Diverse Image Generation via Self-Conditioned GANs

Steven Liu\textsuperscript{1} Tongzhou Wang\textsuperscript{1} David Bau\textsuperscript{1} Jun-Yan Zhu\textsuperscript{2} Antonio Torralba\textsuperscript{1}

\textsuperscript{1}MIT CSAIL \quad \textsuperscript{2}Adobe Research

Abstract

We introduce a simple but effective unsupervised method for generating realistic and diverse images. We train a class-conditional GAN model without using manually annotated class labels. Instead, our model is conditional on labels automatically derived from clustering in the discriminator’s feature space. Our clustering step automatically discovers diverse modes, and explicitly requires the generator to cover them. Experiments on standard mode collapse benchmarks show that our method outperforms several competing methods when addressing mode collapse. Our method also performs well on large-scale datasets such as ImageNet and Places365, improving both image diversity and standard quality metrics, compared to previous methods.

1. Introduction

Despite the remarkable progress of Generative Adversarial Networks (GANs)\textsuperscript{[13, 5, 20]}, there remains a significant gap regarding the quality and diversity between class-conditional GANs trained on labeled data, and unconditional GANs trained without any labels in a fully unsupervised setting\textsuperscript{[30, 26]}. This problem reflects the challenge of mode collapse: the tendency for a generator to focus on a subset of modes to the exclusion of other parts of the target distribution\textsuperscript{[12]}. Both empirical and theoretical studies have shown strong evidence that real data has a highly multi-modal distribution\textsuperscript{[33, 39]}. Unconditional GANs trained on such data distributions often completely miss important modes, e.g., not being able to generate one of ten digits for MNIST\textsuperscript{[37]}, or omitting object classes such as people and cars within synthesized scenes\textsuperscript{[3]}. Class-conditional GANs alleviate...
2. Related Work

Generative Adversarial Networks (GANs). Since the introduction of GANs [13], many variants have been proposed [30, 8, 35, 37, 1, 27, 14, 28], improving both the training stability and image quality. Due to its rapid advance, GANs have been used in a wide range of computer vision and graphics applications [40, 18, 41, 45, 17, 16]. GANs excel at synthesizing photorealistic images for a specific class of images such as faces and cars [19, 20]. However, for more complex datasets such as ImageNet, state-of-the-art models are class-conditional GANs that require ground truth image class labels during training [5]. To reduce the cost of manual annotation, a recent work [26] presents a semi-supervised method based on RotNet [11], a self-supervised image rotation feature learning method. The model is trained with labels provided on only a subset of images. On the contrary, our general-purpose method is not image-specific, and fully unsupervised. Section 4.4 shows that our method outperforms a RotNet-based baseline. A recent method [32] proposes to obtain good clustering using GANs, while we aim to achieve realistic and diverse generation.

Mode collapse. Although GANs are formulated as a minimax game in which each generator is evaluated against a discriminator, during optimization, the generator faces a slowly-changing discriminator that can guide generation to collapse to the point that maximizes the discriminator [29]. Mode collapse does occur in practice, and it is one of the fundamental challenges for GANs training [12, 37]. Several solutions have been proposed, including amending the adversarial loss to look ahead several moves (e.g., Unrolled GAN [29]), jointly training an encoder to recover latent distributions (e.g., VEEGAN [38]), packing the discriminator with sets of points instead of singletons [37, 25], and training a mixture of generators with an auxiliary diversity objective [15, 10]. Different from the above work, our method partitions the real distribution instead of the generated distribution, and devotes a class-conditioned discriminator to each target partition.

Another related line of research trains class-conditional GANs on unlabelled images by clustering on features obtained via unsupervised feature learning methods [26, 36]. Our method directly clusters on discriminator features that inherently exist in GANs, leading to a simpler method and achieving higher quality generation in our experiments (Section 4.4). Mixture Density GAN proposes to use log-likelihoods of a Gaussian mixture distribution in discriminator feature space as the GAN objective [9]. GAN-Tree uses clustering to split a GAN into a tree hierarchy of GANs for better mode coverage [23]. These methods, while also using clustering or mixture models, are mostly orthogonal with our work. Furthermore, the simplicity of our method allows it to be easily combined with a variety of these techniques.
3. Method

One of the core problems in generating diverse outputs in a high-dimensional space such as images is mode collapse: the support of the generator’s output distribution can be much smaller than the support of the real data distribution. One way mode collapse has been empirically lessened is by use of a class-conditional GAN, which explicitly penalizes the generator for not having support on each class.

