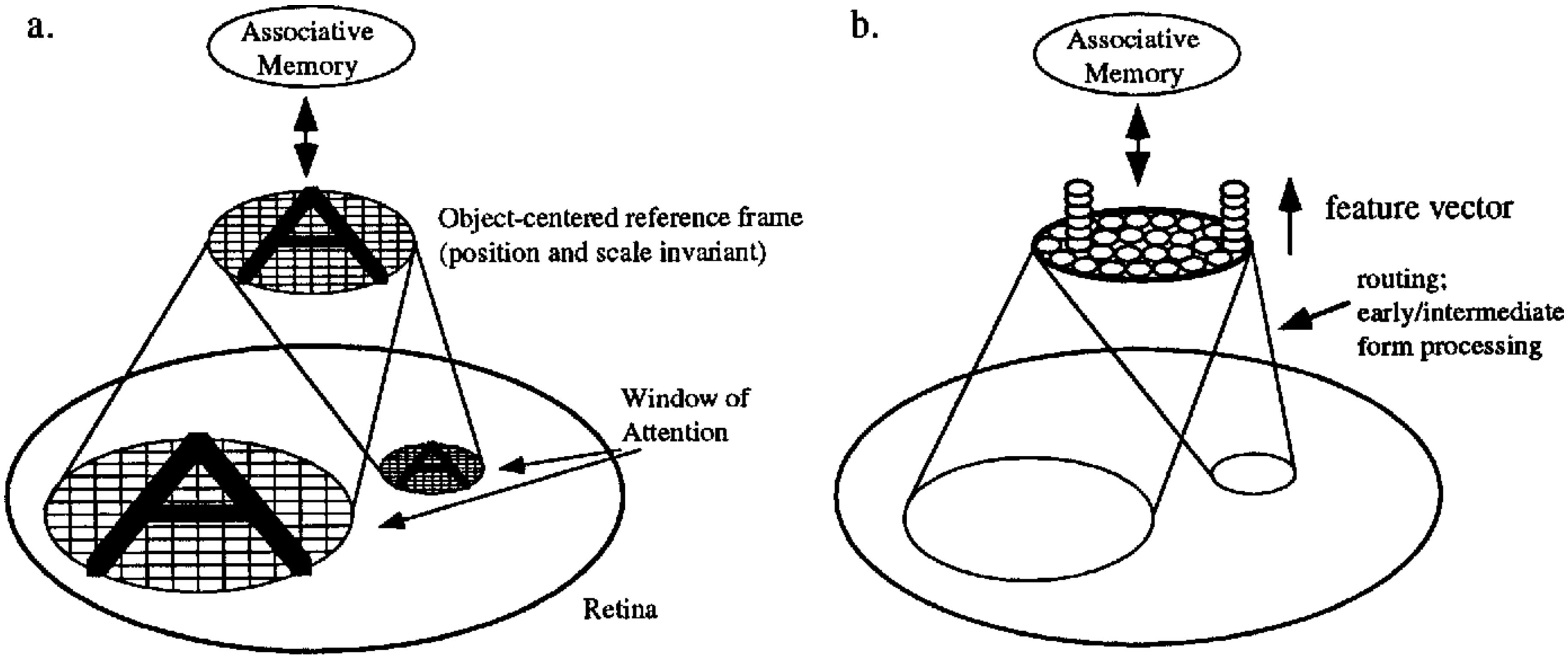
香港中文大學
The Chinese University of Hong Kong

The Glimpse of Detectron: Dynamic Forwarding and Routing in Modern Detectors

Ziwei Liu

Multimedia Lab (MMLAB)
The Chinese University of Hong Kong

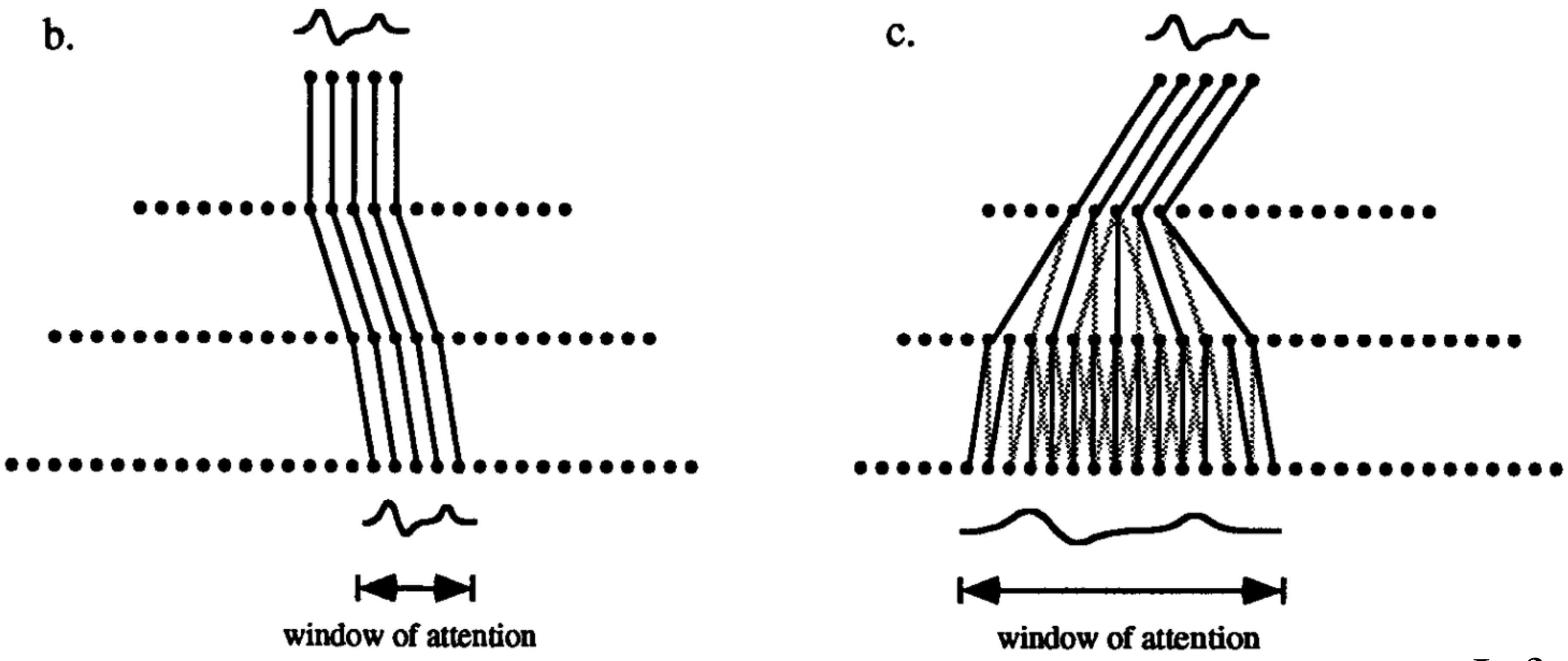
Dynamic Forwarding



- Content-Aware
- Resolution-Adaptive

A neurobiological model of visual attention and invariant pattern recognition based on dynamic routing of information

Dynamic Routing



- Information Flow
- Selection & Fusion

A neurobiological model of visual attention and invariant pattern recognition based on dynamic routing of information

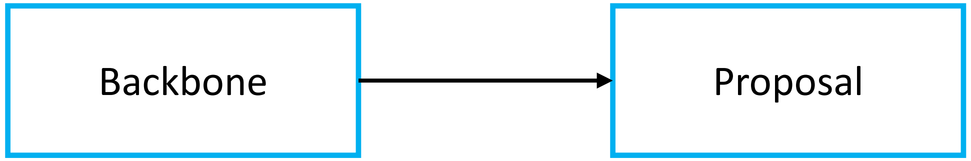
Overview

1. We proposed a new backbone **FishNet**. (NIPS 2018)

Backbone

Overview

1. We proposed a new backbone **FishNet**. (NIPS 2018)
2. We designed a **feature guided anchoring** scheme to improve the average recall (AR) of RPN by 10 points. (CVPR 2019)

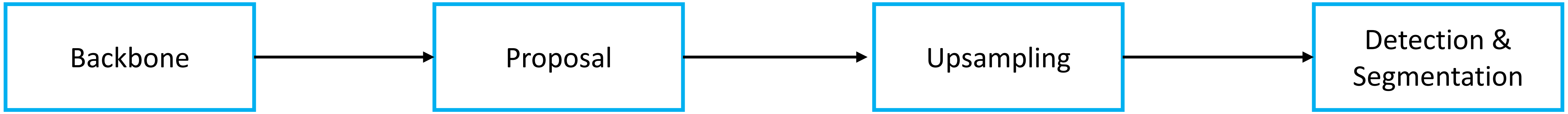


Overview

1. We proposed a new backbone **FishNet**. (NIPS 2018)
2. We designed a **feature guided anchoring** scheme to improve the average recall (AR) of RPN by 10 points. (CVPR 2019)
3. We proposed a new upsampling operator **CARAFE**. (ICCV 2019)

Overview

1. We proposed a new backbone **FishNet**. (NIPS 2018)
2. We designed a **feature guided anchoring** scheme to improve the average recall (AR) of RPN by 10 points. (CVPR 2019)
3. We proposed a new upsampling operator **CARAFE**. (ICCV 2019)
4. We developed a **hybrid cascading and branching** pipeline for detection and segmentation. (CVPR 2019)



香港中文大學
The Chinese University of Hong Kong

FishNet: A Versatile Backbone for Image, Region, and Pixel Level Prediction (NIPS 2018)

FishNet

Motivation

- The basic principles for designing CNN for region and pixel level tasks are **diverging** from the principles for image classification.
- Unify the advantages of networks designed for region and pixel level tasks in obtaining **deep** features with **high-resolution**.

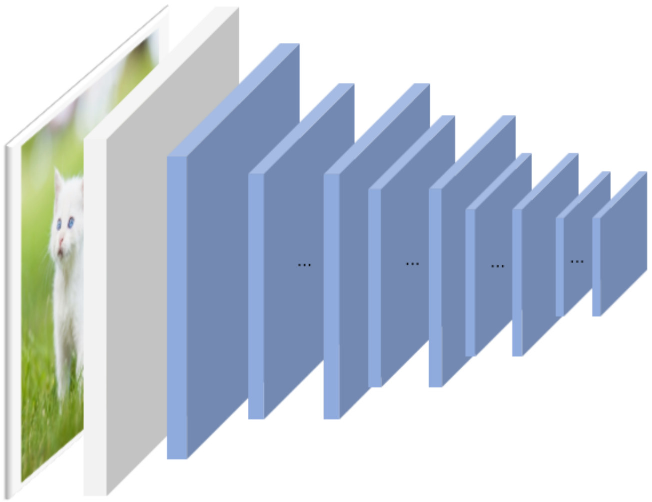
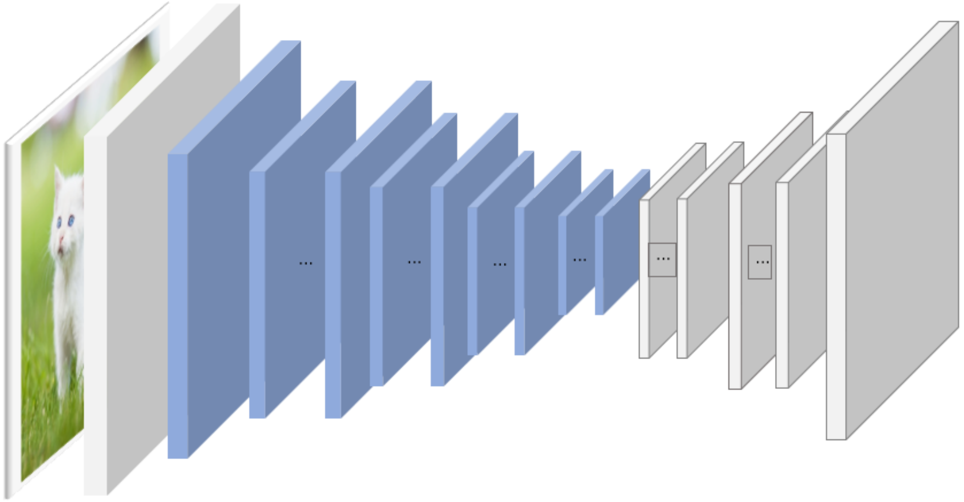


Image classification



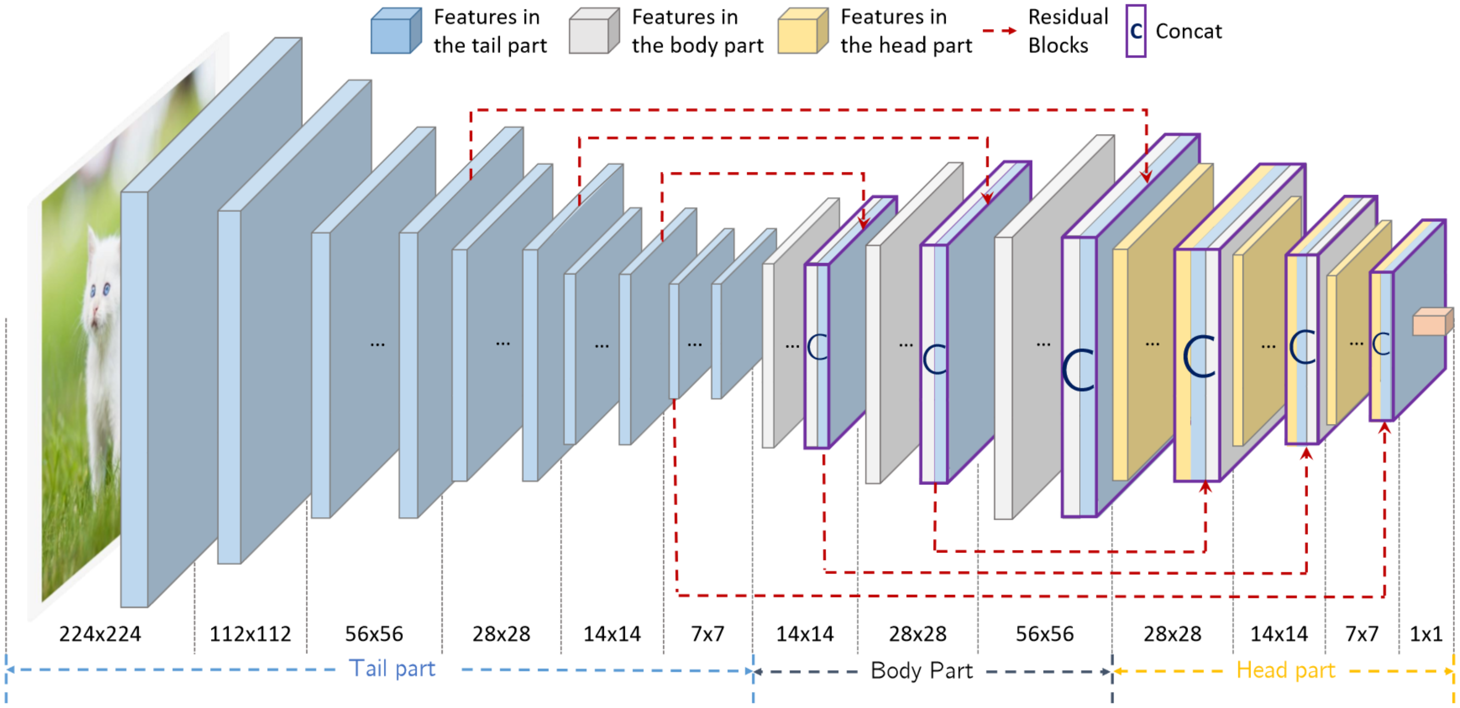
Region and pixel level tasks

Segmentation, pose estimation, detection ...

FishNet

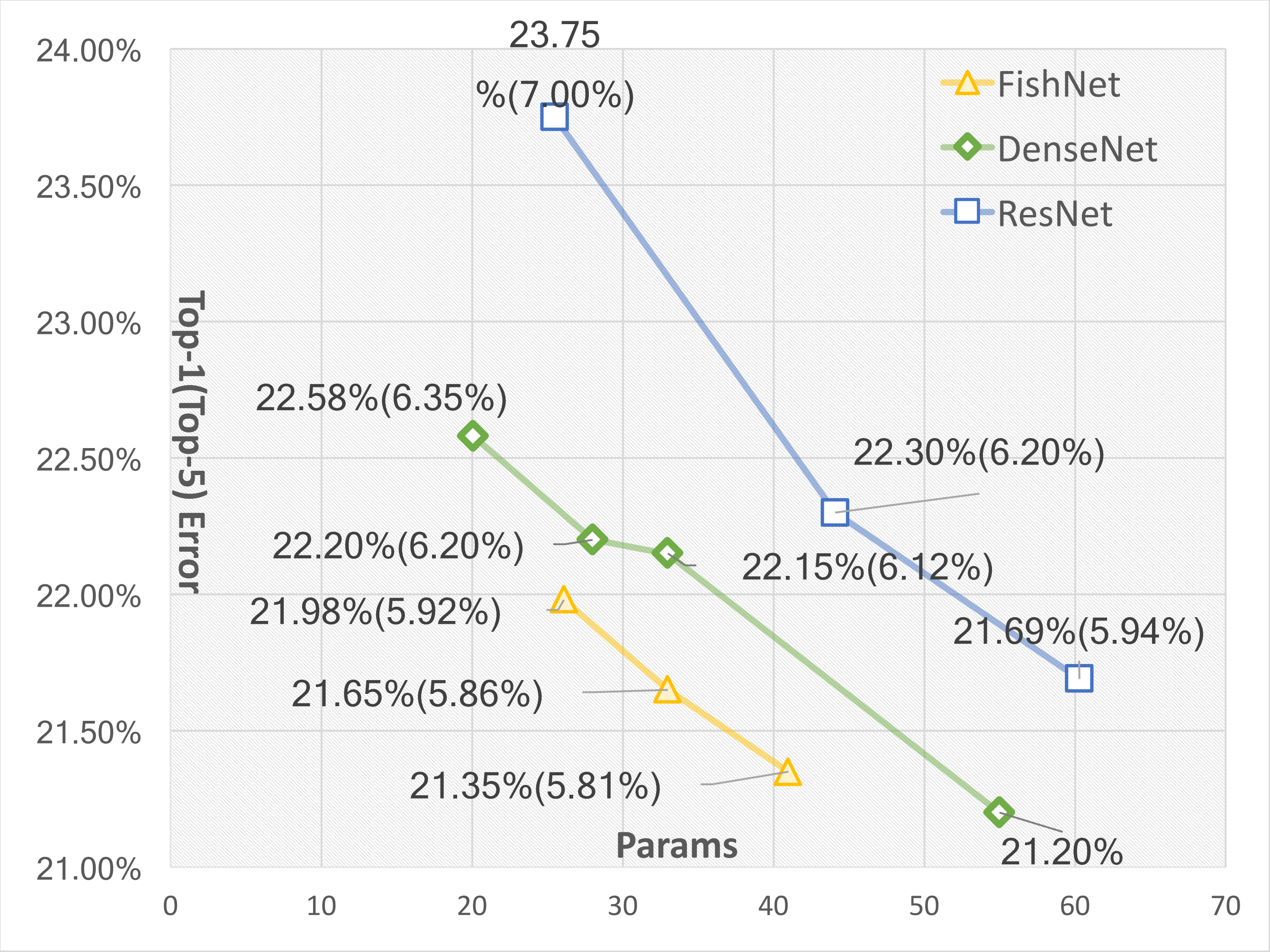
Motivation

- Traditional consecutive down-sampling will prevent the very shallow layers to be directly connected till the end, which may exacerbate the **vanishing gradient problem.**
- Features from varying depths could be used for **refining** each other.



