Diffusion Models for Image-to-Image Translation

Prepared by:

Khanh Duong, Tien-Huy Nguyen, Nhu-Tai Do

taidn@ueh.edu.vn

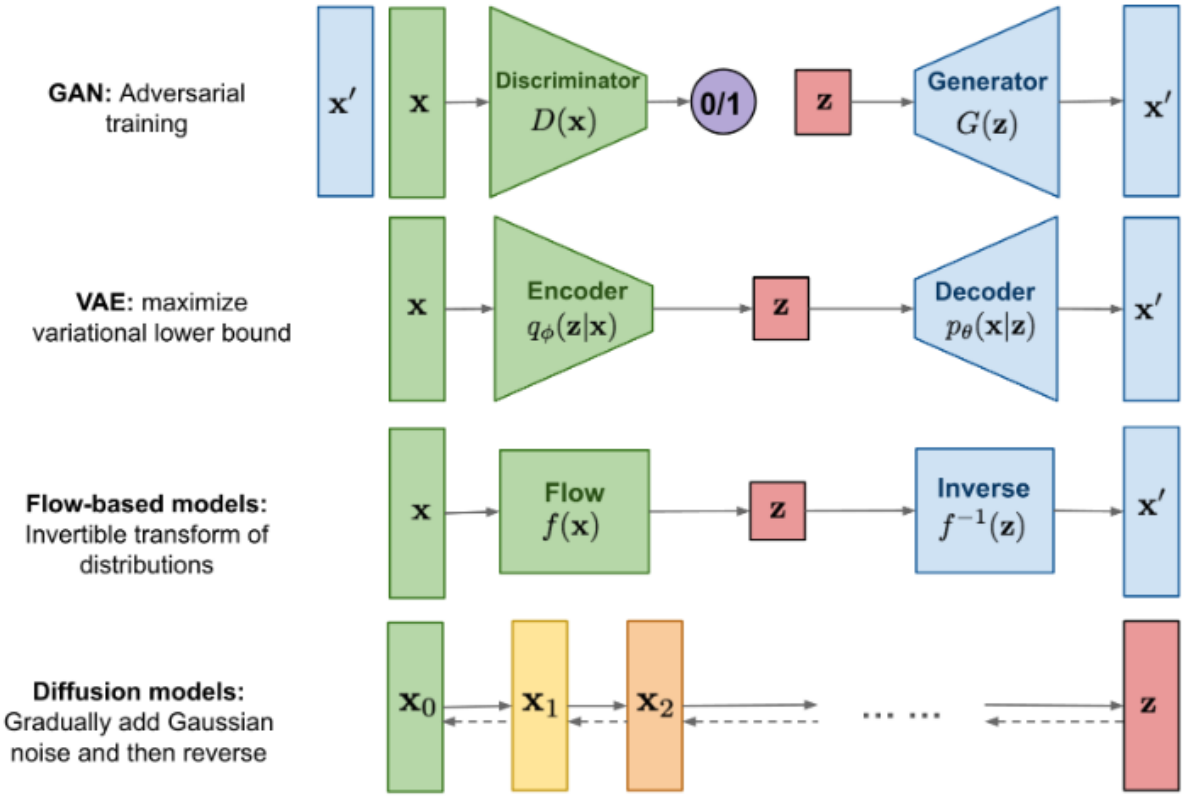
Agenda

1. Latent Diffusion
2. Image-to-Image Translation
3. Diffusion-based Image Colorization

Latent Diffusion

Overview of different types of generative models

- **GAN models**: unstable training and less diversity in generation due to their adversarial training nature.
- **VAE models**: relies on a surrogate loss
- **Flow models**: have to use specialized architectures to construct reversible transform



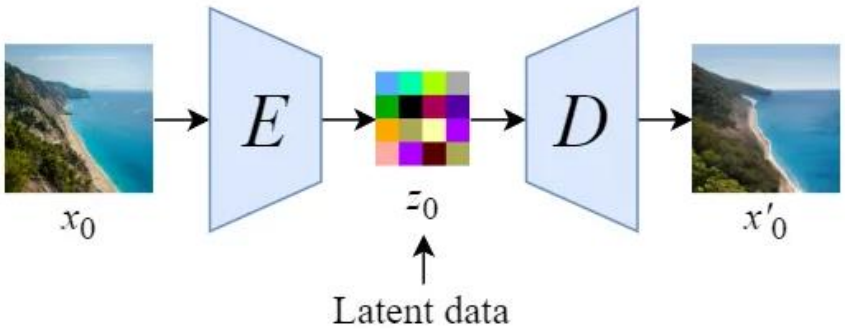
Latent Space vs Latent Diffusion

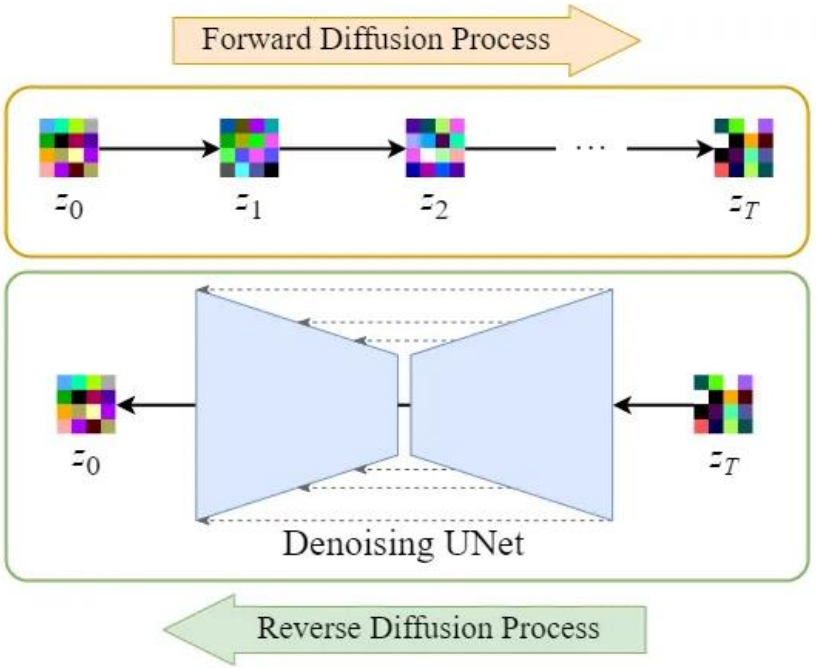
Laten Space

- **Encoder**: ecoding the image to latent data,
- **Decoder**: decoding the latent data back into an image

