Invertible Zero-Shot Recognition Flows

Yuming Shen[1], Jie Qin[2⋆], Lei Huang[2], Li Liu[2], Fan Zhu[2], and Ling Shao[2,3]

[1] eBay
[2] Inception Institute of Artificial Intelligence
[3] Mohamed bin Zayed University of Artificial Intelligence
ymcidence@gmail.com

Abstract. Deep generative models have been successfully applied to Zero-Shot Learning (ZSL) recently. However, the underlying drawbacks of GANs and VAEs (*e.g.*, the hardness of training with ZSL-oriented regularizers and the limited generation quality) hinder the existing generative ZSL models from fully bypassing the seen-unseen bias. To tackle the above limitations, for the first time, this work incorporates a new family of generative models (*i.e.*, flow-based models) into ZSL. The proposed Invertible Zero-shot Flow (IZF) learns factorized data embeddings (*i.e.*, the semantic factors and the non-semantic ones) with the forward pass of an invertible flow network, while the reverse pass generates data samples. This procedure theoretically extends conventional generative flows to a factorized conditional scheme. To explicitly solve the bias problem, our model enlarges the seen-unseen distributional discrepancy based on a negative sample-based distance measurement. Notably, IZF works flexibly with either a naive Bayesian classifier or a held-out trainable one for zero-shot recognition. Experiments on widely-adopted ZSL benchmarks demonstrate the significant performance gain of IZF over existing methods, in both classic and generalized settings.

Keywords: Zero-Shot Learning, Generative Flows, Invertible Networks

1 Introduction

With the explosive growth of image classes, there is an ever-increasing need for computer vision systems to recognize images from never-before-seen classes, a task which is known as Zero-Shot Learning (ZSL) [25]. Generally, ZSL aims at recognizing *unseen* images by exploiting relationships between *seen* and *unseen* images. Equipped with prior semantic knowledge (*e.g.*, attributes [26], word embeddings [37]), traditional ZSL models typically mitigate the *seen-unseen* domain gap by learning a visual-semantic projection between images and their semantics. In the context of deep learning [48,49], the recent emergence of generative models has slightly changed this schema by converting ZSL into supervised learning, where a held-out classifier is trained for zero-shot recognition based on the generated *unseen* images. As both *seen* and synthesized *unseen* images are

⋆ Corresponding author

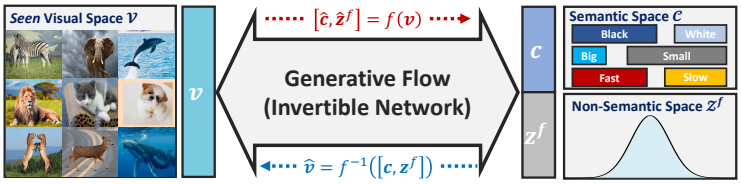

Fig. 1: A brief illustration of IZF for ZSL. We propose a novel factorized conditional **generative flow** with invertible networks.

observable to the model, generative ZSL methods largely favor Generalized ZSL (GZSL) [45] and yet perform well in Classic ZSL (CZSL) [25,36,59]. In practice, Generative Adversarial Networks (GANs) [13], Variational Auto-Encoders (VAEs) [22] and Conditional VAEs (CVAEs) [51] are widely employed for ZSL. Despite the considerable success current generative models [27,38,62,64,71] have achieved, their underlying limitations are still inevitable in the context of ZSL.

First, GANs [13] suffer from mode collapse [5] and instability during training with complex learning objectives. It is usually hard to impose additional ZSL-oriented regularizers to the generative side of GANs other than the real/fake game [46]. Second, the Evidence Lower BOund (ELBO) of VAEs/CVAEs [22,51] requires stochastic approximate optimization, preventing them from generating high-quality *unseen* samples for robust ZSL [64]. Third, as only seen data are involved during training, most generative models are not well-addressing the **seen-unseen bias** problem, *i.e.*, generated *unseen* data tend to have the same distribution as *seen* ones. Though these concerns are as well partially noticed by the recent ZSL research [46,64], they either simply bypass the drawback of GAN in ZSL by resorting to VAE or *vice versa*, which can be yet suboptimal.

Therefore, we ought to seek a novel generative model that can bypass the above limitations to further boost the performance of ZSL. Inspired by the recently proposed Invertible Neural Networks (INNs) [2], we find that another branch of generative models, *i.e.*, flow-based generative models [6,7], align well with our insights into generative ZSL models. Particularly, generative flows adopt an identical set of parameters and built-in network for encoding (*forward pass*) and decoding (*reverse pass*). Compared with GANs/VAEs, the forward pass in flows acts as an additional 'encoder' to fully utilize the semantic knowledge.

In this paper, we fully exploit the advantages of generative flows [6,7], based on which a novel ZSL model is proposed, namely Invertible Zero-shot Flow (IZF). In particular, the forward pass of IZF projects visual features to the semantic embedding space, with the reverse pass consolidating the inverse projection between them. We adopt the idea of factorized representations in [54,57] to disentangle the output of the forward pass into two factors, *i.e.*, semantic and non-semantic ones. Thus, it becomes possible to inject category-wise similarity knowledge into the model by regularizing the semantic factors. Meanwhile, the respective reverse pass of IZF performs conditional data generation with factorized embeddings for

both *seen* and *unseen* data. We visualize this pipeline in Fig. 1. To further accommodate IZF to ZSL, we propose novel bidirectional training strategies to **1)** centralize the *seen* prototypes for stable classification, and **2)** diverge the distribution of synthesized *unseen* data and real *seen* data to explicitly address the bias problem. Our main contributions include:

1. IZF shapes a novel factorized conditional flow structure that supports exact density estimation. This differs from the existing approximated [2] and the non-factorized [3] approach. To the best of our knowledge, IZF is the first generative flow model for ZSL.
2. A novel mechanism tackling the bias problem is proposed with the merits of the generative nature of IZF, *i.e.*, measuring and diversifying the sample-based *seen-unseen* data distributional discrepancy.

2 Related Work

Zero-Shot Learning. ZSL [25] has been extensively studied in recent years [11,12,42,65]. The evaluation of ZSL can be either classic (CZSL) or generalized (GZSL) [45], while recent research also explores the potential in retrieval [32,47]. CZSL excludes *seen* classes during test, while GZSL considers both *seen* and *unseen* classes, being more popular among recent articles [4,8,19,28]. To tackle the problem of *seen-unseen* domain shift, there propose three typical ways to inject semantic knowledge for ZSL, *i.e.*, **(1)** learning visual→semantic projections [1,10,23,26,44], **(2)** learning semantic→visual projections [43,69,67], and **(3)** learning shared features or multi-modal functions [70]. Recently, deep generative models have been adapted to ZSL, subverting the traditional ZSL paradigm to some extent. The majority of existing generative methods employ GANs [27,62,35], CVAEs [24,38,46] or a mixture of the two [18,64] to synthesize *unseen* data points for a successive classification stage. However, as mentioned in Sec. 1, these models suffer from their underlying drawbacks in ZSL.

Generative Flows. Compared with GANs/VAEs, flow-based generative models [6,7,21] have attracted less research attention in the past few years, probably because this family of models require special neural structures that are in principle invertible for encoding and generation. It was not until the first appearance of the coupling layer in NICE [6] and RealNVP [7] that generative flows with deep INNs became practical and efficient. In [29], flows are extended to a conditional scheme, but the density estimation is not deterministic. The Glow architecture [21] is further introduced with invertible 1×1 convolution for realistic image generation. In [3], conditions are injected into the coupling layers. IDF [17] and BipartiteFlow [55] define a discrete case of flows. Flows can be combined with adversarial training strategies [14]. In [41], generative flows have also been successfully applied to speech synthesis.

Literally Invertible ZSL. We also notice that some existing ZSL models involve literally *invertible* projections [23,68]. However, these methods are unable to generate samples, failing to benefit GZSL with the held-out classifier schema [62] and our inverse training objectives. In addition, [23,68] are linear models

and cannot be paralleled as deep neural networks during training. This limits their model capacity and training efficiency on large-scale data.

