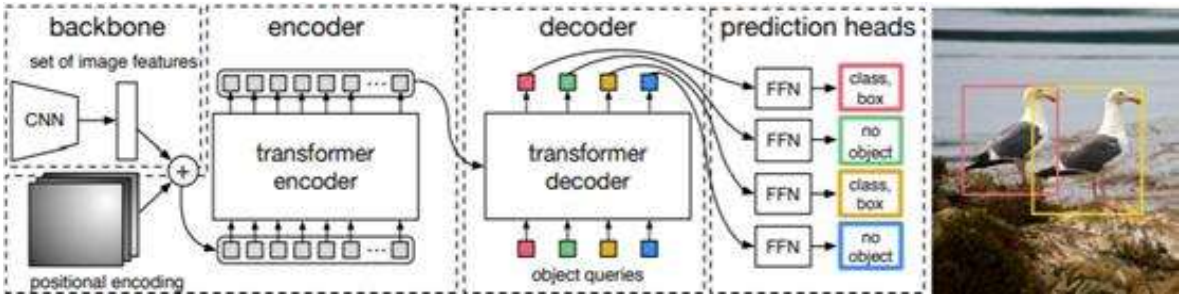
# Deformable DETR

Presentation by Jacob Pfaffenbichler, Sharon Barak, Tristan de Blegiers, Jacob O'Quinn

# Outline

- Introduction
- Related work:
  - Deformable convolution
  - Multi-Scale features
- Method:
  - Deformable attention module
  - Multi-Scale Deformable attention module
- Workflow:
  - Obtaining feature maps
  - Deformable Transformer Encoder/Decoder
- Results

# Introduction




DETR main problems:

- Its inability to converge quickly
- Poor performances on small objects

Solution → Deformable DETR

# Brief Review of DETR Architecture

1) Uses CNN backbone to to produce input feature maps.
2) Multi-Head Self-Attention module in the Transformer Encoder-Decoder.
3) FFN to predict bounding box locations.
4) Loss computed by bipartite matching.

# Related Work

Two main categories:

- A deformable attention module inspired by deformable convolution.
- A multi-scale feature map aggregator attention mechanism.

# Deformable Convolutions

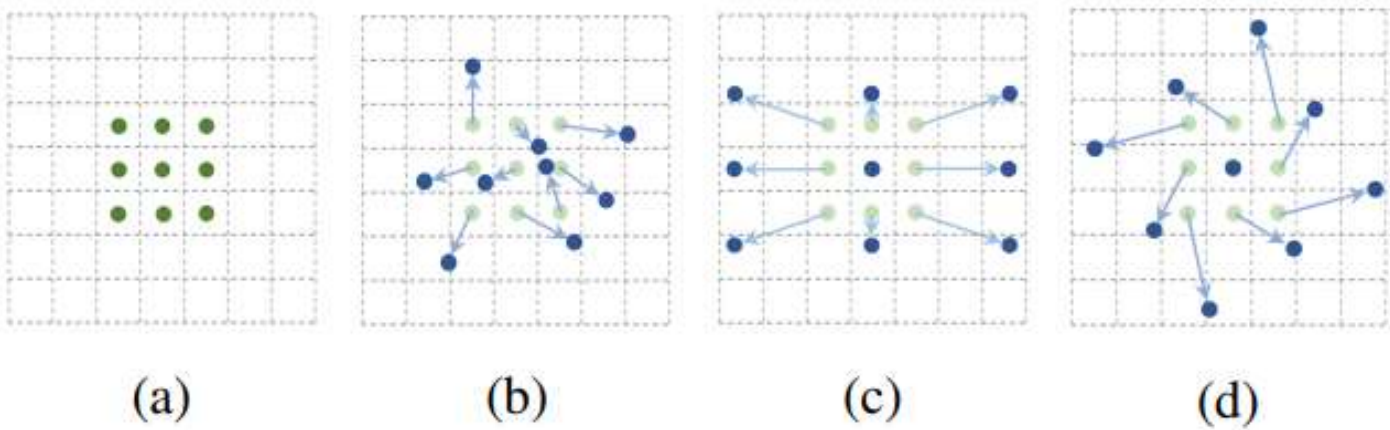
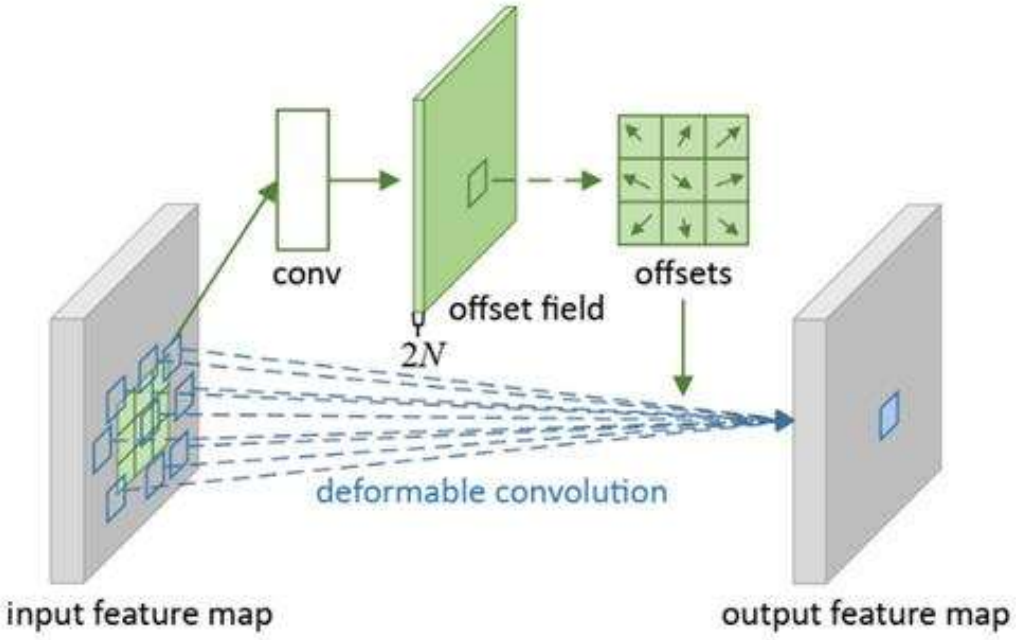
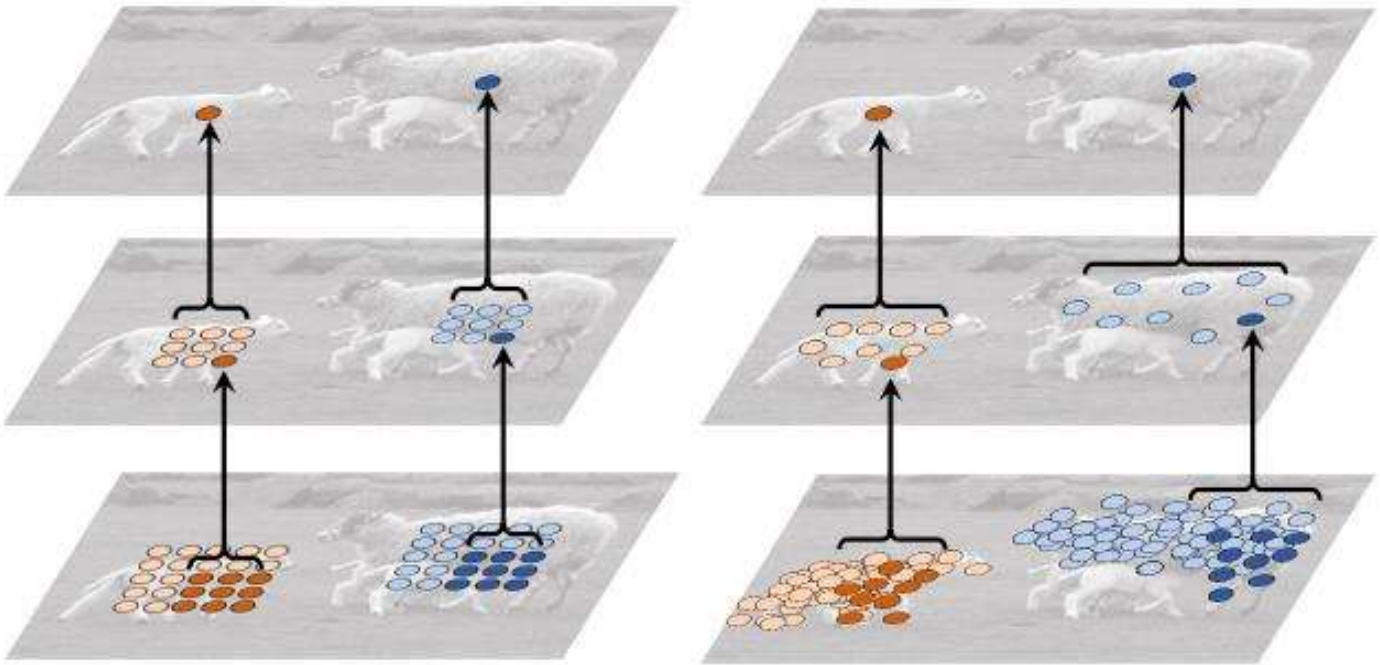



Figure 3. Sampling positions for Normal Convolution Operation (a) compared to Deformable Convolution Operation(b,c,d).

