A Simple Image Segmentation Framework via In-Context Examples

Yang Liu
Zhejiang University, China
yangliu9610@zju.edu.cn

Chenchen Jing
Zhejiang University, China
jingchenchen@zju.edu.cn

Hengtao Li
Zhejiang University, China
liht@zju.edu.cn

Muzhi Zhu
Zhejiang University, China
zhumuzhi@zju.edu.cn

Hao Chen
Zhejiang University, China
haochen.cad@zju.edu.cn

Xinlong Wang
Beijing Academy of Artificial Intelligence
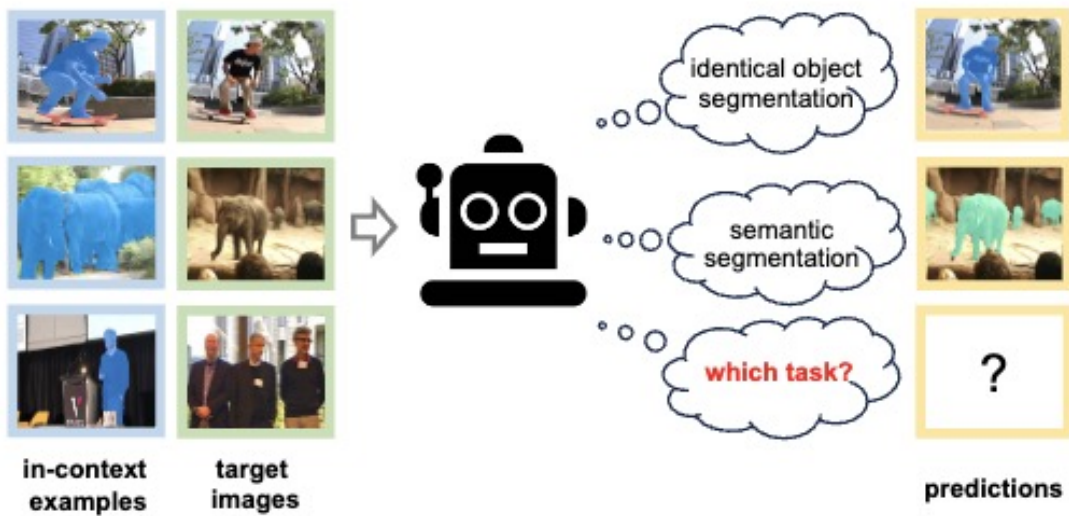wangxinlong@baai.ac.cn

Chunhua Shen
Zhejiang University, China
Ant Group
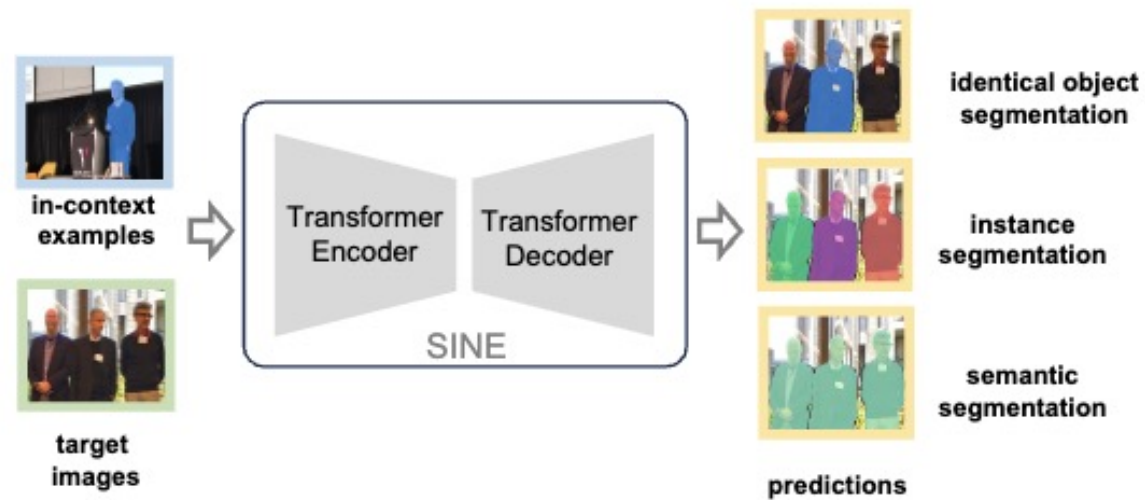chunhuashen@zju.edu.cn

in-context examples | target images

identical object segmentation

semantic segmentation

which task?

