2024-1

Generalised Zero-Shot Learning for Action Recognition fusing text and image GANs

Kaiqiang Huang
Technological University Dublin, d14122793@mytudublin.ie

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

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

Recommended Citation

K. Huang, S. Mckeever and L. Miralles-Pechuán, "Generalised Zero-Shot Learning for Action Recognition fusing text and image GANs," in IEEE Access, doi: 10.1109/ACCESS.2024.3349510.

Funder: Technological University Dublin

IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier 10.1109/ACCESS.2023.0322000

Generalised Zero-Shot Learning for Action Recognition fusing text and image GANs

KAIQIANG HUANG[1], SUSAN MCKEEVER [1] (Senior Member, IEEE), and LUIS MIRALLES-PECHUÁN[1]

[1] Technological University Dublin, School of Computer Science, Grangegorman, Dublin, Ireland

Corresponding author: Luis Miralles-Pechuán (e-mail: luis.miralles@tudublin.ie).

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

ABSTRACT Generalized Zero-Shot Action Recognition (GZSAR) is geared towards recognizing classes that the model has not been trained on, while still maintaining robust performance on the familiar, trained classes. This approach mitigates the need for an extensive amount of labeled training data and enhances the efficient utilization of available datasets. The main contribution of this paper is a novel approach for GZSAR that combines the power of two Generative Adversarial Networks (GANs). One GAN is responsible for generating embeddings from visual representations, while the other GAN focuses on generating embeddings from textual representations. These generated embeddings are fused, with the selection of the maximum value from each array that represents the embeddings, and this fused data is then utilized to train a GZSAR classifier in a supervised manner.

This framework also incorporates a feature refinement component and an out-of-distribution detector to mitigate the domain shift problem between seen and unseen classes. In our experiments, notable improvements were observed. On the UCF101 benchmark dataset, we achieved a 7.43% increase in performance, rising from 50.93% (utilizing images and Word2Vec alone) to 54.71% with the implementation of two GANs. Additionally, on the HMDB51 dataset, we saw a 7.06% improvement, advancing from 36.11% using Text and Word2Vec to 38.66% with the dual-GAN approach. These results underscore the efficacy of our dual-GAN framework in enhancing GZSAR performance. The rest of the paper shows the main contributions to the field of GZSAR and highlights the potential and future lines of research in this exciting area.

INDEX TERMS Generalized Zero-Shot Action Recognition, Generalised Zero-Shot Learning, Generative Adversarial Networks, Human Action Recognition.

I. INTRODUCTION

Artificial General Intelligence (AGI) aims to develop AI systems able to adapt to Machine Learning (ML) models to multiple domains without specific training, emulating the human way of learning [1]. Generalized Zero-Shot Action Recognition (GZSAR) plays an important role in ML by allowing models to classify instances of unseen classes that were not present in the training phase, while still detecting seen classes [2], [3]. This capability is particularly valuable in real-world scenarios where it is impractical and too expensive to acquire labeled data for all possible classes. There have been many approaches to mitigate the problem of training classifiers with data scarcity. For example, using the GANs Synthesis for Oversampling (GANSO) method [4].

GZSAR supports the detection of unseen classes by leveraging the knowledge of relationships between classes, enabling more efficient utilization of available data. Some examples of GZSAR practical applications are recommendation systems [5], fraud detection [6], and medical diagnosis [7]. Moreover, GZSAR enhances the flexibility and adaptability of ML models to handle new classes as they emerge, which is something that frequently happens in many domains such as natural language processing (e.g. detecting new words or sentences) and computer vision (e.g. detecting new animals, objects, or activities) [2], [3].

Traditional GZSAR approaches rely on visual data which can be challenging when there is a scarcity of it. Elhoseiny et al. [8] was the first one to create a methodology to avoid relying on visual data by creating a framework based on textual descriptions. They used solely textual embeddings to train the classifiers to train the associations between textual descriptions and the representations of the images. However, the most promising approaches for achieving higher accuracy in GZSAR are based on Generative Adversarial

Networks (GANs). GANs generate both visual and textual representations from unseen class labels called embeddings. GANs need to be trained using the labels (inputs) and the visual and textual representations of seen classes (outputs) [3], [9]. Embeddings are low-dimensional representations of data samples in the form of arrays with numerical values.

Improving the performance of GZSAR is crucial for several reasons. Firstly, compiling an exhaustive list of all possible action classes is infeasible in realistic settings. Action classes refer to categories or labels that represent different types of actions performed by humans. For example, the class *jumping* represents the action of individuals propelling themselves off the ground using both feet. GZSAR enables models to adapt to new action classes without extensive retraining or fine-tuning [10], [11]. Secondly, the increasing number of classes makes collecting labeled data for each category more challenging because models have to distinguish among more classes some of which can be quite similar. GZSAR models can recognize new classes without the need for additional labeled data [10], [11]. Thirdly, GZSAR reduces the dependency on large volumes of labeled data, resulting in more efficient training processes and decreased computational resource demands [10], [11]. Lastly, GZSAR models exhibit enhanced adaptability to new domains or contexts, increasing their versatility and robustness [10], [11].

GZSAR faces several significant challenges, leading to potential downsides in achieving high levels of accuracy, particularly in unseen classes. The accuracy in the gold standard datasets UCF101 and HMDB51 using GZSAR using the popular framework TF-VAEGAN [12] remains 37.6% and 50.9% respectively as can be seen in subsection V-E, which is still too low for most real-life applications. One of the significant challenges in GZSAR is mitigating the data imbalance issues that arise when dealing with seen and unseen action classes. The prediction results often exhibit bias towards the seen classes, since only seen data are used during the training phase [2], [3], [13].

A major concern is the training complexity and high computational costs associated with GZSAR frameworks. Training such models requires substantial computational power and involves tuning numerous hyper-parameters [2]. In contrast to GANs complex methods, approaches like the one of Romera-Paredes and Torr [14] keep things straightforward. They use a basic linear transformation to map visual features and class attributes, making it computationally efficient. However, accuracy can be compromised. Nevertheless, this simplicity has potential in GZSAR by simplifying training and reducing computational demands.

Furthermore, GZSAR often relies on external sources, such as pre-trained image models (e.g., ResNet101 or GoogLeNet), to generate visual representations of unseen action classes. Resnet101 is a deep convolutional neural network trained on the ImageNet dataset and can be used to extract features from images that are useful to train the classifiers. These models had been used to generate features from images and boost the performance in GZSL [15]. While these external sources can provide valuable information, they also introduce and propagate their limitations and biases into the framework which can hinder the performance of GZSAR [16].

In summary, GZSAR allows models to recognize both seen and unseen actions simultaneously. However, its classification performance faces significant challenges, limiting its practical utility. These obstacles are mainly issues related to data imbalance, model biases, and domain shifts. For example, in team sports [17], actions often involve complex interactions among multiple players, presenting a substantial challenge for automated applications aimed at analysis and explanation generation. Furthermore, GZAR applied to videos is even more difficult due to data scarcity, the need to recognize numerous distinct actions, variations in video lighting and angles, and the necessity to grasp temporal information [18].

The primary aim of this paper is to significantly enhance the performance of Generalized Zero-Shot Action Recognition (GZSAR) models in both known and novel action classes during the testing phase, as highlighted in previous research [2], [3]. This task presents a formidable challenge in practical terms because it necessitates the ability of the model to recognize a diverse array of actions, including those that were not encountered during the training process.

We propose a novel approach to address this challenge named dual-GAN, as detailed in our recent work [3]. The dual-GAN method leverages the power of two Generative Adversarial Networks (GANs) to create visual representations of previously unseen action classes. One GAN is specifically tailored to harness textual information, utilizing class-label texts, while the other GAN capitalizes on visual information, specifically utilizing weights from convolutional neural networks associated with images sourced from Google Images that are relevant to the previously unseen action classes. These two GANs work synergistically to generate comprehensive visual representations, which are subsequently fused and employed to train a classifier dedicated to GZSAR classification.

In addition to the dual-GAN architecture, our framework incorporates two crucial components to further refine the feature representations and enhance model performance. The first component, known as the Feature Refinement Component (FRC), is adopted to elevate the quality and discriminative properties of the features generated by the GANs. This approach is inspired by the work of Chen et al. [13], who have demonstrated its effectiveness in improving feature quality. The second component is an innovative out-of-distribution detector, which is designed to distinguish between features generated by the GANs that correspond to seen and unseen classes. This detector, developed by Mandal et al. [10], plays a pivotal role in the initial classification between known and unknown classes, contributing to the overall robustness of the model.

The main contributions and novelties of our paper can be summarized in the following points:

- **Dual-GAN Framework:** The paper introduces a novel dual-GAN approach for Generalized Zero-Shot Action Recognition (GZSAR) that leverages both textual and visual representations to generate embeddings for action classes. While the constituent elements of our methodology have been developed previously, it is worth noting that our work represents a pioneering effort in the literature. Specifically, this is the first instance in which features from two disparate sources have been systematically combined to tackle the challenges inherent to GZSAR.

