This article has been accepted for publication in IEEE Transactions on Multimedia. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TMM.2023.3318300

Deconfounding Causal Inference for Zero-shot Action Recognition

Junyan Wang, Yiqi Jiang, Yang Long, Xiuyu Sun, Maurice Pagnucco, Yang Song

***Abstract*—Zero-shot action recognition (ZSAR) aims to recognize unseen action categories in the test set without corresponding training examples. Most existing zero-shot methods follow the feature generation framework to transfer knowledge from seen action categories to model the feature distribution of unseen categories. However, due to the complexity and diversity of actions, it remains challenging to generate unseen feature distribution, especially for the cross-dataset scenario when there is potentially larger domain shift. This paper proposes a Deconfounding Causal GAN (DeCalGAN) for generating unseen action video features with the following technical contributions: 1) Our model unifies compositional ZSAR with traditional visual-semantic models to incorporate local object information with global semantic information for feature generation. 2) A GAN-based architecture is proposed for causal inference and unseen distribution discovery. 3) A deconfounding module is proposed to refine representations of local object and global semantic information confounder in the training data. Action descriptions and random object feature after causal inference are then used to discover unseen distributions of novel actions in different datasets. Our extensive experiments on Cross-Dataset Zero-Shot Action Recognition (CD-ZSAR) demonstrate substantial improvement over the UCF101 and HMDB51 standard benchmarks for this problem.**

***Index Terms*—Zero-shot Learning, Action Recognition, Causal Inference.**

I. Introduction

ACTION recognition, also known as video recognition, is a fundamental problem in video understanding. Over the last decade, there has been increasing research attention in video action recognition, with the emergence of high-quality large-scale action recognition datasets. Recently, a wide range of popular and successful model architectures have been designed for action recognition tasks. However, these methods require a large number of training data for each action class, which requires costly and laborious annotations of videos, and the trained model does not generalize to unseen action categories. It is infeasible and extremely expensive to annotate action videos with the ever-increasing need for new categories. To solve this problem, *zero-shot action recognition* has recently drawn considerable interest, with its ability to identify unseen action categories without labeled examples.

Junyan Wang, Yang Song and Maurice Pagnucco are with School of Computer Science and Engineering, University of New South Wales, Australia. E-mail: junyan.wang@unsw.edu.au; yang.song1@unsw.edu.au; morri@unsw.edu.au.

Yiqi Jiang and Xiuyu Sun are with DAMO Academy, Alibaba Group, China. E-mail: yiqi.jyq@alibaba-inc.com; xiuyu.sxy@alibaba-inc.com.

Yang Long is with the Department of Computer Science, Durham University, UK. E-mail: yang.long@ieee.org.

Manuscript submitted 2022.

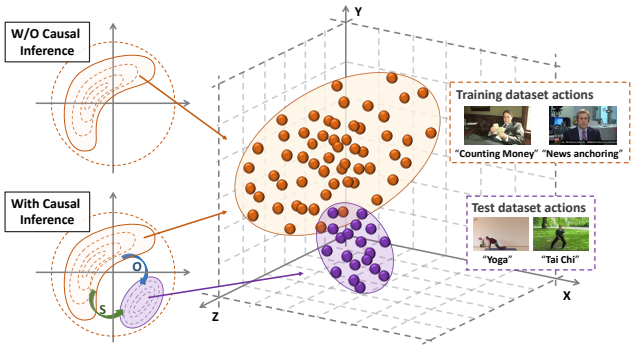


Fig. 1: An illustration of cross-dataset zero-shot action recognition and our proposed causal inference application. With semantic information "S" and object information "O", the proposed DeCalGAN can generate unseen video representations by causal inference. Orange balls represent videos in the training dataset and purple ones denote videos in the test dataset.

Existing studies of ZSAR have mainly focused on inner-dataset seen/unseen splits due to the requirement of human-defined domain attributes [1], [2]. This setting is not very practical since a new dataset could require re-training, as different datasets might exhibit cross-domain issues. Moreover, regardless of the type of side information we adopt, the generalization capability of these approaches could be lacking, due to the higher degree of domain shift across datasets. Recently, a more realistic cross-dataset zero-shot action recognition (CD-ZSAR) task [3] was proposed, which aims to make large-scale pretrained model transfer seamlessly to unseen classes across new datasets, and thus our work focuses on CD-ZSAR scenario.

One of the main challenges of CD-ZSAR is the **weak knowledge representation**. Early research [4], [5] in zero-shot learning focused on developing a compatibility model, and most of these methods are attribute-based. In CD-ZSAR, it is infeasible to design a universal attribute-space that is applicable to every new task and dataset. Therefore, word embedding is currently the most efficient side information for CD-ZSAR. Also, videos are highly complex containing both spatial and temporal information, and hence it is difficult to apply an automatic word-embedding model to represent the global semantic knowledge of a class. Recent studies have investigated how object information [6], [7] or semantic embedding [8] performs in action recognition, and these studies have demonstrated successful outcomes, with object

This article has been accepted for publication in IEEE Transactions on Multimedia. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TMM.2023.3318300

Fig. 2: Illustrations of Elaborative Description in "tai chi", "fencing", "diving", "skiing", "yoga" and "bowling" actions from UCF101 dataset.

information and semantic embedding representing spatial and temporal information, respectively. However, the basic video-based backbone is ineffective in learning different domain knowledge in the zero-shot setting. Another main challenge is the **unseen distribution**. Recently, thanks to advances in generative adversarial networks (GANs), many approaches have been proposed to directly generate unseen samples in zero-shot tasks [9]–[11]. However, although GAN is able to generate data from the distribution of training dataset, it cannot expand the original distribution without seeing novel samples. As shown in Figure 1, part of the feature distribution of unseen action videos is different from the training data, which means the generated unseen action videos by a basic GAN model is difficult to represent the unseen distribution.

The above challenges motivate us to design a new framework for cross-dataset zero-shot learning action recognition with two sub-tasks, *i.e.* **compositional generation** and **distribution inference**. In general, video data distribution is more complex than that of image-level data. Instead of directly generating video data, our focus lies on generating lower-dimensional features extracted by the conventional backbone. Firstly, an action video contains both spatial and temporal information, such as characters, movements and interaction. Weak knowledge representations such as word embedding can be compensated by compositional knowledge which consists of local object information using pretrained detectors and global semantic information using Elaborative Description (ED) [12] as shown in Figure 2. The second task aims to generate unseen action representations that can effectively infer the unseen distribution. We design a **Deconfounding Causal GAN (DeCalGAN)** framework with the following insights: 1) We propose a novel approach for generating compositional features from dual channels, *i.e.,* Elaborative Descriptions (ED) and object detection, based on causal inference. Causal inference has been shown to be useful in compositional zero-shot learning, as it can identify the true causal relationships between variables [13]. Our approach builds a structured causality-inspired generative model that captures the causal relationships between features and actions. Specifically, we use a conditional causal graph to infer action features based on their corresponding semantic and object representations. 2) One of the main challenges in representation learning from videos is the presence of confounding factors that can arise due to the diverse range of latent information. To address this challenge, we propose a deconfounding module that can handle the interference between global semantic and local object features. This is particularly important since each class can have an unlimited number of possible compositions of objects, and distinguishing between confounded object feature dimensions and semantic feature dimensions is critical. By ensuring that each factor is kept independent, our generative model can accurately infer unseen distributions. 3) Our proposed approach achieves zero-shot recognition by generating unseen action features based on random object information and EDs of test actions. Our method outperforms existing approaches on various benchmarks, demonstrating the effectiveness of our causal inference-based generative model for compositional feature generation. This work provides a promising direction for addressing the challenge of zero-shot recognition in video analysis. Our contributions are summarized as follows:

- To the best of our knowledge, we present the first causal inference approach to address the **unseen distribution** problem for cross-dataset zero-shot recognition (CD-ZSAR), and we propose a GAN architecture as a new paradigm for causal inference.
- The proposed Deconfounding Causal GAN (DeCalGAN) consists of a reconstruction module and a deconfounding module that can make confounding features learned from the source domain better generalize to the unseen distribution in the test domain.
- The proposed DeCalGAN is introduced to unify compositional and generative frameworks to tackle the challenging CD-ZSAR problem. Local object information and global semantic descriptions can jointly generate missing distributions across different datasets and achieve state-of-the-art performance.

