Video Audio Domain Generalization via Confounder Disentanglement

Shengyu Zhang[1*], Xusheng Feng[2*], Wenyan Fan[1*], Wenjing Fang[3], Fuli Feng[4], Wei Ji[5], Shuo Li[5], Li Wang[3], Shanshan Zhao[6], Zhou Zhao[1†], Tat-Seng Chua[5], Fei Wu[1,7†]

[1] Zhejiang University
[2] University of Electronic Science and Technology of China
[3] Ant Group
[4] University of Science and Technology of China
[5] National University of Singapore
[6] The University of Sydney
[7] Shanghai AI Laboratory
{sy_zhang, zhaozhou, wufei}@zju.edu.cn {jiwei,dcscts}@nus.edu.sg {engle_xs, wenyan.17}@outlook.com
{bean.fwj,raymond.wangl}@antgroup.com {fulifeng93,weiji0523,sshan.zhao00}@gmail.com

Abstract

Existing video-audio understanding models are trained and evaluated in an intra-domain setting, facing performance degeneration in real-world applications where multiple domains and distribution shifts naturally exist. The key to video-audio domain generalization (VADG) lies in alleviating spurious correlations over multi-modal features. To achieve this goal, we resort to causal theory and attribute such correlation to confounders affecting both video-audio features and labels. We propose a DeVADG framework that conducts uni-modal and cross-modal deconfounding through back-door adjustment. DeVADG performs cross-modal disentanglement and obtains fine-grained confounders at both class-level and domain-level using half-sibling regression and unpaired domain transformation, which essentially identifies domain-variant factors and class-shared factors that cause spurious correlations between features and false labels. To promote VADG research, we collect a VADG-Action dataset for video-audio action recognition with over 5,000 video clips across four domains (*e.g.*, cartoon and game) and ten action classes (*e.g.*, cooking and riding). We conduct extensive experiments, *i.e.*, multi-source DG, single-source DG, and qualitative analysis, validating the rationality of our causal analysis and the effectiveness of the DeVADG framework.

Introduction

Recent AI research (Zhuang et al. 2020) reveals that jointly exploiting video and built-in audio information can facilitate video understanding (Sun et al. 2020; Wu and Yang 2021; Zhao et al. 2020b; Li et al. 2022c,b; Zhang et al. 2020b; Jiang et al. 2022). Most video-audio understanding models are evaluated on independent and identically distributed training/testing data samples. However, real-world platforms (*e.g.*, YouTube) typically have diversified video domains such as advertisement, cartoon, game, and movies.

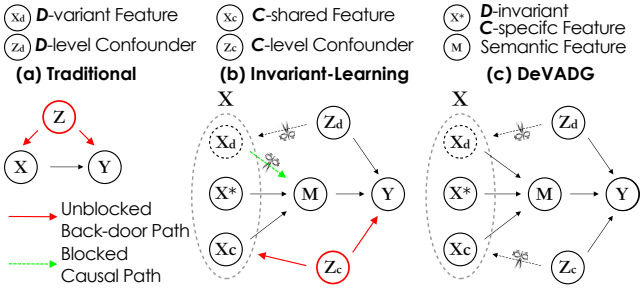


Figure 1: Causal graphs of (a) traditional intra-domain learning; (b) domain-invariant representation learning, which blocks incoming and outgoing effects of domain-variant features; and (c) the proposed DeVADG, which blocks the back-door paths through domain-variant and class-shared features with deconfounding. The transition from a solid line to a dotted line indicates blocking the direct effect.

Existing video-audio understanding models fail to generalize well across diverse domains, encountering a sharp performance drop compared to intra-domain settings (*c.f.*, Table 1). Inferior generalization limits the application of video-audio understanding tools in super video platforms, at the risk of hurting the experience of users with diversified interests beyond major domains. In this paper, we take video-audio action recognition as an example to investigate the video-audio domain generalization problem, where the models are tested in unseen domains.

The open-ended and challenging nature of generic domain generalization has lent itself to diverse models (Li et al. 2022a; Yuan et al. 2021a,c), including domain alignment (Motiian et al. 2017; Zhao et al. 2020a), self-supervised learning (Mahajan, Tople, and Sharma 2021; Kim et al. 2021), and meta-learning based techniques (Du et al. 2021; Xu et al. 2020). Many existing works focus on mining domain-invariant features and thus blocking the impact of domain-variant features that hinder the generalization across domains. However, they are vulnerable to

*These authors contributed equally.
†Corresponding Authors

Intra-domain	$A_r \rightarrow A_e$	$C_r \rightarrow C_e$	$G_r \rightarrow G_e$	$M_r \rightarrow M_e$
SlowFast	56.02	50.14	71.83	60.81
DG	CGM $\rightarrow$ A	AGM $\rightarrow$ C	ACM $\rightarrow$ G	ACG $\rightarrow$ M
SlowFast	23.09	25.25	27.32	17.39

Table 1: Video-audio action recognition with SlowFast model (Xiao et al. 2020; Feichtenhofer et al. 2019) in intra-domain and DG settings with Advertisement, Cartoon, Game, and Movie domains. r and e denote train and test sets.

information loss due to disregarding domain-variant features, which might also affect the label. For example, in playing-instrument videos, there are significantly more pianos in the movie domain, and more guitars in the advertisement domain. As such, instrument-related features might be recognized as domain-variant features and disregarded. Moreover, the existing DG techniques will over-emphasize some domain-invariant features, leading to biased predictions (Tang, Huang, and Zhang 2020). For example, the action feature *put sth. near the mouth* is domain-invariant for the eating action, but models often relate this feature to the playing-instrument (harmonic) action with high confidence (*c.f.*, Figure 8). Besides, most DG techniques are designed for single-modal scenarios, directly applying them to VADG will neglect the cross-modal correlations. Domain generalization with multi-modal data is still ripe for exploration.

We analyze the VADG problem from a causal perspective and abstract the causal relations in Figure 3, where the label (Y) is affected by the semantic features (M) obtained by domain-variant features (X_d) and domain-invariant features (X^* and X_c). We recognize that domain-level confounders (Z_d) affect both video-audio features (X) and labels (Y). For example, the intention of advertisement videos is mostly persuasion or attraction, causing running (Y) videos to have a more colorful appearance and more passionate sounds (X). Domain-level confounders tend to bring spurious correlations between features (*e.g.*, passionate sound) and the label (*e.g.*, running). As the distribution of domain-level confounders varies across domains, such correlations can hardly generalize (*e.g.*, cartoon running videos with peaceful sound). Discarding the domain-variant features is at risk of missing the associated semantic features (M). As such, VADG has to block the back-door path ($X \leftarrow Z_d \rightarrow Y$) while reserving the interaction of features ($X_d \rightarrow M \leftarrow X^*/X_c$). Furthermore, some domain-invariant and class-shared features (X_c, *e.g.*, *put sth. near the mouth* is a feature shared by eating and playing-harmonic actions) might be spuriously correlated with labels due to class-level confounders (Z_c), leading to bias issues (*c.f.*, Figure 8). As such, VADG should also block the back-door path ($X \leftarrow Z_c \rightarrow Y$).

