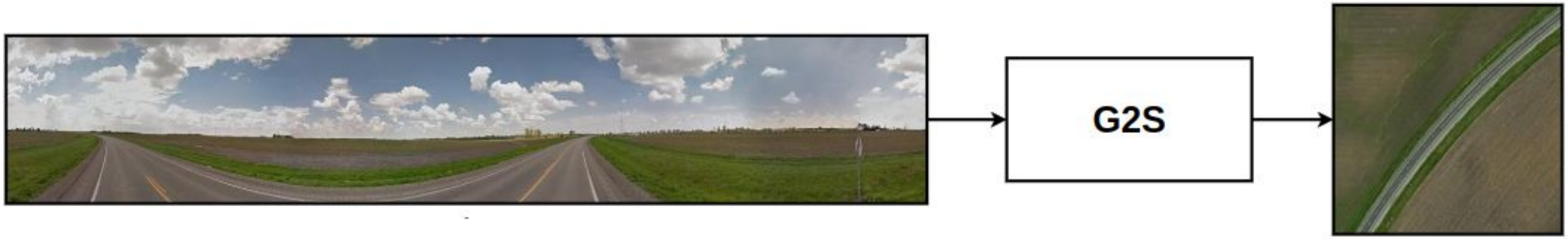
Ground-to-Satellite Image Translation with Diffusion Model

Daniele S. Cardullo - 2127806

The Problem

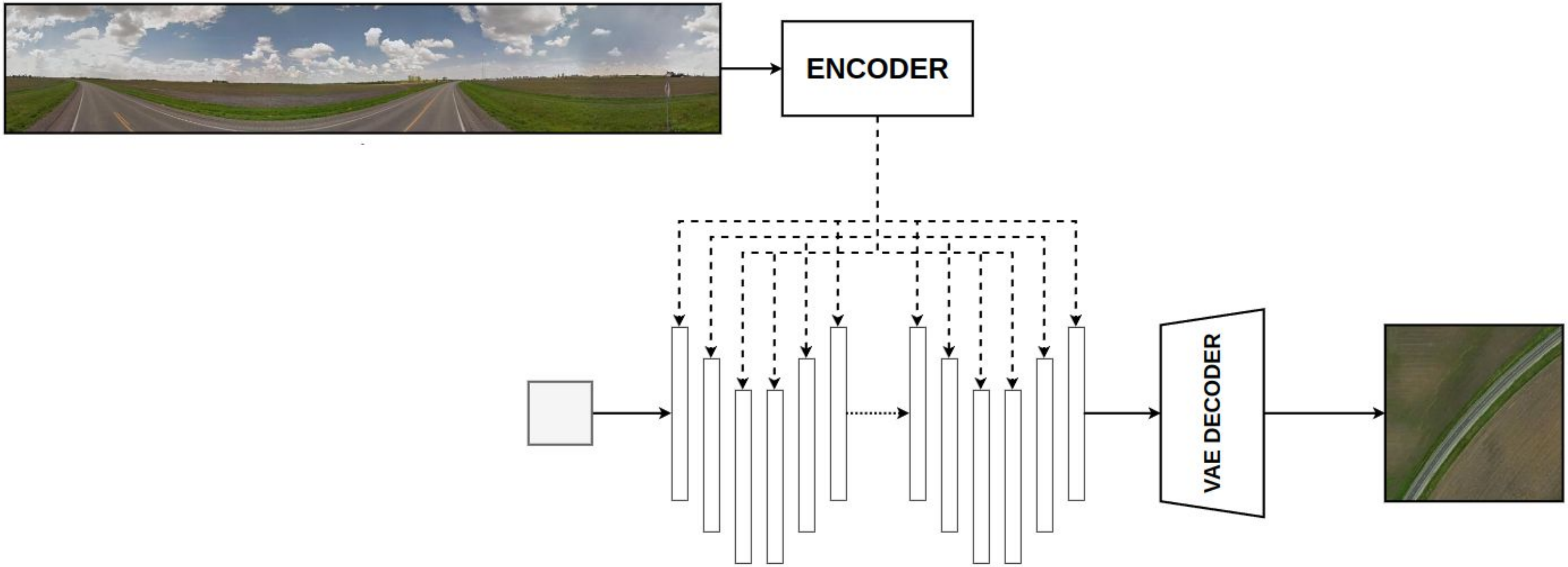
Translating ground spherical images to their corresponding satellite view is a not trivial task to accomplish and involves various steps. Nonetheless it has a practical importance due to its application in mobile robotics, autonomous car driving and localization without GPS.



