

Raw Image

$$\mathbf{x} \in \mathbb{R}^{256 \times 256}$$



$$t \sim \mathcal{T}$$

ResNet50  
encoder

MLP

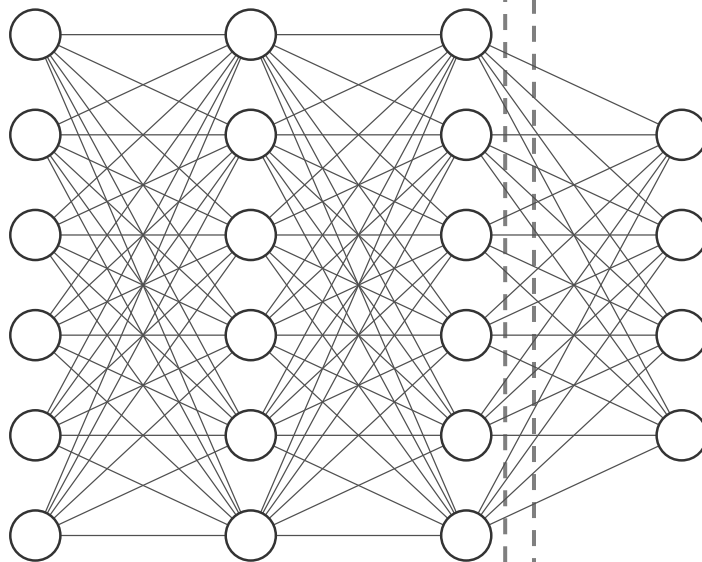


Image  
Embedding

$$\mathbf{z}^{im} \in \mathbb{R}^{128}$$

Pre-trained ResNet50: *frozen*

*trainable*