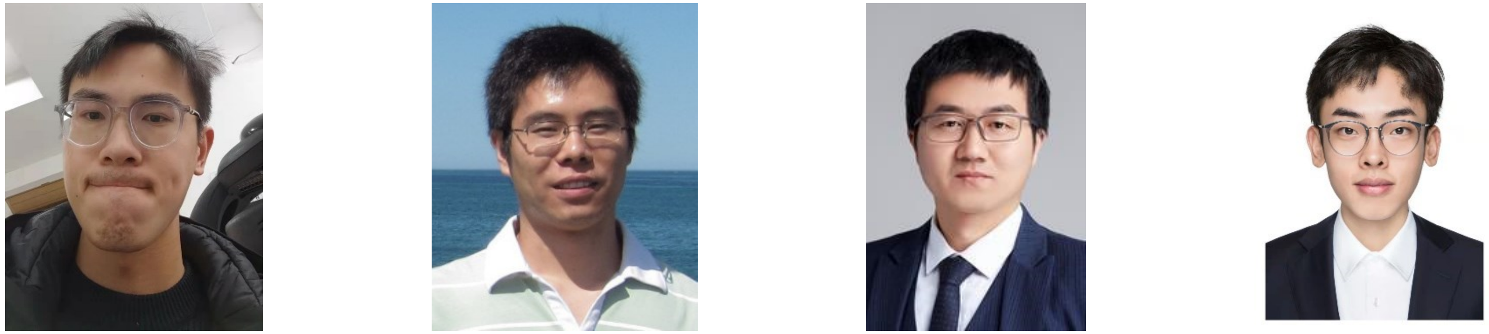
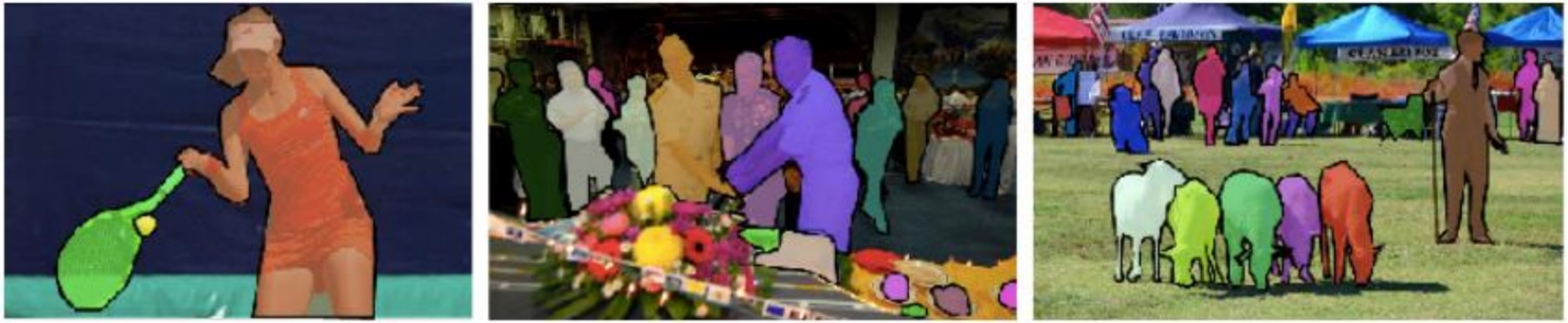
icassp 2022 Singapore