- **Enhanced Classification:** By fusing embeddings from two GANs and incorporating an FRC, the framework improves the quality of embeddings and enhances the classification of both seen and unseen action classes. It also includes an innovative out-of-distribution detector to differentiate features.
- **Performance Improvement:** Experimental results on benchmark datasets (UCF101 and HMDB51) show substantial performance gains, with a 53.03% increase in UCF101 and a 10.56% improvement in HMDB51 over baseline results, demonstrating the effectiveness of the proposed dual-GAN framework in GZSAR applications.

II. RELATED WORK

This section presents several contributions proposing interesting approaches to address some of the current challenges in GZSL, including the out-of-distribution (OOD) detector, and the FR method. It also raises two research questions regarding the effectiveness of the FR component in mitigating bias in GZSAR and the potential use of enriched semantic embeddings to improve GZSL performance.

One of the primary challenges in GZSL is the occurrence of biased predictions and poor performance on unseen classes due to domain shifts between seen and unseen classes, as highlighted by Liu et al. [19]. The domain shift problem occurs when the data from the classes used in the training is very different in terms of patterns or ranges from the classes used in the testing set (the unseen classes) [13]. In this context, it's crucial to note that classes can be predicted as similar by assessing their proximity in a semantic space. The semantic space can be created using attributes or side information of the seen classes [20]. To illustrate, consider a scenario where seen classes pertain to animals like pigs, sheep, and cats. The model will learn certain visual patterns. But if we want to detect a Tiger with white and orange stripes the model is not designed for detecting these stripes. The model most likely will classify the tiger as a cat because of its similarities in body structure.

Lambert et al. [21] have mitigated the problem of domain shift and biased predictions by implementing an approach based on transferring knowledge from attributes. For example, in the case of animal classification, if the animal has wings, or four legs, or a certain color it can be more likely to be one type than the other.

Another frequent issue in GZSL, identified by Mandal et al. [10], is the bias towards seen action classes in learned classifiers. This bias leads to misclassifications of unseen category samples belonging to one of the seen action classes. The authors introduce an OOD detector to tackle this problem. This detector determines whether video instances belong to a seen or unseen action category. The OOD detector is trained using GANs, which synthesize video features for unseen action classes based on the features of seen action categories. The paper by D. Krueger et al. [22] offers a solution to detect OOD data. The authors employ a risk extrapolation method to estimate and effectively manage the uncertainty encountered by the model when it encounters unfamiliar data, enhancing its ability to differentiate between known and unknown samples.

Furthermore, Chen et al. [13] propose a Feature Refinement (FR) method to address the domain shift challenge in GZSL for image classification. The FR method focuses on refining the features of both seen and unseen classes by incorporating semantic-to-visual mapping into a unified generative model. This approach employs an FR module that refines the visual features for both seen and unseen class samples. The authors in their work [13] suggest an enhancement of class-specific and semantically relevant representations within the FR module. They achieve this by introducing a dual-component approach, comprising the Self-Adaptive Margin Center Loss (SAMC-Loss) and a Semantic Cycle-Consistency Loss. This comprehensive feature refinement process is achieved by concatenating the generated features within the FR module.

Based on the literature review, two main research questions can be posed:

1) Can the fusion of GAN-based text and image embeddings enhance GZSAR model accuracy on the UCF101 and HMDB51 gold standard datasets?
2) Does the FR component introduced by Chen et al. [13] effectively address bias concerns in GZSAR, given its focus on image-related aspects? In contrast to Chen et al.'s approach, which solely employs the label word2vec as the semantic embedding in the image domain, our approach encompasses both text and image information. It may be worthwhile to investigate more enriched semantic embeddings, as suggested by Huang et al. [3], including class descriptions, to potentially enhance GZSL performance.

III. METHODOLOGY

In this section, we present a detailed explanation of our novel dual-GAN framework. This framework merges two GANs that create both visual and textual embeddings. We begin by defining the approach and then delve into the TF-VAEGAN framework. Furthermore, we provide an in-depth discussion of the Feature Refinement (FR) component, which enhances the quality of visual representations, and the Out-of-Distribution (OOD) detector, which effectively distinguishes between seen and unseen classes.

Our approach builds upon the VAEGAN model developed by Narayan et al., which is known for generating high-quality visual representations for previously unseen classes [12]. Narayan's model, based on the TF-Vaegan model with an OOD detector, has demonstrated remarkable performance in this field. The primary aim of our dual-GAN approach is to enhance this framework by combining visual and textual representations sourced from two different knowledge bases.

We followed previous studies training and testing approaches [10], [12] to make the experiments comparable. we divided each dataset into 30 parts, with each part having a mix of classes. We made sure that there were 51 seen and 50 unseen classes for UCF101 and 26 seen and 25 unseen classes for HMDB51 in each part. We decided which classes were for training and which were for testing.

A. THE PROPOSED DUAL-GAN FRAMEWORK

As suggested by Mandal et al. [10], we utilize the off-the-shelf Inflated 3D ConvNet (I3D) model for extracting visual features to obtain real visual representations of seen classes. The I3D model extends a 2-dimensional Convolutional Network to 3 dimensions to enable it to process videos. The I3D extracts features from videos and is commonly applied to action recognition in videos [23].

In line with the approach proposed by Huang et al. [3] shown in Figure 1, we generate two types of semantic embeddings, namely text-based and image-based, which are used to condition the VAEGAN. For the HMDB51 dataset, we select the class label Word2Vec embeddings along with the extracted visual features of collected images from ResNet as semantic embeddings. Similarly, for the UCF101 dataset, we choose the description label Word2Vec embeddings and the extracted visual features of collected images from ResNet as semantic embeddings. We adopt the maximum method, wherein the larger value at each position between the two synthesized visual representations is selected to fuse the embeddings.

In Figure 1, the dual-GAN approach is showcased, consisting of two key stages.

Stage 1: The primary objective of this stage is to generate visual representations for unseen classes by leveraging semantic embeddings obtained from two distinct knowledge sources: text-VAEGAN for textual data and image-VAEGAN for visual content. Subsequently, these two sets of unseen class embeddings, one from text and the other from images, are combined using the maximum value. This fusion process results in a novel dataset that comprises the original visual representations of known classes alongside artificially generated visual representations for previously unseen classes, each accompanied by their respective labels.

Stage 2: In this stage, the focus shifts towards training a classifier in a supervised manner using the dataset created in the preceding step. It is noteworthy that the generator of each GAN component is exclusively trained on known data, specifically video instances and their associated labels. Interestingly, the VAEGAN components exhibit the capability to synthesize semantically meaningful visual representations, even when provided solely with semantic embeddings, without access to video instances from unseen classes.

B. FR COMPONENT IN THE TF-VAEGAN FRAMEWORK

For the proposed methodology, the FR component is integrated into the TF-VAEGAN framework proposed by Narayan et al. [12] as illustrated in Figure 2. FR aims to enhance discriminative visual representations for both real and synthesized seen action videos, optimizing them using the Self-Adaptive Margin Center (SAMC) loss [13] and a semantic cycle consistency loss [24]. The SAMC loss encourages FR to learn more discriminative visual features relevant to the classes, promoting intra-class compactness and inter-class separability. Applied to the intermediate encoded features, the loss enhances the distinctiveness of features in the shallower layers of the FR component.

In Figure 2, the detailed process of generating high-quality visual representations for known classes (x) and their corresponding semantic embeddings (a), which can be either text-based or image-based class representations, is illustrated. These paired inputs are then directed to the encoder (E), where a concise latent code (z) is generated through optimization via the Kullback–Leibler divergence. Simultaneously, random noise and semantic embeddings (a) are employed as inputs for the generator (G), which is responsible for the synthesis of visual representations (x'). Subsequently, a comparison between the generated (x') and real (x) visual representations is enabled through binary cross-entropy loss. The pivotal role of distinguishing between real and synthesized visual representations is undertaken by the discriminator (D), utilizing the Wasserstein GAN $(WGAN)$ loss. Furthermore, two crucial components, namely the semantic embedding decoder (SED) and the feedback module (F), collaborate to refine the synthesis of visual representations and mitigate ambiguities in zero-shot classification tasks.

Either real (x) or synthesized (x') visual representations are taken as input by the SED, which then reconstructs corresponding semantic embeddings (a'), learning through a cycle-consistency loss. The feedback module (F) operates by transforming the latent embedding from the SED and feeding it back into the latent representation of the generator (G), thereby refining the synthesized visual representations (x'). It is noteworthy that while the generator (G) transforms semantic embeddings into visual representations, the SED performs the reverse operation by converting visual representations into semantic embeddings. Consequently, supplementary information is provided by these two components that significantly enhance the quality of visual representation synthesis, while ambiguity and misclassification issues among different action classes are concurrently reduced.

Following the FR component, the visual features of both seen and unseen classes are refined, with the refined features used for training the classifier and the refined real unseen features used for testing. As the FR process involves transforming high-dimensional features into low-dimensional ones, there is a loss of some discriminative information.