The Approach

The chosen approach to this problem can be described in three steps:

- Pretraining an autoencoder to encode conditionings (streetview images);
- Training a latent diffusion network to generate meaningful and conditionally driven latents from noise;
- Using a VAE to decode the generated latents and produce the output images.

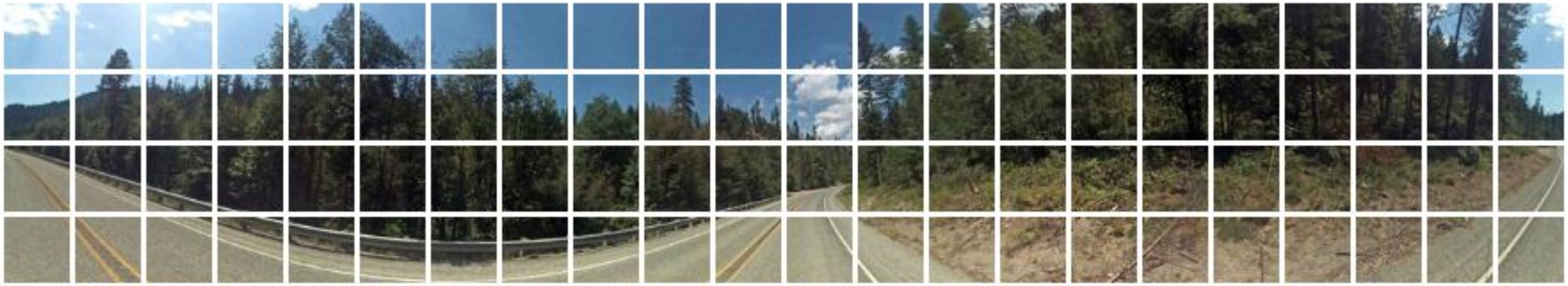


Patched Autoencoder Training

The first step in my approach is to encode the streetview images in order to feed them as conditioning information to the latent diffusion model.

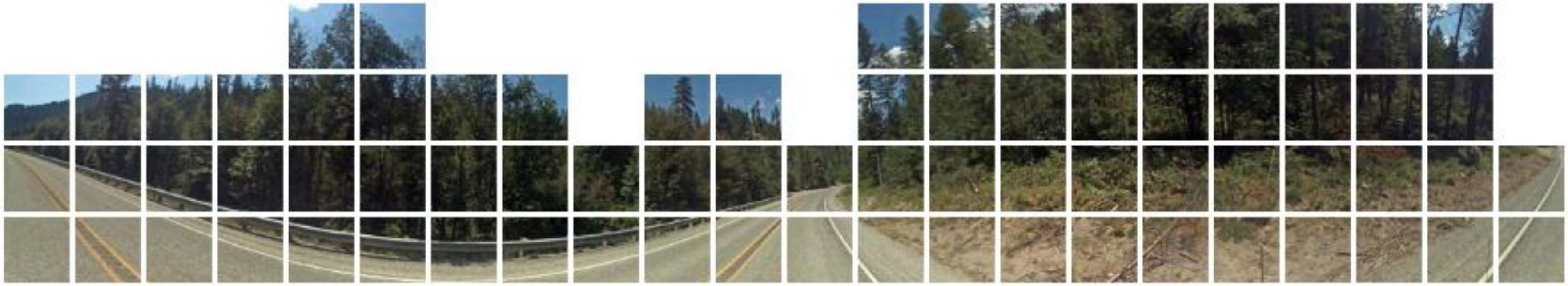
As autoencoder I decided to use a PatchAE trained on masked images.