# Deformable Convolutions



# Deformable Convolutions



(a) standard convolution (b) deformable convolution

# Multi-Scale Features

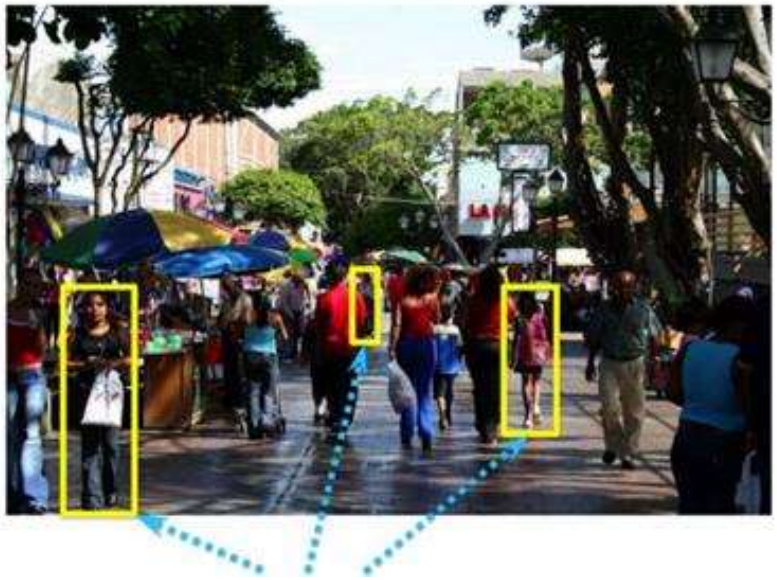
One of the main difficulties in object detection is detecting them at various scales.

Previous works attempt to automatically detect the scale or search at multiple scales.

Examples:

SIFT

FPN



# Multi-Scale Features

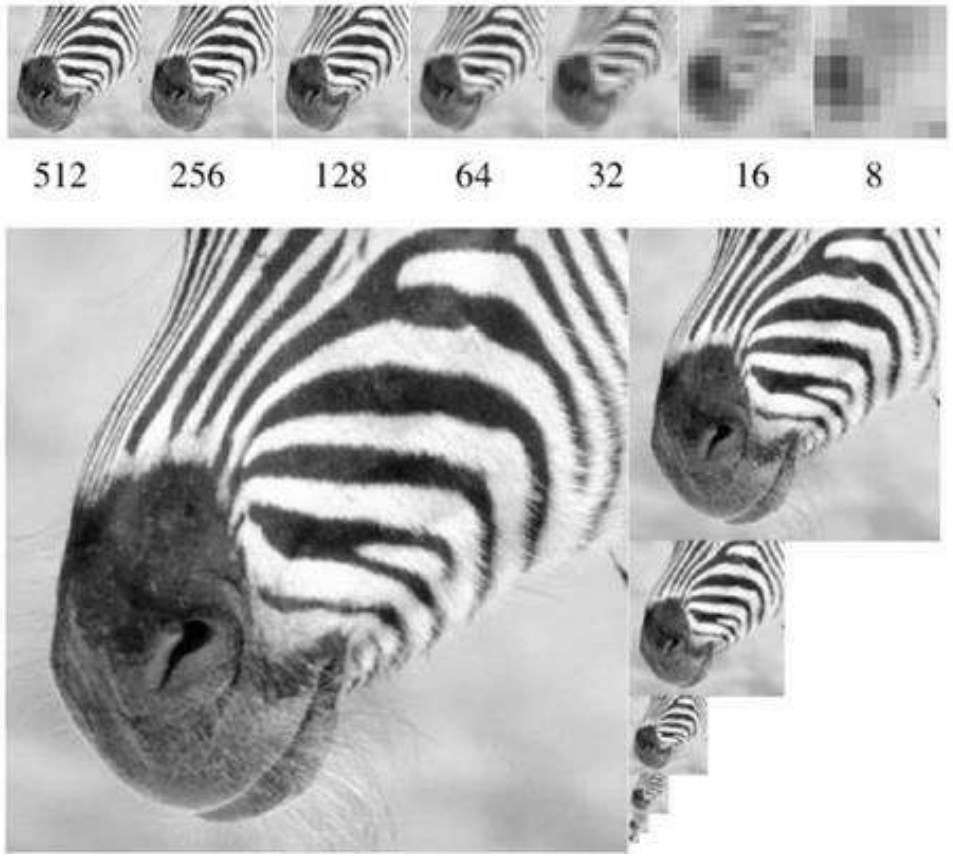



Figure from David Forsyth

# Solution: Deformable DETR

Recreate the transformer encoder/decoder using multi-scale deformable attention modules.

# DETR Encoder Self-Attention

Attention Map Before Training

| | | | | |
|---|---|---|---|---|
| 0.04 | 0.04 | 0.04 | 0.04 | 0.04 |
| 0.04 | 0.04 | 0.04 | 0.04 | 0.04 |
| 0.04 | 0.04 | 0.04 | 0.04 | 0.04 |
| 0.04 | 0.04 | 0.04 | 0.04 | 0.04 |
| 0.04 | 0.04 | 0.04 | 0.04 | 0.04 |

Training



Attention Map After Training

| | | | | |
|---|---|---|---|---|
| 0.02 | 0.27 | 0.02 | 0.01 | 0.00 |
| 0.44 | 0.01 | 0.19 | 0.01 | 0.01 |
| 0.00 | 0.01 | 0.00 | 0.02 | 0.01 |
| 0.01 | 0.01 | 0.00 | 0.02 | 0.02 |
| 0.00 | 0.00 | 0.01 | 0.01 | 0.02 |

# Training Deformable Attention Module

Attention Map Before Training

| | | | | |
|---|---|---|---|---|
| 0 | 0 | 0 | 0 | 0 |
| 0 | 0 | 0 | 0.33 | 0 |
| 0 | 0 | 0 | 0.33 | 0 |
| 0 | 0 | 0 | 0 | 0 |
| 0 | 0.33 | 0 | 0 | 0 |

k = 3

Training

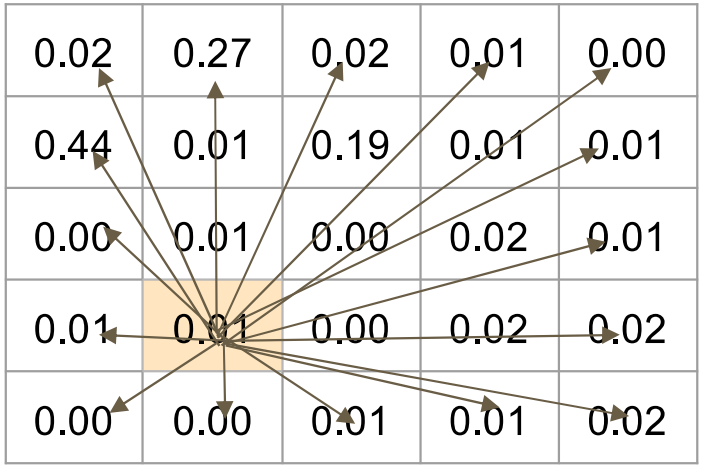
Attention Map After Training

| | | | | |
|---|---|---|---|---|
| 0 | 0.31 | 0 | 0 | 0 |
| 0.50 | 0 | 0.19 | 0 | 0 |
| 0 | 0 | 0 | 0 | 0 |
| 0 | 0 | 0 | 0 | 0 |
| 0 | 0 | 0 | 0 | 0 |

