AdaMixer: A Fast-Converging Query-Based Object Detector

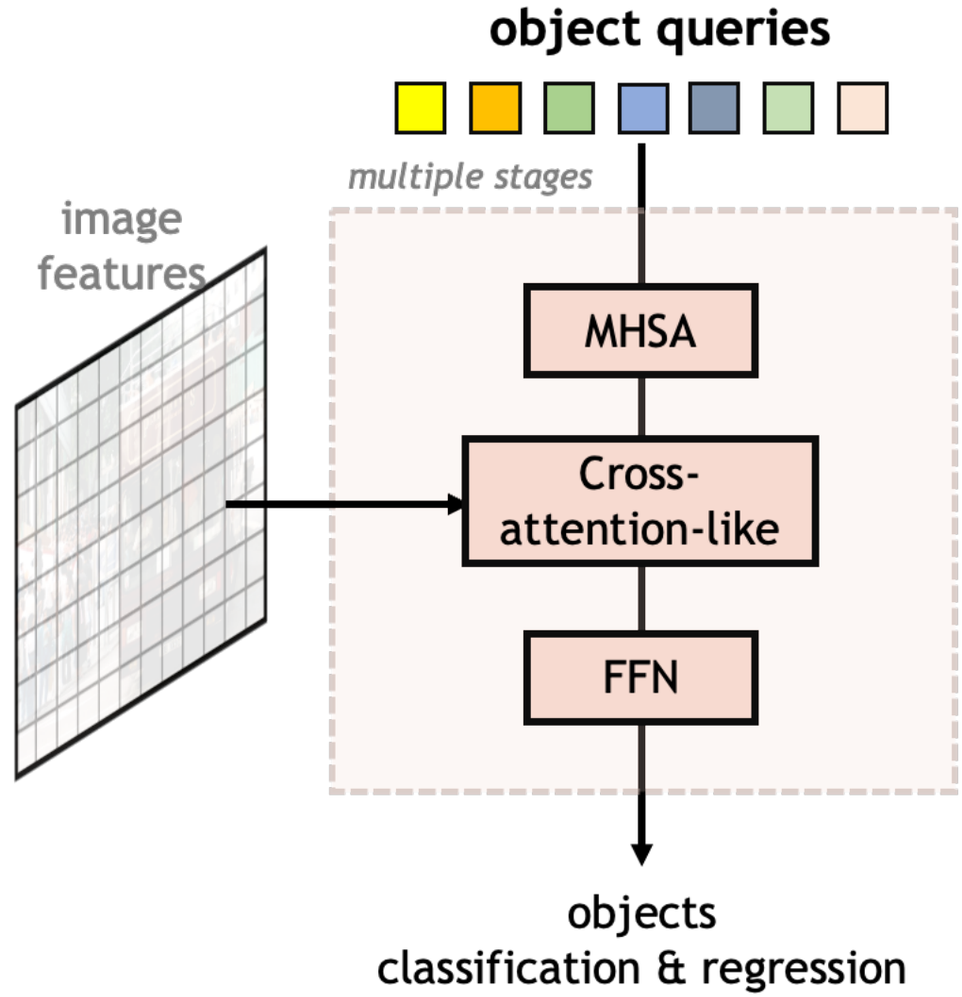
Ziteng Gao[1] Limin Wang[1] Bing Han[2] Sheng Guo[2]