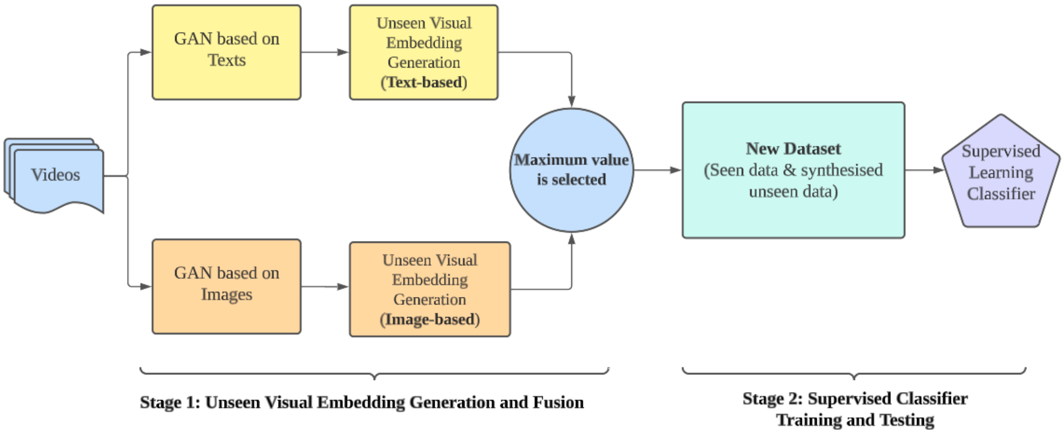

Videos
GAN based on Texts
Unseen Visual Embedding Generation (Text-based)
GAN based on Images
Unseen Visual Embedding Generation (Image-based)
Maximum value is selected
New Dataset (Seen data & synthesised unseen data)
Supervised Learning Classifier
Stage 1: Unseen Visual Embedding Generation and Fusion
Stage 2: Supervised Classifier Training and Testing

FIGURE 1: Architecture of the proposed dual-GAN approach for GZSAR using the TF-VAEGAN framework with the FR component.

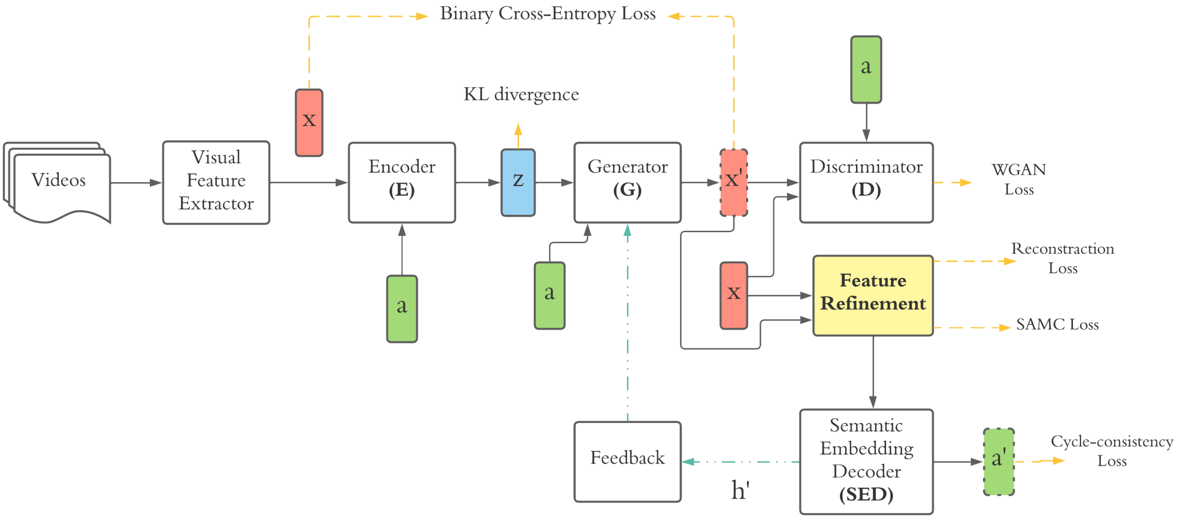


FIGURE 2: Architecture of TF-VAEGAN framework with FR component [12], [13]

The fully refined features are obtained by concatenating the real and synthesized visual features with latent and semantic embeddings to address this issue.

Compared to the framework described in Huang et al. [3], the main difference lies in the inclusion of the FR component to improve the quality of real and synthesized visual features, leading to enhanced GZSAR performance. No further adjustments or modifications are required to combine semantic embeddings from different sources within the framework.

C. OUT-OF-DISTRIBUTION DETECTOR

The OOD component, short for Out-of-Distribution detector, plays a crucial role in identifying classes that lie beyond the scope of the training data. Mandal et al. [10] were pioneers in introducing the concept of an OOD detector within the context of Generalized Zero-Shot Learning (GZSL) for video-based action recognition.

The OOD detector relies on video features generated by Generative Adversarial Networks (GANs), which are trained using features from known classes. Its primary function is to discriminate between features that belong to either known or unknown classes.

This OOD detector is implemented as a fully connected neural network, with an output layer having a dimension equal to the number of known classes. During training, it utilizes real features from the known classes and synthetic features representing the unknown classes. This incorporation of synthetic samples from the unknown classes enables the model to develop more accurate boundaries between the two categories, without making any assumptions about prior data distributions. Essentially, it aids in achieving a better separation between known and unknown classes.

In practical usage during the inference phase, a pre-trained model extracts real features from a test video, which are subsequently fed into the trained OOD detector. If the output falls below a predetermined threshold, the feature is directed to the classifier designed for known classes to predict the label of the test video. Conversely, if the value exceeds the threshold, the test video's label is predicted using the classifier tailored for unknown classes. This intelligent OOD detector effectively ensures a clear distinction in the classification process between known and unknown classes, thereby helping to mitigate any bias that may arise towards known classes.

IV. EXPERIMENTS

This section presents the results of the experiments to compare state-of-the-art approaches on both HMDB51 and UCF101 with the proposed approach with our approach based on fusing information from a GAN based on text and a GAN based on images to improve the accuracy of GZSL models. Additionally, it explores the impact of the FR component on the accuracy of the models. The complexity of O(E * M), where E is the number of training epochs and M is the time it takes to process one mini-batch of data.

The experiments were conducted in a powerful computing system designed for high CPU-consuming tasks for AI and ML. The machine contains a double NVIDIA Tesla V100 16GB PCIe (Volta architecture) GPU. Each GPU has 5,120 CUDA cores and 640 Tensor Cores. The code for the experiments has been made available [1].

A. DATASETS

For the experiments, the same datasets and train/test splits as the literature [10], [12] are utilized. The model is trained and evaluated using the same splits as previous related works [10], [12]. Each dataset consists of 30 independent splits, randomly generated while maintaining a consistent proportion of seen and unseen classes. For UCF101, there are 51 seen and 50 unseen classes, and for HMDB51, there are 26 seen and 25 unseen classes. In each split, each class is predefined as either a seen or unseen class. The model is trained only with seen classes, and both seen and unseen classes are tested under the GZSL setting.

B. EXPERIMENTS CONFIGURATION AND BASELINE

The experiments are mainly designed to test if by fusing using the maximum value, the information of a GAN based on text with a GAN based on images the accuracy of the GSZAR models increases. Experiments also examine the impact of FR on the GZSAR. Table 1 outlines the configurations for all the experiments; which dataset, knowledge source, how the embeddings are generated, and if the FR component is applied. The label Word2Vec is employed as a text-based semantic embedding without incorporating the FR component to establish the baselines for the UCF101 and HMDB51 datasets. Additionally, the image-based semantic embedding (i.e. ResNet-101) is considered. Subsequently, the dual semantic embedding approach - i.e. combining both text and image data by a maximum combination method - is evaluated in order to assess the influence of FR in the context of multiple knowledge sources.

In terms of training and testing approach, we followed that of previous studies [10], [12]. We divided each dataset into 30 parts, with each part having a mix of classes. We made sure that there were 51 seen and 50 unseen classes for UCF101 and 26 seen and 25 unseen classes for HMDB51 in each part. We decided which classes were for training and which were for testing.

The threshold value for the OOD detector in this study was drawn from Mandal et al. [10]. The authors of this paper conducted a series of experiments employing cross-validation, wherein various threshold values were assessed. Ultimately, they selected the threshold value that yielded the highest performance for GZSL.

It is worth noting that the experimentation process, even when utilizing high-powered computing resources, necessitated several days to complete as illustrated in Table 1. Given the substantial time and computational resources required for these experiments, fine-tuning or optimizing this threshold value fell beyond the primary focus of the current paper.

In the original paper introducing the OOD detector [10], the generator is implemented as a three-layer fully connected (FC) network, with an output layer having the same dimension as the video features, and the hidden layers set at 4096 units. Similarly, the decoder also employs a three-layer FC network, with an output size matching the class-embedding dimension and hidden layers containing 4096 units. The discriminator, on the other hand, consists of a two-layer FC network with an output size of 1 and hidden layers comprising 4096 units. As for the OD detector, it employs a three-layer FC network with output and hidden layer sizes aligned with the number of seen classes and set to 512, respectively. In our experiments, we maintain these network configurations for generators, decoders, seen/unseen classifiers, and OOD detectors. However, it is worth noting that these networks are relatively shallow, consisting of no more than three layers, which is a relatively small number compared to other models like the various ResNet versions with 18, 34, 50, 101, and 152 layers [25]. This design choice may impact model performance, particularly in terms of reducing the size of the hidden layers. Changing the network configuration is beyond the scope of this paper.