ISDA: POSITION-AWARE INSTANCE SEGMENTATION WITH DEFORMABLE ATTENTION



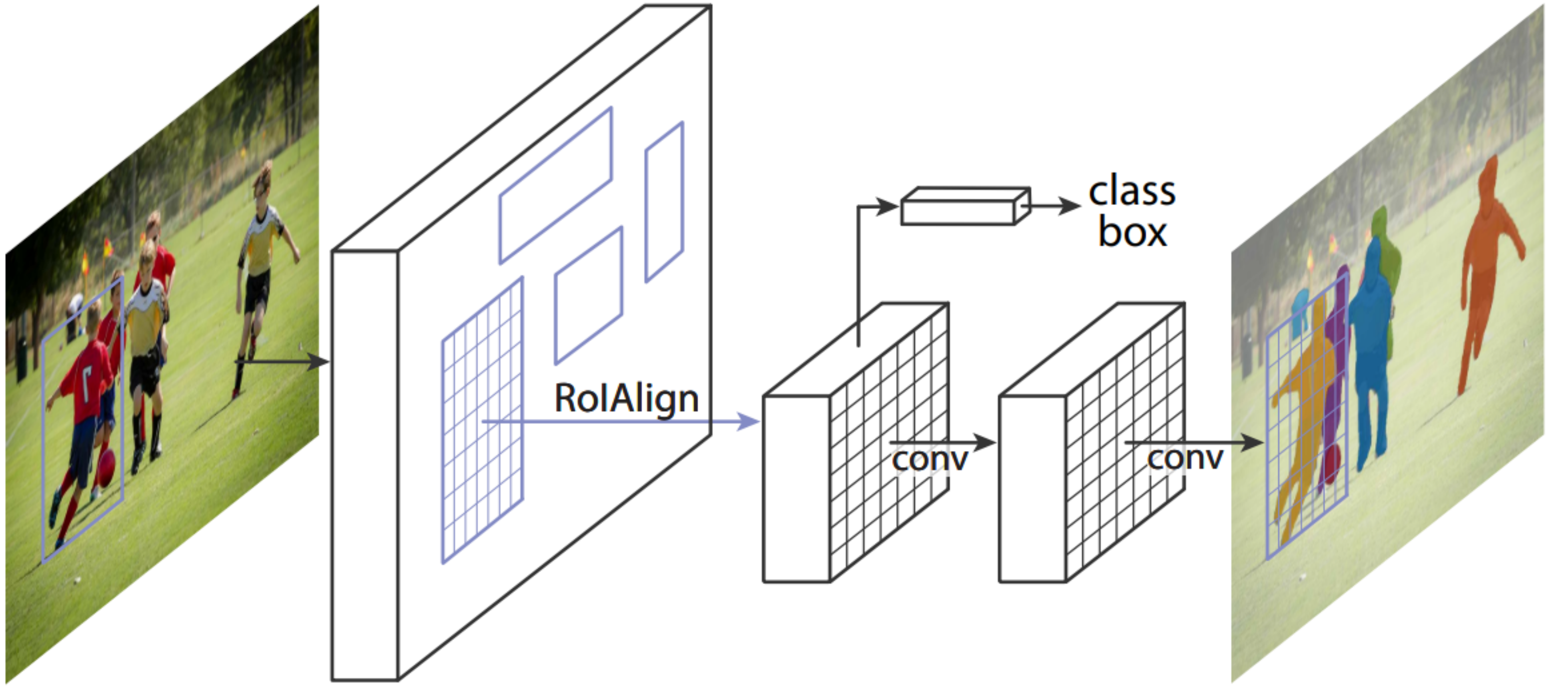
Kaining Ying, Zhenhua Wang*, Cong Bai, Pengfei Zhou
Zhejiang University of Technology
* zhwang@zjut.edu.cn

Instance Segmentation



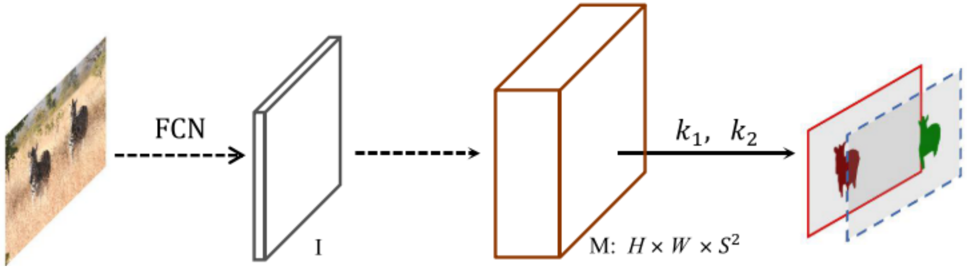
Difficulties: Require both **instance-level (holistic and coarse)** and **pixel-level results (local and fine).** In contrast, object detection only requires instance-level results, and semantic segmentation only requires pixel-level result.