3 Preliminaries: Generative Flows and INNs

Density Estimation with Flows. Generative flows are theoretically based on the *change of variables formula*. Given a d-dimentional datum $\mathbf{x} \in \mathcal{X} \subseteq \mathbb{R}^d$ and a pre-defined prior $p_{\mathcal{Z}}$ supporting a set of latents $\mathbf{z} \in \mathcal{Z} \subseteq \mathbb{R}^d$, the *change of variables formula* defines the estimated density of $p_\theta(\mathbf{x})$ using an invertible (also called *bijective*) transformation $f : \mathcal{X} \to \mathcal{Z}$ as follows:

$$p_\theta(\mathbf{x}) = p_{\mathcal{Z}}\left(f\left(\mathbf{x}\right)\right) \left|\det \frac{\partial f}{\partial \mathbf{x}}\right|. \tag{1}$$

Here θ indicates the set of model parameters and the scalar $\left|\det\left(\partial f/\partial \mathbf{x}\right)\right|$ is the absolute value of the determinant of the Jacobian matrix $(\partial f/\partial \mathbf{x})$. One can refer to [6,7] and our **supplementary material** for more details. The choice of the prior $p_{\mathcal{Z}}$ is arbitrary and a zero-mean unite-variance Gaussian is usually adequate, *i.e.*, $p_{\mathcal{Z}}(\mathbf{z}) = \mathcal{N}(\mathbf{z}|\mathbf{0}, \mathbf{I})$. The respective generative process can be written as $\hat{\mathbf{x}} = f^{-1}\left(\mathbf{z}\right)$, where $\mathbf{z} \sim p_{\mathcal{Z}}$. f is usually called the *forward pass*, with f^{-1} being the *reverse pass*.[4] Stacking a series of invertible functions $f = f_1 \circ f_2 \circ \cdots \circ f_k$ literally complies with the name of *flows*.

INNs with Coupling Layers. Generative flows admit networks with **(1)** exactly invertible structure and **(2)** efficiently computed Jacobian determinant. We adopt a typical type of INNs, called the coupling layers [6], which split network inputs/outputs into two respective partitions: $\mathbf{x} = [\mathbf{x}_a, \mathbf{x}_b]$, $\mathbf{z} = [\mathbf{z}_a, \mathbf{z}_b]$. The computation of the layer is defined as:

$$\begin{aligned} f(\mathbf{x}) &= \left[\mathbf{x}_a, \mathbf{x}_b \odot \exp\left(\mathbf{s}(\mathbf{x}_a)\right) + \mathbf{t}(\mathbf{x}_a)\right], \\ f^{-1}(\mathbf{z}) &= \left[\mathbf{z}_a, (\mathbf{z}_b - \mathbf{t}(\mathbf{z}_a)) \oslash \exp\left(\mathbf{s}(\mathbf{z}_a)\right)\right], \end{aligned} \tag{2}$$

where $\odot$ and $\oslash$ denote element-wise multiplication and division respectively. $\mathbf{s}(\cdot)$ and $\mathbf{t}(\cdot)$ are two arbitrary neural networks with input and output lengths of $d/2$. We show this structure in Fig. 2 (b). Its corresponding log-determinant of Jacobian can be conveniently computed by $\sum |\mathbf{s}|$. Coupling layers usually come together with element-wise permutation to build compact transformation.

4 Formulation: Factorized Conditional Flow

ZSL aims at recognizing *unseen* data. The training set $\mathcal{D}^s = \{(\mathbf{v}^s, y^s, \mathbf{c}^s)\}$ of it is grounded on M^s *seen* classes, *i.e.*, $y^s \in \mathcal{Y}^s = \{1, 2, ..., M^s\}$. Let $\mathcal{V}^s \subseteq \mathbb{R}^{d_v}$ and $\mathcal{C}^s \subseteq \mathbb{R}^{d_c}$ respectively represent the visual space and the semantic space of *seen* data, of which $\mathbf{v}^s \in \mathcal{V}^s$ and $\mathbf{c}^s \in \mathcal{C}^s$ are the corresponding feature

[4] Note that reverse pass and back-propagation are different concepts.

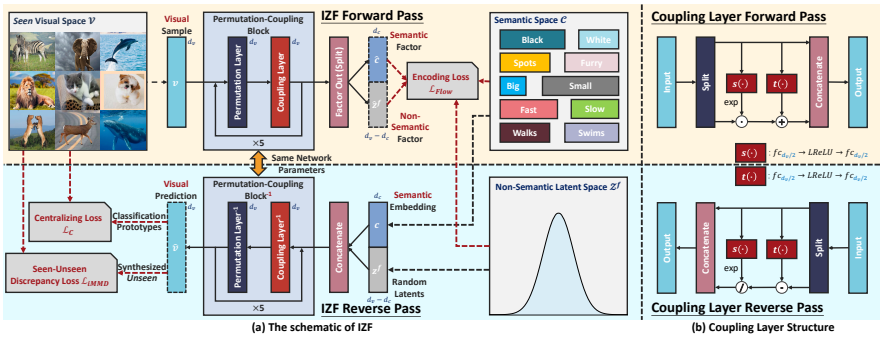

Fig. 2: **(a)** The architecture of the proposed IZF model. The forward pass and reverse pass are indeed sharing network parameters as invertible structures are used. Also note that only *seen* visual samples are accessible during training and IZF is an **inductive** ZSL model. **(b)** A typical illustration of the coupling layer [6] used in our model.

instances. The dimensions of these two spaces are denoted as d_v and d_c. Given an *unseen* label set $\mathcal{Y}^u = \{M^s + 1, M^s + 2, ..., M^s + M^u\}$ of M^u classes, the *unseen* data are denoted with the superscript of $\cdot^u$ as $\mathcal{D}^u = \{(\mathbf{v}^u, y^u, \mathbf{c}^u)\}$, where $\mathbf{v}^u \in \mathcal{V}^u$, $y^u \in \mathcal{Y}^u$ and $\mathbf{c}^u \in \mathcal{C}^u$. In this paper, the superscript are omitted when the referred sample can be both *seen* or *unseen*, *i.e.*, $\mathbf{v} \in \mathcal{V} = \mathcal{V}^s \cup \mathcal{V}^u, y \in \mathcal{Y} = \mathcal{Y}^s \cup \mathcal{Y}^u$ and $\mathbf{c} \in \mathcal{C} = \mathcal{C}^s \cup \mathcal{C}^u$.

The framework of IZF is demonstrated in Fig. 2 (a). IZF factors out the high-level semantic information with its forward pass $f(\cdot)$, equivalently performing visual→semantic projection. The reverse pass handles conditional generation, *i.e.*, semantic→visual projection, with identical network parameters to the forward pass. To reflect label information in a flow, Eq. (1) is slightly extended to a conditional scheme with visual data $\mathbf{v}$ and their labels y:

$$p_\theta(\mathbf{v}|y) = p_{\mathcal{Z}}\left(f\left(\mathbf{v}\right)|y\right)\left|\det\frac{\partial f}{\partial \mathbf{v}}\right|. \tag{3}$$

Detailed proofs are given in the **supplementary material**. Next, we consider reflecting semantic knowledge in the encoder outputs for ZSL. To this end, a factorized model takes its shape.

4.1 Forward Pass: Factorizing the Semantics

High-dimensional image representations contain both high-level semantic-related information and non-semantic information such as low-level image details. As factorizing image features has been proved effective for ZSL in [54], we adopt this spirit, but with different approach to fit the structure of flow. In [54], the factorization is basically only empirical, while IZF derives full likelihood model of a training sample.

