Diffusion Models Chapter 3: Discrete-Time Diffusion Models

Generative AI and Foundation Models

Spring 2024

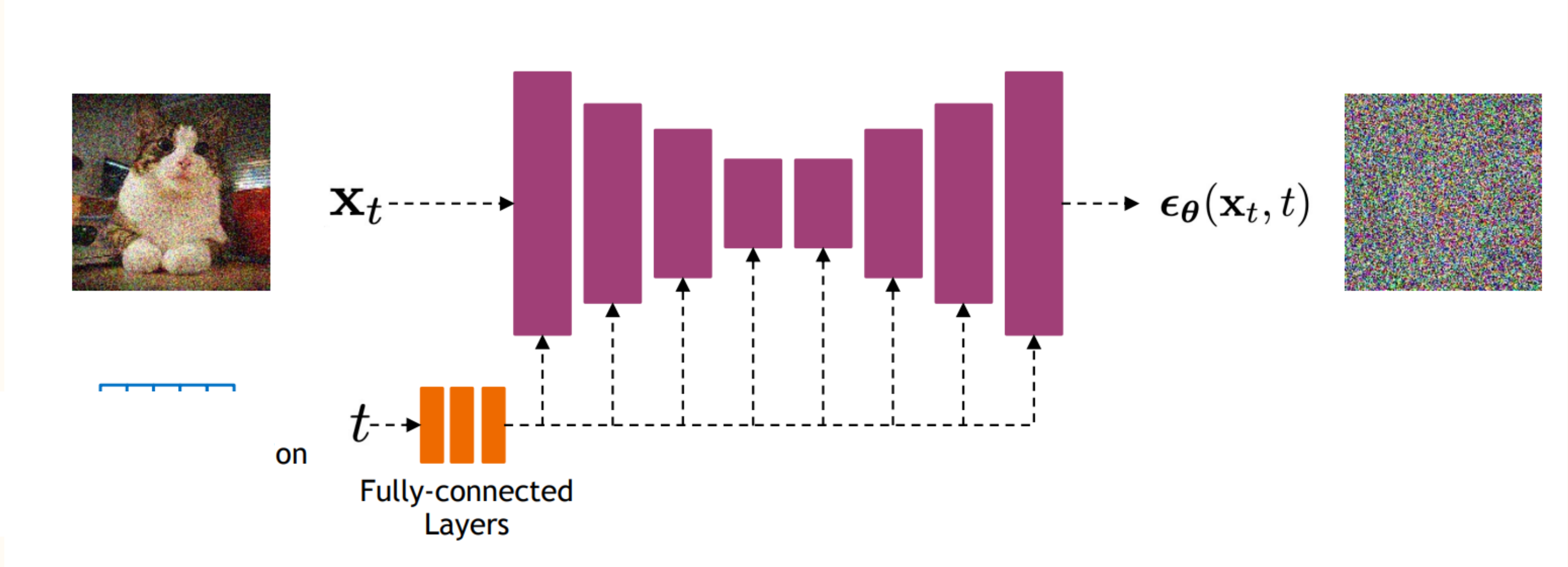
Department of Mathematical Sciences

Ernest K. Ryu

Seoul National University

Score network architecture

$s_\theta(X_t, t)$ is trained as a single U-Net architecture with time t injected into intermediate layers.



J. Ho, A. Jain, and P. Abbeel, Denoising diffusion probabilistic models, *NeurIPS*, 2020.
Y. Song, J. Sohl-Dickstein, D. P. Kingma, A. Kumar, S. Ermon, and B. Poole, Score-based generative modeling through stochastic differential equations, *ICLR*, 2021.
P. Dhariwal and A. Nichol, Diffusion models beat GANs on image synthesis, *NeurIPS*, 2021.

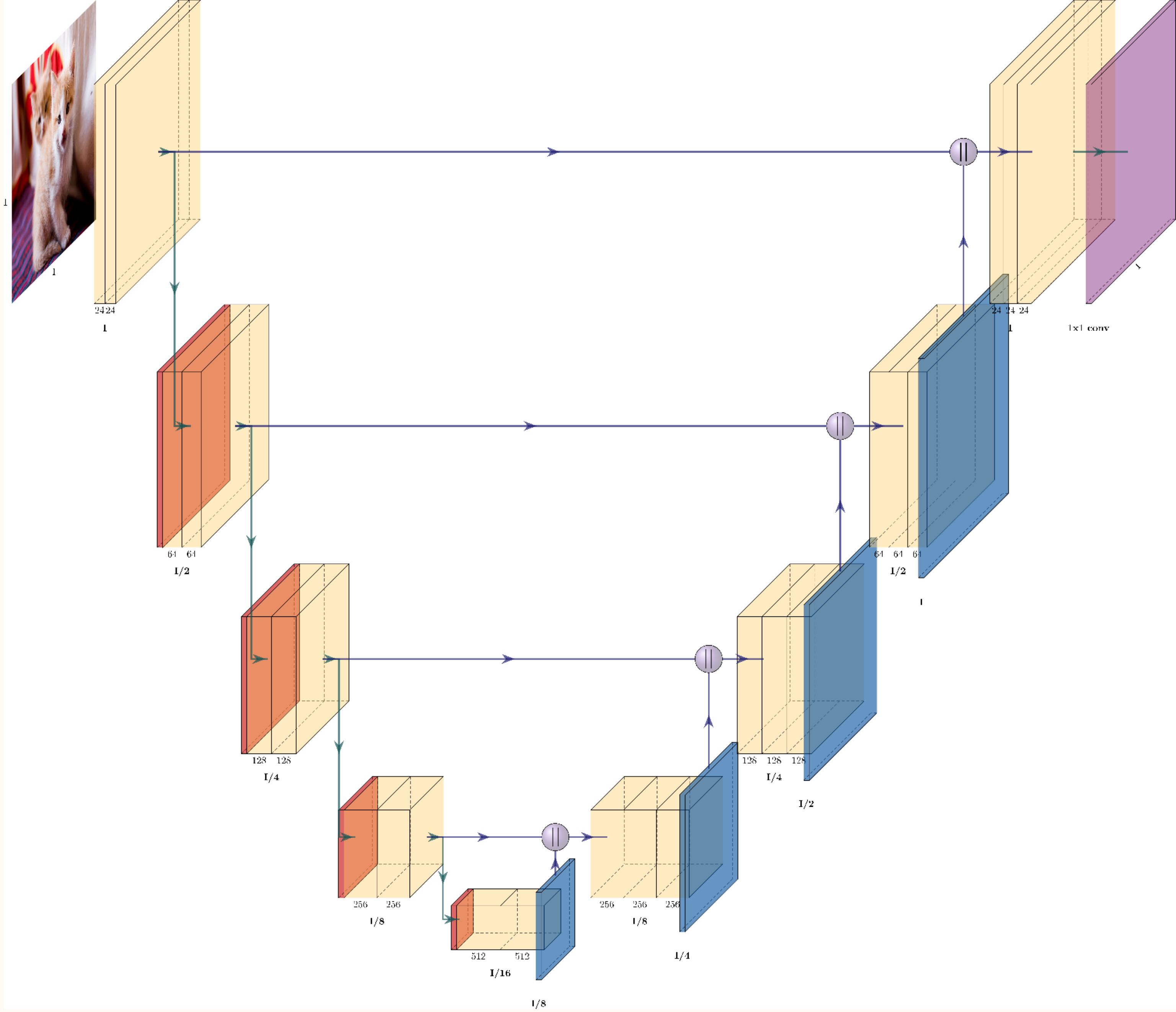
Score network architectural components

- GELU, SiLU, Swish activations
- U-Net
    - Includes convolutional layers and skip connections
- Time embedding
- Additional residual connections in residual blocks
- Attention layers
- GroupNorm

U-Net

The U-Net architecture:

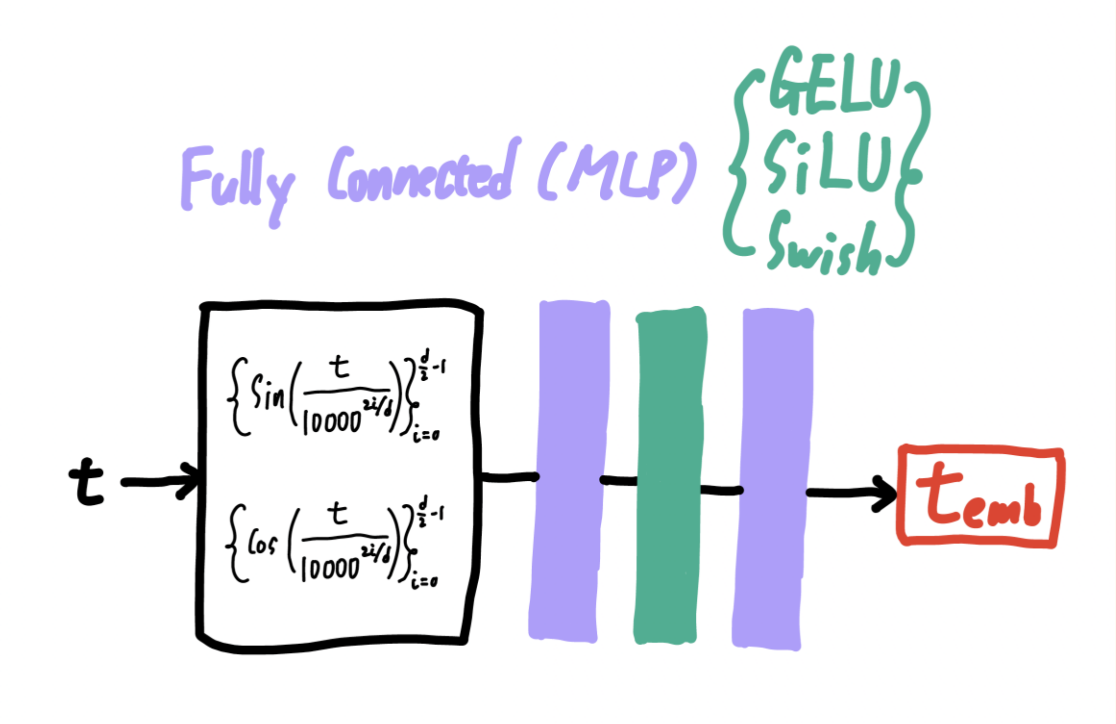
- Reduce the spatial dimension to obtain high-level (coarse scale) features
- Upsample or transpose convolution to restore spatial dimension.
- Use residual connections across each dimension reduction stage.



O. Ronneberger, P. Fischer, and T. Brox, U-Net: Convolutional networks for biomedical image segmentation, *Medical Image Computing and Computer-Assisted Intervention*, 2015.

Time embeddings

Score networks use time embedding similar to the positional encoding of transformer architectures. Time embeddings provide time information and the score networks (the residual blocks) learn to appropritately utilize the informaiton.



A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, I. Polosukhin, Attention is all you need, *NeurIPS*, 2017.

Residual block

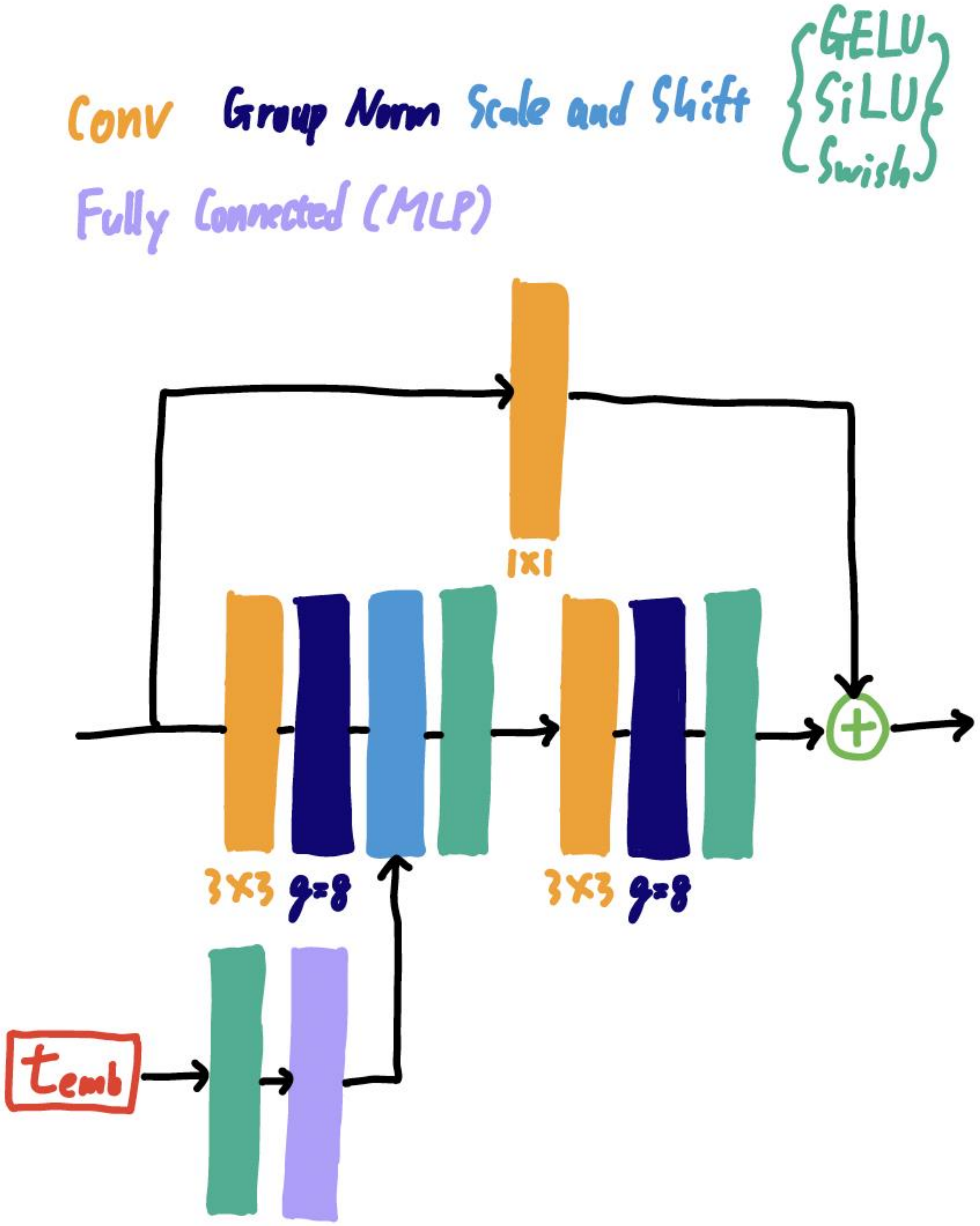
A building block of overall architecture.

Time embedding (t_{emb}) injected into the scale-shift (SS) block. SS performs

$$\text{scale} \odot Y + \text{shift}$$

where Y is the output of GroupNorm and $[\text{scale}; \text{shift}]$ is the output of the MLP processing t_{emb}.

Same t_{emb} is injected into all residual blocks. The different residual blocks can learn to use t_{emb} differently.



Pixel-wise multi-head self-attention

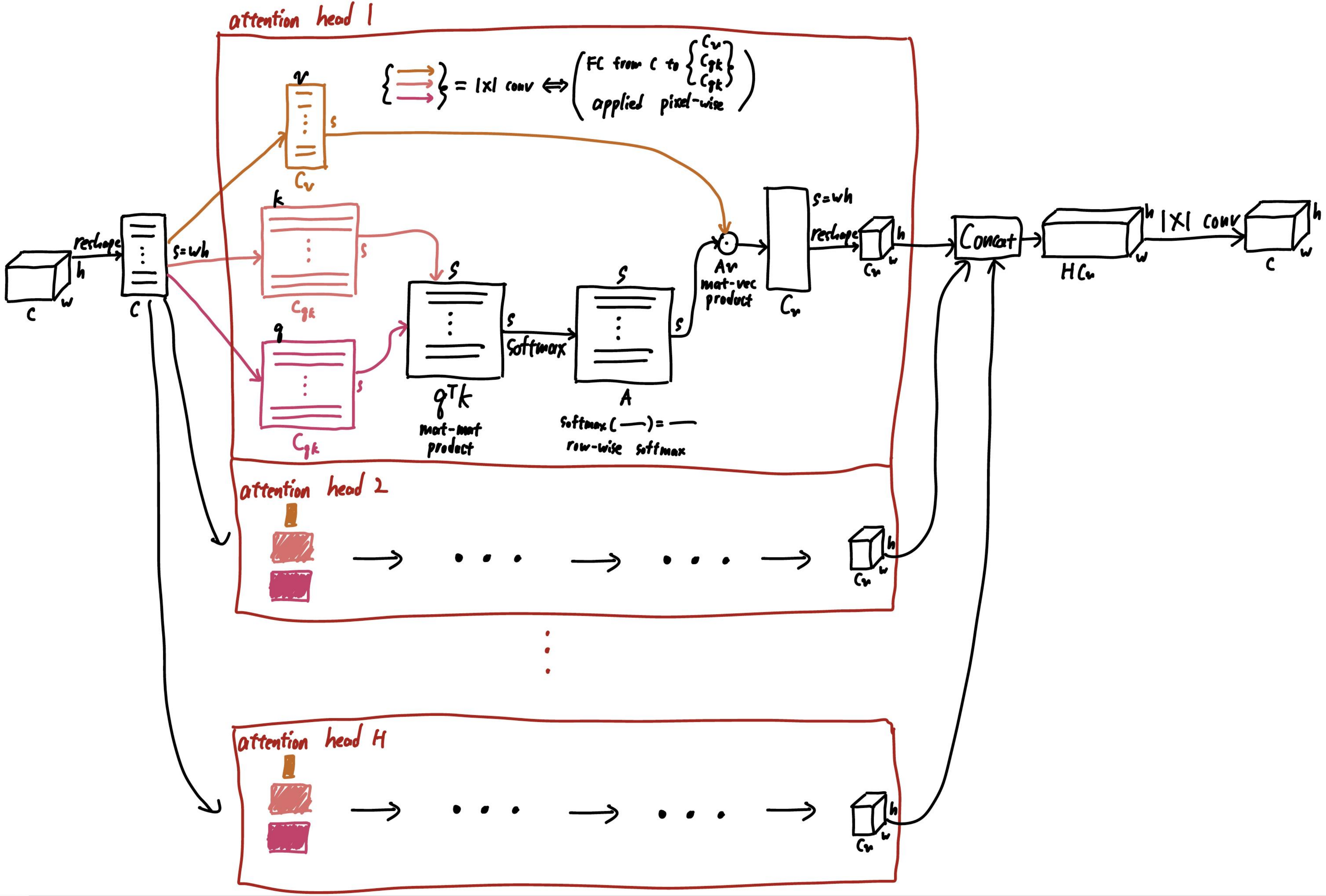
Pixel-wise multi-head encoder-only self-attention layers are used. Layer design due to [#].

