IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 26 May 2024, accepted 27 June 2024, date of publication 9 July 2024, date of current version 18 July 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3425547

RESEARCH ARTICLE

Multi-Label Zero-Shot Learning With Adversarial and Variational Techniques

MUQADDAS GULL[1] **AND OMAR ARIF**[1,2], **(Senior Member, IEEE)**
[1]School of Electrical Engineering and Computer Science, National University of Sciences and Technology (NUST), Islamabad 44000, Pakistan
[2]Department of Computer Science and Engineering, American University of Sharjah, Sharjah, United Arab Emirates

Corresponding author: Muqaddas Gull (Mgull.dphd18seecs@seecs.edu.pk)

This work was supported in part by the Open Access Program from American University of Sharjah.

ABSTRACT Multi-label zero-shot learning expands upon the traditional single-label zero-shot learning paradigm by addressing the challenge of accurately classifying images containing multiple unseen classes, which are not part of the training data. Current techniques rely on attention mechanisms to tackle the complexities of multi-label zero-shot learning (ZSL) and generalized zero-shot learning (GZSL). However, the generation of features, especially within the context of a generative approach, remains an unexplored area. In this paper, we propose a generative approach that leverages the capabilities of Conditional Variational Autoencoder (CVAE) and Conditional Generative Adversarial Network (CGAN) to enhance the quality of generative data for both multi-label ZSL and GZSL. Additionally, we introduce a novel "Regressor" as a supplementary tool to improve the reconstruction of visual features. This Regressor operates in conjunction with a "cycle-consistency loss" to ensure that the generated features preserve the key qualities of the original features even after undergoing transformations. To gauge the efficacy of our proposed approach, we conducted comprehensive experiments on two widely recognized benchmark datasets: NUS-WIDE and MS COCO. Our evaluation spanned both multi-label ZSL and GZSL scenarios. Notably, our approach yielded significant enhancements in mean Average Precision (mAP) for both datasets. Specifically, we observed a 0.2% increase in performance on the NUS-WIDE dataset and a notable 2.6% improvement on the MS COCO dataset in the context of Multi-label ZSL. The results clearly demonstrate that our generative approach outperforms existing methods on these widely-recognized datasets.

INDEX TERMS Conditional variational autoencoder, conditional generative adversarial network, generalized zero-shot learning, regressor, zero-shot learning.

I. INTRODUCTION

In the current era, deep learning models have achieved remarkable performance in various computer vision applications, including medical imaging [1], image classification [2], [3], object detection [4], [5], self-driving cars [6], and agriculture [7], among others. Single-label classification, which involves categorizing a single object in an image, has been extensively studied in image classification tasks. For this purpose, large datasets such as ImageNet 21K [8] and ImageNet 1K [9] have been compiled for evaluating deep learning models. However, it is important to note that natural images frequently feature multiple objects and concepts, underscoring the importance of multi-label classification. In multi-label classification, the objective is to independently classify multiple objects present in an image [10], [11], [12], [13], [14], [15], [16].

Recurrent neural networks [17], [18], attention mechanisms [19], [20], and label correlation [21], [22] have demonstrated significant success in the field of multi-label classification. However, despite these advancements, the challenge of multi-label zero-shot classification remains unresolved. This problem involves classifying images into new and unseen categories during testing, without any visual examples available during training [23], [24]. Multi-label ZSL can be viewed as an extension of multi-label classification. It tackles scenarios where the objective is to classify images into categories that are entirely new

The associate editor coordinating the review of this manuscript and approving it for publication was Yuan Gao.

and unseen during training. Moreover, a broader framework known as GZSL has emerged. In contrast to traditional ZSL, GZSL handles test images from both categories that were encountered during training (seen classes) and those that were not (unseen classes).

Single-label GZSL has garnered significant attention in recent years [24], [25], [26], [27], [28], [29], [30], [31]. These techniques heavily rely on generative models, i.e., Generative Adversarial Network (GAN) [32] and Variational Autoencoder (VAE) [33], to create novel features for classes that were not seen during training. Generative approaches have emerged as the dominant paradigm for single-label GZSL [25], [28], [30], [31], as they can generate synthetic features for unseen classes by learning the underlying feature distribution from the seen classes. However, the development of feature generators for multi-label Zero-Shot Learning (ZSL) and GZSL paradigms has received limited exploration. In this study, we tackle the challenging task of developing a feature generator for multi-label ZSL and GZSL classification.

In recent times, there has been a growing interest in adopting a generative approach for multi-label ZSL and GZSL. Earlier methods leveraged Generative Adversarial Network (GAN) [32] and Variational Autoencoders (VAE) [33] to generate visual features for unseen classes, utilizing side information [34], [35], [36], specifically in the context of ZSL and GZSL. But in the context of multi-label ZSL and GZSL, to the best of our knowledge, the proposed approach is generative in nature and has not been previously explored.

In our prior research [37], we presented a method for generating global image-level embeddings through an identifiable VAE-based generative network. This method involved incorporating two additional VAE networks to map individual modalities to their respective latent spaces. In contrast, our current study introduces an innovative model that combines a CVAE and CGAN, along with a Regressor Network. This extension to work in the context of Multi-label ZSL and GZSL is that real-world images often contain multiple objects and concepts, necessitating the need for Multi-label classification techniques. In contrast to single-label classification where each image is associated with a single class label, Multi-label classification allows images to be associated with multiple class labels simultaneously. Therefore, addressing Multi-label ZSL and GZSL involves handling the complexities of associating semantic information with each of these labels in a coherent manner.

The motivation behind this research lies in addressing the challenging problem of ZSL and GZSL. This research aimed to explore the generative approach that aims to exploit the connection between CVAE and CGAN to convert this conventional Multi-label ZSL problem to a supervised learning problem by generating visual features against all unseen classes using their class semantic attributes as side information. By combining the capabilities of CVAE and CGAN, the proposed generative approach aims to harness the advantages of both to improve the quality of the generated data. This combination leverages the structured latent space of CVAE and the adversarial training of CGAN to generate diverse and high-quality data samples while adhering to specific conditions or class labels. This is motivated by the idea that better-generated data can lead to improved recognition and classification of previously unseen classes. The introduction of a "Regressor" as an additional tool is motivated by the need to enhance the reconstruction of visual features. This Regressor works in conjunction with a "cycle-consistency loss," ensuring that the generated features maintain the essential characteristics of the original features even after transformations. This process aims to improve the overall quality and fidelity of the generated visual features. This innovative methodology transforms the Multi-label ZSL problem into a more familiar supervised learning setting by utilizing class semantic information as side information, ultimately advancing the field by achieving superior performance compared to existing state-of-the-art approaches on standard datasets.

Notably, the sole commonality between our previous and current work is the utilization of the Attribute-level feature fusion technique. This technique is employed to generate global image-level embedding of semantic information. These distinctions underscore the progression of our research and the distinctive contributions made in the present investigation.

However, the primary contributions of this paper are outlined as follows: 1) We have introduced an innovative generative approach that combines the capabilities of two distinct methods, namely CVAE and CGAN, to enhance the quality of generative data for Multi-label ZSL and GZSL.; 2) We have also introduced "Regressor" as an additional component, which ensures that the generated features retain the qualities of the original features even after transformations; 3) Inspired by [38], we leveraged attribute-level feature fusion technique to generate semantically consistent disentangled representations using the corresponding semantic information of the multi-label data, which in result will convert this conventional multi-label ZSL and GZSL problem to a conventional supervised learning problem; and 4) We evaluate our proposed model on two standard datasets: MS COCO [39] and NUS-WIDE [40]. The experimental results demonstrate that our approach achieves significantly better results for both multi-label ZSL and GZSL on both datasets.

II. RELATED WORK

In this section, we will review the literature related to multi-label ZSL and GZSL.

A. TRADITIONAL ZSL AND GZSL

ZSL was first introduced by [41], where attribute-based classification was performed to recognize unseen classes. Earlier works on ZSL mainly focused on learning a mapping function between visual and semantic features [42], [43], [44]. In the proposed work by [42], an attentive region embedding network, is comprised of two branches: the Attentive Region Embedding (ARE) and the Attentive Compressed Second-order Embedding (ACSE). Both ARE and ACSE employ their respective embeddings and map them to the semantic space to calculate the compatibility loss. Reference [45] also defines a compatibility function to perform a mapping from image visual features to its semantic space. They have introduced three types of semantic embeddings i.e., attributes, hierarchical embedding, and unsupervised word embeddings. Among these embeddings, unsupervised word embeddings have achieved good performance on all datasets. A direct method for learning a mapping function [44] employs a convex combination of class label embedding vectors, eliminating the need for additional training. To adopt the structure of the semantic space, an approach proposed by [46] utilizes the semantic relationships between categories. The classes in the given set are categorized into three distinct groups based on their characteristics: identical, semantically similar, and semantically dissimilar relative to a reference class.

Depending on the type of embedding space, conventional methods of ZSL are divided into three categories: (1) visual features space is selected as embedding space and the semantic features are projected to visual features space [47], [48]. To address the hubness problem, they have adopted the visual features space as an embedding space and mapped the semantic features onto these visual features space. They introduced three distinct modalities for semantic embedding, namely single modality, multiple modality, and RNN encoding. (2) Semantic space is considered as an embedding space and visual features are mapped to it [44]. (3) Both visual and semantic features are mapped to a common embedding space [43], [45], [49]. The work proposed by [43] introduces a two-branch framework aimed at achieving high intra-class similarity and low inter-class similarity. This is accomplished by learning a latent embedding space that simultaneously projects visual and semantic features into a joint embedding space. The embedding function that is responsible to perform any sort of mapping can be either linear [29], [41], [44] or nonlinear [50].

Recently, ZSL has been considered as a missing data problem, and the focus is on generating visual representations for unseen classes to perform classification [30], [31], [36], [51]. Generative models are utilized to learn probability distributions of a given dataset, enabling the generation of similar data. In scenarios where visual data is lacking, such as in ZSL, generative models offer a solution by generating the missing visual data based on the established correlation between semantic information and corresponding visual examples. Subsequently, the combination of existing visual data with the generated data is employed in a supervised learning approach. To address the missing data problem, recent methods use the following generative models i.e., Variational Autoencoders (VAE) [33] and Generative Adversarial Network (GAN) [52] for visual features generation. VAE [33] is a probabilistic model that involves a probabilistic Encoder E and a probabilistic Decoder D. GAN [52] is a generative model utilized for data generation. It comprises two fundamental elements: a generator and a Discriminator.

The f-CLSWGAN [31] is a generative model based on GAN, comprising a conditional generator and a conditional Discriminator. The core component of their proposed model is a Discriminator network trained on seen classes, aiming to minimize the classification loss. In [36] another GAN-based approach is proposed that incorporates semantic knowledge in the form of the knowledge graph. The f-VAEGAN [31] represents a generative model designed to address the challenges associated with ZSL and few-shot learning. The proposed framework integrates the strengths of both GAN and VAE networks, to generate the visual features against all unseen classes. The f-VAEGAN [31] operates effectively in both inductive and transductive settings.

