STRUCTURAL ADVERSARIAL OBJECTIVES FOR SELF-SUPERVISED REPRESENTATION LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

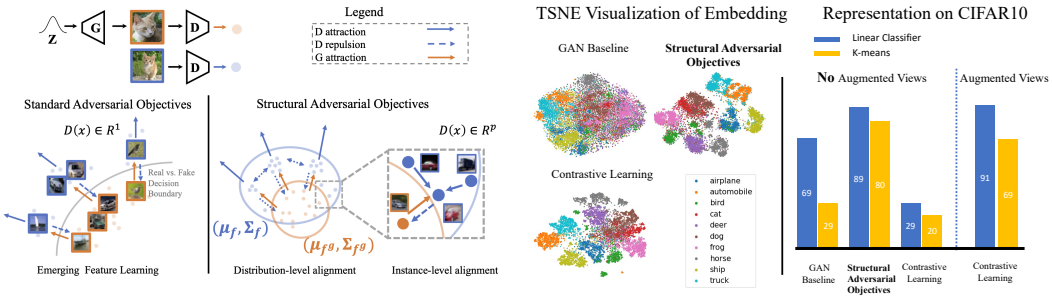
Within the framework of generative adversarial networks (GANs), we propose objectives that task the discriminator for self-supervised representation learning via additional structural modeling responsibilities. In combination with an efficient smoothness regularizer imposed on the network, these objectives guide the discriminator to learn to extract informative representations, while maintaining a generator capable of sampling from the domain. Specifically, our objectives encourage the discriminator to structure features at two levels of granularity: aligning distribution characteristics, such as mean and variance, at coarse scales, and grouping features into local clusters at finer scales. Operating as a feature learner within the GAN framework frees our self-supervised system from the reliance on hand-crafted data augmentation schemes that are prevalent across contrastive representation learning methods. Across CIFAR-10/100 and an ImageNet subset, experiments demonstrate that equipping GANs with our self-supervised objectives suffices to produce discriminators which, evaluated in terms of representation learning, compete with networks trained by contrastive learning approaches.

1 INTRODUCTION

Unsupervised feature learning algorithms aim to directly learn representations from data without reliance on annotations, and have become crucial to efforts to scale vision and language models to handle real-world complexity. Many state-of-the-art approaches adopt a contrastive self-supervised framework, wherein a deep neural network is tasked with mapping augmented views of a single example to nearby positions in a high-dimension embedding space, while separating embeddings of different examples (Wu et al., 2018; He et al., 2020; Chen et al., 2020; Chen & He, 2021; Grill et al., 2020; Zbontar et al., 2021). Though requiring no annotation, and hence unaffected by assumptions baked into any labeling procedure, the invariances learned by these models are still influenced by human-designed heuristic procedures for creating augmented views.

The recent prominence of contrastive approaches was both preceded by and continues alongside a focus on engineering domain-relevant proxy tasks for self-supervised learning. For computer vision, examples include learning geometric layout (Doersch et al., 2015), colorization (Zhang et al., 2016; Larsson et al., 2017), and inpainting (Pathak et al., 2016; He et al., 2022). Basing task design on domain knowledge may prove effective in increasing learning efficiency, but strays further from an alternative goal of developing truly general and widely applicable unsupervised learning techniques.

Another family of approaches, coupling data generation with representation learning, may provide a path toward such generality while also escaping dependence upon the hand-crafted elements guiding data augmentation or proxy task design. Generative adversarial networks (GANs) (Goodfellow et al., 2014; 2020) and variational autoencoders (VAEs) (Kingma & Welling, 2013) are prime examples within this family. Considering GANs, one might expect the discriminator to act as an unsupervised representation learner, driven by the need to model the real data distribution in order to score the generator's output. Indeed, prior work finds that some degree of representation learning occurs within discriminators in a standard GAN framework (Radford et al., 2015). Yet, to improve generator output quality, limiting the capacity of the discriminator appears advantageous (Arjovsky et al., 2017) – a choice potentially in conflict with representation learning. Augmenting the standard GAN framework to separate encoding and discrimination responsibility into different components (Don-



(a) Standard *vs.* proposed structural adversarial objectives for feature learning

(b) Visualizing and quantitatively evaluating discriminator features on CIFAR-10

Figure 1: *(a) Structural GAN Objectives:* In a standard GAN, the discriminator produces a scalar score to discern real and fake samples. As the generator improves, representations produced by the discriminator will update structurally similar data in a similar direction, displayed as solid blue arrows. Our structural adversarial objectives enhance such learning capability by optimizing the feature vectors produced by the discriminator. We achieve this by manipulating mean and variance at a coarser scale and implementing instance-level grouping at a finer scale, allowing the discriminator to explicitly learn semantic representations, in addition to distinguishing between real and fake. *(b) Discriminator as Semantic Representation Learner:* Trained with our new objectives, the discriminator's learned feature embedding reveals category semantics and achieves performance competitive with contrastive learning methods. Unlike self-supervised contrastive methods, our approach *does not* depend upon learning from different views obtained via a data augmentation scheme.

ahue et al., 2017; Dumoulin et al., 2017), along with scaling to larger models (Donahue & Simonyan, 2019), are promising paths forward.

However, it has been unclear whether the struggle to utilize vanilla GANs as effective representation learners stems from inherent limitations of the framework. We provide evidence to the contrary, through an approach that significantly improves representations learned by the discriminator, while maintaining generation quality and operating with a standard pairing of generator and discriminator components. To enhance GANs into effective representation learners, our approach need only modify the training objectives within the GAN framework. Our contributions are as follows:

- We propose adversarial objectives resembling a contrastive clustering target (Figure 1). These self-supervised objectives prompt the discriminator to learn semantic representations, without depending on data augmentation to fuel the learning process.
- We introduce an effective regularization approach that utilizes the approximation of the spectral norm of the Jacobian to regulate the smoothness of the discriminator. This methodology enables the discriminator to strike a balance between its capacity to learn features and its ability to properly guide the generator.
- On representation learning benchmarks, our method achieves competitive performance with recent state-of-the-art contrastive self-supervised learning approaches, even though we do not leverage information from (or even have a concept of) an augmented view. We demonstrate that supplementing a GAN with our proposed objectives not only enhances the discriminator as a representation learner, but also improves the quality of samples produced by the generator.

2 Related Work

2.1 Generative Feature Learning

GANs (Goodfellow et al., 2014; 2020) include two learnable modules: a generator G_ϕ, which produces synthetic data given a sample $\boldsymbol{v}$ from a prior, and a discriminator D_θ, which learns to differentiate between the true data $\boldsymbol{x}$ and generated samples $G_\phi(\boldsymbol{v})$. Here, θ, ϕ denote the trainable parameters. During training, G_ϕ and D_θ are alternatively updated in an adversarial fashion, which

can be formulated as a minimax problem:

$$\min_{G_\phi} \max_{D_\theta} \mathbb{E}_{\boldsymbol{x}\sim p(\boldsymbol{x})}[\log D_\theta(\boldsymbol{x})] - \mathbb{E}_{\hat{\boldsymbol{x}}\sim G_\phi(\boldsymbol{v})}[1 - \log G_\phi(\hat{\boldsymbol{x}})]. \tag{1}$$

Much research on GANs has focused on improving the quality of generated data, yielding significant advances (Karras et al., 2017; 2019; 2020b; 2021; Sauer et al., 2022; Dai et al., 2022). Other efforts have focused on evolving capabilities, including conditional and controllable generation, *e.g.,* text-guided (Zhang et al., 2021; Hinz et al., 2020) or segmentation-guided (Zhu et al., 2017; Chen & Koltun, 2017) generation. In comparison, adopting GANs for unsupervised feature learning has been more scarcely explored. In this area, an adversarial approach dependent upon an additional encoder component (Donahue et al., 2017; Dumoulin et al., 2017; Donahue & Simonyan, 2019; Jahanian et al., 2021) appears most successful to date. Here, the encoder is tasked to invert the generator with a discriminator acting on (data, latent) pairs and representation learning is the responsibility of the encoder, rather than the discriminator.