FishNet: A Versatile Backbone for Image, Region, and Pixel Level Prediction, NIPS 2018.

FishNet



Top-1 Classification Error on ImageNet

FishNet

MS COCO *val-2017* detection and instance segmentation results.

	Instance Segmentation	Object Detection
Backbone	$AP^s/AP^s_S/AP^s_M/AP^s_L$	$AP^d/AP^d_S/AP^d_M/AP^d_L$
ResNet-50 [3]	34.5/15.6/37.1/52.1	38.6/22.2/41.5/50.8
ResNet-50†	34.7/18.5/37.4/47.7	38.7/22.3/42.0/51.2
ResNeXt-50 (32x4d)†	35.7/19.1/38.5/48.5	40.0/23.1/43.0/52.8
FishNet-188	**37.0**/19.8/40.2/50.3	**41.5**/24.1/44.9/55.0
vs. ResNet-50†	**+2.3/+1.3/+2.8/+2.6**	**+2.8/+1.8/+2.9/+3.8**
vs. ResNeXt-50†	**+1.3/+0.7/+1.7/+1.8**	**+1.5/+1.0/+1.9/+2.2**

FishNet

- Fish tail, fish body, fish head
- More flexible information flow
- Adaptive feature resolution reservation

香港中文大學
The Chinese University of Hong Kong

Region Proposal by Guided Anchoring (CVPR 2019)

Overview

香港中文大學
The Chinese University of Hong Kong

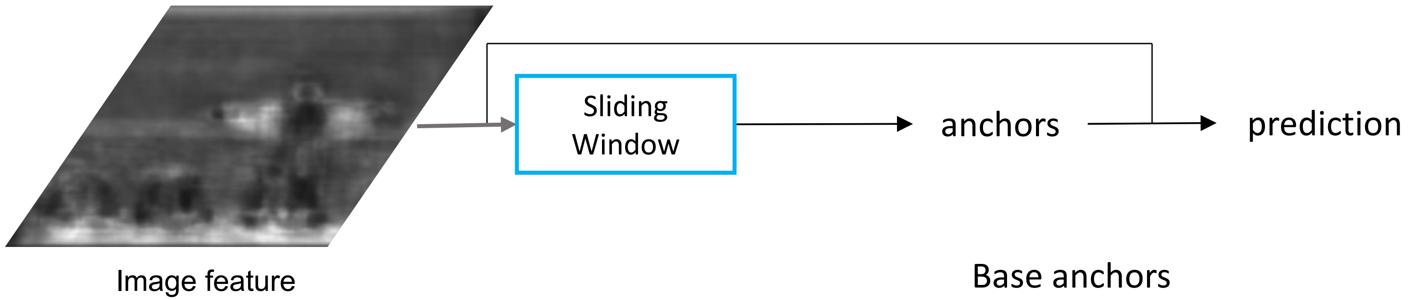
- We introduce a Guided Anchoring Scheme to generate anchors and build up a Guided Anchoring Region Proposal Network (GA-RPN)

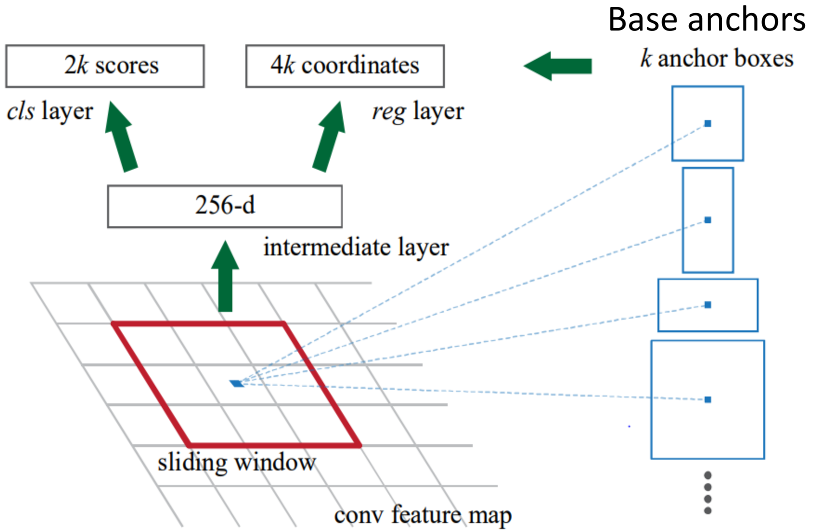
- GA-RPN achieves 9.1% higher average recall (AR) on MS COCO with 90% fewer anchors than the RPN baseline.

- GA-RPN improves Fast R-CNN, Faster R-CNN and RetinaNet by over 2.2%, 2.7% and 1.2%.

Baseline

Region Proposal Network (RPN)





RPN

RPN adopts a ***uniform*** anchoring scheme which ***uniformly*** generates anchors with ***predefined scales*** and ***aspect ratios*** over the whole image.

Ren S, He K, Girshick R, et al. Faster r-cnn: Towards real-time object detection with region proposal networks[C]//Advances in neural information processing systems. 2015: 91-99.

Baseline

Uniform anchoring scheme has intrinsic drawbacks:

- Most of generated anchors are irrelevant to the objects. (less than 0.01% anchors are positive samples)
- The conventional method are unaware of object shapes.

Baseline

How to overcome such drawbacks:

- Anchors should be distributed on feature maps considering how likely the locations contain objects.

- Anchor shapes should be predicted rather than pre-defined.

Guided Anchoring

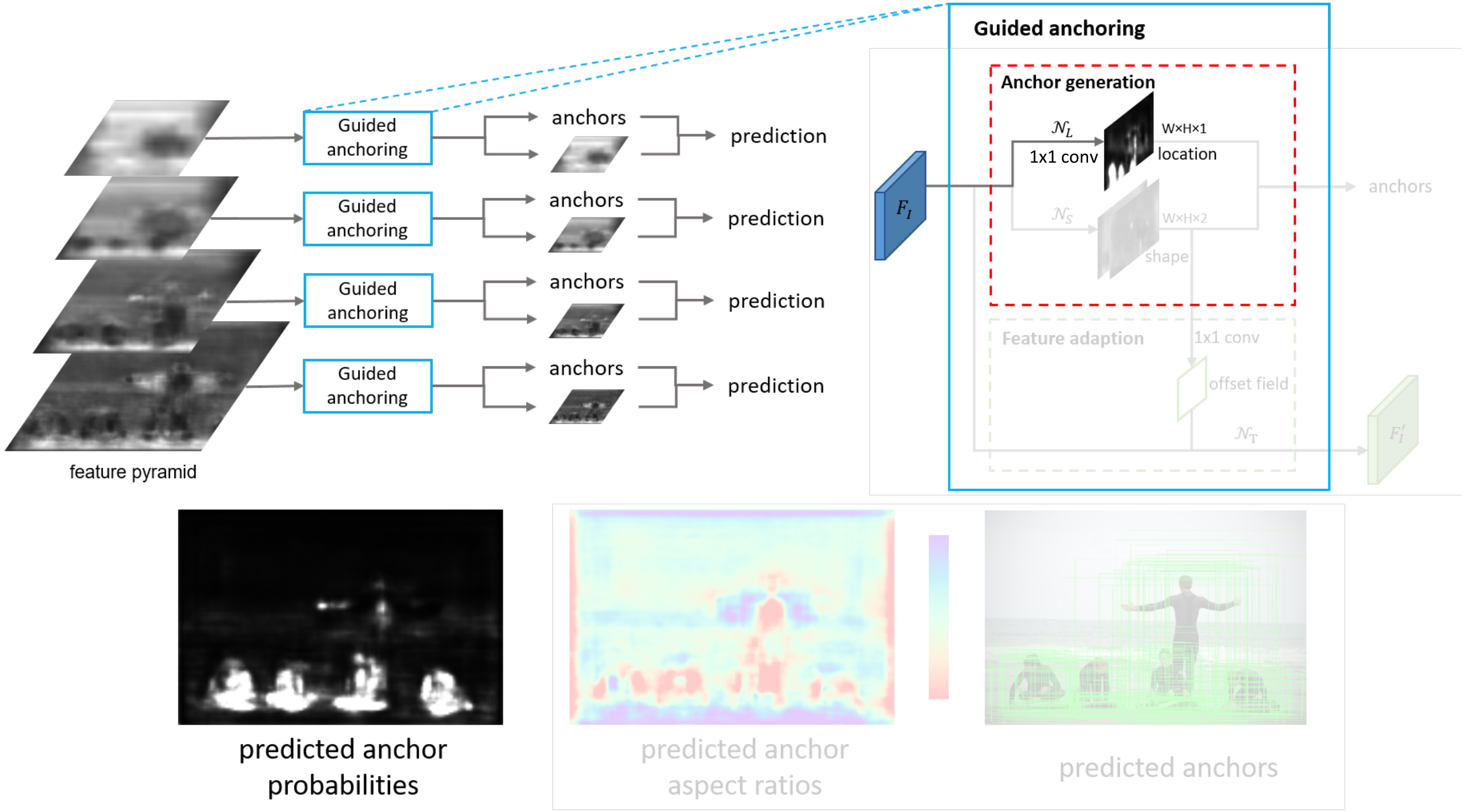
Guided Anchoring Component has following steps:

- The first step identifies the locations where objects are likely to exist.
- The second stage predicts shapes of anchors.
- In addition, we further introduce a feature adaption module to refine the features considering anchor shapes.

Guided Anchoring

香港中文大學
The Chinese University of Hong Kong

Anchor Location Prediction



Guided Anchoring

香港中文大學 The Chinese University of Hong Kong

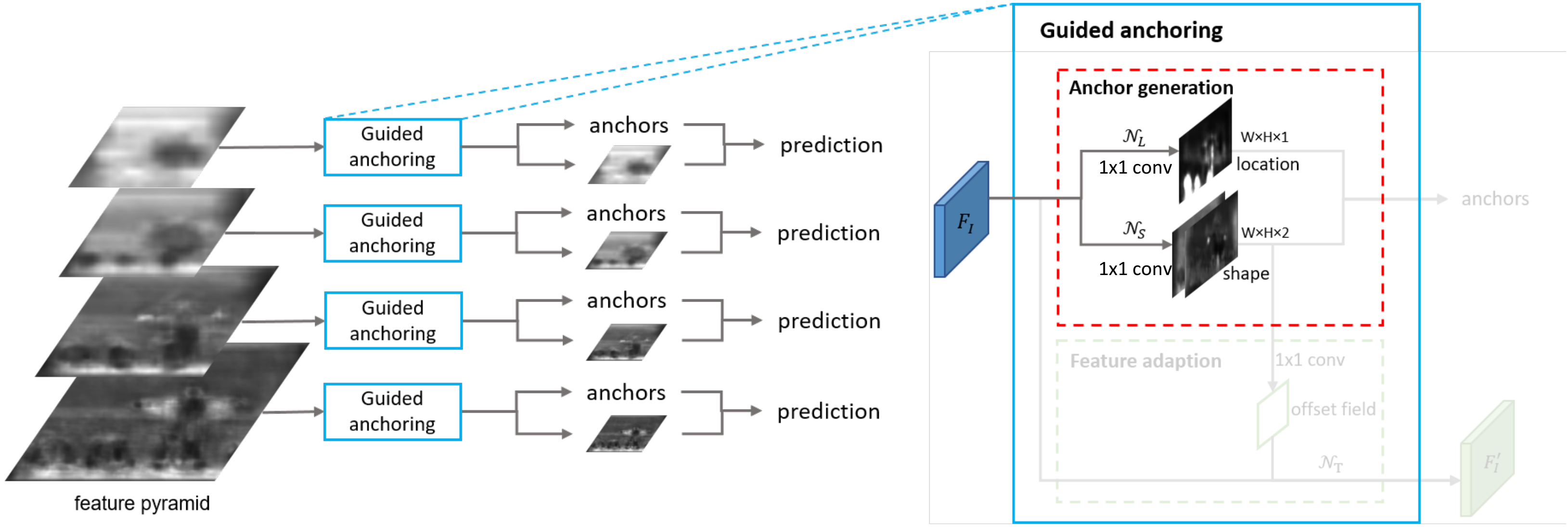
Guided Anchoring Component has following steps:

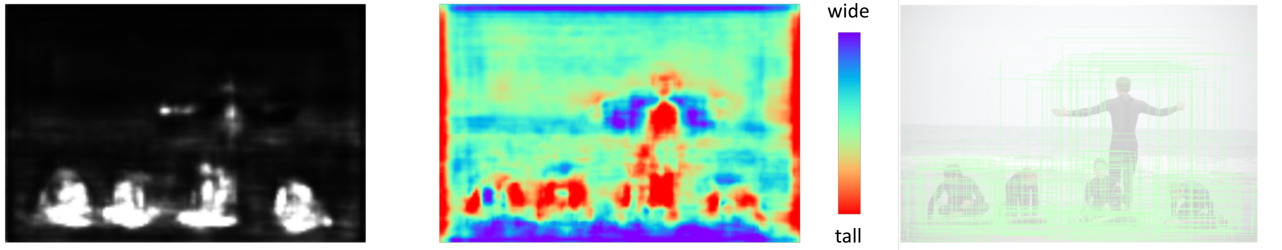
- The first step identifies the locations where objects are likely to exist.
- The second stage predicts shapes of anchors.
- In addition, we further introduce a feature adaption module to refine the features considering anchor shapes.

Guided Anchoring

香港中文大學 The Chinese University of Hong Kong

Anchor Shape Prediction





predicted anchor probabilities

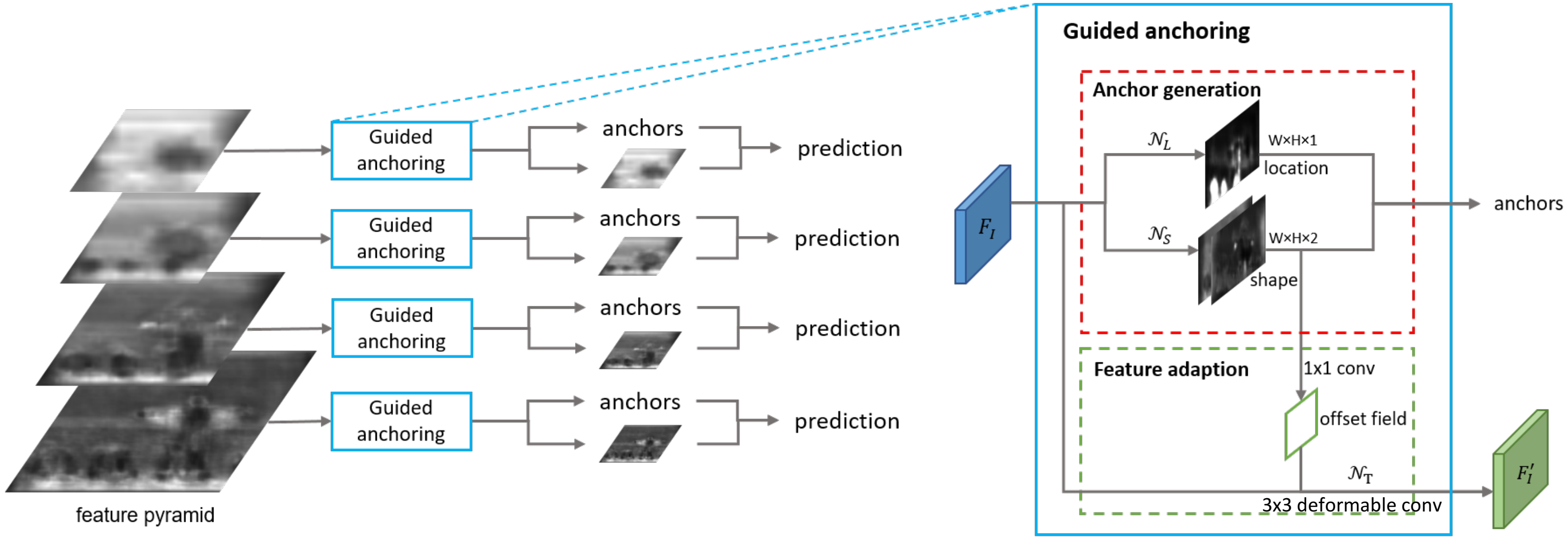
predicted anchor aspect ratios

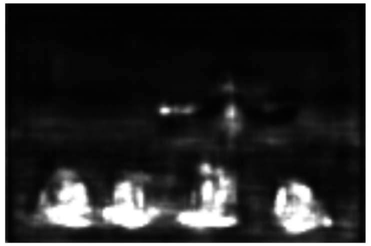
predicted anchors

Guided Anchoring

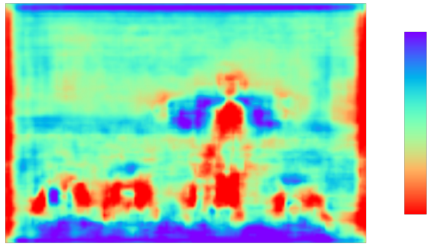
香港中文大學
The Chinese University of Hong Kong

Feature Adaption





predicted anchor probabilities



predicted anchor aspect ratios

predicted anchors

Guided Anchoring

香港中文大學 The Chinese University of Hong Kong

Why feature adaptive?

