Appendix A. Proofs

Theorem 1. *Suppose p_{data} is a distribution supported on a set of disjoint manifolds $\mathcal{M}_1, \dots, \mathcal{M}_k$ in $\mathbb{R}^d$, and $[\pi_1, \dots, \pi_k]$ are the probabilities of being from each manifold. Let $G_{\boldsymbol{\theta}}$ be a c-Lipschitz function, and p_{model} be the distribution of $G_{\boldsymbol{\theta}}(\mathbf{z})$, where $\mathbf{z} \sim \mathcal{N}(0, \boldsymbol{I}_n)$, then:*

$$d_{TV}(p_{\text{data}}, p_{\text{model}}) \geq \sum |\pi_i - p_i| \geq \delta$$

where d_{TV} is the total variation distance and:

$$\begin{aligned}
\pi_i^* &:= \min(\pi_i, 1 - \pi_i) \\
p_i &:= p_{\text{model}}(\mathcal{M}_i) \\
\delta &:= \max_i \{\pi_i^* - \Phi(\Phi^{-1}(\pi_i^*) - d_i/c)\} \\
d_i &:= \inf\{||\boldsymbol{x} - \boldsymbol{y}|| \mid \boldsymbol{x} \in \mathcal{M}_i, \boldsymbol{y} \in \mathcal{M}_j, j \neq i\}
\end{aligned}$$

d_i is the distance of manifold $\mathcal{M}_i$ from the rest, and Φ is the CDF of the univariate standard normal distribution. Note δ is strictly larger than zero iff $\exists i : \; d_i, \pi_i^* \neq 0$.

Proof. We begin by re-stating the definition of Minkowski sum and proceed by proving the theorem for the case where the number of manifold is 2. To extend the theorem form $k = 2$ to the general case, one only needs to consider manifold $\mathcal{M}_i$ as $\mathcal{M}_1$, and $\bigcup_{j \in [1:k] \setminus i} \mathcal{M}_j$ as $\mathcal{M}_2$.

Definition 1 (Minkowski sum)**.** *The Minkowski sum of two sets $U, V \in \mathbb{R}^d$ defined as*

$$U + V := \{u + v | u \in U, v \in V\}$$

and when V is a d dimensional ball with radius r and centered at zero, we use the notation U_r to refer to their Minkowski sum.

If we let $U^{(1)} := G_\theta^{-1}(\mathcal{M}_1), U^{(2)} := G_\theta^{-1}(\mathcal{M}_2)$, then:

$$\forall r_1, r_2 \in \mathbb{R}_+, \text{ if } \; r_1 + r_2 < d_1/c \implies U^{(1)}_{r_1} \cap U^{(2)}_{r_2} = \emptyset$$

that is because if there exists an $\mathbf{x} \in U^{(1)}_{r_1} \cap U^{(2)}_{r_2}$, there would be $\mathbf{u}_1 \in U^{(1)}, \mathbf{u}_2 \in U^{(2)}$ such that:

$$||\mathbf{x} - \mathbf{u}_1|| \leq r_1, \;\; ||\mathbf{x} - \mathbf{u}_2|| \leq r_2 \implies ||\mathbf{u}_1 - \mathbf{u}_2|| < r_1 + r_2 < d_1/c \tag{9}$$

However, due to lipsitz condition of G_θ:

$$||G_\theta(\mathbf{u}_1) - G_\theta(\mathbf{u}_2)|| < c||\mathbf{u}_1 - \mathbf{u}_2|| < c \cdot d_1/c = d_1$$

which contradicts with our assumption that the distance between $\mathcal{M}_1, \mathcal{M}_2$ is d_1. Therefore there is no point in the intersection of $U^{(1)}_{r_1}$ and $U^{(2)}_{r_2}$. The disjointness of this two sets provides us:

$$\gamma_n(U^{(1)}_{r_1}) + \gamma_n(U^{(2)}_{r_2}) \leq \gamma_n(\mathbb{R}^n) = 1$$

where $\gamma_n(.)$ of any set is the probability of a random draw of $\mathcal{N}(0, \boldsymbol{I}_n)$ being from that set. We proceed by using a remark from theorem 1.3 of [42] which restated below:

Lemma 1. *If U is a Borel set in $\mathbb{R}^n$, then:*

$$p \leq \gamma_n(U) \implies \Phi(\Phi^{-1}(p) + r) \leq \gamma_n(U_r).$$

Based on above lemma if we let:

$$p_1 := \gamma_n(U^{(1)}), \;\; p_2 := \gamma_n(U^{(2)})$$

then for $\forall r_1, r_2 \in \mathbb{R}_+$ such that $r_1 + r_2 < d_1/c$, we have:

$$\Phi(\Phi^{-1}(p_1) + r_1) + \Phi(\Phi^{-1}(p_2) + r_2) \leq 1 \tag{10}$$

We can now calculate the total variational distance of the marginal distributions of $p_{\text{data}}, p_{\text{model}}$ on the set $\{G_\theta(U^{(1)}), G_\theta(U^{(2)}), G_\theta\left(\mathbb{R}^n \setminus (U^{(1)} \cup U^{(2)})\right)\}$ as:

$$d_{TV}(p_{\text{data}}^{(\text{marginal})}, p_{\text{model}}^{(\text{marginal})}) = |\pi_1 - p_1| + |\pi_2 - p_2| + |1 - (p_1 + p_2)| \tag{11}$$

and since total variational distance takes a smaller value on marginal distributions than the full distribution, we only need to show that for any i the expression in the equation 11 is larger than $g(\pi_i, d_i, c)$ to prove the theorem 1 for $k = 2$. Here $g(\pi_i, d_i, c) = \pi_i^* - \Phi(\Phi^{-1}(\pi_i^*) - d_i/c)$.

Assume $p_1 \leq \pi_1$, and define $\Delta_1 := \pi_1 - p_1 \geq 0$, based on equation 10 if $r_1 = d_1/c, \;\; r_2 = 0$, we have:

$$\Phi(\Phi^{-1}(\pi_1 - \Delta_1) + d_1/c) + p_2 \leq 1$$

which based on equation 10 implies:

$$r(\Delta_1) := \Phi(\Phi^{-1}(\pi_1 - \Delta_1) + d_1/c) - (\pi_1 - \Delta_1) \leq D_{TV}$$

which D_{TV} refers to total variational distance between the marginal distributions of the data and model. We also know from the equation 10, that $\Delta_1 \leq D_{TV}$, therefore:

$$\max(r(\Delta_1), \Delta_1) \leq D_{TV}$$

for a $\Delta_1 \in [0, \pi_1]$. Therefore

$$\min_{\delta_1 \in [0, \pi_1]} \{\max(r(\delta_1), \delta_1)\} \leq D_{TV}$$

To find the δ_1 which minimize the above equation, we need to check endpoints of the interval $[0, \pi_1]$, points where the curve of two functions $r(\delta_1), \delta_1$ intersects with each other, and points that are the local minimaum of each of them. It can be shown the function $r(\delta_1)$ does not have any local minima when $0 < \pi_1 < 1$ because:

$$r(\delta_1) = P(\mathrm{z} \in [\Phi^{-1}(\pi_1 - \delta_1), \Phi^{-1}(\pi_1 - \delta_1) + d_1/c]) \tag{12}$$

where z is univarite standard normal random variable. Therefore $r(\delta_1)$ is the probablity of a univarite normal being in a fixed length interval d_1/c, and δ_1 only changes the starting point of the interval. By using this fact, it can be easily shown this function does not have any local optima in the

open interval $(0, \pi_1)$. Also the identity function δ_1 also has no local optima inside the interval. The endpoints values are:

$$\max(r(0), 0) = \Phi(\Phi^{-1}(\pi_1) + d_1/c) - \pi_1$$

$$\max(r(\pi_1), \pi_1) = \max(\Phi(\Phi^{-1}(0) + d_1/c), \pi_1) = \pi_1$$

The function curves of r and identity also intersects only when:

$$\Phi(\Phi^{-1}(\pi_1 - \delta_1^*) + d_1/c) = \pi_1$$

which only happens when:

$$\pi_1 - \Phi(\Phi^{-1}(\pi_1) - d_1/c) = \delta_1^*$$

where for this point, $\max(r(\delta_1^*), \delta_1^*) = \delta_1^*$. Therefore based on the above calculations:

$$\min \Bigg\{ \underbrace{\Phi(\Phi^{-1}(\pi_1) + d_1/c) - \pi_1}_{\text{I}}, \pi_1, \underbrace{\pi_1 - \Phi(\Phi^{-1}(\pi_1) - d_1/c)}_{\text{II}} \Bigg\} \leq D_{TV}$$

