VARIATIONAL DISCRIMINATOR BOTTLENECK:
IMPROVING ImitATION LEARNING, INVERSE RL, AND GANs BY
CONSTRaining INFORMATION Flow

Xue Bin Peng & Angjoo Kanazawa & Sam Toyer & Pieter Abbeel & Sergey Levine
University of California, Berkeley
{xbpeng, kanazawa, sdt, pabbeel, svlevine}@berkeley.edu

ABSTRACT

Adversarial learning methods have been proposed for a wide range of applications, but the training of adversarial models can be notoriously unstable. Effectively balancing the performance of the generator and discriminator is critical, since a discriminator that achieves very high accuracy will produce relatively uninformative gradients. In this work, we propose a simple and general technique to constrain information flow in the discriminator by means of an information bottleneck. By enforcing a constraint on the mutual information between the observations and the discriminator’s internal representation, we can effectively modulate the discriminator’s accuracy and maintain useful and informative gradients. We demonstrate that our proposed variational discriminator bottleneck (VDB) leads to significant improvements across three distinct application areas for adversarial learning algorithms. Our primary evaluation studies the applicability of the VDB to imitation learning of dynamic continuous control skills, such as running. We show that our method can learn such skills directly from raw video demonstrations, substantially outperforming prior adversarial imitation learning methods. The VDB can also be combined with adversarial inverse reinforcement learning to learn parsimonious reward functions that can be transferred and re-optimized in new settings. Finally, we demonstrate that VDB can train GANs more effectively for image generation, improving upon a number of prior stabilization methods. (Video 1)

1 INTRODUCTION

Adversarial learning methods provide a promising approach to modeling distributions over high-dimensional data with complex internal correlation structures. These methods generally use a discriminator to supervise the training of a generator in order to produce samples that are indistinguishable from the data. A particular instantiation is generative adversarial networks, which can be used for high-fidelity generation of images (Goodfellow et al., 2014; Karras et al., 2017) and other high-dimensional data (Vondrick et al., 2016; Xie et al., 2018; Donahue et al., 2018). Adversarial methods can also be used to learn reward functions in the framework of inverse reinforcement learning (Finn et al., 2016; Fu et al., 2017), or to directly imitate demonstrations (Ho & Ermon, 2016). However, they suffer from major optimization challenges, one of which is balancing the performance of the generator and discriminator. A discriminator that achieves very high accuracy can produce relatively uninformative gradients, but a weak discriminator can also hamper the generator’s ability to learn. These challenges have led to widespread interest in a variety of stabilization methods for adversarial learning algorithms (Arjovsky et al., 2017; Kodali et al., 2017; Berthelot et al., 2017).

In this work, we propose a simple regularization technique for adversarial learning, which constrains the information flow from the inputs to the discriminator using a variational approximation to the information bottleneck. By enforcing a constraint on the mutual information between the input observations and the discriminator’s internal representation, we can encourage the discriminator to learn a representation that has heavy overlap between the data and the generator’s distribution, thereby effectively modulating the discriminator’s accuracy and maintaining useful and informative gradients.
Figure 1: Our method is general and can be applied to a broad range of adversarial learning tasks. **Left:** Motion imitation with adversarial imitation learning. **Middle:** Image generation. **Right:** Learning transferable reward functions through adversarial inverse reinforcement learning.

...
a reward function. One advantage of adversarial methods is that by leveraging a discriminator in place of a reward function, they can be applied to imitate skills where reward functions can be difficult to engineer. However, the performance of policies trained through adversarial methods still falls short of those produced by manually designed reward functions, when such reward functions are available [Rajeswaran et al., 2017][Peng et al., 2018]. We show that our method can significantly improve upon previous works that use adversarial techniques, and produces results of comparable quality to those from state-of-the-art approaches that utilize manually engineered reward functions.

Our variational discriminator bottleneck is based on the information bottleneck [Tishby & Slavsky, 2015], a technique for regularizing internal representations to minimize the mutual information with the input. Intuitively, a compressed representation can improve generalization by ignoring irrelevant distractors present in the original input. The information bottleneck can be instantiated in practical deep models by leveraging a variational bound and the reparameterization trick, inspired by a similar approach in variational autoencoders (VAE) [Kingma & Welling, 2013]. The resulting variational information bottleneck approximates this compression effect in deep networks [Alemi et al., 2016; Achille & Soatto, 2017]. A similar bottleneck has also been applied to learn disentangled representations [Higgins et al., 2017]. Building on the success of VAEs and GANs, a number of efforts have been made to combine the two. Makhzani et al. [2016] used adversarial discriminators during the training of VAEs to encourage the marginal distribution of the latent encoding to be similar to the prior distribution, similar techniques include [Mescheder et al., 2017] and Chen et al. [2018]. Conversely, Larsen et al. [2016] modeled the generator of a GAN using a VAE. Zhao et al. [2018] used an autoencoder instead of a VAE to model the discriminator, but does not enforce an information bottleneck on the encoding. While instance noise is widely used in modern architectures [Salimans et al., 2016; Sønderby et al., 2016; Arjovsky & Bottou, 2017], we show that explicitly enforcing an information bottleneck leads to improved performance over simply adding noise for a variety of applications.

3 Preliminaries

In this section, we provide a review of the variational information bottleneck proposed by [Alemi et al., 2016] in the context of supervised learning. Our variational discriminator bottleneck is based on the same principle, and can be instantiated in the context of GANs, inverse RL, and imitation learning. Given a dataset \( \{x_i, y_i\} \), with features \( x_i \) and labels \( y_i \), the standard maximum likelihood estimate \( \hat{q}(y|x) \) can be determined according to

\[
\min_q \mathbb{E}_{x,y \sim p(x,y)} [-\log q(y|x)].
\]

Unfortunately, this estimate is prone to overfitting, and the resulting model can often exploit idiosyncrasies in the data [Krizhevsky et al., 2012; Srivastava et al., 2014]. Alemi et al. [2016] proposed regularizing the model using an information bottleneck to encourage the model to focus only on the most discriminative features. The bottleneck can be incorporated by first introducing an encoder \( E(z|x) \) that maps the features \( x \) to a latent distribution over \( Z \), and then enforcing an upper bound \( I_c \) on the mutual information between the encoding and the original features \( I(X, Z) \). This results in the following regularized objective \( J(q, E) \)

\[
J(q, E) = \min_{q,E} \mathbb{E}_{x,y \sim p(x,y)} \mathbb{E}_{z \sim E(z|x)} [-\log q(y|z)]
\]

s.t.
\[
I(X, Z) \leq I_c. \tag{2}
\]

Note that the model \( q(y|z) \) now maps samples from the latent distribution \( z \) to the label \( y \). The mutual information is defined according to

\[
I(X, Z) = \int p(x, z) \log \frac{p(x,z)}{p(x)p(z)} \, dx \, dz = \int p(x) E(z|x) \log \frac{E(z|x)}{p(z)} \, dx \, dz, \tag{3}
\]

where \( p(x) \) is the distribution given by the dataset. Computing the marginal distribution \( p(z) = \int E(z|x) \, dx \) can be challenging. Instead, a variational lower bound can be obtained by using an approximation \( r(z) \) of the marginal. Since \( \text{KL}[p(z)||r(z)] \geq 0 \), \( \int p(z) \log p(z) \, dz \geq \int p(z) \log r(z) \, dz \), an upper bound on \( I(X, Z) \) can be obtained via the KL divergence,

\[
I(X, Z) \leq \int p(x) E(z|x) \log \frac{E(z|x)}{r(z)} \, dx \, dz = \mathbb{E}_{z \sim p(z)} \left[ \text{KL}[E(z|x)||r(z)] \right]. \tag{4}
\]
This provides an upper bound on the regularized objective \( \tilde{J}(q, E) \geq J(q, E) \),
\[
\tilde{J}(q, E) = \min_{q, E} \mathbb{E}_{x, y \sim p(x, y)} \left[ \mathbb{E}_{z \sim E(z|x)} \left[ -\log q(y|z) \right] \right] \\
\text{s.t. } \mathbb{E}_{x \sim p(x)} \left[ \text{KL} [E(z|x)||r(z)] \right] \leq I_c.
\]

To solve this problem, the constraint can be subsumed into the objective with a coefficient \( \beta \)
\[
\min_{q, E} \mathbb{E}_{x, y \sim p(x, y)} \left[ \mathbb{E}_{z \sim E(z|x)} \left[ -\log q(y|z) \right] \right] + \beta \left( \mathbb{E}_{x \sim p(x)} \left[ \text{KL} [E(z|x)||r(z)] \right] - I_c \right).
\]  

Alemi et al. (2016) evaluated the method on supervised learning tasks, and showed that models
trained with a VIB can be less prone to overfitting and more robust to adversarial examples.