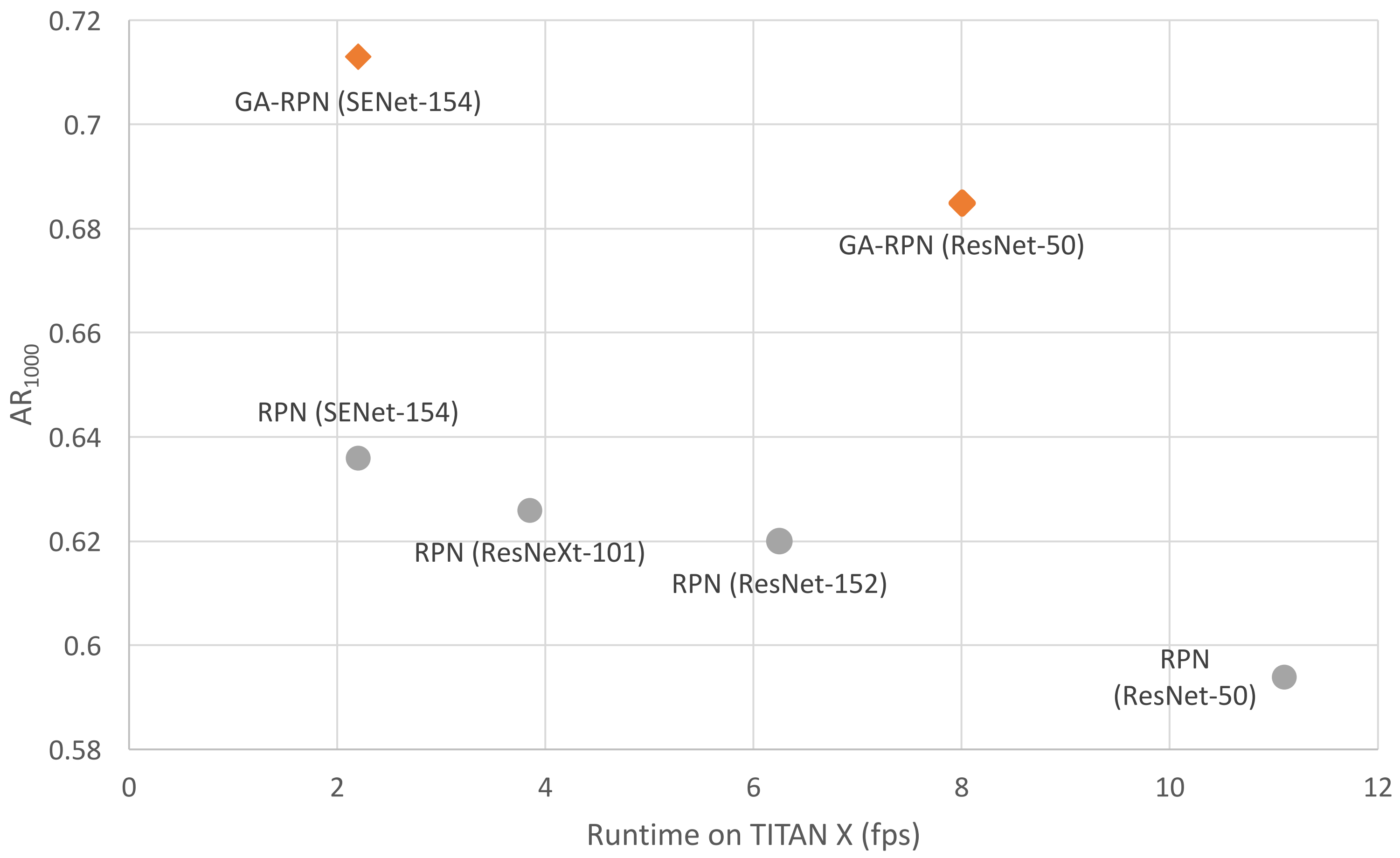
A feature and an anchor on the same location should be consistent.

Method	AR_{100}	AR_{300}	AR_{1000}	AR_S	AR_M	AR_L
RPN	47.5	54.7	59.4	31.7	55.1	64.6
GA-RPN w/o F.A.	54.0	60.1	63.8	36.7	63.1	71.5
GA-RPN + F.A.	59.2	65.2	68.5	40.9	67.8	79.0

Guided Anchoring

香港中文大學
The Chinese University of Hong Kong

Experiment Results



Guided Anchoring

香港中文大學
The Chinese University of Hong Kong

Experiment Results

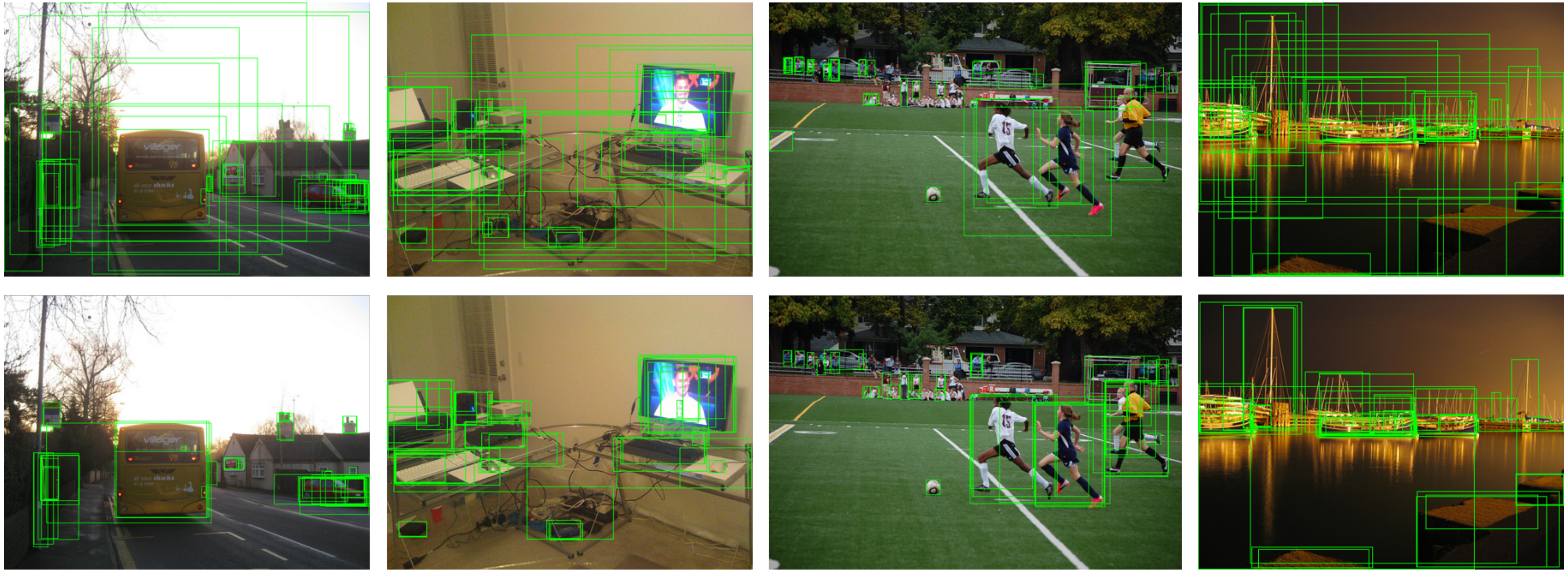
Detector	AP	AR_{50}	AP_{75}	AP_S	AP_M	AP_L
Fast R-CNN	37.1	59.6	39.7	20.7	39.5	47.1
GA-Fast-RCNN	**39.4**	59.4	42.8	21.6	41.9	50.4
Faster R-CNN	37.1	59.1	40.1	21.3	39.8	46.5
GA-Faster-RCNN	**39.8**	59.2	43.5	21.8	42.6	50.7
RetinaNet	35.9	55.4	38.8	19.4	38.9	46.5
GA-RetinaNet	**37.1**	56.9	40.0	20.1	40.1	48.0

Detection results on MS COCO 2017 test-dev with ResNet-50 backbone

Guided Anchoring

香港中文大學
The Chinese University of Hong Kong

Examples



RPN

GA-RPN

Guided Anchoring

- From sliding window to sparse, non-uniform distribution
- From predefined shapes to learnable, arbitrary shapes
- Refine features based on anchor shapes

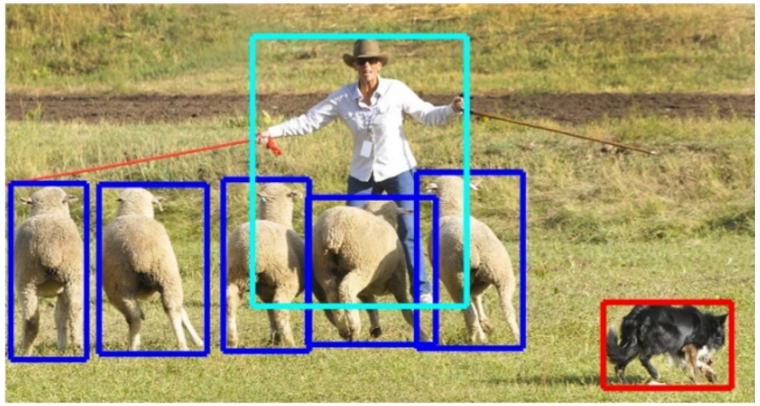
香港中文大學
The Chinese University of Hong Kong

CARAFE: **C**ontent-**A**ware **R**e**A**ssembly of **F**e**a**tures (ICCV 2019 Oral)

Background

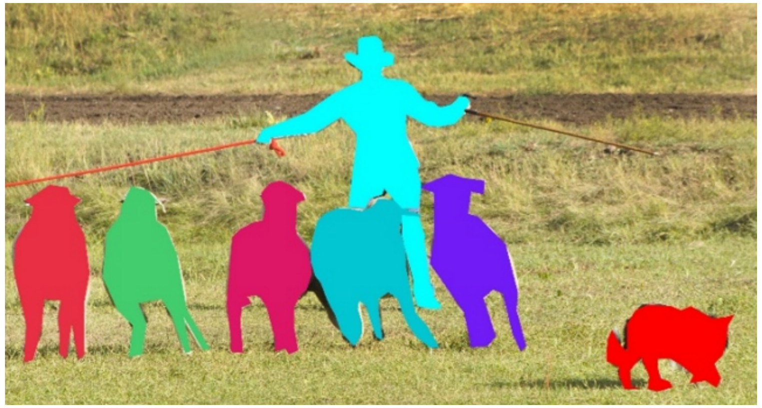
香港中文大學
The Chinese University of Hong Kong

- Feature upsampling is a key operation in a number of modern convolutional network architectures, e.g. Feature Pyramids Networks, U-Net, Stacked Hourglass Networks.
- Its design is critical for dense prediction tasks such as object detection and semantic/instance segmentation.



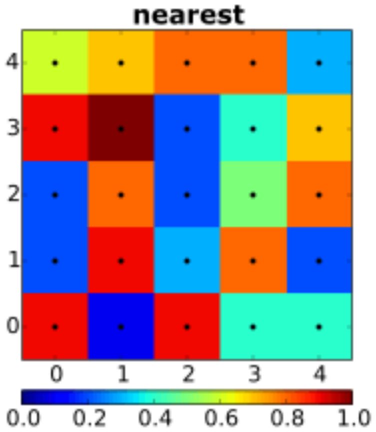
Object detection

Semantic segmentation

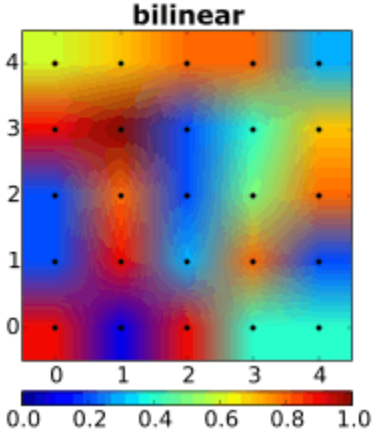


Instance segmentation

Background



Nearest Neighbor (NN)

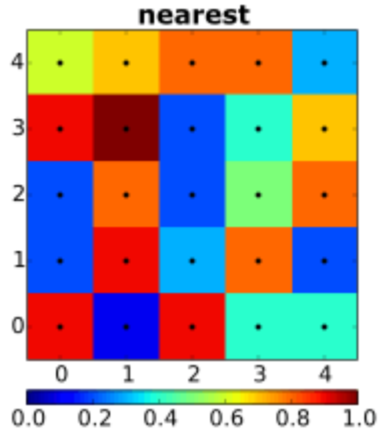


Bilinear

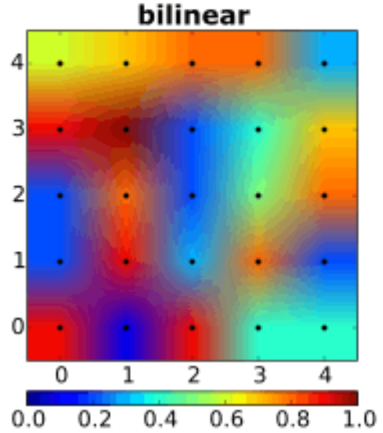
Interpolations leverage distances to measure the correlations between pixels, and hand-crafted upsampling kernels are used.
(Pros: low cost / Cons: hand-crafted upsampling kernels)

Background

香港中文大學
The Chinese University of Hong Kong



Nearest Neighbor (NN)

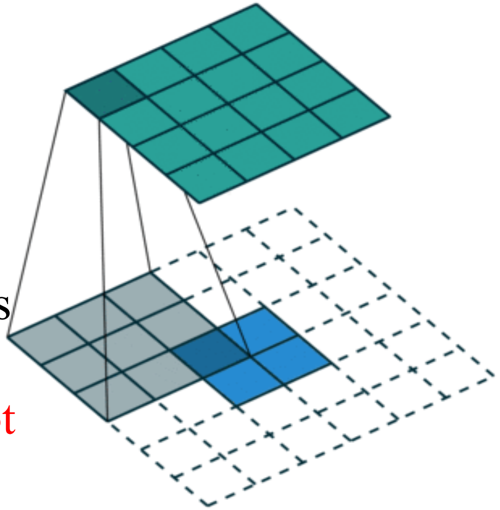


Bilinear

Interpolations leverage distances to measure the correlations between pixels, and hand-crafted upsampling kernels are used.
(Pros: low cost / Cons: hand-crafted upsampling kernels)

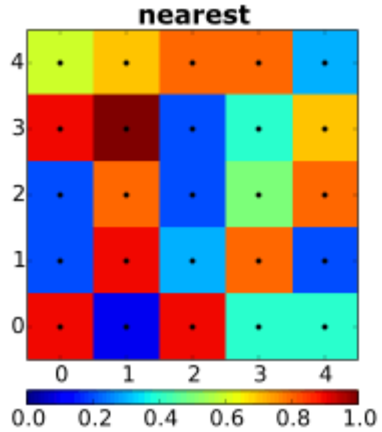
Deconvolution
(Transposed Convolution)

Deconvolution is an inverse operator of a convolution, which uses a fixed kernel for all samples within a limited receptive field.
(Pros: learnable kernel / Cons: not content-aware, limited receptive field)

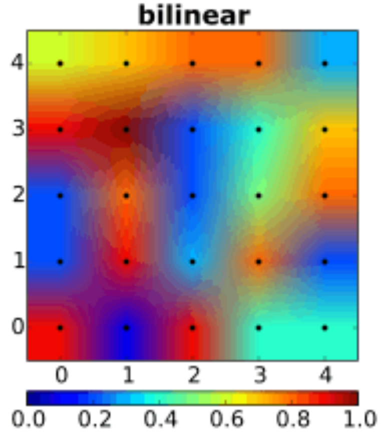


Background

香港中文大學
The Chinese University of Hong Kong



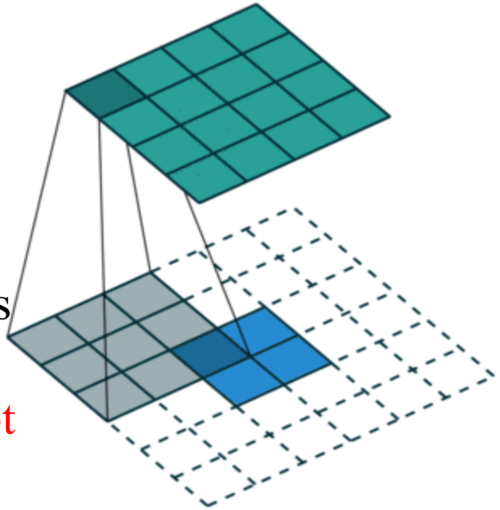
Nearest Neighbor (NN)



Bilinear

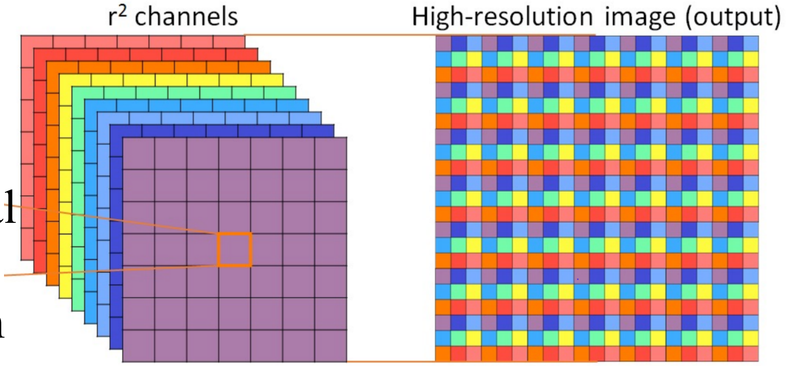
Interpolations leverage distances to measure the correlations between pixels, and hand-crafted upsampling kernels are used.
(Pros: low cost / Cons: hand-crafted upsampling kernels)

Deconvolution (Transposed Convolution)



Deconvolution is an inverse operator of a convolution, which uses a fixed kernel for all samples within a limited receptive field.
(Pros: learnable kernel / Cons: not content-aware, limited receptive field)

Pixel Shuffle



Pixel Shuffle reshapes depth on the channel space into width and height on the spatial space. It brings highly computational overhead when expanding the channel space.

(Pros: learnable kernel/ Cons: not content-aware, limited receptive field, high cost)

Overview

Content-**A**ware **R**e**A**ssembly of **FE**atures (CARAFE) is a universal, lightweight and highly effective upsampling operator.

