Deeper and Wider Siamese Networks for Real-Time Visual Tracking

Zhipeng Zhang and **Houwen Peng**

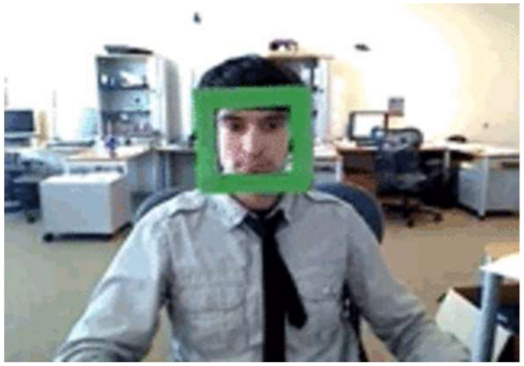
Microsoft Research Asia (MSRA)

CVPR 2019 Oral

June 16th - June 20th, Long Beach, CA

Visual Object Tracking

- Definition
  - It aims to estimate the position of arbitrary targets in a video sequence, given only the location in initial frame.
- Category
  - **Single object tracking**
  - Multiple object tracking

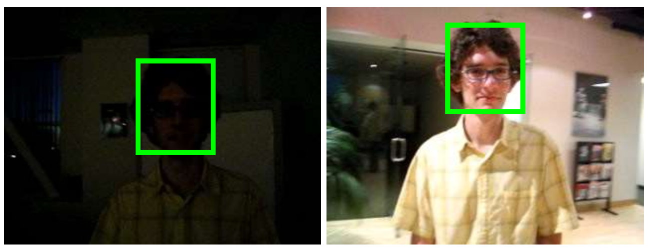


https://www.learnopencv.com/object-tracking-using-opencv-cpp-python/

Visual Object Tracking

- Challenges

Illumination Variation

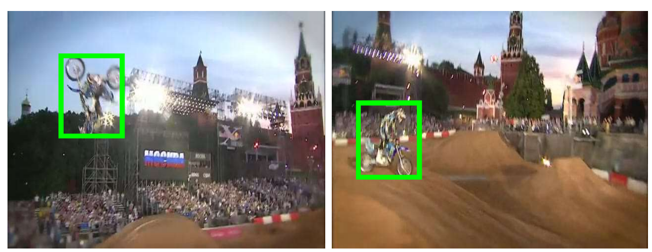


Occlusion

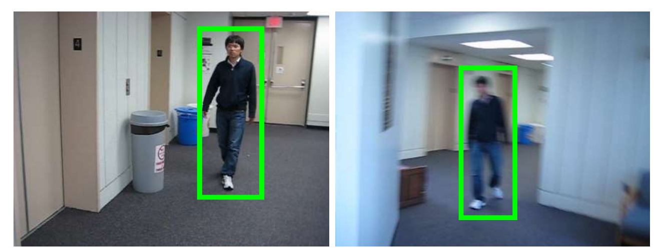
Background Clutters

Scale Variation

Rotation



Motion Blur



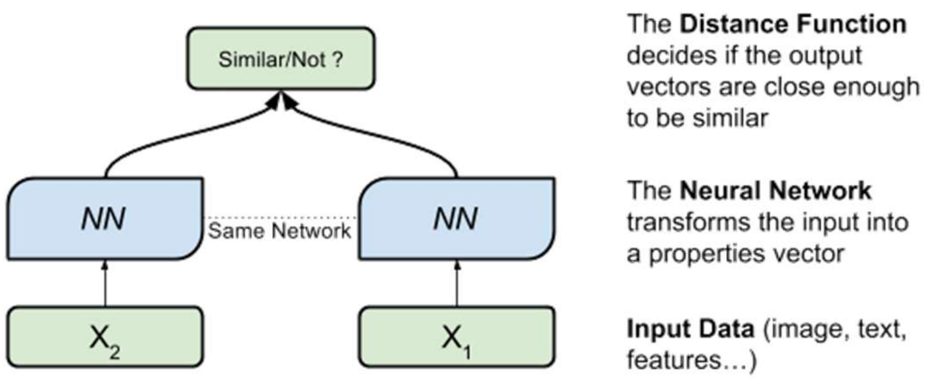
Outline

- Background on Siamese Trackers
- Motivation
- Analysis and Guidelines
- Method
- Experiments

Background on Siamese Trackers

- Siamese network architecture
  - Network and weight sharing
  - Metric learning, loss
  - Increase training samples naturally

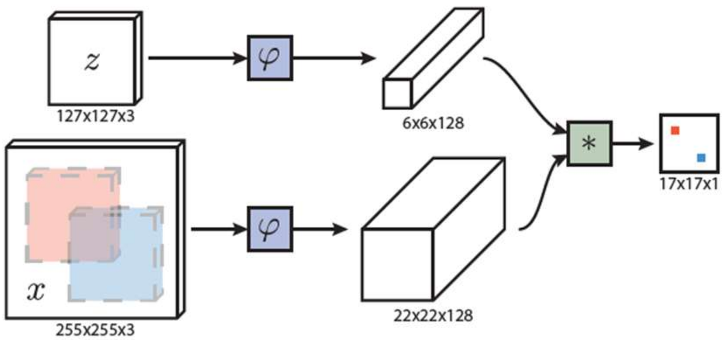
- Applications
  - Face verification
  - Person re-ID



