Inputs

Frame $(t)$ + Boxes $(t-1)$

ROI Extraction

32K params

Fusion & Temporal

196K params

$\mathcal{DCT}(\Delta Y_n^g)$
$80\times80\times C$

spatial regions

**Conv**
$80\times80\times32$

**ROI Pool**
$N\times8\times8\times32$

**DCT Stats**
$N\times32$

**Boxes** $(t-1)$
$N\times4$

**Box Enc**
$N\times32$

spatial regions

$MV_n^g$
$60\times60\times2$

**ROI Pool**
$N\times8\times8\times2$

**MV Stats**
$N\times64$

**Concat**
$N\times128$

**BiLSTM**
$N\times128$

**Δpos**
$N\times2$

**Δsize**
$N\times2$

**Boxes** $(t)$
$N\times4$