CVPR LONG BEACH CALIFORNIA June 16-20, 2019

Electrical & Computer ENGINEERING

Feature Selective Anchor-Free Module for Single-Shot Object Detection

Chenchen Zhu, Yihui He, Marios Savvides

Carnegie Mellon University

04/18/2019

Overview

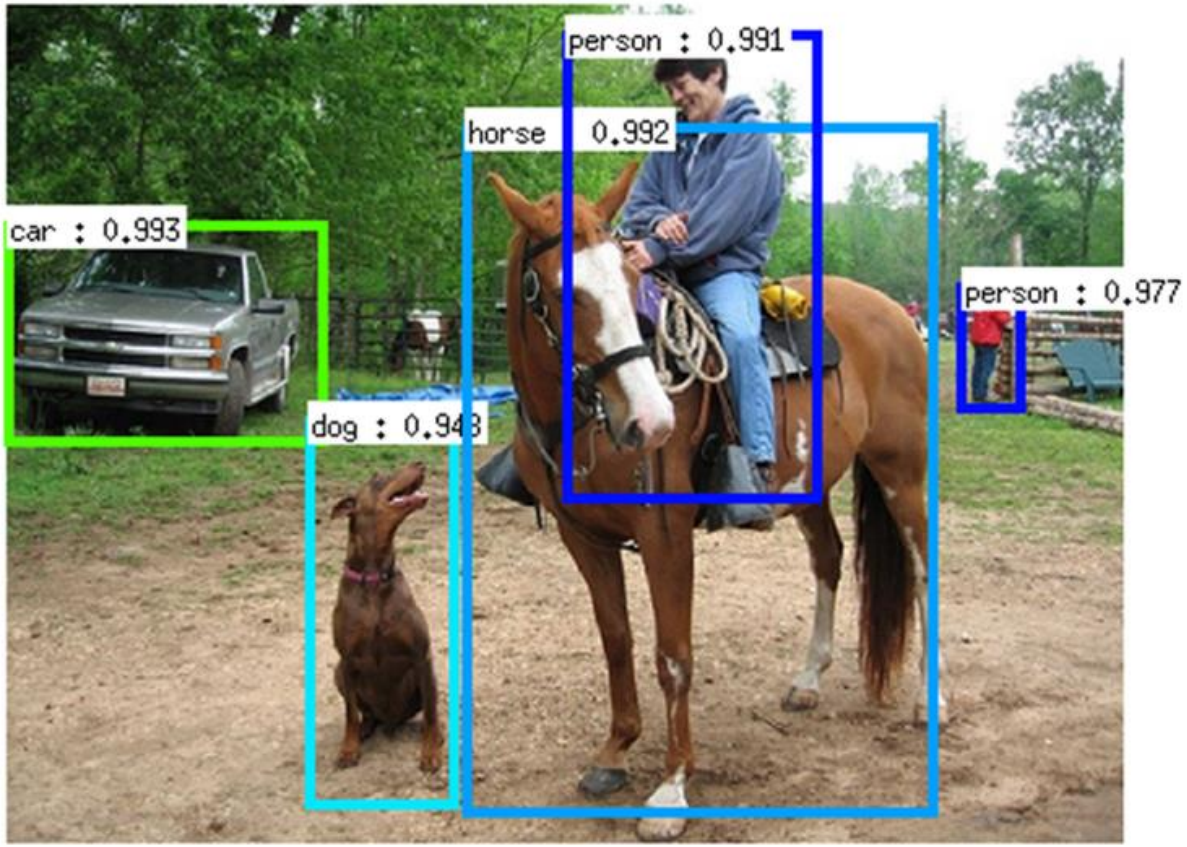
- Background
- Motivation
- Feature Selective Anchor-Free (FSAF) Module
  - General concept
  - Network architecture
  - Ground-truth and loss
  - Online feature selection
- Experiments
- Qualitative Results

Overview

- **Background**
- Motivation
- Feature Selective Anchor-Free (FSAF) Module
  - General concept
  - Network architecture
  - Ground-truth and loss
  - Online feature selection
- Experiments
- Qualitative Results

