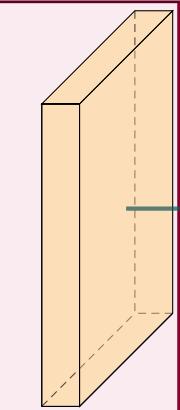


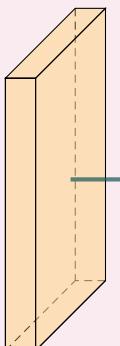
Inputs

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



$DCCT(\Delta Y_n^g)$
 $80 \times 80 \times C$

Box Enc
 $N \times 32$



MV_n^g
 $60 \times 60 \times 2$

spatial regions

spatial
regions
Conv
 $80 \times 80 \times 32$

ROI Pool
 $N \times 8 \times 8 \times 2$

ROI Pool
 $N \times 8 \times 8 \times 32$

DCT Stats
 $N \times 32$

MV Stats
 $N \times 64$

ROI Extraction

32K params

Concat
 $N \times 128$

BiLSTM
 $N \times 128$

Δpos
 $N \times 2$

$\Delta size$
 $N \times 2$

Boxes (t)
 $N \times 4$

196K params

Fusion & Temporal