# Generalized Adversarially Learned Inference

**Yatin Dandi,**[1] **Homanga Bharadhwaj,** [2] **Abhishek Kumar,** [3] **Piyush Rai** [1]

[1] Indian Institute of Technology Kanpur
[2] University of Toronto, Vector Institute
[3] Google Brain

yatind@iitk.ac.in, homanga@cs.toronto.edu, abhishk@google.com, piyush@cse.iitk.ac.in

## Abstract

Allowing effective inference of latent vectors while training GANs can greatly increase their applicability in various downstream tasks. Recent approaches, such as ALI and BiGAN frameworks, develop methods of inference of latent variables in GANs by adversarially training an image generator along with an encoder to match two joint distributions of image and latent vector pairs. We generalize these approaches to incorporate *multiple* layers of feedback on **reconstructions**, **self-supervision**, and other forms of supervision based on prior or **learned knowledge** about the desired solutions. We achieve this by modifying the discriminator's objective to correctly identify more than two joint distributions of tuples of an arbitrary number of random variables consisting of images, latent vectors, and other variables generated through auxiliary tasks, such as reconstruction and inpainting or as outputs of suitable pre-trained models. We design a non-saturating maximization objective for the generator-encoder pair and prove that the resulting adversarial game corresponds to a global optimum that simultaneously matches all the distributions. Within our proposed framework, we introduce a novel set of techniques for providing self-supervised feedback to the model based on properties, such as patch-level correspondence and cycle consistency of reconstructions. Through comprehensive experiments, we demonstrate the efficacy, scalability, and flexibility of the proposed approach for a variety of tasks. The appendix of the paper can be found at the following link: https://drive.google.com/file/d/1i99e682CqYWMEDXlnqkqrctGLVA9viiz/view?usp=sharing

## Introduction

Recent advances in deep generative models have enabled modeling of complex high-dimensional datasets. In particular, Generative Adversarial Networks (GANs) (Goodfellow et al. 2017) and Variational Autoencoders (VAEs) (Kingma and Welling 2013) are broad classes of current state-of-the-art deep generative approaches, providing complementary benefits. VAE based approaches aim to learn an explicit inference function through an encoder neural network that maps from the data distribution to a latent space distribution.

On the other hand, GAN based adversarial learning techniques do not perform inference and directly learn a generative model to construct high-quality data, which are usually much more realistic than those generated by VAEs. However, due to the absence of an efficient inference mechanism it is not possible to learn rich unsupervised feature representations from data.

To address the above issues, recent approaches, in particular, Adversarially Learned Inference (ALI) and Bidirectional GAN (BiGAN) (Dumoulin et al. 2017; Donahue, Krähenbühl, and Darrell 2017) have attempted to integrate an inference mechanism within the GAN framework by training a discriminator to discriminate not only in the data space ($x$ vs $G(z)$), but discriminate *joint* samples of data and encodings $((x, E(x)))$ from joint samples of the generations and latent variables $((G(z), z))$. Here, $E(\cdot)$ denotes the encoder and $G(\cdot)$ denotes the generator. We argue that generalizing such adversarial joint distribution matching to *multiple* distributions and arbitrary number of random variables can unlock a much larger potential for representation learning and generative modeling that has not yet been explored by previous approaches (Dumoulin et al. 2017; Li et al. 2017a; Donahue, Krähenbühl, and Darrell 2017). Unlike other generative models such as VAEs, as we show, GANs can be generalized to match more than two joint distributions of tuples of an arbitrary number of random variables. We demonstrate that this allows integration of self-supervised learning and learned or prior knowledge about the properties of desired solutions.

Unlike previous approaches relying on pixel-level reconstruction objectives (Kingma and Welling 2013; Li et al. 2017a; Ulyanov, Vedaldi, and Lempitsky 2018) which are known to be one of the causes of blurriness (Li et al. 2017a; Zhao, Song, and Ermon 2017), we propose an approach for incorporating multiple layers of reconstructive feedback within the adversarial joint distribution matching framework. Our approach allows incorporation of such task-specific self-supervised feedback and knowledge of equivariance of reconstructions to dataset-dependent transformations within the framework of adversarial joint distribution matching, without requiring careful tuning of weighing terms for different objectives. In particular, we consider a discriminator that classifies joint samples of an arbitrary number of random variables into an arbitrary num-

(a) Chain of reconstructive feedback. The images denote real reconstructions from GALI-4.
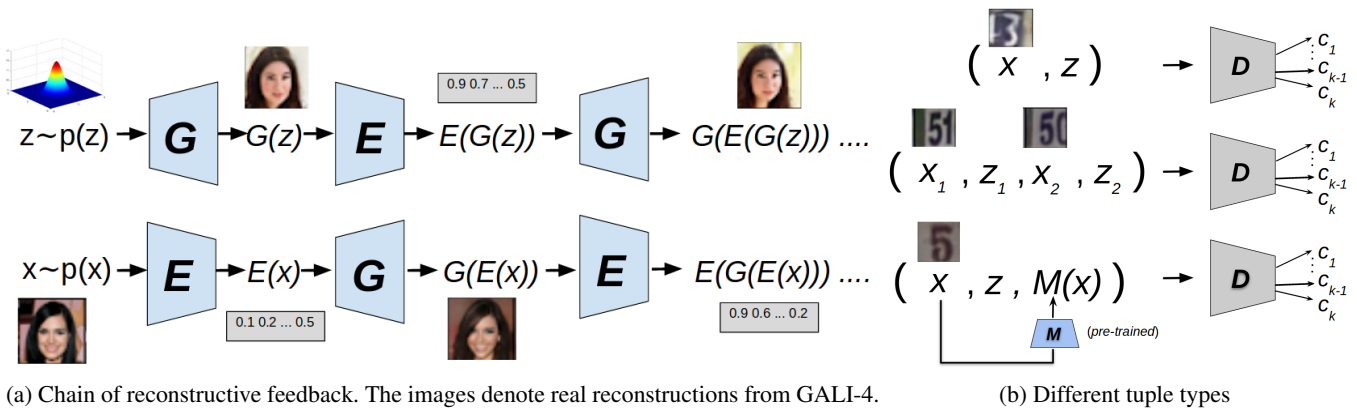
(b) Different tuple types

