Spatial-temporal Causal Inference for Partial Image-to-video Adaptation

Jin Chen,[1] **Xinxiao Wu,**[1*] **Yao Hu,**[2] **Jiebo Luo**[3]
[1]Beijing Laboratory of Intelligent Information Technology
School of Computer Science, Beijing Institute of Technology, Beijing, China
[2]Alibaba Youku Cognitive and Intelligent Lab
[3]Department of Computer Science, University of Rochester, Rochester NY 14627, USA
{chen_jin, wuxinxiao}@bit.edu.cn, yaoohu@alibaba-inc.com, jluo@cs.rochester.edu

Abstract

Image-to-video adaptation leverages off-the-shelf learned models in labeled images to help classification in unlabeled videos, thus alleviating the high computation overhead of training a video classifier from scratch. This task is very challenging since there exist two types of domain shifts between images and videos: 1) spatial domain shift caused by static appearance variance between images and video frames, and 2) temporal domain shift caused by the absence of dynamic motion in images. Moreover, for different video classes, these two domain shifts have different effects on the domain gap and should not be treated equally during adaptation. In this paper, we propose a spatial-temporal causal inference framework for image-to-video adaptation. We first construct a spatial-temporal causal graph to infer the effects of the spatial and temporal domain shifts by performing counterfactual causality. We then learn causality-guided bidirectional heterogeneous mappings between images and videos to adaptively reduce the two domain shifts. Moreover, to relax the assumption that the label spaces of the image and video domains are the same by the existing methods, we incorporate class-wise alignment into the learning of image-video mappings to perform partial image-to-video adaptation where the image label space subsumes the video label space. Extensive experiments on several video datasets have validated the effectiveness of our proposed method.

Introduction

Video recognition has made promising progress in recent years owing to the success of deep neural networks. Training deep video classifiers using large-scale labeled video datasets usually requires high storage resource and incurs heavy computational loads. Moreover, it is a time-consuming and labor-intensive process to annotate a large amount of videos. On the other hand, the computational cost of learning deep classifiers of images is much less and there are also many existing labeled image datasets that can be readily used. It would be highly beneficial to transfer knowledge from images to videos. So the image-to-video adaptation task has been proposed, which leverages off-the-shelf learned models in labeled images (source domain) to help recognition in unlabeled videos (target domain). This task is very challenging since there exist two types of domain shifts: 1) spatial domain shift caused by static appearance variance between images and video frames, and 2) temporal domain shift caused by the absence of dynamic motion in images.

A rich line of prior works attempt to reduce the spatial domain shift by learning a common feature space between images and video frames (Li et al. 2017; Ma et al. 2017; Zhang et al. 2016; Gan et al. 2016b,a; Sun et al. 2015). To reduce both spatial and temporal domain shifts, several recent methods learn domain-invariant features between images and video clips via generative adversarial networks (Yu et al. 2018, 2019), where the two domain shifts are treated equally. It is a fact that for different video classes, the spatial and temporal domain shifts play different roles in the adaptation process. For example, some videos such as "pour" and "kiss" contain very little dynamic motion information and can be easily distinguished by key frames. In this case, the static appearance is more important and the spatial domain shift weighs more in the adaptation. Other videos such as "jump" and "wave" have large variations in motion and the temporal domain shift has the main effect on the adaptation. Hence, it is critical and non-trivial to explore the effects of the two domain shifts to adaptively reduce the two domain shifts.

In this paper, we propose a spatial-temporal causal inference framework for image-to-video adaptation, which infers the effects of the spatial and temporal domain shifts via causal inference and adaptively reduces domain shifts via causality-guided bidirectional heterogeneous mappings. We build a spatial-temporal casual graph to infer the effects of the two domain shifts during adaptation. The causal graph has three nodes, including an appearance feature node A, a motion feature node B and a video class label node Y. The two edges $A \rightarrow Y$ and $B \rightarrow Y$ indicate that the appearance feature A and the motion feature B together affect the class label Y. The counterfactual causality is performed on this graph to reveal the causal relationships among A, B and Y through the counterfactual thinking of "If I had not seen A or B, would I still make the same Y?" Our insight is to manipulate the value of node A or B to infer its effect on video classification by investigating what Y would be. Specifically, we wipe out A and keep B untouched, and then obtain a counterfactual Y that is contrary to the fact that both A and B

*Corresponding author: Xinxiao Wu

affect Y. The difference between the counterfactual Y and the original Y (predicated from A and B) reflects the contribution of A to the video classification, which is actually the effect of the spatial domain shift. In a similar way, we infer the effect of the temporal domain shift via constructing a counterfactual Y by wiping out B and keeping A intact.

To adaptively reduce the two domain shifts, we learn causality-guided bidirectional heterogeneous mapping between images and videos, including the image-to-video mapping and the video-to-image mapping. The image-to-video mapping maps the image feature to the video feature space via adversarial learning. The image feature is coupled with the inferred effects to attend to the spatial and temporal shifts according to the contribution of the appearance and temporal features to video classification, which automatically balances different domain shifts during mapping. To avoid mode collapse (Goodfellow et al. 2014), *i.e.,* all the image features are largely projected into a single data point in the video feature space, an inverse video-to-image mapping is introduced to guarantee that image features mapped into the video feature space can be projected back to their original space.

Existing methods assume that the label spaces between the source and target domains are the same. However, in real-world applications, this assumption may not hold since the classes of the target videos are unknown. To relax this assumption, we perform partial image-to-video adaptation where the target label space is a subspace of the source label space. A class-wise alignment is proposed to match conditional distributions of the source images and the target videos in learning image-video mappings, which ensures that only the images and videos in the same class are aligned with each other.

In summary, the contributions of this paper are as follows:

- We propose a spatial-temporal causal inference framework for partial image-to-video adaptation, where a spatial-temporal causal graph is built to infer the effects of the spatial and temporal domain shifts.
- We propose causality-guided bidirectional heterogeneous mappings to adaptively reduce the spatial and temporal domain shifts under the guidance of counterfactual causality.
- We propose class-wise alignment to address partial image-to-video adaptation, significantly enhancing the ability to exploit images for video recognition.

