HW3 report

1. Comparing to the naïve GPU version,
    a. I store mask into constant memory instead of loading from the global memory.
    b. Create a shared memory used to store input from the global memory, so I can reduce times of accessing global memory
    c. I use 2d structure
2. I select 2d structure for my block
    a. From the slides, with the same kernel size and tile size, 2d convolution has larger bandwidth
    b. Because input and mask are matrix, 2d structure is much easier to implement
3. For boundaries for fetching pixels, It needs to shift from output coordinates to input coordinates by ROW_o minus mask offset (MASK WIDTH / 2). Then, I will check if input row and column are smaller than input width and height. Meanwhile, row and column must not smaller than 0. Otherwise, there would be a zero. (Meaning it is ghost)