ICLR

UNITED IMAGING Intelligence

Order-aware Interactive Image Segmentation

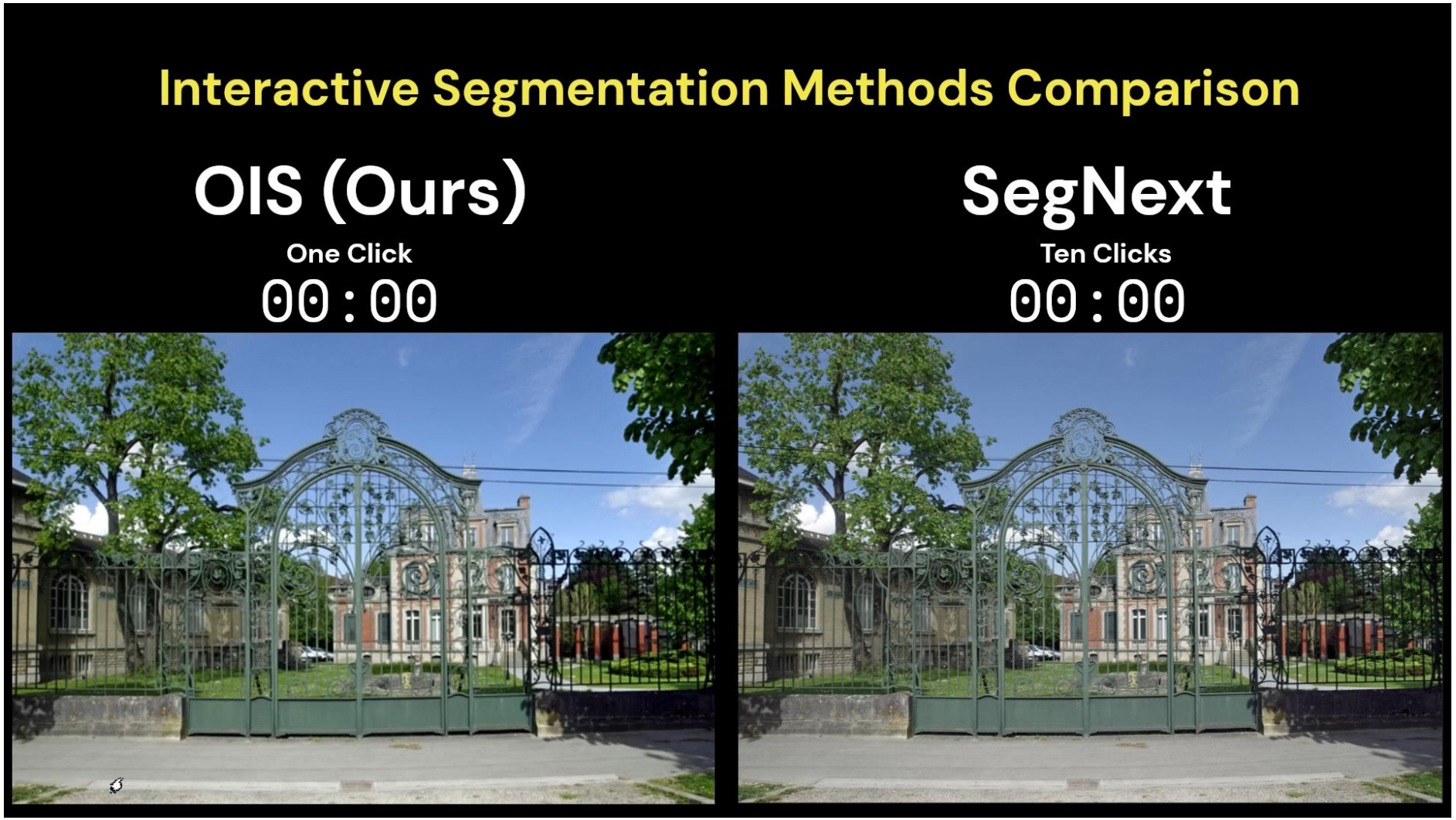
Bin Wang[1,2], Anwesa Choudhuri[1], Meng Zheng[1], Zhongpai Gao[1], Benjamin Planche[1], Andong Deng[1,3], Qin Liu[4], Terrence Chen[1], Ulas Bagci[2], Ziyan Wu[1]