C. EVALUATION METRIC

In the GZSL setting, the evaluation process is not only for unseen classes but also includes seen classes. The harmonic mean (H) based on the work [2] is considered. After calculating the average per-class top-1 accuracy on seen and unseen classes, the harmonic mean (H) for the seen and accuracies of the unseen classes is calculated as:

[1] https://github.com/kaiqiangh/kg_gnn_gan/tree/gzsar

TABLE 1: Experimental configurations for both UCF101 and HMDB51. Experiments 1 and 7 are the baseline for the two datasets, respectively. It can be seen that these experiments take a very long time.

Exp	Dataset	Knowledge Source	Semantic Embedding	FR used?	Time per split
1*	UCF101	Text	Label Word2Vec	No	22 h
2				Yes	
3		Image	ResNet-101	No	24 h
4				Yes	
5		Text + Image	Max. Dual	No	48 h
6				Yes	
7*	HMDB51	Text	Label Word2Vec	No	14h
8				Yes	
9		Image	ResNet-101	No	15 h
10				Yes	
11		Text + Image	Max. Dual	No	30 h
12				Yes	

$$H = \frac{2 * Acc_{seen} * Acc_{unseen}}{Acc_{seen} + Acc_{unseen}} \quad (1)$$

We selected to measure the experiments using the harmonic mean metric because it is commonly used in the GZSL domain for model evaluation e.g., [10], [12], [13], [26], [27]. The harmonic mean balances the model performance on both seen and unseen classes and also makes sure the model will not be biased towards either seen or unseen class

D. IMPLEMENTATION

The implementation of our framework follows a similar approach to the studies by Huang et al. [3], [28]. We use fully connected networks for the discriminator D, encoder E, and generator G, each consisting of two layers with 4,096 hidden units. The semantic embedding decoder SED and feedback module F have the same structures as D, E, and G. We employ the Leaky ReLU activation function for all layers except the output of G. The output of G uses a sigmoid activation function to compute the binary cross-entropy loss [3], [28].

In contrast, the discriminator denoted as D is designed as a two-layer FC network, with an output dimension of 1 and a hidden layer dimension of 4096. Both the classifiers for the seen and unseen classes are implemented as single-layer FC networks, where the input dimension corresponds to the size of the video features. Their respective output dimensions are contingent upon the number of classes for the seen and unseen categories. For the out-of-distribution detector (ODD), it adopts a three-layer FC architecture. The dimensions of its output and hidden layers are contingent upon the quantity of seen classes and are fixed at 512, respectively.

For training the framework, we utilize the Adam optimizer with a learning rate of 10^{-4}. In accordance with Xian et al. [29], α represents the weighting coefficient for the WGAN loss, β is a hyperparameter used to weight the decoder reconstruction error in SED, and σ is employed in the feedback module F to regulate the feedback modulation. We adopt the same hyperparameters as outlined in Narayan et al. [12], where α, β, and σ are set to 10, 0.01, and 1, respectively.

During GAN training, we set the gradient penalty coefficient λ to a fixed value of 10 [3], [28]. Additionally, the feature refinement (FR) module is implemented as a multi-layer perceptron (MLP) with two hidden layers, each consisting of 4,096 units activated by Leaky ReLU. We set the parameters of the SAMC and semantic cycle consistency loss functions to 0.5 and 0.001, respectively [3], [28].

As suggested by Huang et al. [3], we synthesize 600 features, and we adopt the maximum method for combining visual representations from different semantic sources. Furthermore, following Mandal et al. [10], the OOD detector is implemented as a three-layer fully connected network, with output and hidden layer sizes equivalent to the number of seen classes and 512, respectively. The threshold value used in the OOD detector is determined by calculating the mean of the prediction entropies of the seen class features present in the training data.

V. RESULTS & ANALYSIS

In the following section, we analyze the experimental results of our GZSAR model. We explore the impact of different semantic sources and fine-tuning (FR) on two datasets, HMDB51 and UCF101. Additionally, we compare our model's performance to state-of-the-art approaches to highlight its effectiveness.

A. PERFORMANCE EVALUATION

Figure 3 illustrates a comparative analysis of the seen, unseen, and harmonic mean accuracies on the HMDB51 dataset, utilizing the class label Word2Vec as the semantic embedding and investigating the impact of FR on performance. In Experiment 1, the seen accuracy exhibits fluctuations but generally shows an increasing trend over epochs, with a notable improvement observed after Epoch 15. Similarly, the unseen accuracy displays an increasing trend, albeit with occasional fluctuations. The harmonic mean demonstrates a generally increasing trend, accompanied by some fluctuations across epochs. In Experiment 2, the seen accuracy demonstrates a mild increasing trend over epochs despite fluctuations. Conversely, the unseen accuracy follows a clearer increasing trend, particularly after epoch 20. The harmonic mean exhibits an overall increasing trend with fewer fluctuations compared to Experiment 1.

A comparison of experiments 1 and 2 reveals that the seen accuracy exhibits a similar trend in both experiments with no significant difference between them. However, the unseen accuracy in Experiment 2, which utilizes FR, demonstrates a more evident increasing trend compared to Experiment 1. Furthermore, the harmonic mean in Experiment 2 is generally higher and steadier, especially after epoch 20. These results support the hypothesis that leveraging FR can improve the quality of visual features for both seen and unseen classes. Specifically, the consistent improvements in the unseen accuracy and harmonic mean in Experiment 2 compared to Experiment 1, which does not employ FR, provide evidence in support of this hypothesis.

Furthermore, a comparison of seen, unseen, and harmonic mean accuracies is shown in Figure 5, using the combined representation (i.e., text + image) as semantic embedding, whether the component of FR is leveraged on the HMDB51 dataset in Experiment 5 and 6. In general, Experiment 5 has higher seen accuracy across the epochs, while Experiment 6 demonstrates better accuracy for unseen classes, which is similar to the findings from Experiments 1-4. The hypothesis that using FR (Experiment 6) can improve the quality of visual features for seen and unseen classes seems to hold, particularly for unseen classes. The higher harmonic mean in Experiment 6 also indicates better overall performance. Additionally, a summary of experimental results on the HMDB51 is shown in Figure 6.

It is observed that Experiment 8 yields better accuracy for unseen classes while maintaining similar seen accuracy compared to Experiment 7. Similarly, Experiment 10 consistently outperforms Experiment 9 in terms of unseen accuracy while maintaining similar seen accuracy, thereby resulting in higher harmonic mean values. On the other hand, the last comparison shows that both experiments have comparable unseen accuracies, but Experiment 12 demonstrates better accuracy for seen classes, which leads to higher harmonic mean values. These findings across multiple datasets further support the hypothesis that leveraging FR can enhance the performance of GZSAR models, particularly in terms of unseen accuracy and harmonic mean, by improving the quality of visual features for both seen and unseen classes.

B. EXPERIMENTS DISCUSSION

Figure 4 presents a comparison of seen, unseen, and harmonic mean accuracies, utilizing image-based representation (i.e., ResNet) as semantic embedding, and considering whether the component of FR is leveraged on the HMDB51 dataset in Experiment 3 and 4. In general, Experiment 3 has a marginally higher seen accuracy (51.01%) compared to Experiment 4 (49.84%). However, Experiment 4 demonstrates better performance in unseen accuracy (23.26% vs 21.90%) and harmonic mean (29.34% vs 28.79%), indicating the advantage of using FR. It is obvious that FR contributes to better performance: Experiment 4, which employs FR, consistently yields higher accuracies for unseen classes as well as harmonic mean values, compared to Experiment 3. This observation confirms the hypothesis that FR can improve the quality of visual features for both seen and unseen classes. Moreover, while Experiment 4 consistently outperforms Experiment 3 in terms of unseen accuracy and harmonic mean, Experiment 3 achieves higher seen accuracy in most of the epochs. This suggests that the FR might have a slight trade-off with seen class performance, but this trade-off is beneficial for the generalization to unseen classes.

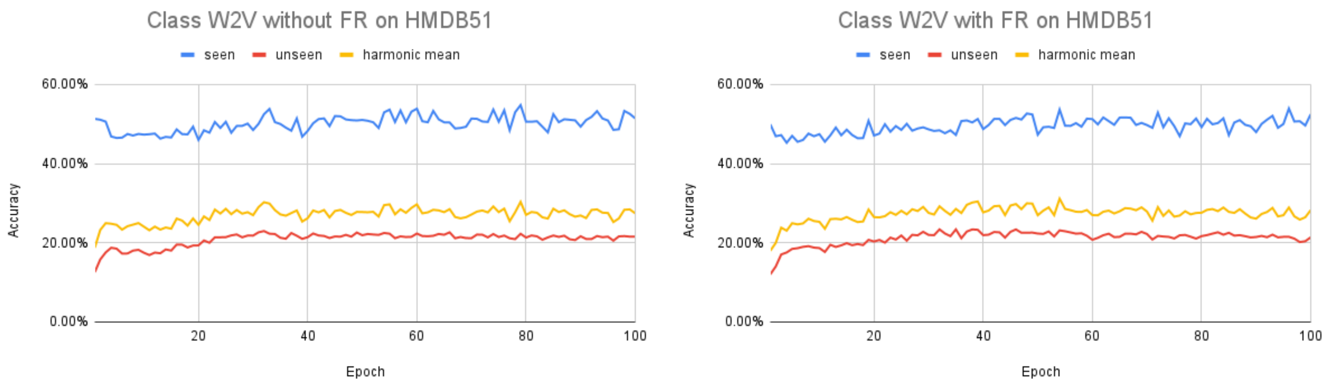
The UCF101 dataset is undergoing three separate comparisons, with and without FR, using different sources of semantic embedding such as text-based description, image-based representation, and combined semantic embeddings. The results of these comparisons are visually presented in Figure 7, Figure 8, and Figure 10.