Note, π_1 is always smaller than term II, and term I (II) is equal to probability of a univariate standard normal random variable being inside the interval $[\Phi^{-1}(\pi_1), \Phi^{-1}(\pi_1) + d_1/c]$ ($[\Phi^{-1}(\pi_1) - d_1/c, \Phi^{-1}(\pi_1)]$). This observation implies that term II is smaller than term I, if and only if $\pi_1 \leq \pi_2$. Based on this fact and symmetry of Φ with respect to zero, it can be easily shown that:

$$g(\pi_1, d_1, c) \leq D_{TV} \tag{13}$$

which proves the theorem. However, we made an assumption that $p_1 \leq \pi_1$, this does not harm the argument because otherwise we would have $p_2 \leq \pi_2$, and we can restate all the above arguments for π_2 instead of π_1. And, since $\pi_2 = 1 - \pi_1$ and $d_1 = d_2$, therefore we can have $g(\pi_1, d_1, c) = g(\pi_2, d_2, c)$, which proves equation 13. □

Theorem 2. *Let $P = \sum_i^k \pi_i p_i$, $Q = \sum_i^k \pi_i q_i$, and $A_1, A_2, ..., A_K$ be a partitioning of the space, such that the support of each distribution p_i and q_i is A_i. Then:*

$$\text{JSD}(P \parallel Q) = \sum_i \pi_i \text{JSD}(p_i \parallel q_i) \tag{3}$$

Proof. Based on the definition of the JSD, we have:

$$\text{JSD}(P \parallel Q) = \frac{1}{2}\text{KL}(P \parallel \frac{P+Q}{2}) + \frac{1}{2}\text{KL}(Q \parallel \frac{P+Q}{2})$$

We also have:

$$\begin{aligned}
&\text{KL}(P \parallel \frac{P+Q}{2}) = \\
&\int_{\mathbb{R}^d} P(x) \log \frac{P(x)}{P(x)+Q(x)}\, dx + \log 2 = \\
&\sum_i \int_{A_i} P(x) \log \frac{P(x)}{P(x)+Q(x)}\, dx + \log 2 = \\
&\sum_i \int_{A_i} \pi_i p_i(x) \log \frac{\pi_i p_i(x)}{\pi_i p_i(x) + \pi_i q_i(x)}\, dx + \log 2 = \\
&\sum_i \pi_i \Big(\int_{A_i} p_i(x) \log \frac{p_i(x)}{p_i(x)+q_i(x)}\, dx + \log 2 \Big) = \\
&\sum_i \pi_i \text{KL}(p_i \parallel \frac{p_i+q_i}{2}).
\end{aligned}$$

Therefore:

$$\text{KL}(P \parallel \frac{P+Q}{2}) = \sum_i \pi_i \text{KL}(p_i \parallel \frac{p_i+q_i}{2}),$$

and similarly:

$$\text{KL}(Q \parallel \frac{P+Q}{2}) = \sum_i \pi_i \text{KL}(q_i \parallel \frac{p_i+q_i}{2})$$

Adding these two terms completes the proof. □

Theorem 3. *Let $\phi(\mathbf{x}) : \mathbb{R}^d \to \mathbb{R}^d$ be a C^1 (differentiable with continuous derivative) function, $W^{partitioner} \in \mathbb{R}^{k \times d}$, and R_i as defined in Eq 6. If there exists $c_0 > 0$, such that:*

$$\forall\, \mathbf{x}, \mathbf{y} \in \mathbb{R}^d, \quad c_0 ||\mathbf{x} - \mathbf{y}|| \leq ||\phi(\mathbf{x}) - \phi(\mathbf{y})||,$$

then for every $i \in [1 : k]$, every local optima of R_i is a global optima, and there exists a positive constant $b_0 > 0$ such that:

$$\forall \mathbf{x} \in \mathbb{R}^d \setminus A_i, \quad b_0 \leq ||\nabla R_i(\mathbf{x})||$$

where $A_i = \{\mathbf{x} | \mathbf{x} \in \mathbb{R}^d, R_i(\mathbf{x}) = 0\}$. Furthermore A_i is a connected set for all i's.

Proof. We start by proving that the Jacobian matrix of function ϕ is invertible for any $\mathbf{x} \in \mathbb{R}^d$. Since $\phi \in C^1$, based on Taylor's expansion theorem for multi-variable vector-valued function ϕ, we can write:

$$\phi(\mathbf{y}) - \phi(\mathbf{x}) = \boldsymbol{J}_\phi(\mathbf{x})\,(\mathbf{y} - \mathbf{x}) + o\,(||\mathbf{y} - \mathbf{x}||)$$

Were $o(\cdot)$ is the Little-o notation. By taking norm from both sides and using triangle inequality, we have:

$$||\phi(\mathbf{y}) - \phi(\mathbf{x})|| \leq ||\boldsymbol{J}_\phi(\mathbf{x})\,(\mathbf{y} - \mathbf{x})\,|| + ||o\,(||\mathbf{y} - \mathbf{x}||)\,||$$