[1]State Key Laboratory for Novel Software Technology, Nanjing University

[2]MYBank, Ant Group

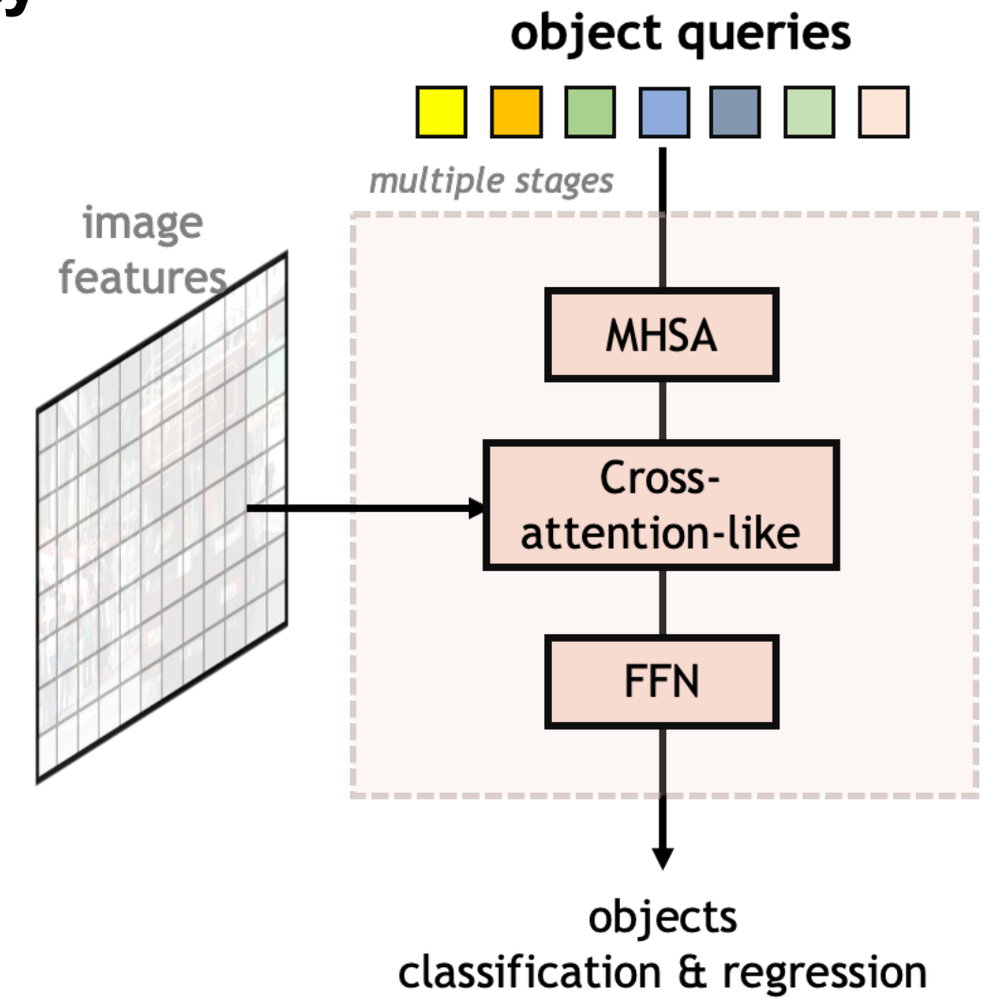
Motivation

- **Query-based object detectors as a new detection paradigm**
  - represent objects by **queries**
  - e.g., DETR, Deformable DETR, Sparse R-CNN
- **Challenges**
  - **slow** convergency
  - architectural & computational **complexity**
  - **poor** small object detection



Motivation

- **We argue that the key is the adaptability**
  - adapt **limited queries** to **varying visual objects** in decoding queries
  - specifically, in cross-attention-like ops between queries and image features
- **Our work improves in two aspects**
  - **positional adaptability**
    - *translation & scale variance*
  - **content adaptability**
    - *visual appearance variance*



Motivation

- **We argue that the key is the adaptability**
  - adapt **limited queries** to **varying visual objects** in decoding queries
  - specifically, in cross-attention-like ops between queries and image features

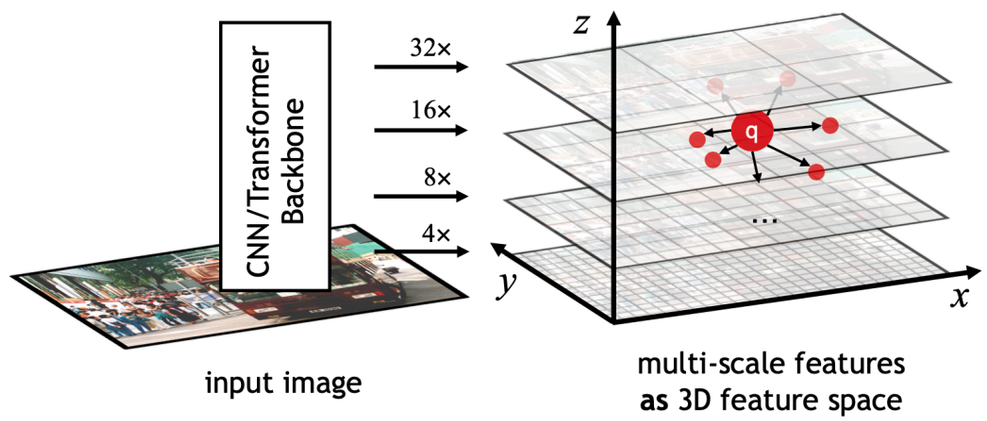

positional adaptability

- **Our work improves this in two aspects**
  - **positional adaptability**
    - *translation & scale variance*
  - **content adaptability**
    - *visual appearance variance*

