AISFormer: Amodal Instance Segmentation with Transformer

Minh Tran[1], Khoa Vo[1], Kashu Yamazaki[1], Arthur Fernandes[2], Michael Kid[3], Ngan Le[1]
[1]Department of CSCE, University of Arkansas [2]Cobb-Vantress, Inc [3]Poultry Science, University of Arkansas

Code

BMVC 2022

Introduction

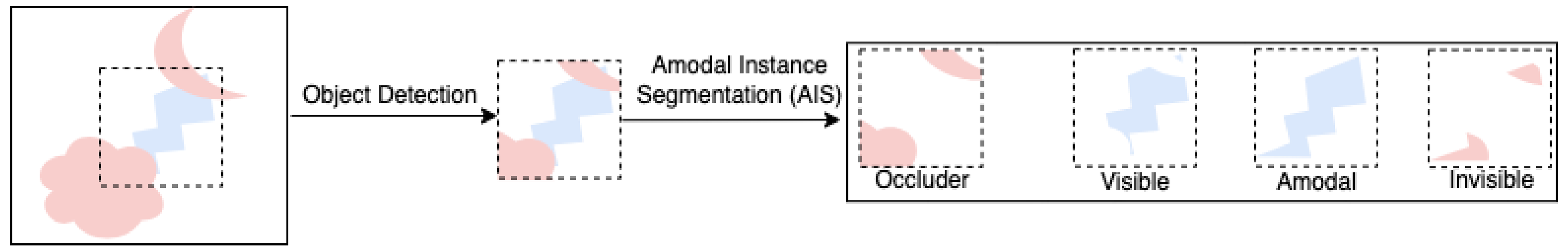

Figure 1: An explanation of different mask instances in Amodal Instance Segmentation (AIS). Given a region of interest (ROI) extracted by an object detector, AIS aims to extract both visible and invisible mask instances including occluder, visible, amodal, and invisible.

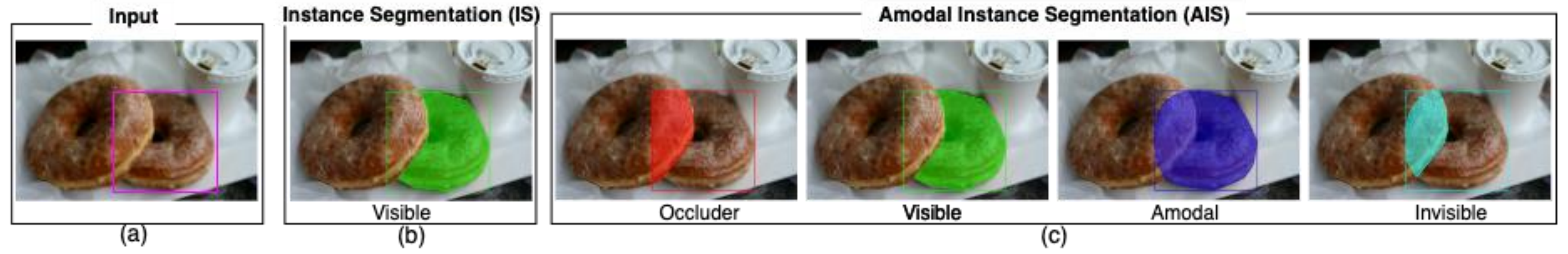

Figure 2: A comparison between Instance Segmentation (IS) and Amodal Instance Segmentation (AIS). Given an image with ROI (a), IS aims to extract the visible mask instance (b) whereas AIS aims to extract both the visible mask and occluded parts (c).

Contributions

- *We propose **AISFormer**, an amodal instance segmentation framework, with a Transformer-based mask head. Our AISFormer can explicitly model the complex coherence between occluder, visible, amodal, and invisible masks within an object's regions of interest by treating them as learnable queries. AISFormer also models the relationship between these embeddings and the corresponding region of interest.*

- *We empirically validate the usefulness of our proposed method by showing that it achieves superior performance to most of the current state-of-the-art methods benchmarked on three amodal datasets, i.e., KINS, COCOA-cls, and D2SA.*