[1]United Imaging Intelligence, Boston MA, USA
[2]Northwestern University, Chicago IL, USA
[3]University of Central Florida, Orlando FL, USA
[4]University of North Carolina at Chapel Hill, Chapel Hill NC, USA

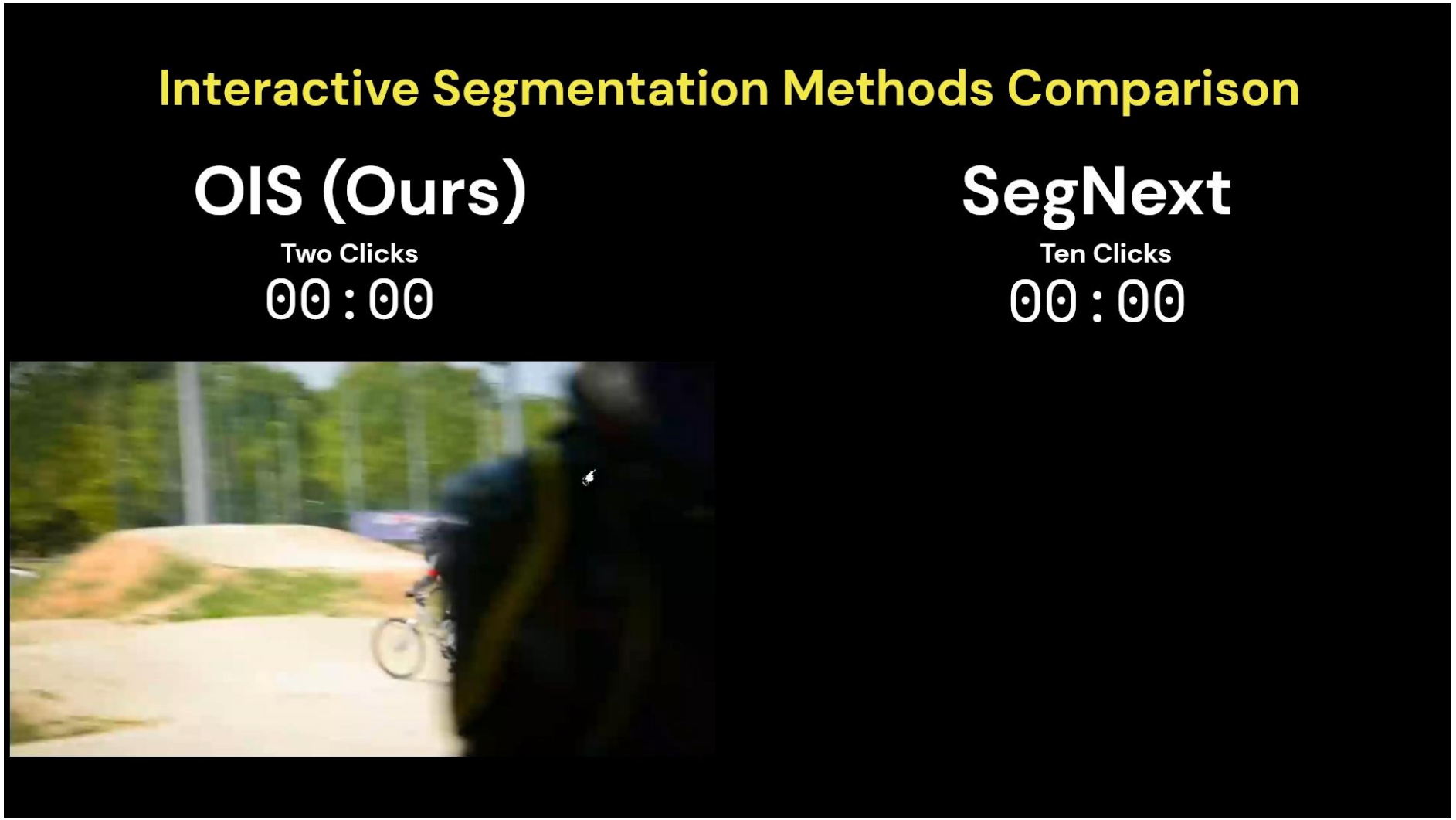
NORTHWESTERN UNIVERSITY

