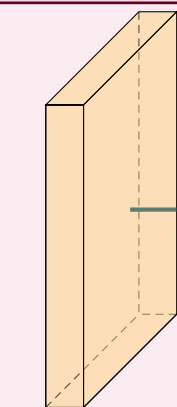


## Inputs

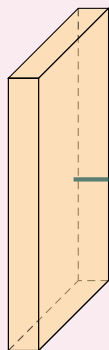
Frame ( $t$ ) + Boxes ( $t-1$ )



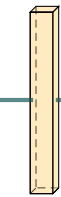
$\mathcal{DCT}(\Delta Y_n^g)$   
 $80 \times 80 \times C$



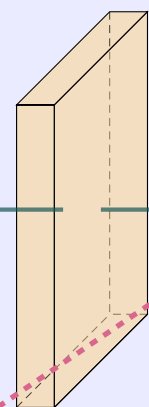
**Boxes**  
( $t-1$ )  
 $N \times 4$



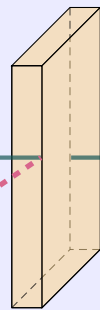
$MV_n^g$   
 $60 \times 60 \times 2$



**Box  
Enc**  
 $N \times 32$



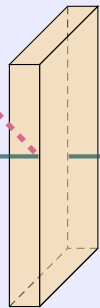
**Conv**  
 $80 \times 80 \times 32$



**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**  
 $N \times 64$

## ROI Extraction

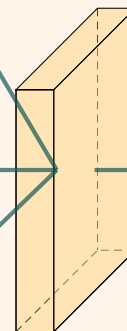
32K params

spatial  
regions

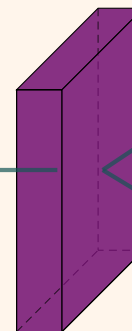
spatial  
regions

## Fusion & Temporal

196K params



**Concat**  
 $N \times 128$



**BiLSTM**  
 $N \times 128$



$\Delta \text{pos}$   
 $N \times 2$



$\Delta \text{size}$   
 $N \times 2$



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