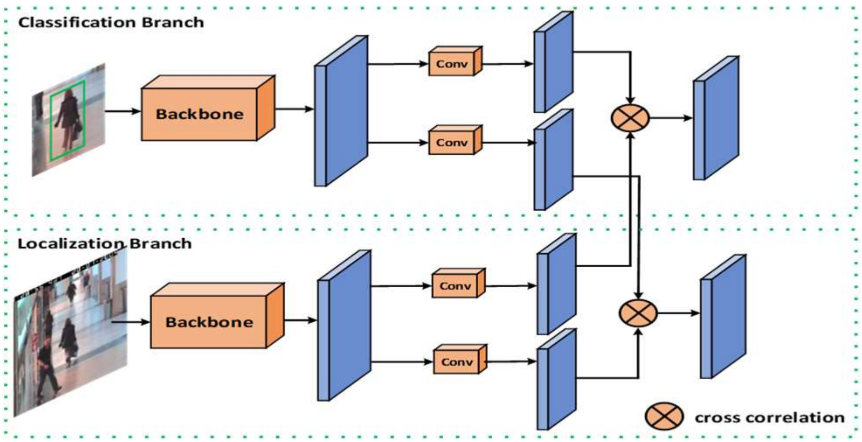
Parkhi, O.M., Vedaldi, A., Zisserman, A.: Deep face recognition. BMVC 2015 1(3) (2015)
https://aws.amazon.com/cn/blogs/machine-learning/combining-deep-learning-networks-gan-and-siamese-to-generate-high-quality-life-like-images

Background on Siamese Trackers

- SiamFC
  - Fully-convolutional networks
  - Similarity learning
  - Offline model



- SiamRPN
  - Region proposal networks
  - More accurate localization



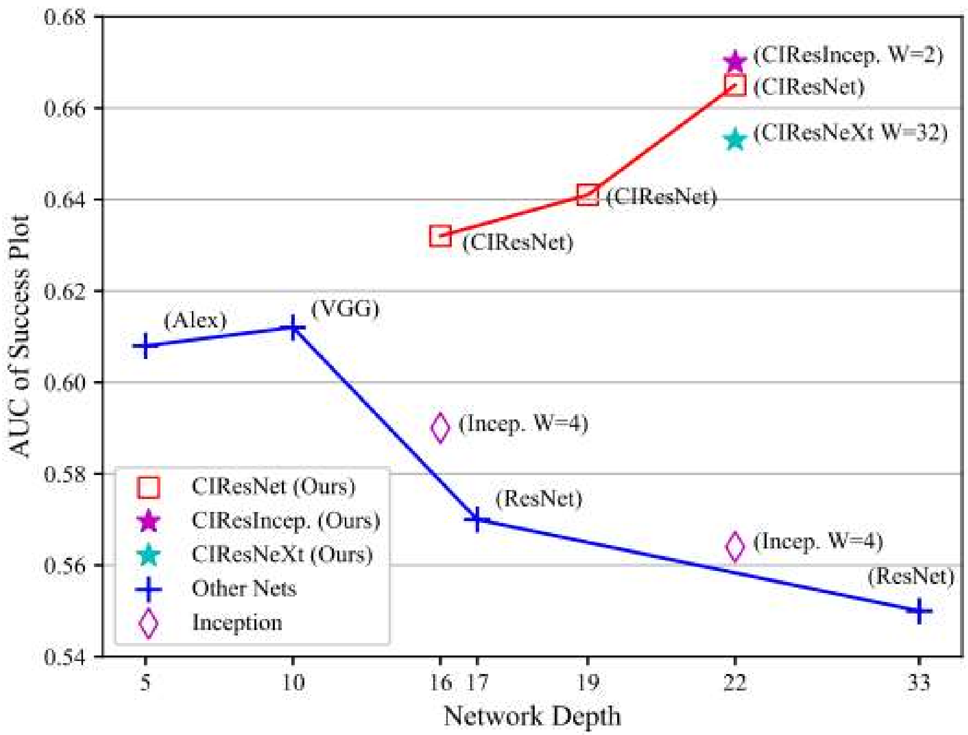
[SiamFC] L. Bertinetto, J. Valmadre, J. F. Henriques, A. Vedaldi, and P. H. Torr. Fully-convolutional siamese networks for object tracking. In ECCV, pages 850–865. Springer, 2016
[SiamRPN] B. Li, J. Yan, W. Wu, Z. Zhu, and X. Hu. High performance visual tracking with siamese region proposal network. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 8971–8980, 2018.

