A. Appendix

This section contains explanatory comments, clarifications and mathematical derivations to support the main text.

A.1. Properties of the softmax function

Here, we discuss properties of the softmax function that are used below. We start with its stationarity:

$$\begin{aligned}\mathrm{smax}(x+d,y+d) &= \frac{\exp(x+d)}{\exp(x+d)+\exp(y+d)} && (26)\\ &= \frac{\exp(d)\exp(x)}{\exp(d)(\exp(x)+\exp(y))} && (27)\\ &= \frac{\exp(x)}{\exp(x)+\exp(y)} && (28)\\ &= \mathrm{smax}(x,y). && (29)\end{aligned}$$

Setting $d=-x-y$, we get the following property:

$$\begin{aligned}\mathrm{smax}(x,y) &= \mathrm{smax}(x-x-y,y-x-y) && (30)\\ &= \mathrm{smax}(-y,-x). && (31)\end{aligned}$$

A.2. Probabilistic derivation of $\mathcal{L}_{\text{log}}$ and $\mathcal{L}_{\text{sub}}$

The log loss has been previously discussed in the literature (e.g. Hoffer and Ailon [9]), but it was reported to perform quite badly. We believe that this is primarily because it was used at the wrong scale, not because it is unsuitable per se (quite the opposite, as we have empirically demonstrated in our experiments). In the following, we give a derivation of both the SSE loss (3) and the log loss (2) where we start from a probabilistic point of view, showing that the log loss in particular has sound theoretical foundations.

Our objective is to encourage that d_p becomes smaller than d_n, where a larger gap is better. This can be expressed with the softmax function:

$$\mathrm{smax}(d_p,d_n) = \frac{\exp(d_p)}{\exp(d_p)+\exp(d_n)} \qquad (32)$$

The softmax function is in $[0,1]$, that is, we can interpret it probabilistically as the uncertainty whether the two patches in the positive pair are indeed more similar than the ones in the negative pair. Our objective is to minimize this uncertainty:

$$\begin{aligned}\mathrm{smax}(d_p,d_n) &\longrightarrow 0, \text{ or, equivalently,} && (33)\\ 1-\mathrm{smax}(d_p,d_n) &\longrightarrow 1. && (34)\end{aligned}$$

To measure the deviation from this objective, we can use the cross-entropy between actual and desired values:

$$\begin{aligned}\mathcal{L}_{\text{log}}(d_p,d_n) =& -0\cdot\log(\mathrm{smax}(d_p,d_n)) && (35)\\ & -1\cdot\log(1-\mathrm{smax}(d_p,d_n)) && (36)\\ =& -\log(1-\mathrm{smax}(d_p,d_n)). && (37)\end{aligned}$$

For the sake of a shorter notation, we rewrite the term under the logarithm:

$$\begin{aligned}1-\mathrm{smax}(d_p,d_n) &= 1-\frac{\exp(d_p)}{\exp(d_p)+\exp(d_n)} && (38)\\ &= \frac{\exp(d_p)+\exp(d_n)-\exp(d_p)}{\exp(d_n)+\exp(d_p)} && (39)\\ &= \frac{\exp(d_n)}{\exp(d_n)+\exp(d_p)} = \mathrm{smax}(d_n,d_p). && (40)\end{aligned}$$

Using (31), this can be further rewritten:

$$\mathrm{smax}(d_n,d_p) = \mathrm{smax}(-d_p,-d_n). \qquad (41)$$

Hence, we get the final formulation of the log loss:

$$\mathcal{L}_{\text{log}}(d_p,d_n) = -\log\mathrm{smax}(-d_p,-d_n). \qquad (42)$$

As an alternative to the cross-entropy, one can simply calculate the squared sum of errors of the two cases (33) and (34):

$$\begin{aligned}\mathcal{L}^*_{\text{sse}}(d_p,d_n) =& (0-\mathrm{smax}(d_p,d_n))^2 && (43)\\ & +(1-(1-\mathrm{smax}(d_p,d_n))^2 && (44)\\ =& \, 2\,\mathrm{smax}(d_p,d_n)^2. && (45)\end{aligned}$$