1. Subdivide the image in 1078 patches of dimension 16 x 16



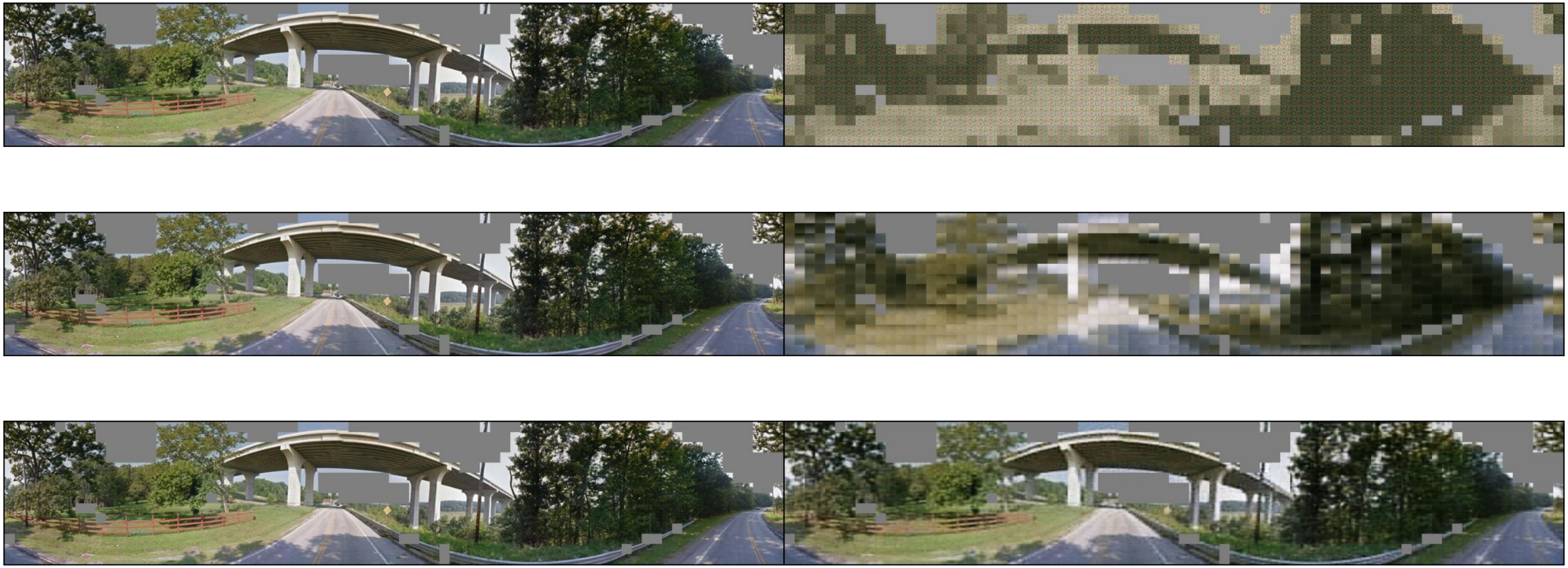
2. Concatenate the patched image with a sinusoidal position encoding

3. Using the corresponding semantic map, mask the patches in which the semantic label is sky (0) for the majority of pixels in that patch.



4. Feed the shuffled remaining patches to the model and produce the decoder output patches (where the position encoding is already removed), perform MSE loss on the patches list.

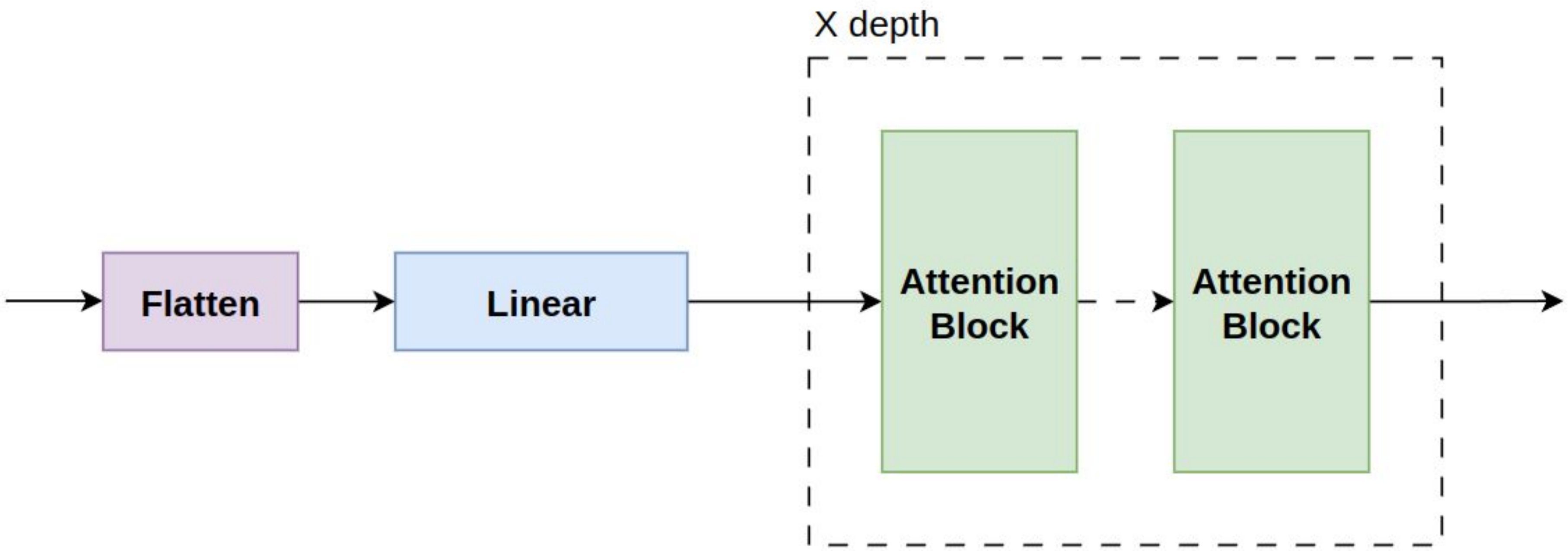
5. Reconstruct the generated masked image.



