Overview of UNet2D Condition Model

1 Introduction

This document outlines the UNet2D Condition Model used for processing four-dimensional tensors represented by `Sample[b, c, w, h]`. Each component of the model, from input embedding to post-processing, is detailed below, with schematics included to aid in understanding.

2 Model Architecture

2.1 Input Embedding

The input tensor undergoes an embedding process to generate time and possibly class embeddings, which are then augmented to enhance the representation:

- Time embedding generation and augmentation
- Class embedding (if applicable)
- Merging and enhancing embeddings

2.2 Pre-processing

The pre-processed tensor is shaped through an initial convolutional layer to prepare it for downstream processing:

$$\text{Output Shape: } [b, 320, w, h]$$

2.3 Down-sampling Path

The down-sampling path consists of several blocks, mainly focusing on reducing dimensions while increasing channels, ensuring feature extraction at various scales:

1. **Cross attention Down Block 2D** repeated three times, including operations of ResNet2D and Transformer2D.
2. **Down Block 2D** that includes traditional convolutional down-sampling.

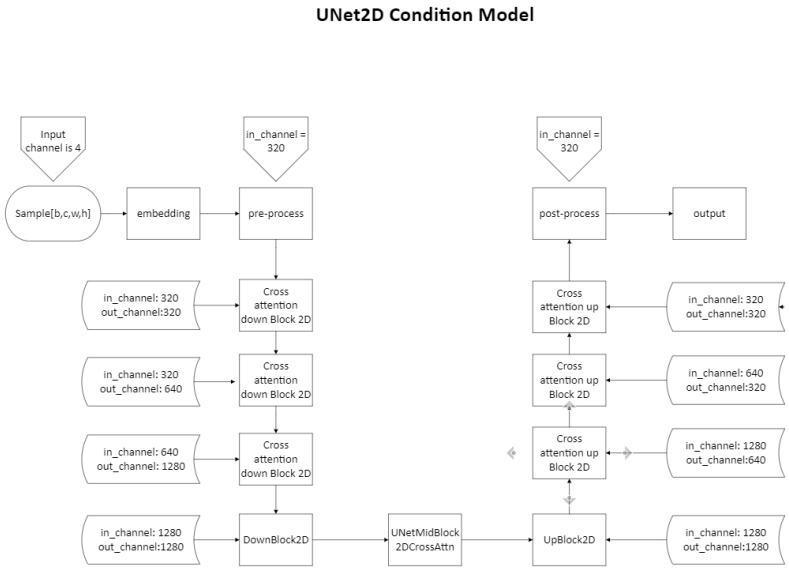

Figure 1: UNet2Dcondition Model

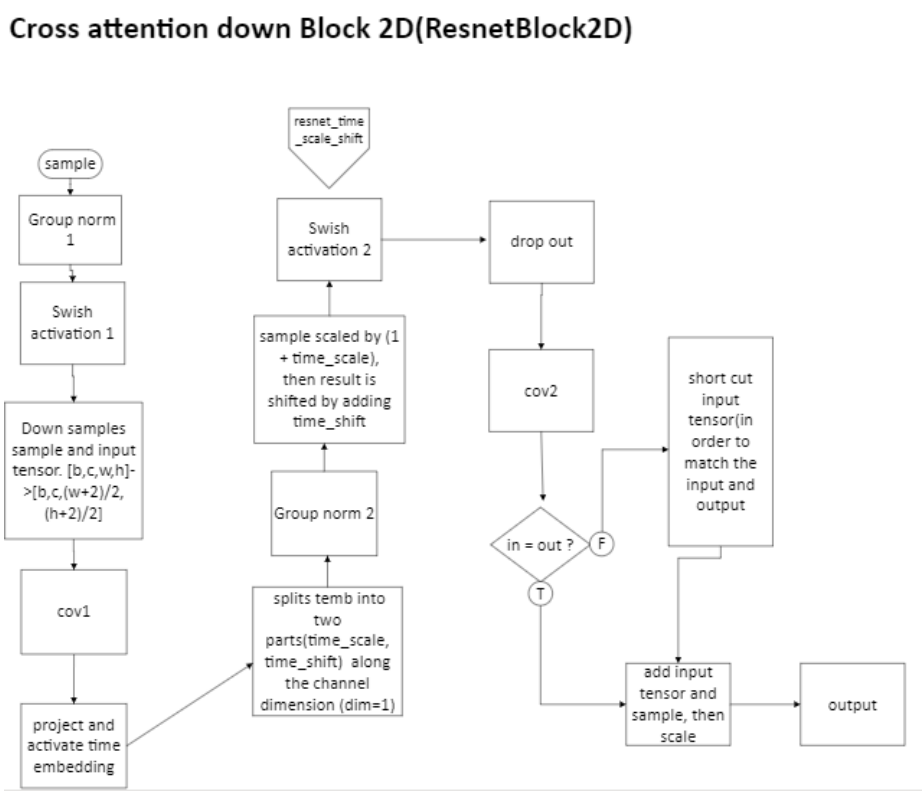

Figure 2: ResNetBlock2D

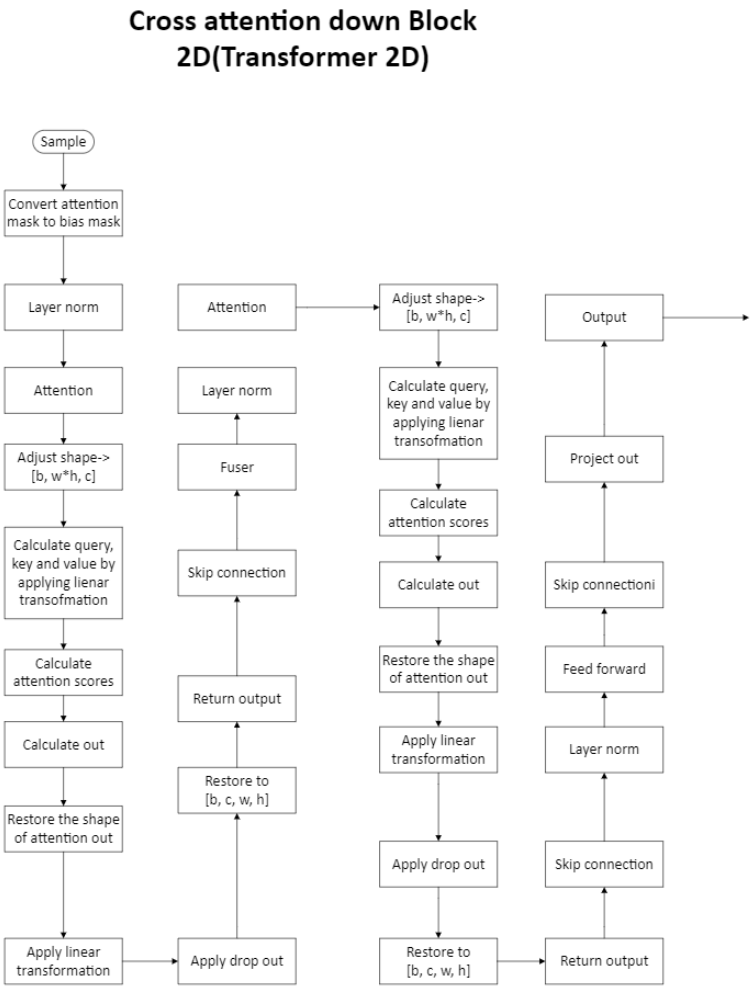

Figure 3: Transformer2D

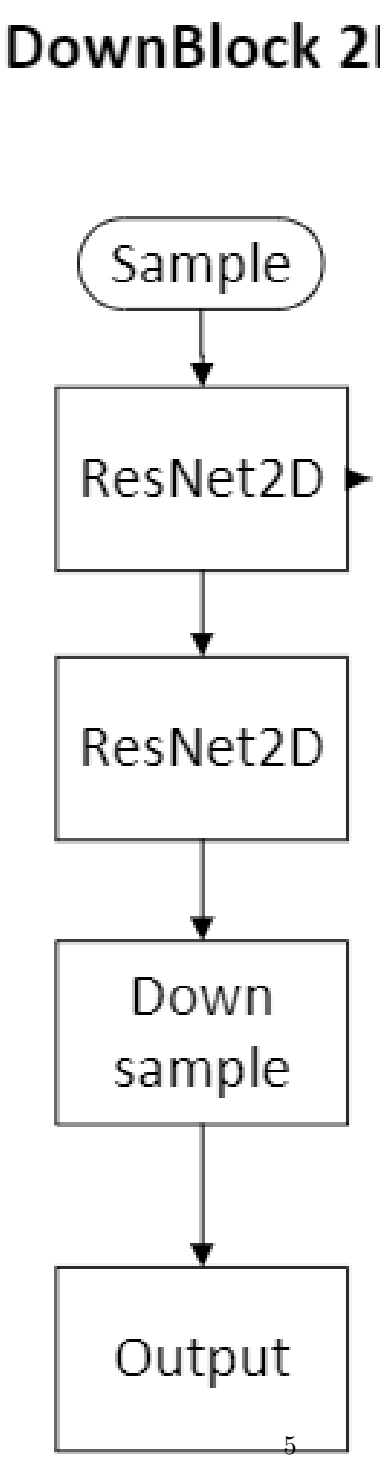

Figure 4: Enter Caption

2.4 Middle Block

The middle of the network, crucial for integrating features at the lowest spatial dimensions:

Composed of: ResNetBlock2D, Transformer2D, and another ResNetBlock2D

2.5 Up-sampling Path

The up-sampling path aims to progressively restore the spatial dimensions while refining the feature maps:

1. **UpBlock2D** includes operations to up-sample and refine features via multiple ResNet blocks.
2. **Cross attention Up Block 2D** repeated three times to enhance feature maps through attention mechanisms and ResNet blocks.

2.6 Post-processing

Finally, the output tensor is post-processed to match the desired output dimensions and channels:

$$\text{Output Shape: } [b, 4, w, h]$$