Overall, it is observed that the models tend to overfit the seen data, as seen accuracy tends to be higher than the unseen accuracy. Nevertheless, the harmonic mean increases over epochs, indicating that the models are improving their performance on both seen and unseen data. However, fluctuations in the harmonic mean values are also noted, indicating that the models may not be learning at a consistent rate or may be sensitive to hyperparameter selection. Figure 11 provides a summary of the results obtained with and without FR and various semantic sources on UCF101, indicating that the highest harmonic mean (54.66%) is achieved using FR and combined semantic embedding.

As mentioned in previous papers [3], the videos comprising the HMDB51 dataset are primarily sourced from movies and YouTube, undergoing minimal modifications like video cropping and centering (Figure 9 shows a few examples.). In contrast, the UCF101 dataset draws heavily from YouTube videos but adheres to stricter video selection criteria, favoring videos with cleaner backgrounds and fewer actors. Furthermore, prior research, as highlighted in [10], [12] has consistently shown that ZSAR performance is notably lower when using the HMDB51 dataset compared to the UCF101 dataset, a trend that our own experimental results corroborate. Hence, we propose that ZSAR performance is closely linked to the clarity and focus of videos concerning their associated action labels.

C. SUMMARY OF THE RESULTS

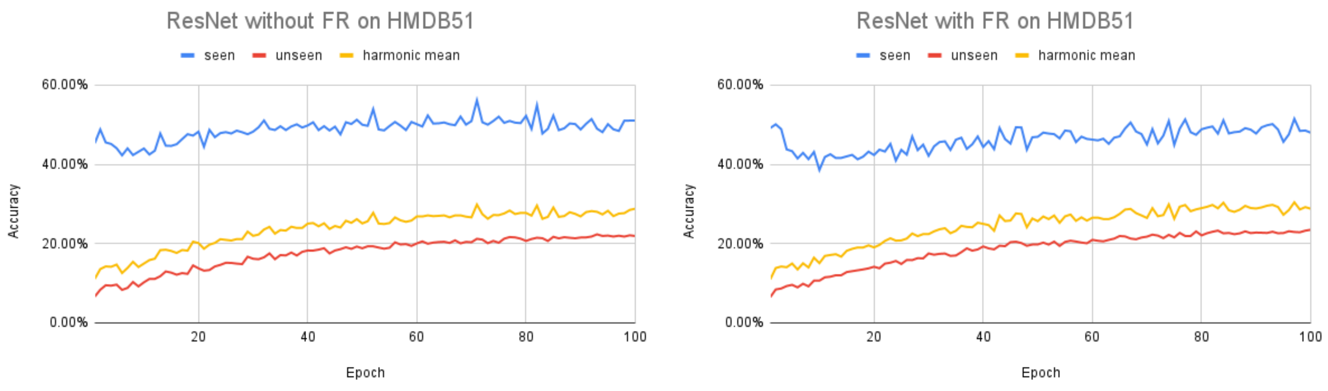
As shown in bold letters in Table 2, the experimental results demonstrated the effectiveness of the dual-GAN framework and its improvement when using the FR component. In the dataset HMDB51 the harmonic mean accuracy with 54.66% and in the UCF101 is 38.66%. Our approach achieved higher accuracy in classifying both seen and unseen action classes compared to baseline methods that used single-source semantic embeddings. The incorporation of the FR component further enhanced the performance, especially when using multiple semantic knowledge sources.



(a) Exp. 1: Class Word2Vec without FR

(b) Exp. 2: Class Word2Vec with FR

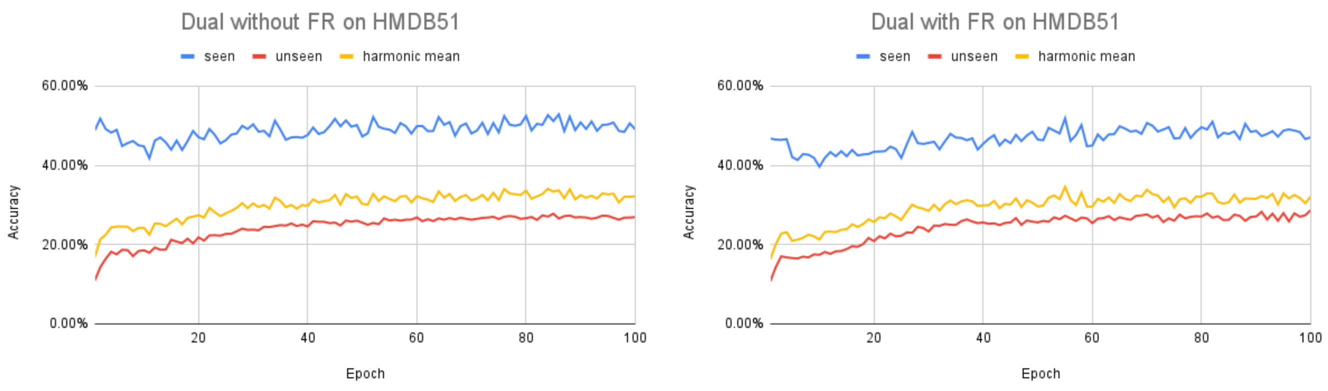
FIGURE 3: Exp. 1 & 2: GZSAR results using class Word2Vec with/without FR on HMDB51.



(a) Exp. 3: ResNet without FR.

(b) Exp. 4: ResNet with FR.

FIGURE 4: Exp. 3 & 4: GZSAR results using ResNet with/without FR on HMDB51.



(a) Exp. 5: Dual embeddings without FR

(b) Exp. 6: Dual embeddings with FR

FIGURE 5: Exp. 5 & 6: GZSAR results of combined embeddings with/without FR on HMDB51.

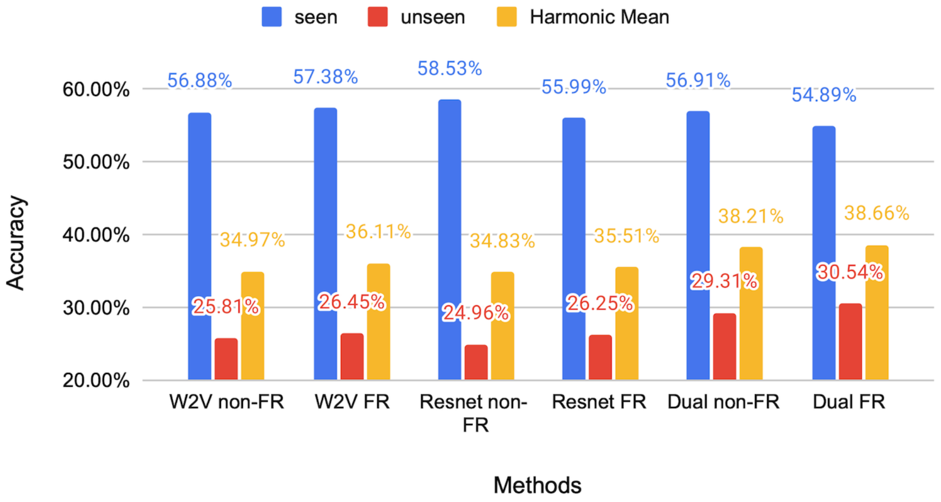
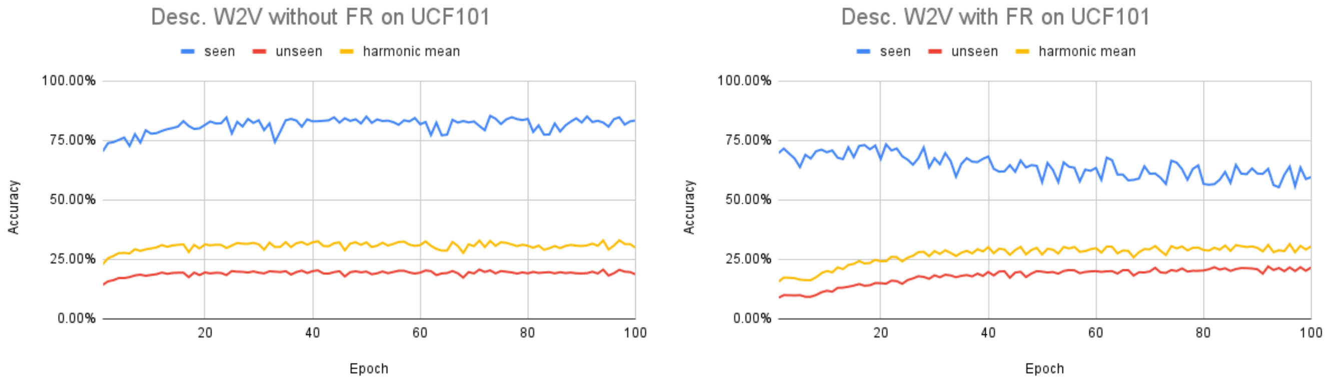