Also because:

$$c_0 ||\mathbf{y} - \mathbf{x}|| \leq ||\phi(\mathbf{y}) - \phi(\mathbf{x})||$$

$$\implies c_0||\mathbf{y}-\mathbf{x}|| \leq ||\boldsymbol{J}_\phi(\mathbf{x})\,(\mathbf{y}-\mathbf{x})\,|| + ||o\,(||\mathbf{y}-\mathbf{x}||)\,|| \tag{14}$$

thus for any fixed $\mathbf{x}$: $\exists\,\epsilon > 0$ such that $\forall\,\mathbf{y} \in \mathbb{R}^d$ where $||\mathbf{y}-\mathbf{x}|| \leq \epsilon$ then:

$$||o\,(||\mathbf{y}-\mathbf{x}||)\,|| \leq \frac{c_0}{2}||\mathbf{y}-\mathbf{x}||$$

which combined with the inequality 14, results in:

$$\frac{c_0}{2}||\mathbf{y}-\mathbf{x}|| \leq ||\boldsymbol{J}_\phi(\mathbf{x})\,(\mathbf{y}-\mathbf{x})\,||$$

For $\mathbf{y} \neq \mathbf{x}$, let $\mathbf{u} := (\mathbf{y}-\mathbf{x})/||\mathbf{y}-\mathbf{x}||$, then by dividing both sides of the above inequality to $||\mathbf{y}-\mathbf{x}||$ we have:

$$\forall \mathbf{u} \in \mathbb{R}^d, ||\mathbf{u}|| = 1 \implies \frac{c_0}{2} \leq ||\boldsymbol{J}_\phi(\mathbf{x})\mathbf{u}||$$

which shows the Jacobian matrix of ϕ is invertiable for any $\mathbf{x}$ and all of its singular values are larger than $c_0/2$. If there is no $\mathbf{x} \in \mathbb{R}^d \setminus A_i$ the proof is complete. Otherwise, consider any $\mathbf{x} \in \mathbb{R}^d \setminus A_i$, for this $\mathbf{x}$ we have:

$$0 < R_i(\mathbf{x}) = \sum_c (f_c(\mathbf{x}) - f_i(\mathbf{x}))_+ = \sum_c ((\boldsymbol{w}_c - \boldsymbol{w}_i)\phi(\mathbf{x}))_+$$

where $\boldsymbol{w}_j$ is the j'th row of the matrix $\boldsymbol{W}^{partitioner}$. Let:

$$I(\mathbf{x}) := \{c | f_c(\mathbf{x}) > f_i(\mathbf{x}), \quad c \in [1:k]\}$$

which is an non-empty set, because $0 < R_i(\mathbf{x})$ and we have

$$0 < R_i(\mathbf{x}) = \left[\sum_{c \in I(\mathbf{x})} (\boldsymbol{w}_c - \boldsymbol{w}_i)\right] \phi(\mathbf{x})$$
$$\implies \boldsymbol{v} := \sum_{c \in I(\mathbf{x})} (\boldsymbol{w}_c - \boldsymbol{w}_i) \neq 0$$

Taking the gradient of the new formulation of R_i, we have:

$$\nabla R_i(\mathbf{x}) = \left[\sum_{c \in I(\mathbf{x})} (\boldsymbol{w}_c - \boldsymbol{w}_i)\right] \nabla(\phi(\mathbf{x})) = \boldsymbol{v}\boldsymbol{J}_\phi(\mathbf{x})$$

but since we showed earlier that all of the singular values of the Jacobian matrix is larger than $c_0/2$, the Jacobian matrix is $d \times d$, and $\boldsymbol{v}$ is not equal to zero, it can be easily shown:

$$||\nabla R_i(\mathbf{x})|| > ||\boldsymbol{v}||\frac{c_0}{2} := b_0$$

Now, we also need to show A_i is connected for any i to complete the proof. To that end, we first show ϕ is a surjective function, which means its image is $\mathbb{R}^d$. To show the ϕ is surjective, we prove its image is both an open and closed set, then since the only sets which are both open and closed (in $\mathbb{R}^d$) are $\mathbb{R}^d, \emptyset$, we can conclude the surejective property. The image of ϕ is an open set due to Inverse Function Theorem [62] for ϕ. We are allowed to use Inverse Function Theorem, since ϕ satisfies both C^1 condition and non zero determinant for all the points in the domain. We will also show that the image of ϕ is a closed set by showing it contains all of its limit points. Let $\boldsymbol{y}$ be a limit point in the image of ϕ, that is there exists $\{\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots\}$ such that $\phi(\boldsymbol{x}_r) \to \boldsymbol{y}$. Since $\mathbb{R}^d$ is complete and we have $c_0||\boldsymbol{x}_r - \boldsymbol{x}_s|| \leq ||\phi(\boldsymbol{x}_r) - \phi(\boldsymbol{x}_s)||$, then $\{\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots\}$ is a Cauchy sequence. Finally since ϕ is a continues function $\phi(\boldsymbol{x}^*) = \boldsymbol{y}$, completing the proof.

The function ϕ is also an invertible function because if $\phi(\boldsymbol{x}) = \phi(\boldsymbol{y})$ then

$$c_0||\mathbf{x}-\boldsymbol{y}|| \leq ||\phi(\boldsymbol{x}) - \phi(\boldsymbol{y})|| = 0$$

which implies $\boldsymbol{x} = \boldsymbol{y}$. Therefore ϕ is in fact a continuous bijecitve function, which means it has a continuous inverse defined on all the space $\mathbb{R}^d$. Furthermore, it can be easily shown R_i for a datapoint is zero iff its transformation by ϕ lies in a polytope (where each of its facets is a hyperplance perpendicular to a $\boldsymbol{w}_c - \boldsymbol{w}_i$). Since convex polytope is a connected set, and by applying ϕ^{-1} (it is well defined everywhere because of bijective property of ϕ) to it, we would have a connected set. That is because a continuous function does not change the connectivity and ϕ^{-1} is continuous. □

Appendix B. Additional qualitative results

We present more samples of our method showing both the partitioner and generative model's performance. Figure 6 and Figure 7 visualize the sample diversity and quality of our method on CIFAR-10 and STL-10.

