Siamese Network based single object tracking

Presented by
Anmol Srivastava, Ashutosh Singh, Nimish Magre

Contents

Introduction

Motivation:

Tracking an object in video is a heavily researched domain.
Industrial utility : Survilance, robotics, autonomous vehicle etc.
Tracking single object in a video is a hard problem. Some of the challenges are listed here:-

Challenges:

- Occlusion.
- Object appearing and disappearing from the frame.
- Multiple Instances of the same object.
- Change of orientation.
- Change in scale and aspect ratio.

We will further discuss some of these in our demos.

Introduction

❑ **Methodology:**

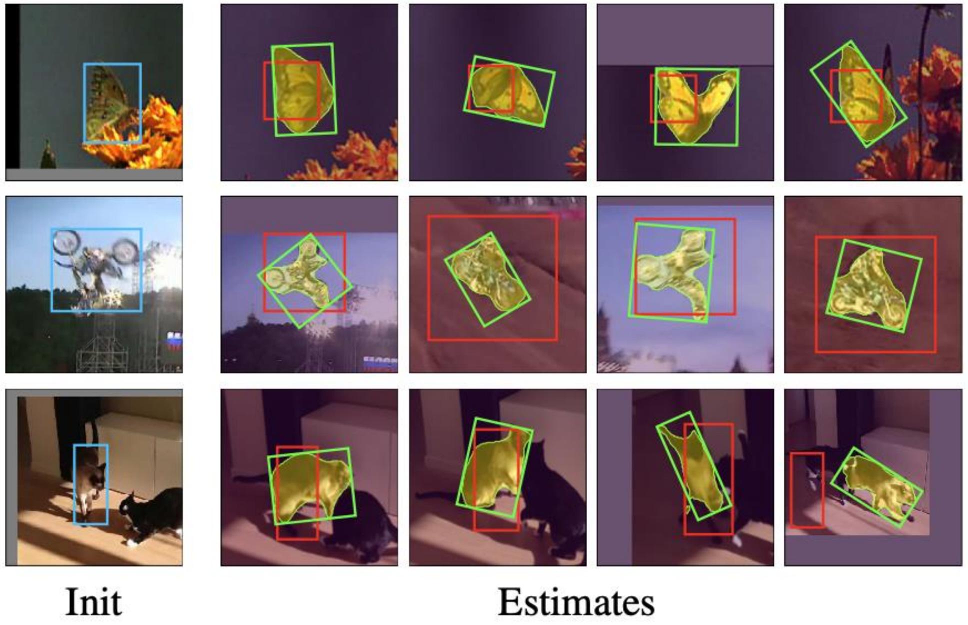
- Study state-of-the-art single object tracking solutions.
- Perform qualitative evaluation for the best performing trackers and identify the strengths and weaknesses
- Discuss potential solutions to the shortcomings
- Make changes to the current SOTA architecture based on the discussed solutions
- Present qualitative + quantitative evaluation

❑ **Aim:**

- Update template patch with each frame to improve tracking performance
- Obtain correlation between search, (t-1) and initial template patch
- Track object segmentation masks along with object orientation

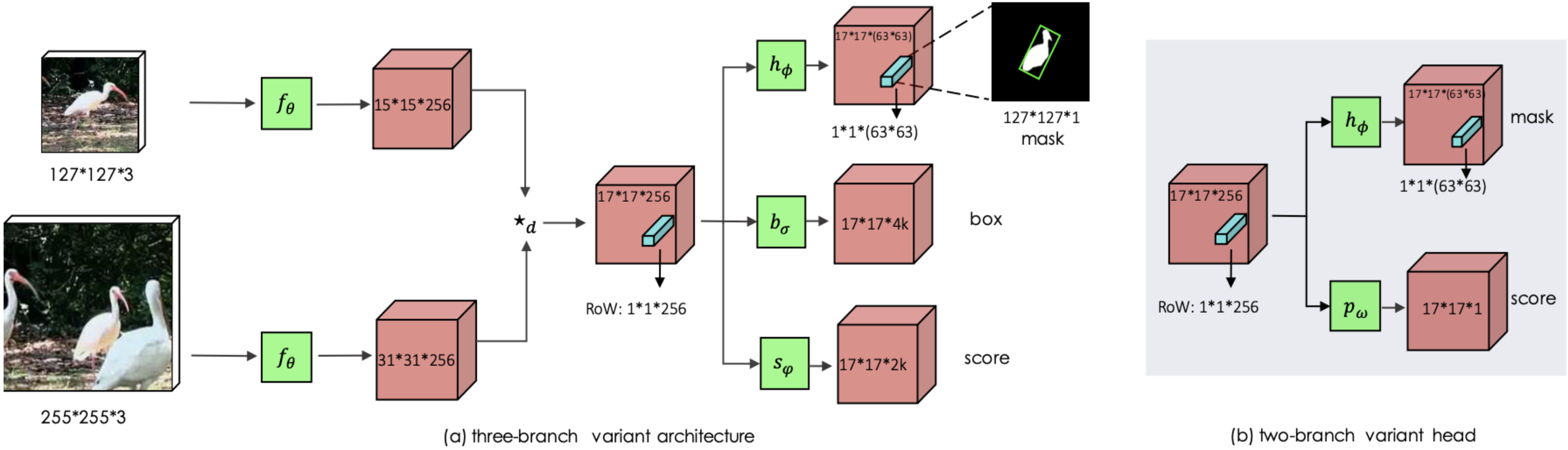
Fast Online Object Tracking and Segmentation (SiamMask)

