Siamese Dynamic Mask Estimation Network for Fast Video Object Segmentation

Dexiang Hong [1], Guorong Li [1, 2], Kai Xu[1], Li Su[1,2], Qingming Huang [1, 2, 3]

1 School of Computer Science and Technology, University of Chinese Academy of Sciences, Beijing, China.
2 Key Laboratory of Big Data Mining and Knowledge Management, CAS, Beijing, China
3 Key Laboratory of Intell. Info. Process. (IIP), Inst. of Computi. Tech., CAS, China.

Motivation

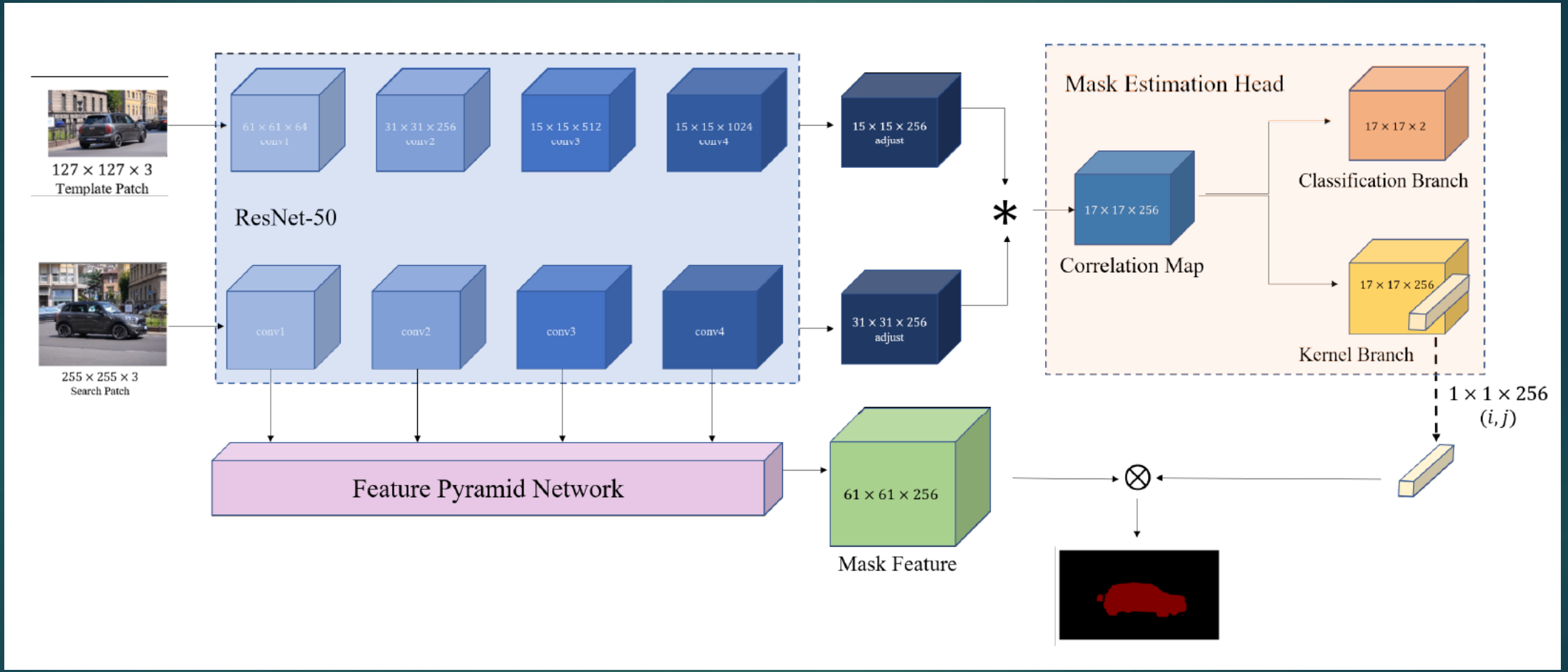
- Problem:
  - Most of previous VOS methods rely on pixel-level matching on the whole image while the targets only occupy a small region. It may cause two problems:
  - 1、Calculating on the entire image brings additional computation cost.
  - 2、The whole image may contain some distracting information resulting in many false-positive matching points.

- Proposed Solution:
  - We search the target only in the region around the target's previous location.
  - We decouple mask learning as the dynamic convolution kernel prediction and mask feature learning problem.

Contribution

- We design an effective video object segmentation framework that can perform end to end offline learning with deep convolution neural networks on well-annotated datasets.
- We reformulate the mask learning problem as a dynamic kernel prediction problem, which avoids pixel-level matching and can calculate in a small region.
- Experiments on DAVIS 2016/2017 datasets prove that the proposed method can run at 35 frames per second on NVIDIA RTX TITAN while preserving competitive accuracy

Method


127 × 127 × 3
Template Patch
255 × 255 × 3
Search Patch
61 × 61 × 64
conv1
31 × 31 × 256
conv2
15 × 15 × 512
conv3
15 × 15 × 1024
conv4
ResNet-50
conv1
conv2
conv3
conv4
15 × 15 × 256
adjust
31 × 31 × 256
adjust
Feature Pyramid Network
61 × 61 × 256
Mask Feature
Mask Estimation Head
17 × 17 × 256
Correlation Map
17 × 17 × 2
Classification Branch
17 × 17 × 256
Kernel Branch
1 × 1 × 256
(i, j)

Experiments

- We evaluate our approach on DAVIS-2016 and DAVIS-2017, and compare our method with the previous state-of-the-art approaches.

PERFORMANCE ON DAVIS 2016.

	$\mathcal{J}$ Mean↑	Recall↑	Decay ↓	$\mathcal{F}$ Mean ↑	Recall↑	Decay ↓	Speed ↑
PLM	70.2	86.3	11.2	62.5	73.2	14.7	6.7
VPN	75.0	82.3	12.4	65.5	69.0	14.4	1.6
CTN	73.5	87.4	15.6	69.3	79.6	12.9	0.03
PML	75.5	89.6	8.5	79.3	93.4	7.8	3.6
VideoMatch	81.0	-	-	-	-	-	3.2
FAVOS	82.4	96.5	4.5	79.5	89.4	5.5	0.8
FEELVOS	81.1	90.5	13.7	82.2	86.6	14.1	6.5
RGMP	81.5	91.7	10.9	82.0	90.8	10.1	8.0
SiamMask	71.7	86.8	3.0	67.8	79.8	2.1	55
Ours	74.3	89.2	4.2	70.2	82.4	3.4	35

PERFORMANCE ON DAVIS 2017.

	$\mathcal{J}$ Mean↑	Recall↑	Decay ↓	$\mathcal{F}$ Mean ↑	Recall↑	Decay ↓	Speed ↑
OSVOS	56.6	63.8	26.1	63.9	73.8	27.0	0.1
ONAVOS	61.6	67.4	27.9	69.1	75.4	26.6	0.1
FAVOS	54.6	61.1	14.1	61.8	72.3	18.0	0.8
OSMN	52.5	60.9	21.5	57.1	66.1	24.3	8.0
SiamMask	54.3	62.8	19.3	58.5	67.5	20.9	55
FEELVOS	69.1	79.1	17.5	74.0	83.8	20.1	6.5
ours	60.4	63.3	15.9	62.3	74.2	18.7	35

Conclusion

- In this work, we propose a Siamese dynamic mask estimation network for video object segmentation, which consists of a Siamese feature extraction network and a mask estimation head.
- We train the whole framework in an end-to-end manner. Our network can run at 35 frames per frame while preserving competitive accuracy.