As shown in Fig. 2 (a), the proposed flow network learns factorized independent image representations $\hat{\mathbf{z}} = [\hat{\mathbf{c}}, \hat{\mathbf{z}}^f] = f(\mathbf{v})$ with its forward pass $f(\cdot)$, where $\hat{\mathbf{c}} \in \mathbb{R}^{d_c}$ denotes the predicted semantic factor of an arbitrary visual sample $\mathbf{v}$ and $\hat{\mathbf{z}}^f \in \mathbb{R}^{d_v - d_c}$ is the low-level non-semantic independent to $\hat{\mathbf{c}}$, *i.e.*, $\hat{\mathbf{z}}^f \perp\!\!\!\perp \hat{\mathbf{c}}$. We assume $\hat{\mathbf{z}}^f$ is not dependent on data label y, *i.e.*, $\hat{\mathbf{z}}^f \perp\!\!\!\perp y$ as it is designed to reflect no high-level semantic/category information. Therefore, we rewrite the conditional probability of Eq. (3) as

$$p_\theta(\mathbf{v}|y) = p_{\mathcal{Z}}\left([\hat{\mathbf{c}}, \hat{\mathbf{z}}^f] = f(\mathbf{v})|y\right) \left|\det \frac{\partial f}{\partial \mathbf{v}}\right| = p_{\mathcal{C}|\mathcal{Y}}(\hat{\mathbf{c}}|y) p_{\mathcal{Z}^f}(\hat{\mathbf{z}}^f) \left|\det \frac{\partial f}{\partial \mathbf{v}}\right|. \tag{4}$$

The conditional independence property gives $p_{\mathcal{Z}}(\hat{\mathbf{c}}, \hat{\mathbf{z}}^f|y) = p_{\mathcal{C}|\mathcal{Y}}(\hat{\mathbf{c}}|y) p_{\mathcal{Z}^f}(\hat{\mathbf{z}}^f)$. According to [16,57], this property is implicitly enforced by imposing fix-formed priors on each variable. In this work, the factored priors are

$$p_{\mathcal{C}|\mathcal{Y}}(\hat{\mathbf{c}}|y) = \mathcal{N}(\hat{\mathbf{c}}|\mathbf{c}(y), \mathbf{I}), \quad p_{\mathcal{Z}^f}(\hat{\mathbf{z}}^f) = \mathcal{N}(\hat{\mathbf{z}}^f|\mathbf{0}, \mathbf{I}), \tag{5}$$

where $\mathbf{c}(y)$ simply denotes the semantic embedding corresponding to y. Similar to the likelihood computation of VAEs [22], we empirically assign a uniformed Gaussian to $p_{\mathcal{C}|\mathcal{Y}}(\hat{\mathbf{c}}|y)$ centered at the corresponding semantic embedding $\mathbf{c}(y)$ of the visual sample so that it can be simply reduced to a $l2$ norm.

The Injected Semantic Knowledge. The benefits of the factorized $p_{\mathcal{C}|\mathcal{Y}}(\hat{\mathbf{c}}|y)$ are two-fold: **1)** it explicitly reflects the degree of similarity between different classes, ensuring smooth *seen-unseen* generalization for ZSL. This is also in line with the main motivation of several existing approaches [23,44]; **2)** a well-trained IZF model with $p_{\mathcal{C}|\mathcal{Y}}(\hat{\mathbf{c}}|y)$ factorizes the semantic meaning from non-semantic information of an image, making it possible to conditionally generate samples with $f^{-1}(\cdot)$ by directly feeding the semantic category embedding (see Eq. (6)).

4.2 Reverse Pass: Conditional Sample Generation

One advantage of deep generative ZSL models is the ability to observe synthesized *unseen* data. IZF fulfills this by

$$\mathbf{c} \in \mathcal{C}, \mathbf{z}^f \sim p_{\mathcal{Z}^f}, \hat{\mathbf{v}} = f^{-1}\left([\mathbf{c}, \mathbf{z}^f]\right). \tag{6}$$

The Use of Reverse Pass. Different from most generative ZSL approaches [38,62] where synthesized *unseen* samples simply feed a held-out classifier, IZF additionally uses these synthesized samples to measure the biased distributional overlap between *seen* and synthesized *unseen* data. We will elaborate the corresponding learning objectives and ideas in Sec. 5.3.

4.3 Network Structure

In the spirits of Eq. (4) and (6), we build the network of IZF as shown in Fig. 2 (a). Concretely, IZF consists of 5 permutation-coupling blocks to shape

a deep non-linear architecture. Inspired by [2,7], we combine the coupling layer with channel-wise permutation in each block. The permutation layer shuffles the elements of an input feature in a random but fixed manner so that the split of two successive coupling layers are different and the encoding/decoding performance is assured. We use identical structure for the built-in neural network $\mathtt{s}(\cdot)$ and $\mathtt{t}(\cdot)$ of the coupling layers in Eq. (2), *i.e.*, $\mathtt{fc}_{d_v/2} \rightarrow \mathtt{LReLU} \rightarrow \mathtt{fc}_{d_v/2}$, where $\mathtt{LReLU}$ is the leaky ReLU activation [33]. In the following, we show how the network is trained to enhance ZSL.

5 Training with the Merits of Generative Flow

To transfer knowledge from *seen* concepts to *unseen* ones, we employ the idea of bi-directional training of INNs [2] to optimize IZF. In principle, generative flows can be trained only with the forward pass (Sec. 5.1). However, considering the fact that the reverse pass of IZF is used for *zero-shot* classification, we impose additional learning objectives to its reverse pass to promote the ability of *seen-unseen* generalization (Sec. 5.2 and 5.3).

5.1 Learning to Decode by Encoding

The first learning objective of IZF comes from the definition of generative flow as depicted in Eq. (1). By analytic log-likelihood maximization of the forward pass, generative flows are ready to synthesize data samples. As only visual features of *seen* categories are observable to IZF, we construct this loss term upon $\mathcal{D}^s$ as

$$\mathcal{L}_{\text{Flow}} = \mathbb{E}_{(\mathbf{v}^s, y^s, \mathbf{c}^s)} \left[-\log p_\theta(\mathbf{v}^s | y^s) \right], \tag{7}$$

where $(\mathbf{v}^s, y^s, \mathbf{c}^s)$ are *seen* samples from the training set $\mathcal{D}^s$ and $p_\theta(\mathbf{v}^s|y^s)$ is computed according to Eq. (4). $\mathcal{L}_{\text{Flow}}$ is not only an encoding loss, but also can legitimate unconditional *seen* data generation due to the invertible nature of IZF. Compared with the training process of GAN/VAE-based ZSL models [38,62], IZF defines an explicit and simpler objective to fulfill the same functionality.

5.2 Centralizing Classification Prototypes

IZF supports naive Bayesian classification by projecting semantic embeddings back to the visual space with its reverse pass. For each class-wise semantic representation, we define a special generation procedure $\hat{\mathbf{v}}_c = f^{-1}([\mathbf{c}, \mathbf{0}])$ as the **classification prototype** of a class. As these prototypes are directly used to classify images by distance comparison, it would be harmful to the final accuracy when the prototypes are too close to unrelated visual samples. To address this issue, f^{-1} is expected to position them close to the centres $\bar{\mathbf{v}}_c$ of the respective classes they belong to. This idea is illustrated in Fig. 3, denoted as $\mathcal{L}_C$. In particular, this centralizing loss is imposed on the *seen* classes as

$$\mathcal{L}_{\text{C}} = \mathbb{E}_{(\mathbf{c}^s, \bar{\mathbf{v}}^s_c)} \left[\| f^{-1}([\mathbf{c}^s, \mathbf{0}]) - \bar{\mathbf{v}}^s_c \|^2 \right], \tag{8}$$

where $\bar{\mathbf{v}}_c^s$ is the corresponding numerical mean of the visual samples that belong to the class with the semantic embedded $\mathbf{c}^s$. Similar to the semantic knowledge loss, we directly apply $l2$ norm to the model to regularize its behavior.

5.3 Measuring the *Seen-Unseen* Bias

Recalling the bias problem in ZSL with generative models, the synthesized *unseen* samples could be unexpectedly too close to the real *seen* ones. This would significantly decrease the classification performance for *unseen* classes, especially in the context of GZSL where *seen* and *unseen* data are both available. We propose to explicitly tackle the bias problem by preventing the **synthesized *unseen*** visual distribution $p_{\hat{\mathcal{V}}^u}$ from colliding with the **real *seen*** one $p_{\mathcal{V}^s}$. In other words, $p_{\mathcal{V}^s}$ is slightly pushed away from $p_{\hat{\mathcal{V}}^u}$.

