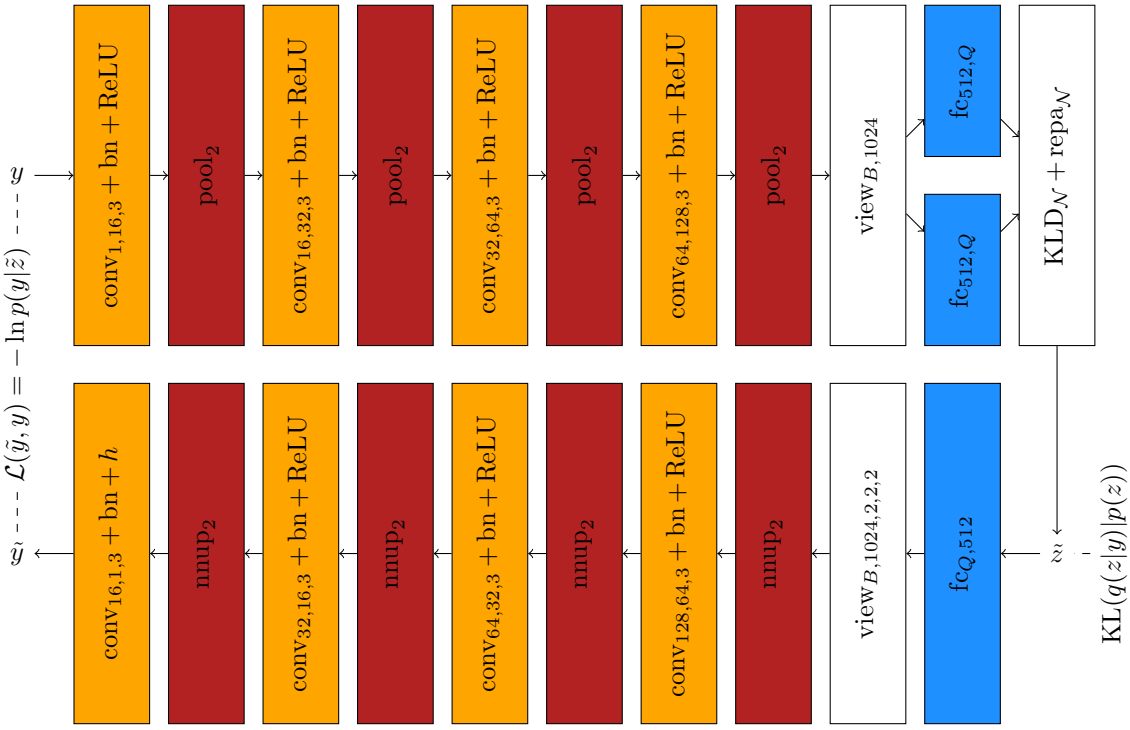

Figure 1: An illustration of a Gaussian variational auto-encoder with four convolutional stages in both encoder and decoder. This is also the architecture used in experiments in Chapter **??** where we assume volumes of size $32 \times 32 \times 32$ such that the spatial size just before the fully connected layers of the encoder is $2 \times 2 \times 2$ resulting in a 1024-dimensional representation when considering 128 channels.