Network Architecture

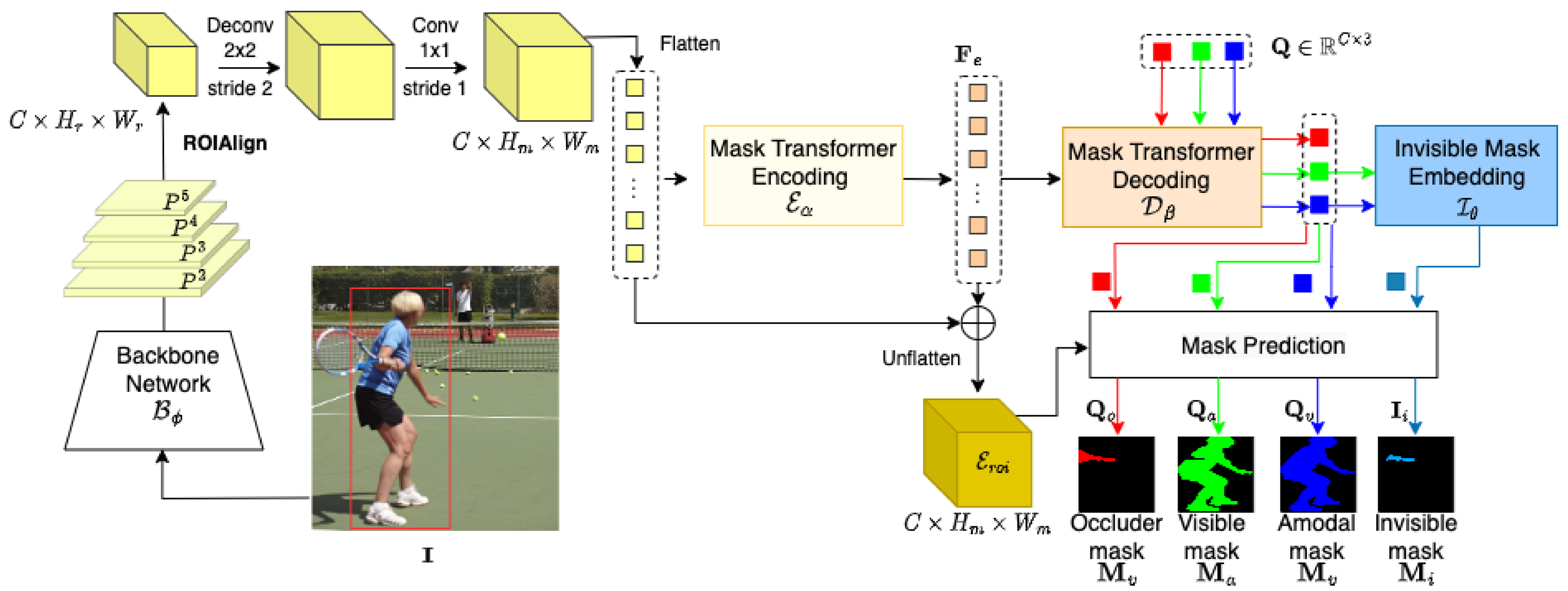

Figure 3: The overall flowchart of our proposed AISFormer. AISFormer consists of four modules corresponding to (i) feature encoding.(ii) mask transformer decoding (iii) invisible mask embedding (iv) segmentation to estimate output masks.

