# Diffusion Models Intro

Alan Yuille

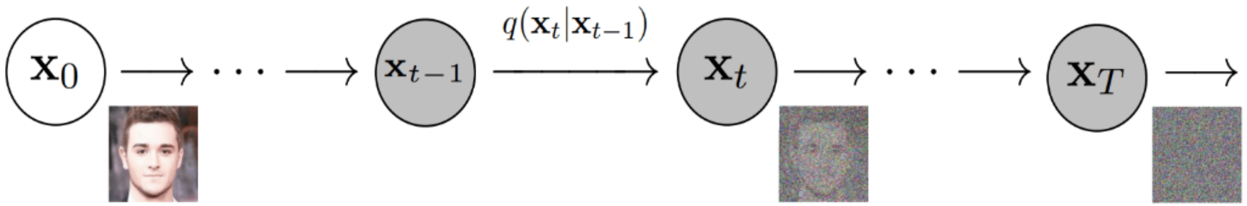
# Diffusion Models conditioned on Text

- Diffusion Models conditioned on text are able to generate create complex and realistic images.
- They can take advantage of the huge advances made by Large Language Models (Auto-Regressive).
- This prize-winning images was created almost entirely by DMs.

*"Théâtre D'opéra Spatial" entry for the Colorado State Fair.*

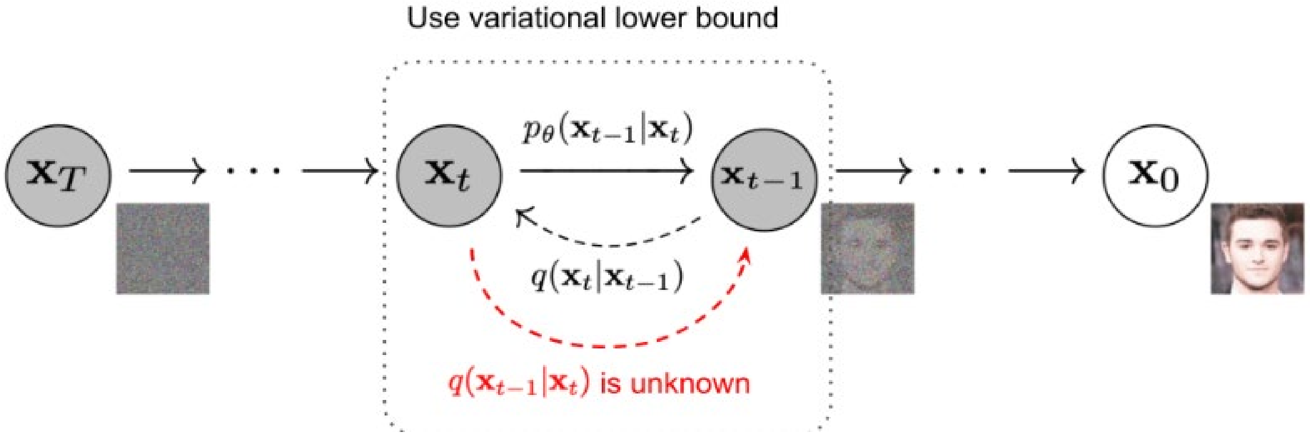
# Diffusion Models : Auto-Encoder

- Forward Diffusion Process: input image, output latent variables.