- Simultaneously produce the per-frame target segmentation along with bounding box.
- **RoW:** $g_\theta(z, x) = f_\theta(z) \star f_\theta(x)$.
- Generates mask from the flatten representation of the object.
- **Loss:** $\mathcal{L}_{3B} = \lambda_1 \cdot \mathcal{L}_{mask} + \lambda_2 \cdot \mathcal{L}_{score} + \lambda_3 \cdot \mathcal{L}_{box}$
- Exemplar Patch: $\mathbf{127} \times \mathbf{127}$
- Search Patch: $\mathbf{255} \times \mathbf{255}$



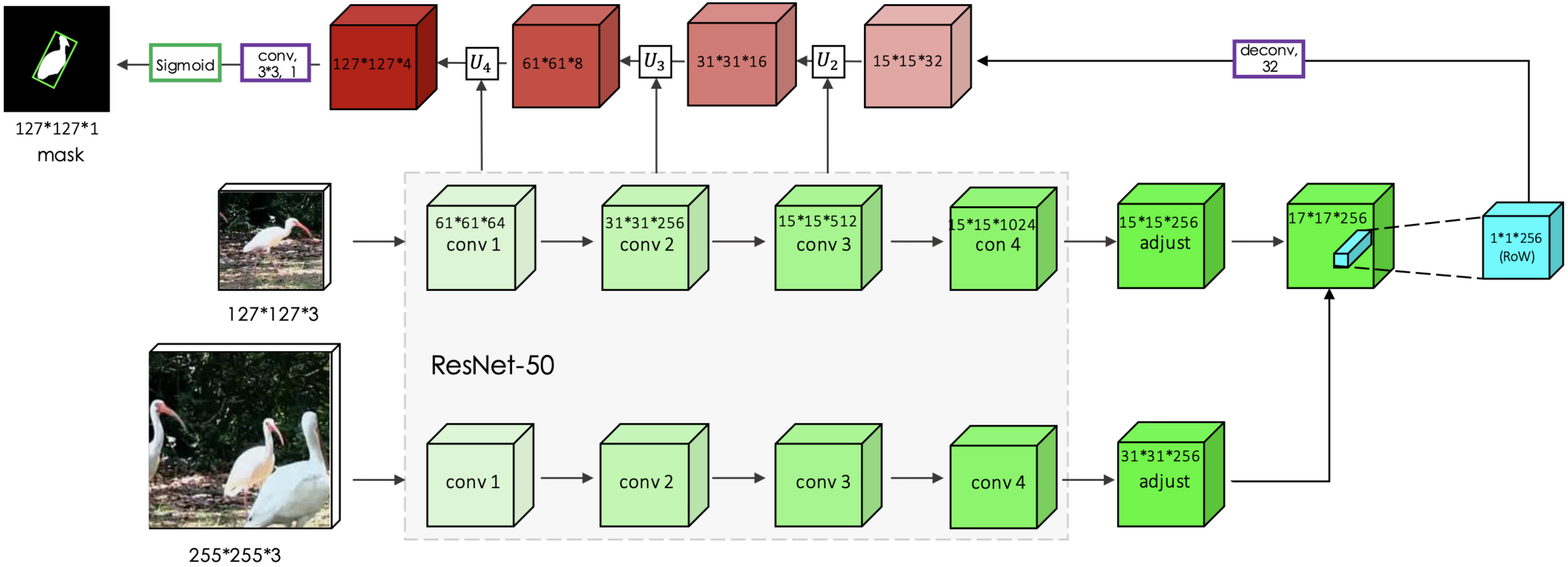
Wang, Q. et al. “Fast Online Object Tracking and Segmentation: A Unifying Approach.” *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2019): 1328-1338.

SiamMask Network Architecture



Wang, Q. et al. "Fast Online Object Tracking and Segmentation: A Unifying Approach." *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2019): 1328-1338.