Figure 1: (a) The two sequences generated by recursively applying the encoder and the generator to images and latent vectors respectively. The top sequence starts from a latent vector drawn from the fixed prior while the bottom sequence starts from a real image. The images correspond to actual reconstructions from the proposed GALI approach. In particular, GALI-4 (i.e. GALI with $n = 4$ as described in "The Proposed Approach" was used to generate the reconstructions. In the figure $G$ denotes the generator, $E$ the encoder (the same set of $G$ and $E$ parameters are shared across each use), $z$ the latent vector, and $x$ the input image. (b) Illustration of the different tuple types that can be input to the discriminator to provide different types of feedbacks via distribution matching. Here, $M$ denotes an external pre-trained neural network and $M(x)$ denotes the features corresponding to image $x$ through model $M$. The multi-class discriminator $D$ outputs one of the $k$ classes.

ber of classes. Each class essentially represents a distribution over tuples of image and latent vectors, defined by recursively computing the encodings and their reconstructions. This provides multiple layers of information for each real image or sampled vector while allowing the generator-encoder pair to gain *explicit* feedback on the quality and relevance of different types of reconstructions. Fig. 1 illustrates this process through a diagram. In the rest of the paper, we refer to our proposed framework as **Generalized Adversarially Learned Inference (GALI)**.

While Adversarially Learned Inference (ALI) can be generalized to multi-class classification within the framework of minimax likelihood based objective, the resulting training procedure is still susceptible to vanishing gradients for the generator-encoder. We illustrate this problem and devise an alternative objective that extends the non saturating GAN objective to multiple distributions. We develop a generalized framework for distribution matching with the following main contributions:

1. We introduce a scalable approach for incorporating multiple layers of knowledge-based, reconstructive and self-supervised feedback in adversarially learned inference without relying on fixed pixel or feature level similarity metrics.
2. We propose a non-saturating objective for training a generator network when the corresponding discriminator performs multi-class classification. We further prove that our proposed objective has the same global optima as the minimax objective which matches all the distributions simultaneously.
3. We demonstrate how the proposed approach can incorporate pre-trained models and can naturally be adapted for particular tasks such as image inpainting by incorporating suitably designed auxiliary tasks within the framework of adversarial joint distribution matching.

## Preliminaries

The following minimax objective serves as the basic framework for optimization in the ALI/BiGAN framework.

$$\min_{G,E} \max_{D} V(D, E, G) \qquad (1)$$

where

$$V(D, E, G) := \mathbb{E}_{\mathbf{x} \sim p_\mathbf{X}} \big[ \underbrace{\mathbb{E}_{\mathbf{z} \sim p_E(\cdot|\mathbf{x})} \big[ \log D(\mathbf{x}, \mathbf{z}) \big]}_{\log D(\mathbf{x}, E(\mathbf{x}))} \big]$$
$$+ \mathbb{E}_{\mathbf{z} \sim p_\mathbf{Z}} \big[ \underbrace{\mathbb{E}_{\mathbf{x} \sim p_G(\cdot|\mathbf{z})} \big[ \log \big( 1 - D(\mathbf{x}, \mathbf{z}) \big) \big]}_{\log(1 - D(G(\mathbf{z}), \mathbf{z}))} \big].$$

Here, the generator $G$ and encoder $E$ can either be both deterministic, such that $G : \Omega_\mathbf{Z} \to \Omega_\mathbf{X}$ with $p_G(\mathbf{x}|\mathbf{z}) = \delta(\mathbf{x} - G(\mathbf{z}))$ and $E : \Omega_\mathbf{X} \to \Omega_\mathbf{Z}$ with $p_E(\mathbf{z}|\mathbf{x}) = \delta(\mathbf{z} - E(\mathbf{x}))$ or the encoder can be stochastic. Deterministic models were used in the BiGAN (Donahue, Krähenbühl, and Darrell 2017) approach, while a stochastic encoder was used in ALI (Dumoulin et al. 2017). For all our experiments and discussions, we use a stochastic encoder following ALI (Dumoulin et al. 2017) but denote samples from $p_E(\mathbf{z}|\mathbf{x})$ as $E(\mathbf{x})$ for notational convenience. Under the assumption of an optimal discriminator, minimization of the generator-encoder pair's objective is equivalent to minimization of the Jensen-Shannon (JS) divergence (Lin 1991) between the two joint distributions. Thus, achieving the global minimum of the objective is equivalent to the two joint distributions becoming equal.

## The Proposed Approach

The proposed approach is based on the two sequences in Fig. 1 : the top sequence starts from a real latent variable and its corresponding generation and contains all the subsequent reconstructions and their encodings while the bottom sequence starts from a real image and contains its corre-

sponding set of reconstructions and encodings. Ideally, with perfect reconstructions, we wish all images and latent vectors within a sequence to be identical to other images and latent vectors respectively. We argue that the optimization objective presented in the previous section is the simplest case of a general family of objectives where the discriminator tries to classify $n$ classes of tuples of size $m$, of images and latent vectors with the variables within a tuple all belonging to one of the sequences in Fig 1 while the generator tries to fool it to incorrectly classify. By including additional latent vectors and images, we allow the generator-encoder pair to receive multiple layers of reconstructive feedback on each latent vector and each generated image while modifying the discriminator to an $n$-way classifier encourages it to perform increasingly fine grained discrimination between different image-latent variable tuples. We experimentally demonstrate results for $(n = 4, m = 2)$, $(n = 8, m = 2)$, and $(n = 8, m = 4)$. Our goal is to design an objective where the discriminator is tasked with discriminating against each of the joint distributions specified by the tuples, and the generator and encoder try to modify the joint distributions such that distributions of all the classes of tuples are indistinguishable from each other. We analyse different alternatives for the same in the subsequent sections.