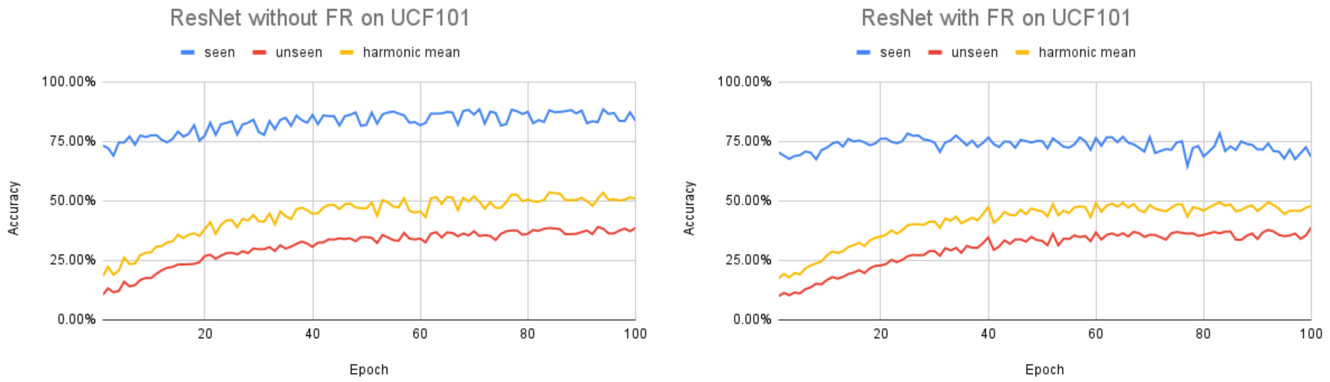
FIGURE 6: Experiments 1-6: Results with/without using FR and various semantic sources on HMDB51.



(a) Exp. 7: Description of Word2Vec without FR.

(b) Exp. 8: Description Word2Vec with FR.

FIGURE 7: Exp. 7 & 8: GZSAR results using description word2Vec with/without FR on UCF101.



(a) Exp. 9: ResNet without FR.

(b) Exp. 10: ResNet with FR.

FIGURE 8: Exp. 9 & 10: GZSAR results using ResNet with/without FR on UCF101.

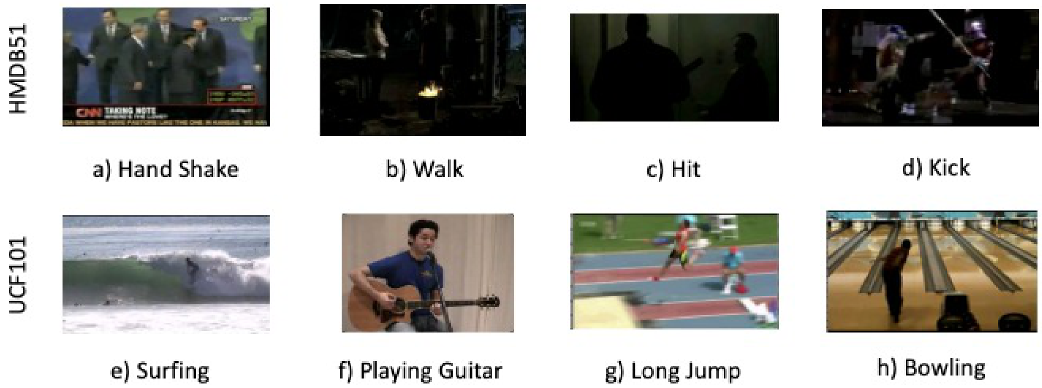
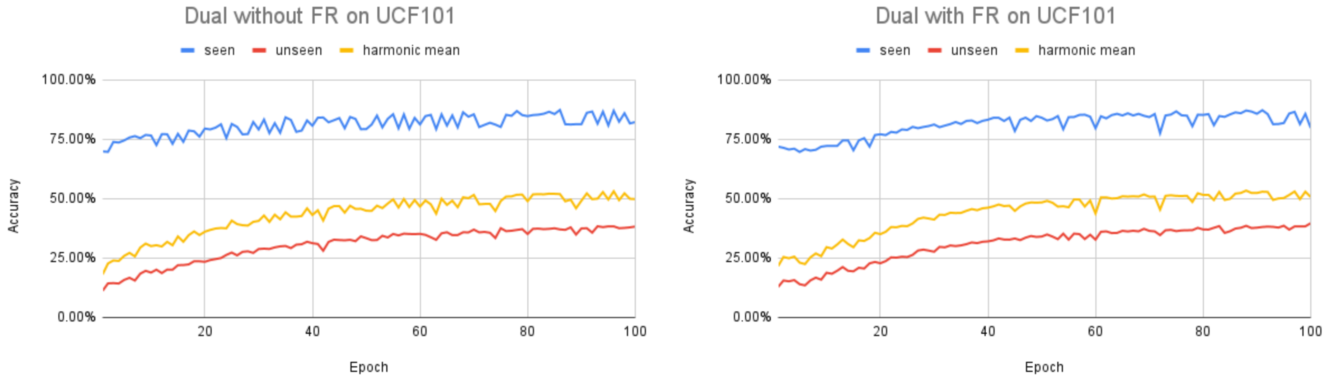


FIGURE 9: As can be observed the quality of the videos in the UCF101 dataset is better than that of the HMDB51 dataset (blurry, poor contrast, and different objects in the background) because they have been selected more carefully and meeting some standard criteria.



(a) Exp. 11: Dual embeddings without FR.

(b) Exp. 12: Dual embeddings with FR.

FIGURE 10: Experiments 11 & 12: GZSAR results using combined embeddings with/without FR on UCF101.

TABLE 2: Summary of the best results for the harmonic mean of seen and unseen accuracy using various types of semantic embedding on both UCF101 and HMDB51.

Datasets	Semantic Embedding	FR Used?	Seen (Avg & STD)	Unseen (Avg & STD)	Harmonic Mean (Avg & STD)
UCF101	Text: Label word2Vec	No	79.00% ± 0.3330	23.84% ± 0.0319	35.70% ± 0.1266
	Text: Label word2Vec	Yes	75.25% ± 0.3717	39.07% ± 0.0472	47.99% ± 0.2069
	Image: Label word2Vec	No	75.39% ± 0.2256	24.68% ± 0.0353	37.04% ± 0.0924
	Image: Label word2Vec	Yes	77.20% ± 0.2138	39.07% ± 0.0369	50.93% ± 0.1260
	Dual: Text + Image	No	83.54% ± 0.3518	39.02% ± 0.0594	52.21% ± 0.1965
	Dual: Text + Image	Yes	**86.08% ± 0.3433**	**40.17% ± 0.0796**	**54.71% ± 0.1974**
HMDB51	Text: Label word2Vec	No	56.88% ± 0.0990	25.81% ± 0.0277	34.97% ± 0.0501
	Text: Label word2Vec	Yes	57.38% ± 0.0348	26.45% ± 0.0265	36.11% ± 0.0250
	Image: Label word2Vec	No	**58.53% ± 0.0316**	24.96% ± 0.0362	34.83% ± 0.0367
	Image: Label word2Vec	Yes	55.99% ± 0.0390	26.25% ± 0.0393	35.51% ± 0.0373
	Dual: Text + Image	No	56.91% ± 0.0994	29.31% ± 0.0284	38.21% ± 0.0544
	Dual: Text + Image	Yes	54.89% ± 0.1003	**30.54% ± 0.0304**	**38.66% ± 0.0593**

The best results for semantic embedding techniques on the UCF101 and HMDB51 datasets. On UCF101, the Dual: Text + Image approach with face recognition achieves the highest performance, with 85.98% seen accuracy, 40.15% unseen accuracy, and a harmonic mean of 54.66%. Similarly, on HMDB51, the Dual: Text + Image approach with face recognition also outperforms other techniques, yielding a higher harmonic mean of 38.66% compared to 34.83% achieved by the Image: Label word2Vec approach without face recognition. These findings emphasize the effectiveness

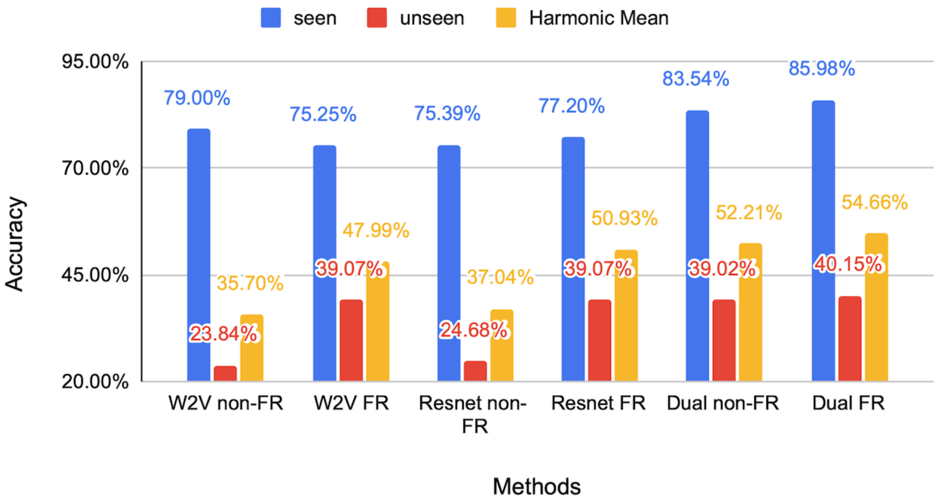


FIGURE 11: Experiments 7-12: Results with/without using FR and various semantic sources on UCF101.

of combining text and image modalities, along with the utilization of face recognition, for achieving superior results in action recognition tasks across both datasets.