predictions
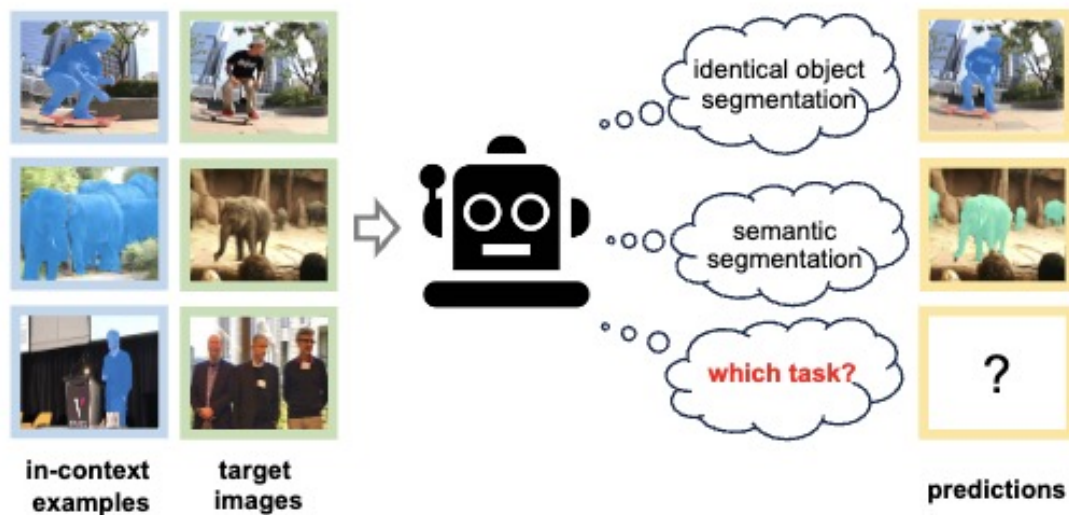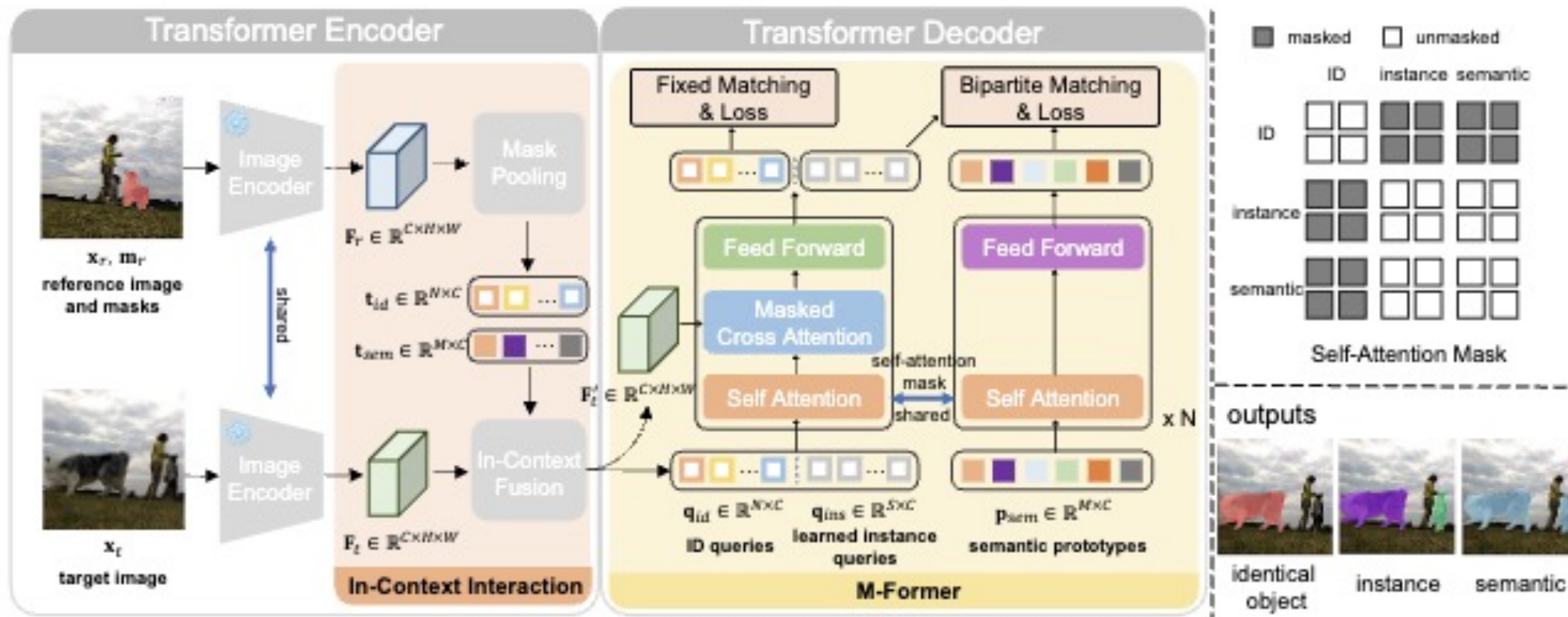
Method

