IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 26 June 2023, accepted 2 August 2023, date of publication 9 August 2023, date of current version 16 August 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3303640

RESEARCH ARTICLE

An Integral Projection-Based Semantic Autoencoder for Zero-Shot Learning

WILLIAM HEYDEN, HABIB ULLAH, MUHAMMAD SALMAN SIDDIQUI, AND FADI AL-MACHOT, (Member, IEEE)
Faculty of Science and Technology (REALTEK), Norwegian University of Life Sciences (NMBU), 1430 Ås, Norway

Corresponding author: William Heyden (william.heyden@nmbu.no)

ABSTRACT Zero-shot Learning (ZSL) classification categorizes or predicts classes (labels) that are not included in the training set (unseen classes). Recent works proposed different semantic autoencoder (SAE) models where the encoder embeds a visual feature vector space into the semantic space and the decoder reconstructs the original visual feature space. The objective is to learn the embedding by leveraging a source data distribution, which can be applied effectively to a different but related target data distribution. Such embedding-based methods are prone to domain shift problems and are vulnerable to biases. We propose an integral projection-based semantic autoencoder (IP-SAE) where an encoder projects a visual feature space concatenated with the semantic space into a latent representation space. We force the decoder to reconstruct the visual-semantic data space. Due to this constraint, the visual-semantic projection function preserves the discriminatory data included inside the original visual feature space. The enriched projection forces a more precise reconstitution of the visual feature space invariant to the domain manifold. Consequently, the learned projection function is less domain-specific and alleviates the domain shift problem. Our proposed IP-SAE model consolidates a symmetric transformation function for embedding and projection, and thus, it provides transparency for interpreting generative applications in ZSL. Therefore, in addition to outperforming state-of-the-art methods considering four benchmark datasets, our analytical approach allows us to investigate distinct characteristics of generative-based methods in the unique context of zero-shot inference.

INDEX TERMS Autoencoder, generative modelling, generative regularisation, latent space, linear transformation, semantic embedding, visual projection, zero-shot learning.

I. INTRODUCTION

In a variety of studies, deep learning-based models have gained human-level abilities. However, these gains are conditional on the availability of high-quality and large-scale data. With the exponential growth of new classes in our real world, gathering enormous amounts of data is prohibitively costly and often infeasible. Additionally, annotating a sufficient amount of the data for training purposes for each class is resource intensive. As a consequence, several learning paradigms based on sparsely labelled data have been proposed, including semi-supervised learning, life-long learning, and active learning. These paradigms, however, are limited in their capacity to investigate changes in a small collection of labelled data.

To address these problems, researchers developed embedding-based zero-shot learning (ZSL) models [1], [2], [3], [4]. They considered pre-trained models to evaluate test data of classes that have not been seen during the training stage. These models typically learn a projection function from a feature space to a semantic embedding space (e.g. attribute space). However, such a projection function is only concerned with predicting the training (seen) class semantic representation (e.g. attribute prediction) or classification. When applied to test data, which in the context of zero-shot contains different classes (unseen), these models typically suffer from the domain shift problem. In addition, the embedding-based model's final classification is subordinate to the nearest

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

neighbour (NN) algorithm in the transformed space. The hubness problem is an inherent property of high dimensional data affecting the distribution of occurrences in this projected embedding space [47].

As a solution to these challenges, generative-based approaches [8], [42], [61], [62] were proposed. By generating class samples from available semantic representations, the NN search is reconditioned into supervised classification. Generative-based models alleviate the domain shift bias by generating authentic prototypes of the disjoint domain [63]. In [5], [7], [8], and [9], the researchers are using Generative Adversarial Networks (GANs) and Variational Autoencoders (VAEs), respectively to acquire prototypes of the unseen distribution. GANs are known to easily diverge due to their adversarial nature [12] and VAEs to create blurry output as a result of the Kullback–Leibler divergence between the data and the distribution [13]. This impacts the generative capacity of prototype creation in a separate domain. In general, they are also not invertible, making them less suitable for downstream inference and reconstruction of unknown distributions [75].

Based on the well-known semantic autoencoder model [10] (SAE), we propose the IP-SAE model to project both the visual feature space and the semantic feature space into the latent representation manifold. In addition, we use this model to demonstrate the adaptability of generative-based models for zero-shot inference. Building on autoencoder architecture, the encoder and decoder in our proposed model are multi-modular and share parameters. The encoder maps data into the integrated visual-semantic manifold. The decoder performs reconstruction of the original visual-semantic features. This projection alleviates the domain shift problem since the reconstructed samples of unseen domains are generated from the domain-invariant manifold. Furthermore, by adopting suitable regularisation parameters for the transformation function, we mitigate the hubness problem. We show that the generative abilities for zero-shot are a tradeoff between stable transferability interpolating domains and performance in specific disjoint spaces. Our overall contributions can be summarised as follows:

- The proposed IP-SAE model uses an analytical solution to tackle the problem of ZSL. It has only one hyperparameter that should be tuned to increase the overall performance. Consequently, our results are reproducible, and the code will be published on GitHub.[1]
- Our IP-SAE model distinguishes from the prior SAE model [10] by reconstructing the latent semantic manifold by enriching the input space and actively using this low-dimension semantic manifold as a regularisation for inference of samples.
- Average per-class accuracy is frequently reported in zero-shot learning state-of-the-art [57], and the obtained results using the IP-SAE model show very high performance. In contrast to the state-of-the-art, we additionally use precision and recall to evaluate the model's end-to-end performance regarding the generalization to unseen classes and distinguish positive instances in unseen domains.
- We propose a methodology to improve the performance of generative ZSL by first, augmenting the input space to encompass multi-modal features; second, engaging regularisation to establish a complete latent manifold and third, leveraging the multi-model embedding of the disjoint data distribution to produce higher quality samples of the unseen classes.

II. RELATED WORKS

We classify the literature of zero-shot learning into two categories: embedding space-based and feature generation-based.

A. EMBEDDING SPACE-BASED METHODS

Embedding space-based methods exploit a transfer function between semantic, visual, and latent feature space to close the gap between seen and unseen domains. Akata et al. [1] cultivated the relationships between features, attributes, and class embeddings. They were amongst the first to rank the compatibility of the visual image feature and the semantic feature embedding for correct classification, in contrast to prior work by researchers in [23] and [68] who used prediction of the class based on a learned mapping function. Xian et al. [25] extended this work by introducing a collection of mapping functions ranked by a compatibility function between image and class embeddings. Their objective was to construct a piece-wise linear factorization of a non-linear compatibility function that learns the latent selection of components of the visual feature space. References [2], [26], and [29] further explored different manifold structures to increase classification accuracy. The methods [14], [15], [16], [17], and [19] advanced the embedding function of the visual features and the semantic descriptors to a latent space. They proposed techniques to assist the transfer and the designs of appropriate embedding manifolds. Their novelty was to compose a discriminatory representation space that aligns the semantic space with the structure of the visual manifold. These models do not provide a natural mechanism for multiple modalities to be fused and optimized jointly in an end-to-end structure. In more recent work, [18] integrated a contrastive embedding model by learning a projection map from the embedding space and a comparative network to align the semantic and visual descriptors. While [70] used the shared label-space actively in training, to reconstruct the semantic representations.

