Figure 1: Implementation of convolution in Caffe, compared with the implementation of segmentation-aware convolution. Convolution involves re-organizing the elements of each (potentially overlapping) patch into a column (i.e., \textit{im2col}), followed by a matrix multiplication with weights. Segmentation-aware convolution works similarly, with an image-to-column transformation on the input, an image-to-distance transformation on the embeddings (i.e., \textit{im2dist}), a pointwise multiplication of those two matrices, and then a matrix multiplication with weights. The variables $H, W$ denote the height and width of the input, respectively; $E$ denotes the number of channels in the input; $K$ denotes the dimensionality of a patch (e.g., $K = 9$ in convolution with a $3 \times 3$ filter); $F$ denotes the number of filters (and the dimensionality of the output). In both cases, an $H \times W \times E$ input is transformed into an $H \times W \times F$ output.