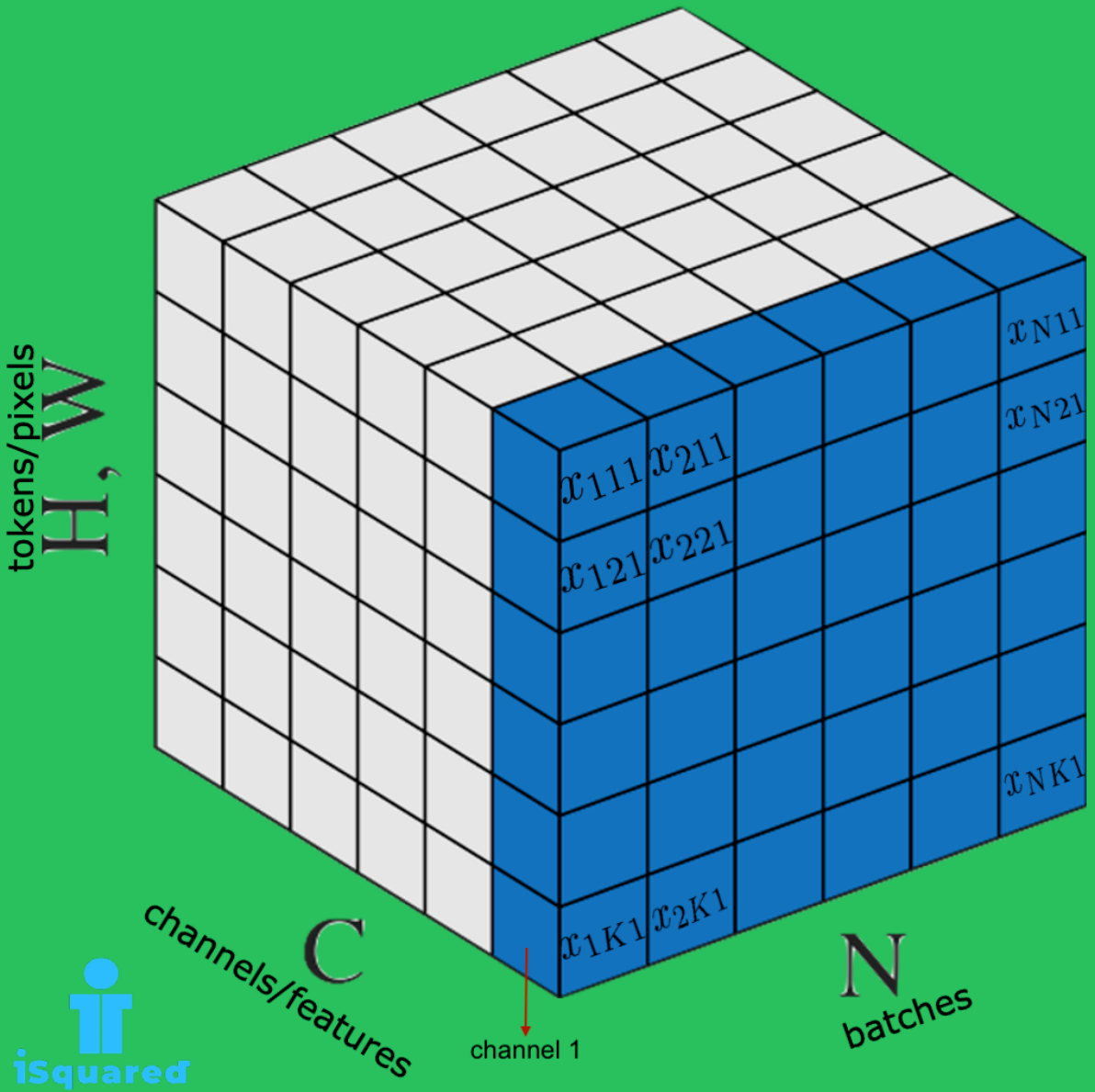
Batch Norm



$$x_{ijk}$$

batch token/pixel channel

Across all pixels/tokens in all batches per channel/feature

$$K = H \times W$$ Number of tokens/pixels

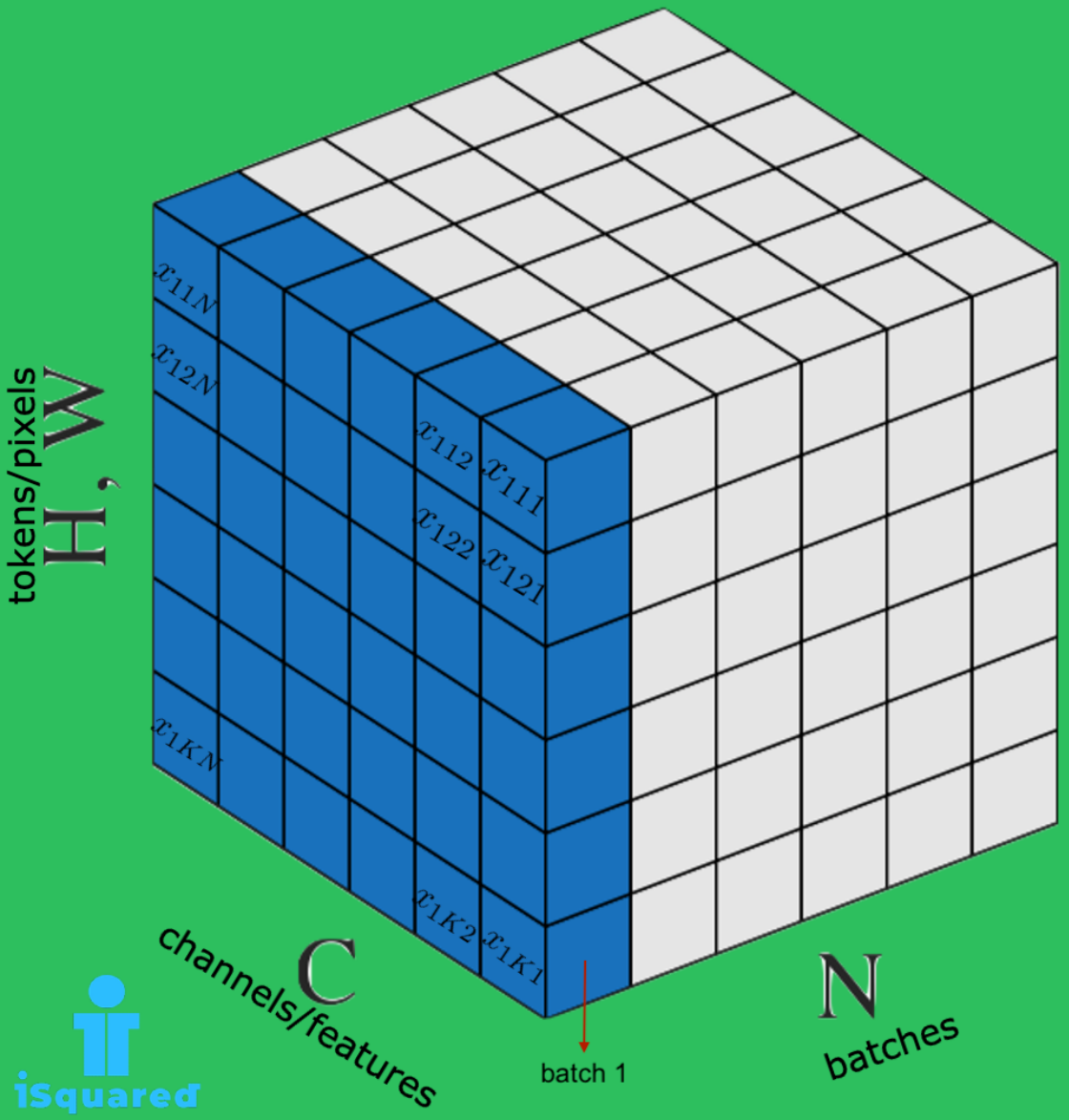
$$\mu_c = \frac{1}{NK} \sum_{i=1}^{N} \sum_{j=1}^{K} x_{i,j,c}$$

$$\sigma_c^2 = \frac{1}{NK} \sum_{i=1}^{N} \sum_{j=1}^{K} (x_{i,j,c} - \mu_c)^2$$

$$\hat{x}_{i,j,c} = \frac{x_{i,j,c} - \mu_c}{\sqrt{\sigma_c^2 + \epsilon}}$$

iSquared

Layer Norm



$$x_{ijk}$$

batch token/pixel channel

Across all pixels/tokens in all channels/features per batch

$$K = H \times W$$ Number of tokens/pixels

$$\mu_b = \frac{1}{KC} \sum_{j=1}^{K} \sum_{k=1}^{C} x_{b,j,k}$$

$$\sigma_b^2 = \frac{1}{K} \sum_{j=1}^{K} \sum_{k=1}^{C} (x_{b,j,k} - \mu_b)^2$$

$$\hat{x}_{b,j,k} = \frac{x_{b,j,k} - \mu_b}{\sqrt{\sigma_b^2 + \epsilon}}$$

iSquared