ISDA: POSITION-AWARE INSTANCE SEGMENTATION WITH DEFORMABLE ATTENTION

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

Keywords: Instance segementation, End-to-End, Transformer

icassp 2022 Singapore

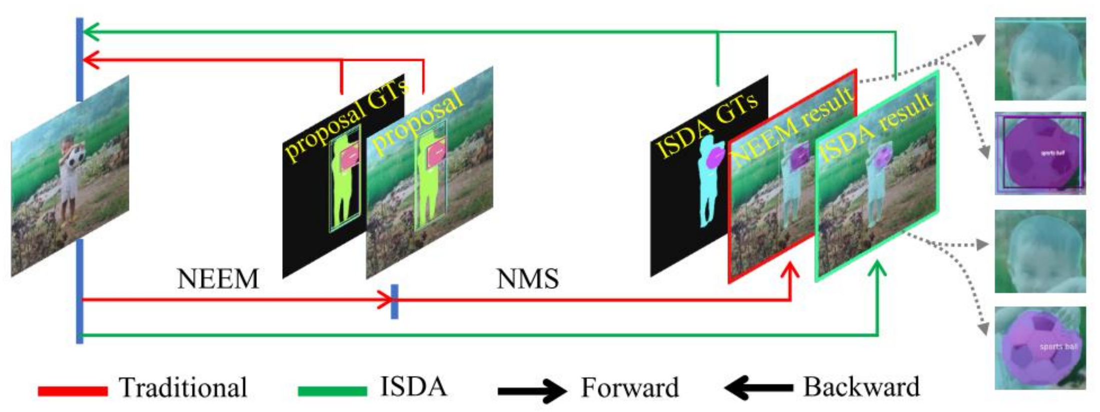
Paper#2331

ABSTRACT

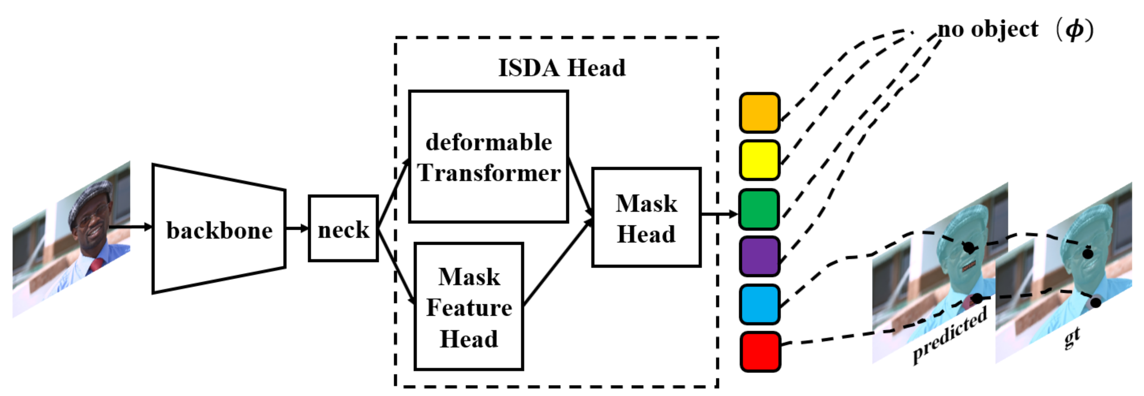
Most instance segmentation models are not end-to-end trainable due to either the incorporation of proposal estimation (RPN) as a pre-processing or non-maximum suppression as a post-processing. Here we propose a novel end-toend instance segmentation method termed ISDA. It reshapes the task into predicting a set of object masks, which are generated via traditional convolution operation with learned position-aware kernels and features of objects. Such kernels and features are learned by leveraging a deformable attention network with multi-scale representation. Thanks to the introduced set-prediction mechanism, the proposed method is NMS-free. Empirically, ISDA outperforms Mask R-CNN (the strong baseline) by 2.6 points on MS-COCO, and achieves leading performance compared with recent models. Code is available at https://github.com/yingkaining/isda.

Motivation

- Traditional instance segmentation models are not end-to-end trainable due to either the incorporation of proposal estimation (RPN) as a pre-processing or non-maximum suppression (NMS) as a post-processing.
- We propose a novel end-to-end instance segmentation method termed ISDA.