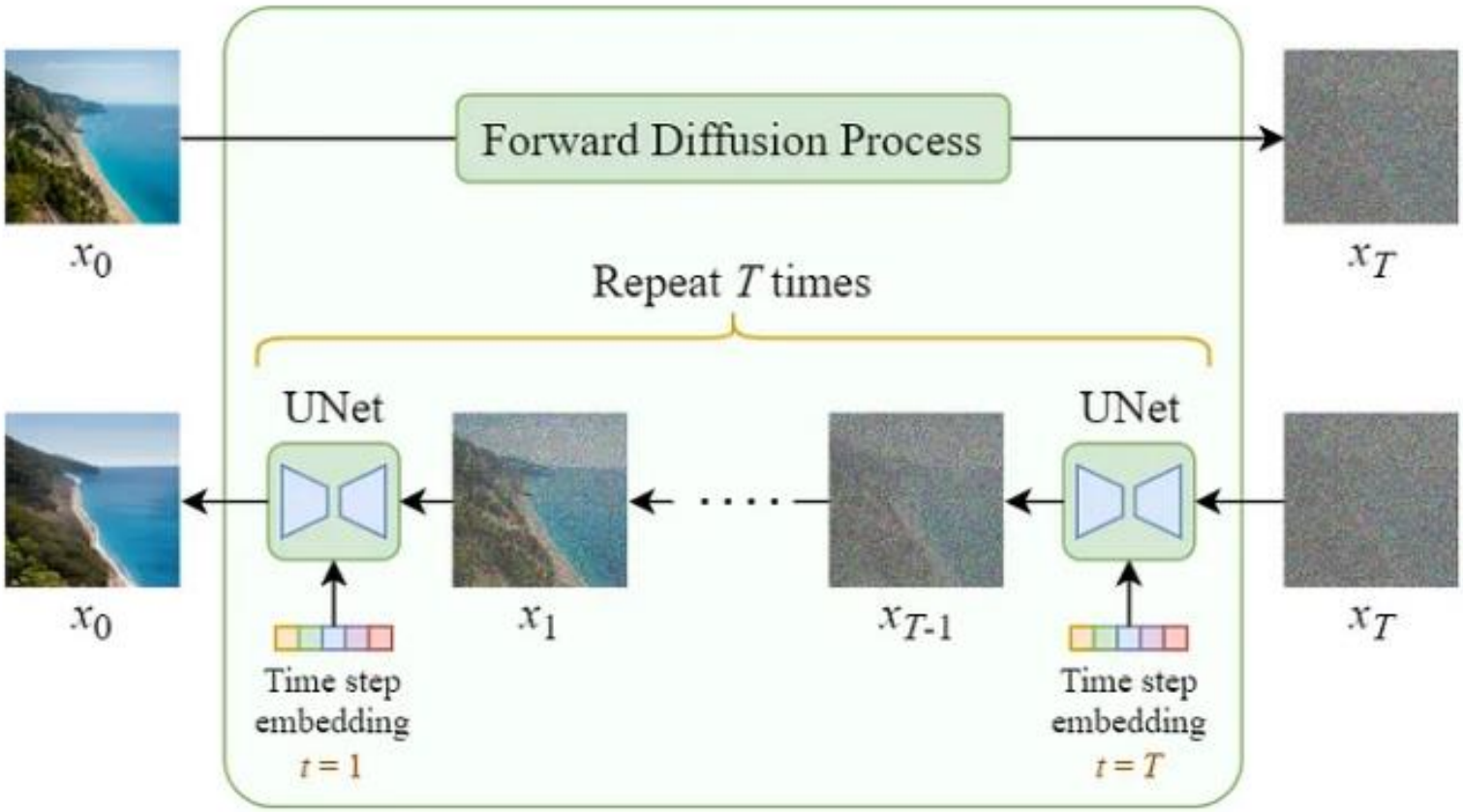
Latent Diffusion

- **Forward Diffusion Process** → add noise to the latent data
- **Reverse Diffusion Process** → remove noise from the latent data



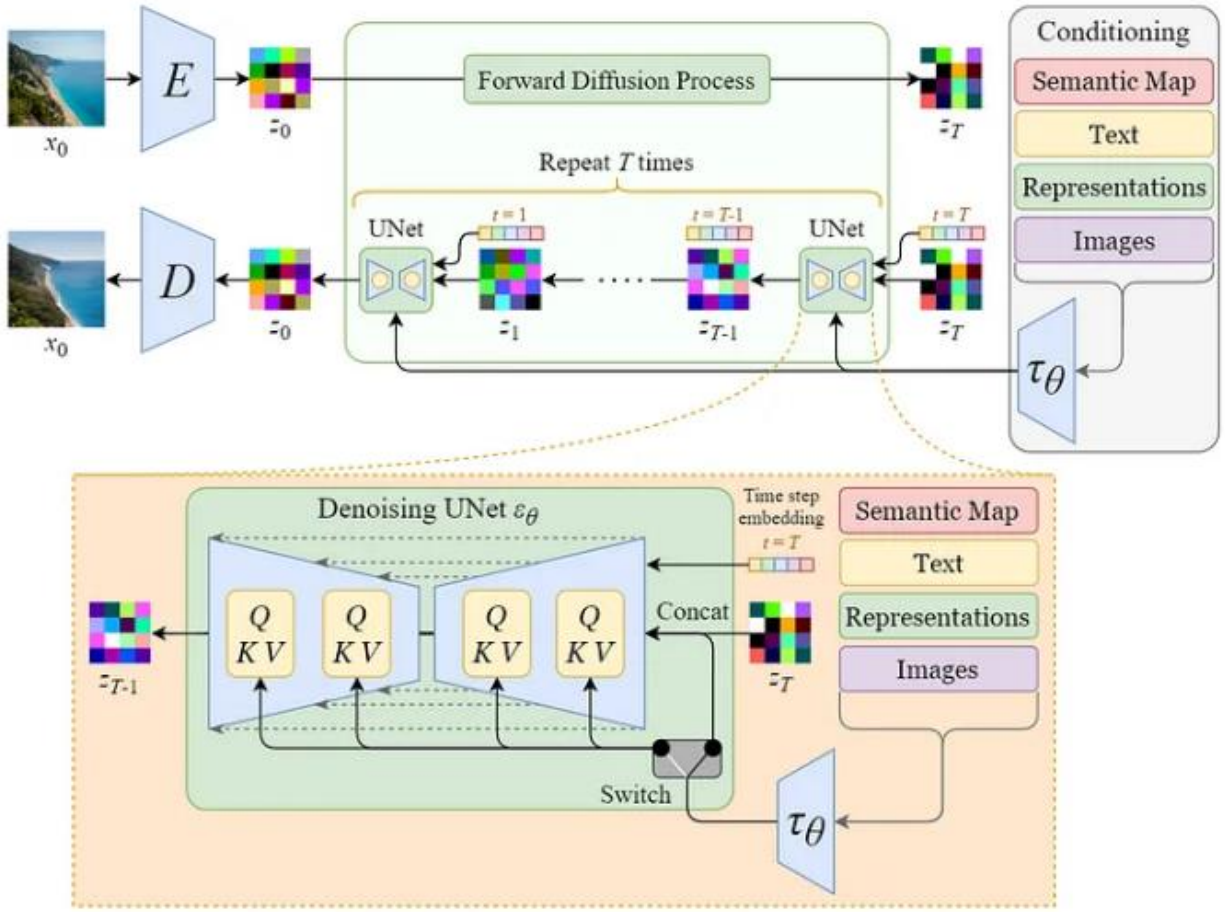


Pure Diffusion Model



Stable Diffusion (Latent Diffusion Model)

- conducts the diffusion process in the latent space
- backbone diffusion model is modified to accept conditioning inputs



