarios.

4.2.2 MMTTA for Semantic Segmentation

Beyond the action recognition task, MMTTA has also been applied to 3D semantic segmentation and other tasks. For example, Shin et al. [54] introduce the intra-modal pseudo-label Generation module, which generates pseudo-labels for each modality independently, and the inter-modal pseudo-label refinement module, which adaptively selects pseudo-labels across modalities. Building on this, Cao et al. [55] explore multi-modal continual test-time adaptation, addressing dynamic target domains over time. It facilitates dynamic adaptation for 3D semantic segmentation by attending to reliable modalities and mitigating catastrophic forgetting through dual-stage mechanisms and class-wise momentum queues designed for continual domain shifts. Recently, Cao et al. [56] further enhanced 3D segmentation by leveraging reliable spatial-temporal correspondences, filtering unstable predictions, and employing cross-modal learning to maintain consistency across consecutive frames.

4.2.3 MMTTA for Other Tasks

Park et al. [57] focuses on MMTTA for depth completion using a single image and an associated sparse depth map. It reduces the domain gap by employing a source-trained embedding module that aligns image and sparse depth features from the target domain in a single pass. Besides, Wang et al. [58] tackles MMTTA for person re-identification, enhancing model generalization by leveraging relationships among heterogeneous modalities. It incorporates a cross-identity inter-modal margin loss and employs a multimodal test-time training strategy for self-supervised adaptation.

5 Multimodal Domain Generalization

In contrast to multimodal domain adaptation and test-time adaptation, multimodal domain generalization (MMDG) presents a more challenging problem setting. In MMDG, the model is trained only on source domains with multiple modalities to generalize across unseen domains, without prior exposure to target domain data during training.

5.1 Problem Definition

In MMDG, we are given D source domains $\mathcal{D}_{src} = \{\mathcal{D}^i \mid i = 1, \cdots, D\}$, where $\mathcal{D}^i = \{(\mathbf{x}_j^i, y_j^i)\}_{j=1}^{n_i}$ denotes the i-th domain with n_i data instances. Each data instance $\mathbf{x}_j^i = \{(\mathbf{x}_j^i)_m \mid m = 1, \cdots, M\} \in \mathbf{X}$ is comprised of M different modalities and $y_j^i \in \mathcal{Y} \subset \mathbb{R}$ denotes the label. The joint distributions between each pair of domains are different: $P_{XY}^i \neq P_{XY}^j, 1 \leq i \neq j \leq D$. The goal of MMDG is to learn a robust and generalizable predictive function $f : \mathbf{X} \rightarrow \mathcal{Y}$ from D source domains $\mathcal{D}_{src}$ and M data modalities to achieve a minimum prediction error on a target domain $\mathcal{D}_{target}$ (*i.e.*, $\mathcal{D}_{target}$ cannot be accessed during training and $P_{XY}^{target} \neq P_{XY}^i$ for $i \in \{1, \cdots, D\}$):

$$f = \arg\min_f \mathbb{E}_{(\mathbf{x},y)\in\mathcal{D}_{target}}[\ell(f(\mathbf{x}), y)], \tag{10}$$

where $\mathbb{E}$ is the expectation and $\ell(\cdot,\cdot)$ is the loss function. The f in MMDG comprises of M feature extractors $g_m(\cdot)$ and a classifier $h(\cdot)$. Each feature extractor $g_m(\cdot)$ extracts an embedding $\mathbf{E}^m$ for its corresponding modality m, and

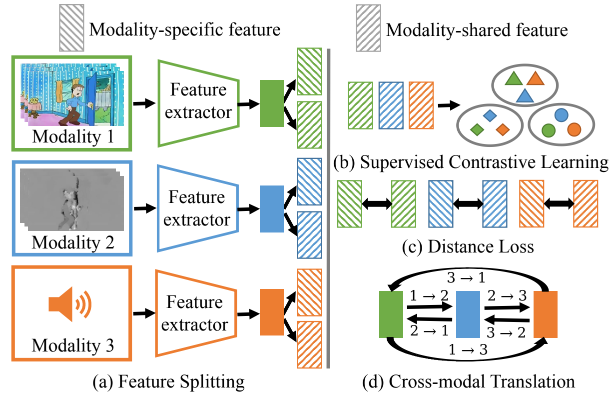


Fig. 6. The SimMMDG [59] architecture for MMDG, which separates features into modality-specific and modality-shared components for better generalization.

the classifier $h(\cdot)$ takes the combined embeddings from all modalities as input and outputs a prediction probability $\hat{p}$:

$$\begin{aligned}\hat{p} = \delta(f(\mathbf{x})) &= \delta(h([g_1(x^1), ..., g_M(x^M)])) \\ &= \delta(h([\mathbf{E}^1, ..., \mathbf{E}^M])).\end{aligned} \tag{11}$$

where $\delta(\cdot)$ is the softmax function.

5.2 Methods for Multimodal Domain Generalization

Similar to MMTTA, MMDG is also a relatively new research direction and few works have been proposed to address this challenging problem.

5.2.1 MMDG for Action Recognition

Planamente et al. [60] introduce the first MMDG approach for egocentric activity recognition by proposing the relative norm alignment loss, which aligns the mean feature norms across different modalities:

$$\mathcal{L}_{align} = \left(\frac{\mathbb{E}[\mathbf{E}^v]}{\mathbb{E}[\mathbf{E}^a]} - 1\right)^2, \tag{12}$$

where $\mathbb{E}[\mathbf{E}^m]$ is the mean feature for the m-th modality in each batch. This loss mitigates the risk of privileging a single modality during multimodal joint training to improve generalization. The following work [184] enhances the relative norm alignment loss by extending it to align class-level feature norms, further improving its effectiveness.

Recently, Dong et al. [59] introduce a unified framework named SimMMDG to address MMDG across diverse scenarios, including multi-source, single-source, and missing-modality settings, as shown in Fig. 6. SimMMDG introduces a feature disentanglement strategy that separates features into modality-specific and modality-shared components for better generalization. For example, given a unimodal embedding $\mathbf{E}$, SimMMDG splits it as $\mathbf{E} = [\mathbf{E}_s; \mathbf{E}_c]$, where $\mathbf{E}_s$ is modality-specific feature and $\mathbf{E}_c$ is modality-shared feature. Such disentanglement is achieved via supervised contrastive learning [185] on modality-shared features with distance constraints on modality-specific features to promote diversity. For a set of N randomly sampled label pairs in a batch, $\{\boldsymbol{x}_j, y_j\}_{j=1,...,N}$, the corresponding batch used for training consists of $M \times N$ pairs, $\{\tilde{\boldsymbol{x}}_k, \tilde{y}_k\}_{k=1,...,M\times N}$, where $\tilde{\boldsymbol{x}}_{M\times j}$, $\tilde{\boldsymbol{x}}_{M\times j-1}$, ... , $\tilde{\boldsymbol{x}}_{M\times j-M+1}$ are data instances from M different modalities in $\boldsymbol{x}_j$ ($j = 1, ..., N$) and

$\tilde{y}_{M\times j} = \tilde{y}_{M\times j-1} = ... = \tilde{y}_{M\times j-M+1} = y_j$. Let $i \in I \equiv \{1,...,M \times N\}$ be the index of an arbitrary uni-modal sample within a batch. We define $A(i) \equiv I \setminus \{i\}$, $P(i) \equiv \{p \in A(i) : \tilde{y}_p = \tilde{y}_i\}$ as the set of indices of all positive samples in the batch which share the same label as i. The cardinality of $P(i)$ is denoted as $|P(i)|$. The multimodal supervised contrastive learning loss can be written as:

$$\mathcal{L}_{con} = \sum_{i\in I} \frac{-1}{|P(i)|} \sum_{p\in P(i)} \log \frac{\exp(\boldsymbol{z}_i \boldsymbol{\cdot} \boldsymbol{z}_p/\tau)}{\sum\limits_{a\in A(i)} \exp(\boldsymbol{z}_i \boldsymbol{\cdot} \boldsymbol{z}_a/\tau)}, \quad (13)$$

with $\boldsymbol{z}_k = Proj(g(\tilde{\boldsymbol{x}}_k)) \in \mathcal{R}^{D_P}$, where $g(\cdot)$ is the feature extractor, that maps $\boldsymbol{x}$ to modality-specific and modality-shared features, $\mathbf{E} = [\mathbf{E}_s; \mathbf{E}_c] = g(\boldsymbol{x})$, where $\mathbf{E}_s, \mathbf{E}_c \in R^{D_E}$, and $Proj(\cdot)$ is the projection network that maps $\mathbf{E}_c$ to a vector $\boldsymbol{z} = Proj(\mathbf{E}_{\boldsymbol{c}}) \in \mathcal{R}^{D_P}$. The inner product between two projected feature vectors is denoted by $\boldsymbol{\cdot}$, and $\tau \in \mathcal{R}^+$ is a scalar temperature parameter. The distance loss is proposed to ensure that the modality-specific features $\mathbf{E}_s$ carry unique and complementary information and is defined as:

$$\mathcal{L}_{dis} = \frac{-1}{M} \sum_{i=1}^{M} ||\mathbf{E}_s^i - \mathbf{E}_c^i||_2^2, \quad (14)$$

where M is the number of modalities, $\mathbf{E}_s^i$ and $\mathbf{E}_c^i$ are the modality-specific and modality-shared features of the i^{th} modality. Finally, a cross-modal translation module is further proposed to ensure the meaningfulness of modality-specific features and improve robustness in missing-modality scenarios:

$$\mathcal{L}_{trans} = \frac{1}{M(M-1)} \sum_{i=1}^{M} \sum_{j\neq i} ||MLP_{\mathbf{E}^i\rightarrow\mathbf{E}^j}(\mathbf{E}^i) - \mathbf{E}^j||_2^2, \quad (15)$$

where MLP is a multi-layer perception to translate the embedding of the i^{th} modality to the j^{th} modality.

Building on SimMMDG, MOOSA [29] extends MMDG to the open-set setting for the first time, introducing the multimodal open-set domain generalization problem. MOOSA leverages self-supervised tasks, such as masked cross-modal translation and multimodal Jigsaw puzzles, along with an entropy-weighting mechanism for balancing losses across modalities. Differently, Fan et al. [61] identify modality competition and discrepant uni-modal flatness as key limitations in MMDG. To address these issues, they construct consistent flat loss regions and enhance knowledge exploitation for each modality through cross-modal knowledge transfer.

5.2.2 MMDG for Semantic Segmentation

Beyond action recognition, MMDG has been applied to 3D semantic segmentation in BEV-DG [62]. BEV-DG employs a BEV-driven domain contrastive learning strategy to optimize the modeling of domain-irrelevant representations and introduces a BEV-based area-to-area fusion mechanism to improve cross-modal learning. These innovations provide greater robustness and fault tolerance compared to traditional point-level alignment approaches.

6 Domain Adaptation and Generalization with the Help of Multimodal Foundation Models

With the recent emergence of large-scale pre-trained multimodal foundation models (MFMs) such as CLIP [14], stable diffusion [19], and segment anything model (SAM) [170], numerous studies have explored leveraging these models to enhance generalization capabilities. These approaches can be categorized into three main directions: data augmentation, knowledge distillation, and learning strategies.

6.1 Multimodal Foundation Models

MFMs are large-scale models trained on extensive datasets, showing strong generalization and requiring minimal fine-tuning for diverse tasks [186].

Contrastive Language–Image Pretraining (CLIP) [14] is a vision-language model comprising an image encoder that maps high-dimensional images to a low-dimensional embedding space and a text encoder that generates text representations from natural language. CLIP, trained on 400 million image-text pairs, aligns image and text embedding spaces using contrastive loss. For a batch of image-text pairs, CLIP maximizes the cosine similarity for matched pairs while minimizing it for unmatched pairs. During testing, the class names of a target dataset are embedded using the text encoder with prompts in the form of "a photo of a [CLASS]", where the class token is replaced with specific class names, such as "cat", "dog" or "car". The text encoder generates text embeddings $\mathbf{T}_c$ for each class, and the prediction probability for an input image $\mathbf{x}$, with embedding $\mathbf{I}_{\mathbf{x}}$, is computed as:

$$P(y|\mathbf{x}) = \frac{\exp\big(\cos(\mathbf{I}_{\mathbf{x}}, \mathbf{T}_y)/\tau\big)}{\sum_{c=1}^{C} \exp\big(\cos(\mathbf{I}_{\mathbf{x}}, \mathbf{T}_c)/\tau\big)}, \quad (16)$$

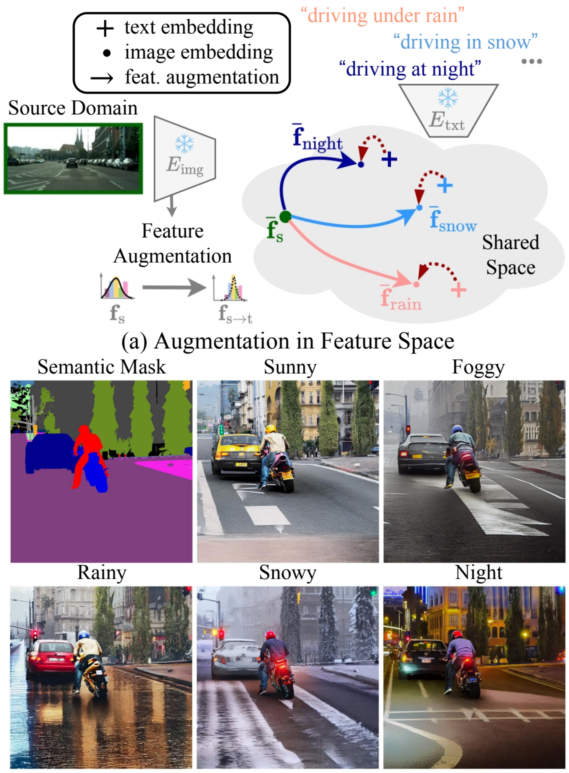
where $\cos(\cdot,\cdot)$ is the cosine similarity between embeddings, and τ is a temperature.