Model Overview

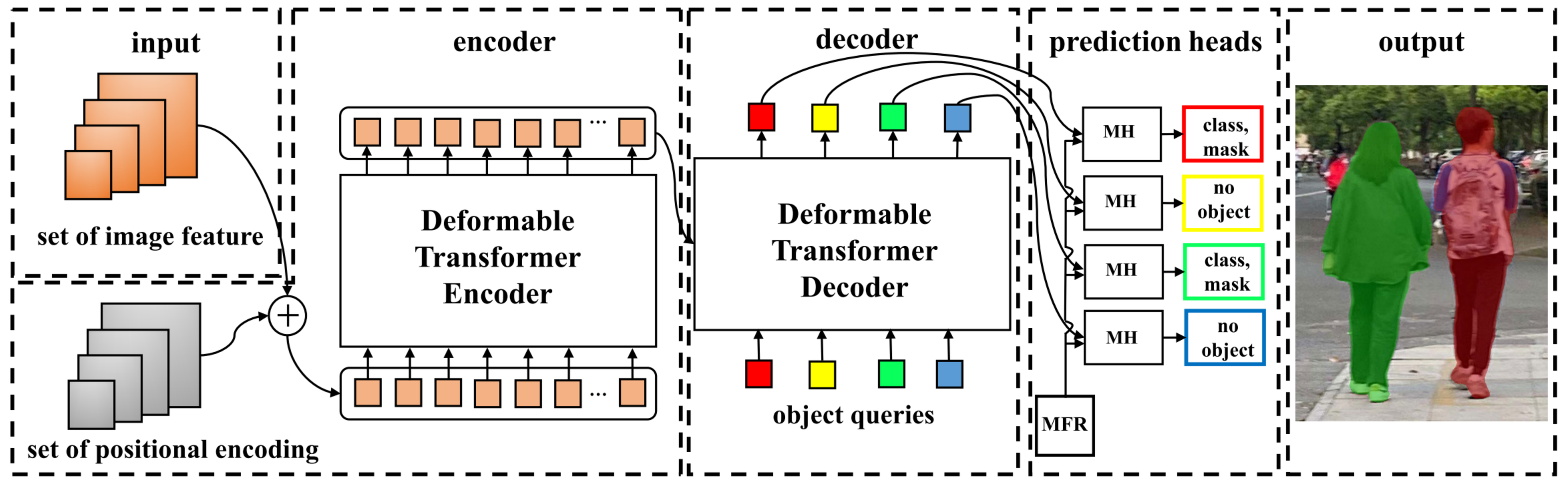


- Our ISDA includes three components
  - The backbone and neck module to extract multi-scale feature
  - The ISDA head which includes a deformable Transformer, a mask feature head and a mask head to predict object masks
  - The Bipartite matching block which associates predictions with ground truth to compute loss

Backbone and neck

Given an image denoted by $x \in R^{3 \times H \times W}$, the CNN backbone extracts four feature maps with different resolutions, denoted *by* $\{C_i \in R^{c_i \times H_i \times W_i}\}_{i=2}^{5}$. Here c_i ,H_i,W_i denote the channel number, the height and the width of the feature map C_i. The neck takes the multi-scale features as the input and then enhances them separately as that done by deformable DETR. Consequently, we get $\{P_i \in R^{256 \times H_i \times W_i}\}_{i=2}^{6}$, where P_6 is down-sampled form C_5.

ISDA Head

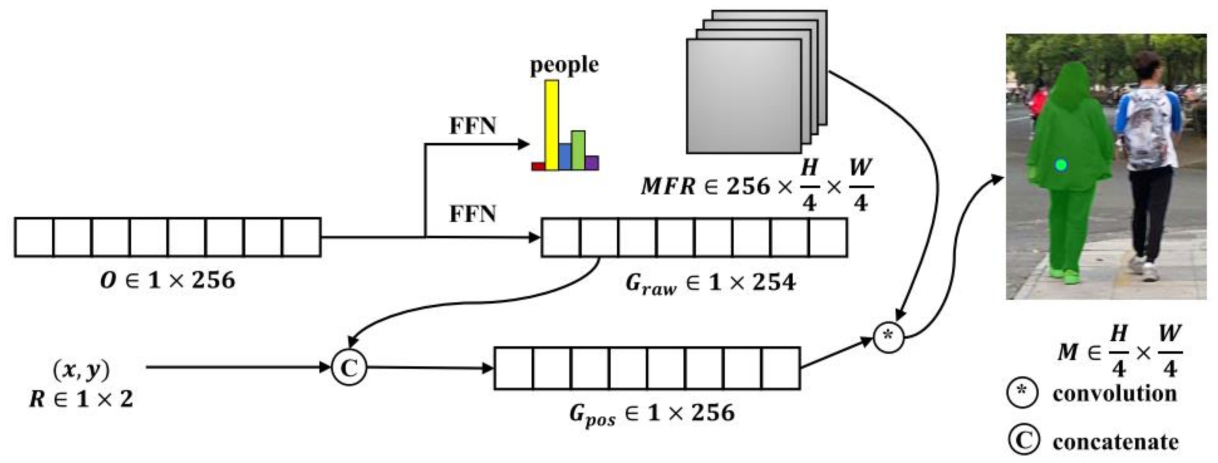


- Our ISDA includes three components
  - An encoder-decoder deformable Transformer
  - A mask feature head used to generate mask feature representations (MFR)
  - a mask head to make final predictions