Related Work

Image-to-video Adaptation

Many methods have been proposed to boost video recognition by leveraging images as auxiliary training data (Duan, Xu, and Chang 2012; Wang, Wu, and Jia 2014; Gan et al. 2016b,a; Li et al. 2017; Ma et al. 2017). Li et al. (2017) leverage labeled web images to train an adaptive classifier for videos, where the video frames are mapped into a low-dimensional feature space to reduce the domain gap between images and video frames. In (Ma et al. 2017), video frames and web images are combined to train a CNN model for action recognition, and discriminative action poses in labeled web images are utilized to highlight the discriminative portions of videos. These methods attempt to reduce the spatial domain shift between images and video frames. In contrast, we focus on adaptively reducing both spatial and temporal domain shifts via exploring the influences of the two domain shifts by causal inference.

Recently, several methods (Yu et al. 2018, 2019) have been proposed to reduce both spatial and temporal domain shifts via learning a domain-invariant feature. Yu et al. (2018) propose hierarchical GAN to learn the mapping from video features to image features. In (Yu et al. 2019), they further propose symmetric GAN for building bidirectional mappings between video and image features to learn augmented domain-invariant features. In (Liu et al. 2020), they utilize video keyframes as a bridge to learn common feature space of images, video keyframes and videos. Different from these methods that treat the two domain shifts equally, our method explores how the two domain shifts affect the adaptation and adpatively reduces them.

Moreover, the aforementioned methods are limited by the fully shared label space assumption while our method relaxes this assumption and focuses on a more general and practical setting, *i.e.,* partial image-to-video adaptation.

Causal Inference

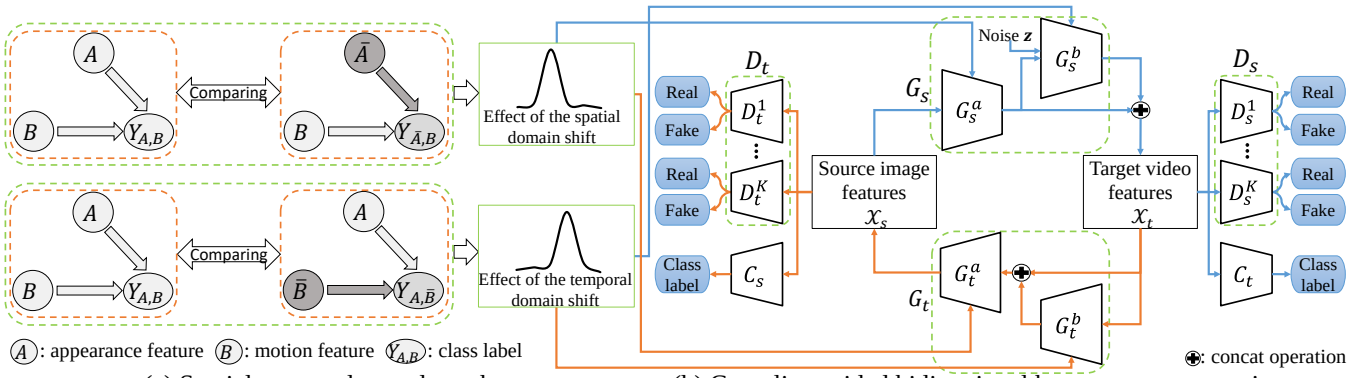
Incorporating causal inference (VanderWeele 2015) into deep learning has attracted more and more attention. It improves the explainability of the deep models and has been explored in several fields, such as scene graph generation (Tang et al. 2020; Chen et al. 2019), image classification (Chalupka, Perona, and Eberhardt 2014; Lopez-Paz et al. 2017), visual question answering (Chen et al. 2020) and object detection (Wang et al. 2020). Tang et al. (2020) present a novel scene graph generation framework to address the biased training data for scene graph generation via counterfactual causality. Lopez-Paz et al. (2017) propose a observational causal discovery technique to reveal the causal relationships between pairs of real entities in the world. Chen et al. (2020) generate counterfactual training samples by masking critical objects in images or words in questions, which enable the visual question answer model to focus on critical objects and words. Wang et al. (2020) propose a visual commonsense region-based network for object detection via causal intervention, which can learn sense-making knowledge.

To the best of our knowledge, we are the first to introduce causal inference into the domain adaptation task to reason about the contributions of different domain shifts via performing counterfactual causality on a spatial-temporal causal graph.

Spatial-Temporal Causal Inference

Problem Definition

The goal of partial image-to-video adaptation is to learn a target video classifier by adapting a source image classifier trained on the labeled images. To bridge the source image



(a) Spatial-temporal causal graph (b) Causality-guided bidirectional heterogeneous mappings

Figure 1: Overview of our method. (a) Spatial-temporal causal graph. We conduct interventions $do(A = \overline{A})$ and $do(B = \overline{B})$ to perform counterfactual thinking to infer the effects of the spatial and temporal domain shifts. (b) Causality-guided bidirectional heterogeneous mappings. The image-to-video mapping G_s and the discriminator D_s are optimized by adversarial learning to map source image features to the target video feature space with the guidance of counterfactual causality. The video-to-image mapping G_t and the discriminator D_t are optimized in a similar way.

domain and the target video domain, we propose a spatial-temporal causal inference framework that first infers the effects of the spatial and temporal domain shifts via a spatial-temporal causal graph and then adaptively reduces the two domain shifts via causality-guided bidirectional heterogeneous mappings between images and videos. A class-wise alignment is incorporated into the learning of the image-video mappings to address the partial image-to-video adaptation. Figure 1 provides an overview of our method.

We are given a labeled source image domain $\mathcal{D}_s = \{(\boldsymbol{x}_s^i, y_s^i)|_{i=1}^{N_s}\}$ and an unlabeled target video domain $\mathcal{D}_t = \{\boldsymbol{x}_t^j|_{j=1}^{N_t}\}$. $\boldsymbol{x}_s^i \in \mathcal{X}_s$ represents the appearance feature of the i-th source image and $y_s^i \in \mathcal{Y}_s$ is the class label of $\boldsymbol{x}_s^i$. $\boldsymbol{x}_t^j \in \mathcal{X}_t$ is the concatenation of the appearance feature $\boldsymbol{x}_{t,a}^j$ and the motion feature $\boldsymbol{x}_{t,b}^j$ of the j-th target video. The source feature space $\mathcal{X}_s$ is different from the target feature space $\mathcal{X}_t$, *i.e.*, $\mathcal{X}_s \neq \mathcal{X}_t$. The target label space $\mathcal{Y}_t$ is a subspace of the source label space $\mathcal{Y}_s$. The classes in $\mathcal{Y}_s$ but not in $\mathcal{Y}_t$ are denoted as outlier classes, and the common classes in $\mathcal{Y}_s$ and $\mathcal{Y}_t$ are denoted as shared classes. We use $k \in \{1, 2, ..., K\}$ to denote the index of the class, where $K = |\mathcal{Y}_s|$ is the number of source classes.