Diffusion Models [19], such as denoising diffusion probabilistic models [187], learn the desired data distribution through a Markov chain of length T. In the forward pass, noise is progressively added to a data sample x_0 to create a sequence of noisy samples $x_t, t \in T$. In the reverse process, a model ϵ_θ, parameterized by θ, predicts the added noise at each step t. Stable diffusion [19] applies this denoising process to the latent representation z of x in the latent space of VQGAN [188] with the learning objective of predicting the added noise at each time step t as:

$$\mathcal{L} = \mathbb{E}_{\mathcal{E}(x),\epsilon\sim\mathcal{N}(0,1),t}\left[\|\epsilon - \epsilon_\theta(z_t, t)\|_2^2\right], \quad (17)$$

where z_t represents the noised latent representation at time step t. During inference, the reverse process starts with a random noise $x_T \sim \mathcal{N}(0, \mathbf{I})$ and iteratively generates an image sample from the noise from step T to 0. Stable diffusion also supports flexible conditional image generation through a cross-attention mechanism [189], enabling models to conditionally learn with various input modalities.

Segment Anything Model (SAM) [170] is a foundation model trained for promptable segmentation tasks, capable of producing high-quality masks for diverse segmentation



(b) Augmentation in Input Space

Fig. 7. Examples of data augmentation using MFMs in feature space [67] and input space [63].

prompts, including points, boxes, text, or masks. SAM consists of an image encoder for extracting image embeddings, a prompt encoder for embedding both sparse (points, boxes, text) and dense (masks) prompts, and a fast mask decoder that efficiently maps image and prompt embeddings to output masks. Trained on over 1 billion masks, SAM demonstrates strong zero-shot segmentation performance.

6.2 Data Augmentation

Several studies have leveraged MFMs to generate additional training data for augmentation, either in the feature space or input space (Fig. 7), to enhance generalization capabilities.

6.2.1 Augmentation in Feature Space

Generating data in feature space is more computationally efficient. For example, Dunlap et al. [20] learn a transformation of image embeddings from the training domain to unseen test domains using text descriptions. A classifier is then trained on both real and augmented embeddings, with domain alignment and class consistency losses ensuring that augmented embeddings remain in the correct domain and retain their class identity. Fahes et al. [67] employ general language descriptions of target domains to optimize affine transformations of source features, steering them towards target text embeddings while preserving semantic content. Given a source feature $\mathbf{f}_s$, they propose to generate stylized target feature $\mathbf{f}_{s\to t}$ by:

$$\mathbf{f}_{s\to t} = \boldsymbol{\sigma}\left(\frac{\mathbf{f}_s - \mu(\mathbf{f}_s)}{\sigma(\mathbf{f}_s)}\right) + \boldsymbol{\mu}, \tag{18}$$

where $\mu(\cdot)$ and $\sigma(\cdot)$ are two functions returning channel-wise mean and standard deviation of input feature, $\boldsymbol{\mu}$ and $\boldsymbol{\sigma}$ are optimizable variables for target style driven by a prompt, which is the description embedding TrgEmb of target domain (e.g., "driving under rain"):

$$\mathcal{L}_{\boldsymbol{\mu},\boldsymbol{\sigma}}(\mathbf{f}_{s\to t}, \mathsf{TrgEmb}) = 1 - \frac{\mathbf{f}_{s\to t} \cdot \mathsf{TrgEmb}}{\|\mathbf{f}_{s\to t}\| \, \|\mathsf{TrgEmb}\|}. \tag{19}$$

Similarly, Yang et al. [69] extend this idea and achieve adaptation to diverse target domains without explicit domain knowledge. Meanwhile, Cho et al. [68] simulate distribution shifts in the joint feature space by synthesizing diverse styles via prompts, eliminating the need for real images. Vidit et al. [190] estimate a set of semantic augmentations using textual domain prompts and source domain images to transform source image embeddings into the target domain specified by the prompts.

6.2.2 Augmentation in Input Space

Some works generate data directly in the image space to simulate domain shifts. For instance, Jia et al. [63] generate a diverse dataset of street scenes using latent diffusion models [19] and train a domain-agnostic semantic segmentation model on this synthetic dataset. Hemati et al. [64] adopt a similar diffusion-based framework for data-centric augmentation, addressing domain generalization by diversifying training data. Similarly, Singha et al. [66] generate proxy images for unknown classes using stable diffusion models and incorporate class-discriminative knowledge into visual embeddings. In contrast, Benigmim et al. [65] integrate multiple foundation models for domain-generalized semantic segmentation. They utilize the CLIP backbone for robust feature representation, diffusion models for generating synthetic images to diversify content, and SAM for iterative prediction refinement.

6.3 Knowledge Distillation

Distilling the rich knowledge of MFMs into smaller, more efficient models has been a key approach for enhancing domain generalization across various tasks. For instance, Zara et al. [70] integrate visual representations from pre-trained VLMs with source model knowledge and target data for source-free video domain adaptation. They use CLIP's extensive prior knowledge for pseudo-labeling and knowledge distillation. In a similar vein, Huang et al. [21] distill CLIP's semantic knowledge into a smaller student model, enforcing the student's representations to align more closely with CLIP's text embeddings, which are concise and domain-invariant. Chen et al. [71] extend this concept to open-set domain generalization by designing perturbations across score, class, and instance levels, enabling lightweight vision models to inherit knowledge from large-scale VLMs. Similarly, Tang et al. [72] customize VLMs for source-free domain adaptation through unsupervised prompt learning, embedding task-specific information, and distilling the customized VLM's knowledge into a target model. Addepalli et al. [73] focus on aligning the vision and language modalities of a teacher model with the vision modality of a pre-trained student model, distilling the aligned representations to the student. Additionally, Li et al. [191] improve lightweight students' generalization by distilling fine-grained visual representations, enhancing vision-language alignment, and enriching teacher models with detailed semantic attributes. Mistretta et al. [192] advance prompt learning in VLMs by distilling knowledge from powerful models without requiring labeled data.

6.4 Learning Strategies

In addition to data augmentation and knowledge distillation, different learning strategies are being developed for domain adaptation and generalization with MFMs.

6.4.1 Prompt-based Strategies

Prompt-based methods try to learn prompts for MFMs to improve DA and DG performances. For example, Zhang et al. [74] extend CLIP for DG by training a lightweight prompt generator that generates domain-specific prompts from input images and append them to label prompts. Ge et al. [75] encode domain-specific information into prompts shared by images from the same domain, dynamically adapting the classifier for each domain. Similarly, Chen et al. [76] address multi-source DA by learning individual prompts for each source-target domain pair, and then mining relationships among prompts to derive a shared, domain-invariant embedding space. Cheng et al. [77] advance this further by introducing a prompt-tuning framework that combines LLM-assisted text prompt disentanglement and text-guided visual representation disentanglement, along with domain-specific prototype learning for better utilization of domain-specific and domain-invariant information. Wang et al. [78] employ vision prompts for domain invariance and language prompts for class separability, balancing the two with adaptive weighting mechanisms. Recently, Xiao et al. [193] introduce any-shift prompting, a probabilistic inference framework using hierarchical architecture and transformer inference networks to construct test prompts that leverage training-test distribution relationships. Additionally, Li et al. [194] develop a dynamic object-centric perception network using prompt learning with object-centric gating and dynamic selective modules to focus on relevant spatial and channel features. Bai et al. [195] presents a generative prompt-learning approach for DG, where domain-specific soft prompts are trained, and instance-specific prompts are generated for unseen target domains.

6.4.2 Prior-based Strategies

Prior-based methods leverage strong prior information from MFMs to improve performances. For instance, Peng et al. [82] address domain adaptation for 3D segmentation by leveraging SAM [170] to incorporate 2D prior knowledge, facilitating the alignment of features across diverse 3D data domains into a unified domain. Similarly, Lim et al. [79] integrate the rich semantic knowledge of VLMs with segment reasoning from traditional domain adaptation methods to relabel new classes in the target domain. This approach enables effective adaptation to extended taxonomies without requiring ground truth labels in the target domain. Xu et al. [80] enhance cross-modal unsupervised domain adaptation by utilizing the prior knowledge encoded in Vision Foundation Models (VFMs) to generate more accurate labels for unlabeled target domains. In a different vein, Zhu et al. [81] employ CLIP to quantify domain divergence through domain-agnostic distributions and calibrate target pseudo-labels with language guidance, effectively reducing the domain gap. Additionally, several works [83], [196], [197] leverage diffusion priors for efficient adaptation.

6.4.3 Refinement-based Strategies

Refinement-based methods use MFMs to refine the features or pseudo-labels for better adaptation and generalization. For example, Wei et al. [84] introduce a robust parameter-efficient fine-tuning method for utilizing VFMs in domain-generalized semantic segmentation. They incorporate learnable tokens directly tied to instances, refining features at the instance level within each layer. Similarly, Lai et al. [85] propose an efficient adaptation of VLMs that preserves their original knowledge while maximizing flexibility for learning new information. They also design a domain-aware pseudo-labeling scheme tailored for VLMs to achieve effective domain disentanglement. In a complementary approach, Hu et al. [86] first mitigate misaligned visual-text embeddings by learning a projection space and generating pseudo-labels. They iteratively apply cross-modality self-training to update visual and text encoders, refine labels, and reduce domain gaps and misalignments. Recently, Xia et al. [87] leverage text-to-image diffusion models pre-trained on large-scale image-text datasets to enhance cross-modal capabilities for semantic segmentation. They achieve state-of-the-art performance across various modality tasks, including images-to-depth, infrared, and event modalities. Zhang et al. [88] integrate CLIP's vision encoder with its zero-shot text-based classifier, refining the fitted classifier for source-free DA.

6.4.4 Other Strategies

There are some other learning strategies for domain adaptation and generalization with MFMs. For instance, Cha et al. [89] introduce mutual information regularization with Oracle, robustly approximating Oracle models using large pre-trained models such as SWAG [198] and CLIP [14]. Shu et al. [90] adapt CLIP to handle domain shifts and open-class scenarios. They introduce margin metric softmax to capture semantic relationships between text classes and combine a zero-shot model with a fine-tuned task-adaptive model using the beta moving average for optimization. Differently, Lai et al. [91] address catastrophic forgetting when fine-tuning CLIP on target domains by introducing catastrophic forgetting measurement to dynamically adjust the learning rate. Zara et al. [199] generate object-centric compositional candidate names for target-private classes, enabling the effective rejection of target-private instances and improving alignment between shared classes across domains. Yu et al. [200] introduce an entropy optimization strategy to support open-set DA by utilizing CLIP outputs. A recent work by Li et al. [201] disentangles CLIP features into language- and vision-specific components, utilizes modality-ensemble training to balance modality-specific nuances and shared information, and employs a modality discriminator for cross-domain alignment.

7 Adaptation of Multimodal Foundation Models

While MFMs exhibit strong zero-shot prediction capabilities, the distribution shifts of images in downstream tasks hinder their generalization. Various transfer learning strategies, such as prompt tuning and feature adapters, have been developed to adapt MFMs to downstream tasks. Fig. 8 illustrates the difference between prompt-based and adapter-based adaptation.

7.1 Prompt-based Adaptation

Prompt-based adaptation modifies input texts or images using a few learnable prompts for parameter-efficient tuning, avoiding the need to fine-tune the entire model.