Besides GANs, other generative models also demonstrate feature learning capability. Recent efforts (Zhang et al., 2022; Ma et al., 2021) discard the low-level structures in VAE and Flow models to improve learned representations. Du et al. (2021) show that an unsupervised energy model can learn semantic structures, *e.g.,* segmentation and viewpoint, from images. Preechakul et al. (2022) attach an encoder to a diffusion model and show that it learns high-level feature representations. We adopt an orthogonal approach that, by imposing structural adversarial objectives in GAN training, tasks the discriminator to learn richer data representations.

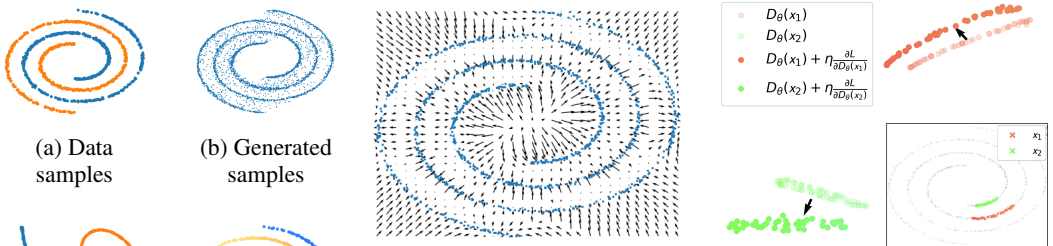
2.2 Contrastive Self-Supervised Learning

Self-supervised learning with a contrastive approach has shown enhanced feature learning capability and has evolved to nearly match the performance of its supervised counterparts. From initial impactful results in vision and language (Wu et al., 2018; He et al., 2020; Chen et al., 2020; Radford et al., 2021), this technique has recently been employed across a variety of domains (Jiang & Willett, 2022; Krishnan et al., 2022; Güldenring & Nalpantidis, 2021). A popular strategy involves using a Siamese architecture to optimize the InfoNCE objective, which aims to maximize the feature similarity across augmented views, while repulsing from all other instances to maintain feature uniformity (Wu et al., 2018; He et al., 2020; Chen et al., 2020; Oord et al., 2018). Another strategy simplifies this pipeline by dropping the negative terms and leveraging specific architectural designs to prevent collapsed solutions (Chen & He, 2021; Grill et al., 2020). As an alternative to operating on an l_2 normalized embedding, other approaches (Caron et al., 2020; 2021; Wang et al., 2021) enforce clustering consistency across views. Inspired by masked language modeling, He et al. (2022) and Bao et al. (2021) propose variants in the image domain by tasking an autoencoder to predict masked pixels.

Though contrastive approaches yield strong benchmark results, Tian et al. (2020) showcase the limitations of view-invariant assumptions and demonstrate their sensitivity to the parameters of augmentation schemes. Zhang & Maire (2020) raise a concern with applying these methods to broader unconstrained datasets, where multiple object instances within the same image should not have mutually invariant representations.

2.3 Stabilizing GAN Training

Despite the ability to generate high-quality samples, successfully training GANs remains challenging due to the adversarial optimization. Several approaches have been proposed to stabilize training and enable scaling to larger models. Heusel et al. (2017) suggest maintaining separate learning rates for the generator and discriminator, in order to maintain local Nash equilibrium. Arjovsky et al. (2017) and Gulrajani et al. (2017) consider constraining the discriminator's Lipschitz constant with gradient clipping and gradient norm penalization. In contrast to regularizing model-wise functionality, Miyato et al. (2018) implement layer-wise spectral normalization schemes by dividing parameters with their leading singular value, which is widely adopted in recent state-of-the-art models. Wu et al. (2021) and Bhaskara et al. (2022) instead propose to build a Lipschitz-constrained function by dividing the output with the gradient norm, and show it can preserve model capacity. However, none of these methods suit our case, since spectral normalization (Miyato et al., 2018) harms model capacity, and gradient-based regularization only works for scalar output, limiting the use of structural objectives.



(a) Data samples

(b) Generated samples

(c) t-SNE visualization

(d) t-SNE palette map

(e) Gradient ($\frac{\partial \mathcal{L}}{\partial x}$) resembles the path of optimal transport, suggesting $D_\theta(x)$ can represent intrinsic structure of the data.

(f) When updated using $\mathcal{L}$, represented by arrow directions, $D_\theta(x)$ will align with semantically similar data, marked by the same color, and diverge from dissimilar data, indicated by different colors.

Figure 2: We train a GAN with our structural objectives on a synthetic *double spiral* dataset. We show: *(a) training data* color-coded based on ground truth assignments; *(b) generated samples*; *(c,d) learned representations* visualized by t-SNE (Van der Maaten & Hinton, 2008), and colored according to ground truth categories *(c)* as well as a 2D palette map *(d)*. Additionally, we highlight *(e) structural correspondence* of $D_\theta(x)$ via $\frac{\partial \mathcal{L}}{\partial x}$, and in *(f)*, we visualize $D_\theta(\boldsymbol{x})$ using t-SNE, showcasing the emerging capability for learning semantic features induced by our loss $\mathcal{L}$ (Eq. 8).

3 Method: Feature Learning with the Discriminator

Our goal is to task D_θ as both a discriminator and a feature extractor that learns semantic representations of real data. We motivate this design from empirical observations of GAN discriminator behavior. Figure 1a conveys some intuition behind our design, while Figure 2 illustrates results, as well as discriminator learning dynamics when applying our method to a synthetic dataset.

As Figure 2f shows, the updating direction induced by our loss enables $D_\theta(\boldsymbol{x})$ to position example $\boldsymbol{x}$ close to similarly structured examples while diverging away from dissimilar ones. Such behavior is not necessarily limited to our system; we hypothesize that it arises in broader contexts due to a Lipschitz-regularized discriminator producing gradients that rearrange the embedding along an optimal transport path, as shown in Figure 2e. As a consequence, structurally similar samples will be updated in a similar direction. Tanaka (2019) establishes this idea in the context of Wasserstein GANs (Arjovsky et al., 2017).

This conjecture suggests that, in a standard GAN, the discriminator implicitly learns some, but perhaps not all, aspects of a semantic representation. We are therefore motivated to propose explicit objectives for the discriminator that are both compatible with its original purpose (providing informative gradients to the generator) and that require it to produce an embedding that captures additional semantic structure of the data distribution.

3.1 Structural Adversarial Objectives

Instead of producing a scalar output, we architect D_θ to learn the mapping from the data space to the feature space, $D_\theta : \mathcal{X} \to \mathcal{Z}$. We denote the output from D_θ on real data and fake (generated) samples as $\boldsymbol{z}$ and $\boldsymbol{z}^g$, respectively. Here $\boldsymbol{z}, \boldsymbol{z}^g \in \mathbf{S}^{p-1}$ are normalized and live in a unit hypersphere. We also maintain unnormalized counterparts $\tilde{\boldsymbol{z}}$ and $\tilde{\boldsymbol{z}}^g$ of $\boldsymbol{z}$ and $\boldsymbol{z}^g$; Section 3.2 explores their utility.