GZSL is considered a more realistic and challenging task as compared to conventional ZSL. In GZSL, the test images can come from both seen and unseen classes. Many of the GZSL methods also learn a mapping function between visual and class-semantic features, where the nearest neighbor classifier is further used to perform classification [46], [53], [54]. Similarly, a popular approach for GZSL is to consider it as a missing data problem [3], [30], [31], [34], [51], [55], [56], [57]. In order to tackle the issue of missing data, contemporary approaches employ generative models such as VAE [33] and GAN [52] to generate visual features. A conditional VAE-based framework is proposed by [55] for visual features generation, where the task is to learn the underlying probability distribution of an image using its semantic information also as an input. A VAE-based method SE-GZSL [56] is composed of an Encoder and a Generator network, leveraging class attributes as semantic information to generate visual features for previously unseen classes. In addition, an attribute Regressor has been introduced to enhance the reconstruction quality of the generated data by offering feedback to the generator.

Another VAE-based generative model [34] uses distribution alignment and cross-reconstruction loss for latent space alignment. In another approach, [3] combines VAE and GAN network along with the Regressor to constrain the generated visual features back to their respective class-semantic information to further improve the features generation process. OCD-CVAE [58] introduces an overcomplete distribution by employing a conditional VAE for both seen and unseen classes. The primary aim of the overcomplete distribution is to enhance the network's generalizability by generating samples that exhibit greater proximity to other classes. To gain the advantages of both generative models, specifically the GAN and VAE, Zero-VAE-GAN [59] propose a joint generative model and employs attributes as class-semantic information to generate features. Boomerang-GAN [60] introduces a novel model utilizing CGAN. It generates unseen visual samples from semantic embeddings and introduces a cycle-consistent loss to translate these visual features back into semantic embeddings. Comprehensive experiments conducted on multiple datasets demonstrate that Boomerang-GAN surpasses previous state-of-the-art methods in both recognition and segmentation tasks within ZSL and Generalized ZSL settings.

In the context of GZSL, where test images can belong to both seen and unseen classes, generative models like VAE and GAN are employed to address the challenge of missing data by generating visual features for the unseen classes.

B. MULTI-LABEL CLASSIFICATION

Multi-label classification is a complex task that involves classifying multiple objects or concepts within a given image. This task is typically more challenging than standard single-label classification. The simplest approach for Multi-label classification is to train a binary classifier against each label present in the training data [12], [61]. Along with this to capture label correlation there are also few graph-based [16], [62], [63] and structure-based learning techniques [13], [64]. To establish connections among multiple labels, [62] utilizes a knowledge graph. Moreover, to adequately capture the interdependencies between seen and unseen class labels, they employ a label propagation mechanism within the semantic space. A Graph Convolutional Network (GCN) based Multi-label classification model [16], builds a directed graph over the object labels, where each label in the graph corresponds to word embeddings of a specific label. To facilitate the learning of structure and parameters, an integrated Bayesian framework rooted in the conditional graphical lasso (CGL) [63] has been devised. This framework presents an efficient approach for acquiring image-dependent label structures. The problem of Multi-label classification is addressed by utilizing a combined framework of recurrent neural network (RNN) and convolutional neural network (CNN) [13]. This framework enables the learning of a shared low-dimensional image-label embedding, effectively capturing the semantic relationship between labels and images. Despite previous efforts, a comprehensive framework was proposed [64] that effectively utilizes both semantic and spatial relationships among labels for Multi-label image classification. This approach incorporates image-level supervision to gain a thorough understanding of the underlying relations between labels.

Vision transformer-based techniques have garnered significant interest due to their exceptional ability to capture global dependencies [65], [66], [67]. In the context of image-label classification, [66] framework employs Transformers to leverage the intricate dependencies existing between labels and visual features. Reference [66] explores the viability of incorporating specialized transformer modules as a means to tackle inherent challenges encountered in CNNs. A novel architecture, termed Multi-label Transformer, has been devised, which integrates window partitioning, in-window pixel attention, and cross-window attention techniques to significantly enhance the performance of Multi-label image classification tasks. Reference [67] have introduced a two-stage framework, called Query2Label, designed for Multi-label classification tasks. In this framework, a label embedding is employed as a query to extract class-specific features from a feature map generated by a vision backbone. These extracted features are then utilized for subsequent binary classifications.

C. MULTI-LABEL ZSL AND GZSL

While successful in multi-label classification, existing models encounter challenges when confronted with unseen classes, thereby restricting their practical usability. ZSL methods, specifically, rely on class semantic information, such as attributes, to identify classes not encountered during training. The concept of attribute-based classification was initially introduced by [68]. Furthermore, [42], [53] aimed to develop a function that maps semantic and visual features for conventional ZSL, where the focus is solely on unseen classes. In the context of GZSL, a more realistic scenario compared to conventional ZSL, a test image may belong to either seen or unseen classes. Numerous GZSL approaches have emerged, aiming to learn a function for mapping both semantic and visual features [53], [69]. GZSL has been treated as a missing data problem, with recent methods incorporating GAN [32] and VAE [33] as generative models for visual feature generation [25], [30], [31]. However, it is crucial to note that existing ZSL and GZSL frameworks are constrained to single-label classification and lack effectiveness in handling multi-label ZSL and GZSL scenarios.

Extensions of multi-label classification, Multi-label ZSL, and GZSL can be considered as natural progressions. These methodologies center around aligning the visual embeddings of images with their corresponding label embeddings and establishing connections between labels, both seen and unseen. Various studies, including [3], [70], [71], delve into identifying label correlations as a means of classification, revealing relationships among different labels. Addressing the semantic diversity between labels and images, Semantic Diversity Learning (SDL) [72] pinpoints principal embedding vectors for images, assigning higher weights to samples with the greatest semantic diversity. Conversely, approaches like LESA [73] and BiAM [74] employ attention modules for multi-label classification, to pinpoint the presence of each label within an image. For the synthesis of multi-label features, GMLZL [38] introduces a GAN-based model that utilizes multi-class semantic information.

The Aligned Dual-modality Classifier (ADDS) [75] incorporates a soft constraint mechanism to effectively align textual and visual features, thereby improving generalization in multi-label classification. ADDS introduces a novel transformer decoder as the dual model decoder and employs Pyramid-Forwarding, a unique adaptation method. Additionally, for unbiased multi-label Zero-Shot Learning (ZSL), [76] addresses the training process of the ML-ZSL classifier. This is achieved by integrating class-specific region information through a channel attention mechanism, establishing a correlation between local and global information of the samples. Notably, this approach is non-generative, as the refined semantic features are mapped into a joint visual-label semantic embedding space. ML-Decoder [77] proposes an attention-based classification head to improve label prediction using queries. To ensure scalability for a large number of classes, they introduce a novel group-decoding scheme that efficiently handles a significant number of classes. These studies collectively contribute to the progress of multi-label ZSL and GZSL by employing a shared projection matrix to propagate information from seen classes to unseen classes, facilitating generalization.

III. METHODOLOGY

We begin by examining the baseline model for feature generation, followed by an explanation of the attribute-level feature fusion module responsible for producing semantically consistent fused multi-label visual features for all unseen classes.

A. PRELIMINARY

The problem formulation for multi-label ZSL and GZSL is outlined as follows. Let Y denote the set of class labels, which is partitioned into two sets as $Y = Y^{\mathrm{s}} \cup Y^{\mathrm{u}}$, and $Y^{\mathrm{s}} \cap Y^{\mathrm{u}} = \phi$, where Y^{s} represents the seen class labels present in the training data, and Y^{u} represents the unseen class labels. Here, $x \in X^{\mathrm{s}}$ is the set of encoded features of multi-label images, and $y \in \{0, 1\}^{\mathrm{s}}$ denotes the corresponding multi-hot labels from the set of seen class labels Y^{s}, with p positive classes present in the image. The category-specific class-level embeddings are utilized as side information and are represented as $a(k) = \{a(k_j), \forall j : y[j] = 1\}$, where $|a(k)| = p$. Given sets U and S, the objective of multi-label ZSL and GZSL is to learn classifiers $f_{zsl} : X \rightarrow \{0, 1\}^{\mathrm{U}}$ and $f_{gzsl} : X \rightarrow \{0, 1\}^{S+U}$ respectively. In ZSL, the exploration domain is confined to unseen classes only, whereas in GZSL, the exploration domain encompasses both seen and unseen classes.

B. FEATURE GENERATION

In this section, we commence the discussion with CVAE [78] and CGAN [32] as the base of our model. CVAE is a type of probabilistic model. It consists of two main parts: an Encoder and a Decoder. The Encoder $q_\phi(z|x, a(y))$ takes as input both an image's visual features x and class-semantic information $a(y)$. It then maps this input to a lower-dimensional space called a latent vector z. The Decoder $p_\theta(\tilde{x}|a(y), z)$ also called a Generator, takes the latent vector z and class-attributes $a(y)$ as input to generate image's visual features $\tilde{x}$. The objective function of CVAE is to minimize the following loss function:

$$\mathcal{L}_{CVAE} = \mathbb{E}_{q_\phi(z|a(y),x)}[logp_\theta(x|z, a(y))] - D_{KL}(q_\phi(z|x, a(y))||p_\theta(z|a(y))) \tag{1}$$

where on the R.H.S of equation 1, the first term represents reconstruction error, i.e., to minimize the difference between image visual features and reconstructed features and the subsequent term represents the Kullback-Leibler divergence between $q_\phi(z|a(y), x)$ and $p_\theta(z|a(y))$. Kullback-Leibler divergence serves as a quantitative measure to assess the dissimilarity between two given distributions.

CGAN is an extension of the traditional GAN that enhances the generative process by introducing conditional information. CGANs are particularly useful for generating data with specific attributes or characteristics, such as generating images based on class attributes, styles, or other conditions. A CGAN consists of two primary components: a conditional Generator G and a conditional Discriminator D. The conditional Generator $G(z, a(y))$ is a neural network that takes two inputs: a random noise vector z and class-attributes $a(y)$. The Generator task is to generate data samples $\tilde{x}$, that adhere to the provided class-attributes and follow the distribution of the real data. By incorporating the class attributes as input, the Generator can produce samples that belong to specific classes or exhibit desired characteristics. While the conditional Discriminator $D(x, a(y))$ is another neural network that takes as input both the class-attributes $a(y)$ and the visual feature x. Its role is to determine whether the given input pair corresponds to real visual features x from the actual dataset or whether it's a generated/fake visual feature $\tilde{x}$ produced by the Generator. The Discriminator's task is to distinguish between real and fake samples based on the provided class attributes.

The training of a CGAN involves a two-player minimax game, similar to the standard GAN framework. The Generator G aims to generate realistic visual features $\tilde{x}$ that can deceive the Discriminator D into believing they are real. It tries to create samples that closely resemble the distribution of real data while satisfying the provided class attributes. Simultaneously, the Discriminator D strives to correctly classify between real visual features x and fake/generated visual features $\tilde{x}$ while considering the associated class-attributes a. It is trained to improve its ability to distinguish between real and fake samples. The iterative process of training involves these two components competing against each other. As training progresses, the Generator becomes better at producing samples that align with the desired attributes, and the Discriminator improves its capability to differentiate between real and generated samples.

