Journal of *Imaging*

MDPI

Article

HFM: A Hybrid Feature Model Based on Conditional Auto Encoders for Zero-Shot Learning

Fadi Al Machot [1], Mohib Ullah [2,*] and Habib Ullah [1]

[1] Faculty of Science and Technology, Norwegian University of Life Science (NMBU), 1430 Ås, Norway; fadi.al.machot@nmbu.no (F.A.M.); habib.ullah@nmbu.no (H.U.)
[2] Department of Computer Science, Norwegian University of Science and Technology, 2819 Gjøvik, Norway
* Correspondence: mohib.ullah@ntnu.no

Abstract: Zero-Shot Learning (ZSL) is related to training machine learning models capable of classifying or predicting classes (labels) that are not involved in the training set (unseen classes). A well-known problem in Deep Learning (DL) is the requirement for large amount of training data. Zero-Shot learning is a straightforward approach that can be applied to overcome this problem. We propose a Hybrid Feature Model (HFM) based on conditional autoencoders for training a classical machine learning model on pseudo training data generated by two conditional autoencoders (given the semantic space as a condition): (a) the first autoencoder is trained with the visual space concatenated with the semantic space and (b) the second autoencoder is trained with the visual space as an input. Then, the decoders of both autoencoders are fed by the test data of the unseen classes to generate pseudo training data. To classify the unseen classes, the pseudo training data are combined to train a support vector machine. Tests on four different benchmark datasets show that the proposed method shows promising results compared to the current state-of-the-art when it comes to settings for both standard Zero-Shot Learning (ZSL) and Generalized Zero-Shot Learning (GZSL).

Keywords: Zero-Shot Learning (ZSL); semantic space; conditional autoencoders; generative models; computer vision

Citation: Al Machot, F.; Ullah, M.; Ullah, H. HFM: A Hybrid Feature Model Based on Conditional Auto Encoders for Zero-Shot Learning. *J. Imaging* **2022**, *8*, 171. https://doi.org/10.3390/jimaging8060171

Academic Editors: Raimondo Schettini, Jérémie Sublime and Hélène Urien

Received: 26 March 2022
Accepted: 9 June 2022
Published: 16 June 2022

1. Introduction

Deep-learning-based models have brought tremendous advancement in different fields, including but not limited to computer vision [1,2], natural language processing [3], and satellite image processing [4]. In these research fields, deep-learning-based models achieved human-level capabilities. In fact, these developments are subject to higher quality and large-scale data. With the exponential growth of new classes in our real world, collecting large amounts of data driven by significant variations requires much cost. It is a key challenge to annotate sufficient training data for each class to exploit supervised learning [4,5]. Therefore, different learning paradigms with limited labeled data have been presented in the literature, namely semi-supervised learning [4], life-long learning [6], and active learning [7]. However, the capabilities of these paradigms are limited in exploring variations in the limited amount of labeled data. Generally, humans can recognize over 30,000 core item types [8] and many more sub-categories. Additionally, humans are also excellent at recognizing items without seeing any visual examples. This capability is the zero-shot learning problem in machine learning.

Zero-shot learning (ZSL) models [9–11] have recently emerged to identify unseen categories with no training data but with semantic descriptions of classes. The ZSL models can take into account situations when data are scarce [12,13]. In general, the ZSL models address this situation by learning either a visual-to-semantic mapping [14,15] or a semantic-to-visual mapping [16,17]. The general assumption is based on the observations that the visual space encodes the semantic space and that the semantic space encodes the visual

space [15,18–20]. However, zero-shot learning is still a challenging research field since we need to predict unseen test categories that are never used when training the models [21–23]. For example, most ZSL methods like Deep Embedding Model (DEM) [24–26] discover direct embeddings from global features to the semantic space. However, the methods cannot capture the appearance relationships between different local regions in this way. The techniques could also ruin the diversity of visual modality due to highly overlapped semantic descriptions of various categories.

To cope with these challenges, we propose a Hybrid Feature Model (HFM) based on conditional autoencoders for zero-shot learning method to identify both seen and unseen classes via transferring knowledge from seen categories to unseen categories. Based on the observations [27] where a single conditional variational autoencoder is used, our method consists of two autoencoders that are depicted in Figure 1. The first autoencoder is provided by the concatenation of the visual and semantic spaces. The second autoencoder is provided by only the visual space. Our proposed method encodes the real data distribution efficiently. Therefore, our approach identifies the unbiased projection toward seen classes and produces close relationships between unseen samples and prototypes.