B. GENERATIVE-BASED METHODS

Generative-based methods work by generating pseudo data of the unseen domain to train a classifier impartial to both domain spaces. Reed et al. [69] predeveloped a GAN architecture effectively allowing for text or descriptive annotations to be translated into visual concepts. They briefly mention the

[1] https://github.com/william-heyden/IP-SAE/

zero-shot capabilities this implies. Mishra et al. [9] employed a similar approach, but expanded to the zero-shot setting exclusively. Using a VAE and conditioning on the semantic feature space, as opposed to the visual space, resulted in improved zero-shot inference. Li et al. [8] took advantage of the assumption that class representation originates from a prototypical space, encoding the relationship. This manifold structure is then learned from the data and used to generate synthetic observations. In Fadi et al. [80] the authors integrated two conditional autoencoders, of both modalities. The hybrid model generates pseudo training data from both decoders, which are then mapped to a final classifier. In a similar fashion, the researchers of the methods [9], [33], [34], [35] learned to consolidate the visual features for unseen classes using semantic information. These methods first learn a generative model considering variational autoencoder (VAE) and Generative adversarial networks (GAN) and then train a classifier using the complete space. With recent advancement made in generative algorithms, the performance of zero-shot architectures structured around generative-based methods has naturally also increased.

Li et al. [37] presented the Boomerang-GAN technique to find bilateral connections in zero-shot learning. They used a multimodal cycle-consistent loss to translate back the engendered features to semantic embeddings. Chou et al. [38] discovered the semantic-to-visual embeddings via a seamless fusion of adaptive and generative learning to investigate the correlation between image features and the corresponding semantic features. They stretched the semantic features of each class by supplementing image-adaptive attention so that the learned embedding could account for inter-class and intra-class variations.

Xian et al. [44] combined VAE and GANs by assembling them into a conditional feature-generating model, called f-VAEGAN-D2, that synthesizes features from class embeddings. The authors [45] proposed the transformation and feedback-VAEGAN model (TF-VAEGAN). In addition to VAEs and GANs, they provided a semantic embedding decoder to reconstruct the embedding space. The decoder is used as a feedback module to improve the output of the Generator of the GAN. Both the f-VAEGAN-D2 model [44] and the TF-VAEGAN model [45] shows competitive performance. However, GANs and their derivatives show training instability, while VAE is more stable [46].

III. THE PROPOSED APPROACH

A. PROBLEM DEFINITION

The fundamental concept of ZSL is to construct a model that learns visual- and/or semantic cues translatable to unseen classes. In other words, generative zero-shot learning is required when all classes under observation lack labelled training instances. As a result, the accessible dataset is divided into two groups: a training subset and a test subset. The training subset represented by seen classes $\mathrm{Y}_{seen} = \{y^1_{seen}, y^2_{seen}, \ldots, y^n_{seen}\}$ and the test subset represented by unseen classes $\mathrm{Y}_{unseen} = \{y^1_{unseen}, y^2_{unseen}, \ldots, y^n_{unseen}\}$. The assumption $\mathrm{Y}_{seen} \cap \mathrm{Y}_{unseen} = \phi$ should hold. In such a situation, the task is to build a model $\mathbb{R}^d \rightarrow \mathrm{Y}_{unseen}$ using only examples of training subsets to classify unseen classes. Afterwards, the trained classifier should be applied to test data of unseen classes under the zero-shot settings $\mathrm{Y}_{seen} \cap \mathrm{Y}_{unseen} = \phi$. As a result, zero-shot learning offers a novel approach to overcome difficulties such as lack of training examples, with the goal of boosting a learning system's capacity to cope with unexpected situations in the same manner that individuals do.

To retain this similarity, most cutting-edge embedding-based solutions handle the ZSL issue by embedding the training data feature space and the semantic representation of class labels in some shared vector space. Unseen classes are then categorized according to a nearest-neighbour search. In the generalized zero-shot case, we seek to design a more generic model $\mathbb{R}^d \rightarrow \mathrm{Y}_{seen} \cup \mathrm{Y}_{unseen}$, that can categorize/classify the seen and unseen classes appropriately. This implies that the test set contains data samples from both the seen and unseen classes [58].

B. MODEL

We present a novel method to ZSL based on learning a Semantic AutoEncoder (SAE) inspired by [10]. The SAE method encodes the visual feature space of the training data into a semantic space. Their work is based on the assumption that a normal autoencoder is unsupervised leading to the fact that the latent space created by the learning process has no meaningful semantic representation. Therefore, they considered that each data point has a semantic representation and they forced the latent space to represent this semantic feature space. Taking advantage of the generative abilities of the autoencoder, the aim is then to learn visual feature prediction of the unseen classes. Classification of synthesised visual features is very challenging. Hence, the objective of the autoencoder generative space is to be close to the visual data space from unseen classes. The latent representation space is very limiting in its expressive power, given the complex distribution of image spaces. Our proposal implements an improved data space in a concatenation of visual- and semantic data space (fig. 1). This novelty helps to better extract class-discriminative components by increasing the expressive power of the latent representation space. We use a symmetric decoder of the encode-decode architecture to reconstruct an enriched visual-semantic sample space. This provides higher separability for unseen classes, accounting for the hubness problem effectively. We formulate the learning aim of the ZSL method as an optimization problem that minimizes the loss formulated as,

$$\underset{W}{\text{minimize}} ||X - W^T WX||^2_F \quad s.t. \;\; WX = S \tag{1}$$

It represents the loss between the image visual space X and the semantic space S. F is the Frobenius norm and W is the weight. Solving equation (1) with such a hard constraint

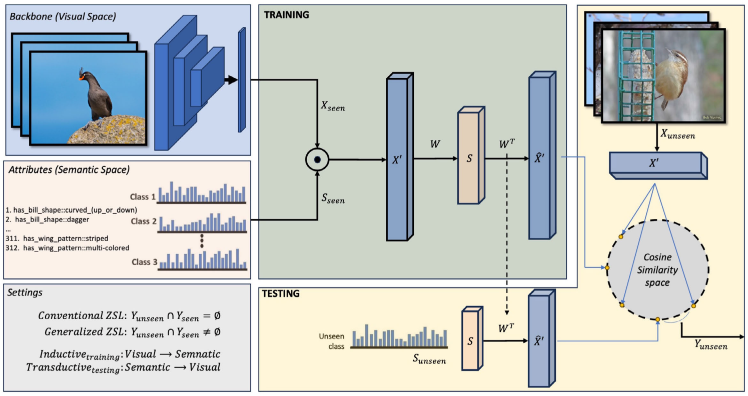
Backbone (Visual Space)
TRAINING
Attributes (Semantic Space)
Class 1
1. has_bill_shape::curved_(up_or_down)
2. has_bill_shape::dagger
...
311. has_wing_pattern::striped
312. has_wing_pattern::multi-colored
Class 2
Class 3
X_{seen}
S_{seen}
X'
W
S
W^T
$\hat{X}'$
X_{unseen}
X'
Cosine Similarity space
Settings
Conventional ZSL: $Y_{unseen} \cap Y_{seen} = \emptyset$
Generalized ZSL: $Y_{unseen} \cap Y_{seen} \neq \emptyset$
$Inductive_{training}$: Visual ⟶ Semnatic
$Transductive_{testing}$: Semantic ⟶ Visual
TESTING
Unseen class
S_{unseen}
S
W^T
$\hat{X}'$
Y_{unseen}
Bob Vuxinic

