Object Classification, Localization and Detection Part 3: SSD and M2DeT

CS8004: Deep Learning and Applications

Today

- Last time YOLO, a single shot detector that trains a single CNN once only for all the objects in the scene.

- Today we shall discuss another single shot detector
  - Single Shot MultiBox Detector by Liu et al. (2015).

SSD: Single Shot MultiBox Detector

- Developed by Liu et al (December 2015) and as reported in their paper –
- Faster than Yolo, as accurate as two stage methods like Faster R-CNN.
- Predicts categories and box offsets.
- Uses small convolutional filters applied to feature maps.
- Makes predictions using feature maps of different scales

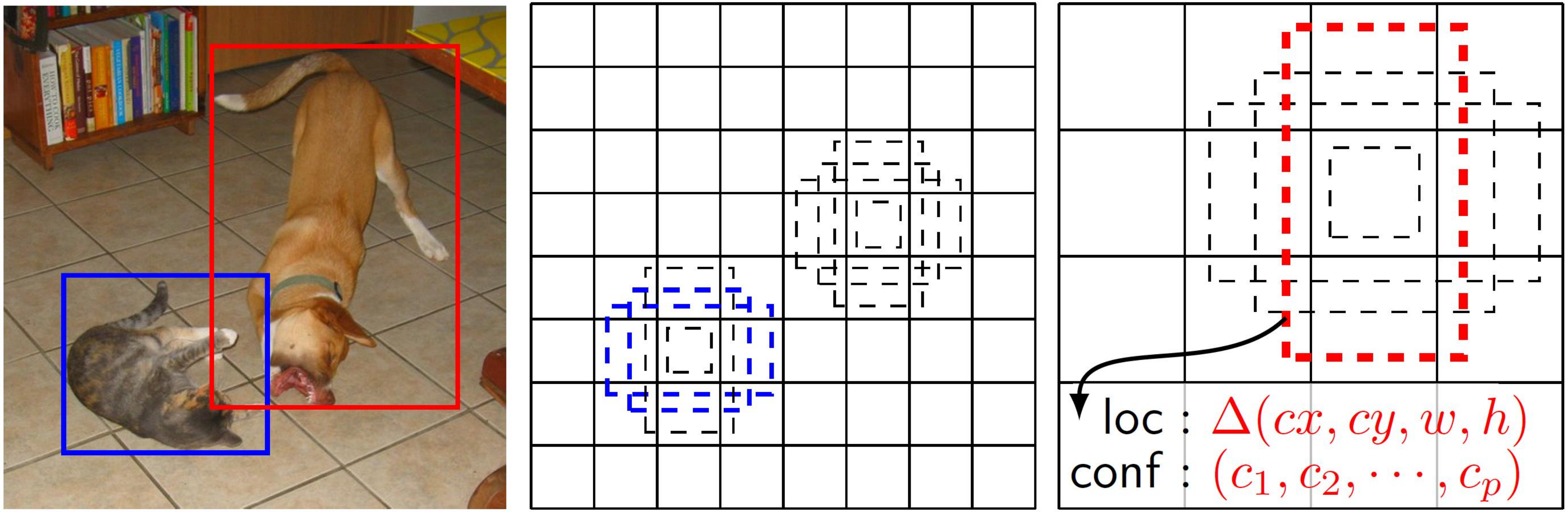
SSD Framework

- SSD only needs an input image and ground truth boxes for each object during training.
- Through CNN a small set (e.g. 4) of default boxes of different aspect ratios is evaluated at each location
- This is done in several feature maps with different scales (e.g. 8 X 8 and 4 X 4).

SSD Framework

- For each default box, both the shape offsets and the confidences for all object categories are predicted.
- At training time, these default boxes are matched with the ground truth boxes.
- The model loss is a weighted sum between localization loss and confidence loss.

SSD Framework



(a) Image with GT boxes (b) 8×8 feature map (c) 4×4 feature map

Two default boxes with the cat and one with the dog are matched, which are treated as positives and the rest as negatives.

SSD Model

- The SSD approach is based on a feed-forward convolutional network that produces a fixed-size collection of bounding boxes and scores for the presence of object class instances in those boxes.

- This is followed by a non-maximum suppression step to produce the final detections.

- The early network layers are based on a standard architecture used for high quality image classification (truncated before any classification layers). This is called base network.

SSD Model …

- Then the auxiliary structure is added to the network to produce detections with the following key features:
  - Multi-scale feature maps for detection,
  - Convolutional predictors for detection,
  - Default boxes and aspect ratios.

SSD Model vs YOLO Model

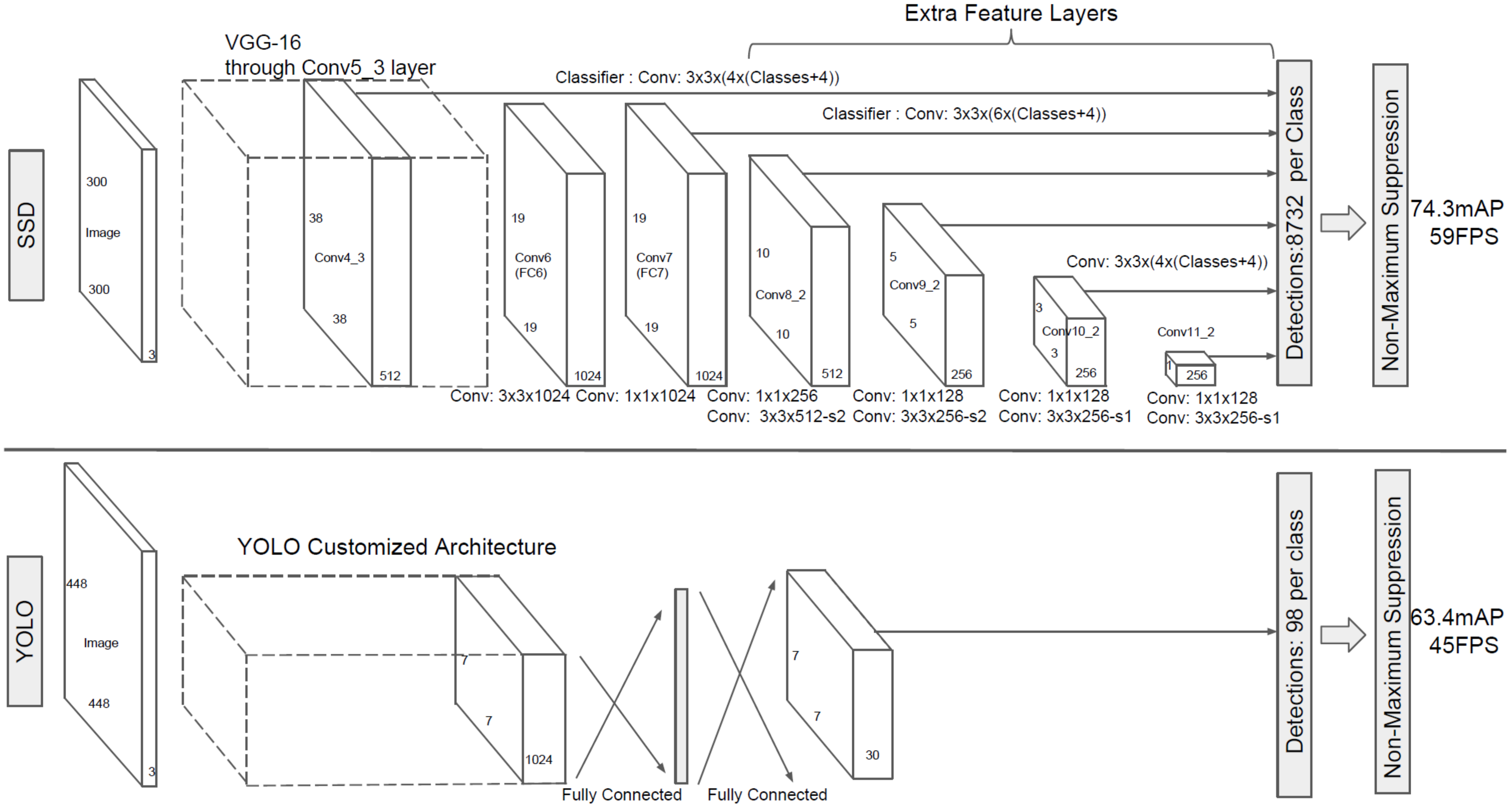


Image source :Liu et al., Single Shot MultiBox Detector, December 2015.

