Background no more: Action recognition across domains by causal interventions

Rastegar, S.; Doughty, H.; Snoek, C.G.M.

DOI
10.1016/j.cviu.2024.103975

Publication date
2024

Document Version
Final published version

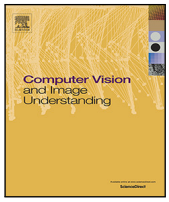
Published in
Computer Vision and Image Understanding

Computer Vision and Image Understanding 242 (2024) 103975

Contents lists available at ScienceDirect

Computer Vision and Image Understanding

journal homepage: www.elsevier.com/locate/cviu



Background no more: Action recognition across domains by causal interventions

Sarah Rastegar *, Hazel Doughty, Cees G.M. Snoek

University of Amsterdam, Science Park 900, Amsterdam 1098 XH, The Netherlands

ARTICLE INFO

Communicated by Yu-Chiang Frank Wang

MSC:
41A05
41A10
65D05
65D17

Keywords:
Causal inference
Domain generalization
Background bias removal

ABSTRACT

We aim to recognize actions under an appearance distribution shift between a source training domain and a target test domain. To enable such video domain generalization, our key idea is to intervene on the action to remove the confounding effect of the domain background on the class label using causal inference. Towards this, we propose to learn a causally debiased model on a source domain that intervenes on the action through three possible *Do*-operators, which separate the action and background. To better align the source and target distributions, we also introduce a test-time action intervention. Experiments on two challenging video domain generalization benchmarks reveal that causal inference is a promising tool for action recognition as it already achieves state-of-the-art results on Kinetics2Mimetics, the benchmark with the largest domain shift.

1. Introduction

In this work, we tackle the challenging problem of video domain generalization. Traditional action recognition works, *e.g.*, Carreira and Zisserman (2017), Lin et al. (2019) and Feichtenhofer et al. (2019) typically assume training and test videos come from the same distribution. Instead, the goal in domain generalization is to train a model on a source dataset such that it can generalize to an unseen target dataset from a different distribution at test time. A large difference between source and target domains is often the environmental context in which actions happen (Choi et al., 2019; Zou et al., 2022; Ding et al., 2022). This means a model may learn shortcuts in the video's background, rather than use an action's appearance to classify actions. These shortcuts are often not applicable in the target domain, meaning the model will generalize poorly.

Some previous research like Kim et al. (2022), Hwang et al. (2023), Varol et al. (2021) and de Souza et al. (2018) employ Probabilistic Graphical Models (PGM) or synthetic video generation to directly create new action videos, but two main limitations exist. First, due to the synthetic nature of the action segments in these datasets, there is an evident gap between the source and the target domains concerning these action parts. Contrarily, our method operates under the presumption that the action components in both domains are analogous, which is a robust assumption when considering real videos in both the source and target domains. Second, the synthetic methodology, despite allowing for numerous video combinations, restricts the diversity of action due to its simplistic rendering when compared to real-world videos. Our approach, which utilizes real datasets, surpasses this limitation, offering enhanced scalability without the necessity of designing intricate prototypes for every new action. Several recent studies, such as Zhang et al. (2023), Wang et al. (2023), Zoetgnande and Dillenseger (2022) and Planamente et al. (2022a), tackle video domain generalization but lean on additionally available modalities for assistance. It is worth noting that the work presented by Hasan et al. (2022) centers on pedestrian detection, which is distinct from our primary objective of action recognition domain generalization. Additionally, Majumdar et al. (2022) present a new dataset and apply image augmentation strategies to address the domain generalization challenge. In general, inferring one modality from the other can be seen as a form of domain generalization. However, note that many of these works rely on the fact that one of the domains usually suffers less from the domain gap. In this scenario, the problem could be considered more in line with cross-modal knowledge transfer when one modality is missing or corrupted at the test time, which has been investigated by Park et al. (2023), Andonian et al. (2022), Xue et al. (2020), Alwassel et al. (2020), Roheda et al. (2018), Rastegar et al. (2016), Sohn et al. (2014) and Ngiam et al. (2011). To the best of our knowledge, there seems to be a scarcity of works specifically focusing on single-modality video domain generalization. To tackle the problem of recognizing actions across domains, we aim to eliminate background bias in the source and target domains using causal interventions where we intervene on an action by changing the background it appears with (see Fig. 1).

* Corresponding author.
E-mail address: s.rastegar2@uva.nl (S. Rastegar).

https://doi.org/10.1016/j.cviu.2024.103975
Received 2 December 2022; Received in revised form 19 January 2024; Accepted 20 February 2024
Available online 28 February 2024

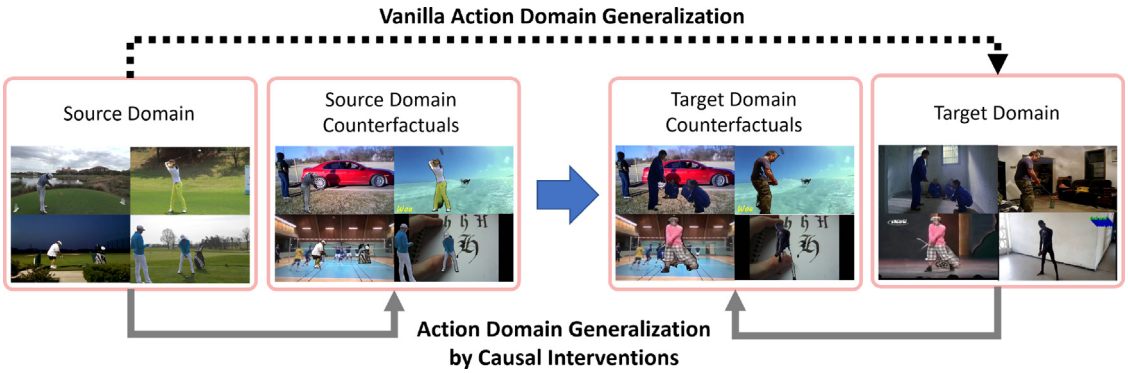


Fig. 1. Action recognition across domains by causal interventions. We create surrogate counterfactual domains for both the source and target domains with the same background distribution while the actions remain domain-specific. The model that is learned on the surrogate source domain generalizes better to the surrogate target domain. We create counterfactuals of actions in training where actions are combined with a variety of different backgrounds. This avoids a model relying on the background to recognize an action, making the model more generalizable to the target domain. By also creating counterfactuals at test time, we reduce the distribution difference between source and target.

There has been a recent surge of interest in adopting causal inference for computer vision *e.g.*, Johnson et al. (2017) and Liu et al. (2022a, 2021b). Some of these works discover causal effects (Johnson et al., 2017), while others estimate and eliminate context features to remove bias (Liu et al., 2022a). We follow the latter approach where causal interventions eliminate spurious correlations between variables. There have been a handful of works that also use causal interventions in the video, although these works focus on other problems such as object grounding (Wang et al., 2022a), moment retrieval (Yang et al., 2021; Nan et al., 2021) and action localization (Liu et al., 2021b).

To the best of our knowledge, we are the first to explore causal inference for video domain generalization and prevent a model from using background shortcuts to recognize actions. Our work offers four main contributions:

- We introduce a Structural Causal Model that identifies background as a confounding variable, which prevents the generalization of action recognition across different domains.
- We present a backbone-agnostic method based on causal relations to reduce this background bias. Specifically, we employ causal interventions during the training phase to achieve background-debiased learning. We accomplish this by intervening on the action variable and introducing three unique operators to implement this intervention.
- We take this a step further by optimizing our method to allow action interventions also at test-time, with no requirement for modifications to the existing model.
- We propose a novel Fourier-based background filtering technique. By considering specific frequency bands as background, we can effectively reduce the background effects and other irrelevant static features like clothing and objects.

