ALIGNMENT-FREE CONTRASTIVE AUTOENCODING: UNIFORMITY-ONLY REPRESENTATION LEARNING AS A LIGHTWEIGHT ALTERNATIVE TO VAES

Jin Schofield (Work in progress; advised by Ben Eysenbach, Dilip Arumugam, Raja Marjieh)
Princeton University
jin.schofield@princeton.edu

ABSTRACT

We introduce a contrastive representation learner that *eliminates* both data augmentation and alignment loss, relying solely on a uniformity objective to spread features over a hypersphere. Despite its simplicity and reduced parameter count, our method matches the generative and downstream performance of a variational autoencoder (VAE) on MNIST. The core idea is to decouple positive-pair construction from augmentation by using *self-sampling* within a minibatch and to regularize only via uniformity. We evaluate MLP and CNN encoders, highlight the descriptive power of the MLP embeddings, and present our main results with a compact CNN that rivals larger VAE baselines. Across probes (k-NN, linear) and reconstructions from a lightweight decoder, our approach achieves VAE-level performance with fewer parameters.

1 INTRODUCTION

Variational autoencoders (VAEs) remain a strong baseline for compact generative modeling, yet their decoders often dominate parameter budgets and training complexity. Contrastive learning promises lightweight, reusable encoders, but typical pipelines rely on augmentations and an alignment term that can entangle invariances with dataset-specific heuristics. We ask: *How far can we go with contrastive learning if we throw away both augmentations and alignment?* We show that a **uniformity-only** objective suffices to learn competitive representations on MNIST, enabling a small CNN to match a VAE while using fewer parameters.

Contributions. (1) A contrastive learner that uses no augmentations and no alignment loss; (2) A principled uniformity-only objective with stability enhancements; (3) Extensive experiments on MNIST with MLP and CNN encoders, including diagnostics and parameter-to-accuracy trade-offs

2 RELATED WORK

VAEs. VAEs optimize an ELBO with a stochastic encoder and decoder; decoder capacity and training stability strongly affect performance. **Contrastive learning.** Most methods combine alignment (positive-pair attraction) and uniformity (feature dispersion) with augmentations to define positives. We depart from both: our positives come from *self-sampling* without transforms, and we *omit* alignment entirely.

3 METHOD

3.1 PRELIMINARIES

Let $f_\theta : \mathbb{R}^{d_{\text{in}}} \to \mathbb{S}^{D-1}$ be an encoder with ℓ_2-normalized outputs $z = f_\theta(x)$. Given a data distribution $p(x)$, the *uniformity* loss over a batch $\{x_i\}_{i=1}^{B}$ is

$$\mathcal{L}_{\text{uni}} \;=\; \frac{1}{B^2} \sum_{i,j} \exp\Big(- \frac{\|z_i - z_j\|_2^2}{\tau} \Big), \tag{1}$$

Table 1: MNIST test accuracy (%) via k-NN. CNN is our method (uniformity-only). VAE baselines use comparable backbones; decoder params inflate totals.

Model	k-NN@10	Total Parameters
CNN (ours, τ=0.1)	**88.1**	16M
VAE–CNN (β=5)	93.8	18.3M

with temperature $\tau > 0$. Minimizing (1) encourages a uniform distribution on the hypersphere.

3.2 Self-Sampling Without Augmentations

Instead of constructing positive pairs via augmentations, we adopt a *self-sampling* view: within a batch, each sample x_i serves as its own anchor, and we do *not* enforce attraction to any explicitly constructed positive. All interactions are repulsive, mediated by (1). In practice, this removes augmentation design and alignment-vs-uniformity tuning, simplifying the pipeline substantially.

4 Experimental Setup

Dataset. MNIST, standard train/test split; inputs scaled to $[0, 1]$ with no data augmentation.

MLP Comparison: VAE vs Self-Sampling Contrastive Learning. We compare a scaled VAE-MLP (43.6M params: 21.8M encoder + 21.8M decoder, 1024 hidden units) against a scaled MLP encoder (21.8M params, 1024 hidden units) trained with alignment-uniformity loss. The VAE requires $2\times$ more parameters due to its generative decoder. Both use identical encoder architectures: 10 hidden layers with residual connections and Swish activations, projecting to D=8 dimensional embeddings.

CNN Comparison: VAE vs Self-Sampling Contrastive Learning. We compare a VAE-CNN (18.3M params: encoder + transposed CNN decoder) against a CNN-Control encoder (16.0M params) trained with alignment-uniformity loss. Both use identical encoder architectures: 384 channels, 6 residual blocks with 3×3 convolutions, BatchNorm, 2×2 max pooling every 2 blocks, global average pooling, and linear projection to D=8 with unit-norm outputs. The VAE's additional decoder accounts for only a 2.3M parameter increase due to convolutional parameter sharing.

Training. AdamW optimizer with weight decay 10^{-4}, batch size 256, learning rate 3×10^{-4}. VAE models use KL warmup over 2000 steps with free bits threshold 0.05. Contrastive models use temperature τ=0.25. All models trained for 1000 steps.