## 4 Variational Discriminator Bottleneck

To outline our method, we first consider a standard GAN framework consisting of a discriminator
\( D \) and a generator \( G \), where the goal of the discriminator is to distinguish between samples from the
target distribution \( p^*(x) \) and samples from the generator \( G(x) \),
\[
\max_{G} \min_{D} \mathbb{E}_{x \sim p^*(x)} \left[ -\log (D(x)) \right] + \mathbb{E}_{x \sim G(x)} \left[ -\log (1 - D(x)) \right].
\]

We incorporate a variational information bottleneck by introducing an encoder \( E \) into the discriminator
that maps a sample \( x \) to a stochastic encoding \( z \sim E(z|x) \), and then apply a constraint \( I_c \) on the mutual information \( I(X, Z) \) between the original features and the encoding. \( D \) is then trained to classify samples drawn from the encoder distribution. A schematic illustration of the framework is available in Figure 2. The regularized objective \( J(D, E) \) for the discriminator is given by
\[
J(D, E) = \min_{D, E} \mathbb{E}_{x \sim p^*(x)} \left[ \mathbb{E}_{z \sim E(z|x)} \left[ -\log (D(z)) \right] + \mathbb{E}_{x \sim G(x)} \left[ -\log (1 - D(z)) \right] \right] \\
\text{s.t. } \mathbb{E}_{x \sim \tilde{p}(x)} \left[ \text{KL} [E(z|x)||r(z)] \right] \leq I_c,
\]

with \( \tilde{p} = \frac{1}{2}p^* + \frac{1}{2}G \) being a mixture of the target distribution and the generator. We refer to this regularizer as the variational discriminator bottleneck (VDB). To optimize this objective, we can introduce a Lagrange multiplier \( \beta \)
\[
J(D, E) = \min_{D, E} \max_{\beta \geq 0} \mathbb{E}_{x \sim p^*(x)} \left[ \mathbb{E}_{z \sim E(z|x)} \left[ -\log (D(z)) \right] + \mathbb{E}_{x \sim G(x)} \left[ -\log (1 - D(z)) \right] \right] \\
+ \beta \left( \mathbb{E}_{x \sim \tilde{p}(x)} \left[ \text{KL} [E(z|x)||r(z)] \right] - I_c \right).
\]

As we will discuss in Section 4.1 and demonstrate in our experiments, enforcing a specific mutual
information budget between \( x \) and \( z \) is critical for good performance. We therefore adaptively update
\( \beta \) via dual gradient descent to enforce a specific constraint \( I_c \) on the mutual information,
\[
D, E \leftarrow \arg\min_{D, E} \mathcal{L}(D, E, \beta) \\
\beta \leftarrow \max (0, \beta + \alpha \beta \left( \mathbb{E}_{x \sim \tilde{p}(x)} \left[ \text{KL} [E(z|x)||r(z)] \right] - I_c \right)),
\]

Figure 2: Left: Overview of the variational discriminator bottleneck. The encoder first maps samples \( x \) to a latent distribution \( E(z|x) \). The discriminator is then trained to classify samples \( z \) from the latent distribution. An information bottleneck \( I(X, Z) \leq I_c \) is applied to \( Z \). Right: Visualization of discriminators trained to differentiate two Gaussians with different KL bounds \( I_c \).
where $\mathcal{L}(D, E, \beta)$ is the Lagrangian

$$
\mathcal{L}(D, E, \beta) = \mathbb{E}_{x \sim p^*(x)} \left[ \mathbb{E}_{z \sim E(x|z)} [-\log (D(z))] \right] + \mathbb{E}_{x \sim G(x)} \left[ \mathbb{E}_{z \sim E(x|z)} [-\log (1 - D(z))] \right] + \beta \left( \mathbb{E}_{x \sim p(x)} [KL (E[z|x] || r(z))] - I_c \right),
$$

and $\alpha_\beta$ is the stepsize for the dual variable in dual gradient descent (Boyd & Vandenberghe 2004). In practice, we perform only one gradient step on $D$ and $E$, followed by an update to $\beta$. We refer to a GAN that incorporates a VDB as a variational generative adversarial network (VGAN).

In our experiments, the prior $r(x) = \mathcal{N}(0, I)$ is modeled with a standard Gaussian. The encoder $E(z|x) = \mathcal{N}(\mu_E(x), \Sigma_E(x))$ models a Gaussian distribution in the latent variables $Z$, with mean $\mu_E(x)$ and diagonal covariance matrix $\Sigma_E(x)$. When computing the KL loss, each batch of data contains an equal number of samples from $p^*(x)$ and $G(x)$. We use a simplified objective for the generator,

$$
\max_G \mathbb{E}_{x \sim G(x)} [-\log (1 - D(\mu_E(x)))].
$$

where the KL penalty is excluded from the generator’s objective. Instead of computing the expectation over $Z$, we found that approximating the expectation by evaluating $D$ at the mean $\mu_E(x)$ of the encoder’s distribution was sufficient for our tasks. The discriminator is modeled with a single linear unit followed by a sigmoid $D(z) = \sigma(w_D^Tz + b_D)$, with weights $w_D$ and bias $b_D$.

### 4.1 Discussion and Analysis

To interpret the effects of the VDB, we consider the results presented by [Arjovsky & Bottou 2017], which show that for two distributions with disjoint support, the optimal discriminator can perfectly classify all samples and its gradients will be zero almost everywhere. Thus, as the discriminator converges to the optimum, the gradients for the generator vanishes accordingly. To address this issue, [Arjovsky & Bottou 2017] proposed applying continuous noise to the discriminator inputs, thereby ensuring that the distributions have continuous support everywhere. In practice, if the original distributions are sufficiently distant from each other, the added noise will have negligible effects. As shown by [Mescheder et al. 2017], the optimal choice for the variance of the noise to ensure convergence can be quite delicate. In our method, by first using a learned encoder to map the inputs to an embedding and then applying an information bottleneck on the embedding, we can dynamically adjust the variance of the noise such that the distributions not only share support in the embedding space, but also have significant overlap. Since the minimum amount of information required for binary classification is 1 bit, by selecting an information constraint $I_c < 1$, the discriminator is prevented from perfectly differentiating between the distributions. To illustrate the effects of the VDB, we consider a simple task of training a discriminator to differentiate between two Gaussian distributions. Figure 2 visualizes the decision boundaries learned with different bounds $I_c$ on the mutual information. Without a VDB, the discriminator learns a sharp decision boundary, resulting in vanishing gradients for much of the space. But as $I_c$ decreases and the bound tightens, the decision boundary is smoothed, providing more informative gradients that can be leveraged by the generator.

Taking this analysis further, we can extend Theorem 3.2 from [Arjovsky & Bottou 2017] to analyze the VDB, and show that the gradient of the generator will be non-degenerate for a small enough constraint $I_c$, under some additional simplifying assumptions. The result in [Arjovsky & Bottou 2017] states that the gradient consists of vectors that point toward samples on the data manifold, multiplied by coefficients that depend on the noise. However, these coefficients may be arbitrarily small if the generated samples are far from real samples, and the noise is not large enough. This can still cause the generator gradient to vanish. In the case of the VDB, the constraint ensures that these coefficients are always bounded below. Due to space constraints, this result is presented in Appendix A.

### 4.2 VAIL: Variational Adversarial Imitation Learning

To extend the VDB to imitation learning, we start with the generative adversarial imitation learning (GAIL) framework [Ho & Ermon 2016], where the discriminator’s objective is to differentiate between the state distribution induced by a target policy $\pi^*(s)$ and the state distribution of the agent’s policy $\pi(s)$,

$$
\max_{\pi} \min_D \mathbb{E}_{a \sim \pi^*(s)} [-\log (D(s))] + \mathbb{E}_{a \sim \pi(s)} [-\log (1 - D(s))].
$$
Figure 3: Simulated humanoid performing various skills. VAIL is able to closely imitate a broad range of skills from mocap data.

