Diffusion Generative Model

Binxu Wang

neuromatch academy

Who is Binxu?

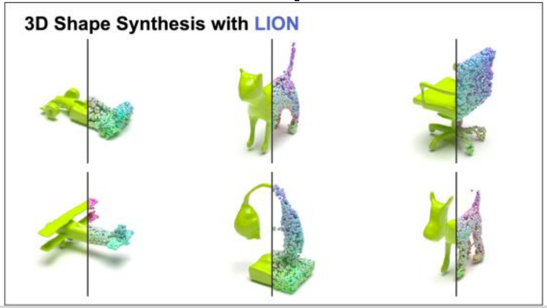
- Graduate Student in Ponce Lab @Harvard
- Incoming research fellow in Kempner Institute

- Generative models, Geometry x Visual neuroscience

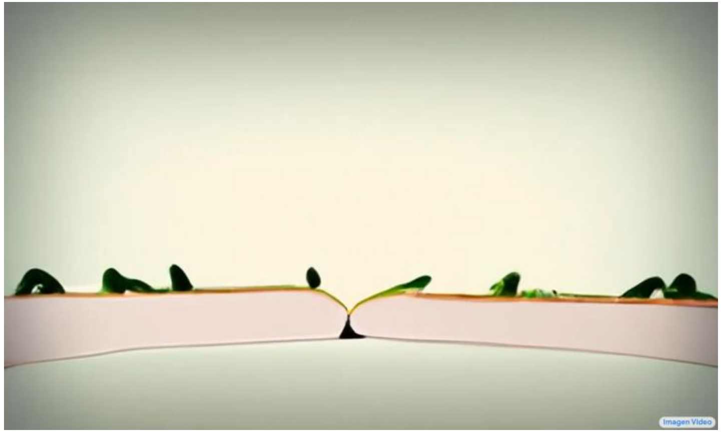
Diffusion models are leading the charge…

Images

3d shapes



Videos



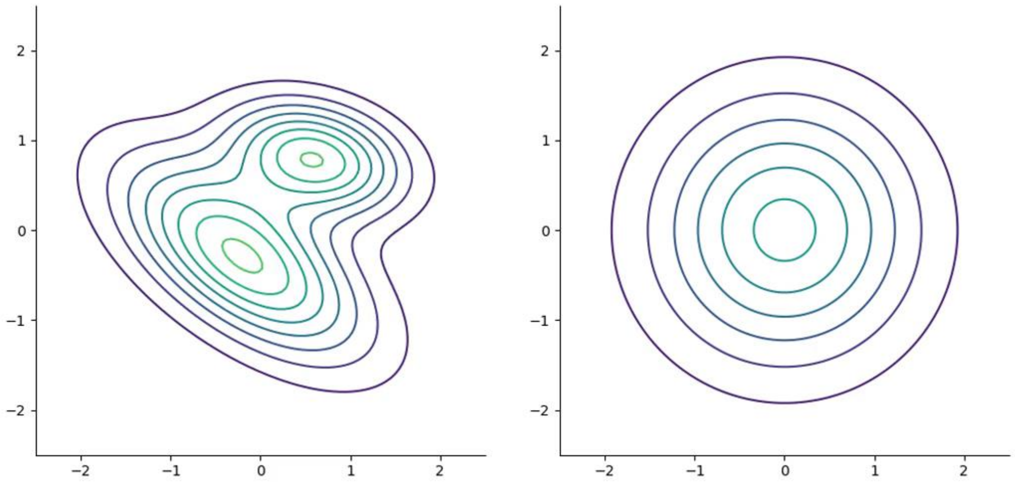
Stable Diffusion
https://imagen.research.google/video/
https://nv-tlabs.github.io/LION/

Principles of Diffusion Model

How does diffusion work at a high level?

Idea of Diffusion Generative Model

- Diffusion model learns a process that connects a simple distribution (e.g. Gaussian) with a complex one.
- To generate samples
  - Sample from the simple distribution
  - Transport it to the complex one through the process.



Denoising diffusion models

- **Forward / noising process**
  - Sample data $p(\mathbf{x}_0)$ → turn to noise

$p_0(\mathbf{x}_0)$  $p_T(\mathbf{x}_T) \sim \mathcal{N}(0, I$

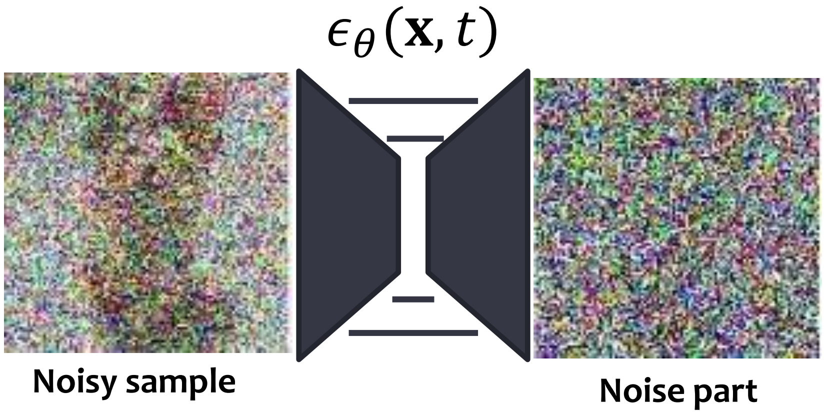
Clean sample $\mathbf{x}_0$ $\mathbf{x}_1$ $\mathbf{x}_{T-1}$ $\mathbf{x}_T$ *Pure noise*