- **Large field of view**. CARAFE can aggregate contextual information within a large receptive field.
- **Content-aware handling.** CARAFE enables instance-specific content-aware handling, which generates adaptive kernels on-the-fly.
- **Lightweight and fast to compute.** CARAFE introduces little computational overhead and can be readily integrated into modern network architectures

Overview

Content-**A**ware **R**e**A**ssembly of **FE**atures (CARAFE) is a universal, lightweight and highly effective upsampling operator.

- **Large field of view**. CARAFE can aggregate contextual information within a large receptive field.
- **Content-aware handling.** CARAFE enables instance-specific content-aware handling, which generates adaptive kernels on-the-fly.
- **Lightweight and fast to compute.** CARAFE introduces little computational overhead and can be readily integrated into modern network architectures

CARAFE shows consistent and substantial gains across **object detection, instance/semantic segmentation** and **inpainting** (1.2%, 1.3%, 1.8%, 1.1db respectively) with negligible computational overhead.

CARAFE

香港中文大學
The Chinese University of Hong Kong

On each location, CARAFE can leverage the **content information** of such location to predict **assembly kernels** and **assemble the features** inside a predefined nearby region.

1) The first step is to predict a reassembly kernel for each destination location according to its content. ($N(\mathcal{X}_l, k)$ is the *k* x *k* sub-region of χ centered at the location l, i.e., the neighbor of X_l.)

$$\mathcal{W}_{l'} = \psi(N(\mathcal{X}_l, k_{encoder})).$$

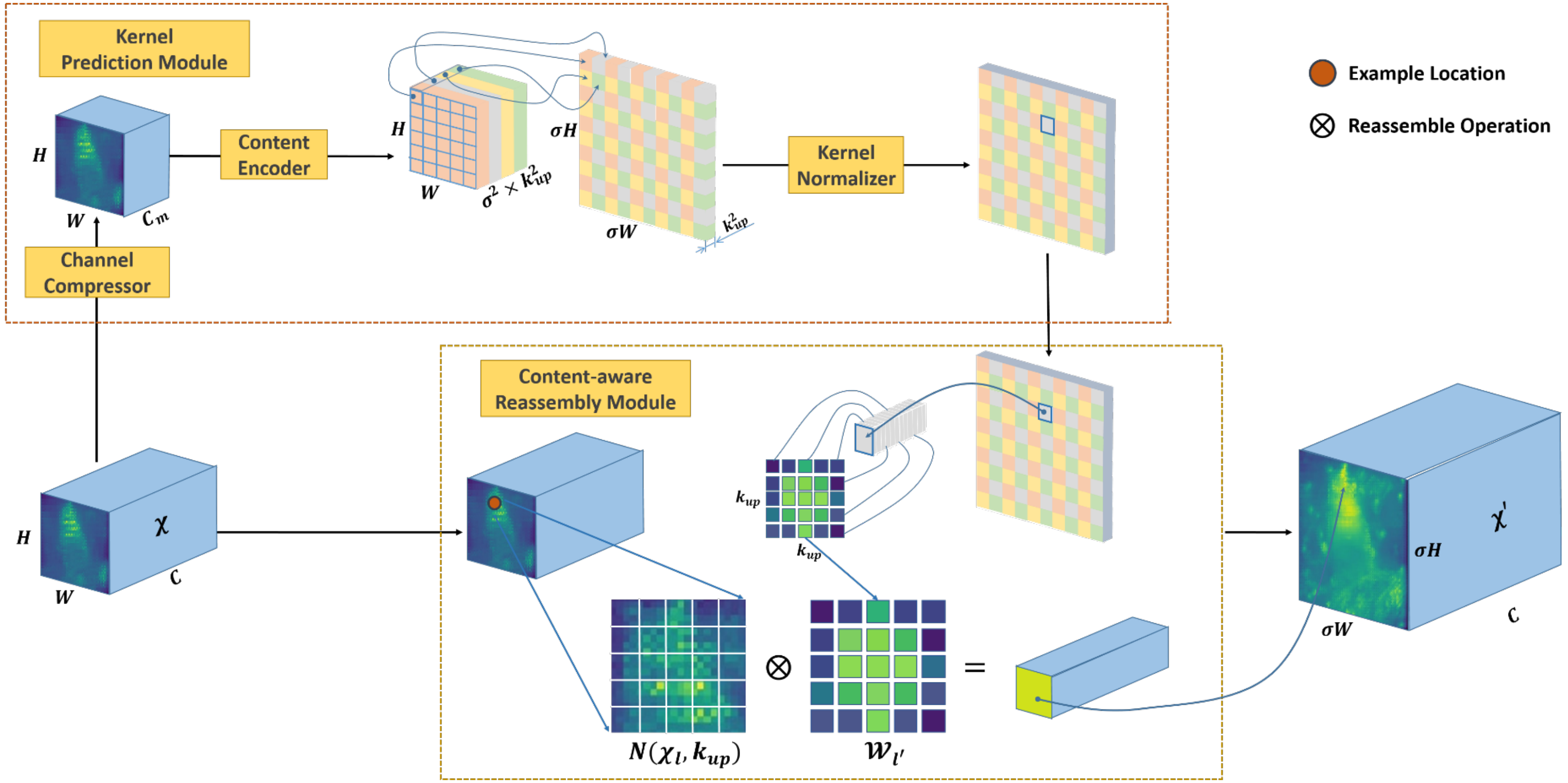
2) The second step is to reassemble the features with predicted kernels.

$$\mathcal{X}'_{l'} = \phi(N(\mathcal{X}_l, k_{up}), \mathcal{W}_{l'}).$$

CARAFE

香港中文大學
The Chinese University of Hong Kong

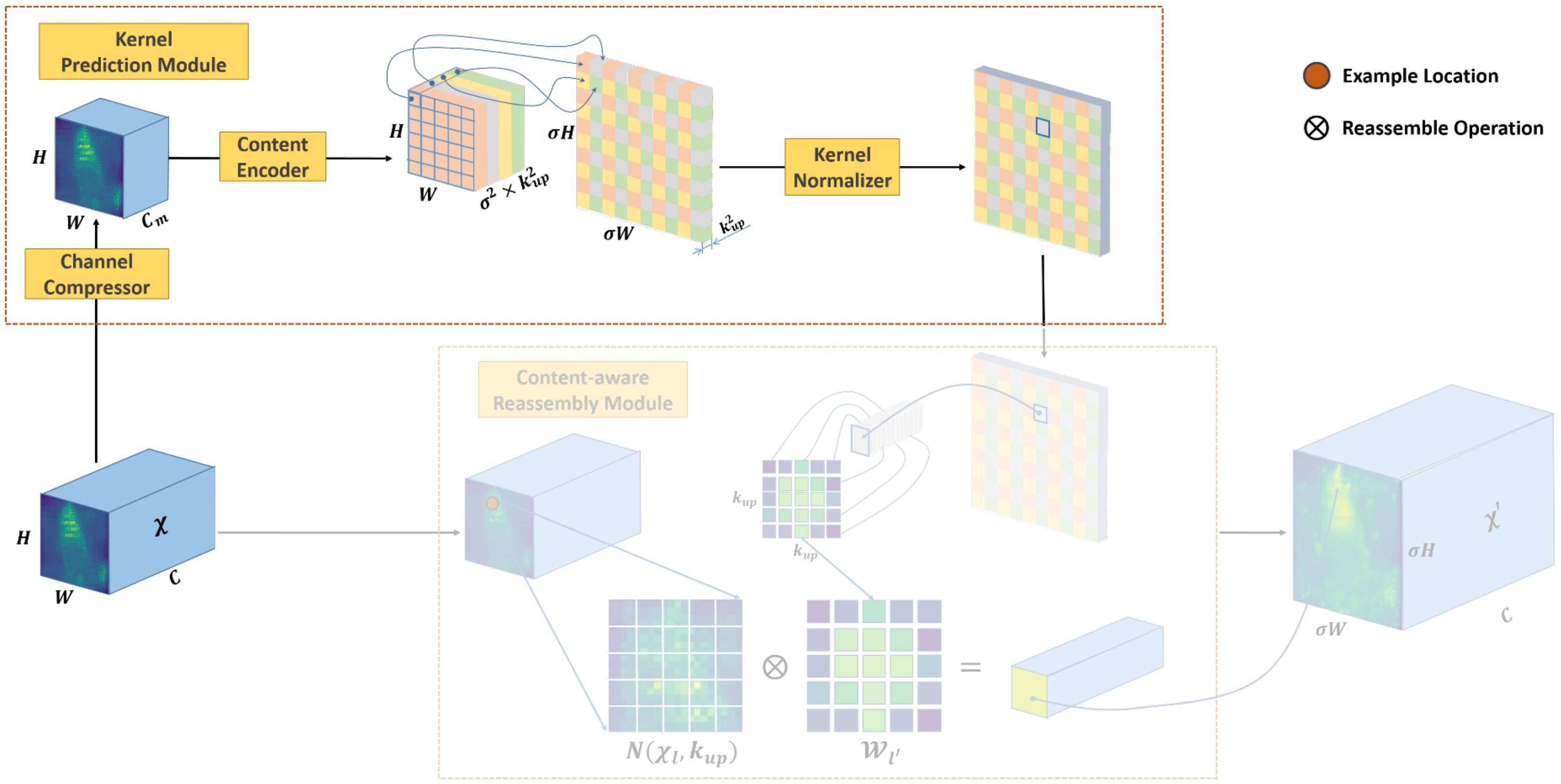
Framework



CARAFE

香港中文大學 The Chinese University of Hong Kong

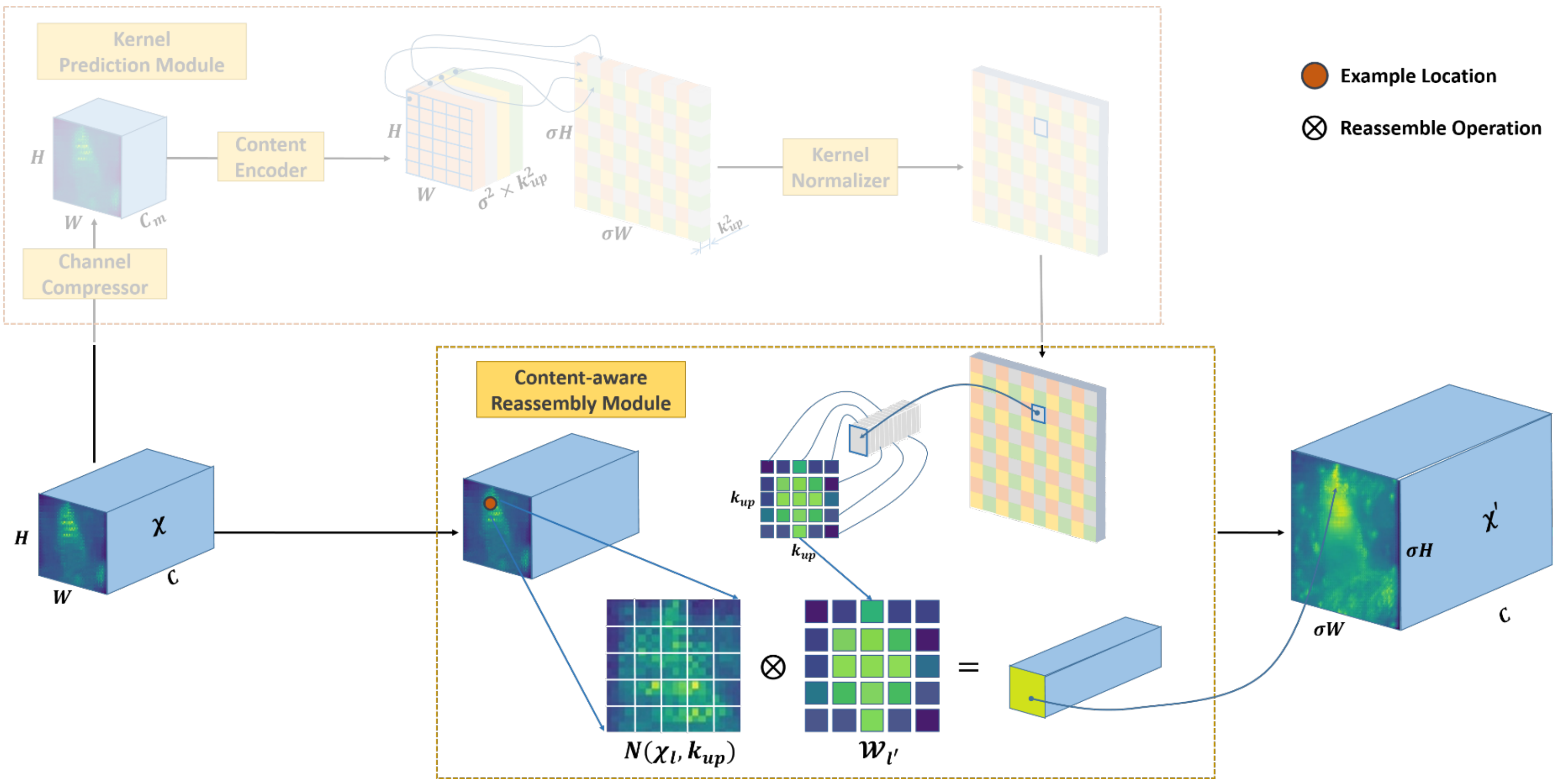
Kernel Predication Module



CARAFE

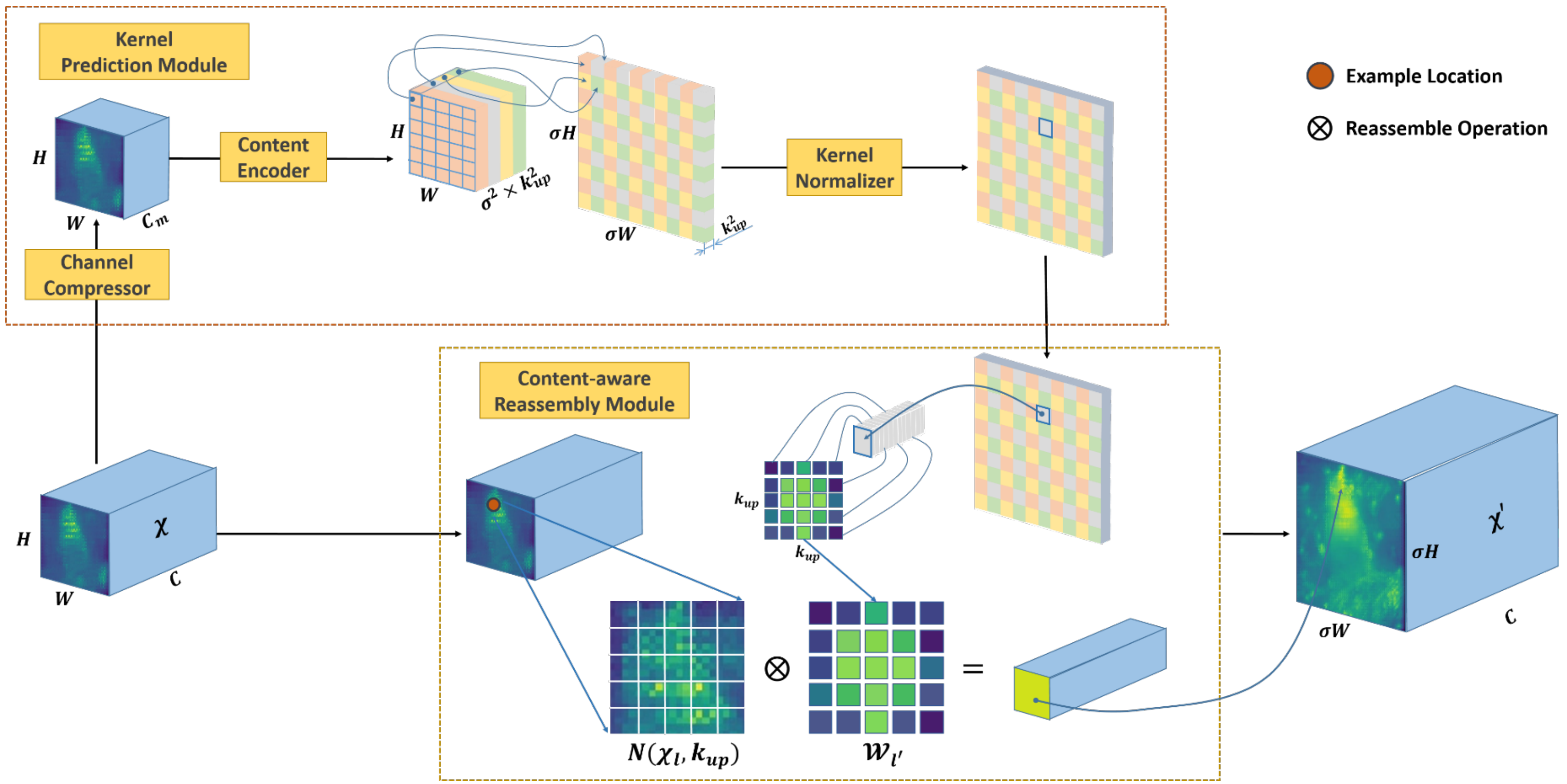
香港中文大學
The Chinese University of Hong Kong

Content-aware Reassembly Module



CARAFE

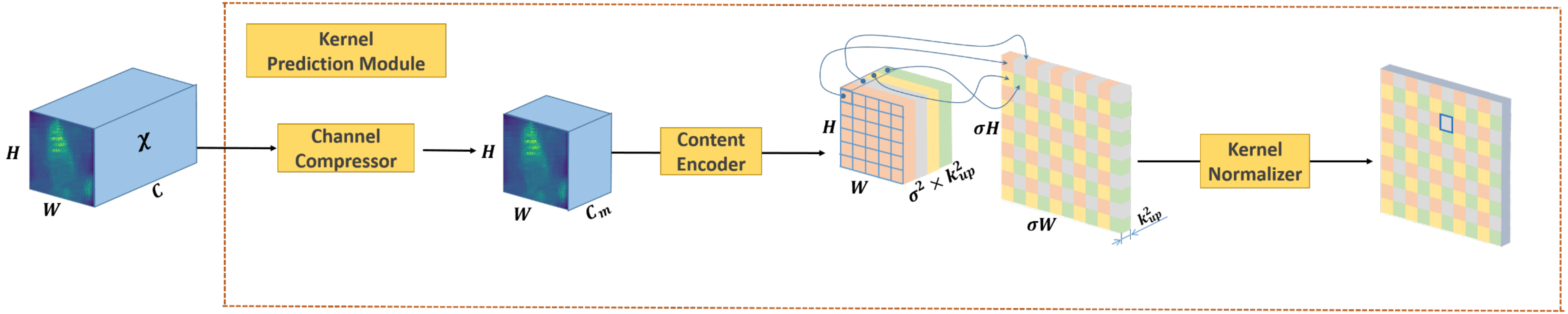
香港中文大學 The Chinese University of Hong Kong



- Each source location on χ corresponds to σ^2 destination locations on χ'.
- Each destination location on χ' requires a k_{up}x k_{up} reassembly kernel. (k_{up} is the reassembly kernel size.)