Mask R-CNN

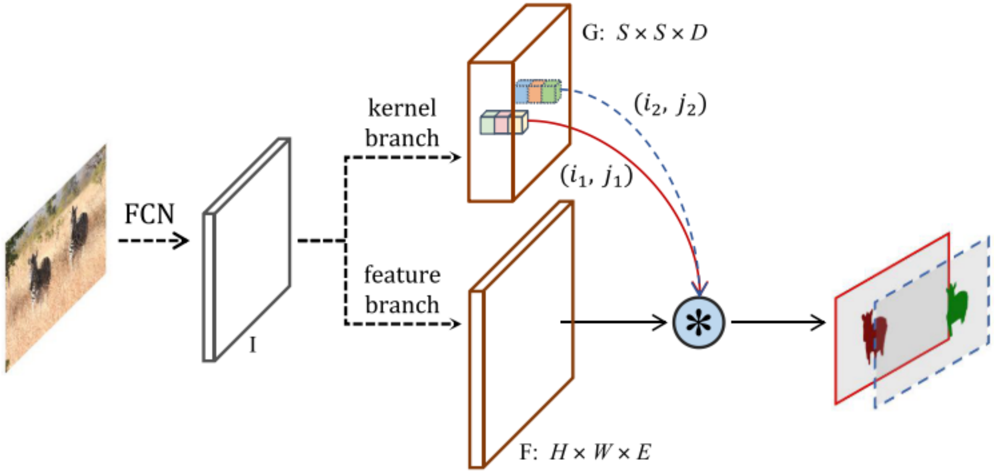


Challenge: RPN and NMS block end-to-end

Segmenting objects by locations