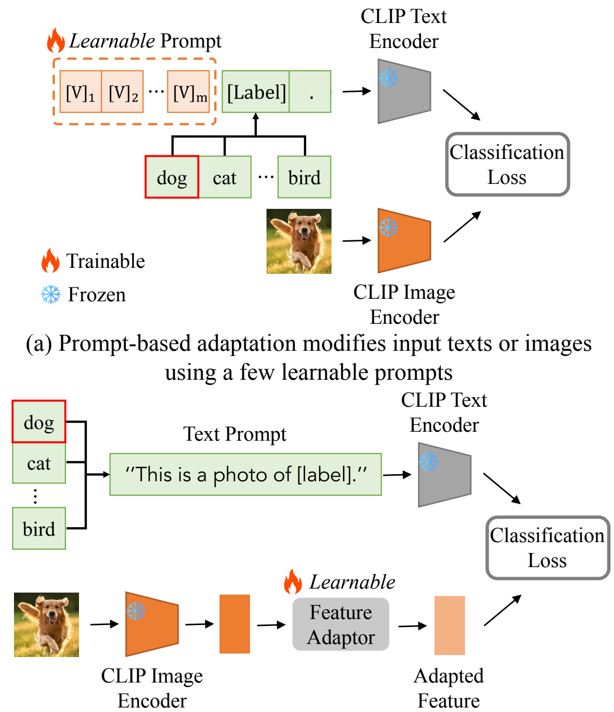
7.1.1 *Text Prompt Tuning*

Text prompt tuning methods modify the text encoder's input to adapt MFMs. For example, CoOp [22] adapts CLIP-like VLMs for downstream image recognition by modeling a prompt's context words as learnable vectors, inspired by prompt learning in NLP [202]. Instead of predefined prompts like "a photo of a [Label]", CoOp represents the prompt as "$[V]_1[V]_2...[V]_M$[Label]", where $[V]_m$ is a vector matching the word embedding dimension. These vectors are optimized using standard cross-entropy loss. Building on CoOp, CoCoOp [94] introduces input-conditional tokens for each image to mitigate CoOp's overfitting to base classes. To further address overfitting, Ma et al. [95] propose projecting gradients onto the low-rank subspace defined by early-stage gradient flow eigenvectors. Bulat et al. [96] take a different approach by ensuring that learned prompts remain close to handcrafted ones to reduce base class overfitting. Lu et al. [97] shift focus to learning the distribution of output embeddings of prompts rather than input embeddings. In a similar vein, Derakhshani et al. [98] model the input prompt space probabilistically as a prior distribution, while Zhu et al. [203] selectively update prompts whose gradients align with general knowledge, preventing knowledge forgetting. He et al. [204] utilize counterfactual generation and contrastive learning for prompt tuning. A novel approach by Chen et al. [205] employs optimal transport to learn diverse prompts capable of capturing the comprehensive characteristics of various categories

Additionally, Sun et al. [206] learn a pair of positive and negative prompts to enhance multi-label recognition, while Guo et al. [207] utilize text descriptions to learn prompts, extracting both coarse- and fine-grained embeddings to improve multi-label recognition performance. Ding et al. [208] propose a multi-task prompt tuning approach by employing a task-shared meta-network to generate task-specific prompt contexts. Yao et al. [209] enhance prompt tuning for unseen classes by minimizing the discrepancy between learnable and hand-crafted prompts. Recently, Wu et al. [210] improved prompt tuning for open-set classification by calibrating predictions across semantic label hierarchies.

7.1.2 *Visual Prompt Tuning*

Visual prompt tuning methods modify the image encoder's input to adapt MFMs. For example, Bahng et al. [99] adapt CLIP by learning a single image perturbation to guide the frozen model for new tasks, achieving competitive performance with linear probes. Rong et al. [100] enhance fine-grained classification in few-shot learning by incorporating a retrieval mechanism that references relevant samples during inference. In contrast, Wu et al. [101] employ a learnable prompt wrapped around a resized image, combined with input diversity and gradient normalization techniques.



(a) Prompt-based adaptation modifies input texts or images using a few learnable prompts

(b) Adapter-based adaptation methods learn lightweight feature adapters

Fig. 8. A comparison of prompt-based and adapter-based adaptation.

7.1.3 *Text-Visual Prompt Tuning*

Text-visual prompt tuning methods modify both text and visual encoder inputs to adapt MFMs. For example, Zang et al. [102] combine text and visual prompt tuning by optimizing prompts across both modalities using a small neural network, achieving improved trade-offs in few-shot learning and domain generalization compared to unimodal approaches. Shen et al. [103] leverage cross-task knowledge to enhance prompt tuning, learning transferable prompts from multiple source tasks and enabling joint adaptation across tasks. Khattak et al. [104] jointly learn dynamic prompts for both vision and language branches, ensuring strong coupling and stage-wise feature alignment for improved generalization. Similarly, Xing et al. [105] improve prompt tuning by jointly learning text and visual prompts, incorporating a class-aware visual prompt tuning scheme that dynamically generates visual prompts through cross-attention between text prompts and image features. Abdul et al. [211] align test sample statistics with source data through prompt tuning, while Khattak et al. [212] introduce a self-regularization framework for CLIP prompt learning, guiding prompts to optimize both task-specific and task-agnostic representations through mutual agreement. Recent developments, such as Wu et al. [106] introduced a two-phase framework that first extracts domain-general knowledge from a larger teacher model using boosting prompts and then adapts prompts for task-specific fine-tuning. Meanwhile, Hao et al. [213] address overfitting and catastrophic forgetting by applying quantization as a lightweight regularization technique.

7.2 Adapter-based Adaptation

Adapter-based adaptation methods leverage lightweight feature adapters to fine-tune MFMs. For example, Clip-Adapter [107] enhances VLMs by fine-tuning feature adapters on either the visual or language branch. It employs a bottleneck layer and residual-style feature blending, outperforming context optimization while maintaining a simple design. The adapter is a two-layer MLP with parameters W_1, b_1, W_2, b_2. Given an input image feature f_c, the adapted feature f_a is:

$$f_a = \varphi(f_c W_1^T + b_1) W_2^T + b_2, \tag{20}$$

where φ denotes the activation function in the MLP. The adapted feature f_a is linearly combined with the f_c with a hyper-parameter $\alpha \in [0, 1]$ to output the final prediction:

$$\hat{p} = \alpha f_a W_c^T + f_c W_c^T, \tag{21}$$

where W_c is the weight of the text classifier. The parameters of MLP are optimized by minimizing the cross-entropy loss between predictions and ground truths. Tip-Adapter [112] improves few-shot classification by generating adapter weights through a key-value cache model from the few-shot training set. This approach eliminates the need for backpropagation and training, achieving performance comparable to or exceeding that of Clip-Adapter. SVL-Adpter [108] combines vision-language pretraining with self-supervised representation learning to enhance low-shot image classification. Besides, Kahana et al. [109] adapt CLIP for regression and classification tasks on unlabeled datasets by incorporating a distributional prior over labels and training an adapter network to minimize prediction changes while aligning with the prior distribution. The work by Karmanov et al. [111] introduces a training-free dynamic adapter for TTA with VLMs. They use a lightweight key-value cache and progressive pseudo-label refinement to adapt efficiently to test data without backpropagation. Lu et al. [214] reduce prediction bias in VLMs by employing variational adapters with learnable textual tokens, which separate base and novel classes in latent space. Additionally, SAM-Adapter [215], MA-SAM [216], and CAT-SAM [217] enhance the performance of SAM by incorporating domain-specific information through simple adapters.

7.3 Other Adaptation Methods

7.3.1 Fine-tuning Methods

Some works fine-tune all parameters of MFMs. For example, Wortsman et al. [113] enhance robustness to distribution shifts in fine-tuning large VLMs by ensembling the weights of zero-shot and fine-tuned models. Li et al. [114] improve CLIP through unsupervised fine-tuning on unlabeled target domain data. They combine pseudo-labeling and regularization to jointly optimize global and local features.

7.3.2 Training-free Methods

Another promising direction aims to adapt MFMs without any parameter tuning. For instance, Udandarao et al. [115] enhance VLMs by constructing dynamic support sets—either via image generation or retrieval—and leveraging image-text distances to refine classification without additional training. Guo et al. [116] enhance CLIP's zero-shot performance by using a parameter-free cross-modal attention module to adaptively align visual and textual features, eliminating the need for additional training or learnable parameters. Zanella et al. [117] jointly optimize view quality assessment and density mode seeking to improve the zero-shot and few-shot performance of VLMs. Zhang et al. [118] introduce dual memory networks, with static memory for caching training data knowledge and dynamic memory for online test feature adaptation. Recently, Wang et al. [119] employ Gaussian assumptions for class features, enabling training-free integration of visual and textual modalities, while Ge et al. [218] improve accuracy by identifying ambiguous predictions through prompt and transformation consistency and augmenting text prompts with semantic labels from the WordNet hierarchy.

7.3.3 LLM-based Methods

Large language models (LLMs) are also helpful for the adaptation of VLMs. For instance, Pratt et al. [120] advance open-vocabulary image classification by utilizing LLMs to generate discriminative prompts, improving accuracy without additional training or task-specific knowledge. Menon et al. [121] enhance VLM-based classification by querying language models for descriptive features. Recently, Parashar et al. [122] leverage LLMs to identify frequent concept synonyms in pretraining data for better prompting.

7.3.4 Test-Time Adaptation of VLMs

Recent works also explore adapting VLMs at test time. For example, Shu et al. [123] adapt prompts dynamically for each test sample by minimizing entropy across augmented views. Differently, Feng et al. [125] employ pre-trained diffusion models for diverse data augmentation and cosine similarity-based filtration to enhance test-time prompt fidelity. Ma et al. [127] use self-supervised contrastive learning with dual prompts—an online prompt and a historical target prompt—alongside a swapped prediction mechanism for improved adaptation. Osowiechi et al. [124] improve prediction accuracy by leveraging weight averaging with different text prompts and incorporating text embedding averaging. A recent work by Farina et al. [126] enhances generalization by augmenting predictions, retaining only the most confident ones, and marginalizing them using a zero Softmax temperature, all without backpropagation.

7.3.5 Dense Prediction Tasks

VLMs can also be extended to dense prediction tasks such as semantic segmentation. Rao et al. [130] introduce a framework for dense prediction tasks by leveraging pre-trained CLIP knowledge, converting image-text matching into pixel-text matching, and utilizing pixel-text score maps alongside contextual prompts for guidance. Similarly, Zhou et al. [128] apply CLIP to pixel-level dense prediction by combining CLIP embeddings with pseudo-labeling and self-training. Recently, Zhang et al. [129] improve the robustness and efficiency of SAM for image segmentation under significant distribution shifts by introducing a weakly supervised self-training strategy with anchor regularization and low-rank fine-tuning.

7.3.6 Other Methods

There are some other methods for the adaptation of MFMs. For instance, Zhang et al. [219] improve CLIP's transfer performance by using visual-guided text features that adaptively explore informative image regions and aggregate visual features through attention, thereby enhancing semantic alignment for downstream classification tasks. Yu et al. [220] enhance VLMs transfer by tuning a residual to the pre-trained text-based classifier, preserving prior knowledge while enabling task-specific adaptation for better performance. Ouali et al. [221] propose a computationally efficient, black-box method for vision-language few-shot adaptation that operates on pre-computed features, aligning image and text representations through a closed-form least-squares initialization and re-ranking loss. Differently, Zanella et al. [222] introduce low-rank adaptation [223] for few-shot learning in VLMs, achieving significant improvements with reduced training times and consistent hyperparameters across multiple datasets. Xuan et al. [224] enhance VLMs by extracting task-agnostic knowledge from the text encoder and injecting it into input image or text features. Recently, Zhang et al. [225] mitigate data misalignment in VLMs by decoupling task-relevant and task-irrelevant knowledge using a structural causal model. Lin et al. [226] learn from visual and non-visual data (e.g., text and audio) and repurpose class names as additional one-shot training samples, enabling a simple linear classifier to achieve strong adaptation performance across vision, language, and audio tasks.

8 Datasets and Applications

Multimodal adaptation and generalization have been studied across many application areas including action recognition, semantic segmentation, image classification, sentiment analysis, person re-identification, depth completion, etc. An overview of common datasets is shown in Tab. 1 and Fig. 9 demonstrates examples of different types of domain shifts.

8.1 Datasets for Action Recognition

The EPIC-Kitchens [16] and Human-Animal-Cartoon (HAC) [59] are two commonly used action recognition datasets for MMDA and MMDG tasks. EPIC-Kitchens includes $10,094$ video clips across eight actions ("put", "take", "open", "close", "wash", "cut", "mix", and "pour"), recorded in three distinct kitchens, forming three separate domains D1, D2, and D3. HAC dataset features seven actions ("sleeping", "watching tv", "eating", "drinking", "swimming", "running", and "opening door") performed by humans, animals, and cartoon figures, spanning three domains H, A, and C with $3,381$ video clips. The video, optical flow, and audio modalities are provided in both datasets. The CharadesEgo [227] and ActorShift [32] datasets are also used in some literature. For MMTTA, UCF [228], HMDB [229], Olympic [230], and Kinetics-600 [241] with corruptions are also used as different types of domain shifts.

