THE HONG KONG POLYTECHNIC UNIVERSITY
香港理工大學

vLAR
Visual Learning And Reasoning

NEURAL INFORMATION PROCESSING SYSTEMS

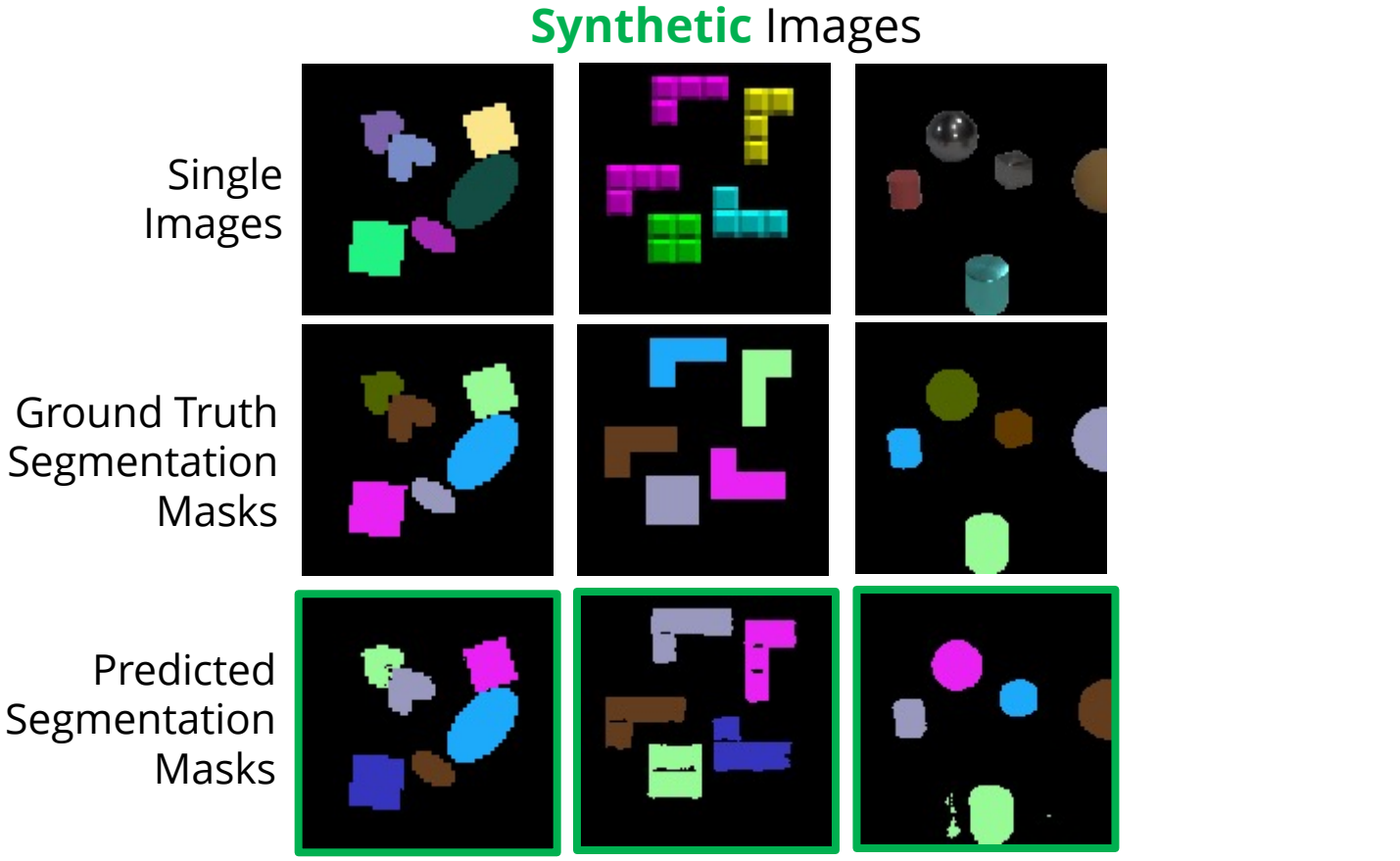
Promising or Elusive? Unsupervised Object Segmentation from Real-world Single Images