Patched Autoencoder Structure

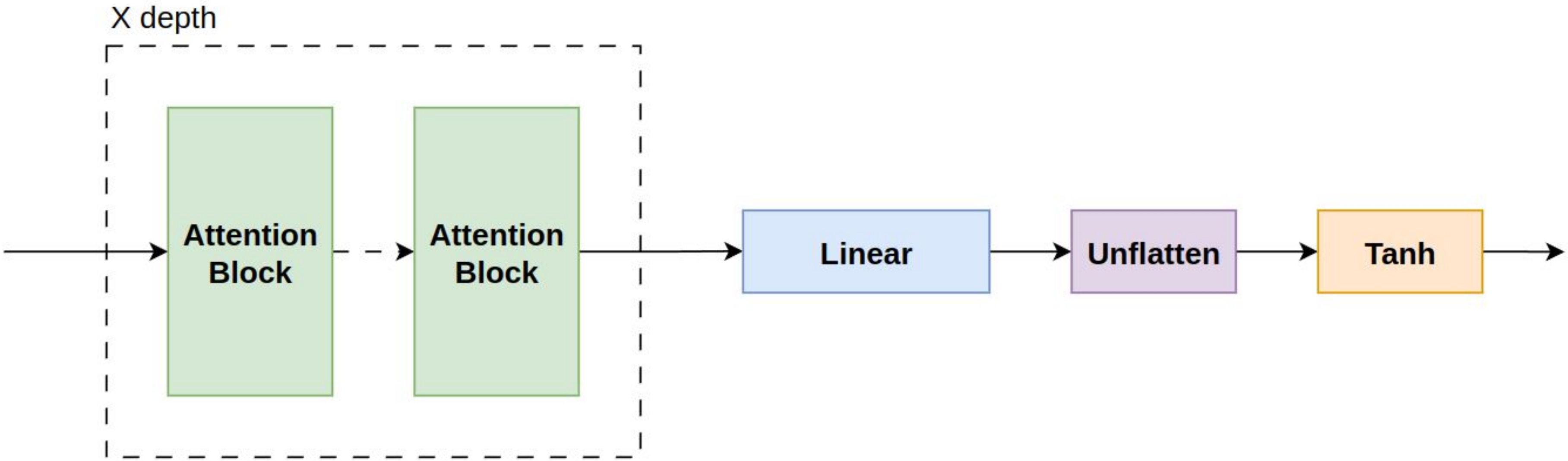
The structure of this autoencoder is inspired by the Masked Autoencoder (K. He, X. Chen et al. - 2021) which was itself inspired by the ViT structure.

