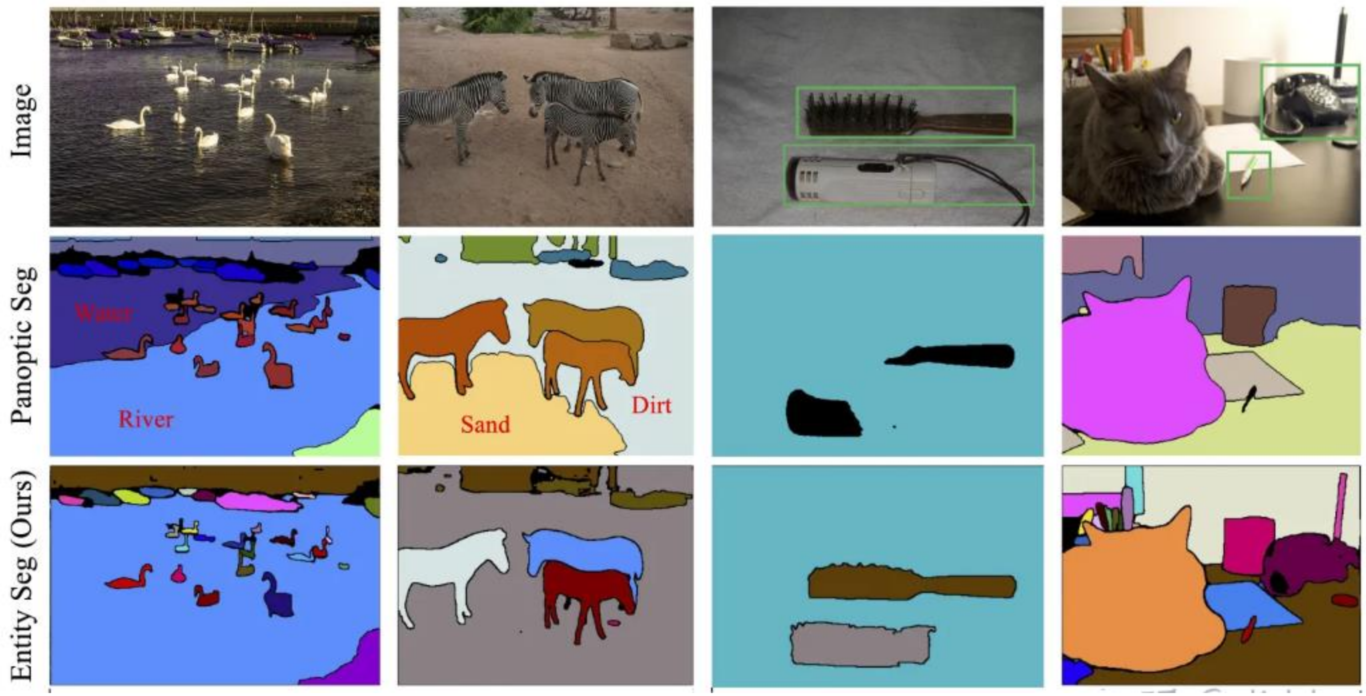
Fine-Grained Entity Segmentation

Lu Qi[1]*, Jason Kuen[2]*, Weidong Guo[3]*, Tiancheng Shen[4], Jiuxiang Gu[2], Wenbo Li[4], Jiaya Jia[4] Zhe Lin[2], Ming-Hsuan Yang[1]