Each pixel (which has many channels) gets its own query, key, and value vectors.

Different from vision transformers (ViT)[%] in 2 main ways.

- ViT are patch-wise self-attention.
- In U-Nets, attention layers are interleaved with convolution layers. ViTs are attention-only architectures.

[#] X. Chen, N. Mishra, N. Rohaninejad, and P. Abbeel, PixelSNAIL: An improved autoregressive generative model, *ICML*, 2018.
[#] J. Ho, A. Jain, and P. Abbeel, Denoising diffusion probabilistic models, *NeurIPS*, 2020.
[%] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, An image is worth 16x16 words: Transformers for image recognition at scale, *ICLR*, 2021.

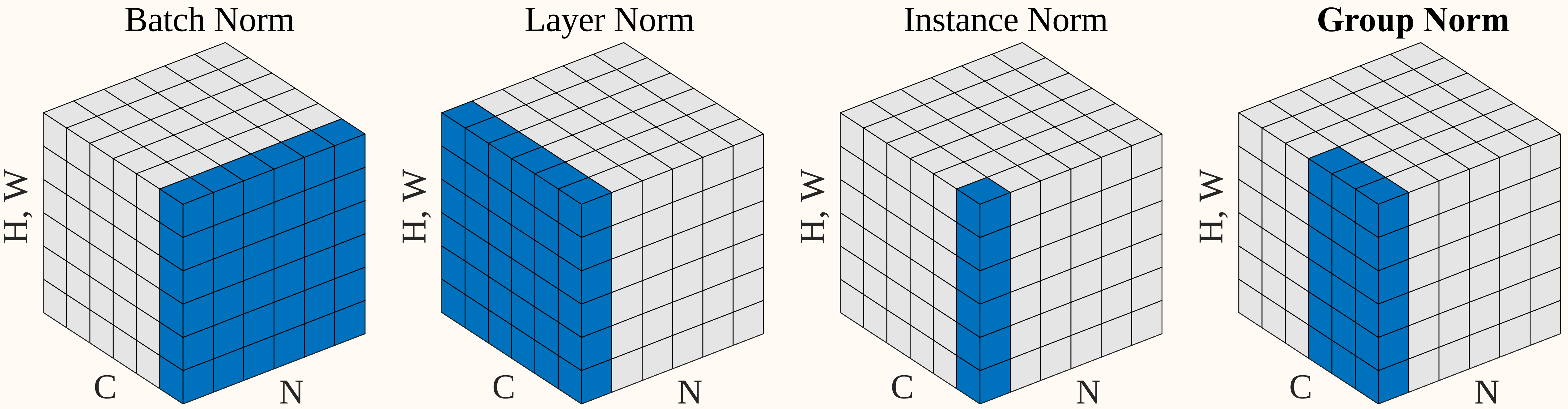


attention head 1
= 1x1 conv ⟺ (FC from c to {c_v, c_{qk}, c_{qk}} applied pixel-wise)
reshape
s=wh
c
h
w
v
c_v
k
c_{qk}
q
c_{qk}
s
$q^T k$
mat-mat product
softmax
A
softmax(—) = —
row-wise softmax
Av
mat-vec product
c_v
s=wh
reshape
Concat
Hc_v
1x1 conv
attention head 2
attention head H

GroupNorm

Batch normalization normalizes across batches and pixels (but not across channels).

Group normalization (GroupNorm) normalizes across a group of channnels and pixels (but not across batch elements).



Y. Wu and K. He, Group normalization, *ECCV*, 2018.

GN for convolutional layers

Input: X $(\text{batch size}) \times (\text{channels}) \times (\text{vertical dim}) \times (\text{horizontal dim})$

output: $\mathrm{GN}_{\beta,\gamma}(X)$. $\text{shape}\left(\mathrm{GN}_{\beta,\gamma}(X)\right) = \text{shape}(X)$

$\mathrm{GN}_{\beta,\gamma}$ for conv. layers acts independently over <u>batch elements</u>. Group count parameter G.

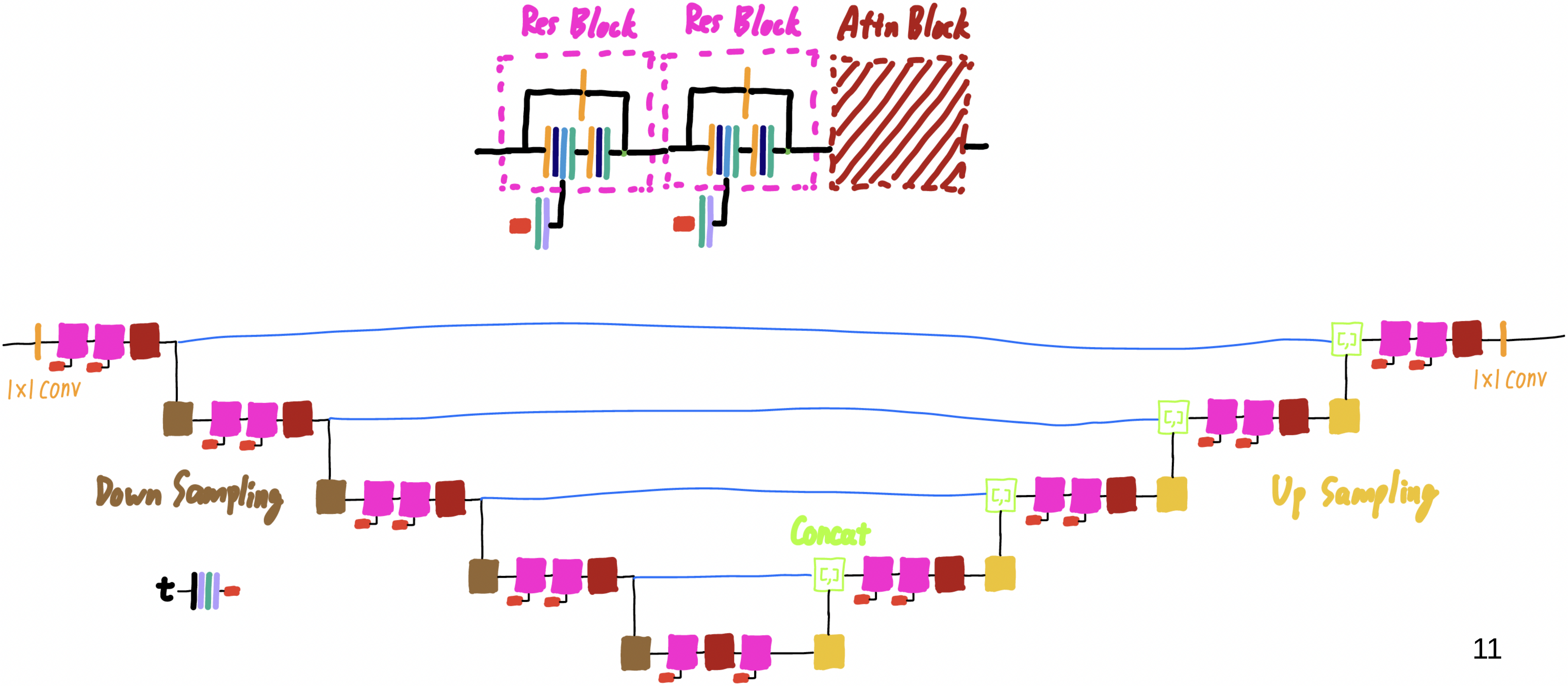
$$\hat{\mu}[:,g] = \frac{1}{(C/G)PQ}\sum_{c=1}^{C/G}\sum_{i=1}^{P}\sum_{j=1}^{Q} X[:,(g-1)C/G+c,i,j] \quad g = 1,\dots,G$$

