SiamMask (CVPR 2019)

Fast Online Object Tracking and Segmentation: A Unifying Approach

Wang, Qiang, et al.

① Problem Description:

VOT (visual object tracking)

VOS (video object segementation)

Initialisation: Tracking / Segementation.

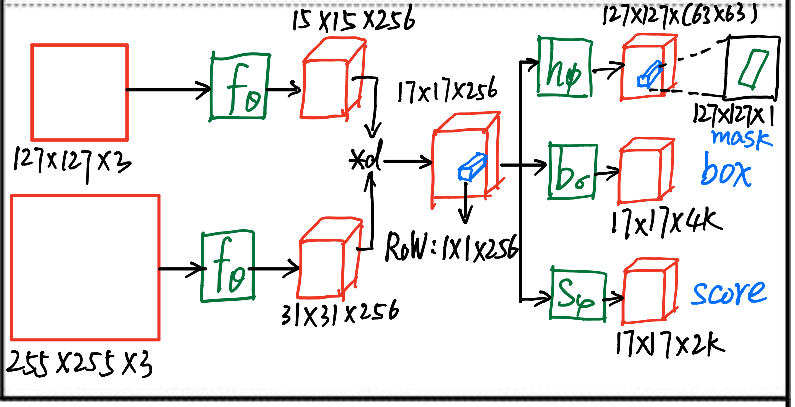
Outputs: axis-aligned bounding box.

② Problem Solution:

Initialisation: a single bounding box for both tracking and segementation.

Outputs: rotated bounding box. and binary segementation mask.

③ Conceptual Understanding



Depth-wise cross correlation: $g_\theta(z, x) = f_\theta(z) * f_\theta(x)$.

Predict n-th RoW mask: $m_n = h_\phi(g_\theta^n(z, x))$.

Loss function: $L_{mask}(\theta, \phi) = \sum_n \left(\frac{1 + \overset{\swarrow \{\pm1\}\,GT}{y_n}}{2wh} \sum_{ij} \log(1 + e^{-\overset{\swarrow \{\pm1\}\,RoW}{c_n^{ij}} m_n^{ij}}) \right)$

multi-task losses: $\begin{cases} L_{2B} = \lambda_1 L_{mask} + \lambda_2 L_{sim}. \\ L_{3B} = \lambda_1 L_{mask} + \lambda_2 L_{box} + \lambda_3 L_{score}. \end{cases}$

Details of implementation

Implementation:

① Network architecture

- backbone: ResNet-50;
- head: conv5 -> Norm, ReLU
  conv6 -> 1×1 conv.
- mask refinement module

② Training

FC: similarity measure learning

RPN: bounding box regression

class-agnostic binary segemetation

③ Inference

evaluated once per frame

output mask with maximum score

binarise with threshold of 0.5.

Architecture:

① backbone

ResNet: before 4-th stage.
(share parameters)

adjust layer: 1×1, 256
(not shared)

depth-wise xcorr: 17×17.

② head

	conv5	conv6
mask	1×1, 256	1×1, (63×63)
box	1×1, 256	1×1, 4k
score	1×1, 256	1×1, 2k

③ Refinement

merge resolution feature:

- upsampling layer
- skip connection

Improvement:

① online learning: Siamese / CF

② accurate output: CornerNet / PoseTrack, ExtremeNet.

③ Network: accuracy (fine-tune) / speed (crop).

④ offline training: similarity measure, step.

⑤ other: fine-grained, generalization, long-term.