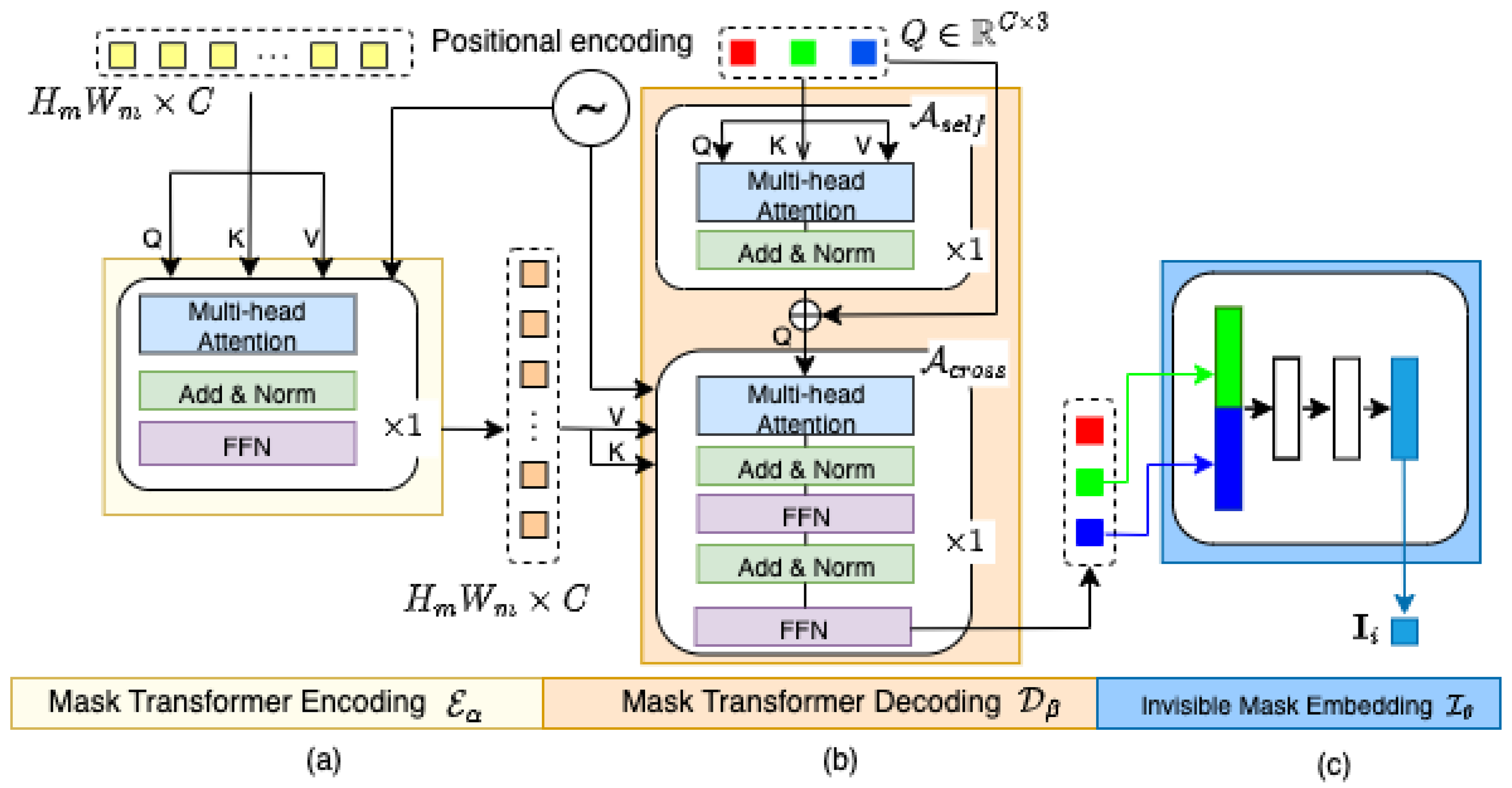

Figure 4: Details of AISFormer. (a): mask transformer encoder is designed as one block of self-attention, (b): mask transformer decoder is designed as a combination of one block of self-attention and one block of cross-attention and (c): invisible embedding is designed as an MLP with two hidden layers.

Experimental Results

Table 1: Performance comparison on KINS dataset. † indicates our reproduced results. On each backbone, the best scores are in **bold** and the second best scores are in underlines.

Sup.	Backbone	Model	Venue	Shape Prior	AP	AP_{50}	AP_{75}	AR
Weakly	ResNet-50	PCNet [44]	CVPR'20	✗	29.1	51.8	29.6	18.3
	ResNet-50	ABSU [31]	ICCV'21	✓	29.3	52.1	29.7	18.4
Fully	ResNet-50	VQVAE [21]	IEEE TMI'20	✓	30.3	–	–	–
	ResNet-50	VRSP-Net [41]	AAAI'21	✓	32.1	55.4	33.3	20.9
	ResNet-50	Mask R-CNN [19]	ICCV'17	✗	30.0	54.5	30.1	19.4
	ResNet-50	ORCNN [16]	WACV'19	✗	30.6	54.2	31.3	19.7
	ResNet-50	ASN [32]	CVPR'19	✗	32.2	–	–	–
	ResNet-50	**AISFormer**	–	✗	**33.8**	**57.8**	**35.3**	**21.1**
	ResNet-101	Mask R-CNN [19]†	ICCV'17	✗	30.2	54.3	30.4	19.5
	ResNet-101	BCNet [23]	CVPR'21	✗	28.9	–	–	–
	ResNet-101	BCNet [23]†	CVPR'21	✗	32.6	57.2	35.4	21.5
	ResNet-101	**AISFormer**	–	✗	**34.6**	**58.2**	**36.7**	**21.9**
	RegNet	APSNet [29]	CVPR'22	✗	**35.6**	–	–	–
	RegNet	**AISFormer**	–	✗	**35.6**	**59.9**	**37.0**	**22.5**