8.2 Datasets for Semantic Segmentation

The nuScenes [177], A2D2 [178], and SemanticKITTI [179] are three commonly used 3D semantic segmentation datasets for MMDA and MMDG tasks. For MMDG, BEV-DG [62] train the model on two datasets and test it on the remaining dataset. For MMDA, xMUDA [17] identifies three adaptation scenarios, including day-to-night (nuScenes Day/Night), country-to-country (nuScenes USA/Singapore), and dataset-to-dataset (A2D2/SemanticKITTI). The Synthia [231] and Waymo [242] datasets are also used in some MMTTA literature. For example, MM-TTA [54] uses the Synthia-to-SemanticKITTI adaptation scenario and CoMAC [55] identifies the SemanticKITTI-to-Synthia and SemanticKITTI-to-Waymo adaptation scenarios. For the 2D semantic segmentation task, Cityscapes [233], GTA5 [232], Synthia [231], Mapillary [235], and ACDC [243] are widely-used datasets. The UrbanSyn [244], Dark Zurich [245], and BDD100K [234] datasets are also used in some literature.

8.3 Datasets for Image Classification

Image classification is the most common task for DA and DG with the help of multimodal foundation models. Among them, PACS [238], VLCS [236], Office-Home [237], DomainNet [239], and Wilds [240] are the most popular datasets. For the adaptation of multimodal foundation models, ImageNet [246], Caltech101 [247], OxfordPets [248], StanfordCars [249], Flowers102 [250], Food101 [251], FGVCAircraft [252], SUN397 [253], DTD [254], EuroSAT [255] and UCF101 [228] are 11 commonly used image classification datasets. ImageNetV2 [256], ImageNet-Sketch [257], ImageNet-A [258] and ImageNet-R [259] are also widely used in the TTA setup.

8.4 Datasets for Other Applications

CMU-MOSEI [260], IEMOCAP [261], MELD [262], MSP-IMPROV [263], and CMU-MOSI [264] are classical datasets for the multimodal sentiment analysis and emotion recognition tasks. RGBNT201 [265], Market1501-MM [266], and RGBNT100 and RGBN300 [267] are datasets for multimodal person/vehicle re-identification. VOID [268], NYUv2 [269], SceneNet [270], and ScanNet [271] are common multimodal depth completion datasets. SceneFlow [272], ETH3D [273], Middlebury [274], and KITTI 2015 [275] are used for domain generalized stereo matching in [92].

9 Future Research Challenges

Multimodal adaptation and generalization remain challenging and unsolved problems. In this section, we provide insights into future research directions, highlighting gaps in the current literature and discussing promising avenues for advancing the field.

TABLE 1
Commonly used multimodal adaptation and generalization datasets. V: video, A: audio, F: optical flow, P: point cloud, I: image.

Dataset	Modality	#Domains	Characterization of domain shift
Action recognition			
- HAC [59]	V+A+F	3	Human, animal, and cartoon
- EPIC-Kitchens [16]	V+A+F	3	Three different kitchens
- CharadesEgo [227]	V+A	2	First and third-person view
- ActorShift [32]	V+A	2	Human, animal
- UCF→{HMDB, Olympic} [228], [229], [230]	V+F	3	Dataset to dataset (see [51])
- Kinetics50-C [18]	V+A	-	Artificial corruptions
- Kinetics-100-C [50]	V+A	-	Artificial corruptions
Semantic segmentation			
- nuScenes Day/Night [177]	P+I	2	Day to Night (see [17])
- nuScenes USA/Singapore [177]	P+I	2	Country to country (see [17])
- A2D2→SemanticKITTI [178], [179]	P+I	2	Dataset to dataset (see [17])
- SYNTHIA→SemanticKITTI [179], [231]	P+I	2	Synthetic to real (see [54])
- SemanticKITTI→SYNTHIA [179], [231]	P+I	2	Real to synthetic (see [55])
- SemanticKITTI→Waymo [179], [231]	P+I	2	Dataset to dataset (see [55])
- GTA5→{Cityscapes, BDD100K, Mapillary} [232], [233], [234], [235]	I	4	Synthetic to real (see [84])
- Cityscapes→{BDD100K, Mapillary} [233], [234], [235]	I	3	Dataset to dataset (see [84])
Image classification			
- VLCS [236]	I	4	Dataset to dataset
- OfficeHome [237]	I	4	Art, clipart, product, real
- PACS [238]	I	4	Photo, art, cartoon, sketch
- DomainNet [239]	I	6	Clipart, infograph, painting, quickdraw, real, sketch
- Wilds [240]	I	-	Camera, hospital, batch, scaffold, location, time, etc

(a) HAC (b) EPIC-Kitchens (c) CharadesEgo (d) nuScenes

Fig. 9. Examples from various datasets manifesting different types of domain shift. In (a), the domain shift mainly corresponds to the person, animals, or cartoon characters that perform actions. In (b), the domain shift is caused by styles in different kitchens. In (c), viewpoint changes are the main reason for the domain shift. In (d), the domain shift is caused by day and night changes.

9.1 Theoretical Analysis

While existing research has largely focused on the theoretical analysis of unimodal DA [276] and DG [277], more rigorous analyses in multimodal scenarios are necessary to provide deeper insights and inspire the development of novel MMDA and MMDG methods. SimMMDG [59] represents an initial effort to explore MMDG theory from both multimodal representation learning and domain generalization perspectives. Future work could benefit from integrating multimodal learning theory [278] with unimodal DG principles to advance the field. Besides, the theoretical analysis of the adaptation of MFMs is also less explored.

9.2 Benchmark and Datasets

Unlike unimodal DA and DG, for which numerous benchmarks [239], [240], [279] exist to facilitate comprehensive evaluation, there is currently no open benchmark for MMDA or MMDG. Additionally, the datasets used in existing MMDA and MMDG studies are significantly smaller in scale compared to their unimodal counterparts. To accelerate progress in these areas, future research should focus on developing comprehensive benchmarks encompassing datasets of varying scales for MMDA and MMDG.

9.3 Open-set Settings

Most existing multimodal adaptation and generalization approaches adhere to the closed-set assumption, which assumes identical label spaces across domains. However, in practical applications, target domains often contain unknown classes, making it crucial to detect these classes effectively. Although some recent studies [29], [50] have addressed the open-set scenario, this challenging yet practical setting requires further investigation. Methods developed for multimodal out-of-distribution detection [280], [281] can potentially be adapted to facilitate unknown class detection in multimodal adaptation and generalization.

9.4 MMTTA and MMDG

While significant efforts have been made to develop methods for MMDA, relatively few studies have targeted MMTTA [18], [50] and MMDG [59], [61] problems. Considering the practical importance of MMTTA and MMDG in real-world scenarios, future research should prioritize addressing these challenging topics.

9.5 Diverse Downstream Tasks

Although multimodal adaptation and generalization have been applied to tasks such as image classification, action recognition, and semantic segmentation, their potential applications in other domains, including regression [282], [283], generative models [284], and image super-resolution [285], remain underexplored and warrant greater attention.

9.6 MFMs for MMDA and MMDG

MFMs have demonstrated their ability to enhance unimodal DA and DG for images [67], [68]. However, leveraging MFMs to complement other modalities, such as audio and optical flow, remains an open challenge. Given the robust feature representations and strong generalization capabilities of MFMs, we remain optimistic about their potential to further improve MMDA and MMDG performance.

10 Conclusion

Adapting pre-trained multimodal models to target domains under distribution shifts represents a critical challenge in machine learning that receives more and more attention these days. This survey provides a comprehensive overview of recent advancements in multimodal domain adaptation, multimodal test-time adaptation, and multimodal domain generalization, highlighting key challenges, methodologies, and applications driving progress in the field. Furthermore, we emphasize the critical role of multimodal foundation models in enhancing domain adaptation and generalization tasks, highlighting their potential to address real-world challenges across diverse modalities. By reviewing existing approaches, datasets, and applications, we identify several key directions for future research, including the development of better benchmarks and datasets, the handling of label shifts in dynamic environments, and further exploration of theoretical analysis. As the field continues to evolve, these insights offer a valuable foundation for advancing the robustness and efficiency of multimodal models in real-world scenarios.

References

[1] M. Wang and W. Deng, "Deep visual domain adaptation: A survey," *Neurocomputing*, vol. 312, pp. 135–153, 2018.

[2] J. Wang, C. Lan, C. Liu, Y. Ouyang, T. Qin, W. Lu, Y. Chen, W. Zeng, and P. Yu, "Generalizing to unseen domains: A survey on domain generalization," *IEEE Transactions on Knowledge and Data Engineering*, vol. 35, no. 8, pp. 8052–8072, 2022.

[3] H. Dong, X. Chen, M. Dusmanu, V. Larsson, M. Pollefeys, and C. Stachniss, "Learning-based dimensionality reduction for computing compact and effective local feature descriptors," in *ICRA*, 2023.

[4] H. Dong, X. Chen, S. Särkkä, and C. Stachniss, "Online pole segmentation on range images for long-term lidar localization in urban environments," *Robotics and Autonomous Systems*, vol. 159, p. 104283, 2023.

[5] D. Damen, H. Doughty, G. M. Farinella, S. Fidler, A. Furnari, E. Kazakos, D. Moltisanti, J. Munro, T. Perrett, W. Price, and M. Wray, "Scaling egocentric vision: The epic-kitchens dataset," in *ECCV*, 2018.

[6] H. Sun, Y. Cao, and O. Fink, "Cut: A controllable, universal, and training-free visual anomaly generation framework," *arXiv preprint arXiv:2406.01078*, 2024.

[7] H. Dong, G. Frusque, Y. Zhao, E. Chatzi, and O. Fink, "NNG-Mix: Improving Semi-supervised Anomaly Detection with Pseudo-anomaly Generation," *arXiv preprint arXiv:2311.11961*, 2023.

[8] M. Long, Y. Cao, J. Wang, and M. Jordan, "Learning transferable features with deep adaptation networks," in *ICML*, 2015.

[9] K. Muandet, D. Balduzzi, and B. Schölkopf, "Domain generalization via invariant feature representation," in *ICML*, 2013.

[10] V. Piratla, P. Netrapalli, and S. Sarawagi, "Efficient domain generalization via common-specific low-rank decomposition," in *ICML*, 2020.

[11] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, "mixup: Beyond empirical risk minimization," in *ICLR*, 2018.

[12] D. Li, Y. Yang, Y.-Z. Song, and T. M. Hospedales, "Learning to generalize: meta-learning for domain generalization," in *AAAI*, 2018.

[13] E. Kazakos, A. Nagrani, A. Zisserman, and D. Damen, "Epic-fusion: Audio-visual temporal binding for egocentric action recognition," in *ICCV*, 2019.

[14] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natural language supervision," in *ICML*, 2021.

[15] H. Dong, X. Zhang, J. Xu, R. Ai, W. Gu, H. Lu, J. Kannala, and X. Chen, "Superfusion: Multilevel lidar-camera fusion for long-range hd map generation," *arXiv preprint arXiv:2211.15656*, 2022.

[16] J. Munro and D. Damen, "Multi-modal domain adaptation for fine-grained action recognition," in *CVPR*, 2020.

[17] M. Jaritz, T.-H. Vu, R. d. Charette, E. Wirbel, and P. Pérez, "xmuda: Cross-modal unsupervised domain adaptation for 3d semantic segmentation," in *CVPR*, 2020.

[18] M. Yang, Y. Li, C. Zhang, P. Hu, and X. Peng, "Test-time adaption against multi-modal reliability bias," in *ICLR*, 2024.

[19] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *CVPR*, 2022.

[20] L. Dunlap, C. Mohri, D. Guillory, H. Zhang, T. Darrell, J. E. Gonzalez, A. Raghunathan, and A. Rohrbach, "Using language to extend to unseen domains," in *ICLR*, 2023.

[21] Z. Huang, A. Zhou, Z. Ling, M. Cai, H. Wang, and Y. J. Lee, "A sentence speaks a thousand images: Domain generalization through distilling clip with language guidance," in *ICCV*, 2023.

[22] K. Zhou, J. Yang, C. C. Loy, and Z. Liu, "Learning to prompt for vision-language models," *International Journal of Computer Vision*, vol. 130, no. 9, pp. 2337–2348, 2022.

[23] J. Zhang, J. Huang, S. Jin, and S. Lu, "Vision-language models for vision tasks: A survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 46, no. 8, pp. 5625–5644, 2024.

[24] F. Qi, X. Yang, and C. Xu, "A unified framework for multimodal domain adaptation," in *ACMMM*, 2018.

[25] W. Zhang, D. Xu, J. Zhang, and W. Ouyang, "Progressive modality cooperation for multi-modality domain adaptation," *IEEE Transactions on Image Processing*, vol. 30, pp. 3293–3306, 2021.

[26] Y. Yin, B. Zhu, J. Chen, L. Cheng, and Y.-G. Jiang, "Mix-dann and dynamic-modal-distillation for video domain adaptation," in *ACMMM*, 2022.

[27] X. Song, S. Zhao, J. Yang, H. Yue, P. Xu, R. Hu, and H. Chai, "Spatio-temporal contrastive domain adaptation for action recognition," in *CVPR*, 2021.

[28] D. Kim, Y.-H. Tsai, B. Zhuang, X. Yu, S. Sclaroff, K. Saenko, and M. Chandraker, "Learning cross-modal contrastive features for video domain adaptation," in *ICCV*, 2021.

