Conditional Generation from Unconditional Diffusion Models using Denoiser Representations

Alexandros Graikos, Srikar Yellapragada, Dimitris Samaras

Stony Brook University

Overview

- The **internal representations** of an **unconditional denoiser** network can be used to **adapt to new conditions** with **limited examples**.
- We verify the **effectiveness** of our approach on **conditional generation** tasks such as semantic mask-conditioned generation.
- Our approach allows us to **cheaply augment** with **synthetic** images to **improve classification accuracy**.

Approach

A trained denoiser network can be interpreted as a learned score function

$$\nabla_{\mathbf{x}_t}\log p_\theta(\mathbf{x}_t) = -\frac{1}{\sqrt{1-\bar{a}_t}}\boldsymbol{\epsilon}_\theta(\mathbf{x}_t,t)$$

For some conditioning **y** we can express the score of the posterior as

$$\begin{aligned}\nabla_{\mathbf{x}_t}\log p_\theta(\mathbf{x}_t\mid\mathbf{y}) &= \nabla_{\mathbf{x}_t}\log p_\theta(\mathbf{x}_t) \;+\; \lambda\nabla_{\mathbf{x}_t}\log p(\mathbf{y}\mid\mathbf{x}_t)\\ &= -\frac{1}{\sqrt{1-\bar{a}_t}}\boldsymbol{\epsilon}_\theta(\mathbf{x}_t,t) \;+\; \lambda\nabla_{\mathbf{x}_t}\log p(\mathbf{y}\mid\mathbf{x}_t)\end{aligned}$$

We propose using intermediate denoiser representations to learn to map the estimate of the final image to the conditioning

$$\underbrace{\hat{\mathbf{x}}_0 = \frac{\mathbf{x}_t-\sqrt{1-\bar{a}_t}\boldsymbol{\epsilon}_\theta(\mathbf{x}_t,t)}{\sqrt{\bar{a}_t}}}_{\text{final image estimate}} \qquad \underbrace{p(\mathbf{y}\mid\hat{\mathbf{x}}_0) = f_\phi(\hat{\mathbf{x}}_0)}_{\text{"few-shot" learned likelihood}}$$

and we can modify sampling as

$$\hat{\boldsymbol{\epsilon}}_\theta(\mathbf{x}_t,t) = \boldsymbol{\epsilon}_\theta(\mathbf{x}_t,t) - \lambda\sqrt{1-\bar{a}_t}\nabla_{\mathbf{x}_t}\log p(\mathbf{y}\mid\hat{\mathbf{x}}_0(\mathbf{x}_t))$$

Using the **unconditional denoiser** as a **feature extractor**:
- Can provide guidance that is **robust** to the initial inaccurate **estimates of $\mathbf{x}_0$**
- Allow **learning the guidance** directions from a **small set of labeled samples**.

Few-shot guidance for image-level attributes

- We perform image-level conditioning with an unconditional CelebA-64 diffusion model and training an attribute classifier with **50 positive** and **50 negative** examples, e.g. *blonde*, *male.*
- DiffAE [1] and D2C [2]: Comparable FID **without learning how to compress** the entire image into a latent representation during training.

Class	Ours	DiffAE	D2C	DDIM-I	NVAE
Male	15.34	11.52	13.44	29.03	41.07
Female	9.94	7.29	9.51	15.17	16.57
Blond	13.07	16.10	17.61	29.09	31.24
Non-Blond	10.97	8.48	8.94	19.76	16.73

Synthetic Data Augmentation

- We fine-tune an unconditional ImageNet model as class conditional on the Tiny-ImageNet dataset.
- We extract features from the unconditional U-Net and train a rejection classifier.
- To sample, discard any class-conditioned image for which the classifier predicts a probability lower than a threshold of 0.2.

- We **augment** Tiny-ImageNet with increasing amounts of diffusion-generated synthetic data.
- Training with our synthetic data **improves accuracy** of ResNet baselines by **9%** on Tiny-ImageNet.
- Augmentation approach **complements image-level augmentations** such as Mixup and Cutmix.