The discriminator is trained to maximize the likelihood assigned to states from the target policy, while minimizing the likelihood assigned to states from the agent’s policy. The discriminator also serves as the reward function for the agent, which encourages the policy to visit states that, to the discriminator, appear indistinguishable from the demonstrations. Similar to the GAN framework, we can incorporate a VDB into the discriminator,

\[
J(D, E) = \min_{D, E} \max_{\beta \geq 0} \mathbb{E}_{s \sim \pi^*} \left[ \mathbb{E}_{z \sim E(z|s)} \left[ -\log \left( D(z) \right) \right] \right] + \mathbb{E}_{s \sim \pi} \left[ \mathbb{E}_{z \sim E(z|s)} \left[ -\log \left( 1 - D(z) \right) \right] \right] \\
+ \beta \left( \mathbb{E}_{s \sim \pi} \left[ KL \left( E(z|s) || r(z) \right) \right] - I_c \right),
\]

where \( \tilde{\pi} = \frac{1}{2}\pi + \frac{1}{2}\pi \) represents a mixture of the target policy and the agent’s policy. The reward for \( \pi \) is then specified by the discriminator \( r^* = -\log \left( 1 - D(\mu_E(s)) \right) \). We refer to this method as variational adversarial imitation learning (VAIL).

4.3 VAIL: Variational Adversarial Inverse Reinforcement Learning

The VDB can also be applied to adversarial inverse reinforcement learning (Fu et al., 2017) to yield a new algorithm which we call variational adversarial inverse reinforcement learning (VAIL). AIRL operates in a similar manner to GAIL, but with a discriminator of the form

\[
D(s, a, s') = \frac{\exp \left( f(s, a, s') \right)}{\exp \left( f(s, a, s') + \pi(a|s) \right)},
\]

where \( f(s, a, s') = g(s, a) + \gamma h(s') - h(s) \), with \( g \) and \( h \) being learned functions. Under certain restrictions on the environment, Fu et al. show that if \( g(s, a) \) is defined to depend only on the current state \( s \), the optimal \( g(s) \) recovers the expert’s true reward function \( r^*(s) \) up to a constant \( g^*(s) = r^*(s) + \text{const} \). In this case, the learned reward can be re-used to train policies in environments with different dynamics, and will yield the same policy as if the policy was trained under the expert’s true reward. In contrast, GAIL’s discriminator typically cannot be re-optimized in this way (Fu et al., 2017). In VAIL, we introduce stochastic encoders \( E_g(z_g|s), E_h(z_h|s) \), and \( g(z_g), h(z_h) \) are modified to be functions of the encoding. We can reformulate Equation 13 as

\[
D(s, a, z) = \frac{\exp \left( f(z, z_h, z'_h) \right)}{\exp \left( f(z, z_h, z'_h) + \pi(a|s) \right)},
\]

for \( z = (z_g, z_h, z'_h) \) and \( f(z_g, z_h, z'_h) = D_g(z_g) + \gamma D_h(z'_h) - D_h(z_h) \). We then obtain a modified objective of the form

\[
J(D, E) = \min_{D, E} \max_{\beta \geq 0} \mathbb{E}_{s, s' \sim \pi^*} \left[ \mathbb{E}_{z \sim E(z|s)} \left[ -\log \left( D(s, a, z) \right) \right] \right] \\
+ \mathbb{E}_{s, s' \sim \pi} \left[ \mathbb{E}_{z \sim E(z|s)} \left[ -\log \left( 1 - D(s, a, z) \right) \right] \right] \\
+ \beta \left( \mathbb{E}_{s, s' \sim \pi} \left[ KL \left( E(z|s) || r(z) \right) \right] - I_c \right),
\]

where \( \pi(s, s') \) denotes the joint distribution of successive states from a policy, and \( E(z|s, s') = E_g(z_g|s)E_h(z_h|s)E_h(z'_h|s') \).
The goal of the motion imitation tasks is to train a simulated character to mimic demonstrations provided in Appendix C. The encoding architecture we employ differs substantially from those used in prior work (Merel et al., 2017), details of which are available in Appendix C. The encoding $Z$ is $128D$ and an information constraint of $I_c = 0.5$ is applied for all skills, with a dual stepsize of $\alpha_{\beta} = 10^{-5}$. All policies are trained using PPO (Schulman et al., 2017).

The motions learned by the policies are best seen in the supplementary video. Snapshots of the character’s motions are shown in Figure 3. Each skill is learned from a single demonstration. VAIL is able to closely reproduce a variety of skills, including those that involve highly dynamics flips and complex contacts. We compare VAIL to a number of other techniques, including state-only GAIL (Ho & Ermon, 2016), GAIL with instance noise applied to the discriminator inputs (GAIL - noise), GAIL with instance noise applied to the last hidden layer (GAIL - noise z), and GAIL with a gradient penalty applied to the discriminator (GAIL - GP) (Mescheder et al., 2018). Since the VDB helps to prevent vanishing gradients, while GP mitigates exploding gradients, the two techniques can be seen as being complementary. Therefore, we also train a model that combines both VAIL and GP (VAIL - GP) (Mescheder et al., 2018). We found that the VDB helps to prevent vanishing gradients, while GP mitigates exploding gradients, the two techniques can be seen as being complementary. Therefore, we also train a model that combines both VAIL and GP (VAIL - GP) (Mescheder et al., 2018).

<table>
<thead>
<tr>
<th>Method</th>
<th>Backflip</th>
<th>Cartwheel</th>
<th>Dance</th>
<th>Run</th>
<th>Spinkick</th>
</tr>
</thead>
<tbody>
<tr>
<td>BC</td>
<td>3.01</td>
<td>2.88</td>
<td>2.93</td>
<td>2.63</td>
<td>2.88</td>
</tr>
<tr>
<td>Merel et al., 2017</td>
<td>1.33 ± 0.03</td>
<td>1.47 ± 0.12</td>
<td>2.61 ± 0.30</td>
<td>0.52 ± 0.04</td>
<td>1.82 ± 0.35</td>
</tr>
<tr>
<td>GAIL</td>
<td>0.74 ± 0.15</td>
<td>0.84 ± 0.05</td>
<td>1.31 ± 0.16</td>
<td>0.17 ± 0.03</td>
<td>1.07 ± 0.03</td>
</tr>
<tr>
<td>GAIL - noise</td>
<td>0.42 ± 0.02</td>
<td>0.92 ± 0.07</td>
<td>0.96 ± 0.08</td>
<td>0.21 ± 0.05</td>
<td>0.95 ± 0.14</td>
</tr>
<tr>
<td>GAIL - noise z</td>
<td>0.67 ± 0.12</td>
<td>0.72 ± 0.04</td>
<td>1.14 ± 0.08</td>
<td>0.14 ± 0.03</td>
<td>0.64 ± 0.09</td>
</tr>
<tr>
<td>GAIL - GP</td>
<td>0.62 ± 0.09</td>
<td>0.69 ± 0.05</td>
<td>0.80 ± 0.32</td>
<td>0.12 ± 0.02</td>
<td>0.64 ± 0.04</td>
</tr>
<tr>
<td>VAIL (ours)</td>
<td><strong>0.36 ± 0.13</strong></td>
<td><strong>0.40 ± 0.08</strong></td>
<td><strong>0.40 ± 0.21</strong></td>
<td><strong>0.13 ± 0.01</strong></td>
<td><strong>0.34 ± 0.05</strong></td>
</tr>
<tr>
<td>VAIL - GP (ours)</td>
<td>0.46 ± 0.17</td>
<td><strong>0.31 ± 0.02</strong></td>
<td><strong>0.15 ± 0.01</strong></td>
<td><strong>0.10 ± 0.01</strong></td>
<td><strong>0.31 ± 0.02</strong></td>
</tr>
</tbody>
</table>

Table 1: Average joint rotation error (radians) on humanoid motion imitation tasks. VAIL outperforms the other methods for all skills evaluated, except for policies trained using the manually-designed reward function from Peng et al., 2018.

