2023

Enhancing Zero-Shot Action Recognition in Videos by Combining GANs with Text and Images

Kaiqiang Huang
Technological University Dublin, Ireland, kaiqiang.huang@tudublin.ie

Luis Miralles-Pechuán
Technological University Dublin, luis.mirlles@tudublin.ie

Susan McKeever
Technological University Dublin, susan.mckeever@tudublin.ie

Funder: This project is funded under the Fiosraigh Scholarship of Technological University Dublin. Open Access funding provided by the IReL Consortium.

SN Computer Science (2023) 4:375
https://doi.org/10.1007/s42979-023-01803-3

ORIGINAL RESEARCH

Enhancing Zero-Shot Action Recognition in Videos by Combining GANs with Text and Images

Kaiqiang Huang[1] · **Luis Miralles-Pechuán**[1] · **Susan Mckeever**[1]

Received: 2 March 2022 / Accepted: 27 March 2023 / Published online: 5 May 2023

Abstract
Zero-shot action recognition (ZSAR) tackles the problem of recognising actions that have not been seen by the model during the training phase. Various techniques have been used to achieve ZSAR in the field of human action recognition (HAR) in videos. Techniques based on generative adversarial networks (GANs) are the most promising in terms of performance. GANs are trained to generate representations of unseen videos conditioned on information related to the unseen classes, such as class label embeddings. In this paper, we present an approach based on combining information from two different GANs, both of which generate a visual representation of unseen classes. Our dual-GAN approach leverages two separate knowledge sources related to the unseen classes: class-label texts and images related to the class label obtained from Google Images. The generated visual embeddings of the unseen classes by the two GANs are merged and used to train a classifier in a supervised-learning fashion for ZSAR classification. Our methodology is based on the idea that using more and richer knowledge sources to generate unseen classes representations will lead to higher downstream accuracy when classifying unseen classes. The experimental results show that our dual-GAN approach outperforms state-of-the-art methods on the two benchmark HAR datasets: HMDB51 and UCF101. Additionally, we present a comprehensive discussion and analysis of the experimental results for both datasets to understand the nuances of each approach at a class level. Finally, we examine the impact of the number of visual embeddings generated by the two GANs on the accuracy of the models.

Keywords Human action recognition · Zero-shot learning · Generative adversarial networks · Semantic knowledge source

Introduction

Human action recognition (HAR) is a discipline within machine learning (ML) that aims to recognise human actions (e.g. jumping, walking, or running) from data samples [1]. The data can be collected in different formats, such as images, videos, or the information collected by sensors installed on the subjects performing those actions [2]. In this paper, we focus on recognising actions from videos. The main approach in the literature for HAR is supervised learning [3–7]. Supervised learning trains models to map each of the videos with an action class (e.g. walking or climbing stairs) resulting in a multi-classification ML task. The results in the literature for identifying an action label, once samples from that label are available, have been quite high, particularly with the application of deep learning network architectures. For example, with the popular Weizmann dataset, 95% accuracy is achieved [8].

In recent years, a more challenging problem called zero-shot action recognition (ZSAR), which aims to recognise actions in videos that are not part of model training has been explored [9–11]. The classes that are not part of the training set, but are predicted and evaluated during the testing phase of the model, are called *unseen classes*. Collecting and labelling samples of the different classes is a laborious and expensive task. If the action is hard to acquire on video, it can be difficult to collect the required volume of

This article is part of the topical collection "Computational Intelligence 2021" guest edited by Juan Julian Merelo, Kevin Warwick, Thomas Bäck, Christian Wagner, Jonathan Garibaldi, H. K. Lam and Marie Cottrell.

✉ Kaiqiang Huang
kaiqiang.huang@tudublin.ie

Luis Miralles-Pechuán
luis.miralles@tudublin.ie

Susan Mckeever
susan.mckeever@tudublin.ie

[1] School of Computing, Technological University Dublin, Central Quad, Grangegorman, Dublin, Ireland

training videos in the first place. ZSAR aims to solve that burden by enabling models to classify video instances of classes that were not used during the training phase [12]. In videos, the HAR of unseen classes can be achieved by transferring semantic knowledge from the seen classes to the unseen ones. According to the literature [12], in 2022, there were more than thirty different approaches for applying ZSAR in videos.

In the early research stage of solving the ZSAR problem, most researchers focussed on projection-based approaches [9, 11, 13–15]. These approaches build functions that map the visual embedding of a given class to its corresponding class semantic embedding (e.g. word2vec representation of the class label). This is first done for all the seen classes and then, the function is used to estimate the visual embedding of the unseen classes. For example, a projection function can be used to map visual features of different classes to the word embedding of the class label. A class label represents any activity a human can perform (e.g. rowing, cycling, or sitting). The learned projection function is then applied to recognise unseen classes using a similarity-based metric that calculates the difference between the ground-truth embeddings and the predicted embeddings for the unseen videos. However, the distributions of visual representations for seen and unseen classes can be different, resulting in a large variational mismatch at the classification stage for unseen classes since the classifier is only trained with seen classes. To mitigate this problem, recent ZSAR approaches have used generative adversarial networks (GANs) [16–19]. GANs can alleviate the discrepancy between the seen and the synthesised unseen data. GANs do so by synthesising visual embeddings of unseen classes using semantic embeddings for conditioning the GAN. Later on, a downstream classifier is trained with the seen and synthesised unseen class representations in a fully supervised manner to make predictions with the testing videos. When applying the ZSAR approach, only unseen classes are tested. As opposed to generalised ZSAR in which seen and unseen classes are tested.

Machine learning approaches to ZSAR struggle under certain scenarios, such as data imbalance, lack of common semantic features, big domain shift, poor knowledge graph coverage, low-quality synthesised embeddings, and too much dependency on semantic embeddings. ZSAR struggles to produce accurate results if the semantic embeddings or the knowledge graph used to describe the seen and unseen classes is incomplete or not representative, or if there is a large domain shift between seen and unseen classes [20, 21]. The quality of the synthesised visual embeddings by GANs is a crucial point for improving the ZSAR performance. This quality is based on how representative and discriminative the generated semantic embeddings for unseen classes are for the classifier. Intuitively, semantic class embeddings with more representative and richer semantic information will produce higher-quality synthesised visual embeddings from the GANs.

Semantic embeddings are vector spaces that represent words, phrases, images, or concepts. Word embeddings are the most common way to generate semantic embeddings in zero-shot learning. Word embeddings use pre-trained models, such as word2vec, GloVe and fastText, to generate semantic embeddings for class labels by averaging the embeddings of the words presented in the label. Generating semantic embeddings involves training a neural network with a large amount of text data. Once trained, neural networks are able to generate, from a class label, a mathematical representation in the form of a dense vector of real numbers known as a feature space. This feature space encodes the semantic content of a word or a phrase. To make sure the correct meaning is captured, we need quality data, preprocessing the data using the right techniques (stemming, removing stop worlds, etc.), using the right architecture for the neural networks, and evaluating the models using the right metrics such as the similarity between the words [22]. Ultimately, the correctness of the semantic embedding meaning as related to the unseen class will manifest in the accuracy of the ZSAR approach.

In our previous work, we investigated how to improve the ZSAR performance by adding visual objects (i.e. represented as text) from the unseen and seen classes as inputs to the GAN [18]. This approach yielded higher ZSAR results than simply using the class label information, suggesting that using additional knowledge sources can lead to a higher quality generation of visual embeddings for unseen classes. The main idea of the paper [18] is to add more semantic context to the GANs by averaging, appending and replacing the name of the class with words that are related to it. On the other hand, the methodology proposed in this paper is about merging information from two GANs (one using visual information and the other textual information) to see if it is possible to boost the performance of the classifier. In this piece of research, we propose a multi-source approach based on GANs under the VAEGAN model [17]. Our underlying assumption is that injecting more knowledge into the semantic embedding(s) used to condition the generation of unseen class visual embeddings from the GAN will result in better downstream ZSAR accuracies. We use semantic embeddings from multiple knowledge sources (text and image) to condition the GAN generation of visual embeddings for unseen classes. We evaluated our approach against two popular benchmark datasets in the HAR field: HMDB51 and UCF101.

Our experiments address two research questions—(1) Are images of the unseen class a richer source of knowledge than text-based sources (class label or text description for conditioning information for the GAN approach)? More specifically, can we achieve higher ZSAR accuracies

using semantic embeddings based on *images* than on *text-based* semantic embeddings of unseen classes to condition the GAN? GANs have previously been applied to achieve the ZSAR [16, 17]. These approaches use basic semantic embedding (i.e. class label word embedding) rather than richer knowledge sources (semantic embeddings based on full-text descriptions or images of the class). (2) Is the combination of image and text sources as conditioning knowledge for the GAN more suitable for ZSAR than using one source alone? Specifically, does our proposed dual-GAN approach of incorporating two knowledge sources for generating semantic embeddings achieve higher ZSAR accuracies than approaches based on a single-source (text or image) single-GAN?

The challenge in zero-shot action recognition is to predict classes that have not been seen in the training phase. This challenge is tackled by utilising synthesised visual embeddings (in the form of a high-dimensional feature space) to represent unseen classes. This involves searching images related to the class label using a search engine. Then, we use trained image ML models (e.g. GoogLeNet and ResNet101) to generate representations for these images. Lastly, we use GANs to learn the relationships between the output of the ImageNet model and the representation vector of the videos. Once the GAN is trained, we can input the semantic embedding of unseen classes, synthesised as explained, to the trained GANs to generate new video representation vectors for the unseen classes. This gives us semantic and visual (synthesised) information about unseen classes. Thus, we convert a zero-shot learning problem into a solvable supervised learning problem. A similar approach has been previously implemented by other researchers [17, 23].