This article has been accepted for publication in IEEE Transactions on Multimedia. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TMM.2023.3318300

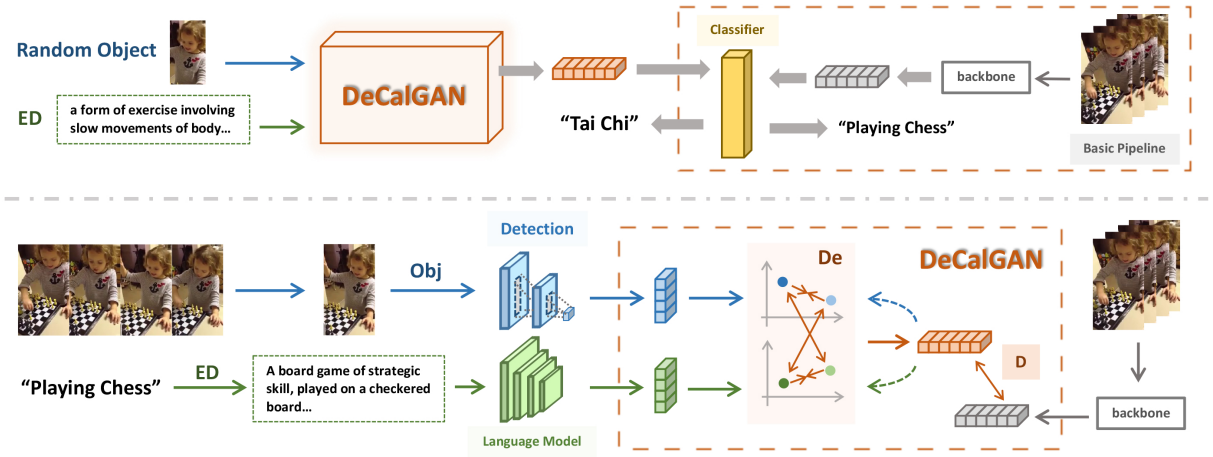


Fig. 3: Architecture of our DeCalGAN enhanced ZSAR model. The bottom part shows our proposed DeCalGAN training process, which incorporates global semantic information and local object information. The top-left part shows our proposed causal inference pipeline, and the top-right represents the basic pipeline following the work of [22].

II. Related Work

Action recognition has drawn a significant amount of attention from the computer vision community in the past few years [14]–[19]. Some attempts have been made to design an efficient method by combining a lightweight temporal module with a conventional 2D CNN-based backbone [16], [20]. For example, Li *et al.* [16] proposed a Temporal Excitation and Aggregation (TEA) block, including a motion excitation module and a multiple temporal aggregation module, specifically designed to capture both short- and long-range temporal evolution. Recent research shows that pure 3D CNNs outperform 2D ones on large-scale benchmarks [21], as 3D CNNs can jointly capture the spatio-temporal features in a unified framework. However, most approaches rely on specific large-scale training video datasets with annotated samples per action class. In this work, we focus on zero-shot action recognition in which test raw data is unavailable.

A. Zero-shot Action Recognition

Many zero-shot action recognition methods have been proposed recently [1], [6], [8], [22], [23]. An initial work [1] used a set of manually defined attributes to describe the spatio-temporal evolution of actions in a video. Other early attempts [8], [23] follow a standard strategy, which first extracts visual features from videos and then trains a joint model that maps the visual embedding to a semantic embedding space. The work of [23] explores word vectors as a shared semantic space to embed labels and videos for zero-shot action recognition. [6] proposed a spatial-aware object embedding for zero-shot action localization and classification. Besides, the work of [24] devises a simple semantic transfer scheme that embeds semantic relatedness information between seen and unseen classes to composite unseen visual prototypes. However, previous studies have typically focused on inner-dataset seen/unseen splits. A recent work [22] proposed to train a 3D CNN to predict word embedding of labels as end-to-end training for CD-ZSAR. In this work, we also follow the cross-dataset protocol of [22] and apply causal inference to generate unseen class representations.

B. Causal Inference

Causality [25], [26] has inspired computer vision researchers to design new methodologies for various tasks such as image recognition [27] and domain adaption [28], [29]. The work of [30] learns a conditional-GAN model jointly with a causal model of label distribution. In contrast, our proposed DeCalGAN jointly learns semantic and object components by causal inference. In addition, [13] formalizes causal inference as a problem of finding the most likely intervention, while another method [31] explicitly promotes the dependency between all primitives and their compositions in the learned graph embedding. Recently, [32] developed a Deconfounded Cross-modal Matching (DCM) method to remove the confounding effects of moment location in the video moment retrieval task. In this work, our proposed method incorporates adversarial training for deconfounding compositional confounders to better generalize to the unseen distribution.

III. Methodology

In cross-dataset zero-shot action recognition (CD-ZSAR), let $D = \{(x_1, y_1), \ldots, (x_N, y_N)\} \subseteq \boldsymbol{X} \times \boldsymbol{Y}$ denote the training dataset that consists of pairs of action videos x and their class labels y, where N is the number of videos. $y \in \{1, \ldots, C\}$ contains C discrete labels of training classes. Given a target dataset D_t, where D_t does not overlap with D ($\boldsymbol{Y} \cap \boldsymbol{Y}_t = \varnothing$), we first train a classification model on D and then test on D_t. To achieve this, we follow the evaluation protocol in [22], using nearest-neighbor search in a semantic class embedding space. However, it is still difficult for zero shot learning classifier to generate representative embedding for the unseen test classes, due to the weak knowledge representation and unseen distribution challenges. Thus, we propose the Deconfounding Causal GAN (DeCalGAN) for unseen action generation to enhance the classifier, details of which are introduced below, and the overall framework is shown in Figure 3.

A. Revisiting Causality

We first give a brief introduction of causality, on which our proposed DeCalGAN is based. In this work, we apply structural causal models (SCMs) [25], which contain structural equations and directed acyclic graphs.

Definition 1: A structural causal model is a triple $\mathcal{M} = (\boldsymbol{V}, \boldsymbol{U}, \boldsymbol{\mathcal{F}})$, where $\boldsymbol{U}$ is a set of exogenous variables, $\boldsymbol{V}$ denotes a set of endogenous variables and $\boldsymbol{\mathcal{F}}$ is a group of deterministic functions.

Concretely, exogenous variables exist outside the model that we do not care about their causes, and each endogenous variable in the model is the child of at least one exogenous variable. Also, exogenous variables cannot be children of other variables, especially endogenous variables. If we know the value of each exogenous variable, we can completely determine the value of each endogenous variable by using the function in $\boldsymbol{\mathcal{F}}$.

Causal Graph. A causal model $\mathcal{M}$ has a corresponding causal graph $\mathcal{G}$. Nodes in the graph represent $\boldsymbol{V}$ and $\boldsymbol{U}$ in the SCM, and edges in the graph represent the functions $\boldsymbol{\mathcal{F}}$. This means if a variable X is the child node of Y, then Y is a direct cause for X. And if X is a parent node of Y, then X is the potential cause of Y.