We propose to exploit this class-conditional architecture, but instead of assuming access to true class labels, we will synthesize labels in an unsupervised way. On a high level, our method dynamically partitions the real data space into different clusters, which are used to train a class-conditional GAN. Because generation conditioned on a cluster index is optimized with respect to the corresponding conditioned discriminator, and each discriminator is responsible for a different subset of the real distribution, our method encourages the generator output distribution to cover all partitions of the real data distribution.

Therefore, to train a GAN to imitate a target distribution \( p_{\text{real}} \), we partition the data set into \( k \) clusters \( \{\pi_1, \ldots, \pi_k\} \) that are determined during training. No ground-truth labels are used; training samples are initially clustered in the randomly initialized discriminator feature space, and the clusters are updated periodically. A class-conditional GAN structure is used to split the discriminator and the generator. Next, we describe two core components of our algorithm:

- Conditional GAN training with respect to cluster labels given by the current partitioning.
- Updating the partition according to the current discriminator features of real data periodically.

**Conditional GAN training.** The GAN consists of a class-conditional generator \( G(z, c) \) associated with a class-conditional discriminator \( D(x, c) \). We denote the internal discriminator feature layers as \( D_f \) and its last layer as \( D_h \), so \( D = D_h \circ D_f \). The generator and discriminator are trained to optimize the following adversarial objective:

\[
\mathcal{L}_{\text{GAN}}(D, G) = \mathbb{E}_{c \sim P_c} \left[ \mathbb{E}_{x \sim \pi_c} \left[ \log(D(x, c)) \right] + \mathbb{E}_{z \sim \mathcal{N}(0, I)} \left[ \log(1 - D(G(z, c), c)) \right] \right],
\]

where the cluster index \( c \) is sampled from the categorical distribution \( P_c \) that weights each cluster proportional to its true size in the training set. Here \( G \) aims to generate images \( G(z, c) \) that look similar to the real images of cluster \( c \) for \( z \sim \mathcal{N}(0, I) \), while \( D(\cdot, c) \) tries to distinguish between such generated images and real images of cluster \( c \). They are jointly optimized in the following minimax fashion:

\[
\min_G \max_D \mathcal{L}_{\text{GAN}}(D, G). \tag{1}
\]

**Algorithm 1** Self-Conditioned GAN Training

Initialize generator \( G \) and discriminator \( D \)

Partition dataset into \( k \) sets \( \{\pi_1, \ldots, \pi_k\} \) using \( D_f \) outputs

for number of training epochs do

// Conditional GAN training based on current partition

for number of training iterations for an epoch do

for \( j \) in \( \{1, \ldots, m\} \) do

Sample cluster \( c^{(j)} \sim P_c \), where \( c \) is chosen with probability proportional to \( |\pi_c| \).

Sample image \( x^{(j)} \sim \pi_{c^{(j)}} \) from cluster \( c^{(j)} \).

Sample latent \( z^{(j)} \sim \mathcal{N}(0, I) \).

end for

Update \( G \) and \( D \) according to \( \min G \max D \mathcal{L}_{\text{GAN}} \) on minibatch \( \{(c^{(j)}, x^{(j)}, z^{(j)})\}_j \). \( \triangleright \) Eqn. (1)

end for

// Clustering to obtain new partitions

Cluster on \( D_f \) outputs of a subset of training set to identify a new partition \( \{\pi_c^{\text{new}}\} \) into \( k \) sets, using previous cluster centroids as initialization.

