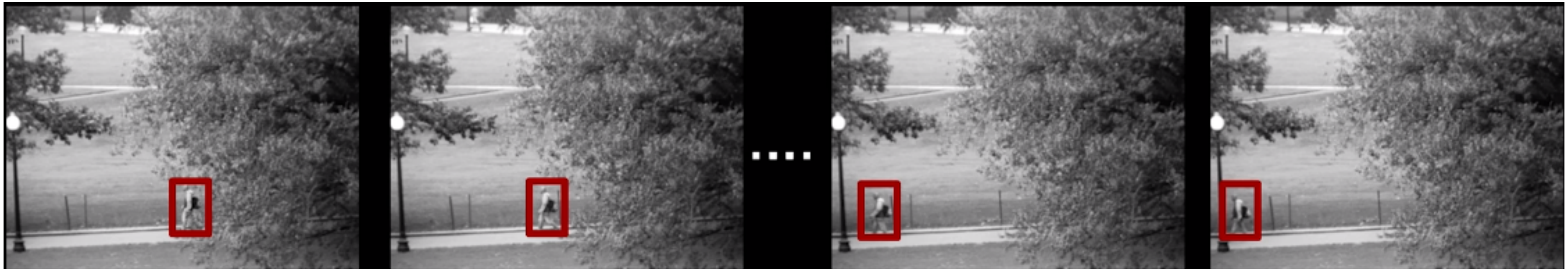
Detect to Track and Track to Detect

Christoph Feichtenhofer, Axel Pinz, Andrew Zisserman, "Detect to Track and Track to Detect", ICCV 2017

2019. 02. 11

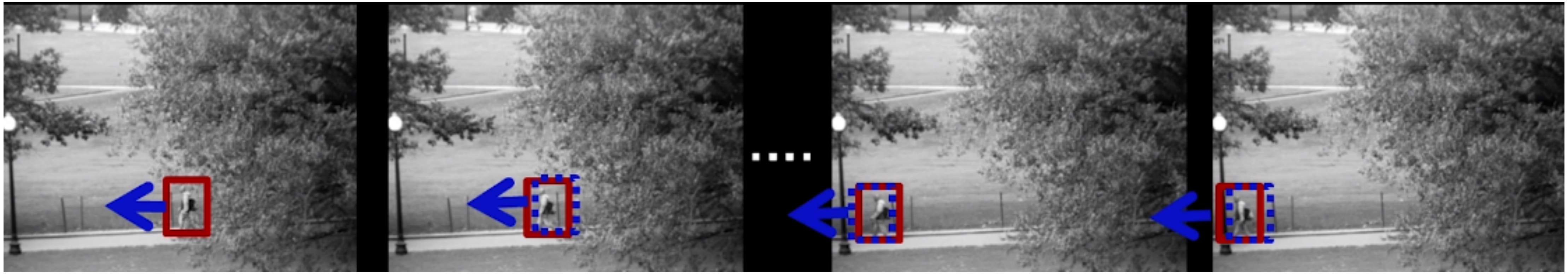
Minji Jo, Hanyang Univ.

Detection vs. tracking



Detection: We detect the object **independently** in each frame

Detection vs. tracking

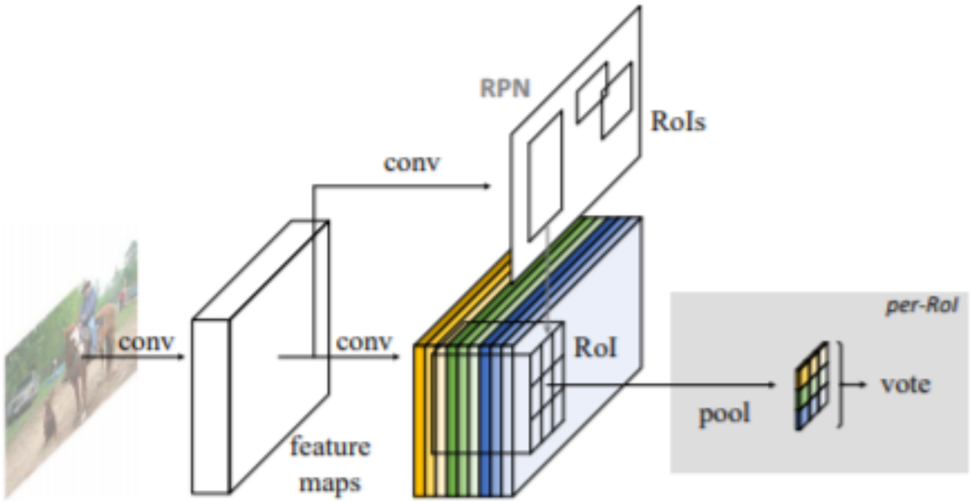


Tracking: We **predict** the new location of the object in the next frame using **estimated dynamics**. Then we **update** based upon measurements.