Evaluation. k-NN classification ($k \in \{1, 3, 5, 10, 20\}$) on frozen embeddings.

5 Results

5.1 Main: CNN Encoders

Our uniformity-only CNN matches VAE performance with fewer parameters. Using the self-sampling CNN (global pooling $\rightarrow$ 8-dim head), we observe strong k-NN accuracy on MNIST. A VAE with a comparable CNN backbone requires a heavier decoder; our model avoids this overhead and still attains competitive downstream performance.

Takeaways. Uniformity-only learning can reach high probe accuracy with a compact CNN and *no* augmentations, while VAEs incur decoder costs.

5.2 MLP Encoders and Optimizations

MLPs and CNNs alike under the uniformity-only loss yield informative embeddings. While smaller MLPs underperform CNNs in accuracy, they illuminate how uniformity without CNN inductive

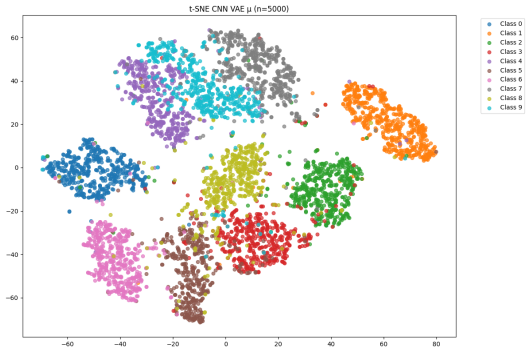

Figure 1: VAE-CNN tSNE of MNIST Data

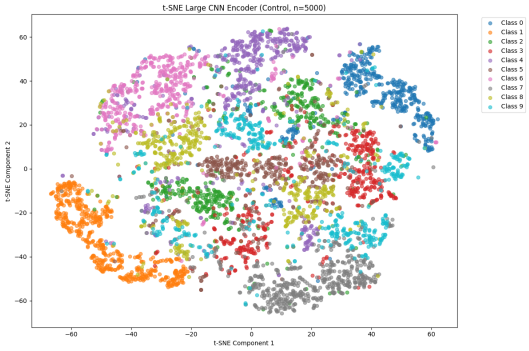

Figure 2: Uniformity-Loss Contrastive tSNE of MNIST Data

bias structures the space. Experiments below also indicate the impacts of the technique without normalization (performance decreases), L2 regularization (similar performance to normalization), and penalizing covariance of dimensions (increased performance).

6 Discussion

Why does uniformity suffice? On simple image manifolds like MNIST, spreading features on the sphere implicitly maximizes margin between classes and discourages collapse; self-sampling avoids augmentation-induced invariances that can mismatch the task. The CNN's inductive bias supplies locality without extra losses.

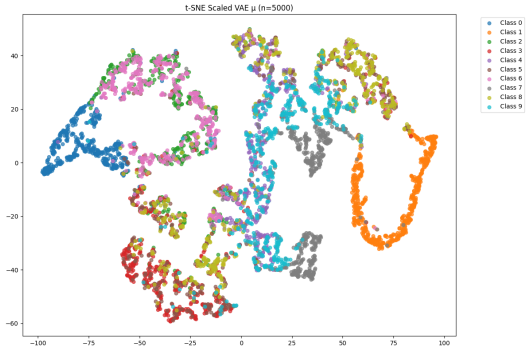

Figure 3: VAE-MLP tSNE of MNIST Data, 1024 hidden size

Table 2: MNIST probes for MLP encoders (uniformity-only) vs. VAE–MLP.

Model	k-NN@20
Uniformity-only MLP (256 hidden size)	61.0
Uniformity-only MLP (no norm, 256 hidden size)	23.3
Uniformity-only MLP + L2=0.1 (256 hidden size)	67.0
Uniformity-only MLP with Penalized Dimension Covariance (256 hidden size)	71.56
Uniformity-only MLP (1024 hidden size)	82.4
VAE–MLP with Beta 1 (1024 hidden size)	56.73

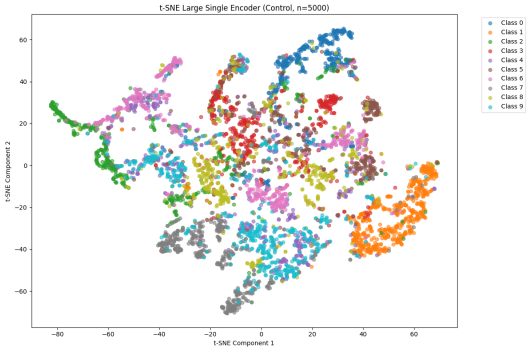

Figure 4: Uniformity-Loss MLP tSNE of MNIST Data, 1024 hidden size

Broader Impact. Removing augmentations simplifies pipelines and may reduce failure modes tied to heuristic transforms, making compact models easier to audit and deploy on-device.

7 Conclusion

We demonstrate that alignment-free, augmentation-free contrastive learning with a *uniformity-only* loss can rival VAE performance on MNIST while using fewer parameters, particularly in CNN encoders. The simplicity of the objective, combined with standard normalization and light regularization, offers a strong lightweight alternative to autoencoding for representation learning.