For simplicity, we leave away the constant factor 2 (this can be compensated by increasing the learning rate), yielding the SSE loss:

$$\mathcal{L}_{\text{sse}}(d_p,d_n) = \mathrm{smax}(d_p,d_n)^2. \qquad (46)$$

A.3. Losses and performance functions

In this section, we show how the five losses discussed in Section 2 can be rewritten in terms of the performance functions given in Equations (7) to (9). The following three cases are straightforward:

$$\begin{aligned}\mathcal{L}_{\text{sub}} &= [d_p-d_n+\alpha]_+ = [\alpha-\rho_{\text{sub}}]_+\,, && (47)\\ \mathcal{L}_{\text{sub2}} &= \left[d_p^2-d_n^2+\alpha\right]_+ = [\alpha-\rho_{\text{sub2}}]_+\,, && (48)\\ \mathcal{L}_{\text{div}} &= \left[1-\frac{d_n}{d_p+\epsilon}\right]_+ = [1-\rho_{\text{div}}]_+\,. && (49)\end{aligned}$$

The two remaining cases, $\mathcal{L}_{\text{log}}$ and $\mathcal{L}_{\text{sse}}$ are less obvious, but using (29) we can write:

$$\begin{aligned}\mathcal{L}_{\text{log}} &= -\log\mathrm{smax}(-d_p,-d_n) && (50)\\ &= -\log\mathrm{smax}(-d_p+d_n,-d_n+d_n) && (51)\\ &= -\log\mathrm{smax}(d_n-d_p,0) && (52)\\ &= -\log\mathrm{smax}(\rho_{\text{sub}},0), && (53)\\ \mathcal{L}_{\text{sse}} &= \mathrm{smax}(d_p,d_n)^2 && (54)\\ &= \mathrm{smax}(d_p-d_n,d_n-d_n)^2 && (55)\\ &= \mathrm{smax}(-(d_n-d_p),0)^2 && (56)\\ &= \mathrm{smax}(-\rho_{\text{sub}},0)^2. && (57)\end{aligned}$$

A.4. Corner Cases of Generalized Scale Log Loss

In the following, we derive the two corner cases of $\tilde{\mathcal{L}}_{\text{log}}$ from Section 3.3.

Case $\delta \to \infty$: Here, we show Equation (18). First, we rewrite the loss (setting $\rho := \rho_{\text{sub}}$):

$$\begin{aligned}
\tilde{\mathcal{L}}_{\text{log}}(\rho) &= \frac{1}{\delta}\mathcal{L}_{\text{log}}(\delta(\rho-\alpha)) && (58)\\
&= -\frac{1}{\delta}\log \operatorname{smax}(\delta(\rho-\alpha),0) && (59)\\
&= -\log \operatorname{smax}(\delta(\rho-\alpha),0)^{1/\delta} && (60)\\
&= -\log\left[\frac{\exp(\delta(\rho-\alpha))}{\exp(\delta(\rho-\alpha))+\exp(0)}\right]^{1/\delta} && (61)\\
&= -\log\left[\frac{\exp(\rho-\alpha)}{\sqrt[\delta]{\exp(\rho-\alpha)^{\delta}+\exp(0)}}\right]. && (62)
\end{aligned}$$

To see what happens as $\delta \to \infty$, we make a case distinction:

- $\rho - \alpha < 0$: in this case, the root in the denominator is dominated by $\exp(0) = 1$ because $\exp(\rho - \alpha)^{\delta}$ disappears. Hence, we have $\lim_{\delta\to\infty} \tilde{\mathcal{L}}_{\text{log}}(\rho) = -\log[\exp(\rho-\alpha)] = \alpha - \rho$.
- $\rho - \alpha \geq 0$: in this case, $\exp(\rho - \alpha)$ dominates the root in the denominator. We have $\lim_{\delta\to\infty} \tilde{\mathcal{L}}_{\text{log}}(\rho) = -\log[\exp(\rho-\alpha)/\exp(\rho-\alpha)] = 0$.