Driving the formulation of our proposed objectives is the idea to require D_θ to model the real and fake distributions (without collapse), while G_ϕ adversarially attempts to align these distributions. As related prior work, OT-GAN (Salimans et al., 2018) proposes explicit optimal-transport adversarial objectives for this purpose, but requires a large batch size (8K) to stabilize. Instead, our objectives operate hierarchically and regularize the learned embeddings at two levels of granularity:

(1) At a coarse level, we align the distribution statistics of the discriminator, focusing on its mean and covariance: $\boldsymbol{\mu}_{\boldsymbol{z}}, \boldsymbol{\mu}_{\boldsymbol{z}^g} \in \mathbb{R}^p$ and $\boldsymbol{\Sigma}_{\boldsymbol{z}}, \boldsymbol{\Sigma}_{\boldsymbol{z}^g} \in \mathbb{R}^{p \times p}$. Here, we simplify the optimization by assuming a diagonal structure of the covariance matrix. This enables efficient alignment of the two distributions with tolerance to finer-grained differences.

(2) At a finer level, we focus on reorganizing embeddings by constructing clusters using local affinity. The corresponding objective tasks D_θ with learning local geometry, further focusing the GAN on feature alignment between real and fake distributions.

Coarse-scale optimization by aligning distributions. To align the distributions in terms of mean and covariance, we can employ a distance function $d(\cdot)$ and optimize the minimax objective:

$$\mathcal{L}_{\text{Gaussian}} := \min_{G_\phi} \max_{D_\theta} d(\boldsymbol{z}, \boldsymbol{z}^g). \tag{2}$$

One widely adopted candidate for $d(\cdot)$ is Jensen-Shannon divergence (JSD) due to its symmetry and stability. For two arbitrary probability distributions P, Q, JSD admits the following form:

$$\text{JSD}(P||Q) = \frac{1}{2}(D_{\text{KL}}(P||\frac{P+Q}{2}) + D_{\text{KL}}(Q||\frac{P+Q}{2})) = \frac{1}{2}(H(\frac{P+Q}{2}) - \frac{1}{2}(H(P) + H(Q))). \tag{3}$$

where D_{KL}, H denote Kullback-Leibler divergence and entropy, respectively. We can compute entropy for Q, P using closed-form expressions. However, entropy for $(P+Q)/2$ is difficult to compute exactly and generally requires Monte Carlo simulation, an infeasible computational approach in high dimensional space. To tackle this problem, we follow Hershey & Olsen (2007) to approximate $\frac{P+Q}{2}$ by a single Gaussian and estimate sample mean and covariance by joint samples of P and Q, which yields an upper bound of $H(\frac{P+Q}{2})$; the bound is tight when $P = Q$. Putting these together, we obtain our distance function for the coarser scale objective[1]:

$$\text{JSD}(\boldsymbol{z}, \boldsymbol{z}^g) \approx \log \frac{\det \boldsymbol{\Sigma}_{\boldsymbol{z}+\boldsymbol{z}^g}}{\sqrt{\det \boldsymbol{\Sigma}_{\boldsymbol{z}} \det \boldsymbol{\Sigma}_{\boldsymbol{z}^g}}}. \tag{4}$$

Another well-established metric between two Gaussian distributions is Bhattacharyya distance D_B:

$$D_B(\boldsymbol{z}, \boldsymbol{z}^g) := \frac{1}{8}(\boldsymbol{\mu}_{\boldsymbol{z}} - \boldsymbol{\mu}_{\boldsymbol{z}^g})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{\mu}_{\boldsymbol{z}} - \boldsymbol{\mu}_{\boldsymbol{z}^g}) + \frac{1}{2} \log \frac{\det \boldsymbol{\Sigma}}{\sqrt{\det \boldsymbol{\Sigma}_{\boldsymbol{z}} \det \boldsymbol{\Sigma}_{\boldsymbol{z}^g}}}, \tag{5}$$

where $\boldsymbol{\Sigma} = \frac{\boldsymbol{\Sigma}_{\boldsymbol{z}} + \boldsymbol{\Sigma}_{\boldsymbol{z}^g}}{2}$. Though having different geometric interpretations, it is notable that D_B and JSD have similar format and, when maximizing $d(\boldsymbol{z}, \boldsymbol{z}^g)$ for $\boldsymbol{z}$, both aim to uniformly repulse z to prevent producing collapsed representations. In experiments, we observe that these two distances yield similar performance and we use JSD as our default choice for $d(\cdot)$ since it has a slightly faster convergence rate and yields better quality for generated images.

Fine-grained optimization via clustering. We perform mean-shift clustering on $\boldsymbol{z}$ by grouping nearby samples. We simplify the clustering process by equally averaging each neighbor sample, rather than using feature similarity to reweight their contribution. To improve nearest neighbor search stability, we maintain a rolling updated memory bank $\boldsymbol{z}^m$ that stores the embedding of all real images as a query pool and use the backbone representation $\boldsymbol{z}^b$, rather than $\boldsymbol{z}$, as the key to computing feature similarity. Denoting $\{\boldsymbol{z}_{i,j}\}_{j=1}^k$ and $\{\boldsymbol{z}_{i,j}^g\}_{j=1}^K$ as the returned K nearest neighbors of real images embedding for $\boldsymbol{z}_i$ and $\boldsymbol{z}_i^g$ respectively, our clustering objective is:

$$\mathcal{L}_{\text{cluster}} := \max_{D_\theta} \frac{1}{NK} \sum_{i=1}^{N} \sum_{j=1}^{K} \boldsymbol{z}_{i,j}^\top \boldsymbol{z}_i + \min_{D_\theta} \max_{G_\phi} \frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{K} \boldsymbol{z}_{i,j}^{g\,\top} \boldsymbol{z}_i^g. \tag{6}$$

IC-GAN (Casanova et al., 2021) implements a similar instance-wise objective. However, they use frozen embeddings from an off-the-shelf model rather than jointly learn an embedding, and their motivation is to improve image generation quality rather than learn semantic features — entirely different from our aim.

3.2 Smoothness Regularization

Besides reformulating adversarial targets for representation learning, we address another common issue in GAN training: balancing the discriminator's capacity and the smoothness constraint. Recent studies demonstrate that regularizing D_θ's smoothness, or its Lipschitz constant, is critical for

[1]Note that though Eq. 4 and MCR in Dai et al. (2022) are constructed similarly, the latter is interpreted from a coding rate reduction perspective.

scaling GANs to large network architectures. Consider a continuous function $F : \mathbb{R}^m \to \mathbb{R}^p$. We can bound its Lipschitz constant by the spectral norm of Jacobian $\boldsymbol{J}_F$:

$$\|\boldsymbol{J}_F(\boldsymbol{x})\|_2 \leq \text{Lip},$$

where $\|\cdot\|_2$ denotes the matrix spectral norm. However, computing the full Jacobian matrix is highly inefficient in standard backpropagation process since each backpropagation call can only compute a single row of the Jacobian matrix, which is impracticable as we usually need large embedding dimension p.

Therefore, we propose to efficiently approximate $\|\cdot\|_2$ using power-iterations. Leveraging the fact that power-iteration is a matrix-free method, we do not need to explicitly compute the Jacobian matrix. Instead, we only need to access the matrix by evaluating the matrix-vector product, which can be efficiently computed by batch-wise VJP and JVP (Jacobian-Vector-Product) subroutine. Algorithm 1 presents the details, where only $(2S+1)$ backpropagation calls are required to approximate $\|\boldsymbol{J}_{D_\theta}(\boldsymbol{x})\|_2$. In experiments, we find that $S = 1$ suffices for a ResNet-18 model.