[29] H. Dong, E. Chatzi, and O. Fink, "Towards multimodal open-set domain generalization and adaptation through self-supervision," *arXiv preprint arXiv:2407.01518*, 2024.

[30] J. Lv, K. Liu, and S. He, "Differentiated learning for multi-modal domain adaptation," in *ACMMM*, 2021.

[31] Y. Huang, X. Yang, J. Zhang, and C. Xu, "Relative alignment network for source-free multimodal video domain adaptation," in *ACMMM*, 2022.

[32] Y. Zhang, H. Doughty, L. Shao, and C. G. Snoek, "Audio-adaptive activity recognition across video domains," in *CVPR*, 2022.

[33] M. Li, Y. Zhang, Y. Xie, Z. Gao, C. Li, Z. Zhang, and Y. Qu, "Cross-domain and cross-modal knowledge distillation in domain adaptation for 3d semantic segmentation," in *ACMMM*, 2022.

[34] Y. Chen, S. Zhao, C. Ding, L. Tang, C. Wang, and D. Tao, "Cross-modal & cross-domain learning for unsupervised lidar semantic segmentation," in *ACMMM*, 2023.

[35] H. Cao, Y. Xu, J. Yang, P. Yin, S. Yuan, and L. Xie, "Mopa: Multi-modal prior aided domain adaptation for 3d semantic segmentation," in *ICRA*, 2024.

[36] A. Cardace, P. Z. Ramirez, S. Salti, and L. Di Stefano, "Exploiting the complementarity of 2d and 3d networks to address domain-shift in 3d semantic segmentation," in *CVPR*, 2023.

[37] C. Simons, D. S. Raychaudhuri, S. Miraj Ahmed, S. You, K. Karydis, and A. K. Roy-Chowdhury, "Summit: Source-free adaptation of uni-modal models to multi-modal targets," in *ICCV*, 2023.

[38] Y. Zhang, M. Li, Y. Xie, C. Li, C. Wang, Z. Zhang, and Y. Qu, "Self-supervised exclusive learning for 3d segmentation with cross-modal unsupervised domain adaptation," in *ACMMM*, 2022.

[39] B. Xing, X. Ying, R. Wang, J. Yang, and T. Chen, “Cross-modal contrastive learning for domain adaptation in 3d semantic segmentation,” in *AAAI*, 2023.

[40] B. Zhang, Z. Wang, Y. Ling, Y. Guan, S. Zhang, and W. Li, “Mx2m: masked cross-modality modeling in domain adaptation for 3d semantic segmentation,” in *AAAI*, 2023.

[41] D. Peng, Y. Lei, W. Li, P. Zhang, and Y. Guo, “Sparse-to-dense feature matching: Intra and inter domain cross-modal learning in domain adaptation for 3d semantic segmentation,” in *ICCV*, 2021.

[42] W. Liu, Z. Luo, Y. Cai, Y. Yu, Y. Ke, J. M. Junior, W. N. Gonçalves, and J. Li, “Adversarial unsupervised domain adaptation for 3d semantic segmentation with multi-modal learning,” *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 176, pp. 211–221, 2021.

[43] Y. Man, L. Gui, and Y.-X. Wang, “Dualcross: Cross-modality cross-domain adaptation for monocular bev perception,” in *IROS*, 2023.

[44] A. Vobecky, D. Hurych, O. Siméoni, S. Gidaris, A. Bursuc, P. Pérez, and J. Sivic, “Drive&segment: Unsupervised semantic segmentation of urban scenes via cross-modal distillation,” in *ECCV*, 2022.

[45] Y. Yin, W. Hu, Z. Liu, G. Wang, S. Xiang, and R. Zimmermann, “Crossmatch: Source-free domain adaptive semantic segmentation via cross-modal consistency training,” in *ICCV*, 2023.

[46] G. Rizzoli, D. Shenaj, and P. Zanuttigh, “Source-free domain adaptation for rgb-d semantic segmentation with vision transformers,” in *WACV*, 2024.

[47] X. Ma, T. Zhang, and C. Xu, “Deep multi-modality adversarial networks for unsupervised domain adaptation,” *IEEE Transactions on Multimedia*, vol. 21, no. 9, pp. 2419–2431, 2019.

[48] Y. Liu, L. Qiao, C. Lu, D. Yin, C. Lin, H. Peng, and B. Ren, “Osan: A one-stage alignment network to unify multimodal alignment and unsupervised domain adaptation,” in *CVPR*, 2023.

[49] X. Zhang, J. Sun, S. Hong, and T. Li, “Amanda: Adaptively modality-balanced domain adaptation for multimodal emotion recognition,” in *ACL*, 2024.

[50] H. Dong, E. Chatzi, and O. Fink, “Towards robust multimodal open-set test-time adaptation via adaptive entropy-aware optimization,” *arXiv preprint arXiv:2501.13924*, 2025.

[51] B. Xiong, X. Yang, Y. Song, Y. Wang, and C. Xu, “Modality-collaborative test-time adaptation for action recognition,” in *CVPR*, 2024.

[52] J. Lei and F. Pernkopf, “Two-level test-time adaptation in multimodal learning,” in *ICMLW*, 2024.

[53] M. Ramazanova, A. Pardo, B. Ghanem, and M. Alfarra, “Test-time adaptation for combating missing modalities in egocentric videos,” in *ICLR*, 2025.

[54] I. Shin, Y.-H. Tsai, B. Zhuang, S. Schulter, B. Liu, S. Garg, I. S. Kweon, and K.-J. Yoon, “Mm-tta: multi-modal test-time adaptation for 3d semantic segmentation,” in *CVPR*, 2022.

[55] H. Cao, Y. Xu, J. Yang, P. Yin, S. Yuan, and L. Xie, “Multi-modal continual test-time adaptation for 3d semantic segmentation,” in *ICCV*, 2023.

[56] H. Cao, Y. Xu, J. Yang, P. Yin, X. Ji, S. Yuan, and L. Xie, “Reliable spatial-temporal voxels for multi-modal test-time adaptation,” *arXiv preprint arXiv:2403.06461*, 2024.

[57] H. Park, A. Gupta, and A. Wong, “Test-time adaptation for depth completion,” in *CVPR*, 2024.

[58] Z. Wang, H. Huang, A. Zheng, and R. He, “Heterogeneous test-time training for multi-modal person re-identification,” in *AAAI*, 2024.

[59] H. Dong, I. Nejjar, H. Sun, E. Chatzi, and O. Fink, “Simmmdg: A simple and effective framework for multi-modal domain generalization,” in *NeurIPS*, 2023.

[60] M. Planamente, C. Plizzari, E. Alberti, and B. Caputo, “Domain generalization through audio-visual relative norm alignment in first person action recognition,” in *WACV*, 2022.

[61] Y. Fan, W. Xu, H. Wang, and S. Guo, “Cross-modal representation flattening for multi-modal domain generalization,” in *NeurIPS*, 2024.

[62] M. Li, Y. Zhang, X. Ma, Y. Qu, and Y. Fu, “Bev-dg: Cross-modal learning under bird’s-eye view for domain generalization of 3d semantic segmentation,” in *ICCV*, 2023.

[63] Y. Jia, L. Hoyer, S. Huang, T. Wang, L. Van Gool, K. Schindler, and A. Obukhov, “Dginstyle: Domain-generalizable semantic segmentation with image diffusion models and stylized semantic control,” in *ECCV*, 2024.

[64] S. Hemati, M. Beitollahi, A. H. Estiri, B. Al Omari, X. Chen, and G. Zhang, “Leveraging generative foundation models for domain generalization,” in *ICMLW*, 2024.

[65] Y. Benigmim, S. Roy, S. Essid, V. Kalogeiton, and S. Lathuilière, “Collaborating foundation models for domain generalized semantic segmentation,” in *CVPR*, 2024.

[66] M. Singha, A. Jha, S. Bose, A. Nair, M. Abdar, and B. Banerjee, “Unknown prompt the only lacuna: Unveiling clip’s potential for open domain generalization,” in *CVPR*, 2024.

[67] M. Fahes, T.-H. Vu, A. Bursuc, P. Pérez, and R. De Charette, “Poda: Prompt-driven zero-shot domain adaptation,” in *ICCV*, 2023.

[68] J. Cho, G. Nam, S. Kim, H. Yang, and S. Kwak, “Promptstyler: Prompt-driven style generation for source-free domain generalization,” in *ICCV*, 2023.

[69] S. Yang, Z. Tian, L. Jiang, and J. Jia, “Unified language-driven zero-shot domain adaptation,” in *CVPR*, 2024.

[70] G. Zara, A. Conti, S. Roy, S. Lathuilière, P. Rota, and E. Ricci, “The unreasonable effectiveness of large language-vision models for source-free video domain adaptation,” in *ICCV*, 2023.

[71] Z. Chen, W. Wang, Z. Zhao, F. Su, A. Men, and H. Meng, “Practicaldg: Perturbation distillation on vision-language models for hybrid domain generalization,” in *CVPR*, 2024.

[72] S. Tang, W. Su, M. Ye, and X. Zhu, “Source-free domain adaptation with frozen multimodal foundation model,” in *CVPR*, 2024.

[73] S. Addepalli, A. R. Asokan, L. Sharma, and R. V. Babu, “Leveraging vision-language models for improving domain generalization in image classification,” in *CVPR*, 2024.

[74] X. Zhang, S. S. Gu, Y. Matsuo, and Y. Iwasawa, “Domain prompt learning for efficiently adapting clip to unseen domains,” *arXiv preprint arXiv:2111.12853*, 2021.

[75] C. Ge, R. Huang, M. Xie, Z. Lai, S. Song, S. Li, and G. Huang, “Domain adaptation via prompt learning,” *IEEE Transactions on Neural Networks and Learning Systems*, vol. 36, no. 1, pp. 1160–1170, 2023.

[76] H. Chen, X. Han, Z. Wu, and Y.-G. Jiang, “Multi-prompt alignment for multi-source unsupervised domain adaptation,” in *NeurIPS*, 2023.

[77] D. Cheng, Z. Xu, X. Jiang, N. Wang, D. Li, and X. Gao, “Disentangled prompt representation for domain generalization,” in *CVPR*, 2024.

[78] L. Wang, Y. Jin, Z. Chen, J. Wu, M. Li, Y. Lu, and H. Wang, “Transitive vision-language prompt learning for domain generalization,” *arXiv preprint arXiv:2404.18758*, 2024.

[79] J. Lim and Y. Kim, “Cross-domain semantic segmentation on inconsistent taxonomy using vlms,” *arXiv preprint arXiv:2408.02261*, 2024.

[80] J. Xu, W. Yang, L. Kong, Y. Liu, R. Zhang, Q. Zhou, and B. Fei, “Visual foundation models boost cross-modal unsupervised domain adaptation for 3d semantic segmentation,” *arXiv preprint arXiv:2403.10001*, 2024.

[81] J. Zhu, Y. Chen, and L. Wang, “Clip the divergence: Language-guided unsupervised domain adaptation,” *arXiv preprint arXiv:2407.01842*, 2024.

[82] X. Peng, R. Chen, F. Qiao, L. Kong, Y. Liu, Y. Sun, T. Wang, X. Zhu, and Y. Ma, “Learning to adapt sam for segmenting cross-domain point clouds,” in *ECCV*, 2024.

[83] M. Prabhudesai, T.-W. Ke, A. C. Li, D. Pathak, and K. Fragkiadaki, “Diffusion-tta: Test-time adaptation of discriminative models via generative feedback,” in *NeurIPS*, 2023.

[84] Z. Wei, L. Chen, Y. Jin, X. Ma, T. Liu, P. Ling, B. Wang, H. Chen, and J. Zheng, “Stronger fewer & superior: Harnessing vision foundation models for domain generalized semantic segmentation,” in *CVPR*, 2024.

[85] Z. Lai, H. Bai, H. Zhang, X. Du, J. Shan, Y. Yang, C.-N. Chuah, and M. Cao, “Empowering unsupervised domain adaptation with large-scale pre-trained vision-language models,” in *WACV*, 2024.

[86] X. Hu, K. Zhang, L. Xia, A. Chen, J. Luo, Y. Sun, K. Wang, N. Qiao, X. Zeng, M. Sun *et al.*, “Reclip: Refine contrastive language image pre-training with source free domain adaptation,” in *WACV*, 2024.

[87] R. Xia, Y. Liang, P.-T. Jiang, H. Zhang, B. Li, Y. Tang, and P. Zhou, “Unsupervised modality adaptation with text-to-image diffusion models for semantic segmentation,” *arXiv preprint arXiv:2410.21708*, 2024.

[88] W. Zhang, L. Shen, and C.-S. Foo, "Source-free domain adaptation guided by vision and vision-language pre-training," *International Journal of Computer Vision*, vol. 133, pp. 844–866, 2025.

[89] J. Cha, K. Lee, S. Park, and S. Chun, "Domain generalization by mutual-information regularization with pre-trained models," in *ECCV*, 2022.

[90] Y. Shu, X. Guo, J. Wu, X. Wang, J. Wang, and M. Long, "Clipood: Generalizing clip to out-of-distributions," in *ICML*, 2023.