2. Related works

Video Domain Generalization. Many works aim to generalize between different image domains (Muandet et al., 2013; Qiao et al., 2020; Zhao et al., 2020; Chuah et al., 2022; Wan et al., 2022; Peng et al., 2022; Zhou et al., 2021). For instance, Qiao et al. (2020) proposed M-ADA, a meta-learning method which uses adversarial domain augmentations. Unlike video domain adaptation (Munro and Damen, 2020; Zhang et al., 2022; Chen et al., 2019; Kim et al., 2021; Sahoo et al., 2021), video domain generalization has seen much less attention. Some works use the more generalizable audio modality to encourage generalization (Planamente et al., 2022a,b). More similar to ours are works that use only the visual modality (Yao et al., 2021; Bahng et al., 2020; Bao et al., 2021; Li and Vasconcelos, 2022). Yao et al. (2021) propose APN, which progressively captures local and global temporal relations in a pyramid network to better align video events in different domains. To further improve robustness, Yao et al. (2021) also propose RADA which generates temporally adversarial augmentations. Rather than using augmentations, Bahng et al. (2020) encourages independence between learned biased and debiased representations. Similarly, DEAR (Bao et al., 2021) contrasts bias representations from single frames with video representations from a 3D CNN. DRL (Li and Vasconcelos, 2022) also compares 2D and 3D models in an adversarial student–teacher framework where a 3D teacher is optimized to give better predictions than the 2D student. Unlike all these works, we introduce a causal approach to generalize between video domains by forcing our model to see the same action with different backgrounds.

Removing Background-Bias in Videos. The effect of background bias in action recognition has recently been gaining more attention. Ilic et al. (2022) showed a notable decrease in performance when training on an appearance-free dataset. Other works (Kim et al., 2022; Thoker et al., 2022) also observe that models trained on datasets with object and scene bias transfer poorly to datasets without these biases. Several prior action recognition works have removed the correlation between background and action (Choi et al., 2019; Ding et al., 2022; Zou et al., 2022). Choi et al. (2019) encourage their model to fail to predict actions on videos where human bounding boxes are masked. Zou et al. (2022) instead paste the contents of human bounding boxes on other videos so actions are seen with varied backgrounds. Ding et al. (2022) propose a segmentation-based background augmentation but use it for self-supervised learning. Gowda et al. (2022) also propose a segmentation-based background augmentation, but sample backgrounds from semantically similar classes, thus maintaining the background bias. We instead aim to eliminate background bias with causal inference as a means to generalize between video domains.

Causal Inference for Vision. Causal inference (Pearl, 2009) has recently attracted increased attention in computer vision (Liu et al., 2022a; Yue et al., 2021a; Mahajan et al., 2021; Yang et al., 2021; Liu et al., 2022; Chen et al., 2022; Dash et al., 2022; Liu et al., 2022c). Since causal effects are invariant for different domains (Peters et al., 2017; Zhang et al., 2015), they are effective for either learning domain invariant representations (Lv et al., 2022; Magliacane et al., 2018; Yue et al., 2021a; Wang et al., 2022b) or augmenting data with counterfactuals (Zhang et al., 2021b; Yue et al., 2021b; Chen et al., 2020). For video, causal inference has been exploited to eliminate semantic bias for action anticipation (Zhang et al., 2021a) relation detection (Li et al., 2021), moment retrieval (Yang et al., 2021; Nan et al., 2021), and action detection (Liu et al., 2021b). To the best of our knowledge, we are the first to explore how causal inference can address the video domain generalization challenge.

Test-Time Adaptation. Recent domain generalization work for images has explored how to adapt a model with target samples at test-time, without ever using target data during training. This may be achieved

by self-supervision (Sun et al., 2020), entropy minimization (Wang et al., 2021) or variational Bayesian inference (Xiao et al., 2022). Notably, Sun et al. (2020) modify their loss function to depend on the test input without looking at the label. This is feasible by casting it as a self-supervised learning problem that updates model parameters before the final classification. Unlike these works, we introduce an (unsupervised) causally-inspired test-time adaptation for action recognition.

3. Approach

For action recognition under domain generalization, we have a source domain S consisting of videos $x \in X^S$ and their corresponding labels $y \in Y^S$. Our goal is to learn an action classifier on these videos and labels such that it can correctly predict the action labels $y \in Y^T$ for videos X^T from an unseen target domain T. We assume the label space is shared between source and target domains; however, the distribution of labels and video appearance will differ. To learn a model that better generalizes between source and target, we want to remove the spurious correlation between the background and action label. Towards this goal, we first define a simple Structural Causal Model for action recognition under domain generalization.

3.1. Structural causal model

Our Structural Causal Model is depicted in Fig. 2. It is a directed acyclic graph indicating the causalities between domain, action labels, and videos, which we decompose into action and background. Nodes represent causal entities, and edges represent the causal effect of the parent node on the child, *e.g.*, Action→Label indicates that changing the action leads to changes in the label. We first explain the rationale behind the nodes and edges before describing how we use them.

Domain→(Action, Background): The domain causes both the appearance of the background and the action (as captured in a video). For instance, in Fig. 2, we have a video of someone performing the 'playing soccer' action in the soccer pitch background from the domain of sports videos.

(Action, Background)→Label: The action variable is the video part with a direct causal effect on the ground-truth label; however, our predicted label is affected by both background and action. For instance, in Fig. 2, a model can create a shortcut from the soccer pitch background to predict 'playing soccer.' There are further elements that can have spurious correlations with the action label, such as clothing and viewpoint. In this paper, we focus on the background confounder exclusively.

In the causal graph, we can see that the domain confounds the action and label by creating a correlation between the label and the background through the backdoor path $Action \leftarrow Domain \rightarrow Background \rightarrow Label$. This path makes our prediction dependent on the domain and its background, which is problematic when generalizing to unseen domains where the background often changes. Next, we explain how we remove this backdoor path by intervening on the action.

3.2. Background-debiased learning

Assume we have a standard supervised action recognition model which we train on the source domain $S=\{(x, y)\}$ using videos $x \in X^S$ and their corresponding ground-truth labels $y \in Y^S$. In training, the goal is to maximize the probability of the model predicting the correct label y given video x, *i.e.*, $P_S(y \mid X{=}x)$. Since a video x consists of a background bg and action a, we can re-write this as maximizing the probability of predicting the label y given action a in video x, *i.e.*, $P_S(y \mid A{=}a)$. This is an Empirical Risk Minimization (ERM) model (Vapnik, 1998) as it aims to minimize the misclassification risk on the source domain. However, such a model is unlikely to generalize well to an unseen target domain with a different distribution.