Related work

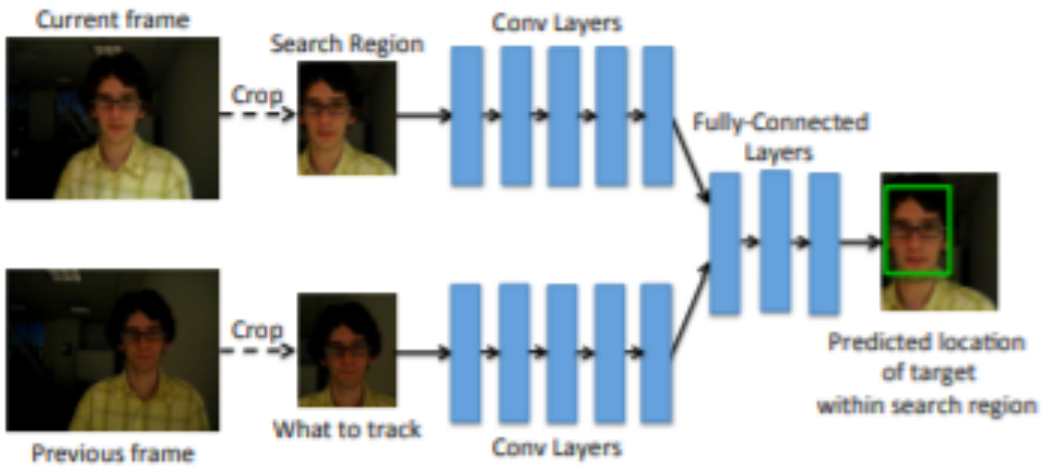
• Object detection:

R-FCN



• Tracking:

Learning to Track at 100 FPS with Deep Regression Networks



Detect & Track Architecture

FCN

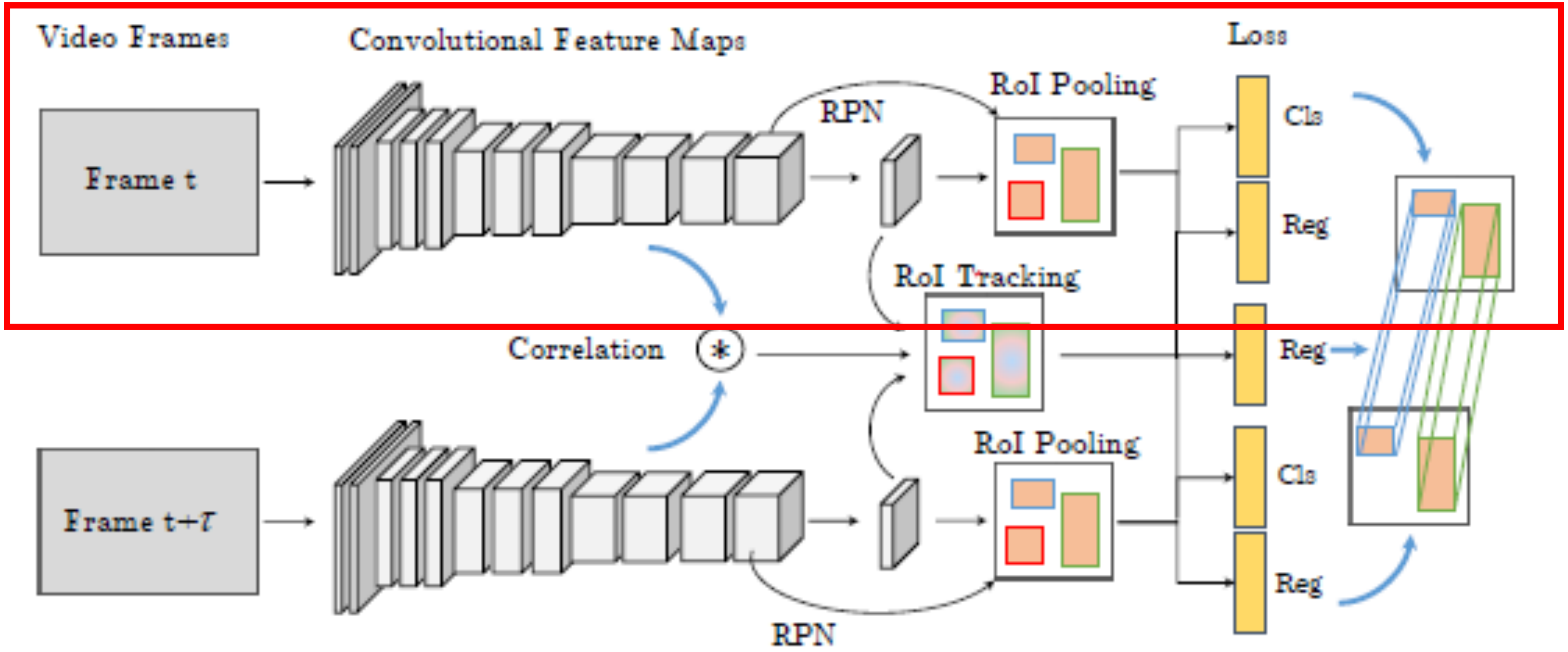


Figure 2. Architecture of our Detect and Track (D&T) approach (see Section 3 for details).