Mode collapse is a significant challenge in the realm of GAN, where the Generator becomes fixated on producing a narrow range of data samples, thereby lacking diversity and variation in its outputs. This phenomenon can lead to a suboptimal generative process, as the Generator fails to capture the entire complexity of the real data distribution. To overcome the mode collapse issue, we rely on the most stable training method i.e., Wasserstein GAN (WGAN) as it provides more stable and meaningful gradients during training, which can lead to improved generation quality. The objective function of the Wasserstein GAN with condition a

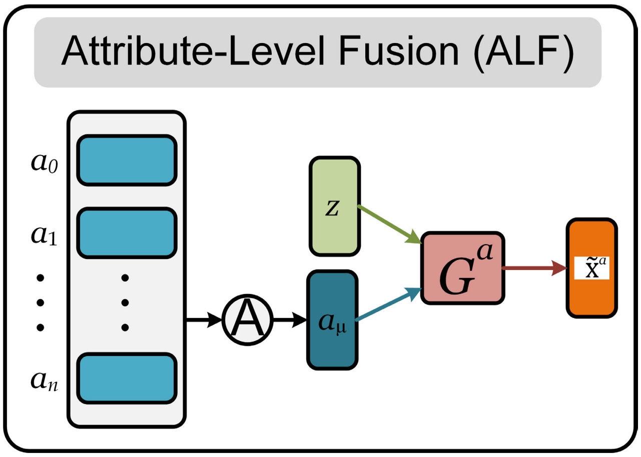


FIGURE 1. Overview of the attribute-level feature fusion process to generate global image-level embedding.

is as follows:

$$\mathcal{L}_{WGAN} = \mathbb{E}[D(x, a(y))] - \mathbb{E}[D(\tilde{x}, a(y))] - \lambda \mathbb{E}[(||\nabla_{\hat{x}} D(\hat{x}, a(y))||_2 - 1)^2] \quad (2)$$

where $\mathbb{E}[.]$ represents the expected value operator, which signifies taking the average value of the expression within the square brackets. While $D(x, a(y))$ and $D(\tilde{x}, a(y))$ represent the joint distribution of visual and semantic features. Specifically, $D(\tilde{x}, a(y))$ corresponds to the distribution of both visual and semantic features for unseen classes, while $D(x, a(y))$ represents the distribution for seen classes. Moreover, $\tilde{x} = G(z, a(y))$, $\hat{x} = \alpha x + (1 - \alpha)\tilde{x}$ serves as a mechanism for generating visual features, where α follows a uniform distribution $\alpha \sim U(0, 1)$. This operation aids in the synthesis of diverse visual attributes. The parameter λ stands as the penalty coefficient, which contributes to the overall optimization process.

C. MULTI-LABEL FEATURE GENERATOR

To produce multi-label features, we have applied a fusion method referred to as attribute-level fusion (ALF), as outlined in the work by [38].

1) ATTRIBUTE-LEVEL FEATURE FUSION

We utilize the attribute-level feature fusion approach as represented in Figure. 1. In the attribute-level fusion approach, the goal is to generate fused multi-label visual features for unseen classes while considering the inter dependencies among the image labels. This approach takes into account the class-semantic information associated with various labels present in the image. To derive a comprehensive image-level global visual features, the attribute-level feature fusion method combines the individual class-semantic information. This fusion process involves aggregating the semantic information associated with positive labels found in the image. One way to achieve this is by averaging the individual class-semantic information, denoted as $a(y_j)$, and the global image-level embedding a_μ is defined as:

$$a_\mu = \frac{1}{n} \sum_{j:y[j]=1} a(y_j), \quad (3)$$

We then integrate this attribute-level feature fusion module to our proposed generative model.

D. THE PROPOSED MODEL

It has been shown that combining CVAE and CGAN [3], [30], [59] approach to generate visually meaningful and semantically coherent features for classes that the model has not seen during training. As shown in Figure. 2 our proposed generative approach for ZSL and GZSL has an Encoder $E(x)\colon X \times A \rightarrow Z$, that Encode the input visual features of an image x along with global image-level embedding a_μ to a latent features space z. This global image-level embedding a_μ is generated by an attribute-level fusion module by utilizing class-level semantic information a(y). A Generator/Decoder $G(z, a_\mu)\colon Z \times A \rightarrow X$, takes the latent features vector z along with global image-level embedding a_μ, and produces the reconstructed visual features representation $\tilde{x}$. A Discriminator $D\colon X \times A \rightarrow R$ evaluates the compatibility between a pair of visual features x and global image-level embedding a_μ and maps them to a compatibility score. The optimization function of both CVAE and CGAN is as follows:

$$\mathcal{L}_{CVAEGAN} = \mathcal{L}_{CVAE} + \gamma \mathcal{L}_{WGAN} \quad (4)$$

where the Decoder of the CVAE and the Generator of the CGAN are designed to share the same set of parameters. This means that the same network architecture is used for both the Decoder and the Generator. This allows the model to combine the strengths of both CVAE and CGAN in generating high-quality visual features for unseen classes. As the CVAE loss involve a reconstruction loss and a regularization term. The reconstruction loss ensures that the generated features are similar to the input features, while the regularization term encourages the latent features to follow a specific probability distribution. While, the WGAN loss focuses on training the Generator and Discriminator in a way that they converge more stably. Instead of using traditional adversarial losses, it employs a Wasserstein distance metric that offers a more meaningful and stable measure of the difference between probability distributions. Standard GANs are known to be sensitive to the choice of hyperparameters and often suffer from issues such as mode collapse and training instability. In the context of CGAN, using WGAN for training the Discriminator can result in a more stable and reliable process of conditional image generation. This approach ensures that the Discriminator provides meaningful feedback to the Generator, enabling it to generate high-quality images that closely match the conditioning information (e.g., class-attributes) while addressing some of the challenges associated with traditional GAN training. The hyperparameter γ is used to control the balance or weighting of losses from both the CVAE and the WGAN components.

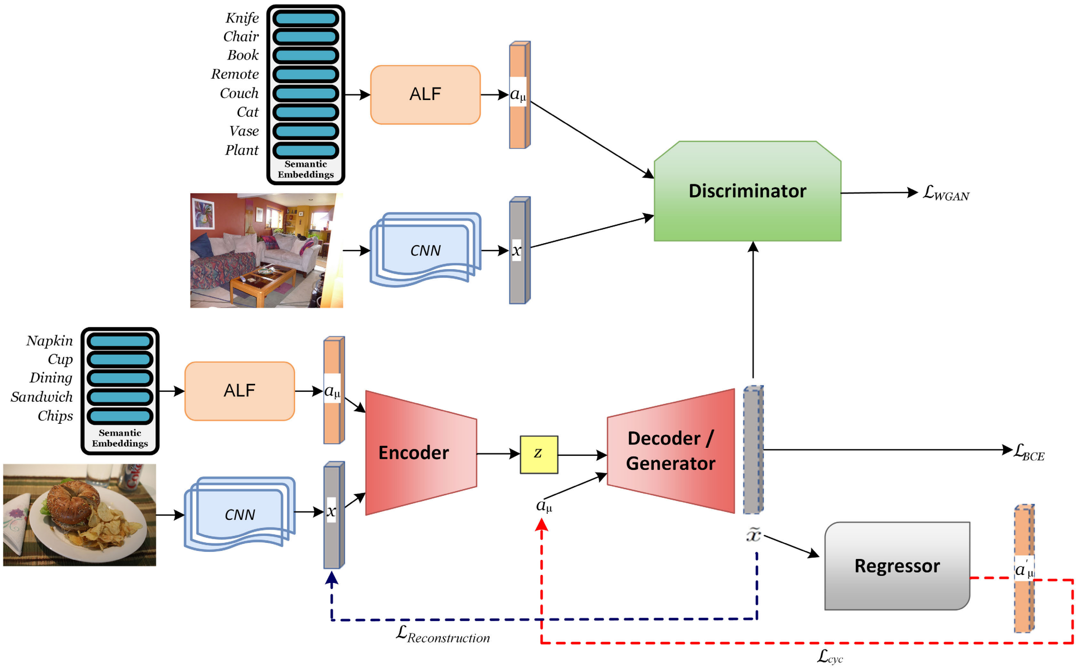


FIGURE 2. Our proposed Generative approach for Multi-label ZSL and GZSL combines CVAE and CGAN. We have used an Encoder E that encodes the visual features of an image x along with global image-level embedding a_μ to a latent feature space z. The Generator G(z,a) generates visual features x, given its global image-level embedding a_μ and latent vector z of the encoder as an input. A Discriminator D maps the pair of visual features and global image-level embedding to a compatibility score. The Regressor works as a regularizer, that will perform visual to-semantic features mapping.

In earlier generative approaches [30], [35], [79] for ZSL and GZSL, faced difficulty in generating visual features that accurately represented the real distribution of training data. This discrepancy between the distribution of generated visual features and the actual distribution of real training data posed a challenge during the training of classifiers, especially when dealing with unseen classes. Essentially, the generated features weren't closely aligned with the characteristics of the real data.

So, to address this issue, we propose to combine CVAE and CGAN along with an additional component, i.e. Regressor. The Regressor in the proposed model serves a role that is conceptually opposite to that of the Generator/Decoder mentioned in previous work by [25]. The Generator/Decoder maps global image-level embedding a_μ to visual features $\tilde{x}$. In contrast, the Regressor in the current model maps the generated or reconstructed visual features $\tilde{x}$ back to their corresponding global image-level embedding a_μ. The main purpose of the Regressor is to enforce a connection between the generated or reconstructed visual features and their associated class attributes. It makes sure that the generated features align with the expected characteristics of the class they are supposed to represent. This process ensures that the generated visual features are not only visually coherent but also semantically meaningful and consistent with their assigned class attributes. The Regressor's role can be likened to the concept of "cycle consistency" loss, which was introduced by [80]. Cycle consistency loss is often used in tasks like image-to-image translation to ensure that an image can be translated to another domain and then back again without losing important information. Similarly, in this proposed model, the Regressor ensures a "cycle" of consistency between visual features and class attributes. This means that the generated visual features should be mapped back to the same class attributes they were generated from, creating a coherent loop of transformation. The Regressor loss is as follows:

$$\mathcal{L}_{cyc} = \mathbb{E}_{a_\mu \sim P_s^{a_\mu}, z \sim P_{E(z|x,a_\mu)}}[||a_\mu - R(G(a_\mu, z))||_2^2] + \mathbb{E}_{a_\mu \sim P_u^{a_\mu}, z \sim P_{E(z|x,a_\mu)}}[||a_\mu - R(G(a_\mu, z))||_2^2] \quad (5)$$

where $P_s^{a_\mu}$ and $P_u^{a_\mu}$ in 5 denote the distribution of seen and unseen classes attributes, respectively. Furthermore, we have a cycle consistency loss in terms of a Regressor $R(G(z, a_\mu))$ that maps the reconstructed visual representations $\tilde{x}$ back to their global image-level embedding a_μ for good visual features generation. So, our unified Generative approach for GZSL trains the following combined loss function:

$$\mathcal{L}_{loss} = \min_G \max_D \mathcal{L}_{CVAEGAN} + \beta \mathcal{L}_{cyc} \quad (6)$$

where β is the hyperparameter that weights the Regressor loss.