5 EXPERIMENTS

We evaluate our method on adversarial learning problems in imitation learning, inverse reinforcement learning, and image generation. In the case of imitation learning, we show that the VDB enables agents to learn complex motion skills from a single demonstration, including visual demonstrations provided in the form of video clips. We also show that the VDB improves the performance of inverse RL methods. Inverse RL aims to reconstruct a reward function from a set of demonstrations, which can then be used to perform the task in new environments, in contrast to imitation learning, which aims to recover a policy directly. Our method is also not limited to control tasks, and we demonstrate its effectiveness for unconditional image generation.

5.1 VAIL: VARIATIONAL ADVERSARIAL IMITATION LEARNING

The goal of the motion imitation tasks is to train a simulated character to mimic demonstrations provided by mocap clips recorded from human actors. Each mocap clip provides a sequence of target states $\{s_0, s_1, ..., s_T\}$ that the character should track at each timestep. We use a similar experimental setup as Peng et al., 2018, with a 34 degrees-of-freedom humanoid character. We found that the discriminator architecture can greatly affect the performance on complex skills. The particular architecture we employ differs substantially from those used in prior work (Merel et al., 2017), details of which are available in Appendix C. The encoding $Z$ is $128D$ and an information constraint of $I_c = 0.5$ is applied for all skills, with a dual stepsize of $\alpha_{\beta} = 10^{-5}$. All policies are trained using PPO (Schulman et al., 2017).

The motions learned by the policies are best seen in the supplementary video. Snapshots of the character’s motions are shown in Figure 3. Each skill is learned from a single demonstration. VAIL is able to closely reproduce a variety of skills, including those that involve highly dynamics flips and complex contacts. We compare VAIL to a number of other techniques, including state-only GAIL (Ho & Ermon, 2016), GAIL with instance noise applied to the discriminator inputs (GAIL - noise), GAIL with instance noise applied to the last hidden layer (GAIL - noise z), and GAIL with a gradient penalty applied to the discriminator (GAIL - GP) (Mescheder et al., 2018). Since the VDB helps to prevent vanishing gradients, while GP mitigates exploding gradients, the two techniques can be seen as being complementary. Therefore, we also train a model that combines both VAIL and GP (VAIL - GP).
Figure 5: **Left:** Snapshots of the video demonstration and the simulated character trained with VAIL. The policy learns to run by directly imitating the video. **Right:** Saliency maps that visualize the magnitude of the discriminator’s gradient with respect to all channels of the RGB input images from both the demonstration and the simulation. Pixel values are normalized between [0, 1].

Figure 6: **Left:** Learning curves comparing policies for the video imitation task trained using a pixel-wise loss as the reward, GAIL, and VAIL. Only VAIL successfully learns to run from a video demonstration. **Middle:** Effect of training with fixed values of $\beta$ and adaptive $\beta$ ($I_c = 0.5$). **Right:** KL loss over the course of training with adaptive $\beta$. The dual gradient descent update for $\beta$ effectively enforces the VDB constraint $I_c$.

GP). Implementation details for combining the VDB and GP are available in Appendix B. Learning curves for the various methods are shown in Figure 10 and Table 1 summarizes the performance of the final policies. Performance is measured in terms of the average joint rotation error between the simulated character and the reference motion. We also include a reimplementation of the method described by Merel et al. (2017). For the purpose of our experiments, GAIL denotes policies trained using our particular architecture but without a VDB, and Merel et al. (2017) denotes policies trained using an architecture that closely mirror those from previous work. Furthermore, we include comparisons to policies trained using the handcrafted reward from Peng et al. (2018), as well as policies trained via behavioral cloning (BC). Since mocap data does not provide expert actions, we use the policies from Peng et al. (2018) as oracles to provide state-action demonstrations, which are then used to train the BC policies via supervised learning. Each BC policy is trained with 10k samples from the oracle policies, while all other policies are trained from just a single demonstration, the equivalent of approximately 100 samples.

VAIL consistently outperforms previous adversarial methods, and VAIL - GP achieves the best performance overall. Simply adding instance noise to the inputs (Salimans et al., 2016) or hidden layer without the KL constraint (Sønderby et al., 2016) leads to worse performance, since the network can learn a latent representation that renders the effects of the noise negligible. Though training with the handcrafted reward still outperforms the adversarial methods, VAIL demonstrates comparable performance to the handcrafted reward without manual reward or feature engineering, and produces motions that closely resemble the original demonstrations. The method from Merel et al. (2017) was able to imitate simple skills such as running, but was unable to reproduce more acrobatic skills such as the backflip and spinkick. In the case of running, our implementation produces more natural gaits than the results reported in Merel et al. (2017). Behavioral cloning is unable to reproduce any of the skills, despite being provided with substantially more demonstration data than the other methods.

**Video Imitation:** While our method achieves substantially better results on motion imitation when compared to prior work, previous methods can still produce reasonable behaviors. However, if the demonstrations are provided in terms of the raw pixels from video clips, instead of mocap data, the imitation task becomes substantially harder. The goal of the agent is therefore to directly im-
averse, the reward function recovered by V AIL can be re-optimized to train new policies to more smoothly imitate the skill depicted in the video. This is also a setting where manually engineering rewards is impractical, since simple losses like pixel distance do not provide a semantically meaningful measure of similarity. Figure 6 compares learning curves of policies trained with VAIL, GAIL, and policies trained using a reward function defined by the average pixel-wise difference between the frame \( M^*_t \) from the video demonstration and a rendered image \( M_t \) of the agent at each timestep \( t \),

\[
r_t = 1 - \frac{1}{\text{pixel}} \| M^*_t - M_t \|^2.
\]

Each frame is represented by a 64 × 64 RGB image.

Both GAIL and the pixel-loss are unable to learn the running gait. VAIL is the only method that successfully learns to imitate the skill from the video demonstration. Snapshots of the video demonstration and the simulated motion are available in Figure 5. To further investigate the effects of the VDB, we visualize the gradient of the discriminator with respect to images from the video demonstration and simulation. Saliency maps for discriminators trained with VAIL and GAIL are available in Figure 5. The VAIL discriminator learns to attend to spatially cohesive image patches around the character, while the GAIL discriminator exhibits less structure. The magnitude of the gradients from VAIL also tend to be significantly larger than those from GAIL, which may suggests that VAIL is able to mitigate the problem of vanishing gradients present in GAIL.

Adaptive Constraint: To evaluate the effects of the adaptive \( \beta \) updates, we compare policies trained with different fixed values of \( \beta \) and policies where \( \beta \) is updated adaptively to enforce a desired information constraint \( I_c = 0.5 \). Figure 6 illustrates the learning curves and the KL loss over the course of training. When \( \beta \) is too small, performance reverts to that achieved by GAIL. Large values of \( \beta \) help to smooth the discriminator landscape and improve learning speed during the early stages of training, but converges to a worse performance. Policies trained using dual gradient descent to adaptively update \( \beta \) consistently achieves the best performance overall.