- **Reverse / denoising process**
  - Sample noise $p_T(\mathbf{x}_T)$ → turn into data

https://yang-song.net/blog/2021/score/

Denoising via neural networks

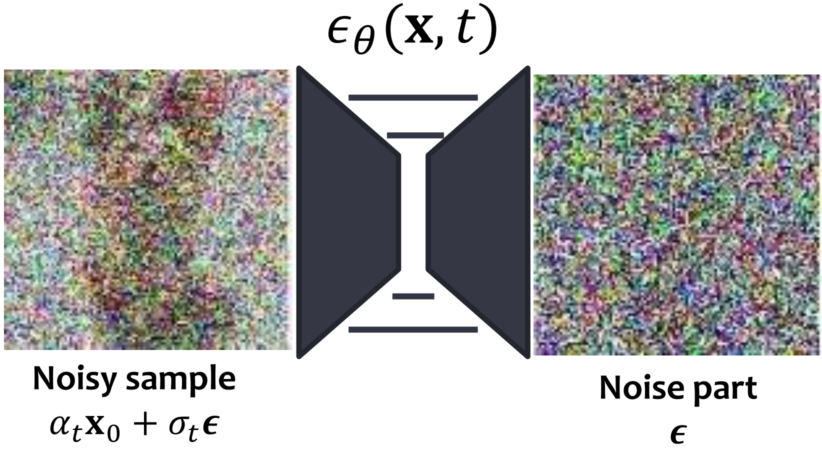
- **Objective**: $\epsilon_\theta(\mathbf{x}, t)$ predicting the noise from a degraded image.
  - a.k.a. Denoising Auto-Encoder
  - It approximates the gradient of data distribution, i.e. score function.



Song, Yang, and Stefano Ermon. "Generative modeling by estimating gradients of the data distribution." *NeurIPS* (2019).

Training Process

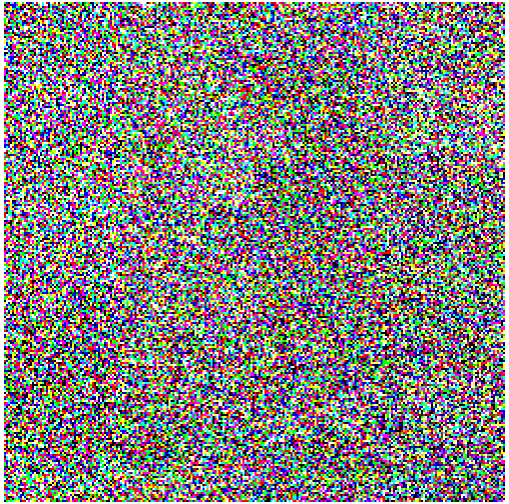
- Sample time t. The noise scale is σ_t, signal scale is α_t.
- Sample a clean example $\mathbf{x}_0 \sim p_0(\mathbf{x})$
- Add Gaussian noise $\boldsymbol{\epsilon}$ at the noise scale σ_t,
$$\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$$
- Train the denoising neural network $\boldsymbol{\epsilon}_\theta$ to infer the noise from the noisy sample
$$\arg\min_\theta \|\boldsymbol{\epsilon}_\theta(\alpha_t \mathbf{x}_0 + \sigma_t \boldsymbol{\epsilon}, t) - \boldsymbol{\epsilon}\|_2^2$$
- Repeat sampling $\mathbf{x}_0, \boldsymbol{\epsilon}, t$



Sampling by Reversing Noising Process

- Iteratively subtracting the "predicted noise" $\epsilon_\theta(\mathbf{x}_t, t)$ at t to turn the noise into data sample

- Various sampling process exists
  - DDPM, DDIM, PNDM, Etc.



https://yang-song.net/blog/2021/score/

Math behind Diffusion Models

What's the theoretical justification of this model?

Score function enables the reverse of forward process

- **Forward / noising process**

$$\dot{\mathbf{x}} = \underbrace{-\beta(t)\mathbf{x}}_{\textit{Decreasing signal}} + \underbrace{g(t)\boldsymbol{\eta}(t)}_{\textit{Adding noise}}$$

t runs forward from $0 \to T$

- **Reverse / denoising process**

$$\dot{\mathbf{x}} = -\beta(t)\mathbf{x} + g(t)\boldsymbol{\eta}(t) \underbrace{- g(t)^2\mathbf{s}(\mathbf{x}, t)}_{\textit{Drifting along score function}}$$