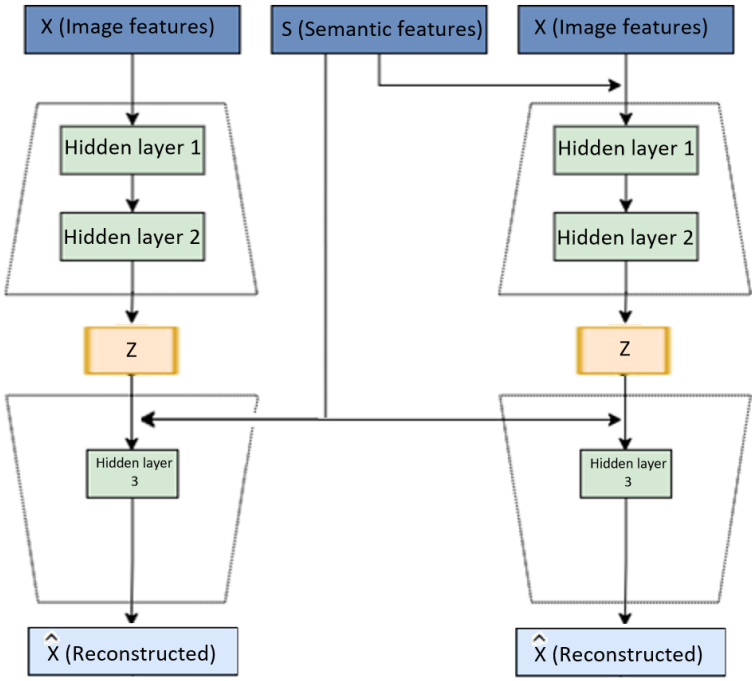

Figure 1. The proposed approach consists of two autoencoders. The first autoencoder is provided by the concatenated vectors of the visual and semantic spaces. The second autoencoder is provided by the visual features vectors only. Both autoencoders have a dense layer, followed by a dropout and a second dense layer. This is followed by another layer, which generates the values z. Activation functions are ReLU, and the activation functions for the last layer for both the encoder and the decoder are linear.

Most techniques fail to consider the discriminative information between the visual and semantic spaces. Thus, the significant insight is that our hybrid autoencoder approach may precisely represent the real data distribution of the query set in a fine-grained and dynamic manner. Especially, when the available samples are not driven by rich discriminative information. This can be exploited to enrich the diversity of data distribution and further improve the model accuracy. Furthermore, we explore both the visual and semantic spaces to encode diversified and discriminative modes of variation for learning a boosted classifier. Therefore, our method alleviates the problems when intra-class diversity and inter-class discriminability are lacking. Consequently, the proposed model presents promising results

using a highly fine-grained dataset (see Section 5). In addition, the work shows that using multiple VAEs generate an improved discriminative image space where data are easier to separate for ZSL classification purposes.

The rest of the paper is divided into the following sections: in Section 2, we present the related works from the literature. In Section 3, we present our proposed method in detail. Experiments and experimental results on four benchmark datasets and a conclusion are presented in Sections 4, 5, and Section 6, respectively.

2. Related Work

We classify the literature into two categories: embedding space-based zero-shot learning and feature generation-based zero-shot learning. In the first category, Lampert et al. [12] presented attribute-based classification based on a high-level description that is phrased in terms of semantic attributes, such as the object's color or shape. Norouzi et al. [13] introduced an image embedding system that mapped images into the semantic embedding space via a convex combination of the class label embedding vectors. However, the methods do not provide a natural mechanism for multiple semantic modalities to be fused and optimized jointly in an end-to-end structure. In [18], authors assumed that unseen categories come from unsupervised text corpora. Their method is based on the distributions of words in texts as a semantic space for understanding what objects look like. The method does not use the distribution information of samples. Therefore, the method cannot discover the cluster structure of samples. The authors [15] presented a visual-semantic embedding model trained to recognize visual objects using both labeled image data as well as semantic features gleaned from the unannotated text. They did not exploit the cluster relationship to rectify the biased sample-prototype relationship. Akata et al. [20] learned a function considering image and class embeddings. They used supervised attributes and unsupervised output embeddings either derived from hierarchies or learned from unlabeled text corpora. Xian et al. [21] introduced a latent embedding model for learning a compatibility function between image and class embeddings. Romera et al. [22] modeled the relationships between features, attributes, and classes as a two linear layers architecture, where the weights of the top layer are not learned but are given by the surrounding features. The researchers [23] embedded each class in the space of attribute vectors. Changpinyo et al. [28] aligned the semantic space to the model space that concerns itself with recognizing visual features. Kodirov et al. [29] presented a ZSL learning model based on a Semantic AutoEncoder (SAE). They projected a visual feature vector into the semantic space. The encoder and decoder may be linear and symmetric, which could not recognize or differentiate multiple features. Zhang et al. [24] used the visual space as the embedding space by considering the subsequent nearest neighbor search. The method [30] introduced an episode-based model for zero-shot learning. They trained their model within a set of episodes, each of which is modeled to simulate a zero-shot classification task. These methods have limited abilities to scale to large numbers of object categories. This limitation is partly due to the increasing complexity of collecting sufficient training data in the form of labeled images as the number of object categories grows.

In the second category, the methods learn to consolidate the visual samples for unseen classes. These methods first learn a conditional generative model considering, for example, Variational Autoencoder (VAE) and Generative Adversarial Networks (GAN). In addition, GAN-based approaches, e.g., f-VAEGAN-D2 [25] and TF-VAEGAN [26] show a competitive performance. In [25], authors proposed f-VAEGAN-D2, which combined VAEs and GANs to learn the marginal feature distribution of unlabeled images through an unconditional discriminator. However, the method cannot discover the class-based feature distribution from the available semantic information. In contrast to f-VAEGAN-D2 model, authors in [26] proposed the TF-VAEGAN model, which combined VAEs and GANs. However, they added a semantic embedding decoder to reconstruct the embedding space, which is used as a feedback module to improve the output of the Generator of the GAN. However, GANs and their derivatives show training instability, while VAE is more stable [31].

