HYBRID MUTUAL INFORMATION LOWER-BOUND ESTIMATORS FOR REPRESENTATION LEARNING

Abhishek Sinha[1]* **Jiaming Song**[1]* **Stefano Ermon**[1]

Department of Computer Science[1]
Stanford University
{a7b23, tsong, ermon}@stanford.edu

ABSTRACT

Self-supervised representation learning methods based on the principle of maximizing mutual information have been successful in unsupervised learning of visual representations. These approaches are low-variance mutual information lower bound estimators, yet the lack of distributional assumptions prevent them from learning certain important information such as texture. Estimators that are based on distributional assumptions bypass this issue with autoencoders but they tend to have worse performance on downstream classification. To this end, we consider a hybrid approach that incorporates both the distribution-free contrastive lower bound and the distribution-based autoencoder lower bound. We illustrate that with one set of representations, the hybrid approach is able to achieve good performance on multiple downstream tasks such as classification, reconstruction, and generation.

1 INTRODUCTION

Unsupervised learning of representations from data is critical to machine learning, since it allows the use of large amounts of unlabeled data to boost performance on downstream tasks and alleviate the use of labels (Mikolov et al., 2013; Devlin et al., 2018; Mnih & Kavukcuoglu, 2013; Radford et al., 2018; Brown et al., 2020). In terms of learning visual representations where the observations are continuous, self-supervised contrastive methods (van den Oord et al., 2018; Belghazi et al., 2018; Devon Hjelm et al., 2018; Poole et al., 2019; Tian et al., 2019) have been more successful than alternatives that consider jigsaw or rotation as self-supervised pre-text tasks. One explanation is that contrastive methods are based on variational mutual information (MI) estimation, so in principle, such representations are encouraged to maximize mutual information with the data.

However, there are fundamental limits to lower bound estimators of mutual information without distributional assumptions. Song & Ermon (2019) have illustrated that unbiased mutual information lower bound estimators can have variances that grows exponentially when measured with samples, and McAllester & Stratos (2020) have shown that any such estimator that reliably produces a lower bound to mutual information with M samples will have values reliably capped at $O(\log M)$.

Most successful contrastive approaches can be treated as an instance of the latter situation where bias is introduced to prevent exponentially large variance. This means that in principle, we are not encouraged to learn more than $\log M$ bits of information of the data with an objective of M samples (positive and negative in the context of contrastive learning). This means that for low-dimensional representations, there are potentially a lot of important information (such as details about texture) that are discarded in favor of better downstream task performance (such as classification). While these representations may perform favorably in tasks such as classification, they may not perform as well in other tasks such as reconstruction or interpolation.

An alternative approach is to bypass these limitations by introducing distributional assumptions. Given that we are trying to maximize mutual information with images, we may leverage inductive

[*] Equal Contribution

biases in deep neural networks, such as convolutional neural networks, to help overcome the fundamental limitations and learn more "informative" representations, *e.g.*, with autoencoders (Kingma & Welling, 2013). Unfortunately, representations learned with autoencoding structures tend to perform worse than ones learned with contrastive methods.

Therefore, we investigate a hybrid approach, where we utilize one estimator without distributional assumptions (such as a contrastive one) and another with distributional assumptions (such as an autoencoder). In principle, this allows us to bring the best of both worlds in terms of estimating mutual information; in practice, we can obtain one set of representations that performs well for multiple downstream tasks including classification and reconstruction.

2 Background and Related Work

The mutual information between two random variables X and Y is the following KL divergence:

$$I(X;Y) = D_{\mathrm{KL}}(P(X,Y) \| P(X)P(Y)) \tag{1}$$

which we wish to estimate using samples from $P(X,Y)$; in certain cases we may know the density of marginals (e.g. $P(X)$). There are a wide range of variational approaches to variational MI estimation. The following lower bound estimator (Barber & Agakov, 2003) integrates distributional assumptions over the conditional distribution $p(\boldsymbol{x}|\boldsymbol{y})$:

$$I_{\mathrm{BA}}(q_\phi) := \mathbb{E}_{P(X,Y)}\left[\log q_\phi(\boldsymbol{x}|\boldsymbol{y}) - \log p(\boldsymbol{x})\right] \tag{2}$$

where $q_\phi : \mathcal{Y} \to \mathcal{P}(\mathcal{X})$ is a valid conditional distribution over $\mathcal{X}$ given $\boldsymbol{y} \in \mathcal{Y}$ and $p(\boldsymbol{x})$ is the probability density function of the marginal distribution $P(X)$. A typical way to implement this is via an autoencoder (Zhao et al., 2018; Song et al., 2018).