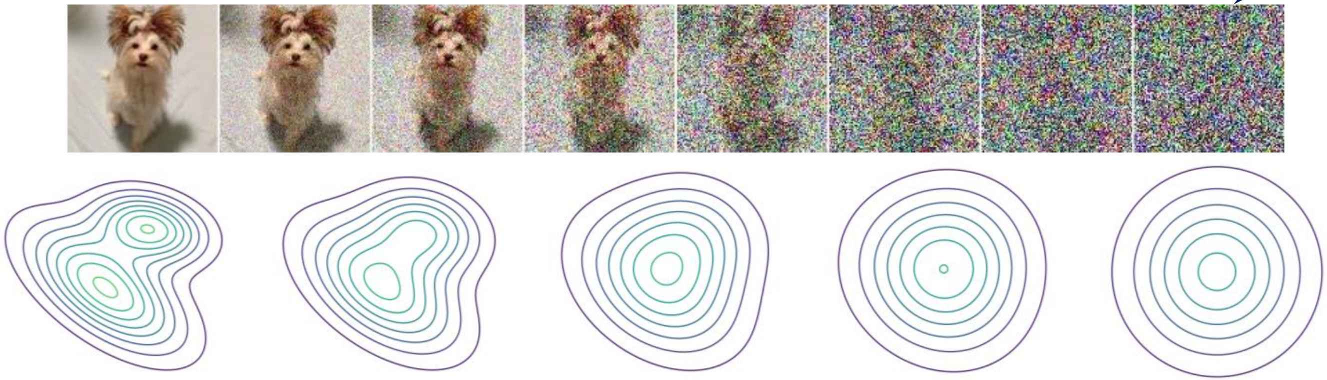
t runs backward from $T \to 0$

- $\mathbf{s}(\mathbf{x}, t) = \nabla \log p_t(\mathbf{x}_t)$

Time-reversal

- **Forward / noising process**

$$\dot{\mathbf{x}} = -\beta(t)\mathbf{x} + g(t)\boldsymbol{\eta}(t)$$



$p_0(\mathbf{x})$ $p_1(\mathbf{x})$ $p_{T-1}(\mathbf{x})$ $p_T(\mathbf{x})$

- **Reverse / denoising process**

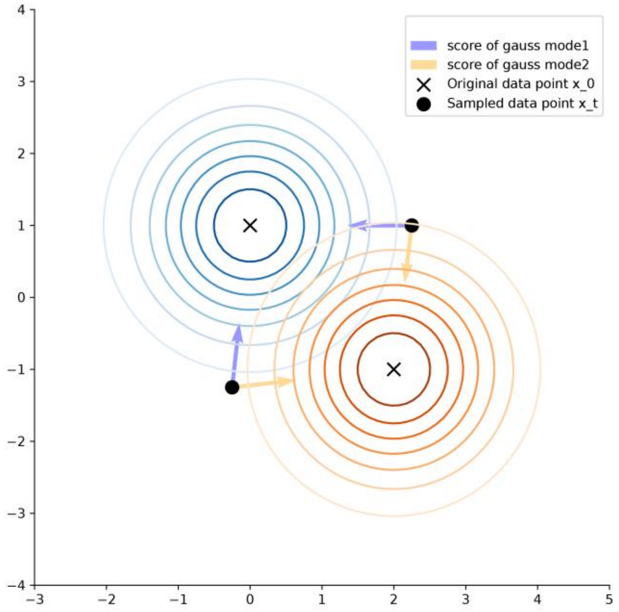
$$\dot{\mathbf{x}} = -\beta(t)\mathbf{x} + g(t)\boldsymbol{\eta}(t) - g(t)^2\mathbf{s}(\mathbf{x}, t)$$

Meaning of the Denoising objective

- Notice that the denoising objective is an **expectation** over noise $\boldsymbol{\epsilon}$, clean sample $\mathbf{x}_0$ and noise scales

$$\arg\min_{\theta} \mathbb{E}_{\substack{\boldsymbol{\epsilon}\sim\mathcal{N}(0,I)\\ \mathrm{x}_0\sim p_0(\mathbf{x})}} \|\boldsymbol{\epsilon}_\theta(\alpha_t\mathbf{x}_0+\sigma_t\boldsymbol{\epsilon},t)-\boldsymbol{\epsilon}\|_2^2$$

- This objective **cannot** be optimized to 0.



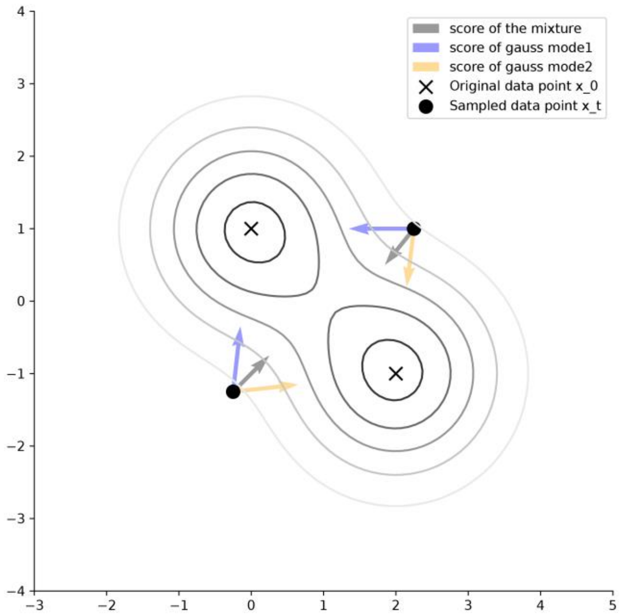
Vincent, P. (2011). A connection between score matching and denoising autoencoders. *Neural computation*

Denoising learns the **gradient of the smoothed data distribution**

- For this objective, the *optimal* $\hat{\boldsymbol{\epsilon}}_\theta(.,t)$ matches the gradient of the smoothed distribution $p_t(\mathbf{x}_t)$, $\mathbf{x}_t = \alpha_t \mathbf{x}_0 + \sigma_t \boldsymbol{\epsilon}$

$$\hat{\boldsymbol{\epsilon}}_\theta(\mathbf{x}, t) \approx -\frac{1}{\sigma_t} \nabla_{\mathbf{x}} \log p_t(\mathbf{x})$$

  - ***Score function:*** Gradient of log density
  - ***a.k.a.*** **Denoising score-matching**



Vincent, P. (2011). A connection between score matching and denoising autoencoders. *Neural computation*