Confounding. The common cause in a pseudo-correlation is known as confounder, also called a bias. The pseudo-correlation caused by confounders is mixed with the real causal effect, which is the case of confounding. One of the goals of causal inference is to try to eliminate the bias caused by confounding, and find the true causal relationship.

Do-operator. Taking variables as conditions change our view of the variable, while intervention changes the variable itself. In a causal model $\mathcal{M}$, intervention $do(\boldsymbol{X} = x)$ is performed by replacing the original function $\boldsymbol{X} = f_x(P_x, \boldsymbol{U}_x)$ with $\boldsymbol{X} = x$, where f_x represents a deterministic function, $P_x \subseteq \boldsymbol{V}/V_i$ and $\boldsymbol{X} \in \boldsymbol{V}$, that the intervention operation will delete all edges pointing to the variable. Thus, the intervention operation changes the distribution of the original data, but does not change the distribution of the original data under the condition of variables.

B. Deconfounding Causal GAN

Although there has been significant research investigating zero-shot learning, learning visual-semantic embedding still remains a challenging issue in video-based tasks. In the CD-ZSAR scenario, a key challenge is to represent the unseen actions that do not exist in the training dataset. We consider that action information is composed of semantic and object features, where semantic information can be used to describe the action progress itself, and object information plays a crucial role in identifying explicit action categories, such as sports and makeup. Since we have action videos and their class labels from the training dataset, we can extract semantic and object information through existing recognition and detection methods. Compared to existing methods that directly learn from seen actions or indirectly learn object/semantic information to recognize unseen actions, our proposed generative model learns the compositional semantic and object representation. To achieve effective learning of such compositional information, we apply causal inference into our adversarial learning methods.

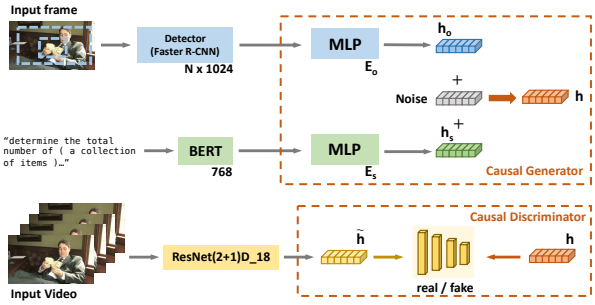


Fig. 4: Details of DeCalGAN architecture, which consists of causal generator and causal discriminator. The causal generator detail is shown in the top and the network detail of the discriminator is shown at the bottom.

In our method, we approach the action recognition problem as modeling video features caused by real-word entities, and we consider two "elementary factors" which are "Semantic" $s \in \boldsymbol{S}$ and "Object" $o \in \boldsymbol{O}$ that are independent in the training data. Thus, our model is designed to estimate $p(h|s, o)$, the likelihood of the feature vector h of a video, conditioned on a tuple (s, o) of semantic-object features. Although we consider the combination of semantic and object information capable of inferring the action class, a video contains much more information than just images, which makes it difficult to learn a comprehensive video representation. For example, action information is also characterized by speed difference, action interaction, trajectory and so on. Therefore, we propose to apply the adversarial training mechanism, for realistic and diverse generation of video representations. In this work, we define our idea as a simple causal graph $\boldsymbol{O} \rightarrow \boldsymbol{X} \leftarrow \boldsymbol{S}$. Based on the observation [30]: *In the GAN training framework, generator neural network connections can be arranged to reflect the causal graph structure*. The generative model can be denoted as $\boldsymbol{O} = f_O(E_O)$, $\boldsymbol{S} = f_S(E_S)$ and $\boldsymbol{X} = f_Z(\boldsymbol{O}, \boldsymbol{S}, E_Z)$, where f_O, f_S, f_Z represent the corresponding generative methods, and E_O, E_S, E_Z are independent variables. Therefore, the essential parts of the DeCalGAN are divided into two components: compositional generation and deconfounding module.

Compositional Generation. For seen compositional learning in the training dataset, we apply the Elaborative Description [12] on each action label y_{ed} for semantic information extraction by a language model as $s = \mathcal{F}_{sem}(y_{ed})$ and object detection for obtaining Top-k object information from the action video x as $o = \mathcal{F}_{obj}(x)$. Following WGAN with gradient penalty [33], the adversarial objective function of generated video feature $h_{\hat{x}} = G(s, o, \boldsymbol{N}_x)$ can be defined as:

$$\begin{aligned} \mathcal{L}_{adv}^{g} &= -\mathbb{E}_{\hat{x}\sim p_{\hat{X}}}[D(h_{\hat{x}})], \\ \mathcal{L}_{adv}^{d} &= -\mathbb{E}_{x\sim p_X}[D(h_x)] + \mathbb{E}_{\hat{x}\sim p_{\hat{X}}}[D(h_{\hat{x}})] \\ &\quad + \lambda_w \, \mathbb{E}_{\bar{x}\sim p_{\bar{X}}}[(\|D(h_{\bar{x}})\|_2 - 1)^2] \, , \end{aligned} \tag{1}$$

where $\boldsymbol{N}_x$ denotes the noise sampled from Gaussian distribution $\mathcal{N}(0,1)$, G and D represent the video feature generator and discriminator, respectively. $h_{\hat{x}}$ is sampled along straight lines between real feature h_x and generated feature $h_{\hat{x}}$. In the video feature generator G, the latent object and semantic features h_s and h_o are extracted by two feed-forward networks E_s and E_o. The generative network can then be used to represent the causal models with graph $\boldsymbol{O} \rightarrow \boldsymbol{X} \leftarrow \boldsymbol{S}$. The detailed architecture is presented in Figure 4. As discussed in previous work, object information and semantic embedding can effectively capture spatial and temporal information. To extract these information, we employ Faster R-CNN [34] for object feature extraction and BERT [35] for semantic feature extraction. Both E_s and E_o are implemented as three layers of a multilayer perceptron (MLP) with 512 dimensions. Additionally, we randomly sample noise from a Gaussian distribution $\mathcal{N}(0,1)$ represented as 768 dimensions. The generated video feature $h_{\hat{x}}$ is then discriminated by the video feature extracted by ResNet(2+1)D_18 [36]. To achieve this, we utilize a fully connected layer-based discriminator.

As our model is designed to estimate $p(h_x|s,o)$, this generative model has two representation distribution spaces: semantic space $\Phi_s \in \mathbb{R}^{d_s}$ and object space $\Phi_o \in \mathbb{R}^{d_o}$, which might be confounded to estimate the video distribution. Therefore, the above structure only constructs the causal correlation, *i.e.,* "conditioning on" operation, but does not solve the confounder problem. Conventionally defined confounder only considered statistical implications, and the actual causal structure is not considered, while confounder is a concept related to real causal structure. To this end, we propose a deconfounding module that overrides the joint distribution to enforce s and o to specific values and propagate them through the causal graph.

Deconfounding Module. With deconfounding, the intervention changes the joint distribution of nodes in the proposed causal graph $\mathcal{G}$. Inspired by [13], we then reconstruct the latent semantic and object features as $h_{\hat{s}}$ and $h_{\hat{o}}$ from the generated video representation $h_{\hat{x}}$ by two feed-forward networks $E_{\hat{s}}$ and $E_{\hat{o}}$ as $h_{\hat{s}} = E_{\hat{s}}(h_{\hat{x}})$ and $h_{\hat{o}} = E_{\hat{o}}(h_{\hat{x}})$. We expect that the reconstructed features $h_{\hat{s}}$ and $h_{\hat{o}}$ maintain approximately the same independence relations, and belong to the same independence space as the original features. To this end, with video feature generator G, the factors s and o are inferred by minimizing the reconstruction loss $\mathcal{L}_{rec}$ as:

$$\begin{aligned}\mathcal{L}_{rec} = & \left\|h_{\hat{s}} - h_s\right\|^2 + \left\|h_{\hat{o}} - h_o\right\|^2 \\ & + \left\|h_x - G(s,o,\boldsymbol{N}_x)\right\|^2 ,\end{aligned} \tag{2}$$

where h_x denotes the video feature extracted from action recognition networks of the given video.