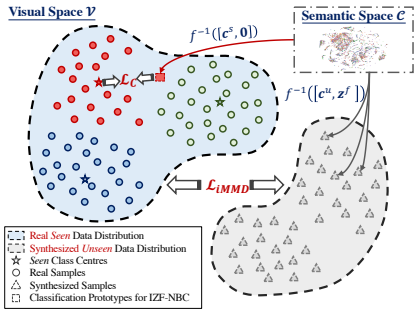

Fig. 3: Typical illustration of the IZF training losses *w.r.t.* the **reverse pass**, *i.e.*, $\mathcal{L}_{\mathrm{C}}$ in Sec. 5.2 and $\mathcal{L}_{\mathrm{iMMD}}$ in Sec. 5.3.

Our key idea is illustrated in Fig. 3, denoted as $\mathcal{L}_{\mathrm{iMMD}}$. With generative models, it is always possible to measure distributional discrepancy without acknowledging the true distribution parameters of $p_{\hat{\mathcal{V}}^u}$ and $p_{\mathcal{V}^s}$ by treating this as a negative two-sample-test problem. Hence, we resort to Maximum Mean Discrepancy (MMD) [2,53] as the measurement. Since we aim to increase the discrepancy, the last loss term of IZF is defined upon the **numerical negation** of MMD $\left(p_{\mathcal{V}^s}||p_{\hat{\mathcal{V}}^u}\right)$ in a batch-wise fashion as

$$\begin{aligned}\mathcal{L}_{\mathrm{iMMD}} = -\,\mathrm{MMD}\left(p_{\mathcal{V}^s}||p_{\hat{\mathcal{V}}^u}\right) &= \tfrac{2}{n^2}\sum_{i,j}\kappa(\mathbf{v}_i^s, \hat{\mathbf{v}}_j^u) \\ &- \tfrac{1}{n(n-1)}\sum_{i\neq j}\left(\kappa(\mathbf{v}_i^s, \mathbf{v}_j^s) + \kappa(\hat{\mathbf{v}}_i^u, \hat{\mathbf{v}}_j^u)\right), \\ \text{where } \mathbf{v}_i^s \in \mathcal{V}^s,\ \mathbf{c}_i^u \in \mathcal{C}^u,\ \mathbf{z}_i^f \sim p_{\mathcal{Z}^f},\ \hat{\mathbf{v}}_i^u &= f^{-1}([\mathbf{c}_i^u, \mathbf{z}_i^f]).\end{aligned} \tag{9}$$

Here n refers to the training batch size, and $\kappa(\cdot)$ is an arbitrary positive-definite reproducing kernel function. Importantly, as only *seen* visual samples $\mathbf{v}_i^s$ are directly used and $\hat{\mathbf{v}}_i^u$ are synthesized, $\mathcal{L}_{\mathrm{iMMD}}$ is indeed an **inductive** objective. The same setting has also been adopted in recent inductive ZSL methods [30,46,51,62], *i.e.*, the names of the *unseen* classes are accessible during training while their visual samples remain inaccessible. We also note that replacing $\mathcal{L}_{\mathrm{iMMD}}$ by simply tuning the values of *unseen* classification templates

$f^{-1}([\mathbf{c}^u, \mathbf{0}])$ is infeasible in inductive ZSL since there exists no *unseen* visual reference sample for direct regularization.

5.4 Overall Objective and Training

By combining the above-discussed losses, the overall learning objective of IZF can be simply written as

$$\mathcal{L}_{\text{IZF}} = \lambda_1 \mathcal{L}_{\text{Flow}} + \lambda_2 \mathcal{L}_{\text{C}} + \lambda_3 \mathcal{L}_{\text{iMMD}}. \tag{10}$$

Three hyper-parameters λ_1, λ_2 and λ_3 are introduced to balance the contributions of different loss terms. IZF is fully differentiable *w.r.t.* $\mathcal{L}_{\text{IZF}}$. Hence, the corresponding network parameters can be directly optimized with Stochastic Gradient Descent (SGD) algorithms.

5.5 Zero-Shot Recognition with IZF

We adopt two ZSL classification strategies (*i.e.*, IZF-NBC and IZF-Softmax) that work with IZF. Specifically, IZF-NBC employs a naive Bayesian classifier to recognize a given test visual sample $\mathbf{v}_q$ by comparing the Euclidean distances between it and the classification prototypes introduced in Sec 5.2. IZF-Softmax leverages a held-out classifier similar to the one used in [62]. The classification processes are performed as

$$\begin{aligned} &\text{IZF-NBC: } \hat{y}^q = \arg\min_y \parallel f^{-1}([\mathbf{c}(y), \mathbf{0}]) - \mathbf{v}^q \parallel, \\ &\text{IZF-Softmax: } \hat{y}^q = \arg\max_y \texttt{softmax}\left(\texttt{NN}(\mathbf{v}^q)\right). \end{aligned} \tag{11}$$

Here $\texttt{NN}(\cdot)$ is a single-layered fully-connected network trained with generated *unseen* data and the softmax cross-entropy loss on top of the softmax activation.

6 Experiments

6.1 Implementation Details

IZF is implemented with the popular deep learning toolbox PyTorch [39]. We build the INNs according to the framework of FrEIA [2,3]. The network architecture is elaborated in Sec. 4.3. The built-in networks $\mathbf{s}(\cdot)$ and $\mathbf{t}(\cdot)$ of all coupling layers of IZF are shaped by $\texttt{fc}_{d_v/2} \rightarrow \texttt{LReLU} \rightarrow \texttt{fc}_{d_v/2}$. Following [2,53], we employ the Inverse Multiquadratic (IM) kernel $\kappa(\mathbf{v}, \mathbf{v}') = 2d_v / \left(2d_v + \parallel \mathbf{v} - \mathbf{v}' \parallel^2\right)$ in Eq. (9) for best performance. We testify the choice of λ_1, λ_2 and λ_3 within $\{0.1, 0.5, 1, 1.5, 2\}$ and report the results of $\lambda_1 = 2, \lambda_2 = 1, \lambda_3 = 0.1$ for all comparisons. The Adam optimizer [20] is used to train IZF with a learning rate of 5×10^{-4} *w.r.t.* $\mathcal{L}_{\text{IZF}}$. The batch size is fixed to 256 for all experiments.

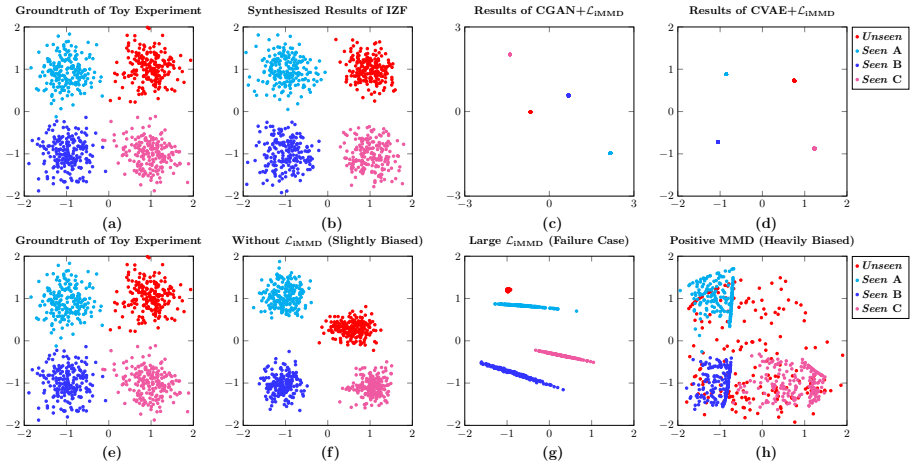

Fig. 4: Illustration of the 4-class toy experiment in Sec. 6.2. **(a, e)** 2-D Ground truth simulation data, with the top-right class being *unseen*. **(b)** Synthesized samples of IZF. **(c, d)** Synthesized results of conditional GAN and CVAE respectively with $\mathcal{L}_{\text{iMMD}}$. **(f)** Results without $\mathcal{L}_{\text{iMMD}}$ of IZF. **(g)** Failure results with extremely and unreasonably large $\mathcal{L}_{\text{iMMD}}$ ($\lambda_3 = 10$) of IZF. **(h)** Results with positive MMD of IZF.

