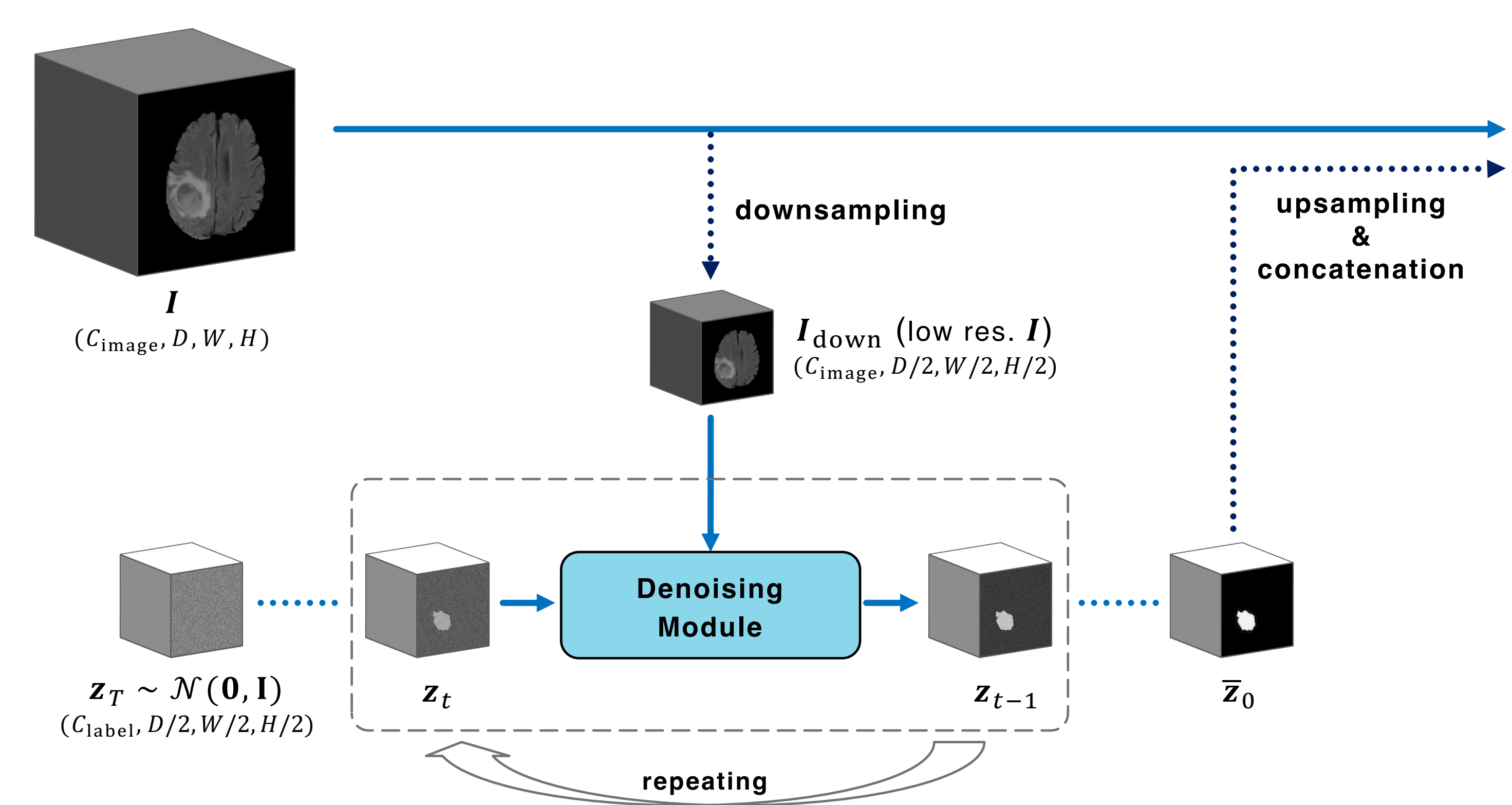**First stage**

Second stage

downsampling

upsampling & concatenation

dividing into patches

using sliding window

I
$(C_{\text{image}}, D, W, H)$

I_{down} (low res. I)
$(C_{\text{image}}, D/2, W/2, H/2)$

I, z_{up}
$(C_{\text{image}} + C_{\text{label}}, D, W, H)$

I', z'_{up}
$(C_{\text{image}} + C_{\text{label}}, 96, 96, 96)$

$\overline{x}_0$
$(C_{\text{label}}, D, W, H)$

Denoising Module

Denoising Module

$z_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
$(C_{\text{label}}, D/2, W/2, H/2)$

z_t

z_{t-1}

$\overline{z}_0$

$x'_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
$(C_{\text{label}}, 96, 96, 96)$

x'_t

x'_{t-1}

$\overline{x}'_0$

repeating

repeating