Algorithm 1: Approximating $\|\boldsymbol{J}_F(\boldsymbol{x})\|_2$ with power iterations

Input: Function $F : \mathbb{R}^m \to \mathbb{R}^p$; Stop gradient operator $sg(\cdot)$; Power iteration steps S; Batch size b; Input data $\boldsymbol{x} \in \mathbb{R}^{b\times m}$;
Init random vector $\boldsymbol{u} \sim \mathcal{N}(0, 1) \in \mathbb{R}^{b\times p}$
for *iter* $= 1 \ldots S$ **do**
  $\boldsymbol{v} = \boldsymbol{u}\boldsymbol{J}_F(\boldsymbol{x})/\|\boldsymbol{u}\boldsymbol{J}_F(\boldsymbol{x})\|_2$ //VJP
  $\boldsymbol{u} = \boldsymbol{J}_F(\boldsymbol{x})\boldsymbol{v}/\|\boldsymbol{J}_F(\boldsymbol{x})\boldsymbol{v}\|_2$ //JVP
end
Return: $\|\boldsymbol{J}_F(\boldsymbol{x})\|_2 \approx sg(\boldsymbol{u})\boldsymbol{J}_F(\boldsymbol{x})sg(\boldsymbol{v})$

We observe that maintaining $\|\tilde{\boldsymbol{z}}\|$ at regular level benefits training stability. To this end, we use hinge loss to regularize the embedding norm and empirically observe it performs better than removing the hinge. Therefore, our smoothness regularization is:

$$\min_{D_\theta} \mathbb{E}_{\boldsymbol{x}}\|\boldsymbol{J}_{D_\theta}(\boldsymbol{x}) - \text{Lip}\|_2 + \lambda_h \mathbb{E}_{\tilde{\boldsymbol{z}}}\| \max(\|\tilde{\boldsymbol{z}}\| - 1, 0)\|_2 \tag{7}$$

where λ_h denotes the ratio for hinge regularization, and Lip denotes the Lipschitz target of D_θ, which is set to 1 by default. Unlike a layer-wise normalization scheme, *e.g.,* Spectral Norm (Miyato et al., 2018), where demanding local regularization hurts the model's capacity, our proposed regularization scheme allows the network to simultaneously fit multiple objectives, *i.e.,* representation learning and smoothness regularization. The model does not have to sacrifice capacity for smoothness. Another benefit of our method is that our proposed term can work with the normalization layer. Spectral Norm cannot, because of the data-dependent scaling term in its normalization layer.

Overall training objective. We define our final objective as:

$$\mathcal{L} := \mathcal{L}_{\text{Gaussian}} + \lambda_c \mathcal{L}_{\text{cluster}} + \lambda_s \mathcal{L}_{\text{reg}}, \tag{8}$$

where λ_c, λ_s control the relative loss weights.

4 Experimental Settings

We train our model for 1000 epochs on CIFAR-10/100 and 500 epochs on ImageNet-10. We use the AdamW optimizer (Loshchilov & Hutter, 2017) with a constant learning rate of 2e-4 for both generator and discriminator. We additionally add 0.1 weight decay to the discriminator. We use batch size 500 on CIFAR-10/100 and 320 on ImageNet-10. We run a small-scale parameter tuning experiment for hyperparameters and find that setting $\lambda_h = 4, \lambda_c = 3, \lambda_s = 5$ yields the best result. For simplicity, we run a single discriminator update before optimizing the generator, *i.e.,* $n_{dis} = 1$.

As a widely adopted GAN training trick, we maintain a momentum-updated discriminator and generator for evaluation purposes and find they produce stable data representations and better quality images. We also try producing $\boldsymbol{z}^b$ from momentum models for nearest neighbor searching, which slightly improves performance in all benchmarks. We set the memory bank size $|\boldsymbol{z}^m|$ = 10240, which is smaller than all datasets, preventing the model from accidentally picking features from augmented versions of the input image. Appendix A.1 provides more model configuration details.

Method	Parameters (M)	CIFAR-10		CIFAR-100		ImageNet-10	
		SVM	K-M	SVM	K-M	SVM	K-M
Supervised	11.5	95.1	95.1	75.9	73.6	96.4	96.3
Random	11.5	42.9	22.0	18.3	8.9	48.2	28.3
DINO (Caron et al., 2021)	11.5	89.7	63.9	65.6	36.7	87.8	68.0
NNCLR (Dwibedi et al., 2021)	11.5	91.7	69.3	69.7	40.4	**91.4**	66.8
SimCLR (Chen et al., 2020)	11.5	90.6	75.3	65.6	41.3	89.0	65.7
BYOL (Grill et al., 2020)	11.5	**93.1**	75.0	**70.6**	**42.8**	90.4	67.3
SWAV (Caron et al., 2020)	11.5	89.1	64.5	65.0	35.2	90.0	61.9
MAE (He et al., 2022)	20.4	82.3	37.0	57.1	17.9	88.4	45.8
DDPM (Ho et al., 2020)	41.8	91.1	78.0	62.5	36.3	-	-
Ours	11.5	89.8	**80.1**	63.3	38.2	91.2	**75.4**

Table 1: *Representation Learning Performance.* We evaluate our trained discriminator by benchmarking its learned representation using linear SVM and K-Means clustering (K-M), reporting average accuracy over 20 runs. Our method, which does not leverage any augmented views, achieves competitive performance with self-supervised approaches across multiple datasets. Compared to denoising autoencoders (shown in the penultimate and antepenultimate rows), our method excels in learning more effective representations while utilizing fewer parameters.

5 Results And Discussion

5.1 Synthetic Data

For illustrative purposes, we first train a GAN using our structural objectives on the synthetic double spirals dataset (Li et al., 2022). Here, we implement discriminator and generator as multi-layer perceptrons and keep all other configuration, *e.g.,* normalization layers, activation functions, objectives, and learning rate, consistent with our settings for experiments on real images.

Figure 2a demonstrates that the generated samples capture all data modes, with few outlier samples between spirals. Besides generation capability, we also visually inspect the discriminator's learned representations using t-SNE (Van der Maaten & Hinton, 2008). Figure 2c shows embeddings of the two categories are substantially separated. Figure 2d colors each data point by projecting its learned representation into a 2d palette map. From this plot, we see that the learned embedding preserves semantic structure within and across groups. Figure 2e shows the gradient of embedding distance approximates the optimal paths between uniform grids and data samples, indicating D_θ learns intrinsic data structure. Figure 2f demonstrates the capability of our structural objectives to learn semantic features: when the embedding is updated via $\mathcal{L}$, data that are semantically similar are updated to align in the same directions, whereas data from different clusters diverge.

5.2 Representation Learning on Real Images

We task the backbone of the discriminator to produce a vector as a data representation and then evaluate its performance on the task of image classification. We compare the results with state-of-the-art contrastive learning approaches under two widely adopted evaluation metrics:

- *Linear Support Vector Machine (SVM)*: We optimize a Linear SVM on top of training feature and report the accuracy on the validation set.
- *K-Means clustering*: We run spherical K-means clustering on the validation set, with K equaling the number of ground-truth categories. We then obtain a prediction on the validation set by solving the optimal assignment problem between the partition produced by clustering and the ground-truth categories. To reduce the randomness in clustering, we repeat this process 20 times and report average performance.

Table 1 reports results and provides comparison with current state-of-the-art methods. For datasets with fewer categories, *i.e.,* CIFAR-10 and ImageNet-10, our method significantly outperforms all contrastive learning approaches on the K-means clustering metric. On CIFAR-10, we achieve 80.1% test accuracy, surpassing the best-competing method, SimCLR, which achieves 75.3% test accuracy. On ImageNet-10, our method reaches 75.4% test accuracy surpasses the best-competed method,

Data Aug	Method	CIFAR-10 KMeans	CIFAR-10 SVM	CIFAR-10 LP	CIFAR-10 1/5 KNN	CIFAR-100 KMeans	CIFAR-100 SVM	CIFAR-100 LP	CIFAR-100 1/5 KNN
None	SimCLR	14.2	21.8	21.8	14.7/15.1	2.1	4.5	5.4	2.7/2.5
None	Ours	**76.5**	**84.5**	**83.2**	**79.7/82.2**	**22.5**	**52.2**	**51.6**	**37.5 / 37.4**
F	SimCLR	17.5	32.8	32.1	23.3 / 25.5	4.1	10.3	10.9	6.0/5.5
F	Ours	**76.2**	**85.7**	**85.8**	**80.9 / 84.5**	**30.8**	**52.0**	**56.5**	**41.0 / 42.9**
$F + C$	SimCLR	27.8	72.7	72.2	64.9 / 67.2	12.2	35.1	34.3	30.8 / 29.1
$F + C$	Ours	**80.0**	**89.3**	**88.4**	**87.7 / 89.2**	**37.4**	**63.2**	**62.0**	**55.0 / 56.1**
$F + C + J$	SimCLR	78.0	90.7	90.2	88.1 / 89.5	41.7	65.2	65.2	59.2 / 61.4

Table 2: *Data Augmentation Dependence.* We compare with SimCLR (Chen et al., 2020) on sensitivity to various data augmentation schemes. In our system, data augmentation is solely employed to enlarge the training dataset; it is not used for achieving view-consistency objectives. F, C, J denote random horizontal flipping, random image cropping, and color jittering, respectively; None means no augmentation is applied during training. For each augmentation scheme, our method outperforms SimCLR across all evaluation metrics (here, LP denotes linear probing). Moreover, we are able to operate even without data augmentation – a regime in which SimCLR fails.