Mishra et al. [27] generated the samples from the given attributes, using a conditional variational autoencoder, and exploited the generated samples to classify the unseen classes.

Our proposed method falls into the feature generation-based zero-shot category driven by stability during training. The approach also encodes complex data distribution efficiently. It demonstrates that for specific test situations (see Section 5), a hybrid model consisting of two VAEs can outperform a GAN-VAE model with less training effort. Excluding the Kullback–Leibler (KL) divergence from the conditional VAE loss yields enhanced discriminative image features for classifying unseen classes in ZSL settings, which is promising. A limitation of the proposed approach is that the proposed model lacks a feedback module that can be coupled with the decoder to improve the reconstructed image space. To show the strength of our proposed method, we perform a comparison with a set of methods [12,13,15,18,20–30]. The reason for choosing these methods for comparison is three-fold. Firstly, they belong to both categories in the literature. Secondly, they represent different techniques. Lastly, these methods represent older and new techniques in the literature. We also compare our method with [19]. The considered approach is reinforcement learning for training image captioning methods. The comparison with this method would highlight the generalization capability of our approach.

3. A Hybrid Feature Model

3.1. Problem Definition

The basic idea of any ZSL approach is to build a model which maps information from the seen to unseen classes based on a semantic description of the unseen classes. In other word, zero-shot learning is needed when there are no labeled training examples for all classes under observation. Therefore, the available dataset is split into two groups, a training subset (seen classes) $Y_{seen} = \{y^1_{seen}, y^2_{seen}, \ldots, y^n_{seen}\}$, and unseen classes $Y_{unseen} = \{y^1_{unseen}, y^2_{unseen}, \ldots, y^m_{unseen}\}$ subset, where n refers to the number of seen classes and m refers to the number of unseen classes. In addition, the assumption $Y_{seen} \cap Y_{unseen} = \phi$ should hold. In such a situation, the task is to build a model $\mathbb{R}^d \rightarrow Y_{unseen}$ using only the training subset and able to classify the unseen classes. Afterward, the trained classifier should be applied on test data of unseen classes under the zero-shot settings $Y_{seen} \cap Y_{unseen} = \phi$. Consequently, zero-shot learning provides a new technique to overcome obstacles, such as the lack of training examples aiming at increasing a learning system's capability to deal with unexpected events in the same way that people do.

Most state-of-the-art techniques solve the ZS problem by embedding the training data feature space and the semantic representation of class labels in some vector space to preserve the similarity. Then, unseen classes can be classified as nearest-neighbor search problems. In the generalized zero-shot case, we seek to design a more generic model $\mathbb{R}^d \rightarrow Y_{seen} \cup Y_{unseen}$, that is able to categorize or classify the seen and unseen classes appropriately.

3.2. Approach

The Variational Autoencoder [32] consists of a decoder and an encoder. The encoder and the decoder are trained to aim at maximizing a goal which is known as the Evidence Lower Bound (ELBo). In both the encoder and the decoder, the variable z represents the hidden, latent space and the variable x represents the data. In addition, the encoder $q_\Phi(z|x)$ consists of parameters Φ and maps from data space to latent space and a decoder $p_\theta(x|z)$ which consists of the parameters θ and maps from latent space to data space. The lower bound for $p(x)$ can be written as:

$$\mathcal{L}(\Phi, \theta; x) = -KL(q_\Phi(z|x)||p_\theta(z)) + \mathbb{E}_{q\Phi(z|x)}[log p_\theta(x|z)] \quad (1)$$

In Equation (1), KL denotes the Kullback–Leibler divergence between the encoder's distribution $q_\Phi(z|x)$ and $p_\theta(z)$.

Conditional Variational Autoencoders (CVAE) [33] consists of the encoder and the decoder that can be conditioned to additional variables like the variable x (data) and the condition variable c. Thus, it is possible to generate samples following desired properties that might be encoded by c also. The loss function can be given as:

$$\mathcal{L}(\Phi, \theta; x, c) = -KL(q_\Phi(z|x,c)||p_\theta(z|c)) + \mathbb{E}_{q\Phi(z|c)}[logp_\theta(x|z,c)] \quad (2)$$

In this work, our loss function considers only the reconstruction term which is the Mean Squared Error (MSE).

We chose to use such a loss function because researchers in [34–36], showed that the KL divergence in the standard conditional variational autoencoder (see Equation (1)) does not allow the model to use the latent variables in many situations effectively. In this paper, we show that dropping the Kullback–Leibler (KL) term from the Variational Autoencoder [32] shows promising performance.

Algorithm 1 shows the training steps. Firstly, the algorithm requires the image features X_{seen}, the labels of the image features (visual space) Y_{seen}, and the vectors of the semantic space S_{seen}. Then the first autoencoder $Autoencoder_1$ is trained using X_{seen} combined with S_{seen} and learns the latent space z to generate $\hat{x}$ given S_{seen}. Then the second autoencoder $Autoencoder_2$ is trained using the X_{seen} and learns the latent space z to generate $\hat{X}_{seen}$ given S_{seen}.