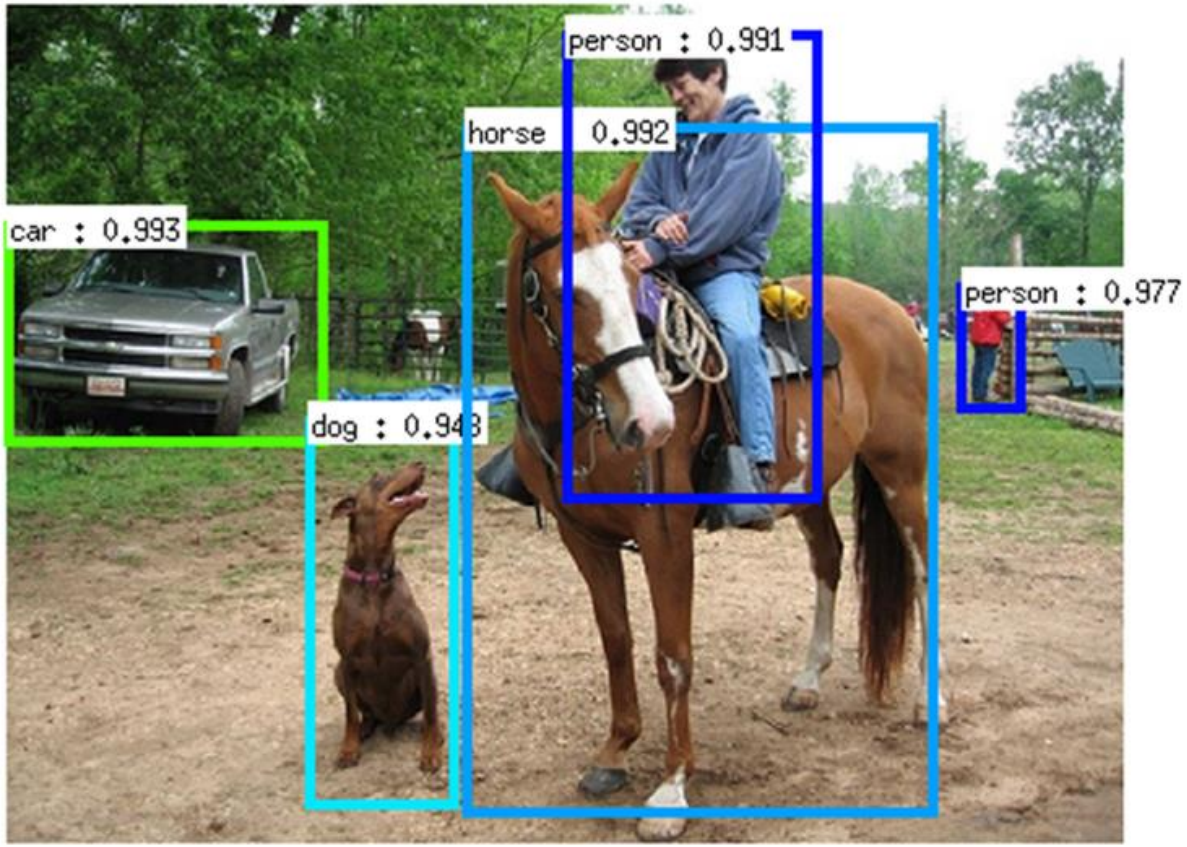
Background

Object detection = localization + classification



Background

A long-lasting challenge: scale variation



Background

Prior methods addressing scale variation

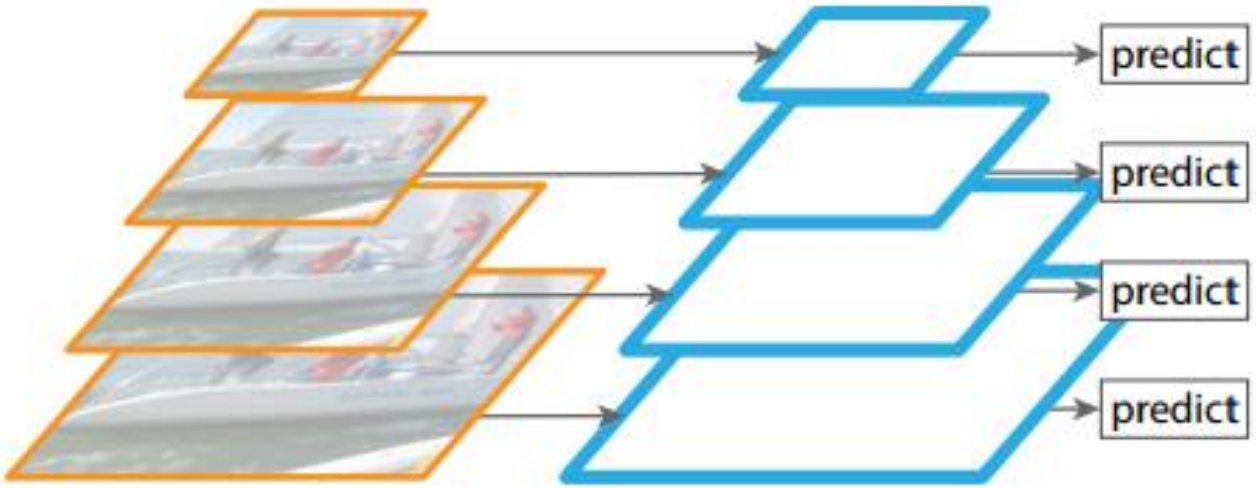
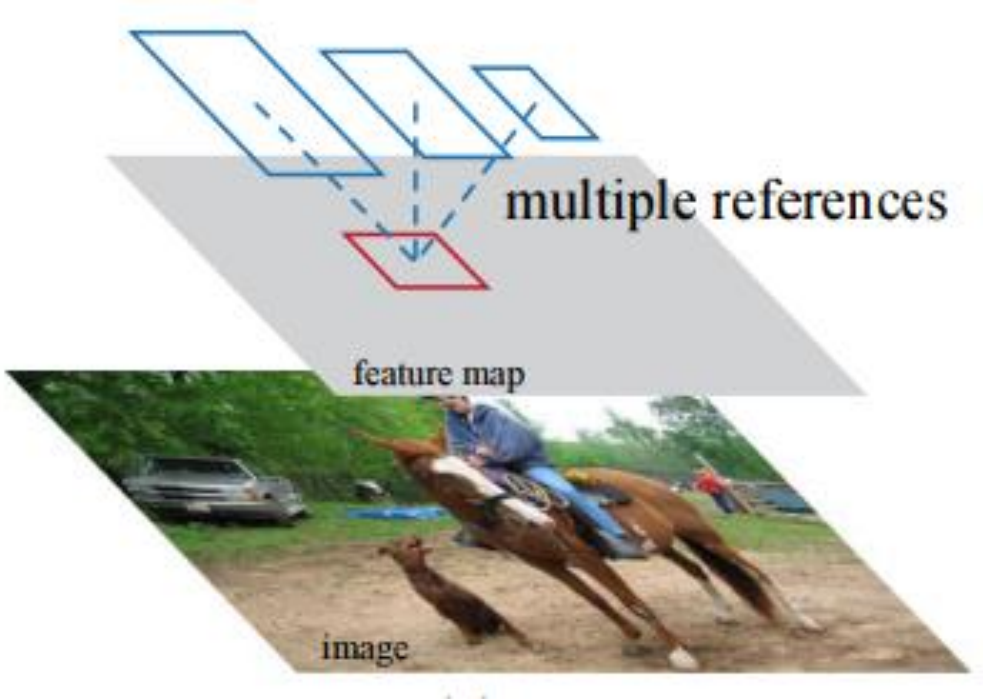


Image pyramid

Background

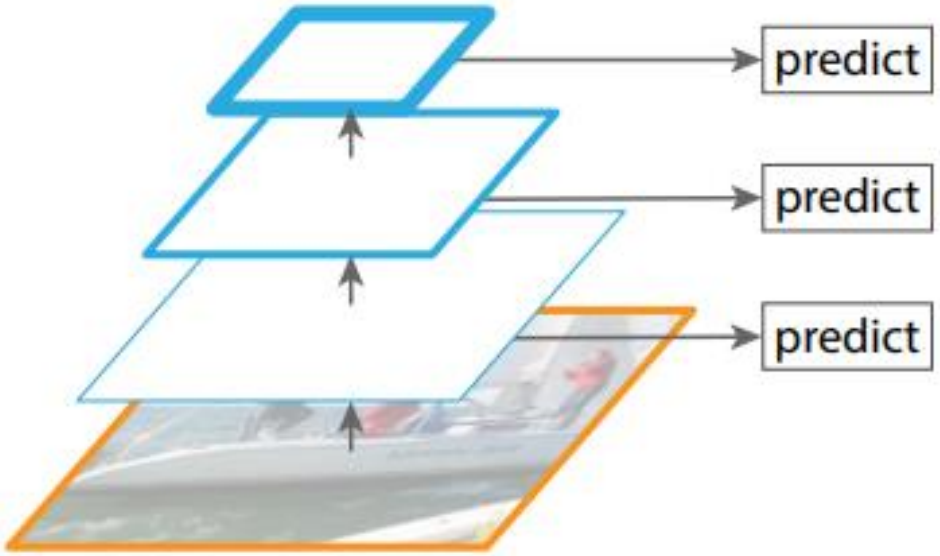
Prior methods addressing scale variation



Anchor boxes [Ren et al, Faster R-CNN]