- Frozen pre-trained image encoder
- In-context interaction module
- Matching Transformer

identical object · instance · semantic

(a) reference · target · SegGPT · SINE

(b) reference · GT · prediction · reference · GT · prediction · reference · GT · prediction

(c) reference · predictions

(d) first frame · predictions

Experiments

Table 1 — Results of few-shot semantic segmentation on COCO-20^i, PASCAL-5^i, and LVIS-92^i.

Methods	Venue	COCO-20^i one-shot	COCO-20^i few-shot	PASCAL-5^i one-shot	PASCAL-5^i few-shot	LVIS-92^i one-shot	LVIS-92^i few-shot
specialist model							
HSNet [41]	ICCV'21	41.7*	50.7*	68.7*	73.8*	17.4	22.9
VAT [20]	ECCV'22	42.9*	49.4*	72.4*	76.3*	18.5	22.7
FPTrans [62]	NeurIPS'22	56.5*	65.5*	77.7*	83.2*	-	-
generalist model							
Painter [54]	CVPR'23	32.8	32.6	64.5	64.6	10.5	10.9
SegGPT [56]	ICCV'23	56.1	67.9	83.2	89.8	18.6	25.4
PerSAM-F [63]	ICLR'24	23.5	-	-	-	18.4	-
Matcher [34]	ICLR'24	52.7	60.7	-	-	33.0	40.0
SINE	this work	64.5	66.1	85.4	86.2	31.2	35.5

Table 1 – Results of few-shot semantic segmentation on COCO-20^i, PASCAL-5^i, and LVIS-92^i. Gray indicates the model is trained by in-domain datasets. ∗ indicates that the categories in training cover the categories in testing within the same dataset.

Table 4 — Results of video object segmentation on DAVIS 2017, DAVIS 2016, and YouTube-VOS 2018.