Find the matching \( \rho(\cdot) \) between \( \{\pi_c^{\text{new}}\}_c \) and \( \{\pi_c\}_c \) that minimizes \( \mathcal{L}_{\text{match}} \). \( \triangleright \) Eqn. (4)

Update all \( \pi_c \leftarrow \pi_c^{\text{new}} \).

end for

When under the condition \( c \), the discriminator is encouraged to give low score for any sample that is not from cluster \( c \) because \( p_{\text{real}}(x \mid c) = 0 \) for all \( x \notin \pi_c \). So the corresponding conditioned generator is penalized for generating points that are not from cluster \( c \), which ultimately prevents the generator from getting stuck on other clusters. The optimization is shown in Figure 2.

**Computing new partition by clustering.** As the training progresses, the shared discriminator layers \( D_f \) learn better representations of the data, so we periodically update \( \pi \) by re-partitioning the target dataset over a metric induced by the current discriminator features. We use \( k \)-means clustering [2] to obtain a new partition into \( k \) clusters \( \{\pi_c\}_{c=1}^k \) according to the \( D_f \) output space, approximately optimizing

\[
\mathcal{L}_{\text{cluster}}(\{\pi_c\}_{c=1}^k) = \mathbb{E}_{c \sim P_c} \left[ \mathbb{E}_{x \sim \pi_c} \left[ \|D_f(x) - \mu_c\|^2 \right] \right], \tag{2}
\]

where \( \mu_c \triangleq \frac{1}{|\pi_c|} \sum_{x \in \pi_c} D_f(x) \) is the mean of each cluster in \( D_f \) feature space.

**Clustering initialization.** For the first clustering, we use the \( k \)-means++ initialization, and when reclustering, we initialize the \( k \)-means algorithm with the means induced by the previous clustering. That is, if \( \{\pi_c\}_{c=1}^k \) denotes the old cluster means and \( \{\mu_c\}_{c=1}^k \) denotes the \( k \)-means
initialization to compute the new clustering, we take
\[
\mu_c^0 = \frac{1}{|\pi_c^{\text{old}}|} \sum_{x \in \pi_c^{\text{old}}} D_f^\text{new}(x),
\]
where \(D_f^\text{new}\) denotes the current discriminator feature space.

**Matching with old clusters.** After repartitioning, to avoid retraining the conditional generator and discriminator from scratch, we match the new clusters \(\{\pi_c^{\text{new}}\}_{c=1}^k\) to the old clusters \(\{\pi_c^{\text{old}}\}_{c=1}^k\) so that the target distribution for each generator does not change drastically. We formulate the task as a min-cost matching problem, where the cost of matching a \(\pi_c^{\text{new}}\) to a \(\pi_c^{\text{old}}\) is taken as \(|\pi_c^{\text{old}} \setminus \pi_c^{\text{new}}|\), the number of samples missing in the new partition. Therefore, we aim to find a permutation \(\rho: [k] \to [k]\) that minimizes the objective:
\[
L_{\text{match}}(\rho) = \sum_c |\pi_c^{\text{old}} \setminus \pi_c^{\text{new}}|.
\]

For a given new partitioning from \(k\)-means++, we solve this matching using the classic Hungarian min-cost matching algorithm [22], and obtain the new clusters to be used for GAN training in future epochs. Algorithm 1 summarizes the entire training method.

**Online clustering.** We also experimented with online \(k\)-means based on gradient descent [4], where we updated the cluster centers and membership using Equation (2) in each iteration. Our online variant achieves comparable results on mode collapse benchmarks (Section 4.2), but performs worse for real image datasets (Section 4.3), potentially due to the training instability caused by frequent clustering updates. Additionally, in Section 4, we also perform an ablation study regarding clustering initialization, online vs. batch clustering, and with or without clustering matching method.

4. Experiments
4.1. Implementation Details

**Network architecture.** For experiments on synthetic data, our unconditional generator and discriminator adopt the structure proposed in PacGAN[25]. To condition the generator, we add a fully connected layer before the input layer of the generator. To condition the discriminator, we increase the dimension of the output layer to be the number of clusters \(k\) we use. We use the outputs of the last hidden layer of the discriminator as features for clustering.

For Stacked-MNIST dataset, we use the DCGAN architecture [35] identical to that used in prior work[25]. For experiments on CIFAR-10, we use the DCGAN architecture identical to that used in SN-GANs [31]. We add conditioning to the GAN similarly to how we add conditioning to the generator of PacGAN. We use the outputs of the last convolutional layer of the discriminator’s features for clustering.

