Deeper and Wider Siamese Networks for Real-Time Visual Tracking

Zhipeng Zhang and Houwen Peng

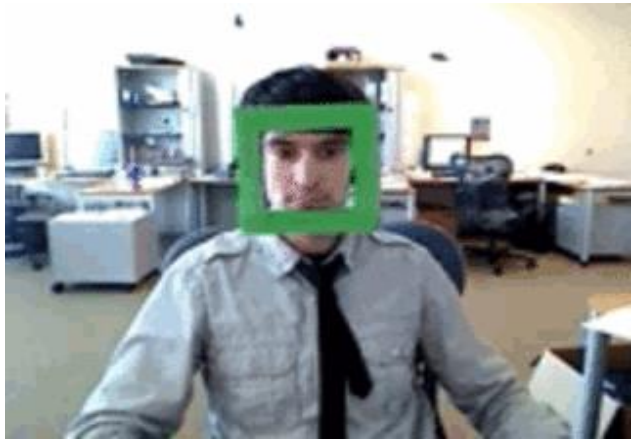
Microsoft Research Asia (MSRA)

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/

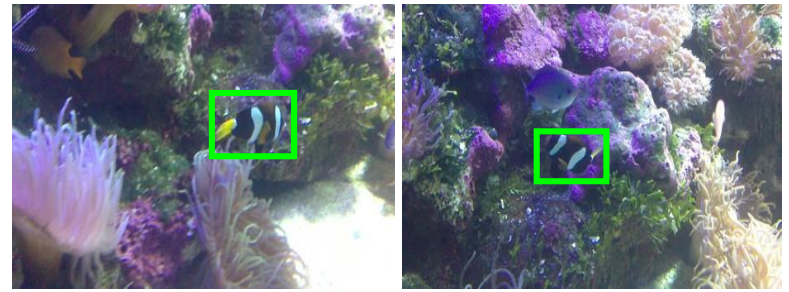
Visual Object Tracking

- Challenges

Illumination Variation

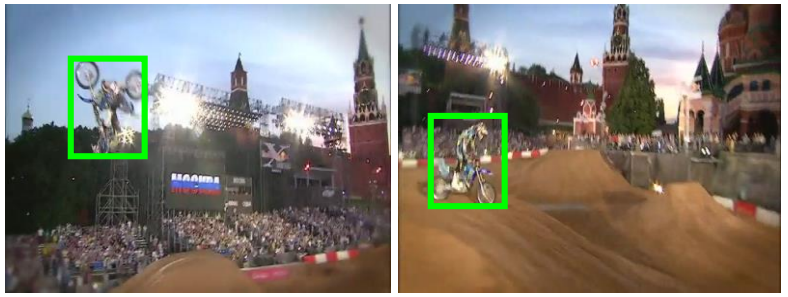
Occlusion

Background Clutters



Scale Variation

Rotation



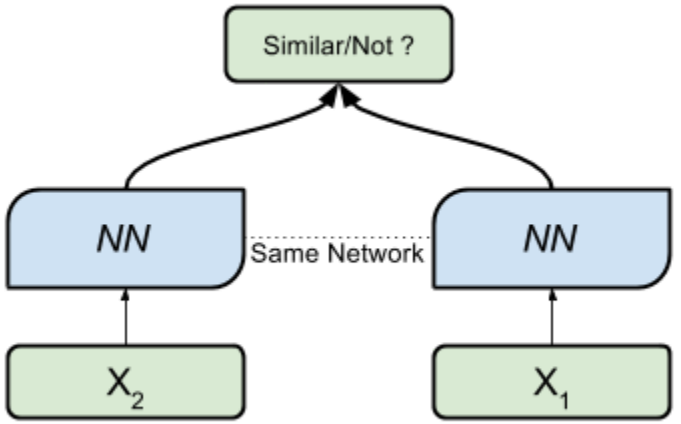
Motion Blur

Outline

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

Background on Siamese Trackers

- Siamese network architecture
  - Network and weight sharing
  - Metric learning, loss
  - Increase training samples naturally
- Applications
  - Face verification
  - Person re-ID



The **Distance Function** decides if the output vectors are close enough to be similar

The **Neural Network** transforms the input into a properties vector

Input Data (image, text, features…)