Yafei YANG , Bo YANG
vLAR Group, Department of Computing
The Hong Kong Polytechnic University

NeurIPS 2022

Unsupervised object segmentation from single images

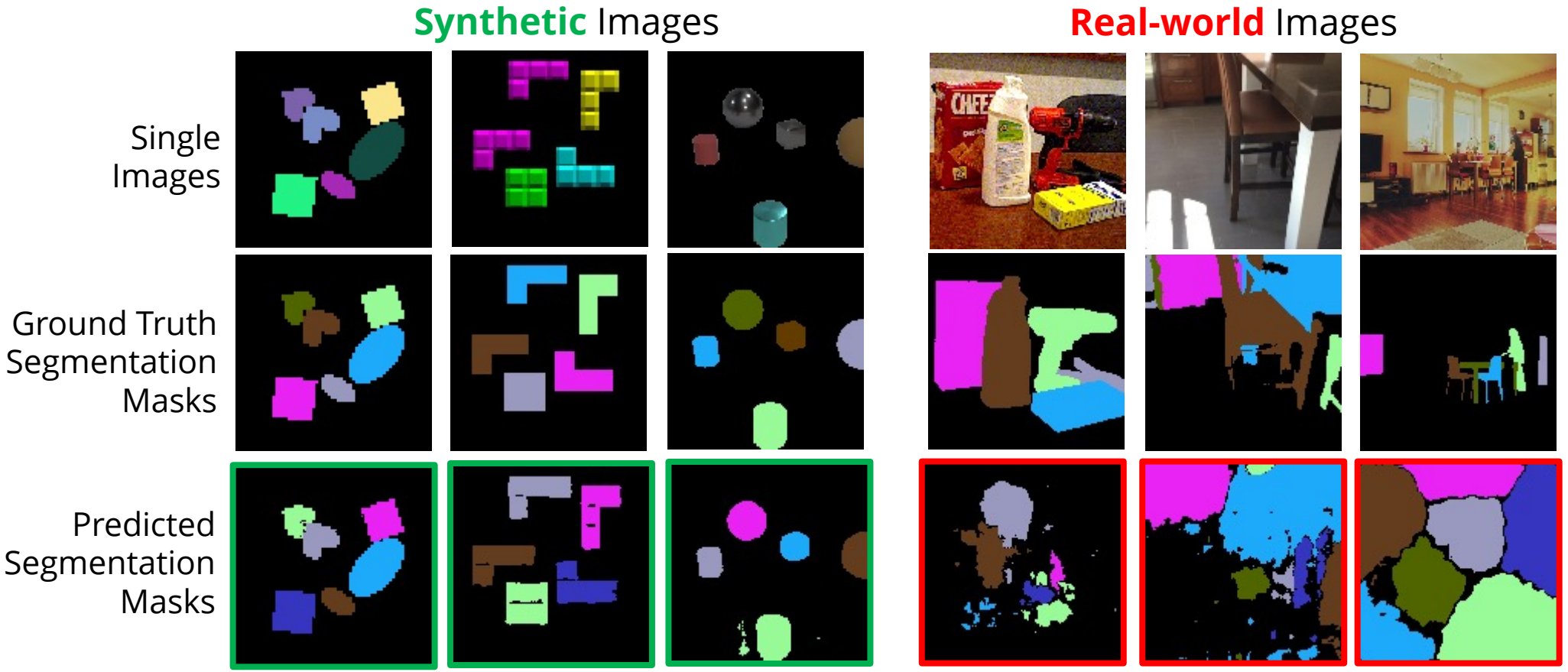
Synthetic Images



Experiment results from SlotAtt*

*Locatello, Francesco, et al. "Object-centric learning with slot attention." *Advances in Neural Information Processing Systems* 33 (2020): 11525-11538.

Unsupervised object segmentation from single images



Experiment results from SlotAtt*

*Locatello, Francesco, et al. "Object-centric learning with slot attention." *Advances in Neural Information Processing Systems* 33 (2020): 11525-11538.

Is it possible to segment generic objects from real-world single images?

What to expect

Is it promising or even possible to segment generic objects from real-world single images using (existing) unsupervised methods?

