Appendix

Assume a setting with two classes C_1 and C_2, with a probability density function $p_1(x)$ and $p_2(x)$, respectively. x_a is sampled from either of the distributions, resulting in $p_a(x_a) = \pi p_1(x_a) + (1-\pi)p_2(x_a)$, with $0 \leq \pi \leq 1$. Set $x = [x_a^{'}\ x_p^{'}\ x_n^{'}]^{'}$ with a probability density function $p(x) = p(x_a, x_p, x_n)$ assumed to be equal to $p(x) = p(x_a, x_p, x_n) = p_a(x_a)p_p(x_p|x_a)p_n(x_n|x_a) = \pi p_1(x_a)p_1(x_p)p_2(x_n) + (1-\pi)p_2(x_a)p_2(x_p)p_1(x_n)$.

The triplet loss between x_a, x_p, x_n can be described using a Euclidean distance function as

$\mathbb{E}_{x\sim p}\mathcal{L}(x_a, x_p, x_n)$ with

$$\mathcal{L}(x_a, x_p, x_n) = \max(0, \alpha + ||f(x_a) - f(x_p)||^2 - ||f(x_a) - f(x_n)||^2),$$

where $f(x)$ is the learnt representation of x.

In the presence of the noise in the sampling process of the triplets (x_a, x_b, x_c) as modelled in the main text, the probability density function becomes

$$\begin{aligned} g(x) = {} & \beta\gamma p_a(x_a)p_p(x_p)p_n(x_n) \\ & + (1-\beta)\gamma p_a(x_a)p_n(x_p)p_n(x_n) \\ & + \beta(1-\gamma)p_a(x_a)p_p(x_p)p_p(x_n) \\ & + (1-\beta)(1-\gamma)p_a(x_a)p_n(x_p)p_p(x_n) \end{aligned} \tag{1}$$

Now the triplet loss becomes

$$\begin{aligned} \mathbb{E}_{x\sim g}\mathcal{L}(x_a, x_p, x_n) = {} & \beta\gamma\mathbb{E}_{x\sim p}\mathcal{L}(x_a, x_p, x_n) \\ & + (1-\beta)\gamma\mathbb{E}_{x\sim h_1}\mathcal{L}(x_a, x_p, x_n) \\ & + \beta(1-\gamma)\mathbb{E}_{x\sim h_2}\mathcal{L}(x_a, x_p, x_n) \\ & + (1-\beta)(1-\gamma)\mathbb{E}_{x\sim p}\mathcal{L}(x_a, x_n, x_p) \end{aligned} \tag{2}$$

with

$$h_1(x_a, x_p, x_n) = \pi p_1(x_a)p_2(x_p)p_2(x_n) + (1-\pi)p_2(x_a)p_1(x_p)p_1(x_n)$$

and

$$h_2(x_a, x_p, x_n) = \pi p_1(x_a)p_1(x_p)p_1(x_n) + (1-\pi)p_2(x_a)p_2(x_p)p_2(x_n)$$

Noting that the expected value of the *gradient* of the loss $\mathcal{L}(x_a, x_p, x_n)$ is zero under h_1 and h_2 (since x_p and x_n are sampled from the same distribution and the loss is symmetric), the impact of those two cases in a gradient based learning scheme is small.

We are then left with a total loss of

$$\begin{aligned} & \beta\gamma\max(0, \alpha + ||f(x_a) - f(x_p)||^2 - ||f(x_a) - f(x_n)||^2) + \\ & \qquad (1-\beta)(1-\gamma)\max(0, \alpha + ||f(x_a) - f(x_n)||^2 - ||f(x_a) - f(x_p)||^2) \end{aligned} \tag{3}$$

under the $p(x)$ probability density function.

Finally, this loss can be compacted to

$$\max\big(0, \beta\gamma(\alpha + ||f(x_a) - f(x_p)||^2 - ||f(x_a) - f(x_n)||^2)\big) + \max\big(0, (1-\beta)(1-\gamma)(\alpha + ||f(x_a) - f(x_n)||^2 - ||f(x_a) - f(x_p)||^2)\big) \quad (4)$$