Spatial-Temporal Causal Graph

Causal graph (Pearl, Glymour, and Jewell 2016) is a directed acyclic graph, represented as $\mathcal{G} = \{\mathcal{N}, \mathcal{E}\}$, which indicates how a set of variables $\mathcal{N}$ interact with each other through the causal links $\mathcal{E}$. Since the appearance feature and the motion feature together affect the video class label, we build a spatial-temporal causal graph to model the causal relationships among the appearance feature, the motion feature and the video class label, as shown in Figure 1(a). The node A denotes the video appearance feature $\boldsymbol{x}_{t,a}^j$, and the node B denotes the video motion feature $\boldsymbol{x}_{t,b}^j$. The node Y denotes the video class label and is represented as the output of the l-layer of C_t, formulated as $\boldsymbol{y}_{\boldsymbol{x}_{t,a}^j, \boldsymbol{x}_{t,b}^j} = C_t^l(\boldsymbol{x}_t^j) = C_t^l([\boldsymbol{x}_{t,a}^j; \boldsymbol{x}_{t,b}^j])$, where $\boldsymbol{y}_{\boldsymbol{x}_{t,a}^j, \boldsymbol{x}_{t,b}^j}$ is a d-dimensional vector, and C_t^l is the l-layer of C_t. The edge $A \rightarrow Y$ denotes predicating the class label Y using the appearance feature A, and the edge $B \rightarrow Y$ indicates predicating the class label Y using the motion feature B. The target classifier C_t is used to learn these edges via the conventional cross-entropy loss of image labels y_s and image features that are mapped into $\mathcal{X}_t$.

We perform counterfactual causality (Roese 1997; Tang et al. 2020) on the spatial-temporal causal graph to infer the effects of the two domain shifts. Specifically, $do(A = a)$ denotes that we assign a certain value a to the node A. Given a video $\boldsymbol{x}_t^j$, we have $A = \boldsymbol{x}_{t,a}^j$, $B = \boldsymbol{x}_{t,b}^j$, and the output Y is denoted as $Y_{A,B} = \boldsymbol{y}_{\boldsymbol{x}_{t,a}^j, \boldsymbol{x}_{t,b}^j}$. A counterfactual scene is defined by performing intervention on the appearance feature to assess its effect, where the appearance feature is wiped out by $do(A = \overline{\boldsymbol{x}}_{t,a}^j)$ and the node B is retained as the original motion feature $\boldsymbol{x}_{t,b}^j$. The wiped-out appearance feature $\overline{A} = \overline{\boldsymbol{x}}_{t,a}^j$ is set to a zero vector with the same dimension as $\boldsymbol{x}_{t,a}^j$. The output Y after intervention is denoted as a counterfactual $Y_{\overline{A},B} = \boldsymbol{y}_{\overline{\boldsymbol{x}}_{t,a}^j, \boldsymbol{x}_{t,b}^j}$, which is counter to $Y_{A,B}$. It is natural to infer the effect of the spatial domain shift on $\boldsymbol{x}_t^j$ by comparing the factual $Y_{A,B}$ and the counterfactual $Y_{\overline{A},B}$, formulated as

$$SE(\boldsymbol{x}_t^j) = Y_{A,B} - Y_{\overline{A},B} = \boldsymbol{y}_{\boldsymbol{x}_{t,a}^j, \boldsymbol{x}_{t,b}^j} - \boldsymbol{y}_{\overline{\boldsymbol{x}}_{t,a}^j, \boldsymbol{x}_{t,b}^j}. \quad (1)$$

Similarly, we wipe out the motion feature by $do(B = \overline{\boldsymbol{x}}_{t,b}^j)$ while keeping the node A as the original appearance feature $\boldsymbol{x}_{t,a}^j$, and obtain the counterfactual $Y_{A,\overline{B}} = \boldsymbol{y}_{\boldsymbol{x}_{t,a}^j, \overline{\boldsymbol{x}}_{t,b}^j}$. Then the effect of the temporal domain shift on $\boldsymbol{x}_t^j$ is formulated by

$$TE(\boldsymbol{x}_t^j) = Y_{A,B} - Y_{A,\overline{B}} = \boldsymbol{y}_{\boldsymbol{x}_{t,a}^j, \boldsymbol{x}_{t,b}^j} - \boldsymbol{y}_{\boldsymbol{x}_{t,a}^j, \overline{\boldsymbol{x}}_{t,b}^j}. \quad (2)$$

We generate the effects of the two domain shifts for each class by averaging the inferred effects $SE(\boldsymbol{x}_t^j)$ and $TE(\boldsymbol{x}_t^j)$, formulated as

$$\begin{aligned} SE_k &= \frac{1}{N_t^k}\sum_{j=1}^{N_t}\mathbb{I}_{\hat{y}_t^j=k}SE(\boldsymbol{x}_t^j), \\ TE_k &= \frac{1}{N_t^k}\sum_{j=1}^{N_t}\mathbb{I}_{\hat{y}_t^j=k}TE(\boldsymbol{x}_t^j), \end{aligned} \tag{3}$$

where SE_k and TE_k denote the effects of the spatial and temporal domain shifts of the k-th class, respectively, $\hat{y}_t^j$ is the pseudo label of $\boldsymbol{x}_t^j$, detailed in the Class-wise Alignment section, and N_t^k is the number of videos classified into the k-th class. $\mathbb{I}_{\hat{y}_t^j=k}$ is an indicator function, meaning that if $\hat{y}_t^j = k$, the value of $\mathbb{I}_{\hat{y}_t^j=k}$ is 1 and 0 otherwise.

Causality-Guided Bidirectional Heterogeneous Mappings

After inferring the effects of the spatial and temporal domain shifts, our method learns causality-guided bidirectional heterogeneous mapping between images and videos to adaptively reduce the two domain shifts, including the image-to-video mapping $G_s : \mathcal{X}_s \rightarrow \mathcal{X}_t$ and the video-to-image mapping $G_t : \mathcal{X}_t \rightarrow \mathcal{X}_s$.