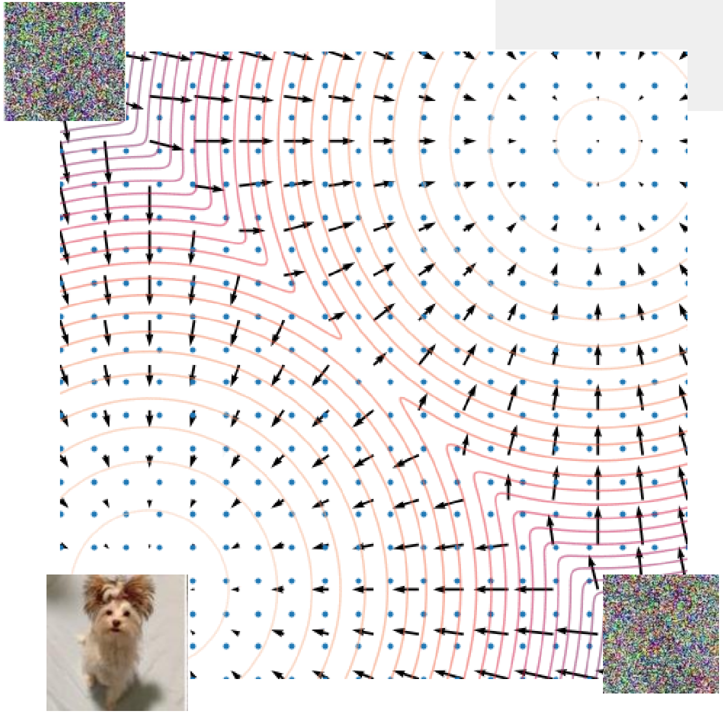
Score function enables sampling

$$\mathbf{s}(\mathbf{x}, t) \coloneqq \nabla \log p_t(\mathbf{x})$$

- Time-varying ***vector field***

  $$\mathbf{s}(\mathbf{x}, t): \mathcal{X} \times [\mathbf{0}, \mathbf{1}] \to \mathcal{X}$$

- Points towards the high-density domains.
- Enables us to ***climb up*** the data distribution to high density region.



https://yang-song.net/blog/2021/score/

Sample generation by solving the reverse ODE / SDE

- There exists a deterministic process that can also reverse the forward process, i.e. probability flow ODE.

- Advanced SDE or ODE solver (e.g. Runge Kutta) can sample diffusion models efficiently.

SDE $\dot{\mathbf{x}} = -\beta(t)\mathbf{x} + g(t)\boldsymbol{\eta}(t) - g(t)^2\mathbf{s}(\mathbf{x}, t)$

ODE $\dot{\mathbf{x}} = -\beta(t)\mathbf{x} - \frac{1}{2}g(t)^2\mathbf{s}(\mathbf{x}, t)$

Liu, L., Ren, Y., Lin, Z. and Zhao, Z., 2022. Pseudo numerical methods for diffusion models on manifolds. ICLR, *arXiv:2202.09778*.

Summary

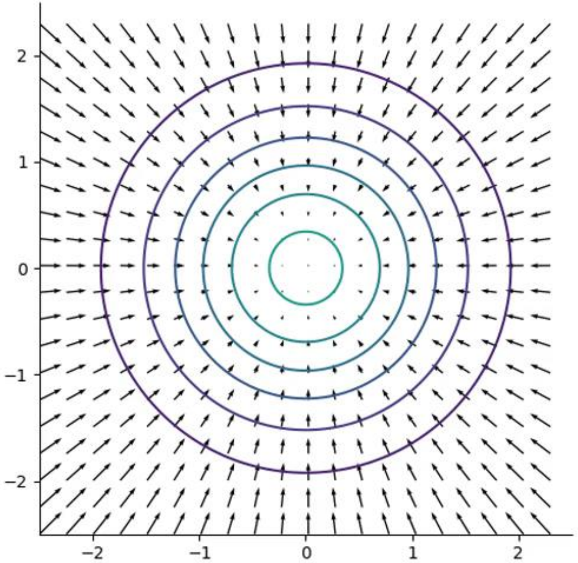
- Score enables the reversal of forward process.
- Diffusion model learns the score function of the data distribution, via denoising.
- Sampling amounts to solving reverse SDE or ODE

Score Network Architecture

How to approximate the score function?

Analysis of the target : score function

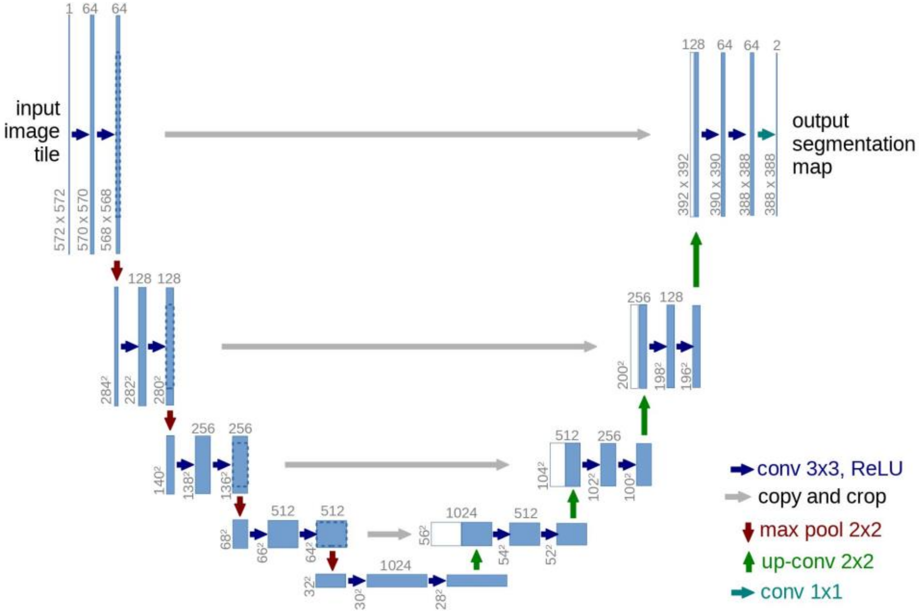
- Score function $s(\mathbf{x}, t) = \nabla_{\mathbf{x}} \log p_t(\mathbf{x})$
  - It's a continuous vector field in the domain of $\mathbf{x}$
  - It's time t dependent.
  - Its magnitude is larger for smaller t and noise scale σ_t
- Score function for specific domains
  - For image domain, it's an image-to-image mapping, modulated by time.