6.2 Toy Experiments: Illustrative Analysis

Before evaluating IZF with real data, we firstly provide a toy ZSL experiment to justify our motivation. Particularly, the following themes are discussed:

1. Why Do We Resort to Flows Instead of GAN/VAE with $\mathcal{L}_{\text{iMMD}}$?
2. The effect of $\mathcal{L}_{\text{iMMD}}$ regarding the bias problem.

Setup. We consider a 4-class simulation dataset with 1 class being *unseen*. The class-wise attributes are defined as $\mathcal{C}^s = \{[0,1],[0,0],[1,0]\}$ for the *seen* classes **A**, **B** and **C** respectively, while the *unseen* class would have attribute of $\mathcal{C}^u = \{[1,1]\}$. The ground truth data are randomly sampled around a linear transformation of the attributes, *i.e.*, $\mathbf{v} \coloneqq 2\mathbf{c} - 1 + \epsilon \in \mathbb{R}^2$, where $\epsilon \sim \mathcal{N}(\mathbf{0}, \frac{1}{3}\mathbf{I})$. To meet the dimensionality requirement, *i.e.*, $d_v > d_c$, we follow the convention of [2] to pad two zeros to data when feeding them to the network, *i.e.*, $\mathbf{v}' \coloneqq [\mathbf{v}, 0, 0]$. The toy data are plotted in Fig. 4 (a) and (e).

Why Do We Resort to Flows Instead of GAN/VAE? We firstly show the synthesized results of IZF in Fig. 4 (b). It can be observed that IZF successfully interprets the relations of the *unseen* class to the *seen* ones, *i.e.*, being closer to **A** and **C** but further to **B**. To legit the use of generative flow, we accordingly build two baselines by combining Conditional GAN (CGAN) and CVAE with our $\mathcal{L}_{\text{iMMD}}$ loss (see our **supplementary document** for implementation details). The respective generated results are shown in Fig. 4 (c) and (d). Aligning with

		AwA1 [26]			AwA2 [26]			CUB [58]			SUN [40]			aPY [9]		
Method	**Reference**	A^s	A^u	H	A^s	A^u	H	A^s	A^u	H	A^s	A^u	H	A^s	A^u	H
DAP [26]	PAMI13	88.7	0.0	0.0	84.7	0.0	0.0	67.9	0.0	0.0	25.1	4.2	7.2	78.3	4.8	9.0
CMT [50]	NIPS13	86.9	8.4	15.3	89.0	8.7	15.9	60.1	4.7	8.7	28.0	8.7	13.3	74.2	10.9	19.0
DeViSE [10]	NIPS13	68.7	13.4	22.4	74.7	17.1	27.8	53.0	23.8	32.8	27.4	16.9	20.9	76.9	4.9	9.2
ALE [1]	CVPR15	16.8	76.1	27.5	81.8	14.0	23.9	62.8	23.7	34.4	33.1	21.8	26.3	73.7	4.6	8.7
SSE [70]	ICCV15	80.5	7.0	12.9	82.5	8.1	14.8	46.9	8.5	14.4	36.4	2.1	4.0	78.9	0.2	0.4
ESZSL [44]	ICML15	75.6	6.6	12.1	77.8	5.9	11.0	63.8	12.6	21.0	27.9	11.0	15.8	70.1	2.4	4.6
LATEM [60]	CVPR16	71.1	7.3	13.3	77.3	11.5	20.0	57.3	15.2	24.0	28.8	14.7	19.5	73.0	0.1	0.2
SAE [23]	CVPR17	77.1	1.8	3.5	82.2	1.1	2.2	54.0	7.8	13.6	18.0	8.8	11.8	**80.9**	0.4	0.9
DEM [69]	CVPR17	84.7	32.8	47.3	86.4	30.5	45.1	57.9	19.6	29.2	34.3	20.5	25.6	11.1	**75.1**	19.4
RelationNet [52]	CVPR18	**91.3**	31.4	46.7	**93.4**	30.0	45.3	61.1	38.1	47.0	-	-	-	-	-	-
DCN [30]	NIPS18	84.2	25.5	39.1	-	-	-	60.7	28.4	38.7	37.0	25.5	30.2	75.0	14.2	23.9
CRNet [67]	ICML19	74.7	58.1	65.4	78.8	52.6	63.1	56.8	45.5	50.5	36.5	34.1	35.3	68.4	32.4	44.0
LFGAA [31]	ICCV19	-	-	-	90.3	50.0	64.4	79.6	43.4	56.2	34.9	20.8	26.1	-	-	-
CVAE-ZSL [38]	ECCVW18	-	-	47.2	-	-	51.2	-	-	34.5	-	-	26.7	-	-	-
SE-GZSL [24]	CVPR18	67.8	56.3	61.5	68.1	58.3	62.8	53.3	41.5	46.7	30.5	40.9	34.9	-	-	-
f-CLSWGAN [62]	CVPR18	61.4	57.9	59.6	-	-	-	57.7	43.7	49.7	36.6	42.6	39.4	-	-	-
LisGAN [27]	CVPR19	76.3	52.6	62.3	-	-	-	57.9	46.5	51.6	37.8	42.9	40.2	-	-	-
SGAL [66]	NIPS19	75.7	52.7	62.2	81.2	55.1	65.6	44.7	47.1	45.9	31.2	42.9	36.1	-	-	-
CADA-VAE [46]	CVPR19	72.8	57.3	64.1	75.0	55.8	63.9	53.5	51.6	52.4	35.7	47.2	40.6	-	-	-
GDAN [18]	CVPR19	-	-	-	67.5	32.1	43.5	66.7	39.3	49.5	**89.9**	38.1	53.4	75.0	30.4	43.4
DLFZRL [54]	CVPR19	-	-	61.2	-	-	60.9	-	-	51.9	-	-	42.5	-	-	38.5
f-VAEGAN-D2 [64]	CVPR19	70.6	57.6	63.5	-	-	-	60.1	48.4	53.6	38.0	45.1	41.3	-	-	-
IZF-NBC	**Proposed**	75.2	57.8	65.4	76.0	58.1	65.9	56.3	44.2	49.5	50.6	44.5	47.4	58.3	39.8	47.3
IZF-Softmax	**Proposed**	80.5	**61.3**	**69.6**	77.5	**60.6**	**68.0**	68.0	**52.7**	**59.4**	57.0	**52.7**	**54.8**	60.5	42.3	**49.8**

Table 1: Inductive GZSL performance of IZF and the state-of-the-art methods with the PS setting [63].

our motivation, $\mathcal{L}_{\text{iMMD}}$ quickly fails the unstable training process of GAN in ZSL. Besides, CVAE+$\mathcal{L}_{\text{iMMD}}$ isn't producing good-quality samples, undergoing the risk of obtaining biased classification hyper-planes of the held-out classifier. The side-effects of $\mathcal{L}_{\text{iMMD}}$ would slightly skew the generated data distributions from being realistic with its negative MMD, which aggravates the drawbacks of unstable training (GAN) and inaccurate ELBO (VAE) discussed in Sec. 1.

Towards the Bias Problem with $\mathcal{L}_{\text{iMMD}}$. We also illustrate the effects of $\mathcal{L}_{\text{iMMD}}$ with more baselines. It is shown in Fig. 4 (f) that the model is biased by the *seen* classes without $\mathcal{L}_{\text{iMMD}}$ (also see `Baseline 4` of Sec. 6.5). The *unseen* generated samples are positioned closely to the *seen* ones. This would be harmful to the employed classifiers when there exist multiple *unseen* categories. Fig. 4 (g) is a failure case with large *seen-unseen* discrepancy loss, which dominates the optimization process and overfits the network to generate unreasonable samples. We also discuss this issue in hyper-parameter analysis (see Fig. 5 (c)). Fig. 4 (h) describes an extreme situation when employing positive MMD to IZF (negative λ_3, `Baseline 5` of Sec. 6.5). The generated *unseen* samples are forced to fit the *seen* distribution and thus, the network is severely biased.

