Re-param for Plain Inference-time Model

In this subsection, we describe how to convert a trained block into a single 3×3 conv layer for inference. Note that we use BN in each branch before the addition (Fig. 4). Formally, we use $\mathrm{W}^{(3)} \in \mathbb{R}^{C_2 \times C_1 \times 3 \times 3}$ to denote the kernel of a 3×3 conv layer with C_1 input channels and C_2 output channels, and $\mathrm{W}^{(1)} \in \mathbb{R}^{C_2 \times C_1}$ for the kernel of 1×1 branch. We use $\boldsymbol{\mu}^{(3)}, \boldsymbol{\sigma}^{(3)}, \boldsymbol{\gamma}^{(3)}, \boldsymbol{\beta}^{(3)}$ as the accumulated mean, standard deviation and learned scaling factor and bias of the BN layer following 3×3 conv, $\boldsymbol{\mu}^{(1)}, \boldsymbol{\sigma}^{(1)}, \boldsymbol{\gamma}^{(1)}, \boldsymbol{\beta}^{(1)}$ for the BN following 1×1 conv, and $\boldsymbol{\mu}^{(0)}, \boldsymbol{\sigma}^{(0)}, \boldsymbol{\gamma}^{(0)}, \boldsymbol{\beta}^{(0)}$ for the identity branch. Let $\mathrm{M}^{(1)} \in \mathbb{R}^{N \times C_1 \times H_1 \times W_1}$, $\mathrm{M}^{(2)} \in \mathbb{R}^{N \times C_2 \times H_2 \times W_2}$ be the input and output, respectively, and $*$ be the convolution operator. If $C_1 = C_2, H_1 = H_2, W_1 = W_2$, we have

$$\begin{aligned}\mathrm{M}^{(2)} &= \mathrm{bn}(\mathrm{M}^{(1)} * \mathrm{W}^{(3)}, \boldsymbol{\mu}^{(3)}, \boldsymbol{\sigma}^{(3)}, \boldsymbol{\gamma}^{(3)}, \boldsymbol{\beta}^{(3)}) \\ &+ \mathrm{bn}(\mathrm{M}^{(1)} * \mathrm{W}^{(1)}, \boldsymbol{\mu}^{(1)}, \boldsymbol{\sigma}^{(1)}, \boldsymbol{\gamma}^{(1)}, \boldsymbol{\beta}^{(1)}) \\ &+ \mathrm{bn}(\mathrm{M}^{(1)}, \boldsymbol{\mu}^{(0)}, \boldsymbol{\sigma}^{(0)}, \boldsymbol{\gamma}^{(0)}, \boldsymbol{\beta}^{(0)}) \,.\end{aligned} \tag{1}$$

Otherwise, we simply use no identity branch, hence the above equation only has the first two terms. Here bn is the inference-time BN function, formally, $\forall 1 \leq i \leq C_2$,

$$\mathrm{bn}(\mathrm{M}, \boldsymbol{\mu}, \boldsymbol{\sigma}, \boldsymbol{\gamma}, \boldsymbol{\beta})_{:,i,:,:} = (\mathrm{M}_{:,i,:,:} - \boldsymbol{\mu}_i)\frac{\boldsymbol{\gamma}_i}{\boldsymbol{\sigma}_i} + \boldsymbol{\beta}_i \,. \tag{2}$$

We first convert every BN and its preceding conv layer into a conv with a bias vector. Let $\{\mathrm{W}', \mathrm{b}'\}$ be the kernel and bias converted from $\{\mathrm{W}, \boldsymbol{\mu}, \boldsymbol{\sigma}, \boldsymbol{\gamma}, \boldsymbol{\beta}\}$, we have

$$\mathrm{W}'_{i,:,:,:} = \frac{\boldsymbol{\gamma}_i}{\boldsymbol{\sigma}_i}\mathrm{W}_{i,:,:,:} \,, \quad \mathrm{b}'_i = -\frac{\boldsymbol{\mu}_i \boldsymbol{\gamma}_i}{\boldsymbol{\sigma}_i} + \boldsymbol{\beta}_i \,. \tag{3}$$

Then it is easy to verify that $\forall 1 \leq i \leq C_2$,

$$\mathrm{bn}(\mathrm{M} * \mathrm{W}, \boldsymbol{\mu}, \boldsymbol{\sigma}, \boldsymbol{\gamma}, \boldsymbol{\beta})_{:,i,:,:} = (\mathrm{M} * \mathrm{W}')_{:,i,:,:} + \mathrm{b}'_i \,. \tag{4}$$

