Generating self-supervised training data

Template-Aware Dynamic Convolution Module (TDCM)
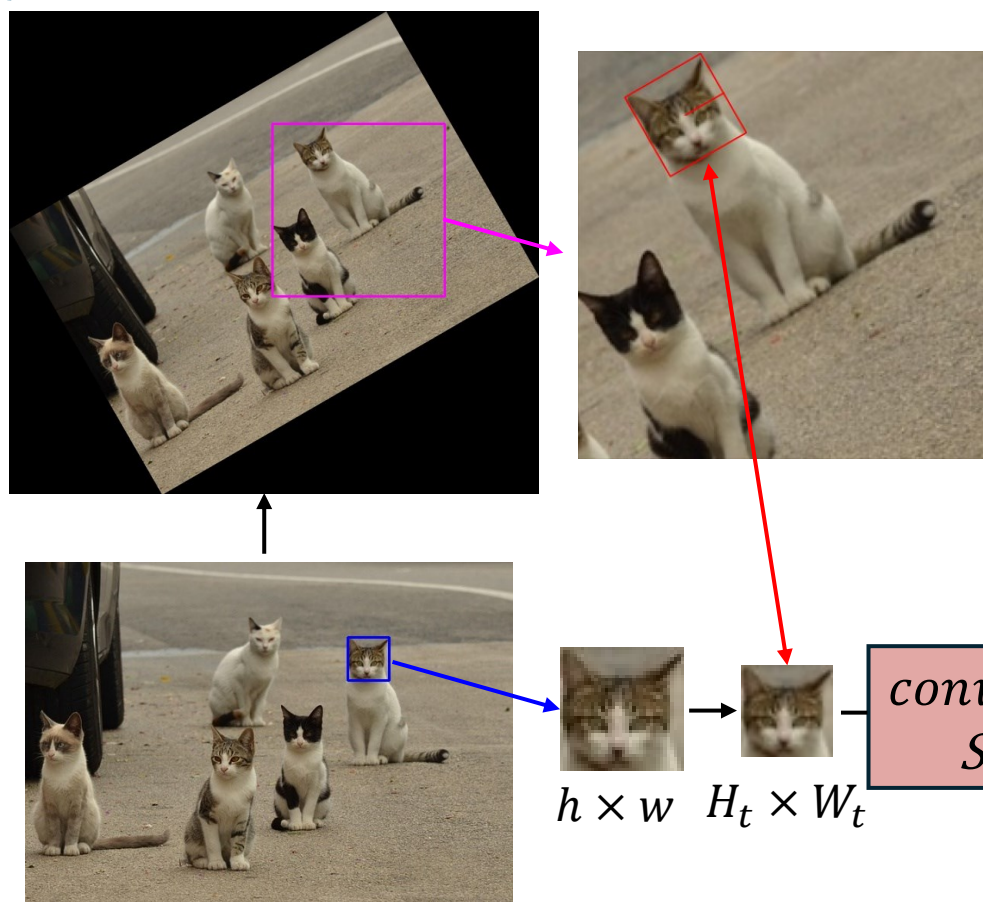
$convNext\text{-}V2$
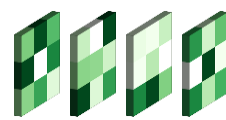$Stage\text{-}1$

$F_i: C \times \frac{H}{4} \times \frac{W}{4}$

$h \times w$ $H_t \times W_t$
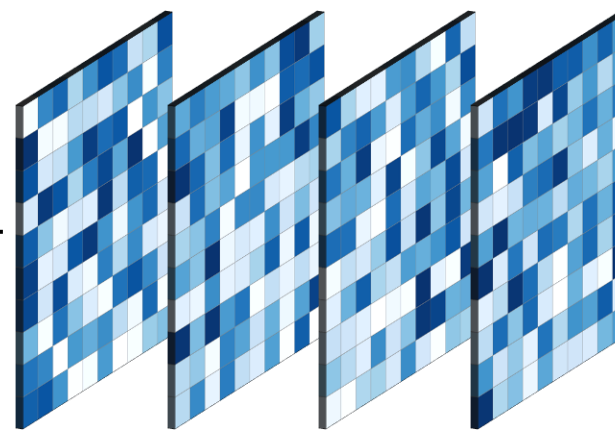
$convNext\text{-}V2$
$Stage\text{-}1$