## Multiclass Classifier Discriminator

We first consider the expected log-likelihood based mini-max objective for the case of $n = 4$ classes and tuples of size $m = 2$. We choose the set of pairs (classes) to be: $(\mathbf{x}, E(\mathbf{x})), (G(\mathbf{z}), \mathbf{z}), (\mathbf{x}, E(G(E(\mathbf{x})))), (G(E(G(\mathbf{z}))), \mathbf{z})$. The discriminator is modified to perform multi-class classification with the output probabilities of input (image, latent vector) $(\mathbf{x}_{in}, \mathbf{z}_{in})$ for the $i^{th}$ class denoted by $D_i(\mathbf{x}_{in}, \mathbf{z}_{in})$. So, the output of $D(\mathbf{x}_{in}, \mathbf{z}_{in})$ is a vector $[D_1(\mathbf{x}_{in}, \mathbf{z}_{in}), ... D_i(\mathbf{x}_{in}, \mathbf{z}_{in}), ..]$ where $D_i(\mathbf{x}_{in}, \mathbf{z}_{in})$ denotes the output probability for the $i^{th}$ class. The minimax objective with a multi-class classifier discriminator, following a straightforward generalization of ALI in Eq. (1) thus becomes:

$$\min_{G,E} \max_D V(D, E, G) \tag{2}$$

where

$$
\begin{aligned}
V(D, E, G) := \; & \mathbb{E}_{\mathbf{x} \sim p_{\mathbf{x}}} \big[ \log \left( D_1(\mathbf{x}, E(\mathbf{x})) \right) \big] \\
& + \mathbb{E}_{\mathbf{z} \sim p_{\mathbf{z}}} \big[ \log \left( D_2(G(\mathbf{z}), \mathbf{z}) \right) \big] \\
& + \mathbb{E}_{\mathbf{x} \sim p_{\mathbf{x}}} \big[ \log \left( D_3(\mathbf{x}, E(G(E(\mathbf{x})))) \right) \big] \\
& + \mathbb{E}_{\mathbf{z} \sim p_{\mathbf{z}}} \big[ \log \left( D_4(G(E(G(\mathbf{z}))), \mathbf{z}) \right) \big].
\end{aligned}
\tag{3}
$$

Although the above adversarial game captures the multiple layers of reconstructive feedback described in Fig. 1, it is insufficient for stable training due to *vanishing gradients* in generator-encoder pair's training. Consider the gradients for the parameters of the generator and the encoder with the above objective. Since the gradient of the softmax activation function w.r.t the logits vanishes whenever one of the logits dominates the rest, when the discriminator is able to classify accurately, the gradients of the generator-encoder's objective nearly vanish and the generator-encoder pair does not receive any feedback. The inability to produce realistic images

due to vanishing gradients is also demonstrated empirically in Figure 5 (Appendix). In order to remedy this, we provide an alternate training objective for the generator-encoder based on products of the likelihoods. We first describe how the vanishing gradients problem cannot be avoided by using an objective based on misclassification likelihood.

**Misclassification Likelihood**   At first it might seem that a natural way to alleviate the vanishing gradient problem is to replace each log likelihood term $\log \left( D_i(\mathbf{x}_{in}, \mathbf{z}_{in}) \right)$ in the mini-max generator-encoder minimization objective with the corresponding misclassification log-likelihood $\log \left( 1 - D_i(\mathbf{x}_{in}, \mathbf{z}_{in}) \right)$ for the given class to construct a maximization objective. This is the approach used while designing the non-saturating objectives for standard GANs (Arjovsky and Bottou 2017), ALI, and BiGAN frameworks (Dumoulin et al. 2017; Donahue, Krähenbühl, and Darrell 2017). However, in the multiclass classification framework, the value and the corresponding gradients for the misclassification objective can vanish even when the discriminator learns to accurately reject many of the incorrect classes as long as it has a low output probability for the true class. For example, for the 4 classes considered above, the discriminator may learn to accurately reject classes 3 and 4 for a pair belonging to class 1 but might still have a high misclassification likelihood if it incorrectly identifies the pair as belonging to class 2. As such an objective does not optimize the individual probabilities for the incorrect classes, the gradient for the generator-encoder pair would provide no feedback for causing an increase in the output probabilities for the remaining classes (3 and 4).

**Product of Terms**   In light of the above, we desire to obtain an objective such that instead of just encouraging the generator-encoder to cause lower discriminator output probability for the *right* class, the generator-encoder's gradient for the modified objective enforces each of the *wrong* classes to have high output probabilities. With this goal, we propose the product of terms objective that explicitly encourages all the distributions to match. The proposed objective for the 4 classes considered above is given below:

$$
\begin{aligned}
\max_{G,E} [ & \mathbb{E}_{\mathbf{x} \sim p_{\mathbf{x}}} [\log \left( D_2(\mathbf{x}, E(\mathbf{x})) D_3(\mathbf{x}, E(\mathbf{x})) D_4(\mathbf{x}, E(\mathbf{x})) \right)] \\
& + \mathbb{E}_{\mathbf{z} \sim p_{\mathbf{z}}} [\log \left( D_1(G(\mathbf{z}), \mathbf{z}) D_3(G(\mathbf{z}), \mathbf{z}) D_4(G(\mathbf{z}), \mathbf{z}) \right)] \\
& + \mathbb{E}_{\mathbf{x} \sim p_{\mathbf{x}}} [\log \left( D_1(\mathbf{x}, E(G(E(\mathbf{x})))) D_2(\mathbf{x}, E(G(E(\mathbf{x})))) \right. \\
& \qquad D_4(\mathbf{x}, E(G(E(\mathbf{x})))) )] \\
& + \mathbb{E}_{\mathbf{z} \sim p_{\mathbf{z}}} [\log \left( D_1(G(E(G(\mathbf{z}))), \mathbf{z}) D_2(G(E(G(\mathbf{z}))), \mathbf{z}) \right. \\
& \qquad D_3(G(E(G(\mathbf{z}))), \mathbf{z}) )]]
\end{aligned}
\tag{4}
$$

