Siamese Network Based Object Tracking

Ashutosh Singh, Anmol Srivastava, Nimish Magre
Northeastern University, Boston, MA
singh.ashu@northeastern.edu, srivastava.anm@northeastern.edu, magre.n@northeastern.edu

Abstract

In this article we discuss the problem statement of single object tracking in video. We discuss the shortcomings of the current state of the art algorithms particularly in the situation where multiple instances of the same class are present. We evaluate and investigate their performance in such scenarios by discussing particular examples where they fail. We extend the current state of art with a new template selection scheme, using hidden markov model, to improve upon their results, (qualitatively) particularly for the above discussed use cases. We further discuss the usability of our proposed model to improve upon the training methodology of these algorithm while making the process of detecting anomaly/occlusion automatic. Hence we provide a two fold solution for the problem statement of video object tracking: a)Improved video object tracking algorithm particularly for the use case where multiple instances of the target class are present, b)present an automatic occlusion/anomaly detection extension.

1. Introduction

There has been multifarious work done in the domain of arbitrary object tracking wherein the arbitrary object could be defined as an object whose mask is not explicitly defined during the initialization of frames. Some of the object tracking frameworks like [1], [14], [9] made use of tracking bounding boxes which do not provide accurate object contours. Moreover, majority of the existing methods require manual first frame initialization of the object to be tracked; thereby making them comparatively inefficient in tracking the arbitrary objects. However, there are few strategies like [9], [12], [8], [2], that link the detection regions of an object in order to form better temporal correlations between frames for the purpose of tracking the object trajectories.

Recent tracking frameworks are built upon Siamese Networks which have been utilized for object detection, tracking and object agnostic segmentation mask production[18] previously. This network has gained exponential popularity over the last few years because of its similarity measurements, end-to-end training capabilities and relatively simple implementation. It has paved the way for several architectural advancements in the object tracking field with networks such as SiameseFC[11] which made use of a Convolutional Siamese Network to compare an exemplar image z against a large search image x in order to obtain a dense response map. The resulting feature maps were then combined together using a cross correlation layer to produce a score map which surpassed the standard efficiency rate at that time. SiameseRPN[3] is the augmented version of SiameseFC and it incorporates a Regional Proposal Network[15] into its backbone architecture for more accurate target bounding box prediction during tracking. This requires a carefully curated anchor box to handle different scales and aspect ratios. Furthermore, SiamMask[18] adds a third branch to the SiameseRPN framework in order to produce the object mask in conjunction with its bounding box. Recently, a flexible and simple yet effective network architecture SiamBAN[4] has been introduced in this family. It directly classifies and regresses the object bounding box in a unified Fully Connected Network (FCN), thereby extinguishing the need for complex heuristics.

2. Related Work

2.1. SiamBAN

Inspired by anchor-free object detection networks [7],[13],[16], Zedu Chen et al [5] propose a simple Siamese network (Siamese Box Adaptive Network-SiamBAN) which performs the single object-tracking task through a foreground-background classification and a bounding box parameter regression module. Specifically, the proposed network improves on the previous state-of-the-art Siamese visual tracking networks by accurately estimating the scale and aspect ratio of a target without the need for predefined anchor boxes or multiscale searching schemes.

As shown in **Figure 3**, the architecture of the proposed network consists of a Siamese network backbone for image feature extraction along with multiple box-adaptive heads

Figure 1. **Row 1**, **Row 4**, and **Row 6** show the demo results of SiamBAN in multiple instances, occlusion environments, out-of-view, scale-change and orientation change use-cases. Similarly, **Row 2**, **Row 3**, and **Row 5** show the demo results of SiamMask network for the same use-cases.

that are responsible for obtaining a foreground-background score classification map and a 4-channel regression map representing multiple bounding-box parameters.

The Siamese network backbone computes the convolutional feature maps of the template and the search patch. To do so, the authors adopt a pretrained ResNet-50 [6] network and remove the down sampling operations from the last 2 convolution blocks to obtain detailed spatial information from the images. Inspired by a previous anchor-based tracker [10] and multi-grid methods [17], the authors adopt atrous convolution with different atrous rates for each of the three convolution blocks to improve the receptive field. Both the branches share parameters and a 1×1 convolution is added at the end of each block to reduce the output feature channels to 256. Also, the authors make use of centre 7×7 region features of the template branch [17].