5.2 VAIL: Variational Adversarial Inverse Reinforcement Learning

Next, we use VAIL to recover reward functions from demonstrations. Unlike the discriminator learned by VAIL, the reward function recovered by VAIL can be re-optimized to train new policies from scratch in the same environment. In some cases, it can also be used to transfer similar behaviour to different environments. In Figure 7 we show the results of applying VAIL to the C-maze from \cite{fu2017learning}, and a more complex S-maze; the simple 2D observation spaces of these tasks make it easy to interpret the recovered reward functions. In both mazes, the expert is trained to navigate from a start position at the bottom of the maze to a fixed target position at the top. We use each method to obtain an imitation policy and to approximate the expert’s reward on the original maze. The recovered reward is then used to train a new policy to solve a left–right flipped version of the training maze. On the C-maze, we found that plain AIRL—without a gradient penalty—would sometimes overfit and fail to transfer to the new environment, as evidenced by the reward visualization in Figure 7 (left) and the higher return variance in Figure 7 (right). In contrast, by incorporating a VDB into AIRL, VAIL learns a substantially smoother reward function that is more suitable for transfer. Furthermore, we found that in the S-maze with two internal walls, AIRL was too unstable to acquire a meaningful reward function. This was true even with the use of a gradient penalty. In contrast, VAIL was able to learn a reasonable reward in most cases without a
<table>
<thead>
<tr>
<th>Method</th>
<th>FID</th>
</tr>
</thead>
<tbody>
<tr>
<td>GAN</td>
<td>63.6</td>
</tr>
<tr>
<td>Inst Noise</td>
<td>30.7</td>
</tr>
<tr>
<td>SN</td>
<td>23.9</td>
</tr>
<tr>
<td>GP</td>
<td>22.6</td>
</tr>
<tr>
<td>WGAN-GP</td>
<td>19.9</td>
</tr>
<tr>
<td>VGAN (ours)</td>
<td>24.8</td>
</tr>
<tr>
<td>VGAN-SN (ours)</td>
<td>71.8</td>
</tr>
<tr>
<td>VGAN-GP (ours)</td>
<td><strong>18.1</strong></td>
</tr>
</tbody>
</table>

Figure 8: Comparison of VGAN and other methods on CIFAR-10, with performance evaluated using the Fréchet Inception Distance (FID).

Figure 9: VGAN samples on CIFAR-10, CelebA 128×128, and CelebAHQ 1024×1024.

5.3 VGAN: Variational Generative Adversarial Networks

Finally, we apply the VDB to image generation with generative adversarial networks, which we refer to as VGAN. Experiment are conducted on CIFAR-10 (Krizhevsky et al., 2015), CelebA (Liu et al., 2015), and CelebAHQ (Karras et al., 2018) datasets. We compare our approach to recent stabilization techniques: WGAN-GP (Gulrajani et al., 2017b), instance noise (Sønderby et al., 2016; Arjovsky & Bottou, 2017), spectral normalization (SN) (Miyato et al., 2018), and gradient penalty (GP) (Mescheder et al., 2018), as well as the original GAN (Goodfellow et al., 2014) on CIFAR-10. To measure performance, we report the Fréchet Inception Distance (FID) (Heusel et al., 2017), which has been shown to be more consistent with human evaluation. All methods are implemented using the same base model, built on the resnet architecture of Mescheder et al. (2018). Aside from tuning the KL constraint $I_c$ for VGAN, no additional hyperparameter optimization was performed to modify the settings provided by Mescheder et al. (2018). The performance of the various methods on CIFAR-10 are shown in Figure 8. While vanilla GAN and instance noise are prone to diverging as training progresses, VGAN remains stable. Note that instance noise can be seen as a non-adaptive version of VGAN without constraints on $I_c$. This experiment again highlights that there is a significant improvement from imposing the information bottleneck over simply adding instance noise. Combining both VDB and gradient penalty (VGAN - GP) achieves the best performance overall with an FID of 18.1. We also experimented with combining the VDB with SN, but this combination is prone to diverging. See Figure 9 for samples of images generated with our approach. Please refer to Appendix B for experimental details and more results.

6 CONCLUSION

We present the variational discriminator bottleneck, a general regularization technique for adversarial learning. Our experiments show that the VDB is broadly applicable to a variety of domains, and yields significant improvements over previous techniques on a number of challenging tasks. While our experiments have produced promising results for video imitation, the results have been primarily with videos of synthetic scenes. We believe that extending the technique to imitating real-world videos is an exciting direction. Another exciting direction for future work is a more in-depth theoretical analysis of the method, to derive convergence and stability results or conditions.
ACKNOWLEDGEMENTS

We would like to thank the anonymous reviewers for their helpful feedback, and AWS and NVIDIA for providing computational resources. This research was funded by an NSERC Postgraduate Scholarship, a Berkeley Fellowship for Graduate Study, BAIR, Huawei, and ONR PECASE N000141612723.

REFERENCES


SUPPLEMENTARY MATERIAL

A ANALYSIS AND PROOFS

In this appendix, we show that the gradient of the generator when the discriminator is augmented with the VDB is non-degenerate, under some mild additional assumptions. First, we assume a pointwise constraint of the form $\text{KL}[E(z|x)||r(z)] \leq I_c$ for all $x$. In reality, we use an average KL constraint, since we found it to be more convenient to optimize, though a pointwise constraint is also possible to enforce by using the largest constraint violation to increment $\beta$. We could likely also extend the analysis to the average constraint, though we leave this to future work. The main theorem can then be stated as follows:

**Theorem A.1.** Let $g(u)$ denote the generator’s mapping from a noise vector $u \sim p(u)$ to a point in $X$. Given the generator distribution $G(x)$ and data distribution $p^*(x)$, a VDB with an encoder $E(z|x) = \mathcal{N}(\mu_E(x), \Sigma)$, and KL$[E(z|x)||r(z)] \leq I_c$, the gradient passed to the generator has the form

$$
\nabla_g \mathbb{E}_{u \sim p(u)} [\log (1 - D^* (\mu_E (g(u))))] = \mathbb{E}_{u \sim p(u)} \left[ a(u) \int E(\mu_E (g(u))\nabla_g ||\mu_E (g(u)) - \mu_E (x)||^2 dp^*(x) - b(u) \int E(\mu_E (g(u))\nabla_g ||\mu_E (g(u)) - \mu_E (x)||^2 dG(x) \right]
$$

where $D^*(x)$ is the optimal discriminator, $a(x)$ and $b(x)$ are positive functions, and we always have $E(\mu_E (g(u))\nabla_g ||\mu_E (g(u)) - \mu_E (x)||^2 dp^*(x) > C(I_c)$, where $C(I_c)$ is a continuous monotonic function, and $C(I_c) \to \delta > 0$ as $I_c \to 0$.

Analysis for an encoder with an input-dependent variance $\Sigma(x)$ is also possible, but more involved. We’ll further assume below for notational simplicity that $\Sigma$ is diagonal with diagonal values $\sigma^2$. This assumption is not required, but substantially simplifies the linear algebra. Analogously to Theorem 3.2 from [Arjovsky & Bottou, 2017], this theorem states that the gradient of the generator points in the direction of points in the data distribution, and away from points in the generator distribution. However, going beyond the theorem in [Arjovsky & Bottou, 2017], this result states that the coefficients on these vectors, given by $E(\mu_E (g(u))\nabla_g ||\mu_E (g(u)) - \mu_E (x)||^2 dp^*(x)$, are always bounded below by a value that approaches a positive constant $\delta$ as we decrease $I_c$, meaning that the gradient does not vanish. The proof of the first part of this theorem is essentially identical to the proof presented by [Arjovsky & Bottou, 2017], but accounting for the fact that the noise is now injected into the latent space of the VDB, rather than being added directly to $x$. This result assumes that $E(z|x)$ has a learned but input-dependent variance $\Sigma = \sigma^2 I$, though the proof can be repeated for an input-dependent or non-diagonal $\Sigma$.

**Proof.** Overloading $p^*(x)$ and $G(x)$, let $p^*(z)$ and $G(z)$ be the distribution of embeddings $z$ under the real data and generator respectively. $p^*(z)$ is then given by

$$
p^*(z) = \mathbb{E}_{x \sim p^*(x)} [E(z|x)] = \int E(z|x) dp^*(x),
$$

and similarly for $G(z)$

$$
G(z) = \mathbb{E}_{x \sim G(x)} [E(z|x)] = \int E(z|x) dG(x),
$$

From [Arjovsky & Bottou, 2017], the optimal discriminator between $p^*(z)$ and $G(z)$ is

$$
D^*(z) = \frac{p^*(z)}{p^*(z) + G(z)}
$$
The gradient passed to the generator then has the form
\[
\nabla_g \mathbb{E}_{u \sim p(u)} [\log (1 - D^*(\mu_E(g(u)))] \\
= \mathbb{E}_{u \sim p(u)} \left[ \nabla_g \log (G(\mu_E(g(u)))) - \nabla_g \log (p^*(\mu_E(g(u))) + G(\mu_E(g(u)))) \right] \\
= \mathbb{E}_{u \sim p(u)} \left[ \nabla_g G(\mu_E(g(u))) - \nabla_g p^*(\mu_E(g(u))) + \nabla_g G(\mu_E(g(u))) \right] \\
= \mathbb{E}_{u \sim p(u)} \left[ \frac{1}{p^*(\mu_E(g(u))) + G(\mu_E(g(u)))} \nabla_g [-p^*(\mu_E(g(u)))] \\
- \frac{p^*(\mu_E(g(u)))}{G(\mu_E(g(u)))} \nabla_g [-G(\mu_E(g(u)))] \right].
\]