Since the vanishing gradients problem only arises in the generator-encoder, we use the same objective as 3 for the discriminator. In Eq. 4, each of the terms of the form $\log \left( D_i(\mathbf{x}_{in}, \mathbf{z}_{in}) D_j(\mathbf{x}_{in}, \mathbf{z}_{in}) D_k(\mathbf{x}_{in}, \mathbf{z}_{in}) \right)$ can be further split as $\log \left( D_i(\mathbf{x}_{in}, \mathbf{z}_{in}) \right) + \log \left( D_j(\mathbf{x}_{in}, \mathbf{z}_{in}) \right) + \log \left( D_k(\mathbf{x}_{in}, \mathbf{z}_{in}) \right)$. The above objective encourages the parameters of the generator and encoder to ensure that none of the incorrect classes are easily rejected by the discriminator. This is because the generator-encoder pair is explic-

itly trained to cause an increase in the discriminator output probabilities for all the *wrong* classes. Moreover, the objective does not lead to vanishing gradients as discarding any of these classes as being true incurs a large penalty in terms of the objective and its gradient. In the next subsection and the appendix we show that the Product of Terms objective has a global optimum which matches all the joint distributions corresponding to the different classes of tuples simultaneously. We further demonstrate the effectiveness of this objective through experiments.

## The Optimal Discriminator

The optimal discriminator $D^*$ for the discriminator's objective in Eq. (3), can be described as $D^* := \arg\max_D V(D, E, G)$, for any $E$ and $G$. Following the derivation in the appendix, we obtain the following functional form of $D^*$: $D_i^*(\mathbf{x}_{in}, \mathbf{z}_{in}) = \frac{p_i(\mathbf{x}_{in}, \mathbf{z}_{in})}{\sum_{j=1}^{4} p_j(\mathbf{x}_{in}, \mathbf{z}_{in})}$.

Here, $D_i$ corresponds to the output probability of the $i^{th}$ class among the four classes: $(\mathbf{x}, E(\mathbf{x})), (G(\mathbf{z}), \mathbf{z}), (\mathbf{x}, E(G(E(\mathbf{x})))), (G(E(G(\mathbf{z}))), \mathbf{z})$. $p_i$ is the joint probability density of the corresponding $\mathbf{x}_{in}$ and $\mathbf{z}_{in}$ in each of the $(\mathbf{x}_{in}, \mathbf{z}_{in})$ pairs above.

## The Optimal Generator-Encoder for the Product of Terms Objective

Following the derivation provided in the appendix and substituting the optimal discriminator found above ($D^* = \arg\max_D V(D, E, G)$) in the Product of Terms objective (Eq. 4) for the generator-encoder leads to the maximization objective: $C(G, E) \leq -\log\left(4^9\right) - JSD_{\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4}}(p_1, p_2, p_3, p_4)$. where $JSD_{\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4}}$ denotes the generalized Jensen–Shannon divergence (Lin 1991) with equal weights assigned to each distribution as described in the appendix. Since the JSD for the four distributions is non negative and vanishes if and only if $p_1 = p_2 = p_3 = p_4$, the global optimum for the product of terms objective is given by: $p_{(\mathbf{x}, E(\mathbf{x}))} = p_{(G(\mathbf{z}), \mathbf{z})} = p_{(\mathbf{x}, E(G(E(\mathbf{x}))))} = p_{(G(E(G(\mathbf{z}))), \mathbf{z})}$ This is the same as the optima of the original generalized objective in Eq. (2) (appendix). However, our proposed objective matches all the distributions simultaneously without suffering from vanishing gradients.

## Extension to Arbitrary Number and Size of Tuples

The analysis presented above can be extended to accommodate any number $n$ of tuples of arbitrary size $m$ from the two chains in 1. We demonstrate this for the case of $n = 8$ and $m = 4$ for the SVHN dataset. We start with one tuple from each chain: $(\mathbf{x}, E(\mathbf{x}), G(E(\mathbf{x})), E(G(E(\mathbf{x}))))$ and $(G(\mathbf{z}), \mathbf{z}, G(E(G(\mathbf{z}))), E(G(\mathbf{z})))$ and construct additional tuples by permuting within the images and latent vectors for both of these tuples to give a total of $4 \times 2$ tuples. These classes of 4-tuples allow the discriminator to directly discriminate between an image and its reconstruction in both image and latent space. The number of pairs $n$ to be considered and the size of tuples is limited only by computational cost, although intuitively we expect to see diminishing returns in terms of performance when increasing $n$ beyond a
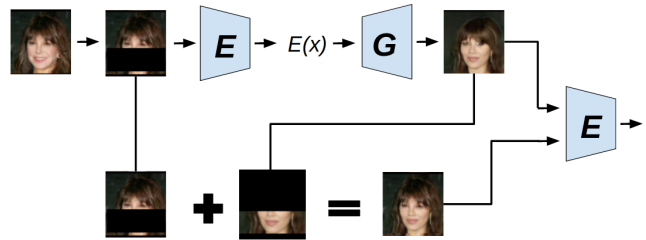


Figure 2: Sections from reconstructions of real images corresponding to missing patches are combined with the original real images to form a set of *mixed* images.

point as matching the distributions for two random variables enforces the matching of the subsequent chains.