Default Box and Class Predictions

- 3X3 convolution applied to each cell shown here. Each cell predicts (say) 4 default box dimensions and a class score (A 21-size for 20 classes, one for no object)

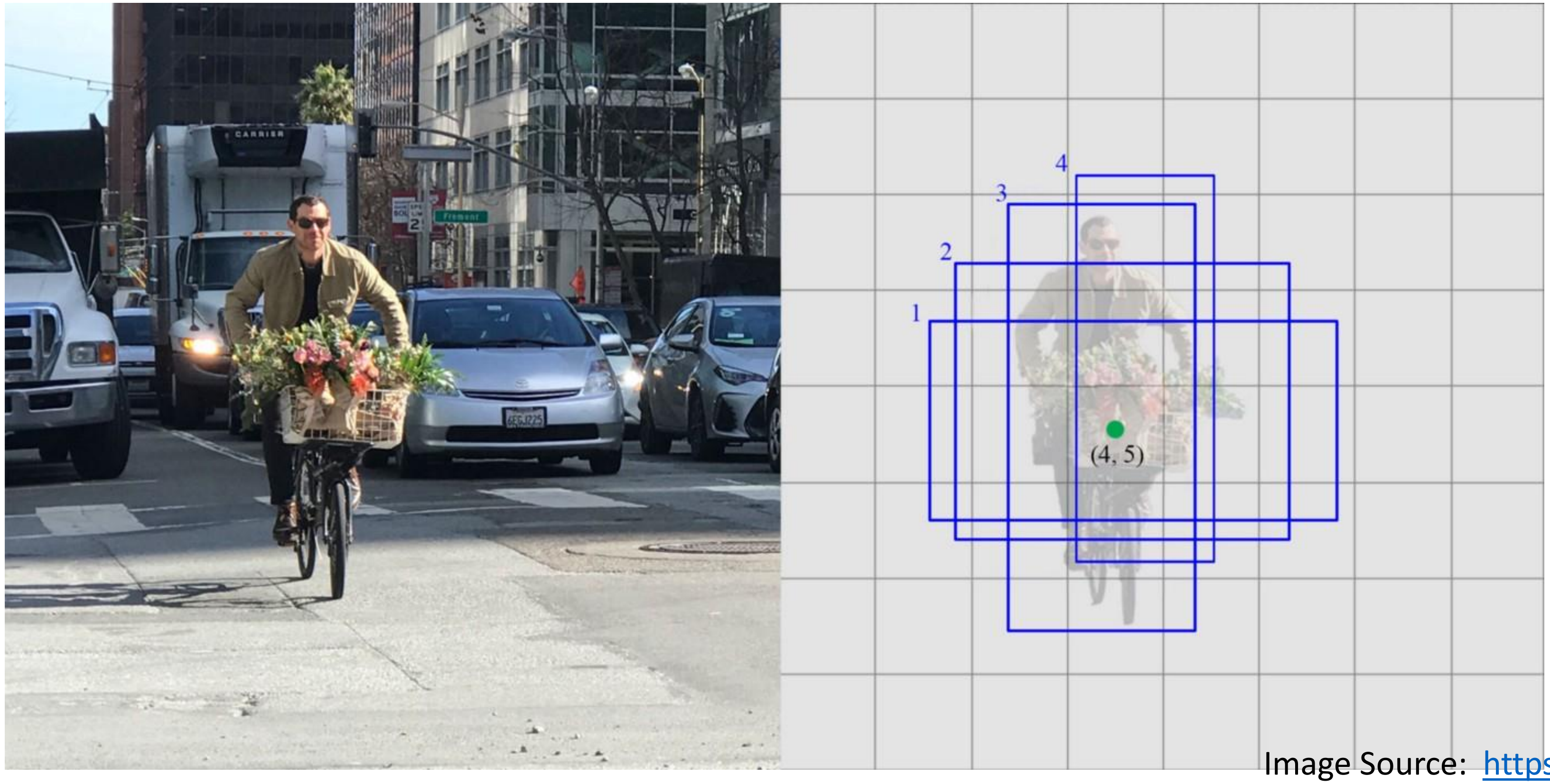
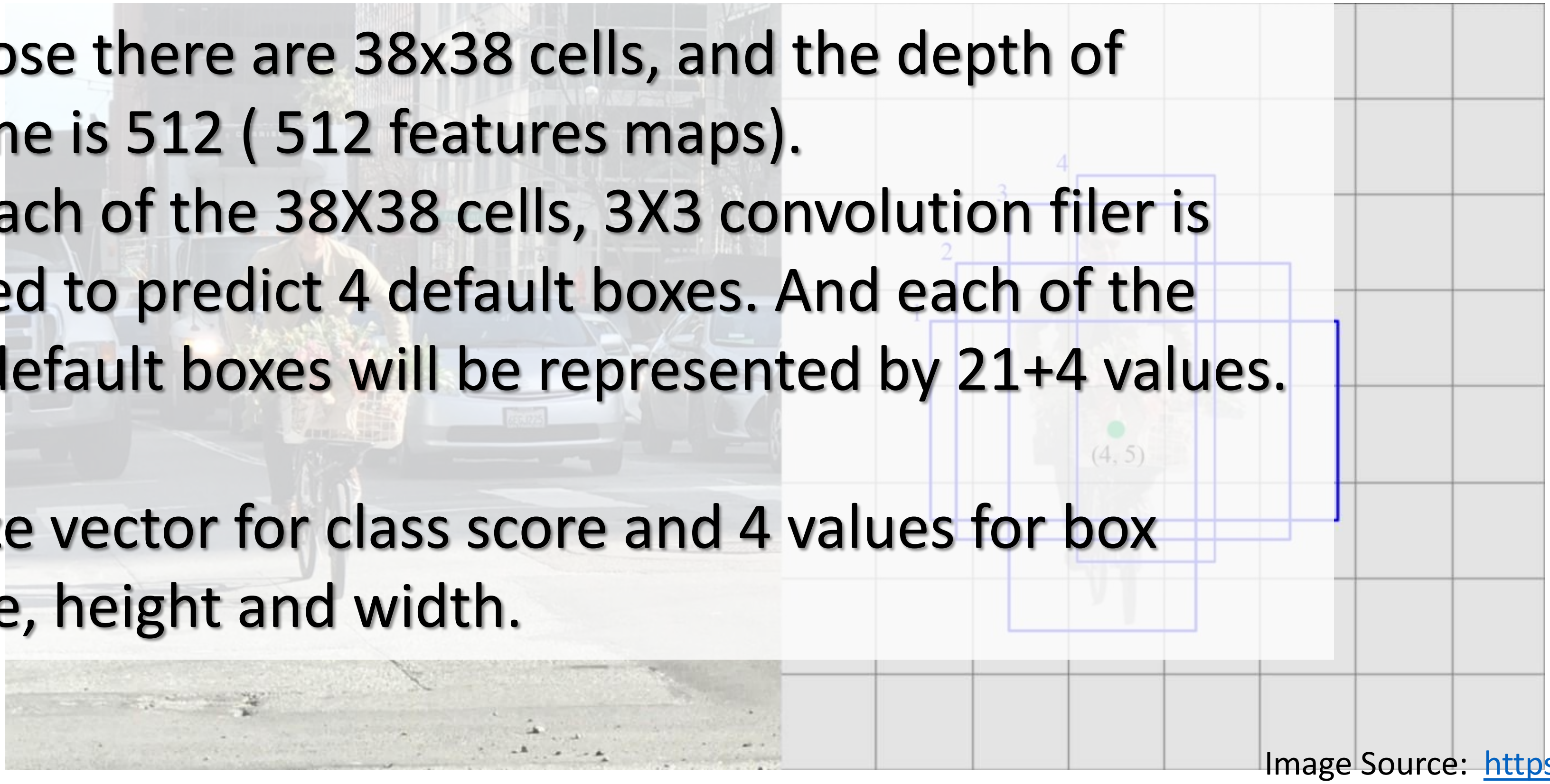


Image Source: https://medium.com/@jonathan_hui

Default Box and Class Predictions

- 3X3 convolution applied to each cell shown here. Each cell predicts (say) 4 default box dimensions and a class score (A 21-size for 20 classes, one for no object)



Suppose there are 38x38 cells, and the depth of Volume is 512 (512 features maps).
On each of the 38X38 cells, 3X3 convolution filer is applied to predict 4 default boxes. And each of the four default boxes will be represented by 21+4 values.

