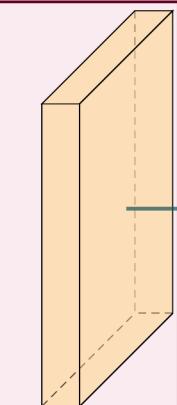
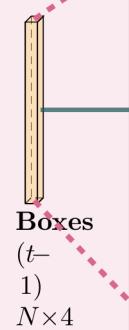


## Inputs

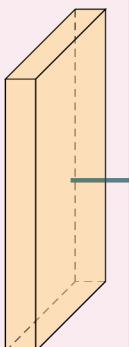
Frame ( $t$ ) + Boxes ( $t-1$ )



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



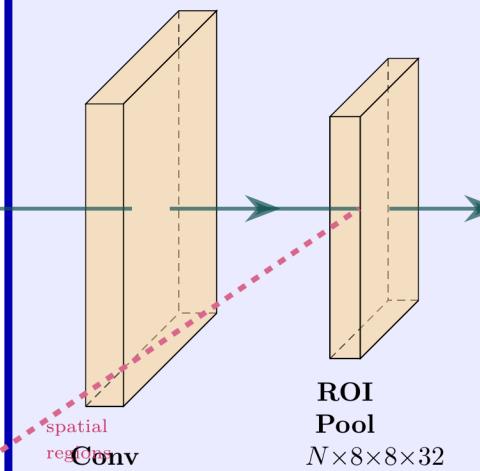
Boxes  
( $t-1$ )  
 $N \times 4$



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

## ROI Extraction

32K params



spatial  
regConv  
80×80×32

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

DCT  
Stats  
 $N \times 32$

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

MV  
Stats  
 $N \times 64$

Concat  
 $N \times 128$

BiLSTM  
 $N \times 128$

$\Delta pos$   
 $N \times 2$

$\Delta size$   
 $N \times 2$

Boxes  
( $t$ )  
 $N \times 4$

196K params

## Fusion & Temporal