Outline

- Background on Siamese Trackers
- **Motivation**
- Analysis and Guidelines
- Method
- Experiments

Motivation

- The backbone network is still the classical AlexNet
- No significant performance improvements on more powerful backbones



Outline

- Background on Siamese Trackers
- Motivation
- **Analysis and Guidelines**
- Method
- Experiments

Analysis and Guidelines

- What is the underlying causes of this phenomenon?

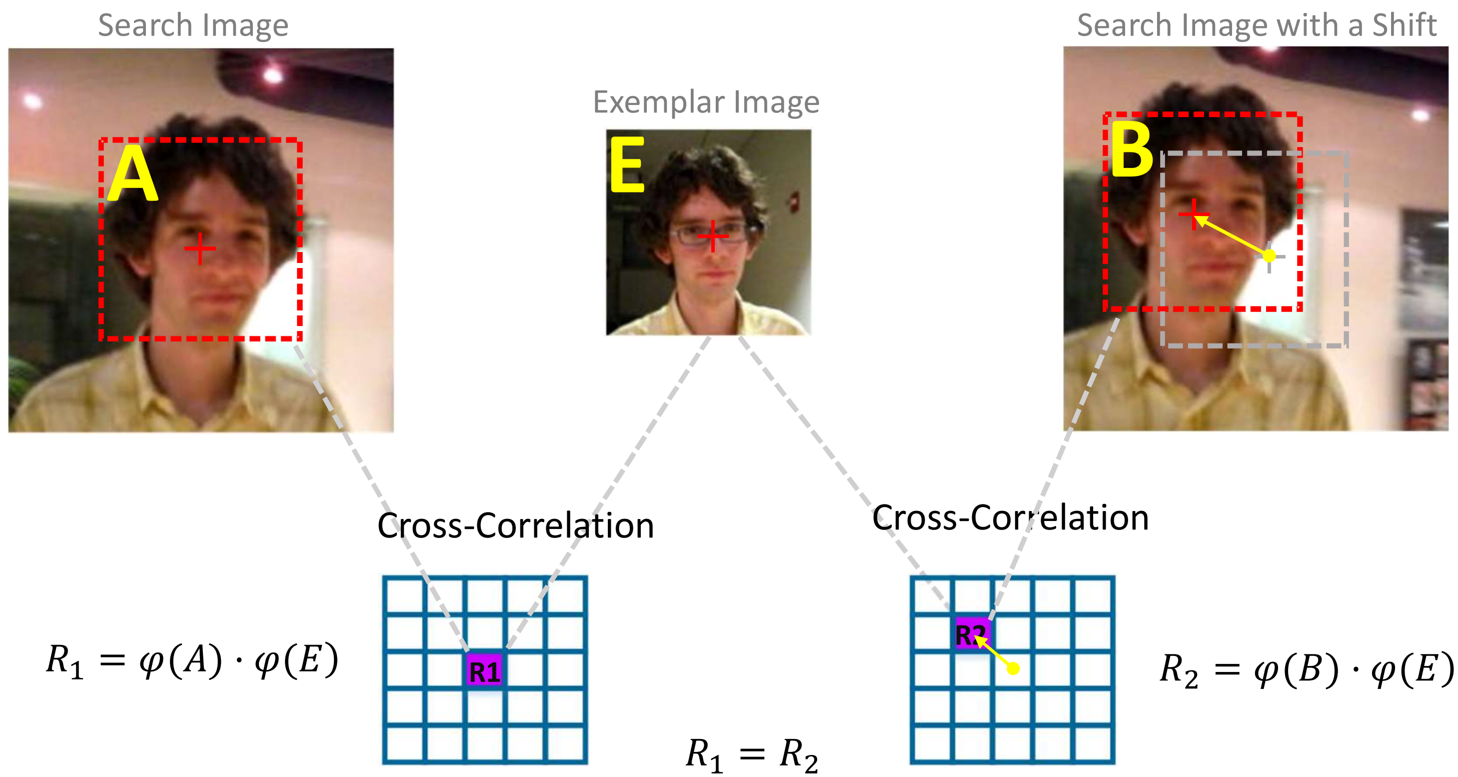
# NUM	①	②	③	④	⑤	⑥	⑦	⑧	⑨	⑩
RF[1]	Max(127)	+24	+16	+8	±0 (87)	±0	-8	-16	+16	+16
STR	8	8	8	8	8	8	8	8	16	4
OFS	1	3	4	5	6	16	7	8	2	7
PAD	✗	✗	✗	✗	✗	✓	✗	✗	✗	✗
Alex	0.56	0.57	0.60	0.60	0.61	0.55	0.59	0.58	0.55	0.59
VGG	0.58	0.59	0.61	0.61	0.62	0.56	0.59	0.58	0.54	0.58

# NUM	①	②	③	④	⑤	⑥	⑦	⑧	⑨
RF	+32	+16	+8	±0 (91)	±0	-8	-16	+16	+16
STR	8	8	8	8	8	8	8	16	4
OFS	1	3	4	5	16	6	7	2	6
PAD	✗	✗	✗	✗	✓	✗	✗	✗	✗
ResNet	0.56	0.59	0.60	0.62	0.56	0.60	0.60	0.54	0.58
Incep.[2]	0.58	0.60	0.61	0.63	0.58	0.62	0.61	0.56	0.59