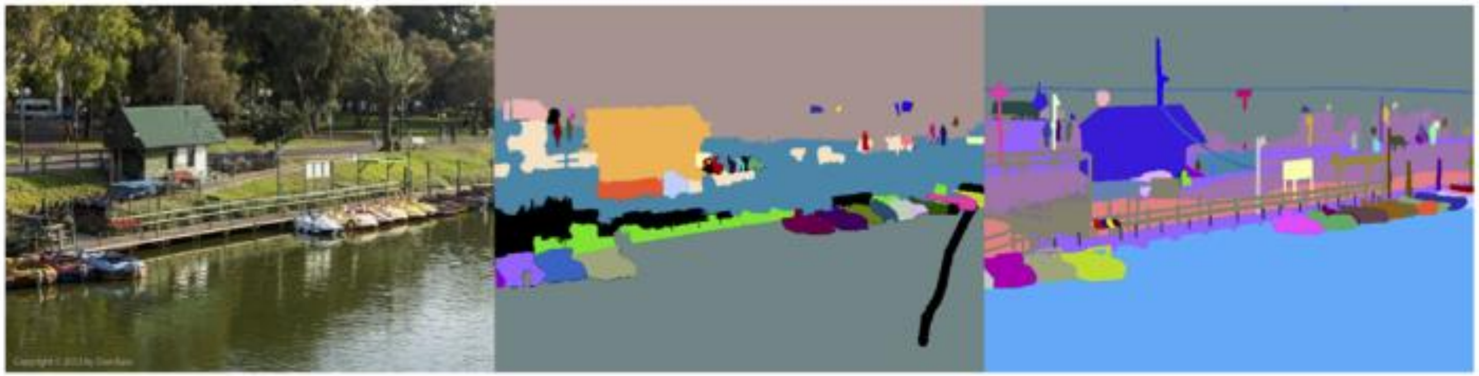
[1]The University of California, Merced
[2]Adobe Research
[3]QQ Browser Lab, Tencent
[4]The Chinese University of Hong Kong

Entity Segmentation



Each entity is a thing or stuff that does not consider category information

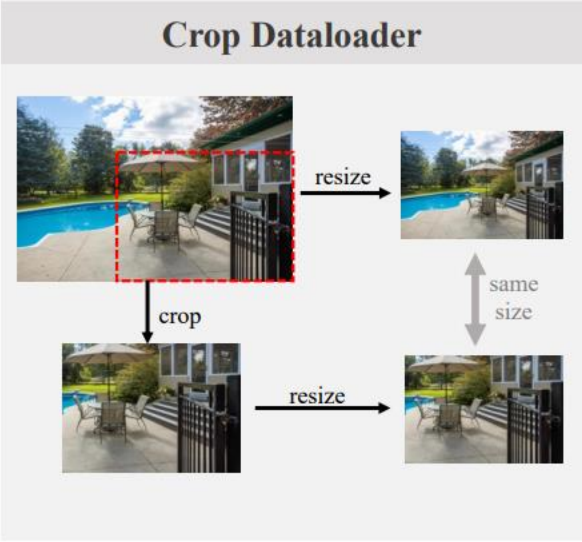
EntitySeg Dataset



- Large-scale
- High-quality
- High-resolution

CropFormer

Add high-resolution crop inputs to improve the quality of fine-grained entity segmentation.