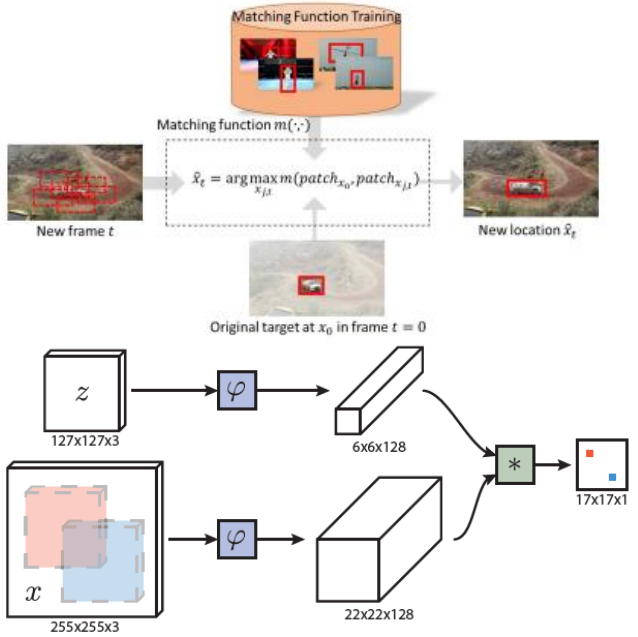
Parkhi, O.M., Vedaldi, A., Zisserman, A.: Deep face recognition. BMVC 2015 1(3) (2015)
https://aws.amazon.com/cn/blogs/machine-learning/combining-deep-learning-networks-gan-and-siamese-to-generate-high-quality-life-like-images

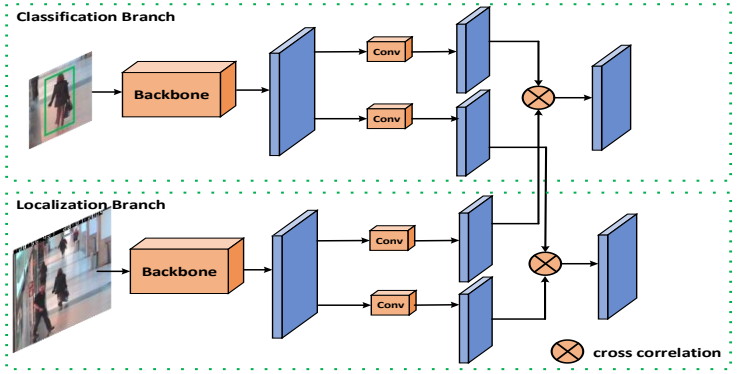
Background on Siamese Trackers

- **SINT**
  - Similarity learning
  - Offline model

- **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

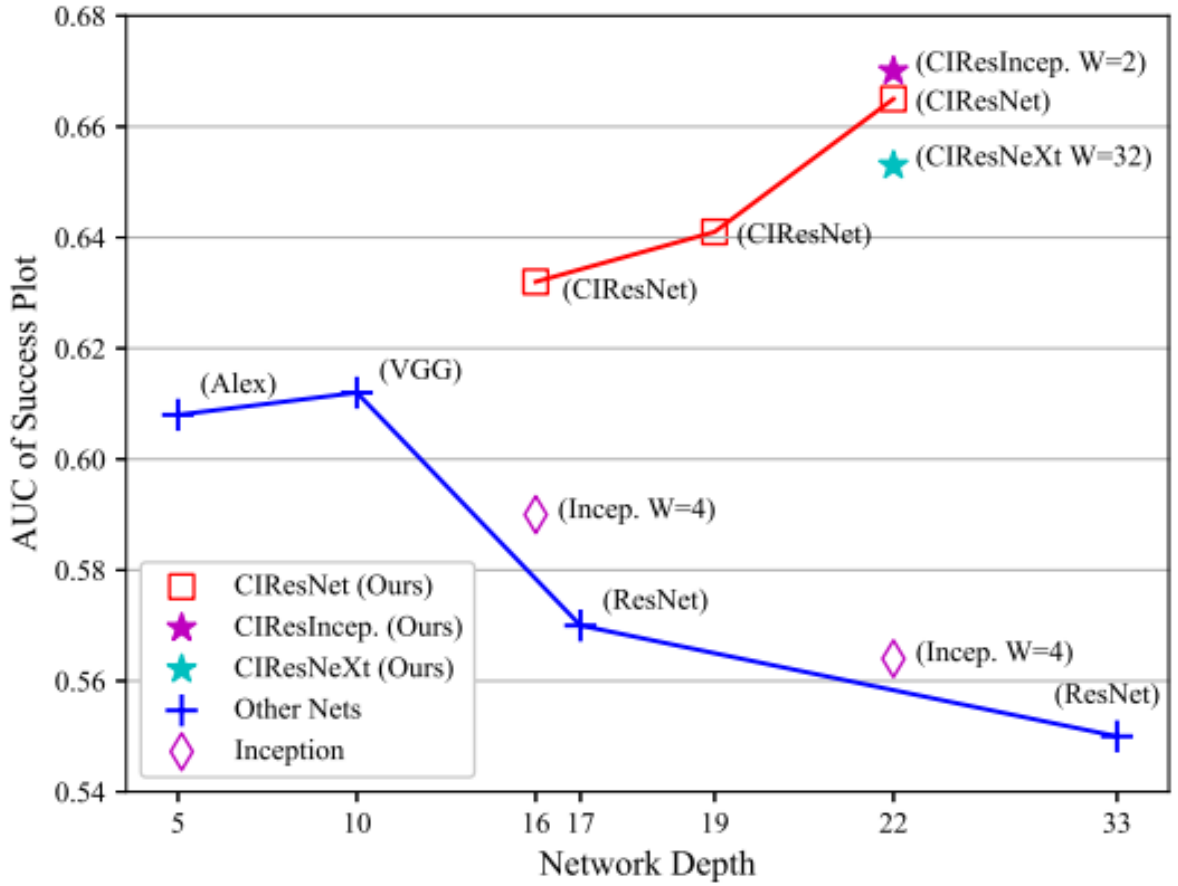
[SiamRPN] B. Li, J. Yan, W. Wu, Z. Zhu, and X. Hu. High performance visual tracking with siamese region proposal network. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 8971–8980, 2018.