We wish to remove the spurious correlation between the background bg in a video x and its action label y to better generalize to new domains during training. To do so, we intervene on the action A using a do-operator $\mathrm{do}(A)$. This do-operator removes the causal effect between the domain D and action A and calculates how changing the action changes the outcome. In our action recognition model we then aim to optimize $P_S(y \mid \mathrm{do}(A{=}a))$ instead of $P_S(y \mid (X{=}x))$. As we cannot optimize $P_S(y \mid \mathrm{do}(A{=}a))$ directly, we use the backdoor adjustment to express it in terms of observable probabilities:

$$P_S(y \mid \mathrm{do}(A{=}a)) = \sum_{bg} P_S(y \mid A{=}a, BG{=}bg) P_S(BG{=}bg). \tag{1}$$

$P_S(y \mid A{=}a, BG{=}bg)$ is the probability of predicting label y when action a appears on background bg. To obtain this probability, we synthesize counterfactual videos with new backgrounds. Let us assume we have a function able to separate video x into action a and background bg and recombine a with a different background. Several such functions are explained in Section 3.3. For each term in the sum in Eq. (1), we sample a random background bg from the distribution of backgrounds in the source domain, $P_S(BG)$. We combine this with action a of video x to create a counterfactual video. $P_S(y \mid A{=}a, BG{=}bg)$ is the prediction of our model on this counterfactual.

$P_S(BG{=}bg)$ is the probability of sampling background bg. This is $1/n$, where n is the total number of videos. We find it helpful to weight the predictions according to the action class so that counterfactuals from under-represented actions have a larger weight. We do this with factor $\frac{\sum n_y/|y|}{n_y}$ where n_y is the number of videos in class y and $|y|$ is the number of classes.

We multiply the prediction for each counterfactual $P_S(y \mid A{=}a, BG = bg)$ by the probability of sampling the background $P_S(BG{=}bg){=}\frac{1}{n}$, note that since $\sum_y n_y{=}n$, the n factor cancels out. It is impractical to create counterfactuals for every combination of action and background. Thus we approximate $P_S(y \mid \mathrm{do}(A{=}a))$ through sampling. Then Eq. (1) becomes:

$$P_S(y \mid \mathrm{do}(A{=}a)) \approx \sum_{bg \in B(x)} P_S(y \mid A{=}a, BG{=}bg) \frac{1}{|y|\ n_y}, \tag{2}$$

where $B(x)$ is the set of backgrounds sampled for video x with action a. In training, a new background is sampled for each video every epoch.

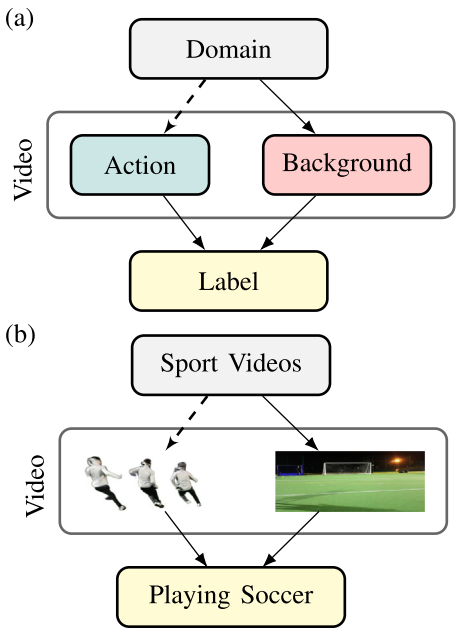
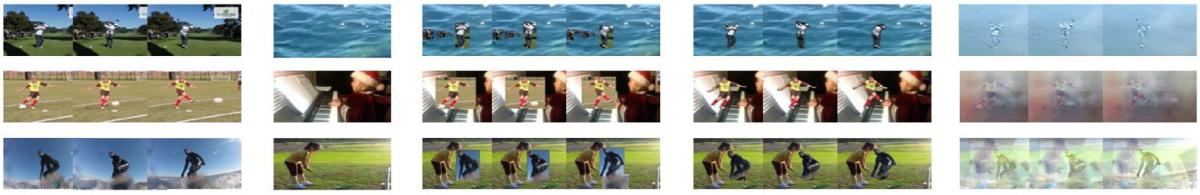


Fig. 2. Causal graph for video domain generalization. (a) The confounded causal graph. The domain has direct causal effects on background and action, which in turn have causal effects on the label. We intervene on the action to remove the causal effect Domain→Action, shown with a dashed line. (b) Causal graph example for 'playing soccer.' We can divide the video into the person 'playing soccer' and the soccer pitch background.



(a) Video (b) Background (c) Do_{Box} (d) Do_{Mask} (e) $Do_{Fourier}$

Fig. 3. Counterfactual examples from different *Do*-operators. (a) original video, (b) random background from a different action, (c-e) counterfactual videos resulting from composing the action from the original video with the random background.

Table 1
Video Domain Generalization benchmarks. Source and Target video datasets for each benchmark and the number of videos per train and test partition. All benchmarks contain a shift in distribution between source and target, but Kinetics2Mimetics has the most severe background shift.

Benchmark	Source domain			Target domain		
	Dataset	Actions	Train	Dataset	Actions	Test
Kinetics2Mimetics (Weinzaepfel and Rogez, 2021)	Kinetics10	10	5,416	Mimetics10	10	136
HMDB2UCF (Chen et al., 2019)	HMDB12	12	1,800	UCF12	12	2,009
UCF2HMDB (Chen et al., 2019)	UCF12	12	2,009	HMDB12	12	1,800

Optimizing our approximation to $P_S(y \mid \mathrm{do}(A{=}a))$ instead of $P_S(y \mid X{=}x)$ reduces the spurious correlation between background and action label in the model. This allows the model to better generalize to new target domains where the backgrounds are different from the source.

3.3. *Do*-Operators for action intervention

We describe three *Do*-operators for creating new videos by separating the background and action, as shown in Fig. 3.

$\mathbf{Do_{Box}}$ is inspired by previous works which remove background bias in action recognition using bounding boxes (Zou et al., 2022; Choi et al., 2019). We first apply an object detector to each video frame to extract bounding boxes of humans. The pixels inside the bounding box are taken as the action part of the video, while a random frame is taken as the background. To create a counterfactual, for each frame of the intervened video, we overlay the action inside the box over a randomly selected background. Do_{Box} accurately detects the part of the video depicting the action since it uses the robustness of object detectors. However, the box inevitably contains a small part of the background meaning we cannot completely remove the effect of the background on the predicted label. One of the key distinctions between Zou et al. (2022) and our approach lies in the treatment of labels in the generated video. Zou et al. (2022) employs a probabilistic approach, using a linear combination of labels for the resulting video. This can create some ambiguities in the final prediction. Our method instead adopts a causal framework. When we intervene on a particular action within the video, the label is specifically assigned to that action segment, irrespective of the background. Importantly, even if the background takes up most of the video frame, it does not influence the final label assignment in our method.

$\mathbf{Do_{Mask}}$ offers a more accurate separation of action and background. We obtain a mask of the action using a model for segmenting humans. As in Do_{Box}, we take the pixels inside the mask to be the action part of the video and a random frame to be the background. To create a counterfactual, we again select a random background and overlay the human mask for each video frame. With Do_{Mask}, the action part contains less background information than in Do_{Box} and thus is more suitable for our causal method. However, the segmentation is computationally expensive and often struggles to detect blurry humans with fast motions, which is crucial for action recognition.