Algorithm 1 Training

Require: $X_{seen}, Y_{seen}, S_{seen}$
Ensure: $Autoencoder_1, Autoencoder_2$
Train the conditional model ($Autoencoder_1, condition\ \ is\ \ S_{seen}$) $(X_{seen}, S_{seen} \rightarrow X_{seen})$
Train the conditional model ($Autoencoder_2, condition\ \ is\ \ S_{seen}$) $(X_{seen} \rightarrow \ \ X_{seen})$

Algorithm 2 shows the detailed steps to classify the unseen classes. The algorithm requires the first autoencoder $Autoencoder_1$, the second autoencoder $Autoencoder_2$, and the semantic vectors of unseen labels S_{unseen}. Then, the encoder of the first autoencoder $Autoencoder_1$ will estimate $q(z^{(i)}|x^{(i)}, S_{Y_i})$ but the input of the encoder is the image feature concatenated with the semantic vectors. Then, the decoder of $Autoencoder_1$ tries to reconstruct x using a sampled z from a standard normal distribution concatenated with S_{unseen}. Then, the encoder of the second autoencoder $Autoencoder_2$ will estimate $q(z^{(i)}|x^{(i)}, S_{Y_i})$ but the input of the encoder is only the image feature space. Then, the decoder of $Autoencoder_2$ tries to reconstruct x using a sampled z from a standard normal distribution concatenated with S_{unseen}. The generated $\hat{x}$ from both autoencoders will be concatenated to form the pseudo training data for a support vector machine. Then, the Support Vector Machine (SVM) is trained, and its parameters are fitted. We use it to predict the performance using the unseen test classes.

Algorithm 2 Unseen classes classification

Require: $Autoencoder_1, Autoencoder_2, X_{unseen}, S_{unseen}, Y_{unseen}$
Ensure: classLabel
$TrainingSet_{Autoenc_1} = \Phi$
for $y_{unseen} \in Y_{unseen}$ **do**
for i $inNumOfSamples$ **do**
$\#$ $sample$ $from$ a $Gaussian$ $distribution$
z $\sim \mathcal{N}(0,1)$
$\#$ $Concanetate$ z and the $unseen$ $semantic$ $class$ $label$
$tmpV_i = S_{unseen} \circ z$
$\#$ $Generate$ a $pseudo-sample$ $from$ the $first$ $autoencoder$
$PseudoX_i \leftarrow Decoder_{Autoencoder_1}(tmpV_i)$
$\#$ Add the $sample$ and the $unseen$ $class$ $label$ to $TrainingSet_{Autoenc_1}$
$TrainingSet_{Autoenc_1} \leftarrow TrainingSet_{Autoenc_1} \cup (PseudoX_i, y_{unseen})$
end for
end for
$TrainingSet_{Autoenc_2} = \Phi$
for $y_{unseen} \in Y_{unseen}$ **do**
for i $inNumOfSamples$ **do**
$\#$ $sample$ $from$ a $Gaussian$ $distribution$
z $\sim \mathcal{N}(0,1)$
$\#$ $Concanetate$ z and the $unseen$ $semantic$ $class$ $label$
$tmpV_i = S_{unseen} \circ z$
$\#$ $Generate$ a $pseudo-sample$ $from$ the $second$ $autoencoder$
$PseudoX_i \leftarrow Decoder_{Autoencoder_2}(tmpV_i)$
$\#$ Add the $sample$ and the $unseen$ $class$ $label$ to $TrainingSet_{Autoenc_2}$
$TrainingSet_{Autoenc_2} \leftarrow TrainingSet_{Autoenc_2} \cup (PseudoX_i, y_{unseen})$
end for
end for
$S_{training} = TrainingSet_{Autoenc_1} \cup TrainingSet_{Autoenc_2}$
fit SVM model using $S_{training}$
Use the trained SVM model
classLabel = SVM(X_{unseen})

4. Experiments

In the field of ZSL, there are well-known benchmark datasets. Therefore, we selected four of them to test the performance of the proposed approach. We used, SUN Attribute (SUN) dataset [37] which consists of 14340 images, 645 classes are seen and 72 unseen. Caltech-UCSD-Birds (CUB) [38] which consists of 11788 images, 150 classes are seen and 50 unseen. In addition, we used Animals with Attributes1 and Animals with Attributes2 (AwA-1) and (AwA-2) [39] datasets. AwA-1 consists of 30475 images, 40 classes are seen and 10 unseen. AwA-2 dataset consists of 37322 images, 40 classes are seen and 10 unseen.

Figures 2–4 show examples from AwA, CUB and SUN datasets, respectively.

Figure 2. Examples from Animals with Attributes (AWA) dataset.

Figure 3. Examples from Caltech-UCSD-Birds (CUB) dataset.

Figure 4. Examples from SUN Attribute (SUN) dataset.