Backbone of Score Network - UNet

- Convolutional architecture for image-to-image mapping (e.g. segmentation, denoising)

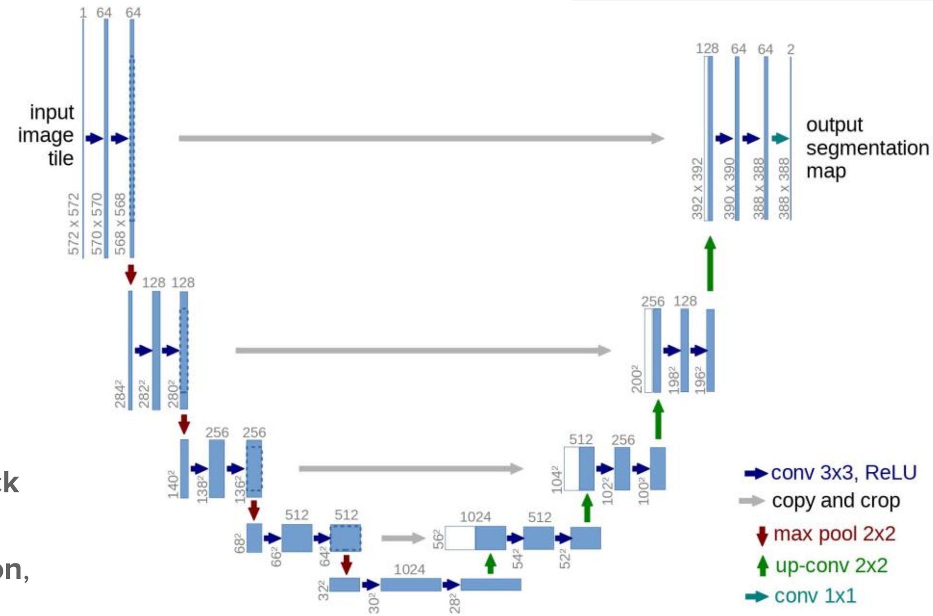
- Key features:
    - Downsampling stream
    - Upsampling stream
    - Skip-Connections



Ronneberger, O., Fischer, P. and Brox, T., 2015. U-net: Convolutional networks for biomedical image segmentation. https://arxiv.org/abs/1505.04597

Comparing with CNN and Autoencoder

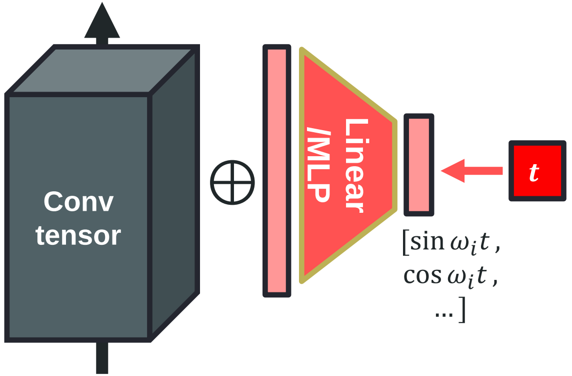
- Downsampling stream is like a normal CNN
  - Extracting features of different scale.
- Upsampling stream is like an inverted CNN, DCGAN
  - Create features of different scale.
- **Skip-connection** is the main difference from **Autoencoder**
  - AE cares about the **latent representation in bottleneck**
  - UNet doesn't care about the **bottleneck representation**, just the output



Ronneberger, O., Fischer, P. and Brox, T., 2015. U-net: Convolutional networks for biomedical image segmentation. https://arxiv.org/abs/1505.04597

Time Modulation of Score Network

- The score function is *time-dependent*.
    - Target: $s(\mathbf{x},t) = \nabla_{\mathbf{x}_t} \log p(\mathbf{x}_t)$
- Add time dependency
    - Assume time dependency is spatially homogeneous.
    - Add one scalar value $f(t)$ per feature channel
    - Parametrize $f(t)$ by MLP / linear of the Fourier basis.



Conditional Diffusion model

How to control the diffusion process

Conditional generative model

- For a paired dataset $\{\mathbf{x}, y\}$, we want to model $p(\mathbf{x}|y)$
- Conditional signal y could be
  - Class of object
  - Artistic style of image
  - Text description of images

"A picture of a cute cat running on a grassland in Van Gogh style"

Conditional diffusion model

- Train a score network to denoise, conditioned on y.

$$\epsilon_\theta(\mathbf{x}, y, t)$$

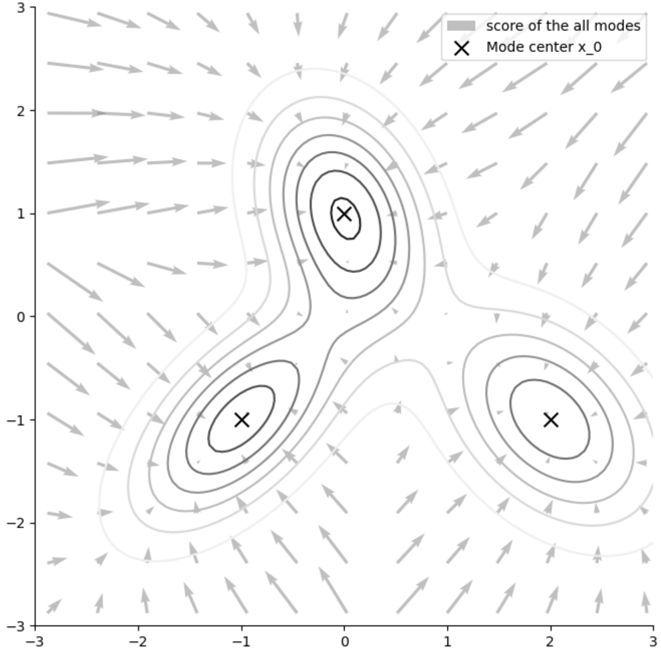
  - Approximates the score function of the conditional distribution

$$\epsilon_\theta(\mathbf{x}, y, t) \propto -\nabla_{\mathbf{x}} \log p(\mathbf{x}|y)$$