6.3 Real Data Experimental Settings

Benchmark Datasets. Five datasets are picked in our experiments. Animals with Attributes (AwA1) [26] contains 30,475 images of 50 classes and 85 attributes, of which AwA2 is a slightly extended version with 37,322 images.

Caltech-UCSD Birds-200-20 (CUB) [58] carries 11,788 images from 200 kinds of birds with 312-attribute annotations. SUN Attribute (SUN) [40] consists of 14,340 images from 717 categories, annotated with 102 attributes. aPascal-aYahoo (aPY) [9] comes with 32 classes with 64 attributes, accounting 15,339 samples. We adopt the **PS** train-test setting [63] for both CZSL and GZSL.

Representations. All images $\mathbf{v}$ are represented using the 2048-D ResNet-101 [15] features and the semantic class embeddings $\mathbf{c}$ are category-wise attribute vectors from [61,63]. We pre-process the image features with min-max rescaling.

6.4 Comparison with the State-of-the-Arts

Baselines. IZF is compared with the state-of-the-art ZSL methods, including DAP [26], CMT [50], SSE [70], ESZSL [44], SAE [23], LATEM [60], ALE [1], DeViSE [10], DEM [69], RelationNet [52], DCN [30], CVAE-ZSL [38], SE-GZSL [24], f-CLSWGAN [62], CRNet [67], LisGAN [27], SGAL [66], CADA-VAE [46], GDAN [18], DLFZRL[54], f-VAEGAN-D2 [64] and LFGAA [31]. We report the official results of these methods from referenced articles with the identical experimental setting used in this paper for fair comparison.

Results. The GZSL comparison results are shown in Tab. 1. It can be observed that deep generative models obtains better on-average ZSL scores than the non-generative ones, while some simple semantic-visual projecting models hit comparable accuracy to them such as CRNet [67]. IZF-Softmax generally outperforms the compared methods, where the performance margins on AwA [26] are significant. IZF-NBC also works well on AwA [26] The proposed model produces balanced accuracy between *seen* and *unseen* data and obtains significant higher *unseen* accuracy. This shows the effectiveness of the discrepancy loss $\mathcal{L}_{\text{iMMD}}$ in solving the bias problem of ZSL. In addition to the GZSL results, we conduct CZSL experiments as well, which is shown in Tab. 2. As a relatively simpler setting, CZSL provides direct clues of the ability to transform knowledge from *seen* to *unseen*.

Method	AwA1	AwA2	CUB	SUN	aPY
DAP [26]	44.1	46.1	40.0	39.9	33.8
CMT [70]	39.5	37.9	34.6	39.9	28.0
SSE [70]	60.1	61.0	43.9	51.5	34.0
ESZSL [44]	58.2	58.6	53.9	54.5	38.3
SAE [23]	53.0	54.1	33.3	40.3	8.3
LATEM [60]	55.1	55.8	49.3	55.3	35.2
ALE [1]	59.9	62.5	54.9	58.1	39.7
DeViSE [10]	54.2	59.7	52.0	56.5	39.8
RelationNet [52]	68.2	64.2	55.6	-	-
DCN [30]	65.2	-	56.2	61.8	43.6
f-CLSWGAN [62]	68.2	-	57.3	60.8	-
LisGAN [27]	70.6	-	58.8	61.7	43.1
DLFZRL [54]	61.2	60.9	51.9	42.5	38.5
f-VAEGAN-D2 [64]	71.1	-	61.0	65.6	-
LFGAA [31]	-	68.1	**67.6**	62.0	-
IZF-NBC	72.7	71.9	59.6	63.0	**45.2**
IZF-Softmax	**74.3**	**74.5**	67.1	**68.4**	44.9

Table 2: CZSL per-class accuracy (%) comparison with the **PS** setting [63].

6.5 Component Analysis

We evaluate the effectiveness of each component of IZF to legitimate our design, including the loss terms and overall network structure. The following baselines

	NBC			Softmax		
Baseline	A^s	A^u	H	A^s	A^u	H
1 CVAE + $\mathcal{L}_{\text{C}}$ + $\mathcal{L}_{\text{iMMD}}$	65.1	30.8	41.8	71.1	36.8	48.5
2 Without $\mathcal{L}_{\text{C}}$ and $\mathcal{L}_{\text{iMMD}}$	66.0	43.4	52.7	78.9	38.1	51.4
3 Without $\mathcal{L}_{\text{C}}$	67.0	41.7	51.4	79.2	60.9	68.8
4 Without $\mathcal{L}_{\text{iMMD}}$	79.6	49.0	60.7	81.3	53.2	64.3
5 Positive MMD	76.2	21.1	33.0	80.7	44.5	57.4
6 IM Kernel→Gaussian Kernel	73.6	54.9	62.9	79.6	61.7	69.5
IZF (full model)	75.2	57.8	65.4	80.5	61.3	69.6

Table 3: Component analysis results on AwA1 [26] (Sec. 6.5). **NBC**: results with distance-based classifier. **Softmax**: results with a held-out trainable classifier.

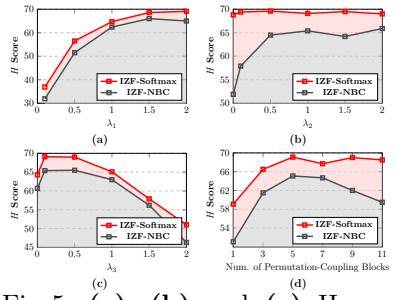

Fig. 5: **(a)**, **(b)** and **(c)** Hyper-parameter analysis for λ_1, λ_2 and λ_3. **(d)** Effect *w.r.t.* numbers of the permutation-coupling blocks.

are proposed. **(1) CVAE + $\mathcal{L}_{\text{C}}$ + $\mathcal{L}_{\text{iMMD}}$.** We firstly replace it with a simple CVAE [51] structure. This baseline uses the semantic representation as condition, and outputs synthesized visual features. $\mathcal{L}_{\text{C}}$ and $\mathcal{L}_{iMMD}$ are applied to this baseline. **(2) Without $\mathcal{L}_{\text{C}}$ & $\mathcal{L}_{\text{iMMD}}$.** All regularization on the reverse pass is omitted. **(3) Without $\mathcal{L}_{\text{C}}$.** The prototype centralizing loss is removed. **(4) Without $\mathcal{L}_{\text{iMMD}}$.** The discrepancy loss to control the *seen-unseen* bias problem of ZSL is deprecated. **(5) Positive MMD.** In Eq. (9), we employ negative MMD to tackle the bias problem. We propose a baseline with a positive MMD version of it to study its influence. This is realized by setting $\lambda_3 = -1$. **(6) IM Kernel→Gaussian Kernel.** Instead of the Inverse Multiquadratic kernel, another widely-used kernel function, *i.e.*, the Gaussian kernel, is tested in implementing Eq. (9).

Results. The above-mentioned baselines are compared in Tab. 3 on AwA1 [26]. The GZSL criteria are adopted here as they are more illustrative metrics for IZF, showing different performance aspects of the model. Through our test, `Baseline 1`, *i.e.*, CVAE+$\mathcal{L}_{\text{C}}$+$\mathcal{L}_{\text{iMMD}}$, is not working well with the distance-based classifier (Eq. (11)). With loss components omitted (`Baseline 2-4`), IZF does not work as expected. In `Baseline 4`, the classification results are significantly biased to the *seen* concepts. When imposing positive MMD to the loss function, the test accuracy of *seen* classes increases while the accuracy of *unseen* data drops quickly. This is because the bias problem gets severer and all generated samples, including the *unseen* classification prototypes, overfit to the *seen* domain. The choice of kernel is not a key factor in IZF, and `Baseline 7` obtains on-par accuracy to IZF. Similar to GAN/VAE-based models [27,38,62], IZF works with a held-out classifier, but it requires additional computational resources.