The training begins with the pre-training of the CVAE. Equation 1 likely refers to the main objective or loss function used to train the CVAE. However, in the context of CVAE, the loss function usually consists of a reconstruction loss and a regularization term. The reconstruction loss ensures that the generated features are similar to the input features, and the regularization term encourages the latent features to follow a specific probability distribution. Pre-training the CVAE allows the model to learn a good initial representation of the data distribution and the latent space. It stabilizes the training process. As, training a CGAN from scratch can be challenging and prone to mode collapse, where the Generator produces limited variety of samples. The pre-trained CVAE provides a more stable starting point for the CGAN. After pre-training the CVAE, the entire Generative model is trained using equation 6. It is used as the objective or loss function for this training and it combines the various components of the model, including the CVAE, CGAN, and Regressor, along with relevant loss terms to guide the training process. This equation ensures that all components work together to generate realistic, semantically rich visual features for both seen and unseen classes.

To enhance the resilience of the visual feature generation process, we have adopted a strategy that involves creating multi-label features to train the ultimate multi-label ZSL and GZSL classifier. This strategy encompasses the formation of multi-label combinations that include both unseen classes and combinations comprising a mix of seen and unseen classes. To achieve this, we begin by assembling multi-label combinations consisting of seen classes from the training dataset. Subsequently, we introduce a degree of randomness, following a method outlined in [38], by associating these seen classes with their nearest unseen counterparts. This proximity is determined based on the embeddings of their respective class representations. As a result, this process yields a collection of diverse multi-label combinations. These combinations include instances exclusively featuring unseen classes as well as combinations that encompass both unseen and seen classes. The primary aim of this approach is to enhance the model's ability to generalize during the training phase. By exposing the model to scenarios where it must recognize objects it has never encountered before, we prepare it for the challenges of ZSL and GZSL. In these scenarios, the model must apply its knowledge to identify entirely new classes, thus requiring robust generalization capabilities."

Further depending on the task to be solved, whether multi-label ZSL or GZSL, both seen and unseen classes samples are utilized to train the final classifier i.e., softmax, along with the Binary cross entropy loss $BCE(f_{zsl}(\tilde{x}^a), y_u)$ and $BCE(f_{gzsl}(\tilde{x}^a), y_{u+s})$ for both multi-label ZSL and GZSL.

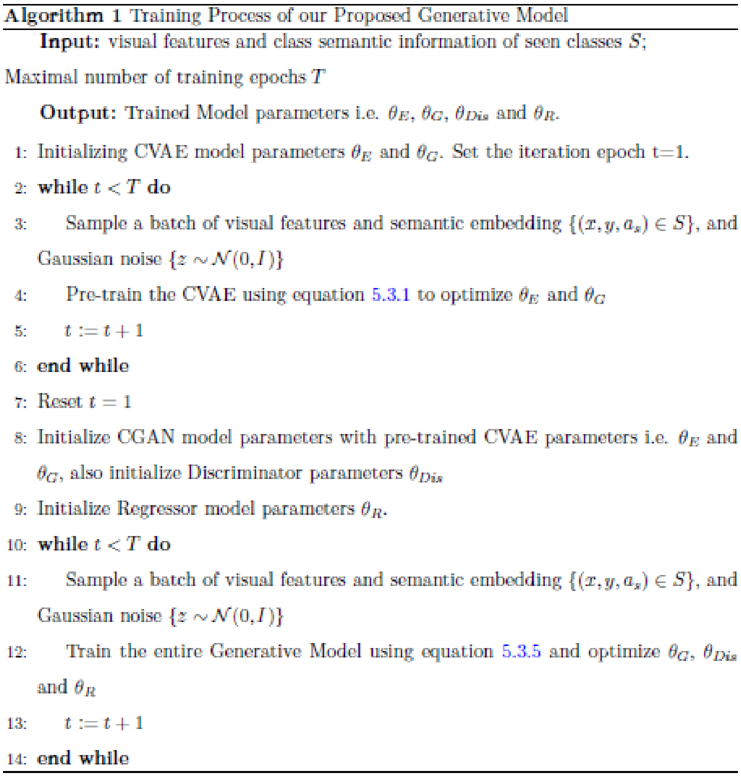
IV. EXPERIMENTS

A. DATASETS

We conducted an extensive evaluation of our proposed approach for multi-label ZSL and GZSL using two widely recognized benchmark datasets: MS COCO [39] and NUS-WIDE [40]. The NUS-WIDE dataset comprises 269,648 images that have been meticulously categorized into 81 classes by human annotators. This dataset encompasses a total of 925 labels, which were derived from tags provided by Flickr users. In line with the approach adopted by previous works such as [73] and [81], we designate the 925 labels as 'seen labels,' while the remaining 81 labels, which have been human-annotated, are considered 'unseen labels'. The MS COCO dataset features 123,287 images spanning 80 distinct categories, including a validation set with 40,504 images and a training set comprising 82,783 images. For our multi-label ZSL and GZSL experiments, we follow the same class split as used in [38], consisting of 15 unseen classes and 65 seen classes.

In these experiments, we employed two common types of data splits [82]: instance-first and label-first. Following established methodologies [73], [81], we chose to utilize the instance-first split for the NUS-WIDE dataset. This decision was made after careful consideration of the label-first protocol's drawbacks on the NUS-WIDE dataset. In the standard NUS-WIDE dataset split, there are 81 human-annotated classes designated as unseen and 925 machine-annotated classes as seen. Implementing a label-first protocol resulted in an imbalanced distribution of training and testing data. This significant data distribution disparity raised concerns about the integrity of the training process. Therefore, we concluded that maintaining the instance-first data split for the NUS-WIDE dataset was crucial to ensure a more balanced distribution of training and testing data, thus facilitating robust model training and evaluation. For the MSCOCO dataset, we adopted a label-first split with a ratio of 65/15 for seen/unseen classes, as proposed by [83]. This decision was driven by specific considerations aimed at preserving the semantic integrity of instances associated with unseen labels throughout the training process. By preserving instances related to unseen labels for testing and utilizing instances with known labels for training and validation, we aimed to mitigate potential biases and ensure a more reliable training process compared to the instance-first split.

In our experiments utilizing the NUSWIDE and MSCOCO datasets, we encountered unique challenges inherent to these datasets that directly influenced the efficacy of our proposed framework. Firstly, both datasets consist of images with varying levels of pixel quality and resolution. The NUSWIDE dataset, in particular, encompasses a wide range of image qualities due to its diverse nature, including both professional and amateur photography. Similarly, the MSCOCO dataset contains images captured under different lighting conditions and environments, leading to variations in pixel quality. Secondly, the presence of small objects within images was prominent in both datasets. In NUSWIDE, images often contain multiple objects of interest, some of which may be small or partially occluded, requiring the model to accurately detect and classify them. Similarly, the MSCOCO dataset features a plethora of object categories, including numerous instances of small objects that pose a challenge to recognition algorithms. Lastly, considering multiple viewpoints was crucial for achieving robust performance on both datasets. Images in NUSWIDE and MSCOCO can depict objects from various angles and perspectives, necessitating the model to generalize effectively across different viewpoints to ensure accurate classification.



```
Algorithm 1 Training Process of our Proposed Generative Model
Input: visual features and class semantic information of seen classes $S$;
Maximal number of training epochs $T$
Output: Trained Model parameters i.e. $\theta_E$, $\theta_G$, $\theta_{Dis}$ and $\theta_R$.
Initializing CVAE model parameters $\theta_E$ and $\theta_G$. Set the iteration epoch t=1.
while $t < T$ do
    Sample a batch of visual features and semantic embedding $\{(x, y, a_s) \in S\}$, and Gaussian noise $\{z \sim \mathcal{N}(0, I)\}$
    Pre-train the CVAE using equation 5.3.1 to optimize $\theta_E$ and $\theta_G$
    $t := t + 1$
end while
Reset $t = 1$
Initialize CGAN model parameters with pre-trained CVAE parameters i.e. $\theta_E$ and $\theta_G$, also initialize Discriminator parameters $\theta_{Dis}$
Initialize Regressor model parameters $\theta_R$.
while $t < T$ do
    Sample a batch of visual features and semantic embedding $\{(x, y, a_s) \in S\}$, and Gaussian noise $\{z \sim \mathcal{N}(0, I)\}$
    Train the entire Generative Model using equation 5.3.5 and optimize $\theta_G$, $\theta_{Dis}$ and $\theta_R$
    $t := t + 1$
end while
```

FIGURE 3. Description of the proposed model combining CVAE and CGAN for ZSL and GZSL.

By addressing these dataset-specific challenges through tailored preprocessing techniques, we were able to enhance the performance of our framework on the NUSWIDE and MSCOCO datasets. Our experimental results underscore the importance of considering factors such as pixel/image quality, small objects, and multiple viewpoints in the context of Multi-label ZSL and GZSL, thereby contributing to the advancement of ZSL methodologies. These benchmark datasets provide a robust foundation for assessing the performance of our proposed approach in the context of multi-label ZSL and GZSL scenarios. Furthermore, they enable us to make meaningful comparisons with state-of-the-art methods. Besides NUSWIDE and MS COCO, we also have the Open Images dataset [84], which is the largest multi-label dataset. However, due to limited resources, we were unable to perform experiments on this dataset.

B. EVALUATION PROTOCOLS

To measure the performance of our proposed generative approach, we employ two performance metrics: F1 score and mean Average Precision (mAP). These metrics are commonly used in the field and have been utilized in previous works such as [73] and [81]. The mAP metric is used to assess the label retrieval accuracy of the model. It measures how well the model ranks the correct labels for each image, indicating the model's ability to assign the most relevant labels. A higher mAP value indicates better performance in terms of label retrieval. On the other hand, the top-k F1 score measures the prediction accuracy of the model. It takes into account both precision and recall and is particularly useful in multi-label classification scenarios. By considering the top-k most probable labels for each image, the top-k F1 score evaluates how well the model predicts the correct labels. Higher values of the top-k F1 score indicate more accurate predictions. By using these evaluation metrics, we can quantitatively assess the performance of our proposed generative approach for multi-label ZSL and GZSL, and compare it with other state-of-the-art methods on standard datasets.

C. IMPLEMENTATION DETAILS

Following the methodologies described in [73] and [81], we employed the pre-trained VGG-19 model to extract features from the NUS-WIDE and MSCOCO datasets. Specifically, we utilize the output of the FC7 layer, consisting of 4096 image-level visual features, as input to our model. We implement the Encoder, Generator, Discriminator, and Regressor network as a feed-forward neural network with a different number of hidden layers. The Encoder and the Discriminator comprises a single hidden layer with 4096 hidden units, while the Generator and the Regressor comprise two hidden layers with 4096 hidden units. The architecture of the Encoder, Generator, Discriminator, and Regressor networks was carefully chosen to strike a balance between model complexity and performance. The specific configurations, such as the number of hidden layers and units, were determined through experimentation and empirical observations. We also considered prior literature and existing best practices in similar tasks to inform our choices. The latent vector z is set to 64 against the Encoder network for all the datasets. To incorporate class semantic information, we employ Glove vectors [85]. These vectors serve as a representation of the semantic information associated with each class. This approach has been widely adopted in existing literature and has shown promising results in handling semantic embeddings for improved performance in Multi-label ZSL and GZSL tasks. We use Adam for optimization and a constant learning rate of $\alpha = 0.001$ for all the datasets. We use $\beta = 0.01$ and $\gamma = 0.01$ for all the datasets. The optimization parameters, including the learning rate (α), regularization parameters (β and γ), and latent vector size (z), play a crucial role in training the model effectively. The determination of these parameters was exclusively conducted through rigorous cross-validation procedures, ensuring that our model's performance is robust and generalizable. Detailed parameter values are provided in Table 1 for reference. In each experiment, we generate 300 visual depictions for each unseen category, facilitating comprehensive evaluation and analysis of our proposed approach. These implementation specifics provide a solid foundation for training and evaluating our proposed model on Multi-label ZSL and GZSL tasks. The model took 3 days to train using a GPU.