In this causal graph, Φ_s and Φ_o are parent nodes of video feature h_x. The reconstructed distribution $\Phi_{\hat{s}}$ and $\Phi_{\hat{o}}$ are estimated from h_x and thus are child nodes of h_x, which as shown in Figure 5. Therefore, they do not immediately follow the conditional relations that Φ_s and Φ_o obey. Since semantic and object representations h_s and h_o are latent and unobserved, they may confound true signals in the generative process. Even though the semantic and object representations are not obviously independent of each other in the representation view, we can also make causal inference if we can find the factors that jointly affect the semantic and object representations and exclude it by some method.

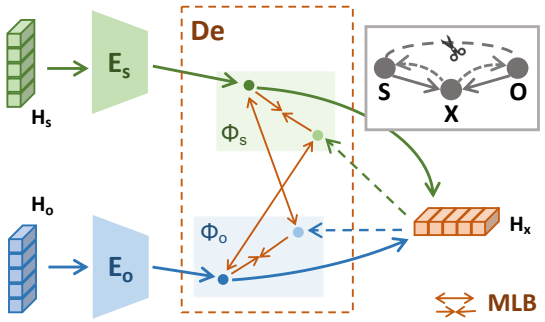


Fig. 5: The causal graph and illustration of deconfounding procedure. The top-right graph represents our designed causal graph and the scissors denote the decondounding operation. Green and blue distributions represent semantic space Φ_s and object space Φ_o, respectively.

To address the challenge of confounding factors between semantic and object representations, we propose a deconfounding module, denoted as De, which compares the original semantic and object feature distribution with the reconstructed distribution. To measure the distance between these two distributions, we employ Multimodel Low-rank Bilinear pooling (MLB) [37] as the distribution score, which has been shown to be effective in multi-modality tasks. The MLB score with given two features (h_1 and h_2) is defined as follows:

$$d(h_1,h_2) = \sigma(\tanh(U^T h_1) \odot \tanh(V^T h_2)), \tag{3}$$

where σ is a linear function, producing values as a one-dimensional score, $\odot$ is the Hadamard product, and both U and V are learnable parameters. Unlike concatenating features, we compare the object and semantic modalities and learn to adaptively weigh them. Our objective is to minimize the distance between similar distributions and increase the distance between dissimilar ones. We achieve this by defining the loss function $\mathcal{L}_{de}$ of the deconfounding module as follows:

$$\begin{aligned}\mathcal{L}^o_{de} &= -d(h_o,h_{\hat{o}}) + d(h_o,h_{\hat{s}}), \\ \mathcal{L}^s_{de} &= -d(h_s,h_{\hat{s}}) + d(h_s,h_{\hat{o}}), \\ \mathcal{L}_{de} &= \mathcal{L}^o_{de} + \mathcal{L}^s_{de},\end{aligned} \tag{4}$$

where minimizing the loss can make the deconfounding module obtain a higher score of the same distribution and a lower score of the different distributions. Thus the following property of causality is encouraged as:

$$\begin{aligned}p^{do(S=s)}(h_{\hat{s}}) &\approx p^{do(S=s,O=o)}(h_{\hat{s}}), \\ p^{do(O=o)}(h_{\hat{o}}) &\approx p^{do(S=s,O=o)}(h_{\hat{o}}).\end{aligned} \tag{5}$$

As confounding factors between training semantic and object sources are latent and unobserved, we assume the notion that the distance between two different distribution spaces can quantify the degree of confounding. We suggest that the closer the distance between the semantic space Φ_s and

the object space Φ_o, the higher the correlation probability between the distributions, indicating a higher probability of confounding. Thus, we model the relationship between Φ_s and Φ_o as follows:

$$\begin{aligned} \Phi_s \perp\!\!\!\perp \Phi_o \mid S = s, \\ \Phi_s \perp\!\!\!\perp \Phi_o \mid O = o. \end{aligned} \tag{6}$$

To this end, the designed deconfounding module is expected to have the ability to maximize the distance, and the experimental results help validate our hypothesis.

Zero-shot Recognition. For the unseen compositional generation, we utilize the above causal inference for generating unseen action representations for zero-shot recognition. According to the proposed deconfounding assumption, we can then apply the "Do-intervention" that overrides the joint distribution to enforce s, o to specific values and propagate them through the causal graph. With this propagation, an intervention can change the joint distribution of nodes in the causal graph, and thus an unseen action representation is generated according to a new joint distribution.

In the action recognition task, we observe that objects in action recognition tasks are similar, *e.g.,* human. Meanwhile, we cannot obtain the object details from the target dataset in the CD-ZSAR setting, and the label distributions between training and target datasets are different as they have "non-overlapping classed". Thus, to achieve the distribution shift, we randomly iterate over all combinations of object variables from training data and ED of test actions to generate the unseen action video features, as follows:

$$h_{\tilde{x}} = G(\tilde{s}, \tilde{o}, \boldsymbol{N}), \tag{7}$$

where $h_{\tilde{x}}$ denotes the generated unseen video features. $\tilde{s}$ and $\tilde{o}$ represent the ED of a test action and a random object variable, respectively. With the proposed deconfounding module, we consider the generated video feature belongs to the given test action class. Finally, after obtaining the unseen video features, we utilize the generated video representation $h_{\tilde{x}}$ and a classifier network R to obtain the test class embedding.

C. Causal Training Strategy

In this work, we aim to train an effective classifier R that has the ability to classify unseen test action classes. The overall training and inference of CD-ZSAR can be described by two pipelines:

Basic Pipeline. Following the work of [22], we use nearest-neighbor search in the semantic class embedding space to obtain zero-shot classification. Given a training set $D = (x_1, y_1), (x_2, y_2) \ldots (x_N, y_N)$ consisting of pairs of video x and its class label y, zero-shot learning classifiers need to generalize to unseen test classes, and we apply the common way [22] to achieve this that uses the nearest-neighbor search in a semantic class embedding space. To do this, we first apply a backbone action recognition network for extracting the video feature h_x, and then use the classifier R to infer the corresponding semantic embedding. The final recognition model $M(\cdot)$ classifies x as the nearest neighbor in the set of embeddings of the classes:

$$M(x) = \arg\min cos(R(h_x), \mathcal{F}_{W2V}(y)), \tag{8}$$

where cos is the cosine distance and the semantic embedding is computed using the Word2Vec function $\mathcal{F}_{W2V}$. Given a video-class pair (x, y) from training dataset, the classifier R is optimized by minimizing the classifier loss $\mathcal{L}_{cls}$:

$$\mathcal{L}_{cls} = \sum \|\mathcal{F}_{W2V}(y) - R(h_x)\|^2. \tag{9}$$

where $\mathcal{L}_{cls}$ denotes the overall loss and h_x is generated by the proposed DeCalGAN.

