Unsupervised Learning of Accurate Siamese Tracking Supplementary Material

Qiuhong Shen[1,2], Lei Qiao[2], Jinyang Guo [3], Peixia Li [3], Xin Li [4],
Bo Li [2], Weitao Feng [3], Weihao Gan [2,5], Wei Wu [2,5], Wanli Ouyang [3,5]
[1] Harbin Institute of Technology (Shenzhen) [2] SenseTime Research
[3] The University of Sydney [4] Peng Cheng Laboratory [5] Shanghai AI Laboratory
{shenqiuhong0905,xinlihitsz}@gmail.com, {qiaolei,libo,ganweihao,wuwei}@sensetime.com,
{jinyang.guo,peixia.li,weitao.feng,wanli.ouyang}@sydney.edu.au

1. Discussion and limitations

Despite proposed framework is effective to learn a better tracker from temporal self-supervision, the pipeline still relies on pseudo labels in initial frames generated by unsupervised optical flow model. As the relationship between a better initialization method and better tracker is a chicken-egg conundrum in this formulation, it's still a remaining problem about how to chain the initialization methods and unsupervised tracking into an end-to-end trainable pipeline.

2. Quantitative analysis of misalignments

Here we conduct experiments to give a quantitative analysis on the misalignments in conventional cycle training of unsupervised visual tracking. In our manuscript, we claim that the misalignment in cycle training severely hinders the performance of unsupervised visual tracking. Detailly, the source of this misalignment is the mismatch between the internal template and search region feature. For delving deep into this insight, we conduct a quantitative analysis of the impact of this misalignment. We choose the classical Siamese tracker, SiamRPN++ [3], as the baseline. For simplification, the mismatch in intermediate frames, which may be produced by forward tracking errors or initialization bias, is simulated as noises added to ground-truth for cropping the template patches in SiamRPN++ [3]. Specifically, we add noises to template patches with the following operation. Let us denote the ground-truth boxes in template frames as (cx, cy, w, h), where (cx, cy) in the center coordinates of bounding boxes, w and h are the width and height of the boxes respectively. The jittered template bounding boxes can be denoted as $(cx{+}\sigma_1 w, cy{+}\sigma_2 h, (1{+}\sigma_3)w, (1{+}\sigma_4)h)$, where σ_k denotes random number between -0.5 and 0.5 generated from uniform distribution.

Table 1. Quantitative analysis on VOT2018 benchmark

template boxes	Acc↑	Rob↓	EAO↑
ground-truth	0.600	0.234	0.414
jittered gt	0.565	0.355	0.298

In training phase, the model trained with noisy template boxes is hard to convergent on the regression branch. We evaluate the tracking performance on VOT2018 [2] benchmark dataset, as shown in Table 1, when template boxes are jittered, the performance of the tracker will drop with a substantial gap in the EAO metric.

3. More discussion about proposed component

3.1. Threshold in region mask

Our proposed region mask is a customized operation for unsupervised visual tracking in cycle training, which penalizes tracking errors on intermediate frames by making the coordinates differentiable. As claimed in our manuscripts, when compared to conventional feature selection operation like PrPool [1], this operation is efficient to select features from last search region feature based on the output of region proposal network (RPN), denoted as P_{cls} and P_{reg}. Specifically, we set boxes number as 3125 ($25 \times 25 \times 5$), the same as the total number of predicted boxes of RPN. And the positive threshold (denoted as TH in the manuscript) is set as 0 for passing all predicted boxes. Besides, we also visualize the regional mask propagation in training samples with different positive thresholds. The regional masks on search region images with four different threshold values are shown in Fig 2. When the value of this threshold increases, the region mask tend to filter out more predicted boxes with low confidence scores, which results in less information propagating between frames. Based on this observation, we always set $TH = 0$ in training phase, for propagating more information between frames. In addition, when proposal boxes with top confidence scores are wrong, region mask with a lower positive threshold is more likely to select the

Table 2. Ablation study of the threshold of region mask

Threshold	ACC ↑	Rob ↓	EAO ↑
0.0	0.560	0.272	0.346
0.5	0.550	0.323	0.327
0.9	0.559	0.342	0.303

Table 3. Ablation study of the CPT module on VOT2018

Settings	ACC ↑	Rob ↓	EAO ↑
LT + ST	0.560	0.272	0.346
LT	0.562	0.318	0.330
ST	0.557	0.328	0.324