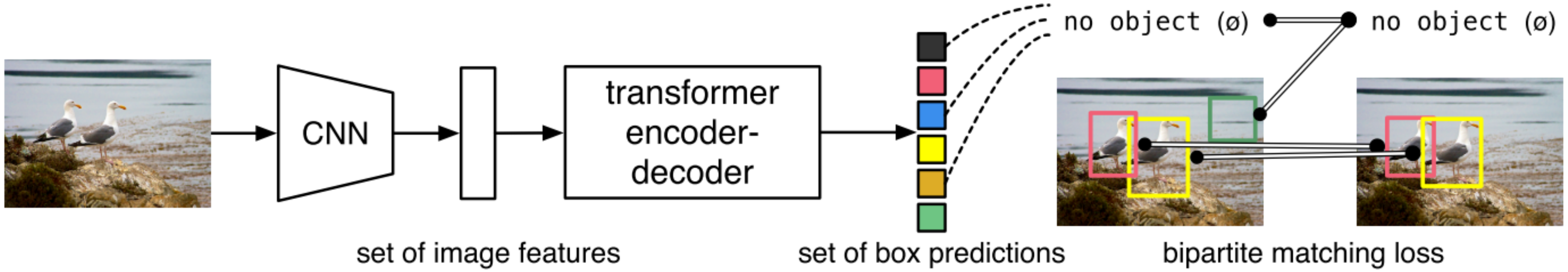
SOLO

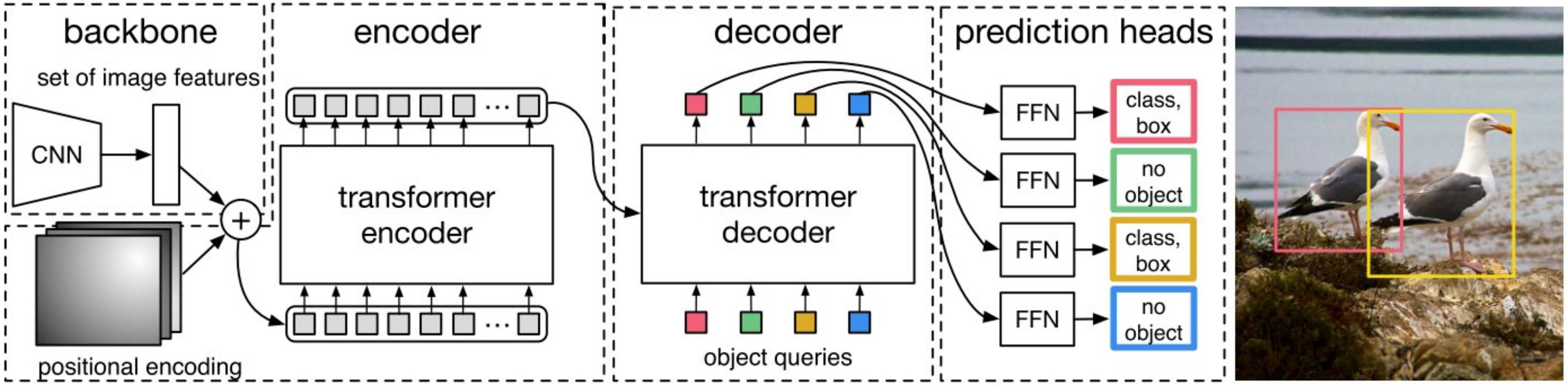


SOLOv2

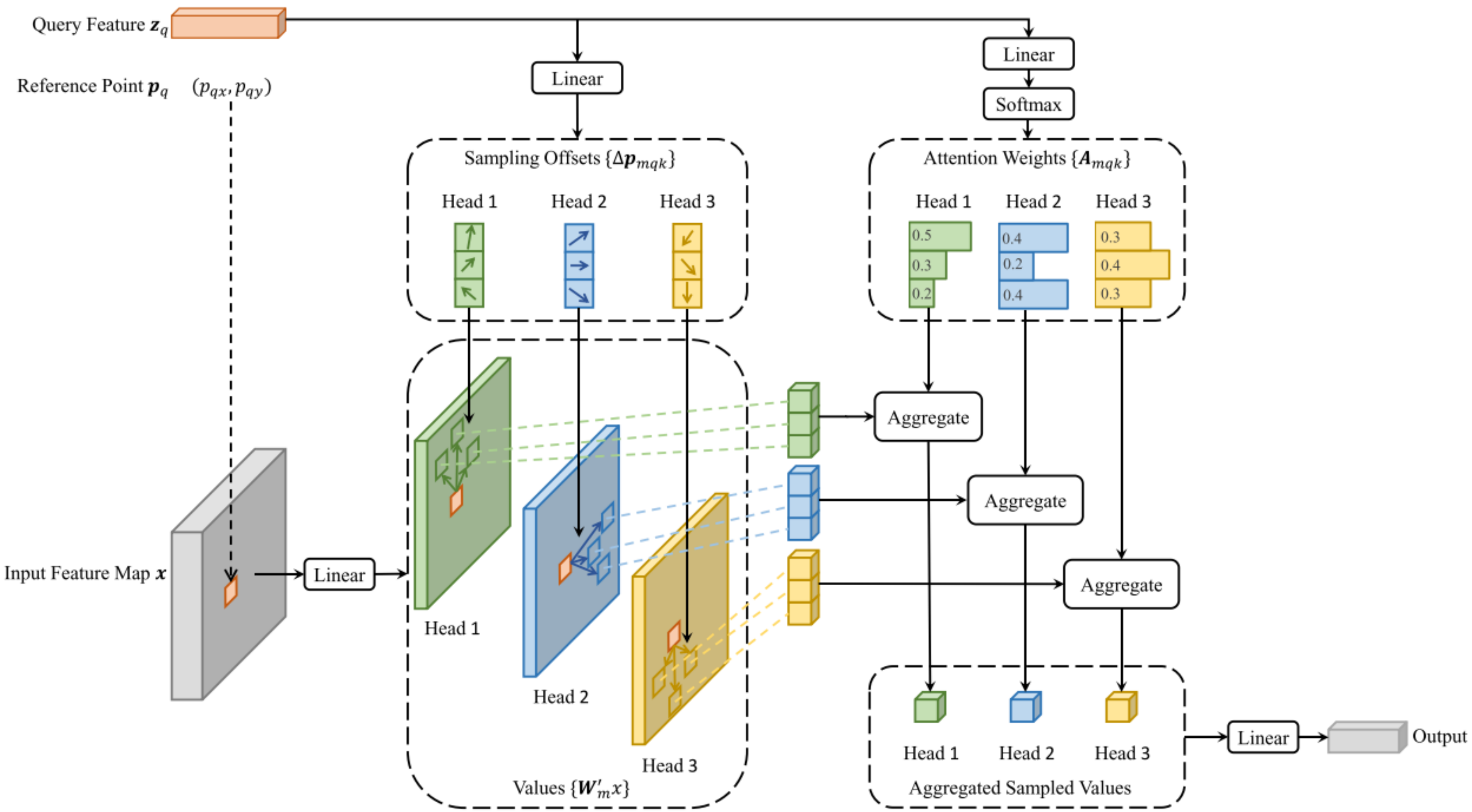
NMS still exits!