Concluding:

$$\begin{aligned}
\lim_{\delta\to\infty} \tilde{\mathcal{L}}_{\text{log}}(\rho) &= \begin{cases} \alpha-\rho & \text{if } \alpha-\rho > 0 \\ 0 & \text{if } \alpha-\rho \leq 0 \end{cases} && (63)\\
&= \max\{\alpha-\rho, 0\} && (64)\\
&= [\alpha-\rho]_{+} = \tilde{\mathcal{L}}_{\text{sub}}(\rho;\alpha). && (65)
\end{aligned}$$

Case $\delta \to 0$: Here, we show Equation (17). Actually, the loss $\mathcal{L}_{\text{log}}$ diverges for $\delta \to 0$, but as we discussed in the text, we are not primarily interested in the actual value of a loss but in its derivative which tells us how strongly the loss acts. Therefore, we consider what happens for very small values of δ by investigating the derivative. It is given below in Equation (76). It is easy to see that $\lim_{\delta\to 0}\exp(\delta(\rho_{\text{sub}} - \alpha)) = 1$, and consequently:

$$\lim_{\delta\to 0}\frac{\partial\mathcal{L}_{\text{log}}(\rho_{\text{sub}})}{\partial\rho_{\text{sub}}} = -\frac{1}{1+1} - \frac{1}{2}. \tag{66}$$

By integrating, we get that the loss behaves asymptotically as the following loss:

$$\lim_{\delta\to 0}\tilde{\mathcal{L}}_{\text{log}}(\rho_{\text{sub}};\alpha,\delta) \propto -0.5\rho_{\text{sub}}. \tag{67}$$

A.5. Antisymmetry

In Section 4.1, we mention that losses in ρ_{sub} are antisymmetric, i.e.,

$$\frac{\partial\mathcal{L}(\rho_{\text{sub}}(d_p,d_n))}{\partial d_p} = -\frac{\partial\mathcal{L}(\rho_{\text{sub}}(d_p,d_n))}{\partial d_n}. \tag{68}$$

Using the definition of the difference performance function in Equation (7) and the chain rule, we can derive:

$$\begin{aligned}
\frac{\partial\mathcal{L}(\rho_{\text{sub}})}{\partial d_p} &= \frac{\partial\mathcal{L}(d_n-d_p)}{\partial d_p} && (69)\\
&= \frac{\partial(d_n-d_p)}{\partial d_p}\frac{\partial\mathcal{L}(d_n-d_p)}{\partial(d_n-d_p)} && (70)\\
&= -\frac{\partial\mathcal{L}(d_n-d_p)}{\partial(d_n-d_p)} = -\frac{\partial\mathcal{L}(\rho_{\text{sub}})}{\partial\rho_{\text{sub}}}, && (71)\\
\frac{\partial\mathcal{L}(\rho_{\text{sub}})}{\partial d_n} &= \frac{\partial\mathcal{L}(d_n-d_p)}{\partial d_n} && (72)\\
&= \frac{\partial(d_n-d_p)}{\partial d_n}\frac{\partial\mathcal{L}(d_n-d_p)}{\partial(d_n-d_p)} && (73)\\
&= +\frac{\partial\mathcal{L}(d_n-d_p)}{\partial(d_n-d_p)} = +\frac{\partial\mathcal{L}(\rho_{\text{sub}})}{\partial\rho_{\text{sub}}}. && (74)
\end{aligned}$$

It is immediately visible that

$$\frac{\partial\mathcal{L}(\rho_{\text{sub}})}{\partial d_p} = -\frac{\partial\mathcal{L}(\rho_{\text{sub}})}{\partial d_n}, \tag{75}$$

i.e., any loss in ρ_{sub} is antisymmetric.

A.6. Derivatives of the losses

For visualizing the losses, we need the derivatives of the losses w.r.t. their performance functions. For the three cases in Figure 4:

$$\begin{aligned}
\frac{\partial\tilde{\mathcal{L}}_{\text{log}}(\rho_{\text{sub}})}{\partial\rho_{\text{sub}}} &= -\frac{1}{\exp(\delta(\rho_{\text{sub}}-\alpha))+1} && (76)\\
\frac{\partial\tilde{\mathcal{L}}_{\text{sse}}(\rho_{\text{sub}})}{\partial\rho_{\text{sub}}} &= -2\frac{\operatorname{smax}(-\delta(\rho_{\text{sub}}-\alpha),0)^2}{\exp(\delta(\rho_{\text{sub}}-\alpha))^2} && (77)\\
\frac{\partial\mathcal{L}_{\text{sub}}(\rho_{\text{sub}})}{\partial\rho_{\text{sub}}} &= \begin{cases} 0 & \text{if } \rho_{\text{sub}} > \alpha \\ -1 & \text{otherwise} \end{cases} && (78)
\end{aligned}$$