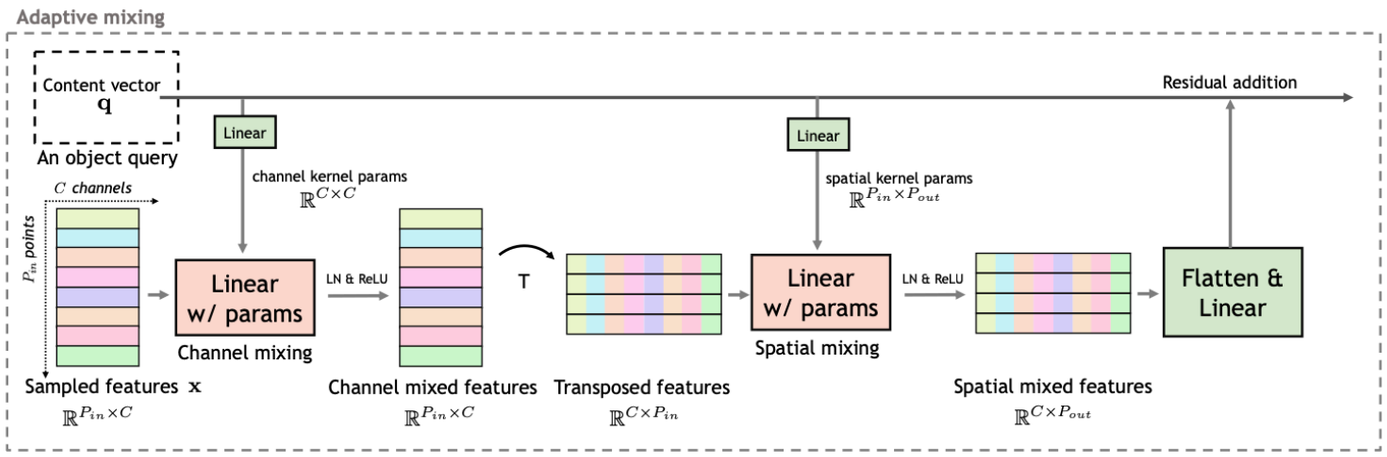

content adaptability

Reforming Cross Attention

- **Regard cross-attention as two steps**
  - sample features
  - decode sampled features
- **Two ingredients making our cross-attention variant**
  - **adaptive 3D feature sampling**
    - for positional adaptability
  - **adaptive mixing**
    - for content adaptability