2025-03-31

Interactive Segmentation Demo



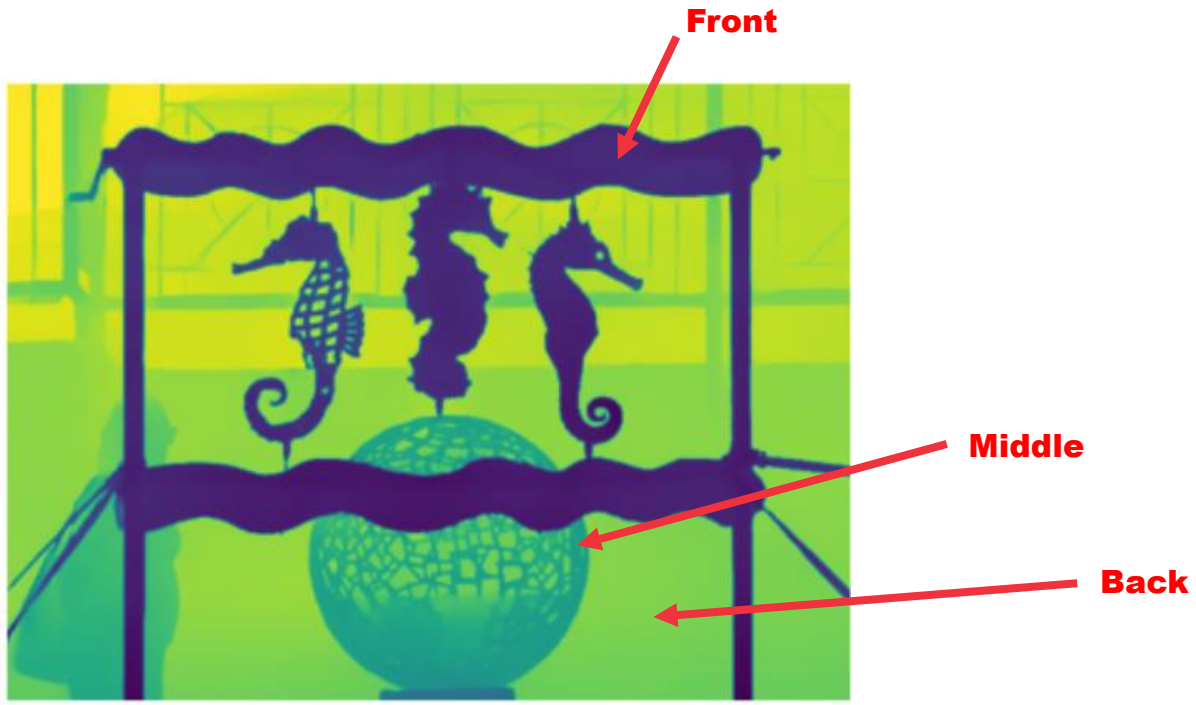
Interactive Segmentation Demo



Motivations

- Supplement 3D information into 2D interactive image segmentation`

Image



Depth

Can we combine prompt click to select order?