Regarding the visual space, we explored the Residual Neural Network 101 (ResNet101) features [39]. Concerning the semantic space, we rely on the semantic space vectors given by the authors of those datasets. Both autoencoders have a dense layer, followed by a dropout and a second dense layer. This is followed by another layer, which generates the values z. Activation functions are ReLU, and the activation functions for the last layer for both the encoder and the decoder are linear. In addition, we use the keras [40] framework in combination with the tensorflow backend [41] for implementation.

In our model, hyper-parameters are divided into two categories. The network hyper-parameters and the Support Vector Machine (SVM) cost parameter. The network hyper-parameters are set to batch size equal to 50, the size of the latent variable is 50, and the optimizer is Adam [42]. The number of generated samples for each class is equal to 200. Cross-validation on training classes is used to determine the latent variable size. The SVM cost parameter is set to 100. To calculate the overall accuracy, we used the per-class average:

$$acc_{average}^{per-class} = \frac{1}{|Y|}\Sigma_{i=0}^{|Y|}\left(\frac{N_{correct_{class}}^{class_i}}{N_{Total}^{class_i}}\right) \quad (3)$$

Regarding the GZSL, we explored the generalized zero-shot situation [43]. We kept aside 20% of the data from the training images and trained the model using the remaining 80% of the data. The SVM is trained using both the seen and the unseen classes to avoid biased performance toward seen classes. For Generalized Zero-Shot Learning (GZLS), we followed the recommendation in [44] to consider the harmonic mean of the accuracy between seen and unseen classes.

5. Results and Discussion

Table 1 shows the state-of-the-art comparison on four datasets using per-class average and the suggested splits from [39]. Our HFM model shows classification scores of 69.5%, 65.0%, 65.5%, and 53.8% on CUB, AwA1, AwA2, and SUN, respectively. For the ZSL settings, Table 1 shows that f-VAEGAN-D2 [25] and TF-VAEGAN [26] performed the best for AwA2 and SUN datasets. However, our model outperforms them using the CUB dataset. This result is promising because our model showed an improved performance using the highly fine-grained CUB dataset, which means that the generated pseudo-images gave separable output space. We attribute this to excluding Kullback Leibler divergence and to

the hybrid nature of our reconstructed image feature space. Unfortunately, the authors of f-VAEGAN-D2 and Tf-VAEGA did not provide any results related to AwA1 dataset.

As shown in Figure 5, we visually inspect the image feature vectors produced by our model for each class using the t-SNE [45] technique, and we compare them to the original test image feature vectors for the AwA-1 dataset. As a result, we could observe that the proposed approach can accurately simulate the underlying images. In addition, we could observe that the reconstructed image features did not exclude many modes compared to the real distribution.

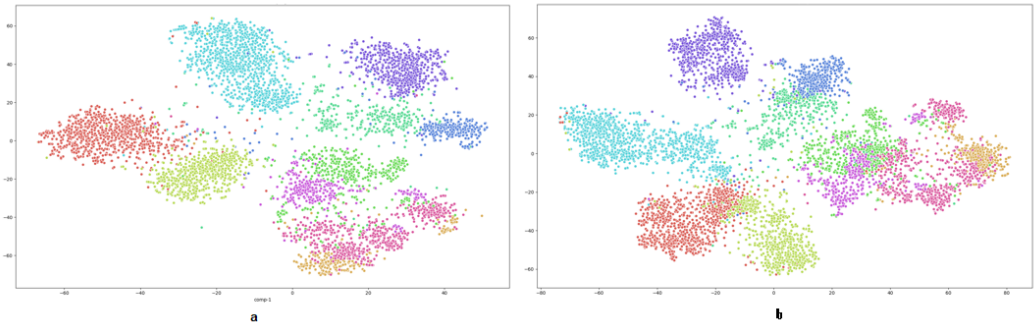

Figure 5. This figure visualizes the image feature space of the AwA-1 dataset (each color denotes a label), (**a**) shows t-SNE real test data visualization and (**b**) shows the test data generated from the proposed approach.

Table 1. State-of-the-art comparison on four datasets using the per-class average under the ZSL setting.

Model	CUB	AwA1	AwA2	SUN
DAP [12]	40.0	44.1	46.1	39.9
IAP [12]	24.0	35.9	35.9	19.4
ConSE [13]	34.3	45.6	44.5	38.8
CMT [18]	34.6	39.5	37.9	39.9
SSE [19]	43.9	60.1	61.0	51.5
DeViSE [15]	52.0	54.2	59.7	56.5
SJE [20]	53.9	65.6	61.9	53.7
LATEM [21]	49.3	55.1	55.8	55.3
ESZSL [22]	53.9	58.2	58.6	54.5
ALE [23]	54.9	59.9	62.5	58.1
SYNC [28]	55.6	54.0	46.6	56.3
SAE [29]	33.3	53.0	54.1	40.3
Relation Net [30]	55.6	68.2	64.2	-
DEM [24]	51.7	68.4	67.1	61.9
f-VAEGAN-D2 [25]	61.0	—	71.1	64.7
TF-VAEGAN [26]	64.9	—	72.2	66.0
CVAE [27]	52.1	71.4	65.8	61.7
HFM (Ours)	69.5	65.0	65.5	53.8

Table 2 shows the result of the Generalized Zero-Shot Learning (GZSL) compared to the well-known state-of-the-art approaches. The table shows comparable performance for the CUB and AwA2 dataset. However, the proposed approach showed better performance using AwA1 dataset. Table 2 shows that our HFM model has a harmonic mean score of 43.4%, 61.6%, 63.4%, and 29.7% on CUB, AwA1, AwA2, and SUN, respectively. The results of the Generalized Zero-Shot learning can be explained because of using ELBo without KL divergence (KL-free) is still theoretically a valid target for generative modeling using VAEs [35].