FIGURE 1. The architecture of our proposed model. By enriching the input data space of the encoder, we are able to ensure surjectivity of the latent representational semantic space. As a result, we achieve enhanced quality and coverage of the reconstructed visual-feature space, leading to improved performance in nearest neighbor classification.

$WX = S$ is not trivial. Therefore, the constraint can be relaxed into a softer one and the objective can be written as,

$$\underset{W}{\text{minimize}} ||X - W^T S||_F^2 + \lambda ||WX - S||_F^2 \tag{2}$$

where λ is a weighting coefficient that controls the importance of the first and second terms. The symmetric terms correspond to the losses of the decoder and encoder, respectively. Equation (2) has a quadratic form that can have a globally optimal solution. Therefore, taking the derivative of equation (2) and setting it to zero, can lead to the analytical solution of this optimization problem [10]. The final solution can be formulated into the well-known Sylvester equation as,

$$AW + WB = C; \tag{3}$$

where $A = SS^T, B = \lambda XX^T$, and $C = (1 + \lambda)SX^T$ such that A and B are positive semi-defined. Equation (3) has an analytical solution that may be solved with efficiency using the Bartels-Stewart method [48].

We propose that the visual representation space X can be replaced by a concatenated version consisting of both the semantic representation space S and the visual representation space X. It can be formulated as,

$$X' = X \oplus S; \tag{4}$$

To take into account the concatenation modelling, we reformulate equation (3), where $A = SS^T, B = \lambda X'X'^T$, and $C = (1 + \lambda)SX'^T$. The importance of the concatenation is evident when we model the projection function (the decoder) as a linear ridge regression resulting in the formulation,

$$\underset{W}{\text{minimize}} ||X' - WS||_F^2 + \lambda ||W||_F^2 \tag{5}$$

The L2 norm calculates the distance of the vector coordinates from the origin of the vector space. It is well known that ridge regression has a closed-form solution $W = X'S^T(SS^T + \lambda I)^{-1}$. Thus, following the matrix norm properties:

$$\begin{aligned} ||WS||_2 &= ||X'S^T(SS^T + \lambda I)^{-1}S||_2 \\ &\leq ||X'||_2 ||S^T(SS^T + \lambda I)^{-1}S||_2 \end{aligned} \tag{6}$$

Using singular value decomposition (SVD), we can write,

$$||S^T(SS^T + \lambda I)^{-1}S||_2 = \frac{\alpha^2}{\alpha^2 + \lambda} \leq 1 \tag{7}$$

where α is the largest singular value of S. So we have $||WS||_2 \leq ||X'||_2$. Consequently, the mapped source data $||WS||_2$ is anticipated to be nearer to the origin of the space in relation to the target data $||X'||_2$ for the decoder, and vice versa for the encoder [27]. Therefore, this would consolidate the grouping of the data around the semantic space of a class and facilitate a clear separation between the classes.

It is important to note that as the attributes of the semantic space (e.g., hand-annotated descriptive features) are sparse, matrix S will be rank-deficient. The enhanced representation space will therefore, by the rank-nullity theorem of inequality $\dim(ker\text{XS}) \leq \dim(ker\text{X}) + \dim(ker\text{S})$, result in the same being true for the visual space matrix X'. Therefore, applying Bartels-Stewart algorithm [48] to equation (3) can no longer guarantee a unique solution, as at least d $- \, rank(\text{S})$ with d representing semantic dimensions, of the similarity matrices will be zero eigenvalues. The enriched space results in a higher-conditioned system sensitive to our choice of λ. This behaviour is evident in figure (5).

1) STANDARD ZERO-SHOT LEARNING

In the standard zero-shot setting our aim is to detect the classes of unseen data. The algorithm's output is the image's class label which is always an unseen class. The first step is to find W which is the result of solving equation (3). Then, we use W^T to project prototypes of unseen enriched visual-semantic space from the latent representations. The final step for classification is then to calculate the cosine similarity between the projected visual space and the true visual space and label according to the most similar (top one) index label. Our approach handles the challenge of disjoint domain raised in the standard setting by enriching the input data space. The projection matrix obtained is therefore consistent with the semantic space provided, irrespective of the domain. Overfitting the data space is challenging for generative models in the convectional ZSL setting. Our proposed model solves this by estimating the lambda attaining optimal orthogonality in the encoded latent representation space. By trading off generative capabilities with information caption our model, and by extension generative models, are able to increase performance in this setting.

2) GENERALIZED ZERO-SHOT LEARNING

In the more realistic generalized zero-shot setting (GZSL) we are presenting samples from both the seen and the unseen domain at testing $\mathrm{Y}_{seen} \cap \mathrm{Y}_{unseen} \neq \phi$. We extract 20% of data samples from the seen classes and mix them with the data samples from the unseen classes. In the generalized setting there is an inherent challenge of seen bias [52], where the classes from the seen domain are significantly more represented in the final classification. Our approach alleviates this challenge also through a regularising lambda. The surjective properties of our proposed encoder ensure that the latent representation space is entirely mapped by every element of the visual feature space. Correcting for desired behaviour through lambda will align the information structure embedded in visual- and semantic- feature space. This structural alignment is directly transferable to the unseen domain, ensuring the quality and completeness of the synthetic visual-feature manifold. In generalized zero-shot learning the alignment of seen and unseen domain corrects for the domain shift issue in the projection space [6]. In our proposed enriched visual-semantic data space the structured projection space is further discriminated through increased class-wise distances. As theoretically proven in equation (7), this is shown in figure (2) with the reduced intra-class space for each cluster (omitting class label 6 and 7).

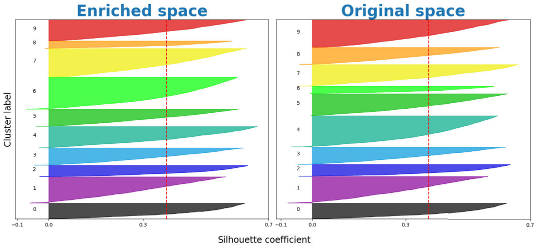

FIGURE 2. Silhouette plotting the discriminative properties of enriched space of AwA2. "Enhanced space" reference our proposed visual-semantic space whereas "original space" corresponds to the visual feature space.