Causal Inference. To improve the feature representation capability of unseen action distributions, we extend the basic pipeline by incorporating DeCalGAN. In the proposed DeCalGAN, semantic and object features h_s and h_o are extracted using E_h and E_o, from the video-class pair (x, y) in the training dataset respectively. As we utilize adversarial training, the generator G network is used to obtain the generated video feature $\hat{x}$ and discriminator D network to discriminate real and fake video features. Therefore, by incorporating the deconfounding operation with the proposed deconfounding loss and reconstruction loss, the overall generator training loss is:

$$\mathcal{L}_{train} = \lambda_1 \mathcal{L}_{rec} + \lambda_2 \mathcal{L}_{de} + \lambda_3 \mathcal{L}_{adv}, \tag{10}$$

where λ_1, λ_2, and λ_3 are hyperparameters that control weights of each loss. Recalling the bias problem in ZSL with generative models, the synthesized unseen samples could be unexpectedly too close to the real seen ones. This would significantly decrease the classification performance for unseen classes. Thus, we infer the unseen video features according to the unseen semantic feature $h_{\tilde{s}}$ extracted by ED of unseen action labels and randomly selected object features h_o from the seen dataset after training the causal generator. Note that ED is only applied for extracting semantic information.

IV. Experiments

In this section, we present our experimental results on two public datasets: UCF101 [38] and HMDB51 [39]. We compare our approach with other state-of-the-art methods and an in-depth ablation analysis is provided to better understand our method. We also discuss the limitations and potential future work in this task.

A. Experimental Setup

Datasets. We employ Kinetics-700 [40] as the source dataset for compositional generation and basic pipeline training, which is the most widely adopted benchmark, covering a wide range of human activities. Kinetics 700 is released in 2019, which has 700 classes with over 500K videos sourced from YouTube. For target datasets to test the zero-shot classifier, there are two commonly used public datasets in zero-shot action recognition: 1) UCF101 is composed of real action videos focused on sports from YouTube, containing 13320 video clips distributed among 101 classes; and, 2) HMDB51 contains 6766 videos divided into 51 human action categories focused around sports and daily activities from commercial videos.

Training Protocol. We first make sure that D_{train} and D_{test} have "non-overlapping classes". The simple solution which just removes the same class names does not work, because

Dataset	Class Name	Elaborative Description (ED)
Kinetics 700	Counting money	Determine the total number of (a collection of items). A current medium of exchange in the form of coins and banknotes; coins and banknotes collectively.
	Eating chips	Put (food) into the mouth and chew and swallow it. A long rectangular piece of deep-fried potato.
	Moon walking	A dance with a gliding motion, in which the dancer appears to be moving forward but in fact is moving backwards.
UCF101	Clean and jerk	A two-movement weightlifting exercise in which a weight is raised above the head following an initial lift to shoulder level.
	Long Jump	An athletic event in which competitors jump as far as possible along the ground in one leap.
HMDB51	Draw sword	Extract (an object) from a container or receptacle. A weapon with a long metal blade and a hilt with a hand guard, used for thrusting or striking and now typically worn as part of ceremonial dress.
	Climb Stairs	Go or come up a staircase. A set of steps leading from one floor of a building to another, typically inside the building.

TABLE I: Examples of Elaborative Descriptions (ED) for action classes in Kinetics-700, UCF101, and HMDB51 datasets. The resource of Elaborate Description is collected from Wikipedia, Dictionary, and Modification [12].

Fig. 6: An example of using object region features as inputs for the "play basketball" class in the Kinetics-700 dataset.

two classes with slightly different names can easily refer to the same concept. Thus, a distance between class names is needed. Equipped with such a metric, we can make sure training and test classes are not too similar. Following the work of [22], we apply the cosine distance as our non-overlapping metric and the distance is defined as:

$$min\ d(y_{train}, y_{test}) > \tau, \tag{11}$$

$$d(y_{train}, y_{test}) = cos(\mathcal{F}_{W2V}(y_{train}), \mathcal{F}_{W2V}(y_{test})), \tag{12}$$

where $\tau \in \mathbb{R}$ denotes a similarity threshold and cos indicates cosine distance. This is consistent with the use of cosine distance in the zero-shot learning setting as we do in Eq. 8. In order for training and test datasets to contain disjoint video sources, we remove classes from Kinetics-700 whose cosine distance to any class in UCF101 and HMDB51 in the word embedding space is smaller than 0.05. These results in a subset of 663 classes as the training set to train our models.

Evaluation Protocol. We test our framework using two evaluation protocols. The first one is compatible with previous work and the second one emulates a true ZSL setting. Both evaluation protocols apply the same model to both UCF101 and HMDB51 datasets. 1) To be fair with previous work, we randomly choose half of the classes of test datasets for evaluation, which are 50 for UCF101 and 25 for HMDB51, and average the results for each test dataset after repeating ten times. 2) Top-1 and Top-5 accuracies (%) are used to evaluate the classifier on all 101 UCF classes and 51 HMDB classes, which is more restrictive than the evaluation protocol of the previous methods [8], [41].

B. Implementation Details

In our experiments, we first utilize R(2+1)D_18 pretrained on Kinetics-400 [42] as our base model. Then we use the classifier R to infer the corresponding semantic embedding of dimension 16×300, where 16 denotes the batch size. Each frame's shortest side is reshaped to 128 pixels, and we crop a random 112×112 patch during training and the center patch during inference. The video clips are 16 frames long and we choose them following the standard protocol established by Wang *et al.* [14]. The feature size of all MLP blocks is 512 and the classifier R is a linear regression model with 512×300 nodes. According to the standard protocol [22], we average multi-word class names by Word2Vec (Python implementation in gensim [43]) into dimension 300. To minimize all losses, we applied the Adam optimizer with ascent learning rates from 1×10^{-3} to 1×10^{-4} for the classifier, and 1×10^{-4} to 1×10^{-5} for the generator and discriminator. All experiments are performed on $8 \times$ Nvidia Tesla P100 GPUs.

Object Detection & Word Embedding. Following the work of [6], we apply Faster R-CNN [34], pretrained on the MS-COCO dataset [44], for detection of local objects, which consist of the person class and 79 objects, such as *snowboard, human* and *horse*. We obtain roughly 50 detections for each object per frame, and extract top 8 objects with a controlled experiment to select the best number of Top-k. In Figure 6, we present an example of using object detection models to extract object features. Here, we select person and basketball as object information. We also follow the standard protocol in computing semantic embedding of action names by using a pretrained Word2Vec model. In rare cases of words not available in the pretrained W2V model (for example, 'rubiks' or 'photobombing') we manually change the words following

the work of [22]. The pre-trained model produces a 300-dimensional representation for each word. If an action class name contains multiple words $c = [c^1, \ldots, c^N]$, we averaged the embedding as $c = \sum_{i=1}^{N} \mathcal{F}_{W2V}(c^i) \in \mathbb{R}^{300}$.

Elaborative Description. Examples of Elaborative Descriptions (ED) in Kinetics-700 [42], UCF101 [38] and HMDB51 [39] are shown in Table I. Chen *et al.* [12] collected ED for action classes by firstly automatically crawling candidate sentences to describe action classes from the Internet; then manually selecting or modifying a minimum set of candidate sentences as the EDs. In the first crawling step, they utilized Wikipedia and online dictionaries. In the second cleaning step, they presented candidate sentences and a video exemplar in a webpage to annotators. As the BERT model has demonstrated excellent capability in implicitly encoding commonsense knowledge, we apply BERT representation as our semantic information source. In this work, denote $d = \{w_1, \ldots, w_{Nd}\}$ as the ED for action y, where w_i is the composed work. The goal of the pre-trained BERT model is to extract semantic features $s \in \mathbb{R}^K$ with dimension of K. Denote $s_i \in \mathbb{R}^{768}$ as the hidden state from the last layer of BERT for word w_i. We apply average pooling to obtain a sentence-level information s:

$$s = \frac{1}{N_d} \sum_{i=1}^{N_d} s_i \ . \tag{13}$$

We then use an MLP model as semantic encoder E to translate s into the joint semantic feature space.