- **4** complexity factors
- **6** benchmark datasets
- **4+1** representative approaches
- **15** types of ablation settings
- **210** experiments

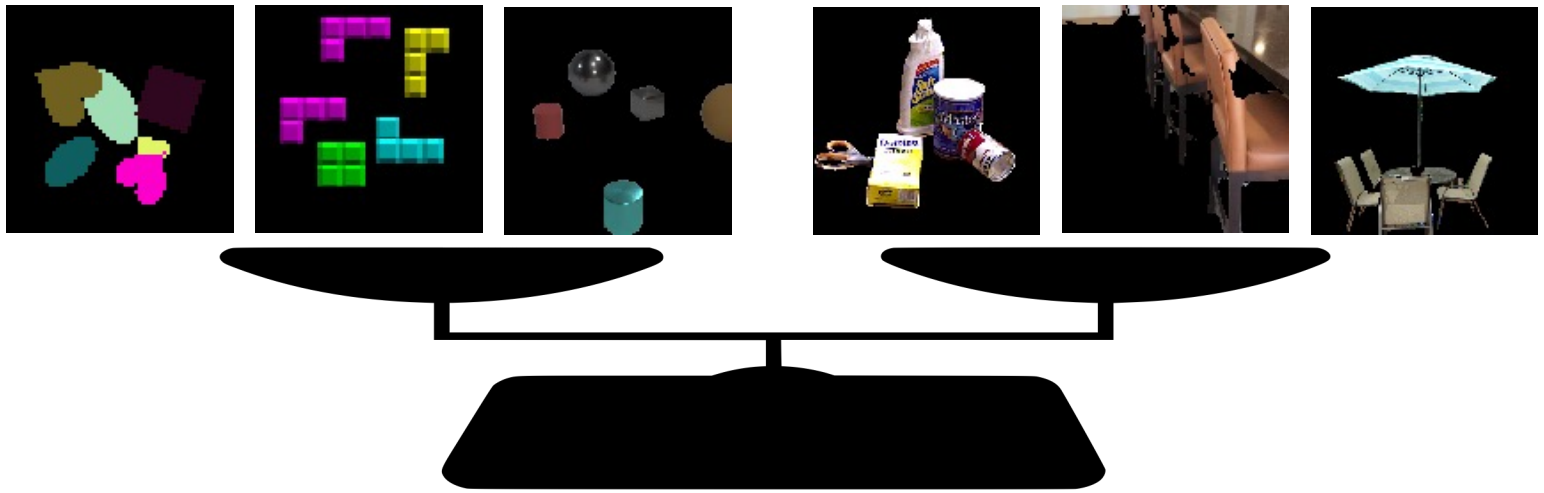
Complexity Factors

- Object Color Gradient
- Object Shape Concavity
- Inter-object Color Similarity
- Inter-object Shape Variation

What is an object?

Synthetic Images

Real-world Images



How to quantify the objectness biases in datasets?

Complexity Factors

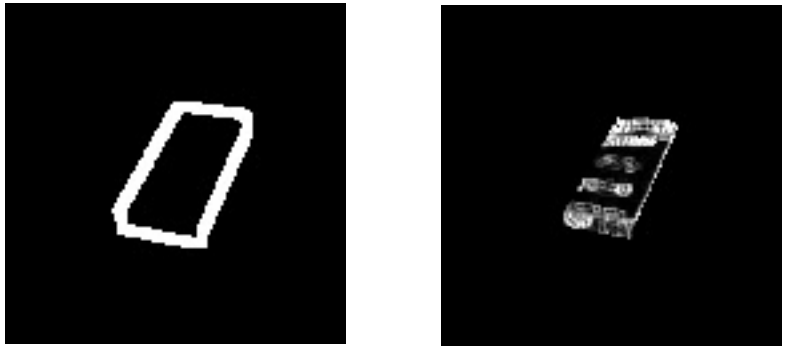
	appearance	**geometry**
object-level	**Object Color** Gradient	**Object Shape** Concavity
scene-level	**Inter-object Color** Gradient	**Inter-object Shape** Variation