The encoder block is formed by a flattening layer and linear embedding, followed by a list of transformer blocks (attention blocks)

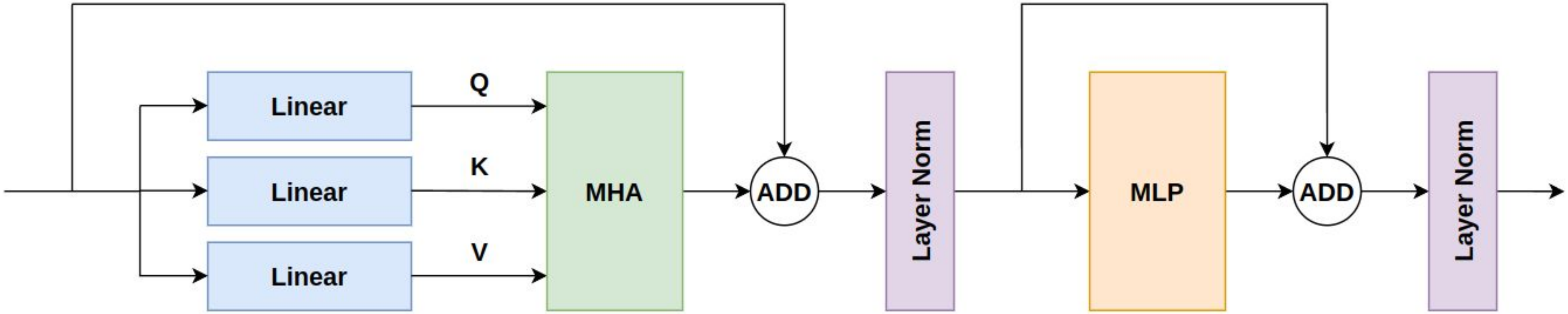


The decoder block is symmetrical to the encoder, it receives the encodings outputted by the encoder and after passing them through a list of attention blocks, they are fed to a linear layer to project them at their original dimension, then they are unflattened and outputted through a Tanh activation.

During inference the decoder part of the AE is not used and the encoder is fed with an unmasked patchified version of the image.



The attention block is an implementation of the standard multi head attention mechanism described in “Attention is all you need”.