$$\hat{\sigma}^2[:,g] = \frac{1}{(C/G)PQ}\sum_{c=1}^{C/G}\sum_{i=1}^{P}\sum_{j=1}^{Q} \left(X[:,(g-1)C/G+c,i,j] - \hat{\mu}[:,g]\right)^2 \quad g = 1,\dots,G$$

$$\mathrm{GN}_{\gamma,\beta}(X)[:,c,i,j] = \gamma[c]\frac{X[:,c,i,j] - \hat{\mu}[:,\lfloor (c-1)/G \rfloor + 1]}{\sqrt{\hat{\sigma}^2[:,\lfloor (c-1)/G \rfloor + 1] + \varepsilon}} + \beta[c] \quad \begin{array}{l} c = 1,\dots,C \\ i = 1,\dots,P \\ j = 1,\dots,Q \end{array}$$

GN normalizes over each group of convolutional filters. So $\hat{\mu}$ and $\hat{\sigma}^2$ per group. However, The mean and variance are explicitly controlled through the per-channel (not per-group) learned parameters β and γ.

Score network architecture



Discrete- to continuous-time diffusion

Publication dates:

- NCSN (NeurIPS 19)
- DDPM (NeurIPS 20)
- DDIM (ICLR 21)
- SDE Diffusion (ICLR 21)

After the dust settled, people now understand that

- NCSN is a discretization of SDE sampling of VE SDE.
- DDPM is a discretization of SDE sampling of VP SDE.
- DDIM is a discretization of ODE sampling of VP SDE. (One specific instance of DDIM.)

Tweedie's formula: 1st order

Consider the random variable

$$Y = X + \sigma Z, \quad X \sim p_X, \quad Z \sim \mathcal{N}(0, I)$$

(We don't assume p_X is Gaussian.) Then,

$$\mathbb{E}[X \mid Y] = Y + \sigma^2 \nabla_Y \log p_Y(Y)$$

If

$$Y = \gamma X + \sigma Z, \quad X \sim p_X, \quad Z \sim \mathcal{N}(0, I)$$

with $\gamma \neq 0$, then

$$\begin{aligned} \mathbb{E}[X \mid Y] &= \frac{1}{\gamma} \mathbb{E}[\gamma X \mid Y] \\ &= \frac{1}{\gamma} \left(Y + \sigma^2 \nabla_Y \log p_Y(Y) \right) \end{aligned}$$

B. Efron, Tweedie's formula and selection bias, *Journal of the American Statistical Association*, 2012.

Tweedie's formula: 2nd order

Consider the random variable

$$Y = X + \sigma Z, \quad X \sim p_X, \quad Z \sim \mathcal{N}(0, I)$$

(We don't assume p_X is Gaussian.) Then,

$$\mathrm{Var}[X \mid Y] = \sigma^2 I + \sigma^4 \nabla_Y^2 \log p_Y(Y)$$

If

$$Y = \gamma X + \sigma Z, \quad X \sim p_X, \quad Z \sim \mathcal{N}(0, I)$$

with $\gamma \neq 0$, then

$$\mathrm{Var}[X \mid Y] = \frac{\sigma^2}{\gamma^2}(I + \sigma^2 \nabla_Y^2 \log p_Y(Y))$$

B. Efron, Tweedie's formula and selection bias, *Journal of the American Statistical Association*, 2012.

Reverse cond. distribution ≈ Gaussian

Consider the random variable

$$Y = X + \sigma Z, \quad X \sim p_X, \quad Z \sim \mathcal{N}(0, I)$$

By definition, $p_{Y|X} = \mathcal{N}(X, \sigma^2 I)$ is Gaussian. (We don't assume p_X is Gaussian.) In general, $p_{X|Y}$ is not a Gaussian, but $p_{X|Y}$ is approximately Gaussian in the limit of $\sigma \to 0$.

$$p_{X|Y}(x \,|\, y) \approx \mathcal{N}\left(y + \sigma^2 \nabla \log p_Y(y), \sigma^2 I\right)$$

If

$$Y = \gamma X + \sigma Z, \quad X \sim p_X, \quad Z \sim \mathcal{N}(0, I)$$

with $\gamma \neq 0$, then, in the limit of $\sigma \to 0$,

$$p_{X|Y}(x \,|\, y) \approx \mathcal{N}\left(\frac{1}{\gamma}(y + \sigma^2 \nabla \log p_Y(y)), \frac{\sigma^2}{\gamma^2} I\right)$$

Reverse cond. distribution ≈ Gaussian

h.o.t. = higher-order term

$$
\begin{aligned}
p_{X|Y}(x \mid y) &= \frac{p_{Y|X}(y \mid x) p_X(x)}{p_Y(y)} \\
&= \frac{\frac{1}{(2\pi\sigma)^{d/2}} \exp\left(-\frac{1}{2\sigma^2}\|y-x\|^2\right) p_X(x)}{\int_{\mathbb{R}^d} \frac{1}{(2\pi\sigma)^{d/2}} \exp\left(-\frac{1}{2\sigma^2}\|y-x\|^2\right) p_X(x) dx} \\
&= \frac{\frac{1}{(2\pi\sigma)^{d/2}} \exp\left(-\frac{1}{2\sigma^2}\|y-x\|^2\right) \left(p_X(y) + \langle \nabla p_X(y), x-y \rangle + O(\|x-y\|^2)\right)}{\mathbb{E}_{x\sim\mathcal{N}(y,\sigma I)}\left[p_X(y) + \langle \nabla p_X(y), x-y \rangle + O(\|x-y\|^2)\right]} \\
&= \frac{\frac{1}{(2\pi\sigma)^{d/2}} \exp\left(-\frac{1}{2\sigma^2}\|y-x\|^2\right) \left(p_X(y) + \langle \nabla p_X(y), x-y \rangle + O(\|x-y\|^2)\right)}{p_X(y) + 0 + O(\sigma^2)} \\
&= \frac{1}{(2\pi\sigma)^{d/2}} \exp\left(-\frac{1}{2\sigma^2}\|y-x\|^2\right) (1 + \langle \nabla \log p_X(y), x-y \rangle + \text{h.o.t.}) \\
&= \frac{1}{(2\pi\sigma)^{d/2}} \exp\left(-\frac{1}{2\sigma^2}\|y-x\|^2\right) \exp\left(\langle \nabla \log p_X(y), x-y \rangle\right) + \text{h.o.t.} \\
&= \frac{1}{(2\pi\sigma)^{d/2}} \exp\left(-\frac{1}{2\sigma^2}\|x-y-\sigma^2\nabla \log p_X(y)\|^2 + \text{h.o.t.}\right) + \text{h.o.t.} \\
&= \frac{1}{(2\pi\sigma)^{d/2}} \exp\left(-\frac{1}{2\sigma^2}\|x-y-\sigma^2\nabla \log p_Y(y)\|^2 + \text{h.o.t.}\right) + \text{h.o.t.} \\
&\approx \mathcal{N}\left(y + \sigma^2 \nabla \log p_Y(y), \sigma^2 I\right)
\end{aligned}
$$

DDPM

Forward model: $X_0 \sim p_0 = p_{\text{data}}$

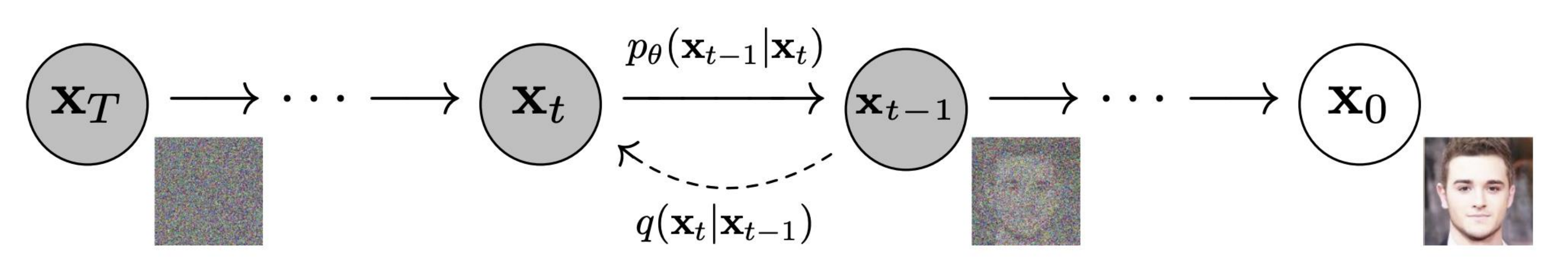
$$X_t | X_{t-1} \sim \mathcal{N}\left(\sqrt{1-\beta_t} X_{t-1}, \beta_t I\right) \quad \text{for} \quad t = 1, \ldots, T \qquad (0 < \beta_t < 1)$$