Latent Diffusion Models

- An **Autoencoder** that performs the compression of the inputted images into a smaller latent representation
- A **Diffusion Model** that will learn the probability data distribution of the latent representations
- A **Text Encoder** which creates an embedding vector that will condition the sampling process

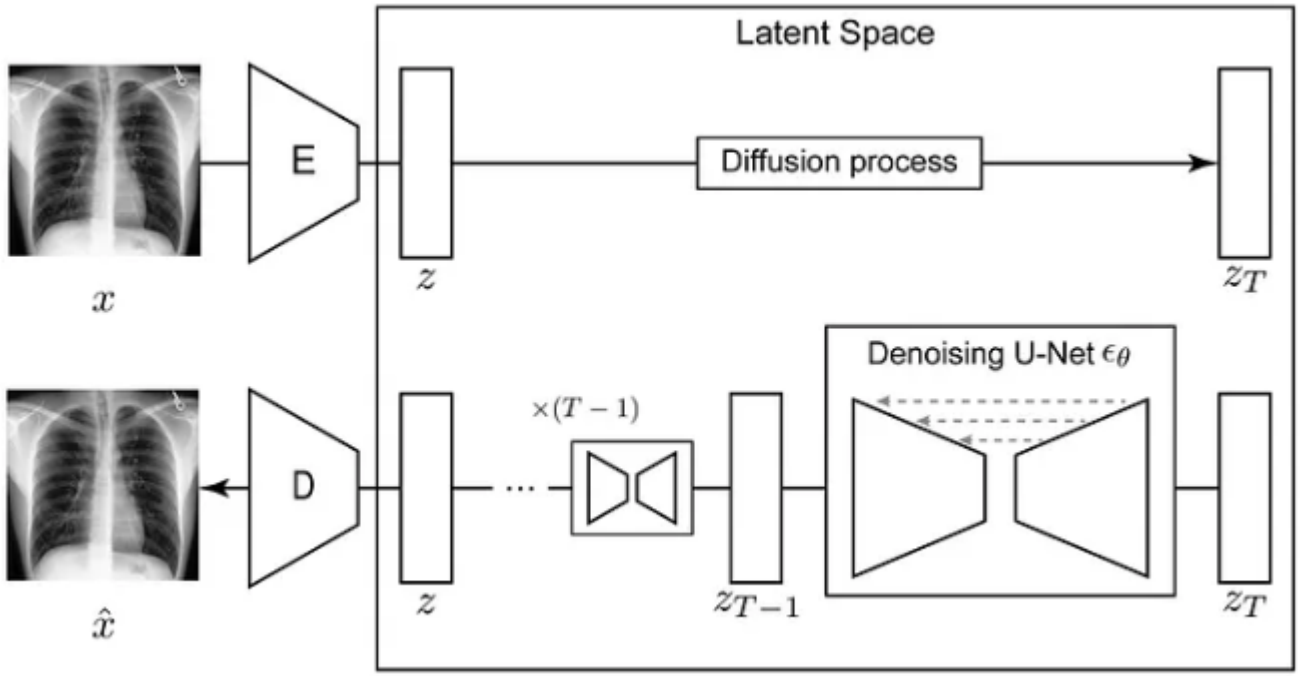


Image-to-Image Translation

Diffusion models for different applications

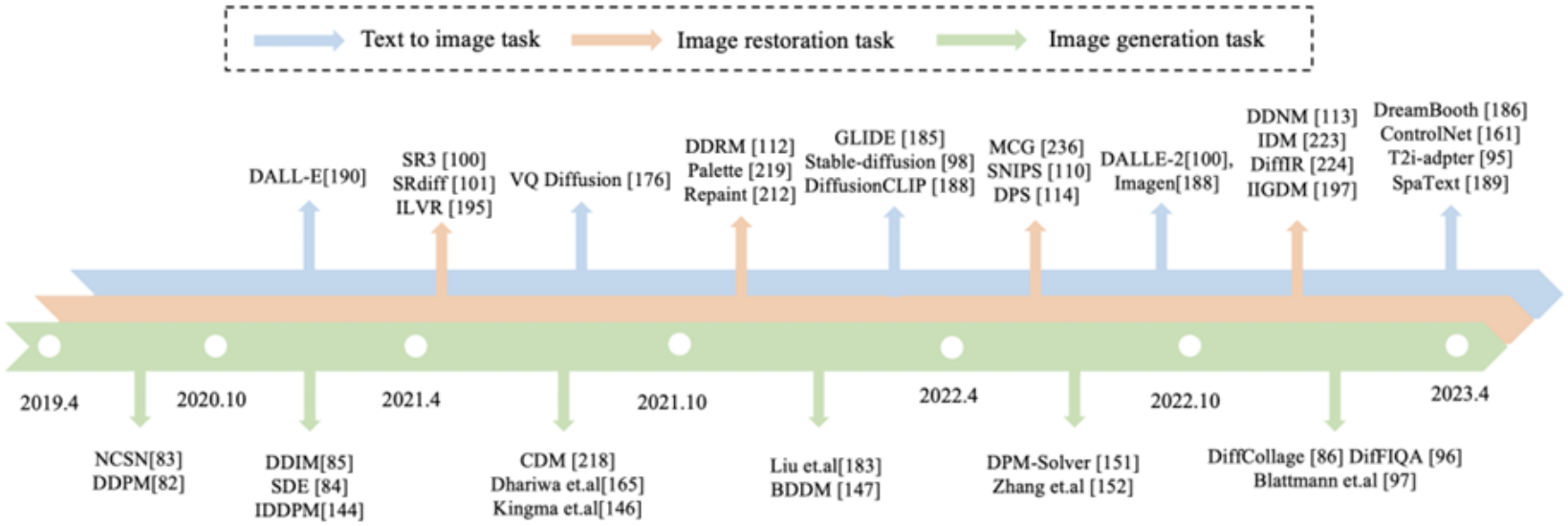
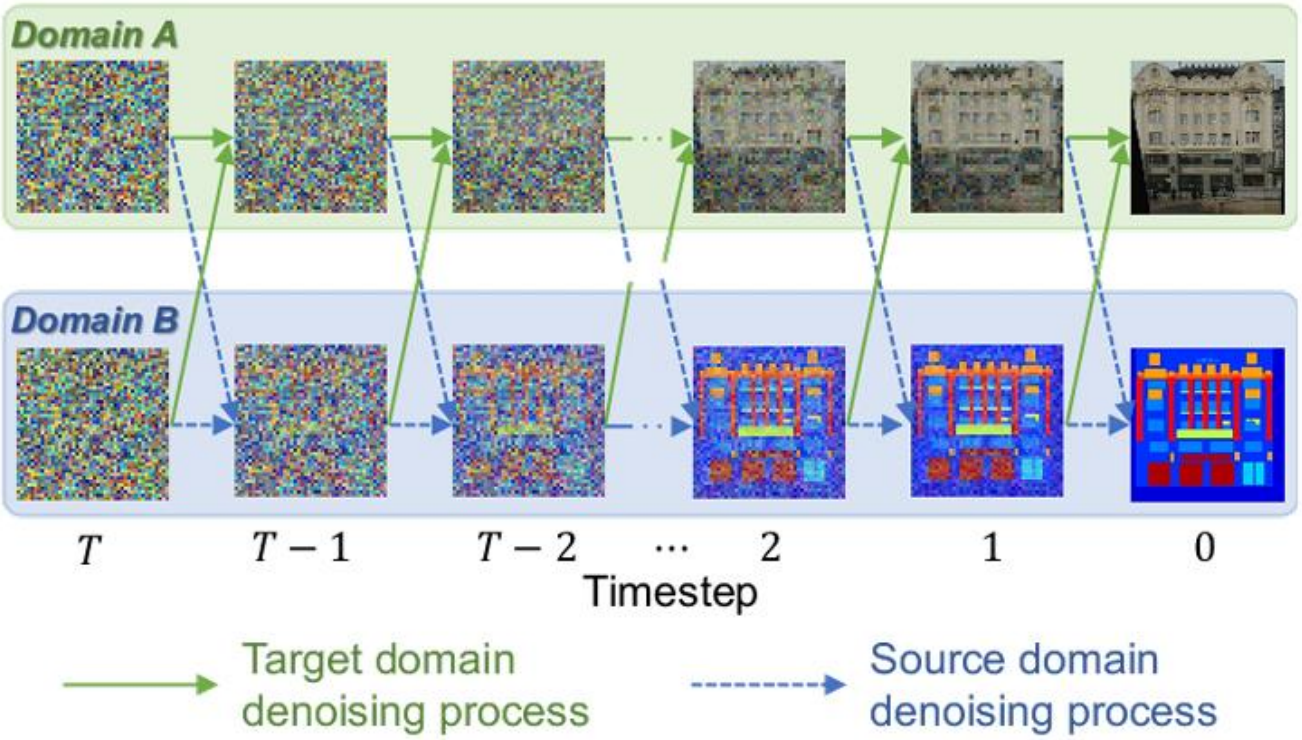