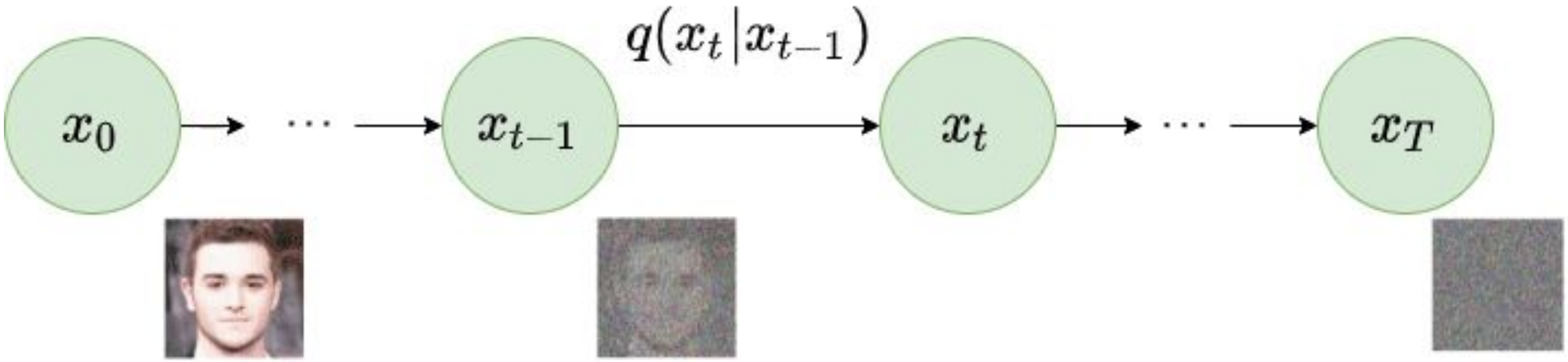


Latent Diffusion Model

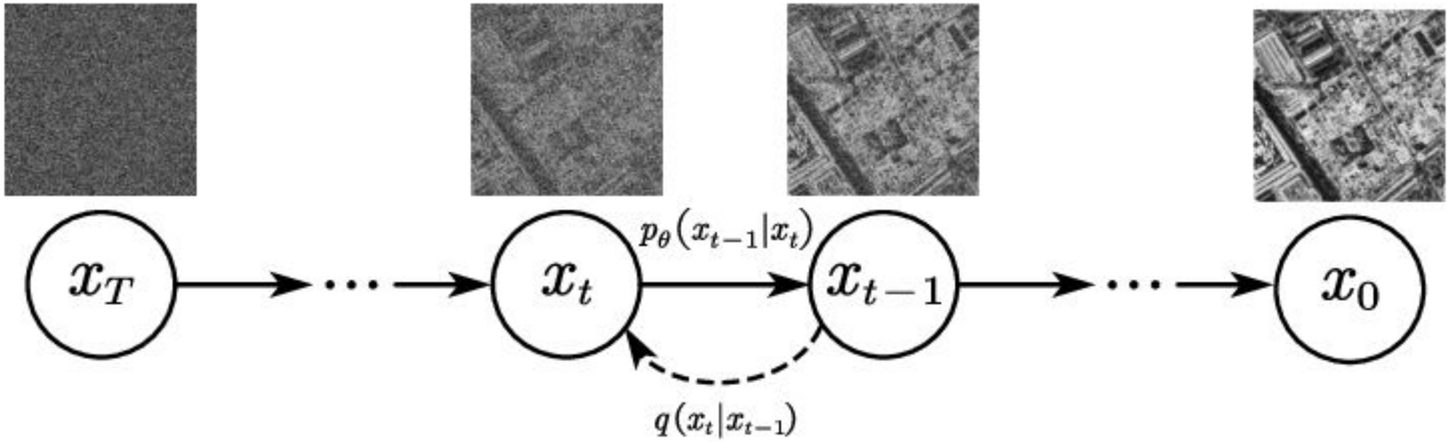
Latent diffusion models are a type of generative models which uses the so called diffusion process to generate new samples.

The diffusion process is divided into two parts:

- Forward diffusion: the latents are step by step added with a white gaussian noise using a noise scheduler, until they become indistinguishable from a gaussian distribution



- Backward diffusion: during backward diffusion we aim at learning the original noise added to the given input, and using the scheduler the model step by step denoise the signal.

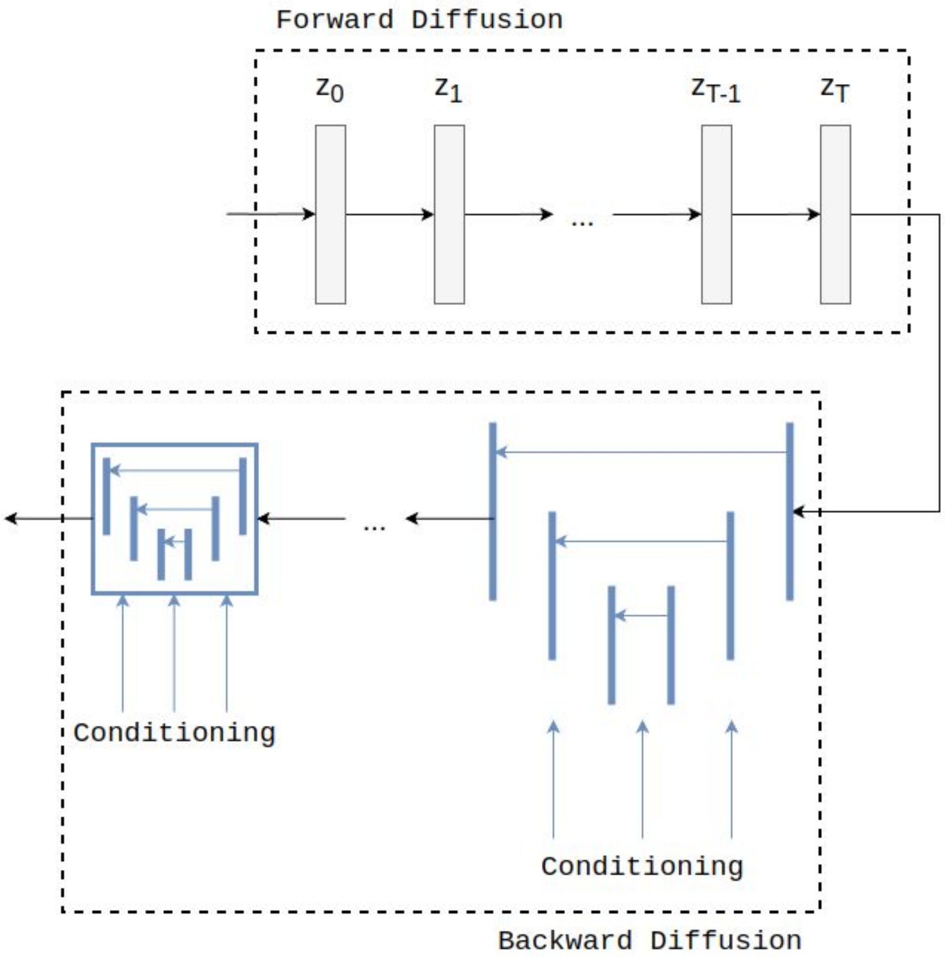


In the specific case of latent diffusion models, the inputs and outputs of the model are latents that are encoded and decoded by a Variational Autoencoder.