Let
\[
a(u) = \frac{1}{2\sigma^2} \frac{1}{p^*(\mu_E(g(u))) + G(\mu_E(g(u)))}, \\
b(u) = \frac{1}{2\sigma^2} \frac{p^*(\mu_E(g(u)))}{G(\mu_E(g(u)))}.
\]
We then have
\[
\nabla_g \mathbb{E}_{u \sim p(u)} [\log (1 - D^*(\mu_E(g(u)))] \\
= \mathbb{E}_{u \sim p(u)} \left[ 2\sigma^2 a(u) \nabla_g [-p^*(\mu_E(g(u)))] - 2\sigma^2 b(u) \nabla_g [-G(\mu_E(g(u)))] \right] \\
= \mathbb{E}_{u \sim p(u)} \left[ 2\sigma^2 a(u) \int -\nabla_g \mu_E(g(u)) \|x\| dp^*(x) \\
- 2\sigma^2 b(u) \int -\nabla_g \mu_E(g(u)) \|x\| dG(x) \right] \\
= \mathbb{E}_{u \sim p(u)} \left[ 2\sigma^2 a(u) \int -\nabla_g \frac{1}{Z} \exp \left(-\frac{1}{2\sigma^2} \|\mu_E(g(u)) - \mu_E(x)\|^2 \right) \|x\| dp^*(x) \\
- 2\sigma^2 b(u) \int -\nabla_g \frac{1}{Z} \exp \left(-\frac{1}{2\sigma^2} \|\mu_E(g(u)) - \mu_E(x)\|^2 \right) \|x\| dG(x) \right] \\
= \mathbb{E}_{u \sim p(u)} \left[ a(u) \int \frac{1}{Z} \exp \left(-\frac{1}{2\sigma^2} \|\mu_E(g(u)) - \mu_E(x)\|^2 \right) \nabla_g \|\mu_E(g(u)) - \mu_E(x)\|^2 dp^*(x) \\
- b(u) \int \frac{1}{Z} \exp \left(-\frac{1}{2\sigma^2} \|\mu_E(g(u)) - \mu_E(x)\|^2 \right) \nabla_g \|\mu_E(g(u)) - \mu_E(x)\|^2 dG(x) \right] \\
= \mathbb{E}_{u \sim p(u)} \left[ a(u) \int E(\mu_E(g(u)) \|x\) \nabla_g \|\mu_E(g(u)) - \mu_E(x)\|^2 dp^*(x) \\
- b(u) \int E(\mu_E(g(u)) \|x\) \nabla_g \|\mu_E(g(u)) - \mu_E(x)\|^2 dG(x) \right].
\]
From the definition of the KL-divergence we can bound the length of all embedding vectors,

\[ \text{KL} \left[ E(z|x)||r(z) \right] = \frac{1}{2} \left( K\sigma^2 + \mu_E(x)^T \mu_E(x) - K - K\log \sigma^2 \right) \leq I_c \]

and similarly for \( ||\mu_E(g(u))||^2 \), with \( K \) denoting the dimension of \( Z \). A lower bound on \( E(\mu_E(g(u)|\mathbf{x})|\mathbf{x}) \), where \( u \sim p(u) \) and \( x \sim p^*(x) \), can then be determined by

\[ \log (E(\mu_E(g(u)|\mathbf{x})|\mathbf{x})) = -\frac{1}{2\sigma^2} (\mu_E(g(u)) - \mu_E(x))^T (\mu_E(g(u)) - \mu_E(x)) - \frac{K}{2} \log \sigma^2 - \frac{K}{2} \log 2\pi \]

Since \( ||\mu_E(x)||^2, ||\mu_E(g(u))||^2 \leq 2I_c - K\sigma^2 + K + K\log \sigma^2 \),

\[ ||\mu_E(g(u)) - \mu_E(x)||^2 \leq 8I_c - 4K\sigma^2 + 4K + 4K\log \sigma^2, \]

and it follows that

\[ -\frac{1}{2\sigma^2} (\mu_E(g(u)) - \mu_E(x))^T (\mu_E(g(u)) - \mu_E(x)) \geq -4\sigma^{-2}I_c + 2K - 2K\sigma^{-2} - 2K\sigma^{-2} \log \sigma^{-2}. \]

The likelihood is therefore bounded below by

\[ \log (E(\mu_E(g(u)|\mathbf{x})|\mathbf{x})) \geq -4\sigma^{-2}I_c + 2K - 2K\sigma^{-2} - 2K\sigma^{-2} \log \sigma^{-2} - \frac{K}{2} \log \sigma^2 - \frac{K}{2} \log 2\pi \]

Since \(-\sigma^{-2} - \sigma^{-2} \log \sigma^{-2} \geq -1, \)

\[ \log (E(\mu_E(g(u)|\mathbf{x})|\mathbf{x})) \geq -4\sigma^{-2}I_c - \frac{K}{2} \log \sigma^2 - \frac{K}{2} \log 2\pi \]

(14)

From the KL constraint, we can derive a lower bound \( \ell(I_c) \) and an upper bound \( \mathcal{U}(I_c) \) on \( \sigma^2 \).

\[ \frac{1}{2} \left( K\sigma^2 + \mu_E(x)^T \mu_E(x) - K - K\log \sigma^2 \right) \leq I_c \]

\[ \sigma^2 - 1 - \log \sigma^2 \leq \frac{2I_c}{K} \]

\[ \log \sigma^2 \geq -\frac{2I_c}{K} - 1 \]

\[ \sigma^2 \geq \exp \left( -\frac{2I_c}{K} - 1 \right) = \ell(I_c) \]

For the upper bound, since \( \sigma^2 - \log \sigma^2 > \frac{1}{2} \sigma^2, \)

\[ \sigma^2 - 1 - \log \sigma^2 \leq \frac{2I_c}{K} \]

\[ \frac{1}{2} \sigma^2 - 1 < \frac{2I_c}{K} \]

\[ \sigma^2 < 4\frac{I_c}{K} + 2 = \mathcal{U}(I_c) \]

Substituting \( \ell(I_c) \) and \( \mathcal{U}(I_c) \) into Equation [14] we arrive at the following lower bound

\[ E(\mu_E(g(u)|x)) > \exp \left( -4I_c \exp \left( \frac{2I_c}{K} + 1 \right) - \frac{K}{2} \log \left( 4\frac{I_c}{K} + 2 \right) - \frac{K}{2} \log 2\pi \right) = C(I_c). \]
B Gradient Penalty

To combine VDB with gradient penalty, we use the reparameterization trick to backprop through the encoder when computing the gradient of the discriminator with respect to the inputs.

\[
J(D, E) = \min_{D, E} \mathbb{E}_{x \sim p^r(x)} \left[ \mathbb{E}_{z \sim E(z|x)} [-\log (D(z))] \right] + \mathbb{E}_{x \sim G(x)} \left[ \mathbb{E}_{z \sim E(z|x)} [-\log (1 - D(z))] \right] + w_{GP} \mathbb{E}_{x \sim p^r(x)} \left[ \mathbb{E}_{\epsilon \sim \mathcal{N}(0, I)} \left[ \frac{1}{2} \left\| \nabla_x D(\mu_E(x) + \Sigma_E(x)\epsilon) \right\|^2 \right] \right]
\]

s.t. \( \mathbb{E}_{x \sim p^r(x)} \left[ \text{KL} \left[ E(z|x) || r(z) \right] \right] \leq I_c, \)

(15)

The coefficient \( w_{GP} \) weights the gradient penalty in the objective, \( w_{GP} = 10 \) for the image generation, \( w_{GP} = 1 \) for motion imitation, and \( w_{GP} = 0.1 \) (C-maze) or \( w_{GP} = 0.01 \) (S-maze) for the IRL tasks. The gradient penalty is applied only to real samples \( p^r(x) \). We have experimented with the penalty to both real and fake samples, but found that performance was worse than penalizing only gradients from real samples. This is consistent with the GP implementation from [Mescheder et al. 2018].