DINO, with 68.0% test accuracy. When evaluating learned representations using linear SVM, our method reaches 89.8% test accuracy, which exceeds SWAV and DINO, with 89.1% and 89.7% test accuracy respectively, but falls slightly behind BYOL (93.1%) and NNCLR (91.7% test accuracy). On ImageNet10, our method's 91.2% accuracy approaches that of the best method (NNCLR with 91.4%) and exceeds the rest.

CIFAR-100 contains fewer training samples per category and operationalizing instance-wise discriminating objectives is thus favorable over clustering objectives or smoothness regularization. Under such case, our method remains competitive on the linear SVM metric, achieving 63.3% test accuracy, which is very close to DINO, SimCLR, and SWAV, which each have around 65% test accuracy. Using K-Means clustering, our method reaches 38.2% test accuracy, outperforming clustering-based contrastive approaches SWAV (35.2%) and DINO (36.7%).

Our quantitative comparison is also qualitatively confirmed by visualizing embeddings using t-SNE. BYOL, as shown in Figure 1, produces isolated and smaller-sized clusters that maintain sufficient space to discern categories under linear transformation. However, those clusters lack sufficient global organization, which is a quality evaluated by the K-Means clustering metric.

In contrast, our approach, as also shown in Figure 1, produces smoother embeddings, which are nearly aligned with the ground-truth partition, and consequently yields good K-Means clustering performance. When compared to denoising autoencoders such as DDPM (Ho et al., 2020) and MAE (He et al., 2022), our model demonstrates superior efficiency by utilizing fewer parameters (11.5M) compared to DDPM (41.8M) and MAE (20.4M). Additionally, our model excels in learning better representations across all evaluated metrics, with the sole exception being a comparison to DDPM on CIFAR-10 (our 89.7% accuracy using SVM *vs.* DDPM's 91.1%).

Method	NMI	Purity
Self-cond GAN	33.26	11.73
Ours	**72.77**	**81.52**

Table 3: *Comparison to Self-conditioned GAN (Liu et al., 2020) on CIFAR-10.* On normalized mutual information (NMI) and purity metrics, our method outperforms self-conditioned GAN (Liu et al., 2020), a generative model which clusters discriminator features iteratively in a self-discovering fashion.

5.3 Ablation Experiments

Sensitivity to data augmentation. Though we adopt some minimal data augmentation in our experiments, our approach is far less sensitive to data augmentation. However, contrastive self-supervised learning approaches, including SimCLR (Chen et al., 2020), require a carefully calibrated augmentation scheme to achieve good performance. Table 2 highlights this discrepancy. Our method demonstrates a clear advantage over SimCLR across all augmentation regimes, and, unlike SimCLR, can still learn useful features when no augmentation applied.

Method / Loss	D regularizer	Parameters (M) D	Parameters (M) G	IS ↑	FID ↓	K-Means	SVM
StyleGAN2-ADA	Grad Penalty	20.7	19.9	**9.82**	**3.60**	28.96	76.50
BigGAN	Spectral Norm	4.2	4.3	8.22	17.50	29.69	69.31
Hinge Loss	$\mathcal{L}_{\text{reg}}$	11.5	4.9	8.13	18.54	36.41	77.19
Eq. 2 only, D_B	$\mathcal{L}_{\text{reg}}$	11.5	4.9	8.39	17.83	70.76	87.9
Eq. 2 only, JSD	$\mathcal{L}_{\text{reg}}$	11.5	4.9	8.55	16.97	**80.55**	88.32
Full Objectives	Spectral Norm	11.5	4.9	7.23	26.41	55.38	83.9
Full Objectives	$\mathcal{L}_{\text{reg}}$	11.5	4.9	8.73	13.63	80.11	**89.76**

Table 4: *Ablation over Loss Function Components on CIFAR-10.* We compare StyleGAN2-ADA (Karras et al., 2020a), BiGAN (Brock et al., 2019) and a GAN baseline using the standard hinge loss to models using ablated variants of our structural objectives. Dicriminators trained using our objectives significantly outperform these baselines (K-Means, SVM metrics), while our corresponding generators also benefit (IS, FID). Including our finer scale clustering objective (last row) improves both representation and image quality over ablated variants using only our coarse scale objective (rows 4 & 5). The benefit observed when using $\mathcal{L}_{\text{reg}}$ over Spectral norm (final to penultimate row) indicates that preserving model capacity is crucial for effective feature learning.

Comparison to other generative feature learners. GenRep (Jahanian et al., 2021) generates images pairs by sampling adjacent features in the latent space of BigBiGAN (Donahue & Simonyan, 2019) and then trains an encoder to optimize contrastive objectives. To compare with GenRep, we train our model on ImageNet-100, following most of the our settings for ImageNet-10, except we extend training to 1000 epochs. For fair comparisons, we utilize their *Tz only* version, a setting where no data augmentation is used, and show the results in Table 5. Our method outperforms GenRep, though we adopt a simpler network architecture for feature learning.

Method	Network	Linear Probing
GenRep(Tz only)	ResNet-50	55.0*
Ours	ResNet-18	**59.9**

Table 5: *ImageNet-100.* Our method outperforms GenRep (Jahanian et al., 2021) though we adopt a simpler network architecture and a more direct training pipeline. *For a fair comparison, this result is from Figure 6 of GenRep (Jahanian et al., 2021), which does not use data augmentation (Tz only).

Self-conditioned GAN (Liu et al., 2020) clusters the discriminator's features iteratively in a self-discovering fashion; cluster information is fed into the GAN pipeline as conditional input. Though this method produces clustering during training, its objective differs entirely from ours: their motivation is to improve the diversity of image generation, rather than learn representations. Table 3 shows that our method outperforms it.

Ablation of system variants. Table 4 provides a quantitative comparison of both generator and discriminator performance across baselines as well as ablated and full variants of our system. Our proposed objectives significantly improve generation and representation quality over the hinge loss baseline. We witness further enhancement in image quality when using our extra instance/clustering-wise objective. Performance drops by replacing $\mathcal{L}_{\text{reg}}$ with spectral norm, indicating the effectiveness of our suggested regularization scheme in preserving model capacity. As an additional advantage over spectral norm, we observed better training stability when using our regularization scheme. Note that while StyleGAN2-ADA achieves state-of-the-art generation quality, it both requires adopting a larger network to do so, and still performs worse at feature learning than our system. Appendix A.3 provides a qualitative comparison with examples of generated images.

6 Conclusion

Our structural adversarial objectives augment the GAN framework for self-supervised representation learning, shaping the discriminator's output at two levels of granularity: aligning features via mean and variance at coarser scale and grouping features to form local clusters at finer scale. Benchmarks across multiple datasets show that training a GAN with these novel objectives suffices to produce data representations competitive with the state-of-the-art self-supervised learning approaches, while also improving the quality of generated images.