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?

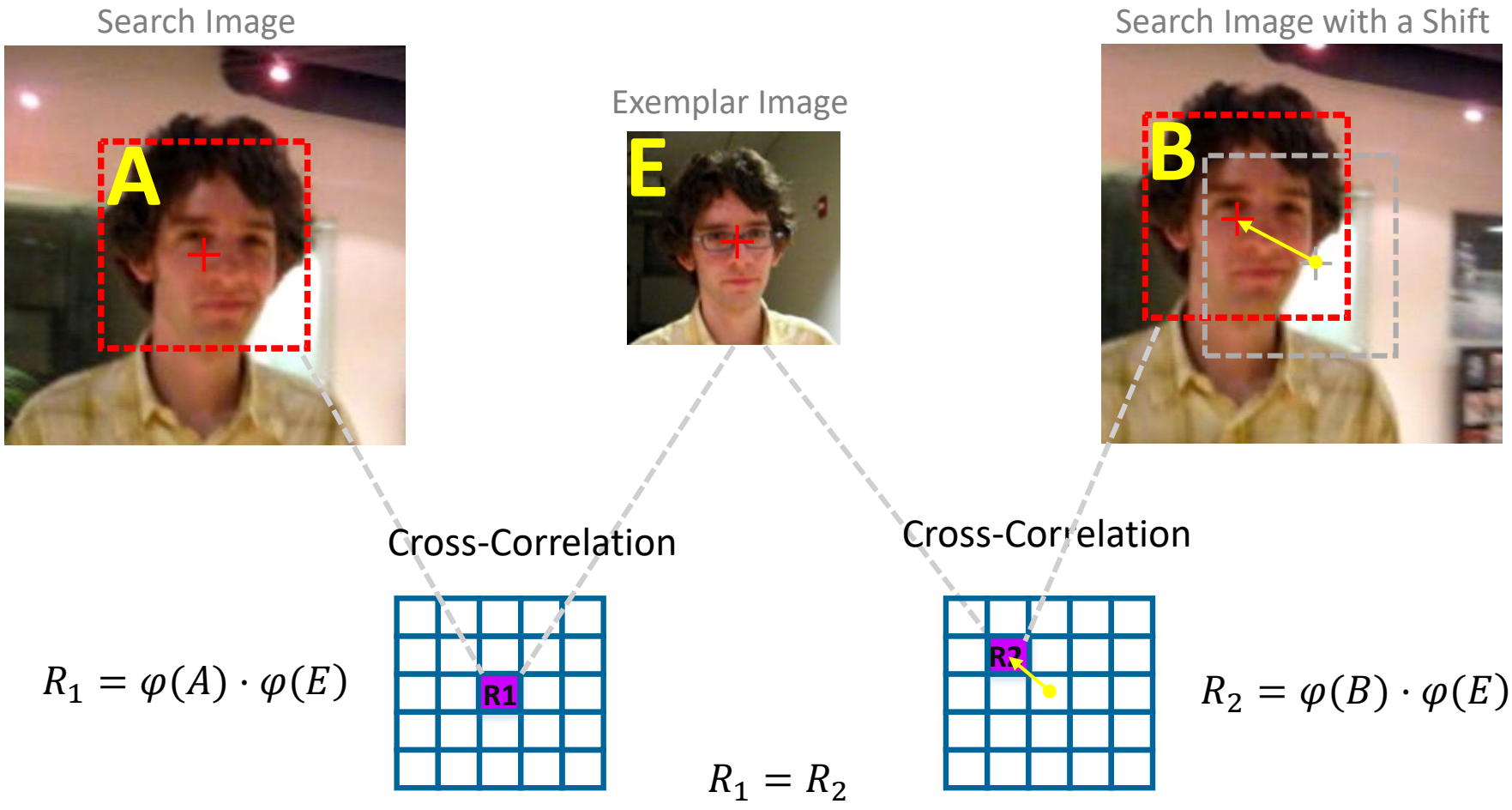
# NUM	①	②	③	④	⑤	⑥	⑦	⑧	⑨	⑩
RF[1]	Max(127)	+24	+16	+8	±0 (87)	±0	-8	-16	+16	+16
STR	8	8	8	8	8	8	8	8	16	4
OFS	1	3	4	5	6	16	7	8	2	7
PAD	✗	✗	✗	✗	✗	✓	✗	✗	✗	✗
Alex	0.56	0.57	0.60	0.60	0.61	0.55	0.59	0.58	0.55	0.59
VGG	0.58	0.59	0.61	0.61	0.62	0.56	0.59	0.58	0.54	0.58