CARAFE

Kernel Predication Module

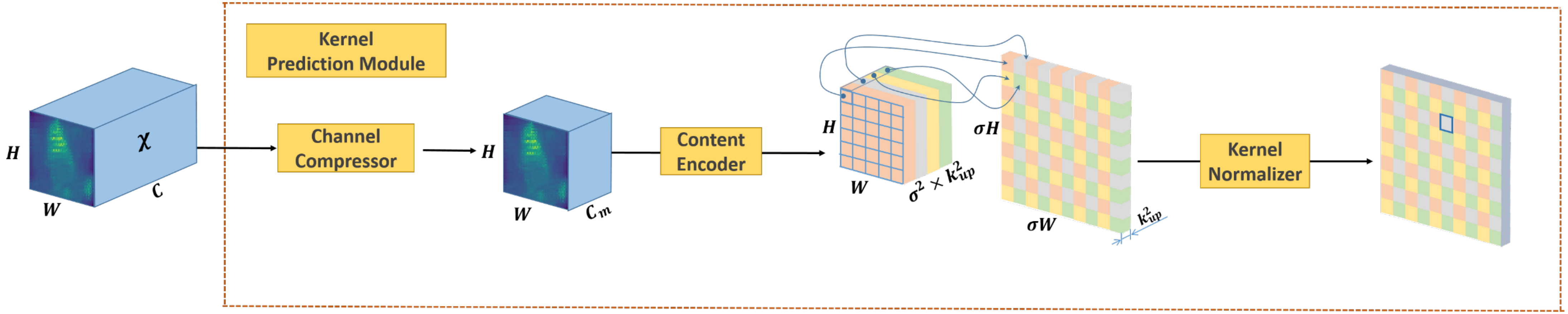


1) **Channel Compressor.** (1 x 1 convolution layer which compresses the input feature channel from C to Cm. The goal of this step is for speed-up without harming the performance.)

CARAFE

香港中文大學 The Chinese University of Hong Kong

Kernel Predication Module

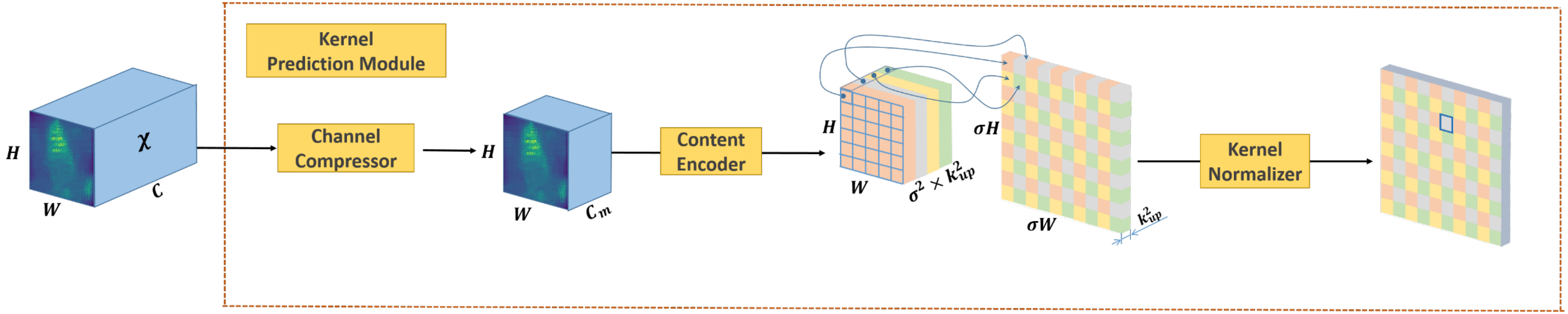


1) **Channel Compressor.** (1 x 1 convolution layer which compresses the input feature channel from C to Cm. The goal of this step is for speed-up without harming the performance.)
2) **Content Encoder.** (Convolution layer of kernel size $k_{encoder}$ to generate reassembly kernels based on the content of input features. An empirical formula $k_{encoder} = k_{up} - 2$ is a good trade-off between performance and efficiency through our study)

CARAFE

香港中文大學
The Chinese University of Hong Kong

Kernel Predication Module

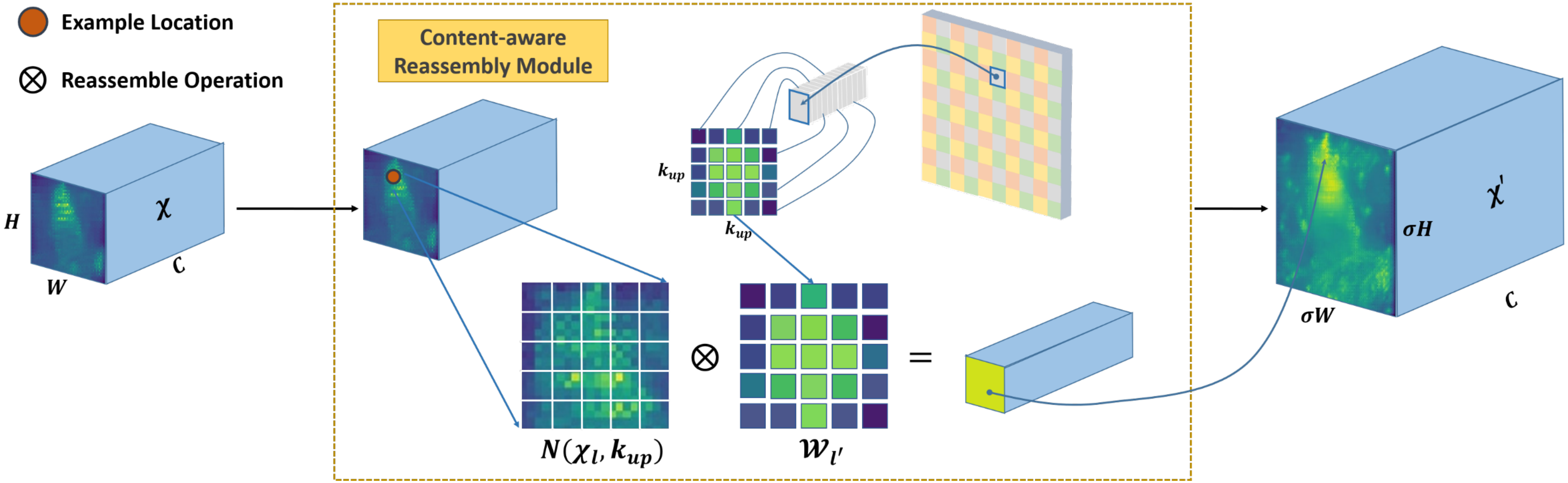


1) **Channel Compressor.** (1 x 1 convolution layer which compresses the input feature channel from C to Cm. The goal of this step is for speed-up without harming the performance.)
2) **Content Encoder.** (Convolution layer of kernel size $k_{encoder}$ to generate reassembly kernels based on the content of input features. An empirical formula $k_{encoder} = k_{up} - 2$ is a good trade-off between performance and effificiency through our study)
3) **Kernel Normalizer.** (Each k_{up} x k_{up} reassembly kernel is normalized with a softmax function.)

CARAFE

香港中文大學
The Chinese University of Hong Kong

Content-aware Reassembly Module



$$\mathcal{X}'_{l'} = \sum_{n=-r}^{r} \sum_{m=-r}^{r} \mathcal{W}_{l'(n,m)} \cdot \mathcal{X}_{(i+n,j+m)}.$$

Applications

香港中文大學
The Chinese University of Hong Kong

CARAFE introduces little computational overhead and can be readily integrated into modern network architectures.

- **Object Detection (Faster R-CNN w/ FPN)**
- **Instance Segmentation (Mask R-CNN w/ FPN)**
- **Semantic Segmentation (UperNet)**
- **Image Inpainting (Global&Local, Partial Conv)**

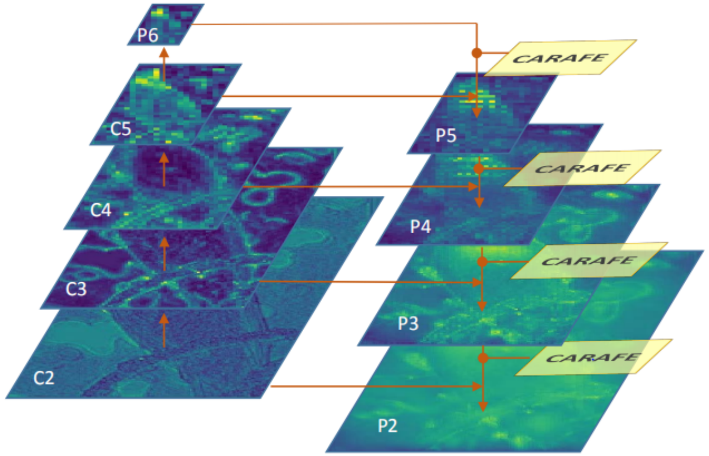
Applications

香港中文大學
The Chinese University of Hong Kong

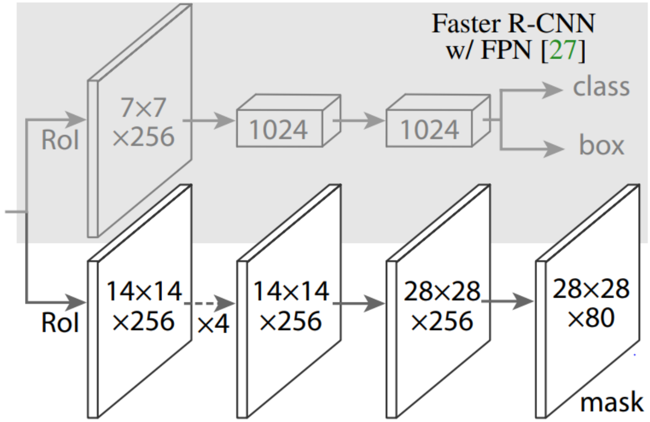
Object Detection & Instance Segmentation

1) **Feature Pyramid Network (Faster R-CNN，Mask R-CNN)**

2) **Mask Head (Mask R-CNN)**



Feature Pyramid Network (FPN)



Mask Head

Experiments

香港中文大學 The Chinese University of Hong Kong

Object Detection & Instance Segmentation:

Table 1: Detection and Instance Segmentation results on MS COCO 2018 *test-dev*.

Method	Backbone	Task	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
Faster R-CNN	ResNet-50	BBox	36.9	59.1	39.7	21.5	40.0	45.6
Faster R-CNN w/ CARAFE	ResNet-50	BBox	**38.1**	**60.7**	**41.0**	**22.8**	**41.2**	**46.9**
Mask R-CNN	ResNet-50	BBox	37.8	59.7	40.8	22.2	40.7	46.8
	ResNet-50	Segm	34.6	56.5	36.8	18.7	37.3	45.1
Mask R-CNN w/ CARAFE	ResNet-50	BBox	**38.8**	**61.2**	**42.1**	**23.2**	**41.7**	**47.9**
	ResNet-50	Segm	**35.9**	**58.1**	**38.2**	**19.8**	**38.6**	**46.5**

Experiments

Semantic Segmentation:

Table 5: Semantic Segmentation results on ADE20k val. Single scale testing is used in our experiments.

Method	Backbone	mIoU	P.A.
PSPNet	ResNet-50	41.68	80.04
PSANet	ResNet-50	41.92	80.17
UperNet[3]	ResNet-50	40.44	79.80
UperNet w/ CARAFE	ResNet-50	**42.23**	**80.34**

Experiments

香港中文大學 The Chinese University of Hong Kong

Image Inpainting:

Table 7: Image inpainting results on Places val.

Method	L1(%)	PSNR(dB)
Global&Local	6.78	19.58
Partial Conv	5.96	20.78
Global&Local w/ CARAFE	6.00	20.71
Partial Conv w/ CARAFE	**5.72**	**20.98**

Experiments

Compare with previous upsamplers:

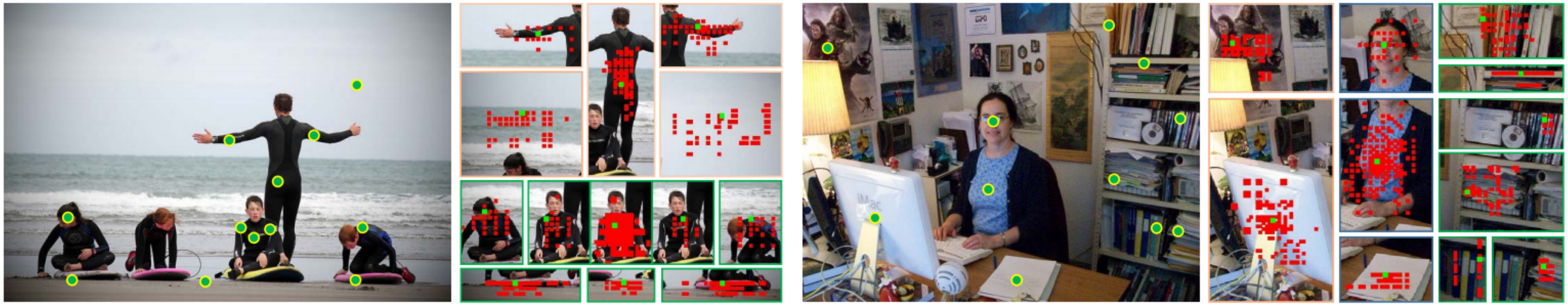
Table 2: Detection results with Faster RCNN. Various upsampling methods are used in FPN.

Method	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L	FLOPs
Nearest	36.5	58.4	39.3	21.3	40.3	47.2	0
Bilinear	36.7	58.7	39.7	21.0	40.5	47.5	8k
Nearest + Conv	36.6	58.6	39.5	21.4	40.3	46.4	4.7M
Bilinear + Conv	36.6	58.7	39.4	21.6	40.6	46.8	4.7M
Deconv [21]	36.4	58.2	39.2	21.3	39.9	46.5	1.2M
Pixel Shuffle[25]	36.5	58.8	39.1	20.9	40.4	46.7	4.7M
GUM[18]	36.9	58.9	39.7	21.5	40.6	48.1	1.1M
S.A.[1]	36.9	58.8	39.8	21.7	40.8	47.0	28k
CARAFE	**37.8**	**60.1**	**40.8**	**23.1**	**41.7**	**48.5**	199k

Experiments

香港中文大學
The Chinese University of Hong Kong

How CARAFE works:



(a) (b)

Example Locations Reassembly Center Reassembled Units

CARAFE

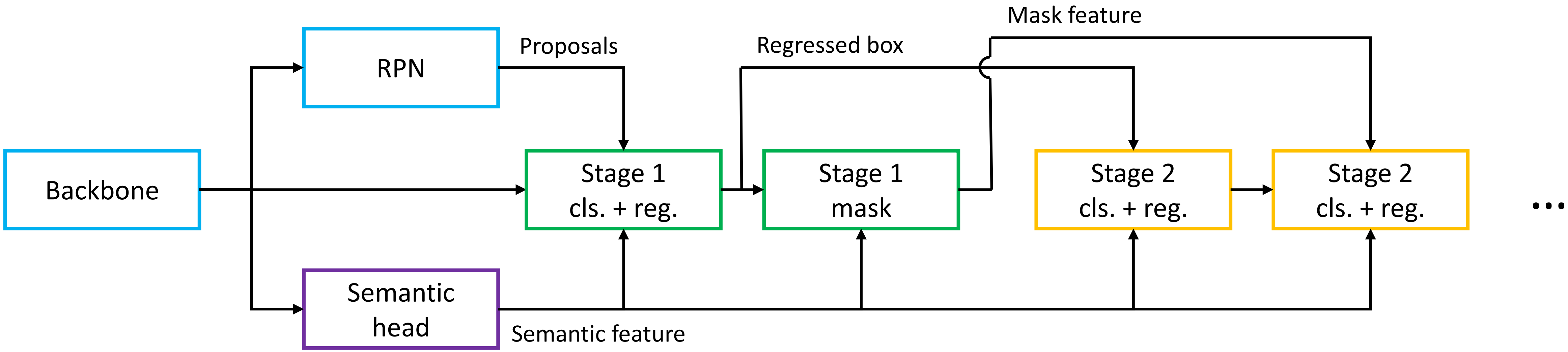
- Universal operator
- Content-aware upsampling
- Fast to compute

香港中文大學
The Chinese University of Hong Kong

Hybrid Task Cascade for Instance Segmentation (CVPR 2019)

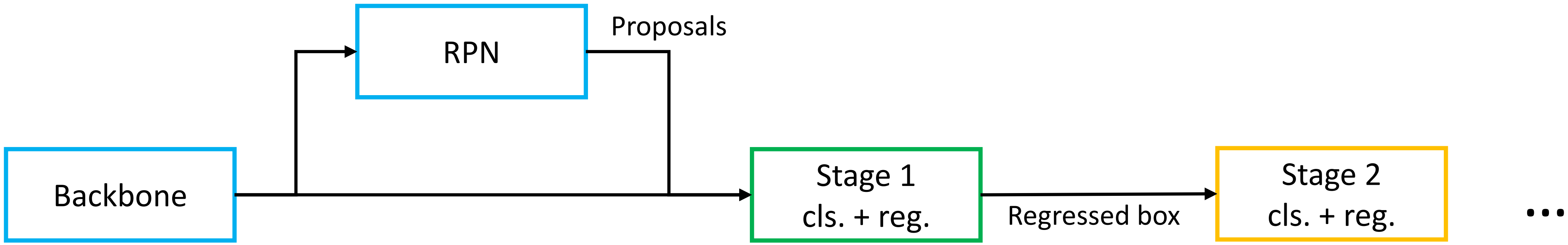
Pipeline

A hybrid architecture with interleaved task branching and cascade.