C Imitation Learning

Experimental Setup: The goal of the motion imitation tasks is to train a simulated agent to mimic a demonstration provided in the form of a mocap clip recorded from a human actor. We use a similar experimental setup as Peng et al. (2018), with a 34 degrees-of-freedom humanoid character. The experimental setup consists of features that represent the configuration of the character’s body (link positions and velocities). We also include a phase variable \( \phi \in [0, 1] \) among the state features, which records the character’s progress along the motion and helps to synchronize the character with the reference motion. With 0 and 1 denoting the start and end of the motion respectively. The action \( a \) sampled from the policy \( \pi(a|s) \) specifies target poses for PD controller positioned at each joint. Given a state, the policy specifies a Gaussian distribution over the action space \( \pi(a|s) = \mathcal{N}(\mu(s), \Sigma) \), with a state-dependent mean \( \mu(s) \) and fixed diagonal covariance matrix \( \Sigma \). \( \mu(s) \) is modeled using a 3-layered fully-connected network with 1024 and 512 hidden units, followed by a linear output layer that specifies the mean of the Gaussian. ReLU activations are used for all hidden layers. The value function is modeled with a similar architecture but with a single linear output unit. The policy is queried at 30Hz. Physics simulation is performed at 1.2kHz using the Bullet physics engine [Bullet 2015].

Given the rewards from the discriminator, PPO (Schulman et al. 2017) is used to train the policy, with a stepsize of \( 2.5 \times 10^{-6} \) for the policy, a stepsize of 0.01 for the value function, and a stepsize of \( 10^{-5} \) for the discriminator. Gradient descent with momentum 0.9 is used for all models. The PPO clipping threshold is set to 0.2. When evaluating the performance of the policies, each episode is simulated for a maximum horizon of 20s. Early termination is triggered whenever the character’s torso contacts the ground, leaving the policy a maximum error of \( \pi \) radians for all remaining timesteps.

Phase-Functioned Discriminator: Unlike the policy and value function, which are modeled with standard fully-connected networks, the discriminator is modeled by a phase-functioned neural network (PFNN) to explicitly model the time-dependency of the reference motion (Holden et al. 2017). While the parameters of a network are generally fixed, the parameters of a PFNN are functions of the phase variable \( \phi \). The parameters \( \theta \) of the network for a given \( \phi \) is determined by a weighted combination of a set of fixed parameters \( \left\{ \theta_0, \theta_1, \ldots, \theta_k \right\} \),

\[
\theta = \sum_{i=0}^{k} w_i(\phi) \theta_i,
\]

where \( w_i(\phi) \) is a phase-dependent weight for \( \theta_i \). In our implementation, we use \( k = 5 \) sets of parameters and \( w_i(\phi) \) is designed to linearly interpolate between two adjacent sets of parameters for each phase \( \phi \), where each set of parameters \( \theta_i \) corresponds to a discrete phase value \( \phi_i \) spaced
Figure 10: Learning curves comparing VAIL to other methods for motion imitation. Performance is measured using the average joint rotation error between the simulated character and the reference motion. Each method is evaluated with 3 random seeds.

Figure 11: Learning curves comparing VAIL with a discriminator modeled by a phase-functioned neural network (PFNN), to modeling the discriminator with a fully-connected network that receives the phase-variable \( \phi \) as part of the input (no PFNN), and a discriminator modeled with a fully-connected network but does not receive \( \phi \) as an input (no phase).

uniformly between \([0, 1]\). For a given value of \( \phi \), the parameters of the discriminator are determined according to

\[
\theta = w_i(\phi)\theta_i + w_{i+1}(\phi)\theta_{i+1}
\]

where \( \theta_i \) and \( \theta_{i+1} \) correspond to the phase values \( \phi_i \leq \phi < \phi_{i+1} \) that form the endpoints of the phase interval that contains \( \phi \). A PFNN is used for all motion imitation experiments, both VAIL and GAIL, except for those that use the approach proposed by Merel et al. (2017), which use standard fully-connected networks for the discriminator. Figure [11] compares the performance of VAIL when the discriminator is modeled with a phase-functioned neural network (with PFNN) to discriminators modeled with standard fully-connected networks. We increased the size of the layers of the fully-connected nets to have a similar number of parameters as a PFNN. We evaluate the performance of fully-connected nets that receive the phase variable \( \phi \) as part of the input (no PFNN), and fully-connected nets that do not receive \( \phi \) as an input. The phase-functioned discriminator leads to significant performance improvements across all tasks evaluated. Policies trained without a phase variable performs worst overall, suggesting that phase information is critical for performance. All methods perform well on simpler skills, such as running, but the additional phase structure introduced by the PFNN proved to be vital for successful imitation of more complex skills, such as the dance and backflip.

Next we compare the accuracy of discriminators trained using different methods. Figure [12] illustrates accuracy of the discriminators over the course of training. Discriminators trained via GAIL quickly overpowers the policy, and learns to accurately differentiate between samples, even when
instance noise is applied to the inputs. VAIL without the KL constraint slows the discriminator’s progress, but nonetheless reaches near perfect accuracy with a larger number of samples. Once the KL constraint is enforced, the information bottleneck constrains the performance of the discriminator, converging to approximately 80% accuracy. Figure 12 also visualizes the value of $\beta$ over the course of training for motion imitation tasks, along with the loss of the KL term in the objective. The dual gradient descent update effectively enforces the VDB constraint $I_c$. 

Video Imitation: In the video imitation tasks, we use a simplified 2D biped character in order to avoid issues that may arise due to depth ambiguity from monocular videos. The biped character has a total of 12 degrees-of-freedom, with similar state and action parameters as the humanoid. The video demonstrations are generated by rendering a reference motion into a sequence of video frames, which are then provided to the agent as a demonstration. The goal of the agent is to imitate the motion depicted in the video, without access to the original reference motion, and the reference motion is used only to evaluate performance. 

D INVERSE REINFORCEMENT LEARNING

D.1 EXPERIMENTAL SETUP

Environments We evaluate on two maze tasks, as illustrated in Figure 13. The C-maze is taken from [Fu et al. (2017)]. In this maze, the agent starts at a random point within a small fixed distance of the mean start position. The agent has a continuous, 2D action space which allows it to accelerate in the $x$ or $y$ directions, and is able to observe its $x$ and $y$ position, but not its velocity. The ground truth reward is $r_t = -d_t - 10^{-3} \|a_t\|^2$, where $d_t$ is the agent’s distance to the goal, and $a_t$ is its action (this action penalty is assumed to be zero in Figure 13). Episodes terminate after 100 steps; for evaluation, we report the undiscounted mean sum of rewards over each episode. The S-maze is a larger variant of the same environment with an extra wall between the agent and its goal. To make the S-maze easier to solve for the expert, we increased the maximum control forces relative to the C-maze to enable more rapid exploration. Environments will be released along with the rest of our VAIRL implementations.

Hyperparameters Policy networks for all methods were two-layer ReLU MLPs with 32 hidden units per layer. Reward and discriminator networks were similar, but with 32-unit mean and standard deviation layers inserted before the final layer for VDB methods. To generate expert demonstrations, we trained a TRPO [Schulman et al. (2015)] agent on the ground truth reward for the training environment for 200 iterations, and saved 107 trajectories from each of the policies corresponding to the five final iterations. TRPO used a batch size of 10,000, a step size of 0.01, and entropy bonus with a coefficient of 0.1 to increase diversity. After generating demonstrations, we trained the IRL and imitation methods on a training maze for 200 iterations; again, our policy optimizer was TRPO with the same hyperparameters used to generate demonstrations. Between each policy update, we did 100 discriminator updates using Adam with a learning rate of $5 \times 10^{-5}$ and batch size of 32. For the C-maze our VAIRL runs used a target KL of $I_C = 0.5$, while for the more complex S-maze we use a tighter target of $I_C = 0.05$. For the test C-maze, we trained new policies against the recovered
D.2 RECOVERED REWARD FUNCTIONS

Figure 14 and 15 show the reward functions recovered by each IRL baseline on the C-maze and S-maze, respectively, along with sample trajectories for policies trained to optimize those rewards. Notice that VAIRL tends to recover smoother reward functions that match the ground truth reward more closely than the baselines. Addition of a gradient penalty enhances this effect for both AIRL and VAIRL. This is especially true in S-maze, where combining a gradient penalty with a variational discriminator bottleneck leads to a smooth reward that gradually increases as the agent nears its goal position at the top of the maze.

