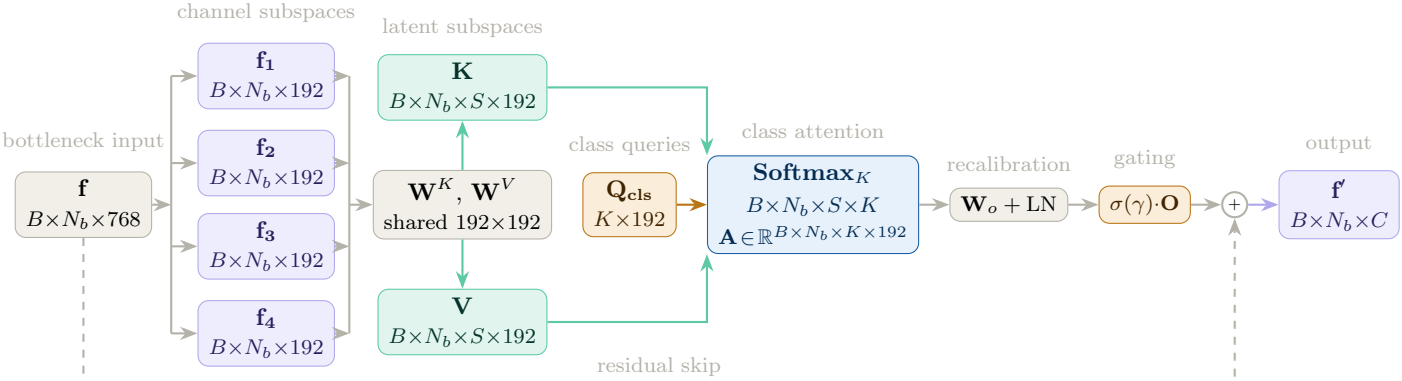

channel subspaces
latent subspaces
bottleneck input
class queries
class attention
recalibration
gating
output
$\mathbf{f}$
$B\times N_b\times 768$
$\mathbf{f_1}$
$B\times N_b\times 192$
$\mathbf{f_2}$
$B\times N_b\times 192$
$\mathbf{f_3}$
$B\times N_b\times 192$
$\mathbf{f_4}$
$B\times N_b\times 192$
$\mathbf{W}^K$, $\mathbf{W}^V$
shared 192×192
$\mathbf{K}$
$B\times N_b\times S\times 192$
$\mathbf{V}$
$B\times N_b\times S\times 192$
$\mathbf{Q_{cls}}$
$K\times 192$
$\mathbf{Softmax}_K$
$B\times N_b\times S\times K$
$\mathbf{A}\in\mathbb{R}^{B\times N_b\times K\times 192}$
$\mathbf{W}_o+\mathrm{LN}$
$\sigma(\gamma)\cdot\mathbf{O}$
$+$
$\mathbf{f'}$
$B\times N_b\times C$
residual skip