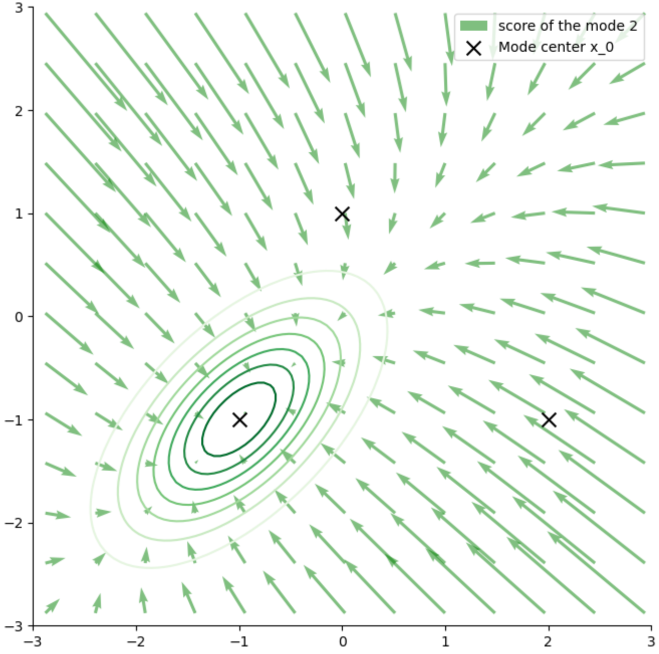
- Same sampling procedure applies.

Conditional Score Function

Unconditional dist. $p(\mathbf{x})$ and score



Conditional dist. $p(\mathbf{x}|y)$ and score



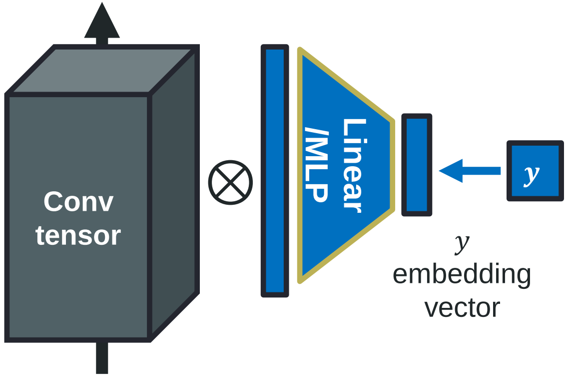
Training of conditional diffusion model

- For each t, find the noise scale σ_t, signal scale α_t.
- Sample a pair of clean example and conditional from the joint $\mathbf{x}_0, y \sim p_0(\mathbf{x}, y)$
- Add Gaussian noise $\boldsymbol{\epsilon}$ at the noise scale σ_t,
  $$\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$$
- Train the conditional denoising neural network $\boldsymbol{\epsilon}_\theta$ to infer the noise from the noisy sample
  $$\arg\min_\theta \|\boldsymbol{\epsilon}_\theta(\alpha_t \mathbf{x}_0 + \sigma_t \boldsymbol{\epsilon}, y, t) - \boldsymbol{\epsilon}\|_2^2$$
- Repeat sampling $\mathbf{x}_0, y, \boldsymbol{\epsilon}, t$

$$\epsilon_\theta(\mathbf{x}, y = \text{"dog"}, t)$$

Condition modulation of Score Network

- The score function is *class-dependent*.
  - Target: $s(\mathbf{x}, y, t) = \nabla_{\mathbf{x}_t} \log p(\mathbf{x}_t | y)$
- If the class variable y is a fixed length variable.
  - Assume condition dependency is spatially homogeneous.
  - Multiply one scalar value $g(y)$ per feature channel
  - Parametrize $g(y)$ by MLP / linear of the y embedding.