Each box-adaptive head consists of a classification and a regression module and both these modules receive features from the template and the search branches. $\mathbf{Eq^n(1)}$ and $\mathbf{Eq^n(2)}$ shown below represents how the 2-channel classification and the 4-channel regression maps are calculated.

$$P^{cls}_{w\times h\times 2} = [\varphi(x)]_{cls} \circledast [\varphi(z)]_{cls} \quad (1)$$

$$P^{reg}_{w\times h\times 4} = [\varphi(x)]_{reg} \circledast [\varphi(z)]_{reg} \quad (2)$$

Each module combines the template patch features $(\phi(z))$ and the search patch features $(\phi(x))$ using depth-wise cross correlation. Each location on the classification or the regression map (i,j) can be mapped to a pixel position on the search patch (p_i, p_j) using $\mathbf{eq^n(3)}$.

$$(i,j) \longleftrightarrow [w_{im} - (\left[\frac{w}{2}\right] - i) \times s, h_{im} - (\left[\frac{h}{2}\right] - j) \times s] = (p_i, p_j) \quad (3)$$

where w_{im} and h_{im} represent the width and height of the input search patch and s represents the total stride of the network.

Features extracted from the earlier layers of a feature extraction network provide low level fine-grained information such as color or shape and these characteristics are essential for localization whereas the deeper layers encode important semantic information that can make the model robust to conditions such a motion blur or object appearance change [8]. Based on this information, the authors make use of multiple box adaptive heads to access effectively access information from multiple layers of. The classification and regression maps obtained from each of the box-adaptive heads are then adaptively fused using hyper-parameters learned during training.

As shown in **Figure 2**, each search patch includes a ground-truth bounding box and ellipses E1 and E2 are calculated for each search patch using $\mathbf{eq^n(4)}$ and $\mathbf{eq^n(5)}$ respectively.

$$\frac{(p_i - g_{xc})^2}{(\frac{g_w}{4})^2} + \frac{(p_j - g_{yc})^2}{(\frac{g_h}{4})^2} = 1 \quad (4)$$

$$\frac{(p_i - g_{xc})^2}{(\frac{g_w}{2})^2} + \frac{(p_j - g_{yc})^2}{(\frac{g_h}{2})^2} = 1 \quad (5)$$

where g_{x_c}, g_{y_c}, g_w and g_h represent the centre location, the width and the height of the ground-truth bounding box.

To further improve the weights obtained in the regression module, the authors only perform bounding-box parameter regression for pixels that lie within ellipse E2. The regression targets are set to be the offset values of the pixel position p_i, p_j w.r.t the 4 sides of the ground-truth bounding box.

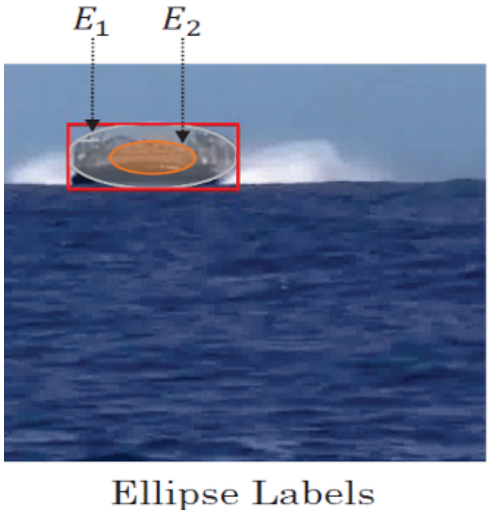

Figure 2. Label assignment-based regression where bounding box parameters are only estimated for pixel positions within ellipse E2

The multi-task loss function $\mathbf{eq^n(6)}$ uses a simple cross-entropy loss for the classification map and an Intersection-over-Union (IoU) loss for the regression map.

$$\mathcal{L} = \lambda_1.\mathcal{L}_{cls} + \lambda_2.\mathcal{L}_{reg} \quad (6)$$

During inference, the template patch is cropped from the first frame and is passed through the feature extraction network. These features are then cached and used along with the features of each search patch to obtain the 2 maps. The features from the search patch are extracted based on the object position predicted in the previous frame.