C. Comparison with State-of-the-art

We compare our model with both inner-dataset methods and cross-dataset inductive zero-shot learning methods, results as shown in Table II. Inner-dataset methods utilize different training and test class in the same dataset, but cross-dataset methods apply training and test class from different dataset.

Inner-dataset Methods. We can observe from Table II that the performance of our DeCalGAN gains large improvement over inner-dataset methods. Even though our method is applied in the cross dataset setting which is more difficult, the results indicate that the essential features can be more effectively obtained by recent state-of-the-art backbones and a large-scale dataset. Compared with O2A [8], TS-GCN [8] and TARN [41], we can observe that incorporating global semantic information and local object information can perform better in actions, and our DeCalGAN enhanced model can effectively infer key cross-domain information from spatio-temporal features. Note that for class labels, our approach follows the conventional training protocol using word embeddings of class names in the final recognition, which contains less semantic information than ED [12], but the work of [12] apply applies ED for both feature learning and zero-shot recognition. Therefore, even though our method utilizes ED for semantic feature learning, our setting is more challenging compared to ED. Moreover, our training and test datasets are different, which will lead to cross-domian issue, yet our model still can obtain close performance on the UCF101 dataset and highest performance on the HMDB51 dataset.

Method	Video	Class	UCF	HMDB
IAP [45]	FV	A	16.7	-
HAA [46]	FV	A	14.9	-
SJE [47]	FV	W_N	9.9	13.3
MTE [48]	FV	W_N	15.8	19.7
GA [49]	C3D*	W_N	22.7	-
O2A [8]	Obj†	W_N	30.3	15.6
CEWGAN [50]	I3D	A	38.3	-
TS-GCN [8]	Obj†	W_N	34.2	23.2
TARN [41]	C3D*	W_N	19.0	28.9
PS-GNN [51]	Obj	W_N	36.1	29.5
DASZL [52]	TSM	A	48.9	-
ED [12]	(ST+Obj)†	ED	**51.8**	35.3
URL [3]	R200	W_N	42.5	28.9
E2E [22]	R(2+1)_18*	W_N	46.1	33.1
Ours	(R(2+1)_18+Obj)*	W_N	51.4	**36.1**

TABLE II: The average top-1 accuracy (%) of state-of-the-art zero-shot learning methods on the UCF and HMDB benchmarks. We evaluate on half test classes following evaluation protocol (1). Visual: Fisher vector (FV), object (Obj), spatio-temporal feature (ST), ResNet200 feature (R200), R(2+1)D_18 feature (R(2+1)_18), *(trained on video dataset) and †(trained on ImageNet dataset); Class: attribute (A), word embedding of class names (W_N) and elaborative description of class names (ED).

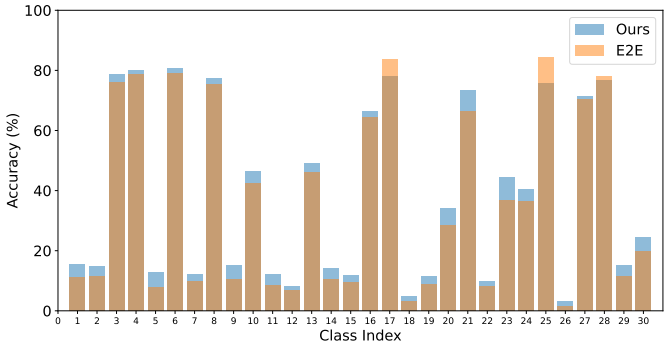


Fig. 7: Per-category performance comparison of DeCalGAN against the baseline E2E on the UCF101 dataset (randomly selected 30 classes).

Cross-dataset Methods. We compare our proposed methods with cross-dataset methods using the same protocol. As shown in Table II, results indicate that E2E and our proposed method outperform the universal-based method URL. This finding suggests that a video-based backbone, such as R(2+1)D, is more effective than an image-based backbone, such as ResNet, in video-level zero-shot learning tasks. We attribute this to the fact that a video-based backbone can more effectively capture motion information compared to an image-based backbone. Our proposed DeCalGAN approach outperforms the E2E method, indicating that a generative model that incorporates both global semantic and local object factors can enable the classifier to learn more action information. Furthermore, the improved performance demonstrates the effectiveness of using causal inference conditioned on joint semantic and object information to generate unseen action representations.

Per-category Improvement Analysis. As shown in Figure 7, the per-category analysis reveals an average improvement of 5.1%. Most of these categories exhibit a wide range of actions and substantial variations, making their improvement highly dependent on semantic reasoning over the global spatial context. For instance, the "Skiing" action is characterized by a long duration and is highly related to object priors and semantic reasoning. Overall, the improvements over the baseline are mainly attributed to the inclusion of both global semantic information and local object information in the causal generation process.

D. Ablation Study

The success of our DeCalGAN can be attributed to both the framework design and technical improvement in each component. To analyze the effect of each component in DeCalGAN, we construct ablation study models including: 1) the basic E2E model without DeCalGAN; 2) "w/o semantic" model without the semantic factor; 3) "w/o object" model without the object factor; 4) "w/o intervention" model without the reconstruction operation and deconfounding loss; 5) "w/o deconfounding" model without deconfounding loss; 6) "Word2Vec" model denotes using Word2Vec embedding in place of BERT model for extracting semantic information of ED; 7) "BERT" model denotes using pretrained BERT model for extracting semantic information of ED; and, 8) "CLIP" model denotes using CLIP pretrained model in place the of BERT model for extracting semantic information of ED. All the ablation studies below are carried out using the second evaluation protocol.

DeCalGAN Type	UCF101		HMDB51	
	Top1	Top5	Top1	Top5
E2E	36.8	61.7	24.1	45.5
w/o semantic	36.9	61.6	24.0	45.7
w/o object	38.0	62.5	25.1	47.2
w/o intervention	38.2	62.8	25.2	47.5
w/o deconfounding loss	38.4	63.2	25.5	47.6
Full model	**39.0**	**64.1**	**27.0**	**50.9**

TABLE III: Ablation study of different module model on UCF101 and HMDB51, following evaluation protocol (2).

Factor Effects. Comparing the results of "w/o semantic" and "w/o object", we can see that the model with semantic information gains better performance, which indicates that the semantic description contains more action information than only applying object embedding. Meanwhile, comparing the results of E2E and "w/o object", there is no obvious change in performance, which indicates that if the generative model only applies object features without any additional side information, it would not help the classifier infer unseen distribution. Comparing the results of the full model and both "w/o semantic" and "w/o object", we observe a notable performance increase on both datasets. We believe that both information is important for unknown actions, and causal inference may not work if a single factor is used.

Deconfounding Effects. Comparing the "full model" with "w/o intervention", we observe that the performance shows a large increase. We think it indicates there exists some latent information in object-semantic joint distribution contains confounders that will confound true signals of generating unseen action representations, and verifies our proposed deconfounding module has the ability of removing some confounders. Moreover, "w/o deconfounding loss" performs better than "w/o intervention", which also proves our proposed deconfounding loss can deconfound confounders in our causal inference setting, whereas reconstruction loss alone cannot remove confounder affect effectively.