To achieve the goal, we propose a deconfounded video-audio domain generalization framework, namely **DeVADG**, which performs back-door adjustment over the confounders Z_d and Z_c to estimate the causal effect $P(Y|do(X))$. To acquire Z_d and Z_c, we propose cross-modal confounder disentanglement at the domain level and class level. Specifically, domain-level confounders should represent the domain characteristics that make them distinct from other domains. To capture domain-specific characteristics, we train domain transformation networks which should adequately modify domain-specific characteristics in order to fool domain-specific discriminators. As such, modified features could serve as domain-level confounders. Furthermore, we refer to factors that cause the features of different classes as the class-level confounders, which are likely to bring spurious correlations features and a false class label. Technically, we leverage half-sibling regression (Schölkopf et al. 2016) for disentanglement. Both half-sibling regression and domain transformation are performed over multi-modal features to achieve uni-modal and cross-modal disentanglement.

To the best of our knowledge, we take the initiative to investigate the video-audio DG problem. There is a lack of public video datasets with both audio information and a clear domain gap. We thus collect a VADG dataset, named VADG-Action, with four distinct domains, advertisement, cartoon, game, and movie. There are 5,318 video clips belonging to ten action classes (*e.g.*, cooking). To summarize, this paper makes the following key contributions:

- We analyze and address the video-audio domain generalization problem from a causal view, which receives little scrutiny in the domain generalization field.
- We disentangle fine-grained confounders at class-level and domain-level for more accurate back-door adjustment.
- We collect the VADG-Action dataset to promote DG research. Extensive experiments reveal the rationality of the analysis and the effectiveness of DeVADG.

Related Works

Domain Generalization. Diverse models are developed to improve model generalization to unseen domains (Matsuura and Harada 2020; Yuan et al. 2022, 2021b), managing novel situations (Qian et al. 2022) and data samples (Chen et al. 2022a,b). Techniques could vary among meta-learning (Du et al. 2020; Li et al. 2019a,b), self-supervised learning (Jeon et al. 2021), domain alignment (Li et al. 2018; Motiian et al. 2017), and regularization (Liu et al. 2021; Wang et al. 2021, 2019). Typically, SelfReg (Kim et al. 2021) minimizes the distance between the inter-domain intra-class data samples, and generates new positive views by feature interpolation. CausalMatch (Mahajan, Tople, and Sharma 2021) employs contrastive learning with matched samples as positives, and updates the matched pairs per epoch. We differ from existing works by alleviating spurious correlations, and incorporating deconfounding techniques from the causal theory. We do not discard domain-variant features, which might help discover post-interaction knowledge (Pan 2020), but identify the causal effects between features and the outcome.

Causality in Vision. Causal inference is a powerful tool for alleviating spurious correlations and identifying the causal effects that are robust and generalizable (Kuang et al. 2020). In computer vision, existing works that exploit causal inference concern object tracking (Xu et al. 2018), recognition (Lopez-Paz et al. 2017), localization (Yang et al. 2021a), question answering (Li et al. 2022d; Zhong et al. 2022), generation (Kocaoglu et al. 2018; Yang, Zhang, and Cai 2020), and representation learning (Wang et al. 2020; Zhang

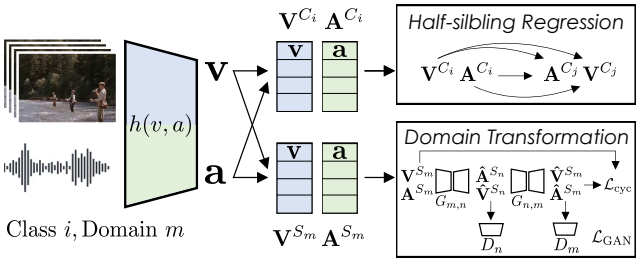


Figure 2: Overall schematic of confounder disentanglement.

et al. 2020a; Yang et al. 2021b). Different from many existing works that exploit off-the-shelf features as confounders, we propose to explicitly disentangle confounders in a cross-modal manner, potentially leading to more fine-grained and accurate deconfounding.

Methods

Problem Definition. Let $\mathcal{X}$ be the video-audio joint feature space, and $\mathcal{Y}$ the outcome (label) space. A *domain* is defined as a joint distribution P_{XY} on $\mathcal{X} \times \mathcal{Y}$. A model is defined as $f : \mathcal{X} \to \mathcal{Y}$, and a loss function is defined as $\ell : \mathcal{Y} \times \mathcal{Y} \to [0, \infty)$. In VADG, we are given K distinct source domains, $\mathcal{S} = \{S_k\}_{k=1}^{K}$, where each source domain accompanies i.i.d. data samples $S_k = \left\{\left(v_i^{(k)}, a_i^{(k)}, y_i^{(k)}\right)\right\}_{i=1}^{N_k}$ drawn from the domain-specific data distribution $P_{XY}^{(k)}$. $v_i^{(k)}$ and $a_i^{(k)}$ are the visual and audio raw information, respectively, in the i-th video of the k-th domain. The goal of VADG is to learn a model f using labeled data in $\mathcal{S}$ such that f can generalize well to an unseen target domain $\mathcal{T}$. Both features and labels in $\mathcal{T}$ are unavailable during training. $P_{XY}^{(k)} \neq P_{XY}^{(k')}, \forall k, k' \in \{1, \ldots, K\}$ and $k \neq k'$. $P_{XY}^{\mathcal{T}} \neq P_{XY}^{(k)}, \forall k \in \{1, \ldots, K\}$.

Back-door Adjustment

As illustrated in the Introduction, one of the evils that hinder the generalization in DG is the confounding effect through the back-door path ($X \leftarrow Z \rightarrow Y$) between the video-audio features X and the label Y, which leads to spurious correlations and is hard to generalize to new domains. To accurately identify the causal effects from features to the outcome, we borrow the back-door adjustment technique from causal theory (Neuberg 2003). The essence of back-door adjustment is to cut the path $X \leftarrow Z$ such that the obtained effects all come from the direct effect $X \rightarrow Y$. The cut operation is formally referred to as the *do*-operation. Originally, the conventional likelihood can be written as:

$$P(Y \mid X) = \sum\nolimits_z P(Y \mid X, z) P(z \mid X), \tag{1}$$

where z indicates a particular confounder sampled from the confounder space $\mathcal{Z}$. We use the Bayes rule to introduce Z in the equation. By using the *do*-operation, the intervened likelihood can be derived as follows:

$$\begin{aligned} P(Y \mid do(X)) &= \sum\nolimits_z P(Y \mid do(X), z) P(z \mid do(X)) \\ &= \sum\nolimits_z P(Y \mid X, z) P(z), \end{aligned} \tag{2}$$