SiamMask Network Architecture

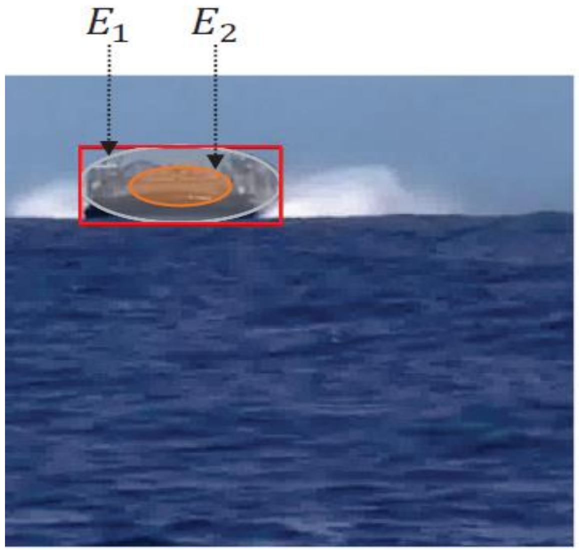


Wang, Q. et al. "Fast Online Object Tracking and Segmentation: A Unifying Approach." *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2019): 1328-1338.

Siamese Box Adaptive Network for Visual Tracking (SiamBAN)

Key Contributions:

- Simple Siamese network with end-to-end offline training capability
- Unified FCN to directly classify target and regress bounding box
  - ✓ making the model robust to varying target objects while running at 40 fps on the testing datasets
  - ✓ No prior anchor box design avoids hyperparameters associated with these boxes
- Use of ellipses for sample label assignment during the training phase

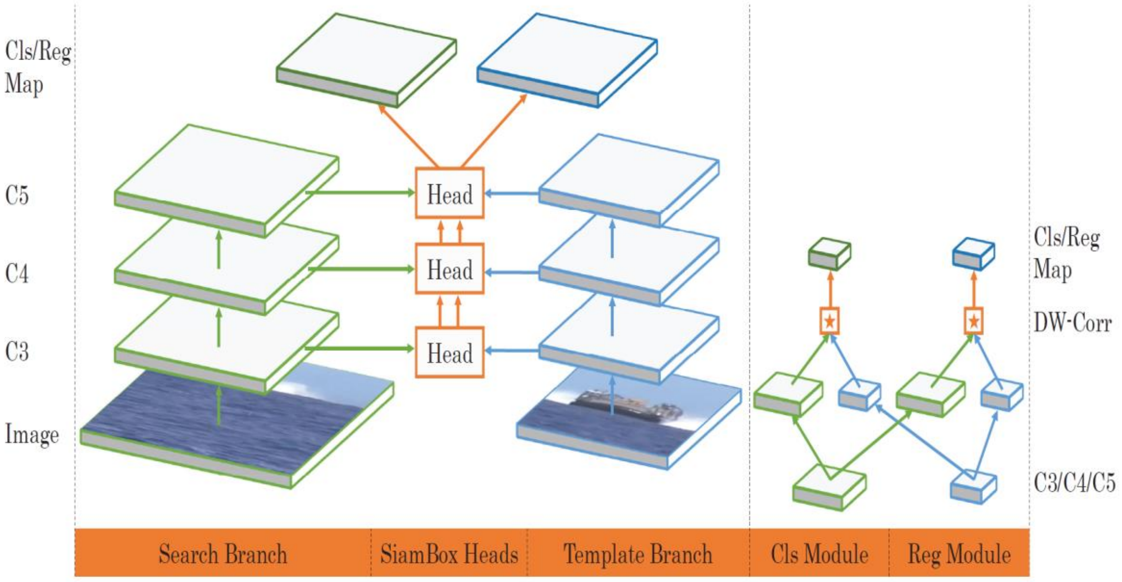


Ellipse Labels

Chen, Zedu et al. "Siamese Box Adaptive Network for Visual Tracking." *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2020): 6667-6676.

Siamese Box Adaptive Network for Visual Tracking (SiamBAN)

Architecture:



Chen, Zedu et al. "Siamese Box Adaptive Network for Visual Tracking." *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2020): 6667-6676.