Detect & Track Architecture

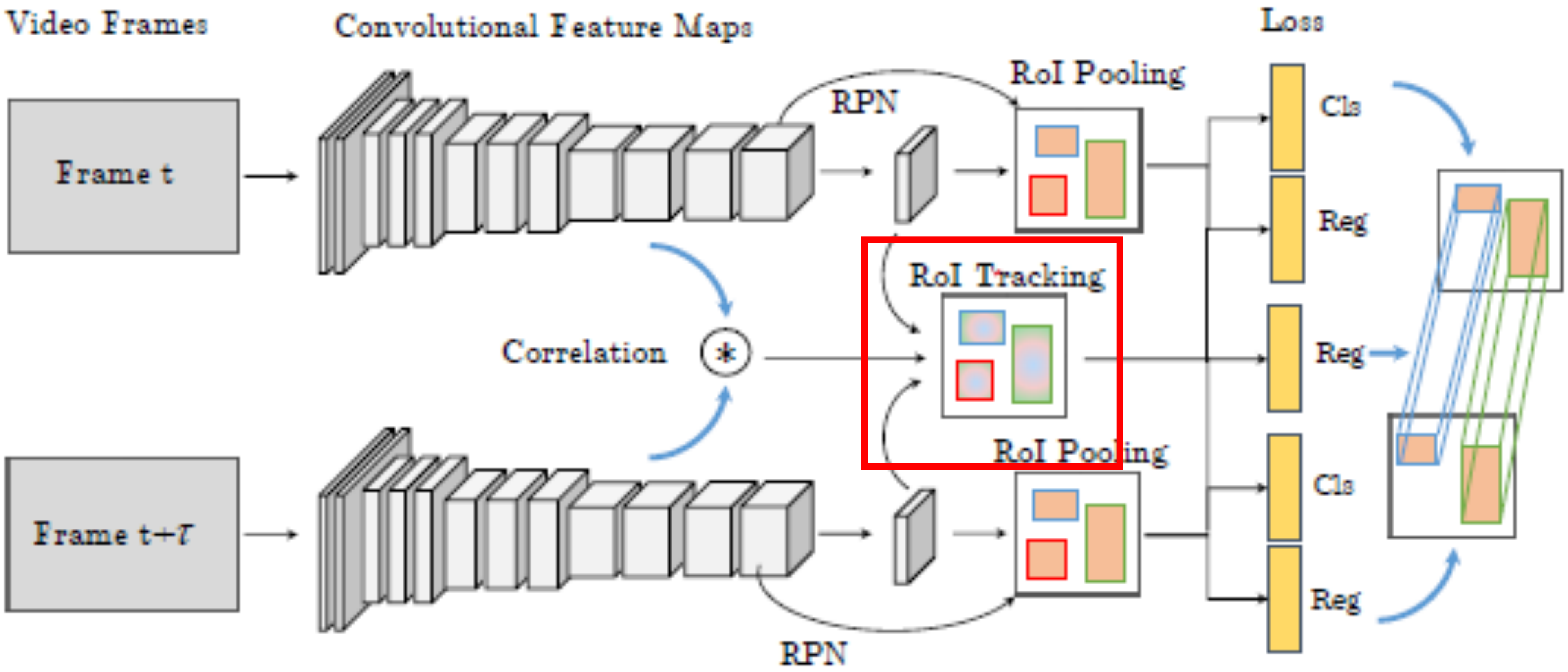


Figure 2. Architecture of our Detect and Track (D&T) approach (see Section 3 for details).