One of the advantages of machine learning is that model can be built from datasets without human intervention, with better performance and cost savings. However, sometimes the task of curating datasets is very time-consuming and expensive, or there may be a lack of suitable data to represent the entire domain space. ZSAR is about recognising new actions or classes based on describing aspects of the classes rather than actual samples, so the tedious and/or difficult task of procuring and manually labelling data can be avoided. Since ZSAR uses information about unseen classes, we thought it is likely that using a model with information from two sources, that is to say, images and text will give more clues to the classifier to recognise an unseen class. We first used textual VAEGANs models which are combinations of variational autoencoder with generative adversarial networks to translate the name of the class of the video into a description. VAEGANs are able to create instances representing unseen classes that can be used to train a classifier. Then, we used an image VAEGAN model for embedding the images that appear in the videos. In the state of the art, there were approaches with good performance using image VAEGANs and other models with good performance using textual VAEGANs, but to our knowledge, there were no authors combining embeddings of both images and text which was a good opportunity for testing this new idea which actually gave a better performance than the other approaches as shown in "Results and analysis".

Our contributions can be summarised as follows. First, we investigate two different knowledge sources (i.e. texts and images) that can be used to represent semantics for action classes. Second, we propose a dual-GAN approach based on the VAEGAN model to generate high-quality visual embeddings for unseen classes and then we combine generated visual embeddings derived from two knowledge sources (i.e. texts and images). The combination methods are synthesised visual embeddings from both text-based and image-based knowledge sources using the average, the summation, the maximum and the minimum. And third, we show that our dual-GAN model outperforms other existing new ZSAR GAN-based approaches. To the best of our knowledge, there are no previous publications that consider combining synthesised unseen visual embeddings derived from two different knowledge sources in the context of the GAN-based frameworks for ZSAR in videos.

The rest of this paper is structured as follows—"Related work" provides a literature review of various approaches for the ZSAR. "Approach" introduces our proposed dual-GAN approach based on the VAEGAN model using multiple knowledge sources for ZSAR. "Methodology" describes the methodology, which includes the process of collecting images and feature combination methods. "Experiments" explains the experimental configurations and implementations in more detail. "Results and analysis" shows the results and key findings. Finally, "Conclusions" concludes the paper and proposes a few ideas for future work.

Related Work

This section reviews the related literature on the main approaches to ZSAR research. Given the role of GANs in our approach, We then examine the most important generative approaches based on GANs. We also summarise the existing works that propose different types of semantic embedding, especially those related to the GAN-based framework.

In the early stages of research on ZSAR, many publications [11, 14, 24–26] proposed projection functions to map the visual representations of the video instances to the semantic representation of the class prototype of that specific video (i.e. typically an embedding space of a class label). These learned projection functions encode the relationship between the visual embeddings and the semantic embeddings using seen data. The learned projection function is then used to recognise unseen classes by measuring

the likelihood between the ground truth and the predicted semantic representations of the video instances in the embedding space. However, classes with similar semantic knowledge may have large variations in the visual space. For example, the classes *Diving* and *Swimming* have very similar descriptions since they both are outdoor activities and include water, but their video samples look very different since *Diving* and *Swimming* have quite different body movements. Therefore, building a high-accuracy projection function is a big challenge, which may cause ambiguity in the visual–semantic mapping due to the large variation in the visual embedding.

Recently, advanced generative-based methods have been used to synthesise visual embeddings of unseen classes, conditioned on semantic embeddings of class information. Some authors [27] proposed a conditional Wasserstein GAN (WGAN) model using the classification loss to synthesise visual embeddings of unseen classes in the image domain. The visual embeddings of the unseen classes are then synthesised using a trained conditional WGAN and used together with the real visual embeddings of the seen classes to train a discriminative classifier in a fully supervised manner.

There are several relevant publications [16, 17, 28–30] that also apply cycle-consistency constraints on the reconstruction of the semantic embeddings from the generated visual embeddings during the training phase. The process of reconstruction is a verification that the generated visual embeddings can be correctly converted back to their corresponding semantic embeddings. The cycle-consistency constraints help to produce a higher quality generator for the synthesis of semantically consistent visual embeddings of the unseen classes. Although these generative-based methods show promising results for the ZSAR task, those methods still struggle to generate discriminative visual embeddings of unseen classes since the performance of seen classes is better than that of unseen ones in the generalised setting.

One of the most widely used semantic embeddings due to its convenience and effectiveness is the word embedding of the action label, typically using a Word2Vec representation [31]. However, Word2Vec is not particularly good at reducing the semantic gap between the visual and the semantic embeddings since the information in the class labels is not discriminative, in the sense that the GANs can find it difficult to be trained with the conditions that are the semantic embeddings from labels.

To alleviate the problem of the semantic gap between the visual and the semantic embeddings, the authors of the paper [32] enhanced the word vectors of the label by collecting and modelling textual descriptions of the action classes. The contextual information (e.g. textual descriptions related to action classes) can remove the ambiguity of the semantics to some extent in the original word vectors of action labels. For example, the class *Haircut* can be described as 'A hairstyle, hairdo, or haircut refers to the styling of hair, usually on the human scalp', as a noun, in terms of parts of speech. *Haircut* also refers to the act of reducing facial or body hair. In that same work [32], the authors proposed a method to collect images related to the action labels for representing visually discriminative semantic embedding. However, the work only evaluated the proposed semantic embeddings in a projection-based approach, not as input for GAN-generated unseen class representations. Other authors [33] also used enriched semantic knowledge for better class-level semantic representations. In particular, they proposed a description text dataset whose definition was taken from the official Wikipedia website for the UCF101 dataset and evaluated it using the GAN-based model.

To summarise, there are a few leading approaches to achieve the problem of ZSAR, to be compared to our proposed dual-GAN model in the experiments, described as follows:

- Gaussian mixture model (GMM): The authors propose a generative model that can synthesise new action instances from a few representative examples by incorporating a GMM to expect the distribution of each action class, and then use these synthetic instances to train a recognition model for achieving the zero-shot action recognition. Note that this approach is not based on GANs [34].
- Classification-loss Wasserstein generative adversarial networks (CLSWGAN): To solve the zero-shot learning, the authors propose a new architecture that contains a WGAN pairing with a classification loss to synthesise visual features conditioned on class-level semantic information [27].
- Out-of-distribution detection: In this approach [16], the authors propose a method for detecting out-of-distribution instances in generalised zero-shot action recognition. This method can accurately recognise if a given video belongs to seen or unseen classes and then apply the corresponding classifiers for recognition. To train the out-of-distribution detector, visual features of unseen action classes are synthesised using GANs trained on seen action classes.
- Feature variational autoencoders and GAN (f-VAEGAN): The authors propose a conditional GAN that combines the strength of variational autoencoder (VAE) and GANs in a unified feature generation framework [23]. The focus of this work is to generate visual features of any class using labelled instances when they are available and generalising them to unknown concepts whose labelled instances are not available.

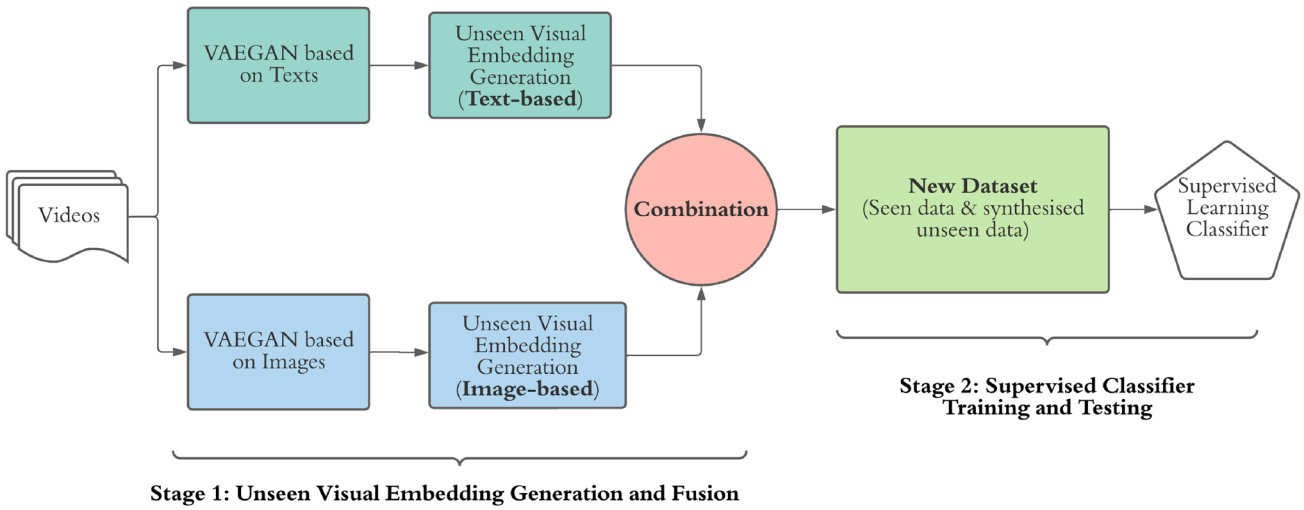

Fig. 1 High-level perspective of the pipeline for the proposed dual-GAN approach based on the VAEGAN model. The semantic embedding derived either from the text source or from the image source is input to the appropriate VAEGAN model, resulting in generating two types of visual embeddings for unseen classes. After that, the generated unseen data is fused to form a new dataset that includes real-seen data and synthesised unseen data to train a supervised learning classifier