21-size vector for class score and 4 values for box centre, height and width.

Image Source: https://medium.com/@jonathan_hui

Convolutional Predictors for Detection

- Each added feature layer (or optionally an existing feature layer from the base network) can produce a fixed set of detection predictions using a set of convolutional filters.

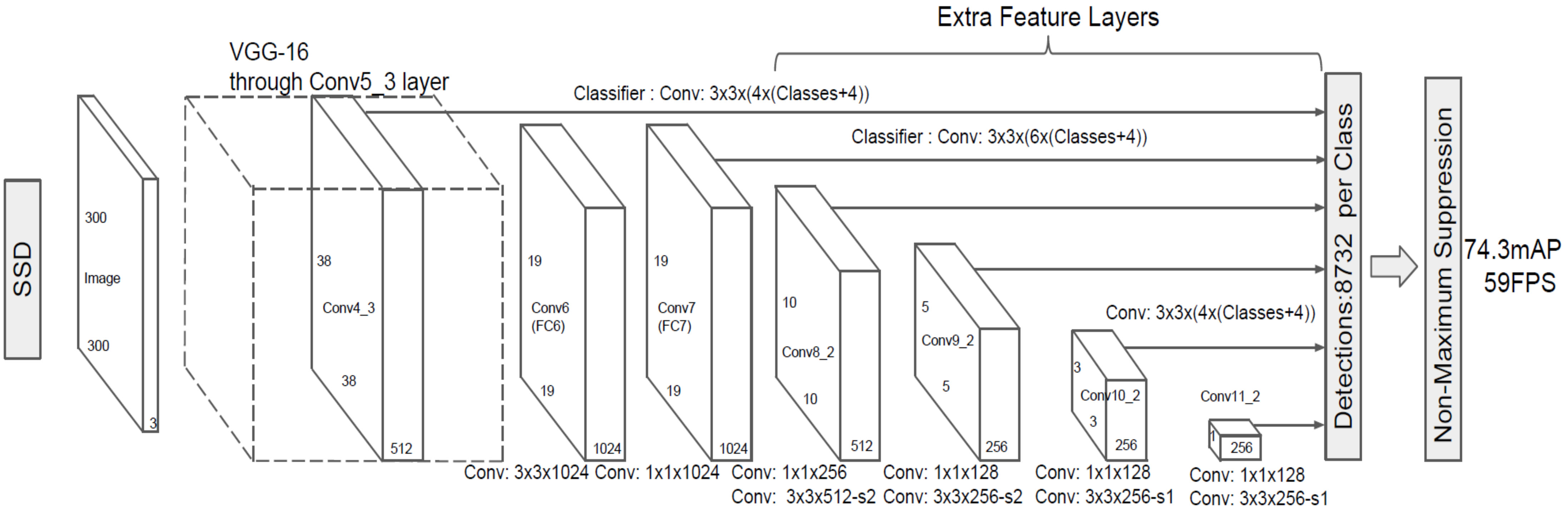


Image source :Liu et al., Single Shot MultiBox Detector, December 2015.

Convolutional Predictors for Detection

- For a feature layer of size m X n with p channels, the basic element for predicting parameters of a potential detection is a 3 X 3 X p small kernel.
- This element produces either a score for a category, or a shape offset relative to the default box coordinates.
- At each of the m X n locations where the kernel is applied, it produces an output value.
- Remember that YOLO uses an intermediate fully connected layer instead of a convolutional filter for this step.

Default Boxes and Aspect Ratios

- A set of default bounding boxes with each feature map cell, for multiple feature maps at the top of the network.

- At each feature map cell, we predict the offsets relative to the default box shapes in the cell, as well as the per-class scores that indicate the presence of a class instance in each of those boxes.

Default Boxes and Aspect Ratios

- For each box at a given location, a c class scores and the 4 offsets relative to the original default box shape are computed (Total k boxes).

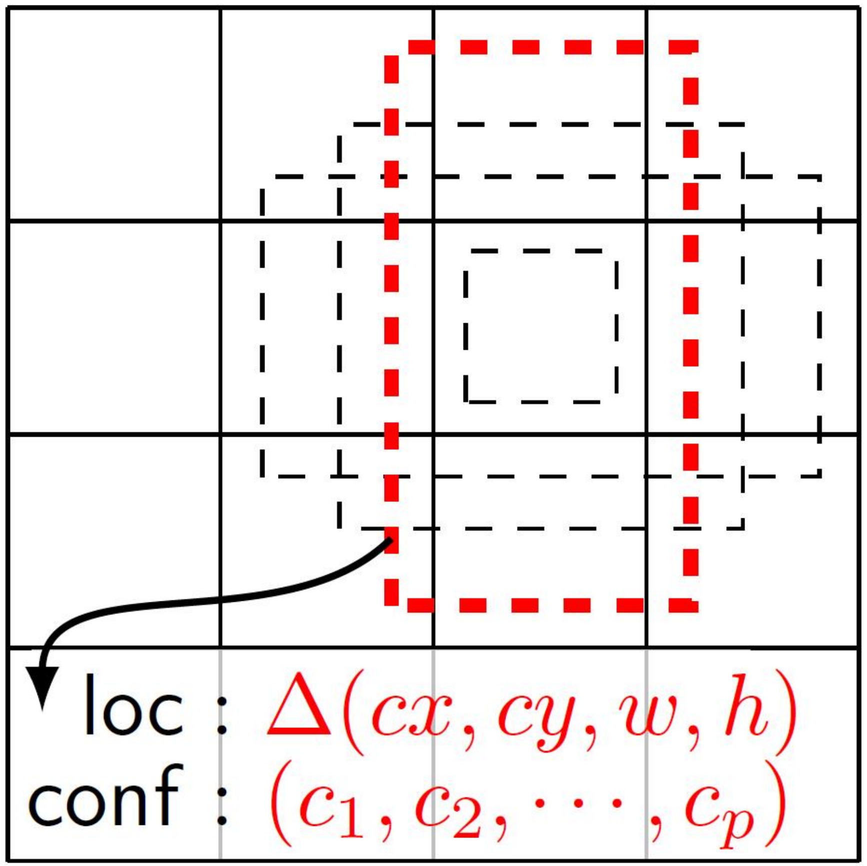
- Default box is like the Anchor box of YOLO.

- This results in a total of $(c + 4)k$ filters that are applied around each location in the feature map

- This yields $(c + 4)kmn$ outputs for a m X n feature map.

Default Boxes and Aspect Ratios

- Allowing different default box shapes in several feature maps let us efficiently discretize the space of possible output box shapes.



(c) 4×4 feature map

Image source :Liu et al., Single Shot MultiBox Detector, December 2015.

Feature Maps of Different Scales

- Lower resolution feature maps detect larger scale objects and higher resolution feature maps detect lower scale objects.