E IMAGE GENERATION

We provide further experiment on image generation and details of the experimental setup.

E.1 EXPERIMENTAL SETUP:

We use the non-saturating objective of Goodfellow et al. (2014) for all models except WGAN-GP. Following Lucic et al. (2017), we compute FID on samples of size 10000. We base our implementation on Mescheder et al. (2018), where we do not use any batch normalization for both the generator and the discriminator. We use RMSprop (Hinton et al.) and a fixed learning rate for all experiments.

For convolutional GAN, variational discriminative bottleneck is implemented as a 1x1 convolution on the final embedding space that outputs a Gaussian distribution over $Z$ parametrized with a mean and a diagonal covariance matrix. For all image experiments, we preserve the dimensionality of the latent space. All experiments use adaptive $\beta$ update with a dual stepsize of $\alpha_\beta = 10^{-5}$. We will make our code public. Similarly to VGAN, instance noise Sønderby et al. (2016); Arjovsky & Bottou (2017) is added to the final embedding space of the discriminator right before applying the classifier. Instance noise can be interpreted as a non-adaptive VGAN without a information constraint.

Architecture: For CIFAR-10, we use a resnet-based architecture adapted from Mescheder et al., 2018 detailed in Tables 2, 3, and 4. For CelebA and CelebAHQ, we use the same architecture used in Mescheder et al., 2018.

E.2 RESULTS

CIFAR-10: We compare our approach with recent stabilization techniques: WGAN-GP Gulrajani et al. (2017b), instance noise Sønderby et al. (2016), Arjovsky & Bottou (2017), spectral nor-
Figure 14: Visualizations of recovered reward functions transferred to the mirrored C-maze. Also shown are trajectories executed by policies trained to maximize the corresponding reward in the new environment.
Figure 15: Visualizations of recovered reward functions transferred to the mirrored S-maze, like Figure 14.
Published as a conference paper at ICLR 2019

<table>
<thead>
<tr>
<th>Layer</th>
<th>Output size</th>
<th>Filter</th>
</tr>
</thead>
<tbody>
<tr>
<td>FC</td>
<td>256·4·4</td>
<td>256 → 256·4·4</td>
</tr>
<tr>
<td>Reshape</td>
<td>256 × 4 × 4</td>
<td></td>
</tr>
<tr>
<td>Resnet-block</td>
<td>128 × 4 × 4</td>
<td>128 → 256</td>
</tr>
<tr>
<td>Upsample</td>
<td>128 × 8 × 8</td>
<td></td>
</tr>
<tr>
<td>Resnet-block</td>
<td>64 × 8 × 8</td>
<td>64 → 128</td>
</tr>
<tr>
<td>Upsample</td>
<td>64 × 16 × 16</td>
<td></td>
</tr>
<tr>
<td>Resnet-block</td>
<td>32 × 16 × 16</td>
<td>32 → 64</td>
</tr>
<tr>
<td>Upsample</td>
<td>32 × 32 × 32</td>
<td></td>
</tr>
<tr>
<td>Resnet-block</td>
<td>32 × 32 × 32</td>
<td>32 → 32</td>
</tr>
<tr>
<td>Conv2D</td>
<td>3 × 32 × 32</td>
<td></td>
</tr>
</tbody>
</table>

Table 2: CIFAR-10 Generator

<table>
<thead>
<tr>
<th>Layer</th>
<th>Output size</th>
<th>Filter</th>
</tr>
</thead>
<tbody>
<tr>
<td>Conv2D</td>
<td>32 × 32 × 32</td>
<td>3 → 32</td>
</tr>
<tr>
<td>Resnet-block</td>
<td>64 × 32 × 32</td>
<td>32 → 64</td>
</tr>
<tr>
<td>AvgPool2D</td>
<td>64 × 16 × 16</td>
<td></td>
</tr>
<tr>
<td>Resnet-block</td>
<td>128 × 16 × 16</td>
<td>128 → 256</td>
</tr>
<tr>
<td>AvgPool2D</td>
<td>128 × 8 × 8</td>
<td></td>
</tr>
<tr>
<td>Resnet-block</td>
<td>256 × 8 × 8</td>
<td>256 → 256</td>
</tr>
<tr>
<td>AvgPool2D</td>
<td>256 × 4 × 4</td>
<td></td>
</tr>
<tr>
<td>Conv2D</td>
<td>1</td>
<td>256 · 4 · 4 → 1</td>
</tr>
</tbody>
</table>

Table 3: CIFAR-10 Discriminator

<table>
<thead>
<tr>
<th>Layer</th>
<th>Output size</th>
<th>Filter</th>
</tr>
</thead>
<tbody>
<tr>
<td>Conv2D</td>
<td>32 × 32 × 32</td>
<td>3 → 32</td>
</tr>
<tr>
<td>Resnet-block</td>
<td>64 × 32 × 32</td>
<td>32 → 64</td>
</tr>
<tr>
<td>AvgPool2D</td>
<td>64 × 16 × 16</td>
<td></td>
</tr>
<tr>
<td>Resnet-block</td>
<td>128 × 16 × 16</td>
<td>128 → 256</td>
</tr>
<tr>
<td>AvgPool2D</td>
<td>128 × 8 × 8</td>
<td></td>
</tr>
<tr>
<td>Resnet-block</td>
<td>256 × 8 × 8</td>
<td>256 → 256</td>
</tr>
<tr>
<td>AvgPool2D</td>
<td>256 × 4 × 4</td>
<td></td>
</tr>
<tr>
<td>1 × 1 Conv2D</td>
<td>256 × 4 × 4</td>
<td>256 → 256</td>
</tr>
<tr>
<td>Sampling</td>
<td>256 × 4 × 4</td>
<td></td>
</tr>
<tr>
<td>FC</td>
<td>1</td>
<td>256 · 4 · 4 → 1</td>
</tr>
</tbody>
</table>

Table 4: CIFAR-10 Discriminator with VDB

CelebA: On the CelebA (Liu et al., 2015) dataset, we generate images of size 128 × 128 with $I_c = 0.2$. On this dataset we do not see a big improvement upon the other baselines. This is likely because the architecture has been effectively tuned for this task, reflected by the fact that even the vanilla GAN trains fine on this dataset. All GAN, GP, and VGAN-GP obtain a similar FID scores of 7.64, 7.76, 7.25 respectively. See Figure 16 for more visual comparisons of randomly generated samples.

CelebAHQ: VGAN can also be trained on CelebAHQ Karras et al. (2018) at 1024 by 1024 resolution directly, without progressive growing (Karras et al., 2018). We use $I_c = 0.1$ and train with VGAN-GP. We train on a single Tesla V100, which fits a batch size of 8 in our experiments. Previous approaches (Karras et al., 2018; Mescheder et al., 2018) use a larger batch size and train over multiple GPUs. The model was trained for 300k iterations.
Figure 16: Random results on CIFAR-10 (Krizhevsky et al.): GAN (Goodfellow et al., 2014) FID: 63.6, instance noise (Sønderby et al., 2016; Arjovsky & Bottou, 2017) FID: 30.7, spectral normalization (SN) (Miyato et al., 2018) FID: 23.9, gradient penalty (GP) (Mescheder et al., 2018) FID: 22.6, WGAN-GP (Gulrajani et al., 2017b) FID: 19.9, and the proposed VGAN-GP FID: 18.1. The samples produced by VGAN-GP (right) look the most realistic where objects like vehicles may be discerned.
Figure 17: Random VGAN samples on CelebA $128 \times 128$ at 300k iterations.
Figure 18: VGAN samples on CelebA HQ [Karras et al., 2018] 1024 × 1024 resolution at 300k iterations. Models are trained from scratch at full resolution, without the progressive scheme proposed by [Karras et al., 2017].