The image-to-video mapping G_s maps the image feature $\boldsymbol{x}_s^i$ to the video feature space $\mathcal{X}_t$ with the guidance of the effects of the two domain shifts, formulated as $\hat{\boldsymbol{x}}_s^i = G_s(\boldsymbol{x}_s^i, SE_{y_s^i}, TE_{y_s^i}, \boldsymbol{z})$, where $SE_{y_s^i}$ and $TE_{y_s^i}$ are the effects of the spatial and temporal domain shifts of the y_s^i-th class, respectively, and $\boldsymbol{z}$ is gussian noise to generate the absent motion feature of image. The mapped image feature $\hat{\boldsymbol{x}}_s^i$ is given by $\hat{\boldsymbol{x}}_s^i = [\hat{\boldsymbol{x}}_{s,a}^i; \hat{\boldsymbol{x}}_{s,b}^i]$, where $\hat{\boldsymbol{x}}_{s,a}^i = G_s^a([\boldsymbol{x}_s^i; SE_{y_s^i}])$ is generated by a spatial module G_s^a and $\hat{\boldsymbol{x}}_{s,b}^i = G_s^b([\hat{\boldsymbol{x}}_{s,a}^i; TE_{y_s^i}; \boldsymbol{z}])$ is generated by a temporal module G_s^b.

To learn the image-to-video mapping G_s, a discriminator D_s is constructed to distinguish the mapped image feature $\hat{\boldsymbol{x}}_s^i$ from the video feature $\boldsymbol{x}_t^j$, while G_s tries to generate $\hat{\boldsymbol{x}}_s^i$ as similar as possible to $\boldsymbol{x}_t^j$. To optimize $G_s$and D_s, the objective function is formulated as

$$\begin{aligned} &\min_{G_s}\max_{D_s}\mathcal{L}_{adv}(G_s, D_s) = \mathbb{E}_{\boldsymbol{x}_t^j}\big(\log D_s(\boldsymbol{x}_t^j)\big) \\ &+ \mathbb{E}_{\boldsymbol{x}_s^i}\Big(\log\Big(1 - D_s\big(G_s(\boldsymbol{x}_s^i, SE_{y_s^i}, TE_{y_s^i}, \boldsymbol{z})\big)\Big)\Big). \end{aligned} \tag{4}$$

The video-to-image mapping G_t attempts to map the video feature $\boldsymbol{x}_t^j$ to the image feature space $\mathcal{X}_s$, formulated as $\hat{\boldsymbol{x}}_t^j = G_t(\boldsymbol{x}_t^j, SE_{\hat{y}_t^j}, TE_{\hat{y}_t^j})$. The mapped video feature $\hat{\boldsymbol{x}}_t$ is generated via a cascade of a spatial module G_t^a and a temporal module G_t^b, formulated as $\hat{\boldsymbol{x}}_t^j = G_t^a\Big(\big[\boldsymbol{x}_{t,a}^j; G_t^b([\boldsymbol{x}_{t,b}^j; TE_{\hat{y}_t^j}]); SE_{\hat{y}_t^j}\big]\Big)$. A discriminator D_t is built to distinguish the the mapped video feature $\hat{\boldsymbol{x}}_t^j$ with the image feature $\boldsymbol{x}_s^i$. G_t and D_t are optimized via adversrial learning: $\min_{G_t}\max_{D_t}\mathcal{L}_{adv}(G_t, D_t)$.

For each image feature $\boldsymbol{x}_s^i$, we expect that its mapped image feature $\hat{\boldsymbol{x}}_s^i$ in the video feature space can be mapped back to the original image feature space via G_t, *i.e.,* $\boldsymbol{x}_s^i \rightarrow \hat{\boldsymbol{x}}_s^i \rightarrow G_t\big(\hat{\boldsymbol{x}}_s^i, SE_{y_s^i}, TE_{y_s^i}\big) \approx \boldsymbol{x}_s^i$. Similarly, for each video feature $\boldsymbol{x}_t^j$, we have $\boldsymbol{x}_t^j \rightarrow \hat{\boldsymbol{x}}_t^j \rightarrow G_s\big(\hat{\boldsymbol{x}}_t^j, SE_{\hat{y}_t^j}, TE_{\hat{y}_t^j}, \boldsymbol{z}\big) \approx \boldsymbol{x}_t^j$. So we propose a cycle consistency loss defined by the distance between the original feature $\boldsymbol{x}_s^i$ (resp. $\boldsymbol{x}_t^j$) and its corresponding reconstructed feature $G_t\big(\hat{\boldsymbol{x}}_s^i, SE_{y_s^i}, TE_{y_s^i}\big)$ (resp. $G_s\big(\hat{\boldsymbol{x}}_t^j, SE_{\hat{y}_t^j}, TE_{\hat{y}_t^j}, \boldsymbol{z}\big)$), formulated as

$$\begin{aligned} \mathcal{L}_{cyc}(G_s, G_t) = &\mathbb{E}_{\boldsymbol{x}_s^i}\Big(\big\|G_t\big(\hat{\boldsymbol{x}}_s^i, SE_{y_s^i}, TE_{y_s^i}\big) - \boldsymbol{x}_s^i\big\|_2\Big) \\ &+ \mathbb{E}_{\boldsymbol{x}_t^j}\Big(\big\|G_s\big(\hat{\boldsymbol{x}}_t^j, SE_{\hat{y}_t^j}, TE_{\hat{y}_t^j}, \boldsymbol{z}\big) - \boldsymbol{x}_t^j\big\|_2\Big), \end{aligned} \tag{5}$$

where $||\cdot||_2$ is the L2-loss to measure the distance of features.