So,

$$X_t \overset{\mathcal{D}}{=} \sqrt{1-\beta_t} X_{t-1} + \sqrt{\beta_t} Z_t, \quad Z_t \sim \mathcal{N}(0, I), \quad \text{for} \quad t = 1, \ldots, T$$

and, after some calculations, this implies

$$X_t | X_0 \sim \mathcal{N}\left(\sqrt{\bar{\alpha}_t} X_0, (1-\bar{\alpha}_t) I\right) \quad \bar{\alpha}_t = \prod_{s=1}^{t} (1 - \beta_s)$$



J. Ho, A. Jain, and P. Abbeel, Denoising diffusion probabilistic models, *NeurIPS*, 2020.

DDPM

Forward model:

$$X_t|X_{t-1} \sim \mathcal{N}\left(\sqrt{1-\beta_t}X_{t-1}, \beta_t I\right) \quad \text{for} \quad t = 1, \ldots, T$$

Reverse model:

(for small β_t)

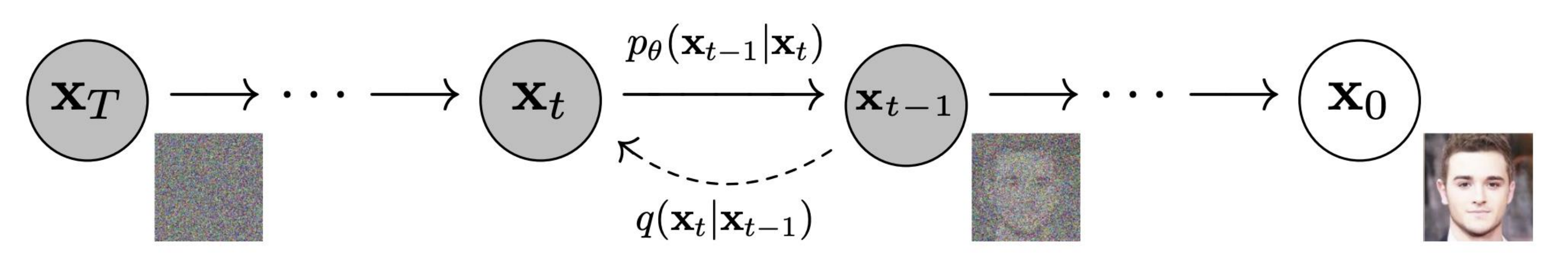
True: $\mathcal{P}(X_{t-1}|X_t) \approx \mathcal{N}\left(\mu(X_t,t), \beta_t I\right)$

Learned: $\mathcal{P}_\theta(X_{t-1}|X_t) = \mathcal{N}(\mu_\theta(X_t,t), \tilde{\beta}_t I)$

$$\mu(X_t,t) = \frac{1}{\sqrt{1-\beta_t}}(X_t + \beta_t \nabla \log p_t(X_t))$$

$$\mu_\theta(X_t,t) = \frac{1}{\sqrt{1-\beta_t}}(X_t + \beta_t s_\theta(X_t,t))$$

$$\tilde{\beta}_t = \begin{cases} \beta_t & \text{or} \\ \frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t}\beta_t & \end{cases}$$



J. Ho, A. Jain, and P. Abbeel, Denoising diffusion probabilistic models, *NeurIPS*, 2020.

Note, for small β_t

$$\frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t}\beta_t = \beta_t + \text{h.o.t.}$$

Choice of $\tilde{\beta}_t$ is further motivated in Ho, Jain, Abbeel paper.

DDPM loss

$$\mathcal{L}(\theta) = \sum_{t=1}^{T} \lambda_t \mathbb{E}_{X_t} \left[\|\mu(X_t, t) - \mu_\theta(X_t, t)\|^2 \right]$$

$$= \sum_{t=1}^{T} \frac{\lambda_t \beta_t^2}{1-\beta_t} \mathbb{E}_{X_t} \left[\|\nabla_{X_t} \log p_t(X_t) - s_\theta(X_t, t)\|^2 \right]$$

$$= \sum_{t=1}^{T} \frac{\lambda_t \beta_t^2}{1-\beta_t} \mathbb{E}_{X_0, X_t} \left[\|\nabla_{X_t} \log p_{t|0}(X_t|X_0) - s_\theta(X_t, t)\|^2 \right] + C$$

$$= \sum_{t=1}^{T} \frac{\lambda_t \beta_t^2}{1-\beta_t} \ \mathbb{E}_{X_0, X_t} \left[\| - \frac{1}{1-\bar{\alpha}_t}(X_t - \sqrt{\bar{\alpha}_t} X_0) - s_\theta(X_t, t)\|^2 \right] + C$$

$$= \sum_{t=1}^{T} \frac{\lambda_t \beta_t^2}{(1-\beta_t)(1-\bar{\alpha}_t)} \mathop{\mathbb{E}}_{\substack{X_0 \sim p_{\mathrm{data}} \\ \varepsilon \sim \mathcal{N}(0,I)}} \|\varepsilon - \varepsilon_\theta(\sqrt{\bar{\alpha}_t} X_0 + \sqrt{1-\bar{\alpha}_t}\varepsilon, t)\|^2 + C$$

$$= \sum_{t=1}^{T} \tilde{\lambda}_t \mathop{\mathbb{E}}_{\substack{X_0 \sim p_{\mathrm{data}} \\ \varepsilon \sim \mathcal{N}(0,I)}} \|\varepsilon - \varepsilon_\theta(\sqrt{\bar{\alpha}_t} X_0 + \sqrt{1-\bar{\alpha}_t}\varepsilon, t)\|^2 + C$$

$$X_t \stackrel{\mathcal{D}}{=} \sqrt{\bar{\alpha}_t} X_0 + \sqrt{1-\bar{\alpha}_t}\varepsilon_t, \qquad \varepsilon_t \sim \mathcal{N}(0, I)$$

$$\varepsilon_\theta \stackrel{\mathrm{def}}{=} -\sqrt{1-\bar{\alpha}_t} s_\theta$$

$$\tilde{\lambda}_t = \frac{\lambda_t \beta_t^2}{(1-\beta_t)(1-\bar{\alpha}_t)}$$

DDPM training

Training is analogous to the continuous-time (SDE) setup.

while (not converged)

$X_0 \sim p_0 = p_{\text{data}}$

$t \sim \text{Uniform}(\{1, \ldots, T\})$

$\varepsilon \sim \mathcal{N}(0, I)$

$X_t = \sqrt{\bar{\alpha}_t} X_0 + \sqrt{1 - \bar{\alpha}_t} \varepsilon$

Call optimizer with $\tilde{\lambda}_t \nabla_\theta \|\varepsilon_\theta(X_t, t) - \varepsilon\|^2$

end

DDPM sampling

The true distribution of X_T is $X_T | X_0 \sim \mathcal{N}\left(\sqrt{\bar{\alpha}_T} X_0, (1-\bar{\alpha}_T) I\right) \quad \bar{\alpha}_T = \prod_{s=1}^{T} (1-\beta_s)$

If T and $\beta_1, \dots, \beta_T$ are chosen such that $\bar{\alpha}_T \approx 0$, then $p_T \approx \mathcal{N}(0, I)$.

Sampling from the learned distribution

$$\mathcal{P}_\theta(X_{t-1} | X_t) = \mathcal{N}(\mu_\theta(X_t, t), \tilde{\beta}_t^2 I) \qquad \mu_\theta(X_t, t) = \frac{1}{\sqrt{1-\beta_t}} (X_t + \beta_t s_\theta(X_t, t))$$

$$\begin{aligned}
&\overline{X}_T \sim \mathcal{N}(0, I) \\
&\text{for } t = T, T-1, \dots, 2, 1 \\
&\quad Z_t \sim \mathcal{N}(0, I) \\
&\quad \overline{X}_{t-1} = \frac{1}{\sqrt{1-\beta_t}} \left(\overline{X}_t - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}} \varepsilon_\theta(\overline{X}_t, t) \right) + \tilde{\beta}_t Z_t \\
&\text{end}
\end{aligned}$$

$$\tilde{\beta}_t = \begin{cases} \beta_t & \text{or} \\ \frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t} \beta_t & \end{cases}$$

Sample X_t via the approximation of $\mathcal{P}(X_t | X_{t-1})$. It is an approximation because $\mathcal{P}(X_t | X_{t-1})$ is not exactly Gaussian and because the scaled score network ε_θ is not exact.