Contrastive methods, on the other hand, do not rely on distributional assumptions. For a batch of n positive pairs $\{(\boldsymbol{x}_i, \boldsymbol{y}_i)\}_{i=1}^n$, the contrastive predictive coding (CPC) objective (van den Oord et al., 2018) is defined as:

$$I_{\mathrm{CPC}}(g) := \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} \log \frac{m \cdot g(\boldsymbol{x}_i, \boldsymbol{y}_i)}{g(\boldsymbol{x}_i, \boldsymbol{y}_i) + \sum_{j=1}^{m-1} g(\boldsymbol{x}_i, \overline{\boldsymbol{y}_{i,j}})}\right] \tag{3}$$

Subsequent works (Poole et al., 2019) have shown that CPC is a lower bound to mutual information.

3 A Hybrid Mutual Information Lower Bound for Representation Learning

Let $\boldsymbol{x}$ denote the unsupervised observations that we intend to learn representations from, and let $\boldsymbol{z} = f_\theta(\boldsymbol{x})$ denote the learned representations that is processed via a neural network with parameters θ. We consider splitting $\boldsymbol{z}$ into two separate components: $\boldsymbol{z} = [\boldsymbol{y}, \boldsymbol{w}]$, where we use the distribution-free lower bound on $\boldsymbol{y}$ and the distribution-based lower bound on the entire $\boldsymbol{y}$.

Concretely, our hybrid objective function then becomes the following, where we optimize $g : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ and $q : \mathcal{Z} \to \mathcal{P}(\mathcal{X})$ as deep neural networks and $\alpha \in (0, 1)$ is a weight that balances the two objectives:

$$I(g,q) = \alpha\mathbb{E}\left[\log q(\boldsymbol{x}|\boldsymbol{z}) - \log p(\boldsymbol{x})\right] + (1-\alpha)\,\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} \log \frac{m \cdot g(\boldsymbol{x}_i, \boldsymbol{y}_i)}{g(\boldsymbol{x}_i, \boldsymbol{y}_i) + \sum_{j=1}^{m-1} g(\boldsymbol{x}_i, \overline{\boldsymbol{y}_{i,j}})}\right]$$

It is straightforward to show that $I(g,q)$ is a lower bound to mutual information (from the fact that each component is a lower bound and that mutual information between $\boldsymbol{x}$ and $\boldsymbol{z}$ is larger than that between $\boldsymbol{x}$ and $\boldsymbol{y}$), so maximizing it amounts to maximizing a lower bound to mutual information in principle. We empirically observe that using the entirely of $\boldsymbol{z}$ in contrastive learning would lead to worse empirical performance in reconstruction due to the loss of information in data augmentation.

In practice, our hybrid method can be illustrated as in Figure 1, where we jointly train an autoencoder (from $\boldsymbol{x}$ to $\boldsymbol{z}$ to $\boldsymbol{x}'$) with a contrastive component ($\boldsymbol{x}$ to $\boldsymbol{y}$ to $g(\boldsymbol{x}, \boldsymbol{y})$). The component for performing contrastive learning with $\boldsymbol{y}$ allows the use of flexible data augmentation techniques to be integrated into the system, whereas the component for autoencoding with $\boldsymbol{z}$ ensures that the learned representations are encouraged to make small errors in reconstruction thus retaining the information that are potentially lost from data augmentation (such as colors). We use the Moco-v2 method (Chen et al., 2020) on contrastive learning and the squared loss for autoencoding (which assumes that $q(\boldsymbol{x}|\boldsymbol{z})$ is Gaussian) in all our experiments.

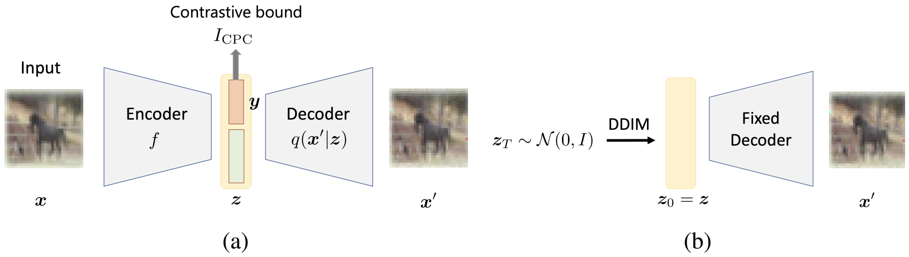


Figure 1: The pipeline for our hybrid approach. a) We apply contrastive learning on top of latent representations of an Auto-Encoder. We only use the top half of the representations for contrastive learning. b) Once trained, we train a DDIM model over the entire latent space for efficient sampling.