Fig. 2: The representative works in diffusion models for different applications(Green denotes image generation tasks, orange denotes image restoration tasks, blue denotes text-to-image tasks)

[1] Li, X., Ren, Y., Jin, X., Lan, C., Wang, X., Zeng, W., ... & Chen, Z. (2023). ***Diffusion Models for Image Restoration and Enhancement--A Comprehensive Survey***. *arXiv preprint arXiv:2308.09388*.

Image Translation with Diffusion Probabilistic Models

- Modeling the joint distribution of both domains as a Markov chain by minimising a denoising score matching objective conditioned on the other domain



[1] Sasaki, Hiroshi, Chris G. Willcocks, and Toby P. Breckon. "***Unit-ddpm: Unpaired image translation with denoising diffusion probabilistic models***." *arXiv preprint arXiv:2104.05358* (2021).

Image Segmentation with Diffusion Probabilistic Models

- Main Idea: use a denoising network that takes in both the input image and the current estimate of the binary segmentation map, and the output is used to update the estimate

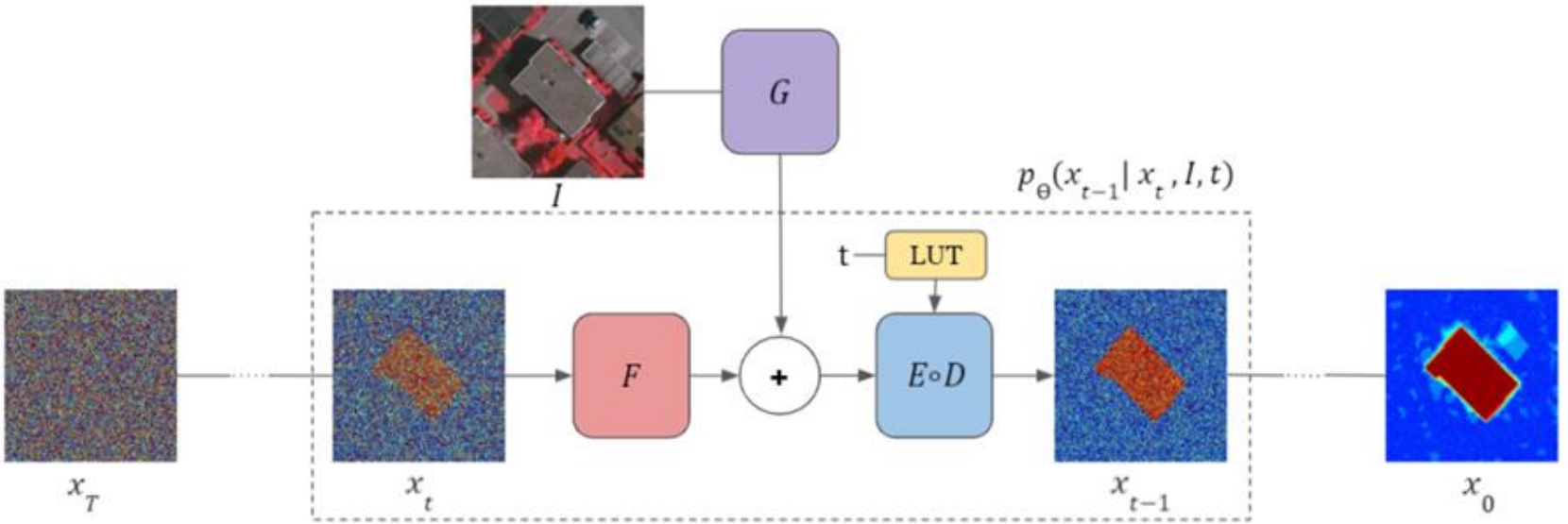


Figure 1. Our proposed diffusion method for image segmentation encodes the input signal, x_t, with F. The extracted features are summed with the feature map of the conditioned image I generated by network G. Networks E and D are a U-net encoder and decoder [35, 38], respectively, that refine the estimated segmentation map, obtaining x_{t-1}.

[1] Amit, Tomer, Tal Shaharbany, Eliya Nachmani, and Lior Wolf. "***Segdiff: Image segmentation with diffusion probabilistic models***." *arXiv preprint arXiv:2112.00390* (2021).