REFERENCES

Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In *International conference on machine learning*, pp. 214–223. PMLR, 2017.

Hangbo Bao, Li Dong, and Furu Wei. Beit: Bert pre-training of image transformers. *arXiv preprint arXiv:2106.08254*, 2021.

Vineeth S Bhaskara, Tristan Aumentado-Armstrong, Allan D Jepson, and Alex Levinshtein. Gran-gan: Piecewise gradient normalization for generative adversarial networks. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 3821–3830, 2022.

Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale GAN training for high fidelity natural image synthesis. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=B1xsqj09Fm.

Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. *Advances in Neural Information Processing Systems*, 33:9912–9924, 2020.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9650–9660, 2021.

Arantxa Casanova, Marlene Careil, Jakob Verbeek, Michal Drozdzal, and Adriana Romero Soriano. Instance-conditioned gan. *Advances in Neural Information Processing Systems*, 34:27517–27529, 2021.

Jianlong Chang, Lingfeng Wang, Gaofeng Meng, Shiming Xiang, and Chunhong Pan. Deep adaptive image clustering. In *Proceedings of the IEEE international conference on computer vision*, pp. 5879–5887, 2017.

Qifeng Chen and Vladlen Koltun. Photographic image synthesis with cascaded refinement networks. In *Proceedings of the IEEE international conference on computer vision*, pp. 1511–1520, 2017.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pp. 1597–1607. PMLR, 2020.

Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15750–15758, 2021.

Djork-Arné Clevert, Thomas Unterthiner, and Sepp Hochreiter. Fast and accurate deep network learning by exponential linear units (elus). *arXiv preprint arXiv:1511.07289*, 2015.

Victor Guilherme Turrisi da Costa, Enrico Fini, Moin Nabi, Nicu Sebe, and Elisa Ricci. solo-learn: A library of self-supervised methods for visual representation learning. *Journal of Machine Learning Research*, 23 (56):1–6, 2022. URL http://jmlr.org/papers/v23/21-1155.html.

Xili Dai, Shengbang Tong, Mingyang Li, Ziyang Wu, Kwan Ho Ryan Chan, Pengyuan Zhai, Yaodong Yu, Michael Psenka, Xiaojun Yuan, Heung-Yeung Shum, and Yi Ma. Closed-loop data transcription to an LDR via minimaxing rate reduction, 2022. URL https://openreview.net/forum?id=s5lIqsrOu3Z.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255, 2009. doi: 10.1109/CVPR.2009.5206848.

Carl Doersch, Abhinav Gupta, and Alexei A Efros. Unsupervised visual representation learning by context prediction. In *CVPR*, 2015.

Jeff Donahue and Karen Simonyan. Large scale adversarial representation learning. *Advances in neural information processing systems*, 32, 2019.

Jeff Donahue, Philipp Krähenbühl, and Trevor Darrell. Adversarial feature learning. In *International Conference on Learning Representations*, 2017. URL https://openreview.net/forum?id=BJtNZAFgg.

Yilun Du, Shuang Li, Yash Sharma, Josh Tenenbaum, and Igor Mordatch. Unsupervised learning of compositional energy concepts. *Advances in Neural Information Processing Systems*, 34:15608–15620, 2021.

Vincent Dumoulin, Ishmael Belghazi, Ben Poole, Alex Lamb, Martin Arjovsky, Olivier Mastropietro, and Aaron Courville. Adversarially learned inference. In *International Conference on Learning Representations*, 2017. URL https://openreview.net/forum?id=B1ElR4cgg.

Debidatta Dwibedi, Yusuf Aytar, Jonathan Tompson, Pierre Sermanet, and Andrew Zisserman. With a little help from my friends: Nearest-neighbor contrastive learning of visual representations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9588–9597, 2021.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems*, 2014.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11): 139–144, 2020.

Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent-a new approach to self-supervised learning. *Advances in neural information processing systems*, 33: 21271–21284, 2020.

Ronja Güldenring and Lazaros Nalpantidis. Self-supervised contrastive learning on agricultural images. *Computers and Electronics in Agriculture*, 191:106510, 2021.

Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron C Courville. Improved training of wasserstein gans. *Advances in neural information processing systems*, 30, 2017.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9729–9738, 2020.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16000–16009, 2022.

John R Hershey and Peder A Olsen. Approximating the kullback leibler divergence between gaussian mixture models. In *2007 IEEE International Conference on Acoustics, Speech and Signal Processing-ICASSP'07*, volume 4, pp. IV–317. IEEE, 2007.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017.

Tobias Hinz, Stefan Heinrich, and Stefan Wermter. Semantic object accuracy for generative text-to-image synthesis. *IEEE transactions on pattern analysis and machine intelligence*, 2020.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International conference on machine learning*, pp. 448–456. PMLR, 2015.

Ali Jahanian, Xavier Puig, Yonglong Tian, and Phillip Isola. Generative models as a data source for multiview representation learning. *arXiv preprint arXiv:2106.05258*, 2021.

Ruoxi Jiang and Rebecca Willett. Embed and emulate: Learning to estimate parameters of dynamical systems with uncertainty quantification. *Advances in Neural Information Processing Systems*, 35:11918–11933, 2022.

Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. *arXiv preprint arXiv:1710.10196*, 2017.

Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 4401–4410, 2019.

Tero Karras, Miika Aittala, Janne Hellsten, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Training generative adversarial networks with limited data. *Advances in neural information processing systems*, 33:12104–12114, 2020a.

Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 8110–8119, 2020b.

Tero Karras, Miika Aittala, Samuli Laine, Erik Härkönen, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Alias-free generative adversarial networks. *Advances in Neural Information Processing Systems*, 34:852–863, 2021.

Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

Rayan Krishnan, Pranav Rajpurkar, and Eric J Topol. Self-supervised learning in medicine and healthcare. *Nature Biomedical Engineering*, 6(12):1346–1352, 2022.

Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

Gustav Larsson, Michael Maire, and Gregory Shakhnarovich. Colorization as a proxy task for visual understanding. In *CVPR*, 2017.

Zengyi Li, Yubei Chen, Yann LeCun, and Friedrich T Sommer. Neural manifold clustering and embedding. *arXiv preprint arXiv:2201.10000*, 2022.

Steven Liu, Tongzhou Wang, David Bau, Jun-Yan Zhu, and Antonio Torralba. Diverse image generation via self-conditioned gans. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 14286–14295, 2020.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

Xuezhe Ma, Xiang Kong, Shanghang Zhang, and Eduard H Hovy. Decoupling global and local representations via invertible generative flows. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=iWLByfvUhN.

Takeru Miyato, Toshiki Kataoka, Masanori Koyama, and Yuichi Yoshida. Spectral normalization for generative adversarial networks. *arXiv preprint arXiv:1802.05957*, 2018.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

Deepak Pathak, Philipp Krahenbuhl, Jeff Donahue, Trevor Darrell, and Alexei A Efros. Context encoders: Feature learning by inpainting. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2536–2544, 2016.

Konpat Preechakul, Nattanat Chatthee, Suttisak Wizadwongsa, and Supasorn Suwajanakorn. Diffusion autoencoders: Toward a meaningful and decodable representation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10619–10629, 2022.

Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*, 2015.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, 2021.

Tim Salimans, Han Zhang, Alec Radford, and Dimitris Metaxas. Improving gans using optimal transport. *arXiv preprint arXiv:1803.05573*, 2018.

Axel Sauer, Katja Schwarz, and Andreas Geiger. Stylegan-xl: Scaling stylegan to large diverse datasets. In *ACM SIGGRAPH 2022 Conference Proceedings*, pp. 1–10, 2022.

Akinori Tanaka. Discriminator optimal transport. *Advances in Neural Information Processing Systems*, 32, 2019.