Complexity Factor - Object Color Gradient

object-level; appearance

Illustration

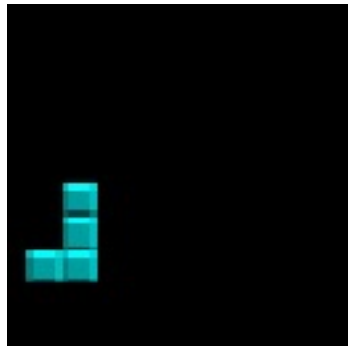
object image | grayscale image | gradient



object boundary | inner gradient

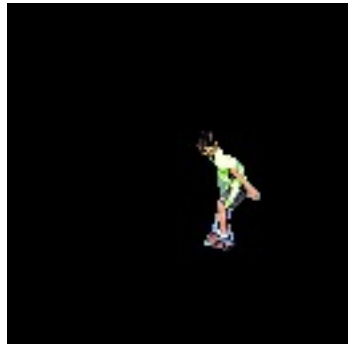
Example images and factor values

0.000



0.165

0.518



0.802

Complexity Factor - Object Shape Concavity

object-level; geometry

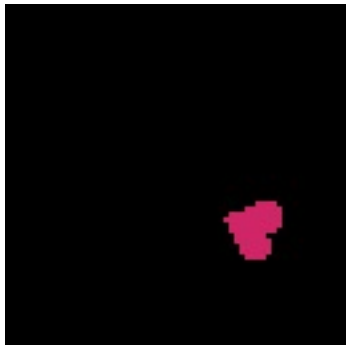
Illustration

object mask

smallest convex polygon mask

Example images and factor values

0.020



0.121

0.526

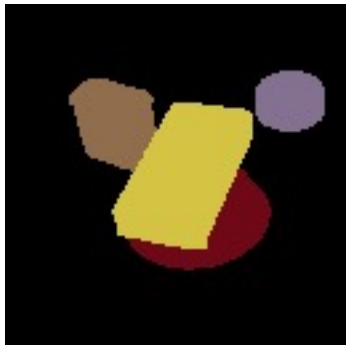
0.750

Complexity Factor – Inter-object Color Similarity

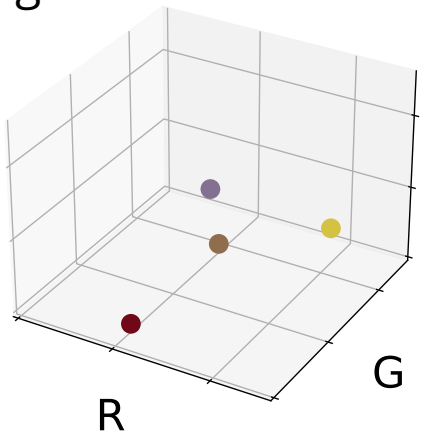
scene-level; appearance