Large-scale GANs are trained on Places365 [43] and ImageNet [7]. In this setting, our GAN adopts the conditional architecture proposed in Mescheder et al. [28], and our unconditional GAN baseline removes all conditioning on the input label. We use the output from the last ResNet block as features for clustering.

**Clustering details.** To compute cluster centers, for Stacked-MNIST and CIFAR experiments, we cluster a random subset of 25,000 images from the dataset, and for ImageNet and Places365 experiments, we cluster a random subset of 50,000 images from the dataset. For \(k\)-means++, we cluster the subset ten times and choose the clustering that obtains the best performance on the clustering objective. We recluster every 25,000 iterations for Stacked-MNIST and CIFAR experiments, and recluster every 75,000 iterations for ImageNet and Places365 experiments. For experiments on synthetic datasets, we cluster a random sample of 10,000 data points and recluster every 10,000 iterations.

**Training details.** For experiments on synthetic datasets, we use 2-dimensional latents, and train for 400 epochs using Adam [21] with a batch size 100 and learning rate \(10^{-3}\). The embedding layer used for conditioning the generator has an output dimension of 32.

For CIFAR-10 and Stacked-MNIST experiments, we use 128 latent dimensions, and Adam with a batch size of 64 and a learning rate of \(1 \times 10^{-4}\) with \(\beta_1 = 0, \beta_2 = 0.99\). We train for 50 epochs on Stacked-MNIST and 400 epochs on CIFAR-10. The embedding layer used for conditioning the generator outputs a 256-dimensional vector.

For experiments on Places365 and ImageNet, our loss function is the vanilla loss function proposed by [13] and use \(R_1\) regularization as proposed by [28]. We find that regularization parameter 0.1 works well with our method. We train our networks from scratch using Adam with a batch size of 128 and a learning rate of \(10^{-4}\), for 200,000 iterations. We choose a 256-dimensional latent space. All metrics are reported using 50,000 samples from the fully trained models, with the Fréchet Inception Distance (FID) computed using the training set. The generator embedding layer outputs a 256-dimensional vector.

4.2. Synthetic Data Experiments

The 2D-ring dataset is a mixture of 8 2D-Gaussians, with means \((\cos \frac{2\pi i}{8}, \sin \frac{2\pi i}{8})\) and variance \(10^{-4}\), for \(i \in \{0, \ldots, 7\}\). The 2D-grid dataset is a mixture of 25 2D-Gaussians, each with means \((2i - 4, 2j - 4)\) and variance 0.0025, for \(i, j \in \{0, \ldots, 4\}\).