To preserve the semantic information of features during mapping, a semantic consistency loss is introduced to ensure that the supervision information (*i.e.,* class labels) in $\mathcal{D}_s$ to be transferred to $\mathcal{D}_t$. Given an image feature $\boldsymbol{x}_s^i$ with its class label y_s^i, the class labels of the mapped image feature $\hat{\boldsymbol{x}}_s^i$ and the reconstructed image feature $G_t\big(\hat{\boldsymbol{x}}_s^i, SE_{y_s^i}, TE_{y_s^i}\big)$ should be the same as y_s^i. Therefore, the semantic consistency loss is given by

$$\begin{aligned} &\mathcal{L}_{sem}(C_s, C_t, G_s, G_t) = \mathbb{E}_{\boldsymbol{x}_s^i}\big(-\sum_{k=1}^{K}\mathbb{I}_{k=y_s^i}\log C_s(\boldsymbol{x}_s^i)\big) \\ &+ \mathbb{E}_{\boldsymbol{x}_s^i}\Big(-\sum_{k=1}^{K}\mathbb{I}_{k=y_s^i}\log C_t\Big(G_s\big(\boldsymbol{x}_s^i, SE_{y_s^i}, TE_{y_s^i}, \boldsymbol{z}\big)\Big)\Big) \\ &+ \mathbb{E}_{\boldsymbol{x}_s^i}\Big(-\sum_{k=1}^{K}\mathbb{I}_{k=y_s^i}\log C_s\Big(G_t\big(\hat{\boldsymbol{x}}_s^i, SE_{y_s^i}, TE_{y_s^i}\big)\Big)\Big). \end{aligned} \tag{6}$$

Class-wise Alignment

Through adversarial learning, the causality-guided bidirectional heterogeneous mappings succeed in reducing the domain gap between source images and target videos in the same label spaces. To relax the same label space assumption, we perform partial image-to-video adaptation and propose class-wise alignment to incorporate the learning of image-video mappings for matching the conditional distributions of the two domains. In this way, only the source image features and target video features in the same class are aligned with each other. As shown in Figure 1(b), the discriminators D_s and D_t are both constructed by K sub-discriminators, denoted as D_s^k and D_t^k, respectively, where $k \in \{1, 2, \cdots, K\}$ and $K = |\mathcal{Y}_s|$. The k-th sub-discriminator is responsible for distinguishing the video features (resp. image features) and the mapped image features (resp. mapped video features) associated with the k-th class, which encourages the image-video mappings to align the source image features and target video features with respect to the class labels.

Since the class labels of videos are unknown, we introduce a self-paced learning strategy to progressively generate pseudo labels of videos, where the class probability is used as prior knowledge to determine whether the target video is used for training at the current time. We compute the class probability $\boldsymbol{p}_t^j$ of the video feature $\boldsymbol{x}_t^j$ by averaging the class probabilities of randomly sampled video frames predicated by C_s. The pseudo label $\hat{y}_t^j$ of $\boldsymbol{x}_t^j$ is then calculated by

$$\hat{y}_t^j = \begin{cases} \arg\max_k \big(\boldsymbol{p}_t^j(k)\big) & \text{if } \max_k \big(\boldsymbol{p}_t^j(k)\big) > \tau \\ 0 & \text{otherwise,} \end{cases} \tag{7}$$

where $\boldsymbol{p}_t^j(k)$ is the k-th element of $\boldsymbol{p}_t^j$ and represents the probability of assigning $\boldsymbol{x}_t^j$ to the k-th class, and τ is the threshold to filter out video features with low confidence in the predicted class probabilities.

With the predicated pseudo label $\hat{y}_t^j$ of $\boldsymbol{x}_t^j$, the objective function of G_s and D_s is formulated by

$$\begin{aligned} &\min_{G_s}\max_{D_s} \mathcal{L}_{adv}(G_s, D_s) = \mathbb{E}_{(\boldsymbol{x}_t^j, \hat{y}_t^j)}\Big(\textstyle\sum_{k=1}^{K} \mathbb{I}_{k=\hat{y}_t^j}\big(\log D_s^k(\boldsymbol{x}_t^j)\big)\Big) \\ &+\mathbb{E}_{(\boldsymbol{x}_s^i, y_s^i)}\left(\textstyle\sum_{k=1}^{K} \mathbb{I}_{k=y_s^i}\left(\log\Big(1 - D_s^k\big(G_s(\boldsymbol{x}_s^i, SE_{y_s^i}, TE_{y_s^i}, \boldsymbol{z})\big)\Big)\right)\right) \end{aligned} \tag{8}$$

Similarly, the discriminator D_t is constructed with K sub-discriminators, denoted as D_t^k. The adversarial loss of learning G_t and D_t is similar as Eq. (8).

Taken together, all the loss functions mentioned above form the complete objective:

$$\begin{aligned} &\min_{\{G_s, G_t, C_s, C_t\}} \max_{\{D_s, D_t\}} \mathcal{L}(G_s, G_t, C_s, C_t, D_s, D_t) \\ &= \mathcal{L}_{adv}(G_s, D_s) + \mathcal{L}_{adv}(G_t, D_t) \\ &+ \lambda\big(\mathcal{L}_{sem}(C_s, C_t, G_s, G_t) + \mathcal{L}_{cyc}(G_s, G_t)\big), \end{aligned} \tag{9}$$

where λ is a trade-off parameter.

Experiments

Datasets

We conduct experiments on two video benchmarks, *i.e.,* UCF101 (U) (Soomro, Zamir, and Shah 2012) and HMDB51 (H) (Kuehne et al. 2011). With the UCF101 dataset as the target domain, we use the Standford40 (S) dataset (Yao et al. 2011) as the source domain. With the HMDB51 dataset as the target domain, we use the EADs (E) dataset (Yu et al. 2018) as the source domain. Therefore, there are two partial image-to-video adaptation tasks: S→U and E→H.

The UCF101 dataset contains $13,000$ videos of 101 action classes collected from YouTube. The HMDB51 dataset has $6,766$ video clips of 51 action classes extracted from commercial movies and public datasets. The Standford40 dataset contains $9,532$ images, collected from Google, Bing and Flickr. It has 40 action classes and each action class has 180 to 300 images with large variations in human pose, appearance and background. The EADs dataset (Yu et al. 2018) consists of the Standford40 and HIIT (Tanisik, Zalluhoglu, and Ikizler-Cinbis 2016) datasets. It has $11,504$ images with 50 action classes. Each action class has at least 150 images.

For the S→U task, there are 12 shared classes between the Standford40 and UCF101 datasets. We use all the images of the Standford40 dataset as the source domain and the videos of 12 shared classes of the UCF101 dataset as the target domain. For the E→H task, there are 13 shared classes between the EADs and HMDB51 datasets. All the images of the EADs dataset are used as the source domain and the videos of 13 shared classes of the HMDB51 dataset are used as the target domain. We use all the labeled source images and unlabeled target videos for training.