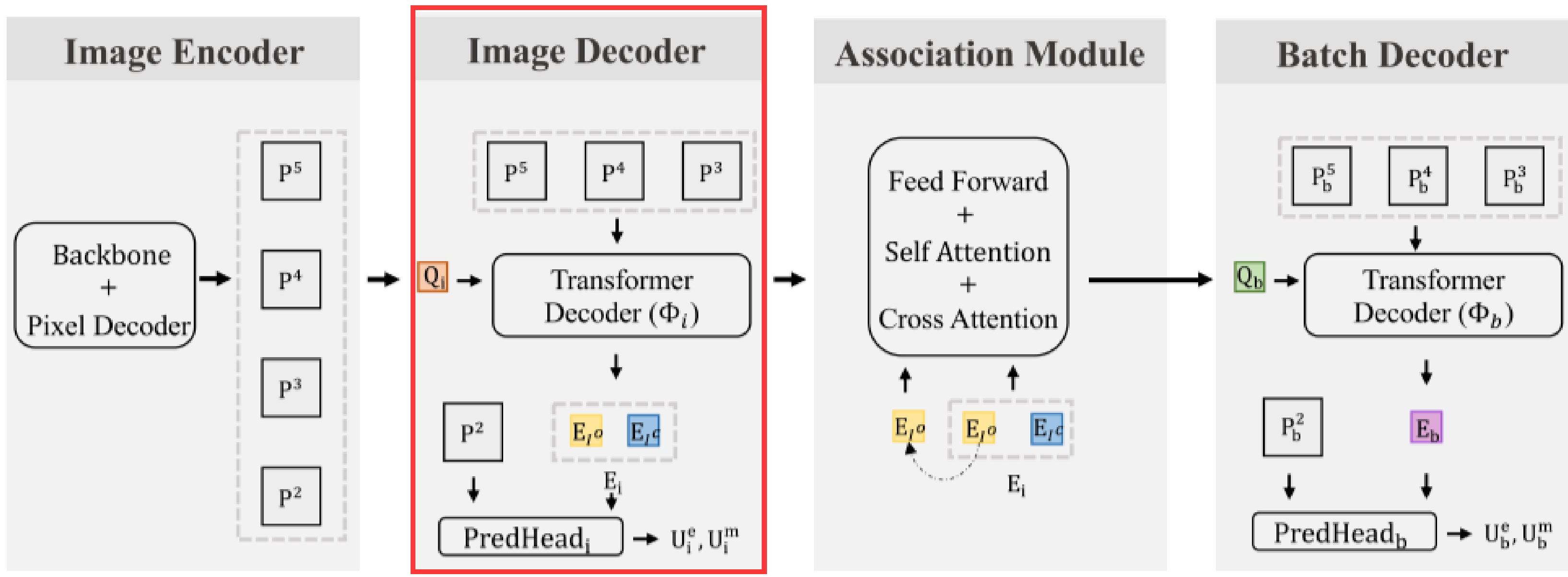


Crop & Resize

$$\downarrow$$

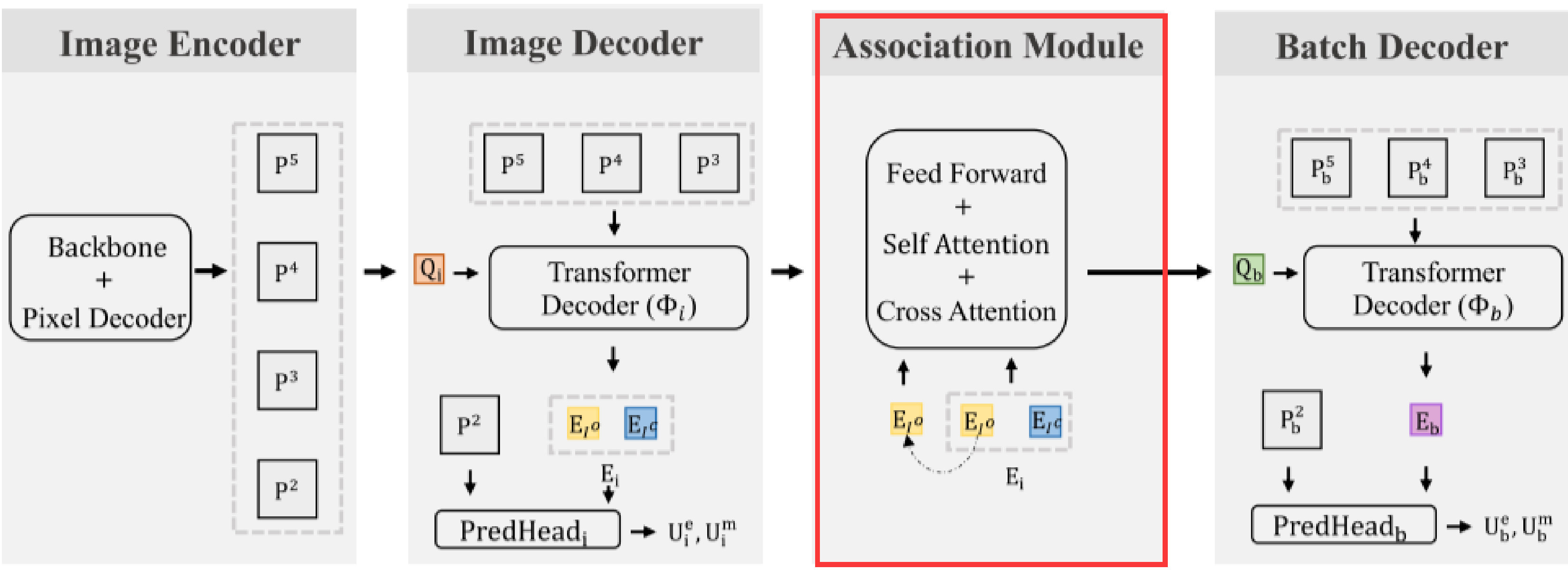
$$\boldsymbol{I} = \{I^O, I^C\} = \Gamma\{I^O, \delta\}$$