Illustration

original

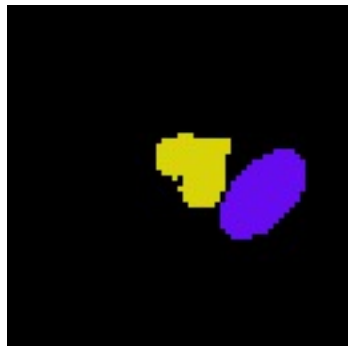


average color

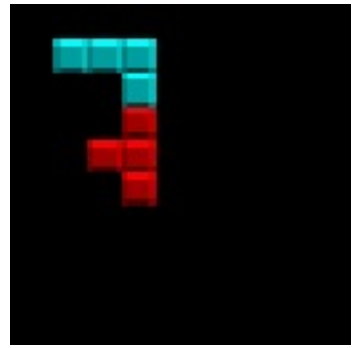


average colors in RGB space

Example images and factor values



0.265



0.359

0.787

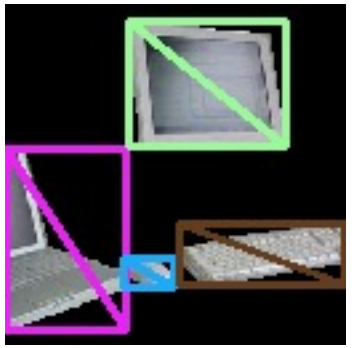
0.936

Complexity Factor – Inter-object Shape Variation

scene-level; geometry

Illustration

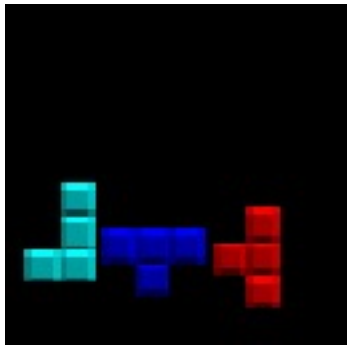
bounding boxes



diagonals

diagonal variation

Example images and factor values



0.005

0.105

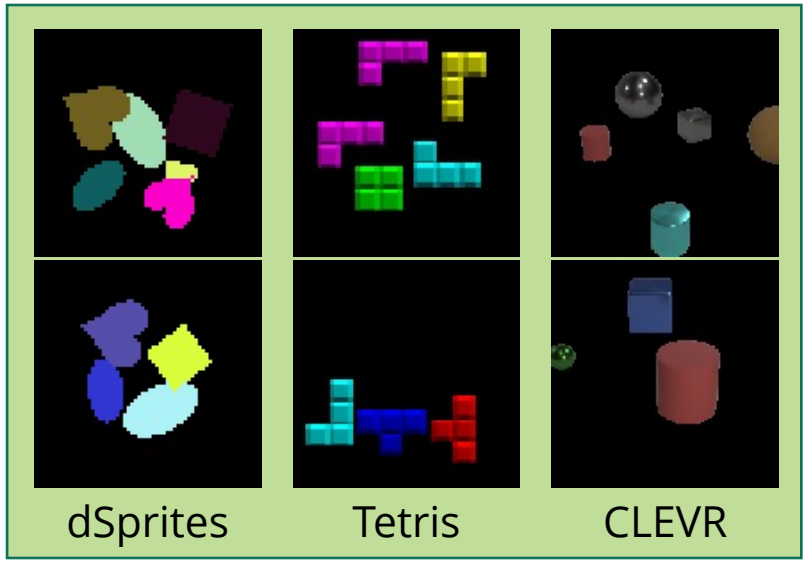
0.257

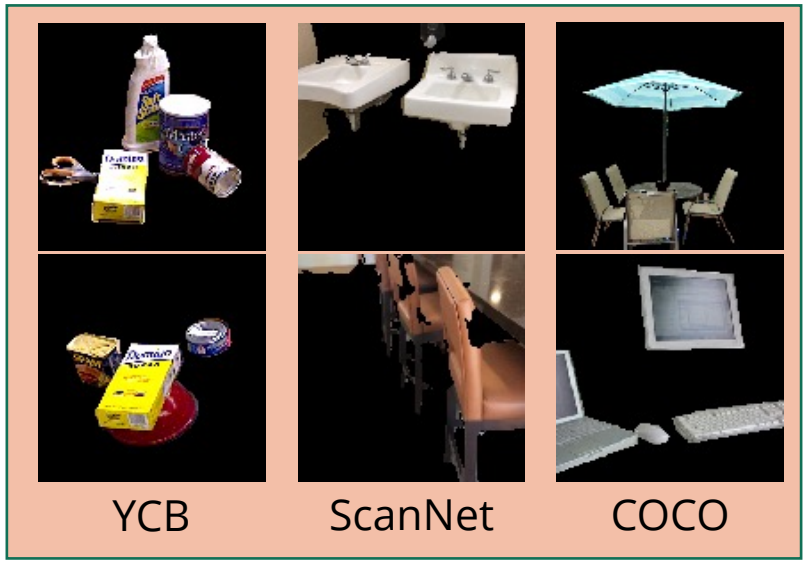
0.565

6 Benchmark Datasets

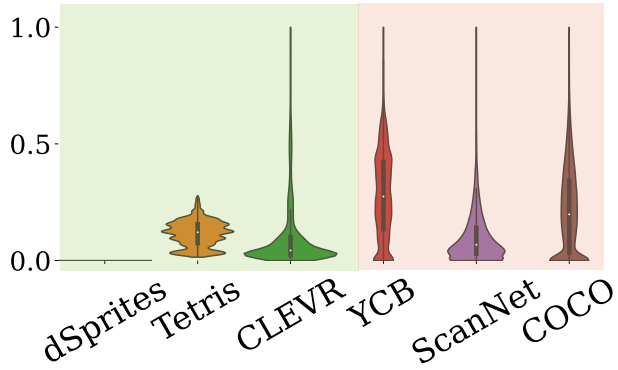
- dSprites
- Tetris
- CLEVR
- YCB
- ScanNet
- COCO