Background

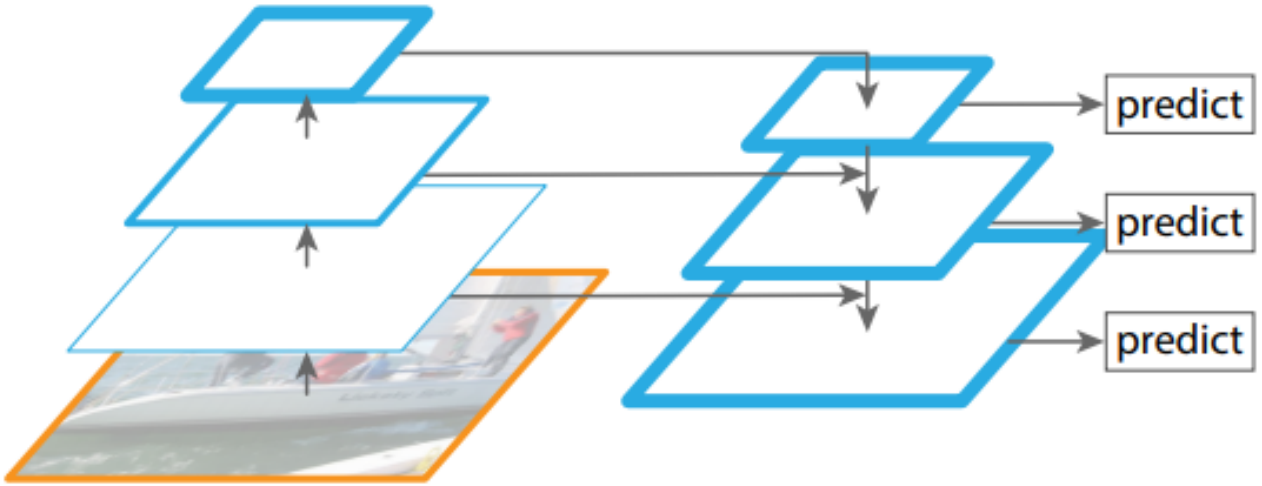
Prior methods addressing scale variation



Pyramidal feature hierarchy, e.g. [Liu et al, SSD]

Background

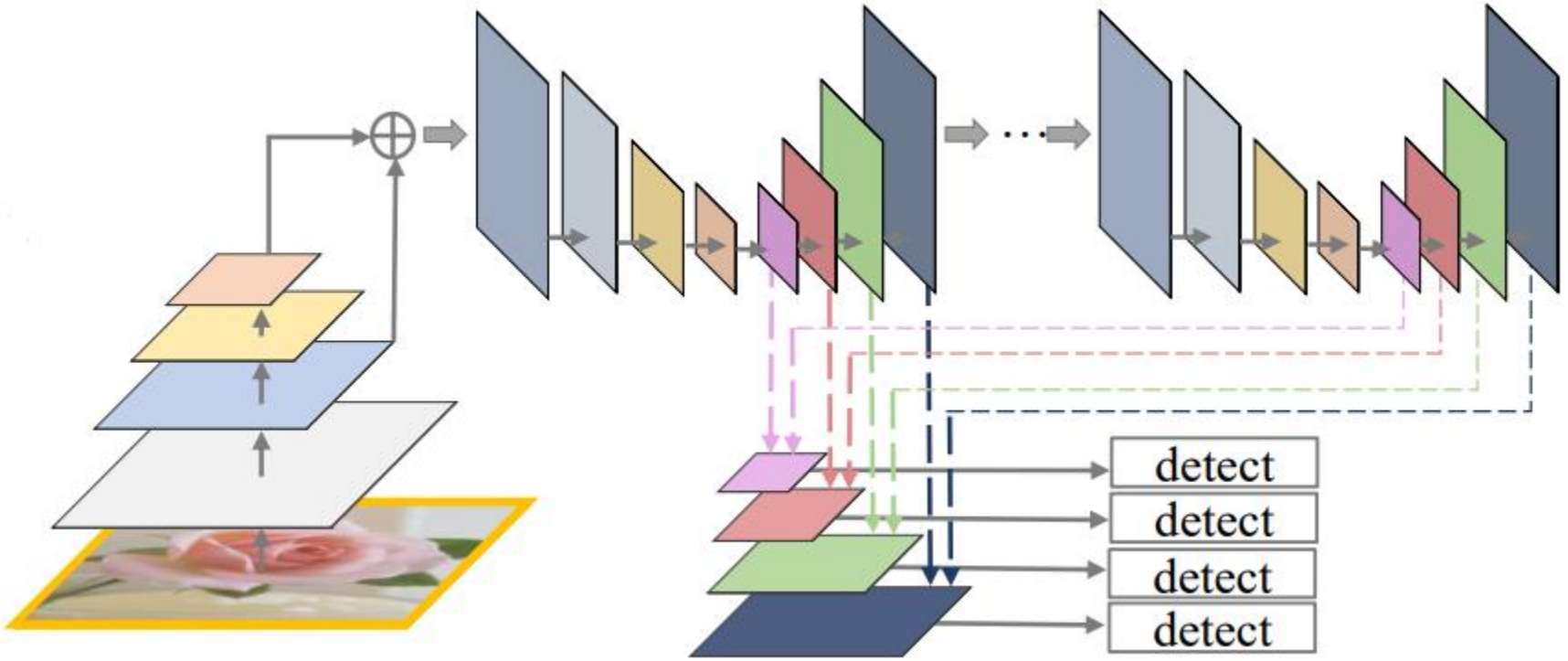
Prior methods addressing scale variation



Feature pyramid network [Lin et al, FPN, RetinaNet]

Background

Prior methods addressing scale variation

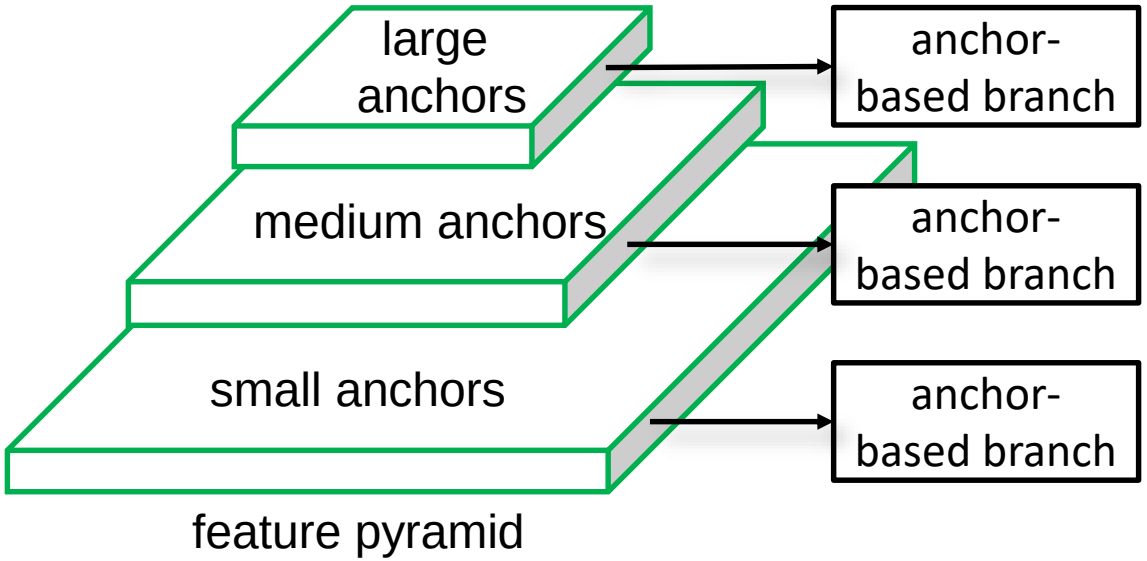


Multi-level feature pyramid network [Zhao et al, M2Det]

Background

Combining feature pyramid with anchor boxes

- **Smaller anchor associated with lower pyramid levels**
- **Larger anchor associated with higher pyramid levels**