# Time Savings

$$\boldsymbol{x} \in \mathbb{R}^{C \times H \times W}$$

DETR encoder self-attention



# Time Savings

$$\boldsymbol{x} \in \mathbb{R}^{C \times H \times W}$$

DETR encoder self-attention

| | | | | |
|---|---|---|---|---|
| 0.02 | 0.27 | 0.02 | 0.01 | 0.00 |
| 0.44 | 0.01 | 0.19 | 0.01 | 0.01 |
| 0.00 | 0.01 | 0.00 | 0.02 | 0.01 |
| 0.01 | 0.01 | 0.00 | 0.02 | 0.02 |
| 0.00 | 0.00 | 0.01 | 0.01 | 0.02 |

1 query => O(*HWC*)

$O(H^2W^2C)$

# Time Savings

$$\boldsymbol{x} \in \mathbb{R}^{C \times H \times W}$$

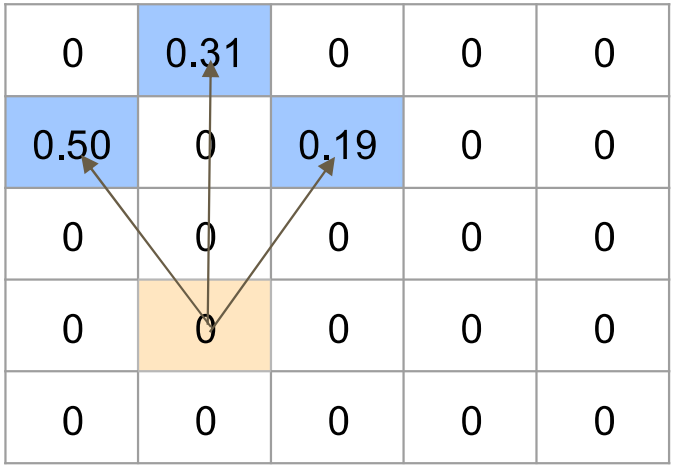
DETR encoder self-attention

| | | | | |
|---|---|---|---|---|
| 0.02 | 0.27 | 0.02 | 0.01 | 0.00 |
| 0.44 | 0.01 | 0.19 | 0.01 | 0.01 |
| 0.00 | 0.01 | 0.00 | 0.02 | 0.01 |
| 0.01 | 0.01 | 0.00 | 0.02 | 0.02 |
| 0.00 | 0.00 | 0.01 | 0.01 | 0.02 |

1 query => O(*HWC*)

O($H^2W^2C$)

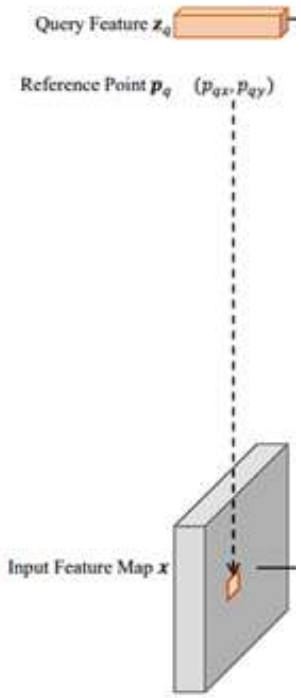
Deformable Cross-Attention



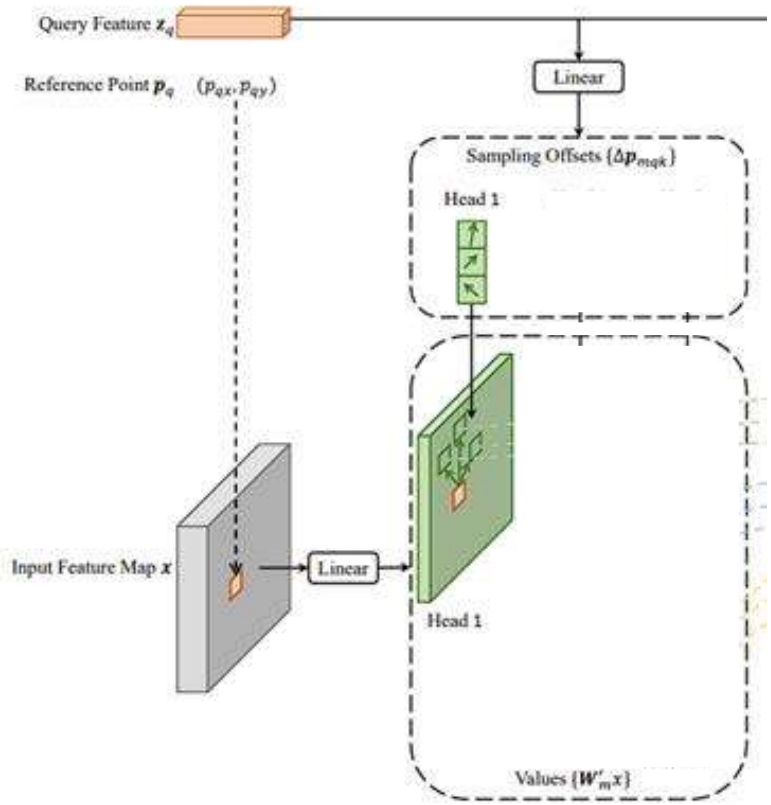

1 query => O(*KC*)

O(*HWKC*)