IV. EXPERIMENTS ANALYSIS

A. DATASETS

To evaluate the performance of our proposed IP-SAE method, we consider four benchmark datasets. SUN Attribute (SUN) [53] dataset, the CUB-200-2011 Bird (CUB) [55] dataset, and the AwA-1 and AwA-2 [54] datasets. The SUN dataset consists of 14340 images where 645 classes are seen and 72 are unseen. The AwA-1 dataset consists of 30475 images where 40 classes are seen and 10 are unseen. The AwA-2 dataset consists of 37322 images with 40 seen and 10 unseen classes. The CUB-200-2011 dataset consists of 11788 images where 150 classes are seen and 50 are unseen bird species. Each image additionally has 312 lower-level binary variables indicating visual properties (colour, pattern, form) of specific regions (beak, wings, tail, etc.). But attribute annotations are noisy. To denoise attributes, we used the concept bottleneck models [56]. We only counted the attributes as present if they were in at least 50% of the images of the same class. Therefore, 200 lower-level features were chosen.

In accordance with the norm of published research in zero-shot learning, we report the average per class top-1 accuracy to calculate the overall accuracy.

$$\mathrm{acc}_{average}^{per\text{-}class} = \frac{1}{|Y|}\sum_{i=0}^{|Y|}\left(\frac{N_{correct\_class}^{class_i}}{N_{Total}^{class_i}}\right) \tag{8}$$

In the conventional setting only the accuracy of unseen $class_i\ \forall i \in Y_{unseen}$ while in the generative setting, we calculate the harmonic mean $H = \frac{2 \cdot A_u \cdot A_s}{A_u + A_s}$ between seen and unseen $class_i\ \forall i \in Y_{seen}$ [35]. In addition, we postulate for generative-based models in zero-shot classification the recall and precision are indispensable performance metrics, verifying a more nuanced evaluation of generative capabilities in the unseen domain.

$$recall = \frac{TP}{TP + FN}, \quad precision = \frac{TP}{TP + FP} \tag{9}$$

B. RESULTS

For the visual space, we explored Resnet101 as the backbone architecture for the extracted features [57]. Concerning the semantic space, we rely on the semantic space vectors given by the authors of respective datasets. Regarding the GZSL, we looked at the empirical situation [58]. To eliminate the performance bias of the mapping, the nearest neighbour is selected using the cosine similarity. We apply equal regularisation value for the parameter λ across all datasets to retain comparative behaviour. It can empirically be shown that by tuning the regulator to specific data manifolds a mutual orthogonal transformation matrix can be derived for the coarse dataset which will achieve near-perfect classification.

Table 1 illustrates the results under the conventional zero-shot setting, where the test data is disjoint from training. The results are reported with the suggested splits from [57]. Our proposed IP-SAE method outperformed the state-of-the-art methods by a high margin considering all four benchmark datasets. In Table 2, we partition the results of our proposed method to consider the generalized zero-shot settings. Among existing methods, [57], the classification accuracy of seen classes only is comparable with state-of-the-art methods across datasets. Considering unseen classes only,

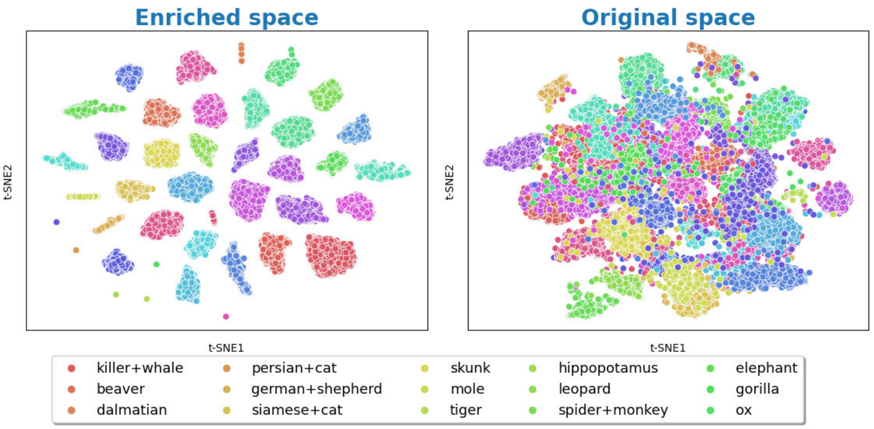

FIGURE 3. Dataset visualization with a selection of labels. We present the visualization of the AwA2 dataset using T-SNE projection of features. The feature-semantic projection is shown on the left side and the semantic projection is shown on the right side. "Enhanced space" reference our proposed visual-semantic space, whereas "original space" corresponds to the visual feature space.

TABLE 1. The results of the disjoint assumption zero-shot setting in conjunction with the per-class accuracy measure.

Model	CUB	AWA1	AWA2	SUN
DAP(PAMI'13) [17]	40.0	44.1	46.1	39.9
IAP(PAMI'13) [17]	24.0	35.9	35.9	19.4
ConSE(arXiv'13) [22]	34.3	45.6	44.5	38.8
CMT(NeurIPS'13) [23]	34.6	39.5	37.9	39.9
SSE(arXiv'17) [65]	43.9	60.1	61.0	51.5
DeViSE(NeurIPS'13) [24]	52.0	54.2	59.7	56.5
SJE(CVPR'15) [2]	53.9	65.6	61.9	53.7
LATEM(CVPR'16) [25]	49.3	55.1	55.8	55.3
ESZSL(ICML'15) [26]	53.9	58.2	58.6	54.5
ALE(PAMI'15) [1]	54.9	59.9	62.5	58.1
SYNC(CVPR'16) [3]	55.6	54.0	46.6	56.3
SAE(CVPR'17) [10]	33.3	53.0	54.1	40.3
Relation Net(CVPR'18) [31]	55.6	68.2	64.2	-
DEM(CVPR'17) [27]	51.7	68.4	67.1	61.9
f-VAEGAN-D2(CVPR'19) [44]	61.0	—-	71.1	64.7
TF-VAEGAN(ECCV'20) [45]	64.9	—-	72.2	66.0
CVAE(CVPR'18) [9]	52.1	71.4	65.8	61.7
GEM-ZSL(CVPR'21) [66]	77.8	—-	67.3	62.8
AFRNet(AAAI'20) [67]	50.3	76.4	75.1	64.0
TransZero(AAAI'22) [77]	76.8	—-	70.1	65.6
JG-ZSL(MDPI'23) [78]	72.5	70.6	69.4	60.3
HRT(ECCV'23) [79]	71.7	—-	67.3	63.9
(Ours)	**80.1**	**92.9**	**82.0**	**94.4**

TABLE 2. Results of Generalized Zero-Shot setting (GZSL) that are calculated based on the accuracy of seen classes, unseen classes, and harmonic mean.