$\mathbf{Do_{Fourier}}$ does not rely on external models as Do_{Box} and Do_{Mask} do. Instead, it exploits there usually being little or no motion in the background of a video. The dominant motion is instead in the action part of a video. We can, therefore, use a 3D Fourier transform to separate the background and action parts by considering the low-frequencies to be the background and the other parts to be the action. To find multidimensional discrete Fourier transform $\mathbf{X}$ for the signal $\mathbf{x}$, we can use following equation:

$$X(K_1, K_2, \dots, K_m) = \sum_{n_1=0}^{N_1-1} \sum_{n_2=0}^{N_2-1} \cdots \sum_{n_m=0}^{N_m-1} x(n_1, n_2, \dots, n_m)\; e^{-i\frac{2\pi}{N_1}n_1K_1 - i\frac{2\pi}{N_2}n_2K_2 - \cdots - i\frac{2\pi}{N_m}n_mK_m}. \quad (3)$$

We can also apply the inverse discrete Fourier transform to find the video values from its spectrum.

$$x(n_1, n_2, \dots, n_m) = \frac{1}{N_1N_2\cdots N_m} \sum_{K_1=0}^{N_1-1} \sum_{K_2=0}^{N_2-1} \cdots \sum_{K_m=0}^{N_m-1} X(K_1, K_2, \dots, K_m)\; e^{i\frac{2\pi}{N_1}n_1K_1 + i\frac{2\pi}{N_2}n_2K_2 + \cdots + i\frac{2\pi}{N_m}n_mK_m}. \quad (4)$$

If we consider the time dimension to have length N_t, the lower frequencies in this dimension demonstrate lower change along the time dimension. This means the lowest or zero frequency is associated with the background information. So let us define the time high pass filter $\mathcal{F}$. We can apply this filter to the frequency spectrum via Hadamard multiplication to remove the background information. Then, we use the complement filter to apply to another video to extract background information. In the end, we create a counterfactual spectrum by overlaying these two masked spectra as $X_{CF} = \mathcal{F} \odot X + (\mathbf{1} - \mathcal{F}) \odot \hat{X}$, in which X_{CF} is the spectrum of counterfactual video, $\mathcal{F}$ is the time high pass filter, $\mathbf{1} - \mathcal{F}$ is the complement filter which is a time low pass filter. X is the video in which we intervene on its action part, and $\hat{X}$ is the video in which we use its background for backdoor adjustment and creating a counterfactual video. Ultimately, we apply inverse Fourier transform to create the counterfactual video. The overall pipeline of our $Do_{Fourier}$ is shown in Fig. 4.

$Do_{Fourier}$ has the advantage that it may also eliminate other static confounders such as clothing or faces. However, the 3D Fourier transform can be sensitive to camera movements and changes in the lighting. With these Do-operators, we can learn a background-debiased model from a source domain through our action intervention.

3.4. Test-time action intervention

While our method has made the model learn to ignore bias from source domain backgrounds during training, it may not be able to ignore the background in the target domain due to the distribution shift between source and target. Thus, we want to remove the background bias in the target domain, as we did for the source domain, but at test

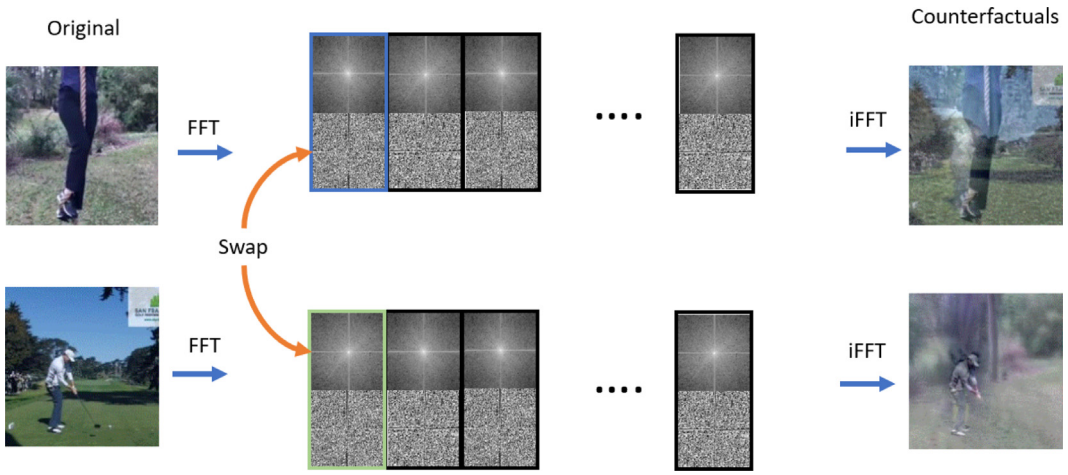


Fig. 4. The pipeline for the $Do_{Fourier}$ Operator. We first convert videos to the frequency domain using the Fourier transform, and then we intervene on frequency bands that have more information about the background than action. We then create counterfactuals by replacing the said frequency bands with their equivalent from another video. Finally, using the inverse Fourier transform, we generate the counterfactual video.

time. To do this, we can intervene on the action at test time and create counterfactuals using backgrounds from the source domain and actions from the target domain. This allows us to minimize the difference in background distribution between source and target videos. Hence, we take the prediction $P_{\mathcal{T}}(y \mid A{=}a)$ for a target domain video x containing background a to be:

$$P_{\mathcal{T}}(y \mid A{=}a) = \frac{1}{|B(x)|} \sum_{bg \in B(x)} P_{\mathcal{S}}(y \mid A{=}a, BG{=}bg), \tag{5}$$

where $P_{\mathcal{S}}(y \mid A{=}a, BG{=}bg)$ is the prediction from a counterfactual created with randomly sampled background $bg \in B(x)$. It is important to note that we do not change the underlying model at test time; we instead perform a simple adjustment to the test data to get more reliable predictions from the underlying model.

4. Experiments

4.1. Benchmarks, evaluation & implementation

We evaluate our approach on two benchmarks. As domain generalization requires one domain for training and a separate domain for testing, benchmarks for our experiments are defined over a source and target dataset-pair that share the same vocabulary of action class labels. Table 1 summarizes the benchmarks we use for evaluation, covering four different datasets.

Kinetics2Mimetics. Weinzaepfel and Rogez (2021) introduced the Kinetics2Mimetics benchmark to highlight the bias of action recognition models to non-motion related clues in the video. It uses Kinetics (Kay et al., 2017), a widely used action recognition dataset, as the source domain and their own Mimetics test set as the target. For domain generalization, a subset of 10 action categories is commonly used (Bao et al., 2021; Li and Vasconcelos, 2022; Bahng et al., 2020). *Kinetics10* consists of 10 s clips taken from YouTube videos. It contains 10 action classes over 5,416 training videos and 449 validation videos. *Mimetics10* consists 136 clips collected from YouTube which are the mimed versions of the 10 action categories in Kinetcs10. Mimetics2Kinetics is an extreme case of video domain generalization, where background, objects and clothing are very different in the source and target domains.

HMDB2UCF and UCF2HMDB. Chen et al. (2019) introduced the HMDB2UCF and UCF2HMDB benchmarks for domain adaptation which where later repurposed for domain generalization (Yao et al., 2021). It uses videos from HMDB12 and UCF12, each containing the same 12 action classes. *HMDB12* is a subset of HMDB51 (Kuehne et al., 2011), which is a collection of action videos from various internet sources.

Table 2
Effect of background-debiased learning and test-time action intervention, where our *Do*-operator is Do_{Mask}. We report average accuracy and standard deviation over 5 runs. Both our background-debiased learning and test-time action intervention contribute to our improved domain generalization accuracy over a standard ERM baseline.