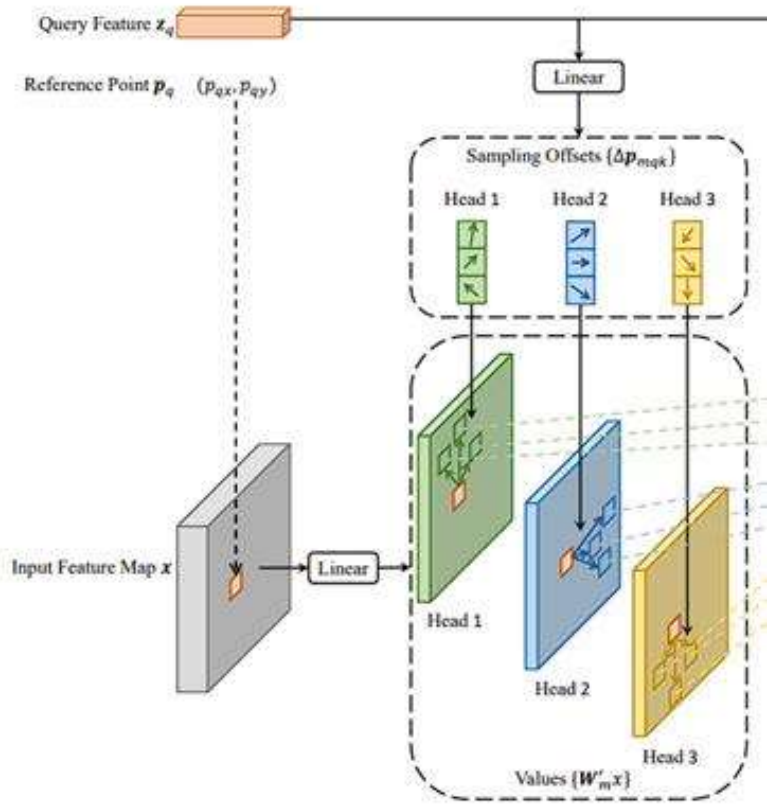
# Methods: Deformable Attention Module



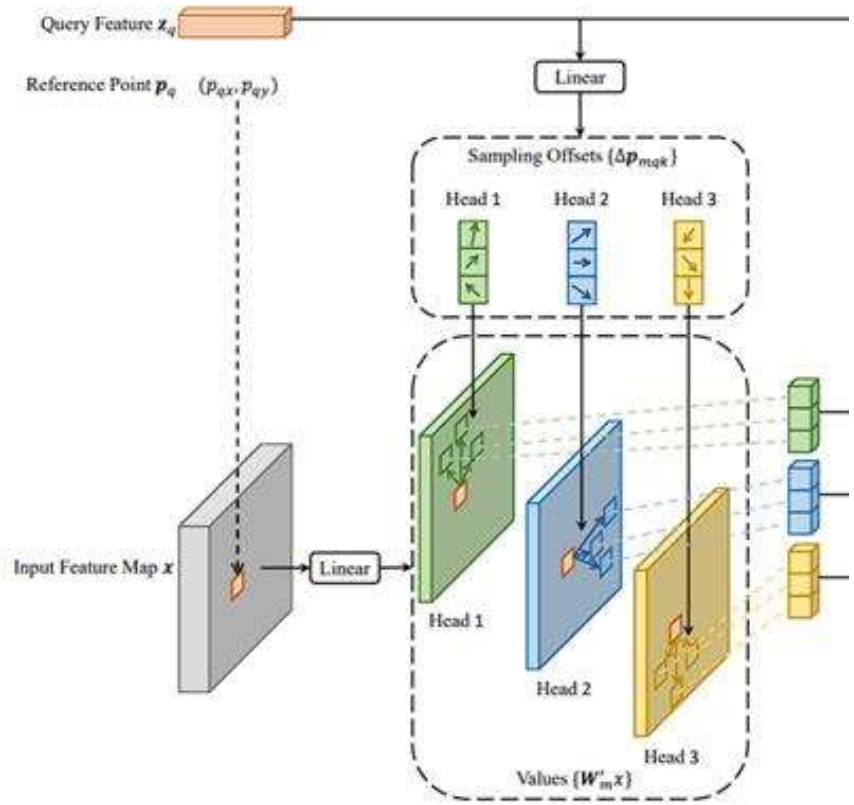
# Methods: Deformable Attention Module



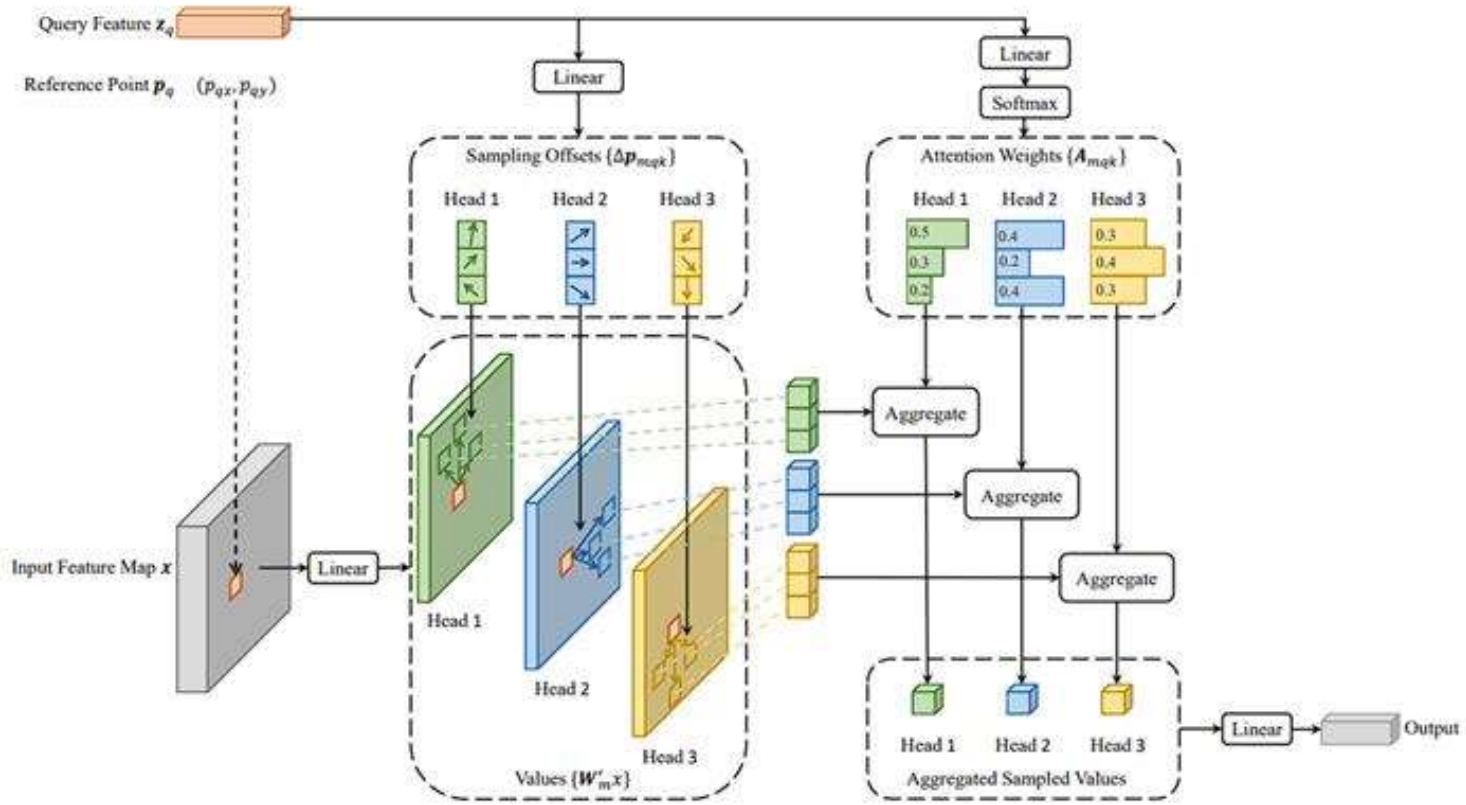
# Methods: Deformable Attention Module



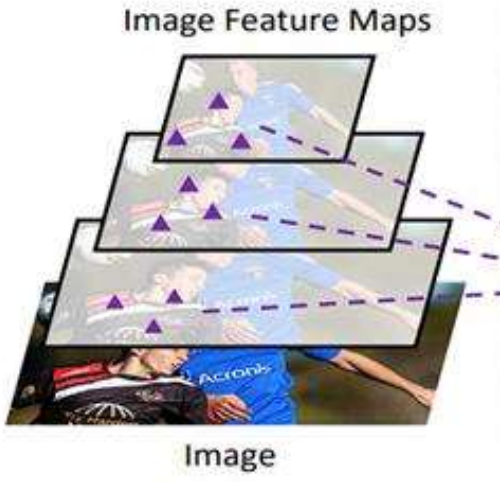
# Methods: Deformable Attention Module



# Methods: Deformable Attention Module

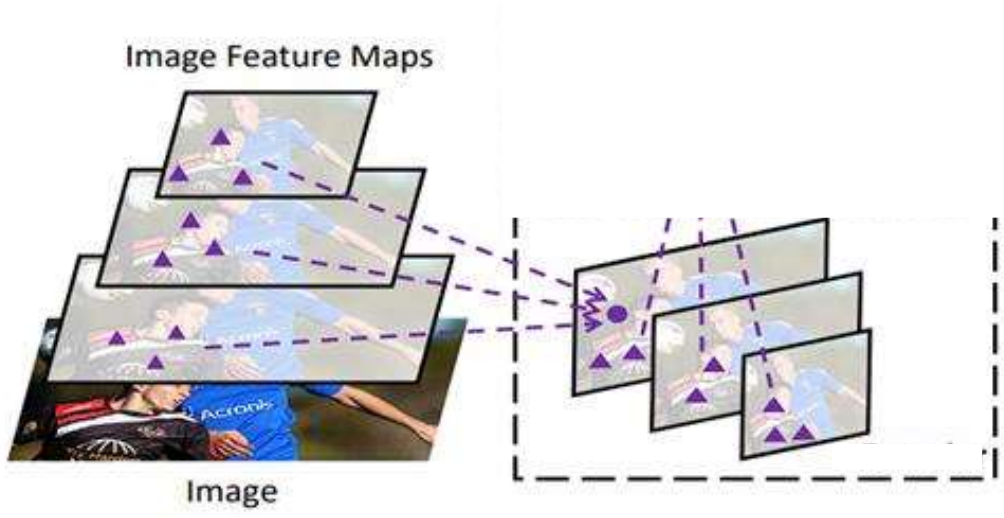