## Self-Supervised Feedback Using Mixed Images

For tasks such as image inpainting and translation, an important desideratum for a reconstructed image is its consistency with parts of original image. We demonstrate that such task specific feedback can be incorporated into the GALI framework through self-supervised learning tasks. We experiment with incorporating one such task into the reconstructive chain of Fig. 1, namely ensuring consistency and quality of *mixed* images constructed by combining the inpainted sections of reconstructions of real images with the other parts of the corresponding original real images. During training, we mask randomly sampled regions of the images input to the encoder, as shown in Fig. 2. Our analysis of the optimal solutions proves that the multi-class classifier based approach can be used to match any number of distributions of (image, latent variable) pairs, as long as each such pair is generated using the real data distribution, fixed prior for the noise or any operation on these such that the output's distribution depends only on generator-encoder parameters. Since the set of mixed images and their encodings satisfy these constraints, introducing them into the set of (image, latent vector) pairs used in the objective leads to matching of the distribution of mixed images with the real image distribution as well as the matching of correspondence between the encodings and reconstructions. Thus, this type of feedback introduces an additional constraint on the model to ensure the consistency of the inpainted patch with the original image. We show this empirically in the experiments section. The self-supervision based model is constructed by introducing 4 additional classes of (image,latent-vector) pairs into the GALI-4 model so that the discriminator is trained to perform an 8-way classification. These additional classes depend on a distribution of masks that are applied to the images. For our experiments, we sample a mask $M$ for CelebA 64X64 images as follows: first, the height $w$ and the width $h$ of the mask are both independently drawn uniformly from the set $1, \cdots, 64$. Subsequently, the $x$ (horizontal) and $y$ (vertical) indices (index origin = 0) of the bottom-left corner of the mask are drawn uniformly from the set $0, \cdots, 63 - h + 1$ and $0, \cdots, 63 - w + 1$. The above procedure thus defines a probability distribution over masks $P(M)$ For every image input to the encoder, a new mask $M$ is sampled independently. Thus

the four classes of distributions considered in GALI-4 are modified to classes whose samples are constructed as : $(\mathbf{x}, E(M_1(\mathbf{x})))$, $(G(\mathbf{z}), \mathbf{z})$, $(\mathbf{x}, E(M_2(G(E(M_1(\mathbf{x}))))))$, $(G((E(M_3(G(\mathbf{z}))), \mathbf{z})$ where $M_1$, $M_2$, and $M_3$ are independently drawn from $P(M)$. For a sampled real image and its corresponding mask $M_1$, a mixed image $Mix(\mathbf{x}, M_1)$ is constructed by replacing the masked region of $\mathbf{x}$ by the corresponding region of $G(E(M_1(\mathbf{x})))$. The additional 4 pairs then correspond to $(Mix(\mathbf{x}, M_1), E(M_1(\mathbf{x})))$, $(Mix(\mathbf{x}, M_1)$ , $E(M_2(G(E(M_1(\mathbf{x}))))))$, $(\mathbf{x}, E(M_4(Mix(\mathbf{x}, M_1)))$ and $(G(E(M_1(\mathbf{x}))), E(M_4(Mix(\mathbf{x}, M_1)))$

## Incorporation of Learned Knowledge

Matching the real and generated data distributions should also result in the matching of the corresponding joint distributions of any of the extractable properties such as attributes, labels, perceptual features or segmentation masks. We argue that utilizing these outputs during training can impose additional constraints and provide additional information to the model similar to the reconstructive and self-supervised feedbacks discussed above. We demonstrate that our approach offers a principled way for incorporating these properties by introducing the final or hidden layer outputs of pretrained models as additional random variables in the tuples. For each class of tuples, these outputs could correspond to input any input image within the same chain as other images and latent variables in the tuple. For experiments and subsequent discussions, we consider the four classes of tuples from the objective in 4 and augment each with the respective outputs from a pre-trained model $M$ to obtain the tuple classes $(\mathbf{x}, E(\mathbf{x}), M(\mathbf{x}))$,$(G(\mathbf{z}), \mathbf{z}, M(G(\mathbf{z})))$, $(\mathbf{x}, E(G(E(\mathbf{x}))), M(G(E(\mathbf{x})))$, $(G(E(G(\mathbf{z}))), \mathbf{z}, M(G(\mathbf{z})))$. While feature level reconstructive feedback can be provided through $L_1$ or $L_2$ reconstruction objectives on the output features, our approach explicitly matches the joint distribution of these features with images and latent vectors of all classes (real, fake, reconstructed, etc.) Unlike, $L_1$ and $L_2$ based reconstruction terms which directly affect only the generator-encoder, our approach distributes the feedback from the pretrained model across all the components of the model including the discriminator. Moreover, as we demonstrate in the ablation study included in the appendix, by jointly providing reconstructive feedback on images, encodings and pretrained model activations, our model with learned knowledge outperforms the model based on $L_2$ reconstruction of features on all metrics. Our approach can also be applied to any arbitrary type of model outputs such as segmentation masks, translations or labels without blurriness effects.

## Related Work

A number of VAE-GAN hybrids have been proposed in recent years. A unified perspective based on Lagrangian dual functions for these hybrids was provided by (Zhao, Song, and Ermon 2018). Adversarial autoencoders (Makhzani et al. 2016) use GANs to match the aggregated posterior of the latent variables with the prior distribution instead of the KL-divergence minimization term in VAEs. VAE-GAN. (Larsen et al. 2016) replaced the pixel-wise reconstruction error term with an error based on the features inferred by the discriminator. AVB (Mescheder, Nowozin, and Geiger 2017) proposed using an auxiliary discriminator to approximately train a VAE model. Adversarial Generator-Encoder Networks (Ulyanov, Vedaldi, and Lempitsky 2018) constructed an adversarial game directly between the generator and encoder without any discriminator to match the distributions in image and latent space. The model still relied on $L_2$ reconstruction to enforce cycle consistency. Unlike the above approaches, BIGAN(Donahue, Krähenbühl, and Darrell 2017), ALI(Dumoulin et al. 2017) and our models do not rely on any reconstruction error term. Although a recent work BigBiGAN (Donahue and Simonyan 2019) demonstrated how the ALI/BiGAN framework alone can allow achieving competitive representation learning, we emphasize that embedding *more information* through multiple layers of feedback would further improve the inference. Some recent approaches (Gan et al. 2017; Li et al. 2017b; Pu et al. 2018; **?**) propose different frameworks for adversarially matching multiple joint distributions for domain transformation, conditional generation, and graphical models. We hypothesize that the use of our proposed product of terms objective and the proposed different types of feedback should lead to further improvements in these tasks similar to the improvements demonstrated for ALI in the experiments.