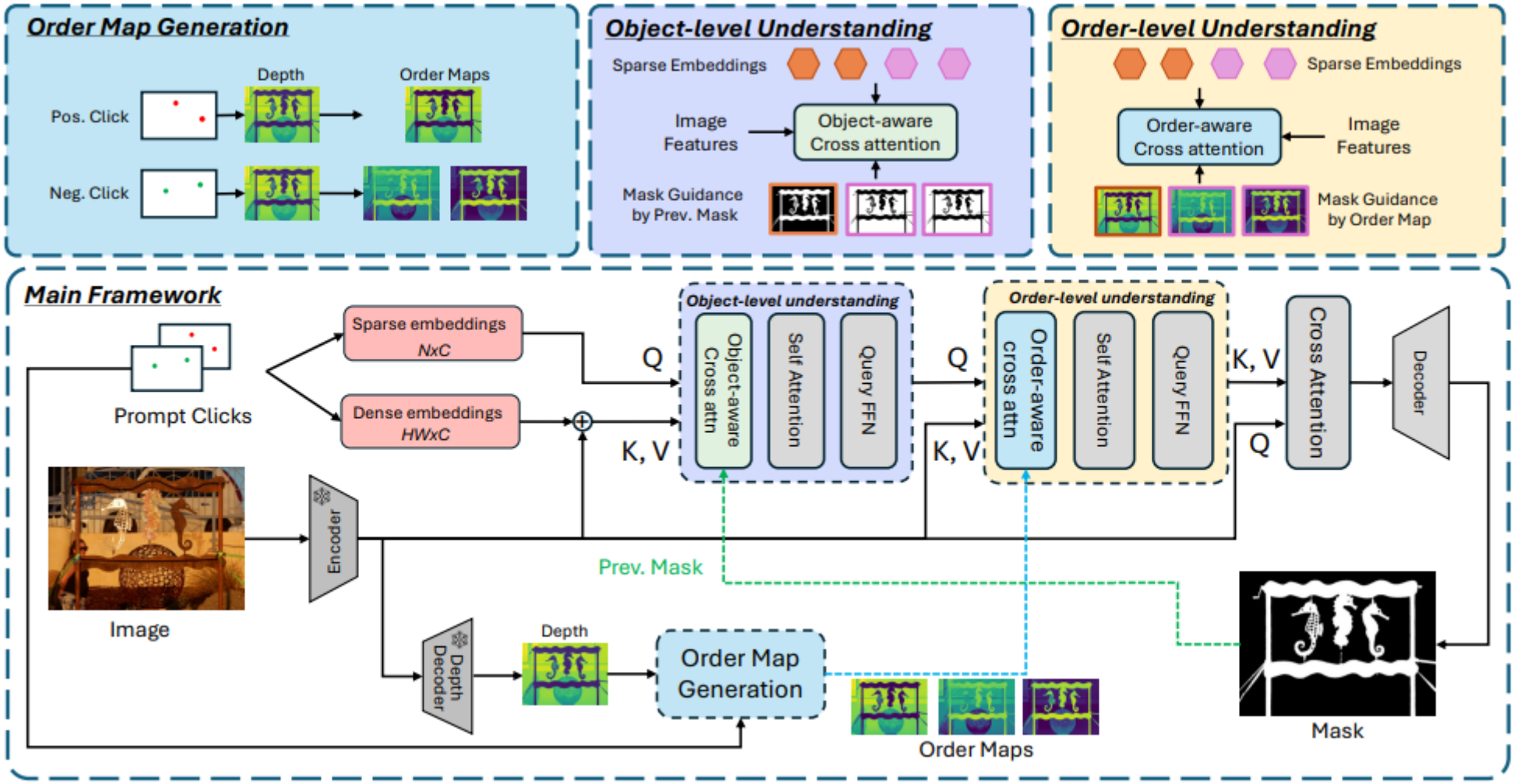
Depth contains strong order information of object in 3D space

Formulate 3D information into Order Map

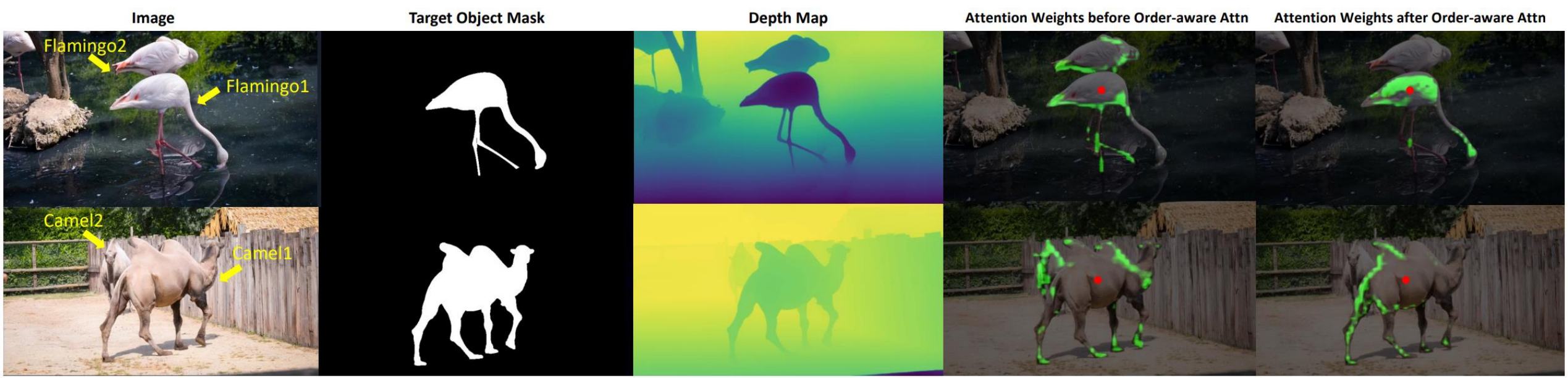
- We define the order map as the relative distances between each pixel in the image and the user-selected object or region