Dataset	%SeenClasses	Seen	Unseen	Harmonic Mean
AWA1	20%	91.6	12.0	21.3
AWA2	20%	89.4	29.2	44.0
SUN	20%	83.7	84.5	84.1
CUB	20%	81.6	67.3	73.7

there is a significant improvement compared to similar methods for the fine-grained dataset CUB and SUN. The lower-level binary attributes depicting the visual characteristics of the data space are captured of higher quality in our proposed enriched visual-feature space. For the coarser annotated dataset AwA1 and AwA2, the performance is comparable to related embedding-based methods.

Furthermore, to highlight the impact of our proposed modelling, we present the data visualization in figure (3). We use the t-SNE [59] approach to visually analyse the image feature vectors generated by our model for each class and compare them to the original image feature vectors for the AwA2 dataset. As can be seen, the original space (right) shows a separable feature space but high overlapping. In contrast to that, projecting the semantic space to the visual−semantic feature representation shows a clear separation between the classes (left). At the same time, the data will be grouped around the semantic space of a class which would overcome the problem of forming hubs. Consequently, data samples that belong to same class are well separable.

The confusion matrix in figure (4) reports a summary of the prediction of unseen classes in a matrix form [74]. We see a distinct main diagonal for our enriched visual-feature space (left) compared to the original space (right). This suggests that our proposed method is able to correctly identify and classify the true label from the projected latent representation space. Given the confusion matrix, we can extract the recall and precision measurements, as displayed in table 3. Here, we show that our proposed model has better precision and excels in classifying the true classes and correctly identifying false classes. This is universal for all benchmark datasets. This demonstrates that our IP-SAE method's capacity to generate a sample of high quality and the latent manifold is indeed expressive enough to cover the unseen distribution.

V. DISCUSSION

A. GENERATIVE PROPERTIES

The generative aptness of autoencoders decreases when the latent representation space is overcomplete (e.g., higher dimension than the data space itself) [71]. Consequently, the composite function of the transformation matrix of eq. 2, $(W^T \circ W)$ the identity map of the visual space itself cannot be bijective. Implying we will never be able to recover the true visual feature space. It follows from functional analysis W is a surjective transformation function in the encoder [49].

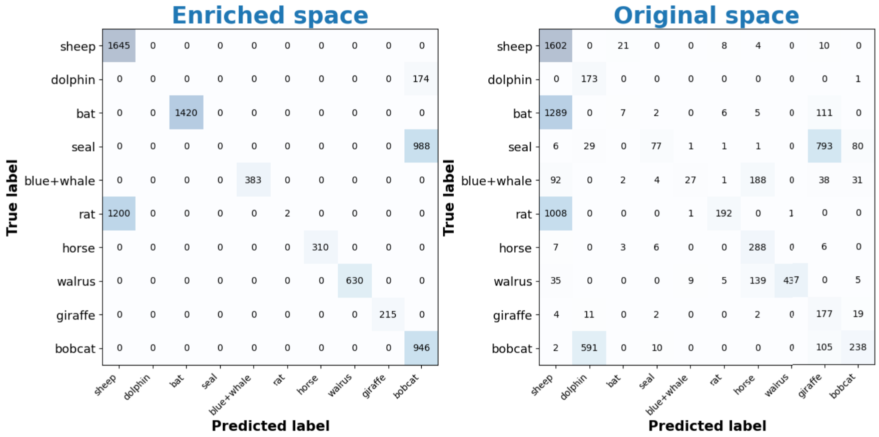

FIGURE 4. confusion matrix of AWA2 dataset. Prediction of labels according to top-1 nearest neighbour. "Enhanced space" reference our proposed visual-semantic space, whereas "original space" corresponds to the visual feature space.

TABLE 3. Accuracy measurement for respective datasets in conventional zero-shot learning setting.

	Enriched			*Original*		
	AWA1	CUB	SUN	AWA1	CUB	SUN
Precision	0.6229	0.8032	0.9166	0.6044	0.3135	0.4805
Recall	0.7359	0.8587	0.9444	0.4571	0.2150	0.3479
F1	0.6568	0.8197	0.9259	0.3902	0.1924	0.3557

Suppose $f : V \rightarrow S$ such that f is the transformation function that maps from the visual feature space V to the semantic representation space S. In our enriched visual feature space, we can show that $\{f^{-1}(s) | s \in S\}$ is none empty. This implies that the columns of the transformation matrix W are precisely the set of linear combinations to form a spanning set of the complete semantic representation space. This implication is not true for the original visual feature space, as the transformation matrix need not be full row rank. The encoder's surjective premise ensures that the enriched visual space image captures all available semantic information [49]. The effect of this is that we *can* always construct a transformation function $g : S \rightarrow V$ satisfying $g(s) = v \in V \; \forall s \in S$, meaning that $g \circ f$ is indeed an identity map. Hence, it is self-evident from the fact that an identity function is bijective [72] the decoder *can* obtain injective properties. Therefore, the encoding ensures that enough information is preserved to recover the visual space, e.g., unique visual properties can be mapped distinctively from the semantic representation space and vice versa [50]. The analytical solution offered by the Sylvester equation (3) finds the optimal transformation matrix between the visual and the semantic space. By enriching the visual space of the seen domain, we are condensing the distance to the subjective image of the unseen domain, hence improving the embedding quality of semantic space for the unseen domain. This is a trade-off between information preservation of the encoding and the generative capabilities of the decoding. The oscillating behaviour seen in figure (5) occurs due to the reconstructed visual space having unique representations in the semantic representation space and multiple reconstructed depictions. However, by the properties of ridge regularisation and the elegant symmetry in the model, we can alleviate this challenge of high variance by designing the coefficients of the transformation matrix through a lambda. The objective is the optimal trade-off between the encoding and decoding of the available information in the seen domain. For generative models, a low-dimension latent space achieves regulatory properties [73]. Since the seen- and unseen domain share semantic feature representation, the design of an elaborate mapping function or embedding space is critical for transferring knowledge between classes [6]. Our results show that the visual-semantic latent space span should be expressive by the mapping function, e.g., the mapping of available knowledge grants surjectivity. The limitation in our proposed model is that we force the same mapping also to be inversely injective, which is contradictory to autoencoders.

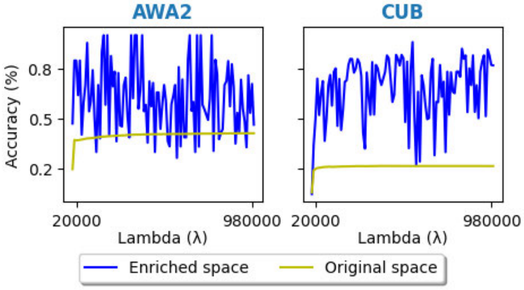

FIGURE 5. Oscillating behaviour in accuracy given variation of lambdas. "Enriched space" reference our proposed visual-semantic space, whereas "original space" corresponds to the visual feature space.