Biases in 6 Datasets - quantitative summary

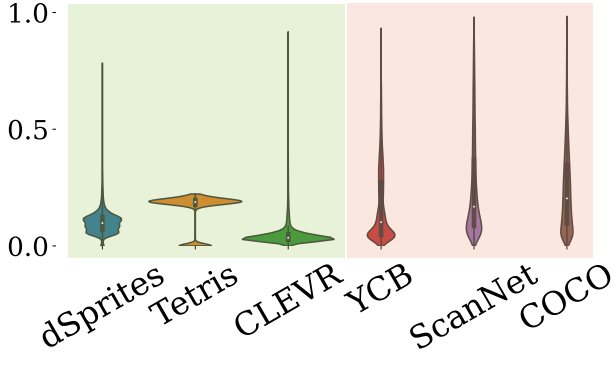




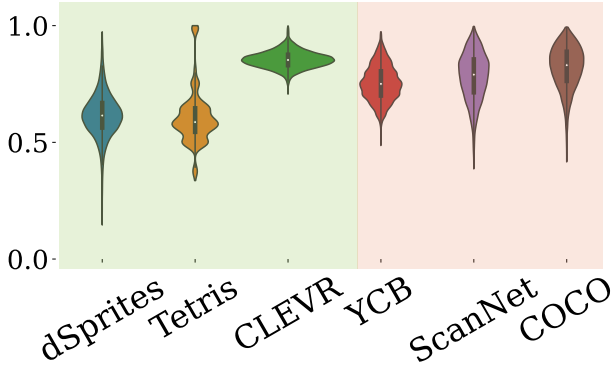
Object Color Gradient



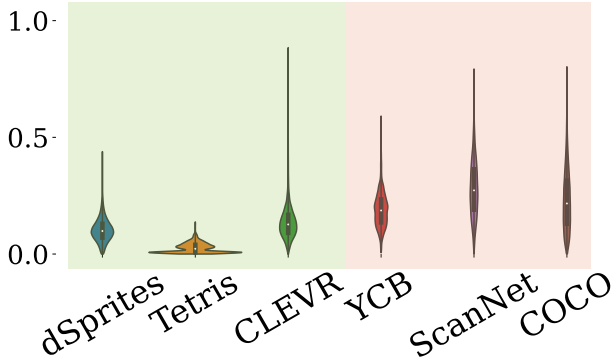
Object shape Concavity



Inter-object Color Similarity



Inter-object Shape Variation

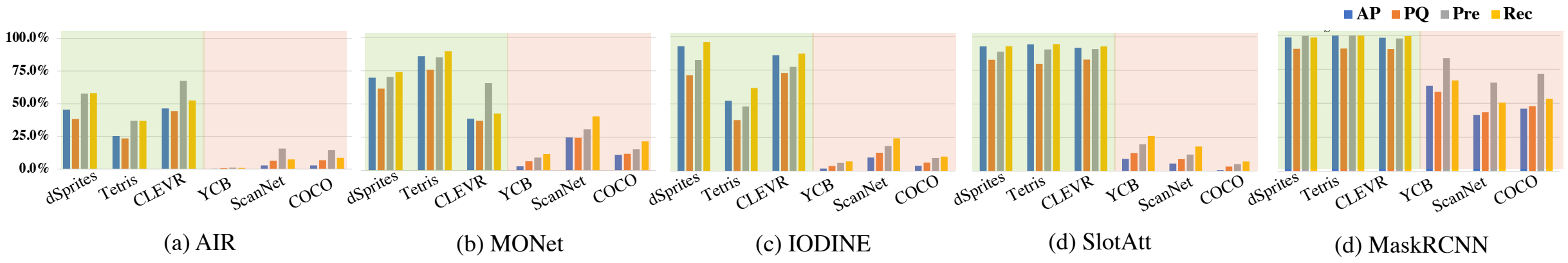


4+1 Representative Methods

- AIR
- MONet
- IODINE
- Slot Attention
- Mask-RCNN*

5 Methods on 6 Datasets

Quantitative Evaluation



Unsupervised

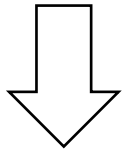
Supervised

15 ablation settings

- C: single color
- S: convex shape
- T: texture replaced
- U: uniform scale
- combinations of above four…