TABLE 1. Summary of the parameter configurations for both multi-label ZSL and GZSL.

Parameters	Values
Visual Features	4096
Number of epochs	70
Number of neurons in each layer	4096
Learning rate	0.001
β	0.01
γ	0.01
Batch size	32
Latent Space	64

V. RESULTS

In this section, we conducted a thorough evaluation of our proposed model in comparison to state-of-the-art techniques, addressing both the multi-label ZSL and GZSL tasks. The outcomes of these assessments, as outlined in Table 2 and Table 3, offer valuable insights into the efficacy of our approach and its superiority over existing methods. To comprehensively gauge our model's performance, we employed two evaluation metrics: the mean Average Precision (mAP) and the F1 score across different K values (Top-K predictions). Furthermore, we computed the Precision (P) and Recall (R) metrics for each F1 score, facilitating a comprehensive analysis of our model's capabilities. Among the spectrum of approaches explored for multi-label ZSL and GZSL, one of the standout state-of-the-art methods is CONSE [69], which leverages a convex combination of class embedding vectors to establish associations between images and their corresponding semantic embedding space.

The LabelEM [86] approach introduces a classification strategy based on attributes, linking each class with its corresponding semantic information. Fast0Tag [81] deals with prioritizing relevant tags over irrelevant ones by identifying primary directions in the word vector space. Attention per Label [87] utilizes bilinear attention networks to effectively incorporate vision-language data. In the context of multi-label ZSL and GZSL classification, both LESA [73] and BiAM [74] embrace a shared multi-attention mechanism, facilitating multi-label recognition, the detection of previously unseen labels within an image, and the identification of relevant regions for each label.

In recent developments, ML-Decoder [77] has made a noteworthy contribution by introducing a novel attention-based classification head applicable to various classification tasks, including multi-label ZSL. GMLZSL [38] presents a GAN-based generative model that leverages multi-class-semantic information through a cross-level feature fusion approach, thereby enhancing visual feature generation. SDL [72] offers a method designed to promote semantic diversity among image labels by assigning higher weights to samples that exhibit greater diversity. In contrast to its counterparts, ML-ZSL [76] pioneers an innovative strategy for unbiased multi-label ZSL. It incorporates distinct class-specific regions into the training process of the classifier, reinforced by the Pyramid Feature Attention (PFA), which effectively bridges global and local information within samples, ensuring a balanced class representation. Lastly, the ADDS framework [75] introduces a flexible constraint to enhance the alignment of visual and textual features in multi-label classification. Within this framework, they introduce the DM-Decoder, a pioneering transformer decoder that facilitates the fusion of semantics from dual-modal sources. Remarkably, their feature generation approach for multi-label ZSL and GZSL demonstrates superior performance compared to other state-of-the-art methods on both datasets.

The comparative analysis conducted on the NUS-WIDE dataset for the conventional multi-label ZSL task reveals noteworthy findings. Initially, ADDS [75] exhibits performance comparable to other methods, achieving an mAP score of 36.5%. Meanwhile, ML-ZSL [76] attains F1 scores of 37.7% and 36.0% at K = 3 and K = 5, respectively. However, our proposed approach surpasses the performance of both ML-ZSL [76] and ADDS [75], alongside TGF [88], achieving F1 scores of 36.8% and 36.6% at K = 3 and K = 5, respectively. This enhancement underscores the efficacy of our model in accurately predicting labels for unseen classes. Transitioning to the GZSL task on the NUS-WIDE dataset, ML-Decoder [77] emerges as a frontrunner, outperforming existing methods with an mAP score of 19.9%. However, our proposed method presents a significant advancement, achieving an mAP score of 22.5%. This represents a substantial absolute gain of 2.6% over ML-Decoder [77]. Furthermore, in terms of F1 scores, our approach excels in comparison to ML-Decoder [77] and other approaches, yielding F1 scores of 24.6% and 28.9% at K = 3 and K = 5, respectively. These outcomes validate the efficacy of our model in precisely predicting labels for both seen and unseen classes, positioning it as a leader in GZSL tasks against other methods.

When evaluating the MS COCO dataset, GMLZSL [38] emerges as a standout among the array of state-of-the-art methods, showcasing remarkable performance across both multi-label ZSL and GZSL tasks. In the realm of conventional multi-label ZSL, GMLZSL achieves an impressive mAP score of 52.2% and a noteworthy F1 score of 43.5% at K = 3. Nevertheless, our proposed approach attains further enhancement over GMLZSL. It secures an mAP score of 53.2% and an elevated F1 score of 46.0% at K = 3, showcasing an absolute gain of 2.5% in F1 score. This distinction underscores the heightened performance of our model. Furthermore, at K = 5, our proposed model maintains an F1 score of 40.4%. In the realm of multi-label GZSL tasks, GMLZSL stands out with an mAP score of 35.3% and F1 scores of 46.7% and 46.4% at K = 3 and K = 5, respectively. However, our proposed approach exhibits superior performance, surpassing GMLZSL's benchmarks. Notably, our model achieves an enhanced F1 score of 46.4% at K = 5.

These findings demonstrate that our proposed approach, leveraging the combined strengths of conditional variational autoencoders (CVAE) and conditional generative adversarial network (CGAN) architectures along with a Regressor network, offers a robust and innovative solution for Multi-label ZSL and GZSL tasks. By integrating generative modeling techniques, our model not only synthesizes samples but also effectively addresses the challenges posed by unseen classes, surpassing the performance of existing generative approaches. Notably, among the methods we compare against, only GMLZSL employs a generative approach similar to ours. However, our proposed model exhibits superior performance even when compared to this established generative method.

Moreover, our research endeavors have culminated in a robust and innovative approach that addresses the challenges of Multi-label ZSL and GZSL. Through the integration of attribute-level fusion and cutting-edge techniques, we have achieved substantial enhancements in performance on both NUS-WIDE and MS COCO datasets. Additionally, careful architectural design choices are made to strike a balance between expressiveness and computational efficiency. Factors such as the number of layers, the dimensionality of latent spaces, and the connectivity patterns within the network are optimized to maximize the model's capacity to capture complex data distributions while minimizing computational overhead. Furthermore, our experimentation involves a broad spectrum of datasets. By subjecting our method to such diverse datasets, we aimed to comprehensively evaluate its ability to generalize across various data types. Our proposed generative approach was employed in these evaluations, showcasing their efficacy in processing a wide array of visual inputs.

We generated multi-label combinations, incorporating both unseen classes and combinations composed of a mixture of seen and unseen classes, for visual feature generation. This involved utilizing global image-level embedding, leading to improved generalization during testing and an enhanced overall performance. Our model, meticulously compared against an array of state-of-the-art methods, consistently demonstrates superior results. To better generalize we Notably, in both conventional multi-label ZSL tasks and the more complex multi-label GZSL tasks, our proposed methodology showcases remarkable proficiency, achieving remarkable gains in mean Average Precision (mAP) and F1 score. This underscores its efficacy in accurately predicting labels for both seen and unseen classes, effectively surmounting the challenges posed by these intricate learning scenarios.

We also provide a nuanced analysis of model performance across individual classes by reporting Average Per-Class Accuracy for each dataset as depicted in Figure.,4 and 5. This analysis offers insights into the model's effectiveness in accurately classifying specific categories within each dataset. Complementing the quantitative analysis, we provide qualitative insights into the performance of our method through a selection of images. Figure. 6 and 7 showcase instances where the proposed method excels in capturing intricate details, leading to accurate classification even in challenging scenarios. These qualitative insights offer a deeper understanding of the proposed method's capabilities, emphasizing its robustness across various real-world scenarios.

Building upon the observed trends in class-wise accuracy and the qualitative assessments, we delve into the implications of our results and discuss how the proposed method contributes to the broader field of Multi-label ZSL and GZSL. The combination of quantitative and qualitative analyses reinforces the robustness and versatility of our approach.

A. ABLATION STUDIES

1) COMPONENT ANALYSIS

In this section, we evaluate the effects of different objective functions of the proposed model, along with their results. We conduct a detailed analysis of individual components of the proposed Generative model to gauge their performance and effectiveness when operating independently. The outcomes of these components on two standard datasets are displayed in Tables 4 and 5, reflecting both conventional and generalized settings. To examine its efficacy in both scenarios, we integrated the Regressor with CVAE and CGAN.

TABLE 2. We conduct a comparative evaluation of the latest techniques for Multi-label ZSL and GZSL on the NUS-WIDE dataset. Our analysis includes the use of mAP and F1 score, with K values chosen from the set 3, 5. The most favorable results are emphasized in bold, while a dash ('-') indicates instances where the methods either do not provide their outcomes or have not conducted experiments with the datasets.

Method	NUS-WIDE													
	ZSL							GZSL						
	K=3			K=5				K=3			K=5			
	P	R	F1	P	R	F1	mAP	P	R	F1	P	R	F1	mAP
Fast0Tag [81]	22.6	36.2	27.8	18.2	48.4	26.4	15.1	18.8	8.3	11.5	15.9	11.7	13.5	3.7
ML-Decoder [77]	-	-	34.1	-	-	30.8	31.1	-	-	23.3	-	-	26.1	19.9
LESA [73]	25.7	41.1	31.6	19.7	52.5	28.7	19.4	23.6	10.4	14.4	19.8	14.6	16.8	5.6
SDL [72]	-	-	30.5	-	-	27.8	25.9	-	-	18.5	-	-	21.0	12.1
Attention per Label [87]	20.9	33.5	25.8	16.2	43.2	23.6	10.4	17.9	7.9	10.9	15.6	11.5	13.2	3.7
GMLZSL [38]	26.6	42.8	32.8	20.1	53.6	29.3	25.7	30.9	13.6	18.9	26.0	19.1	22.0	8.9
BiAM [74]	-	-	33.1	-	-	30.7	26.3	-	-	16.1	-	-	19.0	9.3
ML-ZSL [76]	**34.0**	42.3	**37.7**	26.7	55.3	36.0	28.0	31.2	13.9	19.2	26.4	19.6	22.5	9.3
ADDS [75]	-	-	34.2	-	-	36.0	36.5	-	-	-	-	-	-	-
ADA [89]	26.0	41.1	31.9	19.9	52.3	28.8	26.3	30.2	13.1	18.3	25.2	18.3	21.2	11.0
TGF [88]	29.0	**46.3**	35.6	21.4	56.9	31.1	31.1	33.9	14.9	20.7	29.1	21.4	24.6	15.8
OURS	32.3	42.8	36.8	**26.9**	**57.2**	**36.6**	**36.7**	**34.3**	**19.2**	**24.6**	**30.1**	**27.8**	**28.9**	**22.5**

TABLE 3. We conduct a comparative evaluation of the latest techniques for Multi-label ZSL and GZSL on the MS COCO dataset. Our analysis includes the use of mAP and F1 score, with K values chosen from the set 3, 5. The most favorable results are emphasized in bold, while a dash ('-') indicates instances where the methods either do not provide their outcomes or have not conducted experiments with the datasets.