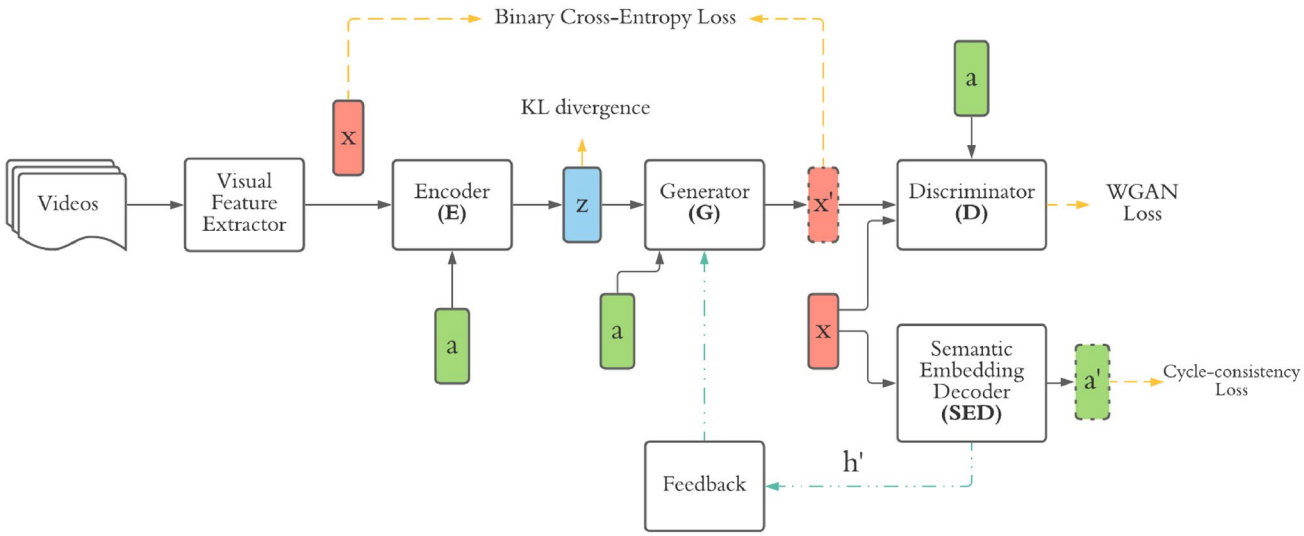

Fig. 2 The details of one VAEGAN component [18], *X* denote the visual embeddings of seen classes that are produced by the feature extractor. *a* denotes the semantic embedding for action classes. *X'* denotes the synthesised visual embeddings of unseen classes from a trained generator (G). *a'* is produced by the component of the semantic embedding decoder (SED) with the input of *X*, which is the process of semantic embedding reconstruction

- Latent embedding feedback and discriminative features: In this investigation [17], the authors propose to enforce semantic consistency for zero-shot action recognition. They introduce a feedback loop, from a semantic embedding decoder, that iteratively refines the generated features during both the training and feature synthesis steps. The synthesised features along with their corresponding latent embeddings from the decoder are then transformed into discriminative features and utilised during classification to reduce ambiguities amongst categories.

Approach

This section explains our dual-GAN approach for zero-shot action recognition and how semantic embeddings from two knowledge sources (text and images) can be fused, as shown

in Fig. 1. Note that, image-based semantic embedding refers to the embeddings extracted from images according to the actions, and is used for conditioning for training GANs.

As shown in Fig. 1, the high-level perspective of the pipeline for the proposed dual-GAN approach contains two steps—*Step 1* aims to synthesise the visual embeddings of unseen classes conditioned on the corresponding semantic embeddings obtained from two different knowledge sources: text-VAEGAN for texts and image-VAEGAN for images. After that, both the image-driven and the text-driven unseen visual embeddings are combined by a combination operation (averaging, summation, minimum, or maximum) to form a new dataset that contains the original visual representations of seen classes and the synthesised visual representations of unseen classes along with their respective labels. *Step 2* focuses on training a classifier in a supervised learning fashion with the new dataset generated in the previous step. Algorithm 1 provides a high-level description of the proposed methodology. It is noted that the generator of each VAEGAN component is only trained with seen data (i.e. video instances and labels). Each VAEGAN component is able to synthesise semantically visual embeddings conditioned on a semantic embedding without having access to any video instances of the unseen classes.

Algorithm 1 Steps for ZSAR using Dual-GANs

1: Splitting the dataset between the training and testing set (removing the unseen classes from the training set and using them on the testing set).
2: Generate visual embeddings of unseen classes using GANs that are trained with text-based semantic sources (e.g. labels or descriptions).
3: Generate visual embeddings of unseen classes using GANs that are trained with image-based semantic sources.
4: Combine the generated unseen visual embeddings from two GANs using the average, summation, minimum, and maximum.
5: Train a classifier using the seen data and the synthesised unseen data.
6: Evaluate the model's performance on the unseen classes using the ZSAR standard metrics.

Combining information from two different GANs can be done in two ways. The first is to create two separate models and combine the output as it is done in ensemble models. However, this methodology requires more computation time since two models need to be trained and tested. The process of ZSAR is computationally very expensive and adding even more complexity did not look like a good option. The second option was to combine the two arrays representing the new instances using different operators, such as the average, the summation and the maximum and the minimum between these two vectors.

To expand the high-level pipeline described in Fig. 2, we implemented the VAEGAN component with a similar structure to the work proposed in [17]. To keep this paper self-contained, we describe in more detail the VAEGAN component, which recently yielded promising results for the ZSAR task. As mentioned in "Introduction" section, GANs can synthesise visual embeddings that are close to the distribution of real instances, but they can suffer from an issue termed *mode collapse* [35, 36], which leads to the problem of having low diversity of synthesised visual embedding.

Similar to GANs, variational autoencoders (VAEs) [37] are another generative model that employs an encoder to represent the input as a latent variable with a Gaussian distribution assumption and a decoder to transform the input from the latent variable. According to previous research articles [38], the generation of unseen visual embeddings with VAE gives more stable outputs than with GANs. Hence, the architecture of the VAEGAN component combines the advantages of VAE with that of GANs by assembling the decoder of the VAE and the generator of the GANs to ultimately synthesise semantically consistent visual representations.

As shown in Fig. 2, the real visual embedding of seen classes x extracted from a deep neural network (e.g. I3D model [39]) along with the semantic embeddings a (can be either text-based or image-based representation of the class) are the input to the encoder E. The output of E is the latent code z that compresses the information from visual representations x, optimised by the Kullback–Leibler divergence. The random noise and semantic embeddings a are the input of the generator G that synthesises the visual representation x', and the synthesised visual representations x'and real visual representations x are compared using a binary cross-entropy loss.

The discriminator D takes either x or x'along with the corresponding semantic embeddings a as the input and determines whether the input is real or synthesised. The WGAN loss is applied to the output of D to distinguish between the real and the synthesised visual representations. Additionally, both the semantic embedding decoder *SED* and the feedback module F improve the process of visual representation synthesis and reduce ambiguities amongst action classes during the zero-shot classification process. The *SED* inputs either x or x'and reconstructs the semantic embedding a', which is trained using a cycle-consistency loss.

The feedback module F transforms the latent embedding of *SED* and puts it back to the latent representation

of G which can refine x' to achieve an enhanced visual representation synthesis. It is worth noting that the generator G transforms the semantic embeddings into visual representations, whilst *SED* transforms the visual representations into semantic embeddings. Consequently, the G and the *SED* include supplementary information regarding visual representation and the supplementary information can assist to improve the quality of the visual representation synthesis and reduce ambiguity and misclassification amongst action classes.

The key approach to achieving ZSAR is to transfer semantic knowledge containing enriched and discriminative information from seen action classes to unseen action classes regardless of using either the project-based method or the GAN-based method. Semantic embeddings derived from multiple knowledge sources can potentially deliver better discriminative representation for the classifier than only using a single source [40]. In this paper, we propose two improvements for ZSAR. First, we believe it is possible to improve the ZSAR performance by introducing a combination of text-based descriptions and images to represent semantic embedding for the corresponding action class. Therefore, we use two GANs rather than one, and we then derive a single visual embedding-based representation for the unseen class by combining the two visual embeddings from the text and image sources to produce a visual embedding of the unseen classes that is calculated by the following methods: average, maximum, minimum, or summation. Second, for extracting textual features, we employ an approach that uses richer textual descriptions for the action rather just than the action class label itself. Intuitively, a textual description should contain more informative and contextual semantic meaning than just the class label. For example, 'Apply Eye Makeup' can be defined as 'cosmetics are products used to enhance or change the appearance of the face, fragrance or texture of the body'. For the visual information, we use images related to the action class that provide enriched visual cues for representing the semantic meaning.

Methodology

In this section, we describe our methodology to perform the ZSAR task based on the proposed dual-GAN model using two HAR benchmark datasets (i.e. HMDB51 and UCF101). We also introduce the method for collecting images for each action class and the method for extracting visual-based and text-based semantic embeddings in more detail.

Datasets We selected two datasets containing human actions, named HMDB51 [41] and UCF101 [42] that are widely used as benchmarks to evaluate the ZSAR performance. The details of the two datasets are described in Table 1. HMDB51 contains 6676 videos divided into 51 action classes, collected from various sources such as movies, YouTube and Google videos. UCF101 contains 101 action classes with a total of 13,320 videos collected from YouTube.

Table 1 Datasets used for experiments

Dataset	#Class	#Instances	Seen/unseen proportion
HMDB51	51	6676	26/25
UCF101	101	13,320	51/50

We used the same split for model training and evaluation as previous related works [16, 17]. Each dataset has 30 independent splits, and each split is randomly generated by keeping the same seen/unseen proportion (51 seen and 50 unseen classes for UCF101, and 26 seen and 25 unseen classes for HMDB51). Each class is predefined as seen or unseen class in each split and seen classes are only for training and unseen classes are only for tests under the ZSL setting. In other words, some classes could be *seen classes* in one split and *unseen classes* in other splits.

Image collection In our approach, we apply a similar strategy to collect images to the one proposed by [32] in which the following steps are followed. First, we consider the action labels as keywords used to search-related images on Image Search Engines (i.e. Google Image Source).[1] For example, we use the keyword *Playing YoYo* for searching images for the class *YoYo*. Then, after collecting the images, we remove the irrelevant and small-size images for each class for both datasets. As a result, we obtain 15,845 images (157 images per class on average) and 6856 images (134 images per class on average) for the UCF101 and the HMDB51, respectively.