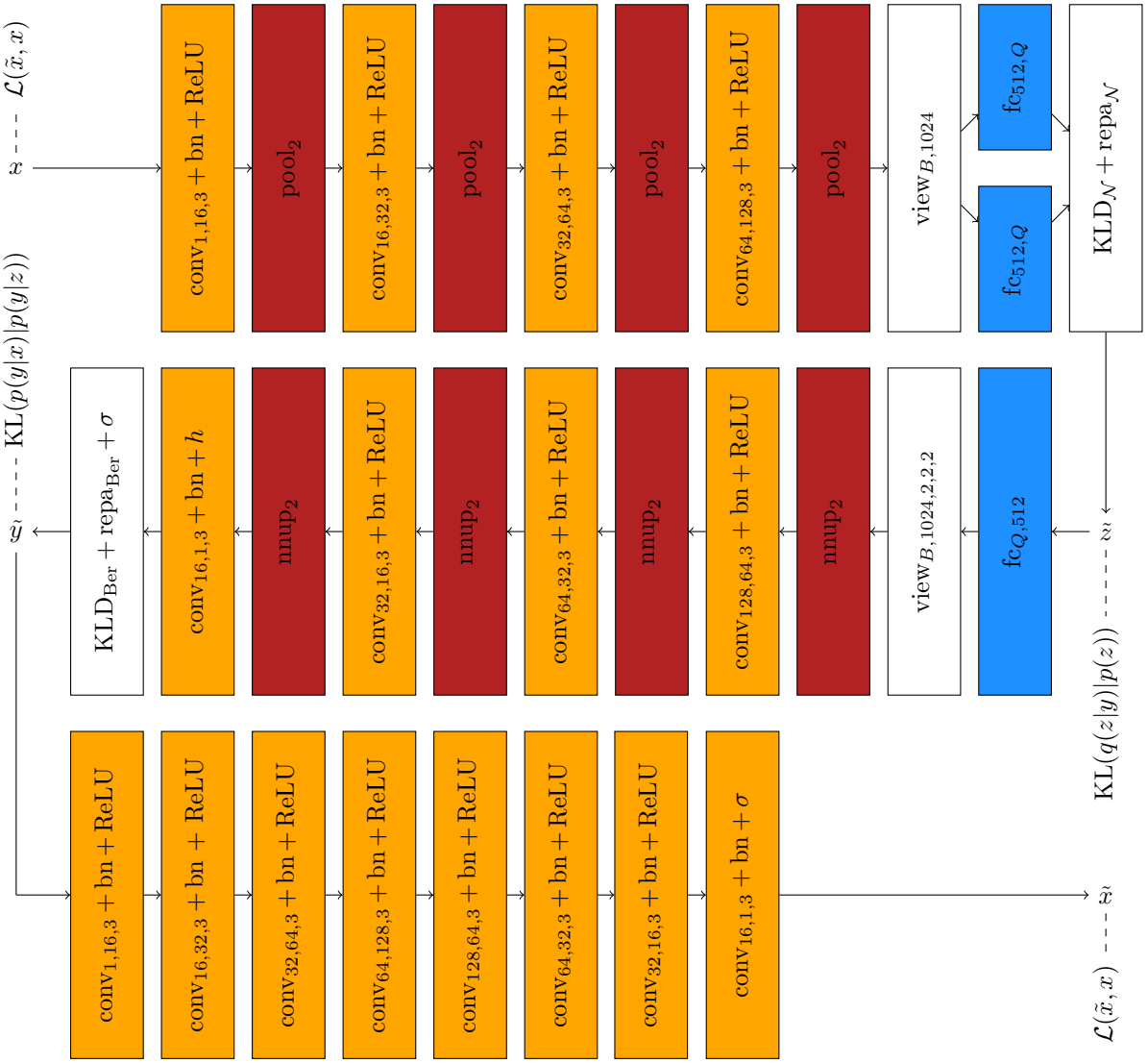

Figure 2: Illustration of the extended variational auto-encoder as also. It is worth comparing the illustrated implementation with Figure 1 to understand that the generative model $p(y|z)$ are taken from the shape prior and the new recognition model $q(z|x)$ follows the architecture of $q(y|z)$. In particular, the decoder, *i.e.* $p(y|z)$, is kept fixed after learning the shape prior. The only completely new part is the convolutional neural network implementing the observation model $p(x|y)$. It consists of of eight convolutional stages including batch normalization and non-linearity. Here, no pooling layers can be found as the size of the output x matches the size of its input y. We additionally make the loss $\mathcal{L}(\tilde{x}, x)$ between reconstructed observation $\tilde{x}$ and original observation x as well as the Kullback-Leibler divergences $\mathrm{KL}(q(z|y)|p(z))$, for the prior $p(z)$, and $\mathrm{KL}(p(y|x)|p(y|z))$, tying observations to predicted shapes, explicit.