A "cell" of a SIFT descriptor, computed on an image \( I \) in a region of size \( \hat{\sigma} \) around a location \( x \), can be written as

\[
h_{\text{SIFT}}(\theta|I, \hat{\sigma})[x] = \int N_\xi(\theta - \xi \nabla I(y)) N_{\hat{\sigma}}(y - x) d\mu(y) \quad (1)
\]

where \( d\mu(y) = \|\nabla I(y)\|dy \), \( \theta \) is the independent variable, ranging from 0 to \( 2\pi \), corresponding to an orientation histogram bin of size \( \varepsilon \), and \( \hat{\sigma} \) is the spatial pooling scale. The kernel \( N_\xi \) is bilinear of size \( \varepsilon \) and \( N_{\hat{\sigma}} \) separable-bilinear of size \( \hat{\sigma} \). Both the location \( x \) and the scale \( \hat{\sigma} \) are typically sampled by a co-variant detector, or regularly as in "dense SIFT." Spatial pooling, interpreted as local marginalization against the kernel \( N_{\hat{\sigma}}(y - x) \), affords insensitivity to small translations around the sampled location \( x \).

But while translations are locally marginalized around the sample \( x \), changes of scale around the sampled \( \hat{\sigma} \) are not.

DSP-SIFT is designed to obviate this asymmetry of treatment, by locally marginalizing scale, in addition to translation. If \( s > 0 \) and \( \varepsilon \) is an exponential or other unilateral density function, the process can be written as

\[
h_{\text{DSP}}(\theta|I)[x] = \int h_{\text{SIFT}}(\theta|I, \sigma)[x] \varepsilon_s(\sigma) d\sigma \quad x \in \Lambda \quad (2)
\]

as illustrated in Fig. 1 and implemented in few lines of code. DSP-SIFT has the same dimension of SIFT and improves its performance by 10% to 40% mean-average precision (mAP) on the datasets we tested. It also outperforms a deep convolutional architecture (CNN) in wide-baseline matching tasks, despite having a considerably smaller size and requiring no training (Fig. 2).

DSP-SIFT pools gradient orientations in regions of different size, hence the name domain-size pooling, in apparent violation of the principles of scale-space theory and scale selection. There, the size of the region where statistics are computed is tied to the spatial frequencies of the image, which facilitates correspondence under changes of scale or distance. However, this does not take into account occlusions: How large a portion of a scene is visible in each corresponding image(s) depends on the shape of the scene, the pose of the two cameras, and the resulting visibility (occlusion) relations, not on the spatial frequencies of the image. Therefore, we unite the size of the domain where the descriptor is computed ("scale") from photometric characteristics of the image (Fig. 3). While somewhat unuitive, as histogram bins mix different regions of the same image, this procedure is rooted in classical sampling theory and the practice of anti-aliasing.

Domain-size pooling can be applied to a range of other low-level vision operations, such as in other histogram-based representations, including the lower layers of convolutional neural networks [2]. A more detailed derivation of DSP-SIFT and its relation to sampling theory is described in [1], and the implementation is available at vision.ucla.edu/dsp-sift.