Table 3: The detailed structure of the upsampling block and face decoder. In the descriptions, Conv $x_{/y,z}$ denotes 2D convolution with kernel size of x, stride length of y, and padding size of z.

Upsampling Block	
Component	**Dimension**
Input	$d_i \times w_i \times w_i$
Upsampling	$d_i \times 2w_i \times 2w_i$
Conv $3_{/2,1}$	$d \times 2w_i \times 2w_i$
ReLU	$d \times 2w_i \times 2w_i$
Batch Norm	$d \times 2w_i \times 2w_i$

Voice Encoder	
Layer	**Dimension**
Input	$512 \times 1 \times 1$
UpBlock 1	$1024 \times 2 \times 2$
UpBlock 2	$512 \times 4 \times 4$
UpBlock 3	$256 \times 8 \times 8$
UpBlock 4	$128 \times 16 \times 16$
UpBlock 5	$64 \times 32 \times 32$
UpBlock 6	$32 \times 64 \times 64$
Conv $1_{/1,0}$	$3 \times 64 \times 64$