Table 2. Results of Generalized Zero-Shot Learning (GZSL) settings.We used the harmonic mean of accuracy on both seen and unseen classes as a measure.

Model	CUB	AwA1	AwA2	SUN
DAP [12]	3.3	0.0	0.0	7.2
IAP [12]	0.4	4.1	1.8	1.8
ConSE [13]	3.1	0.8	1.0	11.6
CMT [18]	8.7	15.3	15.9	13.3
SSE [19]	14.4	12.9	14.8	4.0
DeViSE [15]	32.8	22.4	27.8	20.9
SJE [20]	33.6	19.6	14.4	19.8
LATEM [21]	24.0	13.3	20.0	19.5
ESZSL [22]	21.0	12.1	11.0	15.8
ALE [23]	34.4	27.5	23.9	26.3
SYNC [28]	19.8	16.2	18.0	13.4
SAE [29]	13.6	3.5	2.2	11.8
Relation Net [30]	47.0	46.7	45.3	—
DEM [24]	29.2	47.3	45.1	25.6
f-VAEGAN-D2 [25]	53.6	—	63.5	41.3
TF-VAEGAN [26]	58.1	—	66.6	43.0
CVAE [27]	34.5	47.2	51.2	26.7
HFM (Ours)	43.4	61.6	63.4	29.7

Table 3 shows the results for every autoencoder on four datasets under the ZSL setting. The results of the table confirm that combining the image feature spaces that are generated using both autoencoders improved the overall performance significantly.

Table 3. Results for each autoencoder on four datasets under the ZSL setting. The performance is evaluated using the per-class average.

Dataset	Autoencoder_1	Autoencoder_2	Both
AWA1	63.6	60.0	65.0
AWA2	58.6	58.4	65.5
CUB	68.5	58.9	69.5
SUN	50.6	51.4	53.8

Table 4 shows the results of the Generalized zero-shot setting (GZSL) that are calculated based on per-class average using seen classes, unseen classes, and harmonic mean.

Table 4. Results of Generalized Zero-Shot setting (GZSL) that are calculated based on per-class average using seen classes, unseen classes, and harmonic mean.

Dataset	Seen	Unseen	Harmonic Mean
AWA1	75.7	52.0	61.6
AWA2	80.9	49.7	63.4
CUB	57.9	34.7	43.4
SUN	75.3	18.5	29.7

Furthermore, other recent works, e.g., AFRNet [46] and GEM-ZSL [47] showed competitive results compared to our approach using different experimental settings. In AFRNet [46], authors proposed an adversarial network consisting of a residual generator, a prototype predictor, and a discriminator to synthesize compact semantic visual features for ZSL. Furthermore, authors in GEM-ZSL [47], their goal is the estimation of the real human gaze position to determine the visual attention areas for recognizing an unseen object using the semantic description of attributes. Thus, a feedback module combined with the decoder of each VAE may improve the overall performance of the GZSL problem.

6. Conclusions

Zero-Shot learning is related to building machine learning models that can classify or predict classes (labels) that are not included in the training set. In this work, a generative zero-shot learning model is developed. The model can be extended to different use case scenarios. In addition, this work provided intensive tests and detailed coverage of state-of-the-art technology. According to our results, the model shows promising results in some cases compared to the state-of-the-art methods considering three benchmark datasets, even in the case of generalized zero-shot learning. Our proposed method showed that: (a) excluding the Kullback–Leibler (KL) divergence from the conditional VAE loss synthesizes discriminative image features for classifying unseen classes in ZSL problem settings, (b) Using multiple VAEs generates an improved discriminative image space where data are easier to separate for classification purposes. Moreover, a limitation of the proposed approach is that the proposed model lacks a feedback module that can improve the reconstructed pseudo-image space. In our future work, we will add a feedback module and extend our generative model to combine the generative model with an additional embedding model. It means the model maps both the real and the pseudo-generated samples produced by the generative model into a new embedding space where classes are better separable.