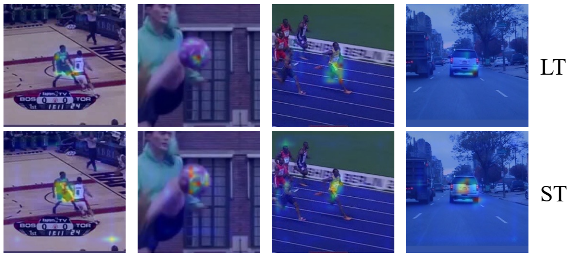


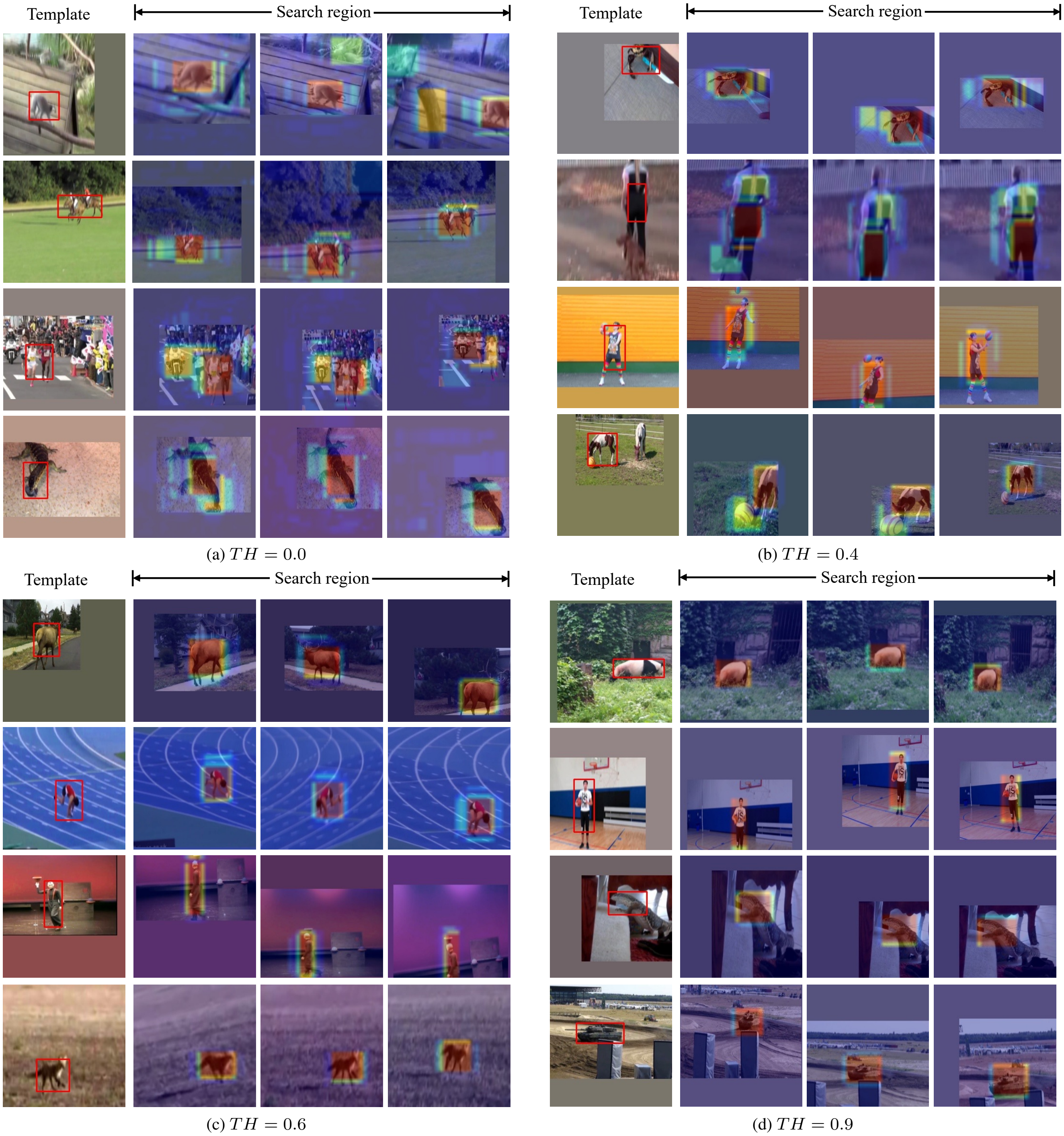
Figure 1. Visualization of attention maps for LT and ST queries.

correct proposal box.

For better understanding, we give a quantitative analysis of this threshold TH in training phase. As ablation studies with other thresholds shown in Table 2, the EAO scores drop significantly on the EAO scores as threshold increases. Every region map is multiplied with its confidence score when generating the region mask, thus the samples with low confidence essentially have smaller gradient. However, considering all boxes (including a large amount of non-target regions) essentially accumulates abundant training samples. As another aspect, such noisy region masks in training enforce the tracker to learn better discriminating abilities, i.e., predict higher confidence scores on the foreground (target) area and lower scores on the background area. While in online tracking phase, we cache search regions features of high confidence and corresponding regional mask for updating in a memory queue. When retrieving the memory kernel with the CPT module, we set a higher threshold to filter similar distractors features.

3.2. Long/short term in CPT module

For training with cycle consistency, it is required to track videos frames as a cycle. Previous works generated template kernels by RoI-Pooling on the top-1 proposal in search frames. If this top-1 proposal is wrong, especially in initial stage, then the generated template kernel becomes too noisy for the tracker to track back. With proposed CPT module, multiple possible matched regions can be used for generating reliable template features between frames. Here we visualize attention maps on search region for long and short term queries in Fig 1, long-term (LT) queries have higher responses on invariant areas of the target, while short-term (ST) queries have higher responses on variant target areas as they are intended for retrieving the most recent target features. Besides, we present ablation study on the long/short term query of CPT module that shown in Table 3. It shows that the combination of long-term and short-term queries performs better on EAO scores than using single term.



(a) $TH = 0.0$

(b) $TH = 0.4$

(c) $TH = 0.6$

(d) $TH = 0.9$

Figure 2. Visualization of regional mask on training samples with different positive threshold value

References

[1] Borui Jiang, Ruixuan Luo, Jiayuan Mao, Tete Xiao, and Yuning Jiang. Acquisition of localization confidence for accurate object detection. In *Proceedings of the European Conference on Computer Vision*, pages 784–799, 2018. 1

[2] Matej Kristan, Ales Leonardis, Jiri Matas, Michael Felsberg, Roman Pflugfelder, Luka ˇCehovin Zajc, Tomas Vojir, Goutam Bhat, Alan Lukezic, Abdelrahman Eldesokey, et al. The sixth visual object tracking vot2018 challenge results. 2018. 1

[3] Bo Li, Wei Wu, Qiang Wang, Fangyi Zhang, Junliang Xing, and Junjie Yan. Siamrpn++: Evolution of siamese visual tracking with very deep networks. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2018. 1