Image | Depth | Order Map for Foreground | Order Map for Background 1 | Order Map for Background 2

Our Order-aware Interactive Segmentation (OIS) framework



Visualization of attention weights before and after applying order-aware attention



Experiments

Datasets:

- public HQSeg44K dataset
- public DAVIS dataset

Compared Methods:

- HR-SAM, HR-SAM++, SegNext (CVPR2024), HQ-SAM (NeurIPS2023), SAM (ICCV2023), InterFormer (ICCV2023), SimpleClick (ICCV2023), FocalClick (CVPR2022), RITM

Evaluation Metrics

- Number of Clicks (NoC)

  NoC85, NoC90, and NoC95, represent the number of clicks required to achieve mIoU thresholds of 90% and 95%

- mean Interaction-over-Union (mIoU)

  1-mIoU and 5-mIoU, denote the average IoU achieved after 1, 5, or 10 consecutive clicks

- Number of Failure (NoF)

  the number of cases requiring more than 20 clicks to achieve 90% IoU

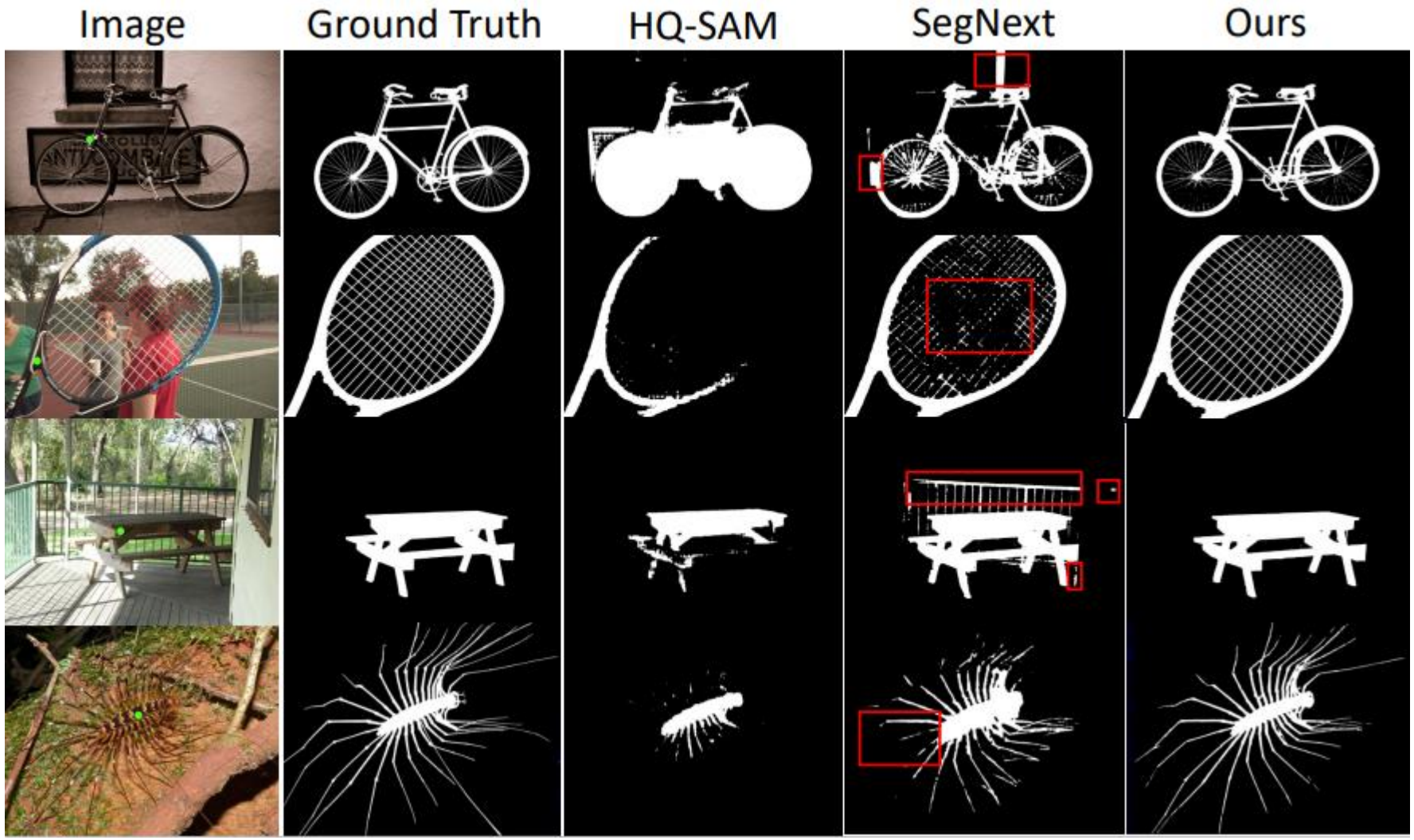
- Seconds Per Click (SPC)