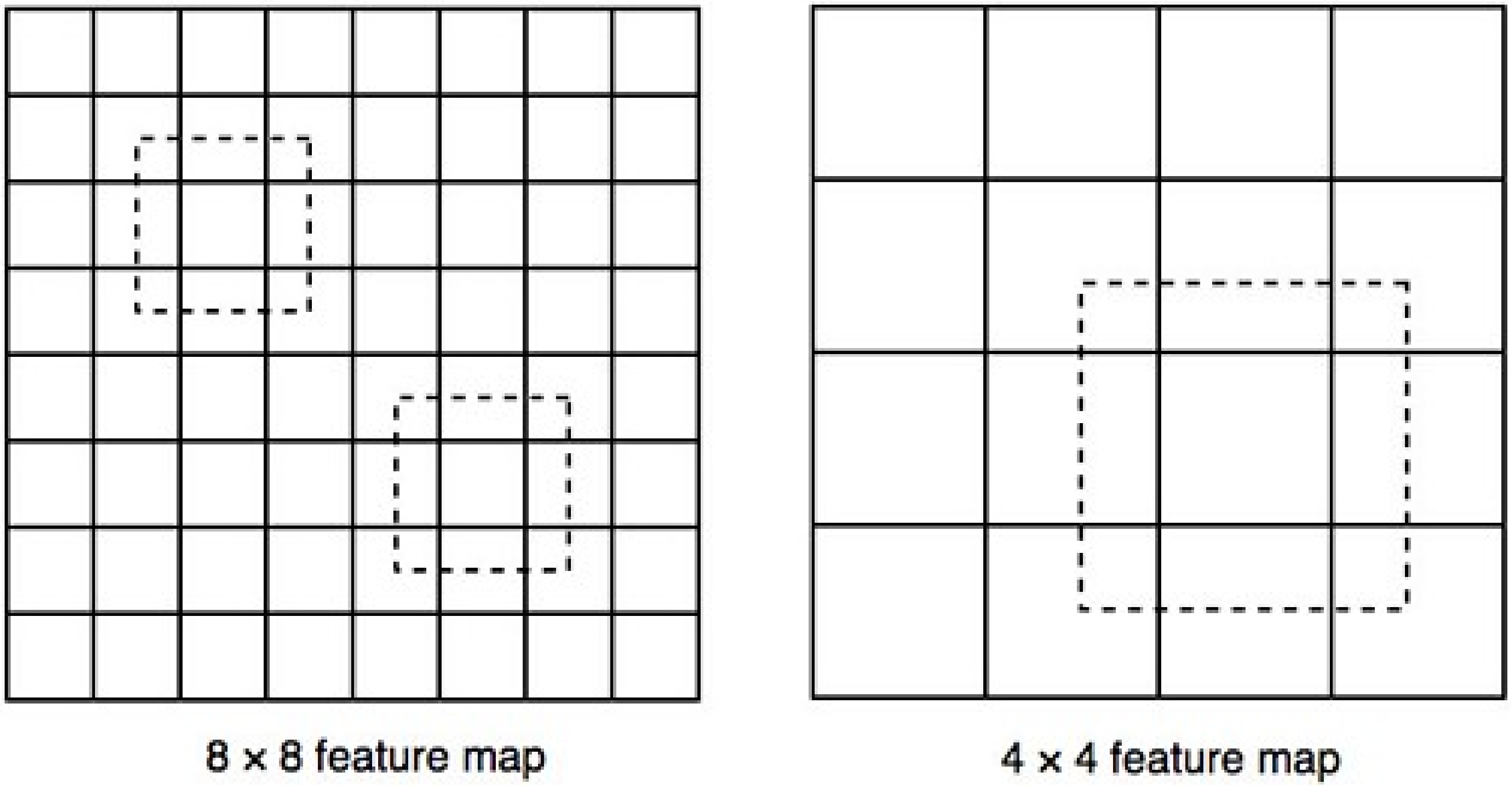


Image Source: https://medium.com/@jonathan_hui

Training

- During training one needs to determine which default boxes correspond to a ground truth detection and train the network accordingly.

- For each ground truth box, selections are made from default boxes that vary over location, aspect ratio, and scale.

- Matches each ground truth box to the default box with the best Jaccard similarity measure (>= 0.5 selected). It is the same as IoU discussed before.

Feature Maps of Different Scales

- Lower resolution feature maps detect larger scale objects and higher resolution feature maps detect lower scale objects.

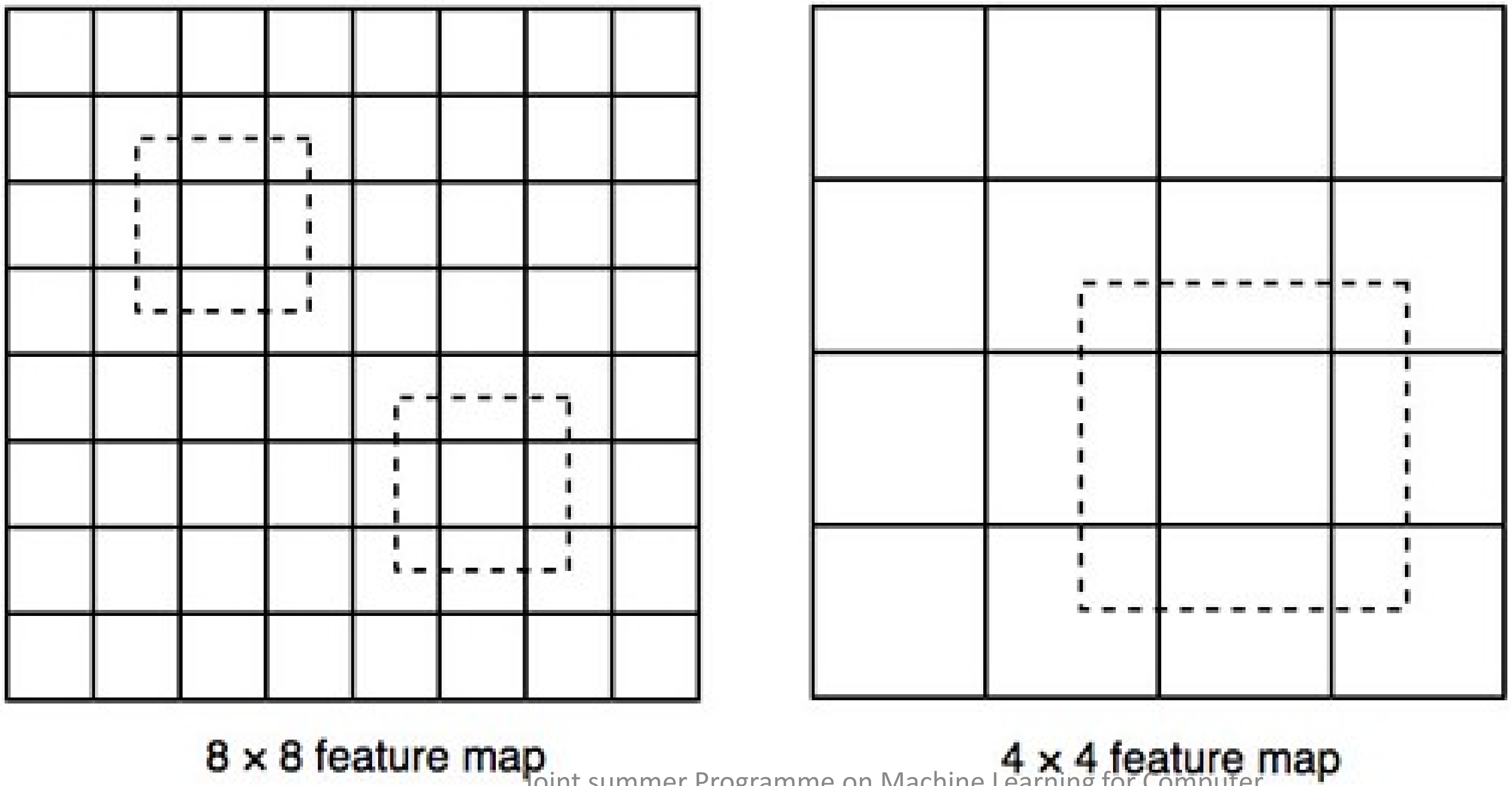
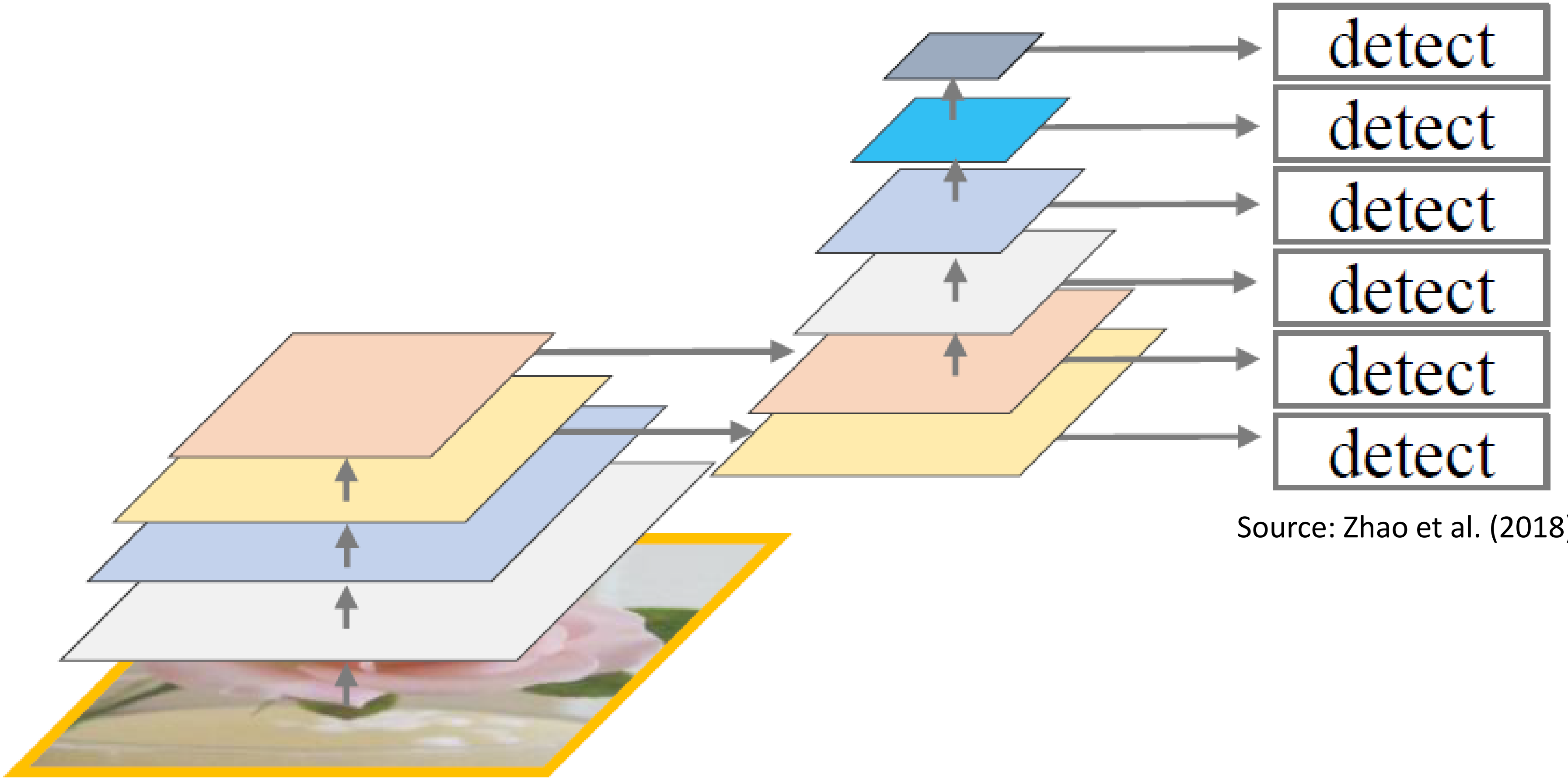