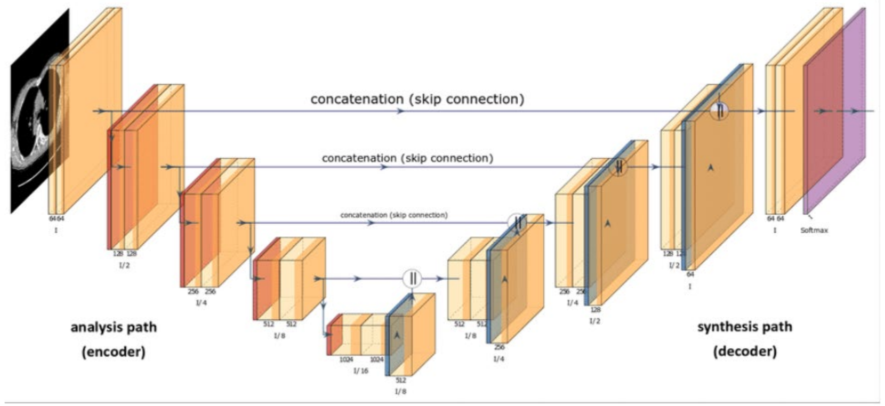
*Forward diffusion process [4]*

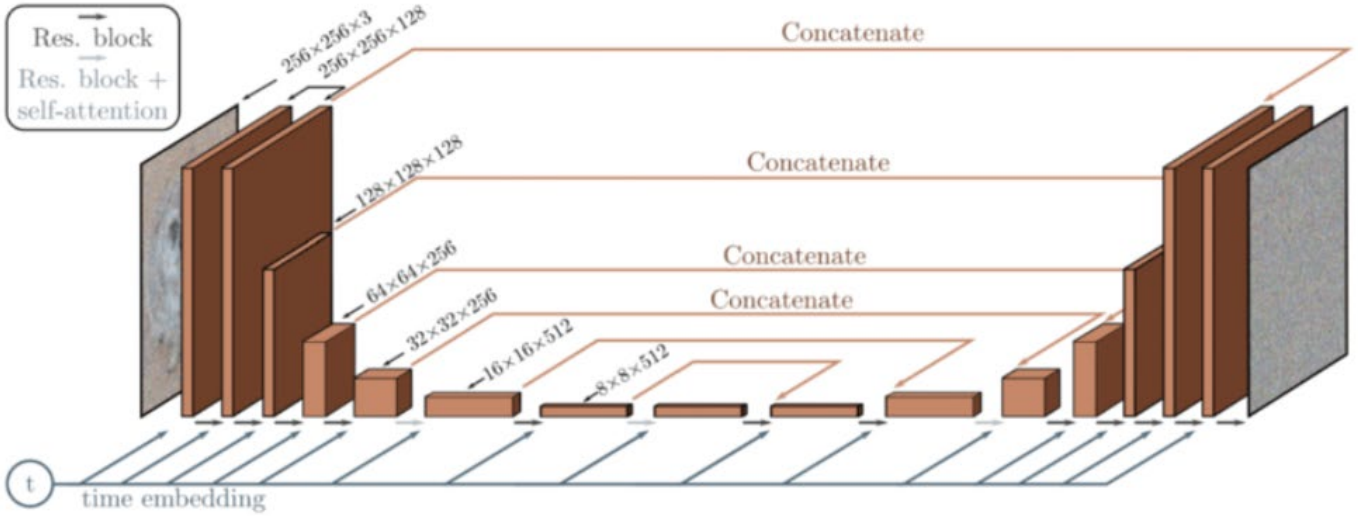
- Reverse Diffusion Process: input latent variables, output image


Use variational lower bound
$\mathbf{x}_T$
$\mathbf{x}_t$
$p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t)$
$\mathbf{x}_{t-1}$
$q(\mathbf{x}_t|\mathbf{x}_{t-1})$
$q(\mathbf{x}_{t-1}|\mathbf{x}_t)$ is unknown
$\mathbf{x}_0$