Training	Test-time	Kinetics2Mimetics
ERM baseline	–	31.7 ± 1.6
Background-Debias	–	43.2 ± 1.3
ERM baseline	Action Intervention	15.5 ± 1.7
Background-Debias	**Action Intervention**	$\mathbf{46.0 \pm 1.4}$

HMDB12 contains 1,800 videos. *UCF12* is a subset of UCF101 (Soomro et al., 2012), which consists of 2,009 action videos from YouTube. For HMDB2UCF we use HMDB12 for the source data and the and UCF12 as the target dataset. For UCF2HMDB, we use UCF12 as the source and HMDB12 as the target. In general UCF2HMDB is more challenging than HMDB2UCF because HMDB contains more variety in backgrounds, giving a model trained on HMDB more ability to generalize to new backgrounds than one trained on UCF.

Evaluation. We report top-1 accuracy for all benchmarks.

Implementation Details. Videos are sampled at 30fps and scaled to 256 × 256px. To facilitate a fair comparison, we use a TSM backbone (Lin et al., 2019) pretrained on Kinetics 400 (Kay et al., 2017) for all ablations. We optimize $P_{\mathcal{S}}(y \mid \mathrm{do}(A{=}a))$ (Eq. (2)) with a cross entropy loss through SGD with an initial learning rate of 0.0001 and momentum of 0.9. The learning rate is reduced by a factor of 10 at epoch 20 and 40 with training ending at epoch 50. For a 3D-ResNet backbone and experiments with HMDB and UCF, we use 500 epochs with an initial learning rate of 0.001, dividing it by 10 every 100 epochs. To perform Do_{Box}, we extract human bounding boxes with a certainty >0.7 using detectron2 (Wu et al., 2019). For Do_{Mask} we obtain the human mask with U^2-Net (Qin et al., 2020). For $Do_{Fourier}$ we consider the lowest frequency in the time dimension to be the background with the remaining frequencies are taken as the action. In our background-debias learning we sample a new counterfactual for each training video every epoch, *i.e.*, $|B(x)| = \#\text{epochs}$. In test-time debiasing we use five counterfactuals per video, *i.e.*, $|B(x)|{=}5$.

4.2. Ablation studies

Effect of background-debiased learning and test-time action intervention. Table 2 shows the contribution of our background-debias

Table 3
Comparing *Do*-operators. The more effectively a *Do*-operator separates the action from the background, the better our model generalizes to new domains. Across the benchmarks we find $Do_{Fourier}$ to be the best at separating action and background. Average accuracy and standard deviation are calculated from 5 runs.

	Kinetics2Mimetics	HMDB2UCF	UCF2HMDB
ERM baseline	31.7 ± 1.6	69.2 ± 0.3	52.2 ± 0.3
Do_{Box}	33.1 ± 0.8	62.5 ± 0.3	47.6 ± 0.3
Do_{Mask}	46.0 ± 1.4	59.2 ± 0.7	44.1 ± 0.3
$Do_{Fourier}$	41.8 ± 1.2	71.3 ± 0.6	53.3 ± 0.4

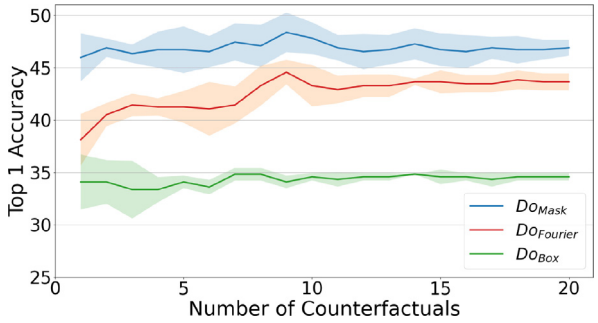


Fig. 5. How many counterfactuals at test time? Increasing the number of counterfactuals has a stabilizing effect on model accuracy.

learning and test-time action intervention for Kinetics2Mimetics using Do_{Mask}. We compare our model with a standard ERM baseline which does not use any debiasing or domain generalization. Both the background-debiased learning and test-time action intervention contribute to the increase in accuracy. Background-debiased learning improves the accuracy considerably from 31.7% to 43.2% while decreasing the standard deviation from 1.6% to 1.3%. This is because our background-debias learning removes the spurious correlation between background and action label, making the model more generalizable to new domains where the actions are found in different backgrounds. On its own, our test-time action intervention does not improve results, however combining it with our background-debiased learning gives a further improvement achieving a +14.3% increase over the ERM baseline.

Comparing *Do*-operators. Next we consider the benefit of the three *Do*-operators we propose. We compare them in Table 3 using our full model with both background-debiased learning and test-time action intervention. The quality of the *Do*-operators varies and also depends on the source and target datasets used. With accurate human mask detection, Do_{Mask} retains the fewest background parts. Therefore, it better generalizes from Kinetics to Mimetics than the other two operators. However, the human segmentation is less accurate on UCF and HMDB, meaning Do_{Mask} does not correctly separately action and background. As expected, Do_{Box} is overall the least effective as it keeps some portion of the background in the box. Since the bounding box contains more background information than our other *Do*-operators, we cannot fully remove the spurious correlation between background and action label. Since these operators are utilized as causal intervention proxies, their precision affects the accuracy of this approximation. $Do_{Fourier}$ is our most effective *Do*-operator, increasing over the ERM baseline on all benchmarks, as it separates the low-frequency background from high-frequency action motions and does not rely on the accuracy of an external model. We use $Do_{Fourier}$ for our state-of-the-art comparison in Section 4.3.

Background removal vs. background-debiased learning. We compare our background-debiased learning to removing the background with Do_{Mask} and training on only the action part of the video rather than our counterfactuals. This gives 34.6% compared to 43.2% with our background-debiased learning.

Table 4
Kinetics2Mimetics. Our method outperforms prior domain generalization methods for 3D-ResNet18 and TSM backbones. When pretrained using Kinetics we further improve over the state-of-the-art considerably. We use $Do_{Fourier}$ for causal intervention in this table.

Model	Pre-training	Acc.
3D-ResNet18 (Hara et al., 2018)		
ERM baseline	None	18.9
Rebias (Bahng et al., 2020)	None	22.4
Ours	None	**22.6**
ERM baseline	Kinetics	23.5
DRL (Li and Vasconcelos, 2022)	Kinetics	26.4
Ours	Kinetics	**29.4**
APN (Yao et al., 2021)		
RADA (Yao et al., 2021)	ImageNet	30.5
TSM (Lin et al., 2019)		
ERM baseline	ImageNet	30.2
DEAR (Bao et al., 2021)	ImageNet	34.4
Ours	ImageNet	**36.0**
ERM baseline	Kinetics	31.7
Ours	Kinetics	**41.8**

Effect of number of counterfactuals at test-time. In Fig. 5 we ablate whether more counterfactuals at test-time leads to better results using Kinetics2Mimetics. We find that using more counterfactuals generally reduces the variance in results for all three *Do*-operators. We also find that using up to 10 counterfactuals can improve results, after this point the accuracy plateaus.

4.3. Comparison with state-of-the-art

For all comparisons with prior works we use $Do_{Fourier}$.