B.1. CIFAR-10

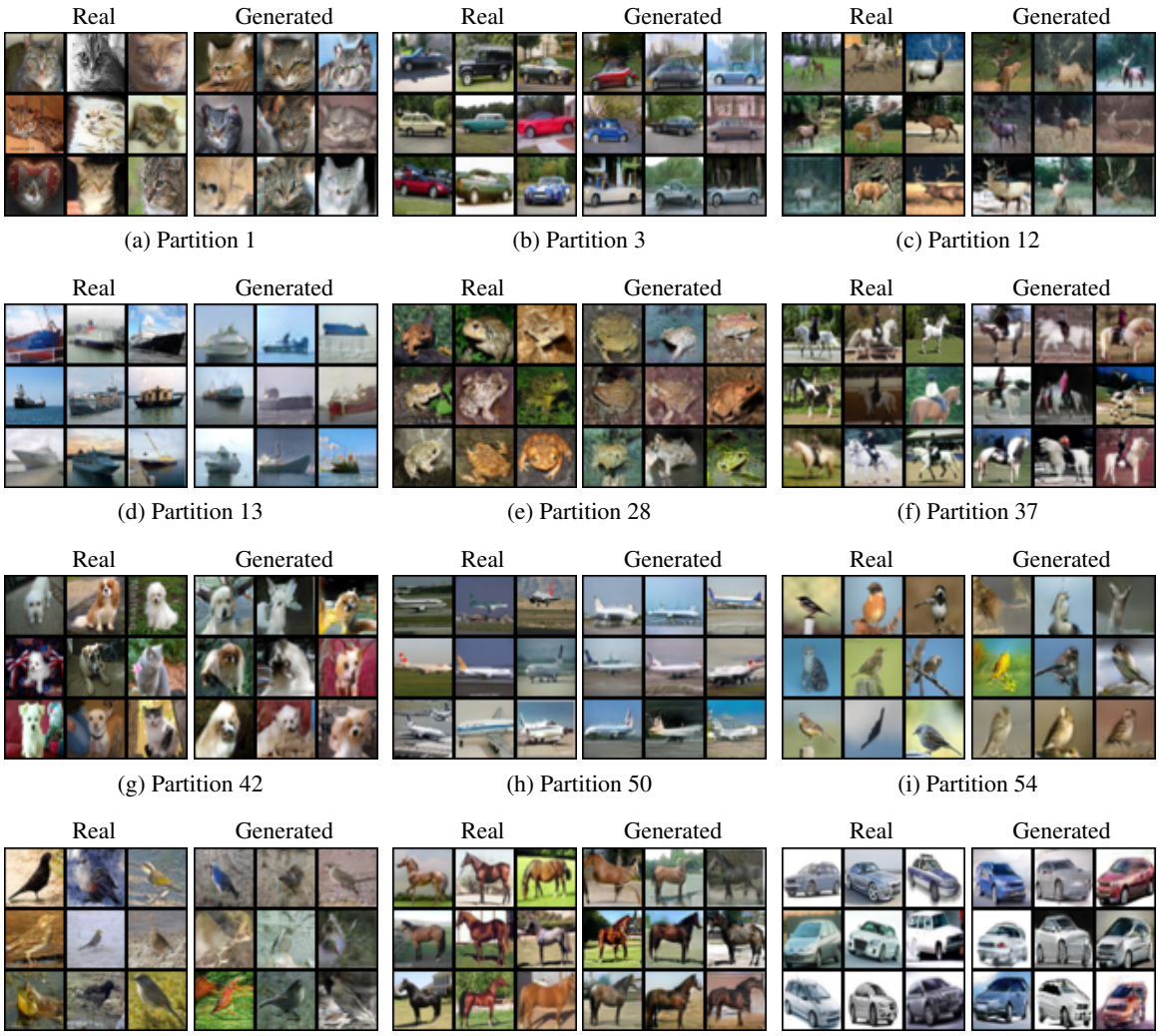
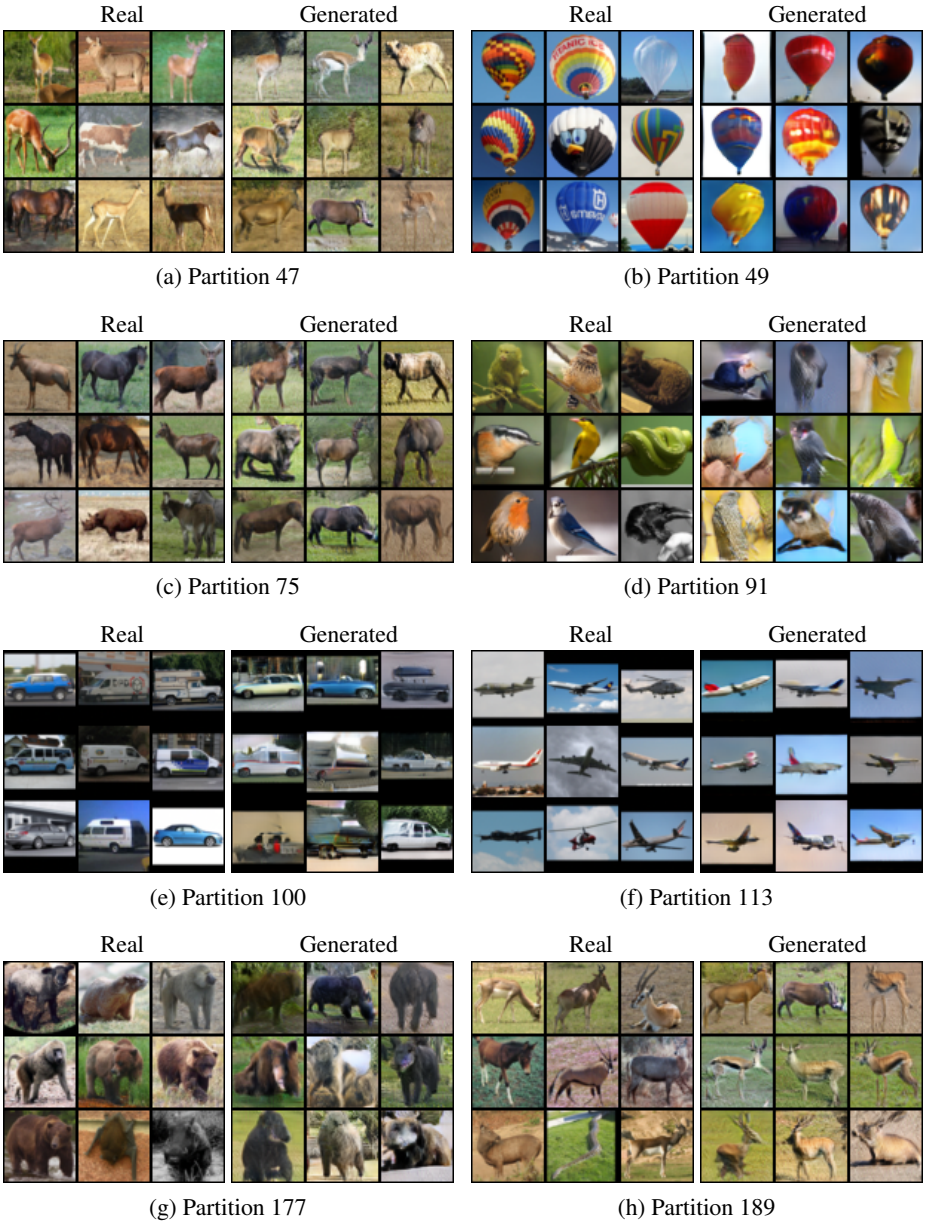


Figure 6: Extra examples of unsupervised partitioning and their corresponding real/generated samples on CIFAR-10 dataset.