$$\boldsymbol{I} \in R^{N \times 2 \times H_I \times W_I \times 3}$$



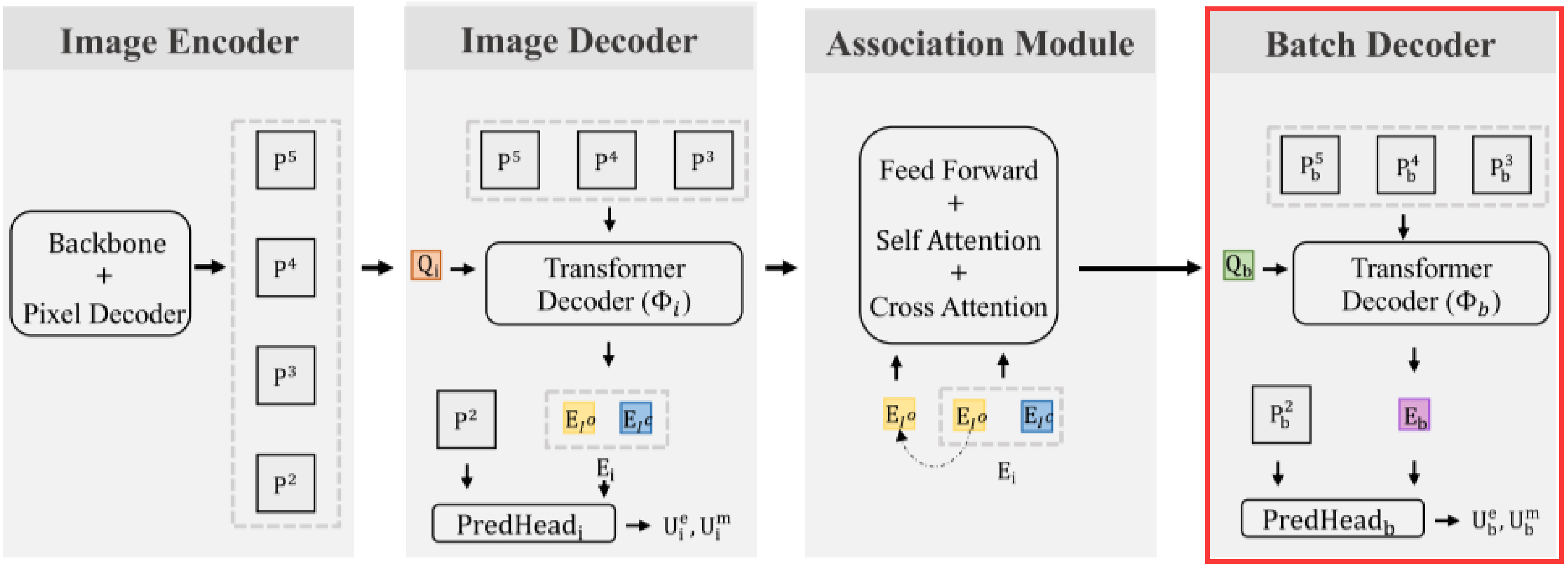
$$\mathbf{E_i} = \Phi_i(\mathbf{Q_i}, \Theta(\mathbf{I})) \qquad \boldsymbol{E_i} \in R^{N\times 2\times 1\times 1\times K}$$

$$\mathbf{U_i^e}, \mathbf{U_i^m} = \mathrm{PredHead_i}(\mathbf{E_i}, \mathbf{P_i^h})$$