Representations from DDPM in Segmentation

- Representations from DDPM can capture high-level semantic information valuable

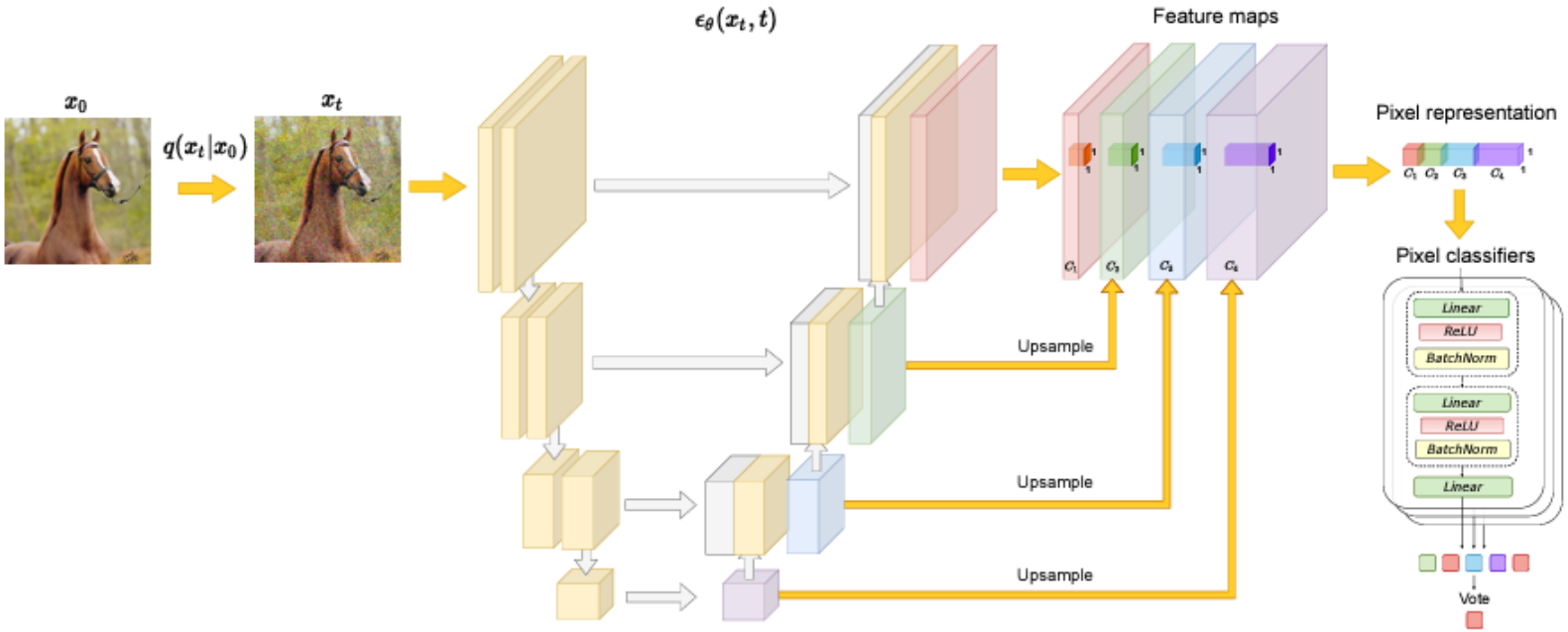


Figure 1: **Overview of the proposed method.** (1) $x_0 \rightarrow x_t$ by adding noise according to $q(x_t|x_0)$. (2) Extracting feature maps from a noise predictor $\epsilon_\theta(x_t, t)$. (3) Collecting pixel-level representations by upsampling the feature maps to the image resolution and concatenating them. (4) Using the pixel-wise feature vectors to train an ensemble of MLPs to predict a class label for each pixel.

[1] Baranchuk er al. (2021). ***Label-efficient semantic segmentation with diffusion models***. *arXiv preprint arXiv:2112.03126*. (ICLR2021)

Medical Image Segmentation with Diffusion Probabilistic Model

- **Dynamic conditional encoding**: establishes the state-adaptive conditions for each sampling step
- **Feature Frequency Parser (FF-Parser)**: to eliminate the negative effect of high-frequency noise component

