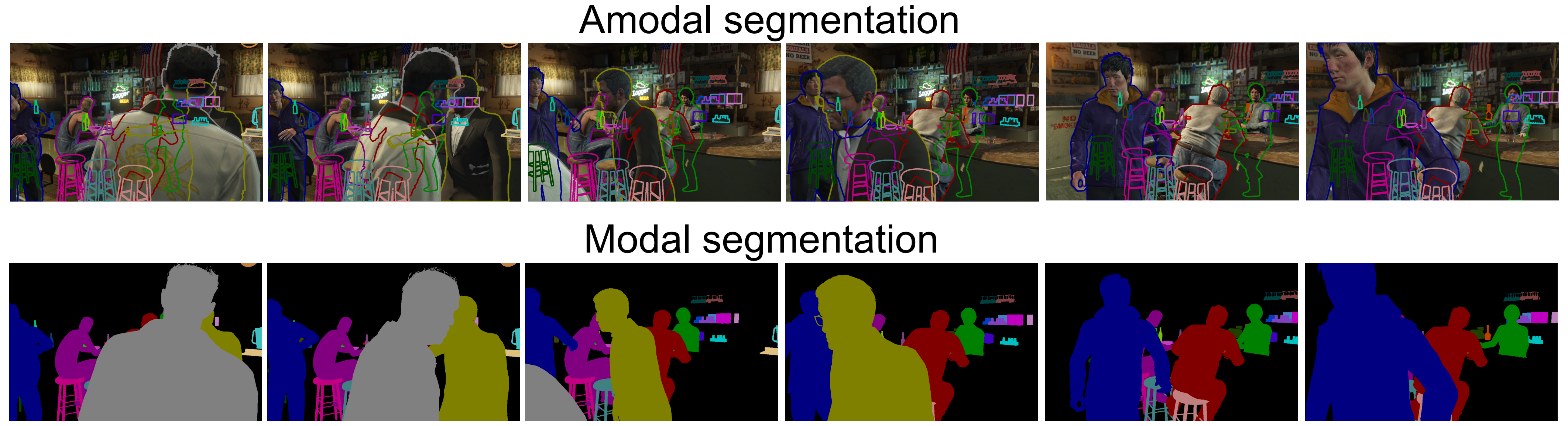
SAIL-VOS: Semantic Amodal Instance Level Video Object Segmentation – A Synthetic Dataset and Baselines

Yuan-Ting Hu[1] Hong-Shuo Chen[1] Kexin Hui[1] Jia-Bin Huang[2] Alexander G. Schwing[1]

[1] University of Illinois Urbana-Champaign [2] Virginia Tech

1. Introduction

Goal: Amodal Instance Level *Video* Segmentation – predicting and forecasting the object extend beyond the visible