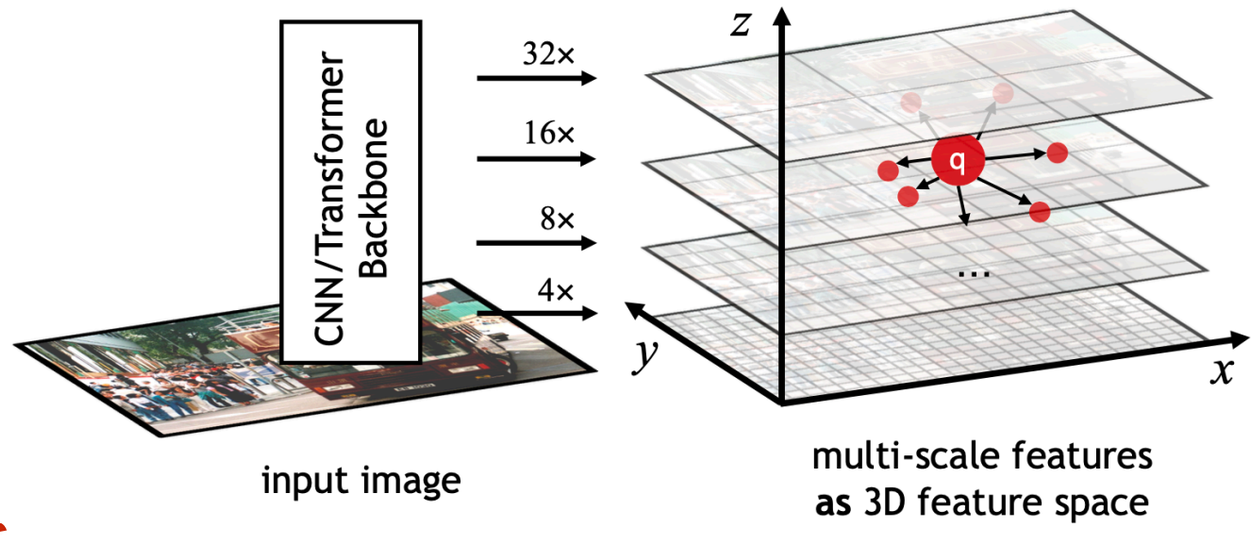
3D feature sampling



adaptive mixing

Feature Sampling

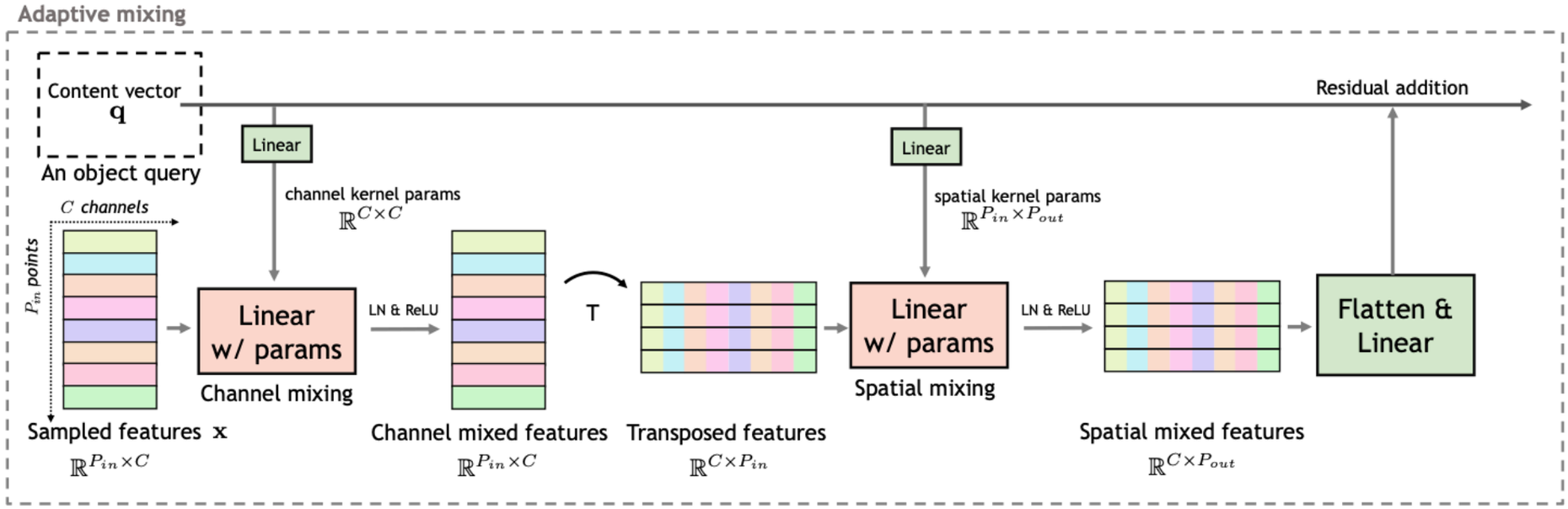
- **Adaptive 3D feature sampling**
  - multi-scale features **directly** from the backbone
  - **3D interpolable** feature space
  - **2D** (x, y)-axes '*translation*' + **1D** z-axis '*scale*'
  - a query generates **adaptive** offsets
  - **learn where to look**
- **Enhance the adaptability**
  - to both **spatial locations** and **scales**
  - **no** extra explicit pyramid networks
  - but multi-scale interaction with the cheap computation