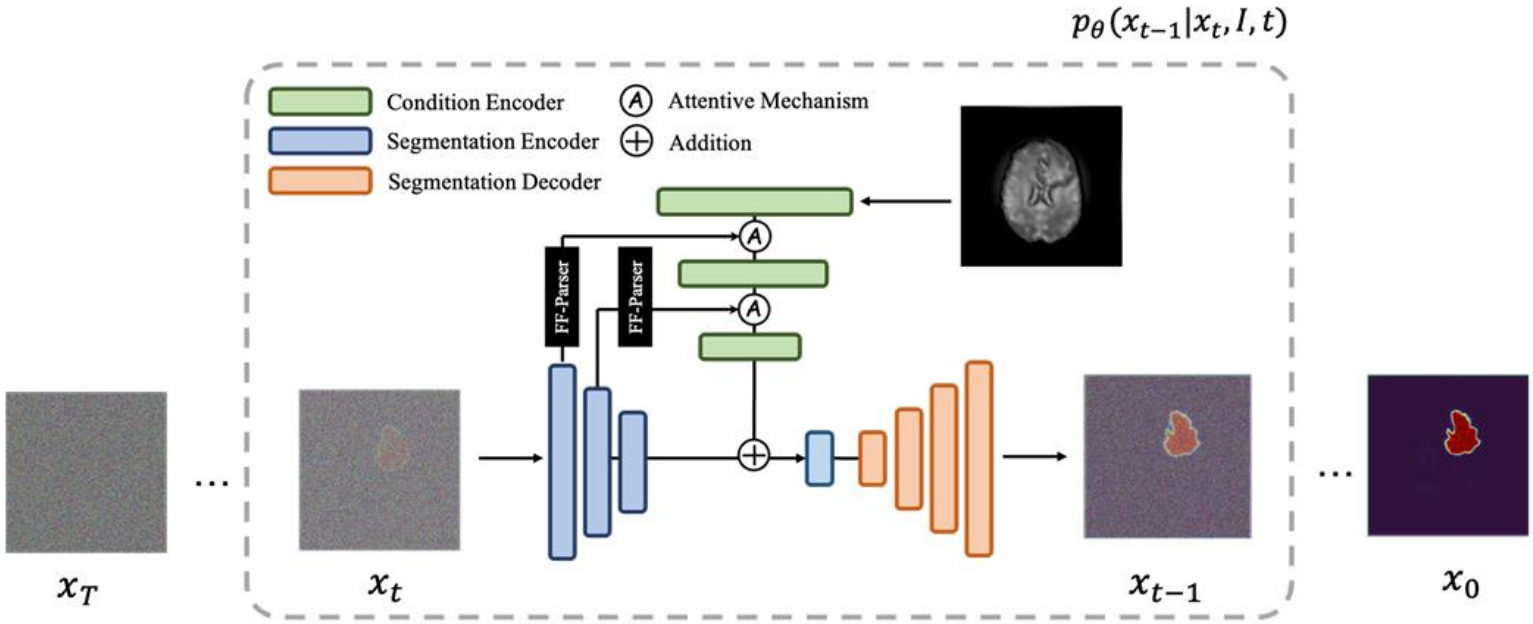


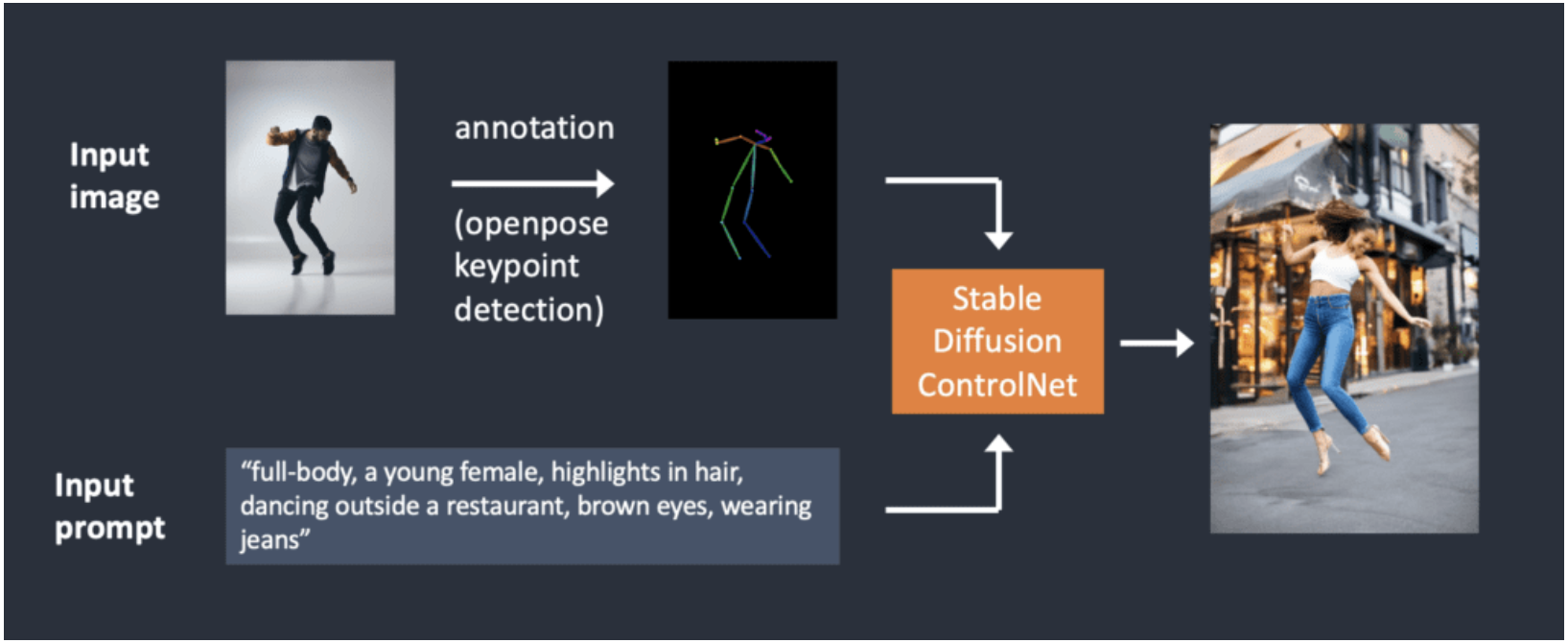
Fig. 1: An illustration of MedSegDiff. For the clarity, the time step encoding is omitted in the figure.

[1] Wu, Junde et al.. "***Medsegdiff: Medical image segmentation with diffusion probabilistic model***." In *Medical Imaging with Deep Learning*, pp. 1623-1639. PMLR, 2024.

Diffusion-based Image Colorization

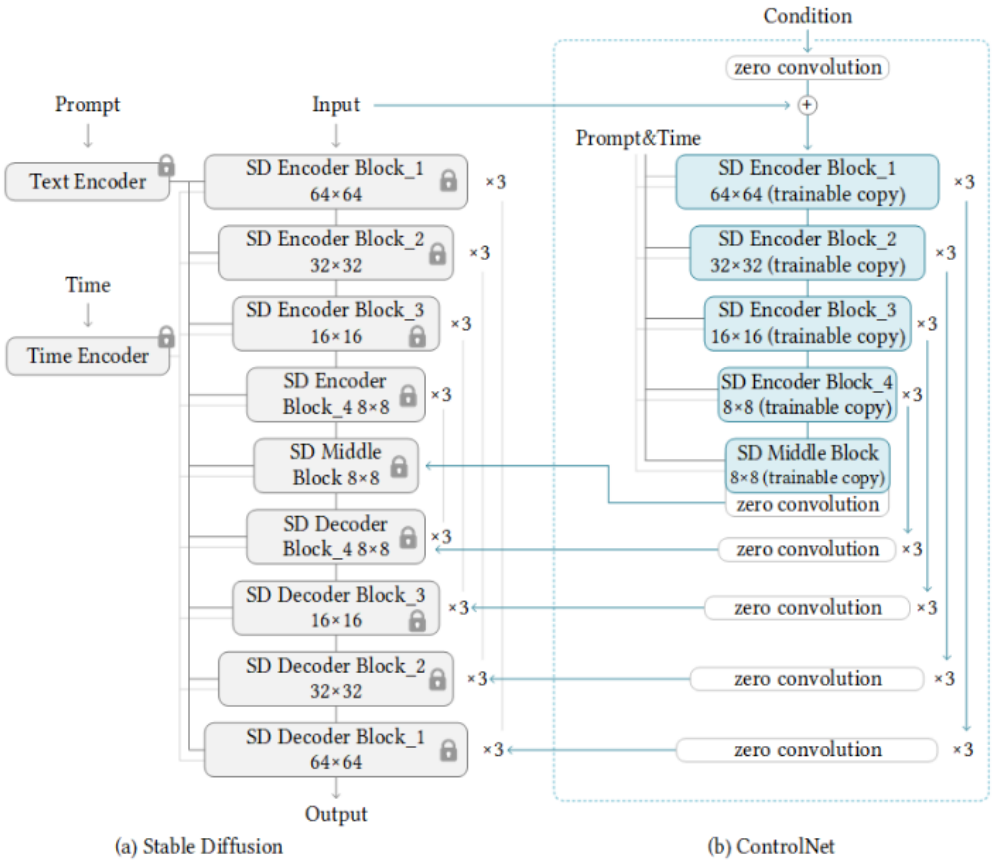
ControlNet

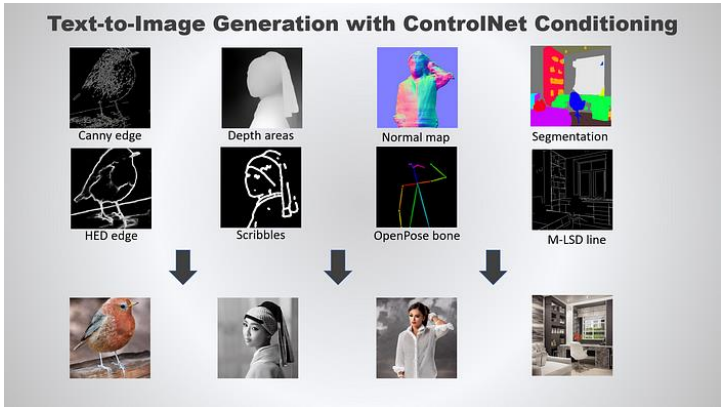
- ControlNet: a neural network structure to control diffusion models by adding extra conditions to augment Stable Diffusion with conditional inputs such as scribbles, edge maps, segmentation maps, pose key points, etc during text-to-image generation