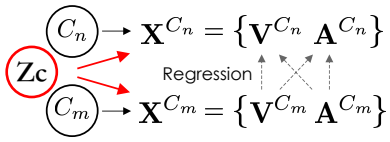


Figure 3: Overall schematic of confounder disentanglement.

where the proof of transitions $P(Y \mid do(X), z) = P(Y \mid X, z)$ and $P(z \mid do(X)) = P(z)$ can be found in (Aronow and Sävje 2020), which is omitted here for brevity. The essence of the transition from Equation (1) to Equation (2) is the transformation of the conditional probability $P(z \mid X)$ to the prior probability of confounders $P(z)$. Intuitively, this transformation cut the path $X \leftarrow Z$ such that $P(Y \mid do(X))$ models the direct effect $X \rightarrow Y$. The expectation calculation is intractable due to the large confounder space $\mathcal{Z}$ and the expensive modeling of $P(Y \mid X, z)$ per z. Therefore, we employ Normalized Weighted Geometric Mean (NWGM) (Wang et al. 2020) for approximation:

$$\begin{aligned} &P(Y \mid \mathrm{do}(X)) = \mathbb{E}_z \left[P(Y \mid X, z)\right] \\ &= \mathbb{E}_z \left[\mathrm{softmax}\left(g(\mathbf{v}, \mathbf{a}, \mathbf{z})\right)\right] \approx \mathrm{softmax}\left(\mathbb{E}_z \left[g(\mathbf{v}, \mathbf{a}, \mathbf{z})\right]\right), \end{aligned} \tag{3}$$

where $P(Y \mid X, z)$ is modeled with a classification network $g(\cdot)$ that takes the video feature $\mathbf{v}$, the audio feature $\mathbf{a}$, and the confounder feature $\mathbf{z}$ as inputs, followed by a softmax layer to obtain the probabilities. Video-audio features are obtained by a video-audio encoding backbone $\mathbf{v}, \mathbf{a} = h(v, a)$, where v and a are raw video and audio input. NWGM essentially moves the expectation operation into the softmax operation. We then model the expectation as follows:

$$\mathbb{E}_z \left[g(\mathbf{v}, \mathbf{a}, \mathbf{z})\right] = \mathbf{W}_1 \left[\mathbf{v} \| \mathbf{a} \| \sum\nolimits_z p(z) \mathbf{z} \phi(\mathbf{z}, \mathbf{v}, \mathbf{a})\right], \tag{4}$$

where $\|$ denotes the concatenation operation, and ϕ is a scoring function for confounder selection, inspired by (Wang et al. 2020). In practice, the prior $p(z)$ is a uniform distribution and $\phi(\mathbf{z}, \mathbf{v}, \mathbf{a})$ is implemented as:

$$\frac{\exp(\mathbf{W}_v \mathbf{v} \cdot \mathbf{W}_{vz} \mathbf{z})}{2 \sum_{z'} \exp(\mathbf{W}_v \mathbf{v} \cdot \mathbf{W}_{vz} \mathbf{z}')} + \frac{\exp(\mathbf{W}_a \mathbf{a} \cdot \mathbf{W}_{az} \mathbf{z})}{2 \sum_{z'} \exp(\mathbf{W}_a \mathbf{a} \cdot \mathbf{W}_{az} \mathbf{z}')}, \tag{5}$$

where $\cdot$ denotes the dot product, and $\mathbf{W}_v, \mathbf{W}_a, \mathbf{W}_{vz}, \mathbf{W}_{az}$ are trainable matrices for feature projection. The remaining problem is how to determine the confounders Z. In this paper, we propose to disentangle potential confounders in a cross-modal manner at the class-level and the domain-level.

Class-level Disentanglement

Recall that class-level confounders lead to spurious correlations between video-audio features and other classes. As a proxy, we propose to disentangle confounders that affect the video-audio features of different classes. Specifically, the causal graph at the class level can be depicted as the top part of Figure **??**, where C_m and C_n represent the class-specific factors of two different classes, $\mathbf{X}^{C_m}$ and $\mathbf{X}^{C_n}$ represent features of data samples from these classes, and Z_c

represents the class-level confounders. $C_m \rightarrow \mathbf{X}^{C_m}$ and $Z_c \rightarrow \mathbf{X}^{C_m}$ mean that features of data samples are affected by the class-specific factors C and class-level confounders Z_c that could affect different classes. We assume that C_m is independent from C_n because they are class-specific factors. With such a causal graph, we can disentangle Z_c from $\mathbf{X}^{C_m}$ and $\mathbf{X}^{C_n}$ using **half-sibling regression (HSR)** (Schölkopf et al. 2016). Technically, we can estimate a weight matrix $\mathbf{W}_{m,n}$ such that:

$$\mathbf{X}^{C_n} \approx \mathbf{X}^{C_m}\mathbf{W}_{m,n}, \tag{6}$$

using ridge regression:

$$\mathbf{W}_{m,n} = \left(\left(\mathbf{X}^{C_m}\right)^\top \mathbf{X}^{C_m} + \alpha_{m,n}\mathbf{I}\right)^{-1}\left(\mathbf{X}^{C_m}\right)^\top \mathbf{X}^{C_n}, \tag{7}$$

where $\mathbf{X}^{C_m}$ denotes the features of all data samples in class m, and $\alpha_{m,n}$ is the regularization constant in ridge regression. Due to the independence of C_m and C_n, the regression results can be identified as confounders according to HSR:

$$\mathbf{Z}_c^{m,n} = \Phi_c(\mathbf{X}^{C_m}, \mathbf{X}^{C_n}) = \mathbf{X}^{C_m}\mathbf{W}_{m,n}. \tag{8}$$

Intuitively, since we can never predict C_n using $\mathbf{X}^{C_m}$ (*e.g.*, predicting *cooking* using a *running* video), the regression results would solely capture Z_c. In video-audio action recognition, the features of data samples (*e.g.*, $\mathbf{X}^{C_m}$) includes video features $\mathbf{V}^{C_m}$ and audio features $\mathbf{A}^{C_m}$. We consider both the uni-modal HSR and the **cross-modal** HSR across classes, as depicted in the bottom part of Figure **??**. Formally, the class-level confounders can be obtained as follows:

$$\mathbf{Z}_c = \{\mathbf{Z}_c^{m,n}; m, n \in [1, J], m \neq n\}, \quad \text{where } \mathbf{Z}_c^{m,n} = \{\Phi_c(\mathbf{U}_1, \mathbf{U}_2); \mathbf{U}_1 \in \{\mathbf{V}^{C_m}, \mathbf{A}^{C_m}\}, \mathbf{U}_2 \in \{\mathbf{V}^{C_n}, \mathbf{A}^{C_n}\}\}. \tag{9}$$

Implementation. In practice, we leverage the video-audio backbone of AVID (Morgado, Vasconcelos, and Misra 2021) that is pretrained on the Kinetics dataset (Kay et al. 2017) to extract video features $\mathbf{V}^C$ and audio features $\mathbf{A}^C$ per class. Upon the confounders $\mathbf{Z}_c$, we perform K-means (KM) clustering and select the N_c clustering centroids as the class-level confounders $\hat{\mathbf{Z}}_c$. The advantage of clustering is to reduce redundancy and obtain representative confounders.