# 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

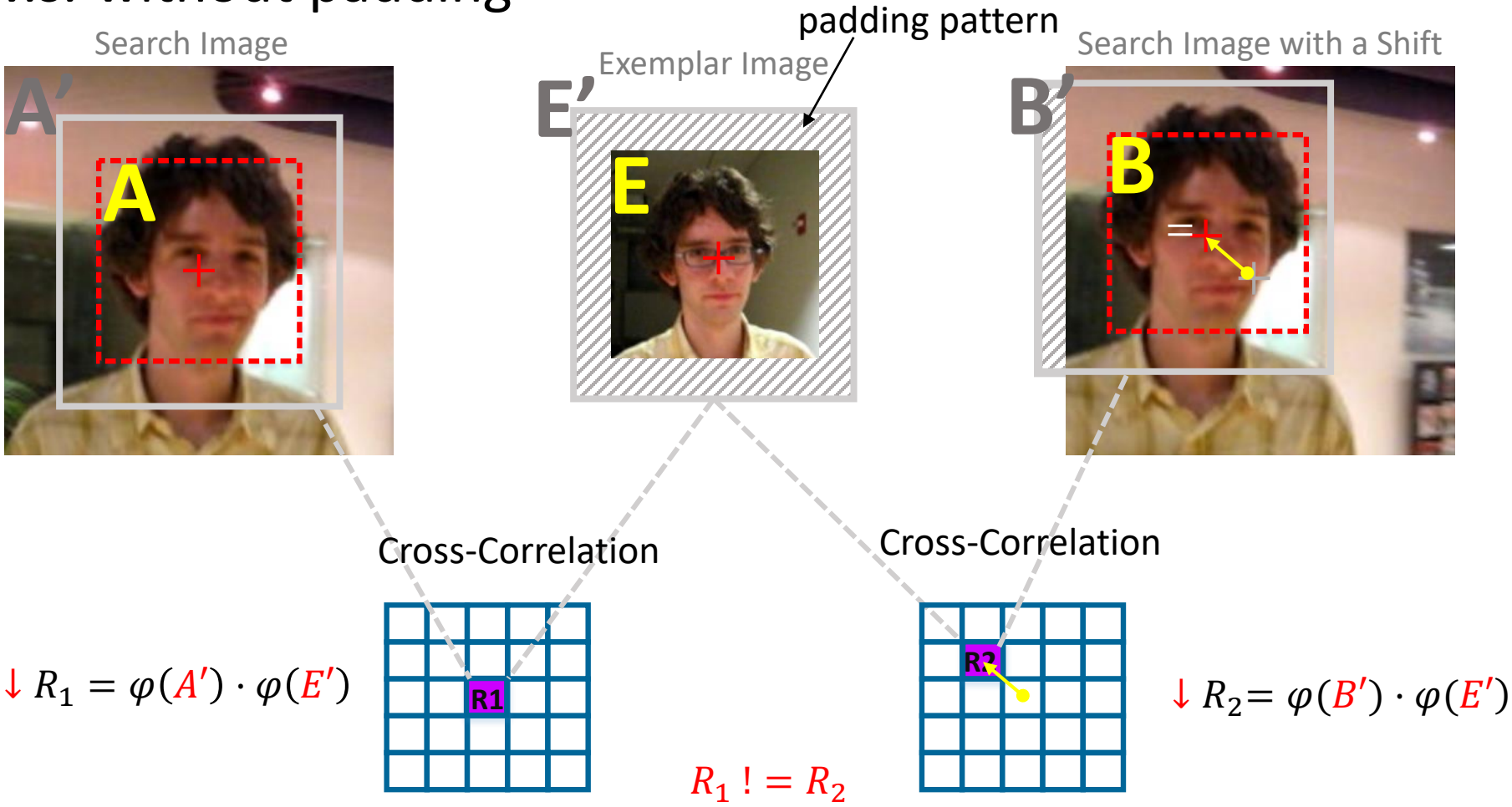
Analysis and Guidelines

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



Analysis and Guidelines

- with *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

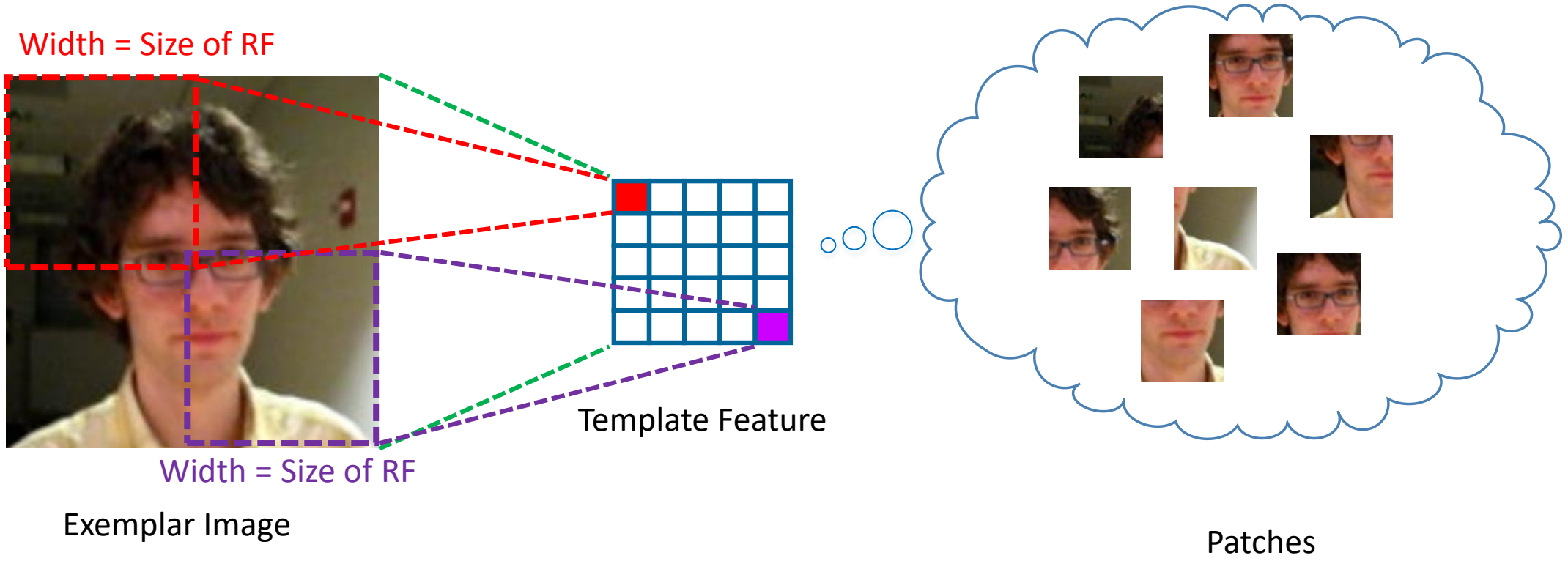
Receptive Field (RF) and Output Feature Size (OFC) Influence: Reasonable RF and OFS are necessary