Padding Influence: Padding causes performance degradation

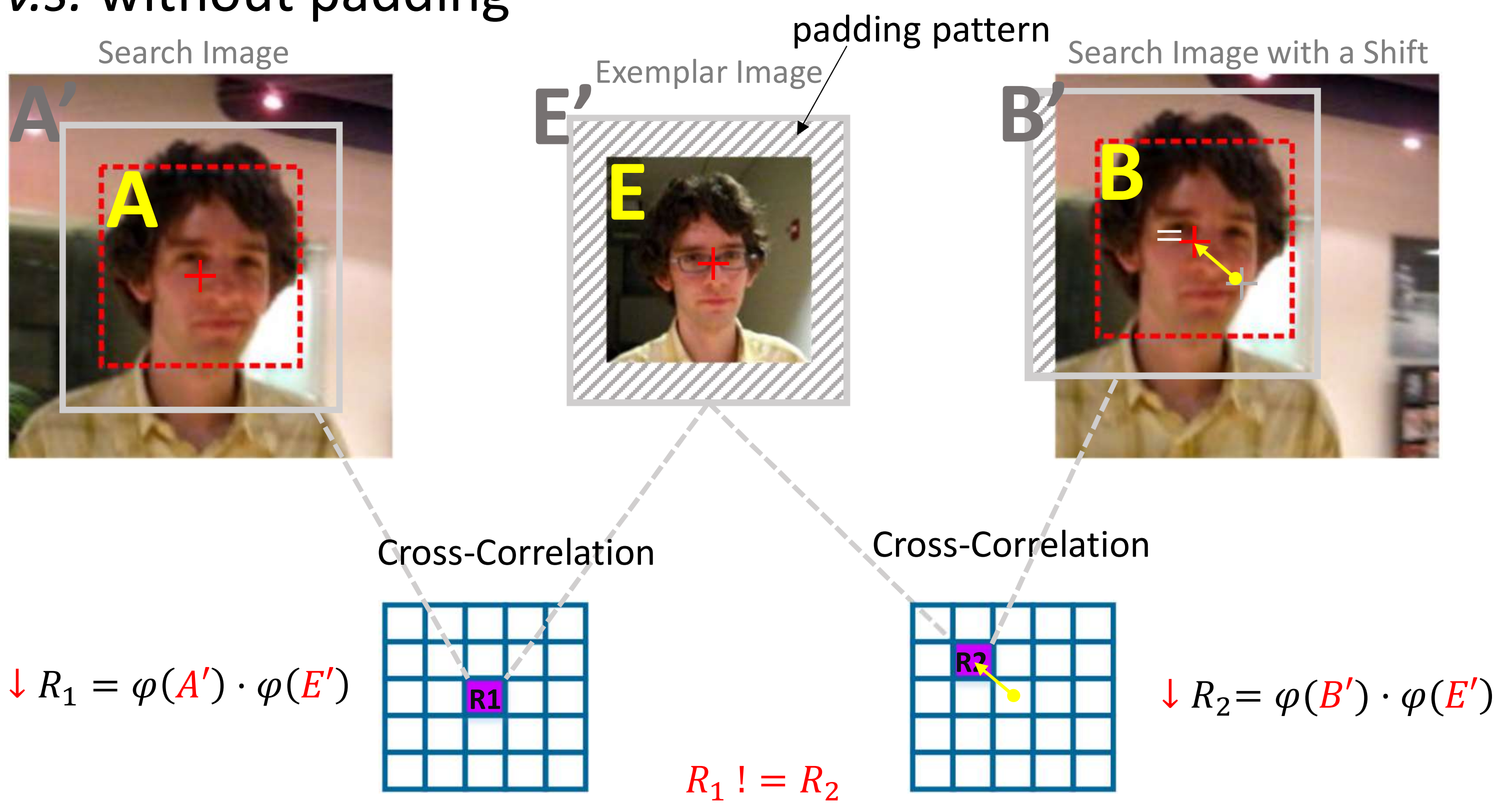
Analysis and Guidelines

- with *v.s.* <u>without</u> padding



Analysis and Guidelines

- <u>with</u> *v.s.* without padding



Analysis and Guidelines

- What is the underlying causes of this phenomenon?