B.2. STL-10



(a) Partition 47 (b) Partition 49

(c) Partition 75 (d) Partition 91

(e) Partition 100 (f) Partition 113

(g) Partition 177 (h) Partition 189

Figure 7: Extra examples of unsupervised partitioning and their corresponding real/generated samples on STL-10 dataset.

Appendix C. Implementation details

We use two RTX 2080 Ti GPUs for experiments on STL-10, eight V-100 GPUs for ImageNet and a single GPU for all other experiments.

Space partitioner. For all experiments we use the same architecture for our space partitioner S. We use pre-activation Residual-Nets with 20 convolutional bottleneck blocks with 3 convolution layers each and kernel sizes of 3×3, 1×1, 3×3 respectively and the ELU [11] nonlinearity. The network has 4 down-sampling stages, every 4 blocks where a dimension squeezing operation is used to decrease the spatial resolution. We use 160 channels for all the blocks. We do not use any initial padding due to our theoretical requirements. The negative slope of LeakyReLU is set as 0.2. In fact we can use a soft version of LeakyReLU if it is critical to guarantee the C^1 constraint of ϕ. We train our pretext network for 500 epochs with momentum SGD and a weight decay of 3e-5, learning rate of 0.4 with cosine scheduling, momentum of 0.9, and batch size of 400 for CIFAR-10 and 200 for STL-10. The final space partitioner is trained for 100 epochs using Adam [38] with a learning rate of 1e-4 and batch size of 128. The weights in equation 5 are set to $\alpha = 5$ and $\beta =$ 1e-3.

Generative model. Following SN-GANs [55] for image generation at resolution 32 or 48, we use the architectures described in Tables 6 and 7. Generators/discriminators are different from each other in first-layer/last-layer by having different partition ID embeddings, (which in fact acts as the condition). We use Adam optimizer with a batch size of 100. For the coefficient of guide λ we utilized linear annealing during training, decreasing form 6.0 to 0.0001. Both G's and D networks are initialized with a normal $\mathcal{N}(0, 0.02\boldsymbol{I})$. For all GAN's experiments, we use Adam optimizer [38] with $\beta_1 = 0$, $\beta_2 = 0.999$ and a constant learning rate 2 for both G and D. The number of D steps per G step training is 4.

For ImageNet experiment, we adopt the full version of BigGAN model architecture [5] described in Table 8. In this experiment, we apply the shared class embedding for each CBN layer in G, and feed noise z to multiple layers of G by concatenating with the partition ID embedding vector. Moreover, we add Self-Attention layer with the resolution of 64, and we employ orthogonal initialization for network parameters [66]. We use batch size of 256 and set the number of gradient accumulations to 8.

Evaluation. It has been shown that [10, 48] when the sample size is not large enough, both FID and IS are biased, therefor we use N=50,000 samples for computing both IS and FID metrics. We also use the official TensorFlow scripts for computing FID.

C.1. Additional Experiments:

Our quantitative results on the (2D-ring, 2D-grid) toy datasets [45] are: recovered modes: (8 , 25), high quality samples: (99.8 , 99.8), reverse KL: (0.0006, 0.0034).

Given the strong performances of recent models (and ours) on these datasets we suffice to these stats. For visualizations of our generated samples related to these two datasets please see Figure 4-left and refer to the appendix of [46] for other methods.

Additional architecture dependent experiment: Since SelfCondGAN [46] uses certain features from the discriminator, it is not trivial to adopt to other architectures. Thus, we trained PGMGAN with the same G/D architecture on CIFAR10 yielding an FID of 10.65.

Partitioning method One way to assess the quality of the space partitioner is by measuring its performance on placing semantically similar images in the same partition. To that end, we use the well-accepted clustering metric Normalized Mutual Information (NMI). NMI is a normalization of the Mutual Information (MI) between the true and inferred labels. This metric is invariant to permutation of the class/partition labels and is always between 0 and 1, with a higher value suggesting a higher quality of partitioning. Table 5 compares the clustering performance of our method to the-stat-of-the-art partition-based GAN method Liu et. [46], which clearly shows superiority of our method.