ControlNet

- Architectures

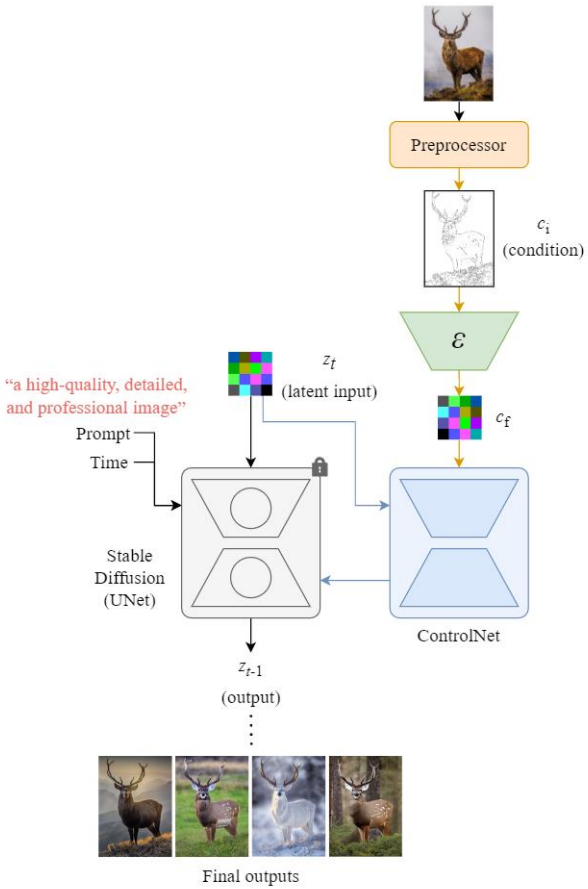


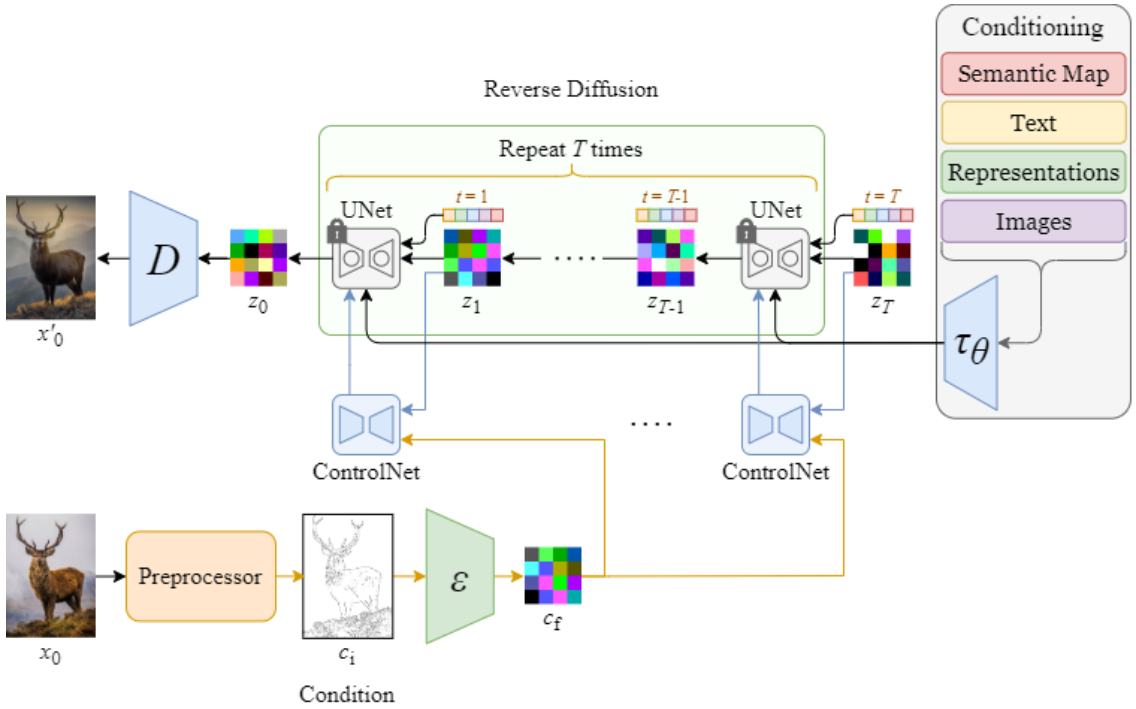


[1] Zhang et al. "***Adding conditional control to text-to-image diffusion models***." In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 3836-3847. 2023.