# NUM	①	②	③	④	⑤	⑥	⑦	⑧	⑨	⑩
RF[1]	Max(127)	+24	+16	+8	±0 (87)	±0	-8	-16	+16	+16
STR	8	8	8	8	8	8	8	8	16	4
OFS	1	3	4	5	6	16	7	8	2	7
PAD	✗	✗	✗	✗	✗	✓	✗	✗	✗	✗
Alex	0.56	0.57	0.60	0.60	0.61	0.55	0.59	0.58	0.55	0.59
VGG	0.58	0.59	0.61	0.61	0.62	0.56	0.59	0.58	0.54	0.58

# NUM	①	②	③	④	⑤	⑥	⑦	⑧	⑨
RF	+32	+16	+8	±0 (91)	±0	-8	-16	+16	+16
STR	8	8	8	8	8	8	8	16	4
OFS	1	3	4	5	16	6	7	2	6
PAD	✗	✗	✗	✗	✓	✗	✗	✗	✗
ResNet	0.56	0.59	0.60	0.62	0.56	0.60	0.60	0.54	0.58
Incep.[2]	0.58	0.60	0.61	0.63	0.58	0.62	0.61	0.56	0.59

Padding Influence: Padding causes performance degradation

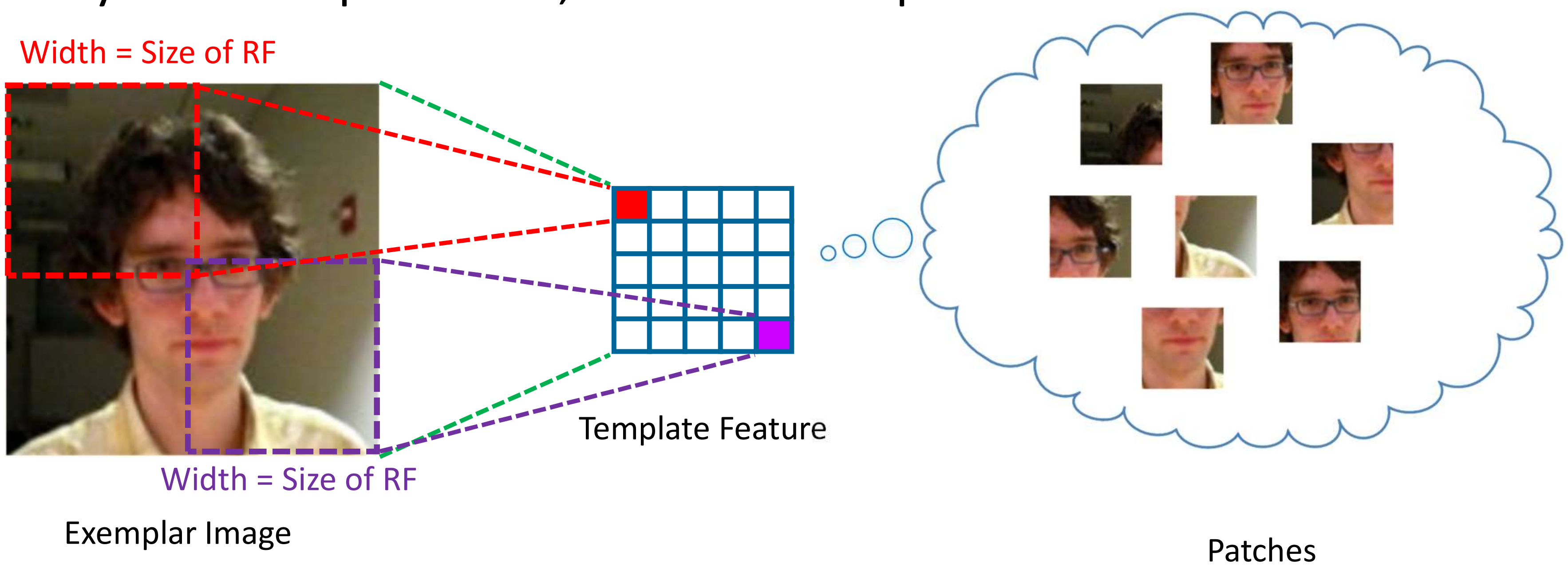
Receptive Field (RF) and Output Feature Size (OFC) Influence: Reasonable RF and OFS are necessary

Stride Influence: Siamese trackers prefer relatively smaller stride

RF, OFS, and stride are not independent of one another. Consider them together.

Analysis and Guidelines

- Analysis of receptive field, stride and output feature size



- Each element in the feature map corresponds to a patch in exemplar image.
- Overlap Ratio = 1 – stride/RF, large overlap ratio will decrease localization precision.