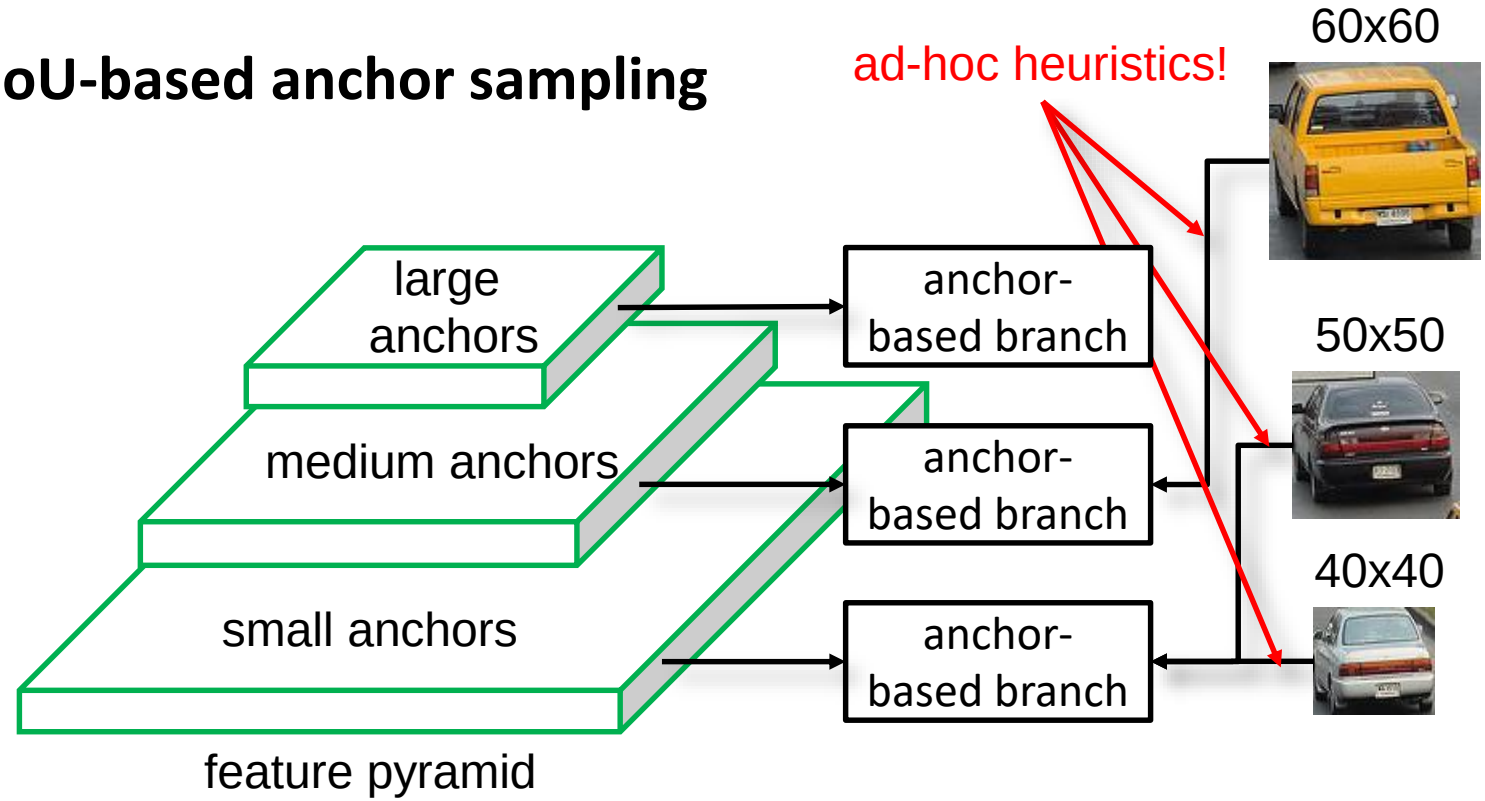
Overview

- **Background**
- **Motivation**
- **Feature Selective Anchor-Free (FSAF) Module**
  - **General concept**
  - **Network architecture**
  - **Ground-truth and loss**
  - **Online feature selection**
- **Experiments**
- **Qualitative Results**

Motivation

Inherent limitations

- **Heuristic-guided feature selection**
- **IoU-based anchor sampling**



Motivation

Problem: feature selection by anchor boxes may not be optimal!

Question: how can we select feature level based on semantic information rather than just box size?

Answer: allowing arbitrary feature assignment by removing the anchor matching mechanism (using anchor-free methods), selecting the most suitable feature level.

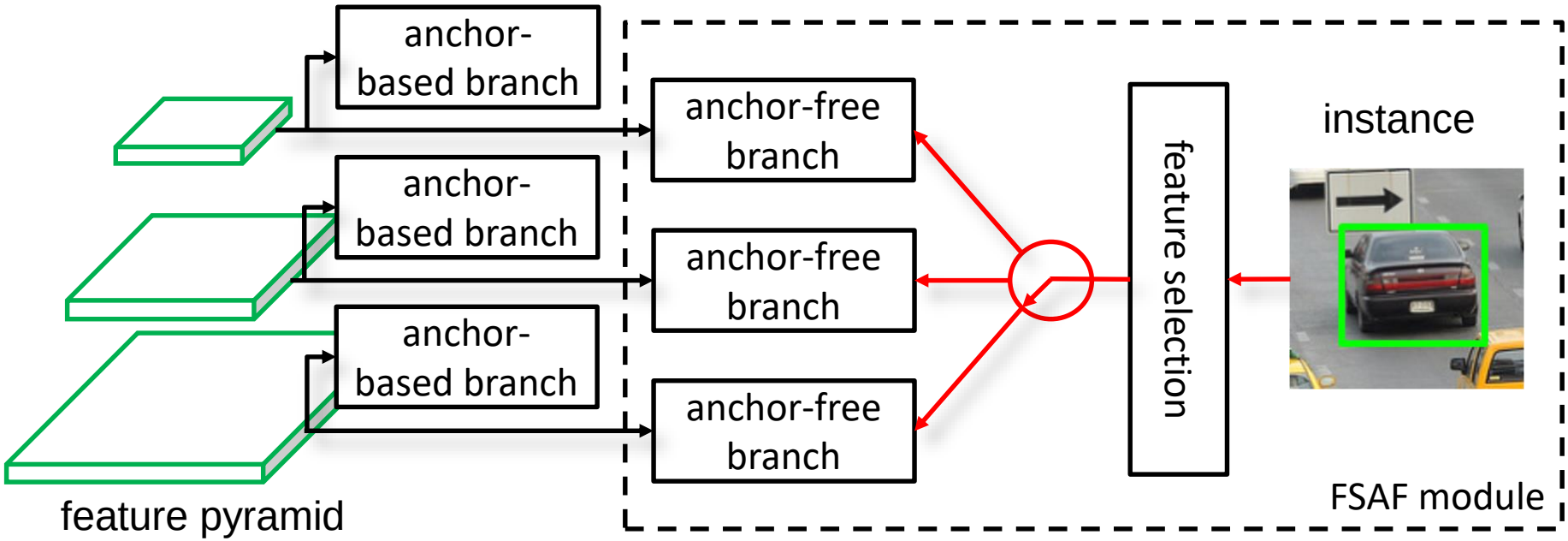
Solution: Feature Selective Anchor-Free (FSAF) Module