Proposed Architecture

The architecture for the LDM designed for this work is formed by PNDM Scheduler for the forward diffusion process, and a UNet denoising network for the backward process.

Training this network means making it learn the total noise that is added at a specific time-step to the input latents.



Problems when dealing with LDM

Some problems arise when training an LDM model due to the high number of computational and data resources required by the approach. In particular to obtain meaningful results it is necessary to use a very large dataset and train for a high number of epochs.

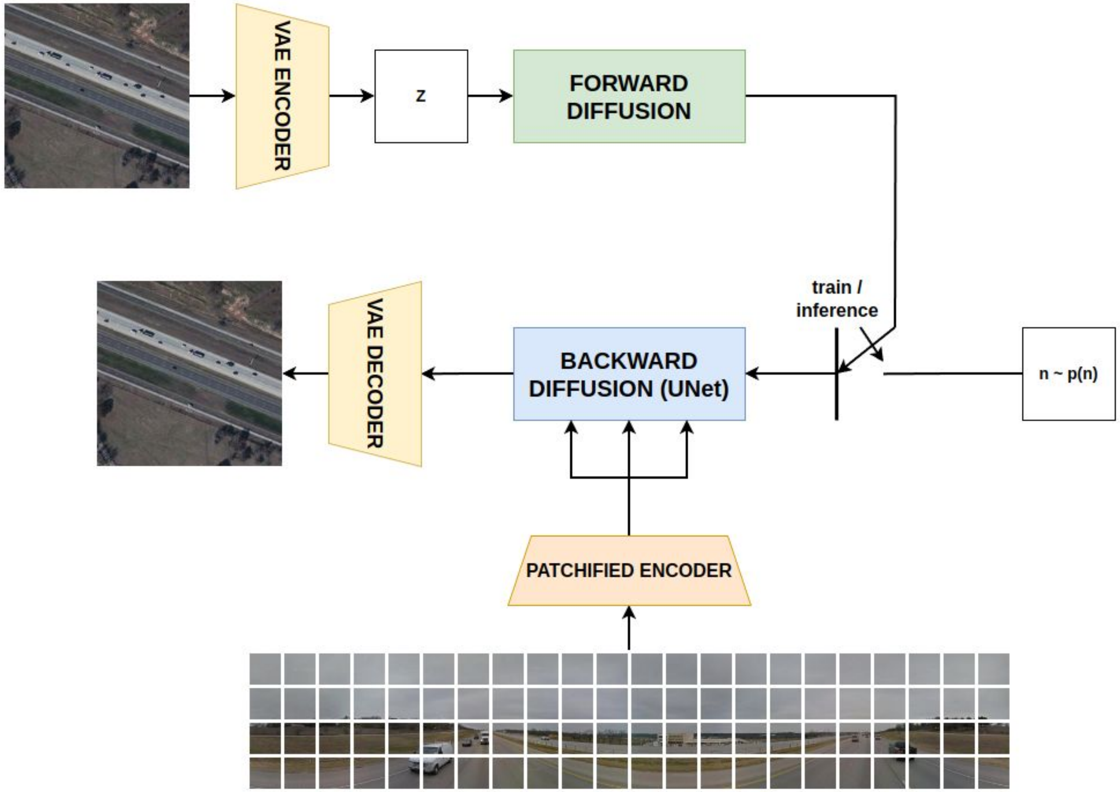
They also require a large capacity in terms of GPU RAM, for these reasons I was not able to perform a complete training on the LDM.

The final architecture

The final architecture has a different behavior during training or inference mode.

During training mode the satellite image is encoded and pass through forward diffusion process.

During inference the input to the model is a noise signal sampled from a Gaussian Distribution.


VAE ENCODER
z
FORWARD DIFFUSION
train / inference
VAE DECODER
BACKWARD DIFFUSION (UNet)
n ~ p(n)
PATCHIFIED ENCODER