Advanced Diffusion Model Architecture

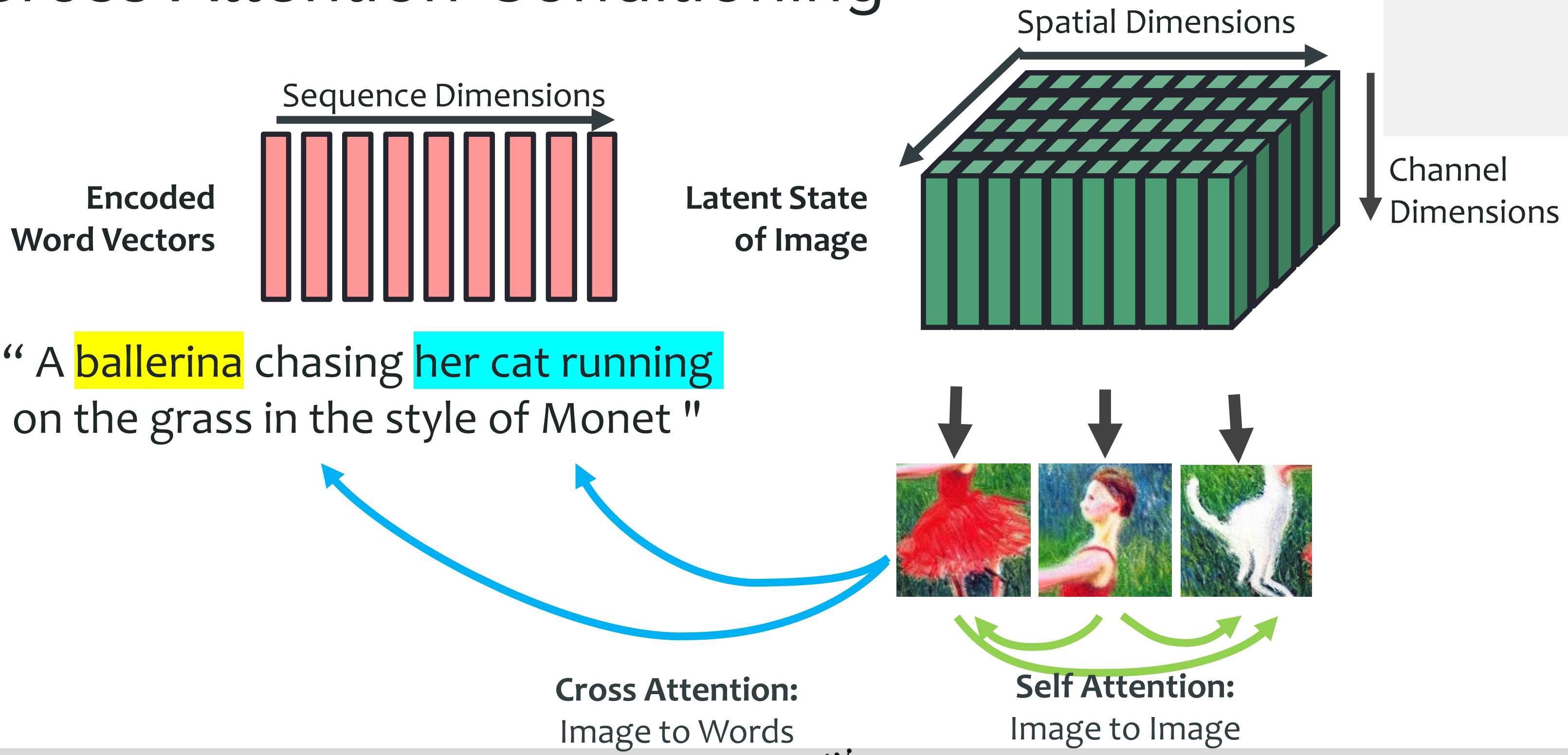
- Use Attention to inject conditional information flexibly

- Compress images into latent space to improve efficiency.

Advanced Conditional Modulation via Attention

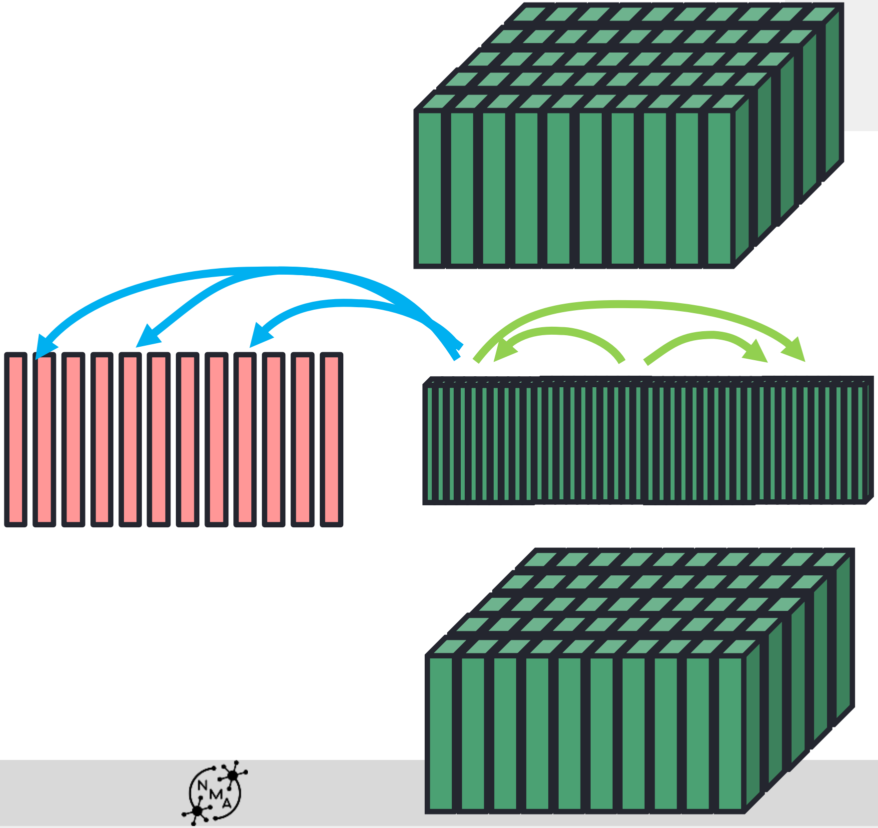
- **Challenge**
  - Conditional signal (e.g. text sentence) could have variable shape / length. A single MLP do not suffice.
  - The conditional modulation of feature is **not homogeneous** over space.
- **Solution**
  - **Cross attention** mechanism

Cross Attention Conditioning



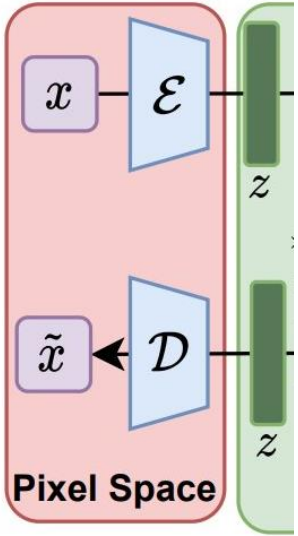
Spatial Transformer

- Rearrange spatial feature tensor to sequence.
- Cross Attention
- Self Attention
- Multi-layer Perceptron
- Rearrange back to spatial tensor (same shape)



Improve Efficiency of Diffusion Model

- **Challenge**
  - Diffusion in pixel space is computationally costly, due to large state space.
- **Solution**: Combine Autoencoder with Diffusion
  - Use Diffusion model to learn low-resolution, high level information, e.g. object, scene.
  - Use Autoencoder to generate high-resolution details, e.g. textures.