Domain-level Disentanglement

Domains are natural confounders that bring spurious correlations between in-domain video-audio features and class labels. We propose to disentangle domain-specific characteristics to represent domains, and refer to them as potential domain-level confounders. Inspired by Generative Adversarial Networks (Goodfellow et al. 2014), we train one discriminator per domain, which evaluates whether a video-audio feature is from this domain according to the domain-specific knowledge. In addition to discriminators, we train one generator per paired domains to conduct unsupervised domain-transformation. In particular, we employ the CycleGAN (Zhu et al. 2017) architecture, where the cycle consistency could help ensure that the transformed sample is mostly similar to the original one. Intuitively, the transformed factors should be domain-specific characteristics, otherwise the generator cannot fool the domain discriminators. The objective for training the generators and discriminators is given by:

$$\mathcal{L}\left(G_{m,n}, G_{n,m}, D_m, D_n\right) = \mathcal{L}_{\text{GAN}}^{S_m \rightarrow S_n}\left(G_{m,n}, D_n, S_m, S_n\right) + \mathcal{L}_{\text{GAN}}^{S_n \rightarrow S_m}\left(G_{n,m}, D_m, S_n, S_m\right) + \lambda \mathcal{L}_{\text{cyc}}(G_{m,n}, G_{n,m}), \tag{10}$$

where S_m and S_n are two different domains. $G_{m,n}$ and $G_{n,m}$ denote the domain transformation generators for $S_m \rightarrow S_n$ and $S_n \rightarrow S_m$, respectively. D_{S_m} and D_{S_n} denote the domain-specific discriminators. In essence, CycleGAN simultaneously performs the forward transformation $S_m \rightarrow S_n$ and the backward transformation $S_n \rightarrow S_m$ while ensuring the data samples $\hat{S}_m$ obtained from the backward transformation are similar to that of the original S_m, *i.e.*, $\mathcal{L}_{\text{cyc}}(G_{m,n}, G_{n,m})$. The details of losses $\mathcal{L}_{\text{GAN}}$ and $\mathcal{L}_{\text{cyc}}$ can be found in (Zhu et al. 2017). We then obtain the difference between the original feature and the feature after transformation as transformed factors:

$$\mathbf{Z}_d^{m,n} = \Phi_d(\mathbf{X}^{S_m}, \mathbf{X}^{S_n}) = G_{m,n}(\mathbf{X}^{S_m}) - \mathbf{X}^{S_m}, \tag{11}$$

where $\mathbf{X}^{S_m}$ denotes the video-audio features from domain S_m, and $\mathbf{Z}_d^{m,n}$ denotes the transformed factors, which are potential domain-specific characteristics necessary to fool D_{S_n}. $\Phi_d(\mathbf{X}^{S_m}, \mathbf{X}^{S_n})$ means that using $\mathbf{X}^{S_m}, \mathbf{X}^{S_n}$ to train domain generators/discriminators, and then obtaining the transformed factors. Similar to class-level disentanglement, we consider both uni-modal transformation and **cross-modal** transformation across domains.

$$\mathbf{Z}_d = \{\mathbf{Z}_d^{m,n}; m, n \in [1, K], m \neq n\}, \quad \text{where } \mathbf{Z}_d^{m,n} = \{\Phi_d(\mathbf{U}_1, \mathbf{U}_2); \mathbf{U}_1 \in \{\mathbf{V}^{S_m}, \mathbf{A}^{S_m}\}, \mathbf{U}_2 \in \{\mathbf{V}^{S_n}, \mathbf{A}^{S_n}\}\}. \tag{12}$$

Implementation. We use the same backbone introduced in Section for obtaining $\mathbf{V}^S$ and $\mathbf{A}^S$ per domain. The CycleGAN network includes MLP-based generators and discriminators with video-audio features as input. Likewise, we perform K-means on $\mathbf{Z}_d$, resulting in total N_m clustering centroids as the disentangled domain-level confounders $\hat{\mathbf{Z}}_d$.

Model Training

In summary, to achieve deconfounded training, we firstly employ a pretrained video-audio backbone for video-audio feature extraction, and then disentangle class-level confounders $\hat{\mathbf{Z}}_c$ using half-sibling regression (*c.f.* Section). We train domain discriminators for each domain and domain transformation generators for each paired domains, and disentangle domain-level confounders $\hat{\mathbf{Z}}_d$ (*c.f.* Section). Upon the disentangled confounders $\mathbf{Z} = \hat{\mathbf{Z}}_c \cup \hat{\mathbf{Z}}_d$, we perform back-door adjustment to rectify the biased video-audio action recognition with Equation (3). We use cross-entropy loss for optimizing the video-audio feature extraction backbone h and the predictor g:

$$g^*, h^* = \arg\min_{g,h} \sum_{(v,a,y)\in\mathcal{S}} \mathcal{L}_{ce}(\hat{y}, y), \tag{13}$$

$$\hat{y} = \text{softmax}(\mathbb{E}_{\mathbf{z}\in\mathbf{Z}}\left[g(\mathbf{v}, \mathbf{a}, \mathbf{z})\right]), \quad \mathbf{v}, \mathbf{a} = h(v, a), \tag{14}$$

where v and a denote the video and audio raw data.

Experiments

Experiment Details

Dataset. To the best of our knowledge, we take the initiative to investigate VADG, and there are no publicly available datasets accommodating this research. To facilitate evaluation, we collect the VADG-Action dataset containing four distinct domains, ten classes, and 5,318 video-audio clips.

Evaluation. We mainly adopt the leave-one-domain-out evaluation protocol widely used in many other multi-source domain generalization tasks (Jeon et al. 2021; Zhou et al. 2021). As for evaluation metrics, we follow (Morgado, Vasconcelos, and Misra 2021) to report the top-1 accuracy and top-5 accuracy for clip-level and video-level predictions, *i.e.*, Clip@1 (C@1), Clip@5 (C@5), Video@1 (V@1), and Video@5 (V@5).