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

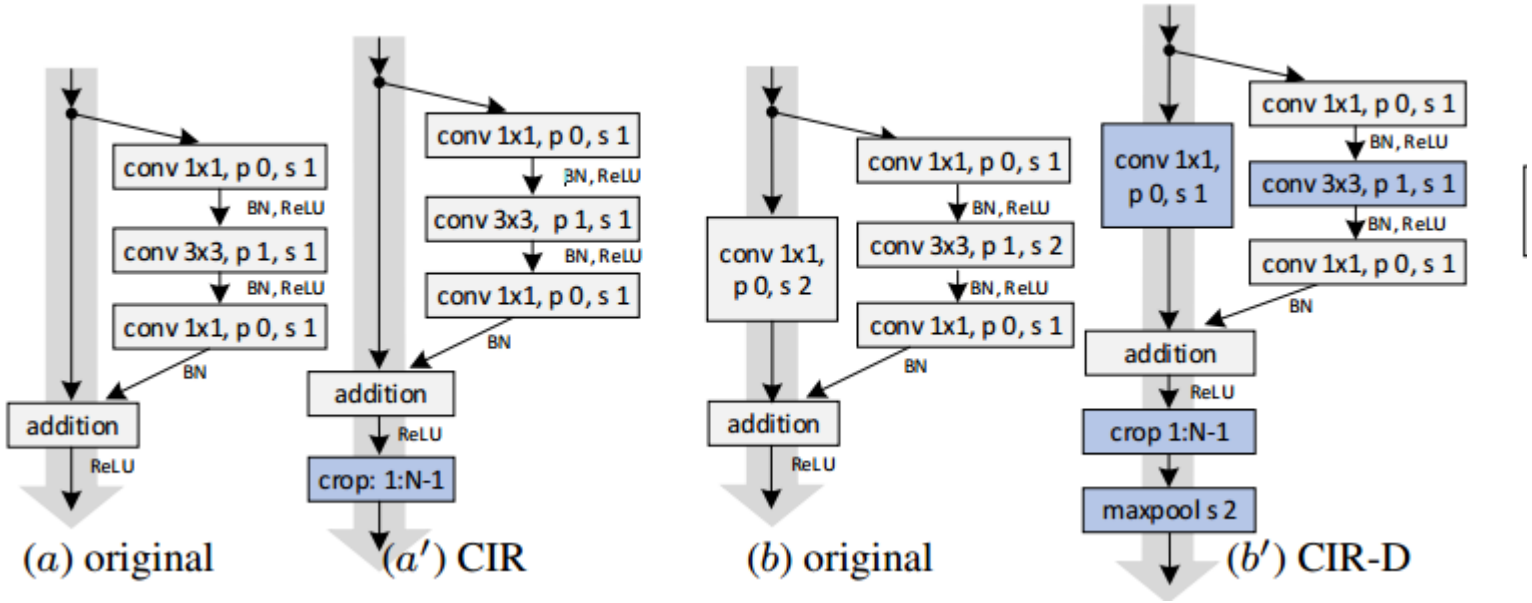
- Guidelines
  - Siamese trackers prefer a relatively small network stride, e.g. 4 or 8.
  - The receptive field of output features should be set based on its ratio to the size of the exemplar image (60%-80%).
  - Network stride, receptive field and output feature size should be considered as a whole 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