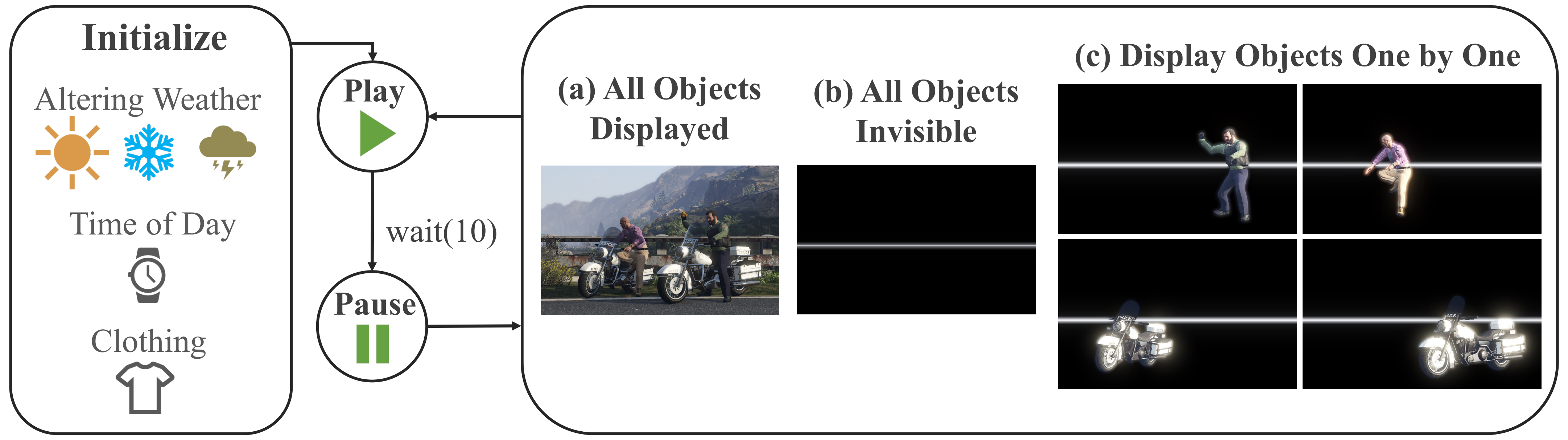


Issue: Only image datasets for amodal segmentation available

Contribution: First *video* dataset & methods that use temporal context

2. Dataset Collection Methodology

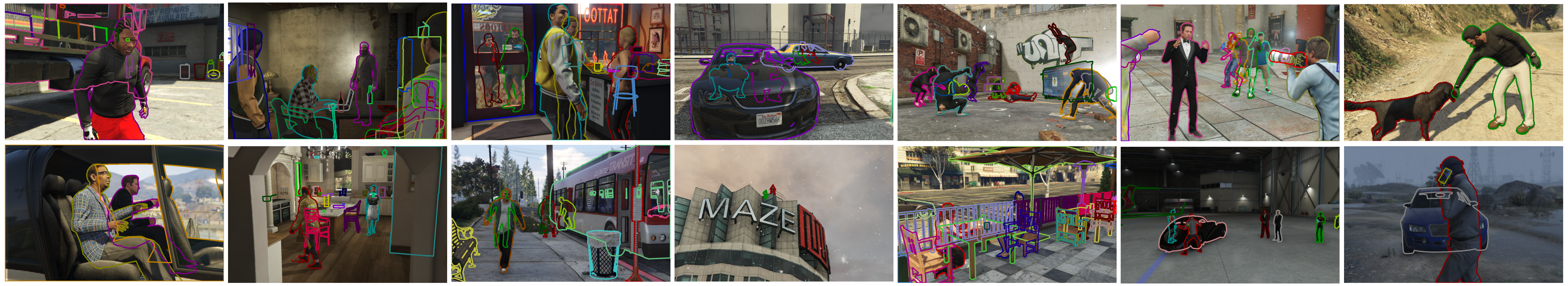
Grand Theft Auto V (GTA-V) is used to automate dataset collection


Initialize
Altering Weather
Time of Day
Clothing
Play
wait(10)
Pause
(a) All Objects Displayed
(b) All Objects Invisible
(c) Display Objects One by One

- We record the RGB image and the corresponding depth and stencil buffer
- **Modal and amodal masks:** computed using depth and stencil buffer
- **Object tracking:** achieved by accessing the rendering resources via the ScriptHookV library
- **Semantic class label:** obtained by grouping the name associated with the 3D model file of each object
- **Other data:** depth ordering, human 2d and 3d pose