  SPC measures the time efficiency, indicating how many seconds each click requires on average

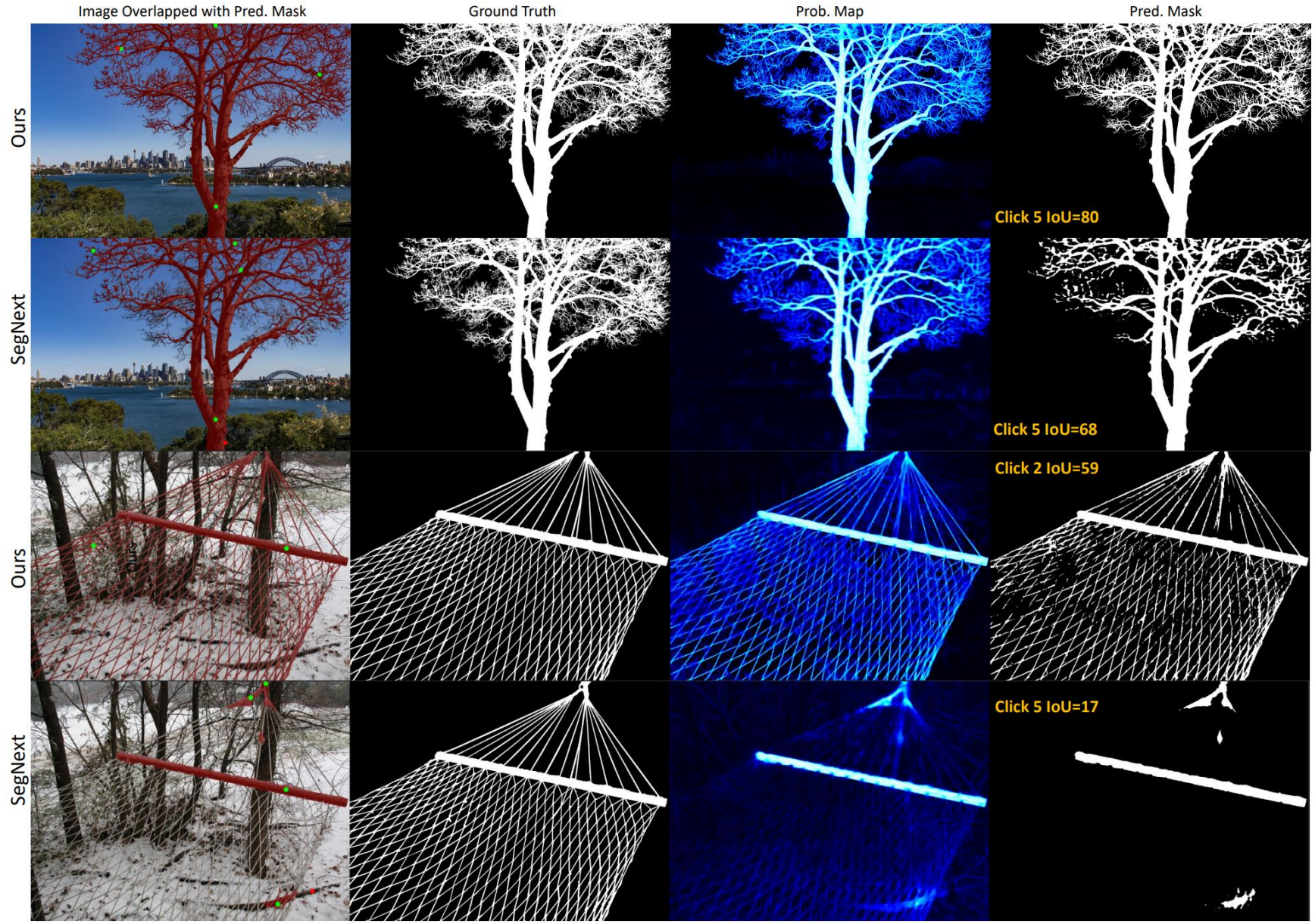
- SAT Latency

  total latency for the Segment Anything Task (prompt with grid of 16×16 points)

Results – Qualitative on HQSeg44K



Results – Qualitative on HQSeg44K



Results – Quantitative on HQSeg44K

Methods	Backbone	NoC90 ↓	NoC95 ↓	1-mIoU ↑	5-mIoU ↑	NoF95 ↓
RITM (Sofiiuk et al., 2022)	HRNet32_{400}	10.01	14.58	36.03	77.72	910
FocalClick (Chen et al., 2022)	SegF-B3-S2_{256}	8.12	12.63	62.89	84.63	835
FocalClick (Chen et al., 2022)	SegF-B3-S2_{384}	7.03	10.74	61.92	85.45	649
SimpleClick (Liu et al., 2023)	ViT-B_{448}	7.47	12.39	65.54	85.11	797
InterFormer (Huang et al., 2023)	ViT-B_{1024}	7.17	10.77	64.40	82.62	658
SAM (Kirillov et al., 2023)	ViT-B_{1024}	7.46	12.42	45.08	86.16	811