Baselines. To have a comprehensive evaluation, we consider the three kinds of state-of-the-art methods as baselines:

- **SlowFast (SF)** (Xiao et al. 2020; Feichtenhofer et al. 2019) is a publicly available video-audio action recognition baseline.
- **Video-audio Representation Learning Methods** learn generalizable representations. We incorporate **AVID**, and **CMA** (Morgado, Misra, and Vasconcelos 2021; Morgado, Vasconcelos, and Misra 2021).
- **Domain Generalization Methods.** 1) Invariant Risk Minimization (**IRM**) (Arjovsky et al. 2019), which learns stable properties and invariant causal predictors from multiple environments. 2) **SelfReg (SR)** (Kim et al. 2021), which pulls samples belonging to the same class, and constructs positive views with feature interpolation; 3) **CausalMatch (CM)** (Mahajan, Tople, and Sharma 2021), which iteratively updates the positive matches for each data sample, and distinguishes them from other data samples with contrastive learning.

Overall Comparison (RQ1)

Multi-source Domain Generalization. Table 2 lists results of DeVADG and baselines across different source-target domain settings. Not surprisingly, the vanilla video-audio action recognizer model (SlowFast) achieves inferior results when tested on unseen domains. Video-audio representation learning techniques (AVID, CMA) substantially outperform SlowFast. Task-agnostic self-supervised learning has the advantage of learning fundamental and potentially generalizable spatial-temporal structures and video-audio matching. However, these models neglect the domain gaps and the domain knowledge, thus failing to identify and alleviate spurious correlations among domain-variant features and labels. By explicitly considering the domain labels, SelfReg and CausalMatch construct domain-related self-supervised learning objectives, such as cross-domain contrastive learning. These two models and IRM learn domain-invariant features and disregard many others in the representation space, which can better generalize across domains.

		SF	AVID	CMA	IRM	SR	CM	DeVADG
A	C@1	15.57	27.72	23.74	28.78	26.85	29.00	**30.33***
	C@5	71.03	79.68	69.17	79.81	78.43	81.20	**82.13***
	V@1	23.09	30.41	28.66	33.09	27.65	32.95	**33.89***
	V@5	73.16	80.99	77.14	81.86	80.04	84.18	**85.41***
	Sum	182.85	218.80	198.71	223.54	212.97	227.33	**231.76***
C	C@1	14.89	26.72	26.08	29.90	31.43	31.7	**32.35**
	C@5	65.57	73.64	68.64	79.12	79.92	80.53	**81.5***
	V@1	25.25	28.13	32.45	34.12	34.82	36.00	**37.19***
	V@5	70.44	74.23	77.51	81.13	81.48	82.17	**83.70***
	Sum	177.15	202.73	204.68	224.27	227.65	230.4	**234.74***
G	C@1	15.67	35.11	29.60	33.76	**35.89**	32.78	35.82
	C@5	71.55	78.40	71.55	80.26	77.11	78.22	**80.80***
	V@1	27.32	36.86	34.5	35.63	37.81	36.48	**38.94***
	V@5	75.24	79.30	75.24	82.14	77.98	79.02	**82.23**
	Sum	189.78	229.67	210.88	231.79	228.78	226.5	**237.79***
M	C@1	16.99	23.24	32.42	31.54	33.54	34.29	**35.11***
	C@5	71.76	74.5	76.36	82.69	82.65	83.61	**84.37***
	V@1	17.39	24.26	38.98	34.00	36.90	36.97	**40.64***
	V@5	72.03	75.60	82.58	84.31	85.42	85.90	**87.97***
	Sum	178.17	197.60	230.34	232.54	238.51	240.77	**248.09***

Table 2: Results of DeVADG and video-audio recognition/ representation-learning/DG baselines in four multi-source DG settings. For each setting, we list the testing domain (*e.g.*, A), and omit the training domains (*e.g.*, CGM). We conduct two-sided t-tests and $*$ indicates that the improvements over the strongest baseline are statistically significant with p-value < 0.05 under 5 independent runs.

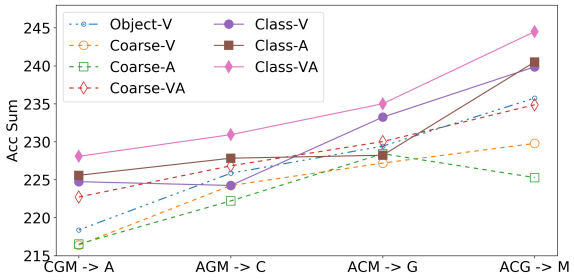


Figure 4: Comparison with coarse-grained deconfounding. Coarse-V/A and Class-V/A denote deconfounding with coarse-grained video/audio confounders and class-level disentangled video/audio confounders. Object-V denotes deconfounding with RoI object confounders.

DeVADG consistently achieves the best results across different source-target domain settings and different metrics in most cases. Different from many state-of-the-art domain generalization baselines, DeVADG does not disregard domain-variant features in video-audio representation, which permits feature interaction and helps to mine action-related high-level semantic knowledge. Instead, DeVADG identifies the causal effects from all features to the outcome (action labels) via deconfounding, where the spurious correlations between the domain-variant features and other labels cannot affect the prediction. These results demonstrate the rationality of our analysis, and the merits of DeVADG on video-audio multi-source domain generalization.

	Acc Sum (Absolute gain from Base)			
	A	M	G	C
A	-	218.0 (+3.27*)	204.9 (+12.31*)	203.2 (+6.83*)
M	211.94 (+ 4.36*)	-	219.2 (-0.24)	209.85 (+3.16*)
G	200.12 (+4.67*)	218.82 (+4.68*)	-	207.99 (+6.07*)
C	207.03 (-1.85)	215.9 (+13.53*)	214.78 (+5.32*)	-

Table 3: Comparison between DeVADG and the Base model in single-source DG settings. Rows and columns represent the source and the target domains, respectively.

Comparison with Coarse-grained Deconfounding. Previous deconfounding methods typically leverage features extracted by off-the-shelf feature extractors as confounders. We construct coarse-grained confounders by extracting video-audio features using the backbone described in Section , and following (Yang et al. 2021b) to perform K-means on these features to have 500 centroids. We also follow (Wang et al. 2020) to construct RoI object confounders. For a fair comparison, we solely consider class-level confounders in DeVADG since domain-level confounder disentanglement requires domain labels. According to Figure 4, we find that DeVADG consistently outperforms the deconfounding with either coarse-grained video/audio confounders or RoI object confounders in most cases, demonstrating the effectiveness of DeVADG. Another finding is that more fine-grained confounders (Videos $\rightarrow$ RoI objects $\rightarrow$ Disentangled factors in DeVADG) mostly lead to better performance, demonstrating the rationality of confounder disentanglement for VADG.