Generate batch queries Q_b that are fully shared by the full image and its crop to represent the same entities consistently

$$\mathbf{Q_b} = \mathrm{FFN}(\mathrm{SAtt}(\mathrm{XAtt}(\underbrace{f_q(\mathbf{E}_{I^o})}_{\text{query}}, \underbrace{f_k(\mathbf{E_i})}_{\text{key}}, \underbrace{f_v(\mathbf{E_i})}_{\text{value}}))))$$



$$\mathbf{E_b} = \Phi_b(\mathbf{Q_b}, \Theta(\mathbf{I}))$$

$\boldsymbol{Q_b} \in R^{N\times1\times1\times1\times K}$, $\boldsymbol{E_b} \in R^{N\times1\times1\times1\times K}$

broadcast

$\boldsymbol{E_b} \in R^{N\times2\times1\times1\times K}$

$$\mathbf{U_b^e}, \mathbf{U_b^m} = \mathrm{PredHead_b}(\mathbf{E_b}, \mathbf{P_b^h})$$

$$\mathcal{L} = \sum_{\mathbf{k} \in \{\mathbf{i}, \mathbf{b}\}} \mathcal{L}_{\mathbf{k}}^{\text{ce}}(\mathbf{U}_{\mathbf{k}}^{\mathbf{e}}, \mathbf{G}_{\mathbf{k}}^{\mathbf{e}}) + \sum_{\mathbf{k} \in \{\mathbf{i}, \mathbf{b}\}} \mathcal{L}_{\mathbf{k}}^{\text{bce}}(\mathbf{U}_{\mathbf{k}}^{\mathbf{m}}, \mathbf{G}_{\mathbf{k}}^{\mathbf{m}}) + \sum_{\mathbf{k} \in \{\mathbf{i}, \mathbf{b}\}} \mathcal{L}_{\mathbf{k}}^{\text{dice}}(\mathbf{U}_{\mathbf{k}}^{\mathbf{m}}, \mathbf{G}_{\mathbf{k}}^{\mathbf{m}}), \tag{7}$$

two separate losses L_i and L_b for image- and batch-level predictions

Method	Decoder	AP^e	AP^e_{50}	AP^e_{75}	RT (ms)
SS-Mask2Former	Image-O	39.5	56.9	40.2	637
SS-Mask2Former($\times\delta$)	Image-O	39.9	57.4	40.3	876
MS-Mask2Former	Image-O	39.2	56.3	39.5	1324
MS-Mask2Former	Batch-OC	39.3	56.4	39.7	2783
CropFormer	Image-O	39.3	56.7	39.8	637
	Batch-O	39.1	56.6	39.7	1514
	Batch-C	40.2	57.5	40.8	1507
	Batch-OC	**41.0**	**58.4**	**41.9**	1545

Table 7: Ablation study on the ensemble strategy on full image and four crops. The 'Decoder' column indicates whether we use the inference result of the full image ('O'), four cropped patches ('C'), or both of them ('OC') from the 'Image' or 'Batch' decoder. Here, the run-time (RT) is the time of network forward except the data processing and calculated on A100 GPU.

δ	AP^e
0.5	38.5
0.6	40.2
0.7	**41.0**
0.8	40.9

(a)

Train	Test	AP^e
Random	Fixed (4)	39.7
Fixed (4)	Fixed (4)	41.0
Fixed (4)	Fixed (8)	**41.3**
Fixed (8)	Fixed (8)	41.0

(b)

XAtt	SAtt	FFN	AP^e
✓	○	○	40.7
✓	○	✓	40.8
✓	✓	○	40.8
✓	✓	✓	**41.0**

(c)

Table 9: Ablation study on the usage of crop ratio δ, crop type and association module in CropFormer. In sub-table (b), 'Random' indicates random crops and 'Fixed (4/8)' indicates 4 or 8 fixed corner crops. In sub-table(c), ✓and ○ means whether we use the module or not.