B. REGULARISATION OF λ FOR ZLS

The regularisation term of the loss function (2) characterises with ridge regression. It reduces the condition number of the near singular similarity matrix XX^T of equation (3) by adding a non-negative element to the diagonalizable matrix. Increasing the singular values in the Schur decomposition (of the Bartels-Stewart algorithm) for the visual space reduces noise in the decoding due to the inverse proportional properties in the reversed operation [51]. The ridge behaviour of lambda ensures discriminating visual and semantic structures are preserved in the transformation matrix. Redundant projections of the high variance found in visual space are replaced with the compact semantic space allowing the projection matrix to learn the direction with the lowest variance by increasing the regularisation. Since we are only concerned with the relative position in the visual semantic feature space, an ambiguous loss of information is adequate in the recovery of the manifold by the decoder. We hypothesize that by optimising for λ, the projection matrix favours shared principle components of the visual and semantic space, a result of minimizing W in the loss function (2).

C. ABLATION STUDY

Table 2 evaluates the transferability of our elevated semantic representation space. The encoded representation space is the source of shared information for the seen and unseen domain. The efficiency of the transformation matrix W introduces a compromise between I) A smooth embedding of the semantic manifold to leverage available domain (seen and unseen) information. II) A stable projection into the visual manifold for discriminative, high-quality sample distribution. This (dis)entanglement of the transformation matrix is due to the symmetric in our model (objective function 2). For the encoder, we see that for the finer-grained datasets SUN and CUB, the discrepancies in the accuracy of seen and unseen classes are less. For coarser datasets (AwA1 and AwA2), the model efficiently captures relevant information but struggles to exploit it. The lower harmonic mean shows this. In terms of generative abilities in a disjoint domain, this trivially implies that the reduced semantic information stimulated by increased regularisation of the surjective mapping prompts an advantage to injective properties of the projection.

Table 3 shows the influence of an embedding function with a spanning set over the semantic representation space. For our enriched space $recall > precision$ while the original space results in $recall < precision$. This suggests a complete visual feature reproduction space. The generative transformation is not guaranteed to be expressive enough to replicate all unseen classes in the high-dimensional visual feature space. However, capturing transferable information between the disjoint domains shows to assist progressively in terms of its position in reproduced space and coverage of the manifold by the generator. Note that overall recall and precision in our enriched space are still greater than in the original space.

D. PERFORMANCE METRICS

In accordance with the progress made within zero-shot learning, there has been extensive research on comparing state-of-the-art algorithms [57], [63]. To evaluate performance and demonstrate abilities of proposed models, the average class accuracy, eq. (8), have exclusively been included in publications. We argue that the classic definition of accuracy to report generative-based zero-shot learning effectiveness is insufficient. This only discloses the average of per class true positive predictions, which is biased towards sample size and favours less exact spatial arrangement of the representational manifold [76]. To comprehensively capture the impact of generative zero-shot learning, it is recommended to consider two crucial aspects: a) whether the manifold of knowledge transfer adequately covers the real distribution and b) if the quality of generated class prototypes reflects reality. Consequently, we advocate for adopting the definitions of recall and precision. By examining recall and precision, we can gain insights into the generative process and accurately interpret the model's performance.

VI. BROADER IMPACT

In the assessment of an analytically solvable algorithm designed for generative zero-shot learning this research show how enriching and regularising multi-modal spaces affects the generative capabilities of the disjoint domains. For zero-shot learning researchers, the conceptual insights derived can be applied to improve performance across various generative-based methodologies. In addition, this work can be a contrivance for improving any larger-scale recognition systems by reducing the dependency on the label of the data. At last, this work enables high-quality samples to be synthesised of unseen distribution, thereby essentially implicating a new aspect of creativity for generative models.

VII. CONCLUSION

In this study, we proposed IP-SAE zero-shot learning model. The proposed work included extensive testing and coverage of generative qualities for zero-shot projections. In the IP-SAE model on four benchmark datasets, our method outperformed the state-of-the-art methods using the precision, the recall, the f-score and the per-class accuracy evaluation metrics. In addition, it showed high performance in both, the conventional ZSL and the generalized zero-shot setting. We selected the analytical solution to show how generative approaches for zero-shot learning can be enhanced by; expanding the data manifold to ensure completeness of the disjoint domains; and by regularising the latent representation to augment the sample manifold.

In future work, we aim to improve the generalized zero-shot learning model by using a generative model on a project matrix of the visual and semantic features. This will improve the transformation function's orthogonality and generate samples into a new embedding space with more distinct classifications.

REFERENCES

[1] 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.

[2] 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.

[3] S. Changpinyo, W.-L. Chao, B. Gong, and F. Sha, "Synthesized classifiers for zero-shot learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 5327–5336.

[4] G. Xie, L. Liu, F. Zhu, F. Zhao, Z. Zhang, Y. Yao, J. Qin, S. Jiem, and L. Shao, "Region graph embedding network for zero-shot learning," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2020, pp. 562–580.

[5] X. Zhao, Y. Shen, S. Wang, and H. Zhang, "Generating diverse augmented attributes for generalized zero shot learning," *Pattern Recognit. Lett.*, vol. 166, pp. 126–133, Feb. 2023.

[6] C. Wang, S. Min, X. Chen, X. Sun, and H. Li, "Dual progressive prototype network for generalized zero-shot learning," in *Proc. Adv. Neural Inf. Process. Syst.*, 2021, pp. 2936–2948.

[7] 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.

[8] Y. Li and D. Wang, "Zero-shot learning with generative latent prototype model," 2017, *arXiv:1705.09474*.

[9] 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.

[10] E. Kodirov, T. Xiang, and S. Gong, "Semantic autoencoder for zero-shot learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 4447–4456.

[11] F. Pourpanah, M. Abdar, Y. Luo, X. Zhou, R. Wang, C. P. Lim, X.-Z. Wang, and Q. M. J. Wu, "A review of generalized zero-shot learning methods," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 4, pp. 4051–4070, Apr. 2023.

[12] M. Lee and J. Seok, "Regularization methods for generative adversarial networks: An overview of recent studies," 2020, *arXiv:2005.09165*.

[13] D. P. Kingma and M. Welling, "An introduction to variational autoencoders," *Found. Trends Mach. Learn.*, vol. 14, no. 4, 2019, pp. 1–18.

[14] D. Huynh and E. Elhamifar, "Fine-grained generalized zero-shot learning via dense attribute-based attention," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 4482–4492.

[15] 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.

[16] S. Min, H. Yao, H. Xie, C. Wang, Z.-J. Zha, and Y. Zhang, "Domain-aware visual bias eliminating for generalized zero-shot learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 12661–12670.

[17] C. H. Lampert, H. Nickisch, and S. Harmeling, "Attribute-based classification for zero-shot visual object categorization," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 3, pp. 453–465, Mar. 2014.