Single-source Domain Generalization. Besides multi-source domain generalization, we are interested in whether DeVADG can be effective in single-source settings where we solely have access to data samples from one domain during training. Table 3 shows the results where rows and columns represent the source and target domains, respectively. Note that the domain-level confounder disentanglement in DeVADG relies on multiple domains. Therefore, we solely leverage the class-level confounders of two modalities for deconfounding. Nevertheless, DeVADG achieves consistent performance gains over the Base model in most cases. The Base model is constructed by removing the deconfounding operation from DeVADG. The results basically reveal that DeVADG is consistently effective in single-source domain generalization settings with class-level deconfounding, which further demonstrates the practical value of DeVADG in complex real-world scenarios.

Model Analysis (RQ2)

Ablation Study. To have a comprehensive understanding of the disentangled confounders, we progressively remove confounders of different levels from DeVADG, and construct different variants, and evaluate their performances. The analysis results of class-level (left) and domain-level (right) multi-modal deconfounding are shown in Figure 5.

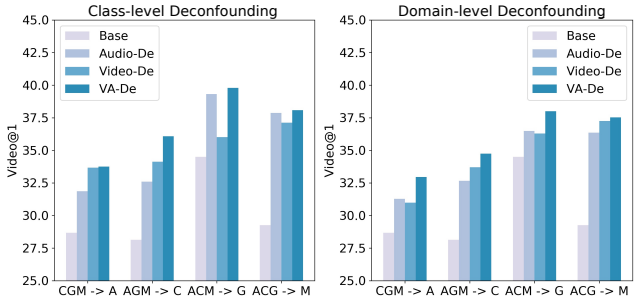


Figure 5: Ablation studies on the disentangled video-audio confounders at the class-level and the domain level.

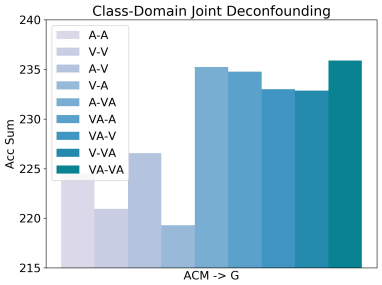


Figure 6: Ablation studies on the disentangled video-audio confounders in class-domain joint deconfounding.

We discuss the empirical performance of deconfounding with uni-modal and multi-modal confounders. For example, **Base** indicates the model without deconfounding, **Audio-De** indicates deconfounding with audio confounders at the class level, **VA-De** indicates deconfounding with video-audio confounders at the class level. According to the results, we observe that 1) deconfounding with any kind of confounders could bring substantial improvements to the Base model, which demonstrates the rationality and effectiveness of confounder disentanglement; 2) each kind of confounders contributes quite differently to the final effectiveness *w.r.t.* different source-target domain settings, which is reasonable since the factors that cause the spurious correlations are different for different domains.

The analysis results of class-domain multi-modal joint deconfounding are shown in Figure 6. For example, **A-V** indicates deconfounding with class-level audio confounders and domain-level video confounders. We find that modeling more confounders (*e.g.*, V-V $\rightarrow$ V-AV) could mostly improve the performance, which again validates the merits of DeVADG. For brevity, we omit the results on their source-target domain settings, where we have similar observations.

Analysis of the Number of Confounders. We are interested in whether and how the number of class-level and domain-level confounders N_c, N_m affects the effectiveness of DeVADG. Towards this end, we keep $N_c = N_m$, vary them in the range $\{5, 20, 50, 200, 500\}$, and obtain their empirical performance in different source-target domain settings. We report the sum of four metrics (Clip1, Clip5,

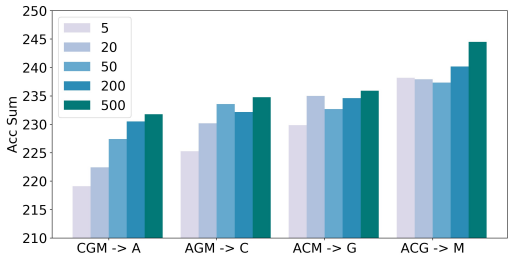


Figure 7: Analysis of how the number of Disentangled Confounders affect the model performance.

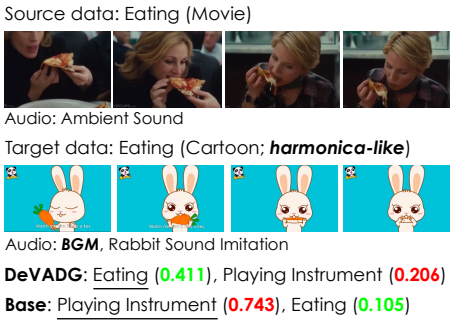
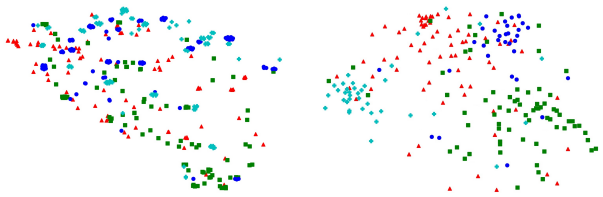


Figure 8: Qualitative examples of Base and DeVADG on the target domain and one source domain.

Video1, Video5), *i.e.*, Acc Sum. Results are depicted in Figure 7. We observe that more confounders generally lead to better performance, which is reasonable since more cluster centroids mean more fine-grained and more comprehensive confounder disentanglement. Comprehensive confounder disentanglement further contributes to an effective estimation of Equation (3). These results further demonstrate the rationality of our confounder disentanglement, which leads to an effective confounder representation space where simple clustering methods could easily find useful confounders for deconfounding.

Qualitative Analysis (RQ3)

Case Study. Figure 8 shows two prediction cases on the testing domains with different source-target settings. In each case, we visualize the sampled frames of the target video and one video sampled from one source domain. We list the Top-2 action classes with the highest probabilities predicted by DeVADG and Base. In summary, without deconfounding, the Base model tends to make false predictions that are spuriously correlated with domain-variant or class-shared video-audio features. DeVADG identifies the causal effects of video-audio features on the outcome (class labels), leading to more accurate predictions. For example, in the left case of Figure 8, the Base model falsely correlates *put sth. near mouth* with the class label *Playing Instrument* with high confidence. In contrast, DeVADG blocks the back-door path from these features to the label and probably identifies that the high-level semantic feature *sth. put near mouth is becoming smaller* is one of the causal features for *Eating*.



(a) Base (b) DeVADG

Figure 9: The visualization displays the t-SNE transformed video-audio joint representation learned by the Base model and DeVADG. Data samples are from four classes in the testing advertisement domain (others for training), and colored according to their action classes.

These results jointly validate that DeVADG achieves better video-audio domain generalization through effective cross-modal deconfounding at the domain level and the class level.