Once the representation model $f(\boldsymbol{x})$ is trained, we can learn a generative model by first freezing the network, and then learn a generative model on top of the entire latent representations (which is the distribution of $\boldsymbol{z} = f(\boldsymbol{x}), \boldsymbol{x} \sim p(\boldsymbol{x})$), similar to the approach used in VQ-VAE-2 (Razavi et al. (2019)). We consider using the Denoising Diffusion Implicit Models (Song et al. (2020), DDIM) in our experiments. Learning a generative model over the latent space, allows us to efficiently sample from the model during test time, and thus our hybrid model can be used to learn one set of representations for both the classification downstream task and generative model downstream task.

4 Experiments and Results

Table 1: Frechet inception distance (FID) and Linear Classification Accuracy (Acc.) results for the hybrid approach and other generative models. We note that BigGAN uses separate representations for the generation and linear classification tasks. “ - ” denotes the results which are not available because the representations are not trained for the tasks in question.

	CIFAR-10			**CIFAR-100**		
	MSE	FID	Acc.	MSE	FID	Acc.
Hybrid	2.6	34.6	77.6	3.3	41.2	46.7
Moco-v2	-	-	78.3	-	-	49.04
NVAE	0.25	36.4	18.8	0.53	42.5	4.1
DDIM	2.5	3.2	22.5	3.2	13.3	2.2
BigGAN	-	18.6	70.8	-	22.2	43.3

We apply our hybrid approach over two datasets - CIFAR-10 (Krizhevsky et al. (2009)) and CIFAR-100 (Krizhevsky et al. (2009)). For both datasets the input and output sizes of the Auto-Encoder are 32*32*3 respectively. We use a modified encoder and decoder network as used in NVAE (Vahdat & Kautz (2020)), with fewer layers in our case. The size of the latent representation is 16*16*6, with the size of the representation used for contrastive learning being 16*16*3. We compare the following methods which can generate images from representations: NVAE, which uses an trained autoencoder for representations; DDIM (Song et al., 2020), which trains a neural ODE model for generative modeling, so the representations are obtained by reversing the ODE; BigGAN (Brock et al., 2018), which uses two different sets of representations for generation and classification.

Quantitative evaluations We evaluate several downstream tasks – image classification (evaluated in accuracy), image generation (evaluated in Frechet inception distance (FID) (Heusel et al. (2017))) and image reconstruction (evaluated in mean squared error).

Empirically, existing methods that use one set of representation either fail on the classification task or fail on the generation / reconstruction task. For example, methods based on autoencoding latent variables, such as NVAE and DDIM, have decent performance in terms of reconstruction and generation, but these representations are largely ineffective for classification; most notably, the performance on CIFAR-100 is not much better than random guessing. Moco-v2 does not have a decoder, so we did not report its generation and reconstruction performance. For BigGAN, there is no decoder and so we do not report its reconstruction performance. In contrast, our hybrid method is able to learn one set of representations that achieve competitive performance in all the scenarios, compared to the BigGAN approach where two separate sets of representations are used to perform each task.

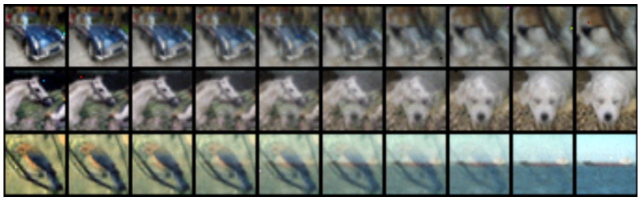


Figure 2: Samples obtained by interpolating across the entire latent code. Each row corresponds to a different set of start and end images.

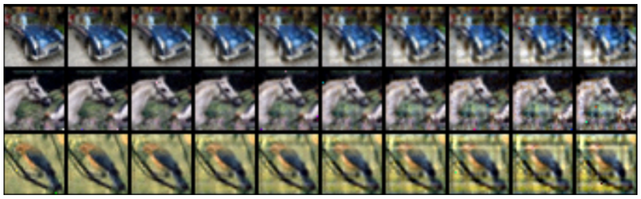


Figure 3: Samples obtained by interpolating across the top half of the latent code. The interpolated samples appear very similar with the start image.

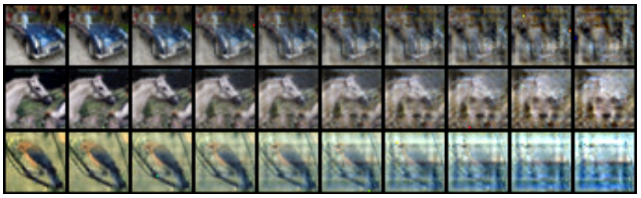


Figure 4: Samples obtained by interpolating across the bottom half of the latent code.