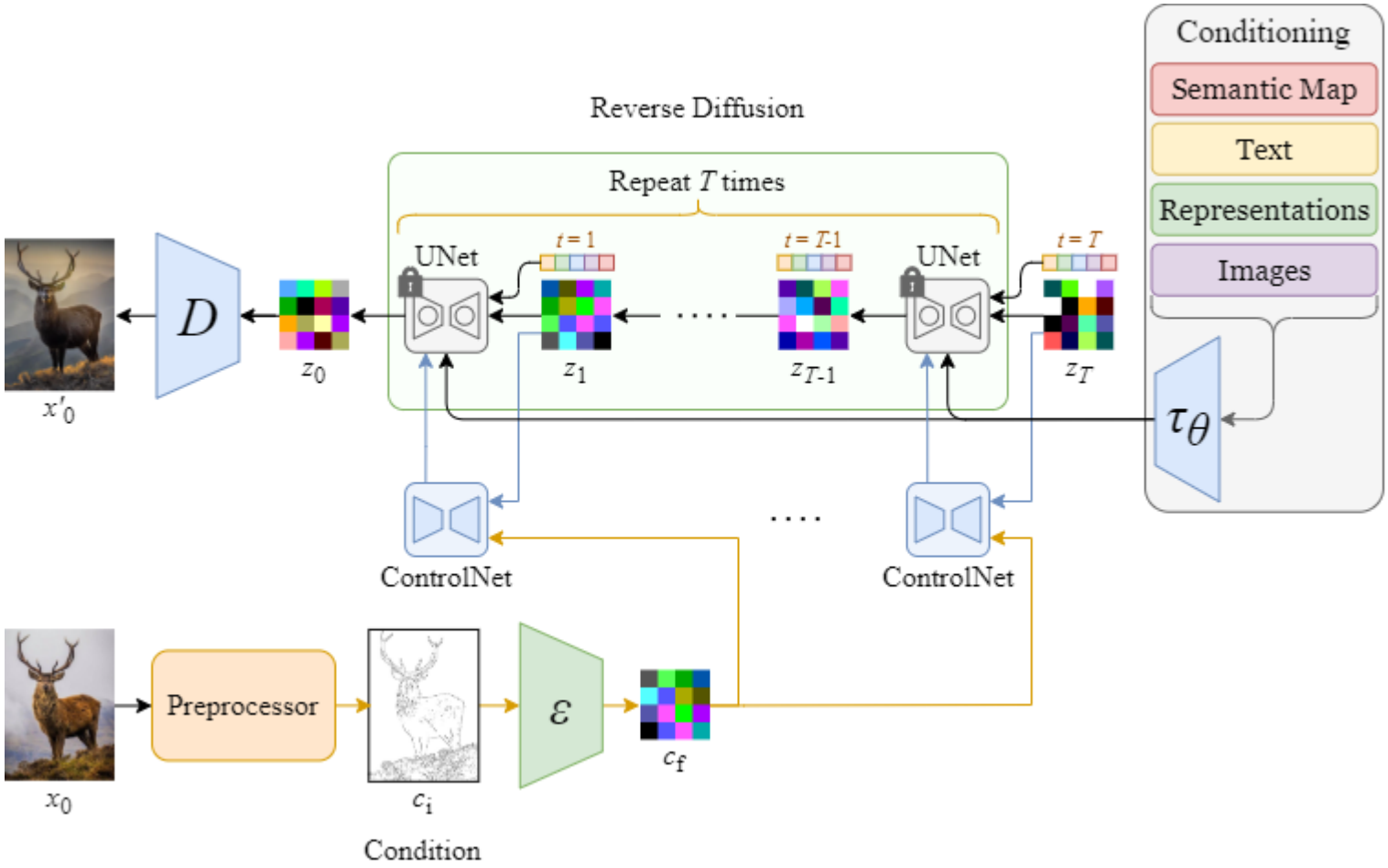
ControlNet

- Reverse diffusion process





ControlNet: Training



Diffusing Colors: Image Colorization with Text Guided Diffusion

- Using VAE Encoder

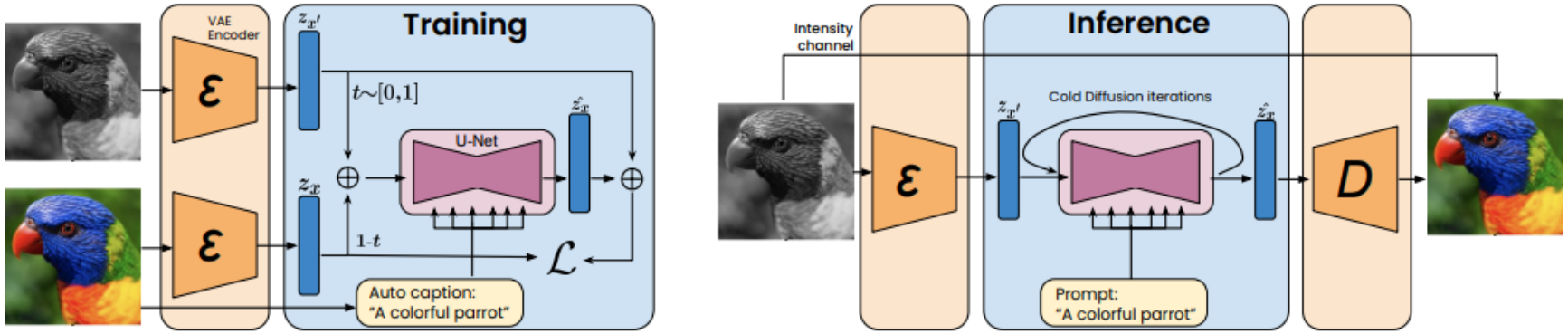
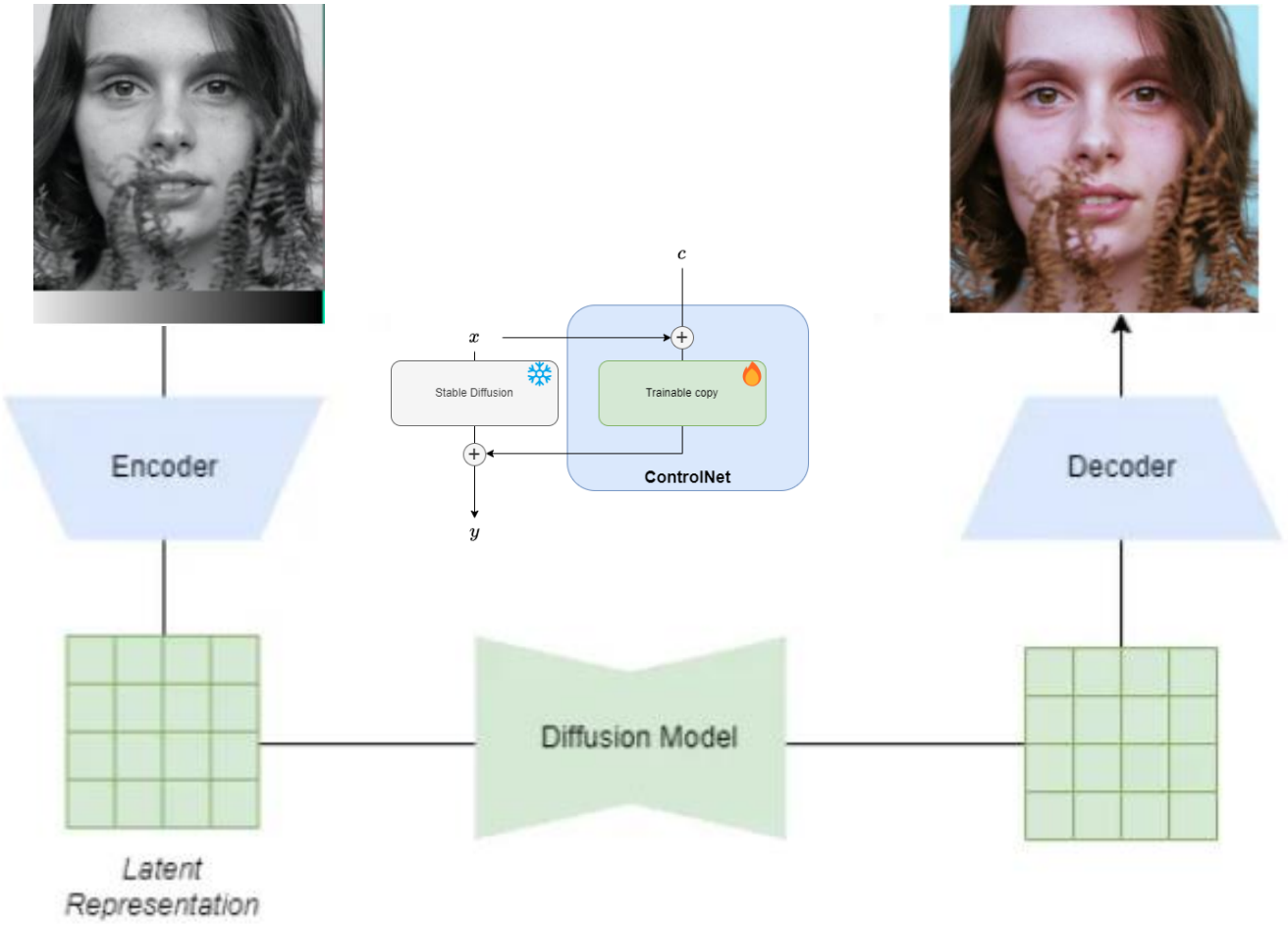


Figure 2: Overview. During training we encode the RGB and grayscale images into the latent space and feed the U-Net with a random convex combination of the two, together with an auto generated caption of the color image, and the timestep t. At inference time we encode the input grayscale image as z_x and iteratively colorize it to get $\hat{z_x}$ which we decode, and combine with the grayscale as a luma channel. Image credit: Unsplash © David Clode.

Zabari, Nir, Aharon Azulay, Alexey Gorkor, Tavi Halperin, and Ohad Fried. "***Diffusing Colors: Image Colorization with Text Guided Diffusion***." In *SIGGRAPH Asia 2023 Conference Papers*, pp. 1-11. 2023.

Diffusion-based Image Colorization using ControlNet

Using ControlNet to learn ab channels from L channel



https://huggingface.co/lllyasviel/ControlNet

THANKS FOR LISTENING!

Waiting for question!