Analysis of the Representation Space. We are interested in how the proposed deconfounding technique facilitates domain-generalizable multi-modal representation learning. As such, we train the Base model, which is DeVADG without deconfounding, and the DeVADG under the CGM (source) $\rightarrow$ A (target) domain setting. We randomly select 400 videos from 4 different classes in the testing domain and extract video-audio joint representations. We perform t-SNE transformation onto these representations, and plot the results in Figure 9a and Figure 9b for Base model and DeVADG, respectively. We color each representation with its action label. According to the results, we observe that the video-audio joint representations of different classes are entangled in the representation space of the Base model. These results probably indicate that video-audio features are easily getting spuriously correlated with the class labels and become less distinguishable when transferring to unseen domains. In contrast, videos of different classes exhibit noticeable clusters in the representation space of DeVADG. These results demonstrate that deconfounding can help to alleviate spurious correlations, and discover more high-level semantic features that are domain-generalizable.

Conclusion

In this paper, we introduce a new problem of video-audio domain generalization, and collect a VADG-Action dataset. We identify that the spurious correlations between video-audio features and class labels are the critical factors hindering generalization. To alleviate spurious correlations, we propose DeVADG which leverages the back-door adjustment and performs deconfouding, which requires knowledge about potential confounders. We disentangle fine-grained class-level and domain-level confounders for both video and audio. Extensive experiments provide insightful analyses of the rationality and effectiveness of DeVADG.

In the future, we plan to investigate unobserved confounders beyond observed features. Furthermore, we will explore disentangling domain-level confounders given one domain via automatic domain partition.

Acknowledgments

The work is supported by the National Natural Science Foundation of China (No. U20A20387, 62037001), National Key R & D Projects of the Ministry of Science and Technology (2020YFC0832500), Zhejiang Natural Science Foundation (No. LR19F020006), and Project by Shanghai AI Laboratory (No. P22KS00111).

References

Arjovsky, M.; Bottou, L.; Gulrajani, I.; and Lopez-Paz, D. 2019. Invariant Risk Minimization. *CoRR*.

Aronow, P. M.; and Sävje, F. 2020. *The Book of Why: The New Science of Cause and Effect: Judea Pearl and Dana Mackenzie. New York: Basic Books, 2018, x+ 418 pp., ISBN: 978-0-46-509760-9.* Taylor & Francis.

Chen, X.; Li, L.; Zhang, N.; Liang, X.; Deng, S.; Tan, C.; Huang, F.; Si, L.; and Chen, H. 2022a. Decoupling Knowledge from Memorization: Retrieval-augmented Prompt Learning. *CoRR*, abs/2205.14704.

Chen, X.; Zhang, N.; Xie, X.; Deng, S.; Yao, Y.; Tan, C.; Huang, F.; Si, L.; and Chen, H. 2022b. KnowPrompt: Knowledge-aware Prompt-tuning with Synergistic Optimization for Relation Extraction. In *WWW*, 2778–2788. ACM.

Du, Y.-J.; Xu, J.; Xiong, H.; Qiu, Q.; Zhen, X.; Snoek, C. G. M.; and Shao, L. 2020. Learning to Learn with Variational Information Bottleneck for Domain Generalization. In *ECCV*.

Du, Z.; Li, J.; Lu, K.; Zhu, L.; and Huang, Z. 2021. Learning Transferrable and Interpretable Representations for Domain Generalization. In *ACM MM*.

Feichtenhofer, C.; Fan, H.; Malik, J.; and He, K. 2019. SlowFast Networks for Video Recognition. In *ICCV*.

Goodfellow, I. J.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A. C.; and Bengio, Y. 2014. Generative Adversarial Nets. In *NIPS*.

Jeon, S.; Hong, K.; Lee, P.; Lee, J.; and Byun, H. 2021. Feature Stylization and Domain-aware Contrastive Learning for Domain Generalization. In *ACM MM*.

Jiang, Z.; Zhang, S.; Yao, S.; Zhang, W.; Zhang, S.; Li, J.; Zhao, Z.; and Wu, F. 2022. Weakly-supervised Disentanglement Network for Video Fingerspelling Detection. In *MM '22: The 30th ACM International Conference on Multimedia, Lisboa, Portugal, October 10 - 14, 2022*, 5446–5455. ACM.

Kay, W.; Carreira, J.; Simonyan, K.; Zhang, B.; Hillier, C.; Vijayanarasimhan, S.; Viola, F.; Green, T.; Back, T.; Natsev, P.; and et al. 2017. The Kinetics Human Action Video Dataset. *CoRR*.

Kim, D.; Yoo, Y.; Park, S.; Kim, J.; and Lee, J. 2021. SelfReg: Self-supervised Contrastive Regularization for Domain Generalization. In *ICCV*.

Kocaoglu, M.; Snyder, C.; Dimakis, A. G.; and Vishwanath, S. 2018. CausalGAN: Learning Causal Implicit Generative Models with Adversarial Training. In *ICLR*.

Kuang, K.; Li, L.; Geng, Z.; Xu, L.; Zhang, K.; Liao, B.; Huang, H.; Ding, P.; Miao, W.; and Jiang, Z. 2020. Causal inference. *Engineering*, 6(3): 253–263.

Li, B.; Shen, Y.; Wang, Y.; Zhu, W.; Reed, C.; Li, D.; Keutzer, K.; and Zhao, H. 2022a. Invariant Information Bottleneck for Domain Generalization. In *AAAI*, 7399–7407.

Li, D.; Zhang, J.; Yang, Y.; Liu, C.; Song, Y.-Z.; and Hospedales, T. M. 2019a. Episodic Training for Domain Generalization. In *ICCV*.

Li, M.; Wang, T.; Zhang, H.; Zhang, S.; Zhao, Z.; Miao, J.; Zhang, W.; Tan, W.; Wang, J.; Wang, P.; Pu, S.; and Wu, F. 2022b. End-to-End Modeling via Information Tree for One-Shot Natural Language Spatial Video Grounding. In *ACL*.

Li, M.; Wang, T.; Zhang, H.; Zhang, S.; Zhao, Z.; Zhang, W.; Miao, J.; Pu, S.; and Wu, F. 2022c. HERO: HiErarchical spatio-tempoRal reasOning with Contrastive Action Correspondence for End-to-End Video Object Grounding. In *ACM MM*.

Li, Y.; Gong, M.; Tian, X.; Liu, T.; and Tao, D. 2018. Domain Generalization via Conditional Invariant Representations. In *AAAI*.

Li, Y.; Wang, X.; Xiao, J.; Ji, W.; and Chua, T.-S. 2022d. Invariant grounding for video question answering. In *CVPR*, 2928–2937.

Li, Y.; Yang, Y.; Zhou, W.; and Hospedales, T. M. 2019b. Feature-Critic Networks for Heterogeneous Domain Generalization. In *ICML*.

Liu, C.; Wang, L.; Li, K.; and Fu, Y. 2021. Domain Generalization via Feature Variation Decorrelation. In *ACM MM*.

Lopez-Paz, D.; Nishihara, R.; Chintala, S.; Schölkopf, B.; and Bottou, L. 2017. Discovering Causal Signals in Images. In *CVPR*, 58–66.