Furthermore, the model adopts a proposal selection strategy from the SiamRPN [9] paper wherein:

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

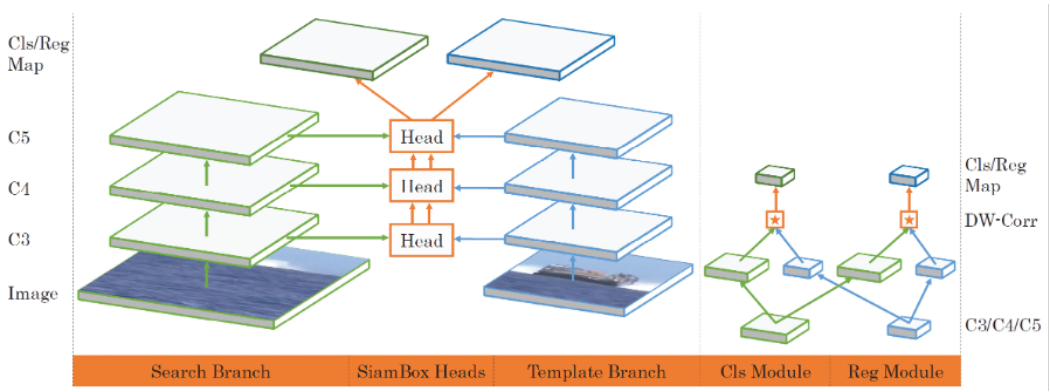

Figure 3. SiamBAN architecture: Siamese network backbone for feature extraction and multiple box-adaptive heads that include a classification module for foreground-background score and a regression module to regress bounding box parameters

The authors compare their model with previous SOTA tracking networks on 6 tracking benchmarks and use the Expected Average Overlap metric that estimates the average per-frame overlap of the predicted and ground-truth bounding boxes. On each of the benchmarks, the proposed model shows competitive results while running a 40 fps.

The importance of features from each of the 3 convolutional blocks is evaluated by using their individual features to obtain the classification and regression maps and estimating their performance on the OTB100 dataset [19] and the aggregation of features from all 4 layers shows the best results. The authors also experiment with using a circle and a rectangle for label assignment but upon evaluation, they find that using ellipses provides the best results.

To further analyze and identify the sources of issues with the SiamBAN model, we performed qualitative evaluation of the model on videos where the object faces constraints such as occlusion, change in orientation and disappearance-appearance process in subsequent frames.

2.2. SiamMask

The 3 branch Siamese convolution network (SiamMask) builds upon earlier work with a single branch Siamese network (SiamFC) and a 2 branch Siamese network (SiamRPN). SiamFC compares an exemplar image $\mathbf{z}$ against a larger search image $\mathbf{x}$ to obtain a dense response map,

$$g_\theta(\mathbf{z}, \mathbf{x}) = f_\theta(\mathbf{z}) * f_\theta(\mathbf{x})$$

$$\mathcal{L}_{3B} = \lambda_1.\mathcal{L}_{mask} + \lambda_2.\mathcal{L}_{score} + \lambda_3.\mathcal{L}_{box} \qquad (7)$$

The network produces 'n' such maps corresponding to the n^{th} candidate window in $\mathbf{x}$. Consequently, the highest response map value is supposed to correspond to the target location in search area $\mathbf{x}$. SiamMask replaces the cross-correlation used in SiamFC with depth-wise cross-correlation and uses logistic loss for training $\mathbf{eq^n(7)}$. SiamRPN (Regional Proposal Network) improves on SiamFC performance by estimating a bounding box along with the target location using a 2 branch Siamese network. It uses the L1 loss for the bounding box branch and the cross-entropy loss for the classification score.

Out-of-view object environment is the challenging scenario where the target objects disappear for some time duration and reappears after a certain sequence of frames. It could be inferred from the *row 3* of **Figure 1** that siamMask performs comparatively well in segmenting and tracking target vehicle after it disappears for a couple of frames. On the contrary, the adaptive siamese network is abortive in handling out-of-scope objects as illustrated in *row 4*. Therefore, the results suggest that simultaneous segmentation of the object could bolster the robustness of the tracking algorithm.

Furthermore, it could be observed from the last two rows that during scale change environments, the siamMask architecture was able to track the biker for the first initial frames with significant accuracy. Yet, exhibits the instant decrease in tracking after immediate scale and orientation change of an object. Thereby, resulting in deviated bounding boxes. However, on the other hand, siamBAN performed exceptionally well throughout all the frames and adapts well to the continuous scale change and aspect ratio of the target.

Consequentially, all these demo experiments highlight the inefficacy of siamese-based visual trackers in an unstructured video where objects undergo various challenging situations which further exacerbates their robustness.