# Diffusion Architecture

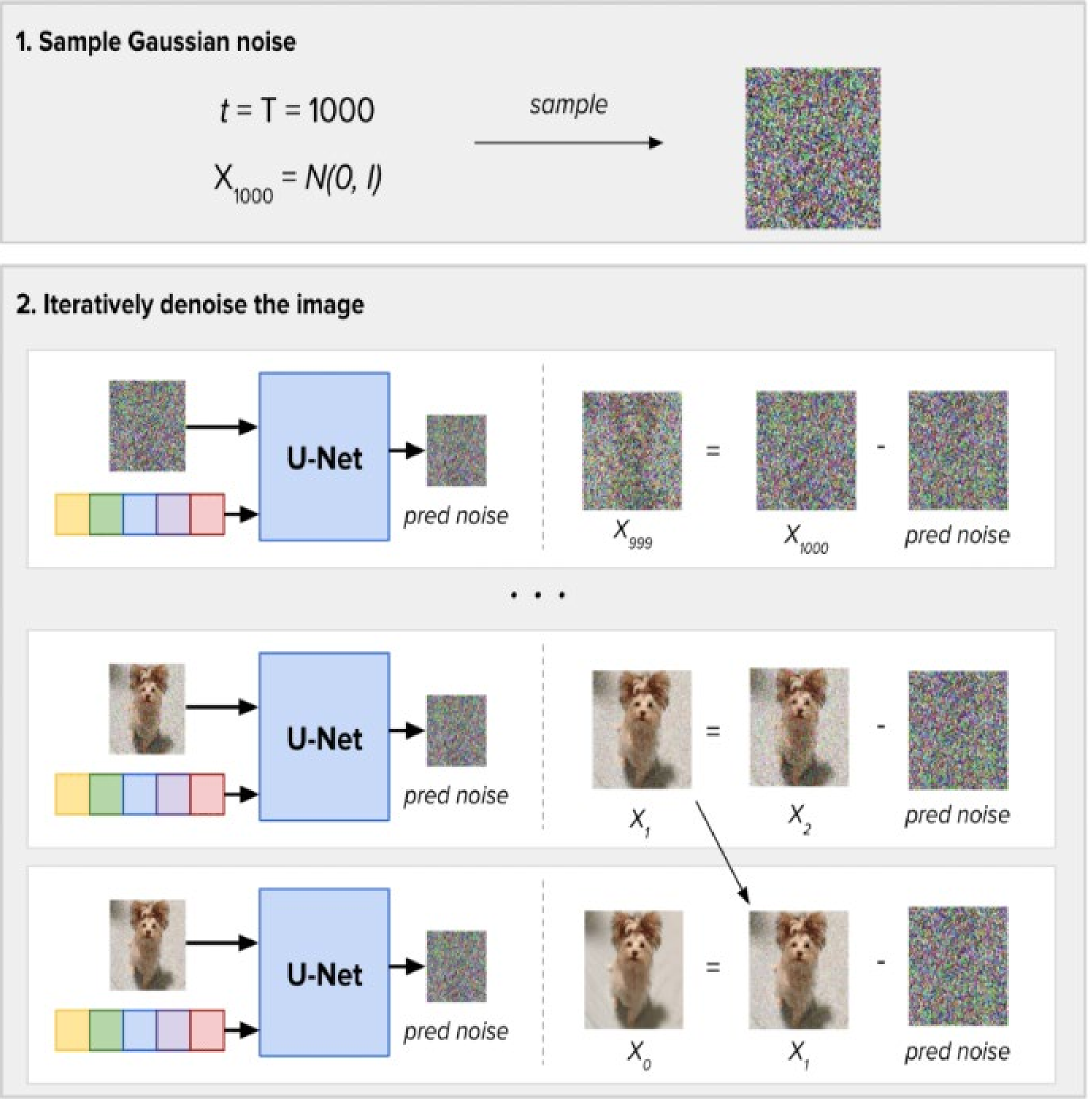
- These are variants of Unet.
- Out-of-scope of course.





# Generate Images

- Sample the latent variables -- random gaussian noise.
- Iterative sampling generates an image.



# Stable Diffusion

- Stable Diffusion performs diffusion in the latent space.