[91] Z. Lai, N. Vesdapunt, N. Zhou, J. Wu, C. P. Huynh, X. Li, K. K. Fu, and C.-N. Chuah, "Padclip: Pseudo-labeling with adaptive debiasing in clip for unsupervised domain adaptation," in *ICCV*, 2023.

[92] Y. Zhang, L. Wang, K. Li, Y. Wang, and Y. Guo, "Learning representations from foundation models for domain generalized stereo matching," in *ECCV*, 2024.

[93] X. Yu, S. Yoo, and Y. Lin, "Clipceil: Domain generalization through clip via channel refinement and image-text alignment," in *NeurIPS*, 2024.

[94] K. Zhou, J. Yang, C. C. Loy, and Z. Liu, "Conditional prompt learning for vision-language models," *arXiv preprint arXiv:2203.05557*, 2022.

[95] C. Ma, Y. Liu, J. Deng, L. Xie, W. Dong, and C. Xu, "Understanding and mitigating overfitting in prompt tuning for vision-language models," *arXiv preprint arXiv:2211.02219*, 2022.

[96] A. Bulat and G. Tzimiropoulos, "Language-aware soft prompting for vision & language foundation models," *arXiv preprint arXiv:2210.01115*, 2022.

[97] Y. Lu, J. Liu, Y. Zhang, Y. Liu, and X. Tian, "Prompt distribution learning," in *CVPR*, 2022.

[98] M. M. Derakhshani, E. Sanchez, A. Bulat, V. G. T. da Costa, C. G. Snoek, G. Tzimiropoulos, and B. Martinez, "Bayesian prompt learning for image-language model generalization," in *ICCV*, 2023.

[99] H. Bahng, A. Jahanian, S. Sankaranarayanan, and P. Isola, "Exploring visual prompts for adapting large-scale models," *arXiv preprint arXiv:2203.17274*, 2022.

[100] J. Rong, H. Chen, T. Chen, L. Ou, X. Yu, and Y. Liu, "Retrieval-enhanced visual prompt learning for few-shot classification," *arXiv preprint arXiv:2306.02243*, 2023.

[101] J. Wu, X. Li, C. Wei, H. Wang, A. Yuille, Y. Zhou, and C. Xie, "Unleashing the power of visual prompting at the pixel level," *arXiv preprint arXiv:2212.10556*, 2022.

[102] Y. Zang, W. Li, K. Zhou, C. Huang, and C. C. Loy, "Unified vision and language prompt learning," *arXiv preprint arXiv:2210.07225*, 2022.

[103] S. Shen, S. Yang, T. Zhang, B. Zhai, J. E. Gonzalez, K. Keutzer, and T. Darrell, "Multitask vision-language prompt tuning," *arXiv preprint arXiv:2211.11720*, 2022.

[104] M. U. Khattak, H. Rasheed, M. Maaz, S. Khan, and F. S. Khan, "Maple: Multi-modal prompt learning," *arXiv preprint arXiv:2210.03117*, 2022.

[105] Y. Xing, Q. Wu, D. Cheng, S. Zhang, G. Liang, P. Wang, and Y. Zhang, "Dual modality prompt tuning for vision-language pre-trained model," *IEEE Transactions on Multimedia*, vol. 26, pp. 2056–2068, 2023.

[106] G. Wu, X. Zhang, Z. Li, Z. Chen, J. Liang, J. Yang, and X. Li, "Cascade prompt learning for vision-language model adaptation," in *ECCV*, 2024.

[107] P. Gao, S. Geng, R. Zhang, T. Ma, R. Fang, Y. Zhang, H. Li, and Y. Qiao, "Clip-adapter: Better vision-language models with feature adapters," *arXiv preprint arXiv:2110.04544*, 2021.

[108] O. Pantazis, G. Brostow, K. Jones, and O. Mac Aodha, "Svl-adapter: Self-supervised adapter for vision-language pretrained models," *arXiv preprint arXiv:2210.03794*, 2022.

[109] J. Kahana, N. Cohen, and Y. Hoshen, "Improving zero-shot models with label distribution priors," *arXiv preprint arXiv:2212.00784*, 2022.

[110] F. Peng, X. Yang, and C. Xu, "Sgva-clip: Semantic-guided visual adapting of vision-language models for few-shot image classification," *arXiv preprint arXiv:2211.16191*, 2022.

[111] A. Karmanov, D. Guan, S. Lu, A. El Saddik, and E. Xing, "Efficient test-time adaptation of vision-language models," in *CVPR*, 2024.

[112] R. Zhang, R. Fang, P. Gao, W. Zhang, K. Li, J. Dai, Y. Qiao, and H. Li, "Tip-adapter: Training-free clip-adapter for better vision-language modeling," *arXiv preprint arXiv:2111.03930*, 2021.

[113] M. Wortsman, G. Ilharco, J. W. Kim, M. Li, S. Kornblith, R. Roelofs, R. G. Lopes, H. Hajishirzi, A. Farhadi, H. Namkoong *et al.*, "Robust fine-tuning of zero-shot models," in *CVPR*, 2022.

[114] J. Li, S. Savarese, and S. C. Hoi, "Masked unsupervised self-training for zero-shot image classification," *arXiv preprint arXiv:2206.02967*, 2022.

[115] V. Udandarao, A. Gupta, and S. Albanie, "Sus-x: Training-free name-only transfer of vision-language models," *arXiv preprint arXiv:2211.16198*, 2022.

[116] Z. Guo, R. Zhang, L. Qiu, X. Ma, X. Miao, X. He, and B. Cui, "Calip: Zero-shot enhancement of clip with parameter-free attention," *arXiv preprint arXiv:2209.14169*, 2022.

[117] M. Zanella and I. Ben Ayed, "On the test-time zero-shot generalization of vision-language models: Do we really need prompt learning?" in *CVPR*, 2024.

[118] Y. Zhang, W. Zhu, H. Tang, Z. Ma, K. Zhou, and L. Zhang, "Dual memory networks: A versatile adaptation approach for vision-language models," in *CVPR*, 2024.

[119] Z. Wang, J. Liang, L. Sheng, R. He, Z. Wang, and T. Tan, "A hard-to-beat baseline for training-free clip-based adaptation," *arXiv preprint arXiv:2402.04087*, 2024.

[120] S. Pratt, R. Liu, and A. Farhadi, "What does a platypus look like? generating customized prompts for zero-shot image classification," *arXiv preprint arXiv:2209.03320*, 2022.

[121] S. Menon and C. Vondrick, "Visual classification via description from large language models," *arXiv preprint arXiv:2210.07183*, 2022.

[122] S. Parashar, Z. Lin, T. Liu, X. Dong, Y. Li, D. Ramanan, J. Caverlee, and S. Kong, "The neglected tails in vision-language models," in *CVPR*, 2024.

[123] M. Shu, W. Nie, D.-A. Huang, Z. Yu, T. Goldstein, A. Anandkumar, and C. Xiao, "Test-time prompt tuning for zero-shot generalization in vision-language models," in *NeurIPS*, 2022.

[124] D. Osowiechi, M. Noori, G. A. V. Hakim, M. Yazdanpanah, A. Bahri, M. Cheraghalikhani, S. Dastani, F. Beizaee, I. B. Ayed, and C. Desrosiers, "Watt: Weight average test-time adaption of clip," *arXiv preprint arXiv:2406.13875*, 2024.

[125] C.-M. Feng, K. Yu, Y. Liu, S. Khan, and W. Zuo, "Diverse data augmentation with diffusions for effective test-time prompt tuning," *arXiv preprint arXiv:2308.06038*, 2023.

[126] M. Farina, G. Franchi, G. Iacca, M. Mancini, and E. Ricci, "Frustratingly easy test-time adaptation of vision-language models," *arXiv preprint arXiv:2405.18330*, 2024.

[127] X. Ma, J. Zhang, S. Guo, and W. Xu, "Swapprompt: Test-time prompt adaptation for vision-language models," in *NeurIPS*, 2024.

[128] C. Zhou, C. C. Loy, and B. Dai, "Extract free dense labels from clip," in *ECCV*, 2022.

[129] H. Zhang, Y. Su, X. Xu, and K. Jia, "Improving the generalization of segmentation foundation model under distribution shift via weakly supervised adaptation," in *CVPR*, 2024.

[130] Y. Rao, W. Zhao, G. Chen, Y. Tang, Z. Zhu, G. Huang, J. Zhou, and J. Lu, "Denseclip: Language-guided dense prediction with context-aware prompting," in *CVPR*, 2022.

[131] Y. Ganin and V. Lempitsky, "Unsupervised domain adaptation by backpropagation," in *ICML*, 2015.

[132] J. Hoffman, E. Tzeng, T. Park, J.-Y. Zhu, P. Isola, K. Saenko, A. Efros, and T. Darrell, "Cycada: Cycle-consistent adversarial domain adaptation," in *ICML*, 2018.

[133] M. Ghifary, W. B. Kleijn, M. Zhang, D. Balduzzi, and W. Li, "Deep reconstruction-classification networks for unsupervised domain adaptation," in *ECCV*, 2016.

[134] T. Sun, C. Lu, T. Zhang, and H. Ling, "Safe self-refinement for transformer-based domain adaptation," in *CVPR*, 2022.

[135] Z. Cao, L. Ma, M. Long, and J. Wang, "Partial adversarial domain adaptation," in *ECCV*, 2018.

[136] I. Nejjar, H. Dong, and O. Fink, "Recall and refine: A simple but effective source-free open-set domain adaptation framework," *arXiv preprint arXiv:2411.12558*, 2024.

[137] K. Saito, D. Kim, S. Sclaroff, and K. Saenko, "Universal domain adaptation through self supervision," in *NeurIPS*, 2020.

[138] J. Tobin, R. Fong, A. Ray, J. Schneider, W. Zaremba, and P. Abbeel, "Domain randomization for transferring deep neural networks from simulation to the real world," in *IROS*, 2017.

[139] E. Tzeng, J. Hoffman, N. Zhang, K. Saenko, and T. Darrell, "Deep domain confusion: Maximizing for domain invariance," *arXiv preprint arXiv:1412.3474*, 2014.