3D feature sampling

Decoding the Sampled

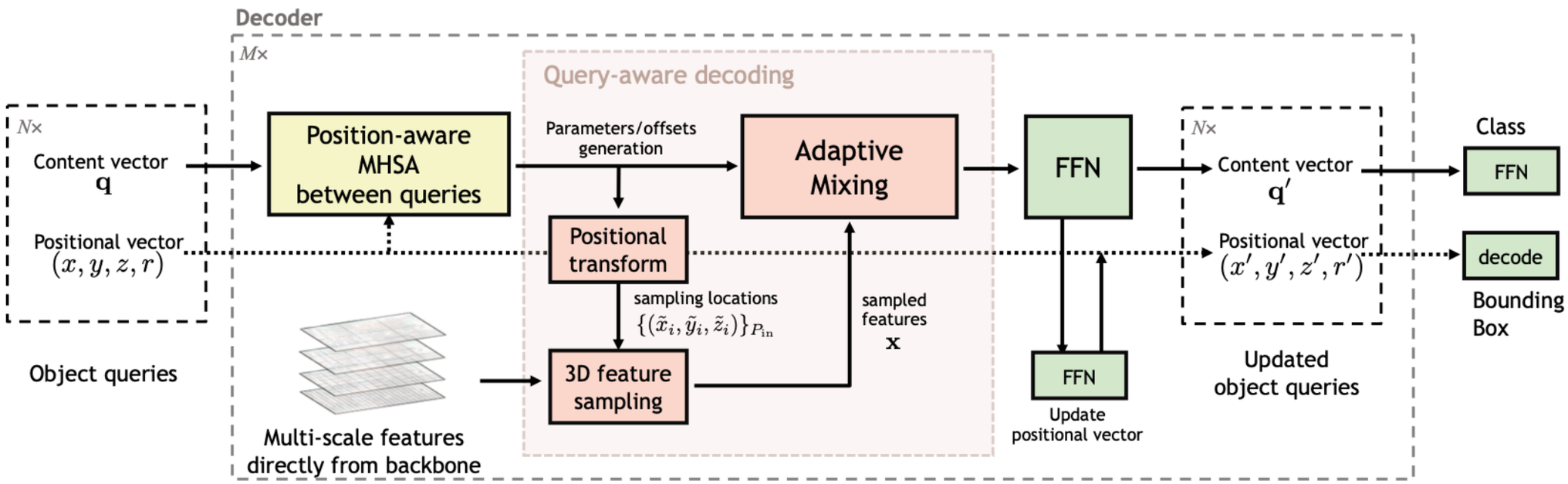
- **Adaptive mixing**
  - an **adaptive** variant of MLP-Mixer
  - channel mixing & spatial mixing
  - involve both **adaptive channel semantics** and **spatial structures**
  - **learn what expected to see** and **what to do after**
- **Enhance the adaptability to the object content**



adaptive mixing

AdaMixer Decoder

- **Our cross attention variant**
  - = adaptive 3D feature sampling + adaptive mixing
- **The overall query AdaMixer decoder structure**
  - generally follows the DETR decoder, multi-stage
  - includes self-attention, our cross-attention variant and FFN layers sequentially



adaptive mixing

Overall AdaMixer

and eventually, AdaMixer is

- simply **a backbone** and our AdaMixer query **decoder**
- with **no** extra attention encoders or explicit pyramid networks **required**

Experiments

- **Training with 1x schedule**
  - 12 training epochs
  - only data augmentation: horizontal flipping
- **Leading performance with**
  - **fast** convergency
  - training sample **efficiency**
    - *no heavy data augmentation*
  - **well-performed** small detections
    - *compared with traditional ones*

detector	epochs	AP	AP_{50}	AP_{75}	AP_s	AP_m	AP_l
FCOS [36]	12	38.7	57.4	41.8	22.9	42.5	50.1
Cascade R-CNN [3]	12	40.4	58.9	44.1	22.8	43.7	54.0
GFocalV2 [19]	12	41.1	58.8	44.9	23.5	44.9	53.3
BorderDet [29]	12	41.4	59.4	44.5	23.6	45.1	54.6
Dynamic Head [6]	12	42.6	60.1	**46.4**	**26.1**	**46.8**	56.0
DETR [4]	12	20.0	36.2	19.3	6.0	20.5	32.2
Deformable DETR [51]	12	35.1	53.6	37.7	18.2	38.5	48.7
Sparse R-CNN [35]	12	37.9	56.0	40.5	20.7	40.0	53.5
AdaMixer (N=100)	12	**42.7**	**61.5**	45.9	24.7	45.4	**59.2**
AdaMixer (N=300)	12	**44.1**	**63.4**	**47.4**	**27.0**	**46.9**	**59.5**
AdaMixer (N=500)	12	**45.0**	**64.2**	**48.6**	**27.9**	**47.8**	**61.1**