ALICE (Li et al. 2017a) illustrated how the ALI objective with stochasticity in both the generator and the encoder can easily result in an optimal solution where cycle consistency is violated ($x \neq \tilde{x}$) due to the non-identifiability of solutions. The analysis however does not apply to our approach as our optimal solution explicitly matches additional joint distributions which involve reconstructions and their corresponding encodings.

In augmented BiGAN (Kumar, Sattiger, and Fletcher 2017), instead of generalizing the discriminator to perform multiclass classification, the fake distribution is divided into two sources (one of generated images and latent vectors and the other of encodings of real images and their reconstructions) and a weighted average of the likelihood of these two parts is used for the dicriminator and generator's objective. Unlike our optimal solution which matches all the distributions simultaneously, the augmented BiGAN's optimal solution matches the (real image, encoding) distribution with the average of the two "fake" distributions. This solution causes a trade-off between good reconstructions and good generation, which is avoided in our method since all the distributions are enforced to match simultaneously.

## Experiments

Through experiments on two benchmark datasets, SVHN (Netzer et al. 2011) and CelebA (Liu et al. 2015), we aim to assess the reconstruction quality, meaningfulness of the learned representations for use in downstream tasks and generation, effects of extending the approach to more classes of tuples and larger tuple size, the ability of the proposed approach to incorporate knowledge from pretrained models trained for a different task, and its

| MODEL | PIXEL-MSE | FEATURE-MSE |
|---|---|---|
| ALI | $0.023 \pm 0.0024$ | $0.0857 \pm 0.0075$ |
| ALICE | $0.0096 \pm 0.00067$ | $0.0736 \pm 0.0058$ |
| GALI-4 | $0.0132 \pm 0.0011$ | $0.0717 \pm 0.0067$ |
| GALI-8 | $\mathbf{0.0095} \pm 0.0016$ | $\mathbf{0.066} \pm 0.0060$ |
| GALI-PT | $\mathbf{0.0093} \pm 0.00099$ | $\mathbf{0.041} \pm 0.0046$ |

| MODEL | PIXEL-MSE | FEATURE-MSE |
|---|---|---|
| ALI | $0.074 \pm 0.004$ | $0.307 \pm 0.018$ |
| ALICE | $0.042 \pm 0.002$ | $0.248 \pm 0.013$ |
| GALI-4 | $\mathbf{0.036} \pm 0.002$ | $\mathbf{0.201} \pm 0.012$ |
| GALI-PT | $\mathbf{0.032} \pm 0.0019$ | $\mathbf{0.131} \pm 0.008$ |

Table 1: Pixel-wise Mean-Squared Error (MSE) and Feature level Mean Squared Error (MSE) on the test set for SVHN (left table) and CelebA (right table). Lower is better.

| Model | Misc rate (%) |
|---|---|
| VAE (Kingma and Welling 2013) | 36.02 |
| DCGAN + L2-SVM | 22.18 |
| SDGM (Hayashi and Uchida 2019) | 16.61 |
| NC (Belghazi, Oquab, and Lopez-Paz 2019) | 17.12 |
| ALI (Dumoulin et al. 2017) | $19.11 \pm 0.50$ |
| ALI (baseline) | $19.05 \pm 0.53$ |
| ALICE (Li et al. 2017a) | $18.89 \pm 0.52$ |
| GALI-4 | $\mathbf{16.58} \pm \mathbf{0.38}$ |
| GALI-8 | $\mathbf{15.82} \pm \mathbf{0.43}$ |
| GALI-PT (supervised) | $\mathbf{11.43} \pm \mathbf{0.30}$ |

Table 2: Missclassification rate of GALI-4 and various baselines on the test set of SVHN dataset demonstrating the usefulness of learned representations. The results for baselines are from (Dumoulin et al. 2017). Lower is better.

## Notation and Setup

For all the experiments,following the description in the section "The Proposed Approach", GALI-4 is the proposed GALI model with 4 terms, GALI-8 is the proposed GALI model with 8 terms and GALI-PT is the GALI-4 model augmented with a pretrained network $M$.

For all our proposed models and both ALI (Dumoulin et al. 2017) and ALICE (Li et al. 2017a) (ALI + $L_2$ reconstruction error) baselines, we borrow the architectures from (Dumoulin et al. 2017) with the discriminator using spectral normalization (Miyato et al. 2018) instead of batch normalization and dropout (Srivastava 2013). As shown in Table 3 and Figure 3, our baseline obtains similar representation learning scores and reconstruction quality as reported in ALI (Dumoulin et al. 2017). All the architectural details and hyper-parameters considered are further described in the appendix.

## Reconstruction Quality

We evaluate the reconstruction quality on test images using both pixel level and semantic similarity metrics. For pixel level similarity, we report the average pixel wise mean squared error on test datasets for SVHN and CelebA. For semantic similarity, we use the mean squared error of the features from pre-trained multi-digit classification model (Goodfellow et al. 2014) trained on SVHN and a pretrained attribute classification model for SVHN and CelebA datasets respectively. Further details of these models are provided in the appendix.

The test set reconstruction errors on SHVN and CelebA ($64 \times 64$) datasets for the proposed models, the ALI baseline, and the ALICE baseline are shown in Table 1. As reported in (Li et al. 2017a), the improvements in reconstructions for ALICE are obtained at the cost of blurriness in images. The blurriness is visible in the reconstructions shown in Fig. 3a and Fig. 3b and quantitatively verified through the higher

FID score of ALICE in Table 3 as well as the higher values of the feature level MSE for both datasets. Moreover, we observe that the introduction of L2 reconstruction error does not lead to significant gains in the usefulness of the learned representations for downstream tasks. Our models achieve significant improvements in the quality of reconstructions and representations over the ALI baseline without succumbing to the above drawbacks.