Analysis and Guidelines

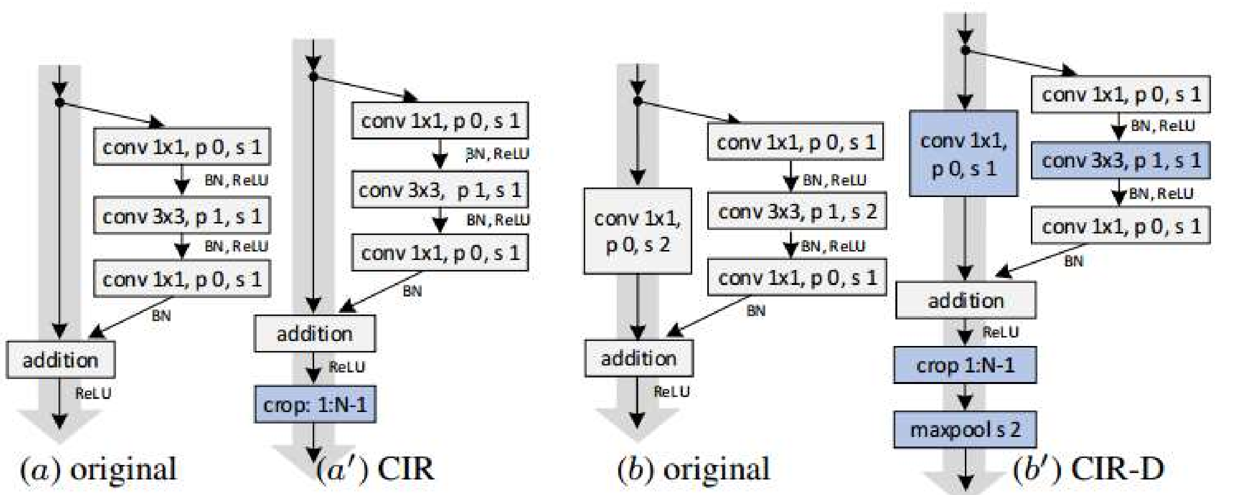
- Guidelines
  - Siamese trackers prefer a relatively small network stride, e.g. 4 or 8.
  - The receptive field of output features should be set based on its ratio to the size of the exemplar image (60%-80%).
  - Network stride, receptive field and output feature size should be considered as a whole when designing a network architecture.
  - For a fully convolutional Siamese matching network, it is critical to handle the problem of perceptual inconsistency between the two network streams.

Outline

- Background on Siamese Trackers
- Motivation
- Analysis and Guidelines
- **Method**
- Experiments

Method

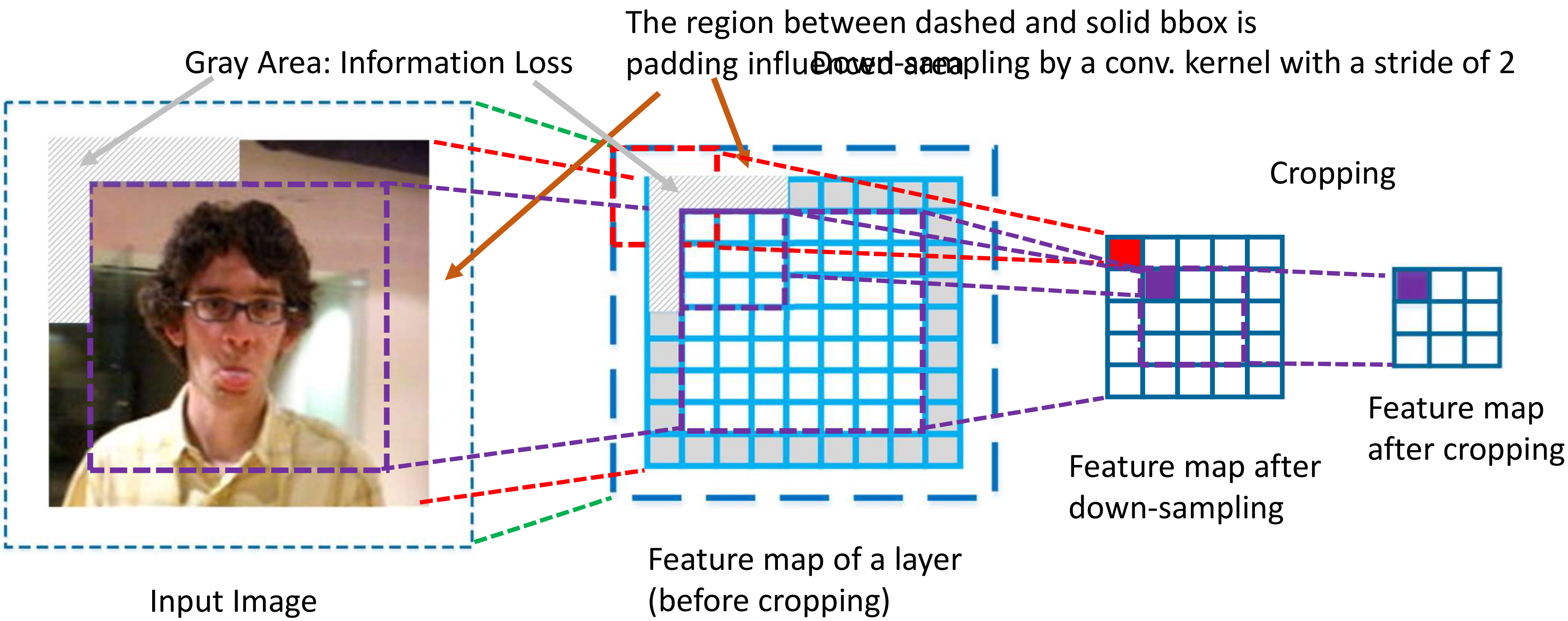
- Cropping-inside residual unit
  - **CIR Module:** center crop not only remove padding influence but also accelerate training and testing
  - **CIR-Downsampling Module**: reduce the spatial size of feature maps while doubling the number of feature channels



(a) original (a') CIR (b) original (b') CIR-D

Method

- Why we need CIR-Downsampling?



