Denoising Diffusion Implicit Models (DDIM)

Implementation Details

U-Net Architecture

U-Net consists of a contracting path and an expansive path. The contracting path is a series of convolutional layers and pooling layers, where the resolution of the feature map gets progressively reduced. Expansive path is a series of up-sampling layers and convolutional layers where the resolution of the feature map gets progressively increased. At every step in the expansive path the corresponding feature map from the contracting path concatenated with the current feature map.

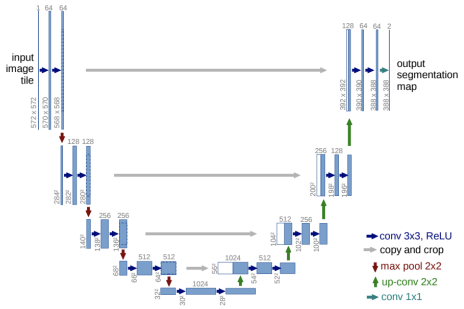


Fig. 1. U-net architecture (example for 32x32 pixels in the lowest resolution). Each blue box corresponds to a multi-channel feature map. The number of channels is denoted on top of the box. The x-y-size is provided at the lower left edge of the box. White boxes represent copied feature maps. The arrows denote the different operations.

DDIM Sampler

1. Noise Term (σ_i):

$$\sigma_i = \eta \sqrt{\frac{(1-\alpha_{\tau_{i-1}})}{(1-\alpha_{\tau_i})} \cdot \left(1 - \frac{\alpha_{\tau_i}}{\alpha_{\tau_{i-1}}}\right)}$$

2. Predicted Image (x_0):

$$x_0 = \frac{1}{\sqrt{\alpha_{\tau_i}}} x_t - \frac{1}{\sqrt{\alpha_{\tau_{i-1}}}} \epsilon_\theta(x_t)$$

3. Next Sample (x_{t-1}):

$$x_{t-1} = \text{mean} + \sigma_i \cdot \text{noise}$$

4. Mean Calculation:

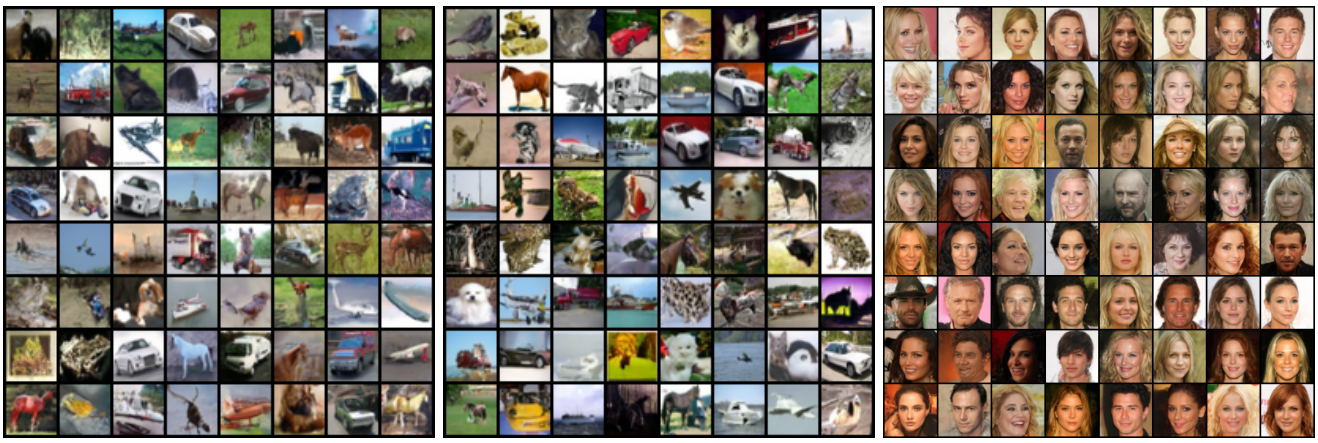
$$\text{mean} = \sqrt{\alpha_{\tau_{i-1}}} x_0 + \text{coeff}_i \cdot \epsilon_\theta(x_t)$$

5. Coefficient Calculation (coeff_i):

$$\text{coeff}_i = \sqrt{1 - \alpha_{\tau_{i-1}} - \sigma_i^2}$$

Results

Images



Cifar10 64-dimension(left), Cifar10 128-dimension(middle), Celeba hq (right)

FID score

Dataset	FID score	Reconstruction loss
Cifar10 64-dimension	11.81	0.03-0.04
Cifar10 128-dimension	8.31	0.02-0.03
CelebA-HQ	11.97	0.03-0.04

Dataset Description

Cifar 10

The CIFAR-10 dataset consists of 60000 32x32 colour images in 10 classes, with 6000 images per class. There are 50000 training images and 10000 test images.

The dataset is divided into five training batches and one test batch, each with 10000 images. The test batch contains exactly 1000 randomly-selected images from each class. The training batches contain the remaining images in random order, but some training batches may contain more images from one class than another. Between them, the training batches contain exactly 5000 images from each class.

CelebA-HQ

A dataset containing 30,000 high-quality celebrity faces, resampled to 256px.