Detection Transformer

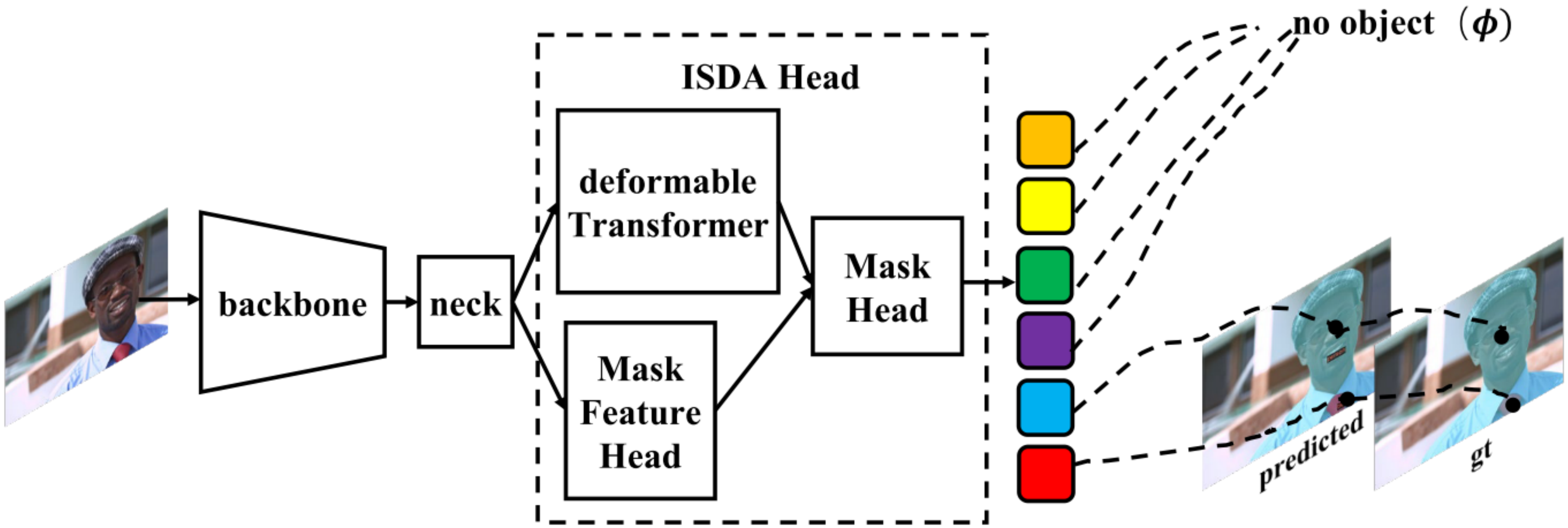




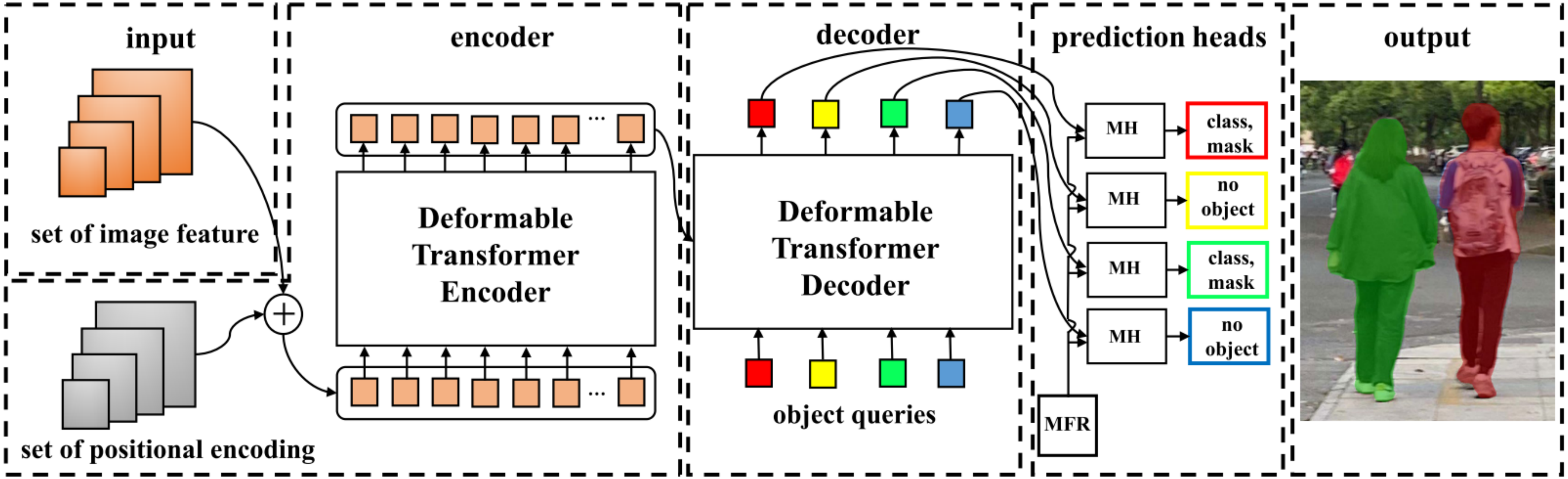
Deformable DETR



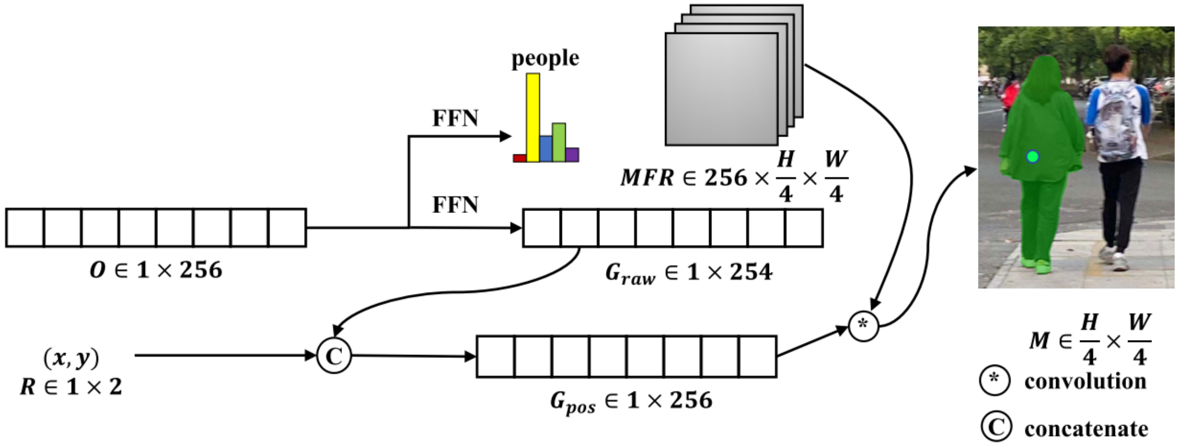
ISDA Overview



ISDA Head

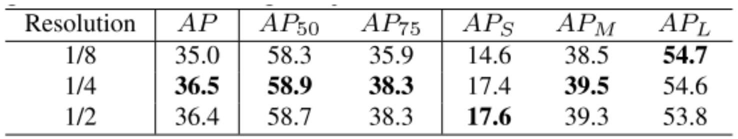

input
set of image feature
set of positional encoding
encoder
Deformable Transformer Encoder
...
decoder
Deformable Transformer Decoder
object queries
prediction heads
MH
MH
MH
MH
class, mask
no object
class, mask
no object
MFR
output

Mask Head


people
FFN
FFN
$MFR \in 256 \times \frac{H}{4} \times \frac{W}{4}$