Method	MS COCO													
	ZSL							GZSL						
	K=3			K=5				K=3			K=5			
	P	R	F1	P	R	F1	mAP	P	R	F1	P	R	F1	mAP
Fast0Tag [81]	-	-	37.5	-	-	-	43.3	-	-	33.8	-	-	34.6	27.9
LabelEM [86]	-	-	10.3	-	-	-	9.6	-	-	6.7	-	-	7.9	4.0
LESA [73]	-	-	33.6	-	-	-	31.8	-	-	26.7	-	-	28.0	17.7
CONSE [68]	-	-	18.4	-	-	-	13.2	-	-	19.6	-	-	18.9	7.7
GMLZSL [38]	-	-	43.5	-	-	-	52.2	-	-	44.1	-	-	43.4	33.2
TGF [88]	27.8	77.9	41.0	-	-	-	49.3	**49.3**	45.4	**47.3**	38.3	**58.7**	46.3	**40.3**
OURS	**32.3**	**79.8**	**46.0**	**26.4**	**86.3**	**40.4**	**53.2**	45.1	**48.5**	46.7	**39.8**	55.7	**46.4**	35.3

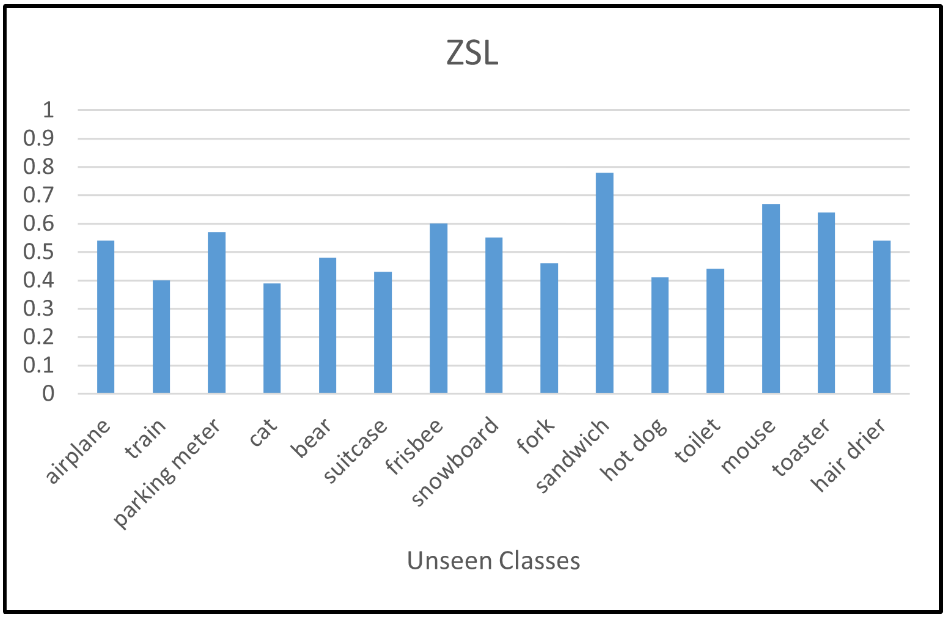


FIGURE 4. Class-wise prediction accuracies on MS COCO Dataset.

In our pursuit, we undertake the fusion of the Regressor with two diverse components: CVAE and CGAN. The initial case in this exploration involves the combination of $\mathcal{L}_{CVAE}+\mathcal{L}_{CYC}$, wherein the CVAE is harmonized with the Regressor.

This amalgamation utilizes the Encoder to generate a compact latent vector z, derived from a fusion of image visual features x and global image-level embedding a_{μ}, which the Decoder subsequently utilizes to reconstruct the datapoint $\tilde{x}$ along with global image-level embedding a_{μ}.

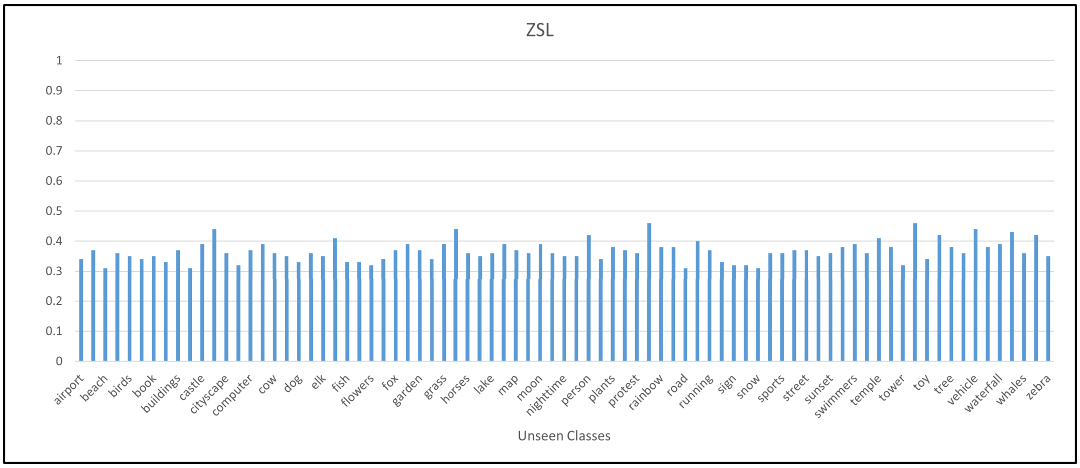


FIGURE 5. Class-wise prediction accuracies on NUS-WIDE Dataset.

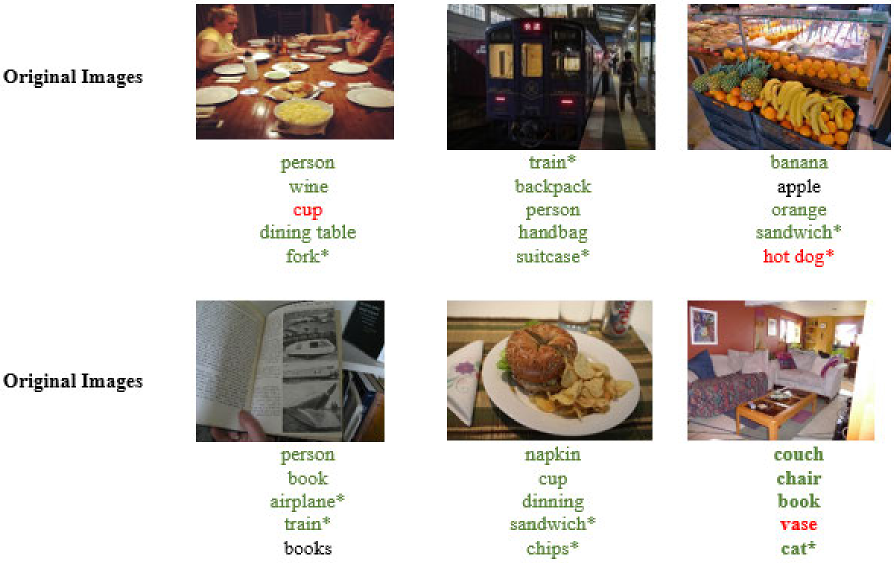


FIGURE 6. Comparison of predictions on test samples from the MS COCO dataset. The results depict the Top-5 predictions for Multi-label GZSL. '*' denotes unseen labels, with green text indicating True Positive predictions and red text indicating apparent incorrect predictions.

The second conjuncture unfolds as the amalgamation of the Regressor with CGAN, represented as $\mathcal{L}_{WGAN}+\mathcal{L}_{CYC}$. This configuration involves latent feature vector z and global image-level embedding a_{μ} as inputs, culminating in the generation of the reconstructed visual features $\tilde{x}$.

By individually investigating three distinct variants involving the Regressor coupled with CVAE and CGAN, we gain valuable insights into their performance across both conventional and generalized ZSL settings. On evaluating these combinations within both the conventional and generalized contexts, it becomes evident that their isolated efficacy falls short of optimal performance. Remarkably, the third synthesis stands as the vanguard, outperforming its predecessors by seamlessly integrating all three components, namely $\mathcal{L}_{WGAN}$, $\mathcal{L}_{CVAE}$, and $\mathcal{L}_{CYC}$. Notably, the fusion of $\mathcal{L}_{WGAN}$, $\mathcal{L}_{CVAE}$, and $\mathcal{L}_{CYC}$ emerges as the pinnacle, showcasing superior performance due to the comprehensive incorporation of all components. This observation underscores the pivotal role each component plays in collectively enhancing performance, a finding consistently resonant in both conventional and generalized contexts. This insight advances our understanding of the intricate interplay of objective functions and their cumulative impact on the effectiveness of the proposed generative model for ZSL.

2) ANALYZING γ ON NUS-WIDE DATASET

In our study, we systematically manipulate the value of aparameters, γ, to comprehensively assess their influence on the overall system performance. Through empirical analysis, we have carefully examined the consequences of varying

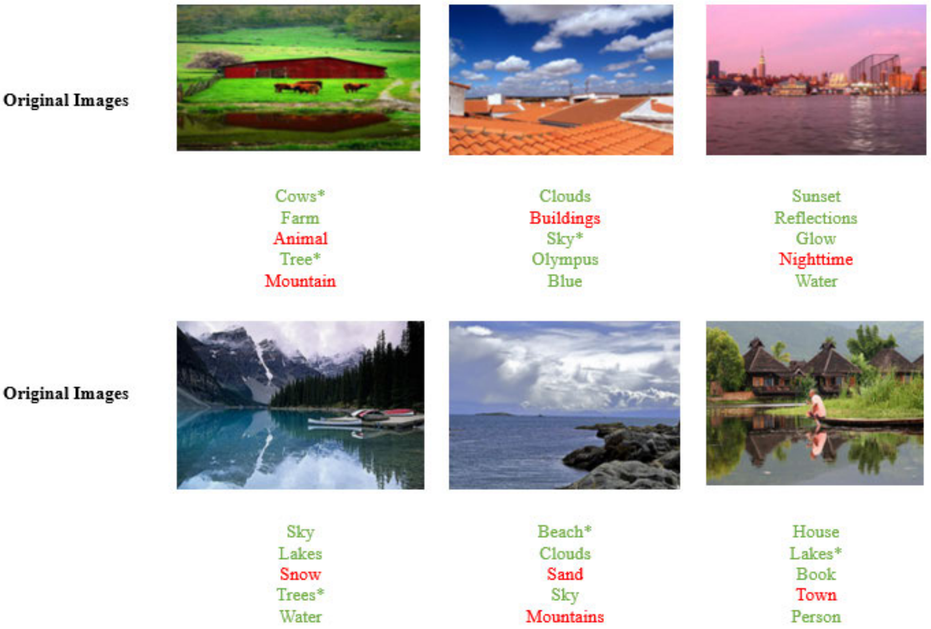


FIGURE 7. Comparison of predictions on test samples from the NUSWIDE dataset. The results depict the Top-5 predictions for Multi-label GZSL. '*' denotes unseen labels, with green text indicating True Positive predictions and red text indicating apparent incorrect predictions.

TABLE 4. Ablation Study: Analyzing the components of our proposed model on the NUS-WIDE dataset for Multi-label ZSL and GZSL.