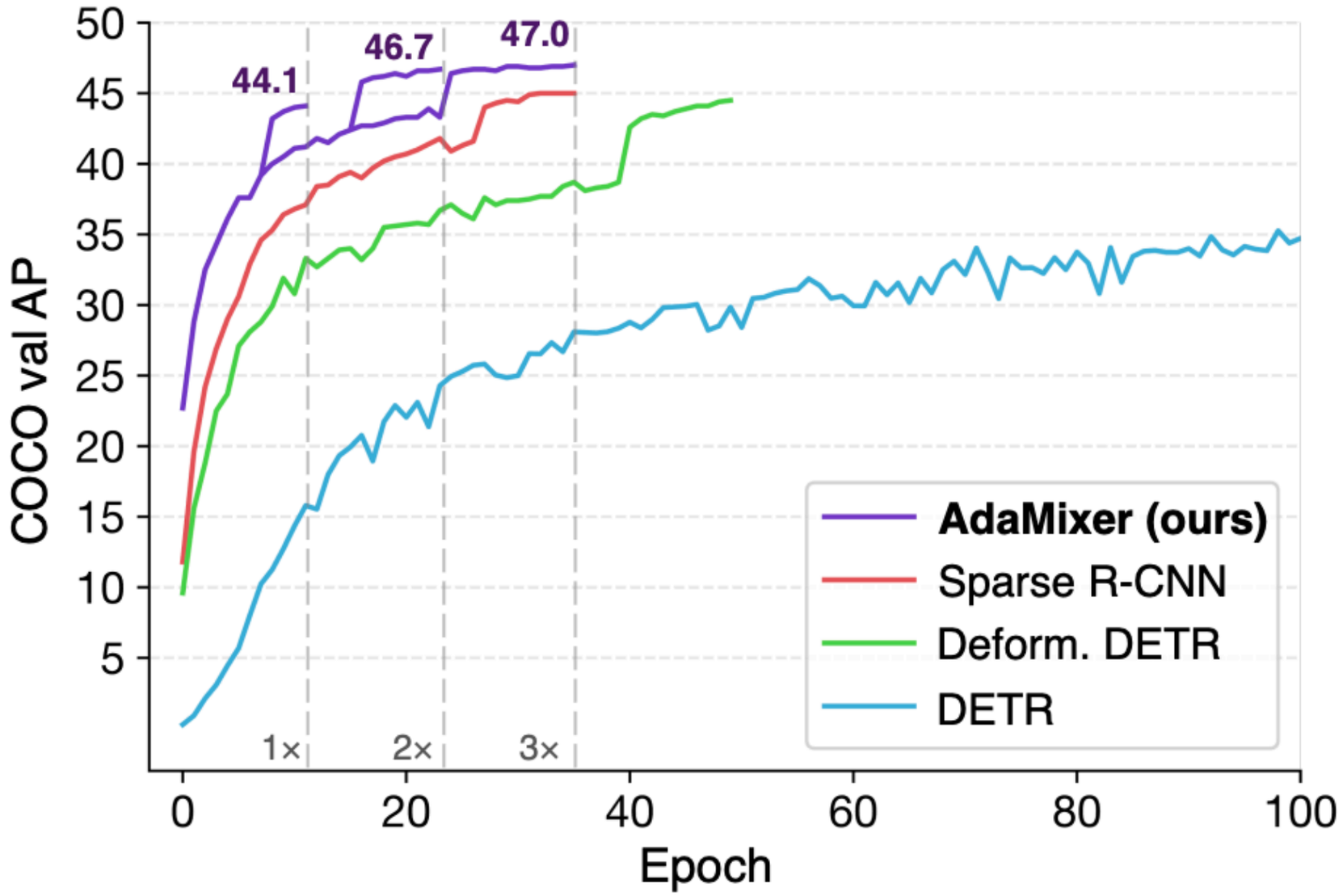
Experiments

- **Training with longer schedules**
  - 36 training epochs & data augmentation aligned to DETR & 300 queries
- **Leading performance**
  - among query-based detectors with lower computational costs

detector	backbone	encoder/pyramid net	#epochs	GFLOPs	AP	AP_{50}	AP_{75}	AP_s	AP_m	AP_l
DETR [4]	ResNet-50-DC5	TransformerEnc	500	187	43.3	63.1	45.9	22.5	47.3	61.1
SMCA [13]	ResNet-50	TransformerEnc	50	152	43.7	63.6	47.2	24.2	47.0	60.4
Deformable DETR [56]	ResNet-50	DeformTransEnc	50	173	43.8	62.6	47.7	26.4	47.1	58.0
Sparse R-CNN [39]	ResNet-50	FPN	**36**	174	45.0	63.4	48.2	26.9	47.2	59.5
Efficient DETR [49]	ResNet-50	DeformTransEnc	**36**	210	45.1	63.1	49.1	28.3	48.4	59.0
Conditional DETR [31]	ResNet-50-DC5	TransformerEnc	108	195	45.1	65.4	48.5	25.3	49.0	**62.2**
Anchor DETR [46]	ResNet-50-DC5	DecoupTransEnc	50	151	44.2	64.7	47.5	24.7	48.2	60.6
AdaMixer (ours)	ResNet-50	-	**12**	132	44.1	63.1	47.8	29.5	47.0	58.8
AdaMixer (ours)	ResNet-50	-	**24**	132	46.7	65.9	50.5	29.7	49.7	61.5
AdaMixer (ours)	ResNet-50	-	**36**	132	**47.0**	**66.0**	**51.1**	**30.1**	**50.2**	61.8
DETR [4]	ResNet-101-DC5	TransformerEnc	500	253	44.9	64.7	47.7	23.7	49.5	62.3
SMCA [13]	ResNet-101	TransformerEnc	50	218	44.4	65.2	48.0	24.3	48.5	61.0
Sparse R-CNN [39]	ResNet-101	FPN	**36**	250	46.4	64.6	49.5	28.3	48.3	61.6
Efficient DETR [49]	ResNet-101	DeformTransEnc	**36**	289	45.7	64.1	49.5	28.2	49.1	60.2
Conditional DETR [31]	ResNet-101-DC5	TransformerEnc	108	262	45.9	66.8	49.5	27.2	50.3	63.3
AdaMixer (ours)	ResNet-101	-	**36**	208	**48.0**	**67.0**	**52.4**	**30.0**	**51.2**	**63.7**
AdaMixer (ours)	ResNeXt-101-DCN	-	**36**	214	**49.5**	**68.9**	**53.9**	**31.3**	**52.3**	**66.3**
AdaMixer (ours)	Swin-S	-	**36**	234	**51.3**	**71.2**	**55.7**	**34.2**	**54.6**	**67.3**