We follow the metrics used in prior work [38, 25]. A generated point is deemed high-quality if it is within three standard deviations from some mean [38]. The number of modes covered by a generator is the number of means that have at least one corresponding high-quality point. To
Table 1: Number of modes recovered, percent high quality samples, and reverse KL divergence metrics for 2D-Ring and 2D-Grid experiments. Results are averaged over ten trials, with standard error reported.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Number of Clusters</th>
<th>Modes (Max 8)</th>
<th>Reverse KL (Max 100)</th>
<th>% High Quality Samples</th>
</tr>
</thead>
<tbody>
<tr>
<td>2D-Ring</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GAN [13]</td>
<td>6.3 ± 0.5</td>
<td>98.2 ± 0.2</td>
<td>0.45 ± 0.09</td>
<td></td>
</tr>
<tr>
<td>PacGAN2 [25]</td>
<td>7.9 ± 0.1</td>
<td>95.6 ± 2.0</td>
<td>0.07 ± 0.03</td>
<td></td>
</tr>
<tr>
<td>PacGAN3 [25]</td>
<td>7.8 ± 0.1</td>
<td>97.7 ± 0.3</td>
<td>0.10 ± 0.02</td>
<td></td>
</tr>
<tr>
<td>PacGAN4 [25]</td>
<td>7.8 ± 0.1</td>
<td>95.9 ± 1.4</td>
<td>0.07 ± 0.02</td>
<td></td>
</tr>
<tr>
<td>MGAN (k = 50)</td>
<td>7.6 ± 0.2</td>
<td>80.3 ± 1.7</td>
<td>0.32 ± 0.06</td>
<td></td>
</tr>
<tr>
<td>Random Labels (k = 50)</td>
<td>7.9 ± 0.1</td>
<td>96.3 ± 1.1</td>
<td>0.07 ± 0.02</td>
<td></td>
</tr>
<tr>
<td>Our Method (k = 50)</td>
<td>8.0 ± 0.0</td>
<td>99.5 ± 0.3</td>
<td><strong>0.0014 ± 0.0002</strong></td>
<td></td>
</tr>
<tr>
<td>2D-Grid</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GAN [13]</td>
<td>7.3 ± 0.9</td>
<td>94.8 ± 0.7</td>
<td>0.70 ± 0.07</td>
<td></td>
</tr>
<tr>
<td>PacGAN2 [25]</td>
<td>23.8 ± 0.7</td>
<td>91.3 ± 0.8</td>
<td>0.13 ± 0.04</td>
<td></td>
</tr>
<tr>
<td>PacGAN3 [25]</td>
<td>24.6 ± 0.4</td>
<td>94.2 ± 0.4</td>
<td>0.06 ± 0.02</td>
<td></td>
</tr>
<tr>
<td>PacGAN4 [25]</td>
<td>24.8 ± 0.2</td>
<td>93.6 ± 0.6</td>
<td>0.04 ± 0.01</td>
<td></td>
</tr>
<tr>
<td>MGAN (k = 50)</td>
<td>22.5 ± 0.6</td>
<td>51.1 ± 1.6</td>
<td>0.36 ± 0.06</td>
<td></td>
</tr>
<tr>
<td>Random Labels (k = 50)</td>
<td>16.0 ± 1.0</td>
<td>90.6 ± 1.6</td>
<td>0.57 ± 0.07</td>
<td></td>
</tr>
<tr>
<td>Our Method (k = 50)</td>
<td>25.0 ± 0.0</td>
<td>99.5 ± 0.1</td>
<td><strong>0.0063 ± 0.0007</strong></td>
<td></td>
</tr>
</tbody>
</table>

Figure 3: The dependence of our method and MGAN on the choice of number of clusters k for 2D-grid (mixture of 25 Gaussians) experiment. MGAN is sensitive to the value k and performance degrades if k is too large. Our method is generally more stable and scales well with k.

We compare the reverse KL divergence between the generated distribution and the real distribution, we assign each point to its closest mean in Euclidean distance, and compare the empirical distributions [25].

Our results are displayed in Table 1. We compare our method to PacGAN [25] for various packing factors, on PacGAN’s mother architecture. We observe that our method generates both higher quality and more diverse samples than competing unsupervised methods. The difference in quality can be observed in Table 1. We also see that the success of our method is not solely due to the addition of class conditioning, as the conditional architecture with random labels still fails to capture diversity and quality.

Our online clustering variant, with k = 50 on the 2D-grid dataset is able to cover 25 modes with 99.1% high quality samples and a reverse KL of 0.00122 and on the 2D-ring dataset, is able to cover 8 modes with 99.9% high quality samples and a reverse KL of 0.00349.

We also compare our method’s sensitivity regarding the number of clusters k to existing clustering-based mode collapse methods [15]. For k chosen to be larger than the ground truth number of modes, we observe that our method levels off and generates both diverse and high-quality samples. For k chosen to be smaller than the ground truth number of modes, our results worsen. Figure 3 plots the sensitivity of k on the 2D-grid dataset, and the same trend holds for the 2D-ring dataset. Therefore, we lean towards using a larger number of clusters to ensure coverage of all modes.

MGAN [15] learns a mixture of k generators that can be compared to our k-way conditional generator. In Table 2, we see that MGAN performs worse as k increases. We hypothesize that when k is large, multiple generators must contribute to a single mode, and MGAN’s auxiliary classification loss, which encourages each generator’s output to be distinguishable, makes it harder for the generators to cover a mode collaboratively. On the other hand, our method scales well with k, because it dynamically updates cluster weights, and does not explicitly require the conditioned generators to output distinct distributions.