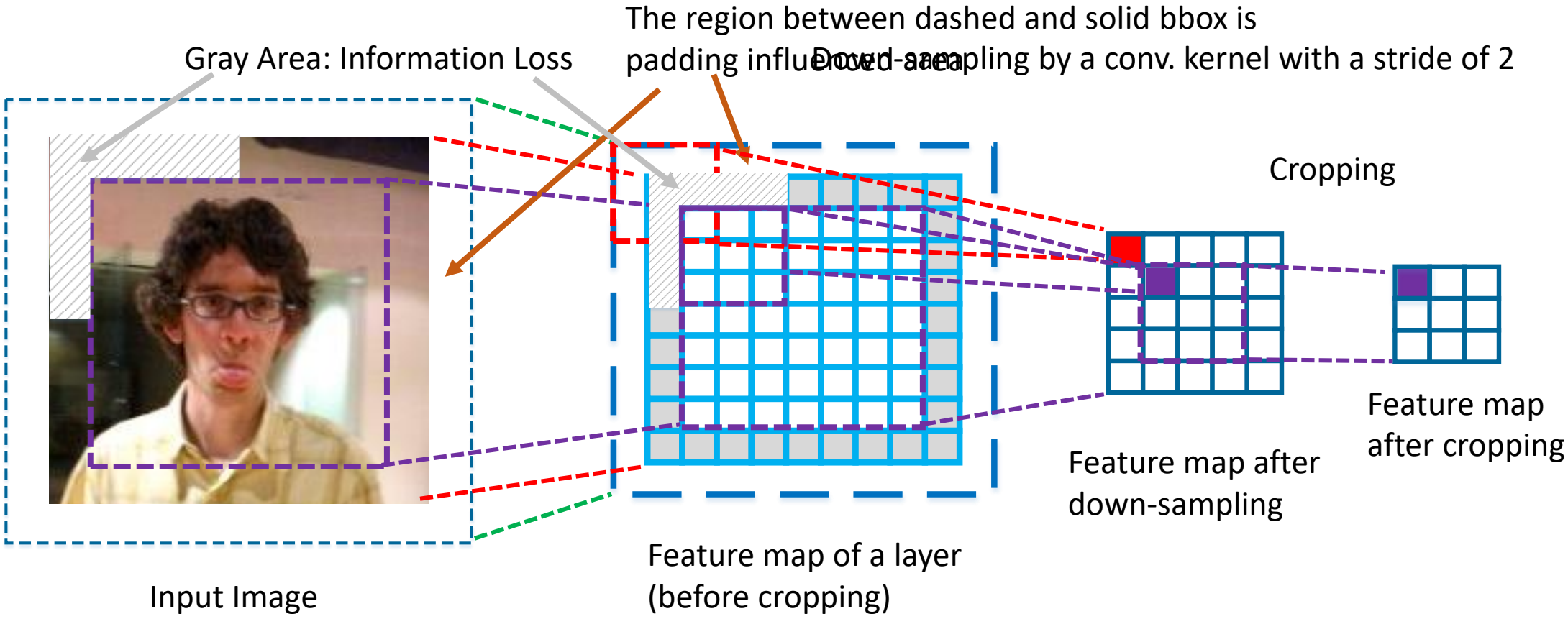
Method

- 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



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.