Visual and semantic embeddings To extract real visual embedding of seen classes x in Fig. 2, we adopted the off-the-shelf I3D model for visual feature extraction provided by [16]. I3D was originally proposed by [39] and it contains RGB and Inflated 3D networks to generate appearance and flow features from the *Mixed_5c* layer. For each video instance, the outputs from the *Mixed_5c* layer for both networks are averaged through a temporal dimension, pooled in the spatial dimension, and then flattened to obtain a 4096-dimensional vector for appearance and flow features. In the end, both appearance and flow features are

[1] Image scraping tool is available at https://github.com/Joeclinton1/google-images-download.git.

Table 2 The details of knowledge sources and semantic embeddings

Semantics	Source	Embedding	Dimensions
Labels	Text	Word2Vec	300
Descriptions	Text	Word2Vec	300
Collected images	Image	GoogleNet	1024
Collected images	Image	ResNet101	2048

The embeddings, derived from two different types of semantic sources with different dimensions, are used as the input for the GANs' training and evaluation

concatenated to represent a video with an 8192-dimensional vector.

We produced four types (two for text and two for image) of semantic embedding a that can be used to condition the VAEGAN as shown in Fig. 2. The summary of semantic embedding is given in Table 2. The semantic embedding of action labels is transformed using Word2Vec. Word2Vec [31], which is built upon a skip-gram model that was pre-trained on a large-scale text corpus (i.e. Google News Dataset), is used to deliver a 300-dimensional vector for each action class label. The text-based description per class for both datasets is provided by the work [32], motivated by the fact that a class label is not adequate to represent the complex concepts in human actions. The idea is that each label is transformed into a description of that label and then, we use Word2vec to represent each word of that description. Then, we simply average all the Word2Vec arrays, which also deliver a 300-dimensional vector for each class.

To generate the visual representation of an image, we apply two off-the-shelf models: GoogLeNet [43] and ResNet101 [44] which were both pre-trained on the ImageNet dataset. The average pooling layer that is before the last fully connected layer is used as the deep image features for both pre-trained models. As a result, for the collected images, the visual features are represented as a 2048-dimension vector and 1024-dimension vector by GoogLeNet and ResNet-101, respectively and we average them for each action class.

Embeddings combination As shown in Step 1 in Fig. 1, we aim to synthesise and combine visual embeddings for unseen classes using various knowledge sources (i.e. textual descriptions and collected images related to class labels) in the proposed dual-GAN approach. We have considered four methods to combine the two pseudo-unseen visual embeddings conditioned by the text-based and the image-based knowledge sources using the following operations: averaging, summation, maximum and minimum. For averaging, we calculate the mean of the unseen visual embedding from the text-based semantic knowledge source and the unseen visual embedding from the image-based semantic knowledge source. For summation, the same position of each element for both synthesised unseen visual embeddings is summed up. For maximum, the larger value in each position between two synthesised visual embeddings is selected. Similarly, for minimum, the smaller value in each position is selected. All four embedding combination methods will be empirically evaluated on the datasets using the proposed dual-GAN approach.

We generate two visual embeddings per unseen class instance (i.e. based on text-GAN image-GAN). Both visual embeddings represent the unseen class. Therefore, we want to merge these two multi-dimensional representations of the class. The idea of merging vectors using different operations is inspired by the work in word embedding space, where the distance between embeddings reflects equivalent semantic distance. For example, in Word2Vec where words are represented as one-hot encoding in a high-dimensional space, we can use the vectors computationally for word relationships, such as King − Man + Women = Queen. The average is a typical operation for representing a document. The idea behind this is that the average of all the Word2Vec representations of the words of a document would be a centroid in the feature space that will likely represent that document. We have done the same whereby the average will represent a better representation of the visual embeddings. Likewise, the maximum, minimum and summation can also work very well for representing the meaning of combining embeddings.

Evaluation metrics Class accuracy is a standard metric in the ZSAR field. To represent the performance of the methodologies, we use the average per-class accuracy [20] defined by the following equation:

$$\mathrm{ACC}_{\mathrm{class}} = \frac{1}{N_{\mathrm{class}}} \sum_{C=1}^{N_{\mathrm{class}}} \frac{\#\text{ correct predictions in Class } C}{\#\text{ instances in Class } C} \quad (1)$$

The mean per-class accuracy averaged over 30 independent splits will be reported along with the standard deviation.

Experiments

In this section, we present the experimental configurations for comparing our proposed dual-GAN approach that incorporates two knowledge sources (i.e. texts and images) with other state-of-the-art methodologies on the UCF101 and HMDB51 datasets. Their implementations are described in detail.

Experiments and baseline To the first research question described in "Introduction" section on whether the

Table 3 The methods used in experiments for comparing text-driven semantic embedding to image-driven semantic embedding in the single-GAN model

Dataset	Knowledge source	Semantic embedding
UCF101	Text (baseline)	Action class Word2Vec
	Text	Description Word2Vec
	Image	GoogLeNet
	Image	ResNet101
HMDB51	Text (baseline)	Action class Word2Vec
	Text	Description Word2Vec
	Image	GoogLeNet
	Image	ResNet101

information provided from images was more suitable for GANs than that from text, we investigated whether the synthesised visual embeddings conditioned on the image-driven knowledge source can lead to better ZSAR accuracies than those from the text-driven knowledge source using a single-GAN model. The single-GAN model follows only one line of the dual-GAN pipeline using either text-VAEGAN or image-VAEGAN depending on which knowledge source is used without the process of embedding combination illustrated in Fig. 1. Table 3 shows that two text-driven knowledge sources (i.e. class label and description) and two image-driven knowledge sources (i.e. GoogLeNet and ResNet101) are evaluated for each dataset. As a baseline, we use the Word2Vec of action class label to represent the semantic embedding for the UCF101 and the HMDB51, respectively.

For answering the second research question introduced in "Introduction" section on whether two sources can work better than just one, we investigate and evaluate which embedding combination method is best for both datasets. The embedding combination methods are averaging Avg.), summation (Sum.), maximum (Max.) and minimum (Min.). The results from dual-GAN experiments are compared to the results from the single-GAN to investigate whether the dual-GAN approach can deliver better ZSAR performance than the single-GAN approach.

Methodology implementation Similar to our last work [18], the structures of the discriminator D, the encoder E and the generator G are designed as fully connected networks in two layers along with 4096 hidden units. The semantic embedding decoder SED and the feedback module F have the same structure as D, E and G. Leaky ReLU is used for each activation function, except in the output of G, where a sigmoid activation is applied to calculate the binary cross-entropy loss. The whole framework is trained using an Adam optimiser with a learning rate of 10^{-4}. The supervised-learning classifier is a single-layer fully connected network with equal output units to the number of unseen classes. We apply the same hyper-parameters as in the work [17], where α, β and σ are set to 10, 0.01 and 1, respectively. As explained in the work [23], α is the coefficient for weighting the WGAN loss, β is a hyper-parameter for weighting the decoder reconstruction error in the semantic decoder embedding SED, and σ is used in the feedback module F to control the feedback modulation. The gradient penalty coefficient λ is initially set to 10 for training a GAN. All experiments were conducted on Google Colab which provides a Tesla P100 GPU with 25 GB memory use.

Table 4 Comparing our results to the TF-VAEGAN

Model	TF-VAEGAN [17]	single-GAN (ours)
Dataset		
HMDB51	33.00%	31.75%
UCF101	41.00%	38.42%

Additionally, the number of synthesised visual embeddings is a hyper-parameter in the experiments. Therefore, to efficiently conduct the initial experiments, we synthesised 400 and 800 visual embeddings for each unseen class on UCF101 and HMDB51, respectively. Afterwards, we empirically evaluate how the number of synthesised unseen visual embeddings can influence the ZSAR performance. Our code is available online and is compatible with Pytorch 1.9.0 and CUDA 11.1 version.[2]

Results and Analysis

In this section, we present and analyse the results of the conducted experiments for all configurations described in "Experiments". For each configuration, the mean average accuracy is reported along with the standard deviation.

Verification of experimental baseline Our first experimental run is to confirm that we have set up the TF-VAEGAN experimental pipeline correctly. We compare our results to the work [17] from which our model is built using identical semantic embeddings for both datasets. The results are shown in Table 4. For the HMDB51, the Word2Vec of each action label is used as semantic embedding. Our result is degraded by 1.25% against the TF-VAEGAN. Similarly, for the UCF101, the annotated class-level attributes provided by the work [9] are used and our result is decreased by 2.58%. Note that, due to the scaling limit of using annotated attributes in other datasets, attribute-based semantic information will not be used for further experiments and comparisons.

[2] https://github.com/kaiqiangh/kg_gnn_gan.

Table 5 Results from the single-GAN approach for both UCF101 and HMDB51 datasets

Dataset	Semantic embedding	Acc	Std (%)
HMDB51	Action class W2V	**31.75%**	3.30%
	Description W2V	27.21%	3.48%
	GoogLeNet	29.52%	4.26%
	ResNet-101	31.41%	3.86%
UCF101	Action class W2V	28.02%	3.04%
	Description W2V	29.09%	2.61%
	GoogLeNet	44.35%	3.15%
	ResNet-101	**45.87%**	3.42%

The numbers in bold denote the best result for each dataset

Acc mean average accuracy, *Std* standard deviation, *W2V* Word2Vec

Comparing image-source approaches with text-source approaches for GANs Table 5 shows the results of evaluating the text-based (i.e. action classes and textual descriptions) and image-based (GoogLeNet and ResNet101) semantic embeddings on our single-GAN implementations for both datasets. As can be seen, in the HMDB51 the case of using action class Word2Vec outperforms other semantic embeddings, such as 'description Word2Vec' description, GoogLeNet and ResNet101 by a margin of 4.54%, 2.23% and 0.34%, respectively. Admittedly, the results are against one of our hypotheses that image-based semantic embeddings have higher performance than text-based ones. We randomly selected 40 videos across 4 action classes in the HMDB51 as manual checking, with examples shown in Fig. 3. We observed that the majority of videos examined have many non-class-related objects and visual contents in the background. In other words, the clips from the videos for a given labelled action have many other objects that were not filtered, making it more difficult for the algorithm to classify. Furthermore, some of the collected images based on the class labels could not be related to the video samples from the perspective of the visual content of those images for conditioning GANs training. Those images do not represent relevant semantic information for such action classes (e.g. the frame of class *Walk* in the HMDB51 is not highly related to the label), which results in performance degradation. Also, the HMDB51 has 25 unseen classes for testing with a random guess is 4% (i.e 1/25) and the UCF101 has 50 unseen classes for testing with a random guess is 2% (i.e. 1/50), lower than the HMDB51. However, the results show that the GAN-based framework can yield better ZSAR performance on the UCF101 than the HMDB51, even though UCF101 has more testing classes. As UCF101 instances have fewer unrelated visual contents, this is likely to improve the ZSAR performance in this regard. Furthermore, although textual descriptions contain much more semantic information than just the labels, their performance is lower since classes in HMDB51 are difficult to accurately describe by a set of sentences.