Detect & Track Architecture

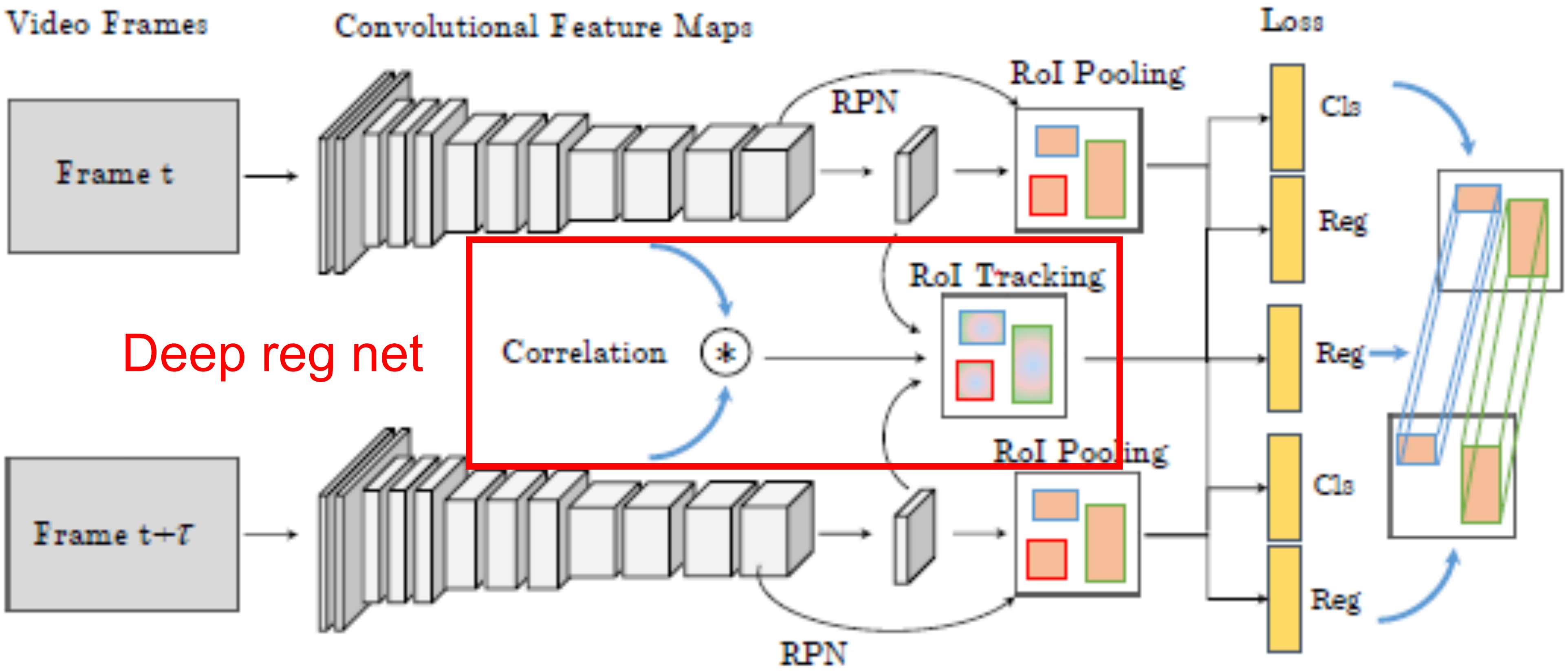