$O \in 1 \times 256$
$G_{raw} \in 1 \times 254$
(x, y)
$R \in 1 \times 2$
C
$G_{pos} \in 1 \times 256$
*
$M \in \frac{H}{4} \times \frac{W}{4}$
* convolution
C concatenate

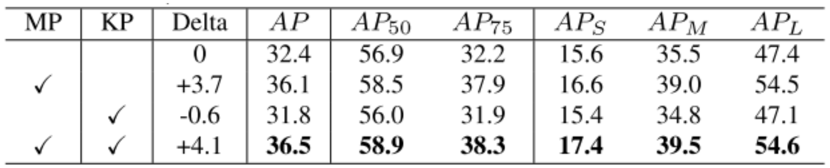
Quantitative results

Ablation Study on mask resolution



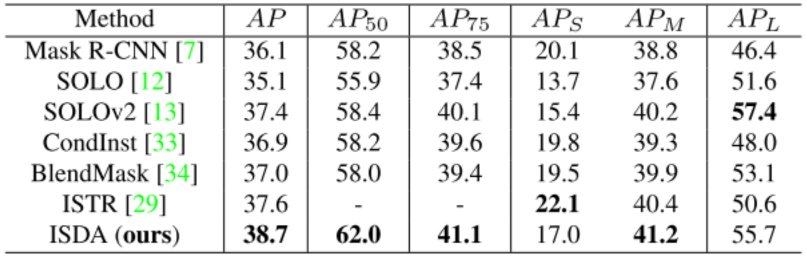
Resolution	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
1/8	35.0	58.3	35.9	14.6	38.5	**54.7**
1/4	**36.5**	**58.9**	**38.3**	17.4	**39.5**	54.6
1/2	36.4	58.7	38.3	**17.6**	39.3	53.8