Method

- Modules: Cropping-inside residual units
  - Remove padding

- Design:
  - First, we determine the network stride.
  - Then, we stack CIR units.
  - When network depth increases, the receptive field may exceed this range. Therefore, we halve the stride to 4 to control the receptive field.

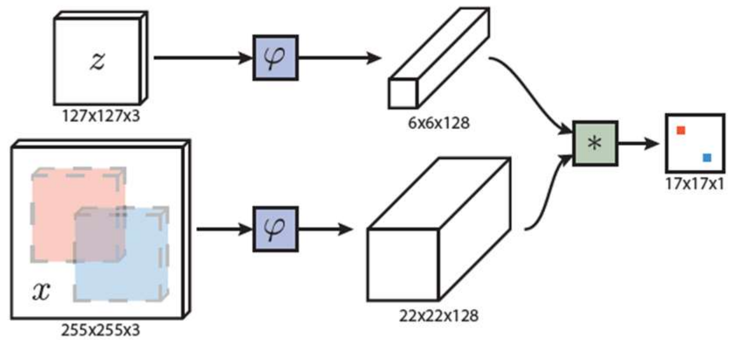
Method

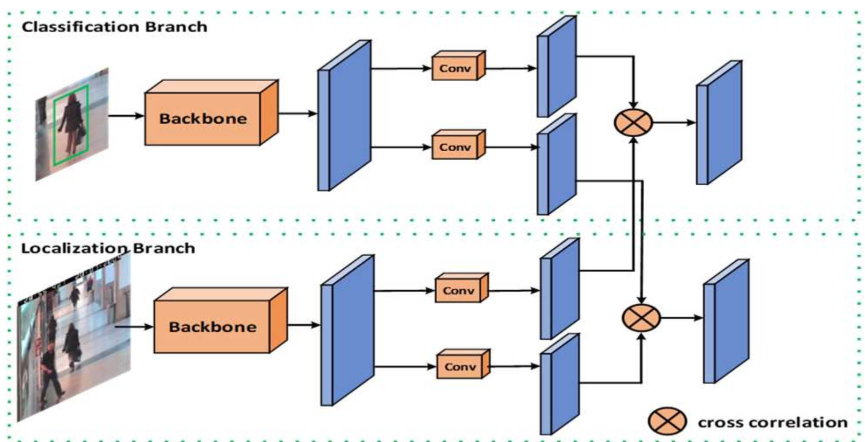
- Network Architecture

Stage	CIResNet-16	CIResNet-19	CIResNet-22	CIResInception-22	CIResNeXt-22	CIResNet-43
conv1	7×7, 64, stride 2					
conv2	2×2 max pool, stride 2					
	$\begin{bmatrix}1\times1,64\\3\times3,64\\1\times1,256\end{bmatrix}\times 1$	$\begin{bmatrix}1\times1,64\\3\times3,64\\1\times1,256\end{bmatrix}\times 2$	$\begin{bmatrix}1\times1,64\\3\times3,64\\1\times1,256\end{bmatrix}\times 3$	$\begin{bmatrix}1\times1,64\\3\times3,64\\1\times1,256\end{bmatrix}\times 3$ $[1\times1,64]\times 3$	$\begin{bmatrix}1\times1,64\\3\times3,64,C=32\\1\times1,256\end{bmatrix}\times 3$	$\begin{bmatrix}1\times1,64\\3\times3,64\\1\times1,256\end{bmatrix}\times 14$
conv3	$\begin{bmatrix}1\times1,128\\3\times3,128\\1\times1,512\end{bmatrix}\times 4$	$\begin{bmatrix}1\times1,128\\3\times3,128\\1\times1,512\end{bmatrix}\times 4$	$\begin{bmatrix}1\times1,128\\3\times3,128\\1\times1,512\end{bmatrix}\times 4$	$\begin{bmatrix}1\times1,128\\3\times3,128\\1\times1,512\end{bmatrix}\times 4$ $[1\times1,128]\times 4$	$\begin{bmatrix}1\times1,128\\3\times3,128,C=32\\1\times1,512\end{bmatrix}\times 4$	
	cross correlation Eq. 1					
# RF	77	85	93	13∼93	93	105
# OFS	7	6	5	5	5	6
# Params	1.304 M	1.374 M	1.445 M	1.695 M	1.417 M	1.010 M
# FLOPs	2.43 G	2.55 G	2.65 G	2.71 G	2.52 G	6.07 G

Method

- Applications
  - **SiamFC**
    - Fully-convolutional networks
    - Similarity learning
    - Offline model
  - **SiamRPN**
    - Region proposal networks
    - More accurate localization





[SiamFC] L. Bertinetto, J. Valmadre, J. F. Henriques, A. Vedaldi, and P. H. Torr. Fully-convolutional siamese networks for object tracking. In ECCV, pages 850–865. Springer, 2016
[SiamRPN] B. Li, J. Yan, W. Wu, Z. Zhu, and X. Hu. High performance visual tracking with siamese region proposal network. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 8971–8980, 2018.