Reinterpreting DDPM sampling

Consider the case $\tilde{\beta}_t = \frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t}\beta_t$. We can equivalently express DDPM sampling as

$$
\begin{aligned}
&\overline{X}_T \sim \mathcal{N}(0, I)\\
&\text{for } t = T, T-1, \ldots, 2, 1\\
&\qquad \hat{X}_0 = \frac{1}{\sqrt{\bar{\alpha}_t}}\overline{X}_t - \frac{\sqrt{1-\bar{\alpha}_t}}{\sqrt{\bar{\alpha}_t}}\varepsilon_\theta(\overline{X}_t, t)\\
&\qquad Z_t \sim \mathcal{N}(0, I)\\
&\qquad \overline{X}_{t-1} = \frac{\sqrt{\bar{\alpha}_t}\beta_t}{1-\bar{\alpha}_t}\hat{X}_0 + \frac{\sqrt{1-\beta_t}(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t}\overline{X}_t + \sqrt{\frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t}\beta_t}Z_t\\
&\text{end}
\end{aligned}
$$

Equivalence follows from direct calculations.

Reinterpreting DDPM sampling

Since, $X_t | X_0 \sim \mathcal{N}\left(\sqrt{\bar{\alpha}_t} X_0, (1-\bar{\alpha}_t) I\right)$, Tweedie's formula tells us

$$\begin{aligned}
\mathbb{E}[X_0 \mid X_t] &= \frac{1}{\sqrt{\bar{\alpha}_t}} X_t + \frac{1-\bar{\alpha}_t}{\sqrt{\bar{\alpha}_t}} \nabla_{X_t} \log p_{X_t}(X_t) \\
&\approx \frac{1}{\sqrt{\bar{\alpha}_t}} X_t + \frac{1-\bar{\alpha}_t}{\sqrt{\bar{\alpha}_t}} s_\theta(X_t, t) \\
&= \frac{1}{\sqrt{\bar{\alpha}_t}} X_t - \frac{\sqrt{1-\bar{\alpha}_t}}{\sqrt{\bar{\alpha}_t}} \varepsilon_\theta(X_t, t)
\end{aligned}$$

Also, using

$$p(x_{t-1}|x_t, x_0) = \frac{p(x_t|x_{t-1}, x_0)p(x_{t-1}|x_0)}{p(x_t|x_0)} = \frac{p(x_t|x_{t-1})p(x_{t-1}|x_0)}{p(x_t|x_0)}$$

we can compute

$$\mathcal{P}(X_{t-1} \mid X_t, X_0) = \mathcal{N}\left(\frac{\sqrt{\bar{\alpha}_t}\beta_t}{1-\bar{\alpha}_t} X_0 + \frac{\sqrt{1-\beta_t}(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t} X_t, \frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t} \beta_t I\right)$$

Reinterpreting DDPM sampling

Using these identities, we can reinterpret DDPM sampling as

$\overline{X}_T \sim \mathcal{N}(0, I)$

for $t = T, T-1, \ldots, 2, 1$

$$\hat{X}_0 = \frac{1}{\sqrt{\bar{\alpha}_t}}\overline{X}_t - \frac{\sqrt{1-\bar{\alpha}_t}}{\sqrt{\bar{\alpha}_t}}\varepsilon_\theta(\overline{X}_t, t) \quad \texttt{\# } \mathbb{E}[X_0 \,|\, X_t]\ \texttt{Unbiased estimator of } X_0$$

$$\overline{X}_{t-1} \sim \mathcal{P}\left(\overline{X}_{t-1} \,|\, \overline{X}_t, \overline{X}_0 = \hat{X}_0\right)$$

end

At each step, (i) estimate X_0 and (ii) sample $\bar{X}_{t-1}$ conditioned on $\bar{X}_t$ and $\bar{X}_0 = \hat{X}_0$.

DDPM = discretization of VP SDE

DDPM forward process in the limit $\beta_t \to 0$

$$X_{t+1} = \sqrt{1-\beta_t}X_t + \sqrt{\beta_t}Z_t \approx \left(1 - \frac{\beta_t}{2}\right)X_t + \sqrt{\beta_t}Z_t$$

Consider the general VP forward-time SDE

$$dX_t = -\frac{\beta(t)}{2}X_t dt + \sqrt{\beta(t)}dW_t$$

With $\Delta t = 1$, the Euler–Maruyama discretization is

$$X_{t+1} = \left(1 - \frac{\beta(t)}{2}\right)X_t + \sqrt{\beta(t)}Z_t$$

and the two agree.

DDPM = discretization of VP SDE

DDPM sampling in the limit of slowly varying β_t and $\beta_t \to 0$

$$\begin{aligned}\overline{X}_{t-1} &= \frac{1}{\sqrt{1-\beta_t}}\left(\overline{X}_t - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}}\varepsilon_\theta(\overline{X}_t, t)\right) + \sigma_t Z_t \\ &\approx \left(1 + \frac{\beta_t}{2}\right)\overline{X}_t + \frac{\beta_t}{\sqrt{1 - \exp(-\int_0^t \beta(s)\, ds)}}\varepsilon_\theta(\overline{X}_t, t) + \sigma_t Z_t\end{aligned}$$

Here, we identify $\beta(t) = \beta_t$ and argue that

$$\bar{\alpha}_t = \prod_{s=0}^{t}(1-\beta_s) \approx \prod_{s=0}^{t}\exp(-\beta_s) = \exp\left(-\sum_{s=0}^{t}\beta_s\right) \approx \exp\left(-\int_0^t \beta(s)\, ds\right)$$

DDPM = discretization of VP SDE

Reverse-time VP SDE

$$d\overline{X}_t = \left(\frac{\beta(t)}{\sigma_t}\varepsilon_\theta(\overline{X}_t, t) - \frac{\beta(t)}{2}\overline{X}_t\right) dt + \sqrt{\beta(t)} d\bar{W}_t$$

With $\Delta t = -1$, the Euler–Maruyama discretization is

$$\overline{X}_{t-1} = \overline{X}_t - \left(\frac{\beta(t)}{\sqrt{1 - \exp(-\int_0^t \beta(s)\ ds)}}\varepsilon_\theta(\overline{X}_t, t) + \frac{\beta(t)}{2}\overline{X}_t\right) - \sqrt{\beta(t)} Z_t$$

and the two agree.

DDPM loss via variational lower bound

The score-matching DDPM loss can be obtained as a variational lower (upper) bound.

Let $X_{i:j}$ denote $(X_i, X_{i+1}, \dots, X_j)$. Let q denote the forward process and p_θ the learned reverse process. Then,

$$X_0 \sim q(X_0)$$

$$q(X_{1:T}|X_0) = \prod_{t=1}^{T} q(X_t \mid X_{t-1})$$

$$p_\theta(X_{0:T}) = p(X_T) \prod_{t=1}^{T} p_\theta(X_{t-1} \mid X_t)$$

$$p_\theta(X_0) = \int p_\theta(X_{0:T}) dX_{1:T}$$

DDPM loss via VLB

Instead of minimizing the negative log-likelihood, minimize a variational lower (upper) bound (VLB). Follow the VLB derivation using Jensen's inequality, standard for VAEs, to get the upper bound:

$$\begin{aligned} -\log p_\theta(X_0) &= -\log\left[\int p_\theta(X_{0:T})dX_{1:T}\right] \\ &= -\log\left[\int \frac{p_\theta(X_{0:T})}{q(X_{1:T}\,|\,X_0)}q(X_{1:T}\,|\,X_0)dX_{1:T}\right] \\ &= -\log \mathbb{E}_{X_{1:T}\sim q(X_{1:T}\,|\,X_0)}\left[\frac{p_\theta(X_{0:T})}{q(X_{1:T}\,|\,X_0)}\,\middle|\,X_0\right] \\ &\leq \mathbb{E}_{X_{1:T}\sim q(X_{1:T}\,|\,X_0)}\left[-\log\left(\frac{p_\theta(X_{0:T})}{q(X_{1:T}\,|\,X_0)}\right)\,\middle|\,X_0\right] \end{aligned}$$

Next, take the expectation with respect to X_0 on both sides.