Overview

- Background
- Motivation
- Feature Selective Anchor-Free (FSAF) Module
  - General concept
  - Network architecture
  - Ground-truth and loss
  - Online feature selection
- Experiments
- Qualitative Results

FSAF Module

The general *concept*



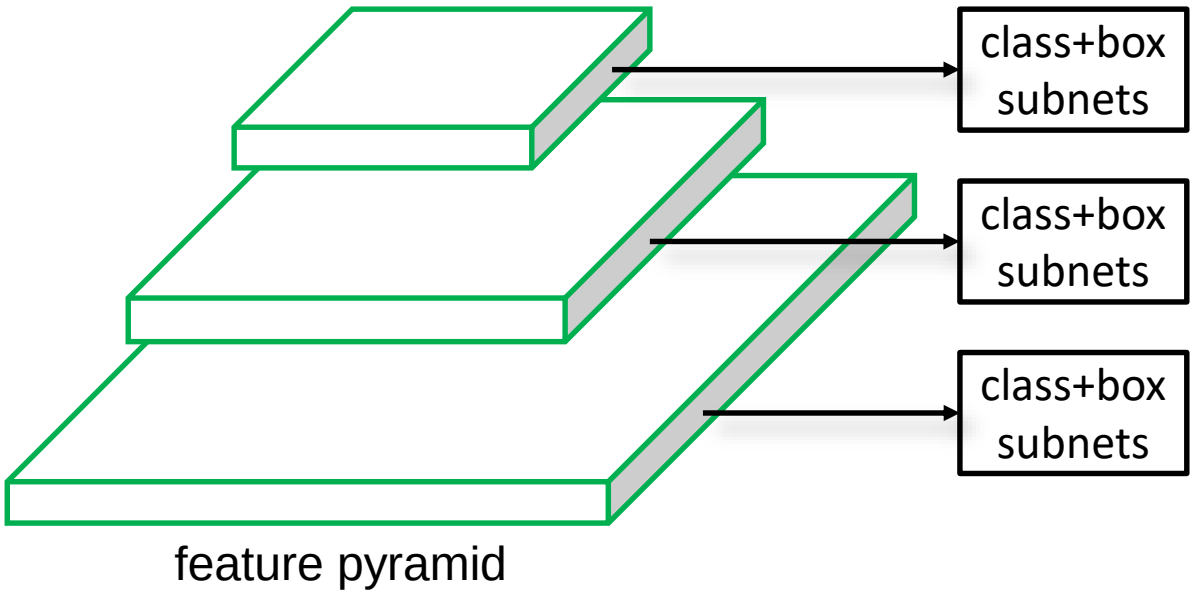
FSAF Module

Instantiation

- Network architecture
- Ground-truth and loss
- Online feature selection

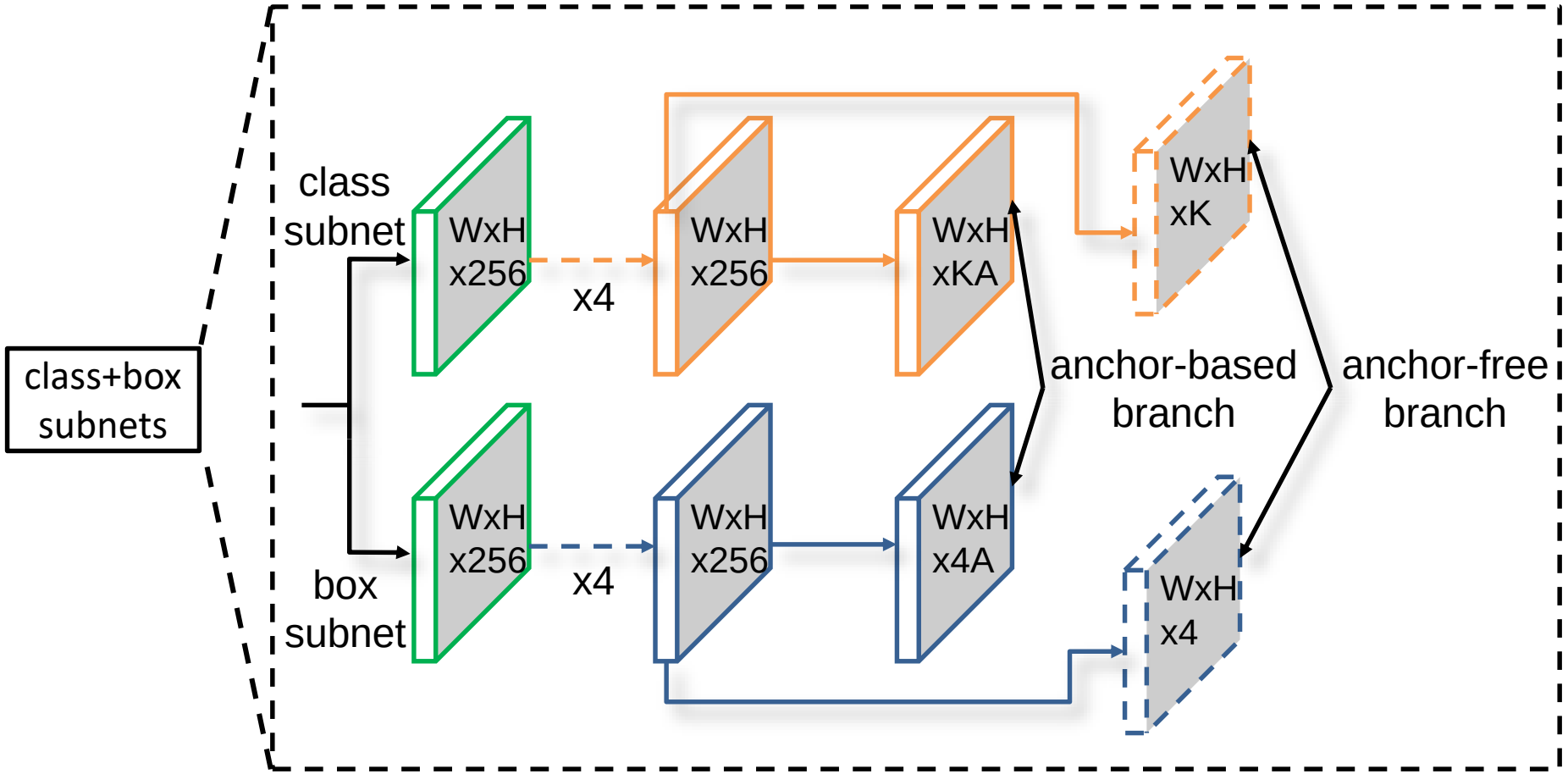
FSAF Module

Network architecture (on RetinaNet)



FSAF Module

Network architecture (on RetinaNet)