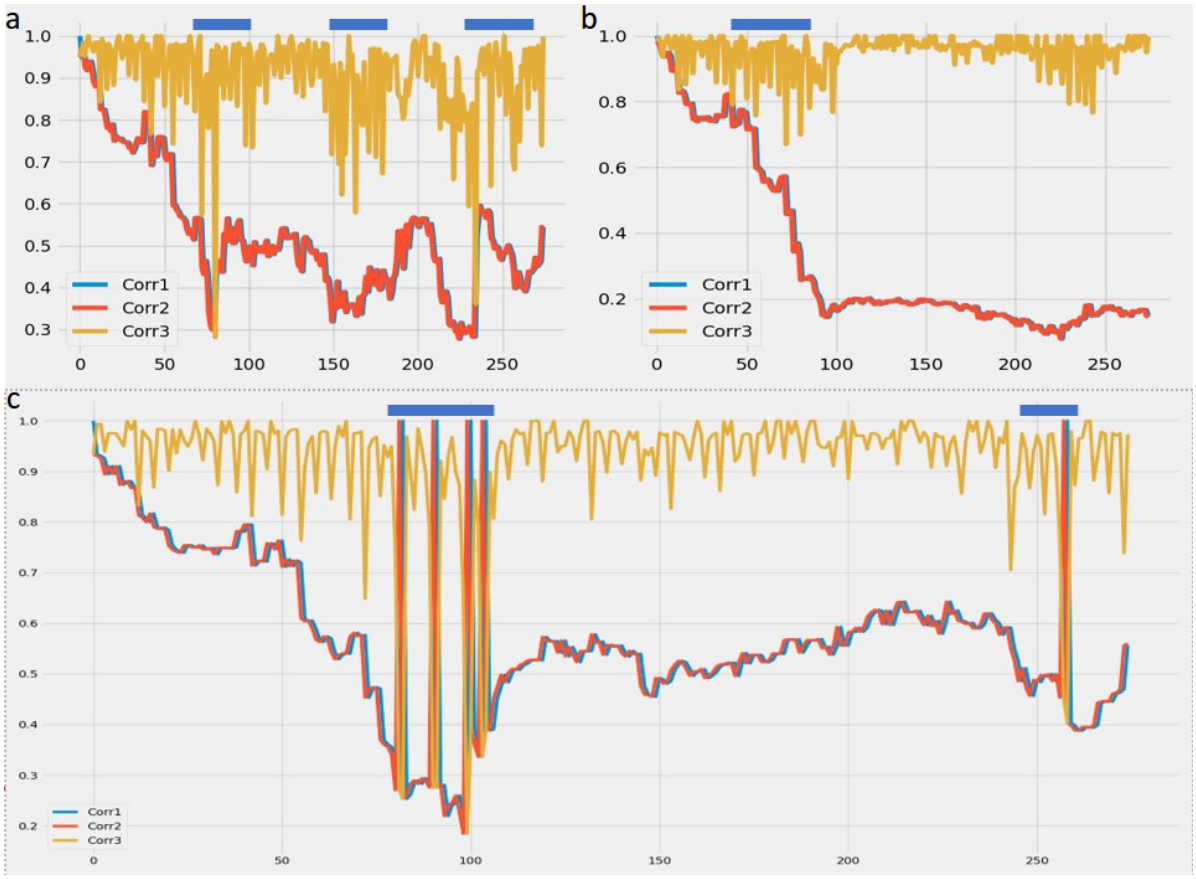

Figure 4. Correlation results for a) original SiamBAN model, b) SiamBAN model output using updated template (using template patch from frame at $t-1$ to get feature vector of the target object) and c) Correlation signals for modified siamBAN framework with guided template feature extraction using state switching based on HMM. Here corr1, corr2 represents the correlation measurements between feature map acquired from template at initial frame and the template patch from frame at t and $t-1$ respectively. Corr3 represents the correlation between the feature map of template at t and that at $t-1$. The blue bar represents the state identified at occlusion.)

3. Method

3.1. Updating Template Patch

In siamban the feature map from the template patch is only generated at the initial frame and then these features are used for the subsequent frames. Through experimentation we found that this proves to be ineffective particularly in the case where there are multiple instances of the target class eg. a video with multiple people (target class human), or in cases where the target object deforms or get occluded in subsequent frames. To particularly improve upon this drawback of siamban we proposed to update the feature vector of the template patch using the frame at time $t-1$ for tracking in frame at time t.