6.6 Hyper-Parameters

IZF involves 3 hyper-parameters in balancing the contribution of different loss items, shown in Eq. (10). The influences of the values of them on AwA1 are

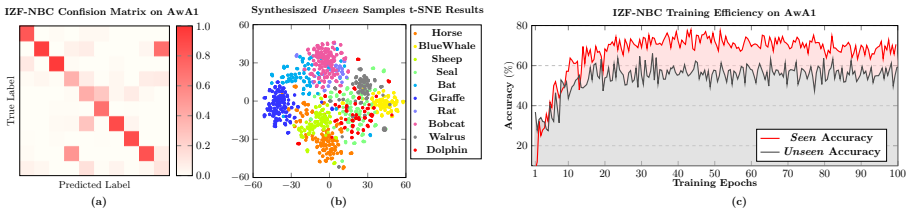

Fig. 6: **(a)** Confusion matrix of IZF on AwA1 with the CZSL setting. The order of labels is identical to the t-SNE legend. **(b)** t-SNE [34] results of the synthesized *unseen* samples on AwA1. **(c)** Training efficiency of IZF-NBC on AwA1.

plotted in Fig. 5 (a), (b) and (c) respectively. A large weight is imposed to the semantic knowledge loss $\mathcal{L}_{\text{Flow}}$, *i.e.*, $\lambda_1 = 2$, for best performance, as it plays an essential role in formulating the normalizing flow structure that ensures data generation with the sampled conditions and latents. A well-regressed visual-semantic projection necessitates conditional generation and, hence, bi-directional training. On the other hand, it is notable that a large value of λ_3 fails IZF overall. A heavy penalty to $\mathcal{L}_{\text{iMMD}}$ overfits the network to generate unreasonable samples to favour large *seen-unseen* distributional discrepancy, and further prevents the encoding loss $\mathcal{L}_{\text{Flow}}$ from functioning. We observe significant increase of $\mathcal{L}_{\text{Flow}}$ throughout the training steps with $\lambda_3 = 2$, though $\mathcal{L}_{\text{iMMD}}$ decreases quickly. We further report the training efficiency of IZF in Fig.6 (c), where IZF only requires $\sim$20 epochs to obtain best-performing parameters.

6.7 Discriminability on *Unseen* Classes

We intuitively analyze the discriminability and generation quality of IZF on *unseen* data by plotting the generated samples. The t-SNE [34] visualization of synthesized *unseen* data on AwA1 [26] is shown in Fig. 6 (b). Although no direct regularization loss is applied to *unseen* classes, IZF manages to generate distinguishable samples according to their semantic meanings. In addition, the CZSL confusion matrix on AwA1 is reported in Fig. 6 (a) as well.

7 Conclusion

In this paper, we proposed Invertible Zero-shot Flow (IZF), fully leveraging the merits of generative flows for ZSL. The invertible nature of flows enabled IZF to perform bi-directional mapping between the visual space and the semantic space with identical network parameters. The semantic information of a visual sample was factored-out with the forward pass of IZF. To handle the bias problem, IZF penalized *seen-unseen* similarity by computing kernel-based distribution discrepancy with the generated data. The proposed model consistently outperformed state-of-the-art baselines on benchmark datasets.

References

1. Akata, Z., Reed, S., Walter, D., Lee, H., Schiele, B.: Evaluation of output embeddings for fine-grained image classification. In: CVPR (2015) 3, 11, 12
2. Ardizzone, L., Kruse, J., Wirkert, S., Rahner, D., Pellegrini, E.W., Klessen, R.S., Maier-Hein, L., Rother, C., Köthe, U.: Analyzing inverse problems with invertible neural networks. In: ICLR (2019) 2, 3, 7, 8, 9, 10
3. Ardizzone, L., Lüth, C., Kruse, J., Rother, C., Köthe, U.: Guided image generation with conditional invertible neural networks. arXiv preprint arXiv:1907.02392 (2019) 3, 9
4. Cacheux, Y.L., Borgne, H.L., Crucianu, M.: Modeling inter and intra-class relations in the triplet loss for zero-shot learning. In: ICCV (2019) 3
5. Che, T., Li, Y., Jacob, A.P., Bengio, Y., Li, W.: Mode regularized generative adversarial networks. In: ICLR (2017) 2
6. Dinh, L., Krueger, D., Bengio, Y.: Nice: Non-linear independent components estimation. In: ICLR Workshops (2014) 2, 3, 4, 5
7. Dinh, L., Sohl-Dickstein, J., Bengio, S.: Density estimation using real NVP. In: ICLR (2017) 2, 3, 4, 7
8. Elhoseiny, M., Elfeki, M.: Creativity inspired zero-shot learning. In: ICCV (2019) 3
9. Farhadi, A., Endres, I., Hoiem, D., Forsyth, D.A.: Describing objects by their attributes. In: CVPR (2009) 11, 12
10. Frome, A., Corrado, G.S., Shlens, J., Bengio, S., Dean, J., Ranzato, M., Mikolov, T.: Devise: A deep visual-semantic embedding model. In: NeurIPS (2013) 3, 11, 12
11. Gao, R., Hou, X., Qin, J., Chen, J., Liu, L., Zhu, F., Zhang, Z., Shao, L.: Zero-vae-gan: Generating unseen features for generalized and transductive zero-shot learning. IEEE Transactions on Image Processing **29**, 3665–3680 (2020) 3
12. Gao, R., Hou, X., Qin, J., Liu, L., Zhu, F., Zhang, Z.: A joint generative model for zero-shot learning. In: ECCV Workshops (2018) 3
13. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. In: NeurIPS (2015) 2
14. Grover, A., Dhar, M., Ermon, S.: Flow-gan: Combining maximum likelihood and adversarial learning in generative models. In: AAAI (2018) 3
15. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR (2016) 12
16. Higgins, I., Matthey, L., Pal, A., Burgess, C., Glorot, X., Botvinick, M.M., Mohamed, S., Lerchner, A.: beta-VAE: Learning basic visual concepts with a constrained variational framework. In: ICLR (2017) 6
17. Hoogeboom, E., Peters, J.W., Berg, R.v.d., Welling, M.: Integer discrete flows and lossless compression. In: NeurIPS (2019) 3
18. Huang, H., Wang, C., Yu, P.S., Wang, C.D.: Generative dual adversarial network for generalized zero-shot learning. In: CVPR (2019) 3, 11, 12
19. Jiang, H., Wang, R., Shan, S., Chen, X.: Transferable contrastive network for generalized zero-shot learning. In: ICCV (2019) 3
20. Kingma, D., Ba, J.: Adam: A method for stochastic optimization. In: ICLR (2015) 9
21. Kingma, D., Dhariwal, P.: Glow: Generative flow with invertible 1x1 convolutions. In: NeurIPS (2018) 3
22. Kingma, D., Welling, M.: Auto-encoding variational Bayes. In: ICLR (2014) 2, 6