The single-GAN results for the UCF101 are in line with our hypothesis since the image-based ResNet101 semantic embedding outperforms the 'action class Word2Vec', the 'description Word2Vec' and 'image-based GoogLeNet' by large margins of 17.85%, 16.78% and a small margin of 1.52%, respectively. We also randomly selected 80 videos across 4 action classes in the UCF101 (see examples in Fig. 3. We discovered that compared to the videos examined in HMDB51, they have a clean background—in the sense that no non-class-related objects appeared in the background to confuse the classification algorithm. The dataset has single and centred actors, which can be accurately represented by either textual descriptions or relevant images. To our best understanding, the videos from the HMDB51 are collected from movies and YouTube videos without much modification, such as video cropping and centring, whilst the videos from the UCF101 are largely collected from YouTube videos but with video selection standards such as picking videos that have a relatively clean background with fewer actors. Additionally, previous works [16, 17] have indicated that ZSAR performance using the HMDB51 dataset is poorer than using the UCF101 dataset and our experimental results are in line with this finding. Therefore, we suggest that the ZSAR performance is closely related to the clarity and focus of the videos with regard to the associated action label. Note that, for both datasets, ResNet101 can deliver a slight boost over GoogLeNet probably due to the better model capability of representing image features.

Comparison between the dual-GAN approach and the single-GAN approach Table 6 shows the results of evaluating the dual-GAN model by employing four embedding combination methods on the HMDB51. Note that, the highest mean average accuracy for each case of *Dual Semantic Embeddings* is highlighted in bold and the best result for all cases is marked with the symbol (*). ZSAR performance from all cases using the dual-GAN model outperforms the best single-GAN case (i.e. action class Word2Vec) by a margin of 4.30%. However, no embedding combination method dominates over others. As illustrated in Fig. 4, the *Max.* combination method yields the best performance at 36.05% when action class Word2Vec is involved, but the *Min.* combination method also delivers promising performance when including the information of class descriptions. Similar to the findings from the single-GAN experiment, there is still poor performance when including the textual descriptions in the dual-GAN model.

As can be seen in Table 7, the *Max.* combination method surpasses the other methods for all the *dual semantic embeddings* cases in the UCF101 dataset. The combination of 'Description Word2vec' and ResNet101 using the *Max.* operator delivers

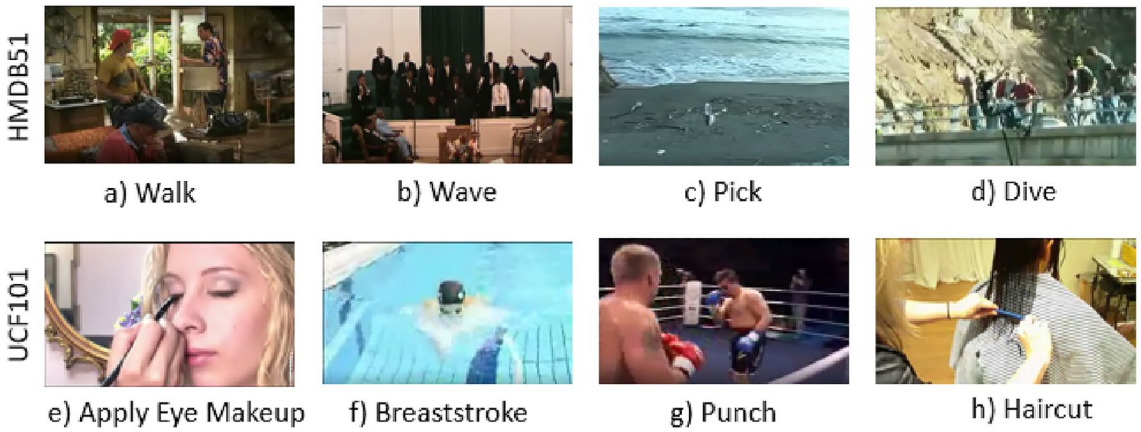

Fig. 3 The frame examples of videos from two datasets show that the background quality in the HMDB51 is not clear and contains many irrelevant visual contents, but the background in the UCF101 is relatively clear without much noisy data. In addition, some frames are not related to the corresponding labels in the HMDB51, such as Walk and Wave

Table 6 A comparison of dual-GAN model with different combination methods for HMDB51

Dual semantic embedding	Avg		Sum		Max		Min	
	Acc	Std	Acc	Std	Acc	Std	Acc	Std
Action Class Word2Vec and GoogLeNet	35.15%	3.26%	34.91%	3.02%	**35.39%**	3.11%	35.36%	2.89%
Action Class Word2Vec and ResNet101	35.94%	3.28%	35.59%	3.25%	**36.05%** *	3.38%	35.93%	3.11%
Description Word2Vec and GoogLeNet	32.72%	3.41%	32.15%	3.80%	32.46%	3.81%	**32.83%**	3.60%
Description Word2Vec and ResNet101	33.37%	3.71%	33.71%	3.66%	33.14%	3.67%	**33.88%**	3.67%

The numbers in bold denote the best result for each type of semantic embedding

An asterisk * denotes the best result amongst all cases

Acc average accuracy, *Std* standard deviation (in %)

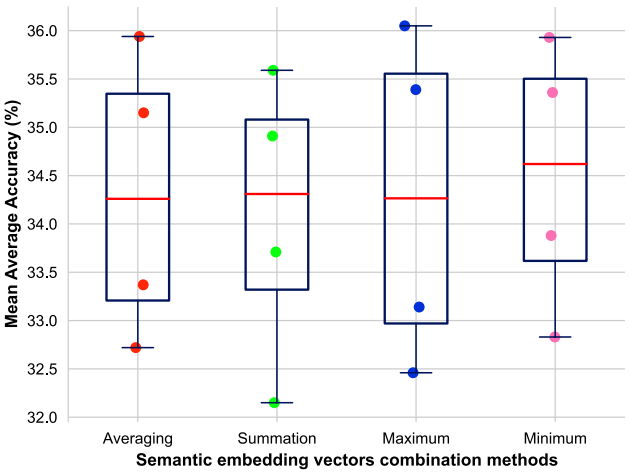

Fig. 4 A comparison of dual-GAN using different combination methods in HMDB51. The *Max* has the best performance, but the *Min* delivers the best-averaged performance

the best performance with an accuracy of 46.37%, which surpasses the baseline (i.e. action class Word2Vec in the single-GAN model) by a large margin of 18.35%. As opposed to the HMDB51, where the approaches that included textual descriptions delivered better results than the cases that included action labels. We suggest that using textual descriptions used to represent the semantic embedding of the class in addition to visual embeddings in a clean and less-noisy action dataset (e.g. less background clutter and less unrelated contents appearing in videos) has a positive impact on the performance of the ZSAR. Additionally, as shown in Fig. 5, the *Max.* operator also performs the best at an average level. To this end, Fig. 6 reports the comparison of the best results for each semantic embedding case (i.e. text-only, image-only, text & image).

Comparison between our proposed dual-GAN approach and other published approaches For further investigations, we compare our best results to the existing approaches that follow the GAN-based framework on both datasets. As presented in Table 8, our dual-GAN model outperforms other approaches by at least 3.05% and 5.37% for HMDB51 and UCF101, respectively. There is no doubt that combining embeddings derived from different knowledge sources (i.e. texts and images) delivers a performance boost in the ZSAR. Note that, we do not implement and

Table 7 A comparison of the dual-GAN model with different combination methods for UCF101

Dual semantic embedding	Avg		Sum		Max		Min	
	Acc	Std	Acc	Std	Acc	Std	Acc	Std
Action class Word2Vec and GoogLeNet	41.20%	3.21%	41.14%	3.17%	**41.84%**	3.22%	41.06%	3.19%
Action class Word2Vec and ResNet101	41.29%	3.34%	41.05%	3.38%	**41.95%**	3.37%	41.24%	3.33%
Description Word2Vec and GoogLeNet	45.01%	2.78%	44.73%	2.71%	**45.59%**	2.77%	44.85%	2.66%
Description Word2Vec and ResNet101	45.58%	3.00%	45.57%	3.12%	**46.37%** *	3.10%	45.37%	3.00%

The numbers in bold denote the best result for each type of semantic embedding

Acc and Std denote mean average accuracyand standard deviation (in %), respectively

*Denotes the best result among all cases.