Image Source: https://medium.com/@jonathan_hui

Multi-Scale Feature Maps for Detection

- Convolutional feature layers are added to the end of the truncated base network.
- These layers decrease in size progressively and allow predictions of detections at multiple scales.
- The convolutional model for predicting detections is different for each feature layer.

Feature Pyramid of Different Scales



Source: Zhao et al. (2018)

Results

Method	mAP	FPS	batch size	# Boxes	Input resolution
Faster R-CNN (VGG16)	73.2	7	1	~ 6000	$\sim 1000 \times 600$
Fast YOLO	52.7	155	1	98	448×448
YOLO (VGG16)	66.4	21	1	98	448×448
SSD300	74.3	46	1	8732	300×300
SSD512	76.8	19	1	24564	512×512
SSD300	74.3	59	8	8732	300×300
SSD512	76.8	22	8	24564	512×512

Source: https://towardsdatascience.com

Real Time Performance Evaluation

- Let us check out their real time performance

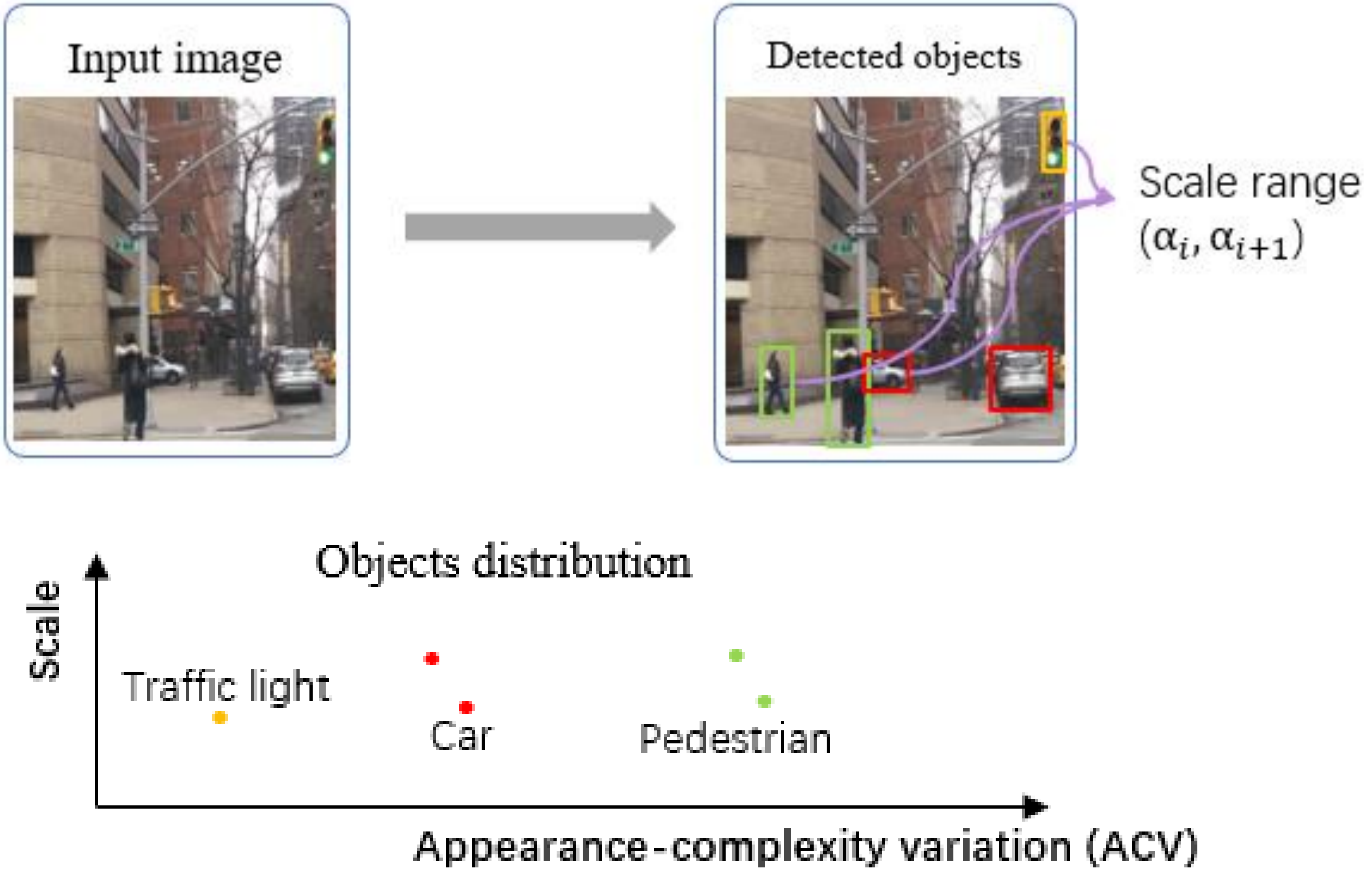
Recent Single Shot Detectors

- Li et al., FSSD: Feature Fusion Single Shot Multibox Detector
- Fu et al., DSSD, 2017: Deconvolutional Single Shot Detector
- Lin, Goyal, et al., 2017, RetinaNet
- **Zhao et al, 2018, M2Det**
- **Tan et al., EfficientDet, 2020**