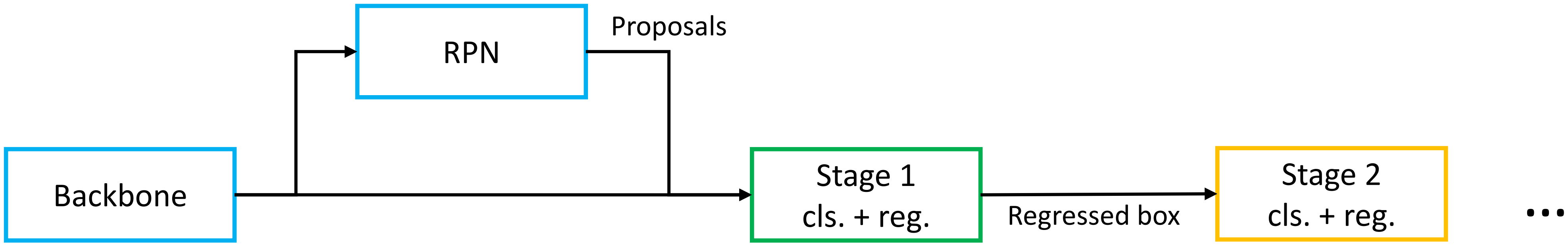
Pipeline

Baseline: Cascade R-CNN



Pipeline

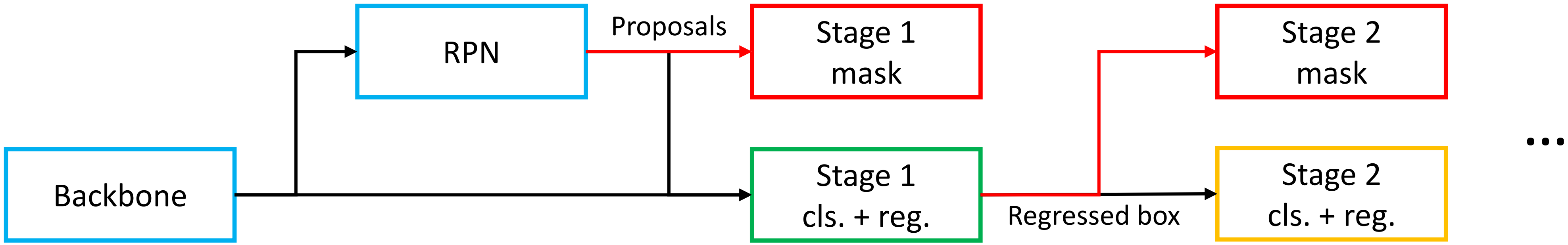
Baseline: Cascade R-CNN



Problem: designed for detection, not segmentation

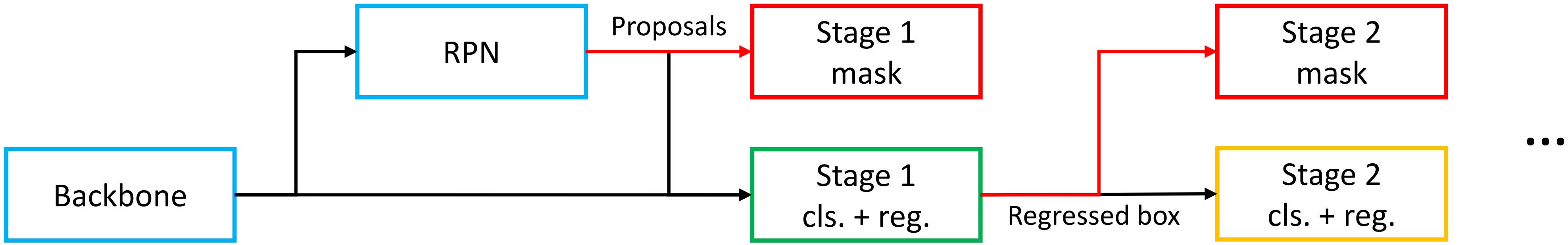
Pipeline

Baseline: Cascade R-CNN + Mask R-CNN



Pipeline

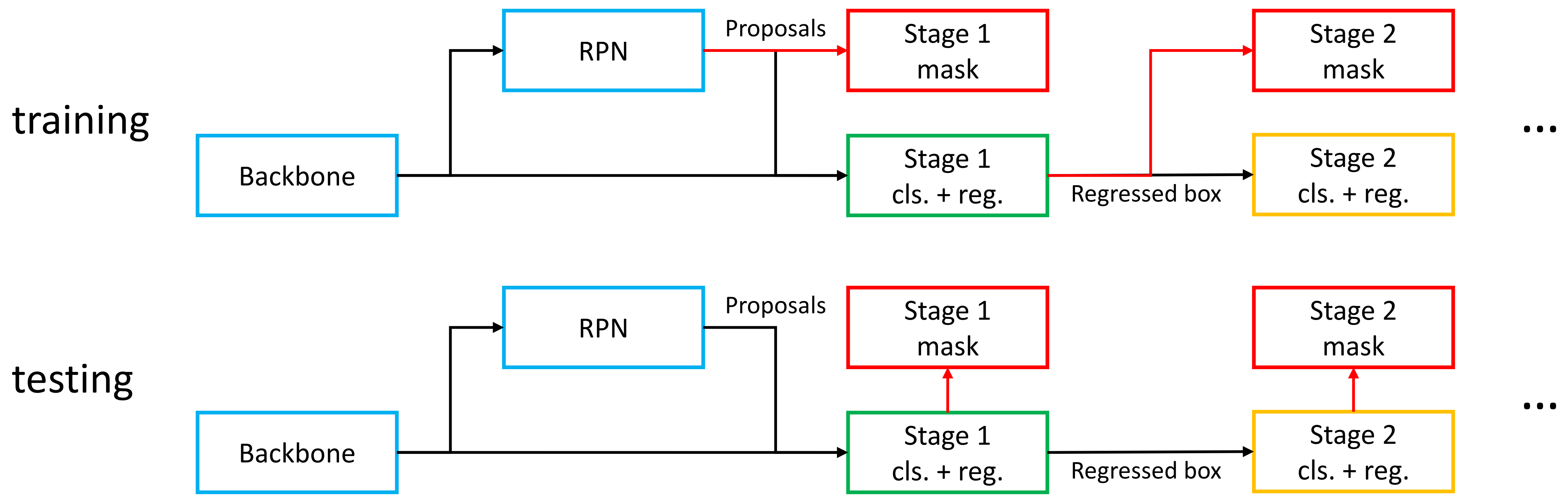
Baseline: Cascade R-CNN + Mask R-CNN



Problem: mismatch of training and testing pipeline

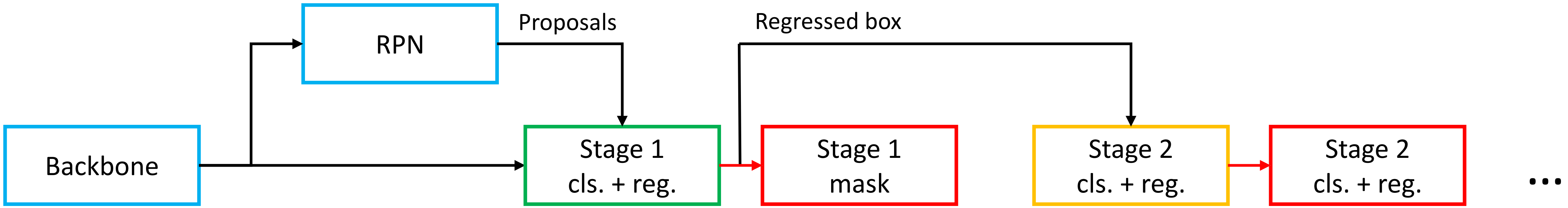
Pipeline

Problem: mismatch of training and testing pipeline



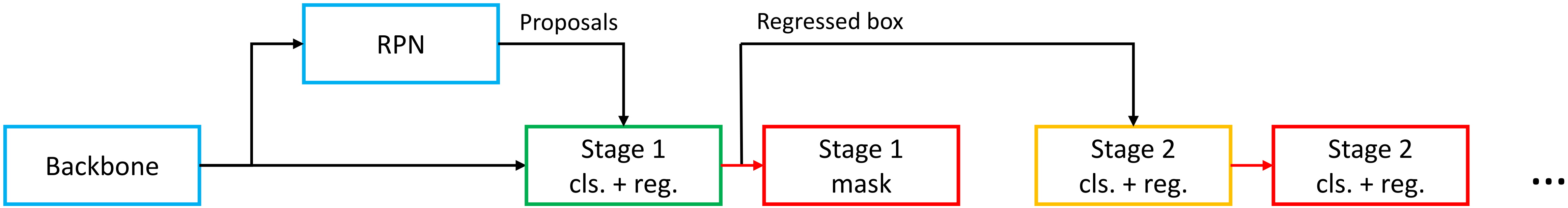
Pipeline

Task cascade: ordinal bbox prediction and mask prediction



Pipeline

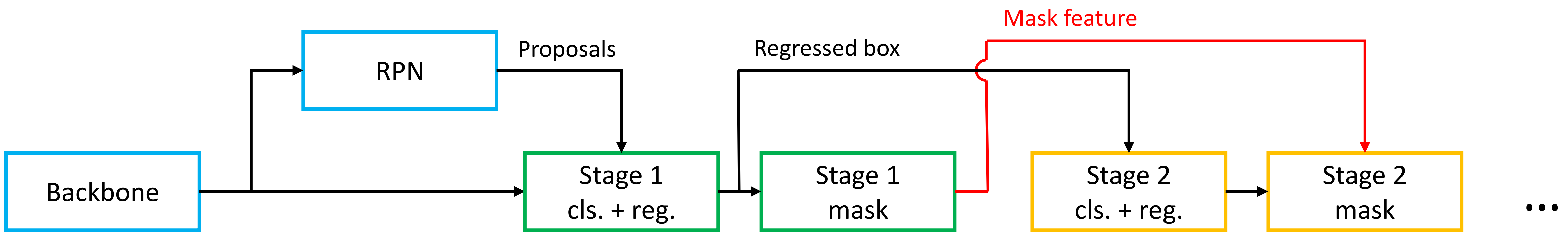
Task cascade: ordinal bbox prediction and mask prediction



Problem: no connection between mask branches of different stages

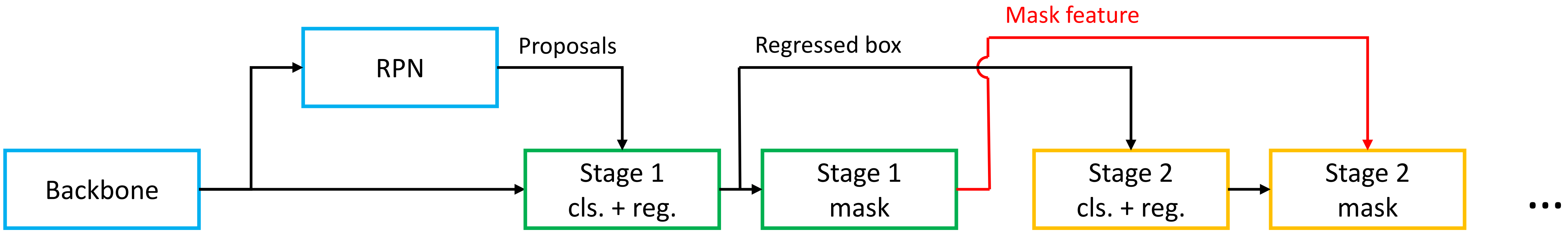
Pipeline

Interleaved execution: box cascade & mask cascade



Pipeline

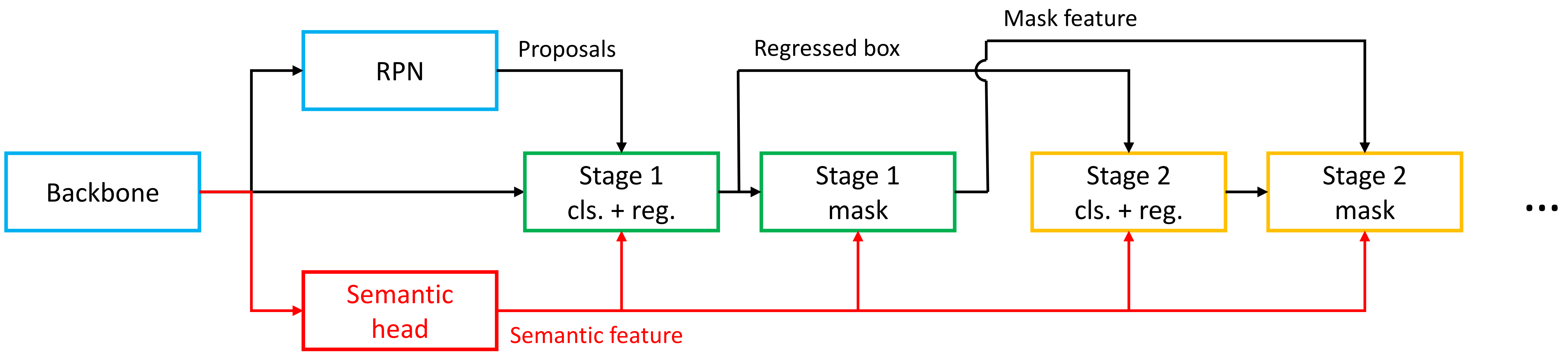
Interleaved execution: box cascade & mask cascade



Problem: contextual information is not much explored

Pipeline

Hybrid branching: additional semantic segmentation branch

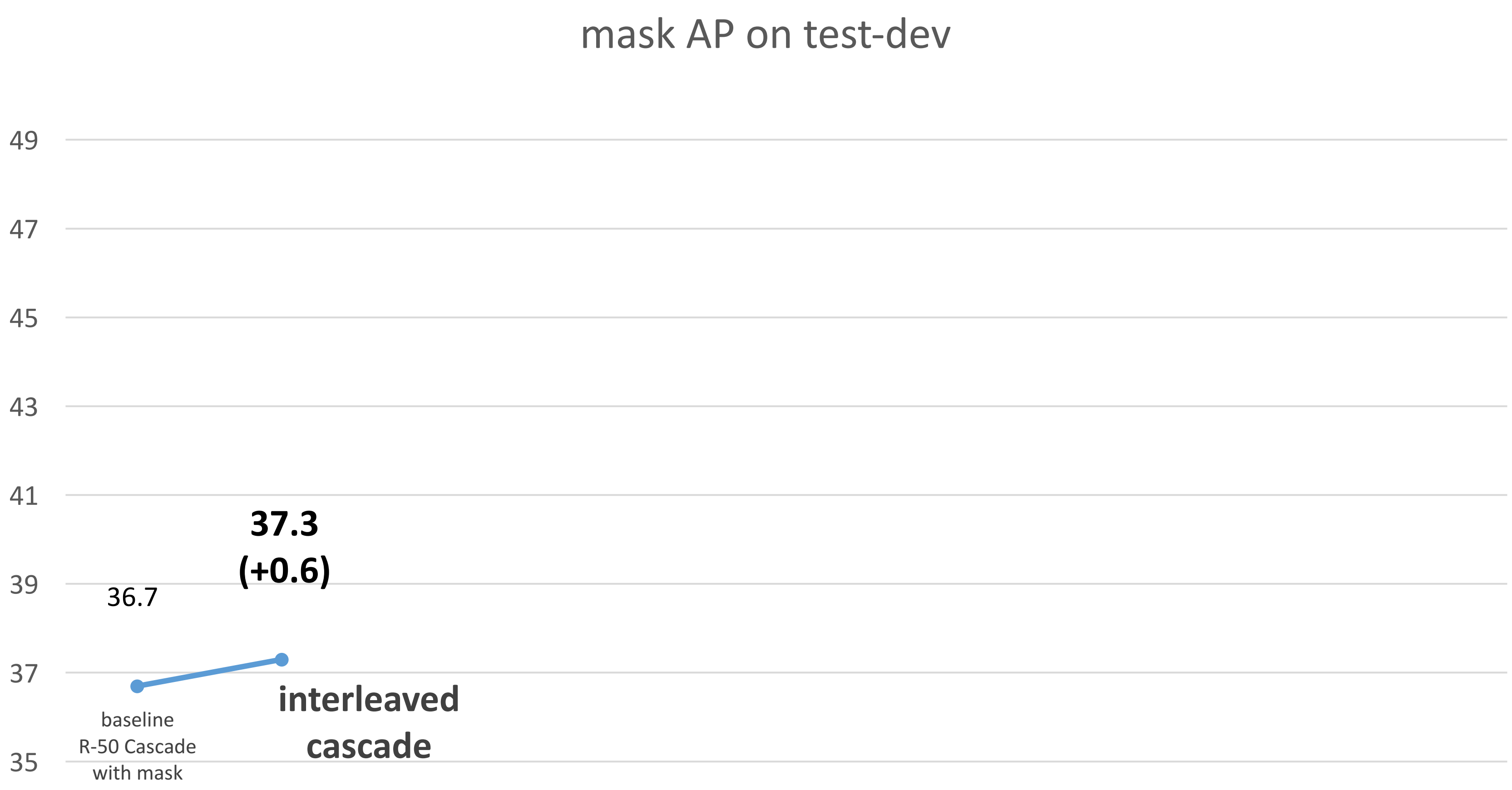


Hybrid Task Cascade

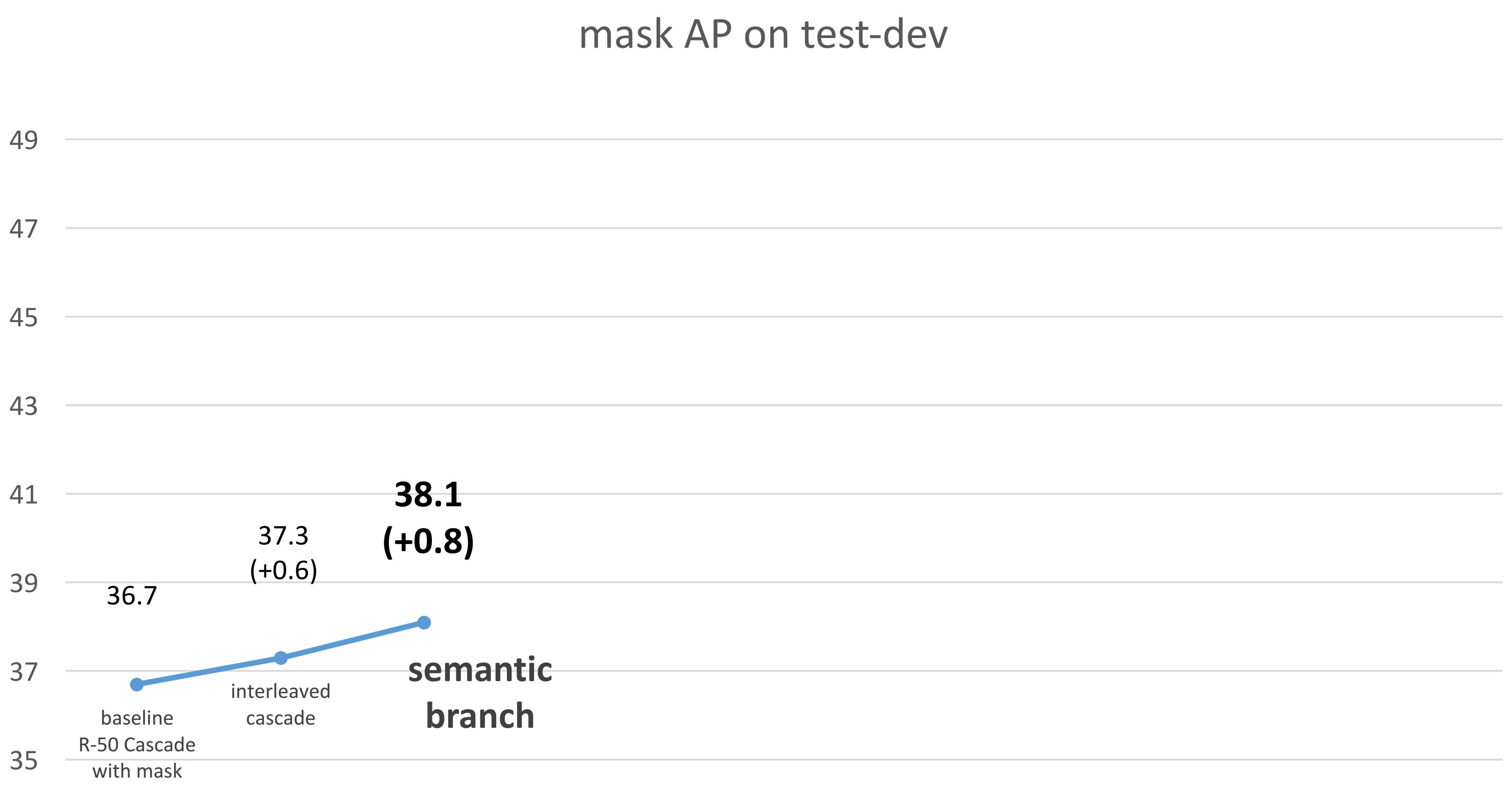
- Cascade between different tasks
- Interleaved execution
- Contextual information fusion