## Representation Learning

In Table 2, we evaluate the representation learning capabilities of the encoder by training a linear SVM model on features corresponding to 1000 labelled images from the training set. Following (Dumoulin et al. 2017), the feature vectors are obtained by concatenating the last three hidden layers of the encoder as well as its output.The hyperparameters of the SVM model are selected using a held-out validation set. We report the average test set misclassification rate for 100 different SVMs trained on different random 1000-example training sets. We also report the misclassification rate for other semi-supervised learning approaches for comparison.

## Image Generation Quality

It is important to ensure that improved reconstructions do not come at the cost of poor generation quality. Such a trade-off is possible when the encoder-generator pair learns to encode pixel level information. It is instead desired that variations in the latent space of the generator correspond to variations in explanatory factors of variation of the data. We evaluate our model's image generation ability using the Frechet Inception Distance (FID) metric (Heusel et al. 2017) on the CelebA dataset in Table 3. We provide visualizations of the generated samples in the appendix.

## Image Inpainting

In Fig. 4, we evaluate our mixed images based model on an image inpainting task for the CelebA dataset through comparisons with ALI and ALICE baselines trained with the same procedure of masking out inputs. The ALICE based
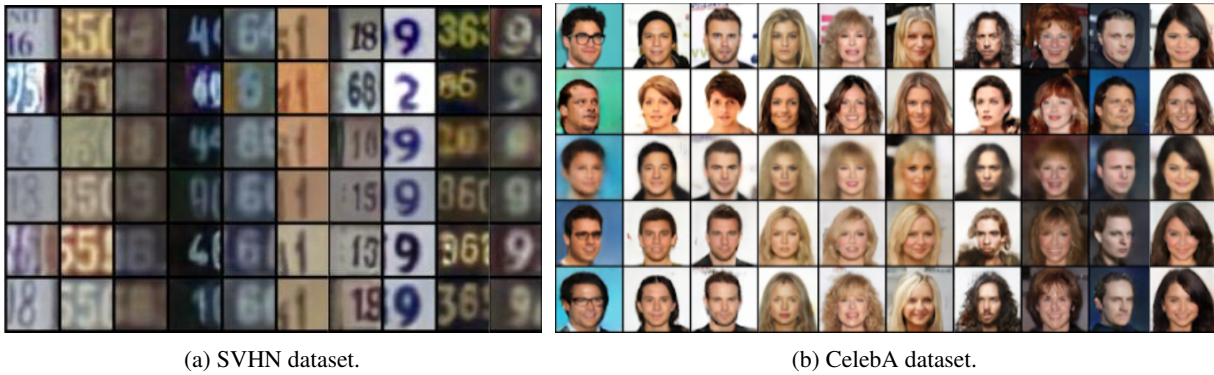
(a) SVHN dataset.



(b) CelebA dataset.

Figure 3: *Top to bottom*: Original images, images reconstructed by ALI, ALICE, GALI-4, GALI-8, and GALI-PT for (a) and ALI, ALICE, GALI-4, and GALI-PT for (b).

| MODEL | FID |
|---|---|
| ALI | 24.49 |
| ALICE | 36.91 |
| GALI-4 | **23.11** |
| GALI-PT | **10.13** |

| MODEL | PIXEL-MSE | FEATURE-MSE |
|---|---|---|
| ALI | $0.060 \pm 0.0080$ | $0.38 \pm 0.016$ |
| ALICE | $0.047 \pm 0.0069$ | $0.32 \pm 0.015$ |
| GALI-4 | $0.046 \pm 0.0066$ | $0.285 \pm 0.012$ |
| GALI-MIX | $\mathbf{0.031} \pm 0.0050$ | $\mathbf{0.164} \pm 0.010$ |

Table 3: *Top*: FID scores on CelebA demonstrating image generation quality of GALI-4, GALI-pretrain and the baselines. Lower is better.
*Bottom*: Pixel-wise Mean-Squared Error (MSE) and Feature level Mean Squared Error (MSE) on the image-inpainting task for CelebA. Lower is better.

model leads to blurriness whereas ALI suffers from poor consistency with the original images. Our approach alleviates both these issues. Quantitative comparisons for this task are also described in Fig. 4. Similar to the evaluation of reconstruction quality, we quantitatively evaluate image inpainting capabilities of the models using both pixel and feature level metric. In Table 3(bottom), pixel-level MSE denotes the average pixel wise squared difference of the inpainted region with the original image, while the feature level MSE is calculated as the average squared difference between the feature vectors of the inpainted image and the original image for each model. The feature vectors are calculated using the same pretrained classifier as used for the reconstruction quality. In the table, we denote our proposed mixed images based model as GALI-mix while ALI and ALICE denote models obtained by utilizing the same distribution over masks as GALI-mix while feeding images to the encoder in the respective baselines. GALI-4 similarly denotes the GALI-4 model augmented with masked inputs. Results for GALI-4 without self-supervision are included as part of an ablation study in the appendix.



Figure 4: *Top to bottom*: Original images, incomplete input images with the blackened region denoting an applied random occlusion mask, inpainted images from ALI , ALICE and GALI-4 with mixed images.

## Utilization of Pretrained Models

For SVHN, the pretrained model M outputs features from the pre-trained multi-digit classification model used for the feature-level MSE above while for CelebA, M outputs the features from the pretrained inception net (Szegedy et al. 2017) used for calculating the FID scores. We emphasize that our goal is to demonstrate the ability of our approach to incorporate *learned knowledge* from other models. These approaches however do not correspond to truly unsupervised settings as the pretrained models utilize supervision in the form of labelled SVHN digits and imagenet labels for SVHN and CelebA datasets respectively. This leads to expected yet significant improvements in metrics based on the output features of or the same training tasks as the pre-trained models such as misclassification rate and feature-MSE for SVHN and FID for CelebA. The above results demonstrate that the improvements further carry over to other independent metrics such as feature and pixel-MSE for CelebA and pixel-MSE for SVHN, indicating significant improvements in the overall reconstruction and image generation quality.