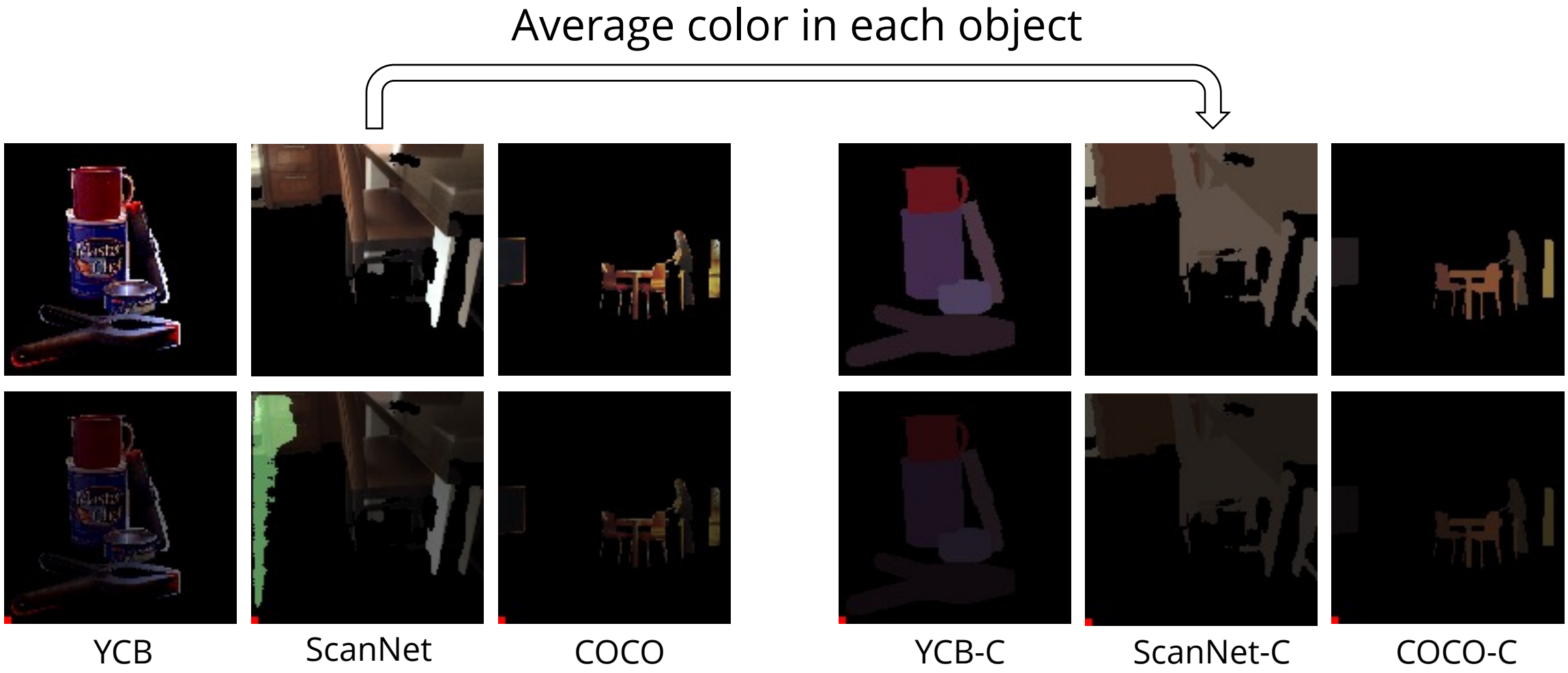
From complexity factors to ablation settings

Complexity Factor		
	appearance	**geometry**
object-level	**Object Color** Gradient	**Object Shape** Concavity
scene-level	**Inter-object Color** Gradient	**Inter-object Shape** Variation