$F_t: C \times \frac{H_t}{4} \times \frac{W_t}{4}$

as dynamic conv kernels

Dynamic
Depthwise
Convolution

$C \times \frac{H}{4} \times \frac{W}{4}$

$\frac{C}{4} \times H \times W$

$C \times \frac{H}{4} \times \frac{W}{4}$

1×1
$Conv$

*Pixel
Shuffling*

$Conv1$

*Pixel
Shuffling*

$Conv2$

$C \times \frac{H}{4} \times \frac{W}{4}$ $\frac{C}{4} \times \frac{H}{2} \times \frac{W}{2}$ $\frac{C}{2} \times \frac{H}{2} \times \frac{W}{2}$ $\frac{C}{8} \times H \times W$ $\frac{C}{4} \times H \times W$

regression

$Conv1$ $BN1$ $ReLU$ $Conv2$

$\frac{C}{4} \times H \times W$ $\frac{C}{8} \times H \times W$ $\frac{C}{16} \times H \times W$

$Sigmoid$ $Conv3$ $ReLU$ $BN2$

$1 \times H \times W$ $1 \times H \times W$
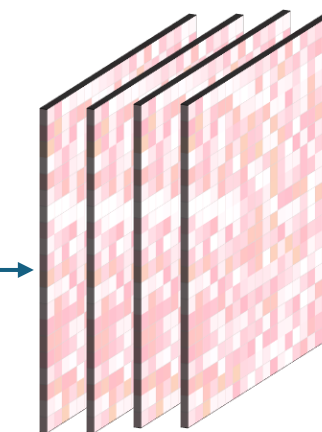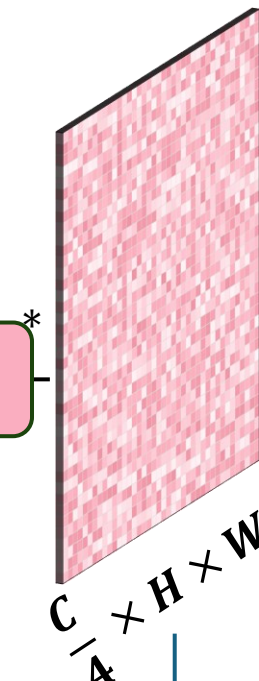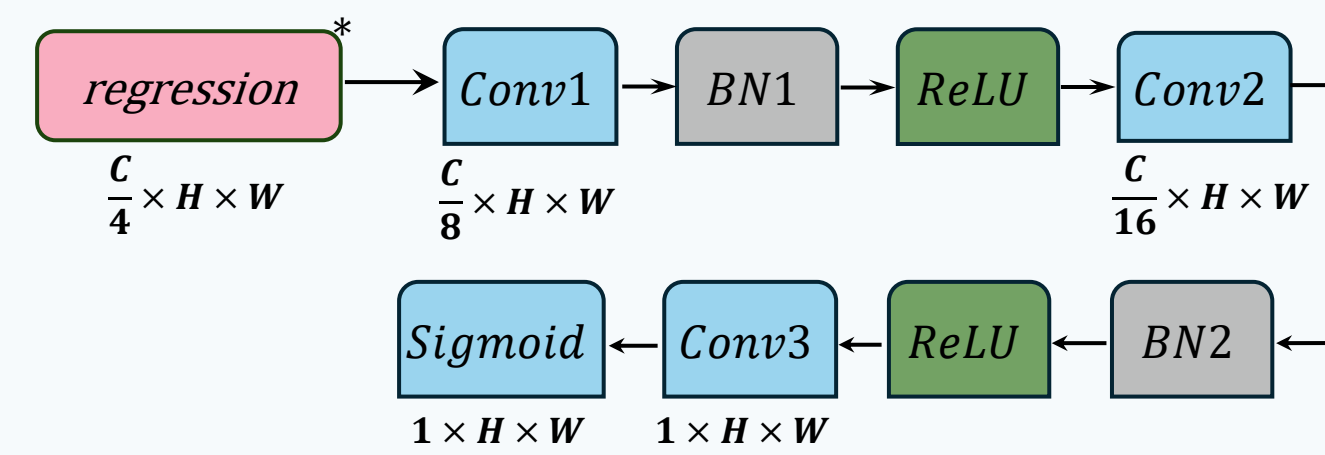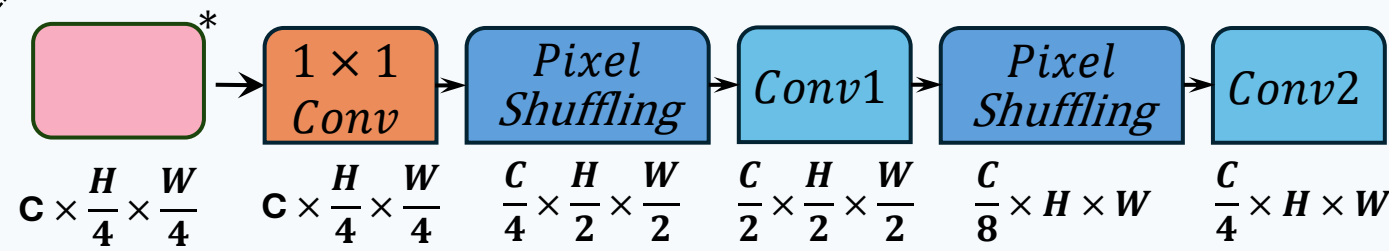
$\Phi^*: (x_c^*, y_c^*, s_x^*, s_y^*, \theta^*)$
$\Phi: (x_c, y_c, s_x, s_y, \theta)$