Kinetics2Mimetics. We compare against prior works which aim for domain generalization (Li and Vasconcelos, 2022) or learn an unbiased domain representation (Bahng et al., 2020; Bao et al., 2021) on Kinetics2Mimetics. To obtain a fair comparison we run our approach with the same backbone as each prior work. Table 4 shows our approach outperforms other methods for every combination of backbone and pre-training. For instance, using a 3D-ResNet18 backbone with Kinetics pretraining we have a 3.0% improvement over the most recent prior approach DRL (Li and Vasconcelos, 2022). Note that DRL is trained on the entire Kinetics400 dataset and then tested on Mimetics allowing the model to see more varied backgrounds. For consistency with our ablations and other state-of-the-art comparisons, we report our results by fine-tuning on Kinetics10. With a TSM backbone and ImageNet pretraining, our causal model gives +1.6% improvement over DEAR (Bao et al., 2021). We also compare to prior work (Yao et al., 2021) which proposes the domain generalization backbone APN in combination with the RADA method. Since APN+RADA was not originally tested on Kinetics2Mimetics we obtain it using their publicly available code. Using a non-specialized TSM backbone our approach outperforms APN+RADA by 5.5%. Using our approach with the TSM backbone and Kinetics pretraining outperforms the prior state-of-the-art even further to obtain 41.8% accuracy. Kinetics2Mimetics is a challenging domain generalization benchmark as backgrounds in Kinetics are highly correlated with the action label, while backgrounds in Mimetics are very different and unrelated to the action. Our approach outperforms prior works on this benchmark as our background-debias learning allows our model to remove the spurious correlation between background and action label in the source domain.

The assumption of completely unbiased data, like Mimetics, is not realistic. However, current appearance-based datasets with the same

Table 5
HMDB2UCF and UCF2HMDB with TSM and APN backbones. All baseline numbers come from Yao et al. (2021). We are the only approach that always improves over the ERM baseline. We further obtain best (**bold**) or second best (underlined) performance on each benchmark, demonstrating that our method is effective and not backbone specific. We use $Do_{Fourier}$ for causal intervention in this table.

Model	HMDB2UCF		UCF2HMDB	
	TSM (Lin et al., 2019)	APN (Yao et al., 2021)	TSM (Lin et al., 2019)	APN (Yao et al., 2021)
ERM baseline	69.2±0.3	71.4 ± 0.3	52.2 ± 0.3	54.3 ± 0.3
Jigsaw (Carlucci et al., 2019)	68.9 ± 0.3	72.4 ± 0.3	52.5±0.3	55.2 ± 0.3
M-ADA (Qiao et al., 2020)	69.1 ± 0.3	71.5 ± 0.3	52.5±0.2	56.9 ± 0.3
RADA (Yao et al., 2021)	69.2±0.2	**74.9 ± 0.3**	51.3 ± 0.2	**59.1 ± 0.3**
Ours	**71.3 ± 0.6**	73.3±0.6	**53.3 ± 0.4**	57.2±0.4

Fig. 6. Qualitative results for Kinetics2Mimetics. The baseline model relies on the background to classify the action. Since our model is robust to changes in background, it can correctly predict the action when the background is irrelevant to the target action.

train and test distribution are also unrealistic. The real world is between these extremes. This paper aims to address the other often ignored extreme. With Kinetics2Mimetics we show that removing the background bias is especially helpful when source and target datasets are very different. Next we demonstrate that our model also provides some improvement when source and target datasets have similar backgrounds.

HMDB2UCF and UCF2HMDB. Since video domain generalization is a nascent field, we only found one published work, RADA (Yao et al., 2021), reporting on HMDB2UCF and UCF2HMDB for domain generalization. Thus, we also compare to image domain generalization methods Jigsaw (Carlucci et al., 2019) and M-ADA (Qiao et al., 2020). We do this comparison for the common action recognition backbone TSM (Lin et al., 2019) as well as APN (Yao et al., 2021) which was proposed for domain generalization alongside RADA. Table 5 demonstrates that for both HMDB2UCF and UCF2HMDBF our approach outperforms prior works when using the TSM backbone, where prior works do not manage to improve over the ERM baseline. Our approach also improves over the ERM baseline, Jigsaw and M-ADA with the APN backbone and is only outperformed by RADA, the method designed to work with the APN backbone. As expected, the UCF2HMDB benchmark is more challenging than HMDB2UCF for all models, but the behavior of the models is the same.

We conclude that despite the backgrounds in HMDB and UCF being more similar to each other than Kinetics and Mimetics, our method is still able to eliminate much of the correlation between background and action label. This results in a domain generalization model that is effective with different backbones.

Our method consistently enhances domain generalization performance, particularly for models such as TSM and APN. These models inherently treat the temporal dimension distinctively from the spatial dimensions. Conversely, in the case of models like 3D-ResNet, which lacks specific mechanisms for distinguishing the temporal dimension, the observed improvement is less pronounced. It is essential to highlight that the improvement observed on the Kinetics2Mimetics benchmark stands out significantly compared to other benchmarks. This disparity can be attributed to this benchmark's substantial distribution gap between source and target domains. In contrast, HMDB and UCF benchmarks exhibit a lesser gap between their source and target domains. Our Structural Causal Model (SCM) analysis indicates that the ERM baseline performs comparably to the causal approach in scenarios where this source–target domain gap is less pronounced.

It is important to highlight that while many current techniques are based on probabilistic methods, they can gain insights from background cues, especially when the background provides context about the action. Yet, the emphasis shifts towards motion cues in motion-intensive datasets, such as the *something-something* (Yao et al., 2021) dataset. For these reasons, it is predictable that the ERM baseline outperforms the causal approach in these scenarios. Nonetheless, even in such datasets, object-related cues can provide spurious information about the associated action.

4.4. Qualitative results

Fig. 6 shows qualitative results for Kinetics2Mimetics. Each example shows the predictions of the baseline model on the original video and our model on counterfactual videos from the three *Do*-operators. We see that the baseline is often reliant on the background to predict the action, for instance it can correctly predict 'surfing water' when there is a water background (row 3), but struggles to predict the mimed version of the action in row 1. On the other hand, our approach can predict the correct action when the background is irrelevant to the target action. We find our model struggles to predict the correct action when the action is small (row 4) and when the *Do*-operators remove relevant objects such as the surfboard (row 3).

5. Conclusion

This paper presents a method for recognizing actions in new domains through causal intervention. With our background-debias learning, we can train a model where the spurious correlation between

background and action label is removed through a causal intervention on the action through three possible *Do*-operators. We can then further improve the generalization of our model to new domains through a test-time action intervention, which reduces the difference in distribution between source and target domains. We have tested our approach on different video domain generalization benchmarks where it constantly improves over an empirical risk minimization baseline, independent of the underlying action recognition backbone. The promise of causal interventions is further demonstrated by the new state-of-the-art on the challenging Kinetics2Mimetics benchmark. Future studies can extend to address eliminating additional bias factors, including objects, clothing, and viewpoints. It is important to note that our current techniques, $\mathbf{Do}_{\mathbf{Box}}$ and $\mathbf{Do}_{\mathbf{Mask}}$, are primarily effective in scenarios where there is no domain gap in temporal aspects like speed and movement. However, an intriguing future application of $\mathbf{Do}_{\mathbf{Fourier}}$ operators is to tackle the variations in speed and movement across diverse frequency bands. This will enhance the adaptability and applicability of our method in a broader range of domain gaps.

CRediT authorship contribution statement

Sarah Rastegar: Conceptualization, Methodology, Writing – original draft. **Hazel Doughty:** Supervision, Validation, Writing – reviewing & editing. **Cees G.M. Snoek:** Supervision, Validation, Writing – reviewing & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgments

This work is part of the project Real-Time Video Surveillance Search with project number 18038, which is (partly) financed by the Dutch Research Council (NWO) domain Applied and Engineering/ Sciences (TTW).

Appendix A. Further ablations

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.cviu.2024.103975.