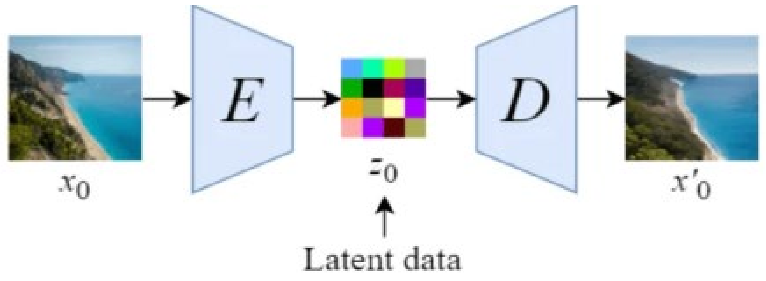
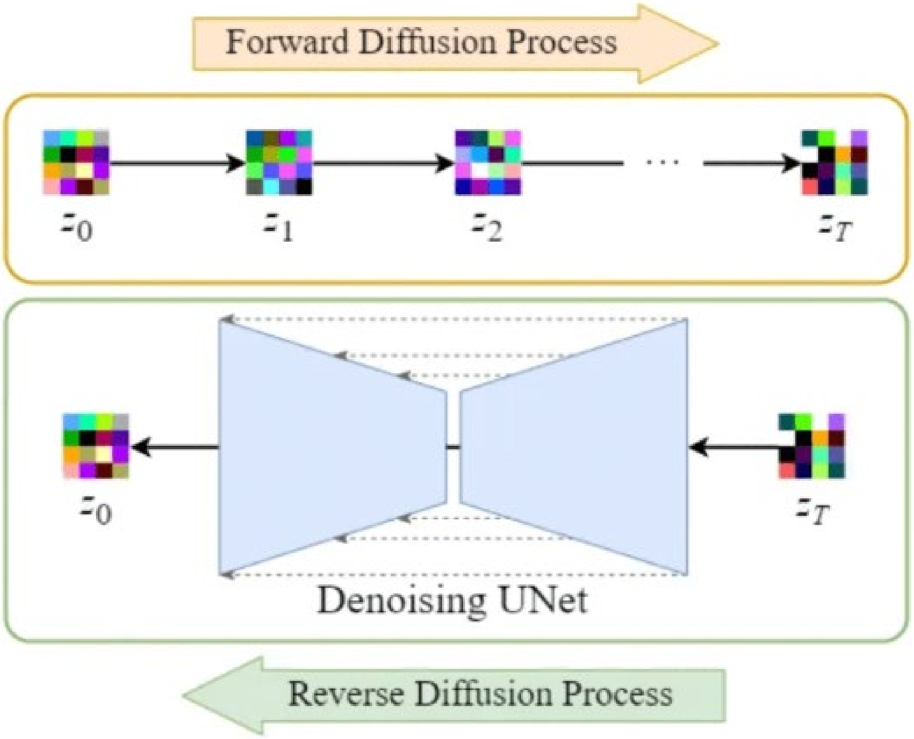
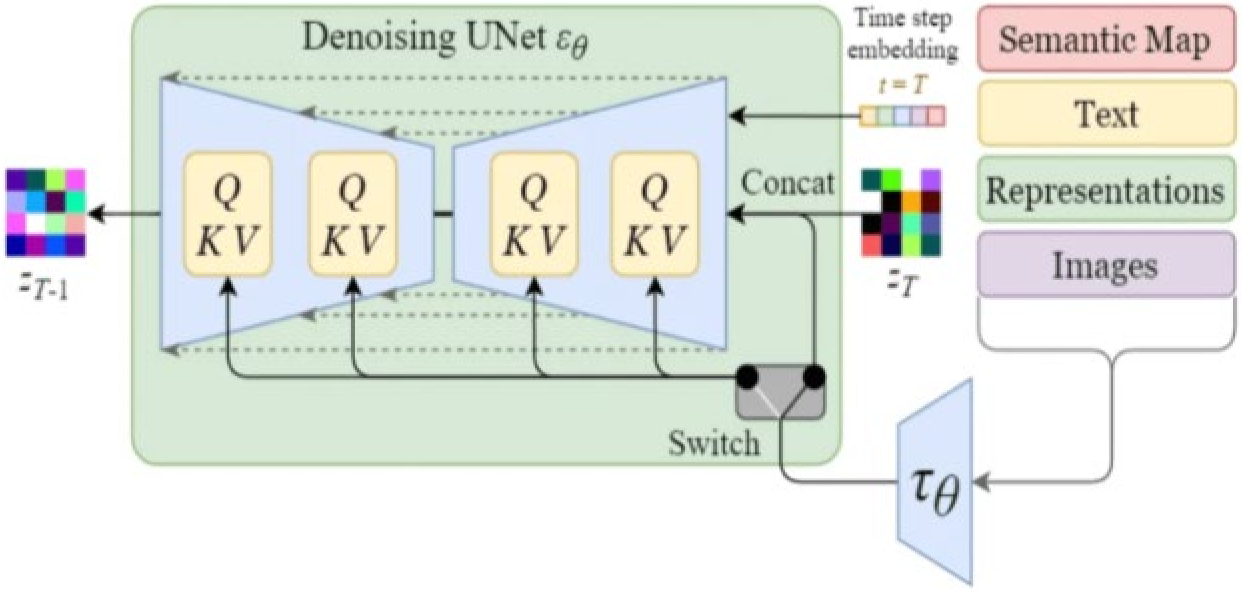