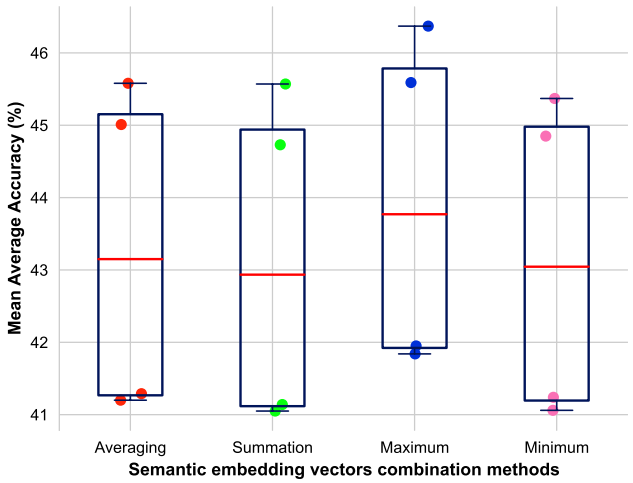

Fig. 5 A comparison of dual-GAN using different combination methods in UCF101. The *Max* delivers the highest performance and also yields the best-averaged performance

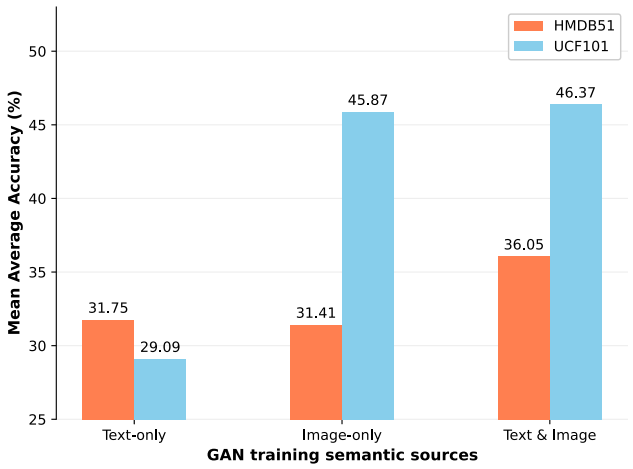

Fig. 6 A comparison of the best result of each semantic embedding case for both datasets

evaluate other approaches, but directly report the results from the work [17]. As shown in Fig. 7, we also report the best-performance methods for both datasets UCF101 and HMDB51 over epochs, respectively.

As a result, we summarise our main findings as follows: (1) The image-driven semantic embedding is only better than the text-driven one on the UCF101 dataset. (2) All cases using the dual-GAN model outperform their counterpart cases when using a single-GAN on both datasets, indicating that the fused semantic embedding obtained from two knowledge sources is more representative of the classes. (3) The *Max.* combination method performs better than the other methods in most cases. Additionally, we could improve this approach in the future by fine-tuning the hyper-parameters of the proposed dual-GAN model.

Analysis of the results at class-level It is necessary to explore how classifications are performed amongst action classes and to see which classes have better performance. We present per-class accuracies over single-GAN and dual-GAN methods for both datasets in Figs. 8 and 9. Note that, each action class and GAN-based method are represented on X-axis and Y-axis, respectively and also the deeper colour represents the higher accuracy. For example, *class* in X-axis denotes the single-GAN model with the semantic embedding of action class Word2Vec. *class_googlenet_avg* denotes the dual-GAN model with the averaged fused dual semantic embedding of action classes Word2Vec and GoogLeNet. In Fig. 8, the classes of *cart wheel*, *dribble*, *shake hand* and *ride bike* have higher accuracies across most GAN-based methods. However, there are some classes with poor accuracies, such as *talk*, *turn* and *pick*. It is noted that the labels of those poor-accuracy classes are general and less discriminative for the classifier. For example, for action classes that may show up as part of other actions, such as *turn* and *talk*, it is difficult to have clear unambiguous visual representations of these classes.

In addition, we show the segments of confusion matrices for both datasets to further explore class-level ZSAR results. As shown in Fig. 10, 25% instances of *kick ball* are misclassified as the class *catch*. Also, 38% and 43% of the instances of *cart wheel* and *somersault* are misclassified as *flic flac*. Also, as shown in Fig. 11, 40%, 46%, 47% and 58% of the clips of

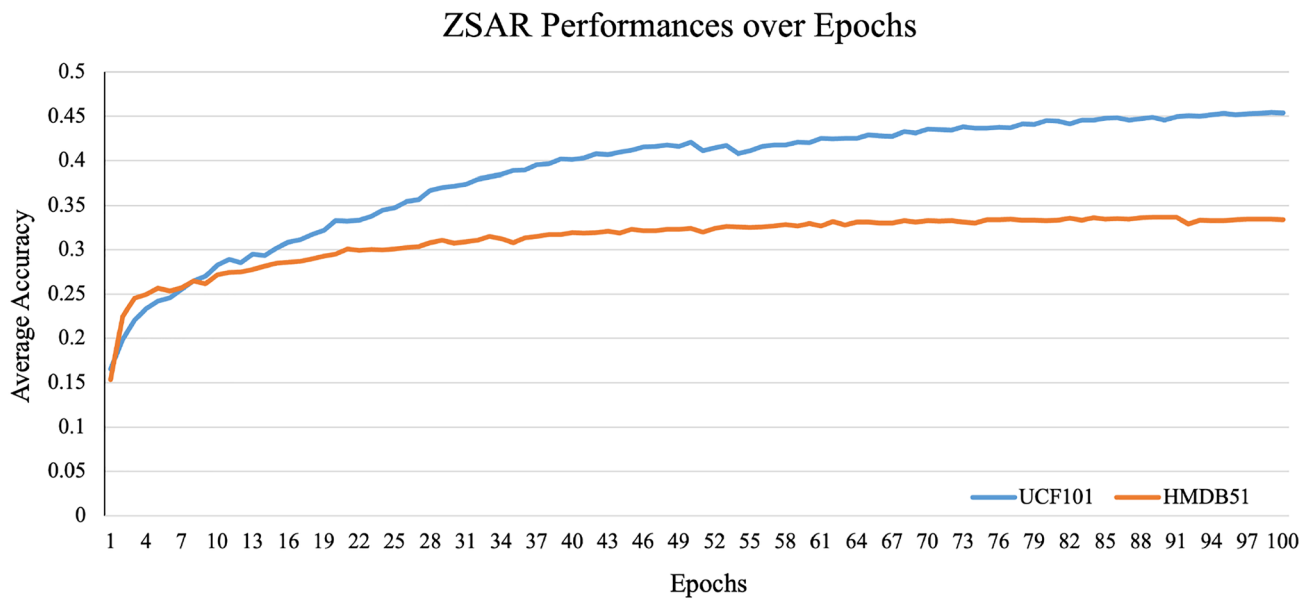

Fig. 7 ZSAR performance with the best methods for UCF101 and HMDB51 over epochs

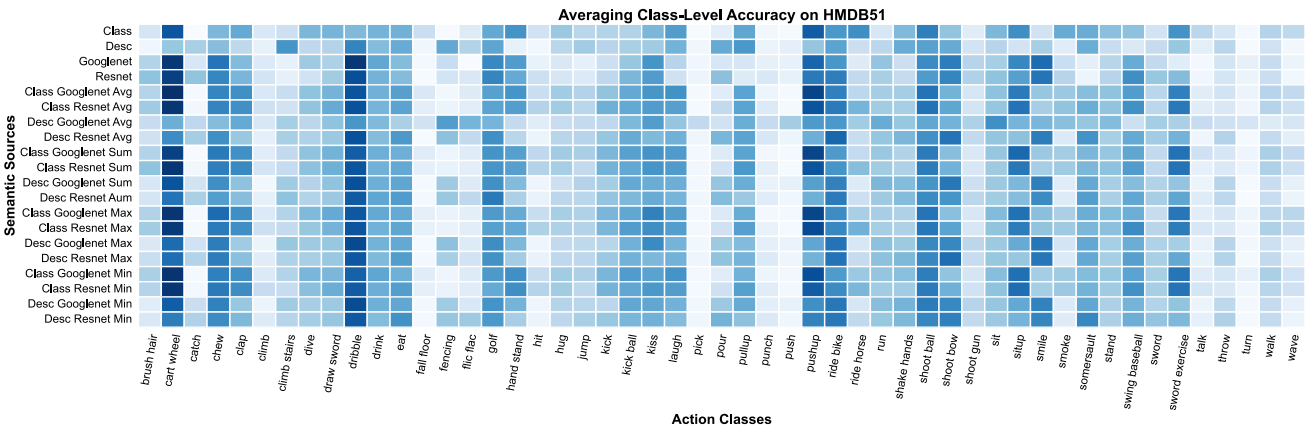

Fig. 8 Heatmap for per-class accuracy for each method for the HMDB51 dataset. Classes are listed along the horizontal axis, with methods along the vertical. Accuracy is indicated by the depth of colour: The darker, the higher the performance

Punch, *Shaving Beard*, *Playing Violin* and *Balance Beam* are misclassified as *Box Punching Bag*, *Blow Dry Hair*, *Playing Dhol* and *Parallel Bars*, respectively. As a result, the videos with similar visual cues (i.e. *Punch* class and *Box Punching Bag* class have similar body movements, such as arms and fists moving) from different action classes could be easily confused with the classification model in the GAN-based framework.

Predicting classes in the context of ZSAR is very challenging since the model has not seen any of the classes with which it is tested. Additionally, the fact that we are applying a multiclass classification approach with 50 unseen classes for the UCF101 dataset and 25 classes for the HMDB51 dataset makes the predictions for the model even more difficult. Therefore, the expectations of accuracy in this domain are much lower than in other domains (e.g. action recognition) where the models are trained with all the tested classes. As shown in Table 8, our results outperform the results in the state-of-the-art. The advantage of ZSAR is that manual labelling is not necessary. The downsides are that it requires higher computation and that the results are not as accurate as when models are trained with those classes. Also, we reported more evaluation metrics such as precision, recall and F1 score for our best model in Table 9.