Ablation Study on positional information



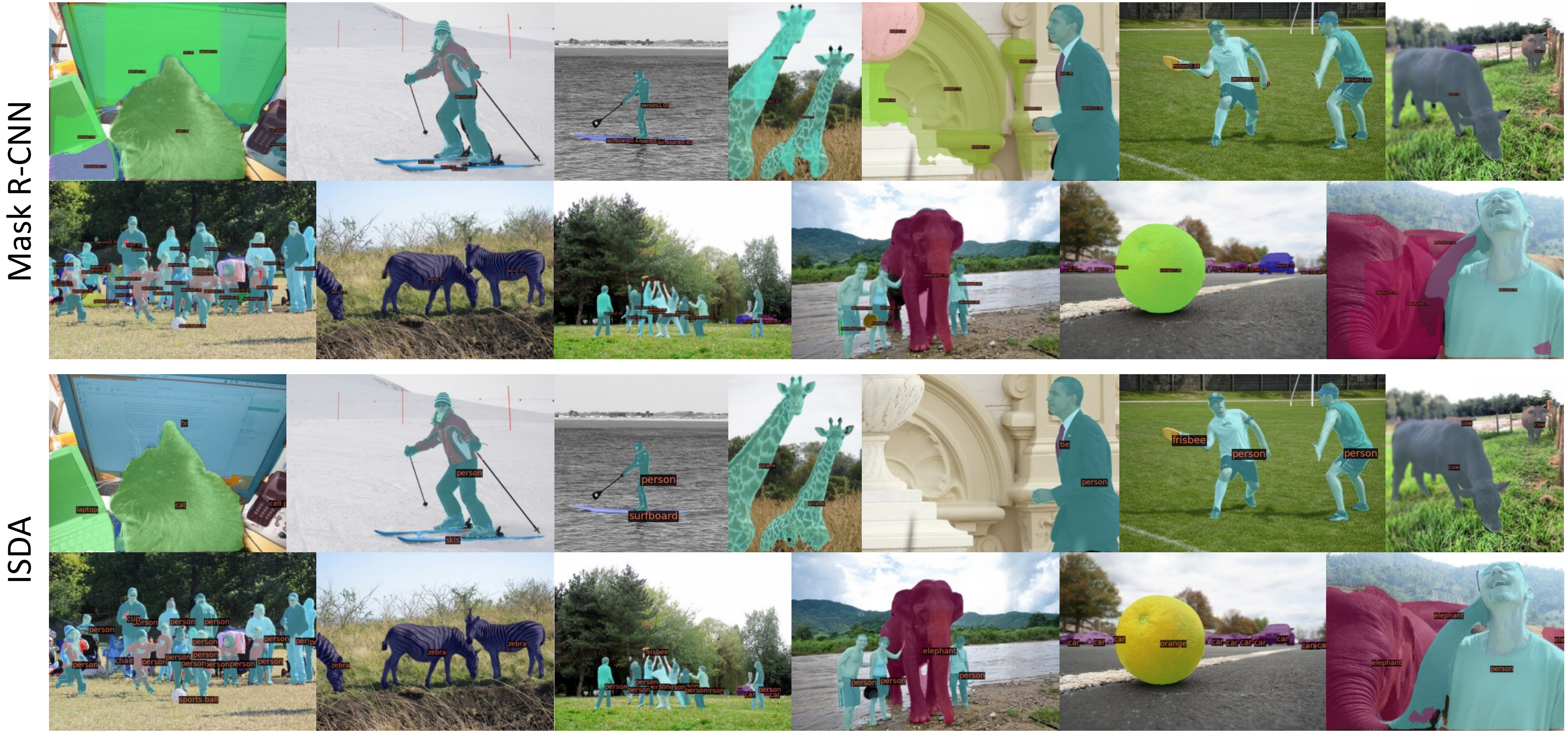
MP	KP	Delta	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
		0	32.4	56.9	32.2	15.6	35.5	47.4
✓		+3.7	36.1	58.5	37.9	16.6	39.0	54.5
	✓	-0.6	31.8	56.0	31.9	15.4	34.8	47.1
✓	✓	+4.1	**36.5**	**58.9**	**38.3**	**17.4**	**39.5**	**54.6**