## Conclusion

In this paper, we proposed a novel framework for incorporating different types of generalized feedbacks in adversarially learned inference along with a non-saturating objective for adversarially matching multiple distributions. Through experiments on two benchmark datasets, SVHN and CelebA, we demonstrated the efficacy of the proposed framework in terms of improvements in reconstruction quality, representation learning, image generation, and image inpainting as compared to previous approaches.

# References

Arjovsky, M.; and Bottou, L. 2017. Towards Principled Methods for Training Generative Adversarial Networks. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net. URL https://openreview.net/forum?id=Hk4\_qw5xe.

Belghazi, M.; Oquab, M.; and Lopez-Paz, D. 2019. Learning about an exponential amount of conditional distributions. In *Advances in Neural Information Processing Systems*, 13703–13714.

Donahue, J.; Krähenbühl, P.; and Darrell, T. 2017. Adversarial feature learning. In *Proc. ICLR*.

Donahue, J.; and Simonyan, K. 2019. Large scale adversarial representation learning. In *Advances in Neural Information Processing Systems*, 10541–10551.

Dumoulin, V.; Belghazi, I.; Poole, B.; Arjovsky, A. L. M.; Mas-tropietro, O.; and Courville, A. C. 2017. Adversarial feature learning. In *Proc. ICLR*.

Gan, Z.; Chen, L.; Wang, W.; Pu, Y.; Zhang, Y.; Liu, H.; Li, C.; and Carin, L. 2017. Triangle generative adversarial networks. In *Advances in neural information processing systems*, 5247–5256.

Goodfellow, I.; Bulatov, Y.; Ibarz, J.; Arnoud, S.; and Shet, V. 2014. Multi-digit Number Recognition from Street View Imagery using Deep Convolutional Neural Networks. In *ICLR2014*.

Goodfellow, I. J.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y. 2017. Generative Adversarial Networks. In *Advances in Neural Information Processing Systems, 2014*.

Hayashi, H.; and Uchida, S. 2019. SDGM: Sparse Bayesian Classifier Based on a Discriminative Gaussian Mixture Model. *arXiv preprint arXiv:1911.06028* .

Heusel, M.; Ramsauer, H.; Unterthiner, T.; Nessler, B.; and Hochreiter, S. 2017. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in neural information processing systems*, 6626–6637.

Kingma, D. P.; and Welling, M. 2013. Auto-encoding variational bayes. *International Conference on Learning Representations (ICLR)* .

Kumar, A.; Sattiger, P.; and Fletcher, T. 2017. Semi-supervised Learning with GANs: Manifold Invariance with Improved Inference. In *Advances in Neural Information Processing Systems, 2017*.

Larsen, A. B. L.; Sønderby, S. K.; Larochelle, H.; and Winther, O. 2016. Autoencoding beyond pixels using a learned similarity metric. In *Proceedings of The 33rd International Conference on Machine Learning*, 1558–1566. PMLR.

Li, C.; Liu, H.; Chen, C.; Pu, Y.; Chen, L.; Henao, R.; and Carin, L. 2017a. ALICE: Towards Understanding Adversarial Learning for Joint Distribution Matching. In *Advances in Neural Information Processing Systems, 2017*.

Li, C.; Xu, K.; Zhu, J.; and Zhang, B. 2017b. Triple Generative Adversarial Nets. In *Advances in neural information processing systems*, 5247–5256.

Lin, J. 1991. Divergence measures based on the Shannon entropy. *IEEE Transactions on Information Theory* 37(1): 145–151. ISSN 1557-9654. doi:10.1109/18.61115.

Liu, Z.; Luo, P.; Wang, X.; and Tang, X. 2015. Deep Learning Face Attributes in the Wild. In *Proceedings of International Conference on Computer Vision (ICCV)*.

Makhzani, A.; Shlens, J.; Jaitly, N.; and Goodfellow, I. J. 2016. Adversarial autoencoders. In *Proc. ICLR*.

Mescheder, L.; Nowozin, S.; and Geiger, A. 2017. Adversarial Variational Bayes: Unifying Variational Autoencoders and Generative Adversarial Networks. In *International Conference on Machine Learning (ICML)*.

Miyato, T.; Kataoka, T.; Koyama, M.; and Yoshida, Y. 2018. Spectral normalization for generative adversarial networks. *arXiv preprint arXiv:1802.05957* .

Netzer, Y.; Wang, T.; Coates, A.; Bissacco, A.; Wu, B.; and Ng, A. Y. 2011. Reading Digits in Natural Images with Unsupervised Feature Learning. In *NIPS Workshop on Deep Learning and Unsupervised Feature Learning 2011*. URL http://ufldl.stanford.edu/housenumbers/nips2011_housenumbers.pdf.

Pu, Y.; Dai, S.; Gan, Z.; Wang, W.; Wang, G.; Zhang, Y.; Henao, R.; and Carin, L. 2018. Jointgan: Multi-domain joint distribution learning with generative adversarial nets. *arXiv preprint arXiv:1806.02978* .

Srivastava, N. 2013. Improving neural networks with dropout. *University of Toronto* 182(566): 7.

Szegedy, C.; Ioffe, S.; Vanhoucke, V.; and Alemi, A. A. 2017. Inception-v4, inception-resnet and the impact of residual connections on learning. In *Thirty-first AAAI conference on artificial intelligence*.

Ulyanov, D.; Vedaldi, A.; and Lempitsky, V. 2018. It Takes (Only) Two: Adversarial Generator-Encoder Networks. In *AAAI Conference on Artificial Intelligence*.

Zhao, S.; Song, J.; and Ermon, S. 2017. Towards Deeper Understanding of Variational Autoencoding Models. *arXiv preprint arXiv:1702.08658, 2017* .

Zhao, S.; Song, J.; and Ermon, S. 2018. The information autoencoding family: A lagrangian perspective on latent variable generative models. *arXiv preprint arXiv:1806.06514* .