References

Alwassel, H., Mahajan, D., Korbar, B., Torresani, L., Ghanem, B., Tran, D., 2020. Self-supervised learning by cross-modal audio-video clustering. Adv. Neural Inf. Process. Syst. 33, 9758–9770.

Andonian, A., Chen, S., Hamid, R., 2022. Robust cross-modal representation learning with progressive self-distillation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 16430–16441.

Bahng, H., Chun, S., Yun, S., Choo, J., Oh, S.J., 2020. Learning de-biased representations with biased representations. In: International Conference on Machine Learning. PMLR.

Bao, W., Yu, Q., Kong, Y., 2021. Evidential deep learning for open set action recognition. In: Proceedings of the IEEE/CVF International Conference on Computer Vision.

Carlucci, F.M., D'Innocente, A., Bucci, S., Caputo, B., Tommasi, T., 2019. Domain generalization by solving jigsaw puzzles. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.

Carreira, J., Zisserman, A., 2017. Quo vadis, action recognition? A new model and the kinetics dataset. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.

Chen, M.H., Kira, Z., AlRegib, G., Woo, J., Chen, R., Zheng, J., 2019. Temporal attentive alignment for large-scale video domain adaptation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision.

Chen, Z., Tian, Z., Zhu, J., Li, C., Du, S., 2022. C-CAM: Causal CAM for weakly supervised semantic segmentation on medical image. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.

Chen, L., Yan, X., Xiao, J., Zhang, H., Pu, S., Zhuang, Y., 2020. Counterfactual samples synthesizing for robust visual question answering. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.

Choi, J., Gao, C., Messou, J.C., Huang, J.B., 2019. Why can't I dance in the mall? Learning to mitigate scene bias in action recognition. In: Advances in Neural Information Processing Systems.

Chuah, W., Tennakoon, R., Hoseinnezhad, R., Bab-Hadiashar, A., Suter, D., 2022. ITSA: An information-theoretic approach to automatic shortcut avoidance and domain generalization in stereo matching networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.

Dash, S., Balasubramanian, V.N., Sharma, A., 2022. Evaluating and mitigating bias in image classifiers: A causal perspective using counterfactuals. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision.

de Souza, C.R., Gaidon, A., Cabon, Y., López, A.M., 2018. Procedural generation of videos to train deep action recognition networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.

Ding, S., Li, M., Yang, T., Qian, R., Xu, H., Chen, Q., Wang, J., Xiong, H., 2022. Motion-aware contrastive video representation learning via foreground-background merging. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.

Feichtenhofer, C., Fan, H., Malik, J., He, K., 2019. Slowfast networks for video recognition. In: Proceedings of the IEEE/CVF International Conference on Computer Vision.

Gowda, S.N., Rohrbach, M., Keller, F., Sevilla-Lara, L., 2022. Learn2Augment: Learning to composite videos for data augmentation in action recognition. In: European Conference on Computer Vision. Springer, pp. 242–259.

Hara, K., Kataoka, H., Satoh, Y., 2018. Can spatiotemporal 3D CNNs retrace the history of 2D CNNs and ImageNet? In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.

Hasan, I., Liao, S., Li, J., Akram, S.U., Shao, L., 2022. Pedestrian detection: Domain generalization, CNNs, transformers and beyond. arXiv preprint arXiv:2201.03176.

Hwang, H., Jang, C., Park, G., Cho, J., Kim, I.J., 2023. Eldersim: A synthetic data generation platform for human action recognition in eldercare applications. IEEE Access 11, 9279–9294. http://dx.doi.org/10.1109/ACCESS.2021.3051842.

Ilic, F., Pock, T., Wildes, R.P., 2022. Is appearance free action recognition possible? In: European Conference on Computer Vision. Springer, pp. 156–173.

Johnson, J., Hariharan, B., Van Der Maaten, L., Fei-Fei, L., Lawrence Zitnick, C., Girshick, R., 2017. CLEVR: A diagnostic dataset for compositional language and elementary visual reasoning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.

Kay, W., Carreira, J., Simonyan, K., Zhang, B., Hillier, C., Vijayanarasimhan, S., Viola, F., Green, T., Back, T., Natsev, P., Suleyman, M., Zisserman, A., 2017. The kinetics human action video dataset. arXiv preprint arXiv:1705.06950.

Kim, Y.-w., Mishra, S., Jin, S., Panda, R., Kuehne, H., Karlinsky, L., Saligrama, V., Saenko, K., Oliva, A., Feris, R., 2022. How transferable are video representations based on synthetic data? In: Advances in Neural Information Processing Systems Datasets and Benchmarks Track.

Kim, D., Tsai, Y.H., Zhuang, B., Yu, X., Sclaroff, S., Saenko, K., Chandraker, M., 2021. Learning cross-modal contrastive features for video domain adaptation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision.

Kuehne, H., Jhuang, H., Garrote, E., Poggio, T., Serre, T., 2011. HMDB: a large video database for human motion recognition. In: Proceedings of the IEEE International Conference on Computer Vision.

Li, Y., Vasconcelos, N., 2022. Improving video model transfer with dynamic representation learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.

Li, Y., Yang, X., Shang, X., Chua, T.S., 2021. Interventional video relation detection. In: Proceedings of the ACM International Conference on Multimedia.

Lin, J., Gan, C., Han, S., 2019. Tsm: Temporal shift module for efficient video understanding. In: Proceedings of the IEEE/CVF International Conference on Computer Vision.

Liu, Y., Chen, J., Chen, Z., Deng, B., Huang, J., Zhang, H., 2021b. The blessings of unlabeled background in untrimmed videos. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.

Liu, R., Liu, H., Li, G., Hou, H., Yu, T., Yang, T., 2022a. Contextual debiasing for visual recognition with causal mechanisms. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.

Liu, B., Wang, D., Yang, X., Zhou, Y., Yao, R., Shao, Z., Zhao, J., 2022. Show, deconfound and tell: Image captioning with causal inference. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.

Liu, Y., Wei, Y.S., Yan, H., Li, G.B., Lin, L., 2022c. Causal reasoning meets visual representation learning: A prospective study. Mach. Intell. Res. 1–27.

Lv, F., Liang, J., Li, S., Zang, B., Liu, C.H., Wang, Z., Liu, D., 2022. Causality inspired representation learning for domain generalization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.

Magliacane, S., Van Ommen, T., Claassen, T., Bongers, S., Versteeg, P., Mooij, J.M., 2018. Domain adaptation by using causal inference to predict invariant conditional distributions. In: Advances in Neural Information Processing Systems.

Mahajan, D., Tople, S., Sharma, A., 2021. Domain generalization using causal matching. In: International Conference on Machine Learning. PMLR.

Majumdar, S.S., Jain, S., Tourni, I.C., Mustafin, A., Lteif, D., Sclaroff, S., Saenko, K., Bargal, S.A., 2022. Ani-GIFs: A benchmark dataset for domain generalization of action recognition from GIFs. Front. Comput. Sci. 4, 876846.

Muandet, K., Balduzzi, D., Schölkopf, B., 2013. Domain generalization via invariant feature representation. In: International Conference on Machine Learning. PMLR.

Munro, J., Damen, D., 2020. Multi-modal domain adaptation for fine-grained action recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.

Nan, G., Qiao, R., Xiao, Y., Liu, J., Leng, S., Zhang, H., Lu, W., 2021. Interventional video grounding with dual contrastive learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.