Implementation Details

For the source images, we use ResNet-50 (He et al. 2016) pre-trained on the ImageNet dataset (Deng et al. 2009) to extract a 2048-dimensional vector from the *pool5* layer as the appearance feature. For the target videos, we use two-stream I3D networks (Carreira and Zisserman 2017) pre-trained on the Kinetics dataset (Kay et al. 2017) to extract 1024-dimensional optical flow and 1024-dimensional RGB features that represent the motion and appearance information of the videos, respectively.

We employ the Adam solver (Kingma and Ba 2015) with the batch size of 16, including 8 source images and 8 target videos. The trade-off parameter λ in Eq. (9) is set to 100. The dimension of the noise $\boldsymbol{z}$ is set to 1024. We set the threshold τ in Eq. (7) as 0.5, as it achieves the best performance on the two datasets. All the networks are trained from scratch with 400 epochs. We keep the same learning rate for the first 200 epochs and linearly decay the rate to zero for the next 200 epochs. The initial learning rate of the S→U and E→H tasks are set to 0.0001 and 0.00005, respectively. The code is available at https://github.com/ChenJinBIT/HPDA.

Compared Methods

Our method addresses the new problem of partial image-to-video adaptation where $\mathcal{X}_s \neq \mathcal{X}_t$ and $\mathcal{Y}_t \subset \mathcal{Y}_s$. We compare our method with traditional heterogeneous image-to-video adaptation methods that make the fully shared label space assumption (*i.e.,* $\mathcal{X}_s \neq \mathcal{X}_t$ and $\mathcal{Y}_s = \mathcal{Y}_t$): Hierarchical Generative Adversarial Networks (HiGAN) (Yu et al. 2018) and Symmetric Generative Adversarial Networks (SymGAN) (Yu et al. 2019). Our method is also compared with partial homogeneous domain adaptation methods that require the same feature spaces of different domains (*i.e.,* $\mathcal{X}_s = \mathcal{X}_t$ and $\mathcal{Y}_t \subset \mathcal{Y}_s$): ResNet (He et al. 2016), Importance Weighted Adversarial Nets (IWAN) (Zhang et al. 2018), Selective Adversarial Networks (SAN) (Cao et al. 2018a), Partial Adversarial Domain Adaptation (PADA) (Cao et al. 2018b), Example Transfer Network (ETN) (Cao et al. 2019), Adaptive Feature Norm (AFN) (Xu et al. 2019), and Universal Source-Free Domain Adaptation (USFDA) (Kundu et al. 2020).

For all the compared methods, we use ResNet-50 (He et al. 2016) as the backbone which is pre-trained on the ImageNet dataset. Note that ResNet, IWAN, SAN, PADA, ETN, AFN and USFDA require images as input, and we take source images and target video frames as input to train the networks followed by (Yu et al. 2018, 2019). During test-

Method	H	P	S→U	E→H
HiGAN (Yu et al. 2018)	✓		63.50	24.06
SymGAN (Yu et al. 2019)	✓		57.08	23.94
ResNet (He et al. 2016)		✓	70.22	24.85
IWAN (Zhang et al. 2018)		✓	72.52	23.53
SAN (Cao et al. 2018a)		✓	73.00	28.00
PADA (Cao et al. 2018b)		✓	70.50	32.80
ETN (Cao et al. 2019)		✓	80.10	26.50
AFN (Xu et al. 2019)		✓	84.63	29.59
USFDA (Kundu et al. 2020)		✓	84.38	25.00
Our method	✓	✓	**95.15**	**52.66**

Table 1: Classification accuracies (%) on the S→U and E→H tasks. H and P indicate the heterogeneous feature spaces and the partial label spaces between images and videos, respectively.

ing, the decision scores of the video frames are averaged to determine the final class label of the video.

Results

Table 1 shows the classification accuracies of different methods on both S→U and E→H tasks. The first part shows the results of traditional heterogeneous image-to-video adaptation methods, and the second part shows the results of partial homogeneous domain adaptation methods. From the results, we have several interesting observations as follows.

Our method achieves much better performance than traditional heterogeneous image-to-video adaptation methods, probably due to the following reasons. First, our method infers how the spatial and temporal domain shifts affect the adaptation via causal inference to adaptively reduce the two domain shifts via causality-guided image-video mappings, instead of treating the two domain shifts equally. Second, the class-wise alignment incorporated into the learning of image-video mappings effectively matches the conditional distributions of the source and target domains, thus avoiding the false alignment of source images in the outlier classes and target videos.

Our method performs much better than partial homogeneous domain adaptation methods. The reasons are as followings. First, our method maps the image feature to the video feature space under the guidance of the inferred effect of the spatial domain shift which reflects the causal relationships in video classification. Second, the motion information is captured by learning bidirectional image-video mappings between heterogeneous feature spaces. In contrast, the partial homogeneous domain adaptation methods represent videos as a bag of images, thus ignoring the dynamic motion information in videos.

Ablation Studies

To better understand the effect of each component, we conduct ablation experiments on both S→U and E→H tasks, as shown in Table 2. Our method is compared with several variations: without effects inferred by causal inference ("w/o causality"), without the effect of the spatial domain shift ("w/o spatial causality"), without the effect of the temporal domain shift ("w/o temporal causality"), without the video-to-image mapping ("w/o bidirection"), without the cycle consistency loss ("w/o cycle consistency"), removing the third term from Eq.(6) ("w/o semantic consistency"), and without self-paced learning ("w/o self-paced learning").

Method	S→U	E→H
w/o causality	90.74	43.79
w/o spatial causality	91.92	48.40
w/o temporal causality	93.85	45.33
w/o bidirection	43.13	38.05
w/o cycle consistency	34.80	32.96
w/o semantic consistency	94.16	48.93
w/o self-paced learning	93.69	48.34
Our method	**95.15**	**52.66**

Table 2: Classification accuracies (%) of ablation studies on both S→ U and E→H tasks.