Author Contributions: F.A.M. conceptualization and methodology; H.U. and M.U. formal analysis, F.A.M., H.U. and M.U. wrote the paper. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Yan, T.; Li, H.; Sun, B.; Wang, Z.; Luo, Z. Discriminative Feature Mining and Enhancement Network for Low-resolution Fine-grained Image Recognition. *IEEE Trans. Circuits Syst. Video Technol.* Available online: https://ieeexplore.ieee.org/document/9684445 (accessed on 2 February 2022).
2. Shagdar, Z.; Ullah, M.; Ullah, H.; Cheikh, F.A. Geometric Deep Learning for Multi-Object Tracking: A Brief Review. In Proceedings of the 2021 9th European Workshop on Visual Information Processing (EUVIP), Paris, France, 23–25 June 2021; pp. 1–6.
3. Wu, C.; Li, X.; Guo, Y.; Wang, J.; Ren, Z.; Wang, M.; Yang, Z. Natural language processing for smart construction: Current status and future directions. *Autom. Constr.* **2022**, *134*, 104059. [CrossRef]
4. Ullah, H.; Ahmed, T.U.; Ullah, M.; Cheikh, F.A. IR-SSL: Improved Regularization Based Semi-Supervised Learning For Land Cover Classification. In Proceedings of the 2021 IEEE International Conference on Image Processing (ICIP), Anchorage, AK, USA, 19–22 September 2021; IEEE: Piscataway, NJ, USA, 2021; pp. 874–878.
5. Aljaloud, A.S.; Ullah, H. IA-SSLM: Irregularity-Aware Semi-Supervised Deep Learning Model for Analyzing Unusual Events in Crowds. *IEEE Access* **2021**, *9*, 73327–73334. [CrossRef]
6. Zhao, T.; Wang, Z.; Masoomi, A.; Dy, J. Deep Bayesian Unsupervised Lifelong Learning. *Neural Netw.* **2022**, *149*, 95–106. [CrossRef] [PubMed]
7. Hunter, R.A.; Pompano, R.R.; Tuchler, M.F. Alternative Assessment of Active Learning. In *Active Learning in the Analytical Chemistry Curriculum*; ACS Publications: New York, NY, USA, 2022; pp. 269–295.
8. Biederman, I. Recognition-by-components: A theory of human image understanding. *Psychol. Rev.* **1987**, *94*, 115. [CrossRef] [PubMed]
9. Min, S.; Yao, H.; Xie, H.; Wang, C.; Zha, Z.J.; Zhang, Y. Domain-aware visual bias eliminating for generalized zero-shot learning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 12664–12673.
10. Han, Z.; Fu, Z.; Chen, S.; Yang, J. Contrastive embedding for generalized zero-shot learning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 2371–2381.
11. Zhang, J.; Li, Q.; Geng, Y.A.; Wang, W.; Sun, W.; Shi, C.; Ding, Z. A zero-shot learning framework via cluster-prototype matching. *Pattern Recognit.* **2022**, *124*, 108469. [CrossRef]
12. Lampert, C.H.; Nickisch, H.; Harmeling, S. Attribute-based classification for zero-shot visual object categorization. *IEEE Trans. Pattern Anal. Mach. Intell.* **2013**, *36*, 453–465. [CrossRef] [PubMed]
13. Norouzi, M.; Mikolov, T.; Bengio, S.; Singer, Y.; Shlens, J.; Frome, A.; Corrado, G.S.; Dean, J. Zero-shot learning by convex combination of semantic embeddings. *arXiv* **2013**, arXiv:1312.5650.