# Methods: Multi-scale Deformable Attention Module

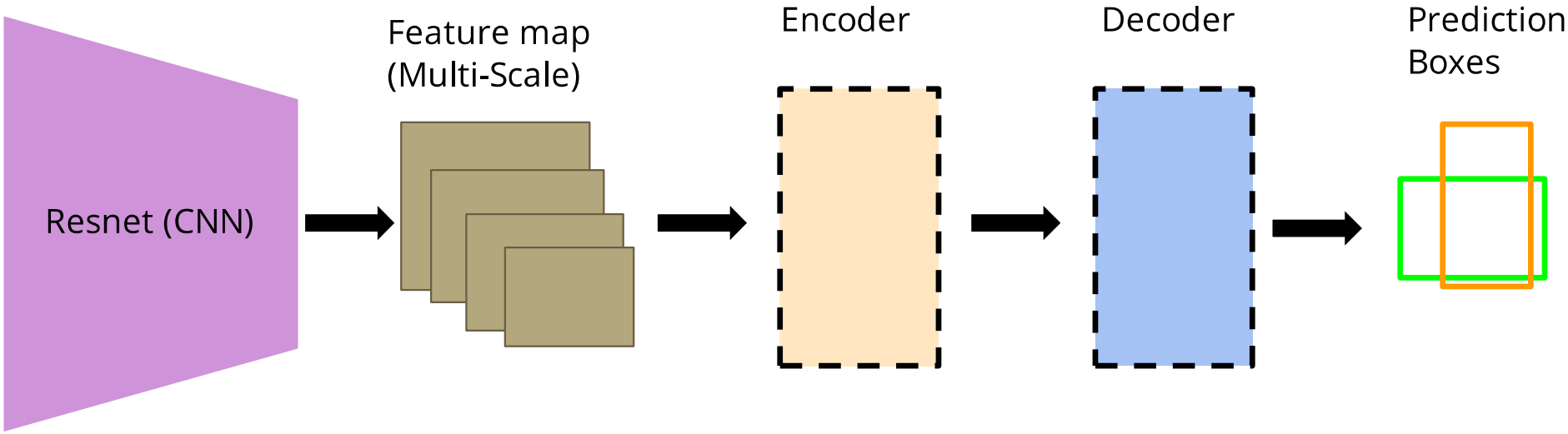


# Methods: Multi-scale Deformable Attention Module

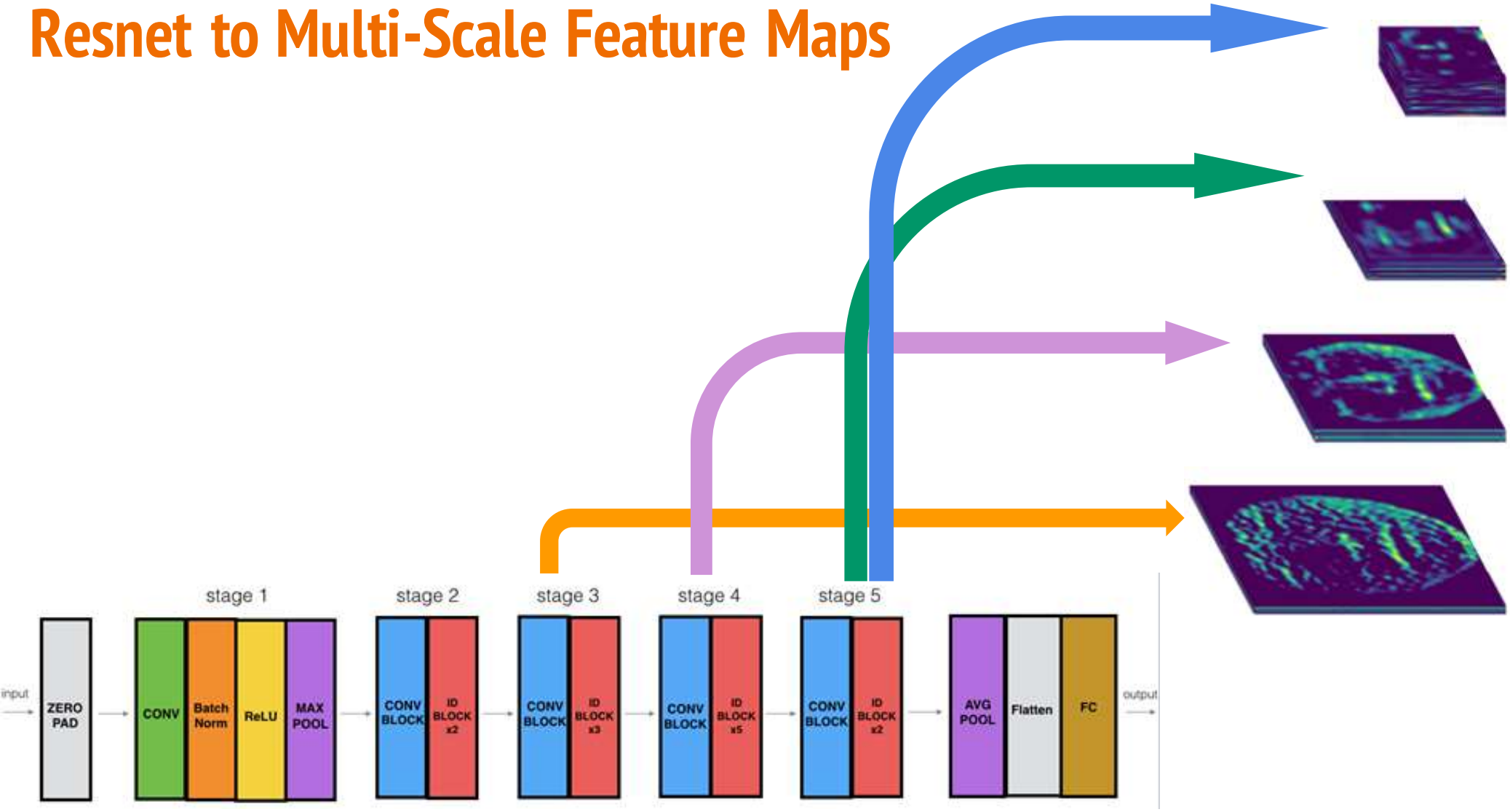
k = 3



# Deformable DETR - Workflow

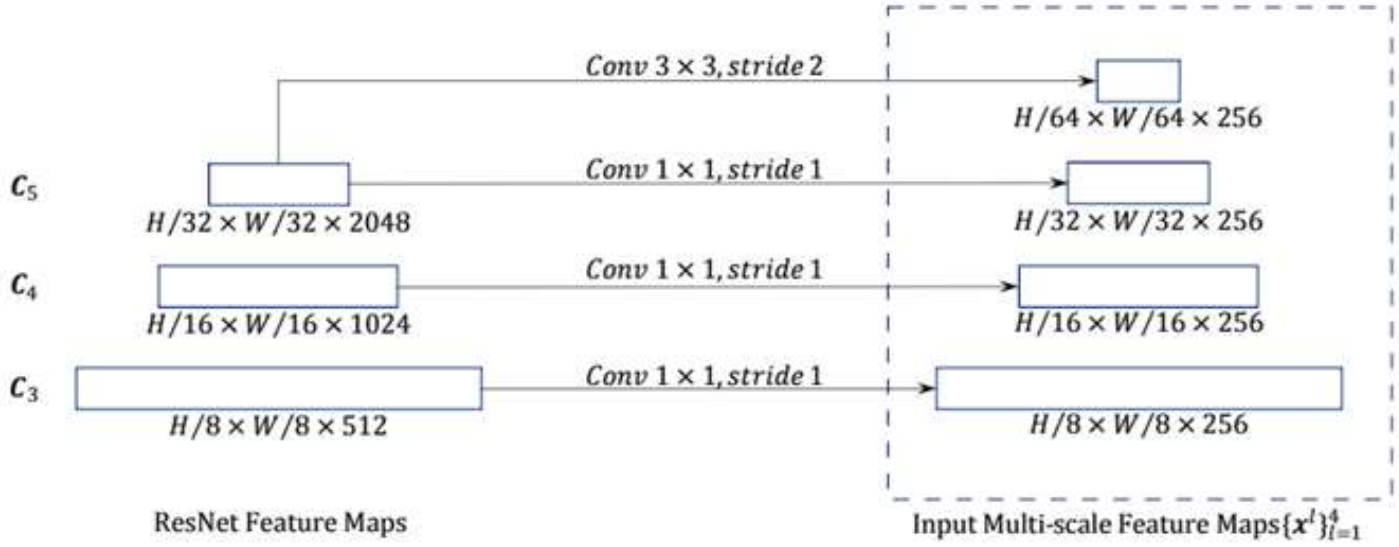


# Resnet to Multi-Scale Feature Maps



# Resnet to Multi-Scale Feature Maps

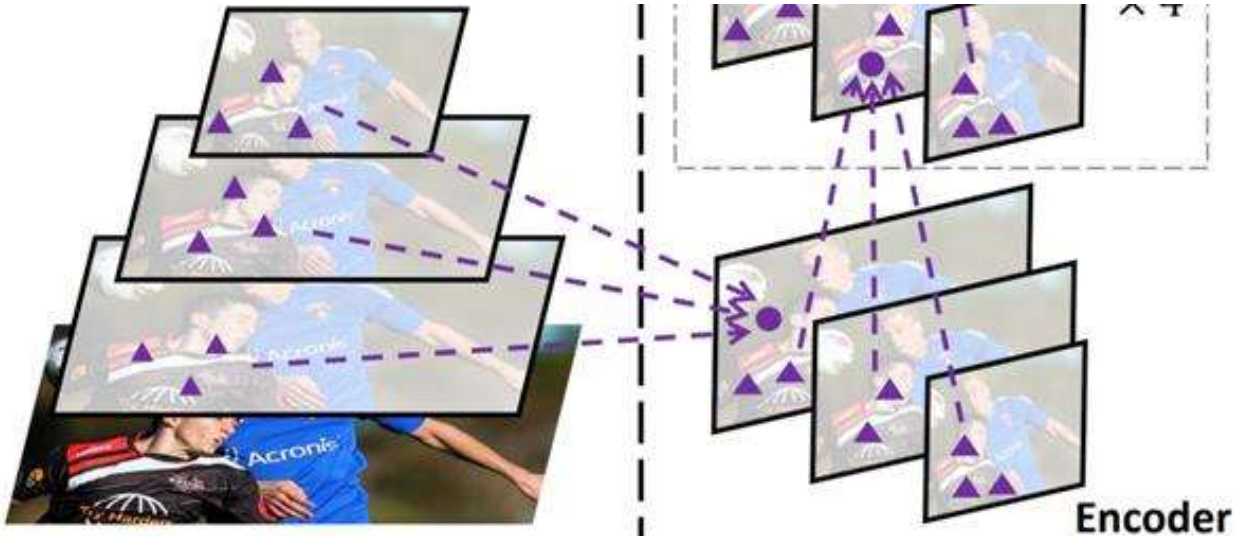
Different convolutions from different stages of Restnet make multi-scale feature maps