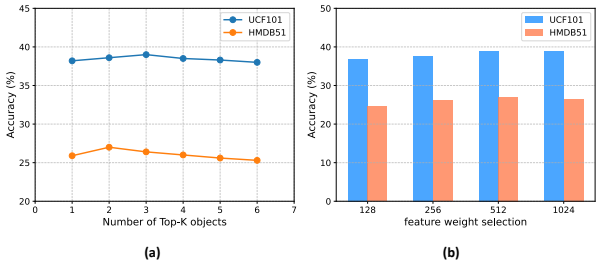


Fig. 8: Comparison results among (a) different number of selected objects (b) different feature weight selection on UCF101 and HMDB51 datasets. Both follow evaluation protocol (2).

Semantic	UCF101		HMDB51	
	Top1	Top5	Top1	Top5
Word2Vec	38.7	64.0	27.1	50.8
BERT	39.0	64.1	27.0	50.9
CLIP	41.2	65.8	28.5	53.1

TABLE IV: Semantic representation selection on UCF101 and HMDB51 datasets, following evaluation protocol (2).

Semantic Representation. We also evaluated different methods for extracting semantic information from Event Descriptions (ED). Recent works on open-vocabulary learning have started using multi-modality pretrained models such as CLIP [53] for the extraction of semantic features. In this ablation analysis, we also apply the pretrained CLIP model as another semantic feature extractor, employing '*a video of* {*category*}' as the text prompt. The results, as shown in Table IV, demonstrate that the selection of the semantic representation methodology does not much influence the overall performance of our proposed framework when using text-only embedding techniques. However, with CLIP, a substantial performance improvement is observed. However, using multi-modality pretrained models could lead to unfair comparisons with other zero-shot learning approaches, since the text embedding integrates visual information into the semantic representation. Therefore, we selected BERT as our semantic representation learning approach, which has shown to be effective in capturing semantic information.

Top-k objects. We conducted experiments to investigate the impact of the number of Top-k objects on the quality of generated unseen action features. Results show that the best performance occurs when k = 3 for the UCF101 dataset and k = 2 for the HMDB51 dataset. Our findings suggest that inferring actions in UCF101 requires more object information

compared to HMDB51. Additionally, our results indicate that including too many object features can make it more challenging for our deconfounding module to remove confounding effects effectively, thus highlighting the importance of selecting an optimal number of objects for optimal performance in compositional zero-shot learning.

Feature Channel Selection. To explore the best representation of semantic and object information, we conduct ablation study of different feature dimensions as shown in Figure 8. We can observe that feature dimension of 512 can achieve the highest performance. It indicates that small channel networks lack in learning semantic and object information effectively, which might lose the important factor information for action feature generation. Meanwhile, large channels might learn more confounding information which is difficult for our proposed deconfounding module to remove latent confounders.

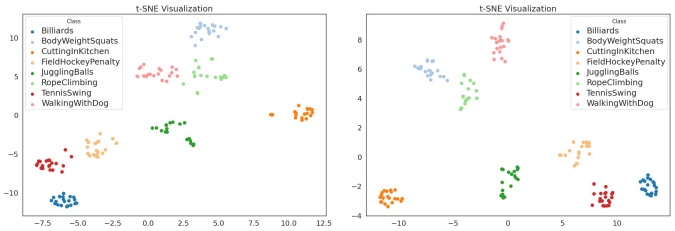
Ratio	UCF101		HMDB51	
	Top1	Top5	Top1	Top5
1:1.0:1.0	37.2	61.9	24.2	46.3
1:0.6:0.6	38.6	63.6	25.7	48.7
1:0.6:0.3	37.5	61.6	25.1	47.2
1:0.3:0.6	**39.0**	**64.1**	**27.0**	**50.9**
1:0.3:0.3	38.1	62.7	25.3	47.5
1:0.1:0.1	38.5	63.2	25.9	48.2

TABLE V: Ablation study of loss weight selection on UCF101 and HMDB51 datasets, following evaluation protocol (2). The different ratio models denotes models utilize different loss weight training.

Loss Weight Selection. When training the causal generator, we try to obtain optimal performance by tuning the training loss weight hyperparameters λ_1, λ_2 and λ_3 according to Eq. 10. According to Table V, we observe that when the ratio is 1:0.3:0.6, the performance is the best. It indicates that the deconfounding module can be more effective when the generator is well trained. Also, when deconfounding loss becomes relatively small, the performance drops, which implies that our deconfounding module is important and necessary.

E. Qualitative Evaluation via t-SNE Visualization

We employ t-SNE visualization to compare the performance of E2E and our proposed DeCalGAN approach. We randomly select 20 samples from eight actions and use the extracted 300-dimensional features to visualize t-SNE, as shown in Figure 9. Our visualization reveals that the distribution using the E2E is sparser compared to DeCalGAN, and all samples are closer to the center when applying our proposed approach. This observation suggests that leveraging both local object information and global semantic descriptions can jointly benefit causal inference, and the deconfounding module can enable confounding features learned from the source domain to better generalize the unseen distribution in the test domain. However, the basic video-based backbone struggles to distinguish complex actions, such as "rope climbing", which may mislead our proposed causal generator. In future work, we will explore methods to enhance the video representations to better capture the complexity of human actions.



(a) E2E **(b) Ours**

Fig. 9: t-SNE visualization of video representation extracted by E2E baseline and our DeCalGAN in eight video actions from the UCF101 dataset.

V. Conclusion

This paper proposed a DeCalGAN model to address the problems of weak knowledge representation and unseen distribution in CD-ZSAR. Class word embedding is enhanced by local object information that unifies the compositional and traditional visual-semantic frameworks. A deconfounding module is proposed to refine global semantic and local object features by reconstruction and deconfounding constraints. The proposed method is able to transfer the large-scale pretrained model on Kinectcs-700 to two ZSAR benchmark datasets, UCF101 and HMDB51. Extensive results validate that DeCalGAN can successfully infer novel samples with unseen distributions in new datasets.

References

[1] J. Liu, B. Kuipers, and S. Savarese, "Recognizing human actions by attributes," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2011, pp. 3337–3344.

[2] Z. Zhang, C. Wang, B. Xiao, W. Zhou, and S. Liu, "Robust relative attributes for human action recognition," *Pattern Analysis and Applications*, vol. 18, no. 1, pp. 157–171, 2015.

[3] Y. Zhu, Y. Long, Y. Guan, S. Newsam, and L. Shao, "Towards universal representation for unseen action recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 9436–9445.

[4] Z. Akata, F. Perronnin, Z. Harchaoui, and C. Schmid, "Label-embedding for attribute-based classification," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2013, pp. 819–826.

[5] B. Romera-Paredes and P. Torr, "An embarrassingly simple approach to zero-shot learning," in *International conference on machine learning*. PMLR, 2015, pp. 2152–2161.

[6] P. Mettes and C. G. Snoek, "Spatial-aware object embeddings for zero-shot localization and classification of actions," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 4443–4452.

[7] P. Mettes, W. Thong, and C. G. Snoek, "Object priors for classifying and localizing unseen actions," *International Journal of Computer Vision*, vol. 129, no. 6, pp. 1954–1971, 2021.

[8] J. Gao, T. Zhang, and C. Xu, "I know the relationships: Zero-shot action recognition via two-stream graph convolutional networks and knowledge graphs," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 33, no. 01, 2019, pp. 8303–8311.

[9] Y. Long, L. Liu, L. Shao, F. Shen, G. Ding, and J. Han, "From zero-shot learning to conventional supervised classification: Unseen visual data synthesis," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 1627–1636.

[10] Y. Zhu, M. Elhoseiny, B. Liu, X. Peng, and A. Elgammal, "A generative adversarial approach for zero-shot learning from noisy texts," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 1004–1013.

[11] J. Li, M. Jing, K. Lu, Z. Ding, L. Zhu, and Z. Huang, "Leveraging the invariant side of generative zero-shot learning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 7402–7411.