[18] Z. Han, Z. Fu, S. Chen, and J. Yang, "Contrastive embedding for generalized zero-shot learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 2371–2381.

[19] Y. Fu, T. M. Hospedales, T. Xiang, Z. Fu, and S. Gong, "Transductive multi-view embedding for zero-shot recognition and annotation," in *Proc. Comput. Vis. 13th Eur. Conf.* Zurich, Switzerland: Springer, Sep. 2014, pp. 584–599.

[20] E. Kodirov, T. Xiang, Z. Fu, and S. Gong, "Unsupervised domain adaptation for zero-shot learning," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 2452–2460.

[21] J. Zhang, Q. Li, Y.-A. Geng, W. Wang, W. Sun, C. Shi, and Z. Ding, "A zero-shot learning framework via cluster-prototype matching," *Pattern Recognit.*, vol. 124, Apr. 2022, Art. no. 108469.

[22] 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," 2013, *arXiv:1312.5650*.

[23] R. Socher, M. Ganjoo, C. D. Manning, and A. Ng, "Zero-shot learning through cross-modal transfer," in *Proc. Adv. Neural Inf. Process. Syst.*, 26, 2013, pp. 1–10.

[24] A. Frome, G. S. Corrado, J. Shlens, S. Bengio, J. Dean, M. A. Ranzato, and T. Mikolov, "DeViSE: A deep visual-semantic embedding model," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 26, 2013, pp. 1–11.

[25] 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.

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

[27] 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.

[28] Q. Li, M. Hou, H. Lai, and M. Yang, "Cross-modal distribution alignment embedding network for generalized zero-shot learning," *Neural Netw.*, vol. 148, pp. 176–182, Apr. 2022.

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

[30] Y. Yu, Z. Ji, J. Han, and Z. Zhang, "Episode-based prototype generating network for zero-shot learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 14032–14041.

[31] F. Sung, Y. Yang, L. Zhang, T. Xiang, P. H. S. Torr, and T. M. Hospedales, "Learning to compare: Relation network for few-shot learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 1199–1208.

[32] S. Badirli, Z. Akata, G. Mohler, C. Picard, and M. M. Dundar, "Fine-grained zero-shot learning with DNA as side information," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 34, 2021, pp. 19352–19362.

[33] 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.

[34] 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.

[35] 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.

[36] B. Xu, Z. Zeng, C. Lian, and Z. Ding, "Generative mixup networks for zero-shot learning," *IEEE Trans. Neural Netw. Learn. Syst.*, early access, Jan. 28, 2022, doi: 10.1109/TNNLS.2022.3142181.

[37] 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.

[38] Y. Y. Chou, H. T. Lin, and T. L. Liu, "Adaptive and generative zero-shot learning," in *Proc. Int. Conf. Learn. Represent.*, 2021, pp. 1–11.

[39] P. Ma, H. Lu, B. Yang, and W. Ran, "GAN-MVAE: A discriminative latent feature generation framework for generalized zero-shot learning," *Pattern Recognit. Lett.*, vol. 155, pp. 77–83, Mar. 2022.

[40] Y. Ye, T. Pan, T. Luo, J. Li, and H. T. Shen, "Learning latent representations for generalized zero-shot learning," *IEEE Trans. Multimedia*, vol. 25, pp. 2252–2265, 2023, doi: 10.1109/TMM.2022.3145237.

[41] Z. Lu, Z. Lu, Y. Yu, and Z. Wang, "Learn more from less: Generalized zero-shot learning with severely limited labeled data," *Neurocomputing*, vol. 477, pp. 25–35, Mar. 2022.

[42] 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.

[43] M. R. Vyas, H. Venkateswara, and S. Panchanathan, "Leveraging seen and unseen semantic relationships for generative zero-shot learning," in *Proc. Comput. Vis. 16th Eur. Conf.* Glasgow, U.K.: Springer, Aug. 2020, pp. 70–86.

[44] 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.

[45] S. Narayan, A. Gupta, F. S. Khan, C. G. Snoek, and L. Shao, "Latent embedding feedback and discriminative features for zero-shot classification," in *Proc. Comput. Vis. 16th Eur. Conf.* Glasgow, U.K.: Springer, Aug. 2020, pp. 479–495.

[46] T. Zhang, Z. Yang, and D. Li, "Stochastic simulation of deltas based on a concurrent multi-stage VAE-GAN model," *J. Hydrol.*, vol. 607, Apr. 2022, Art. no. 127493.

[47] M. Radovanović, A. Nanopoulos, and M. Ivanović, "Hubs in space: Popular nearest neighbors in high-dimensional data," *J. Mach. Learn. Res.*, vol. 11, pp. 2487–2531, 2010.

[48] R. H. Bartels and G. W. Stewart, "Solution of the matrix equation AX+ XB= C [F4]," *Commun. ACM*, vol. 15, no. 9, pp. 820–826, 1972.

[49] O. Bretscher, *Linear Algebra With Applications*. vol. 52, Eaglewood Cliffs, NJ, USA: Prentice-Hall, 1997, Ch. 3.

[50] M. Artin, *Algebra*. London, U.K.: Pearson Education. 2011.

[51] G. Strang, *Linear Algebra and Its Applications*. Belmont, CA, USA: Thomson, Brooks/Cole, 2006, p. 260.

[52] J. Song, C. Shen, Y. Yang, Y. Liu, and M. Song, "Transductive unbiased embedding for zero-shot learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 1024–1033.

[53] G. Patterson and J. Hays, "SUN attribute database: Discovering, annotating, and recognizing scene attributes," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 2751–2758.

[54] B. Zhao, Y. Fu, R. Liang, J. Wu, Y. Wang, and Y. Wang, "A large-scale attribute dataset for zero-shot learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2019, pp. 398–407.

[55] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie, "The caltech-UCSD birds-200–2011 dataset," California Inst. Technol., Tech. Rep. CNS-TR-2010-001, 2010.

[56] P. W. Koh, T. Nguyen, Y. S. Tang, S. Mussmann, E. Pierson, B. Kim, and P. Liang, "Concept bottleneck models," in *Proc. Int. Conf. Mach. Learn.*, 2020, pp. 5338–5348.

[57] 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)*, Jul. 2017, pp. 3077–3086.

[58] W. L. Chao, S. Changpinyo, B. Gong, and F. Sha, "An empirical study and analysis of generalized zero-shot learning for object recognition in the wild," in *Proc. Comput. Vis.-ECCV 14th Eur. Conf.* Amsterdam, The Netherlands: Springer, Oct. 2016, pp. 52–68.

[59] L. Van der Maaten and G. Hinton, "Visualizing data using t-SNE," *J. Mach. Learn. Res.*, vol. 9, pp. 1–27, Nov. 2008.