# Deformable Transformer Encoder

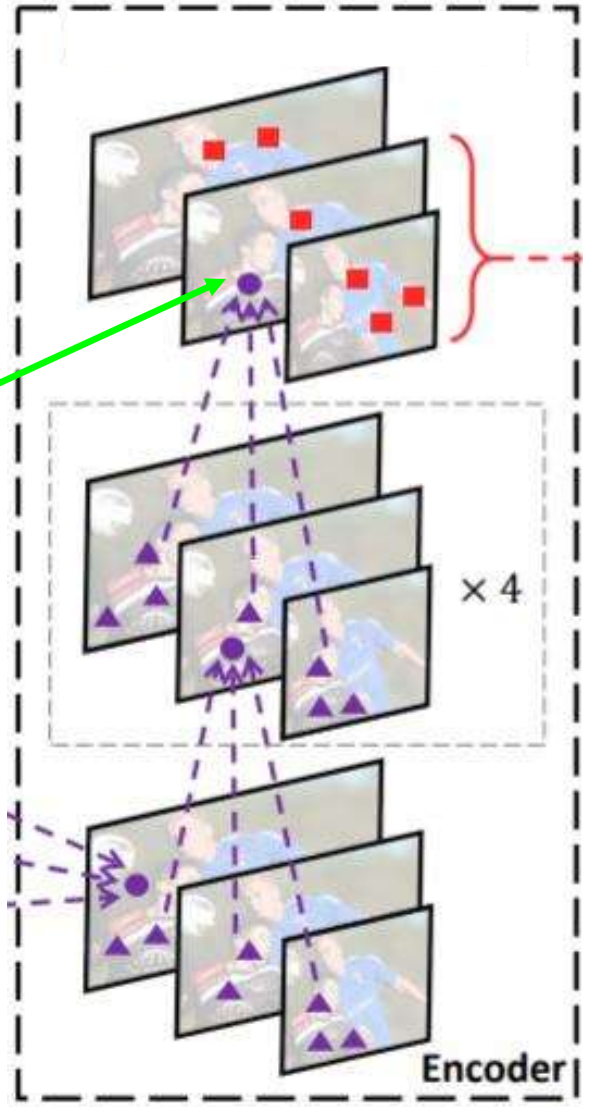
Input → Multi-Scale feature maps

Encoder → Perform Self-Attention on features



# Deformable Transformer Encoder

Self - attention is repeated until multi-scale feature maps are obtained



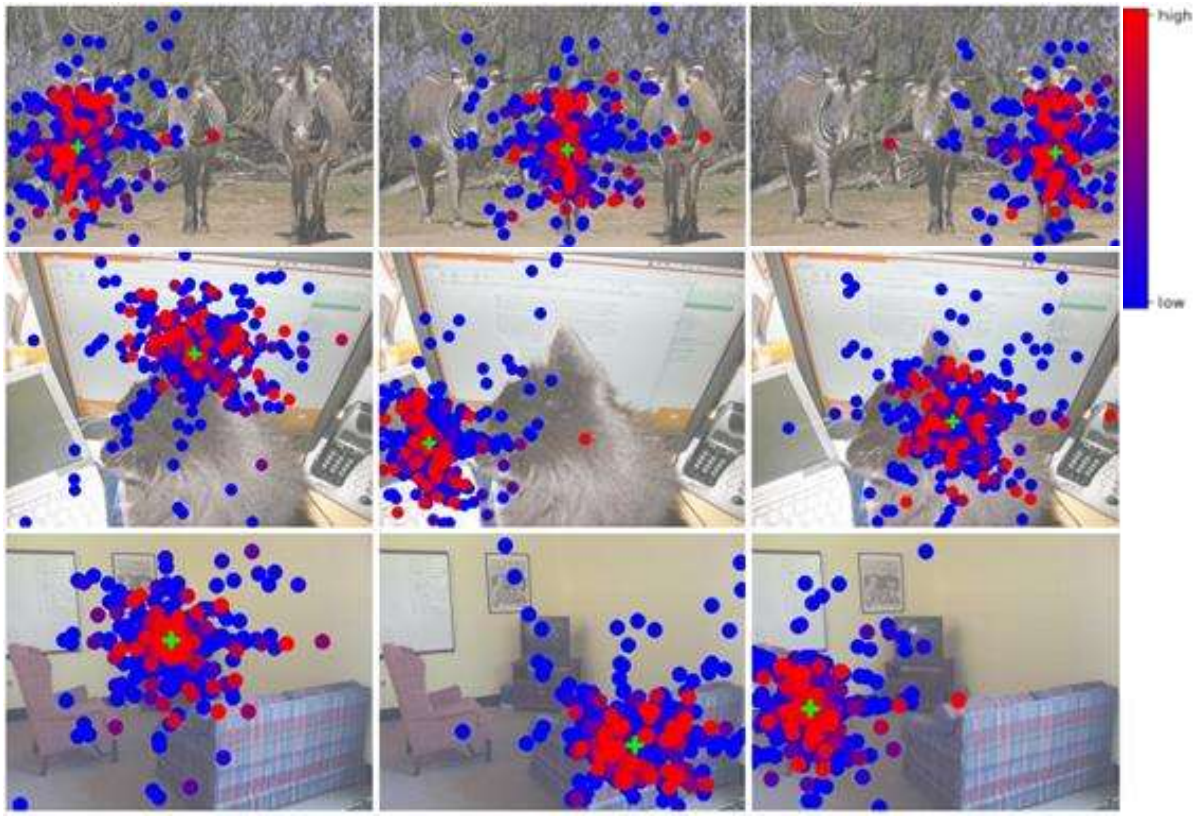

Multi-scale Deformable Self-Attention in Encoder

# Multi-Scale Deformable Self-Attention in Encoder

**Reference point**

Offsets , colored by attention weights

(Simplified to single scale)



# Deformable Transformer Decoder

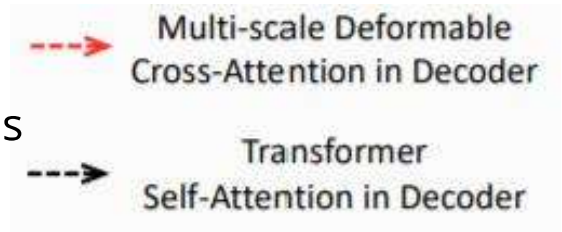
Cross-attention modules:
Features to Object Queries