Experiments

- and, fast convergency again



Key Ablations

adaptive loc.	adaptive cont.	AP	AP_{50}	AP_{75}	AP_s	AP_m	AP_l
		35.7	55.2	37.8	20.1	38.1	48.8
✓		37.3	55.8	39.7	20.7	40.1	50.9
	✓	40.4	60.5	43.4	23.0	42.5	56.7
✓	✓	**42.7**	**61.5**	**45.9**	**24.7**	**45.4**	**59.2**

adaptability of decoding sampling locations and sampled content

mixing	AP	AP_{50}	AP_{75}	AP_s	AP_m	AP_l
ACM	ACM	41.5	60.5	44.3	23.5	44.1	57.4
ASM	ASM	39.8	58.8	42.6	22.8	42.4	56.1
ACM	ASM	**42.7**	**61.5**	**45.9**	**24.7**	**45.4**	**59.2**
ASM	ACM	41.5	60.4	44.5	23.9	44.4	57.1

design in our adaptive mixing procedure
ACM for adaptive channel mixing
ASM for adaptive spatial mixing

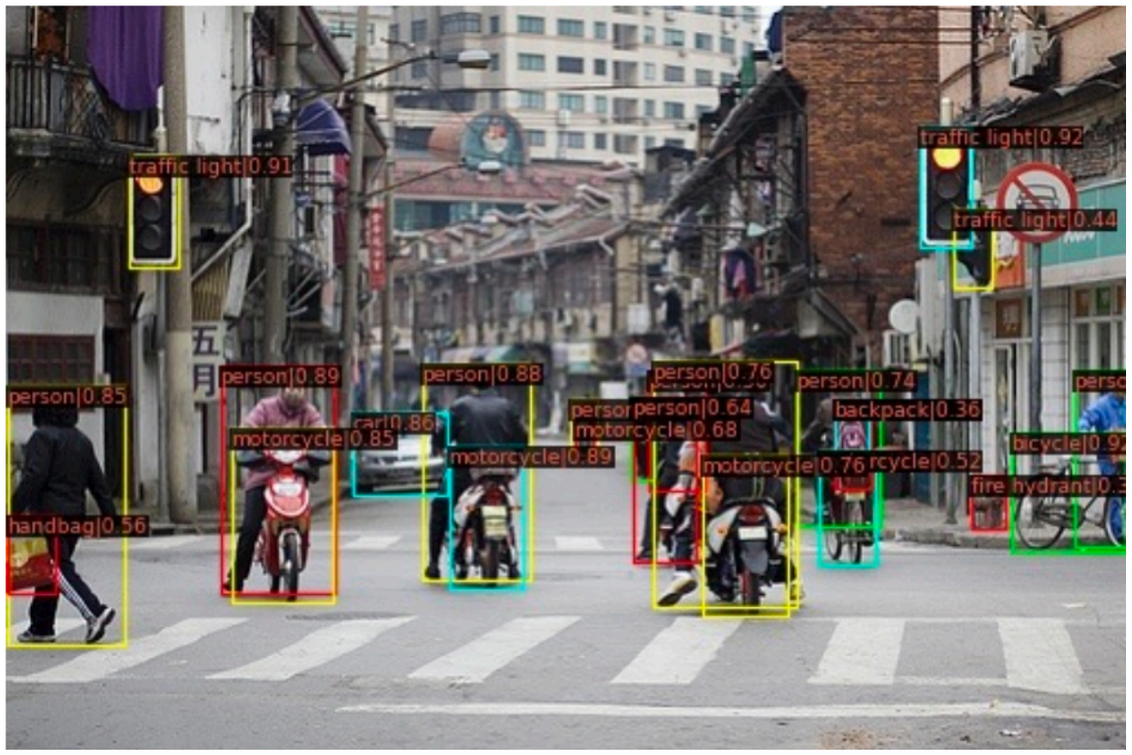
Key Ablations

sampling	AP	AP_{50}	AP_{75}	AP_s	AP_m	AP_l
only C_3 feature	26.2	42.0	27.3	15.8	28.7	34.1
only C_4 feature	38.3	57.2	41.0	20.0	41.9	54.1