Methods	Venue	DAVIS 2017 J&F	J	F	DAVIS 2016 J&F	J	F	YouTube-VOS 2018 G	J_s	F_s	J_u	F_u
with video data												
AGAME [22]	CVPR'19	70.0	67.2	72.7	-	-	-	66.0	66.9	-	61.2	-
AGSS [30]	ICCV'19	67.4	64.9	69.9	-	-	-	71.3	71.3	65.5	75.2	73.1
AFB-URR [28]	NeurIPS'20	74.6	73.0	76.1	-	-	-	79.6	78.8	83.1	74.1	82.6
AOT [60]	NeurIPS'21	85.4	82.4	88.4	92.0	90.7	93.3	84.5	84.3	89.3	77.9	86.4
SWEM [32]	CVPR'22	84.3	81.2	87.4	91.3	89.9	92.6	82.8	82.4	86.9	77.1	85.0
XMem [6]	ECCV'22	87.7	84.0	91.4	92.0	90.7	93.2	86.1	85.1	89.8	80.3	89.2
without video data												
Painter [54]	CVPR'23	34.6	28.5	40.8	70.3	69.6	70.9	24.1	27.6	35.8	14.3	18.7
SegGPT [56]	ICCV'23	75.6	72.5	78.6	83.7	83.6	83.8	74.7	75.1	80.2	67.4	75.9
SEEM [68]	NeurIPS'23	58.9	55.0	62.8	-	-	-	50.0	57.2	38.2	61.3	43.3
DINOv [27]	CVPR'24	73.3	71.0	75.7	-	-	-	60.9	65.3	70.0	52.3	57.9
PerSAM-F [63]	ICLR'24	76.1	74.9	79.7	-	-	-	54.4	53.9	56.4	50.7	56.6
Matcher [34]	ICLR'24	79.5	76.5	82.6	86.1	85.2	86.7	-	-	-	-	-
SINE	this work	77.0	72.6	81.3	82.3	81.4	83.2	66.2	69.1	57.6	71.7	66.5

Table 4 – Results of video object segmentation on DAVIS 2017, DAVIS 2016, and YouTube-VOS 2018. Gray indicates the model is trained on target datasets with video data. G is the average score over the "seen" and "unseen" classes in YouTube-VOS 2018.

Table 2

Methods	Venue	Det. 1	Det. 5	Segm. 1	Segm. 5
specialist model					
Meta R-CNN [59]	ICCV'19	-	3.5	-	2.8
MTFA [12]	CVPR'21	2.5	6.6	2.7	6.6
iMTFA	CVPR'21	3.3	6.2	2.8	5.2
Meta-DETR [61]	TPAMI'22	-	15.4	-	8.1
RefT [17]	arXiv'23	-	15.0	-	14.2
generalist model					
SINE	this work	18.0	22.5	16.9	21.8

Table 2 – Results (AP) of few-shot object detection and instance segmentation on COCO-NOVEL with $K = \{1, 5\}$.

Table 3

Methods	Venue	Det. AP	Det. AP50	Segm. AP	Segm. AP50
1-shot					
FGN [11]	CVPR'20	-	30.8	-	16.2
MTFA [12]	CVPR'21	10.0	21.7	9.5	19.3
iMTFA	CVPR'21	11.5	22.4	8.6	16.3
SINE	this work	35.9	51.9	27.6	47.6
5-shot					
SINE	this work	42.8	62.1	33.3	57.7

Table 3 – Results (AP and AP50) of few-shot object detection and instance segmentation on COCO2VOC with $K = \{1, 5\}$.

Table 5

Methods	Venue	mIoU
specialist model		
FCN [35]	CVPR'15	29.4
RefineNet [29]	CVPR'17	40.7
DPT [48]	ICCV'21	49.2
Mask2Former [5]	CVPR'22	57.7
generalist model		
Painter [54]	CVPR'23	49.9
SegGPT [56]	ICCV'23	39.6
SINE	this work	54.1

Table 5 – Transfer performance on ADE20K semantic segmentation.

Table 6

Methods	Venue	PQ	PQ^{Th}	PQ^{St}
specialist model				
PanopticFPN [24]	CVPR'19	40.8	48.3	29.4
SOLOv2 [55]	NeurIPS'20	42.1	49.6	30.7
Mask2Former [5]	CVPR'22	57.8	64.2	48.1
UViM [26]	NeurIPS'22	45.8	-	-
generalist model				
Painter [54]	CVPR'23	43.4	-	-
SegGPT [56]	ICCV'23	34.4	-	-
SINE	this work	51.0	57.8	40.8

Table 6 – Transfer performance on COCO panoptic segmentation.

Thanks for listening.