Experiments

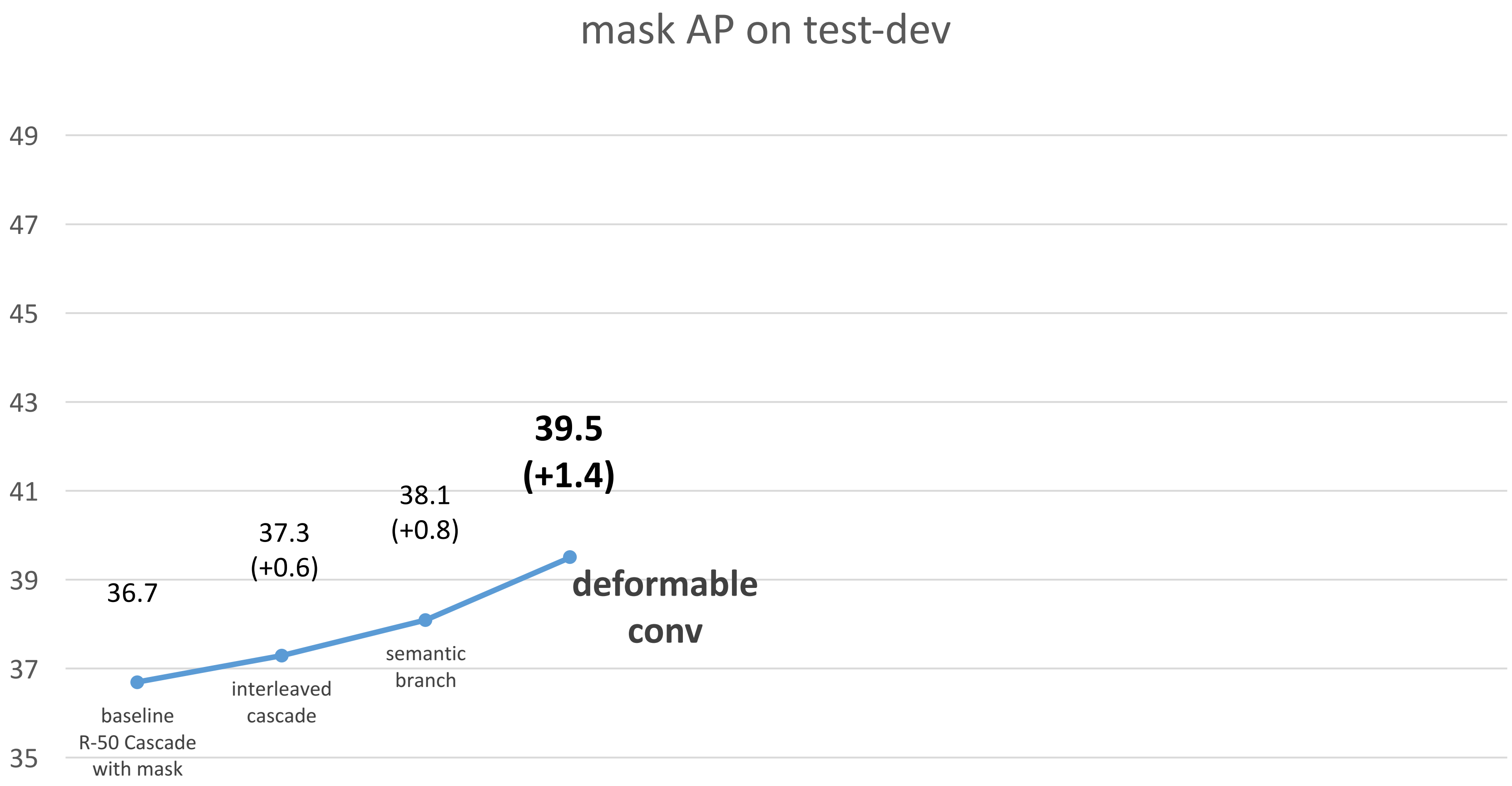
Experiments



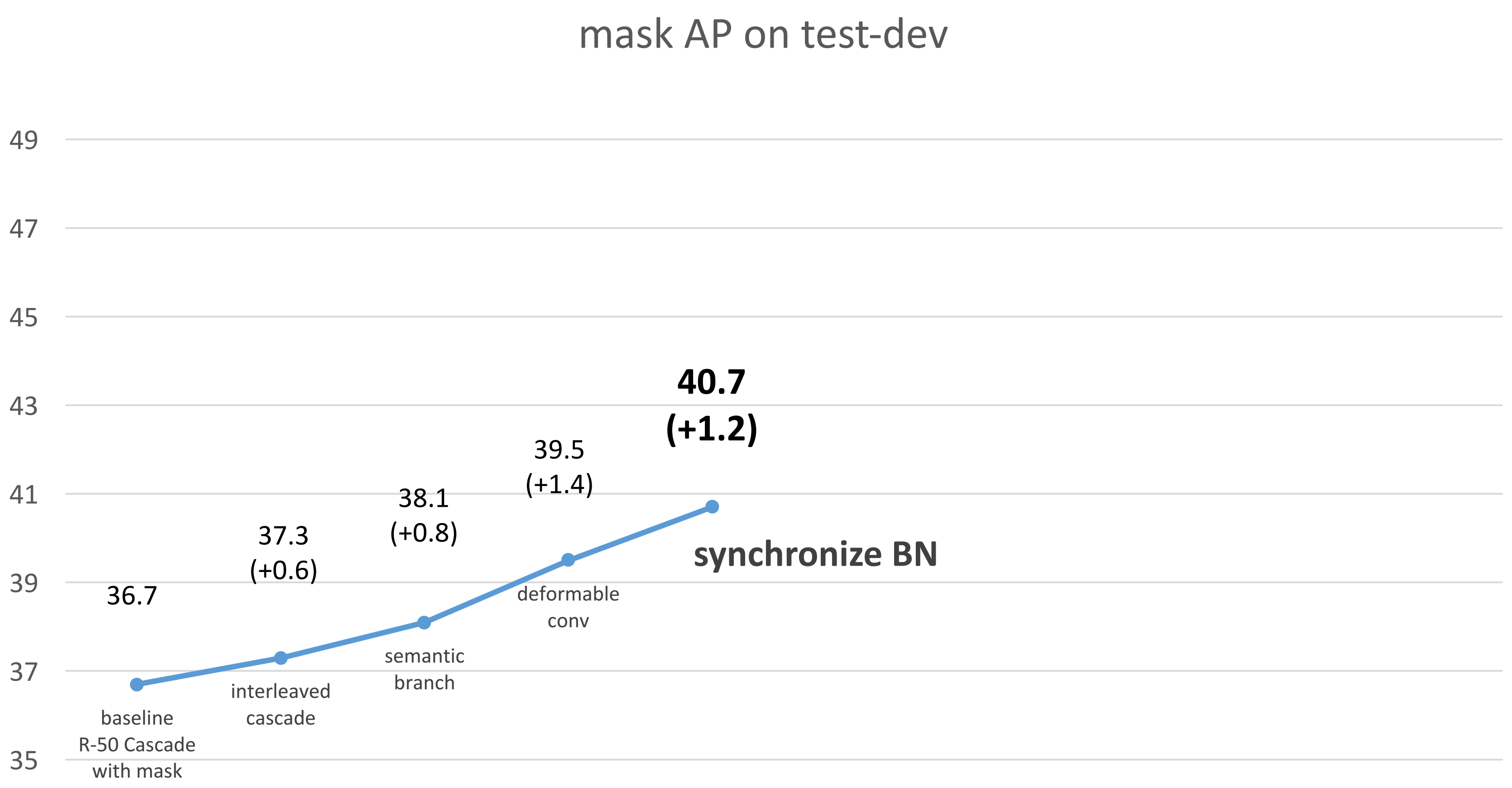
Experiments



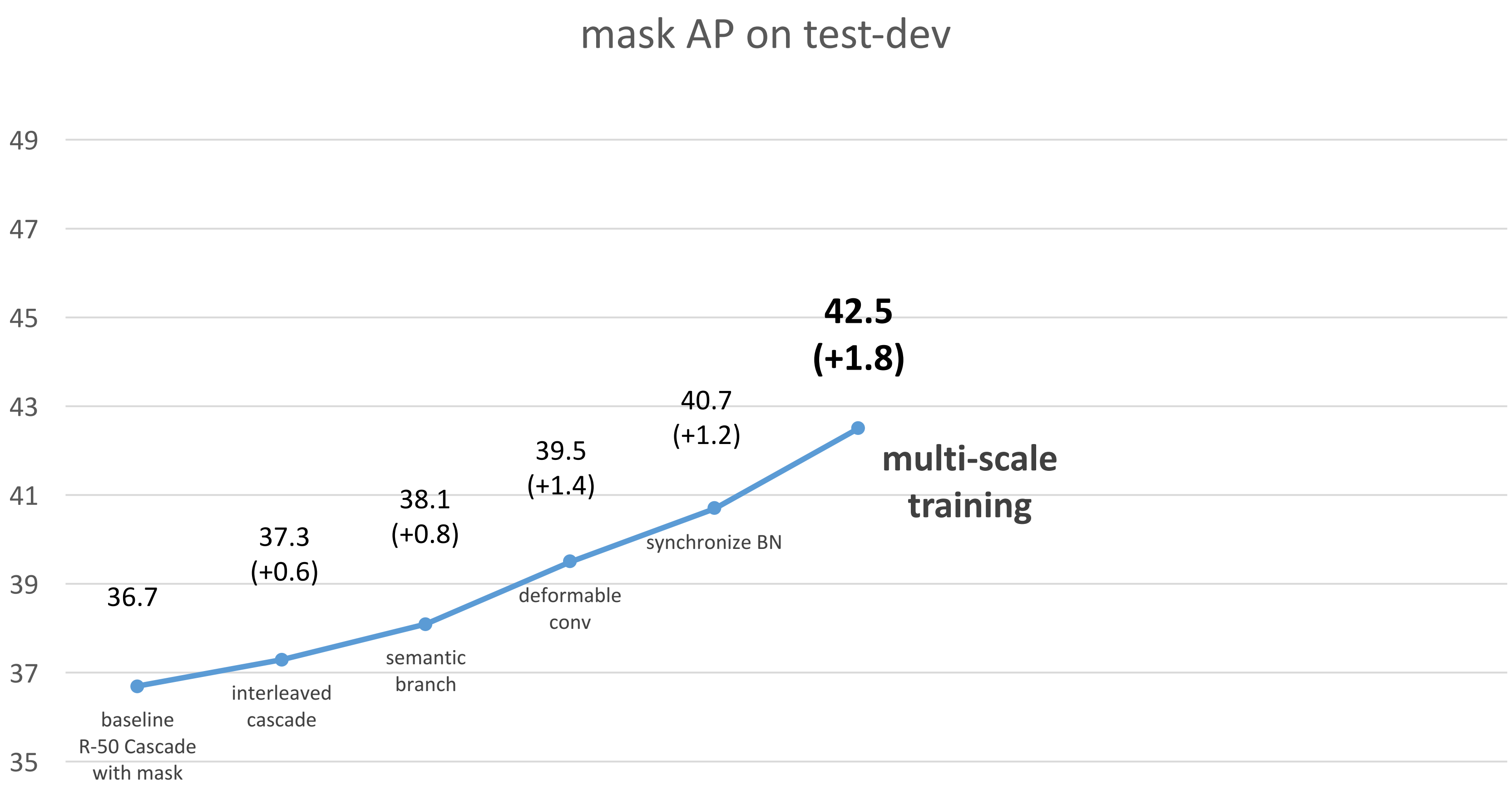
Experiments



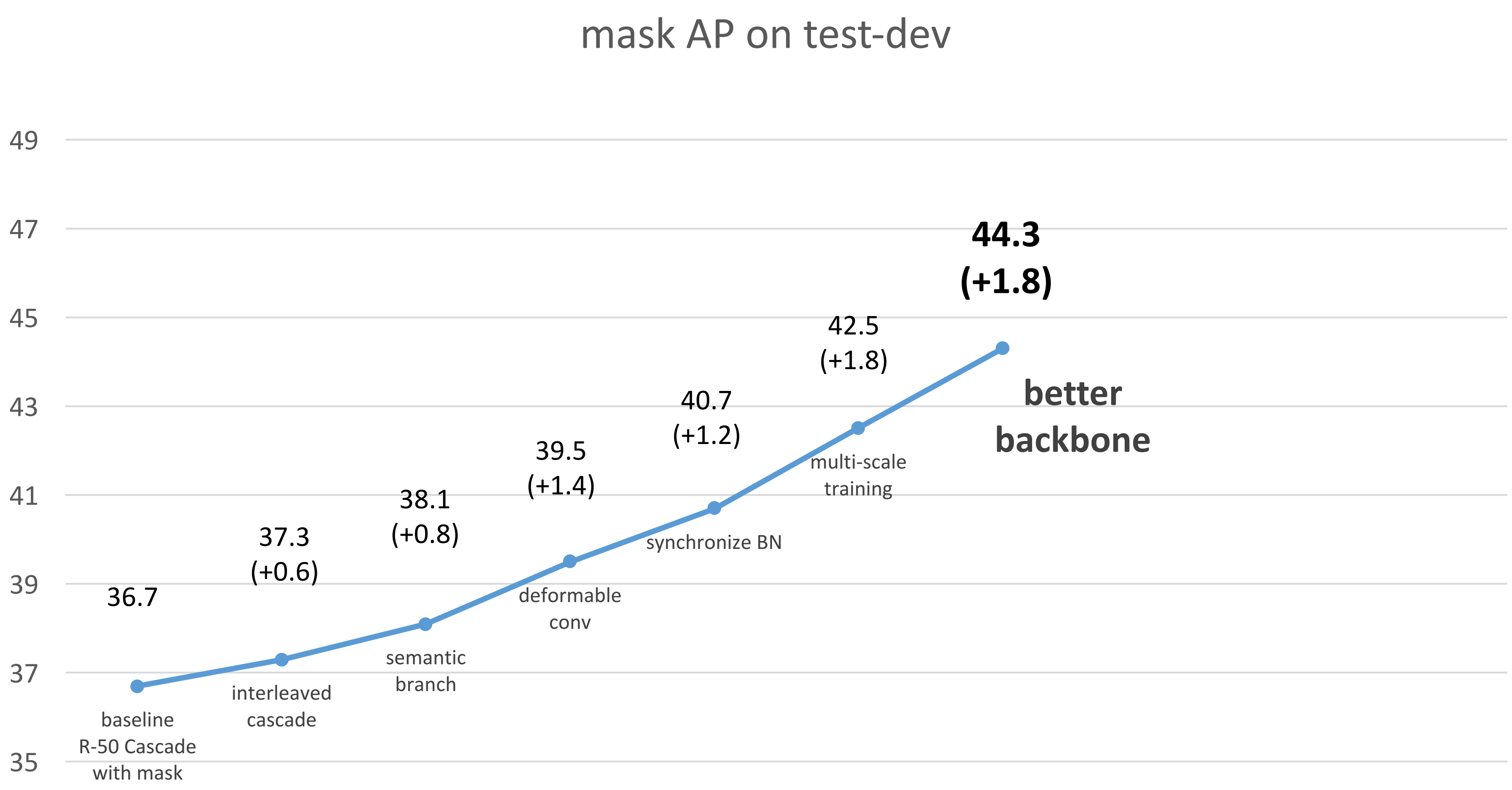
Experiments



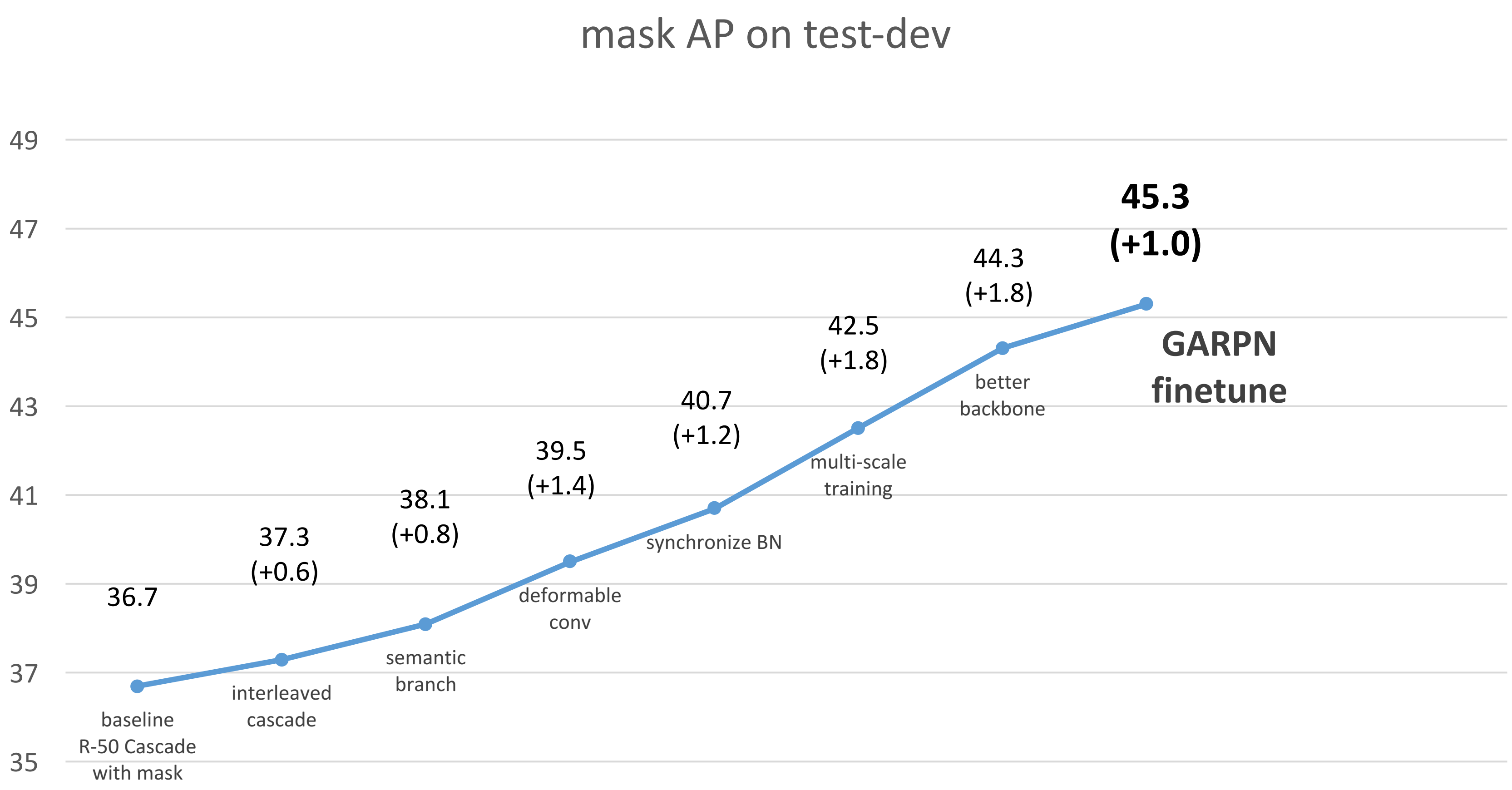
Experiments



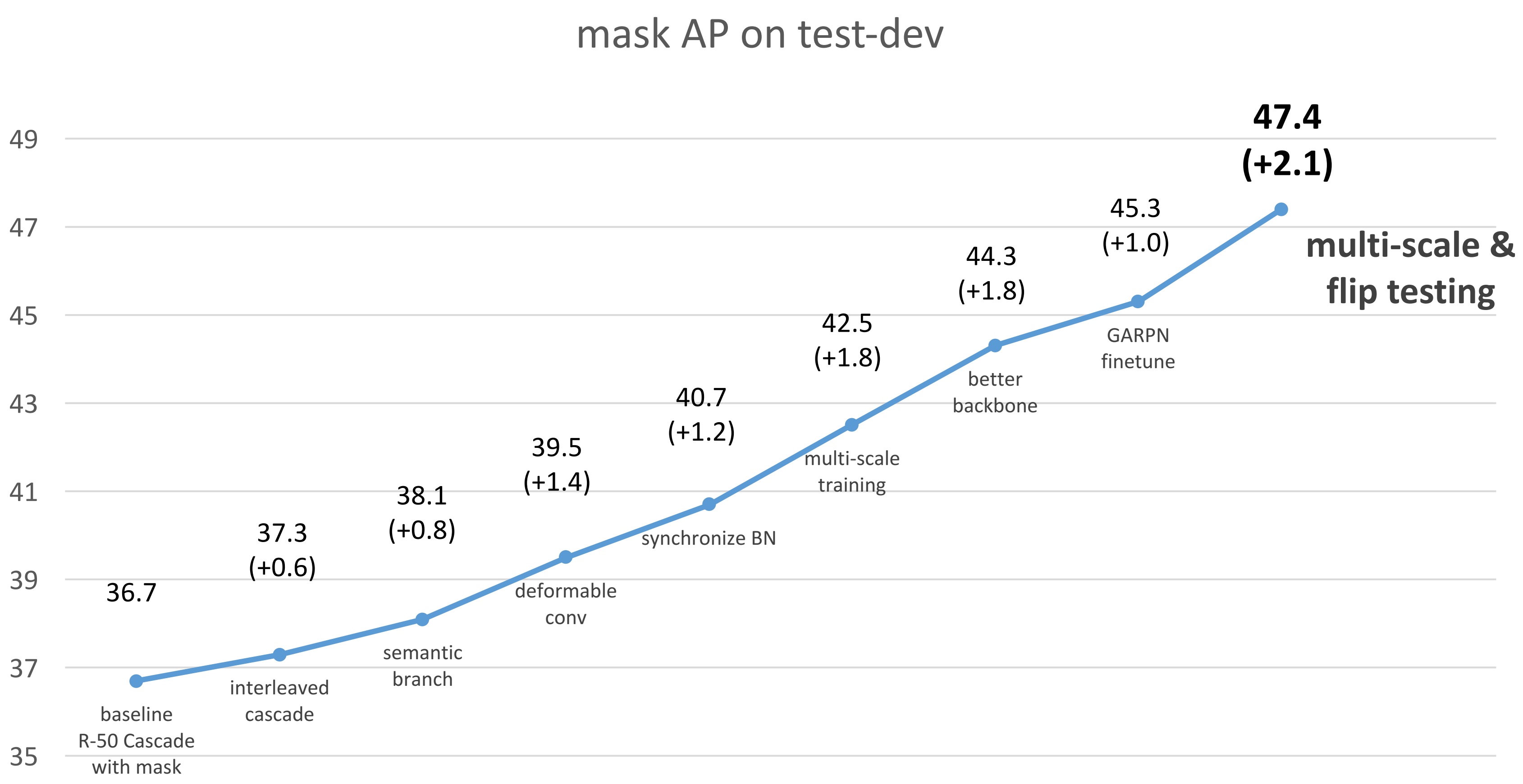
Experiments



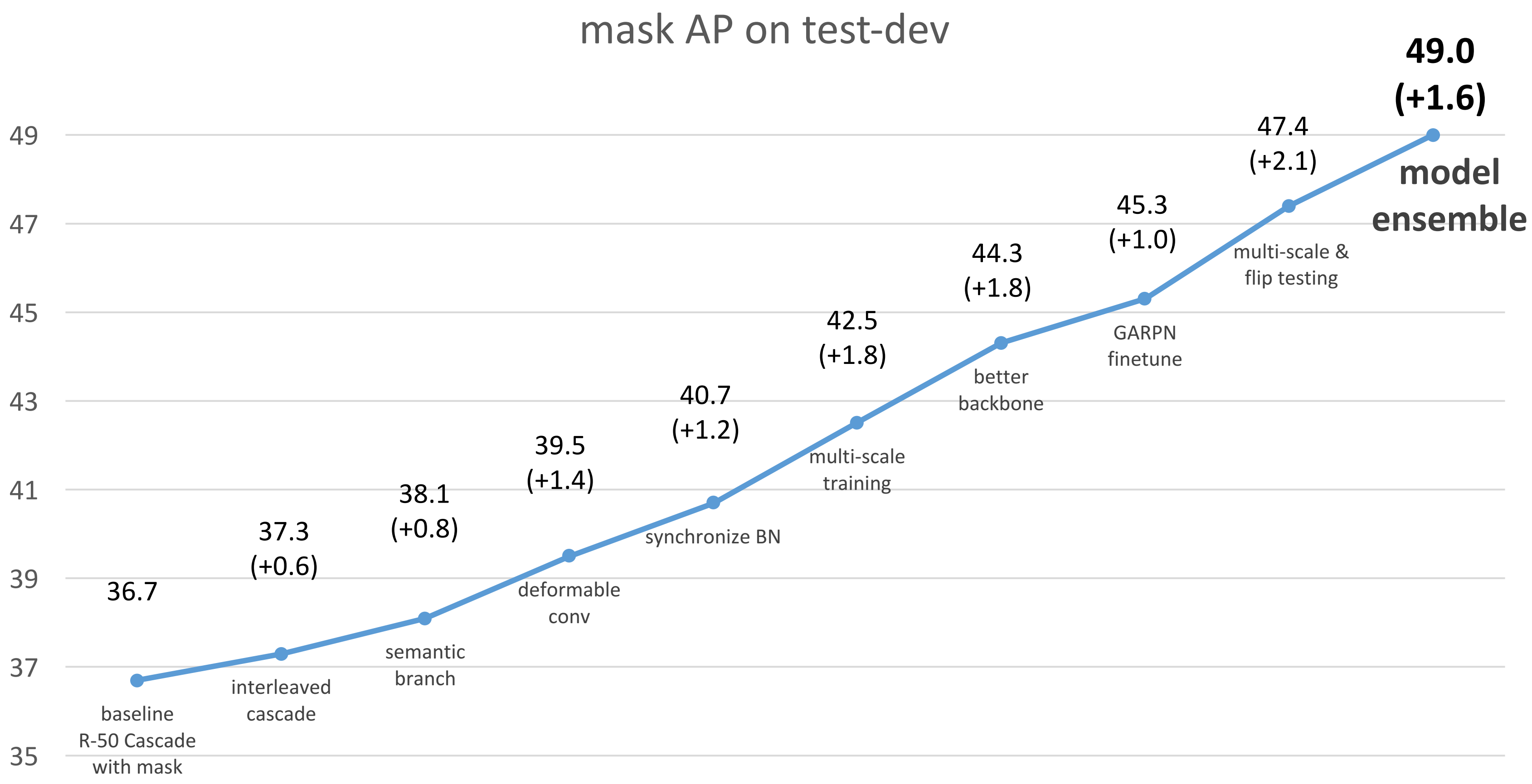
Experiments



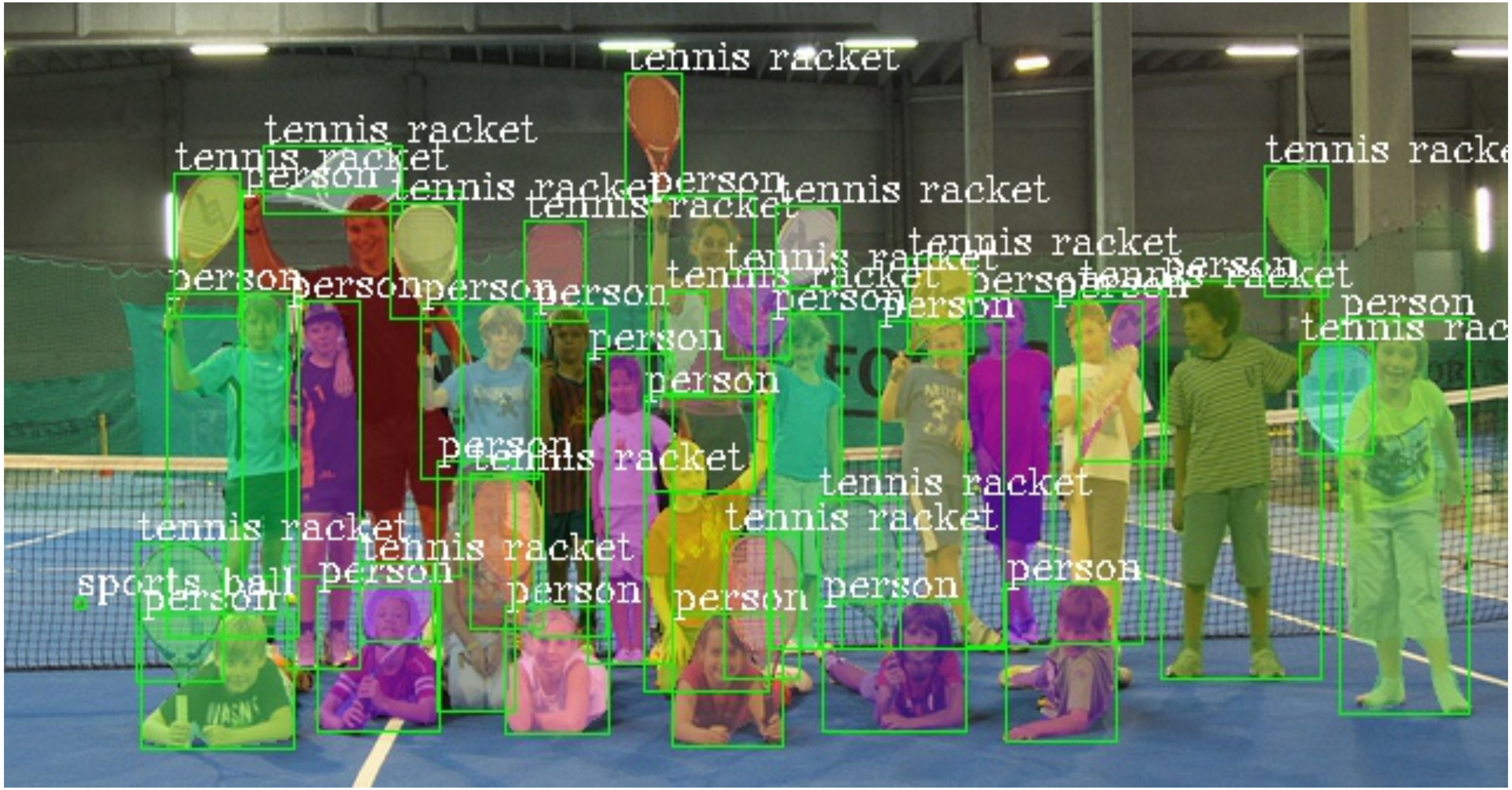
Experiments



Experiments



Visualization



香港中文大學
The Chinese University of Hong Kong

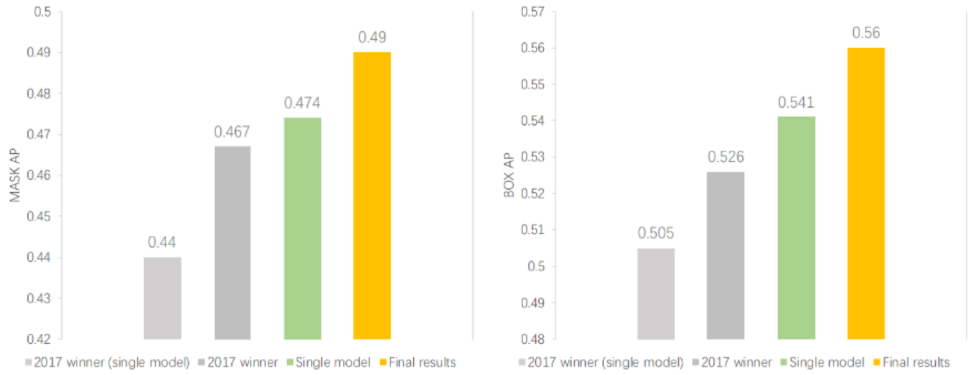
mmdetection (Open-MMLAB)

Codebase

open-mmlab / **mmdetection**

Watch 244 | Star 5,890 | Fork 1,705

	MMDetection	maskrcnn-benchmark	Detectron	SimpleDet
Fast R-CNN	✓	✓	✓	✓
Faster R-CNN	✓	✓	✓	✓
Mask R-CNN	✓	✓	✓	✓
RetinaNet	✓	✓	✓	✓
DCN	✓	✓	✓	✓
DCNv2	✓	✓		
Mixed Precision Training	✓	✓		✓
Cascade R-CNN	✓		*	✓
Weight Standardization	✓	*		
Mask Scoring R-CNN	✓	*		
FCOS	✓	*		
SSD	✓			
R-FCN	✓			
M2Det	✓			
GHM	✓			
ScratchDet	✓			
Double-Head R-CNN	✓			
Grid R-CNN	✓			
FSAF	✓			
Hybrid Task Cascade	✓			
Guided Anchoring	✓			
Libra R-CNN	✓			
Generalized Attention	✓			
GCNet	✓			
HRNet	✓			