We run both our method and MGAN with varying k values on the 2D-grid dataset. The results summarized in Figure 3 confirm our hypothesis that MGAN is sensitive to the choice of k, while our method is more stable and scales well with k. In our arXiv version, we further present results showing that our method is much less sensitive to the variance of each Gaussian than MGAN.

4.3. Stacked-MNIST and CIFAR-10 Experiments

The Stacked-MNIST dataset [24, 38, 25] is produced by stacking three randomly sampled MNIST digits into an RGB image, one per channel, generating 10^3 modes with high probability. To calculate our reverse KL metric, we use pre-trained MNIST and CIFAR-10 classifiers to classify and count the occurrences of each mode. For these experiments, we use k = 100. We also compare our method against an online variant, as well as a variant where we start with
Table 2: Number of modes recovered, reverse KL divergence, and Inception Score (IS) metrics for Stacked MNIST and CIFAR-10 experiments. Results are averaged over five trials, with standard error reported. Results of PacGAN on Stacked MNIST are taken from [25]. For CIFAR-10, all methods recover all 10 modes.

<table>
<thead>
<tr>
<th></th>
<th>Stacked MNIST</th>
<th>CIFAR-10</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Modes (Max 1000) ↑</td>
<td>Reverse KL ↓</td>
</tr>
<tr>
<td>GAN [13]</td>
<td>133.4 ± 17.70</td>
<td>2.97 ± 0.216</td>
</tr>
<tr>
<td>PacGAN2 [25]</td>
<td>1000.0 ± 0.00</td>
<td>0.06 ± 0.003</td>
</tr>
<tr>
<td>PacGAN3 [25]</td>
<td>1000.0 ± 0.00</td>
<td>0.06 ± 0.003</td>
</tr>
<tr>
<td>PacGAN4 [25]</td>
<td>1000.0 ± 0.00</td>
<td>0.07 ± 0.005</td>
</tr>
<tr>
<td>Logo-GAN-AE [36]</td>
<td>1000.0 ± 0.00</td>
<td>0.00 ± 0.005</td>
</tr>
<tr>
<td>Logo-GAN-RC [36]</td>
<td>1000.0 ± 0.00</td>
<td>0.08 ± 0.006</td>
</tr>
<tr>
<td>Random Labels</td>
<td>240.0 ± 12.02</td>
<td>2.90 ± 0.192</td>
</tr>
<tr>
<td>Online Clustering</td>
<td>995.8 ± 0.86</td>
<td>0.17 ± 0.027</td>
</tr>
<tr>
<td>Ours + Supervised Init.</td>
<td>1000.0 ± 0.00</td>
<td>0.08 ± 0.014</td>
</tr>
<tr>
<td>Our Method</td>
<td>1000.0 ± 0.00</td>
<td>0.08 ± 0.009</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Places365</th>
<th>ImageNet</th>
</tr>
</thead>
<tbody>
<tr>
<td>GAN [13]</td>
<td>14.21 ± 0.00</td>
<td>8.7122 ± 0.014</td>
</tr>
<tr>
<td>PacGAN2 [25]</td>
<td>18.02 ± 0.00</td>
<td>8.5765 ± 0.014</td>
</tr>
<tr>
<td>PacGAN3 [25]</td>
<td>22.00 ± 0.00</td>
<td>8.5637 ± 0.014</td>
</tr>
<tr>
<td>MGN [15]</td>
<td>15.78 ± 0.00</td>
<td>8.4141 ± 0.014</td>
</tr>
<tr>
<td>RotNet Feature Clustering</td>
<td>14.88 ± 0.00</td>
<td>8.5436 ± 0.014</td>
</tr>
<tr>
<td>Logo-GAN-AE [36]</td>
<td>14.51 ± 0.00</td>
<td>8.1936 ± 0.014</td>
</tr>
<tr>
<td>Logo-GAN-RC [36]</td>
<td>8.66 ± 0.00</td>
<td>10.5474 ± 0.014</td>
</tr>
<tr>
<td>Random Labels</td>
<td>14.20 ± 0.00</td>
<td>8.8154 ± 0.014</td>
</tr>
<tr>
<td>Our Method</td>
<td>9.76 ± 0.00</td>
<td>8.8186 ± 0.014</td>
</tr>
<tr>
<td>Class Conditional GAN [30]</td>
<td>8.12 ± 0.00</td>
<td>10.9738 ± 0.014</td>
</tr>
</tbody>
</table>

Table 3: Fréchet Inception Distance (FID), Inception Score (IS), and reverse KL divergence metrics for Places365 and ImageNet experiments. Our method improves in both quality and diversity over previous methods but still fails to reach the quality of fully-supervised class conditional ones.