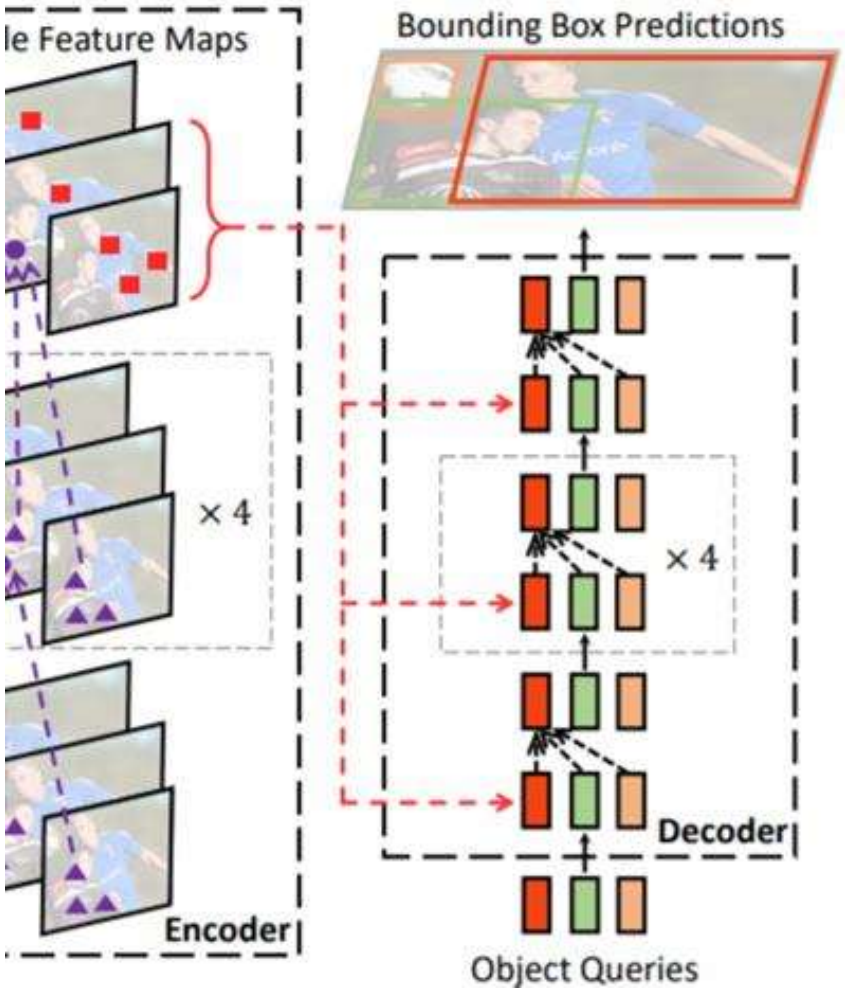
Difference → becomes multi-scale

In self-attention modules:
Object Queries tuning

No Difference

Output → Bounding boxes



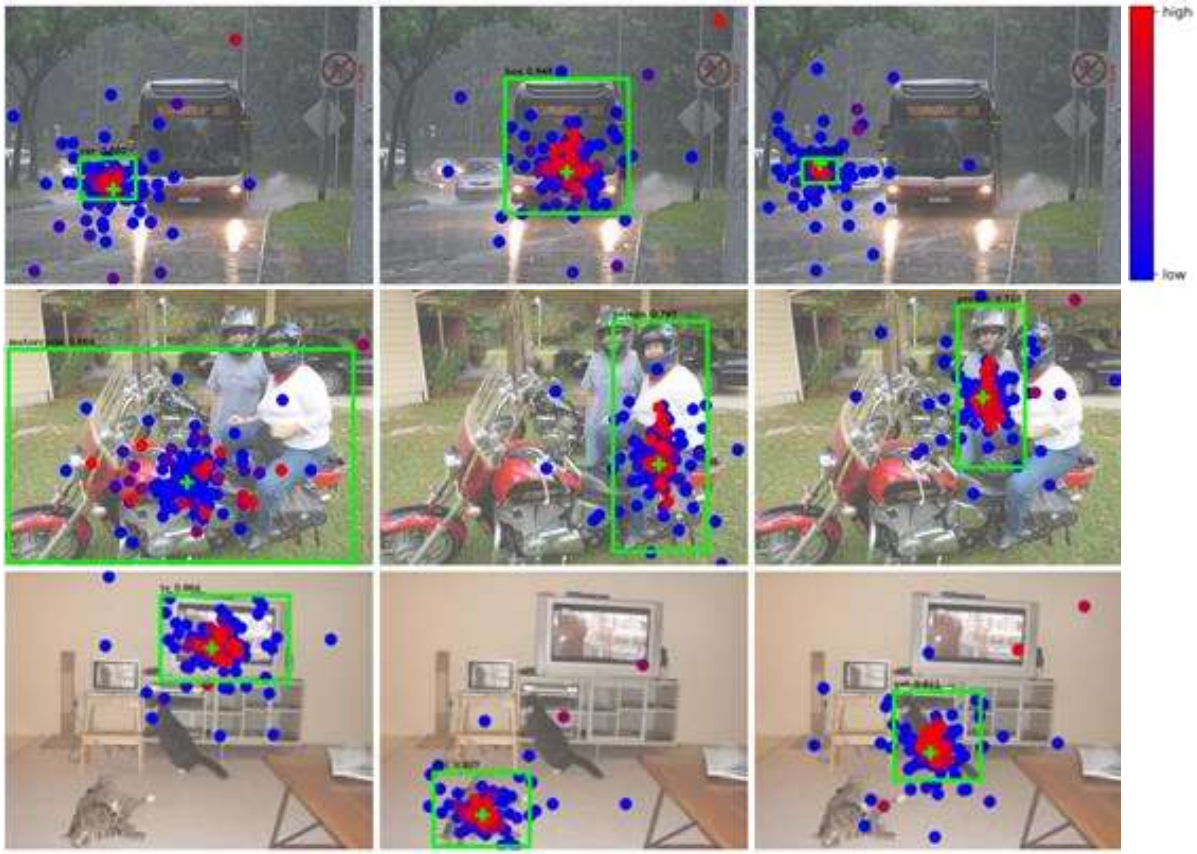
# Multi-Scale Deformable Self-Attention in Decoder

**Reference point**

Offsets , colored by attention weights

(Simplified to single scale)

**(New) Prediction Box**

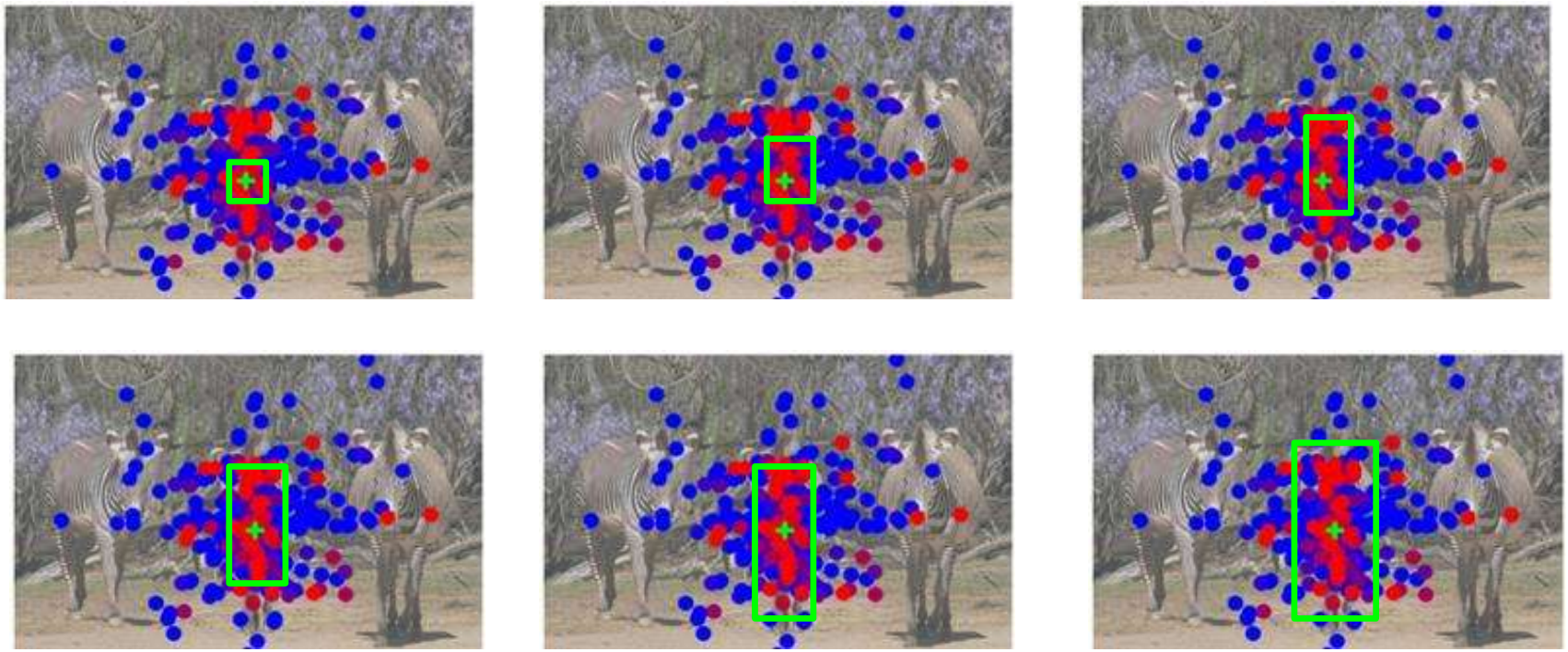


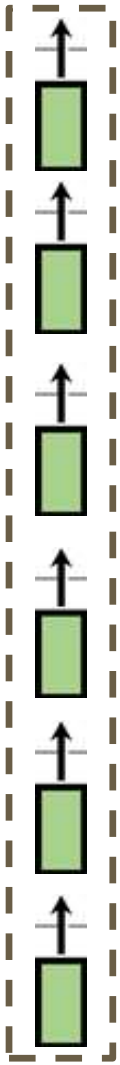
# Additional Improvements and Variants

- Iterative Bounding Box Refinement

- Two-Stage Deformable DETR

# Iterative Bounding Box Refinement (Decoder)





decoder

# Additional Improvements and Variants

Two-Stage Deformable DETR:

Encoder:

→ Pixel → Predicted Bounding Box

Decoder:

→ Iterative Bounding Box Refinement

# Experiments: Background

Dataset:

- COCO 2017 Dataset

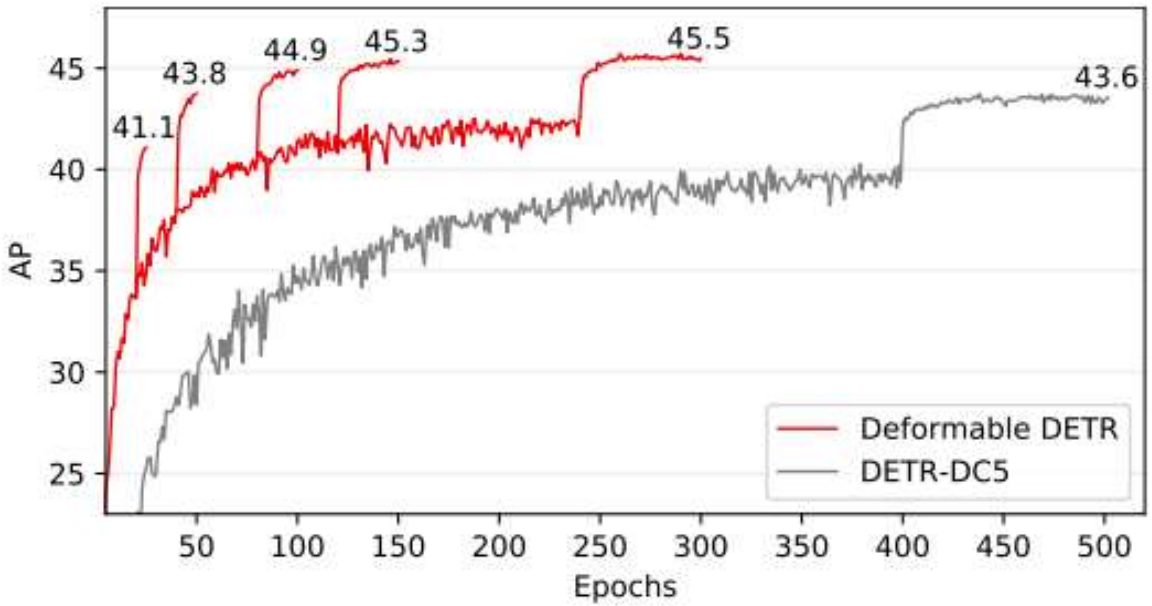
Implementation:

- M = 8, K = 4
- DETR Hyperparameters
- 50 Epochs
- ResNet50 backbone
- NVIDIA Tesla V100 GPU

# Experiments: Comparison with DETR

| Method | Epochs | AP | $AP_{50}$ | $AP_{75}$ | $AP_S$ | $AP_M$ | $AP_L$ | params | FLOPs | Training GPU hours | Inference FPS |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Faster R-CNN + FPN | 109 | 42.0 | 62.1 | 45.5 | 26.6 | 45.4 | 53.4 | 42M | 180G | 380 | 26 |
| DETR | 500 | 42.0 | 62.4 | 44.2 | 20.5 | 45.8 | 61.1 | 41M | 86G | 2000 | 28 |
| DETR-DC5 | 500 | 43.3 | 63.1 | 45.9 | 22.5 | 47.3 | 61.1 | 41M | 187G | 7000 | 12 |
| DETR-DC5 | 50 | 35.3 | 55.7 | 36.8 | 15.2 | 37.5 | 53.6 | 41M | 187G | 700 | 12 |
| DETR-DC5+ | 50 | 36.2 | 57.0 | 37.4 | 16.3 | 39.2 | 53.9 | 41M | 187G | 700 | 12 |
| Deformable DETR | 50 | 43.8 | 62.6 | 47.7 | 26.4 | 47.1 | 58.0 | 40M | 173G | 325 | 19 |
| + iterative bounding box refinement | 50 | 45.4 | 64.7 | 49.0 | 26.8 | 48.3 | 61.7 | 40M | 173G | 325 | 19 |
| ++ two-stage Deformable DETR | 50 | 46.2 | 65.2 | 50.0 | 28.8 | 49.2 | 61.7 | 40M | 173G | 340 | 19 |

# Experiments: Comparison with DETR cont'd



# Experiments: Ablation Study

| MS inputs | MS attention | K | FPNs | AP | $AP_{50}$ | $AP_{75}$ | $AP_S$ | $AP_M$ | $AP_L$ |
|---|---|---|---|---|---|---|---|---|---|
| ✓ | ✓ | 4 | FPN (Lin et al., 2017a) | 43.8 | 62.6 | 47.8 | 26.5 | 47.3 | 58.1 |
| ✓ | ✓ | 4 | BiFPN (Tan et al., 2020) | 43.9 | 62.5 | 47.7 | 25.6 | 47.4 | 57.7 |
| | | 1 | w/o | 39.7 | 60.1 | 42.4 | 21.2 | 44.3 | 56.0 |
| ✓ | | 1 | | 41.4 | 60.9 | 44.9 | 24.1 | 44.6 | 56.1 |
| ✓ | | 4 | | 42.3 | 61.4 | 46.0 | 24.8 | 45.1 | 56.3 |
| ✓ | ✓ | 4 | | 43.8 | 62.6 | 47.7 | 26.4 | 47.1 | 58.0 |

# Experiments: State-of-the-Art Models

| Method | Backbone | TTA | AP | $AP_{50}$ | $AP_{75}$ | $AP_S$ | $AP_M$ | $AP_L$ |
|---|---|---|---|---|---|---|---|---|
| FCOS (Tian et al., 2019) | ResNeXt-101 | | 44.7 | 64.1 | 48.4 | 27.6 | 47.5 | 55.6 |
| ATSS (Zhang et al., 2020) | ResNeXt-101 + DCN | ✓ | 50.7 | 68.9 | 56.3 | 33.2 | 52.9 | 62.4 |
| TSD (Song et al., 2020) | SENet154 + DCN | ✓ | 51.2 | 71.9 | 56.0 | 33.8 | 54.8 | 64.2 |
| EfficientDet-D7 (Tan et al., 2020) | EfficientNet-B6 | | 52.2 | 71.4 | 56.3 | - | - | - |
| Deformable DETR | ResNet-50 | | 46.9 | 66.4 | 50.8 | 27.7 | 49.7 | 59.9 |
| Deformable DETR | ResNet-101 | | 48.7 | 68.1 | 52.9 | 29.1 | 51.5 | 62.0 |
| Deformable DETR | ResNeXt-101 | | 49.0 | 68.5 | 53.2 | 29.7 | 51.7 | 62.8 |
| Deformable DETR | ResNeXt-101 + DCN | | 50.1 | 69.7 | 54.6 | 30.6 | 52.8 | 64.7 |
| Deformable DETR | ResNeXt-101 + DCN | ✓ | 52.3 | 71.9 | 58.1 | 34.4 | 54.4 | 65.6 |

# Conclusion

- More efficient and fast converging end-to-end object detector.
- Multi-scale deformable attention modules.
- Improves on DETR object detection

# References

Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. “Deformable DETR: Deformable Transformers For End-To-End Object Detection.” In ICLR, 2021.

Jifeng Dai, Haozhi Qi, Yuwen Xiong, Yi Li, Guodong Zhang, Han Hu, and Yichen Wei. “Deformable Convolutional Networks.” In ICCV, 2017.

Divyanshu Mishra. “Deformable Convolutions Demystified”. Web. Sep 30, 2020.

# References cont.

"Object Detection". Dr. Yogesh Singh Rawat. CAP5415. Class lecture. 2021. UCF.