[140] F. M. Carlucci, A. D'Innocente, S. Bucci, B. Caputo, and T. Tommasi, "Domain generalization by solving jigsaw puzzles," in *CVPR*, 2019.
[141] Y. Shu, Z. Cao, C. Wang, J. Wang, and M. Long, "Open domain generalization with domain-augmented meta-learning," in *CVPR*, 2021.
[142] D. Wang, E. Shelhamer, S. Liu, B. Olshausen, and T. Darrell, "Tent: Fully test-time adaptation by entropy minimization," in *ICLR*, 2021.
[143] L. Yuan, B. Xie, and S. Li, "Robust test-time adaptation in dynamic scenarios," in *CVPR*, 2023.
[144] S. Niu, J. Wu, Y. Zhang, Y. Chen, S. Zheng, P. Zhao, and M. Tan, "Efficient test-time model adaptation without forgetting," in *ICML*, 2022.
[145] Z. Zhou, L.-Z. Guo, L.-H. Jia, D. Zhang, and Y.-F. Li, "Ods: Test-time adaptation in the presence of open-world data shift," in *ICML*, 2023.
[146] Q. Wang, O. Fink, L. Van Gool, and D. Dai, "Continual test-time domain adaptation," in *CVPR*, 2022.
[147] Y. Gan, Y. Bai, Y. Lou, X. Ma, R. Zhang, N. Shi, and L. Luo, "Decorate the newcomers: Visual domain prompt for continual test time adaptation," in *AAAI*, 2023.
[148] J. Liang, R. He, and T. Tan, "A comprehensive survey on test-time adaptation under distribution shifts," *International Journal of Computer Vision*, vol. 133, pp. 1–34, 2025.
[149] Z. Wang, Y. Luo, L. Zheng, Z. Chen, S. Wang, and Z. Huang, "In search of lost online test-time adaptation: A survey," *International Journal of Computer Vision*, pp. 1–34, 2024.
[150] Y. Peng, J. Qi, X. Huang, and Y. Yuan, "Ccl: Cross-modal correlation learning with multigrained fusion by hierarchical network," *IEEE Transactions on Multimedia*, vol. 20, no. 2, pp. 405–420, 2017.
[151] P. Bachman, R. D. Hjelm, and W. Buchwalter, "Learning representations by maximizing mutual information across views," in *NeurIPS*, 2019.
[152] Z. Liu, Y. Shen, V. B. Lakshminarasimhan, P. P. Liang, A. Zadeh, and L.-P. Morency, "Efficient low-rank multimodal fusion with modality-specific factors," *arXiv preprint arXiv:1806.00064*, 2018.
[153] X. Bai, Z. Hu, X. Zhu, Q. Huang, Y. Chen, H. Fu, and C.-L. Tai, "Transfusion: Robust lidar-camera fusion for 3d object detection with transformers," in *CVPR*, 2022.
[154] V. Cirik, L.-P. Morency, and T. Berg-Kirkpatrick, "Visual referring expression recognition: What do systems actually learn?" *arXiv preprint arXiv:1805.11818*, 2018.
[155] S. Haresh, S. Kumar, H. Coskun, S. N. Syed, A. Konin, Z. Zia, and Q.-H. Tran, "Learning by aligning videos in time," in *CVPR*, 2021.
[156] P. Xu, X. Zhu, and D. A. Clifton, "Multimodal learning with transformers: A survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 10, pp. 12 113–12 132, 2023.
[157] P. P. Liang, A. Zadeh, and L.-P. Morency, "Foundations & trends in multimodal machine learning: Principles, challenges, and open questions," *ACM Computing Surveys*, vol. 56, no. 10, pp. 1–42, 2024.
[158] C. Doersch, A. Gupta, and A. A. Efros, "Unsupervised visual representation learning by context prediction," in *ICCV*, 2015.
[159] R. Zhang, P. Isola, and A. A. Efros, "Colorful image colorization," in *ECCV*, 2016.
[160] P. Li, J. Gu, J. Kuen, V. I. Morariu, H. Zhao, R. Jain, V. Manjunatha, and H. Liu, "Selfdoc: Self-supervised document representation learning," in *CVPR*, 2021.
[161] J. Lu, C. Clark, R. Zellers, R. Mottaghi, and A. Kembhavi, "Unified-io: A unified model for vision, language, and multi-modal tasks," in *ICLR*, 2022.
[162] Y. Tian, D. Krishnan, and P. Isola, "Contrastive multiview coding," in *ECCV*, 2020.
[163] X. Yuan, Z. Lin, J. Kuen, J. Zhang, Y. Wang, M. Maire, A. Kale, and B. Faieta, "Multimodal contrastive training for visual representation learning," in *CVPR*, 2021.
[164] R. Arandjelovic and A. Zisserman, "Look, listen and learn," in *ICCV*, 2017.
[165] J.-B. Alayrac, A. Recasens, S. Schneider, R. Arandjelovic, A. Zisserman, and A. Jaegle, "Self-supervised multimodal versatile networks," in *NeurIPS*, 2020.
[166] A. Owens, P. Isola, J. McDermott, A. Torralba, E. Adelson, and W. Freeman, "Ambient sound provides supervision for visual learning," in *ECCV*, 2016.
[167] J. Gui, T. Chen, J. Zhang, Q. Cao, Z. Sun, H. Luo, and D. Tao, "A survey on self-supervised learning: Algorithms, applications, and future trends," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 46, no. 12, pp. 9052–9071, 2024.
[168] L. Jing and Y. Tian, "Self-supervised visual feature learning with deep neural networks: A survey," *IEEE transactions on pattern analysis and machine intelligence*, vol. 43, no. 11, pp. 4037–4058, 2020.
[169] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell *et al.*, "Language models are few-shot learners," in *NeurIPS*, 2020.
[170] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo *et al.*, "Segment anything," in *ICCV*, 2023.
[171] M. Oquab, T. Darcet, T. Moutakanni, H. Vo, M. Szafraniec, V. Khalidov, P. Fernandez, D. Haziza, F. Massa, A. El-Nouby *et al.*, "Dinov2: Learning robust visual features without supervision," *arXiv preprint arXiv:2304.07193*, 2023.
[172] J.-B. Alayrac, J. Donahue, P. Luc, A. Miech, I. Barr, Y. Hasson, K. Lenc, A. Mensch, K. Millican, M. Reynolds *et al.*, "Flamingo: a visual language model for few-shot learning," in *NeurIPS*, 2022.
[173] C. Zhou, Q. Li, C. Li, J. Yu, Y. Liu, G. Wang, K. Zhang, C. Ji, Q. Yan, L. He *et al.*, "A comprehensive survey on pretrained foundation models: A history from bert to chatgpt," *International Journal of Machine Learning and Cybernetics*, pp. 1–65, 2024.
[174] Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. March, and V. Lempitsky, "Domain-adversarial training of neural networks," *Journal of machine learning research*, vol. 17, no. 59, pp. 1–35, 2016.
[175] P. Khosla, P. Teterwak, C. Wang, A. Sarna, Y. Tian, P. Isola, A. Maschinot, C. Liu, and D. Krishnan, "Supervised contrastive learning," in *NeurIPS*, 2020.
[176] L. Yang, Y. Huang, Y. Sugano, and Y. Sato, "Interact before align: Leveraging cross-modal knowledge for domain adaptive action recognition," in *CVPR*, 2022.
[177] H. Caesar, V. Bankiti, A. H. Lang, S. Vora, V. E. Liong, Q. Xu, A. Krishnan, Y. Pan, G. Baldan, and O. Beijbom, "nuscenes: A multimodal dataset for autonomous driving," in *CVPR*, 2020.
[178] J. Geyer, Y. Kassahun, M. Mahmudi, X. Ricou, R. Durgesh, A. S. Chung, L. Hauswald, V. H. Pham, M. Mühlegg, S. Dorn *et al.*, "A2d2: Audi autonomous driving dataset," *arXiv preprint arXiv:2004.06320*, 2020.
[179] J. Behley, M. Garbade, A. Milioto, J. Quenzel, S. Behnke, C. Stachniss, and J. Gall, "Semantickitti: A dataset for semantic scene understanding of lidar sequences," in *ICCV*, 2019.
[180] S. Yun, D. Han, S. Chun, S. J. Oh, Y. Yoo, and J. Choe, "Cutmix: Regularization strategy to train strong classifiers with localizable features," in *ICCV*, 2019.
[181] A. Nekrasov, J. Schult, O. Litany, B. Leibe, and F. Engelmann, "Mix3d: Out-of-context data augmentation for 3d scenes," in *3DV*, 2021.
[182] Y. Wu, M. Xing, Y. Zhang, Y. Xie, J. Fan, Z. Shi, and Y. Qu, "Cross-modal unsupervised domain adaptation for 3d semantic segmentation via bidirectional fusion-then-distillation," in *ACMMM*, 2023.
[183] S. Bultmann, J. Quenzel, and S. Behnke, "Real-time multi-modal semantic fusion on unmanned aerial vehicles with label propagation for cross-domain adaptation," *Robotics and Autonomous Systems*, vol. 159, p. 104286, 2023.
[184] M. Planamente, C. Plizzari, S. A. Peirone, B. Caputo, and A. Bottino, "Relative norm alignment for tackling domain shift in deep multi-modal classification," *International Journal of Computer Vision*, vol. 132, pp. 2618–2638, 2024.
[185] P. Khosla, P. Teterwak, C. Wang, A. Sarna, Y. Tian, P. Isola, A. Maschinot, C. Liu, and D. Krishnan, "Supervised contrastive learning," in *NeurIPS*, 2020.
[186] C. Li, Z. Gan, Z. Yang, J. Yang, L. Li, L. Wang, J. Gao *et al.*, "Multimodal foundation models: From specialists to general-purpose assistants," *Foundations and Trends® in Computer Graphics and Vision*, vol. 16, no. 1-2, pp. 1–214, 2024.
[187] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," in *NeurIPS*, 2020.
[188] P. Esser, R. Rombach, and B. Ommer, "Taming transformers for high-resolution image synthesis," in *CVPR*, 2021.
[189] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *NeurIPS*, 2017.

[190] V. Vidit, M. Engilberge, and M. Salzmann, "Clip the gap: A single domain generalization approach for object detection," in *CVPR*, 2023.
[191] X. Li, Y. Fang, M. Liu, Z. Ling, Z. Tu, and H. Su, "Distilling large vision-language model with out-of-distribution generalizability," in *ICCV*, 2023.
[192] M. Mistretta, A. Baldrati, M. Bertini, and A. D. Bagdanov, "Improving zero-shot generalization of learned prompts via unsupervised knowledge distillation," in *ECCV*, 2024.
[193] Z. Xiao, J. Shen, M. M. Derakhshani, S. Liao, and C. G. Snoek, "Any-shift prompting for generalization over distributions," in *CVPR*, 2024.
[194] D. Li, A. Wu, Y. Wang, and Y. Han, "Prompt-driven dynamic object-centric learning for single domain generalization," in *CVPR*, 2024.
[195] S. Bai, Y. Zhang, W. Zhou, Z. Luan, and B. Chen, "Soft prompt generation for domain generalization," in *ECCV*, 2024.
[196] Z. Du and J. Li, "Diffusion-based probabilistic uncertainty estimation for active domain adaptation," in *NeurIPS*, 2023.
[197] J. Gao, J. Zhang, X. Liu, T. Darrell, E. Shelhamer, and D. Wang, "Back to the source: Diffusion-driven adaptation to test-time corruption," in *CVPR*, 2023.
[198] M. Singh, L. Gustafson, A. Adcock, V. de Freitas Reis, B. Gedik, R. P. Kosaraju, D. Mahajan, R. Girshick, P. Dollár, and L. Van Der Maaten, "Revisiting weakly supervised pre-training of visual perception models," in *CVPR*, 2022.
[199] G. Zara, S. Roy, P. Rota, and E. Ricci, "Autolabel: Clip-based framework for open-set video domain adaptation," in *CVPR*, 2023.
[200] Q. Yu, G. Irie, and K. Aizawa, "Open-set domain adaptation with visual-language foundation models," *Computer Vision and Image Understanding*, vol. 250, p. 104230, 2025.
[201] X. Li, Y. Li, Z. Du, F. Li, K. Lu, and J. Li, "Split to merge: Unifying separated modalities for unsupervised domain adaptation," in *CVPR*, 2024.
[202] P. Liu, W. Yuan, J. Fu, Z. Jiang, H. Hayashi, and G. Neubig, "Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing," *ACM Computing Surveys*, vol. 55, no. 9, pp. 1–35, 2023.
[203] B. Zhu, Y. Niu, Y. Han, Y. Wu, and H. Zhang, "Prompt-aligned gradient for prompt tuning," *arXiv preprint arXiv:2205.14865*, 2022.
[204] X. He, D. Yang, W. Feng, T.-J. Fu, A. Akula, V. Jampani, P. Narayana, S. Basu, W. Y. Wang, and X. E. Wang, "Cpl: Counterfactual prompt learning for vision and language models," *arXiv preprint arXiv:2210.10362*, 2022.
[205] G. Chen, W. Yao, X. Song, X. Li, Y. Rao, and K. Zhang, "Prompt learning with optimal transport for vision-language models," *arXiv preprint arXiv:2210.01253*, 2022.
[206] X. Sun, P. Hu, and K. Saenko, "Dualcoop: Fast adaptation to multi-label recognition with limited annotations," in *NeurIPS*, 2022.
[207] Z. Guo, B. Dong, Z. Ji, J. Bai, Y. Guo, and W. Zuo, "Texts as images in prompt tuning for multi-label image recognition," *arXiv preprint arXiv:2211.12739*, 2022.
[208] K. Ding, Y. Wang, P. Liu, Q. Yu, H. Zhang, S. Xiang, and C. Pan, "Prompt tuning with soft context sharing for vision-language models," *arXiv preprint arXiv:2208.13474*, 2022.
[209] H. Yao, R. Zhang, and C. Xu, "Visual-language prompt tuning with knowledge-guided context optimization," in *CVPR*, 2023.
[210] T.-Y. Wu, C.-H. Ho, and N. Vasconcelos, "Protect: Prompt tuning for taxonomic open set classification," in *CVPR*, 2024.
[211] J. Abdul Samadh, M. H. Gani, N. Hussein, M. U. Khattak, M. M. Naseer, F. Shahbaz Khan, and S. H. Khan, "Align your prompts: Test-time prompting with distribution alignment for zero-shot generalization," in *NeurIPS*, 2023.
[212] M. U. Khattak, S. T. Wasim, M. Naseer, S. Khan, M.-H. Yang, and F. S. Khan, "Self-regulating prompts: Foundational model adaptation without forgetting," in *ICCV*, 2023.
[213] T. Hao, X. Ding, J. Feng, Y. Yang, H. Chen, and G. Ding, "Quantized prompt for efficient generalization of vision-language models," in *ECCV*, 2024.
[214] Z. Lu, F. Shen, M. Liu, Y. Yu, and X. Li, "Improving zero-shot generalization for clip with variational adapter," in *ECCV*, 2024.
[215] T. Chen, L. Zhu, C. Deng, R. Cao, Y. Wang, S. Zhang, Z. Li, L. Sun, Y. Zang, and P. Mao, "Sam-adapter: Adapting segment anything in underperformed scenes," in *ICCVW*, 2023.
[216] C. Chen, J. Miao, D. Wu, A. Zhong, Z. Yan, S. Kim, J. Hu, Z. Liu, L. Sun, X. Li *et al.*, "Ma-sam: Modality-agnostic sam adaptation for 3d medical image segmentation," *Medical Image Analysis*, vol. 98, p. 103310, 2024.
[217] A. Xiao, W. Xuan, H. Qi, Y. Xing, R. Ren, X. Zhang, L. Shao, and S. Lu, "Cat-sam: Conditional tuning for few-shot adaptation of segment anything model," in *ECCV*, 2024.
[218] Y. Ge, J. Ren, A. Gallagher, Y. Wang, M.-H. Yang, H. Adam, L. Itti, B. Lakshminarayanan, and J. Zhao, "Improving zero-shot generalization and robustness of multi-modal models," in *CVPR*, 2023.
[219] R. Zhang, L. Qiu, W. Zhang, and Z. Zeng, "Vt-clip: Enhancing vision-language models with visual-guided texts," *arXiv preprint arXiv:2112.02399*, 2021.
[220] T. Yu, Z. Lu, X. Jin, Z. Chen, and X. Wang, "Task residual for tuning vision-language models," in *CVPR*, 2023.
[221] Y. Ouali, A. Bulat, B. Matinez, and G. Tzimiropoulos, "Black box few-shot adaptation for vision-language models," in *ICCV*, 2023.
[222] M. Zanella and I. Ben Ayed, "Low-rank few-shot adaptation of vision-language models," in *CVPR*, 2024.
[223] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, "Lora: Low-rank adaptation of large language models," *arXiv preprint arXiv:2106.09685*, 2021.
[224] S. Xuan, M. Yang, and S. Zhang, "Adapting vision-language models via learning to inject knowledge," *IEEE Transactions on Image Processing*, vol. 33, pp. 5798–5809, 2024.
[225] Y. Zhang, J. Li, L. Liu, and W. Qiang, "Rethinking misalignment in vision-language model adaptation from a causal perspective," *arXiv preprint arXiv:2410.12816*, 2024.
[226] Z. Lin, S. Yu, Z. Kuang, D. Pathak, and D. Ramanan, "Multimodality helps unimodality: Cross-modal few-shot learning with multimodal models," in *CVPR*, 2023.
[227] G. A. Sigurdsson, A. Gupta, C. Schmid, A. Farhadi, and K. Alahari, "Actor and observer: Joint modeling of first and third-person videos," in *CVPR*, 2018.
[228] K. Soomro, A. R. Zamir, and M. Shah, "Ucf101: A dataset of 101 human actions classes from videos in the wild," *arXiv preprint arXiv:1212.0402*, 2012.
[229] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, and T. Serre, "Hmdb: a large video database for human motion recognition," in *ICCV*, 2011.
[230] J. C. Niebles, C.-W. Chen, and L. Fei-Fei, "Modeling temporal structure of decomposable motion segments for activity classification," in *ECCV*, 2010.
[231] G. Ros, L. Sellart, J. Materzynska, D. Vazquez, and A. M. Lopez, "The synthia dataset: A large collection of synthetic images for semantic segmentation of urban scenes," in *CVPR*, 2016.
[232] S. R. Richter, V. Vineet, S. Roth, and V. Koltun, "Playing for data: Ground truth from computer games," in *ECCV*, 2016.
[233] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, "The cityscapes dataset for semantic urban scene understanding," in *CVPR*, 2016.
[234] F. Yu, H. Chen, X. Wang, W. Xian, Y. Chen, F. Liu, V. Madhavan, and T. Darrell, "Bdd100k: A diverse driving dataset for heterogeneous multitask learning," in *CVPR*, 2020.
[235] G. Neuhold, T. Ollmann, S. Rota Bulo, and P. Kontschieder, "The mapillary vistas dataset for semantic understanding of street scenes," in *ICCV*, 2017.
[236] C. Fang, Y. Xu, and D. N. Rockmore, "Unbiased metric learning: On the utilization of multiple datasets and web images for softening bias," in *ICCV*, 2013.
[237] H. Venkateswara, J. Eusebio, S. Chakraborty, and S. Panchanathan, "Deep hashing network for unsupervised domain adaptation," in *CVPR*, 2017.
[238] D. Li, Y. Yang, Y.-Z. Song, and T. M. Hospedales, "Deeper, broader and artier domain generalization," in *ICCV*, 2017.
[239] X. Peng, Q. Bai, X. Xia, Z. Huang, K. Saenko, and B. Wang, "Moment matching for multi-source domain adaptation," in *ICCV*, 2019.
[240] P. W. Koh, S. Sagawa, H. Marklund, S. M. Xie, M. Zhang, A. Balsubramani, W. Hu, M. Yasunaga, R. L. Phillips, I. Gao *et al.*, "Wilds: A benchmark of in-the-wild distribution shifts," in *ICML*, 2021.
[241] J. Carreira, E. Noland, A. Banki-Horvath, C. Hillier, and A. Zisserman, "A short note about kinetics-600," *arXiv preprint arXiv:1808.01340*, 2018.