[12] S. Chen and D. Huang, "Elaborative rehearsal for zero-shot action recognition," in *Proceedings of the IEEE International Conference on Computer Vision*, 2021, pp. 13 638–13 647.

[13] Y. Atzmon, F. Kreuk, U. Shalit, and G. Chechik, "A causal view of compositional zero-shot recognition," in *34th Conference on Neural Information Processing Systems*, 2020.

[14] L. Wang, Y. Xiong, Z. Wang, Y. Qiao, D. Lin, X. Tang, and L. Van Gool, "Temporal segment networks: Towards good practices for deep action recognition," in *Proceedings of the European Conference on Computer Vision*. Springer, 2016, pp. 20–36.

[15] J. Lin, C. Gan, and S. Han, "TSM: Temporal shift module for efficient video understanding," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 7083–7093.

[16] Y. Li, B. Ji, X. Shi, J. Zhang, B. Kang, and L. Wang, "TEA: Temporal excitation and aggregation for action recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2020, pp. 909–918.

[17] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3d convolutional networks," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 4489–4497.

[18] J. Carreira and A. Zisserman, "Quo vadis, action recognition? A new model and the kinetics dataset," in *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 6299–6308.

[19] C. Feichtenhofer, H. Fan, J. Malik, and K. He, "Slowfast networks for video recognition," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 6202–6211.

[20] L. Wang, Z. Tong, B. Ji, and G. Wu, "TDN: Temporal difference networks for efficient action recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2021, pp. 1895–1904.

[21] K. Hara, H. Kataoka, and Y. Satoh, "Can spatiotemporal 3D CNNs retrace the history of 2D CNNs and ImageNet?" in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 6546–6555.

[22] B. Brattoli, J. Tighe, F. Zhdanov, P. Perona, and K. Chalupka, "Rethinking zero-shot video classification: End-to-end training for realistic applications," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2020, pp. 4613–4623.

[23] X. Xu, T. Hospedales, and S. Gong, "Transductive zero-shot action recognition by word-vector embedding," *International Journal of Computer Vision*, vol. 123, no. 3, pp. 309–333, 2017.

[24] C.-C. Lin, K. Lin, L. Wang, Z. Liu, and L. Li, "Cross-modal representation learning for zero-shot action recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 19 978–19 988.

[25] J. Pearl *et al.*, "Models, reasoning and inference," *Cambridge, UK: CambridgeUniversityPress*, vol. 19, 2000.

[26] D. B. Rubin, "Causal inference using potential outcomes: Design, modeling, decisions," *Journal of the American Statistical Association*, vol. 100, no. 469, pp. 322–331, 2005.

[27] B. Schölkopf, D. Janzing, J. Peters, E. Sgouritsa, K. Zhang, and J. Mooij, "On causal and anticausal learning," in *29th International Conference on Machine Learning (ICML 2012)*. International Machine Learning Society, 2012, pp. 1255–1262.

[28] M. Gong, K. Zhang, T. Liu, D. Tao, C. Glymour, and B. Schölkopf, "Domain adaptation with conditional transferable components," in *International conference on machine learning*. PMLR, 2016, pp. 2839–2848.

[29] Y. Li, X. Tian, M. Gong, Y. Liu, T. Liu, K. Zhang, and D. Tao, "Deep domain generalization via conditional invariant adversarial networks," in *Proceedings of the Proceedings of the European Conference on Computer Vision*, 2018, pp. 624–639.

[30] M. Kocaoglu, C. Snyder, A. G. Dimakis, and S. Vishwanath, "CausalGAN: Learning causal implicit generative models with adversarial training," in *International Conference on Learning Representations*, 2018.

[31] M. F. Naeem, Y. Xian, F. Tombari, and Z. Akata, "Learning graph embeddings for compositional zero-shot learning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2021, pp. 953–962.

[32] X. Yang, F. Feng, W. Ji, M. Wang, and T.-S. Chua, "Deconfounded video moment retrieval with causal intervention," *arXiv preprint arXiv:2106.01534*, 2021.

[33] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. C. Courville, "Improved training of wasserstein gans," *Advances in neural information processing systems*, vol. 30, 2017.

[34] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," *Advances in neural information processing systems*, vol. 28, pp. 91–99, 2015.

[35] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, "Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter," *arXiv preprint arXiv:1910.01108*, 2019.

[36] D. Tran, H. Wang, L. Torresani, J. Ray, Y. LeCun, and M. Paluri, "A closer look at spatiotemporal convolutions for action recognition," in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2018, pp. 6450–6459.

[37] J.-H. Kim, K.-W. On, W. Lim, J. Kim, J.-W. Ha, and B.-T. Zhang, "Hadamard product for low-rank bilinear pooling," *International Conference on Learning Representations*, 2016.

[38] K. Soomro, A. R. Zamir, and M. Shah, "Ucf101: A dataset of 101 human actions classes from videos in the wild," *arXiv preprint arXiv:1212.0402*, 2012.

[39] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, and T. Serre, "HMDB: a large video database for human motion recognition," in *Proceedings of the IEEE International Conference on Computer Vision*. IEEE, 2011, pp. 2556–2563.

[40] J. Carreira, E. Noland, C. Hillier, and A. Zisserman, "A short note on the kinetics-700 human action dataset," *arXiv preprint arXiv:1907.06987*, 2019.

[41] M. Bishay, G. Zoumpourlis, and I. Patras, "TARN: Temporal attentive relation network for few-shot and zero-shot action recognition," *British Machine Vision Conference*, 2019.

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

[43] R. Řehřek, P. Sojka *et al.*, "Gensim—statistical semantics in python," *Retrieved from genism. org*, 2011.

[44] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft COCO: Common objects in context," in *Proceedings of the European Conference on Computer Vision*. Springer, 2014, pp. 740–755.

[45] C. H. Lampert, H. Nickisch, and S. Harmeling, "Learning to detect unseen object classes by between-class attribute transfer," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2009, pp. 951–958.

[46] J. Liu, B. Kuipers, and S. Savarese, "Recognizing human actions by attributes," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition 2011*. IEEE, 2011, pp. 3337–3344.

[47] Z. Akata, S. Reed, D. Walter, H. Lee, and B. Schiele, "Evaluation of output embeddings for fine-grained image classification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 2927–2936.

[48] X. Xu, T. M. Hospedales, and S. Gong, "Multi-task zero-shot action recognition with prioritised data augmentation," in *Proceedings of the European Conference on Computer Vision*. Springer, 2016, pp. 343–359.

[49] A. Mishra, V. K. Verma, M. S. K. Reddy, S. Arulkumar, P. Rai, and A. Mittal, "A generative approach to zero-shot and few-shot action recognition," in *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 2018, pp. 372–380.

[50] D. Mandal, S. Narayan, S. K. Dwivedi, V. Gupta, S. Ahmed, F. S. Khan, and L. Shao, "Out-of-distribution detection for generalized zero-shot action recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 9985–9993.

[51] J. Gao, T. Zhang, and C. Xu, "Learning to model relationships for zero-shot video classification," *IEEE transactions on pattern analysis and machine intelligence*, vol. 43, no. 10, pp. 3476–3491, 2020.

[52] T. S. Kim, J. Jones, M. Peven, Z. Xiao, J. Bai, Y. Zhang, W. Qiu, A. Yuille, and G. D. Hager, "Daszl: Dynamic action signatures for zero-shot learning," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 35, no. 3, 2021, pp. 1817–1826.

[53] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natural language supervision," in *International Conference on Machine Learning*. PMLR, 2021, pp. 8748–8763.