Model Outputs:

- $P^{cls}_{w\times h\times 2} = [\varphi(x)]_{cls} * [\varphi(z)]_{cls}$
- $P^{reg}_{w\times h\times 4} = [\varphi(x)]_{reg} * [\varphi(z)]_{reg}$
- Regression map:

$d_l = p_i - g_{x_1},$	$p_{x_1} = p_i - d^{reg}_l,$
$d_t = p_j - g_{y_1},$	$p_{y_1} = p_j - d^{reg}_t,$
$d_r = g_{x_2} - p_i,$	$p_{x_2} = p_i + d^{reg}_r,$
$d_b = g_{y_2} - p_j$	$p_{y_2} = p_j + d^{reg}_b$

- Classification map:

$$P^{cls}_{w\times h\times 2} = \begin{matrix} Foreground\ classification\ score \\ Background\ classification\ score \end{matrix}$$

Siamese Box Adaptive Network for Visual Tracking (SiamBAN)

Proposal Selection[1]

- A Cosine window is added to suppress large displacement
- A Penalty is added to suppress large change in size and ratio:

$$penalty = e^{k*max(\frac{r}{r'},\frac{r'}{r})*max(\frac{s}{s'},\frac{s'}{s})}$$

- Proposals are ranked by multiplying classification scores by their temporal penalty
- Proposal box with the best score is selected and linear interpolation with the previous frame is applied to update its size

B. Li, J. Yan, W. Wu, Z. Zhu and X. Hu, "High Performance Visual Tracking with Siamese Region Proposal Network," 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018, pp. 8971-8980, doi: 10.1109/CVPR.2018.00935.