only C_5 feature	37.8	58.3	39.5	18.0	41.2	51.7
RoIAlign	37.2	58.5	39.0	19.0	39.3	55.6
A2DS	41.3	61.0	44.4	23.3	43.8	57.8
A3DS	**42.7**	**61.5**	**45.9**	**24.7**	**45.4**	**59.2**

sampling methods
A3DS = adaptive 3D sampling
A2DS = A3DS minus z-axis sampling offsets
no pyramid networks used through this table

Visualizations

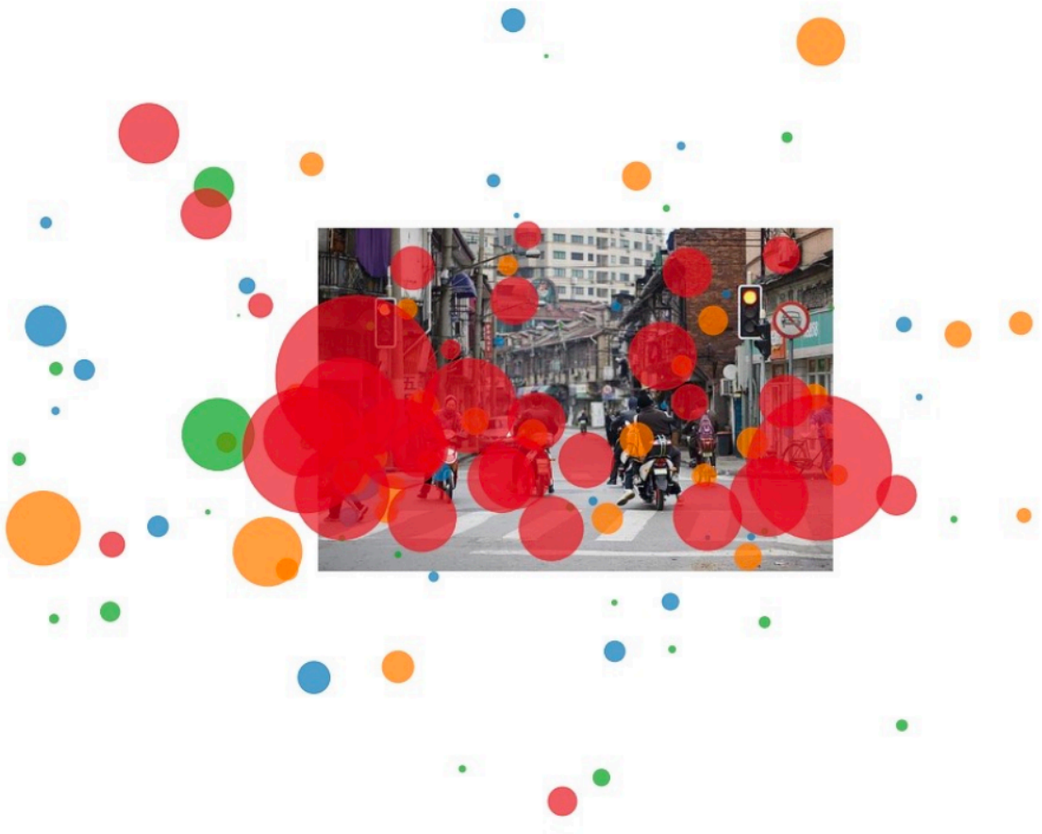
input image



detection results

Visualizations

sampling points
for person #1



stage #1

stage #2

Visualizations

sampling points
for person #1

stage #5

stage #6

Recap

- **AdaMixer, a query-based detector, enjoys**
  - a **simple backbone-and-decoder-only** architecture
  - **fast** convergency
  - **leading** performance
  - **lower** FLOPs & **actual satisfying** FPS
  - **well** small object detections

AdaMixer: A Fast-Converging Query-Based Object Detector

Code is available at

https://github.com/MCG-NJU/AdaMixer/