From Table 2, it is noteworthy to make several observations. First, our method achieves 4.41% and 8.87% gains over "w/o causality" on the S→U and E→H tasks, respectively, and also outperforms "w/o spatial causality" and "w/o temporal causality". This clearly demonstrates that both the spatial and temporal domain shifts should be inferred to boost the performance and explainability of image-to-video adaptation. Second, when removing the video-to-image mapping ("w/o bidirection") or the cycle consistency loss ("w/o cycle consistency"), the classification accuracies substantially degrade due to the problem of model collapse, which validates the effectiveness of bidirectional image-video mappings for heterogeneous domain adaptation. Third, our method outperforms "w/o semantic consistency", showing that it is beneficial to capture the semantic information during mapping. Finally, "w/o self-paced learning" works worse than our method, which clearly verifies the benefit of progressive matching between target video features and source image features on reducing the misclassification of target video features.

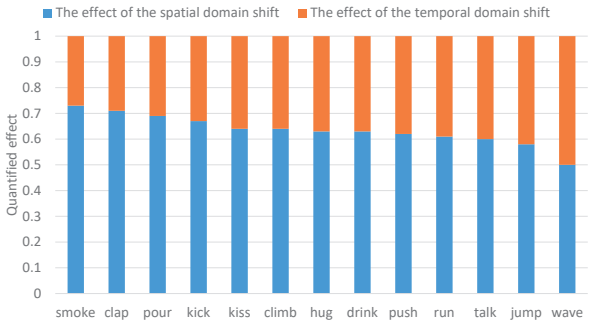
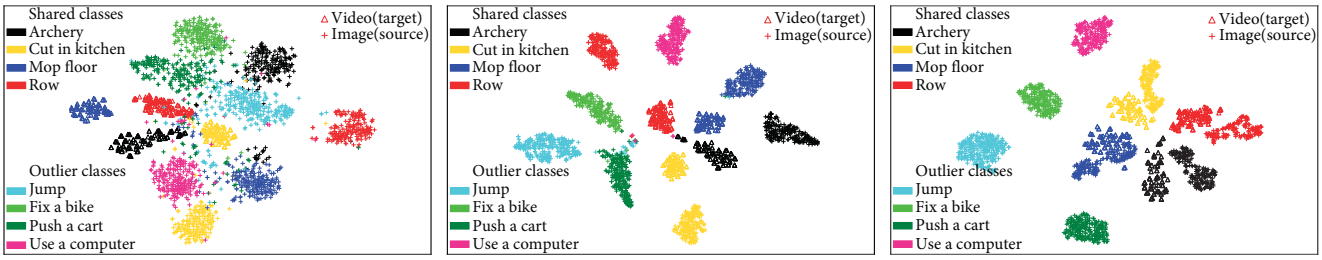


Figure 2: Quantified effects of the spatial and temporal domain shifts on the E→H task. The blue and orange bars denote the effects of the spatial and temporal domain shifts, respectively. The horizontal axis is the target video classes and the vertical axis is the value of the quantified effect.

Method	pour	kiss	push	hug	climb	drink	kick	clap	run	smoke	talk	jump	wave	Avg
w/o causality	69.23	76.85	37.20	0.00	33.77	43.85	71.57	80.19	45.69	43.10	23.85	50.83	0.00	43.79
Our method	77.69	87.04	44.51	6.78	31.13	58.46	83.33	86.79	48.28	50.86	58.72	54.17	10.58	52.66

Table 3: Classification accuracies (%) of "w/o causality" and our method for each class on the E→H task.



(a) Original (b) w/o causality (c) Our method

Figure 3: Feature visualization on the S→U task. "△" and " + " denote the target video feature and the source image feature, respectively. Different colors denote different classes as shown in the legend.

Causality Analysis

To further analyze the effectiveness of casual inference, we report the classification accuracies of "w/o causality" and our method for each class on the E→H task in Table 3, and illustrate the effects of the two domain shifts quantified by entropy in Figure 2.

From results in Table 3, it is interesting to observe that our method outperforms "w/o causality" not only in motion-light classes (8.46% gains on "smoke", 10.19% gains on "kiss") but also on motion-rich classes (10.58% gains on "wave" and 3.34% gains on "jump"). The promising performances validate the effectiveness of counterfactual causality on revealing the causal relationships in image-to-video adaptation.

In Figure 2, the larger quantified effect of the spatial domain shift (blue bars) indicates that the spatial domain shift is more critical to adaptation, and the large quantified effect of the temporal domain shift (orange bars) indicates that the temporal domain shift is more important. From the results, we observe that the learned effects correctly reflect the importance of different domain shifts. For example, "wave" has a larger quantified effect of the temporal domain shift than "smoke" since videos of "wave" contain more motion information than videos of "smoke", and the temporal domain shift in "wave" needs much more attention than that in "smoke". This shows that our spatial-temporal causal graph can infer the contributions or weights of the spatial and temporal domain shifts for image-to-video adaptation.

Feature Visualization

To further evaluate the effectiveness of the causality-guided bidirectional heterogeneous mappings, we visualize the distributions of the original features ("original"), the learned features by image-video mappings without causal inference ("w/o causality") and the learned features by image-video mappings with causal inference ("Our method") in the video feature space on the S→U task, as shown in Figure 3(a), 3(b) and 3(c), respectively. For clarity, we visualize four shared classes and four outlier classes using t-SNE embeddings (Donahue et al. 2014).

From Figure 3, we make several interesting observations. First, there is a large domain gap between the two domains as shown in Figure 3(a), and even some source image features and target video features of the same class fall into different clusters. Second, our method aligns the learned source image features with the target video features better than "w/o causality" owing to the guidance of counterfactual causality. Third, in Figure 3(c), the learned source image features in the shared classes are aligned with the target video features, clearly demonstrating that our method can successfully reduce the heterogeneous domain shift. The learned source image features in the outlier classes are not aligned with the target video features, validating the superiority of class-wise alignment to avoid the false alignment of the target classes with the outlier classes.

Conclusion

We have presented a spatial-temporal causal inference framework for partial image-to-video adaptation. The proposed spatial-temporal causal graph can help infer how the spatial and temporal domain shifts affect the adaptation via counterfactual causality. With the guidance of the effects of the two domain shifts, the learned causality-guided bidirectional heterogeneous mappings can succeed in adaptively reducing the two domain shifts and enhancing the explainability of the image-to-video adaptation. The class-wise alignment incorporated in the image-video mappings is capable of alleviating the false alignment of the target classes and outlier classes by matching the conditional distributions of the two domains. Extensive experiments on two benchmark datasets have validated the effectiveness of our method.