Qualitative Results Comparison

Multiple Instances

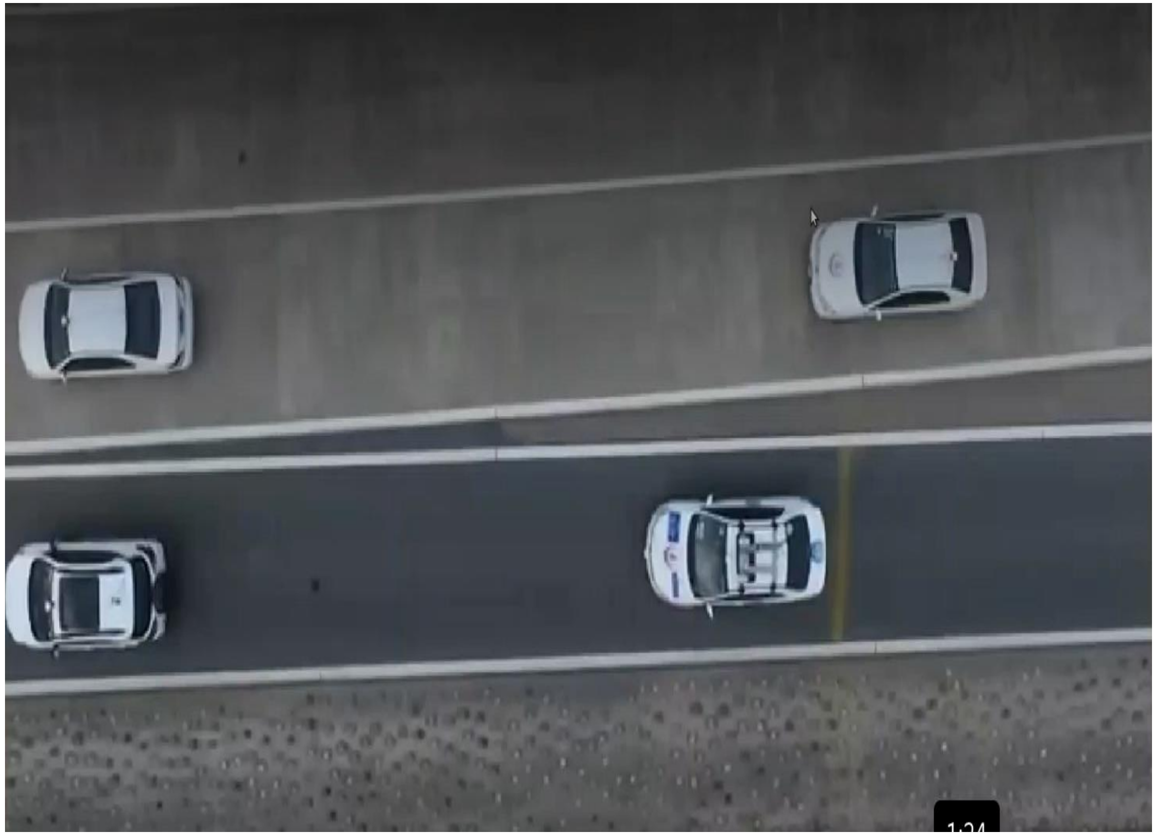
SiamBAN Performance

SiamMask Performance

Qualitative Results Comparison

Object moving in and out of frame

SiamBAN Performance



SiamMask Performance

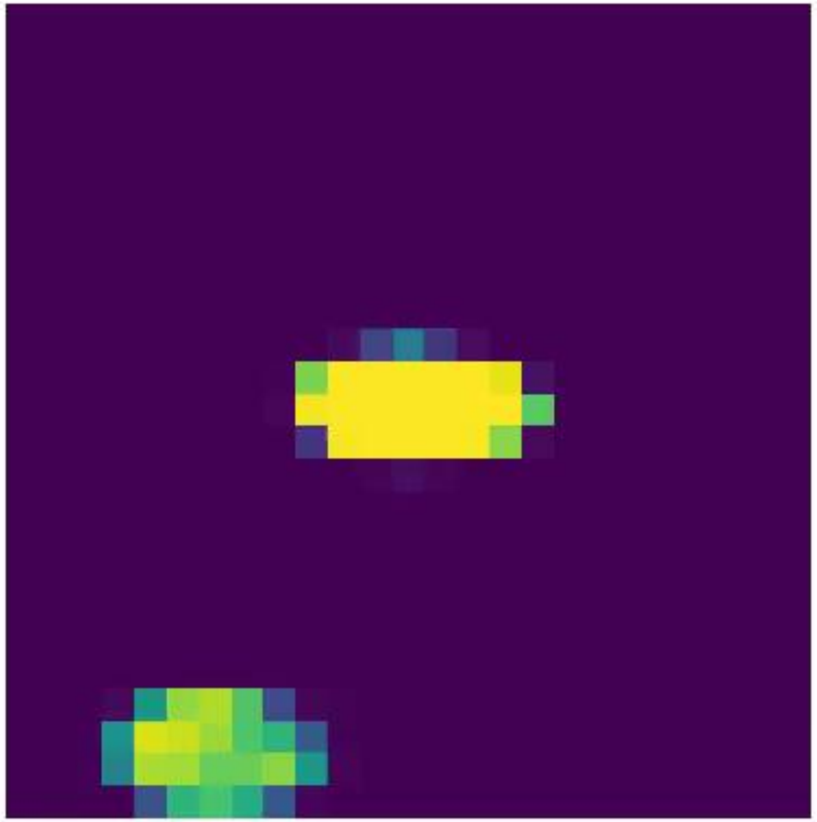
Qualitative Results Comparison

Change in scale and orientation of object

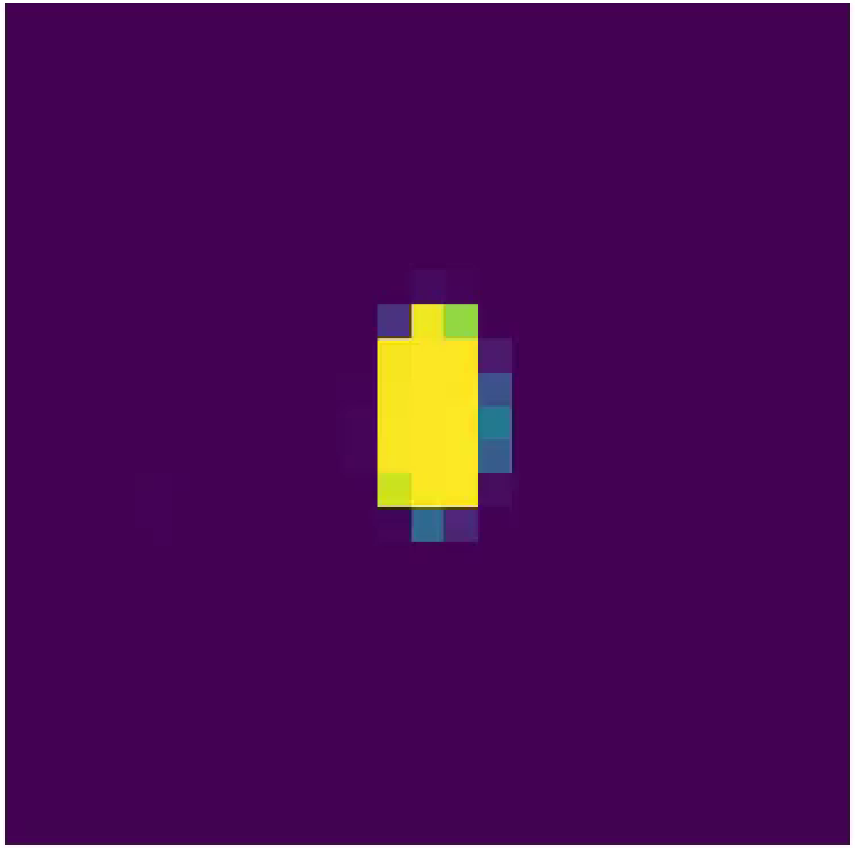
SiamBAN Performance

SiamMask Performance

Heatmap Generation



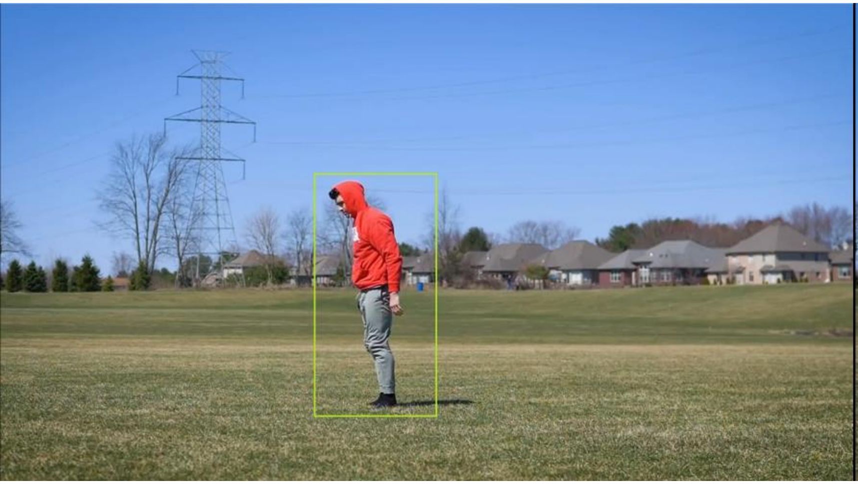
Heatmap Generation



(t-1) Template Update

Motivation:

To help model perform better when object deforms



Search frame | (t-1) template | initial frame template

(t-1) Template Update

Demo Results

(t-1) Template Update

Quantitative Evaluation

(Results based VOT2018 dataset)