TridentNet [17]				✓



PyTorch @PyTorch · 12 Oct 2018
{**mmdetection**, mmcv} by Multimedia Lab @ CUHK
- a modular, object detection and segmentation framework
- fast state-of-the-art models like Mask RCNN, RetinaNet, etc.
- powered the winning entry of COCO Detection 2018 challenge.
github.com/open-mmlab/mmd...
mmcv.readthedocs.io/en/latest/

95 232

- 10+ research institutes
- 20+ supported methods
- 200+ pre-trained models

GitHub: mmdet

Codebase

[Update] 1st place solution with code

posted in iMaterialist (Fashion) 2019 at FGVC6 24 days ago

95

Miras Amir

1st place

Hi Kagglers,

My solution is based on the COCO challenge 2018 winners arti: 901.07518.

Code:

https://github.com/amirassov/kaggle-imaterialist

Model:

Hybrid Task Cascade with ResNeXt-101-64x4d-FPN backbone. This model has a metric Mask mAP = 43.9 on COCO dataset. This is SOTA for instance segmentation.

GitHub: mmdet

The entries ranking 1, 2, and 3 of iMaterialist (Fashion) 2019 at FGVC6 (CVPR 2019 Workshop) are based on HTC. Here is the post of the winner.

香港中文大學
The Chinese University of Hong Kong

Thank you!

Dynamic forwarding and routing as a computational strategy for detection and beyond