M2DeT

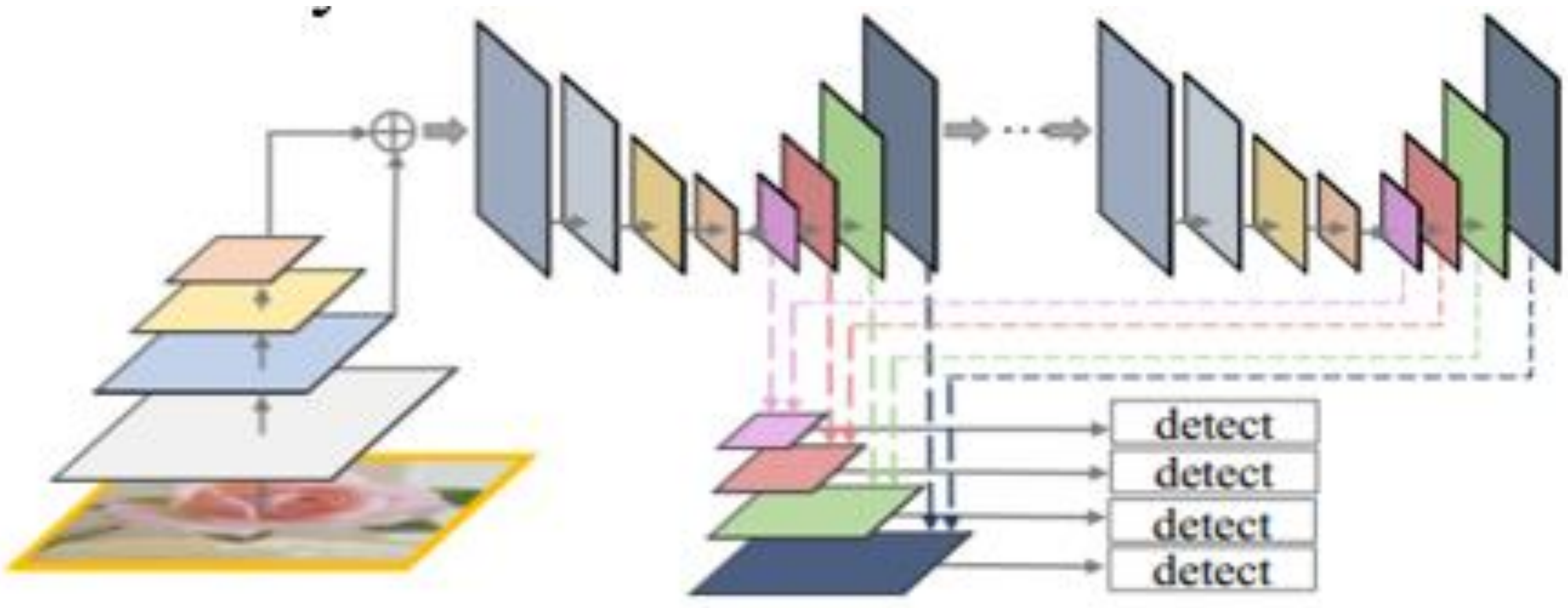
- Zhao et al. (2018) introduced a new single shot object detector based on multi-level feature pyramid network.
- Apart from scale variation, appearance-complexity variation should also be considered for the object detection task.
- Object instances with similar size can be quite different.
- M2Det adds a new dimension to multi-scale detection - **multi-level learning**.
- Deeper level learns features for objects with more appearance-complexity variation (e.g., pedestrian in a road), while shallower level learns features for more simplistic objects(e.g., traffic light).

M2DeT



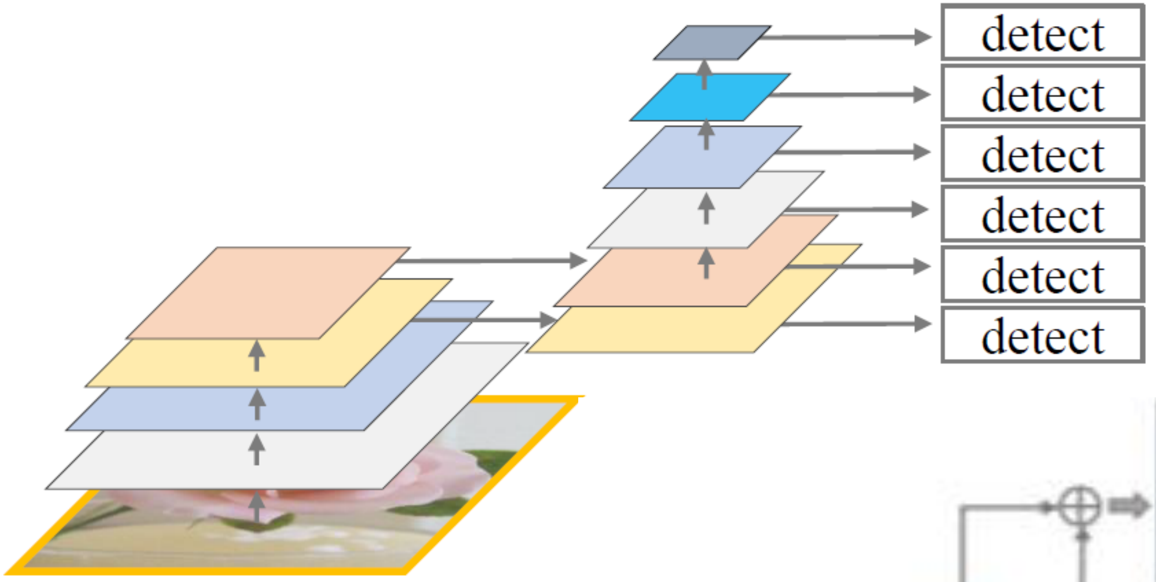
Source: Zhao et al. (2018)

m2DeT: Multi Level Features

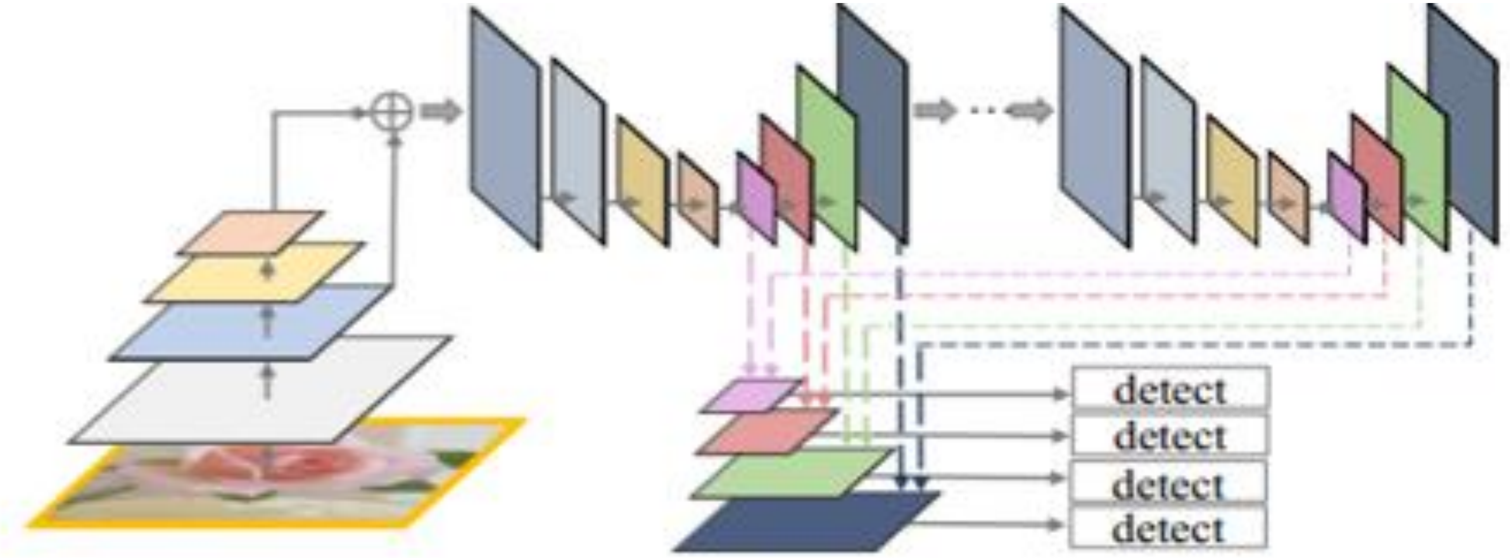


Source: Zhao et al. (2018)