In Figure 12, we have presented a visual representation of examples from two datasets, UCF101 and HMDB51. The visualization showcases both correctly and incorrectly classified samples, where the instances with incorrect predictions are highlighted in red, and those with correct predictions are enclosed within green frames.

Upon visual inspection, it becomes apparent that the misclassified instances, marked in red, are often images and video frames that pose a significant challenge to human recognition. On the other hand, the correctly classified instances, denoted by green frames, are notably clearer and more discernible.

It is important to note that due to space constraints, we could not display complete sequences of images and videos in this paper. Nevertheless, this illustration demonstrates how image quality significantly influences the performance of classifiers in predictive tasks.

D. STATISTICAL TESTS

Table 3 displays the statistical significance comparison of various methods using p-values. The table exhibits p-values calculated for different combinations of methods. Each cell represents the result of a two-sample t-test. The methods listed in the header are abbreviated as follows: "W2v Free" for Word2vect using Free, "W2v" for Word2vect not using Free, "Resnet Free" for using Resnet and Free, "Resnet" for Resnet and not using Ffree, "Dual Free" for Dual using two sources and Free, and "Dual" using two sources but not using Free.

The asterisk (*) indicates p-values below 0.05, which implies statistically significant differences between the corresponding methods. When the p-values are below 0.05, the null hypothesis, which assumes no difference between the methods, can be rejected in favor of the alternative hypothesis, indicating a significant difference.

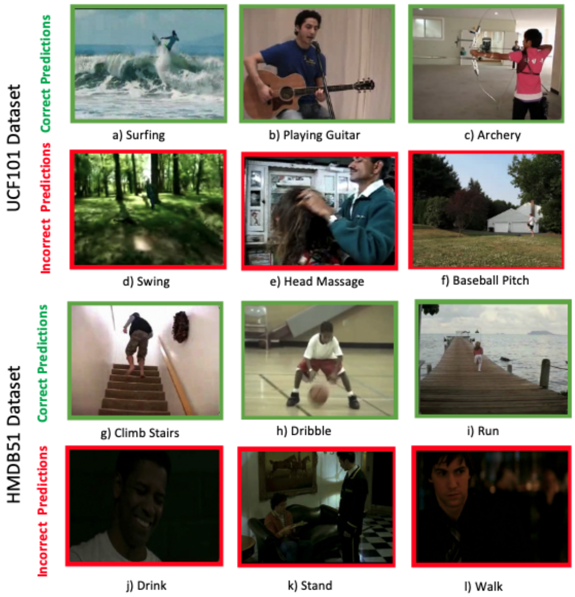


FIGURE 12: Visualization of Classifications in UCF101 and HMDB51 Datasets: Distinguishing Correct and Incorrect Predictions.

Table 3 and 3 show the statistical significance of the disparities between the methods. It is important to analyze the performance differences among the methods in our experiments to validate the effectiveness of our Dual-GAN approach.

TABLE 3: Pairwise comparison of statistical significance using p-values in the HMDB51 Dataset.

	W2v	W2v Free	Resnet	Resnet Free	Dual	Dual Free
W2v	-	0.2557*	0.9479	0.4577*	0.0195*	0.0118*
W2v Free	0.2557*	-	0.1300*	0.7267	0.0640*	0.0371*
Resnet	0.9479	0.1300*	-	0.3418*	0.0075*	0.0045*
Resnet Free	0.4577*	0.7267	0.3418*	-	0.0590*	0.0351*
Dual	0.0195*	0.0640*	0.0075*	0.0590*	-	0.7636
Dual Free	0.0118*	0.0371*	0.0045*	0.0351*	0.7636	-

The p-values in Table 3 reveal a significant contrast between using "Free" and "Non-Free". For example, when comparing "W2v" to "W2v Free," a p-value of 0.2557* indicates a statistically significant difference, implying that the "Free" version performs differently. The same thing shows the p-value of 0.3418* between "Resnet" and "Resnet Free". However, this does not happen between "Dual Free" and using "Dual". Probably because the classification algorithms now have two different sources and the classes are much more spread in terms of their values.

Both the "Dual" and "Dual Free" methods stand out prominently in Table 3. They exhibits statistically significant differences when compared to all other methods, with p-values consistently below 0.05, except when compared between themselves, the p-value is not significant (0.7636), indicating a similarity in performance between these two variants. Overall, the statistical analysis shows that the proposed Dual-GAN approach has significant differences and is better in terms of performance than the other approaches.

TABLE 4: Pairwise comparison of statistical significance using p-values in the UCF101 Dataset.

	W2v	W2v Free	Resnet	Resnet Free	Dual	Dual Free
W2v	-	0.2314*	0.0056*	0.0000*	0.0026*	0.0050*
W2v Free	0.2314*	-	0.0285*	0.0000*	0.0149*	0.0270*
Resnet	0.0056*	0.0285*	-	0.1218*	0.9001	0.9593
Resnet Free	0.0000*	0.0000*	0.1218*	-	0.1456*	0.0963*
Dual	0.0026*	0.0149*	0.9001	0.1456*	-	0.8563
Dual Free	0.0050*	0.0270*	0.9593	0.0963*	0.8563	-

In Table 4, the distinction between "Free" and "Non-Free" methods remains evident. It can be observed that there is a statistical difference in results in the p-value of 0.2314 when comparing Word2vec with and without Free. And the p-value for the ResNet approach using and not using Free is 0.1218. The differences in this approach when using Dual-GAN are not significant.

The "Dual" method continues to exhibit substantial differences in performance compared to the rest of the methods in the updated table. Significantly low p-values (e.g., 0.0026* with "W2v," 0.0149* with "W2v Free") indicate that the "Dual" method performs distinctly in comparison. However, it's important to note that when compared to "Dual Free," the p-value is not significant (0.8563), suggesting that these two variants have similar performance. The same happens with comparing both "Dual" and "Dual Free" with Resnet, the value is 0.9001 (not far from 0.5) but the difference is still not significant enough. The conclusion is that these results reinforce the uniqueness of the "Dual" and "Dual Free" methods and highlight the importance of their selection in this context.

E. STATE OF THE ART COMPARISON

The highest accuracy results for seen, unseen, and harmonic mean are extracted for each experiment on both datasets to evaluate the capacity of the proposed model, as demonstrated in Table 2. It is observed that the proposed method achieves promising results, outperforming the existing methods reported in the literature. The greatest improvement is achieved by combining the embeddings of a GAN based on text with one based on images.

Specifically, the proposed method achieves an optimal harmonic mean of 54.66% and 38.66% for UCF101 and HMDB51, respectively, as presented in Table 5. This outcome has significant implications for various practical applications, including human-computer interaction, surveillance, and robotics, where reliable and accurate action recognition is of utmost importance. It is worth noting that the improved performance of the proposed method is attributed to its ability to leverage various sources of semantic embeddings, such as text-based descriptions and image-based representations, along with FR techniques. The effectiveness of the model in this regard is evidenced by the observed increase in the harmonic mean over epochs, indicating that the model is being learned and its performance is being improved on both seen and unseen data.

VI. CONCLUSION

In summary, this paper addressed the challenge of recognizing actions in a GZSAR setting. We introduced a new approach called dual-GAN, which combines information from two GANs. This approach aims to use textual descriptions or class labels and images from Google Images to create visual representations of actions that have never been seen before. These representations are then used to train a classifier for GZSAR.

We also introduced an FR component with an out-of-distribution detector to handle the domain shift between seen and unseen classes and mitigate biases in the trained model. The FRC improves the quality and discriminative properties of the visual features generated by the GANs, while the OOD detector helps distinguish between seen and unseen action categories.

The proposed dual-GAN framework contributes to the field of GZSAR by addressing the challenges of data imbalance, biases, and domain shift. It enables models to recognize both seen and unseen action classes simultaneously, reducing the need for extensive retraining or fine-tuning when new classes emerge. Our approach also reduces the dependency on large volumes of labeled data, making the training process more efficient and computationally feasible.

In conclusion, our study demonstrates the effectiveness of the dual-GAN framework, especially when combined with face recognition (FR) components, for improving action recognition in the UCF101 and HMDB51 datasets. The results presented in Table 2 clearly show that the Dual: Text + Image approach with FR achieves the highest performance, with a remarkable harmonic mean of 54.66% on UCF101 and 38.66% on HMDB51. These results outperform existing approaches in the literature, as highlighted in Table 5. The incorporation of both text-based and image-based semantic embeddings, along with FR, significantly enhances the model's ability to classify both seen and unseen action classes. Moreover, our statistical analysis (Tables 3 and 4) confirms the significance of the differences between our pro-

TABLE 5: Comparison of GZSAR performance (harmonic mean) among the best results and the existing approaches in the literature for both HMDB51 and UCF101. The improvement goes from 50.93% (using image and word2vec) to 54.66% using two GANs) and an from 36.11% (using Text and word2vec to 38.66% (using two GANs) in the HMDB51.