23. Kodirov, E., Xiang, T., Gong, S.: Semantic autoencoder for zero-shot learning. In: CVPR (2017) 3, 6, 11, 12
24. Kumar Verma, V., Arora, G., Mishra, A., Rai, P.: Generalized zero-shot learning via synthesized examples. In: CVPR (2018) 3, 11, 12
25. Lampert, C.H., Nickisch, H., Harmeling, S.: Learning to detect unseen object classes by between-class attribute transfer. In: CVPR (2009) 1, 2, 3
26. Lampert, C.H., Nickisch, H., Harmeling, S.: Attribute-based classification for zero-shot visual object categorization. IEEE Transactions on Pattern Analysis and Machine Intelligence **36**(3), 453–465 (2013) 1, 3, 11, 12, 13, 14
27. Li, J., Jing, M., Lu, K., Ding, Z., Zhu, L., Huang, Z.: Leveraging the invariant side of generative zero-shot learning. In: CVPR (2019) 2, 3, 11, 12, 13
28. Li, K., Min, M.R., Fu, Y.: Rethinking zero-shot learning: A conditional visual classification perspective. In: ICCV (2019) 3
29. Liu, R., Liu, Y., Gong, X., Wang, X., Li, H.: Conditional adversarial generative flow for controllable image synthesis. In: CVPR (2019) 3
30. Liu, S., Long, M., Wang, J., Jordan, M.I.: Generalized zero-shot learning with deep calibration network. In: NeurIPS (2018) 8, 11, 12
31. Liu, Y., Guo, J., Cai, D., He, X.: Attribute attention for semantic disambiguation in zero-shot learning. In: ICCV (2019) 11, 12
32. Long, Y., Liu, L., Shen, Y., Shao, L.: Towards affordable semantic searching: Zero-shot retrieval via dominant attributes. In: AAAI (2018) 3
33. Maas, A.L., Hannun, A.Y., Ng., A.Y.: Rectifier nonlinearities improve neural network acoustic models. In: ICML (2013) 7
34. Maaten, L.v.d., Hinton, G.: Visualizing data using t-SNE. Journal of Machine Learning Research **9**(Nov), 2579–2605 (2008) 14
35. Mandal, D., Narayan, S., Dwivedi, S.K., Gupta, V., Ahmed, S., Khan, F.S., Shao, L.: Out-of-distribution detection for generalized zero-shot action recognition. In: CVPR (2019) 3
36. Mensink, T., Verbeek, J., Perronnin, F., Csurka, G.: Distance-based image classification: Generalizing to new classes at near-zero cost. IEEE transactions on pattern analysis and machine intelligence **35**(11), 2624–2637 (2013) 2
37. Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J.: Distributed representations of words and phrases and their compositionality. In: NeurIPS (2013) 1
38. Mishra, A., Krishna Reddy, S., Mittal, A., Murthy, H.A.: A generative model for zero shot learning using conditional variational autoencoders. In: CVPR Workshops (2018) 2, 3, 6, 7, 11, 12, 13
39. Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., (Facebook), L.F., Chintala, S.: PyTorch: An imperative style, high-performance deep learning library. In: NeurIPS (2019) 9
40. Patterson, G., Hays, J.: Sun attribute database: Discovering, annotating, and recognizing scene attributes. In: CVPR (2012) 11, 12
41. Prenger, R., Valle, R., Catanzaro, B.: Waveglow: A flow-based generative network for speech synthesis. In: ICASSP (2019) 3
42. Qin, J., Liu, L., Shao, L., Shen, F., Ni, B., Chen, J., Wang, Y.: Zero-shot action recognition with error-correcting output codes. In: CVPR (2017) 3
43. Radovanović, M., Nanopoulos, A., Ivanović, M.: Hubs in space: Popular nearest neighbors in high-dimensional data. Journal of Machine Learning Research **11**(Sep), 2487–2531 (2010) 3

44. Romera-Paredes, B., Torr, P.: An embarrassingly simple approach to zero-shot learning. In: ICML (2015) 3, 6, 11, 12
45. Scheirer, W.J., de Rezende Rocha, A., Sapkota, A., Boult, T.E.: Toward open set recognition. IEEE transactions on pattern analysis and machine intelligence **35**(7), 1757–1772 (2012) 2, 3
46. Schonfeld, E., Ebrahimi, S., Sinha, S., Darrell, T., Akata, Z.: Generalized zero- and few-shot learning via aligned variational autoencoders. In: CVPR (2019) 2, 3, 8, 11, 12
47. Shen, Y., Liu, L., Shen, F., Shao, L.: Zero-shot sketch-image hashing. In: CVPR (2018) 3
48. Shen, Z., Lai, W.S., Xu, T., Kautz, J., Yang, M.H.: Exploiting semantics for face image deblurring. International Journal of Computer Vision (2020) 1
49. Shen, Z., Wang, W., Lu, X., Shen, J., Ling, H., Xu, T., Shao, L.: Human-aware motion deblurring. In: ICCV (2019) 1
50. Socher, R., Ganjoo, M., Sridhar, H., Bastani, O., Manning, C.D., Ng, A.Y.: Zero-shot learning through cross-modal transfer. In: NeurIPS (2013) 11, 12
51. Sohn, K., Lee, H., Yan, X.: Learning structured output representation using deep conditional generative models. In: NeurIPS (2015) 2, 8, 13
52. Sung, F., Yang, Y., Zhang, L., Xiang, T., Torr, P.H., Hospedales, T.M.: Learning to compare: Relation network for few-shot learning. In: CVPR (2018) 11, 12
53. Tolstikhin, I., Bousquet, O., Gelly, S., Schoelkopf, B.: Wasserstein auto-encoders. In: ICLR (2018) 8, 9
54. Tong, B., Wang, C., Klinkigt, M., Kobayashi, Y., Nonaka, Y.: Hierarchical disentanglement of discriminative latent features for zero-shot learning. In: CVPR (2019) 2, 5, 11, 12
55. Tran, D., Vafa, K., Agrawal, K.K., Dinh, L., Poole, B.: Discrete flows: Invertible generative models of discrete data. In: ICLR Workshops (2019) 3
56. Tsai, Y.H.H., Huang, L.K., Salakhutdinov, R.: Learning robust visual-semantic embeddings. In: ICCV (2017)
57. Tsai, Y.H.H., Liang, P.P., Zadeh, A., Morency, L.P., Salakhutdinov, R.: Learning factorized multimodal representations. In: ICLR (2019) 2, 6
58. Wah, C., Branson, S., Welinder, P., Perona, P., Belongie, S.: The Caltech-UCSD Birds-200-2011 Dataset. Tech. Rep. CNS-TR-2011-001, California Institute of Technology (2011) 11, 12
59. Wang, Q., Chen, K.: Zero-shot visual recognition via bidirectional latent embedding. International Journal on Computer Vision **124**(3), 356–383 (2017) 2
60. Xian, Y., Akata, Z., Sharma, G., Nguyen, Q., Hein, M., Schiele, B.: Latent embeddings for zero-shot classification. In: CVPR (2016) 11, 12
61. Xian, Y., Lampert, C.H., Schiele, B., Akata, Z.: Zero-shot learning—a comprehensive evaluation of the good, the bad and the ugly. IEEE transactions on pattern analysis and machine intelligence **41**(9), 2251–2265 (2018) 12
62. Xian, Y., Lorenz, T., Schiele, B., Akata, Z.: Feature generating networks for zero-shot learning. In: CVPR (2018) 2, 3, 6, 7, 8, 9, 11, 12, 13
63. Xian, Y., Schiele, B., Akata, Z.: Zero-shot learning-the good, the bad and the ugly. In: CVPR (2017) 11, 12
64. Xian, Y., Sharma, S., Schiele, B., Akata, Z.: f-VAEGAN-D2: A feature generating framework for any-shot learning. In: CVPR (2019) 2, 3, 11, 12
65. Xie, G.S., Liu, L., Jin, X., Zhu, F., Zhang, Z., Qin, J., Yao, Y., Shao, L.: Attentive region embedding network for zero-shot learning. In: CVPR (2019) 3
66. Yu, H., Lee, B.: Zero-shot learning via simultaneous generating and learning. In: NeurIPS (2019) 11, 12

67. Zhang, F., Shi, G.: Co-representation network for generalized zero-shot learning. In: ICML (2019) 3, 11, 12
68. Zhang, H., Koniusz, P.: Zero-shot kernel learning. In: CVPR (2018) 3
69. Zhang, L., Xiang, T., Gong, S.: Learning a deep embedding model for zero-shot learning. In: CVPR (2017) 3, 11, 12
70. Zhang, Z., Saligrama, V.: Zero-shot learning via semantic similarity embedding. In: ICCV (2015) 3, 11, 12
71. Zhu, Y., Xie, J., Liu, B., Elgammal, A.: Learning feature-to-feature translator by alternating back-propagation for generative zero-shot learning. In: ICCV (2019) 2