Qualitative evaluations Since our model has an autoencoder that encodes $\boldsymbol{x}$ into $\boldsymbol{z} = f(\boldsymbol{x})$ with two separate components (which we call $\boldsymbol{y}$ and $\hat{\boldsymbol{y}}$), we are interested in seeing how these representations are interpreted by the decoder on a high level. For two images $\boldsymbol{x}_0$ and $\boldsymbol{x}_1$, we evaluate their representations $\boldsymbol{z}_0$ and $\boldsymbol{z}_1$ and consider the following interpolations:

- Interpolate the entire latent code: $\boldsymbol{z}_t^1 = t\boldsymbol{z}_0 + (1-t)\boldsymbol{z}_1$
- Interpolate only the part for $\boldsymbol{y}$: $\boldsymbol{z}_t^2 = [t\boldsymbol{y}_0 + (1-t)\boldsymbol{y}_1, \hat{\boldsymbol{y}}_0]$
- Interpolate only the part for $\hat{\boldsymbol{y}}$: $\boldsymbol{z}_t^3 = [\boldsymbol{y}_0, t\hat{\boldsymbol{y}}_0 + (1-t)\hat{\boldsymbol{y}}_1]$

We then visualize the interpolations by using the decoder. Figures 2, 3, 4 illustrates these cases respectively; we show additional pairs in the appendix. These results suggest that the component of $\boldsymbol{y}$ appears to contain little visual information that are used by the decoder as the images appears very similar with the same $\hat{\boldsymbol{y}}$ whereas $\hat{\boldsymbol{y}}$ tend to contain most of the information needed to reconstruct the image. These results suggest that contrastive learning alone might not be suitable for image reconstruction or generation and may be prone to generative adversarial examples.

REFERENCES

David Barber and Felix V Agakov. The IM algorithm: a variational approach to information maximization. In *Advances in neural information processing systems*, pp. None. researchgate.net, 2003.

Mohamed Ishmael Belghazi, Aristide Baratin, Sai Rajeswar, Sherjil Ozair, Yoshua Bengio, Aaron Courville, and R Devon Hjelm. MINE: Mutual information neural estimation. *arXiv preprint arXiv:1801.04062*, January 2018.

Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale GAN training for high fidelity natural image synthesis. *arXiv preprint arXiv:1809.11096*, September 2018.

Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*, 2020.

Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*, March 2020.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

R Devon Hjelm, Alex Fedorov, Samuel Lavoie-Marchildon, Karan Grewal, Phil Bachman, Adam Trischler, and Yoshua Bengio. Learning deep representations by mutual information estimation and maximization. *arXiv preprint arXiv:1808.06670*, August 2018.

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. *arXiv preprint arXiv:1706.08500*, June 2017.

Diederik P Kingma and Max Welling. Auto-Encoding variational bayes. *arXiv preprint arXiv:1312.6114v10*, December 2013.

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

David McAllester and Karl Stratos. Formal limitations on the measurement of mutual information. In *International Conference on Artificial Intelligence and Statistics*, pp. 875–884, 2020.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In C J C Burges, L Bottou, M Welling, Z Ghahramani, and K Q Weinberger (eds.), *Advances in Neural Information Processing Systems 26*, pp. 3111–3119. Curran Associates, Inc., 2013.

Andriy Mnih and Koray Kavukcuoglu. Learning word embeddings efficiently with noise-contrastive estimation. In *Advances in neural information processing systems*, pp. 2265–2273, 2013.

Ben Poole, Sherjil Ozair, Aaron van den Oord, Alexander A Alemi, and George Tucker. On variational bounds of mutual information. *arXiv preprint arXiv:1905.06922*, May 2019.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. Improving language understanding by generative pre-training. *URL https://s3-us-west-2. amazonaws. com/openai-assets/researchcovers/languageunsupervised/language understanding paper. pdf*, 2018.

Ali Razavi, Aaron van den Oord, and Oriol Vinyals. Generating diverse High-Fidelity images with VQ-VAE-2. *arXiv preprint arXiv:1906.00446*, June 2019.

Jiaming Song and Stefano Ermon. Understanding the limitations of variational mutual information estimators. *arXiv preprint arXiv:1910.06222*, October 2019.

Jiaming Song, Pratyusha Kalluri, Aditya Grover, Shengjia Zhao, and Stefano Ermon. Learning controllable fair representations. *arXiv preprint arXiv:1812.04218*, December 2018.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020.