Methods / Datasets	Gaussian Mixture Model [30]	Classification-Loss Wasserstein GAN [26]	CEWGAN [10]	f-VAEGAN [29]	TF-VAEGAN [12]	dual-FR-VAEGAN (Proposed model)
HMDB51	20.1%	32.7%	36.1%	35.6%	37.6%	**38.66%**
UCF101	17.5%	44.4%	49.4%	47.2%	50.9%	**54.66%**

posed Dual-GAN approach and other methods, emphasizing its uniqueness and superiority in action recognition tasks. Overall, our findings underscore the potential of combining multiple semantic knowledge sources and FR for advancing the state of the art in GZSL.

In future work, we aim to explore additional semantic knowledge sources and investigate the generalizability of our approach to other domains beyond action recognition. We also plan to evaluate the framework on larger and more diverse datasets to further validate its effectiveness and robustness.

The potential applications of GZSAR are vast, including recommendation systems, fraud detection, and medical diagnosis. By improving the performance of GZSAR, our framework opens up new opportunities for utilizing ML models in real-world scenarios where it is difficult to collect training data for all possible classes.

REFERENCES

[1] B. Goertzel, "Artificial general intelligence: concept, state of the art, and future prospects," *Journal of Artificial General Intelligence*, vol. 5, no. 1, p. 1, 2014.

[2] Y. Xian, B. Schiele, and Z. Akata, "Zero-shot learning-the good, the bad and the ugly," in *Proceedings of the IEEE Conference on CVPR*, 2017, pp. 4582–4591.

[3] K. Huang, L. Miralles-Pechuán, and S. Mckeever, "Enhancing zero-shot action recognition in videos by combining gans with text and images," *SN Computer Science*, vol. 4, no. 4, p. 375, 2023.

[4] A. Salazar, L. Vergara, and G. Safont, "Generative adversarial networks and markov random fields for oversampling very small training sets," *Expert Systems with Applications*, vol. 163, p. 113819, 2021.

[5] H. Ding, Y. Ma, A. Deoras, Y. Wang, and H. Wang, "Zero-shot recommender systems," *arXiv preprint arXiv:2105.08318*, 2021.

[6] G. Kwon and G. Al Regib, "A gating model for bias calibration in generalized zero-shot learning," *IEEE Transactions on Image Processing*, 2022.

[7] D. Mahapatra, B. Bozorgtabar, and Z. Ge, "Medical image classification using generalized zero shot learning," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 3344–3353.

[8] M. Elhoseiny, B. Saleh, and A. Elgammal, "Write a classifier: Zero-shot learning using purely textual descriptions," in *Proceedings of the IEEE International Conference on Computer Vision*, 2013, pp. 2584–2591.

[9] H. Liu, L. Yao, Q. Zheng, M. Luo, H. Zhao, and Y. Lyu, "Dual-stream generative adversarial networks for distributionally robust zero-shot learning," *Information Sciences*, vol. 519, pp. 407–422, 2020.

[10] 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 CVPR*, 2019, pp. 9985–9993.

[11] H. Xiang, C. Xie, T. Zeng, and Y. Yang, "Multi-knowledge fusion for new feature generation in generalized zero-shot learning," *arXiv preprint arXiv:2102.11566*, 2021.

[12] S. Narayan, A. Gupta, F. S. Khan, C. G. Snoek, and L. Shao, "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, pp. 479–495.

[13] S. Chen, W. Wang, B. Xia, Q. Peng, X. You, F. Zheng, and L. Shao, "Free: Feature refinement for generalized zero-shot learning," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 122–131.

[14] 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.

[15] Y. Xian, C. H. Lampert, B. Schiele, and Z. Akata, "Zero-shot learning—a comprehensive evaluation of the good, the bad and the ugly," *IEEE transactions on pattern analysis and machine intelligence*, vol. 41, no. 9, pp. 2251–2265, 2018.

[16] F. Pourpanah, M. Abdar, Y. Luo, X. Zhou, R. Wang, C. P. Lim, X.-Z. Wang, and Q. J. Wu, "A review of generalized zero-shot learning methods," *IEEE transactions on pattern analysis and machine intelligence*, 2022.

[17] F. Wu, Q. Wang, J. Bian, N. Ding, F. Lu, J. Cheng, D. Dou, and H. Xiong, "A survey on video action recognition in sports: Datasets, methods and applications," *IEEE Transactions on Multimedia*, 2022.

[18] V. Estevam, H. Pedrini, and D. Menotti, "Zero-shot action recognition in videos: A survey," *Neurocomputing*, vol. 439, pp. 159–175, 2021.

[19] S. Liu, M. Long, J. Wang, and M. I. Jordan, "Generalized zero-shot learning with deep calibration network," *Advances in neural information processing systems*, vol. 31, 2018.

[20] Z. Zhang and V. Saligrama, "Zero-shot learning via semantic similarity embedding," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 4166–4174.

[21] C. H. Lampert, H. Nickisch, and S. Harmeling, "Learning to detect unseen object classes by between-class attribute transfer," in *2009 IEEE conference on computer vision and pattern recognition*. IEEE, 2009, pp. 951–958.

[22] D. Krueger, E. Caballero, J.-H. Jacobsen, A. Zhang, J. Binas, D. Zhang, R. Le Priol, and A. Courville, "Out-of-distribution generalization via risk extrapolation (rex)," in *International Conference on Machine Learning*. PMLR, 2021, pp. 5815–5826.

[23] Q. Wu, A. Zhu, R. Cui, T. Wang, F. Hu, Y. Bao, and H. Snoussi, "Pose-guided inflated 3d convnet for action recognition in videos," *Signal Processing: Image Communication*, vol. 91, p. 116098, 2021.

[24] R. Felix, I. Reid, G. Carneiro *et al.*, "Multi-modal cycle-consistent generalized zero-shot learning," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 21–37.

[25] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of CVPR*, 2016, pp. 770–778.

[26] Y. Xian, T. Lorenz, B. Schiele, and Z. Akata, "Feature generating networks for zero-shot learning," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 5542–5551.

[27] O.-B. Mercea, L. Riesch, A. Koepke, and Z. Akata, "Audio-visual generalised zero-shot learning with cross-modal attention and language," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 10 553–10 563.

[28] K. Huang, L. Miralles-Pechuán, and S. Mckeever, "Zero-shot action recognition with knowledge enhanced generative adversarial networks," in *In Proceedings of the 13th International Joint Conference on Computational Intelligence*, 2021, pp. 254–264.

[29] Y. Xian, S. Sharma, B. Schiele, and Z. Akata, "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, pp. 10 275–10 284.

[30] 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 WACV*. IEEE, 2018, pp. 372–380.

KAIQIANG HUANG is a dedicated researcher specializing in computer vision, particularly in the area of human action recognition. With a primary focus on leveraging zero-shot learning techniques, he aims to advance the state-of-the-art in this field. Currently pursuing his PhD at Technological University Dublin, he is actively involved in developing innovative computer vision algorithms to enable machines to accurately recognize and comprehend human actions in visual data. Before embarking on his PhD journey, Kaiqiang obtained a Master's in Data Analytics from the School of Computing at Technological University Dublin. This educational background has provided him with a solid foundation in data analysis and computational techniques, equipping him with the necessary skills to tackle complex challenges within computer vision research. Kaiqiang Huang's research demonstrates his passion for pushing the boundaries of computer vision, particularly in the realm of human action recognition. His dedication and expertise make him a valuable contributor to the field, and he continues to strive for advancements that enhance machines' ability to understand and interpret human actions in visual content.

DR. LUIS MIRALLES-PECHUÁN is an assistant lecturer at TU Dublin. Upon completing his PhD, Luis continued his research journey as a postdoctoral researcher at CeADAR, UCD. During this period, he showcased his expertise by publishing at the Digital Forensic conference and was honoured with the Best Student Paper award. Currently, his primary interest lies in the application of reinforcement learning to combat the COVID-19 pandemic, specifically in devising strategies that strike a balance between public health and economic considerations. Luis Miralles-Pechuán's dedication to academic research and his focus on utilizing machine learning and reinforcement learning techniques in various domains demonstrate his commitment to addressing complex real-world challenges. With his expertise and passion for interdisciplinary research, he continues to make valuable contributions to the field while actively working towards innovative solutions with societal impact.

DR. SUSAN MCKEEVER holds the position of senior lecturer at the School of Computer Science within Technological University Dublin. Her research expertise lies in the field of applied machine learning, with a strong focus on harnessing technology for the betterment of society. Her specific areas of interest include health, accessibility, and smart cities, where she strives to leverage cutting-edge technologies to bring about positive societal impact. Dr. McKeever has made significant contributions to these domains and has a distinguished record of publications that reflect her commitment to advancing knowledge and understanding in her field. Her work embodies the intersection of technology and social benefit, paving the way for transformative applications that improve the lives of individuals and communities.