Table 2: Performance comparison on the D2SA and COCOA-cls datasets with ResNet-50 as the backbone. † indicates our reproduced results. In the category of without shape prior, the best scores are in **bold** and the second best scores are in underlines.

Model	Venue	Shape Prior	D2SA				COCOA-cls			
			AP	AP_{50}	AP_{75}	AR	AP	AP_{50}	AP_{75}	AR
VRSP-Net [41]	AAAI'21	✓	70.27	85.11	75.81	69.17	35.41	56.03	38.67	37.11
Mask R-CNN [19]	ICCV'17	✗	63.57	83.85	68.02	65.18	28.03	53.68	25.36	29.83
ORCNN [16]	WACV'19	✗	64.22	83.55	69.12	65.25	28.03	53.68	25.36	29.83
ASN [32]†	CVPR'19	✗	63.94	84.35	69.57	65.20	35.33	58.82	37.10	35.50
BCNet [23]†	CVPR'21	✗	65.97	84.23	72.74	66.90	35.14	58.84	36.65	35.80
AISFormer	–	✗	**68.36**	**85.08**	**74.58**	**68.64**	**37.27**	**59.69**	**40.70**	**37.40**

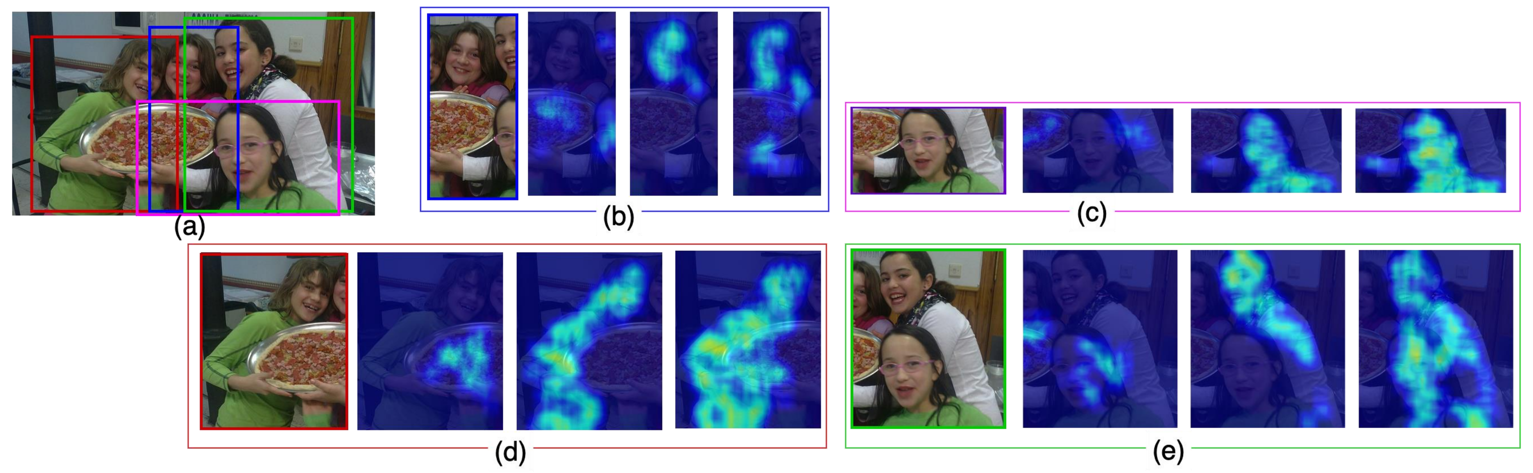

Figure 5: Attention visualization of query embeddings. (a): Input image with four ROIs. (b), (c), (d), (e): attention feature maps of queries in each ROI. For each ROI, from left-right: ROI, occluder query embedding, visible query embedding, and amodal query embedding.