[242] P. Sun, H. Kretzschmar, X. Dotiwalla, A. Chouard, V. Patnaik, P. Tsui, J. Guo, Y. Zhou, Y. Chai, B. Caine *et al.*, "Scalability in perception for autonomous driving: Waymo open dataset," in *CVPR*, 2020.
[243] C. Sakaridis, D. Dai, and L. Van Gool, "Acdc: The adverse conditions dataset with correspondences for semantic driving scene understanding," in *ICCV*, 2021.
[244] J. L. Gómez, M. Silva, A. Seoane, A. Borrás, M. Noriega, G. Ros, J. A. Iglesias-Guitian, and A. M. López, "All for one, and one for all: Urbansyn dataset, the third musketeer of synthetic driving scenes," *arXiv preprint arXiv:2312.12176*, 2023.
[245] C. Sakaridis, D. Dai, and L. V. Gool, "Guided curriculum model adaptation and uncertainty-aware evaluation for semantic nighttime image segmentation," in *ICCV*, 2019.
[246] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *CVPR*, 2009.
[247] L. Fei-Fei, R. Fergus, and P. Perona, "Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories," in *CVPRW*, 2004.
[248] O. M. Parkhi, A. Vedaldi, A. Zisserman, and C. Jawahar, "Cats and dogs," in *CVPR*, 2012.
[249] J. Krause, M. Stark, J. Deng, and L. Fei-Fei, "3d object representations for fine-grained categorization," in *ICCVW*, 2013.
[250] M.-E. Nilsback and A. Zisserman, "Automated flower classification over a large number of classes," in *ICVGIP*, 2008.
[251] L. Bossard, M. Guillaumin, and L. Van Gool, "Food-101–mining discriminative components with random forests," in *ECCV*, 2014.
[252] S. Maji, E. Rahtu, J. Kannala, M. Blaschko, and A. Vedaldi, "Fine-grained visual classification of aircraft," *arXiv preprint arXiv:1306.5151*, 2013.
[253] J. Xiao, J. Hays, K. A. Ehinger, A. Oliva, and A. Torralba, "Sun database: Large-scale scene recognition from abbey to zoo," in *CVPR*, 2010.
[254] M. Cimpoi, S. Maji, I. Kokkinos, S. Mohamed, and A. Vedaldi, "Describing textures in the wild," in *CVPR*, 2014.
[255] P. Helber, B. Bischke, A. Dengel, and D. Borth, "Eurosat: A novel dataset and deep learning benchmark for land use and land cover classification," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 12, no. 7, pp. 2217–2226, 2019.
[256] B. Recht, R. Roelofs, L. Schmidt, and V. Shankar, "Do imagenet classifiers generalize to imagenet?" in *ICML*, 2019.
[257] H. Wang, S. Ge, Z. Lipton, and E. P. Xing, "Learning robust global representations by penalizing local predictive power," in *NeurIPS*, 2019.
[258] D. Hendrycks, K. Zhao, S. Basart, J. Steinhardt, and D. Song, "Natural adversarial examples," in *CVPR*, 2021.
[259] D. Hendrycks, S. Basart, N. Mu, S. Kadavath, F. Wang, E. Dorundo, R. Desai, T. Zhu, S. Parajuli, M. Guo *et al.*, "The many faces of robustness: A critical analysis of out-of-distribution generalization," in *ICCV*, 2021.
[260] A. B. Zadeh, P. P. Liang, S. Poria, E. Cambria, and L.-P. Morency, "Multimodal language analysis in the wild: Cmu-mosei dataset and interpretable dynamic fusion graph," in *ACL*, 2018.
[261] C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, and S. S. Narayanan, "Iemocap: Interactive emotional dyadic motion capture database," *Language resources and evaluation*, vol. 42, pp. 335–359, 2008.
[262] S. Poria, D. Hazarika, N. Majumder, G. Naik, E. Cambria, and R. Mihalcea, "Meld: A multimodal multi-party dataset for emotion recognition in conversations," *arXiv preprint arXiv:1810.02508*, 2018.
[263] C. Busso, S. Parthasarathy, A. Burmania, M. AbdelWahab, N. Sadoughi, and E. M. Provost, "Msp-improv: An acted corpus of dyadic interactions to study emotion perception," *IEEE Transactions on Affective Computing*, vol. 8, no. 1, pp. 67–80, 2016.
[264] A. Zadeh, R. Zellers, E. Pincus, and L.-P. Morency, "Mosi: multimodal corpus of sentiment intensity and subjectivity analysis in online opinion videos," *arXiv preprint arXiv:1606.06259*, 2016.
[265] A. Zheng, Z. Wang, Z. Chen, C. Li, and J. Tang, "Robust multimodality person re-identification," in *AAAI*, 2021.
[266] Z. Wang, C. Li, A. Zheng, R. He, and J. Tang, "Interact, embed, and enlarge: Boosting modality-specific representations for multi-modal person re-identification," in *AAAI*, 2022.
[267] H. Li, C. Li, X. Zhu, A. Zheng, and B. Luo, "Multi-spectral vehicle re-identification: A challenge," in *AAAI*, 2020.
[268] A. Wong, X. Fei, S. Tsuei, and S. Soatto, "Unsupervised depth completion from visual inertial odometry," *IEEE Robotics and Automation Letters*, vol. 5, no. 2, pp. 1899–1906, 2020.
[269] N. Silberman, D. Hoiem, P. Kohli, and R. Fergus, "Indoor segmentation and support inference from rgbd images," in *ECCV*, 2012.
[270] J. McCormac, A. Handa, S. Leutenegger, and A. J. Davison, "Scenenet rgb-d: 5m photorealistic images of synthetic indoor trajectories with ground truth," *arXiv preprint arXiv:1612.05079*, 2016.
[271] A. Dai, A. X. Chang, M. Savva, M. Halber, T. Funkhouser, and M. Nießner, "Scannet: Richly-annotated 3d reconstructions of indoor scenes," in *CVPR*, 2017.
[272] N. Mayer, E. Ilg, P. Hausser, P. Fischer, D. Cremers, A. Dosovitskiy, and T. Brox, "A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation," in *CVPR*, 2016.
[273] T. Schops, J. L. Schonberger, S. Galliani, T. Sattler, K. Schindler, M. Pollefeys, and A. Geiger, "A multi-view stereo benchmark with high-resolution images and multi-camera videos," in *CVPR*, 2017.
[274] D. Scharstein, H. Hirschmüller, Y. Kitajima, G. Krathwohl, N. Nešić, X. Wang, and P. Westling, "High-resolution stereo datasets with subpixel-accurate ground truth," in *GCPR*, 2014.
[275] M. Menze, C. Heipke, and A. Geiger, "Object scene flow," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 140, pp. 60–76, 2018.
[276] S. Ben-David, J. Blitzer, K. Crammer, A. Kulesza, F. Pereira, and J. W. Vaughan, "A theory of learning from different domains," *Machine Learning*, vol. 79, pp. 151–175, 2010.
[277] H. Ye, C. Xie, T. Cai, R. Li, Z. Li, and L. Wang, "Towards a theoretical framework of out-of-distribution generalization," in *NeurIPS*, 2021.
[278] Z. Lu, "A theory of multimodal learning," in *NeurIPS*, 2023.
[279] X. Zhang, Y. He, R. Xu, H. Yu, Z. Shen, and P. Cui, "Nico++: Towards better benchmarking for domain generalization," in *CVPR*, 2023.
[280] H. Dong, Y. Zhao, E. Chatzi, and O. Fink, "MultiOOD: Scaling Out-of-Distribution Detection for Multiple Modalities," *arXiv preprint arXiv:2405.17419*, 2024.
[281] S. Li, H. Gong, H. Dong, T. Yang, Z. Tu, and Y. Zhao, "Dpu: Dynamic prototype updating for multimodal out-of-distribution detection," *arXiv preprint arXiv:2411.08227*, 2024.
[282] C. Cortes and M. Mohri, "Domain adaptation in regression," in *International Conference on Algorithmic Learning Theory*, 2011.
[283] I. Nejjar, Q. Wang, and O. Fink, "Dare-gram: Unsupervised domain adaptation regression by aligning inverse gram matrices," in *CVPR*, 2023.
[284] C. Yang, Y. Shen, Z. Zhang, Y. Xu, J. Zhu, Z. Wu, and B. Zhou, "One-shot generative domain adaptation," in *ICCV*, 2023.
[285] Z. Deng, Z. Chen, S. Niu, T. Li, B. Zhuang, and M. Tan, "Efficient test-time adaptation for super-resolution with second-order degradation and reconstruction," in *NeurIPS*, 2023.