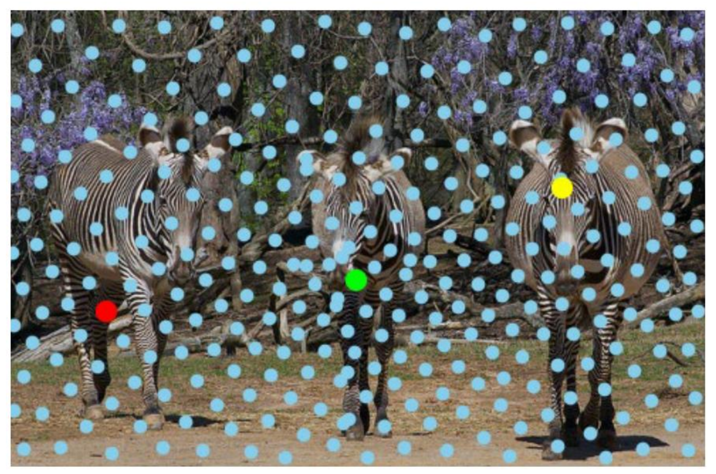
Mask Feature Representation

Inspired by SOLOv2[],ISDA learns a compact and high-resolution mask feature representation (MFR) with feature pyramid. After repeated stages of 3×3 Conv, group-norm, ReLU and $2 \times$ bilinear upsampling, the neck features $\{P_i\}_{i=2}^{5}$ are fused (via elementwise summation) to create one single output at 1/4 scale. It is worth noting that normalized pixel coordinates are fed into the smallest feature map (at $1/32$ scale) before convolution and upsampling.

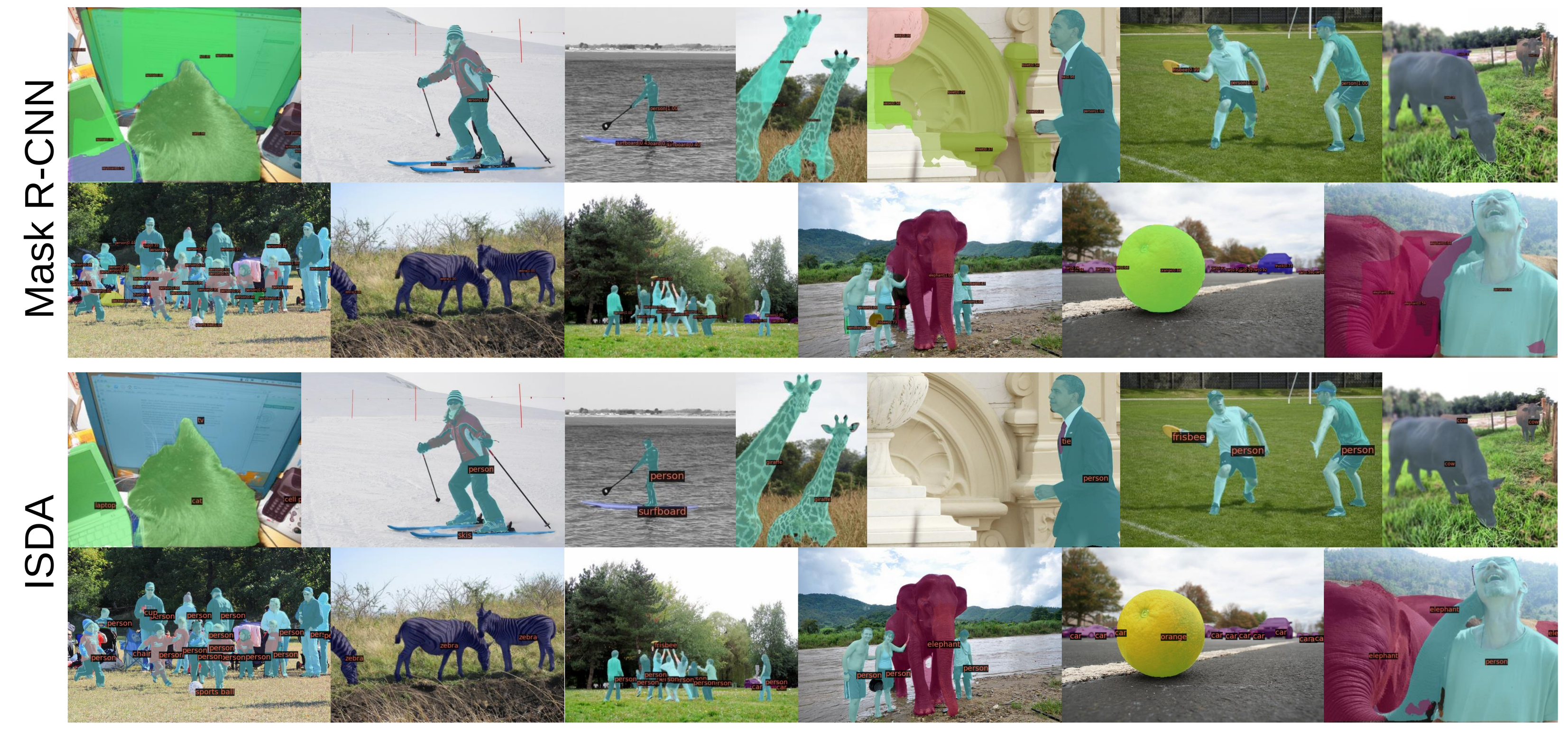
Mask Head



Visualization of reference points



Experimental Result



Ablation 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 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**

Results 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