EifficientSAM (Xiong et al., 2024)	ViT-T_{1024}	10.11	14.60	-	77.90	-
EifficientSAM (Xiong et al., 2024)	ViT-S_{1024}	8.84	13.18	-	79.01	-
MobileSAM (Zhang et al., 2023a)	ViT-T_{1024}	8.70	13.83	53.20	81.98	951
HQ-SAM (Ke et al., 2024)	ViT-B_{1024}	6.49	10.79	42.38	89.85	671
HR-SAM (Huang et al., 2024)	ViT-B_{1024}	5.42	9.27	-	91.81	-
HR-SAM++ (Huang et al., 2024)	ViT-B_{1024}	5.32	9.18	-	91.84	-
SegNext (Liu et al., 2024a)	ViT-B_{1024}	5.32	9.42	81.79	91.75	583
Ours	ViT-B_{1024}	**3.95**	**7.50**	**89.40**	**93.78**	**485**
EifficientSAM (Xiong et al., 2024)	ViT-T_{2048}	9.47	13.13	-	74.20	-
EifficientSAM (Xiong et al., 2024)	ViT-S_{2048}	8.27	11.97	-	74.91	-
HR-SAM (Huang et al., 2024)	ViT-B_{2048}	4.37	7.86	-	93.34	-
HR-SAM++ (Huang et al., 2024)	ViT-B_{2048}	4.20	7.79	-	93.32	-
Ours	ViT-B_{2048}	**3.47**	**6.63**	**89.57**	**94.45**	**398**