<table>
<thead>
<tr>
<th></th>
<th>Places365</th>
<th>ImageNet</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>FID ↓</td>
<td>IS ↑</td>
</tr>
<tr>
<td>GAN [13]</td>
<td>14.21 ± 0.00</td>
<td>8.7122 ± 0.014</td>
</tr>
<tr>
<td>PacGAN2 [25]</td>
<td>18.02 ± 0.00</td>
<td>8.5765 ± 0.014</td>
</tr>
<tr>
<td>PacGAN3 [25]</td>
<td>22.00 ± 0.00</td>
<td>8.5637 ± 0.014</td>
</tr>
<tr>
<td>MGN [15]</td>
<td>15.78 ± 0.00</td>
<td>8.4141 ± 0.014</td>
</tr>
<tr>
<td>RotNet Feature Clustering</td>
<td>14.88 ± 0.00</td>
<td>8.5436 ± 0.014</td>
</tr>
<tr>
<td>Logo-GAN-AE [36]</td>
<td>14.51 ± 0.00</td>
<td>8.1936 ± 0.014</td>
</tr>
<tr>
<td>Logo-GAN-RC [36]</td>
<td>8.66 ± 0.00</td>
<td>10.5474 ± 0.014</td>
</tr>
<tr>
<td>Random Labels</td>
<td>14.20 ± 0.00</td>
<td>8.8154 ± 0.014</td>
</tr>
<tr>
<td>Our Method</td>
<td>9.76 ± 0.00</td>
<td>8.8186 ± 0.014</td>
</tr>
<tr>
<td>Class Conditional GAN [30]</td>
<td>8.12 ± 0.00</td>
<td>10.9738 ± 0.014</td>
</tr>
</tbody>
</table>

manual class labels as the first clustering. Details of these variants are described in our arXiv version.

Our results are shown in Table 2. On both datasets, we are able to achieve large gains in diversity, significantly improve Fréchet Inception Distance (FID) and Inception Score (IS) over baselines and even the supervised baselines (e.g., Class-conditional GAN). Interestingly, starting with supervised class labels seems to slightly degrade performance compared to starting with clusters from random discriminator features.

We find that our cluster initialization step is necessary: without it, on CIFAR, we obtain an FID score of 19.85 and Inception Score of 7.41. Furthermore, our cluster rematching step is necessary: without it, on CIFAR, we obtain an FID score of 20.42 and Inception Score of 7.31.

4.4. Large-Scale Image Datasets Experiments

Lastly, we measure the quality and diversity of images for GANs trained on large-scale datasets Places365 ImageNet. We trained using all 1.2 million ImageNet challenge images across all 1000 classes and all 1.8 million Places365 images across 365 classes. No class labels were revealed to the model. For both datasets, we choose $k = 100$.

Across datasets, our method improves the quality of generated images in terms of FID, IS, and reverse KL divergence, outperforming all baselines in most cases.

Figure 4 shows samples of the generated output, comparing our model to a vanilla GAN and a PacGAN model and a sample of training images. To highlight differences in diversity between the different GAN outputs, Figure 4a shows a sample of 20 images from each method taken from a single Places365 class. To predict the class label from the generated output, we classify each image using a standard ResNet18 scene classifier [43], and we show the 20 highest-scoring members of the class from a sample of 50,000 generated images. Similarly, Figure 4b shows samples from ImageNet generators filtered to a single ImageNet class, as classified by a standard ResNet50 classifier [34]. We observed that across classes, vanilla GAN and PacGAN trend to synthesize less diverse samples, repeating many similar images. On the other hand, our method improves diversity significantly and does not produce visually similar images.

4.4.1 Reconstruction of Real Images

While random samples reveal the positive output of the generator, reconstructions of training set images [44, 6] can be used to visualize the omissions of the generator [3].

