A Score-based representation diffusion models

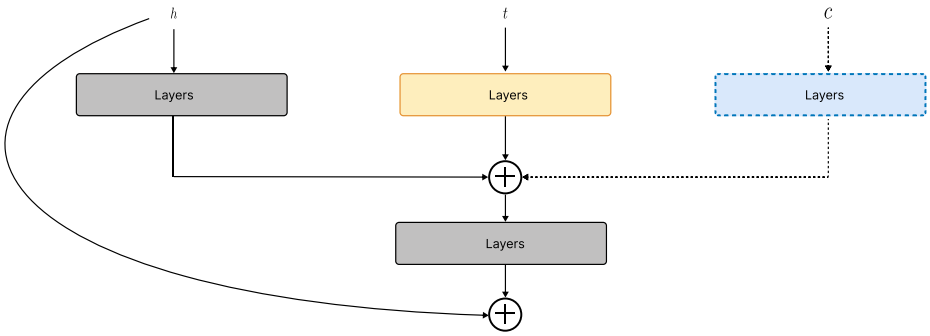

Fig. 1. Structure of each residual block. h, t, c are hidden feature, time and class condition, respectively.

The model used in our study is a score-based diffusion model, employed to estimate the density of the extracted representation $\mathbf{z}$. The forward-time diffusion process projects the representation distribution $p_0(z)$ to the noise distribution $p_1(z)$, which follows a stochastic differential equation (SDE),

$$d\mathbf{z} = \mathbf{f}(\mathbf{z}, t)dt + g(t)d\mathbf{w}, \tag{1}$$

where $\mathbf{w}$ is the Brownian motion, $\mathbf{f}(\mathbf{z}, t)$ is a vector-valued function, and $g(t)$ is a scalar function known as the diffusion coefficient. Sampling from the prior noise distribution p_t and reversing the diffusion process, we can obtain a sample from the data distribution p_0. This reverse diffusion process is given by the reverse-time SDE,

$$d\mathbf{z} = [\mathbf{f}(\mathbf{z}, t) - g^2(t)\nabla_{\mathbf{z}} \log p_t(\mathbf{z})]dt + g(t)d\bar{\mathbf{w}}. \tag{2}$$

A time-dependent score network $s_\theta(\mathbf{z}, t)$ can be trained to approximate the score $\nabla_{\mathbf{z}} \log p_t(\mathbf{z})$, using the weighted sum of denoising score matching objectives,

$$\min_\theta \mathbb{E}_{t\sim\mathcal{U}(0,1)}[\lambda(t)\mathbb{E}_{\mathbf{z}(0)}\mathbf{E}_{\mathbf{z}(t)}[\|s_\theta(\mathbf{z}(t), t) - \nabla_{\mathbf{z}(t)} \log p_{0t}(\mathbf{z}(t) \mid \mathbf{z}(0))\|_2^2]]. \tag{3}$$

Where $\mathbf{z}(0) \sim p_0(\mathbf{z})$, $\mathbf{z}(t) \sim p_{0t}(\mathbf{z}(t) \mid \mathbf{z}(0))$, $\mathcal{U}(0, 1)$ is a uniform distribution over $[0, 1]$, $p_{0t}(\mathbf{z}(t) \mid \mathbf{z}(0))$ denotes the transition probability from $\mathbf{z}(0)$ to $\mathbf{z}(t)$, and $\lambda(t)$ denotes a positive weighting function.

The $s_\theta(\mathbf{z}, t)$ is parameterized by a residual MLP network, each block is shown in Fig. 1. The class condition layer is only used for ConRDM and encoded by Fourier features up to 256 dimension. In practice, the training objective of our model is determined by selecting one of the Variance Exploding (VE), Variance

Preserving (VP), or subVP forms for the SDE, for $\sigma_{min}, \sigma_{max}$ and β_{min}, β_{max}, we use the default setting, as outlined in Table 1.

SDEs	**Formulation**	**Setting**
VE SDE	$d\mathbf{z} = \sqrt{\frac{[\sigma^2(t)]}{t}} d\mathbf{w}$	$\sigma_{min} = 0.01, \sigma_{max} = 50$
VP SDE	$d\mathbf{z} = -\frac{1}{2}\beta(t)\mathbf{z}\ dt + \sqrt{\beta(t)}\ d\mathbf{w}$	$\beta_{min} = 0.2, \beta_{max} = 20$
subVP SDE	$d\mathbf{z} = -\frac{1}{2}\beta(t)\mathbf{z}\ dt + \sqrt{\beta(t)(1 - e^{-2\int_0^t \beta(s)ds})} d\mathbf{w}$	$\beta_{min} = 0.2, \beta_{max} = 20$

Table 1: Formulation of 3 different SDEs, including hyperparameters that are used in our method.

B Residual setting on PCam benchmark

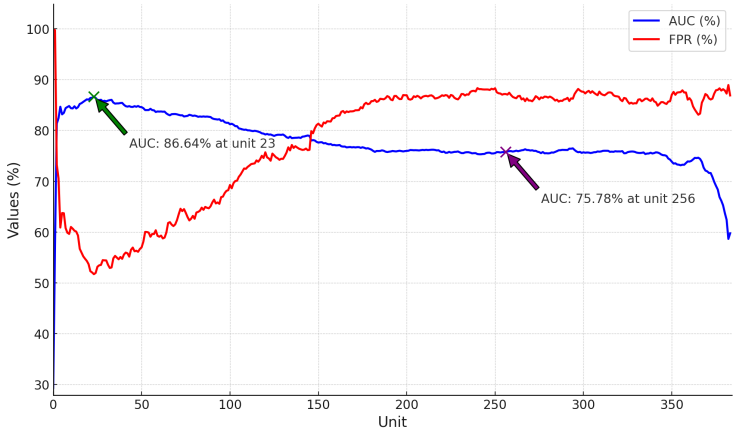

Fig. 2. For the PCam dataset, AUC and FPR95 are reported using different units in the residual score calculation.

AUC and FPR are calculated using varying numbers of residual units, ranging from 0 to 384, corresponding to the highest to lowest principal units, as illustrated in Fig. 2. The official implementation does not specify the number of residual units to use when the dimensionality of the representation is lower than 512. To maintain consistency, we follow the same ratio as used in ViT (starting from unit 512 to the last 768 units). Since the dimensionality in Pathology-SSL is 384, we keep the same 1/3 ratio and calculate Residual score starting from unit 256. However, the best OOD detection performance is observed from around unit 23, which use almost the whole principle space in OOD score calculation.