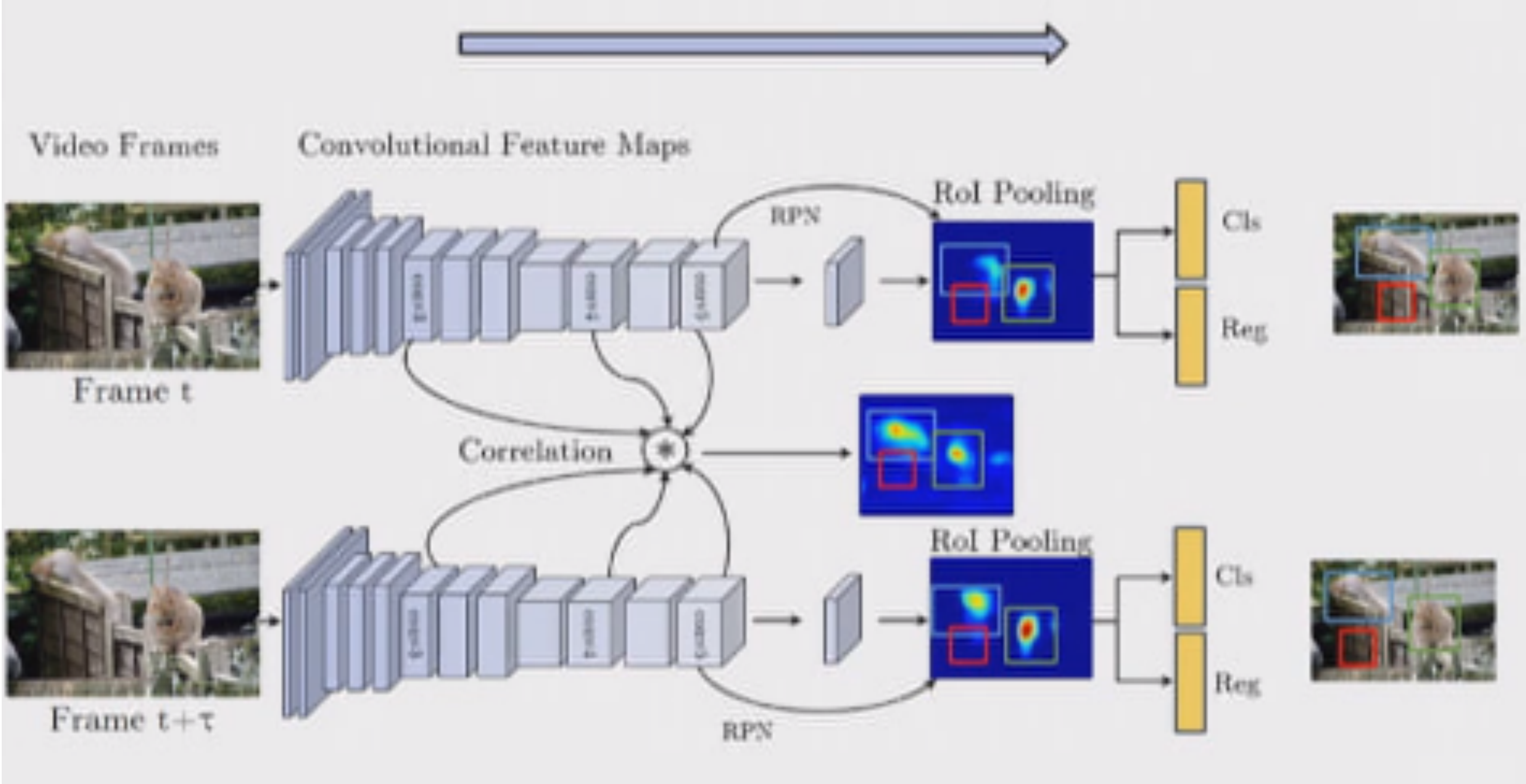
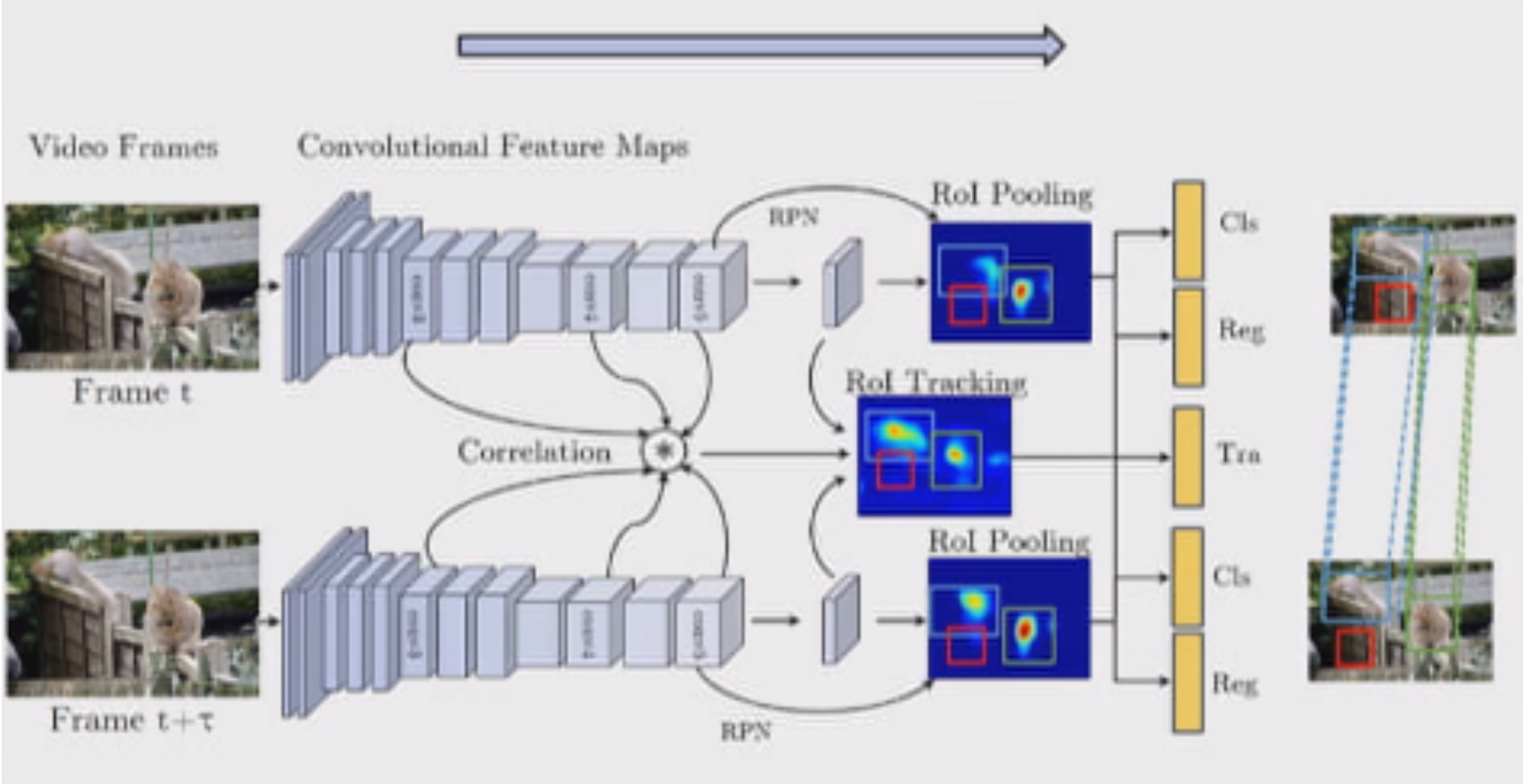