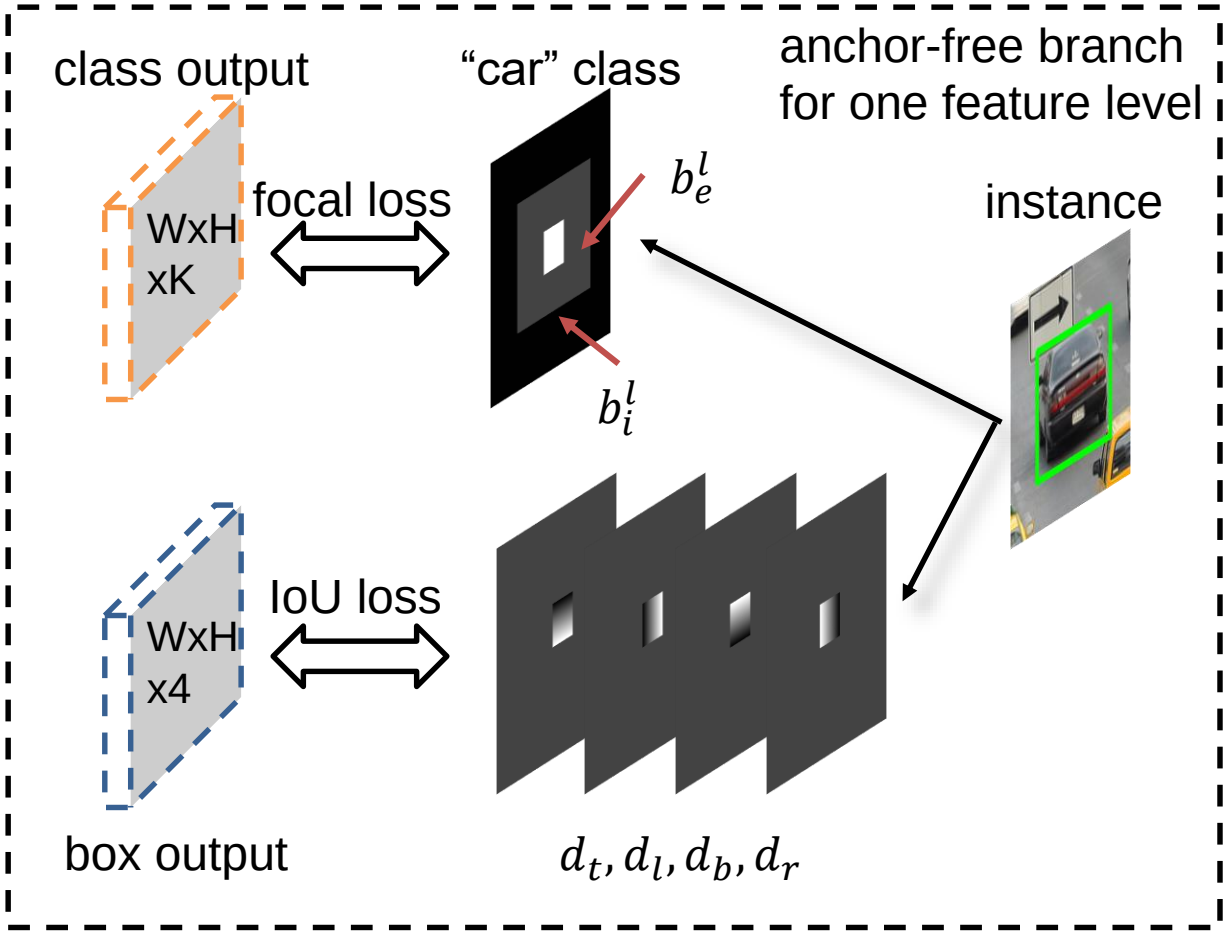
FSAF Module

Ground-truth and loss

- **Definitions**
  - Instance box: $b = [x, y, w, h]$
  - Projected box on P_l: $b_p^l = \left[x_p^l, y_p^l, w_p^l, h_p^l\right] = b/2^l$
  - Effective box on P_l: $b_e^l = \left[x_p^l, y_p^l, \epsilon_e w_p^l, \epsilon_e h_p^l\right]$
  - Ignoring box on P_l: $b_i^l = \left[x_p^l, y_p^l, \epsilon_i w_p^l, \epsilon_i h_p^l\right]$
  - For pixel (i, j) in b_e^l, $[d_{t_{i,j}}^l, d_{l_{i,j}}^l, d_{b_{i,j}}^l, d_{r_{i,j}}^l]$ are distances of (i, j) to the top, left, bottom, right boundaries of b_p^l, respectively.

FSAF Module

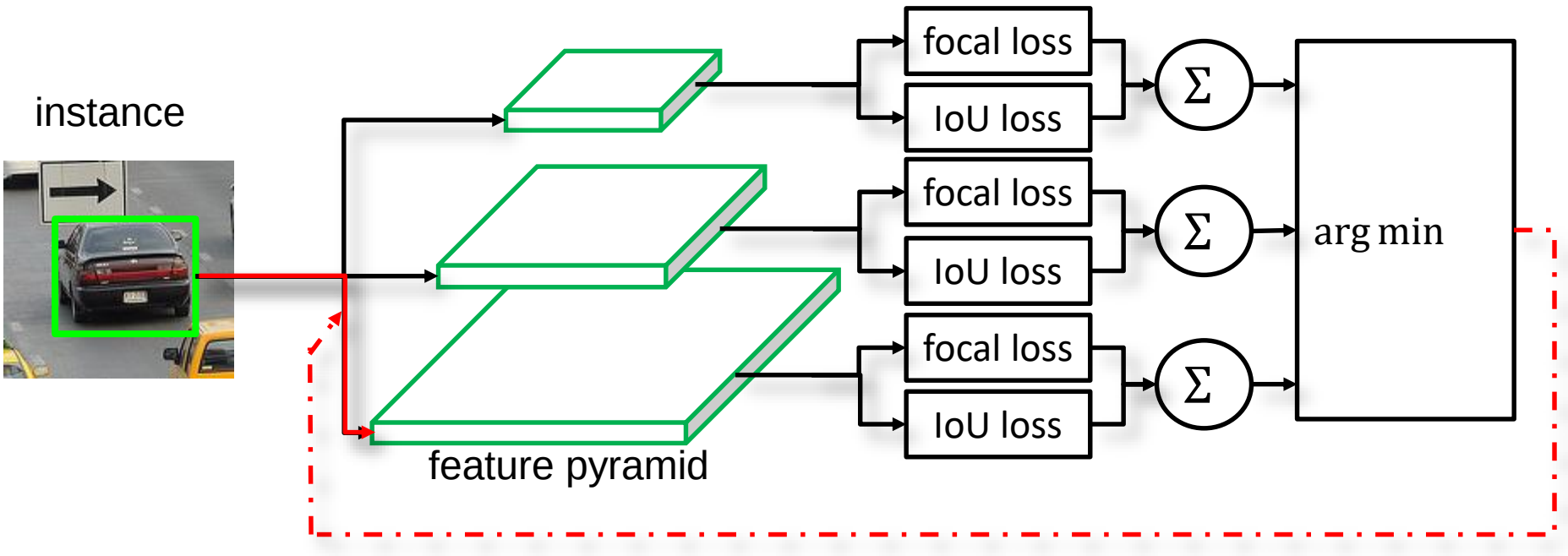
Ground-truth and loss



FSAF Module

Online feature selection on anchor-free branches

$$l^* = \arg\min_l L_{FL}^I(l) + L_{IoU}^I(l)$$



FSAF Module

Heuristic feature selection (for comparison)

$$l' = \left\lfloor l_0 + \log_2(\sqrt{wh}/224) \right\rfloor$$

where l_0 is the target level to which an instance with $w \times h = 224^2$ is mapped [Lin et al, FPN].