*Illustration of an autoencoder as proposed by the Stable Diffusion paper [14]*



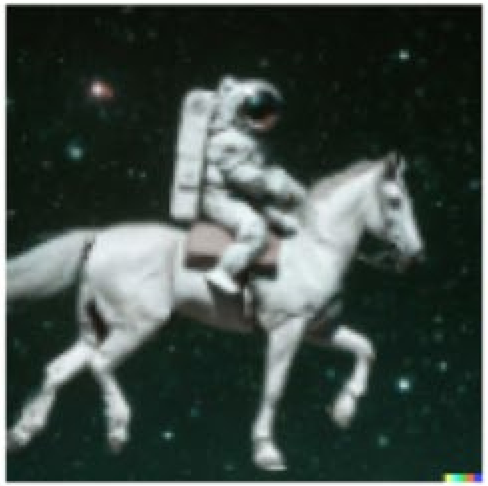
# Stable Diffusion is conditioned on text.

- This is performed by a cross-attentional mechanism.
- This enables prompting on Text.

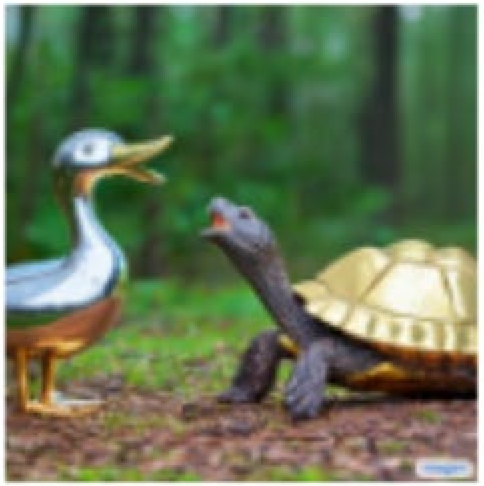


# Generation by Text Prompts

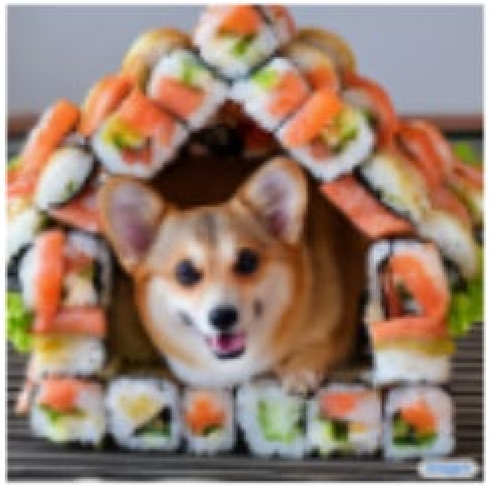
- A few examples.



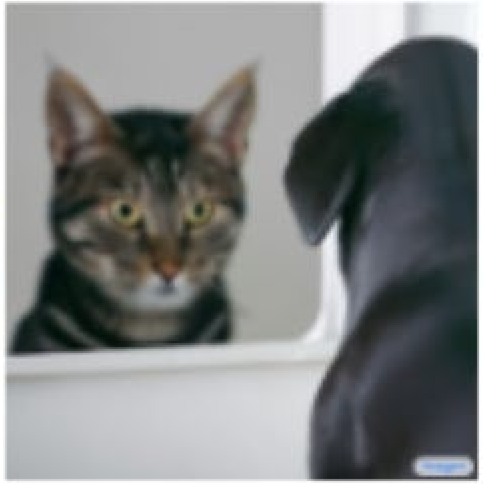
An astronaut riding a horse in photorealistic style.



A chrome-plated duck with a golden beak arguing with an angry turtle in a forest.



A cute corgi lives in a house made out of sushi.



A dog looking curiously in the mirror, seeing a cat.

# Limitations

- DMs can generate a very rich variety of realistic images controlled by text prompts. And can be extended to generate videos.
- But, for computer vision, these are lacking as generative models. From analysis by synthesis perspective we would like generative models that are conditioned on the world state.
- DMs are conditioned on latent variables, which are hard to understand, and on text prompts. This limits their usefulness.