3.2. Correlation Analysis

Using the template updating framework has an inherent drawback of error accumulation. Especially in the case where the object gets partially or completely occluded, we find that the learned features map from the frame at time $t-1$ starts to instead represent the object occluding the target object. Though we did see that updating the template results in better tracking for the case when multiple instances of the same class are present. To study exactly when the template updating fails and tracking based on feature map from initial frame works better we measure the pairwise Pearson correlation coefficient between feature map acquired from initial template, that acquired from the template at time $t-1$ and the new predicted template patch.

3.3. Automatic Occlusion/Anomaly Prediction

We found that the correlation signals had multiple states/behaviours. One such state represented instances where occlusion happens. To identify these states we develop a generative approach based on hidden markov model (HMM) that uses these correlation signals as input and gives out the state sequences. To keep our approach simple we assume the emission probability distribution to be gaussian.

a

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

b

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

c

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

Figure 5. Quantitative results for modified siamBAN models. a) represents the results for orginal SiamBAN model, b) represents the siamBAN model while updating the template at each frame. c) represents the results for modified siamBAN framework with guided template feature extraction using state switching based on HMM.

We use kfold cross validation to find the most optimum number of states i.e. the number of states that gives us the highest log-likelihood. We use Dirichlet distribution as priors over the transition probabilities and initial state probabilities. We train the HMM model and perform parameter estimation using the Baum-welch algorithm. We train our model based on the results of the modified siamBAN on the testing dataset. We generate 150x3 correlation signals where each test case gives us 3 signals as discussed in previous subsection.

3.4. Automatic Switching

Based on the results of the previous sections we identify the state the represents the initiation of the failure our modified siamBAN model i.e. instances that represent occlusion in the video. We couple this information with set thresholds to switch between using the feature map from initial frame and that acquired from the template in frame at time $t-1$. We see that this approach gives better results when compared to just using the updated template patch.

4. Results

In this section we will discuss the quantitative and qualitative results of our modified model and draw comparison with the original model. In figure 4 we see that our anomaly/occlusion detection does pretty good job at identifying these instances from patterns in the correlation signals. Here we have only illustrated the state that represents occlusion. This state is particularly characterised by huge drop in the value of $corr3$ and $corr1$. The end of this state is characterised by a low $corr3$ value and an increment in $corr1$ value. It is pretty intuitive that when an object is just getting occluded the correlation value between the feature map of previous and current frame would drastically drop as the later will contain features representing the object occluding the target object. When the target object reemerges the correlation between the feature map of current and initial template would increase while the feature map from the previous frame is misguided due to error accumulation whilst object in occlusion. Hence the switching framework would guide the tracking algorithm to identify the occlusion and choose the best template patch.

In figure 5 we present the quantitative result of our model. Figure 5.c) represents the evaluation based on the modified model with guided template selection. It is clear from figure 5.b and figure 5.c that using a guided template selection scheme works better when compared with just using the template at time $t-1$. The lower performance when compared to original siamBAN approach could be attributed to the dataset on which testing and evaluation are performed. These datasets do not contain cases where there are multiple instances of the same object are present or rigorous occlusion is happening for example when two objects of same instance cross path. Furthermore the accuracy of our proposed method could be further improved by training it to discriminate the false tracking. Figure 6 represents qualitative comparison for a use case outside of the testing dataset which represent two instances of the same class crossing their path. On comparing our result one can clearly see that the modified framework works much better that original model and the intermediate model without the switching capabilities. On observing carefully one can

Figure 6. Qualitative results based on the use case of multiple instances of the same class crossing each other. The first row shows the results from siamBAN (original model), the second row represents the result of the modified siamBAN model but without switching capabilities and the thrid row represents the result for our modified siamBAN approach with automatic switching between templates patch of the targeted object.

see that the original tracking algorithm gets confused when there are multiple instances of the target object present in the video. It also performs poorly when the the target object gets occluded. When we just use the frame at time $t-1$ to generate the template patch we see that after the target object gets occluded the the tracker is unable to correctly track the object which is due to the accumulation of error as discussed earlier. In the last row, which represents the results obtained with the modified siamBAN model with switching capabilities, we see that it performs, qualitatively speaking, much better than the previous two iterations of the model.

