Supplementary Materials

$$\mathcal{D}_{KL}(u||\hat{y}) = \sum_{c=0}^{K-1} u^{(x,y,z)}(c) \log \frac{u^{(x,y,z)}(c)}{\hat{y}^{(x,y,z)}(c)} \tag{1}$$

$$\mathcal{L}_{mse}(\hat{y}, y) = \sum_{c=0}^{K-1} ||\hat{y}^{(x,y,z)}(c) - y^{(x,y,z)}(c)||_2^2 \tag{2}$$

Eq. (1) represents the KL divergence for each voxel, and Eq. (2) represents the mean square error for each voxel. In the code we use a variant of KL divergence, assuming that for a voxel, the u component is u_c, in addition $u_{\frac{1}{K}}, u_1$ stand for $u_c = \frac{1}{K}$ or 1, respectively. The following proof procedure exists:

$$H(\hat{p}) = -\sum_{c=0}^{K} \hat{p}_c log \hat{p}_c, \qquad \text{(where } c \text{ is } c^{th}\text{-component)} \tag{3}$$

$$H(u, \hat{p}) = -\sum_{c=0}^{K} u_c log \hat{p}_c, \tag{4}$$

$$\begin{aligned} & KL(\hat{p}||u_{\frac{1}{K}}) \\ =& \sum_{c=0}^{K} \hat{p}_c log(\frac{\hat{p}_c}{u_c}) \\ =& \sum_{c=0}^{K} \hat{p}_c log \hat{p}_c - \sum_{c=0}^{K} \hat{p}_c log u_c \quad \text{Have, } \sum_{c=0}^{K} \hat{p}_c = 1, log u_c = -log K \\ =& -H(\hat{p}) + log K \stackrel{K}{=} -H(\hat{p}) \end{aligned} \tag{5}$$

Symbol $\stackrel{K}{=}$ denotes equality up to an additive or multiplicative constant.

$$\begin{aligned} & KL(u_{\frac{1}{K}}||\hat{p}) \\ =& \sum_{c=0}^{K} u_c log(\frac{u_c}{\hat{p}_c}) \\ =& \sum_{c=0}^{K} u_c log u_c - \sum_{c=0}^{K} u_c log \hat{p}_c \quad \text{Have, } \sum_{c=0}^{K} u_c = 1, log u_c = -log K \\ =& -log K - \frac{1}{K} \sum_{c=0}^{K} log \hat{p}_c \stackrel{K}{=} H(u_{\frac{1}{K}}, \hat{p}). \end{aligned} \tag{6}$$

When $u_c = 1$, exist:

$$
\begin{aligned}
&KL(u_1||\hat{p})\\
=&-\sum_{c=0}^{K} u_c log\hat{p}_c = -\sum_{c=0}^{K} log\hat{p}_c\\
=&H(u_1,\hat{p}) = K\cdot H(u_{\frac{1}{K}},\hat{p}) \overset{K}{=} K\cdot KL(u_{\frac{1}{K}}||\hat{p}).
\end{aligned}
\tag{7}
$$

In Torch, if have a tensor $T \in \mathcal{R}^{B\times K\times W\times H}$, use $torch.kl(u_1, T)$ get Output $O \in \mathcal{R}^{B\times K\times W\times H}$, then conduct $\mathbf{M} = torch.mean(O, dim = 1)$, $\mathbf{M} = -\frac{1}{K}\sum_{c=0}^{K} log\hat{p}_c$ for each voxel.

So $torch.mean(torch.kl(u_1, T), dim = 1) = H(u_{\frac{1}{K}},\hat{p}) \overset{K}{=} KL(u_{\frac{1}{K}}||\hat{p})$.

Then, $KL(u_{\frac{1}{K}}||\hat{p}) = -logK - \frac{1}{K}\sum_{c=0}^{K} log\hat{p}_c = -logK + H(u_{\frac{1}{K}},\hat{p})$. Taking the derivative of it with respect to the parameters θ being optimized:

$$
\begin{aligned}
&\frac{\partial KL(u_{\frac{1}{K}}||\hat{p})}{\partial\theta}\\
=&\frac{\partial(-logK + H(u_{\frac{1}{K}},\hat{p}))}{\partial\theta}\\
=&\frac{\partial H(u_{\frac{1}{K}},\hat{p})}{\partial\theta}.
\end{aligned}
\tag{8}
$$

Eventually we use $H(u_{\frac{1}{K}},\hat{p})$ as a substitute for $KL(u_{\frac{1}{K}}||\hat{p})$.