Tracker Name	Accuracy	Robustness	Lost Number	EAO
model	0.586	0.229	49.0	0.378

Tracker Name	Accuracy	Robustness	Lost Number	EAO
model	0.470	0.375	80.0	0.253

Correlation based Template Update

Aim:

To study the error accumulation and detect the point of failure in the tracking result.
Using this information to improve upon the tracking by switching between the predicted template feature vector.

Method:

We track three correlation values across the video.

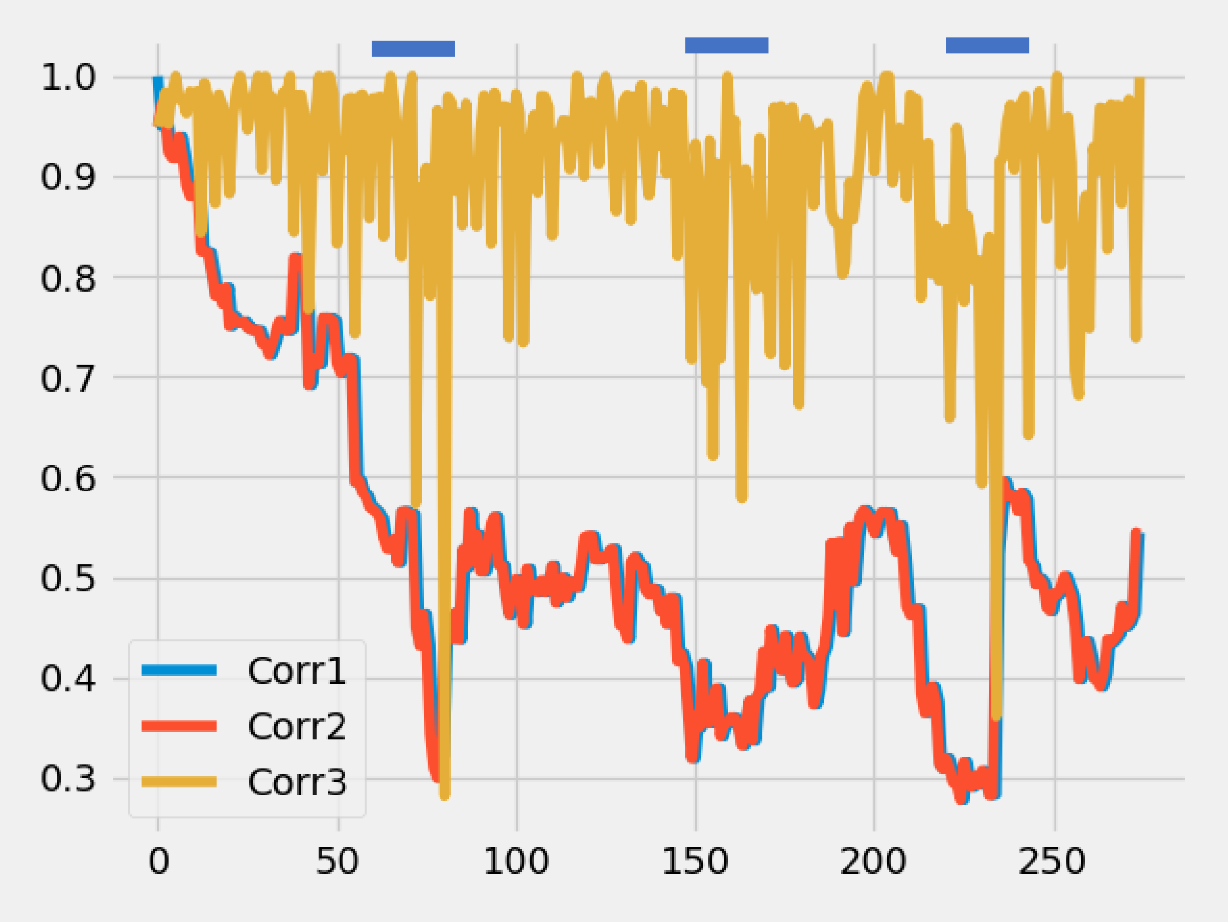
1. Correlation between template feature vector of initial frame and that of the frame at time (t-1)
2. Correlation between template feature vector of the frame at (t-1) and that of the frame at (t-search)
3. Correlation between template feature vector of the initial frame and that of the frame at (t-search)