3. SAIL-VOS Dataset

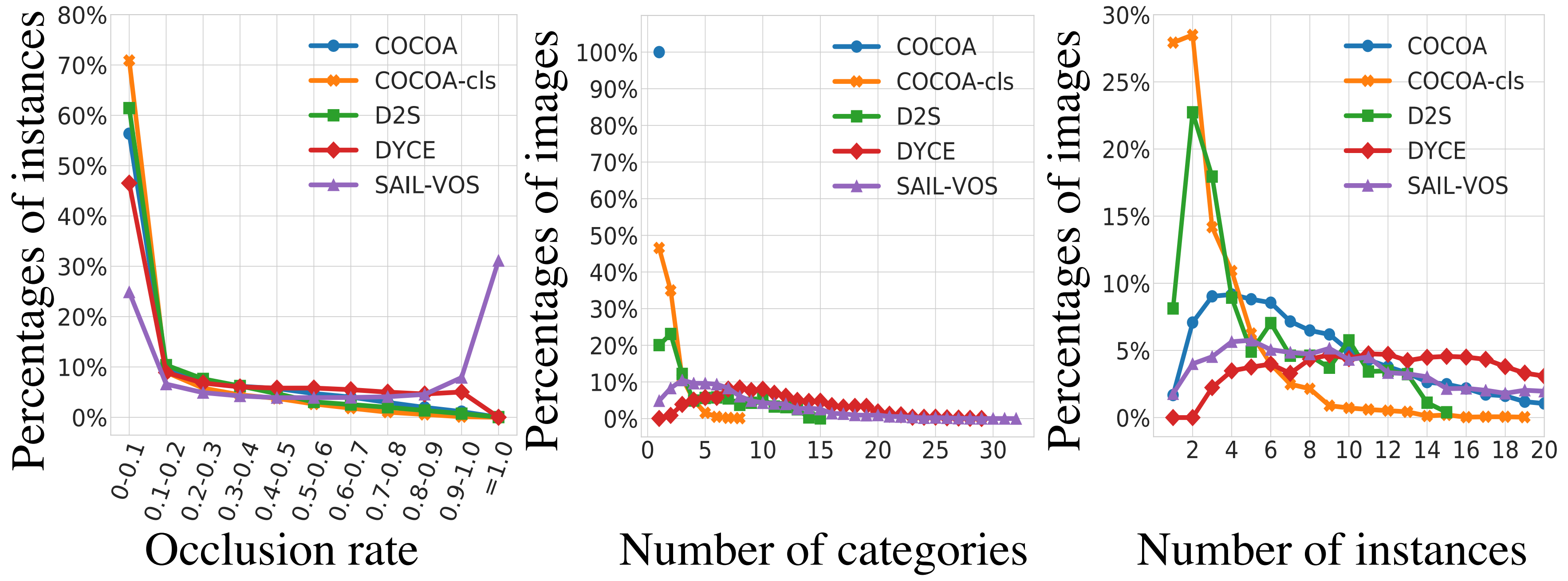
- Contains diverse scenes (outdoor/indoor), different weather (sunny/rainy/storm), different lighting conditions (day/night)
- Provides annotations for modal segmentation, amodal segmentation, depth ordering and 2d/3d human pose



4. Statistics

Comparisons with other datasets:

Dataset	COCOA	COCOA-cls	D2S	DYCE	Ours
Image/Video	Image	Image	Image	Image	Video
Resolution	275K pix	275K pix	3M pix	1M pix	1M pix
	-	-	1440×1920	1000×1000	800×1280
Synthetic/Real	Real	Real	Real	Synthetic	Synthetic
# of images	5,073	3499	5,600	5,500	111,654
# of classes	-	80	60	79	162
# of instances	46,314	10,562	28,720	85,975	1,896,295
# of occluded instances	28,106	5,175	16,337	70,766	1,653,980
Avg. occlusion rate	18.8%	10.7%	15.0%	27.7%	56.3%