m2DeT vs SSD : Feature Maps



SSD

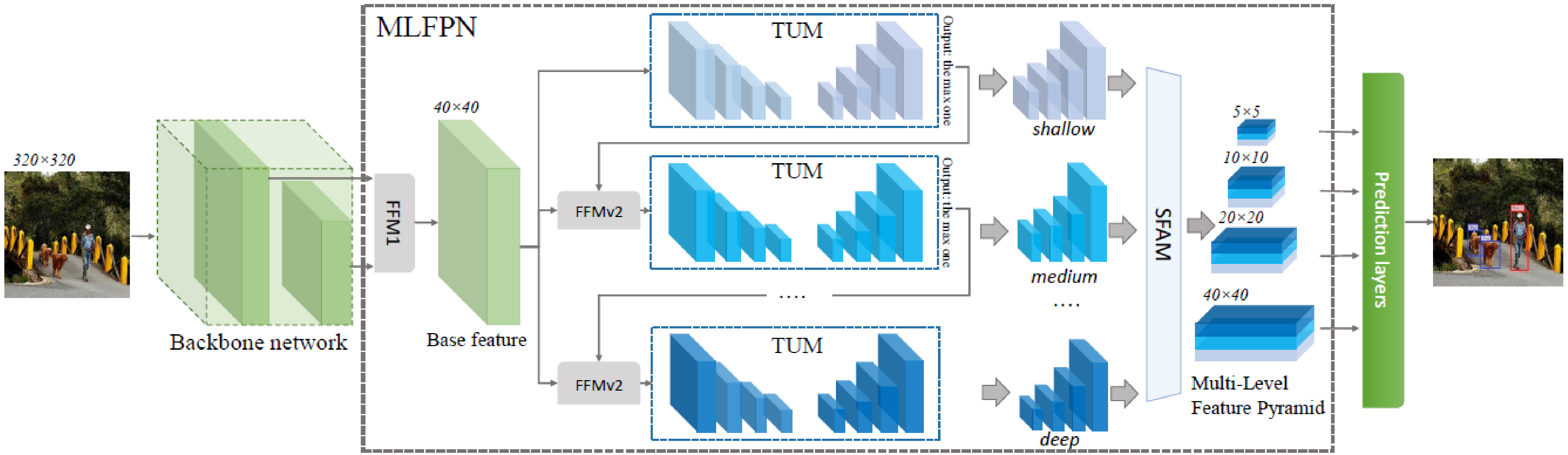


Source: Zhao et al. (2018)

M2DeT

M2DeT

- Three modules : Feature Fusion Module (FFM), Thinned U-shape Module (TUM) and Scale-wise Feature Aggregation Module (SFAM).



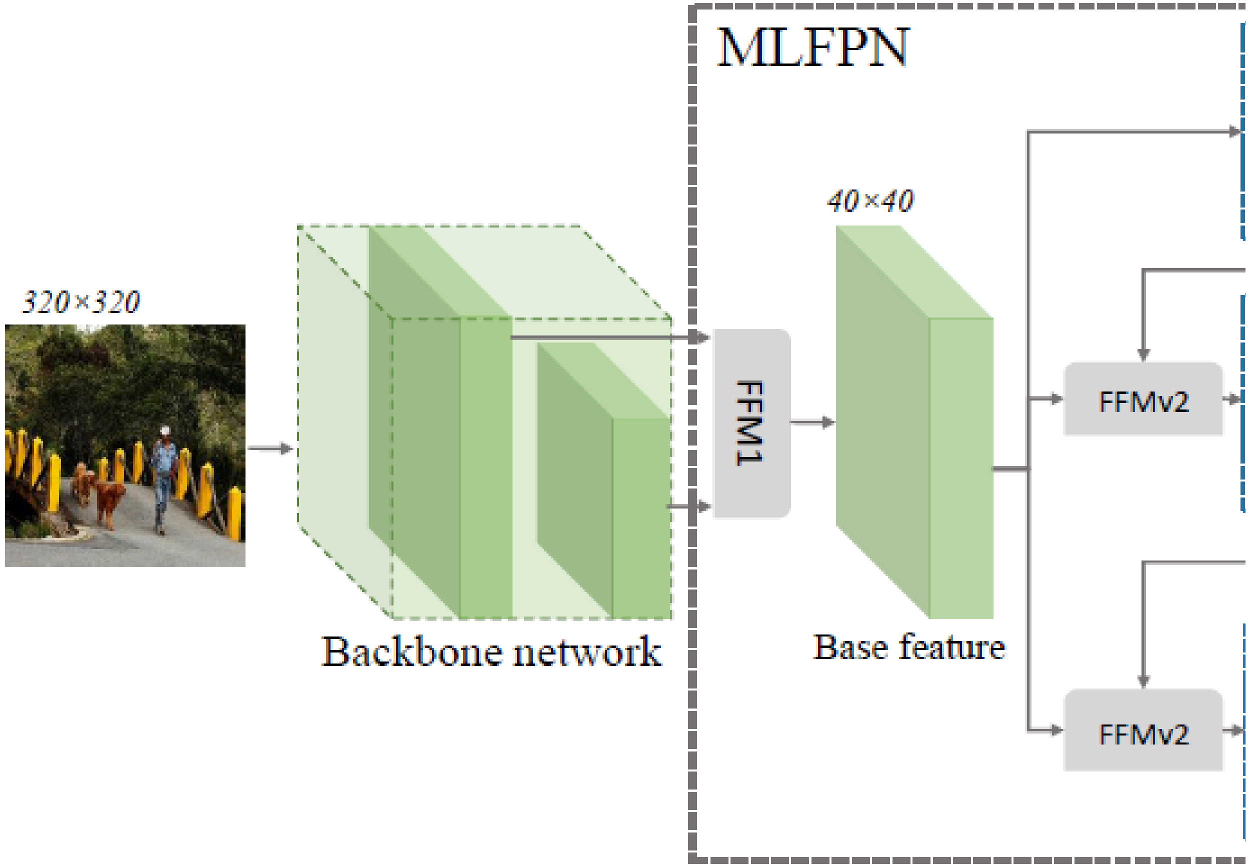
Source: Zhao et al. (2018)

[Real Time Performance](#)

M2DeT: Feature Fusion Module

- **FFMv1** enriches semantic information into base features by fusing feature maps of the backbone.

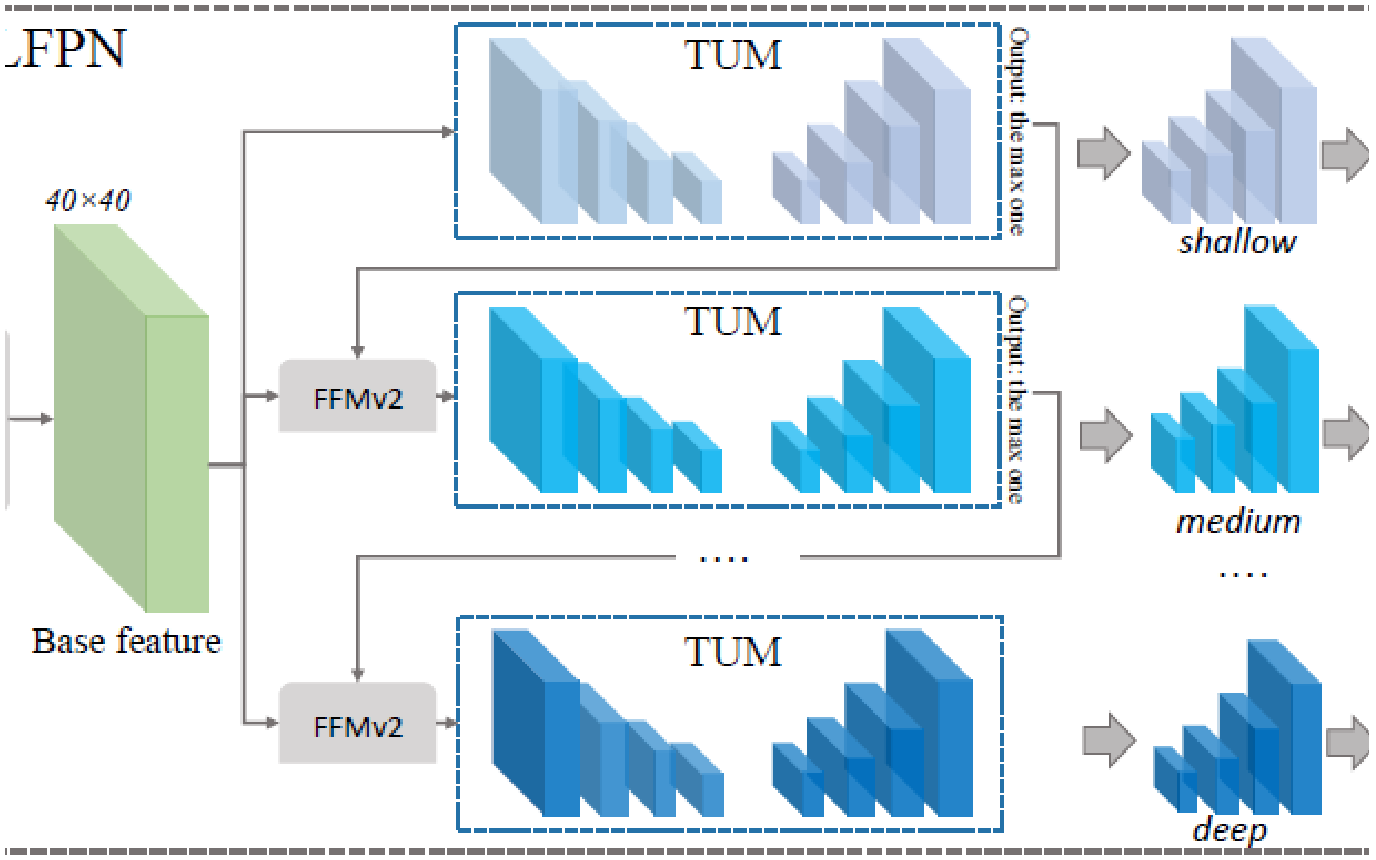
- **FFMv2** modules extract multi-level multiscale features together with TUMs.