Yonglong Tian, Chen Sun, Ben Poole, Dilip Krishnan, Cordelia Schmid, and Phillip Isola. What makes for good views for contrastive learning? *Advances in Neural Information Processing Systems*, 33:6827–6839, 2020.

Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.

Feng Wang, Tao Kong, Rufeng Zhang, Huaping Liu, and Hang Li. Self-supervised learning by estimating twin class distributions. *arXiv preprint arXiv:2110.07402*, 2021.

Yi-Lun Wu, Hong-Han Shuai, Zhi-Rui Tam, and Hong-Yu Chiu. Gradient normalization for generative adversarial networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 6373–6382, 2021.

Yuxin Wu and Kaiming He. Group normalization. In *Proceedings of the European conference on computer vision (ECCV)*, pp. 3–19, 2018.

Zhirong Wu, Yuanjun Xiong, Stella X Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance discrimination. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3733–3742, 2018.

Weilai Xiang, Hongyu Yang, Di Huang, and Yunhong Wang. Denoising diffusion autoencoders are unified self-supervised learners. *arXiv preprint arXiv:2303.09769*, 2023.

Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stéphane Deny. Barlow twins: Self-supervised learning via redundancy reduction. In *International Conference on Machine Learning*. PMLR, 2021.

Han Zhang, Jing Yu Koh, Jason Baldridge, Honglak Lee, and Yinfei Yang. Cross-modal contrastive learning for text-to-image generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 833–842, 2021.

Mingtian Zhang, Tim Z Xiao, Brooks Paige, and David Barber. Improving vae-based representation learning. *arXiv preprint arXiv:2205.14539*, 2022.

Richard Zhang, Phillip Isola, and Alexei A Efros. Colorful image colorization. In *ECCV*, 2016.

Xiao Zhang and Michael Maire. Self-supervised visual representation learning from hierarchical grouping. *Advances in Neural Information Processing Systems*, 33:16579–16590, 2020.

Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pp. 2223–2232, 2017.

A APPENDIX

A.1 DETAILS OF DATASET AND MODEL

Datasets. We focus on three benchmark datasets: CIFAR-10, CIFAR-100 (Krizhevsky et al., 2009) and ImageNet-10.

ImageNet-10: We follow Chang et al. (2017) to select 10 categories from the ImageNet dataset (Deng et al., 2009), resulting in 13,000 training images and 500 validation images. During training, we only perform spatial augmentation, including random spatial cropping and horizontal flipping, followed by resizing images to 128x128 resolution to match the generated images. During testing, we resize the images to align the smaller edge to 144 pixels, followed by central cropping to produce a 128x128 output.

CIFAR-10/100: During training, we apply the same augmentation strategy as in ImageNet-10 but produce 32x32 images. During testing, we do not perform cropping.

For compared methods, we keep their default augmentation strategy. On ImageNet-10, we resized their augmented images to 128x128. For all methods, we learn in an unsupervised manner on the training split and evaluate on the validation split.

In CIFAR-10/100 experiments, we use default configurations from SOLO-Learns (da Costa et al., 2022), an open source library providing heavily tuned configurations for multiple state-of-the-art self-supervised methods. In ImageNet-10 experiments, we train competing approaches using the suggested hyperparameters for ImageNet-100, but extend the total epochs to 1000 for sufficient convergence. For fair comparison, we run these methods with our modified backbone and resize input images to 128x128.

Model details: discriminator. We construct our discriminator using ResNet-18 (He et al., 2016) and perform several modifications to make it cooperate reasonably with the generator. Inspired by the discriminator configuration in BigGAN (Brock et al., 2019), we perform spatial reduction only within the residual block and replace all stride two convolution layers with average pooling followed by stride one convolution. We remove the first max-pooling layer and switch the first convolution layer to a 3x3 kernel with a 1x1 stride to keep the resolution unchanged before the residual block.

To maintain a substantial downsample rate in ImageNet-10 images, we duplicate the first residual block and enable a spatial reduction in all blocks to reach a 32x downsampling. On CIFAR-10/100, we preserve the default setting for residual blocks. As our proposed smoothness term regularizes each sample, we replace all BatchNorm layers (Ioffe & Szegedy, 2015) with GroupNorm (Wu & He, 2018), specifying 16 channels as a single group; this prevents batch-wise interaction. We also remove the first normalization layer in each block, as doing so produces better results. We replace ReLU with ELU (Clevert et al., 2015) activations for broader non-linear support on negative values.

Model details: generator. We adapt the generator configuration from BigGAN-deep(Brock et al., 2019). Specifically, we take their model for 32x32 images on CIFAR, and additionally increase the base channels to 128 to prevent image generation from being the system bottleneck. For ImageNet-10, we replicate their settings for 128x128 images.

A.2 COMPARED SELF-SUPERVISED LEARNING METHODS

We evaluate the representations produced by our method in comparison to those produced by the following state-of-the-art self-supervised learning methods:

- SimCLR (Chen et al., 2020) optimizes the InfoNCE loss, maximizing feature similarity across views while repulsing all the images.
- NNCLR (Dwibedi et al., 2021) samples nearest neighbors from the data set using cross-view features and treats them as positives for InfoNCE objectives. We additionally run a baseline, denoted NNCLR (same views) in Figure 1, by removing the augmented view and directly maximizing the similarity between image features and their nearest neighbor.

- SWAV (Caron et al., 2020) maximizes view consistent objectives using clustering-based targets; it balances the categorical assignment using sinkhorn iterations.
- DINO (Caron et al., 2021) optimizes clustering-based across-views objectives via knowledge distillation and proposes sharpening and centering techniques to prevent collapsing.
- BYOL (Grill et al., 2020) only contains the maximizing term and adopts a momentum-updated Siamese model to process augmented input to prevent collapsed solutions.

In MAE (He et al., 2022), we employ a VIT-small model, training it with default masking ratio and a patch-size of 4 for CIFAR experiments and 8 for ImageNet-10 experiments.

In DDPM (Ho et al., 2020), we use unconditional model and train it with default hyper-parameters. Feature are extracted from the second decoder block with noise level at t = 11, following the optimal configurations of Xiang et al. (2023).

A.3 Qualitative Comparison

We provide visualization of generated images for the following configurations:

- Figure 3: Results of training with our full objectives (our method):

$$\mathcal{L}^{\text{Full}} := \mathcal{L}_{\text{Gaussian}} + \lambda_c \mathcal{L}_{\text{cluster}} + \lambda_s \mathcal{L}_{\text{reg}}.$$

- Figure 4: Results of training with Equation 2 only, JSD:

$$\mathcal{L}^{\text{JSD}} := \mathcal{L}_{\text{Gaussian}} + \lambda_s \mathcal{L}_{\text{reg}}.$$

- Figure 5: Results of training with Hinge Loss.
  To train with Hinge loss, we change discriminator to output a scalar: $D_\theta(\boldsymbol{x}) \in \mathbb{R}$ and optimize the hinge loss defined as follows:

$$\begin{aligned}
\mathcal{L}^{\text{Hinge}} &:= \mathcal{L}^{\text{Hinge}}_{D_\theta} + \mathcal{L}^{\text{Hinge}}_{G_\phi} + \lambda_s \mathcal{L}_{\text{reg}}, \\
\mathcal{L}^{\text{Hinge}}_{D_\theta} &:= \max_{D_\theta} \left(\min\left(0, -1 + D_\theta\left(\boldsymbol{x}\right)\right) - \min\left(0, -1 - D_\theta\left(\hat{\boldsymbol{x}}\right)\right)\right), \\
\mathcal{L}^{\text{Hinge}}_{G_\phi} &:= \min_{G_\phi} -D_\theta(\hat{\boldsymbol{x}}).
\end{aligned}$$

- Figure 6: Results of BigGAN (Brock et al., 2019).