Mahajan, D.; Tople, S.; and Sharma, A. 2021. Domain Generalization using Causal Matching. In *ICML*.

Matsuura, T.; and Harada, T. 2020. Domain Generalization Using a Mixture of Multiple Latent Domains. In *AAAI*, 11749–11756.

Morgado, P.; Misra, I.; and Vasconcelos, N. 2021. Robust Audio-Visual Instance Discrimination. In *CVPR*.

Morgado, P.; Vasconcelos, N.; and Misra, I. 2021. Audio-Visual Instance Discrimination with Cross-Modal Agreement. In *CVPR*.

Motiian, S.; Piccirilli, M.; Adjeroh, D. A.; and Doretto, G. 2017. Unified Deep Supervised Domain Adaptation and Generalization. In *ICCV*.

Neuberg, L. G. 2003. Causality: models, reasoning, and inference, by judea pearl, cambridge university press, 2000. *Econometric Theory*, 19(4): 675–685.

Pan, Y. 2020. Multiple knowledge representation of artificial intelligence. *Engineering*, 6(3): 216–217.

Qian, X.; Xu, Y.; Lv, F.; Zhang, S.; Jiang, Z.; Liu, Q.; Zeng, X.; Chua, T.; and Wu, F. 2022. Intelligent Request Strategy Design in Recommender System. In *KDD '22: The 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 3772–3782. ACM.

Schölkopf, B.; Hogg, D. W.; Wang, D.; Foreman-Mackey, D.; Janzing, D.; Simon-Gabriel, C.-J.; and Peters, J. 2016. Modeling confounding by half-sibling regression. *Proc. Natl. Acad. Sci. USA*.

Sun, Z.; Sarma, P. K.; Sethares, W. A.; and Liang, Y. 2020. Learning Relationships between Text, Audio, and Video via Deep Canonical Correlation for Multimodal Language Analysis. In *AAAI*, 8992–8999.

Tang, K.; Huang, J.; and Zhang, H. 2020. Long-Tailed Classification by Keeping the Good and Removing the Bad Momentum Causal Effect. In *NeurIPS*.

Wang, H.; He, Z.; Lipton, Z. C.; and Xing, E. P. 2019. Learning Robust Representations by Projecting Superficial Statistics Out. In *ICLR*.

Wang, T.; Huang, J.; Zhang, H.; and Sun, Q. 2020. Visual Commonsense R-CNN. In *CVPR*.

Wang, Y.; Li, H.; Chau, L.-P.; and Kot, A. C. 2021. Embracing the Dark Knowledge: Domain Generalization Using Regularized Knowledge Distillation. In *ACM MM*.

Wu, Y.; and Yang, Y. 2021. Exploring Heterogeneous Clues for Weakly-Supervised Audio-Visual Video Parsing. In *CVPR*.

Xiao, F.; Lee, Y. J.; Grauman, K.; Malik, J.; and Feichtenhofer, C. 2020. Audiovisual SlowFast Networks for Video Recognition. *CoRR*.

Xu, H.; Xie, H.; Zha, Z.-J.; Liu, S.; and Zhang, Y. 2020. March on Data Imperfections: Domain Division and Domain Generalization for Semantic Segmentation. In *ACM MM*.

Xu, Y.; Qin, L.; Liu, X.; Xie, J.; and Zhu, S.-C. 2018. A Causal And-Or Graph Model for Visibility Fluent Reasoning in Tracking Interacting Objects. In *CVPR*, 2178–2187.

Yang, X.; Feng, F.; Ji, W.; Wang, M.; and Chua, T.-S. 2021a. Deconfounded video moment retrieval with causal intervention. In *SIGIR*, 1–10.

Yang, X.; Zhang, H.; and Cai, J. 2020. Deconfounded Image Captioning: A Causal Retrospect. *CoRR*.

Yang, X.; Zhang, H.; Qi, G.; and Cai, J. 2021b. Causal Attention for Vision-Language Tasks. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*.

Yuan, J.; Ma, X.; Chen, D.; Kuang, K.; Wu, F.; and Lin, L. 2021a. Collaborative Semantic Aggregation and Calibration for Separated Domain Generalization. *arXiv e-prints*, arXiv–2110.

Yuan, J.; Ma, X.; Chen, D.; Kuang, K.; Wu, F.; and Lin, L. 2021b. Domain-Specific Bias Filtering for Single Labeled Domain Generalization. *arXiv preprint arXiv:2110.00726*.

Yuan, J.; Ma, X.; Chen, D.; Kuang, K.; Wu, F.; and Lin, L. 2022. Label-Efficient Domain Generalization via Collaborative Exploration and Generalization. *ACM MM*.

Yuan, J.; Ma, X.; Kuang, K.; Xiong, R.; Gong, M.; and Lin, L. 2021c. Learning Domain-Invariant Relationship with Instrumental Variable for Domain Generalization. *ArXiv*, abs/2110.01438.

Zhang, S.; Jiang, T.; Wang, T.; Kuang, K.; Zhao, Z.; Zhu, J.; Yu, J.; Yang, H.; and Wu, F. 2020a. DeVLBert: Learning Deconfounded Visio-Linguistic Representations. In *ACM MM*.

Zhang, S.; Tan, Z.; Zhao, Z.; Yu, J.; Kuang, K.; Jiang, T.; Zhou, J.; Yang, H.; and Wu, F. 2020b. Comprehensive Information Integration Modeling Framework for Video Titling. In *KDD '20: The 26th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2744–2754. ACM.

Zhao, S.; Gong, M.; Liu, T.; Fu, H.; and Tao, D. 2020a. Domain Generalization via Entropy Regularization. In *NeurIPS*.

Zhao, S.; Ma, Y.; Gu, Y.; Yang, J.; Xing, T.; Xu, P.; Hu, R.; Chai, H.; and Keutzer, K. 2020b. An End-to-End Visual-Audio Attention Network for Emotion Recognition in User-Generated Videos. In *AAAI*, 303–311.

Zhong, Y.; Ji, W.; Xiao, J.; Li, Y.; Deng, W.; and Chua, T.-S. 2022. Video Question Answering: Datasets, Algorithms and Challenges. *EMNLP*.

Zhou, K.; Liu, Z.; Qiao, Y.; Xiang, T.; and Loy, C. C. 2021. Domain generalization in vision: A survey. *arXiv preprint arXiv:2103.02503*.

Zhu, J.-Y.; Park, T.; Isola, P.; and Efros, A. A. 2017. Unpaired Image-to-Image Translation Using Cycle-Consistent Adversarial Networks. In *ICCV*.

Zhuang, Y.; Cai, M.; Li, X.; Luo, X.; Yang, Q.; and Wu, F. 2020. The next breakthroughs of artificial intelligence: The interdisciplinary nature of AI. *Engineering*, 6(3): 245.