14. Gao, R.; Hou, X.; Qin, J.; Shen, Y.; Long, Y.; Liu, L.; Zhang, Z.; Shao, L. Visual-Semantic Aligned Bidirectional Network for Zero-Shot Learning. *IEEE Trans. Multimed.* Available online: https://ieeexplore.ieee.org/document/9693152 (accessed on 2 February 2022).
15. Frome, A.; Corrado, G.S.; Shlens, J.; Bengio, S.; Dean, J.; Ranzato, M.; Mikolov, T. Devise: A deep visual-semantic embedding model. *Adv. Neural Inf. Process. Syst.* **2013**, *26*, 1–9.
16. Annadani, Y.; Biswas, S. Preserving semantic relations for zero-shot learning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 7603–7612.
17. Vyas, M.R.; Venkateswara, H.; Panchanathan, S. Leveraging Seen and Unseen Semantic Relationships for Generative Zero-Shot Learning. In Proceedings of the European Conference on Computer Vision, Glasgow, UK, 23–28 August 2020; Springer: Cham, Switzerland, 2020; pp. 70–86.
18. Socher, R.; Ganjoo, M.; Manning, C.D.; Ng, A. Zero-shot learning through cross-modal transfer. *Adv. Neural Inf. Process. Syst.* **2013**, *26*, 1–10.
19. Zhang, L.; Sung, F.; Liu, F.; Xiang, T.; Gong, S.; Yang, Y.; Hospedales, T.M. Actor-critic sequence training for image captioning. *arXiv* **2017**, arXiv:1706.09601.
20. Akata, Z.; Reed, S.; Walter, D.; Lee, H.; Schiele, B. Evaluation of output embeddings for fine-grained image classification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 2927–2936.
21. Xian, Y.; Akata, Z.; Sharma, G.; Nguyen, Q.; Hein, M.; Schiele, B. Latent embeddings for zero-shot classification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 69–77.
22. Romera-Paredes, B.; Torr, P. An embarrassingly simple approach to zero-shot learning. In Proceedings of the International Conference on Machine Learning, Lille, France, 6–11 July 2015; PMLR: New York, NY, USA, 2015; pp. 2152–2161.
23. Akata, Z.; Perronnin, F.; Harchaoui, Z.; Schmid, C. Label-embedding for image classification. *IEEE Trans. Pattern Anal. Mach. Intell.* **2015**, *38*, 1425–1438. [CrossRef] [PubMed]
24. Zhang, L.; Xiang, T.; Gong, S. Learning a deep embedding model for zero-shot learning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 2021–2030.
25. Xian, Y.; Sharma, S.; Schiele, B.; Akata, Z. f-vaegan-d2: A feature generating framework for any-shot learning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 10275–10284.
26. Narayan, S.; Gupta, A.; Khan, F.S.; Snoek, C.G.; Shao, L. Latent embedding feedback and discriminative features for zero-shot classification. In Proceedings of the European Conference on Computer Vision, Glasgow, UK, 23–28 August 2020; Springer: Cham, Switzerland, 2020; pp. 479–495.
27. Mishra, A.; Krishna Reddy, S.; Mittal, A.; Murthy, H.A. A generative model for zero shot learning using conditional variational autoencoders. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, Salt Lake City, UT, USA, 18–22 June 2018; pp. 2188–2196.
28. Changpinyo, S.; Chao, W.L.; Gong, B.; Sha, F. Synthesized classifiers for zero-shot learning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 5327–5336.
29. Kodirov, E.; Xiang, T.; Gong, S. Semantic autoencoder for zero-shot learning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 3174–3183.
30. Sung, F.; Yang, Y.; Zhang, L.; Xiang, T.; Torr, P.H.; Hospedales, T.M. Learning to compare: Relation network for few-shot learning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 1199–1208.
31. Zhang, T.; Yang, Z.; Li, D. Stochastic simulation of deltas based on a concurrent multi-stage VAE-GAN model. *J. Hydrol.* **2022**, *607*, 127493. [CrossRef]
32. Kingma, D.P.; Welling, M. Auto-encoding variational bayes. *arXiv* **2013**, arXiv:1312.6114.
33. Sohn, K.; Lee, H.; Yan, X. Learning structured output representation using deep conditional generative models. *Adv. Neural Inf. Process. Syst.* **2015**, *28*, 1–9.
34. Bowman, S.R.; Vilnis, L.; Vinyals, O.; Dai, A.M.; Jozefowicz, R.; Bengio, S. Generating sentences from a continuous space. *arXiv* **2015**, arXiv:1511.06349.
35. Zhao, S.; Song, J.; Ermon, S. Towards deeper understanding of variational autoencoding models. *arXiv* **2017**, arXiv:1702.08658.
36. Chen, R.T.; Li, X.; Grosse, R.B.; Duvenaud, D.K. Isolating sources of disentanglement in variational autoencoders. *Adv. Neural Inf. Process. Syst.* **2018**, *31*, 1–11.
37. Patterson, G.; Hays, J. Sun attribute database: Discovering, annotating, and recognizing scene attributes. In Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition, Providence, RI, USA, 16–21 June 2012; IEEE: Piscataway, NJ, USA, 2012; pp. 2751–2758.
38. Wah, C.; Branson, S.; Welinder, P.; Perona, P.; Belongie, S. *The Caltech-Ucsd Birds-200-2011 Dataset*; California Institute of Technology: Pasadena, CA, USA, 2011.
39. Xian, Y.; Schiele, B.; Akata, Z. Zero-shot learning-the good, the bad and the ugly. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 211–26 July 2017; pp. 4582–4591.
40. Bursztein, E.; Chollet, F.; Jin, H.; Watson, M.; Zhu, Q.S. Keras: The Python Deep Learning API. Available online: https://keras.io (accessed on 2 February 2022).

41. Abadi, M.; Agarwal, A.; Barham, P.; Brevdo, E.; Chen, Z.; Citro, C.; Corrado, G.S.; Davis, A.; Dean, J.; Devin, M.; et al. TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems. 2015. Available online: tensorflow.org (accessed on 2 February 2022).
42. Kingma, D.P.; Ba, J. Adam: A method for stochastic optimization. *arXiv* **2014**, arXiv:1412.6980.
43. Chao, W.L.; Changpinyo, S.; Gong, B.; Sha, F. An empirical study and analysis of generalized zero-shot learning for object recognition in the wild. In Proceedings of the European Conference on Computer Vision, Amsterdam, The Netherlands, 8–16 October 2016; Springer: Amsterdam, The Netherlands, 2016; pp. 52–68.
44. Xian, Y.; Lorenz, T.; Schiele, B.; Akata, Z. Feature generating networks for zero-shot learning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 5542–5551.
45. Van der Maaten, L.; Hinton, G. Visualizing data using t-SNE. *J. Mach. Learn. Res.* **2008**, *9*, 2579–2605.
46. Liu, B.; Dong, Q.; Hu, Z. Zero-shot learning from adversarial feature residual to compact visual feature. In Proceedings of the AAAI Conference on Artificial Intelligence, New York, NY, USA, 1–12 February 2020; Volume 34, pp. 11547–11554.
47. Liu, Y.; Zhou, L.; Bai, X.; Huang, Y.; Gu, L.; Zhou, J.; Harada, T. Goal-oriented gaze estimation for zero-shot learning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 3794–3803.