$$
\begin{aligned}
&\mathbb{E}_{X_0\sim q}\left[-\log p_\theta(X_0)\right]\\
&\le \mathbb{E}_{X_{0:T}\sim q}\left[-\log\left(\frac{p_\theta(X_{0:T})}{q(X_{1:T}\,|\,X_0)}\right)\right]\\
&= \mathbb{E}_{X_{0:T}\sim q}\left[-\log p(X_T) - \sum_{t=1}^{T}\log\frac{p_\theta(X_{t-1}\,|\,X_t)}{q(X_t\,|\,X_{t-1})}\right]\\
&= \mathbb{E}_{X_{0:T}\sim q}\left[-\log p(X_T) - \sum_{t=2}^{T}\log\frac{p_\theta(X_{t-1}\,|\,X_t)}{q(X_t\,|\,X_{t-1})} - \log\frac{p_\theta(X_0\,|\,X_1)}{q(X_1\,|\,X_0)}\right]\\
&\overset{\text{(i)}}{=} \mathbb{E}_{X_{0:T}\sim q}\left[-\log p(X_T) - \sum_{t=2}^{T}\log\frac{p_\theta(X_{t-1}\,|\,X_t)}{q(X_{t-1}\,|\,X_t,X_0)}\cdot\frac{q(X_{t-1}\,|\,X_0)}{q(X_t\,|\,X_0)} - \log\frac{p_\theta(X_0\,|\,X_1)}{q(X_1\,|\,X_0)}\right]\\
&= \mathbb{E}_{X_{0:T}\sim q}\left[-\log\frac{p(X_T)}{q(X_T\,|\,X_0)} - \sum_{t=2}^{T}\log\frac{p_\theta(X_{t-1}\,|\,X_t)}{q(X_{t-1}\,|\,X_t,X_0)} - \log p_\theta(X_0\,|\,X_1)\right]\\
&= \mathbb{E}_{X_0\sim q}\left[\mathbb{E}_{X_{1:T}\,|\,X_0}\left[-\log\frac{p(X_T)}{q(X_T\,|\,X_0)}\,\middle|\,X_0\right] - \sum_{t=2}^{T}\mathbb{E}_{X_{1:T}\,|\,X_0}\left[\log\frac{p_\theta(X_{t-1}\,|\,X_t)}{q(X_{t-1}\,|\,X_t,X_0)}\,\middle|\,X_0\right] - \mathbb{E}_{X_{1:T}\,|\,X_0}\left[\log p_\theta(X_0\,|\,X_1)\,\middle|\,X_0\right]\right]\\
&= \mathbb{E}_{X_0\sim q}\left[\mathbb{E}_{X_T\,|\,X_0}\left[-\log\frac{p(X_T)}{q(X_T\,|\,X_0)}\,\middle|\,X_0\right] - \sum_{t=2}^{T}\mathbb{E}_{X_{t-1:t}\,|\,X_0}\left[\log\frac{p_\theta(X_{t-1}\,|\,X_t)}{q(X_{t-1}\,|\,X_t,X_0)}\,\middle|\,X_0\right] - \mathbb{E}_{X_1\,|\,X_0}\left[\log p_\theta(X_0\,|\,X_1)\,\middle|\,X_0\right]\right]\\
&= \mathbb{E}_{X_0\sim q}\left[\mathbb{E}_{X_T\,|\,X_0}\left[-\log\frac{p(X_T)}{q(X_T\,|\,X_0)}\,\middle|\,X_0\right] - \sum_{t=2}^{T}\mathbb{E}_{X_t\,|\,X_0}\left[\mathbb{E}_{X_{t-1}\,|\,X_0,X_t}\left[\log\frac{p_\theta(X_{t-1}\,|\,X_t)}{q(X_{t-1}\,|\,X_t,X_0)}\,\middle|\,X_0,X_t\right]\,\middle|\,X_0\right] - \mathbb{E}_{X_1\,|\,X_0}\left[\log p_\theta(X_0\,|\,X_1)\,\middle|\,X_0\right]\right]\\
&= \mathbb{E}_{X_0\sim q}\left[D_{\mathrm{KL}}(q(X_T\,|\,X_0)\,\|\,p(X_T)) + \sum_{t=2}^{T}\mathbb{E}_{X_t\,|\,X_0}\left[D_{\mathrm{KL}}(q(X_{t-1}\,|\,X_t,X_0)\,\|\,p_\theta(X_{t-1}\,|\,X_t))\,\middle|\,X_0\right] + \mathbb{E}_{X_1\,|\,X_0}\left[-\log p_\theta(X_0\,|\,X_1)\,\middle|\,X_0\right]\right]\\
&= \mathbb{E}_{X_{0:T}\sim q}\left[\underbrace{D_{\mathrm{KL}}(q(X_T\,|\,X_0)\,\|\,p(X_T))}_{L_T} + \sum_{t=2}^{T}\underbrace{D_{\mathrm{KL}}(q(X_{t-1}\,|\,X_t,X_0)\,\|\,p_\theta(X_{t-1}\,|\,X_t))}_{L_{t-1}}\underbrace{-\log p_\theta(X_0\,|\,X_1)}_{L_0}\right]
\end{aligned}
$$

$$
\text{(i)}: q(X_t\,|\,X_{t-1})q(X_{t-1}\,|\,X_0) = q(X_t,X_{t-1}\,|\,X_0) = q(X_{t-1}\,|\,X_t,X_0)q(X_t\,|\,X_0)
$$

DDPM loss via VLB

So we arrive at

$$\mathbb{E}_{X_0 \sim q}\left[-\log p_\theta(X_0)\right] \leq \mathbb{E}_{X_{0:T} \sim q}\left[\underbrace{D_{\mathrm{KL}}(q(X_T \mid X_0) \parallel p(X_T))}_{L_T} + \sum_{t=2}^{T} \underbrace{D_{\mathrm{KL}}(q(X_{t-1} \mid X_t, X_0) \parallel p_\theta(X_{t-1} \mid X_t))}_{L_{t-1}} \underbrace{- \log p_\theta(X_0 \mid X_1)}_{L_0}\right]$$

Note that L_T is independent of θ. L_0 is often ignored because it is cumberson and it does not seem to significantly affect the results. So we consider the loss

$$L = \sum_{t=2}^{T} L_{t-1}$$

DDPM loss via VLB

In the homework, you will show

$$q(X_{t-1}|X_t, X_0) = \mathcal{N}\left(\mu_t(X_t|X_0), \tilde{\beta}_t I\right)$$

$$\mu_t(X_t|X_0) = \frac{1}{\sqrt{1-\beta_t}}(X_t + \beta_t \nabla_{X_t} \log p_t(X_t|X_0))$$

Remember that

$$p_\theta(X_{t-1}|X_t) = \mathcal{N}\left(\mu_\theta(X_t, t), \tilde{\beta}_t I\right)$$

$$\mu_\theta(X_t, t) = \frac{1}{\sqrt{1-\beta_t}}(X_t + \beta_t s_\theta(X_t, t))$$

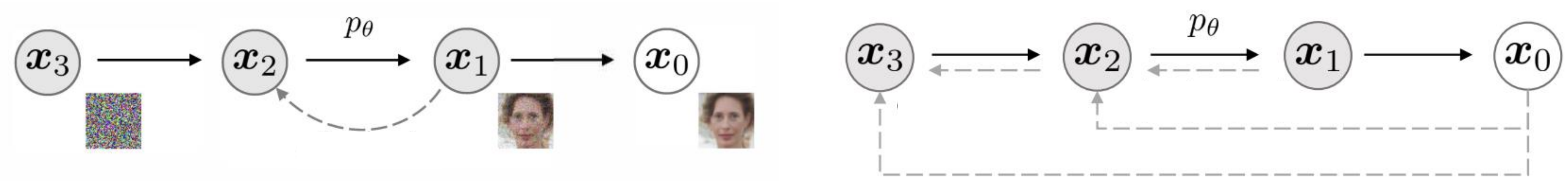
Using KL-divergence calculations that you will carry out in the homework, we have

$$\begin{aligned}
L &= \sum_{t=2}^{T} \frac{1}{2\beta_t} \|\mu_t(X_t|X_0) - \mu_\theta(X_t, X_0)\|^2 \\
&= \sum_{t=2}^{T} \frac{1}{2\beta_t} \|\nabla \log p_t(X_t|X_0) - s_\theta(X_t, X_0)\|^2 \\
&= \sum_{t=2}^{T} \frac{\beta_t}{2(1-\beta_t)} \left\| \frac{X_t - \sqrt{\bar{\alpha}_t} X_0}{1-\bar{\alpha}_t} - \frac{\varepsilon_\theta(X_t, t)}{\sqrt{1-\bar{\alpha}_t}} \right\|^2 \\
&= \sum_{t=2}^{T} \frac{\beta_t}{2(1-\beta_t)(1-\bar{\alpha}_t)} \left\|\varepsilon_t - \varepsilon_\theta\left(\sqrt{\bar{\alpha}_t} X_0 + \sqrt{1-\bar{\alpha}_t}\varepsilon_t, t\right)\right\|^2
\end{aligned}$$

$$\varepsilon_\theta = \sqrt{1-\bar{\alpha}_t} s_\theta$$

$$X_t = \sqrt{\bar{\alpha}_t} X_0 - \sqrt{1-\bar{\alpha}_t}\varepsilon_t, \quad \varepsilon_t \sim \mathcal{N}(0, I)$$

DDIM



The graphical models of DDPM generation (left) and DDIM generation (right).