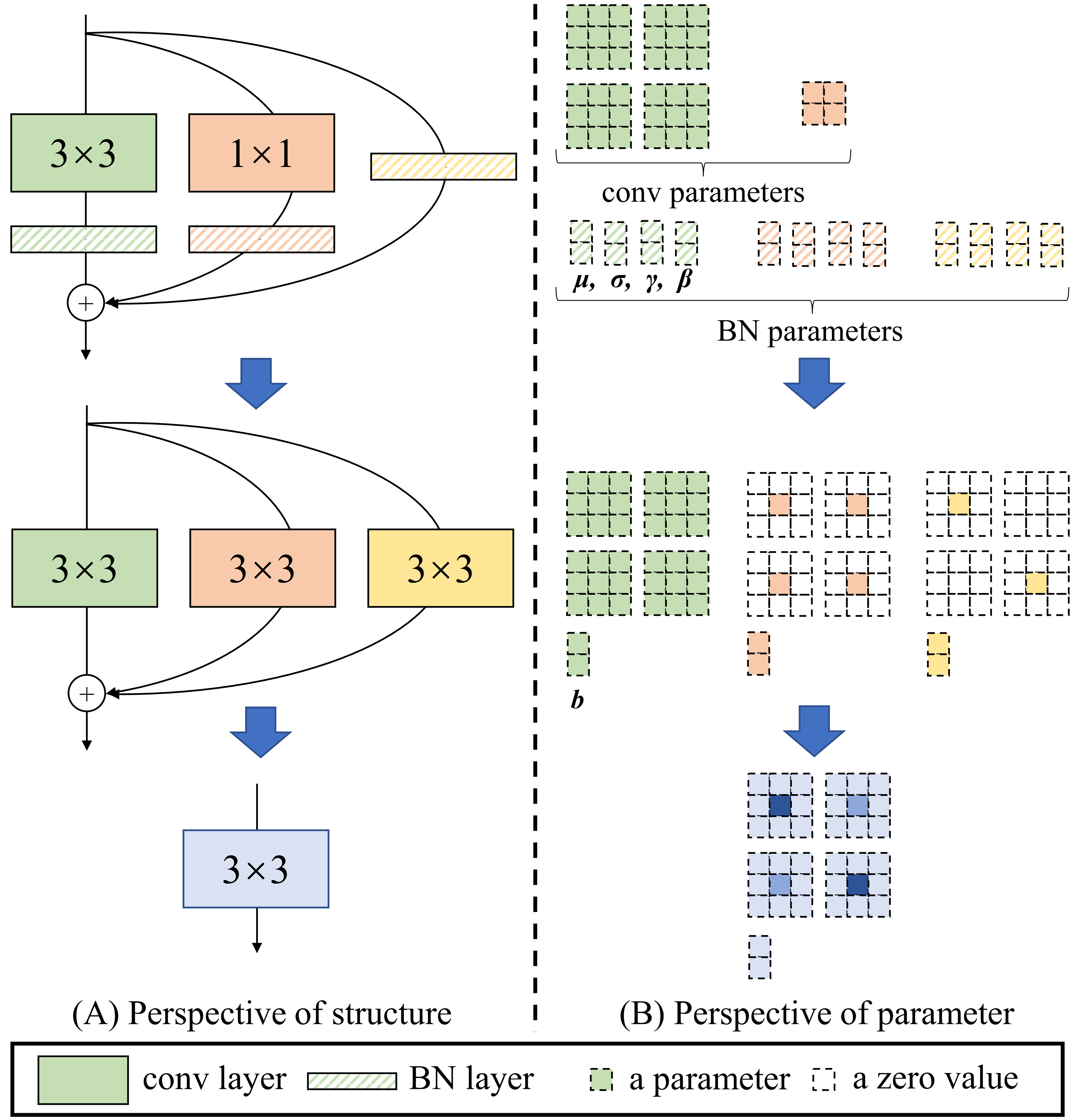

Figure 4: Structural re-parameterization of a RepVGG block. For the ease of visualization, we assume $C_2 = C_1 = 2$, thus the 3×3 layer has four 3×3 matrices and the kernel of 1×1 layer is a 2×2 matrix.

The above transformation also applies to the identity branch because an identity mapping can be viewed as a 1×1 conv with an identity matrix as the kernel.

After such transformations, we will have one 3×3 kernel, two 1×1 kernels, and three bias vectors.

Then we obtain the final bias by adding up the three bias vectors, and the final 3×3 kernel by adding the 1×1 kernels onto the central point of 3×3 kernel, which can be easily implemented by first zero-padding the two 1×1 kernels to 3×3 and adding the three kernels up, as shown in Fig. 4.

Note that the equivalence of such transformations requires the 3×3 and 1×1 layer to have the same stride, and the padding configuration of the latter shall be one pixel less than the former. *E.g.*, for a 3×3 layer that pads the input by one pixel, which is the most common case, the 1×1 layer should have padding = 0.