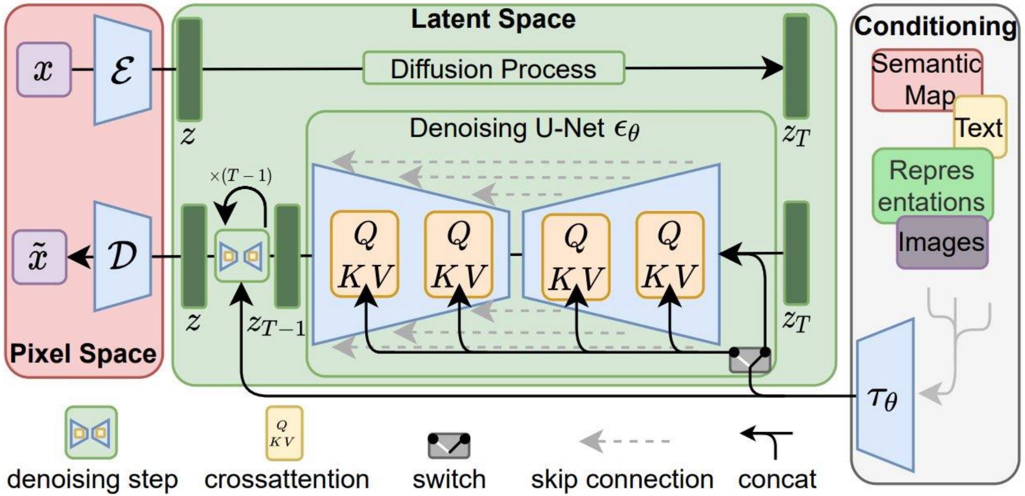
Diffusion Modelling for $p(z)$

Latent Space

Rombach, R., Blattmann, A., Lorenz, D., Esser, P. and Ommer, B., 2022. High-resolution image synthesis with latent diffusion models. CVPR

Latent Diffusion Model

- Train an autoencoder $\mathcal{E}, \mathcal{D}$ to compress image into a compact latent space
- Learn a diffusion model for the latent vectors in the latent space
- Example:
  - Stable diffusion



Rombach, R., Blattmann, A., Lorenz, D., Esser, P. and Ommer, B., 2022. High-resolution image synthesis with latent diffusion models. CVPR

Ethical Considerations

- Unleash the creativity
- Copyright and IP issue
- Spread of misinformation
- Fairness and bias

Unleash Creativity to the People

- Normal people can create personalized and artistic posters, illustrations, storybooks etc. efficiently and almost for free.

https://ai-book-colah-video.tarwin.art

[Creating a children's book with AI]()

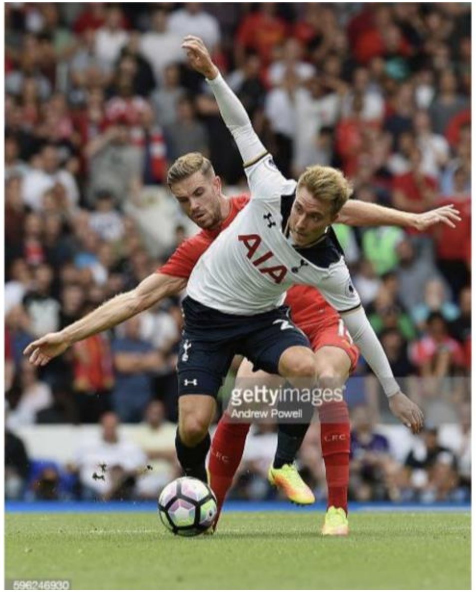
Threat to the Art Community

- Dispute
  - **Credit Assignment**: Is the person entering prompt regarded *artist*? Who receives credit for AI generated images? Is AI stealing credit from the artist?
  - **Job loss**: Will AI put currently active artists out of job ? (esp. by learning from their styles)

[An A.I.-Generated Picture Won an Art Prize. Artists Aren't Happy.] **NY times, Sep. 2022**

Copyright and Intellectual Property

Original from GettyImages



Generated by Diffusion

- Ongoing lawsuits between GettyImage and Stability AI (Stable Diffusion), arguing that SD can generate images substantially similar to its copyrighted training data.
    - Can we train generative model on “copyright” images?

[Getty Images sues AI art generator Stable Diffusion in the US for copyright infringement.](#) **Feb. 2023**

Fairness and Biases

Portrait of a librarian, SD2

Portrait of a designer, DallE 2

- By training on text-image pairs over the internet, models learn existing biased associations regarding gender, race, profession, etc.

https://huggingface.co/spaces/society-ethics/DiffusionBiasExplorer

Luccioni AS, Akiki C, Mitchell M, Jernite Y. Stable Bias: Analyzing Societal Representations in Diffusion Models. arXiv:2303.11408.

Misinformation

- Generative models can make “fake news” far more convincing, by producing photorealistic images guided by text.
- Generated video, audio could be used in scam.

https://twitter.com/EliotHiggins/status/1637927681734987777?s=20

Ongoing efforts on research and legislation

- Technology to protect images or art styles, e.g.
  - Provable Copyright Protection for Generative Models 2302.10870
  - GLAZE: Protecting Artists from Style Mimicry by Text-to-Image Models 2302.04222
- Legislation to enforce disclosure of AI generated content.
  - Clarke introduces legislation to regulate AI in political advertisements May 2nd 2023