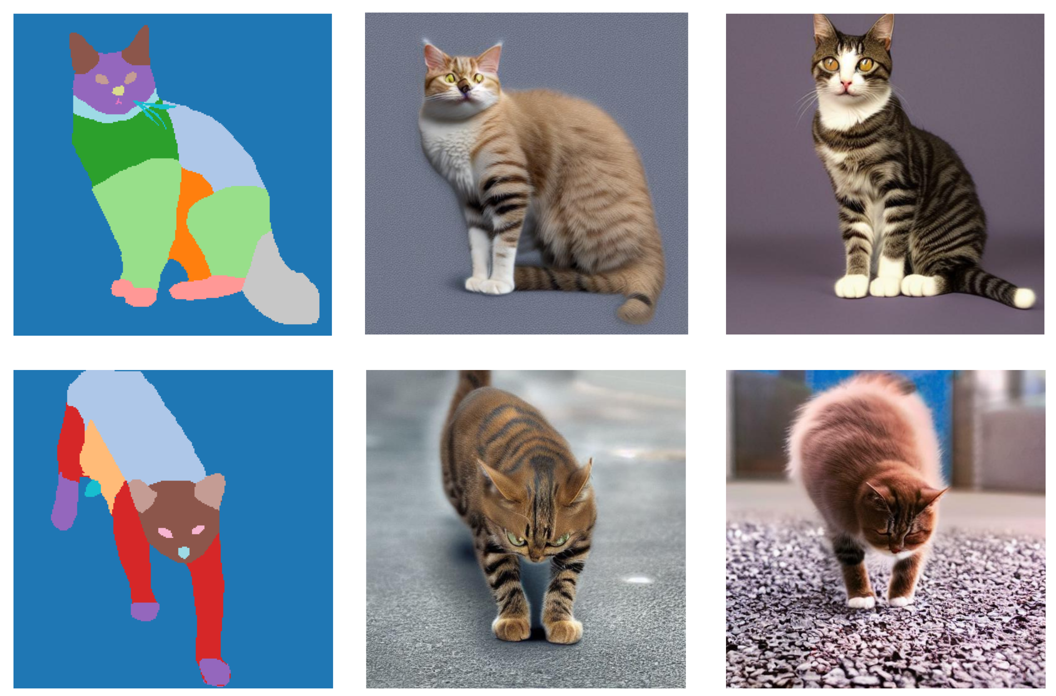
Architecture	Mixup+ Cutmix	Real only	Real+ 1x Generated	Real+ 2x Generated	Real+ 3x Generated
Resnet-18	No	52.24	56.13	58.13	**59.37**
	Yes	52.9	58.9	62.01	**62.75**
Wide-ResNet-50	No	53.27	58.57	61.71	**62.82**
	Yes	56.56	62.71	66.42	**66.82**
ResNeXt-50	No	53.98	59.33	62.27	**63.15**
	Yes	57.98	64.4	66.85	**67.05**

Few-shot guidance for semantic segmentations

- We can generate conditional samples from a small set of image-segmentation pairs
- We use a pre-trained diffusion model on FFHQ-256 and adapt it to generate conditionally with just **20 examples**
  - DiffAE [1]: The latent representation over-compresses the image and fails to accurately reproject the per-pixel segmentation
  - DDIM-I: Providing guidance with a network trained only on "clean" images does not work. The intermediate denoiser representations are more robust to the inaccurate estimates of the final image.

- We compare with ControlNet [3] which fails to work in low-data regimes.
  - We exploit the fact that the **information is highly correlated** to the existing unconditional **denoiser representations**. This allows us to **learn guidance** even in these extremely **constrained settings**.

- We showcase our ability to work with large models; we **condition** a **Stable Diffusion** model on segmentations with **30 examples**.



[1] Konpat Preechakul, et al. Diffusion autoencoders: Toward a meaningful and decodable representation, CVPR, 2022
[2] Abhishek Sinha et Al. D2C: diffusion-decoding models for few-shot conditional generation, NeurIPS 2021
[3] Lvmin Zhang et Al. Adding conditional control to text-to-image diffusion models, ICCV 2023

Acknowledgements: This work is supported by NSF grants IIS-2123920 and IIS-2212046, Stony Brook Profund 2022 seed funding and a gift from Adobe Research.

BMVC 2023