To identify the different states based on these correlation time series we then perform event segmentation using HMM. From these states we were able to identify frames where occlusion or misclassification started to happen.

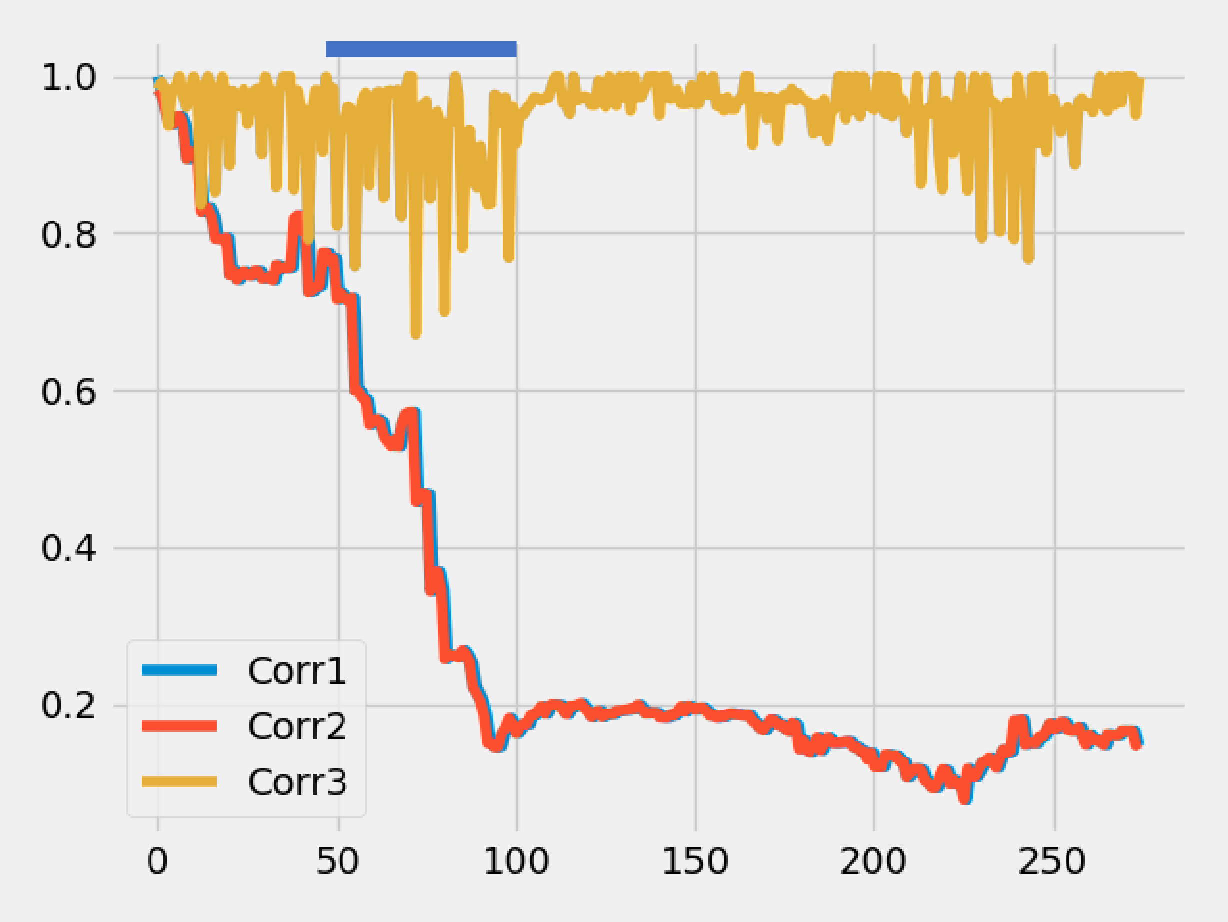
Correlation based Template Update

Results:

Initial frame template

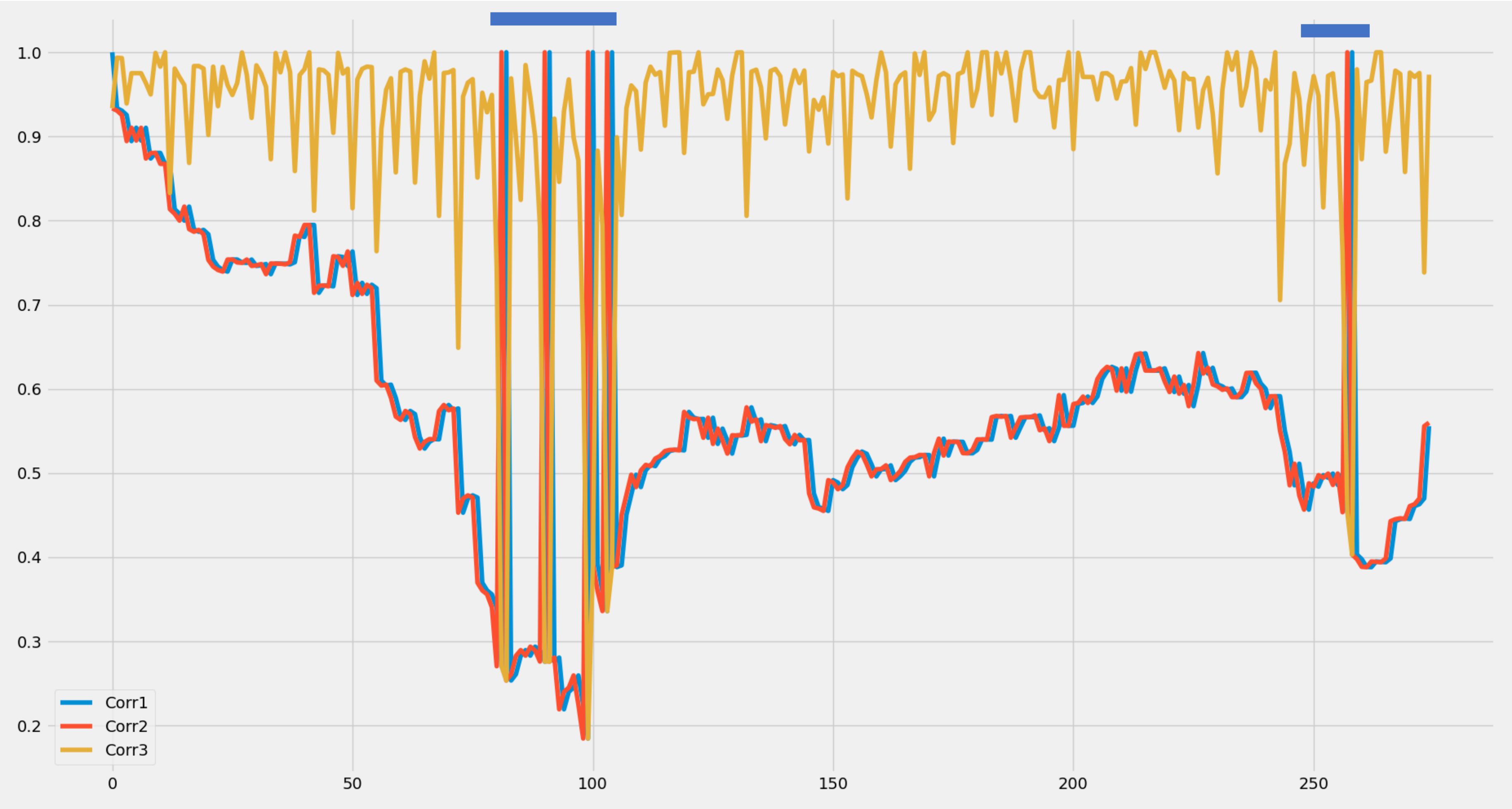


(t-1) used as template



Correlation based Template Update

Results:



Northeastern University

Correlation based Template Update

Qualitative Results:

Correlation based Template Update

Quantitative Results:

Tracker Name	Accuracy	Robustness	Lost Number	EAO
model	0.586	0.229	49.0	0.378

Initial frame template (based on our testing)

Tracker Name	Accuracy	Robustness	Lost Number	EAO
model	0.470	0.375	80.0	0.253

(t-1) used as template

Tracker Name	Accuracy	Robustness	Lost Number	EAO
model	0.504	0.304	65.0	0.296

Correlation based template update

(Results based VOT2018 dataset)

Future Work

Thank you

Questions?