[60] W. Wang, Y. Pu, V. Verma, K. Fan, Y. Zhang, C. Chen, P. Rai, and L. Carin, "Zero-shot learning via class-conditioned deep generative models," in *Proc. AAAI Conf. Artif. Intell.*, Apr. 2018, vol. 32, no. 1, pp. 1–8.

[61] V. K. Verma and P. Rai, "A simple exponential family framework for zero-shot learning," in *Proc. Mach. Learn. Knowl. Discovery Databases, Eur. Conf. (ECML PKDD)*. Skopje, Macedonia: Springer, Sep. 2017, pp. 792–808.

[62] T. Mukherjee and T. Hospedales, "Gaussian visual-linguistic embedding for zero-shot recognition," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2016, pp. 912–918.

[63] X. Sun, J. Gu, and H. Sun, "Research progress of zero-shot learning," *Appl. Intell.*, vol. 51, pp. 3600–3614, Nov. 2021.

[64] L. Zhang, F. Sung, F. Liu, T. Xiang, S. Gong, Y. Yang, and T. M. Hospedales, "Actor-critic sequence training for image captioning," 2017, *arXiv:1706.09601*.

[65] 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.

[66] Y. Liu, L. Zhou, X. Bai, Y. Huang, L. Gu, J. Zhou, and T. Harada, "Goal-oriented gaze estimation for zero-shot learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 3793–3802.

[67] B. Liu, Q. Dong, and Z. Hu, "Zero-shot learning from adversarial feature residual to compact visual feature," in *Proc. AAAI Conf. Artif. Intell.*, vol. 34, no. 7, Apr. 2020, pp. 11547–11554.

[68] M. Palatucci, D. Pomerleau, G. E. Hinton, and T. M. Mitchell, "Zero-shot learning with semantic output codes," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 22, 2009, pp. 1–9.

[69] S. Reed, Z. Akata, X. Yan, L. Logeswaran, B. Schiele, and H. Lee, "Generative adversarial text to image synthesis," in *Proc. Int. Conf. Mach. Learn.*, 2016, pp. 1060–1069.

[70] Y. Liu, X. Gao, Q. Gao, J. Han, and L. Shao, "Label-activating framework for zero-shot learning," *Neural Netw.*, vol. 121, pp. 1–9, Jan. 2020.

[71] Y. Bengio, "Learning deep architectures for AI," *Found. Trends Mach. Learn.*, vol. 2, no. 1, pp. 1–127, 2009.

[72] S. K. Mapa, *Higher Algebra Abstract and Linear*, 11th ed. Kolkata, India: Sarat Book House, 2011, p. 36.

[73] A. Bora, A. Jalal, E. Price, and A. G. Dimakis, "Compressed sensing using generative models," in *Proc. Int. Conf. Mach. Learn.*, Jul. 2017, pp. 537–546.

[74] A. Tiwari, *Artificial Intelligence and Machine Learning for EDGE Computing11*. New York, NY, USA: Academic Press, 2022, Ch 2, pp. 23–32.

[75] K. Kothari, A. Khorashadizadeh, M. D. Hoop, and I. Dokmanic, "Trumpets: Injective flows for inference and inverse problems," in *Proc. Uncertainty Artif. Intell.*, 2021, pp. 1269–1278.

[76] T. Kynkäänniemi, T. Karras, S. Laine, J. Lehtinen, and T. Aila, "Improved precision and recall metric for assessing generative models," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 32, 2019, pp. 1–12.

[77] S. Chen, Z. Hong, Y. Liu, G. S. Xie, B. Sun, H. Li, Q. Peng, K. Lu, and X. You, "Transzero: Attribute-guided transformer for zero-shot learning," in *Proc. AAAI Conf. Artif. Intell.*, vol. 36, no. 1, Jun. 2022, pp. 330–338.

[78] M. Zhang, X. Wang, Y. Shi, S. Ren, and W. Wang, "Zero-shot learning with joint generative adversarial networks," *Electronics*, vol. 12, no. 10, p. 2308, May 2023.

[79] D. Cheng, G. Wang, B. Wang, Q. Zhang, J. Han, and D. Zhang, "Hybrid routing transformer for zero-shot learning," *Pattern Recognit.*, vol. 137, May 2023, Art. no. 109270.

[80] F. A. Machot, M. Ullah, and H. Ullah, "HFM: A hybrid feature model based on conditional auto encoders for zero-shot learning," *J. Imag.*, vol. 8, no. 6, p. 171, Jun. 2022.

WILLIAM HEYDEN received the B.Sc. degree (Hons.) in finance from Grenoble Ecole de Management, Grenoble, France, in 2018, and the M.Sc. degree (Hons.) in data science and business analytics from the University of Westminister, London, U.K., in 2019. He is currently pursuing the Ph.D. degree with the Norwegian University of Life Sciences (NMBU), Ås, Norway, working on the research area of zero-shot learning, deep learning, and representation learning. He is a Teacher Assistant at computer science courses taught at NMBU.

HABIB ULLAH received the M.S. degree in electronics and computer engineering from Hanyang University, Seoul, South Korea, in 2009, and the Ph.D. degree in information and communication technology from the University of Trento, Trento, Italy, in 2015. From 2015 to 2016, he was an Assistant Professor of electrical engineering with COMSATS University Islamabad, Wah Campus, Pakistan. From 2016 to 2020, he was an Assistant Professor with the College of Computer Science and Engineering, University of Ha'il, Ha'il, Saudi Arabia. In 2020, he was a Postdoctoral Researcher with The Arctic University of Norway, Tromsø, Norway. He is currently an Associate Professor with the Norwegian University of Life Sciences, Ås, Norway. His research interests include computer vision and machine learning.

MUHAMMAD SALMAN SIDDIQUI received the bachelor's and master's degrees in mechanical engineering from the National University of Science and Technology, Islamabad, in 2010 and 2012, respectively, and the Ph.D. degree in applied mathematics from the Norwegian University of Science and Technology, Trondheim, in 2018. He was a part-time Research Scientist with the SINTEF Department of Mathematics and Cybernetics, from 2015 to 2019. He held a postdoctoral position with the Building Physics Department, Norwegian University of Science and Technology, from 2019 to 2021. He is currently an Associate Professor with the Department of Mechanical Engineering, Norwegian University of Life Sciences. His research interests include developing novel methods and numerical solvers for engineering, scientific, and technological applications.

FADI AL-MACHOT (Member, IEEE) received the German Diploma degree in computer science from the University of Potsdam, in 2010, the Ph.D. degree in computer science from Klagenfurt University, Austria, in 2013, and the Habilitation degree in applied computer science from the University of Lübeck, Germany, in 2020. He is currently an Associate Professor of machine learning (ML) with the Norwegian University of Life Sciences (NMBU), Norway. His work has been patented and published in peer-reviewed international conferences and journals. His research interests include deep learning, neural-symbolic learning, video understanding, cognitive modeling, and zero/few-shot learning. He is an active reviewer of well-known journals.