Figure 2. Architecture of our Detect and Track (D&T) approach (see Section 3 for details).

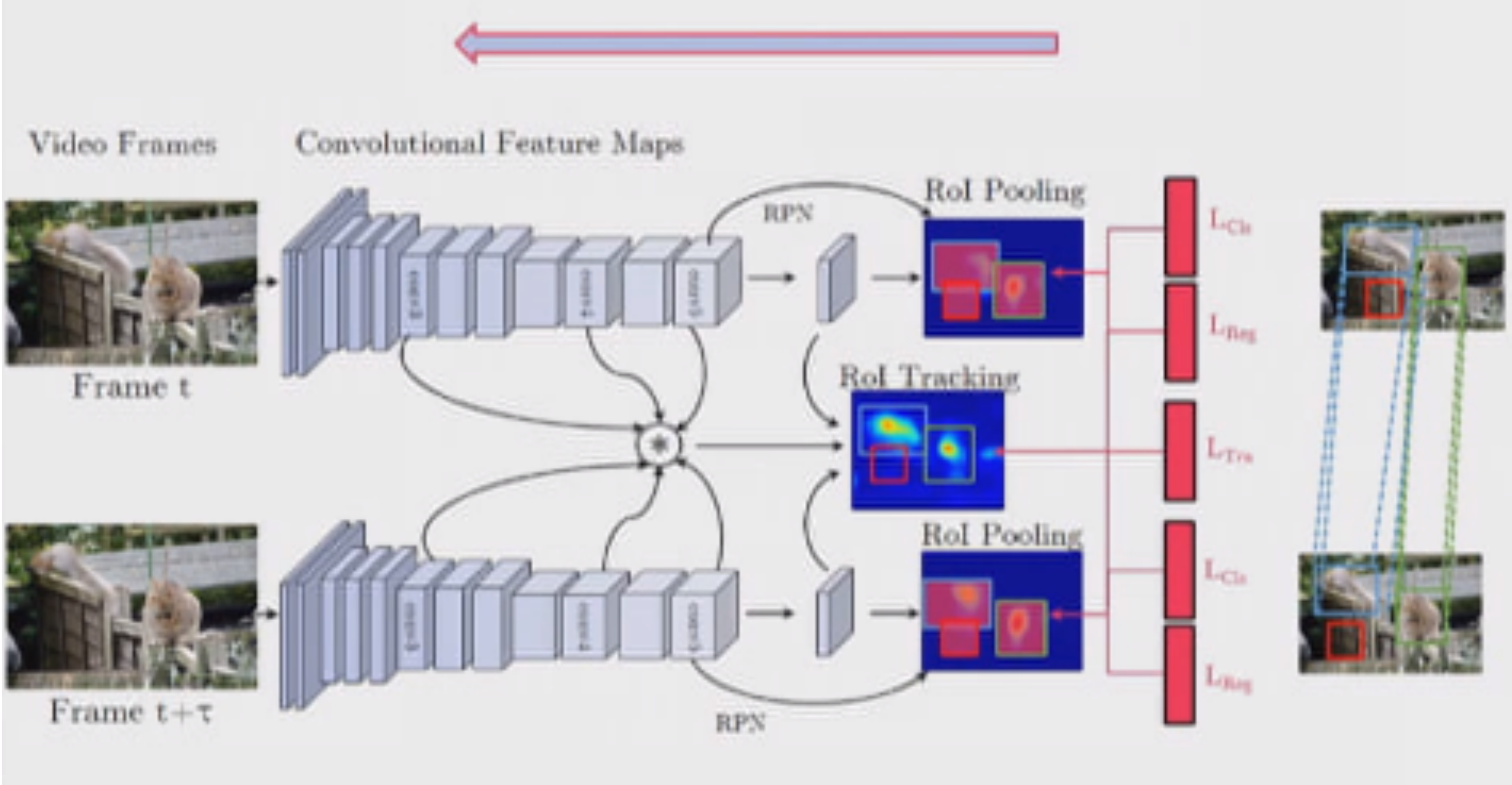
Detect & Track Training: Forward pass



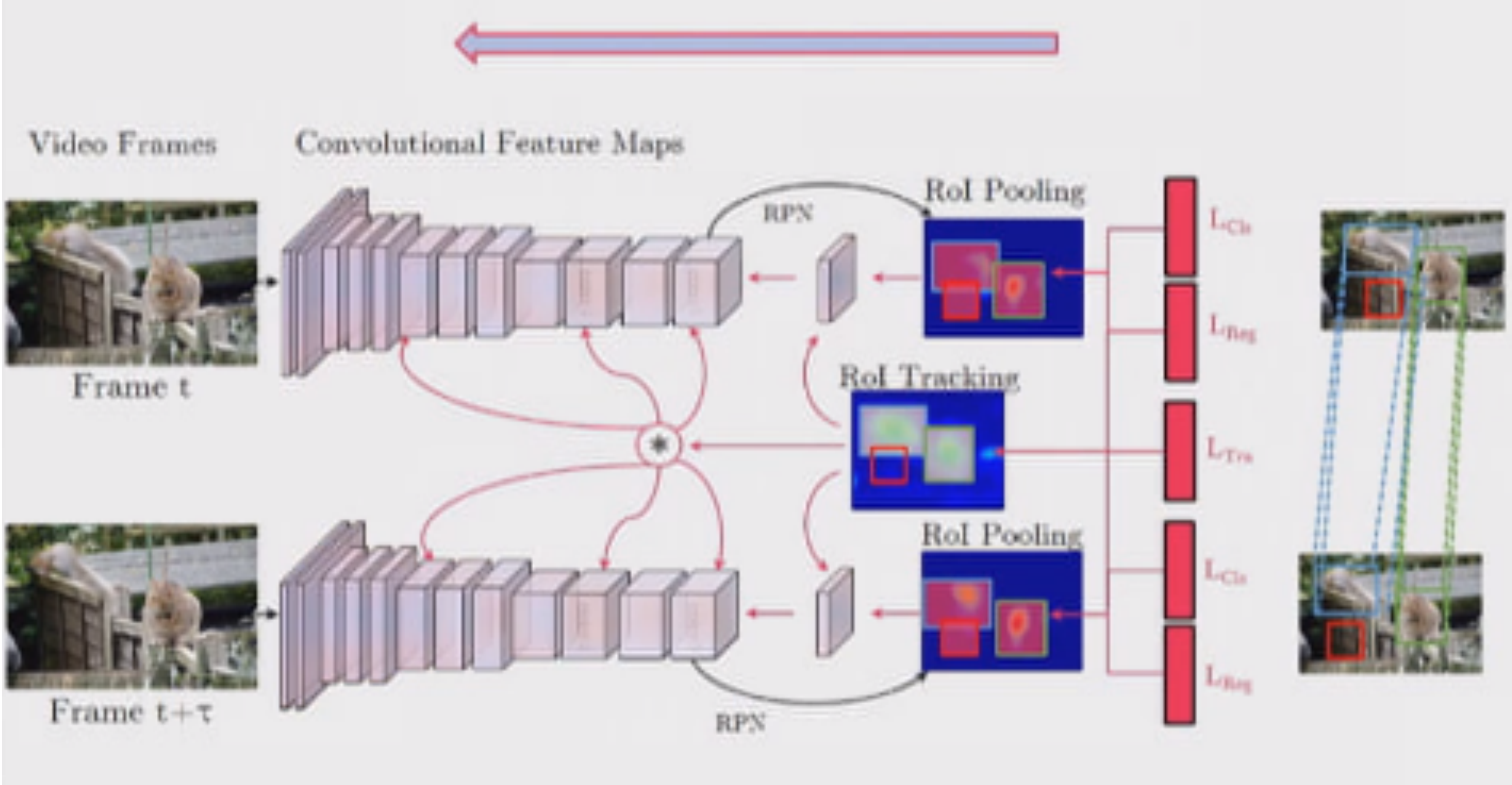
Detect & Track Training: Forward pass



Detect & Track Training: Backward pass



Detect & Track Training: Backward pass



objective

$$L(\{p_i\},\{t_i\}) = \frac{1}{N_{cls}}\sum_i L_{cls}(p_i, p_i^*) + \lambda\frac{1}{N_{reg}}\sum_i p_i^* L_{reg}(t_i, t_i^*)$$

$$L(\{p_i\},\{b_i\},\boxed{\{\Delta_i\}}) = \frac{1}{N}\sum_{i=1}^{N} L_{cls}(p_{i,c^*})$$

$$+\lambda\frac{1}{N_{fg}}\sum_{i=1}^{N}[c_i^* > 0] L_{reg}(b_i, b_i^*)$$

$$\boxed{+\lambda\frac{1}{N_{tra}}\sum_{i=1}^{N_{tra}} L_{tra}(\Delta_i^{t+\tau}, \Delta_i^{*,t+\tau}).}$$

bounding box regression features $\{\mathbf{x}_{reg}^{t}, \mathbf{x}_{reg}^{t+\tau}\}$

$$\Delta^{t+\tau} = (\Delta_x^{t+\tau}, \Delta_y^{t+\tau}, \Delta_w^{t+\tau}, \Delta_h^{t+\tau})$$

$$\Delta_x^{*,t+\tau} = \frac{B_x^{t+\tau} - B_x^t}{B_w^t} \qquad \Delta_y^{*,t+\tau} = \frac{B_y^{t+\tau} - B_y^t}{B_h^t}$$

$$\Delta_w^{*,t+\tau} = \log(\frac{B_w^{t+\tau}}{B_w^t}) \qquad \Delta_h^{*,t+\tau} = \log(\frac{B_h^{t+\tau}}{B_h^t})).$$

$$\Delta^{*,t+\tau} = \{\Delta_x^{*,t+\tau}, \Delta_y^{*,t+\tau}, \Delta_w^{*,t+\tau}, \Delta_h^{*,t+\tau}\}$$