Overview

- Background
- Motivation
- Feature Selective Anchor-Free (FSAF) Module
  - General concept
  - Network architecture
  - Ground-truth and loss
  - Online feature selection
- Experiments
- Qualitative Results

Experiment

- **Data**
  - **COCO Dataset, train set:** trainval35k**, validation set:** minival**, test set:** test-dev
- **Ablation study**
  - **Train on** trainval35k**, evaluate on** minival
  - **ResNet-50 as backbone network**
- **Runtime analysis**
  - **Train on** trainval35k**, evaluate on** minival
  - **Run on a single Titan X with CUDA 9 and CUDNN 7**
- **Compare to state of the art**
  - **Train on** trainval35k **with 1.5x iterations, evaluate on** minival

Experiments

Ablation study

	Anchor-based	Anchor-free		AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
		Heuristic feature selection	Online feature selection						
RetinaNet	✓			35.7	54.7	38.5	19.5	39.9	47.5
Ours		✓		34.7	54.0	36.4	19.0	39.0	45.8
			✓	35.9	55.0	37.9	19.8	39.6	48.2
	✓	✓		36.1	55.6	38.7	19.8	39.7	48.9
	✓		✓	**37.2**	**57.2**	**39.4**	**21.0**	**41.2**	**49.7**

Experiments

Ablation study

	Anchor-based	Anchor-free		AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
		Heuristic feature selection	Online feature selection						
RetinaNet	✓			35.7	54.7	38.5	19.5	39.9	47.5
Ours		✓		34.7	54.0	36.4	19.0	39.0	45.8
			✓	35.9	55.0	37.9	19.8	39.6	48.2
	✓	✓		36.1	55.6	38.7	19.8	39.7	48.9
	✓		✓	**37.2**	**57.2**	**39.4**	**21.0**	**41.2**	**49.7**

Anchor-free branches only with heuristic feature selection are not able to compete with anchor-based counterparts due to less parameters.

Experiments

Ablation study

	Anchor-based	Anchor-free		AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
		Heuristic feature selection	Online feature selection						
RetinaNet	✓			35.7	54.7	38.5	19.5	39.9	47.5
Ours		✓		34.7	54.0	36.4	19.0	39.0	45.8
			✓	35.9	55.0	37.9	19.8	39.6	48.2
	✓	✓		36.1	55.6	38.7	19.8	39.7	48.9
	✓		✓	**37.2**	**57.2**	**39.4**	**21.0**	**41.2**	**49.7**

Online feature selection is essential to overcome the parameter disadvantage!

Experiments

Ablation study

	Anchor-based	Anchor-free		AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
		Heuristic feature selection	Online feature selection						
RetinaNet	✓			35.7	54.7	38.5	19.5	39.9	47.5
Ours		✓		34.7	54.0	36.4	19.0	39.0	45.8
			✓	35.9	55.0	37.9	19.8	39.6	48.2
	✓	✓		36.1	55.6	38.7	19.8	39.7	48.9
	✓		✓	**37.2**	**57.2**	**39.4**	**21.0**	**41.2**	**49.7**

Online feature selection also guarantees anchor-free and anchor-based branches to work well together.

Experiments

Ablation study

Class name	AP improvement
Sports ball	+8.4
Tie	+5.9
Hair drier	+5.2
Kite	+5.1
Snowboard	+4.6
Skis	+4.3
Toothbrush	+3.9
Carrot	+3.8
Keyboard	+3.5

Experiments

Runtime analysis

Backbone	Method	AP	AP_{50}	Runtime (ms/im)
ResNet-50	RetinaNet	35.7	54.7	131
	Ours(FSAF)	35.9	55.0	107
	Ours(AB+FSAF)	37.2	57.2	138
ResNet-101	RetinaNet	37.7	57.2	172
	Ours(FSAF)	37.9	58.0	148
	Ours(AB+FSAF)	39.3	59.2	180
ResNeXt-101	RetinaNet	39.8	59.5	356
	Ours(FSAF)	41.0	61.5	288
	Ours(AB+FSAF)	41.6	62.4	362

Experiments

Runtime analysis

Backbone	Method	AP	AP_{50}	Runtime (ms/im)
ResNet-50	RetinaNet	35.7	54.7	131
	Ours(FSAF)	35.9	55.0	107
	Ours(AB+FSAF)	37.2	57.2	138
ResNet-101	RetinaNet	37.7	57.2	172
	Ours(FSAF)	37.9	58.0	148
	Ours(AB+FSAF)	39.3	59.2	180
ResNeXt-101	RetinaNet	39.8	59.5	356
	Ours(FSAF)	41.0	61.5	288
	Ours(AB+FSAF)	41.6	62.4	362

Experiments

Runtime analysis

Backbone	Method	AP	AP_{50}	Runtime (ms/im)
ResNet-50	RetinaNet	35.7	54.7	131
	Ours(FSAF)	35.9	55.0	107
	Ours(AB+FSAF)	37.2	57.2	138
ResNet-101	RetinaNet	37.7	57.2	172
	Ours(FSAF)	37.9	58.0	148
	Ours(AB+FSAF)	39.3	59.2	180
ResNeXt-101	RetinaNet	39.8	59.5	356
	Ours(FSAF)	41.0	61.5	288
	Ours(AB+FSAF)	41.6	62.4	362

Experiments

Comparison with state-of-the-art single-shot detectors

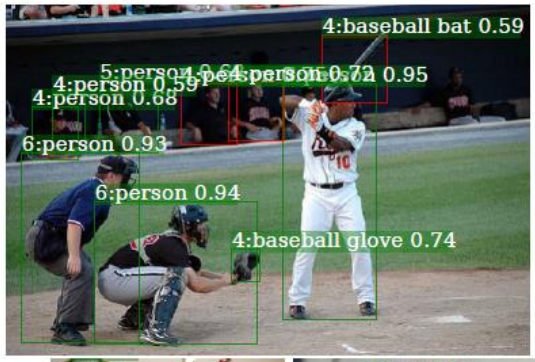
Method	Backbone	AP	AP_{50}	AP_S	AP_M	AP_L
YOLOv2	DarkNet-19	21.6	44.0	5.0	22.4	35.5
SSD513	ResNet-101	31.2	50.4	10.2	34.5	49.8
RefineDet512		36.4	57.5	16.6	39.9	51.4
RefineDet(ms)		41.8	62.9	25.6	45.1	**54.1**
RetinaNet800		39.1	59.1	21.8	42.7	50.2
Ours800		40.9	61.5	24.0	44.2	51.3
Ours(ms)		**42.8**	**63.1**	**27.8**	**45.5**	53.2
CornerNet511	Hourglass-104	40.5	56.5	19.4	42.7	53.9
CornerNet(ms)		42.1	57.8	20.8	44.8	56.7
Ours800	ResNeXt-101	42.9	63.8	26.6	46.2	52.7
Ours(ms)		44.6	65.2	29.7	47.1	54.6