5. Discussion

In this article we presented a tracking algorithm modified to automatically identify occlusion and used this information to guide its template patch selection. Our proposed method for automatic occlusion/anomaly detection could be further used to develop a training dataset aimed at better discriminating the false tracking situations. Hence we can further improve our model by forcing it to learn from its mistakes. Other future direction for our model could be to learn and enrich the latent representation of the tracked object at each frame. This could be done using an auto-encoder framework. This would make the model robust and able to track deformable objects. Having a latent representation of the object could further be used to simulate synthetic pose of the object. Siammask algorithm is good at segmentation and orientation tracking. This could be further combined with the modified siamban architecture to develop a tracking algorithm that is able to handle : scale and aspect ration variation, multiple instances of the target class, orientation change and deformation in the shape of the tracked object.

References

[1] Stefan Roth Anton Andriyenko, Konrad Schindler. Discrete-continuous optimization for multi-target tracking. *Computer Vision and Pattern Recognition (CVPR)*, pages 1926–1933, 2012. 1

[2] Kap Luk Chan Bing Wang, Gang Wang and Li Wan. Tracklet association with online target-specific metric learning. *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, page 1234–1241, 2014. 1

[3] Wei Wu Zheng Zhu Bo Li, Junjie Yan and Xiaolin Hu. High performance visual tracking with siamese region proposal network. *In IEEE Conference on Computer Vision and Pattern Recognition*, page 8971–8980, 2018. 1

[4] Zedu Chen, Bineng Zhong, Guorong Li, Shengping Zhang, and Rongrong Ji. Siamese box adaptive network for visual tracking, 2020. 1

[5] Zedu Chen, Bineng Zhong, Guorong Li, Shengping Zhang, and Rongrong Ji. Siamese box adaptive network for visual tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6668–6677, 2020. 1

[6] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 2

[7] Lichao Huang, Yi Yang, Yafeng Deng, and Yinan Yu. Densebox: Unifying landmark localization with end to end object detection. *arXiv preprint arXiv:1509.04874*, 2015. 1

[8] Cheng-Hao Kuo and Ram Nevatia. How does person identity recognition help multi-person tracking? *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, page 1217–1224, 2011. 1

[9] Konrad Schindler Daniel Cremers Ian Reid Stefan Rot Laura Leal-Taixe , Anton Milan. Tracking the trackers: An analysis of the state of the art in multiple object tracking. *arXiv preprint arXiv:1704.02781*, 2017. 1

[10] Bo Li, Wei Wu, Qiang Wang, Fangyi Zhang, Junliang Xing, and Junjie Yan. Siamrpn++: Evolution of siamese visual tracking with very deep networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4282–4291, 2019. 2

[11] Joao F Henriques Andrea Vedaldi Luca Bertinetto, Jack Valmadre and Philip HS Torr. Fully-convolutional siamese networks for object tracking. *In European Conference on Computer Vision*, page 850–865, 2016. 1

[12] Yu Zhongjie Bjoern Andres Thomas Brox Margret Keuper, Siyu Tang and Bernt Schiele. A multi-cut formulation for joint segmentation and tracking of multiple objects. *arXiv*. 1

[13] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 779–788, 2016. 1

[14] Kuk-Jin Yoon Seung-Hwan Bae. Robust online multi-object tracking based on tracklet confidence and online discriminative appearance learning. *IEEE Computer Vision and Pattern Recognition (CVPR)*, pages 1218–1225, 2014. 1

[15] Ross Girshick Shaoqing Ren, Kaiming He and Jian Sun. Faster r-cnn: Towards real-time object detection with re-gion proposal networks. *In Advances in neural information processing systems*, page 91–99, 2015. 1

[16] Zhi Tian, Chunhua Shen, Hao Chen, and Tong He. Fcos: Fully convolutional one-stage object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9627–9636, 2019. 1

[17] Panqu Wang, Pengfei Chen, Ye Yuan, Ding Liu, Zehua Huang, Xiaodi Hou, and Garrison Cottrell. Understanding convolution for semantic segmentation. In *2018 IEEE winter conference on applications of computer vision (WACV)*, pages 1451–1460. IEEE, 2018. 2

[18] Qiang Wang, Li Zhang, Luca Bertinetto, Weiming Hu, and Philip H. S. Torr. Fast online object tracking and segmentation: A unifying approach. *CoRR*, abs/1812.05050, 2018. 1

[19] Yi Wu, Jongwoo Lim, and Ming-Hsuan Yang. Online object tracking: A benchmark. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2411–2418, 2013. 4