Correlation feature for object tracking

$$\mathbf{x}_{corr}^{t,t+\tau}(i,j,p,q) = \left\langle \mathbf{x}_l^t(i,j), \mathbf{x}_l^{t+\tau}(i+p, j+q) \right\rangle$$

where $-d \le p \le d$ and $-d \le q \le d$

defined by the maximum displacement, d

RoI Tracking's input $\{\mathbf{x}_{corr}^{t,t+\tau}, \mathbf{x}_{reg}^{t}, \mathbf{x}_{reg}^{t+\tau}\}$.

Linking tracklets to object tubes

$$D_i^{t,c} = \{x_i^t, y_i^t, w_i^t, h_i^t, p_{i,c}^t\}$$

$$T_i^{t,t+\tau} = \{x_i^t, y_i^t, w_i^t, h_i^t; x_i^t + \Delta_x^{t+\tau}, y_i^t + \Delta_y^{t+\tau}, w_i^t + \Delta_w^{t+\tau}, h_i^t + \Delta_h^{t+\tau}\}$$

$$\Downarrow$$

$$s_c(D_{i,c}^t, D_{j,c}^{t+\tau}, T^{t,t+\tau}) = p_{i,c}^t + p_{j,c}^{t+\tau} + \psi(D_i^t, D_j, T^{t,t+\tau})$$

$$\bar{D}_c^\star = \underset{\bar{D}}{\operatorname{argmax}} \frac{1}{\mathcal{T}} \sum_{t=1}^{\mathcal{T}-\tau} s_c(D^t, D^{t+\tau}, T^{t,t+\tau}).$$

Result

Methods	airplane	antelope	bear	bicycle	bird	bus	car	cattle	dog	d. cat	elephant	fox	g. panda	hamster	horse	lion
TCN [18]	72.7	75.5	42.2	39.5	25.0	64.1	36.3	51.1	24.4	48.6	65.6	73.9	61.7	82.4	30.8	34.4
TPN+LSTM [16]	84.6	78.1	72.0	67.2	68.0	80.1	54.7	61.2	61.6	78.9	71.6	83.2	78.1	91.5	66.8	21.6
Winner ILSVRC'15 [17]	83.7	85.7	84.4	74.5	73.8	75.7	57.1	58.7	72.3	69.2	80.2	83.4	80.5	93.1	84.2	67.8
D (R-FCN)	87.4	79.4	84.5	67.0	72.1	84.6	54.6	72.9	70.9	77.3	76.7	89.7	77.6	88.5	74.8	57.9
D (& T loss)	89.4	80.4	83.8	70.0	71.8	82.6	56.8	71.0	71.8	76.6	79.3	89.9	83.3	91.9	76.8	57.3
D&T ($\tau = 1$)	90.2	82.3	87.9	70.1	73.2	87.7	57.0	80.6	77.3	82.6	83.0	97.8	85.8	96.6	82.1	66.7
D&T ($\tau = 10$)	89.1	79.8	87.5	68.8	72.9	86.1	55.7	78.6	76.4	83.4	82.9	97.0	85.0	96.0	82.2	66.0

Methods	lizard	monkey	motorcycle	rabbit	red panda	sheep	snake	squirrel	tiger	train	turtle	watercraft	whale	zebra	mAP (%)
TCN [18]	54.2	1.6	61.0	36.6	19.7	55.0	38.9	2.6	42.8	54.6	66.1	69.2	26.5	68.6	47.5
TPN+LSTM [16]	74.4	36.6	76.3	51.4	70.6	64.2	61.2	42.3	84.8	78.1	77.2	61.5	66.9	88.5	68.4
Winner ILSVRC'15 [17]	80.3	54.8	80.6	63.7	85.7	60.5	72.9	52.7	89.7	81.3	73.7	69.5	33.5	90.2	73.8
Winner ILSVRC'16 [39]	(single model performance)														76.2
D (R-FCN)	76.8	50.1	80.2	61.3	79.5	51.9	69.0	57.4	90.2	83.3	81.4	68.7	68.4	90.9	74.2
D (& T loss)	79.0	54.1	80.3	65.3	85.3	56.9	74.1	59.9	91.3	84.9	81.9	68.3	68.9	90.9	75.8
D&T ($\tau = 1$)	83.4	57.6	86.7	74.2	91.6	59.7	76.4	68.4	92.6	86.1	84.3	69.7	66.3	95.2	**79.8**
D&T ($\tau = 10$)	83.1	57.9	79.8	72.7	90.0	59.4	75.6	65.4	90.5	85.6	83.3	68.3	66.5	93.2	78.6

Table 1. Performance comparison on the ImageNet VID validation set. The average precision (in %) for each class and the mean average precision over all classes is shown. τ corresponds to the temporal sampling stride.