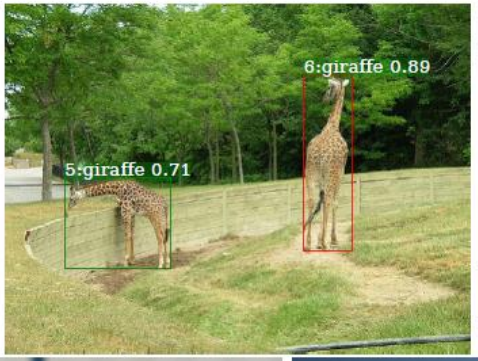
Overview

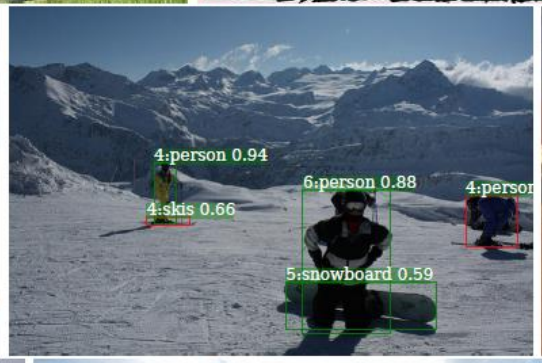
- Background
- Motivation
- Feature Selective Anchor-Free (FSAF) Module
  - General concept
  - Network architecture
  - Ground-truth and loss
  - Online feature selection
- Experiments
- Qualitative Results

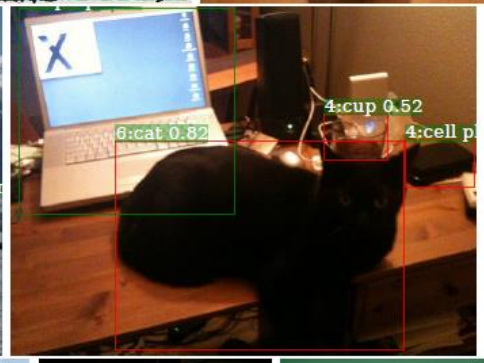
Qualitative Results

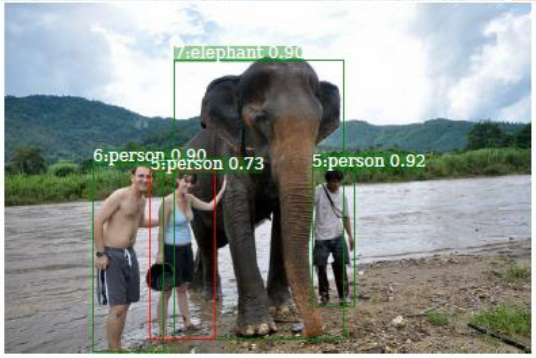
Qualitative Results

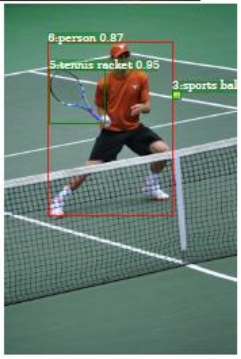












References

Ren, Shaoqing, et al. "Faster r-cnn: Towards real-time object detection with region proposal networks." *Advances in neural information processing systems*. 2015.

Liu, Wei, et al. "Ssd: Single shot multibox detector." *European conference on computer vision*. Springer, Cham, 2016.

Lin, Tsung-Yi, et al. "Feature pyramid networks for object detection." *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2017.

Lin, Tsung-Yi, et al. "Focal loss for dense object detection." *Proceedings of the IEEE international conference on computer vision*. 2017.

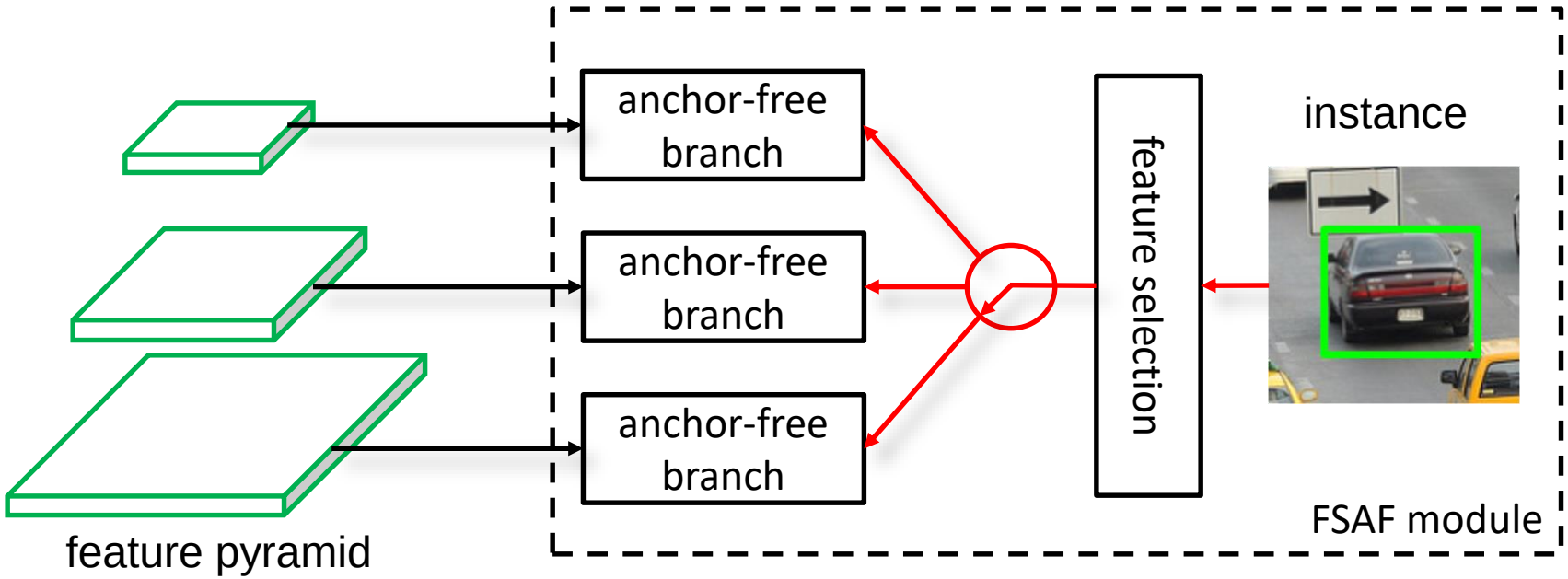
Huang, Lichao, et al. "Densebox: Unifying landmark localization with end to end object detection." *arXiv preprint arXiv:1509.04874* (2015).

Yu, Jiahui, et al. "Unitbox: An advanced object detection network." *Proceedings of the 24th ACM international conference on Multimedia*. ACM, 2016.

Zhao, Qijie, et al. "M2Det: A Single-Shot Object Detector based on Multi-Level Feature Pyramid Network." *arXiv preprint arXiv:1811.04533* (2018).

Takeaway

Feature selection based on semantics is the key!



Q&A