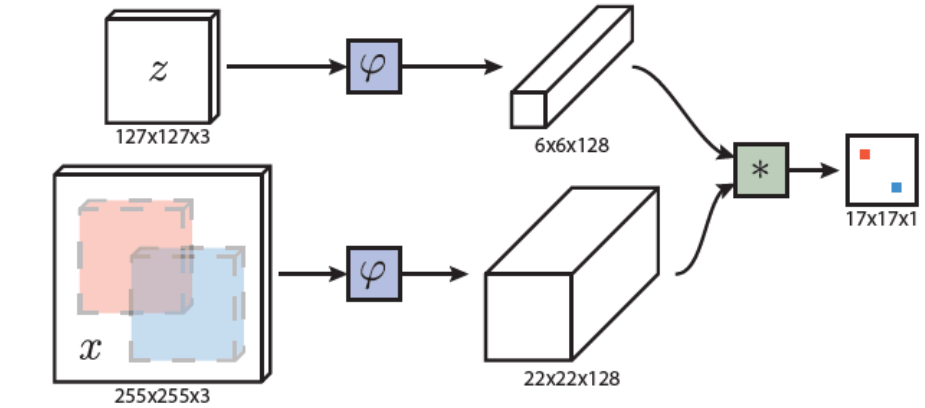
Method

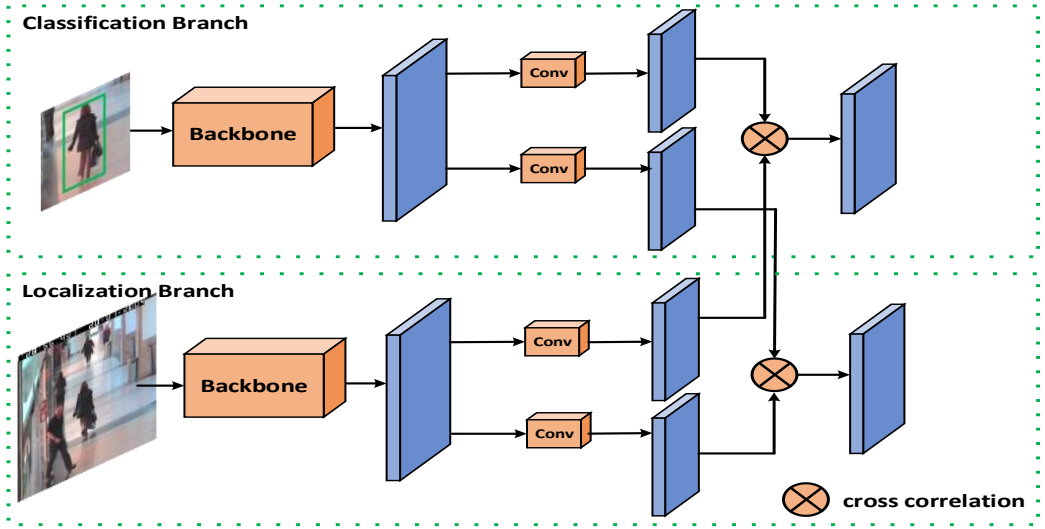
- Network Architecture

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					
	$\begin{bmatrix} 1\times1, 64 \\ 3\times3, 64 \\ 1\times1, 256 \end{bmatrix} \times 1$	$\begin{bmatrix} 1\times1, 64 \\ 3\times3, 64 \\ 1\times1, 256 \end{bmatrix} \times 2$	$\begin{bmatrix} 1\times1, 64 \\ 3\times3, 64 \\ 1\times1, 256 \end{bmatrix} \times 3$	$\begin{bmatrix} 1\times1, 64 \\ 3\times3, 64 \\ 1\times1, 256 \end{bmatrix} \times 3$	$\begin{bmatrix} 1\times1, 64 \\ 3\times3, 64, C=32 \\ 1\times1, 256 \end{bmatrix} \times 3$	$\begin{bmatrix} 1\times1, 64 \\ 3\times3, 64 \\ 1\times1, 256 \end{bmatrix} \times 14$
				$[1\times1, 64]\times3$		
conv3	$\begin{bmatrix} 1\times1, 128 \\ 3\times3, 128 \\ 1\times1, 512 \end{bmatrix} \times 4$	$\begin{bmatrix} 1\times1, 128 \\ 3\times3, 128 \\ 1\times1, 512 \end{bmatrix} \times 4$	$\begin{bmatrix} 1\times1, 128 \\ 3\times3, 128 \\ 1\times1, 512 \end{bmatrix} \times 4$	$\begin{bmatrix} 1\times1, 128 \\ 3\times3, 128 \\ 1\times1, 512 \end{bmatrix} \times 4$	$\begin{bmatrix} 1\times1, 128 \\ 3\times3, 128, C=32 \\ 1\times1, 512 \end{bmatrix} \times 4$	
				$[1\times1, 128]\times4$		
	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.

Updated Results

Train on VID/(VID and YTB for RPN)

Models	OTB13	OTB15	VOT15	VOT16	VOT17
Alex-FC	0.608	0.579	0.289	0.235	0.188
Alex-RPN	-	0.637	0.349	0.344	0.244
CIResNet22-FC	0.663	0.644	0.318	0.303	0.234
CIResIncep22-FC	0.662	0.642	0.310	0.295	0.236
CIResNext23-FC	0.659	0.633	0.297	0.278	0.229
CIResNext22-RPN	0.674	0.666	0.381	0.376	0.294