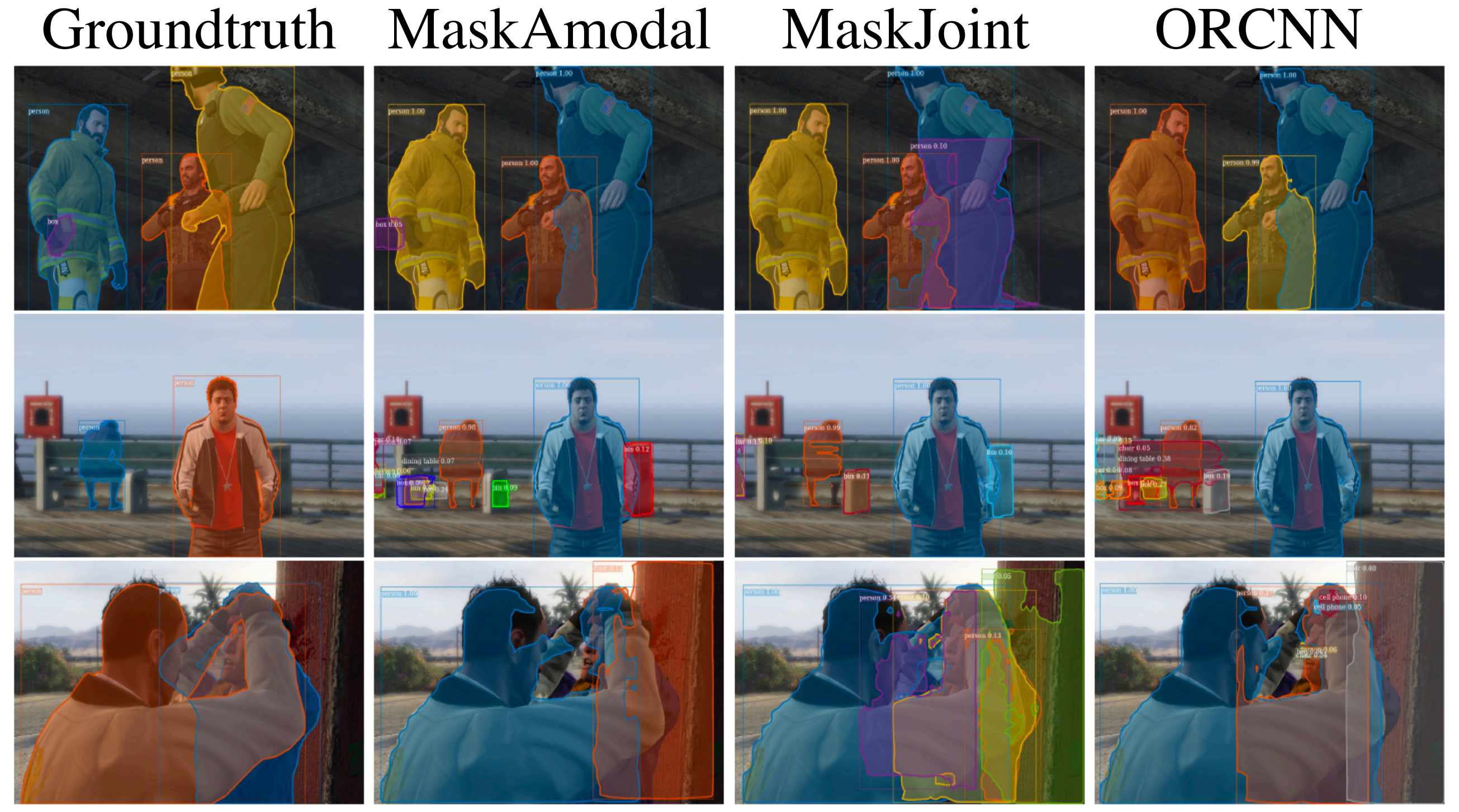


5. Baselines and Results

Evaluation on the SAIL-VOS dataset in the **class agnostic** setting:

	Modal mask							Amodal mask						
	AP_{50}	AP	AP^P_{50}	AP^H_{50}	AP^L_{50}	AP^M_{50}	AP^S_{50}	AP_{50}	AP	AP^P_{50}	AP^H_{50}	AP^L_{50}	AP^M_{50}	AP^S_{50}
MaskRCNN	**40.6**	**28.0**	**51.2**	**13.5**	**74.6**	**20.2**	**5.6**	-	-	-	-	-	-	-
MaskAmodal	-	-	-	-	-	-	-	40.4	**26.6**	**51.2**	14.8	72.9	**20.6**	6.8
MaskJoint	38.8	26.0	49.5	11.9	70.4	17.4	6.4	**40.8**	26.4	**51.2**	**15.8**	**73.1**	19.6	7.5
ORCNN	37.3	24.3	49.0	9.8	68.2	16.5	6.3	40.1	25.5	**51.2**	14.2	71.9	19.5	**7.6**

Qualitative Results:



Video Object Segmentation:

DAVIS results with and without pretraining on SAIL-VOS:

IoU on the DAVIS validation set.

DAVIS fraction	0%	10%	20%	30%	50%	100%
VideoMatch-pretrain	**0.74**	**0.77**	**0.78**	**0.78**	**0.78**	0.79
VideoMatch	0.55	0.66	0.73	0.74	**0.78**	**0.81**

More details and results on: http://sailvos.web.illinois.edu