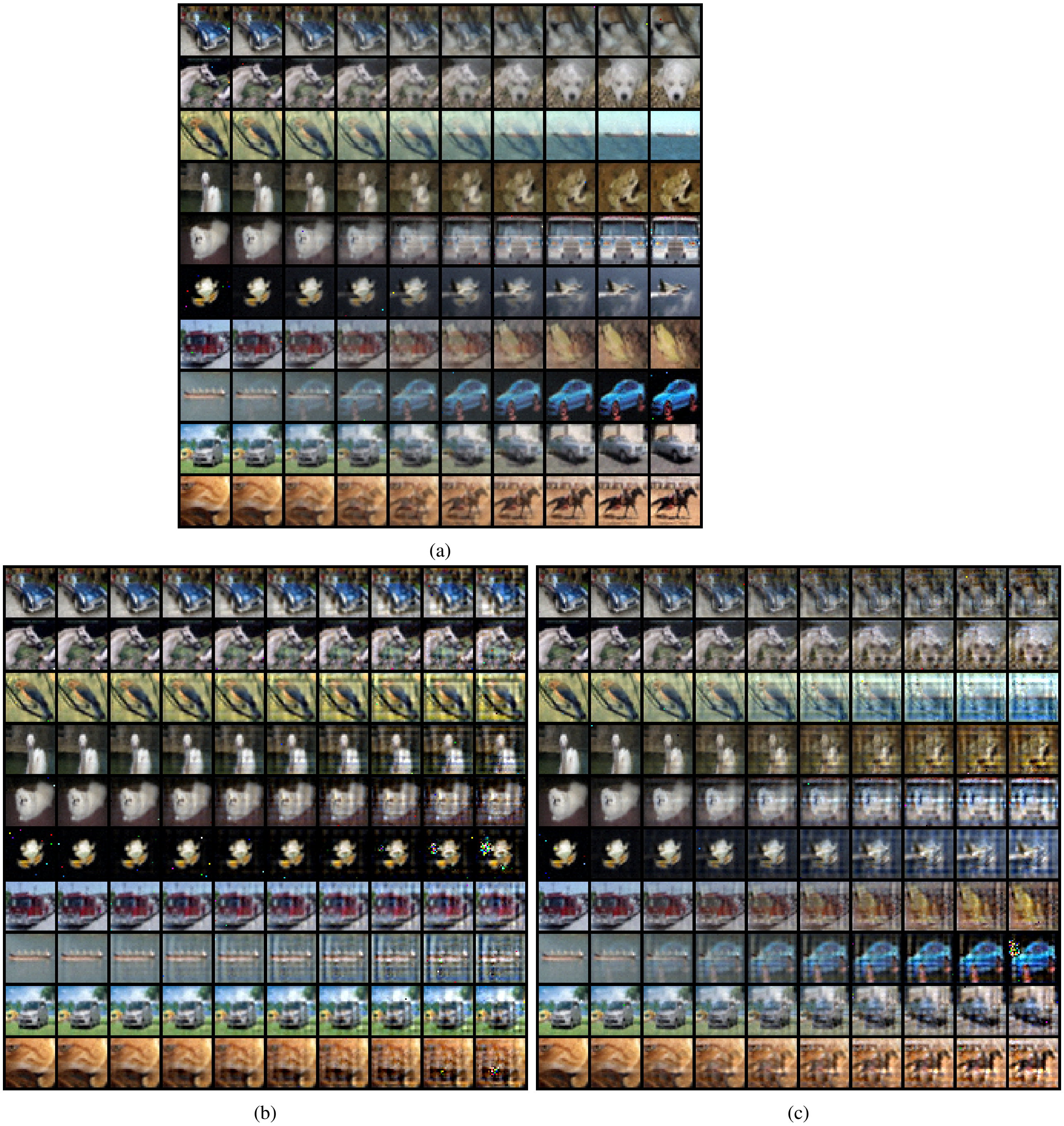
Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive representation distillation. *arXiv preprint arXiv:1910.10699*, 2019.

Arash Vahdat and Jan Kautz. Nvae: A deep hierarchical variational autoencoder. *arXiv preprint arXiv:2007.03898*, 2020.

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

Shengjia Zhao, Jiaming Song, and Stefano Ermon. The information autoencoding family: A lagrangian perspective on latent variable generative models. *arXiv preprint arXiv:1806.06514*, June 2018.

A Additional Images



(a)

(b)

(c)

Figure 5: Additional Interpolation results. Entire latent code is used for interpolation in (a), whereas the top and bottom half are used for interpolation in (b) and (c) respectively.