Previous GAN inversion methods do not account for class conditioning, so we extend the encoder+optimization hybrid method of [44, 3]. We first train an encoder backbone $F : x \rightarrow r$ jointly with a classifier $F_c : r \rightarrow c$ and a reconstruction network $F_r : r \rightarrow z$ to recover the class $c$ and the original $z$ of a generated image. We then optimize $z$ to match the pixels of the query image $x$ as well as encoder features extracted by $F$:

$$L_{\text{reconst}}(z, c) = \|G(z, c) - x\|_1 + \lambda_f \|F(G(z, c)) - F(x)\|_2^2,$$
We set $\lambda_f = 5 \times 10^{-4}$. When initialized using $F_z$ and $F_c$, this optimization faithfully reconstructs images generated by $G$, and reconstruction errors of real images reveal cases that $G$ omits. More details regarding image reconstruction can be found in our arXiv version.

To evaluate how well our model can reconstruct the data distribution, we compute the average LPIPS perceptual similarity score \[\text{LPIPS}\] between 50,000 ground truth images and their reconstructions. Between two images, a low LPIPS score suggests the reconstructed images are similar to target real images. We find that on Places365, our model is able to better reconstruct the real images, with an average LPIPS score of 0.433, as compared to the baseline score of 0.528.

Figure 5 visually compares reconstructions of Places365 training set images created by our model with those of the same architecture trained without self-conditioning and clustering algorithm. Reconstructions done by previous training iterations are also compared to the final models. Visualizations show distinctive features that can be reconstructed by our model that are not present in the baseline model, including improved forms for cars, buildings, and indoor objects. These features correspond to common visual features that appear in the clusters. Self-conditioned classes include classes of images with cars, buildings, indoor rooms, and scenes from specific viewpoints.

We apply a similar procedure to evaluate reconstructions of a particular cluster, obtaining the average LPIPS score between real images in the cluster and their reconstructed images. Figure 6 shows quantitatively that per cluster LPIPS of our method are noticeably better than those of the unconditional GANs baseline, which is consistent with Figure 5.

4.5. Clustering Metrics

We measure the quality of our clustering through Normalized Mutual Information (NMI) and clustering purity across all experiments in Table 4.

NMI is defined as $\text{NMI}(X, Y) = \frac{2I(X; Y)}{H(X) + H(Y)}$, where $I$ is mutual information and $H$ is entropy. NMI lies in $[0, 1]$, and higher NMI suggests higher quality of clustering. Purity is defined as $\frac{1}{N} \sum_y \max_k |\pi_c \cap \pi^*_y|$, where $\{\pi_c\}_{c=1}^k$ is the partition of inferred clusters and $\{\pi^*_y\}_{y=1}^g$ is the partition given by the true classes. Higher purity suggests higher clustering quality. Purity is close to 1 when each cluster has a large majority of points from some true class set.
We observe that many of our clusters in large-scale datasets do not correspond directly to true classes, but instead corresponded to object classes. For example, we saw that many clusters corresponded to people and animals, none of which are part of a true class, which is an explanation for low clustering metric scores.

Though our clustering scores are low, they are significantly better than a random clustering. Randomly clustering ImageNet to 100 clusters gives an NMI of 0.0069 and a purity of 0.0176; randomly clustering Places to 100 clusters gives an NMI of 0.0019 and a purity of 0.0137.

5. Conclusion

We have found that when a conditional GAN is trained with clustering labels derived from discriminator features, it is effective at reducing mode collapse, outperforming several previous approaches. We observe that the method continues to perform well when the number of synthesized labels exceeds the number of modes in the data. Furthermore, our method scales well to large-scale datasets, improving Fréchet Inception Distance and Inception Score measures on ImageNet and Places365 generation, and generating images that are qualitatively more diverse than an unconditional GAN.

Acknowledgments. We thank Phillip Isola, Bryan Russell, Richard Zhang, and our anonymous reviewers for helpful comments. We are grateful for the support from the DARPA XAI program FA8750-18-C000, NSF 1524817, NSF BIGDATA 1447476, and a GPU donation from NVIDIA.
References