Impact of the number of synthesised embeddings on the accuracy The number of unseen visual embeddings being

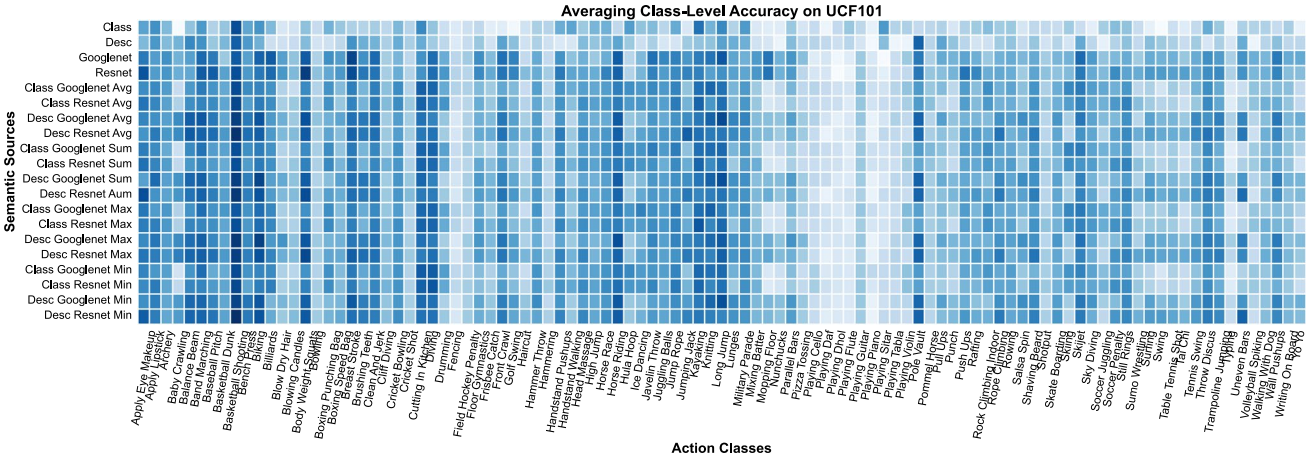

Fig. 9 Heatmap of per-class accuracy for each method on the UCF101 dataset. Classes are listed along the horizontal axis and methods along the vertical. The darker the colour, the more accurate that class using the associated method has been predicted

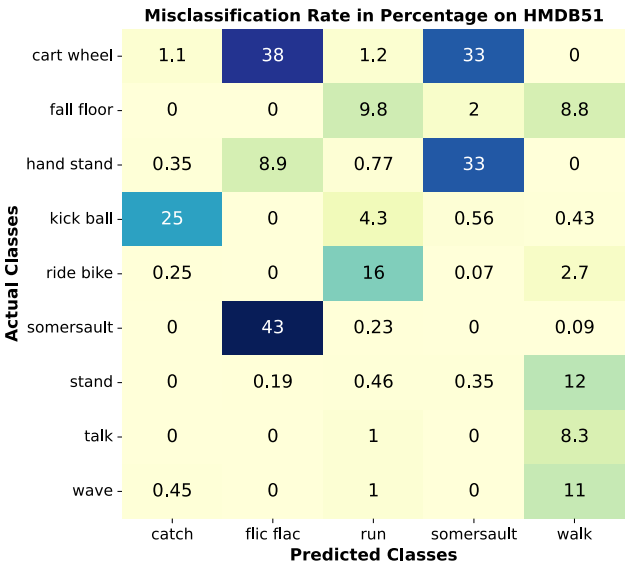

Fig. 10 This is a selection of the most representative elements of the confusion matrix for the HMDB51 dataset. The elements represent the percentage of misclassifications of the actual classes (vertical axis) that were incorrectly predicted to another class (horizontal axis). For example, 38% instances of 'cart wheel' were incorrectly classified as 'flic flac'

synthesised by the generator (G) could be an important factor in the ZSAR performance. As a further exploration, we selected the best result for UCF101 (i.e. Action Class Word2Vec & ResNet101 with max combination) and HMDB51 (i.e. Description Word2Vec & ResNet101 with max combination). The number of unseen visual embeddings was initially set to 200 and then it was increased by 200, within the range from 200 to 1600. As shown in Fig. 12, the best performance is delivered when using 1200 and 200 unseen visual embeddings for UCF101 and HMDB51, respectively. It is not guaranteed that more generated embeddings can yield higher accuracy in this regard. Therefore, it is suggested that choosing a smaller number of unseen embeddings would be efficient.

Conclusions

In this work, we have empirically evaluated the ZSAR performance using text-driven and image-driven semantic embeddings related to the action classes in the GAN-based framework on the HMDB51 and UCF101 datasets. We also have investigated the impact of combining both text and image knowledge by applying different combination methods (i.e. averaging, summation, maximum and minimum).

We have proven that by applying image-driven semantic embeddings, we can deliver significant increments in performance when compared to the text-driven one within a range between 15.26% (GoogLeNet against Description) and 17.85% (ResNet101 against Action Class) in the single-GAN framework for UCF101. However, HMDB51 does not follow this pattern, which we reason may be due to its video samples containing videos with different objects in the background. Similarly, in the perspective of employing different types of text-driven semantic embedding, the descriptions yield slightly better performance than the class labels by 1.07% in the UCF101 dataset (but not in HMDB51) since descriptions contain more enriched information. Furthermore, our proposed dual-GAN model outperforms the

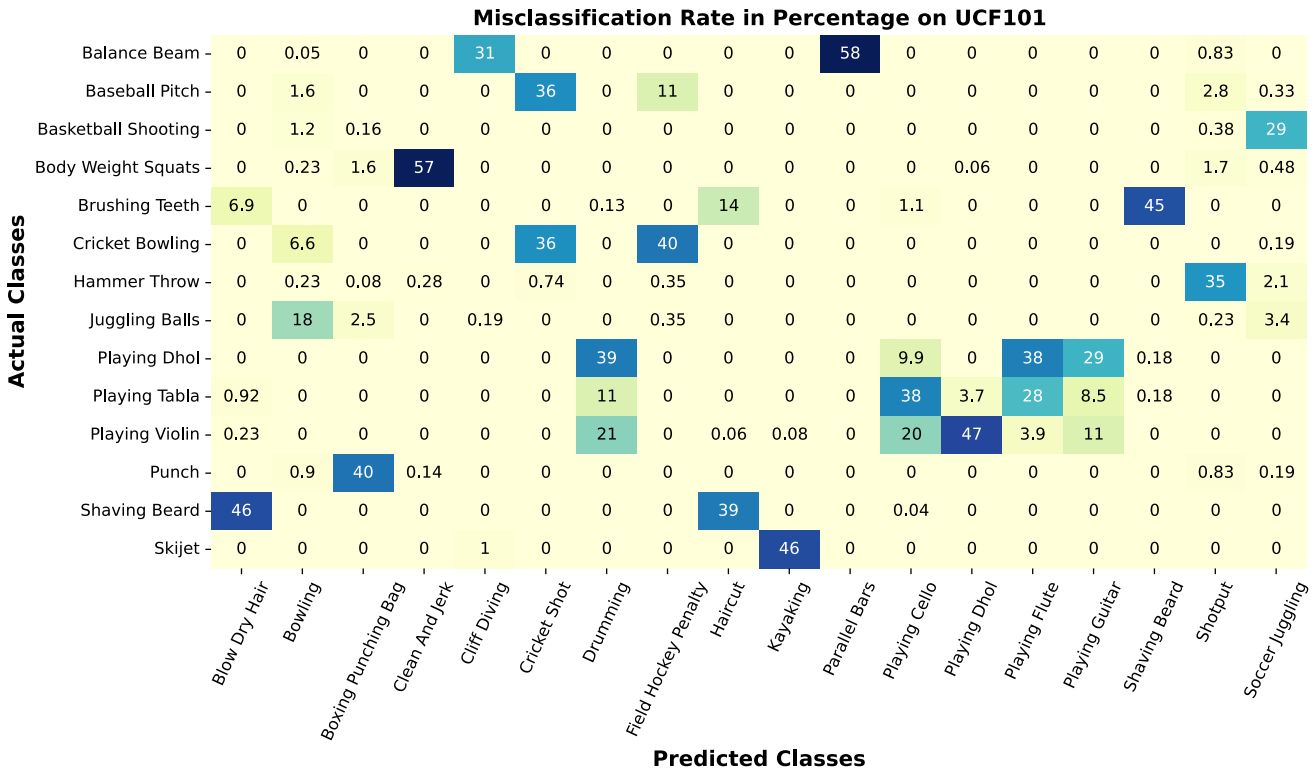

Fig. 11 This is a selection of the most representative elements of the confusion matrix for the UCF101 dataset. Each element represents the percentage of the actual classes (vertical axis) that were misclassified as a different class (horizontal axis)

Table 8 A comparison of the ZSAR performance represented as the mean average accuracy of our dual-GAN model with the existing approaches (generative-based) in the literature for the HMDB51 and the UCF101 datasets

Methods	Gaussian mixture model [34]	Classification-loss Wasserstein generative adversarial networks [27]	Out-of-distribution detection [16]	Feature variational autoencoders and GAN [23]	Latent embedding feedback and discriminative feature [17]	dual-GAN (our approach)
Datasets						
HMDB51	20.7%	29.1%	30.2%	31.1%	33.0%	**36.05%**
UCF101	20.3%	37.5%	38.3%	38.2%	41.0%	**46.37%**

The numbers in bold denote the best result for each dataset

The compared methods are shown in the related work

Table 9 A comparison of more evaluation metrics of our dual-GAN model (best model) for the HMDB51 and the UCF101 datasets

Metrics	Accuracy	Precision	Recall	F1 score
Datasets				
HMDB51	36.05%	35.17%	31.99%	33.50%
UCF101	46.37%	44.94%	44.21%	44.57%

baseline (i.e. action class in the single-GAN model) by large margins of 4.30% and 18.35% and also outperforms the existing GAN-based approaches by a minimum of 3.05% and 5.37% in the datasets HMDB51 and UCF101, respectively.