Train on GOT10K

Models	OTB13	OTB15	VOT15	VOT16	VOT17
CIResNet22-FC	0.664	0.654	0.361	0.335	0.266
CIResNet22W-FC	**0.689**	**0.664**	**0.368**	**0.352**	**0.269**
CIResIncep22-FC	0.673	0.650	0.332	0.305	0.251
CIResNext22-FC	0.668	0.651	0.336	0.304	0.246

Paper and Code

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

I “hate” Siamese Tracking

False Prosperity

- Truly hard to reproduce for many works
- Unstable to dataset
- Easy to explode when using deeper network

Hard to reproduce

1. Do not try to reproduce all works!! We don’t know which work really works

2. Do not try to start from scratch. Many good demos are presented on github. You can refer my code for further improvement

3. Do not waste too much time on a work. Performance lies.

4. Work with others. Siamese is full of tricks.

Instabilities

1. Do not deny this to lie to pretend peace. Talkers try to boast their works.

2. Gods and Ghosts are in same room. Siamese is stable for similar scene. This is a superiority for building specific trackers, eg. Car trackers.

3. Generic tracker is a ill-posed problem. No best tracker !!! Be smart.

To fresher

1. Be patient. Start from Alex-SiamFC.

2. Accumulate experience. Many tricks in tracking.

3. Focus on real problem. Do not simply pile up works.

4. Don't trust others easily. Believe in your experiments. Behind a good job are hundreds of experiments.

5. Cooperate with others.

6. Hold your sudden inspiration.

Thanks!

We are hiring research interns.

houwen.peng@microsoft.com (for interns)
zhangzhipeng2017@ia.ac.cn (for paper or discussion)