Method	NUS-WIDE													
	ZSL							GZSL						
	K=3			K=5				K=3			K=5			
	P	R	F1	P	R	F1	mAP	P	R	F1	P	R	F1	mAP
$\mathcal{L}_{CVAE} + \mathcal{L}_{CYC}$	27.6	36.5	31.4	21.3	49.9	29.8	30.7	27.2	16.8	20.7	25.3	23.1	24.1	18.6
$\mathcal{L}_{WGAN} + \mathcal{L}_{CYC}$	29.3	38.3	33.2	22.6	51.5	31.4	32.7	29.6	17.8	22.2	27.7	24.6	26.1	20.8
$\mathcal{L}_{WGAN} + \mathcal{L}_{CVAE} + \mathcal{L}_{CYC}$	**32.3**	**42.8**	**36.8**	**26.9**	**56.0**	**36.3**	**36.7**	**33.4**	**19.2**	**24.4**	**30.1**	**27.8**	**28.9**	**22.5**

TABLE 5. Ablation Study: Analyzing the components of our proposed model on the MS COCO dataset for Multi-label ZSL and GZSL.

Method	MS COCO													
	ZSL							GZSL						
	K=3			K=5				K=3			K=5			
	P	R	F1	P	R	F1	mAP	P	R	F1	P	R	F1	mAP
$\mathcal{L}_{CVAE} + \mathcal{L}_{CYC}$	28.3	75.4	41.1	21.2	82.6	33.7	48.7	42.9	39.7	41.2	33.3	48.2	39.4	31.8
$\mathcal{L}_{WGAN} + \mathcal{L}_{CYC}$	30.4	77.8	43.7	23.6	84.8	36.9	50.5	44.2	41.7	42.9	36.1	50.5	42.1	33.8
$\mathcal{L}_{WGAN} + \mathcal{L}_{CVAE} + \mathcal{L}_{CYC}$	**32.3**	**79.8**	**46.0**	**26.4**	**86.3**	**40.4**	**53.2**	**45.1**	**44.5**	**44.8**	**38.7**	**53.6**	**44.9**	**35.3**

TABLE 6. Analyzing the impact of hyperparameter γ on NUS-WIDE dataset for Multi-label ZSL and GZSL.

Method	NUS-WIDE													
	ZSL							GZSL						
	K=3			K=5				K=3			K=5			
	P	R	F1	P	R	F1	mAP	P	R	F1	P	R	F1	mAP
$\gamma = 1$	28.3	37.6	32.3	22.7	51.6	31.5	30.3	29.1	15.2	19.9	25.6	22.0	23.7	17.3
$\gamma = 0.05$	31.6	40.9	35.7	25.2	54.7	34.5	34.3	32.2	17.6	22.8	28.5	25.2	26.8	20.7
$\gamma = 0.01$	**32.3**	**42.8**	**36.8**	**26.9**	**56.0**	**36.3**	**36.7**	**33.4**	**19.2**	**24.4**	**30.1**	**27.8**	**28.9**	**22.5**

these parameters and have made noteworthy observations. As γ is a scalar value that we can adjust. By changing the value of γ we can control the contribution of the $\mathcal{L}_{WGAN}$ to the total loss that is defined in Equation 4. If γ is large, it means that the $\mathcal{L}_{WGAN}$ loss has a higher weight, and the model will be more influenced by the adversarial training aspect of the WGAN. On the other hand, if γ is small, the reconstruction loss from the $\mathcal{L}_{CVAE}$ will have a larger impact on the total loss.

Our investigations have led us to a pivotal finding. Specifically, when we set $\gamma = 0.01$, a remarkable stabilization of the training process occurs on the NUS-WIDE dataset. These results are succinctly summarized in Table 6, where we present the outcomes of our experimentation. This configuration of γ appears to yield a favorable balance, conducive to reliable and consistent training outcomes. In practical terms, the model's training will be more focused on the reconstruction aspect of the $\mathcal{L}_{CVAE}$, and the adversarial aspect introduced by the $\mathcal{L}_{CGAN}$ will have less influence. Furthermore, an intriguing pattern surfaces from our experimentation. We have scrutinized the impact of deviating from the aforementioned parameter value and found a consistent trend. Notably, for values other than 0.01, we did not observe any discernible improvement in performance. This underscores the significance of the specific parameter value we have identified, emphasizing their efficacy in optimizing system performance.

VI. CONCLUSION

Our study presents an innovative generative framework designed to tackle the challenges of both multi-label ZSL and GZSL. This approach leverages the combined strengths of CVAE and CGAN to generate visual representations for previously unseen classes. By utilizing latent space vectors from the CVAE and global image-level embeddings as inputs for the Generator/Decoder, we establish a robust foundation for visual feature synthesis. To ensure the stability of the CGAN during training, we implement the Wasserstein GAN technique, which guides the generation process towards features that are conducive to accurate classification. Additionally, a complementary Regressor functions as a regularizer, enhancing the fidelity of feature reconstructions through cycle consistency loss by mapping the generated visual representations back to their corresponding global image-level embeddings. The integration of Softmax classifier training for both seen and unseen classes further refines the system for classification tasks after augmenting the unseen class data. Our approach incorporates attributes as supplementary information for visual feature generation and undergoes empirical evaluation on two benchmark datasets: NUS-WIDE and MS COCO. Notably, our method proves effective in both multi-label ZSL and GZSL scenarios, highlighting its versatility and potential to advance state-of-the-art solutions in the field.

ACKNOWLEDGMENT

This article represents the opinions of the author(s) and does not mean to represent the position or opinions of American University of Sharjah.

REFERENCES

[1] Q. Li, W. Cai, X. Wang, Y. Zhou, D. D. Feng, and M. Chen, "Medical image classification with convolutional neural network," in *Proc. 13th Int. Conf. Control Autom. Robot. Vis. (ICARCV)*, Dec. 2014, pp. 844–848.

[2] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[3] H. Huang, C. Wang, P. S. Yu, and C.-D. Wang, "Generative dual adversarial network for generalized zero-shot learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 801–810.

[4] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 28, 2015, pp. 1–9.

[5] Y. Shen, J. Qin, L. Huang, L. Liu, F. Zhu, and L. Shao, "Invertible zero-shot recognition flows," in *Computer Vision—ECCV*. Glasgow, U.K.: Springer, 2020, pp. 614–631.

[6] H. Fujiyoshi, T. Hirakawa, and T. Yamashita, "Deep learning-based image recognition for autonomous driving," *IATSS Res.*, vol. 43, no. 4, pp. 244–252, Dec. 2019.

[7] A. Kamilaris and F. X. Prenafeta-Boldú, "Deep learning in agriculture: A survey," *Comput. Electron. Agricult.*, vol. 147, pp. 70–90, Apr. 2018.

[8] T. Ridnik, E. Ben-Baruch, A. Noy, and L. Zelnik-Manor, "ImageNet-21k pretraining for the masses," 2021, *rXiv:2104.10972*.

[9] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 248–255.

[10] L. Feng, B. An, and S. He, "Collaboration based multi-label learning," in *Proc. AAAI Conf. Artif. Intell.*, 2019, vol. 33, no. 1, pp. 3550–3557.

[11] Y. Gong, Y. Jia, T. Leung, A. Toshev, and S. Ioffe, "Deep convolutional ranking for multilabel image annotation," 2013, *arXiv:1312.4894*.

[12] G. Tsoumakas and I. M. Katakis, "Multi-label classification: An overview," *Int. J. Data Warehousing Mining*, vol. 3, pp. 1–13, 2007. [Online]. Available: https://api.semanticscholar.org/CorpusID:11608263

[13] J. Wang, Y. Yang, J. Mao, Z. Huang, C. Huang, and W. Xu, "CNN-RNN: A unified framework for multi-label image classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2285–2294.

[14] H. Yang, J. T. Zhou, Y. Zhang, B.-B. Gao, J. Wu, and J. Cai, "Exploit bounding box annotations for multi-label object recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 280–288.

[15] H.-F. Yu, P. Jain, P. Kar, and I. S. Dhillon, "Large-scale multi-label learning with missing labels," 2013, *arXiv:1307.5101*.

[16] Z.-M. Chen, X.-S. Wei, P. Wang, and Y. Guo, "Multi-label image recognition with graph convolutional networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 5172–5181.

[17] J. Nam, E. L. Mencia, H. J. Kim, and J. Fürnkranz, "Maximizing subset accuracy with recurrent neural networks in multi-label classification," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 1–11.

[18] V. O. Yazici, A. Gonzalez-Garcia, A. Ramisa, B. Twardowski, and J. van de Weijer, "Orderless recurrent models for multi-label classification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 13437–13446.

[19] J. Ye, J. He, X. Peng, W. Wu, and Y. Qiao, "Attention-driven dynamic graph convolutional network for multi-label image recognition," in *Computer Vision—ECCV*. Glasgow, U.K.: Springer, 2020, pp. 649–665.

[20] R. You, Z. Guo, L. Cui, X. Long, Y. Bao, and S. Wen, "Cross-modality attention with semantic graph embedding for multi-label classification," in *Proc. AAAI Conf. Artif. Intell.*, vol. 34, 2020, pp. 12709–12716.

[21] T. Durand, N. Mehrasa, and G. Mori, "Learning a deep ConvNet for multi-label classification with partial labels," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 647–657.

[22] J. Weston, S. Bengio, and N. Usunier, "Wsabie: Scaling up to large vocabulary image annotation," in *Proc. 22nd Int. Joint Conf. Artif. Intell.*, 2011, pp. 1–7.

[23] W. Wang, V. W. Zheng, H. Yu, and C. Miao, "A survey of zero-shot learning: Settings, methods, and applications," *ACM Trans. Intell. Syst. Technol.*, vol. 10, no. 2, pp. 1–37, Mar. 2019.

[24] Y. Xian, B. Schiele, and Z. Akata, "Zero-shot learning—The good, the bad and the ugly," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Los Alamitos, CA, USA, Jul. 2017, pp. 3077–3086.

[25] R. Felix, B. G. Vijay Kumar, I. Reid, and G. Carneiro, "Multi-modal cycle-consistent generalized zero-shot learning," in *Computer Vision—ECCV*, V. Ferrari, M. Hebert, C. Sminchisescu, and Y. Weiss, Eds., Cham, Switzerland: Springer, 2018, pp. 21–37.

[26] Y. Fu, T. M. Hospedales, T. Xiang, and S. Gong, "Transductive multi-view zero-shot learning," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 11, pp. 2332–2345, Nov. 2015.

[27] D. Jayaraman and K. Grauman, "Zero-shot recognition with unreliable attributes," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 27, 2014, pp. 1–9.

[28] J. Li, M. Jing, K. Lu, Z. Ding, L. Zhu, and Z. Huang, "Leveraging the invariant side of generative zero-shot learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 7394–7403.

[29] B. Romera-Paredes and P. Torr, "An embarrassingly simple approach to zero-shot learning," in *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 2152–2161.

[30] Y. Xian, T. Lorenz, B. Schiele, and Z. Akata, "Feature generating networks for zero-shot learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 5542–5551.

[31] Y. Xian, S. Sharma, B. Schiele, and Z. Akata, "F-VAEGAN-d2: A feature generating framework for any-shot learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 10267–10276.

[32] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial networks," *Commun. ACM*, vol. 63, no. 11, pp. 139–144, 2020.

[33] D. P. Kingma and M. Welling, "Auto-encoding variational Bayes," 2013, *arXiv:1312.6114*.