Conclusion. We observe that training with full objectives (our method) achieves the best quality and diversity in generated images.

Figure 3: Randomly generated images from GAN trained with our full objectives.

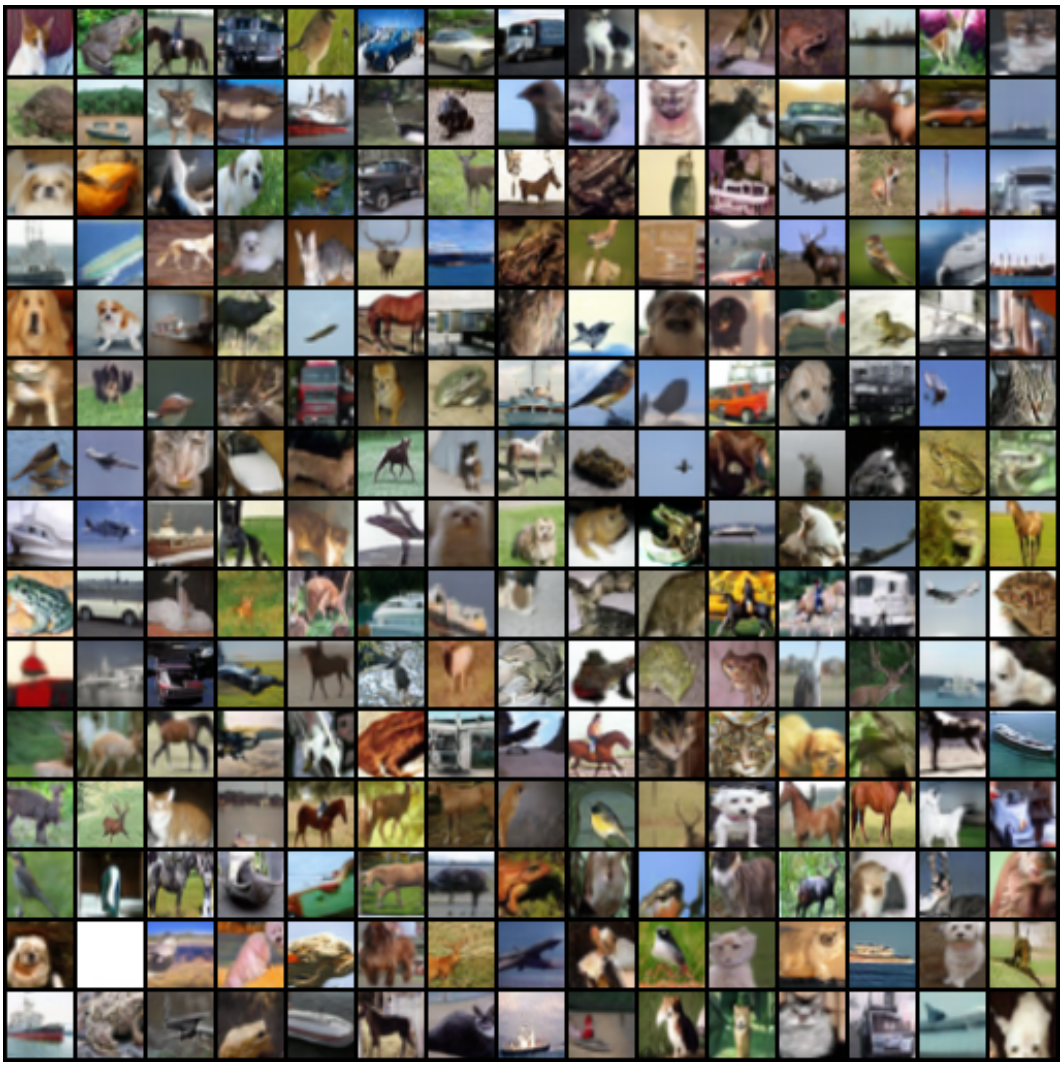


Figure 4: Randomly generated images from GAN trained with Eq. 2 only, JSD.

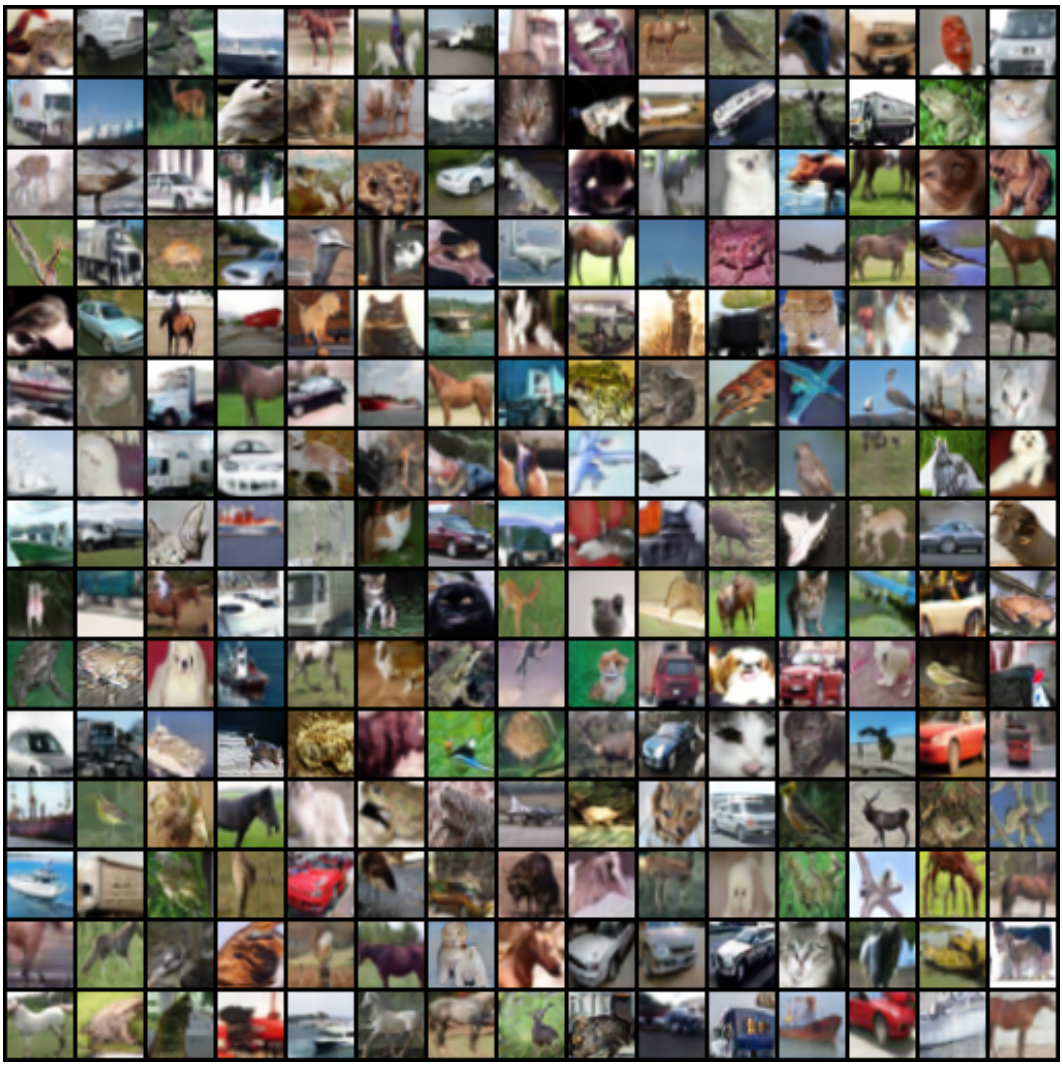


Figure 5: Randomly generated images from GAN trained with Hinge Loss.

Figure 6: Randomly generated images from unconditional BigGAN (Brock et al., 2019).