Results – Qualitative on DAVIS

UNITED IMAGING Intelligence

Results – Quantitative on DAVIS

UNITED IMAGING Intelligence

Methods	Backbone	NoC90 ↓	NoC95 ↓	1-mIoU ↑	5-mIoU ↑	NoF95 ↓
RITM (Sofiiuk et al., 2022)	HRNet32_{400}	5.34	11.45	72.53	89.75	139
FocalClick (Chen et al., 2022)	SegF-B3-S2_{256}	5.17	11.42	76.28	90.82	155
FocalClick (Chen et al., 2022)	SegF-B3-S2_{384}	4.90	10.40	76.35	91.22	123
SimpleClick (Liu et al., 2023)	ViT-B_{448}	5.06	10.37	72.90	90.73	107
InterFormer (Huang et al., 2023)	ViT-B_{1024}	5.45	11.88	64.40	87.79	150
SAM (Kirillov et al., 2023)	ViT-B_{1024}	5.14	10.74	48.66	90.95	154
MobileSAM (Zhang et al., 2023a)	ViT-T_{1024}	5.83	12.74	61.69	89.18	196
EifficientSAM (Xiong et al., 2024)	ViT-T_{1024}	7.37	14.28	-	85.26	-
EifficientSAM (Xiong et al., 2024)	ViT-S_{1024}	6.37	12.26	-	87.55	-
HQ-SAM (Ke et al., 2024)	ViT-B_{1024}	5.26	10.00	45.75	91.77	136
HR-SAM (Huang et al., 2024)	ViT-B_{1024}	4.82	11.86	-	91.34	-
HR-SAM++ (Huang et al., 2024)	ViT-B_{1024}	5.02	11.64	-	91.25	-
SegNext (Liu et al., 2024a)	ViT-B_{1024}	4.43	10.73	85.97	91.87	123
Ours	ViT-B_{1024}	**3.80**	**8.59**	**87.29**	**92.76**	114
EifficientSAM (Xiong et al., 2024)	ViT-T_{2048}	8.00	14.37	-	84.10	-
EifficientSAM (Xiong et al., 2024)	ViT-S_{2048}	6.86	12.49	-	85.17	-
HR-SAM (Huang et al., 2024)	ViT-B_{2048}	4.22	8.83	-	92.63	-
HR-SAM++ (Huang et al., 2024)	ViT-B_{2048}	4.12	8.72	-	92.73	-
Ours	ViT-B_{2048}	**3.48**	**8.42**	**88.05**	**92.90**	**105**

Results – Efficiency Analysis

- Our model achieves the best balance, offering low latency with superior segmentation accuracy

Methods	Parameters (M)	SPC (ms) ↓	SAT Latency (s) ↓	NoC90 ↓	5-mIoU ↑
SimpleClick	96.46	55	81.3	7.47	85.11
HQ-SAM	94.81	**10**	**5.1**	6.49	89.85
SegNext	113.79	58	20.6	<u>5.32</u>	<u>91.75</u>
Ours	107.88	<u>31</u>	<u>9.2</u>	**3.95**	**93.78**

Results – Ablation Study

UNITED IMAGING Intelligence

DAVIS Dataset

Methods	NoC90 ↓	5-mIoU ↑	NoF95 ↓
Full	**3.80**	**92.76**	**114**
w/o order	4.84 (+1.04)	91.61 (-1.15)	171
w/o object	3.92 (+0.12)	92.52 (-0.24)	125
w/o sparse	5.18 (+1.38)	89.81 (-2.95)	167
w/o dense	4.63 (+0.83)	91.31 (-1.45)	188

HQSeg44K Dataset

Methods	NoC90 ↓	5-mIoU ↑
Full	3.95	93.78
w/o order	4.87 (+0.98)	92.49 (-1.29)
w/o object	4.23 (+0.28)	93.28 (-0.5)
w/o sparse	5.23 (+1.28)	90.80 (-2.98)
w/o dense	4.97 (+1.02)	91.75 (-2.03)

Conclusion

- Introduced Order-Aware Interactive Segmentation (OIS), which incorporates 3D spatial context through the concept of order into 2D interactive segmentation.
- Proposed order-aware attention enables the model to better distinguish objects based on their relative depths.
- Introduced object-aware attention to enhance our model’ s ability to differentiate objects within the same depth level.
- Integrated user clicks using both sparse and dense representations, improving segmentation accuracy and computational efficiency.
- Experimental results validated that OIS significantly improves segmentation accuracy and speed as compared to prior methods.

UNITED IMAGING
Intelligence

Leading

CHANGE

引　领　改　变