[34] E. Schönfeld, S. Ebrahimi, S. Sinha, T. Darrell, and Z. Akata, "Generalized zero- and few-shot learning via aligned variational autoencoders," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 8239–8247.

[35] F. Jurie, M. Bucher, and S. Herbin, "Generating visual representations for zero-shot classification," in *Proc. IEEE Int. Conf. Comput. Vis. Workshops (ICCVW)*, Oct. 2017, pp. 2666–2673.

[36] Y. Geng, J. Chen, Z. Chen, Z. Ye, Z. Yuan, Y. Jia, and H. Chen, "Generative adversarial zero-shot learning via knowledge graphs," 2020, *arXiv:2004.03109*.

[37] M. Gull and O. Arif, "Multi-label generalized zero-shot learning using identifiable variational autoencoders," in *Proc. Int. Conf. Extended Reality*. Switzerland: Springer, 2023, pp. 35–50.

[38] A. Gupta, S. Narayan, S. Khan, F. S. Khan, L. Shao, and J. Van De Weijer, "Generative multi-label zero-shot learning," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 12, pp. 14611–14624, Dec. 2023.

[39] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft COCO: Common objects in context," in *Computer Vision—ECCV*. Zürich, Switzerland: Springer, 2014, pp. 740–755.

[40] T.-S. Chua, J. Tang, R. Hong, H. Li, Z. Luo, and Y. Zheng, "NUS-WIDE: A real-world web image database from national University of Singapore," in *Proc. ACM Int. Conf. Image Video Retr.*, Jul. 2009, pp. 1–9.

[41] Z. Akata, F. Perronnin, Z. Harchaoui, and C. Schmid, "Label-embedding for attribute-based classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 819–826.

[42] G.-S. Xie, L. Liu, X. Jin, F. Zhu, Z. Zhang, J. Qin, Y. Yao, and L. Shao, "Attentive region embedding network for zero-shot learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 9376–9385.

[43] L. Zhang, P. Wang, L. Liu, C. Shen, W. Wei, Y. Zhang, and A. van den Hengel, "Towards effective deep embedding for zero-shot learning," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 30, no. 9, pp. 2843–2852, Sep. 2020.

[44] M. Norouzi, T. Mikolov, S. Bengio, Y. Singer, J. Shlens, A. Frome, G. S. Corrado, and J. Dean, "Zero-shot learning by convex combination of semantic embeddings," 2014, *arXiv:1312.5650*.

[45] Z. Akata, S. Reed, D. Walter, H. Lee, and B. Schiele, "Evaluation of output embeddings for fine-grained image classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 2927–2936.

[46] Y. Annadani and S. Biswas, "Preserving semantic relations for zero-shot learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7603–7612.

[47] L. Zhang, T. Xiang, and S. Gong, "Learning a deep embedding model for zero-shot learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 3010–3019.

[48] 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 Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 9, pp. 2251–2265, Sep. 2019.

[49] Z. Zhang and V. Saligrama, "Zero-shot learning via semantic similarity embedding," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 4166–4174.

[50] Y. Xian, Z. Akata, G. Sharma, Q. Nguyen, M. Hein, and B. Schiele, "Latent embeddings for zero-shot classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 69–77.

[51] R. Felix, I. Reid, and G. Carneiro, "Multi-modal cycle-consistent generalized zero-shot learning," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 21–37.

[52] M. Mirza and S. Osindero, "Conditional generative adversarial nets," 2014, *arXiv:1411.1784*.

[53] A. Frome, G. S. Corrado, J. Shlens, S. Bengio, J. Dean, M. Ranzato, and T. Mikolov, "DeViSE: A deep visual-semantic embedding model," in *Proc. NIPS*, 2013, pp. 2121–2129.

[54] Z. Han, Z. Fu, and J. Yang, "Learning the redundancy-free features for generalized zero-shot object recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2021, pp. 12865–12874.

[55] A. Mishra, S. K. Reddy, A. Mittal, and H. A. Murthy, "A generative model for zero shot learning using conditional variational autoencoders," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2018, pp. 2269–22698.

[56] V. K. Verma, G. Arora, A. Mishra, and P. Rai, "Generalized zero-shot learning via synthesized examples," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 4281–4289.

[57] H. Yu and B. Lee, "Zero-shot learning via simultaneous generating and learning," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 32, 2019, pp. 1–11.

[58] R. Keshari, R. Singh, and M. Vatsa, "Generalized zero-shot learning via over-complete distribution," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 13297–13305.

[59] R. Gao, X. Hou, J. Qin, J. Chen, L. Liu, F. Zhu, Z. Zhang, and L. Shao, "Zero-VAE-GAN: Generating unseen features for generalized and transductive zero-shot learning," *IEEE Trans. Image Process.*, vol. 29, pp. 3665–3680, 2020.

[60] J. Li, M. Jing, K. Lu, L. Zhu, and H. T. Shen, "Investigating the bilateral connections in generative zero-shot learning," *IEEE Trans. Cybern.*, vol. 52, no. 8, pp. 8167–8178, Aug. 2022.

[61] J. Read, B. Pfahringer, G. Holmes, and E. Frank, "Classifier chains for multi-label classification," *Mach. Learn.*, vol. 85, no. 3, pp. 333–359, Dec. 2011.

[62] C.-W. Lee, W. Fang, C.-K. Yeh, and Y. F. Wang, "Multi-label zero-shot learning with structured knowledge graphs," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 1576–1585.

[63] Q. Li, M. Qiao, W. Bian, and D. Tao, "Conditional graphical lasso for multi-label image classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2977–2986.

[64] F. Zhu, H. Li, W. Ouyang, N. Yu, and X. Wang, "Learning spatial regularization with image-level supervisions for multi-label image classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2027–2036.

[65] X. Cheng, H. Lin, X. Wu, D. Shen, F. Yang, H. Liu, and N. Shi, "MLTR: Multi-label classification with transformer," in *Proc. IEEE Int. Conf. Multimedia Expo (ICME)*, Jul. 2022, pp. 1–6.

[66] J. Lanchantin, T. Wang, V. Ordonez, and Y. Qi, "General multi-label image classification with transformers," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 16473–16483.

[67] S. Liu, L. Zhang, X. Yang, H. Su, and J. Zhu, "Query2Label: A simple transformer way to multi-label classification," 2021, *arXiv:2107.10834*.

[68] C. H. Lampert, H. Nickisch, and S. Harmeling, "Learning to detect unseen object classes by between-class attribute transfer," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 951–958.

[69] M. Norouzi, T. Mikolov, S. Bengio, Y. Singer, J. Shlens, A. Frome, G. S. Corrado, and J. Dean, "Zero-shot learning by convex combination of semantic embeddings," in *Proc. Int. Conf. Learn. Represent.*, 2013, pp. 1–9.

[70] Z.-M. Chen, Q. Cui, X.-S. Wei, X. Jin, and Y. Guo, "Disentangling, embedding and ranking label cues for multi-label image recognition," *IEEE Trans. Multimedia*, vol. 23, pp. 1827–1840, 2021.

[71] M. Shi, Y. Tang, X. Zhu, and J. Liu, "Multi-label graph convolutional network representation learning," *IEEE Trans. Big Data*, vol. 8, no. 5, pp. 1169–1181, Oct. 2022.

[72] A. Ben-Cohen, N. Zamir, E. B. Baruch, I. Friedman, and L. Zelnik-Manor, "Semantic diversity learning for zero-shot multi-label classification," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 620–630.

[73] D. Huynh and E. Elhamifar, "A shared multi-attention framework for multi-label zero-shot learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 8773–8783.

[74] S. Narayan, A. Gupta, S. Khan, F. S. Khan, L. Shao, and M. Shah, "Discriminative region-based multi-label zero-shot learning," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 8711–8720.

[75] S. Xu, Y. Li, J. Hsiao, C. Ho, and Z. Qi, "Open vocabulary multi-label classification with dual-modal decoder on aligned visual-textual features," 2022, *arXiv:2208.09562*.

[76] Z. Liu, S. Guo, J. Guo, Y. Xu, and F. Huo, "Towards unbiased multi-label zero-shot learning with pyramid and semantic attention," *IEEE Trans. Multimedia*, vol. 25, pp. 7441–7455, 2022.

[77] T. Ridnik, G. Sharir, A. Ben-Cohen, E. Ben-Baruch, and A. Noy, "ML-decoder: Scalable and versatile classification head," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2023, pp. 32–41.

[78] K. Sohn, H. Lee, and X. Yan, "Learning structured output representation using deep conditional generative models," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 28, 2015, pp. 1–9.

[79] Y. Long, L. Liu, F. Shen, L. Shao, and X. Li, "Zero-shot learning using synthesised unseen visual data with diffusion regularisation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 10, pp. 2498–2512, Oct. 2018.

[80] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2242–2251.

[81] Y. Zhang, B. Gong, and M. Shah, "Fast zero-shot image tagging," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 5985–5994.

[82] Q. Wang and K. Chen, "Multi-label zero-shot human action recognition via joint latent ranking embedding," *Neural Networks*, vol. 122, pp. 1–23, Feb. 2019.

[83] N. Hayat, M. Hayat, S. Rahman, S. Khan, S. W. Zamir, and F. S. Khan, "Synthesizing the unseen for zero-shot object detection," in *Computer Vision—ACCV*, H. Ishikawa, C.-L. Liu, T. Pajdla, and J. Shi, Ed., Cham, Switzerland: Springer, 2021, pp. 155–170.

[84] A. Kuznetsova, H. Rom, N. Alldrin, J. Uijlings, I. Krasin, J. Pont-Tuset, S. Kamali, S. Popov, M. Malloci, A. Kolesnikov, T. Duerig, and V. Ferrari, "The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale," *Int. J. Comput. Vis.*, vol. 128, no. 7, pp. 1956–1981, Jul. 2020.

[85] J. Pennington, R. Socher, and C. Manning, "Glove: Global vectors for word representation," in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, 2014, pp. 1532–1543.

[86] Z. Akata, F. Perronnin, Z. Harchaoui, and C. Schmid, "Label-embedding for image classification," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 7, pp. 1425–1438, Jul. 2016.

[87] J.-H. Kim, J. Jun, and B.-T. Zhang, "Bilinear attention networks," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 31, 2018, pp. 1–11.

[88] P. Ma, Z. He, W. Ran, and H. Lu, "A transferable generative framework for multi-label zero-shot learning," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 34, no. 5, pp. 3409–3423, May 2024.

[89] K.-Y. Chen and M.-C. Yeh, "Generative and adaptive multi-label generalized zero-shot learning," in *Proc. IEEE Int. Conf. Multimedia Expo (ICME)*, Jul. 2022, pp. 1–6.

MUQADDAS GULL received the B.S. degree in software engineering and the M.S. degree in computer science from the University of Sargodha, in 2013 and 2015, respectively, and the Ph.D. degree in computer science from the National University of Sciences and Technology (NUST). Her research interests include machine learning, computer vision, and deep learning.

OMAR ARIF (Senior Member, IEEE) received the B.E. degree in software engineering from the National University of Sciences and Technology (NUST), Islamabad, Pakistan, and the M.S. and Ph.D. degrees in electrical and computer engineering from Georgia Institute of Technology, Atlanta, GA, USA. He is currently a tenured Associate Professor with NUST and a Visiting Professor with American University of Sharjah. His current research interests include machine learning and computer vision.