Table 5: Comparison of the clustering performance in term of Normalized Mutual Information (NMI), higher is better.

	Stacked MNIST	CIFAR-10	STL-10	ImageNet
Self-Cond-GAN [46]	0.3018	0.3326	-	17.39
PGMGAN	0.4805	0.4146	0.3911	68.57

Table 6: GANs architecture for 32×32 images.

$z \in \mathbb{R}^{128} \sim \mathcal{N}(0, I)$ Embed(*Partition$_{ID}$*) $\in \mathbb{R}^{128}$
dense, $4 \times 4 \times 256$
ResBlock up 256
ResBlock up 256
ResBlock up 256
BN, ReLU, 3×3 Conv, 3 Tanh

(a) Generator

RGB image $x \in \mathbb{R}^{32 \times 32 \times 3}$
ResBlock down 128
ResBlock down 128
ResBlock 128
ResBlock 128
ReLU, Global sum pooling
Embed(*Partition$_{ID}$*)$\cdot \boldsymbol{h}$ + (linear $\rightarrow$ 1)

(b) Discriminator

Table 7: GANs architecture for 48×48 images.

$z \in \mathbb{R}^{128} \sim \mathcal{N}(0, I)$ Embed(*Partition$_{ID}$*) $\in \mathbb{R}^{128}$
dense, $3 \times 3 \times 1024$
ResBlock up 512
ResBlock up 256
ResBlock up 128
ResBlock up 64
BN, ReLU, 3×3 Conv, 3 Tanh

(a) Generator

RGB image $x \in \mathbb{R}^{48 \times 48 \times 3}$
ResBlock down 64
ResBlock down 128
ResBlock down 256
ResBlock down 512
ResBlock 1024
ReLU, Global sum pooling
Embed(*Partition$_{ID}$*)$\cdot \boldsymbol{h}$ + (linear $\rightarrow$ 1)

(b) Discriminator

Table 8: GANs architecture for 128×128 images. "ch" represents the channel width multiplier and is set to 96.

$z \in \mathbb{R}^{120} \sim \mathcal{N}(0, I)$ Embed(*Partition$_{ID}$*) $\in \mathbb{R}^{128}$
Linear $(20 + 128) \rightarrow 4 \times 4 \times 16ch$
ResBlock up $16ch \rightarrow 16ch$
ResBlock up $16ch \rightarrow 8ch$
ResBlock up $8ch \rightarrow 4ch$
ResBlock up $4ch \rightarrow 2ch$
Non-Local Block (64×64)
ResBlock up $2ch \rightarrow ch$
BN, ReLU, 3×3 Conv $ch \rightarrow 3$
Tanh

(a) Generator

RGB image $x \in \mathbb{R}^{128 \times 128 \times 3}$
ResBlock down $ch \rightarrow 2ch$
Non-Local Block (64×64)
ResBlock down $2ch \rightarrow 4ch$
ResBlock down $4ch \rightarrow 8ch$
ResBlock down $8ch \rightarrow 16ch$
ResBlock down $16ch \rightarrow 16ch$
ResBlock $16ch \rightarrow 16ch$
ReLU, Global sum pooling
Embed(*Partition$_{ID}$*)$\cdot \boldsymbol{h}$ + (linear $\rightarrow$ 1)

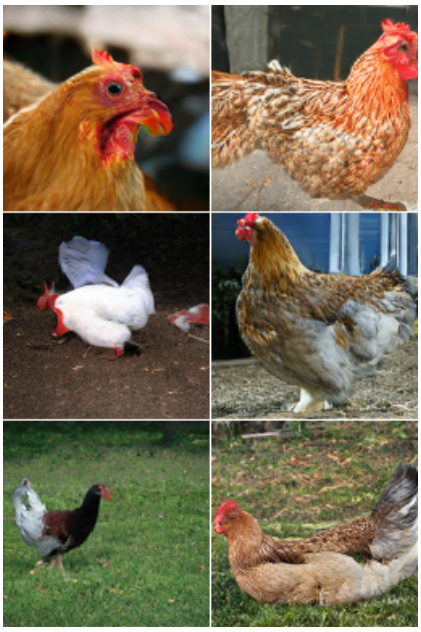
(b) Discriminator

C.2. ImageNet

(a) Partition 034

(b) Partition 159

(c) Partition 455



(d) Partition 642

Figure 8: Examples of generated samples on unsupervised ImageNet 128×128 dataset.