Source: Zhao et al. (2018)

M2DeT: Thinned U-shape Module

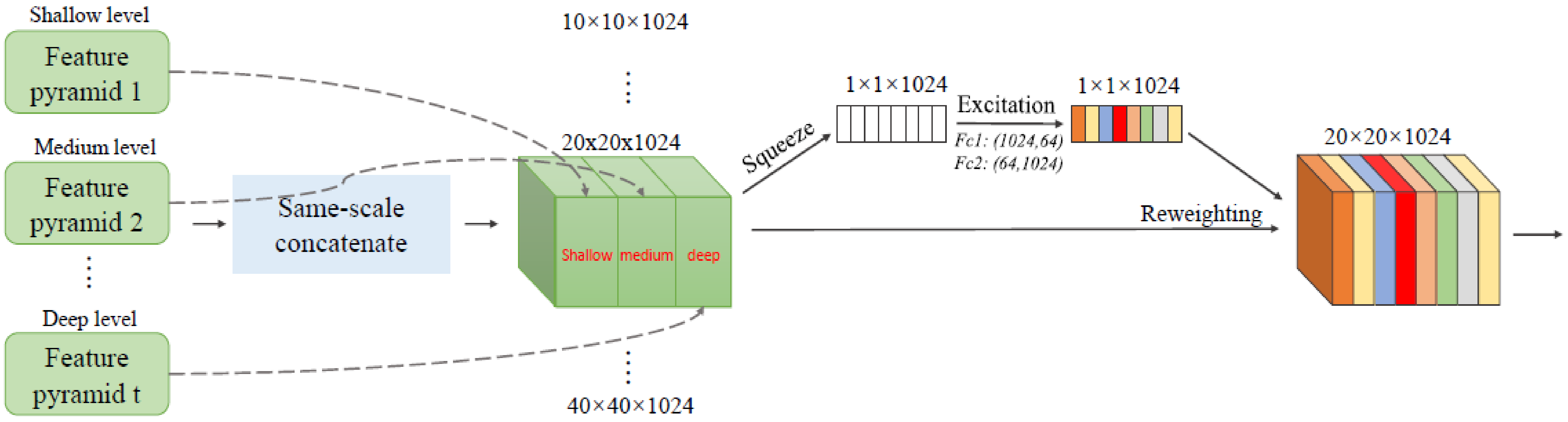
- Each **TUM** generates a group of multi-scale features.
- **TUM**s and **FFMv2**s together extract multi-level multiscale features.



Source: Zhao et al. (2018)

M2DeT: Scale-wise Feature Aggregation Module

- SFAM aggregates the multi-level multiscale features generated by TUMs into a multi-level feature pyramid



Source: Zhao et al. (2018)

M2DeT Performance

Method	Backbone	Input size	MultiScale	FPS	Avg. Precision, IoU: 0.5:0.95	0.5	0.75	Avg. Precision, Area: S	M	L
YOLOv3 (Redmon and Farhadi 2018)	DarkNet-53	608×608	False	19.8	33.0	57.9	34.4	18.3	35.4	41.9
SSD512* (Liu et al. 2016)	VGG-16	512×512	False	22	28.8	48.5	30.3	10.9	31.8	43.5
DSSD513 (Fu et al. 2017)	ResNet-101	513×513	False	5.5	33.2	53.3	35.2	13.0	35.4	51.1
RetinaNet500 (Lin et al. 2017b)	ResNet-101	~832×500	False	11.1	34.4	53.1	36.8	14.7	38.5	49.1
RefineDet512 (Zhang et al. 2018)	VGG-16	512×512	False	22.3	33.0	54.5	35.5	16.3	36.3	44.3
RefineDet512 (Zhang et al. 2018)	ResNet-101	512×512	True	-	41.8	62.9	45.7	25.6	45.1	54.1
CornerNet (Law and Deng 2018)	Hourglass	512×512	False	4.4	40.5	57.8	45.3	20.8	44.8	56.7
CornerNet (Law and Deng 2018)	Hourglass	512×512	True	-	42.1	57.8	45.3	20.8	44.8	56.7
M2Det (Ours)	VGG-16	512×512	False	18.0	37.6	56.6	40.5	18.4	43.4	51.2
M2Det (Ours)	VGG-16	512×512	True	-	42.9	62.5	47.2	28.0	47.4	52.8
M2Det (Ours)	ResNet-101	512×512	False	15.8	38.8	59.4	41.7	20.5	43.9	53.4
M2Det (Ours)	ResNet-101	512×512	True	-	**43.9**	**64.4**	**48.0**	**29.6**	**49.6**	**54.3**
RetinaNet800 (Lin et al. 2017b)	Res101-FPN	~1280×800	False	5.0	39.1	59.1	42.3	21.8	42.7	50.2
M2Det (Ours)	VGG-16	800×800	False	11.8	41.0	59.7	45.0	22.1	46.5	53.8
M2Det (Ours)	VGG-16	800×800	True	-	**44.2**	**64.6**	**49.3**	**29.2**	**47.9**	**55.1**

Source: Zhao et al. (2018)

Summary

- Object Detection is a growing area of research.
- YOLO and SSD are quite promising. RetinaNet, M2Det improve the performance for object detection in different scales.
- Still there are challenges –
  - Small objects
  - Irregularly shape objects
  - Applications to motion estimation, activity detection, pose detection, salient object detection etc.

References

- Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C. Y., & Berg, A. C. (2016, October). Ssd: Single shot multibox detector. In *European conference on computer vision* (pp. 21-37). Springer, Cham.
- Zhao, Q., Sheng, T., Wang, Y., Tang, Z., Chen, Y., Cai, L., & Ling, H. (2019, July). M2det: A single-shot object detector based on multi-level feature pyramid network. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 33, pp. 9259-9266).

References

- Girshick, R., Donahue, J., Darrell, T., & Malik, J. (2015). Region-based convolutional networks for accurate object detection and segmentation. *IEEE transactions on pattern analysis and machine intelligence*, *38*(1), 142-158. (first appeared in 2014).
- Girshick, R. (2015). Fast r-cnn. In Proceedings of the IEEE international conference on computer vision (pp. 1440-1448).
- Ren, S., He, K., Girshick, R., & Sun, J. (2015). Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems* (pp. 91-99).
- Redmon, J., Divvala, S., Girshick, R., & Farhadi, A. (2016). You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 779-788).
- Andrew Ng Course on Convolutional Neural Networks, Coursera (deeplearning. ai)