Φ^* *Refine* Φ
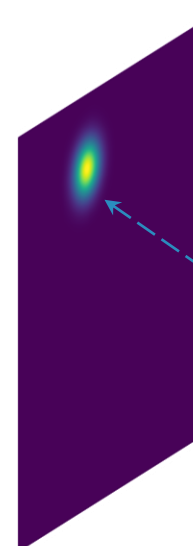
classification

$\mathcal{L}_{\text{center}}$

Extract ROI

$\mathcal{L}_{\text{sign}(\theta)}^{CE}$
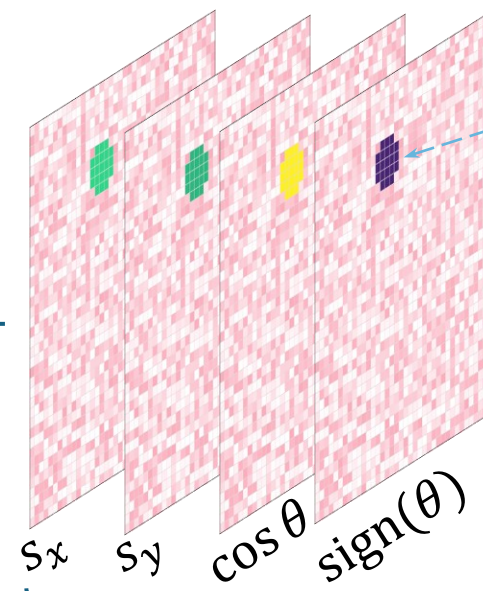
$\mathcal{L}_{geom}$ *regression*

s_x s_y $\cos\theta$ $\text{sign}(\theta)$

Center localization & geometry regression