Experiment

- Comparison with baselines

Backbone	OTB(AUC)		VOT-17(EAO)	
	SiamFC	SiamRPN	SiamFC	SiamRPN
AlexNet	0.608[2]	0.637[20]	0.188[17]	0.244[20]
CIResNet-16	0.632	0.651	0.202	0.260
CIResNet-19	0.640	0.660	0.225	0.279
CIResNet-22	0.665	0.665	**0.234**	**0.301**
CIResIncep.-22	**0.666**	**0.673**	0.215	0.296
CIResNeXt-22	0.654	0.660	0.230	0.285
CIResNet-43	0.638	0.652	0.207	0.265

Experiment

- Comparison to state-of-the-arts

Table 5: Performance comparisons on five tracking benchmarks. Red, Green and Blue fonts indicate the top-3 trackers, respectively.

Tracker	Year	OTB-2013		OTB-2015		VOT15			VOT16			VOT17		
		AUC	Prec.	AUC	Prec.	A	R	EAO	A	R	EAO	A	R	EAO
SRDCF [5]	2015	0.63	0.84	0.60	0.80	0.56	1.24	0.29	0.54	0.42	0.25	0.49	0.97	0.12
SINT [34]	2016	0.64	0.85	-	-	-	-	-	-	-	-			
Staple [1]	2016	0.60	0.80	0.58	0.78	0.57	1.39	0.30	0.54	0.38	0.30	0.52	0.69	0.17
SiamFC [2]	2016	0.61	0.81	0.58	0.77	0.53	0.88	0.29	0.53	0.46	0.24	0.50	0.59	0.19
ECO-HC [4]	2017	0.65	0.87	0.64	0.86	-	-	-	0.54	0.3	0.32	0.49	0.44	0.24
PTAV [8]	2017	0.66	0.89	0.64	0.85	-	-	-	-	-	-	-	-	-
DSiam [12]	2017	0.64	0.81	-	-	-	-	-	-	-	-	-	-	-
CFNet [35]	2017	0.61	0.80	0.59	0.78	-	-	-	-	-	-	-	-	-
StructSiam [40]	2018	0.64	0.88	0.62	0.85	-	-	-	-	-	0.26	-	-	-
TriSiam [7]	2018	0.62	0.82	0.59	0.78	-	-	-	-	-	-	-	-	0.20
SiamRPN [20]	2018	-	-	0.64	0.85	0.58	1.13	0.35	0.56	0.26	0.34	0.49	0.46	0.24
SiamFC+	Ours	0.67	0.88	0.64	0.85	0.57	1.18	0.31	0.54	0.38	0.30	0.50	0.49	0.23
SiamRPN+	Ours	0.67	0.87	0.67	0.86	0.59	1.08	0.38	0.58	0.24	0.37	0.52	0.41	0.30

Our SiamFC+ and SiamRPN+ obtain up to **9.8%/5.7% (AUC), 23.3%/8.8% (EAO) and 24.4%/25.0% (EAO)** relative improvements over the original versions on the **OTB-15, VOT-16 and VOT-17** datasets, respectively **solely** due to the proposed backbone.

Visual Comparison

Paper and Code

- https://arxiv.org/pdf/1901.01660.pdf
- https://github.com/researchmm/SiamDW

Thanks!

We are hiring research interns.

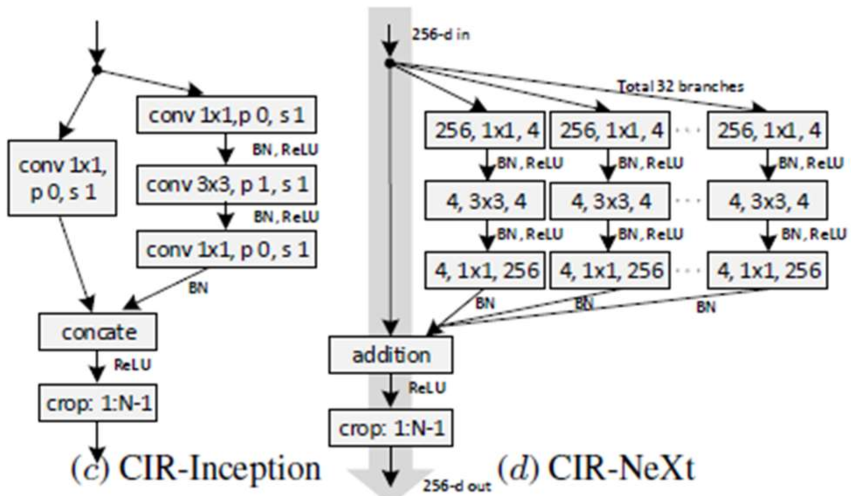
houwen.peng@microsoft.com

True Problems and Future Work

- Deeper Networks
- Online model update
- Instance-level representation

Backup

- Wider modules



(c) CIR-Inception (d) CIR-NeXt