Ngiam, J., Khosla, A., Kim, M., Nam, J., Lee, H., Ng, A.Y., 2011. Multimodal deep learning. In: Proceedings of the 28th International Conference on Machine Learning. pp. 689–696.

Park, Y., Woo, S., Lee, S., Nugroho, M.A., Kim, C., 2023. Cross-modal alignment and translation for missing modality action recognition. Comput. Vis. Image Underst. 236, 103805.

Pearl, J., 2009. Causality. Cambridge University Press.

Peng, D., Lei, Y., Hayat, M., Guo, Y., Li, W., 2022. Semantic-aware domain generalized segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.

Peters, J., Janzing, D., Schölkopf, B., 2017. Elements of Causal Inference: Foundations and Learning Algorithms. The MIT Press.

Planamente, M., Plizzari, C., Alberti, E., Caputo, B., 2022a. Domain generalization through audio-visual relative norm alignment in first person action recognition. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision.

Planamente, M., Plizzari, C., Caputo, B., 2022b. Test-time adaptation for egocentric action recognition. In: International Conference on Image Analysis and Processing.

Qiao, F., Zhao, L., Peng, X., 2020. Learning to learn single domain generalization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.

Qin, X., Zhang, Z., Huang, C., Dehghan, M., Zaiane, O., Jagersand, M., 2020. U2-Net: Going deeper with nested U-structure for salient object detection. Pattern Recognit. 106.

Rastegar, S., Soleymani, M., Rabiee, H.R., Shojaee, S.M., 2016. Mdl-cw: A multimodal deep learning framework with cross weights. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2601–2609.

Roheda, S., Riggan, B.S., Krim, H., Dai, L., 2018. Cross-modality distillation: A case for conditional generative adversarial networks. In: 2018 IEEE International Conference on Acoustics, Speech and Signal Processing. ICASSP, IEEE, pp. 2926–2930.

Sahoo, A., Shah, R., Panda, R., Saenko, K., Das, A., 2021. Contrast and mix: Temporal contrastive video domain adaptation with background mixing. In: Advances in Neural Information Processing Systems.

Sohn, K., Shang, W., Lee, H., 2014. Improved multimodal deep learning with variation of information. Adv. Neural Inf. Process. Syst. 27.

Soomro, K., Zamir, A.R., Shah, M., 2012. UCF101: A dataset of 101 human actions classes from videos in the wild. arXiv preprint arXiv:1212.0402.

Sun, Y., Wang, X., Liu, Z., Miller, J., Efros, A., Hardt, M., 2020. Test-time training with self-supervision for generalization under distribution shifts. In: International Conference on Machine Learning. PMLR.

Thoker, F.M., Doughty, H., Bagad, P., Snoek, C.G., 2022. How severe is benchmark-sensitivity in video self-supervised learning? In: European Conference on Computer Vision. Springer, pp. 632–652.

Vapnik, V.N., 1998. Statistical learning theory. J. Wiley.

Varol, G., Laptev, I., Schmid, C., Zisserman, A., 2021. Synthetic humans for action recognition from unseen viewpoints. Int. J. Comput. Vis. 129 (7), 2264–2287.

Wan, C., Shen, X., Zhang, Y., Yin, Z., Tian, X., Gao, F., Huang, J., Hua, X.S., 2022. Meta convolutional neural networks for single domain generalization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.

Wang, W., Gao, J., Xu, C., 2022a. Weakly-supervised video object grounding via causal intervention. IEEE Trans. Pattern Anal. Mach. Intell..

Wang, D., Shelhamer, E., Liu, S., Olshausen, B., Darrell, T., 2021. Tent: Fully test-time adaptation by entropy minimization. In: International Conference on Learning Representations.

Wang, R., Yi, M., Chen, Z., Zhu, S., 2022b. Out-of-distribution generalization with causal invariant transformations. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.

Wang, Z., Zhao, Y., Huang, H., Xia, Y., Zhao, Z., 2023. Scene-robust natural language video localization via learning domain-invariant representations. In: Findings of the Association for Computational Linguistics. ACL 2023, pp. 144–160.

Weinzaepfel, P., Rogez, G., 2021. Mimetics: Towards understanding human actions out of context. Int. J. Comput. Vis. 129 (5), 1675–1690.

Wu, Y., Kirillov, A., Massa, F., Lo, W.Y., Girshick, R., 2019. Detectron2. https://github.com/facebookresearch/detectron2.

Xiao, Z., Zhen, X., Shao, L., Snoek, C.G.M., 2022. Learning to generalize across domains on single test samples. In: International Conference on Learning Representations.

Xue, H., Jiang, W., Miao, C., Ma, F., Wang, S., Yuan, Y., Yao, S., Zhang, A., Su, L., 2020. Deepmv: Multi-view deep learning for device-free human activity recognition. Proc. ACM Interact. Mob. Wearable Ubiquitous Technol. 4 (1), 1–26.

Yang, X., Feng, F., Ji, W., Wang, M., Chua, T.S., 2021. Deconfounded video moment retrieval with causal intervention. In: Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval.

Yao, Z., Wang, Y., Wang, J., Yu, P., Long, M., 2021. Videodg: Generalizing temporal relations in videos to novel domains. IEEE Trans. Pattern Anal. Mach. Intell..

Yue, Z., Sun, Q., Hua, X.S., Zhang, H., 2021a. Transporting causal mechanisms for unsupervised domain adaptation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision.

Yue, Z., Wang, T., Sun, Q., Hua, X.S., Zhang, H., 2021b. Counterfactual zero-shot and open-set visual recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.

Zhang, Y., Doughty, H., Shao, L., Snoek, C.G.M., 2022. Audio-adaptive activity recognition across video domains. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.

Zhang, S., Feng, X., Fan, W., Fang, W., Feng, F., Ji, W., Li, S., Wang, L., Zhao, S., Zhao, Z., et al., 2023. Video-audio domain generalization via confounder disentanglement. In: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 37, No. 12. pp. 15322–15330.

Zhang, K., Gong, M., Schölkopf, B., 2015. Multi-source domain adaptation: A causal view. In: Proceedings of the AAAI Conference on Artificial Intelligence.

Zhang, T., Min, W., Yang, J., Liu, T., Jiang, S., Rui, Y., 2021a. What if we could not see? Counterfactual analysis for egocentric action anticipation.. In: Proceedings of the International Joint Conferences on Artificial Intelligence Organization.

Zhang, X., Wong, Y., Wu, X., Lu, J., Kankanhalli, M., Li, X., Geng, W., 2021b. Learning causal representation for training cross-domain pose estimator via generative interventions. In: Proceedings of the IEEE/CVF International Conference on Computer Vision.

Zhao, L., Liu, T., Peng, X., Metaxas, D., 2020. Maximum-entropy adversarial data augmentation for improved generalization and robustness. In: Advances in Neural Information Processing Systems.

Zhou, K., Yang, Y., Qiao, Y., Xiang, T., 2021. Domain adaptive ensemble learning. IEEE Trans. Image Process. 30, 8008–8018.

Zoetgnande, Y., Dillenseger, J.L., 2022. Domain generalization for activity recognition: Learn from visible, infer with thermal. In: 11th International Conference on Pattern Recognition Applications and Methods. SCITEPRESS-Science and Technology Publications, pp. 722–729.

Zou, Y., Choi, J., Wang, Q., Huang, J.B., 2022. Learning representational invariances for data-efficient action recognition. Comput. Vis. Image Underst. 103597.