Acknowledgments

This work was supported in part by the Natural Science Foundation of China (NSFC) under Grants No. 62072041 and No. 61673062, and Alibaba Group through Alibaba Innovative Research Program.

References

Cao, Z.; Long, M.; Wang, J.; and Jordan, M. I. 2018a. Partial Transfer Learning With Selective Adversarial Networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2724–2732.

Cao, Z.; Ma, L.; Long, M.; and Wang, J. 2018b. Partial Adversarial Domain Adaptation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 139–155.

Cao, Z.; You, K.; Long, M.; Wang, J.; and Yang, Q. 2019. Learning to Transfer Examples for Partial Domain Adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2985–2994.

Carreira, J.; and Zisserman, A. 2017. Quo vadis, action recognition? a new model and the kinetics dataset. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 6299–6308.

Chalupka, K.; Perona, P.; and Eberhardt, F. 2014. Visual Causal Feature Learning. *Computer Science* .

Chen, L.; Yan, X.; Xiao, J.; Zhang, H.; Pu, S.; and Zhuang, Y. 2020. Counterfactual Samples Synthesizing for Robust Visual Question Answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 6979–6897.

Chen, L.; Zhang, H.; Xiao, J.; He, X.; and Chang, S. F. 2019. Counterfactual Critic Multi-Agent Training for Scene Graph Generation. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 4613–4623.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. Imagenet: A large-scale hierarchical image database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 248–255.

Donahue, J.; Jia, Y.; Vinyals, O.; Hoffman, J.; Zhang, N.; Tzeng, E.; and Darrell, T. 2014. Decaf: A deep convolutional activation feature for generic visual recognition. In *Proceedings of the International Conference on Machine Learning (ICML)*, 647–655.

Duan, L.; Xu, D.; and Chang, S.-F. 2012. Exploiting web images for event recognition in consumer videos: A multiple source domain adaptation approach. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 1338–1345.

Gan, C.; Sun, C.; Duan, L.; and Gong, B. 2016a. Webly-supervised video recognition by mutually voting for relevant web images and web video frames. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 849–866.

Gan, C.; Yao, T.; Yang, K.; Yang, Y.; and Mei, T. 2016b. You lead, we exceed: Labor-free video concept learning by jointly exploiting web videos and images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 923–932.

Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y. 2014. Generative adversarial nets. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2672–2680.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 770–778.

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

Kingma, D. P.; and Ba, J. 2015. Adam: A method for stochastic optimization. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Kuehne, H.; Jhuang, H.; Garrote, E.; Poggio, T.; and Serre, T. 2011. HMDB: a large video database for human motion recognition. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2556–2563.

Kundu, J. N.; Venkat, N.; M V, R.; and Babu, R. V. 2020. Universal Source-Free Domain Adaptation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 4544–4553.

Li, J.; Wong, Y.; Zhao, Q.; and Kankanhalli, M. S. 2017. Attention transfer from web images for video recognition. In *Proceedings of the ACM International Conference on Multimedia (ACM MM)*, 1–9.

Liu, Y.; Lu, Z.; Li, J.; Yang, T.; and Yao, C. 2020. Deep Image-to-Video Adaptation and Fusion Networks for Action Recognition. *IEEE Transactions on Image Processing (TIP)* 29: 3168–3182.

Lopez-Paz, D.; Nishihara, R.; Chintala, S.; Scholkopf, B.; and Bottou, L. 2017. Discovering Causal Signals in Images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 58–66.

Ma, S.; Bargal, S. A.; Zhang, J.; Sigal, L.; and Sclaroff, S. 2017. Do less and achieve more: Training cnns for action recognition utilizing action images from the web. *Pattern Recognition* 68: 334–345.

Pearl, J.; Glymour, M.; and Jewell, N. P. 2016. *Causal inference in statistics: A primer*. John Wiley & Sons.

Roese, N. J. 1997. Counterfactual thinking. *Psychological bulletin* 121(1): 133.

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

Sun, C.; Shetty, S.; Sukthankar, R.; and Nevatia, R. 2015. Temporal localization of fine-grained actions in videos by domain transfer from web images. In *Proceedings of the ACM International Conference on Multimedia (ACM MM)*, 371–380.

Tang, K.; Niu, Y.; Huang, J.; Shi, J.; and Zhang, H. 2020. Unbiased Scene Graph Generation from Biased Training. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 3713–3722.

Tanisik, G.; Zalluhoglu, C.; and Ikizler-Cinbis, N. 2016. Facial descriptors for human interaction recognition in still images. *Pattern Recognition Letters* 73: 44–51.

VanderWeele, T. 2015. *Explanation in causal inference: methods for mediation and interaction*. Oxford University Press.

Wang, H.; Wu, X.; and Jia, Y. 2014. Video annotation via image groups from the web. *IEEE Transactions on Multimedia (TMM)* 16(5): 1282–1291.

Wang, T.; Huang, J.; Zhang, H.; and Sun, Q. 2020. Visual Commonsense R-CNN. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 10760–10770.

Xu, R.; Li, G.; Yang, J.; and Lin, L. 2019. Larger Norm More Transferable: An Adaptive Feature Norm Approach for Unsupervised Domain Adaptation. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 1426–1435.

Yao, B.; Jiang, X.; Khosla, A.; Lin, A. L.; Guibas, L.; and Fei-Fei, L. 2011. Human action recognition by learning bases of action attributes and parts. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 1331–1338.

Yu, F.; Wu, X.; Chen, J.; and Duan, L. 2019. Exploiting Images for Video Recognition: Heterogeneous Feature Augmentation via Symmetric Adversarial Learning. *IEEE Transactions on Image Processing (TIP)* 28(11): 5308–5321.

Yu, F.; Wu, X.; Sun, Y.; and Duan, L. 2018. Exploiting Images for Video Recognition with Hierarchical Generative Adversarial Networks. In *International Joint Conference on Artificial Intelligence (IJCAI)*, 1107–1113.

Zhang, J.; Ding, Z.; Li, W.; and Ogunbona, P. 2018. Importance Weighted Adversarial Nets for Partial Domain Adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 8156–8164.

Zhang, J.; Han, Y.; Tang, J.; Hu, Q.; and Jiang, J. 2016. Semi-supervised image-to-video adaptation for video action recognition. *IEEE Transactions on Cybernetics (T-CYB)* 47(4): 960–973.