For completeness, we also give the derivatives of the other two losses:

$$\begin{aligned}
\frac{\partial\mathcal{L}_{\text{sub2}}(\rho_{\text{sub2}})}{\partial\rho_{\text{sub2}}} &= \begin{cases} 0 & \text{if } \rho_{\text{sub2}} > \alpha \\ -1 & \text{otherwise} \end{cases} && (79)\\
\frac{\partial\mathcal{L}_{\text{div}}(\rho_{\text{div}})}{\partial\rho_{\text{div}}} &= \begin{cases} 0 & \text{if } \rho_{\text{div}} > 1 \\ -1 & \text{otherwise} \end{cases} && (80)
\end{aligned}$$

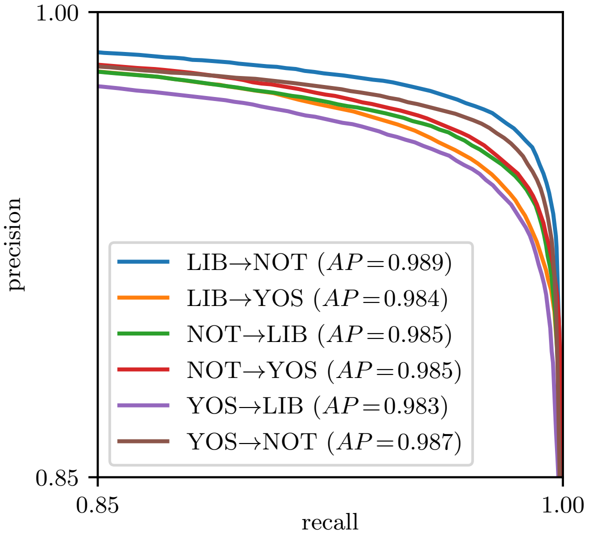

Figure 8: PR curves and their APs on the mixed-context loss. Model is trained on each of the three scenes: `liberty`, `notredame`, and `yosemite` and then tested on the other datasets.

A.7. Full PR curves and their AP

In this section, we report the full PR curves and their AP in Figure 8 on the UBC benchmark [7] by following the procedures described in [22]. The PR curve is computed from the default 100'000 evaluation patch pairs of each dataset. The values are analogous to the last row of Table 1

A.8. Scale-aware sampling and L2-Net

Based on our discussion of the importance of scale in descriptor learning, we introduced scale-aware sampling in Section 4.3 and empirically demonstrated in Section 5 that it indeed brings a considerable boost in performance. In fact, if we look at the losses that were presented in key papers in the last two years, we can see that they followed a similar idea, only less consequently. One example is L2-Net [23] whose loss can be written as follows (it differs in details from the formulation in the original paper, but is true to the main idea):

$$\mathcal{L}_{\text{L2}} = -\sum_{i=1}^{N} \log\left[\frac{\exp(-d_p^i)}{\exp(-d_p^i)+\sum_j \exp(-d_n^{i,j})}\right]. \tag{81}$$

At first, glance, it is difficult to see the relationship with scale-aware sampling. However, if we use the fact that $\sum_j \exp(-d_n^{i,j}) = \exp\log\sum_j \exp(-d_n^{i,j})$, we can understand the fraction as a softmax function and hence rewrite the loss in terms of the log loss $\mathcal{L}_{\text{log}}$ (Equation (2)):

$$\mathcal{L}_{\text{L2}} = \sum_{i=1}^{N} -\log \operatorname{smax}\left(-d_p^i, \log\sum\nolimits_j \exp\left(-d_n^{i,j}\right)\right) \tag{82}$$

$$= \sum_{i=1}^{N} \mathcal{L}_{\text{log}}\left(d_p^i, -\log\sum\nolimits_j \exp\left(-d_n^{i,j}\right)\right). \tag{83}$$

In other words, we can interpret the loss of L2-Net as a kind triplet loss where many negative distances are amalgamated into one single value by means of a log-sum-exp expression (note that it therefore still suffers from the localized-context problem). This gives more weight to the smaller negative distances than to larger ones, that is, we can understand L2-Net as an approximation to our scale-aware sampling scheme where we simply pick the smallest negative distance $\min_j\{d_n^{i,j}\}$. The corresponding loss then is the one we presented in Equation (23), with $\mathcal{L}_{\text{triplet}} := \mathcal{L}_{\text{log}}$:

$$\sum_{i=1}^{N} \mathcal{L}_{\text{log}}\left(d_p^i, \min_j\{d_n^{i,j}\}\right). \tag{84}$$