Denoising Diffusion Implicit Models (DDIM) is a discrete-time diffusion probabilistic model based on non-Markovian “forward” process.

Specifically we have

$$q(X_1, \ldots, X_T \mid X_0) = q(X_T \mid X_0) \prod_{t=1}^{T-1} q(X_t \mid X_{t+1}, X_0)$$

$$q(X_T \mid X_0) = \mathcal{N}\left(\sqrt{\bar{\alpha}_T} X_0, (1 - \bar{\alpha}_T) I\right)$$

$$q(X_t \mid X_{t+1}, X_0) = \mathcal{N}\left(\sqrt{\bar{\alpha}_t} X_0 + \frac{\sqrt{1 - \bar{\alpha}_t - \rho_{t+1}^2}}{\sqrt{1 - \bar{\alpha}_{t+1}}}(X_{t+1} - \sqrt{\bar{\alpha}_{t+1}} X_0), \rho_{t+1}^2 I\right)$$

For us, the $\rho_t = 0$ case is most interesting as it corresponds to ODE sampling.

J. Song, C. Meng, and S. Ermon, Denoising diffusion implicit models, *ICLR*, 2021.

DDIM marginals = DDPM marginals

The transition kernel $X_0 \mapsto X_T$ and $(X_0, X_{t+1}) \mapsto X_t$ are chosen so that the marginals of DDIM match the marginals of DDPM:

$$q(X_t \mid X_0) = \mathcal{N}(\sqrt{\bar{\alpha}_t} X_0, (1 - \bar{\alpha}_t) I), \qquad t = 0, \dots, T$$

Proof by induction:

$$q(X_{t+1} \mid X_0) = \mathcal{N}(\sqrt{\bar{\alpha}_{t+1}} X_0, (1 - \bar{\alpha}_{t+1}) I)$$

$$q(X_t \mid X_{t+1}, X_0) = \mathcal{N}\left(\sqrt{\bar{\alpha}_t} X_0 + \frac{\sqrt{1 - \bar{\alpha}_t - \rho_{t+1}^2}}{\sqrt{1 - \bar{\alpha}_{t+1}}} (X_{t+1} - \sqrt{\bar{\alpha}_{t+1}} X_0), \rho_{t+1}^2 I\right)$$

$$q(X_t \mid X_0) = \int q(X_{t+1} \mid X_0) q(X_t \mid X_{t-1}, X_0)\, dX_{t+1}$$

$$\overset{\text{calculations}}{=} \mathcal{N}(\sqrt{\bar{\alpha}_t} X_0, (1 - \bar{\alpha}_t) I)$$

To be precise, this shows that the conditional marginals, conditioned on X_0, match.
This implies that the marginals, conditioned on nothing, also match.

DDIM training = DDPM training

Since DDIM and DDPM have the same conditional marginals, their conditional and unconditional score functions are the same.

DDIM trains the error (score) network $\varepsilon_\theta(X_t, t)$ that predicts ε_t given

$$X_t \overset{\mathcal{D}}{=} \sqrt{\bar{\alpha}_t} X_0 + \underbrace{\sqrt{1 - \bar{\alpha}_t} Z_t}_{= \varepsilon_t}, \qquad t = 0, \ldots, T$$

where $Z_1, \ldots, Z_t$is are IID unit Gaussians.

Training of DDPM and DDIM are identical.
(Training requires “forward-time” corruption.)

Sampling of DDPM and DDPM differ.
(Sampling refers to “reverse-time” sampling.)

DDIM sampling

Unbiased estimator of X_0 given X_t:

$$\hat{X}_0 = \mathbb{E}[X_0 \,|\, X_t] \approx \frac{1}{\sqrt{\bar{\alpha}_t}}(X_t - \sqrt{1-\bar{\alpha}_t}\varepsilon_\theta(X_t, t))$$

DDIM sampling is done with

$$p_\theta(X_t \,|\, X_{t+1}) = q(X_t \,|\, X_{t+1}, X_0 = \hat{X}_0)$$

$\overline{X}_T \sim \mathcal{N}(0, I)$

for $t = T, T-1, \ldots, 2, 1$

$$\hat{X}_0 = \frac{1}{\sqrt{\bar{\alpha}_t}}\overline{X}_t - \frac{\sqrt{1-\bar{\alpha}_t}}{\sqrt{\bar{\alpha}_t}}\varepsilon_\theta(\overline{X}_t, t) \quad \texttt{\#}\ \mathbb{E}[X_0 \,|\, X_t]\ \texttt{Unbiased estimator of}\ X_0$$

$$Z_t \sim \mathcal{N}(0, I)$$

$$\overline{X}_{t-1} = \sqrt{\bar{\alpha}_{t-1}}\hat{X}_0 + \sqrt{1-\bar{\alpha}_{t-1}-\rho_t^2}\varepsilon_\theta(\overline{X}_t, t) + \rho_t Z_t$$

end

Deterministic DDIM sampling

When $\rho_t = 0$,only generation of $\bar{X}_T$ is random, and the subsequent steps are deterministic.

$$\overline{X}_T \sim \mathcal{N}(0, I)$$

for $t = T, T-1, \ldots, 2, 1$

$$\hat{X}_0 = \frac{1}{\sqrt{\bar{\alpha}_t}}\overline{X}_t - \frac{\sqrt{1-\bar{\alpha}_t}}{\sqrt{\bar{\alpha}_t}}\varepsilon_\theta(\overline{X}_t, t) \quad \texttt{\# } \mathbb{E}[X_0 \,|\, X_t]\, \texttt{Unbiased estimator of } X_0$$

$$\overline{X}_{t-1} = \sqrt{\bar{\alpha}_{t-1}}\hat{X}_0 + \sqrt{1-\bar{\alpha}_{t-1}}\varepsilon_\theta(\overline{X}_t, t)$$

end

Deterministic DDIM sampling

We can express the $\sigma_t = 0$ generation in one line as follows.

$$
\begin{aligned}
&\overline{X}_T \sim \mathcal{N}(0, I) \\
&\text{for } t = T, T-1, \ldots, 2, 1 \\
&\qquad \overline{X}_{t-1} = \frac{1}{\sqrt{1-\beta_t}} \overline{X}_t - \left(\frac{\sqrt{1-\bar{\alpha}_t}}{\sqrt{1-\beta_t}} - \sqrt{1 - \frac{\bar{\alpha}_t}{1-\beta_t}} \right) \varepsilon_\theta(\overline{X}_t, t) \\
&\text{end}
\end{aligned}
$$

Equivalence follows from direct calculations.

DDIM = discretization of VP ODE

Consider the general VP forward-time SDE

$$dX_t = -\frac{\beta(t)}{2} X_t dt + \sqrt{\beta(t)} dW_t$$

Since DDIM and DDPM share the same marginals, the forward process of DDIM can also be viewed as a discretization of VP ODE.

DDIM = discretization of VP ODE

DDIM sampling in the limit of slowly varying β_t and $\beta_t \to 0$

$$
\begin{aligned}
\overline{X}_{t-1} &= \frac{1}{\sqrt{1-\beta_t}}\overline{X}_t - \left(\frac{\sqrt{1-\bar{\alpha}_t}}{\sqrt{1-\beta_t}} - \sqrt{1-\frac{\bar{\alpha}_t}{1-\beta_t}}\right)\varepsilon_\theta(\overline{X}_t, t) \\
&\approx \left(1+\frac{\beta_t}{2}\right)\overline{X}_t - \frac{\beta_t}{2\sqrt{1-\bar{\alpha}_t}}\varepsilon_\theta(\overline{X}_t, t) \\
&\approx \left(1+\frac{\beta_t}{2}\right)\overline{X}_t - \frac{\beta_t}{2\sqrt{1-\exp(-\int_0^t \beta(s)\ ds)}}\varepsilon_\theta(\overline{X}_t, t)
\end{aligned}
$$

DDIM = discretization of VP ODE

The corresponding reverse-time VP ODE is

$$d\overline{X}_t = \left(\frac{\beta(t)}{\sigma_t} \varepsilon_\theta(\overline{X}_t, t) - \frac{\beta(t)}{2} \overline{X}_t \right) dt, \qquad \sigma_t^2 = 1 - e^{-\int_0^t \beta(s)\, ds}$$

With $\Delta t = -1$, the Euler discretization is

$$\overline{X}_{t-1} = \left(1 + \frac{\beta(t)}{2}\right) \overline{X}_t - \frac{\beta(t)}{2\sqrt{1 - \exp(-\int_0^t \beta(s)\, ds)}} \varepsilon_\theta(\overline{X}_t, t)$$

and the two agree.