Result on MS COCO



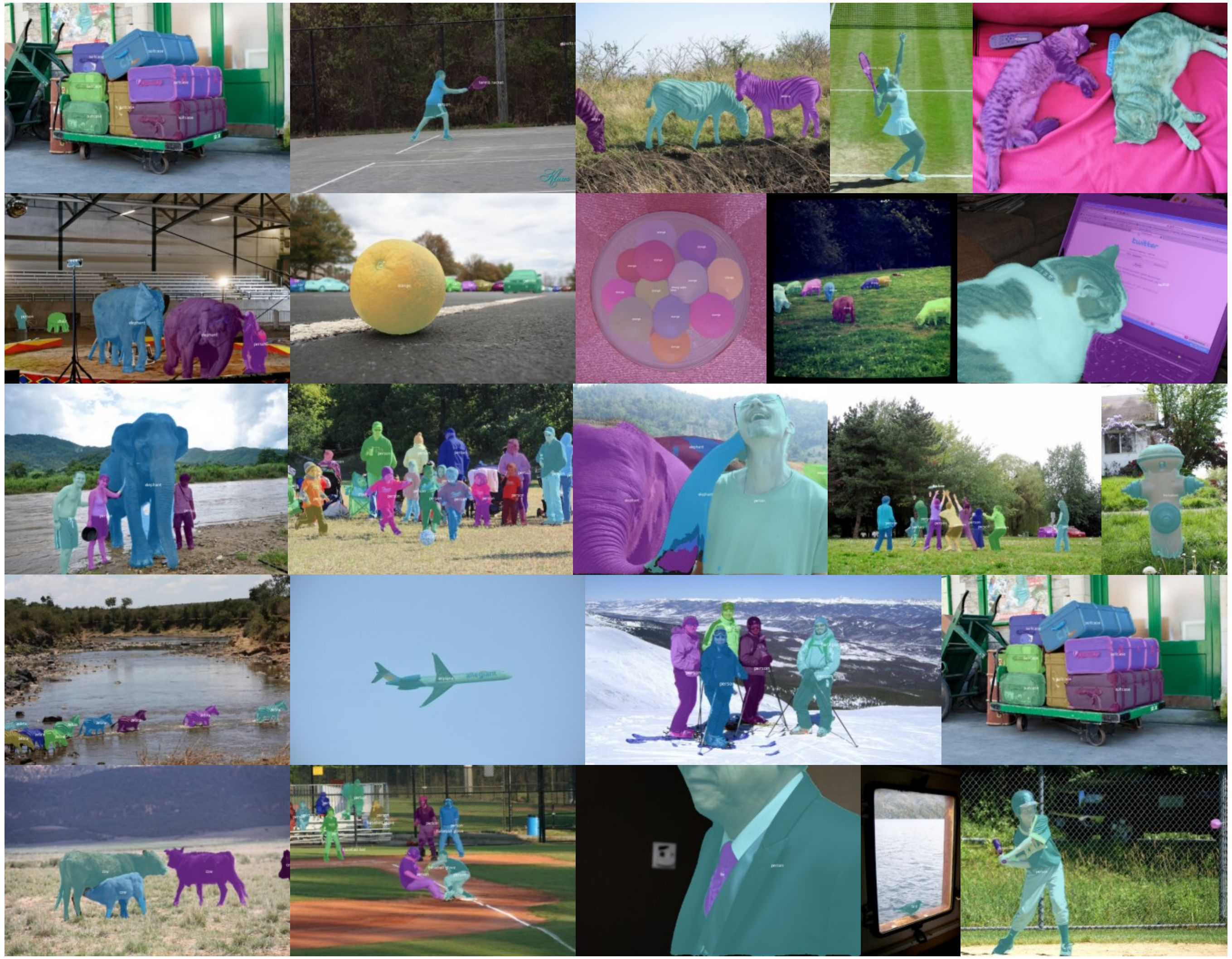
Method	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
Mask R-CNN [7]	36.1	58.2	38.5	20.1	38.8	46.4
SOLO [12]	35.1	55.9	37.4	13.7	37.6	51.6
SOLOv2 [13]	37.4	58.4	40.1	15.4	40.2	**57.4**
CondInst [33]	36.9	58.2	39.6	19.8	39.3	48.0
BlendMask [34]	37.0	58.0	39.4	19.5	39.9	53.1
ISTR [29]	37.6	-	-	**22.1**	40.4	50.6
ISDA (**ours**)	**38.7**	**62.0**	**41.1**	17.0	**41.2**	55.7

Qualitative results



Compare with Mask R-CNN

Qualitative results



More results

Conclusion

- ISDA introduced a Transformer-style framework for instance segmentation, which effectively removed NMS and achieved end-to-end training and inference
- ISDA is able to distinguish similar objects better by learning extra positional features
- ISDA gives SOTA results

Thank you!

浙江工业大学
ZHEJIANG UNIVERSITY OF TECHNOLOGY
1953