For future work, we could investigate generalised ZSAR which is a more challenging task that tests both seen and unseen classes together in the classification stage. Also, we plan to explore other approaches to produce more enriched and meaningful semantic embeddings that can also mitigate

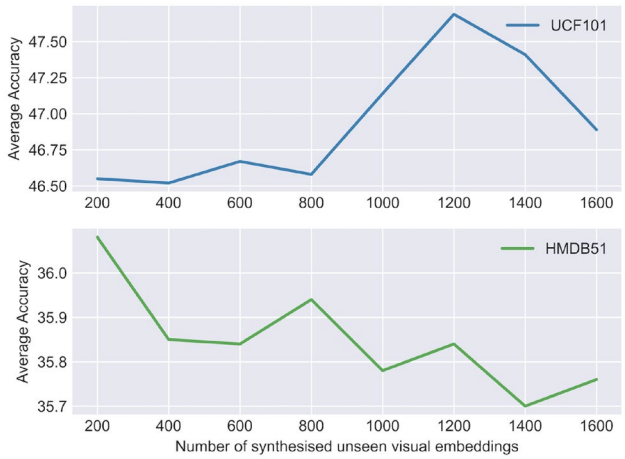

Fig. 12 Influence of the different numbers of synthesised unseen visual embeddings on HMDB51 and UCF101

the problem of the semantic gap between classes and video samples.

We are also planning to use other combination methods such as concatenation. Additionally, we could also use two different classifiers and calculate the predicted class as a combination of both classifiers. We also plan to use other supervised methods such as random forest, support vector machines, or deep learning to see if they can deliver better results. Lastly, an interesting line of research is improving the performance of the proposed approach by fine-tuning the parameters of the pipeline.

Acknowledgements This project is funded under the Fiosraigh Scholarship of Technological University Dublin.

Funding Open Access funding provided by the IReL Consortium.

Data availability Additionally, the link to access all the data used in this paper is provided- https://drive.google.com/drive/folders/1dQhiquJzaeA1wKD2gFBsGYlJq7OQLBPW?usp=sharing.

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

References

1. Sahoo SP, Ari S, Mahapatra K, Mohanty SP. HAR-depth: a novel framework for human action recognition using sequential learning and depth estimated history images. In: IEEE transactions on emerging topics in computational intelligence. 2020.
2. Ponce H, Martínez-Villaseñor MDL, Miralles-Pechuán L. A novel wearable sensor-based human activity recognition approach using artificial hydrocarbon networks. Sensors. 2016;16(7):1033.
3. Wang H, Schmid C. Action recognition with improved trajectories. In: Proceedings of IEEE ICCV. 2013; p. 3551–3558.
4. Karpathy A, Toderici G, Shetty S, Leung T, Sukthankar R, Fei-Fei L. Large-scale video classification with convolutional neural networks. In: Proceedings of IEEE CVPR. 2014; p. 1725–1732.
5. Simonyan K, Zisserman A. Two-stream convolutional networks for action recognition in videos. In: Advances in NIPS. 2014; p. 568–576.
6. Feichtenhofer C, Pinz A, Zisserman A. Convolutional two-stream network fusion for video action recognition. In: Proceedings of CVPR. 2016. p. 1933–1941.
7. Hara K, Kataoka H, Satoh Y. Can spatiotemporal 3d cnns retrace the history of 2d cnns and imagenet? In: Proceedings of CVPR. 2018. p. 6546–6555.
8. Khan MA, Zhang YD, Khan SA, Attique M, Rehman A, Seo S. A resource conscious human action recognition framework using 26-layered deep convolutional neural network. Multimed Tools Appl. 2021;80(28):35827–49.
9. Liu J, Kuipers B, Savarese S. Recognizing human actions by attributes. In: CVPR 2011. IEEE. 2011. p. 3337–3344.
10. Jain M, Van Gemert JC, Mensink T, Snoek CG. Objects2action: classifying and localizing actions without any video example. In: Proceedings of the IEEE international conference on computer vision. 2015. p. 4588–4596.
11. Wang Q, Chen K. Zero-shot visual recognition via bidirectional latent embedding. IJCV. 2017;124(3):356–83.
12. Estevam V, Pedrini H, Menotti D. Zero-shot action recognition in videos: a survey. Neurocomputing. 2021;439:159–75.
13. Xian Y, Akata Z, Sharma G, Nguyen Q, Hein M, Schiele B. Latent embeddings for zero-shot classification. In: Proceedings of CVPR. 2016; p. 69–77.
14. Xu X, Hospedales T, Gong S. Transductive zero-shot action recognition by word-vector embedding. Int J Comput Vis. 2017;123(3):309–33.
15. Huang K, Delany SJ, McKeever S. Fairer evaluation of zero shot action recognition in videos. In: VISIGRAPP (5: VISAPP). 2021. p. 206–215.
16. Mandal D, Narayan S, Dwivedi SK, Gupta V, Ahmed S, Khan FS, et al. Out-of-distribution detection for generalized zero-shot action recognition. In: Proceedings of CVPR. 2019. p. 9985–9993.
17. Narayan S, Gupta A, Khan FS, Snoek CG, Shao L. Latent embedding feedback and discriminative features for zero-shot classification. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXII 16. Springer; 2020. p. 479–495.
18. Huang K, Luis Miralles-Pechuán, Mckeever S. Zero-shot action recognition with knowledge enhanced generative adversarial networks. In: Proceedings of the 13th international joint conference on computational intelligence. 2021. p. 254–264
19. Liu H, Yao L, Zheng Q, Luo M, Zhao H, Lyu Y. Dual-stream generative adversarial networks for distributionally robust zero-shot learning. Inf Sci. 2020;519:407–22.
20. Xian Y, Schiele B, Akata Z. Zero-shot learning-the good, the bad and the ugly. In: Proceedings of the IEEE conference on CVPR. 2017. p. 4582–4591.

21. Bansal A, Sikka K, Sharma G, Chellappa R, Divakaran A. Zero-shot object detection. In: Proceedings of the European conference on computer vision (ECCV). 2018. p. 384–400.
22. Wang X, Ye Y, Gupta A. Zero-shot recognition via semantic embeddings and knowledge graphs. In: Proceedings of the IEEE conference on computer vision and pattern recognition. 2018. p. 6857–6866.
23. Xian Y, Sharma S, Schiele B, Akata Z. f-vaegan-d2: a feature generating framework for any-shot learning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2019; p. 10275–10284.
24. Xu X, Hospedales T, Gong S. Semantic embedding space for zero-shot action recognition. In: 2015 IEEE international conference on image processing (ICIP). IEEE; 2015. p. 63–67.
25. Li Y, Hu S, Li B. Recognizing unseen actions in a domain-adapted embedding space. In: 2016 IEEE international conference on image processing (ICIP). IEEE; 2016. p. 4195–4199.
26. Xu X, Hospedales TM, Gong S. Multi-task zero-shot action recognition with prioritised data augmentation. In: European conference on computer vision. Springer; 2016. p. 343–359.
27. Xian Y, Lorenz T, Schiele B, Akata Z. Feature generating networks for zero-shot learning. In: Proceedings of the IEEE conference on computer vision and pattern recognition. 2018. p. 5542–5551.
28. Felix R, Reid I, Carneiro G, et al. Multi-modal cycle-consistent generalized zero-shot learning. In: Proceedings of the European conference on computer vision (ECCV). 2018. p. 21–37.
29. Huang H, Wang C, Yu PS, Wang CD. Generative dual adversarial network for generalized zero-shot learning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2019. p. 801–810.
30. Mishra A, Pandey A, Murthy HA. Zero-shot learning for action recognition using synthesized features. Neurocomputing. 2020;390:117–30.
31. Mikolov T, Sutskever I, Chen K, Corrado GS, Dean J. Distributed representations of words and phrases and their compositionality. In: Advances in neural information processing systems. 2013. p. 3111–3119.
32. Wang Q, Chen K. Alternative semantic representations for zero-shot human action recognition. In: Joint European conference on machine learning and knowledge discovery in databases. Springer; 2017. p. 87–102.
33. Hong M, Li G, Zhang X, Huang Q. Generalized zero-shot video classification via generative adversarial networks. In: Proceedings of the 28th ACM international conference on multimedia. 2020. p. 2419–2426.
34. Mishra A, Verma VK, Reddy MSK, Arulkumar S, Rai P, Mittal A. A generative approach to zero-shot and few-shot action recognition. In: 2018 IEEE Winter conference on WACV. IEEE. 2018. p. 372–380.
35. Arjovsky M, Bottou L. Towards principled methods for training generative adversarial networks. 2017. arXiv:1701.04862.
36. Walia M, Tierney B, Mckeever S. Synthesising tabular data using Wasserstein conditional GANs with gradient penalty (WCGAN-GP). In: AICS. 2020.
37. Kingma DP, Welling M. Auto-encoding variational Bayes. 2013. arXiv:1312.6114.
38. Verma VK, Arora G, Mishra A, Rai P. Generalized zero-shot learning via synthesized examples. In: Proceedings of the IEEE conference on computer vision and pattern recognition. 2018. p. 4281–4289.
39. Carreira J, Zisserman A. Quo vadis, action recognition? A new model and the kinetics dataset. In: proceedings of the IEEE conference on computer vision and pattern recognition. 2017. p. 6299–6308.
40. Xiang H, Xie C, Zeng T, Yang Y. Multi-knowledge fusion for new feature generation in generalized zero-shot learning. 2021. arXiv:2102.11566.
41. Kuehne H, Jhuang H, Garrote E, Poggio T, Serre T, HMDB: a large video database for human motion recognition. In: International conference on computer vision. IEEE. 2011. p. 2556–63.
42. Soomro K, Zamir AR, Shah M. UCF101: a dataset of 101 human actions classes from videos in the wild. 2012. arXiv:1212.0402.
43. Szegedy C, Liu W, Jia Y, Sermanet P, Reed S, Anguelov D, et al. Going deeper with convolutions. In: Proceedings of the IEEE conference on computer vision and pattern recognition. 2015. p. 1–9.
44. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. 2016. p. 770–778.
45. Huang K, Miralles-Pechuán L, Mckeever S. Combining text and image knowledge with GANs for zero-shot action recognition in videos. In: VISIGRAPP (5: VISAPP). 2022. p. 623–631.