Ablation Setting		
	appearance	**geometry**
object-level	C: single **C**olor	S: convex **S**hape
scene-level	T: **T**exture replaced	U: **U**niform scale

C - single **C**olor ablation



S - convex **S**hape ablation

Change shape each object to be convex

YCB | ScanNet | COCO | YCB-S | ScanNet-S | COCO-S

T – **T**exture replaced ablation

Change objects appearance with distinctive texture

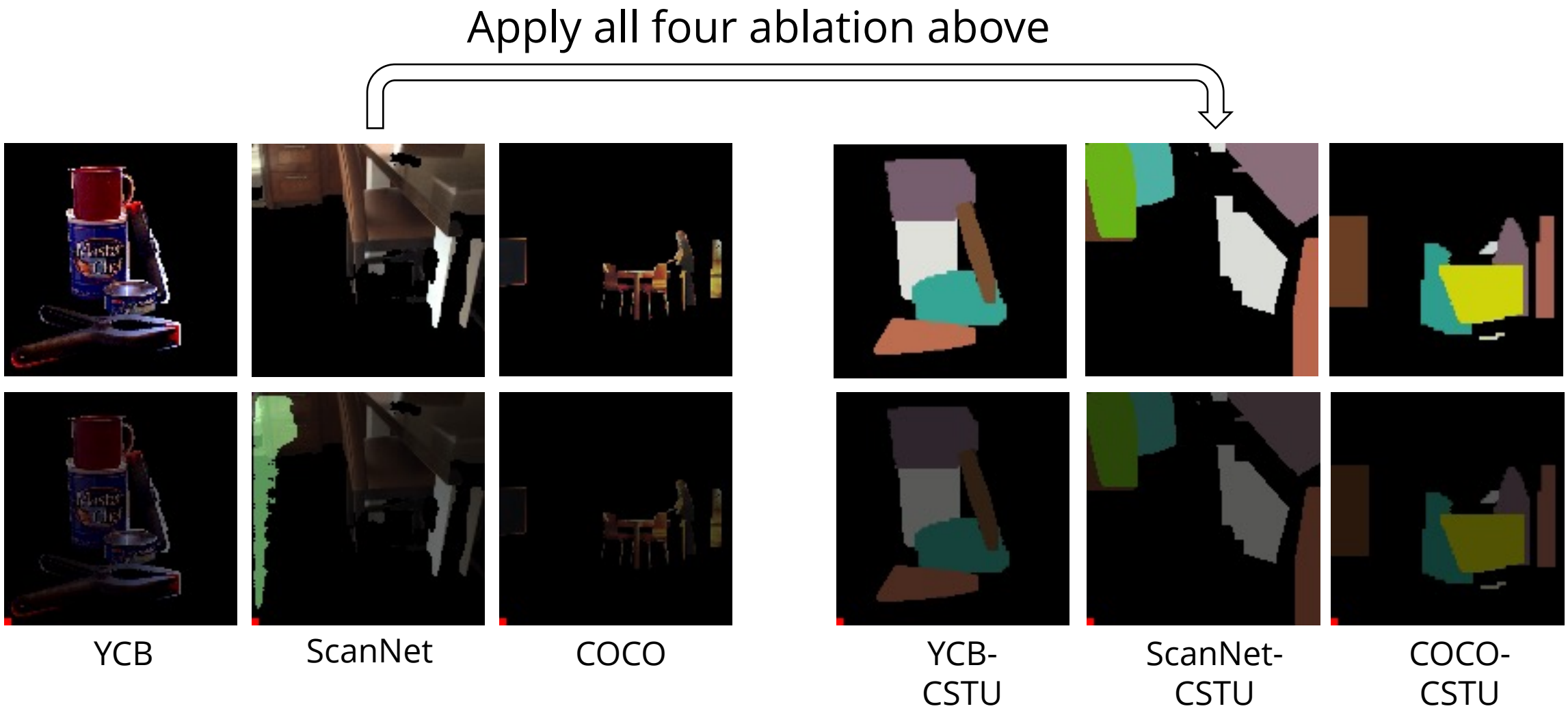
YCB ScanNet COCO YCB-T ScanNet-T COCO-T

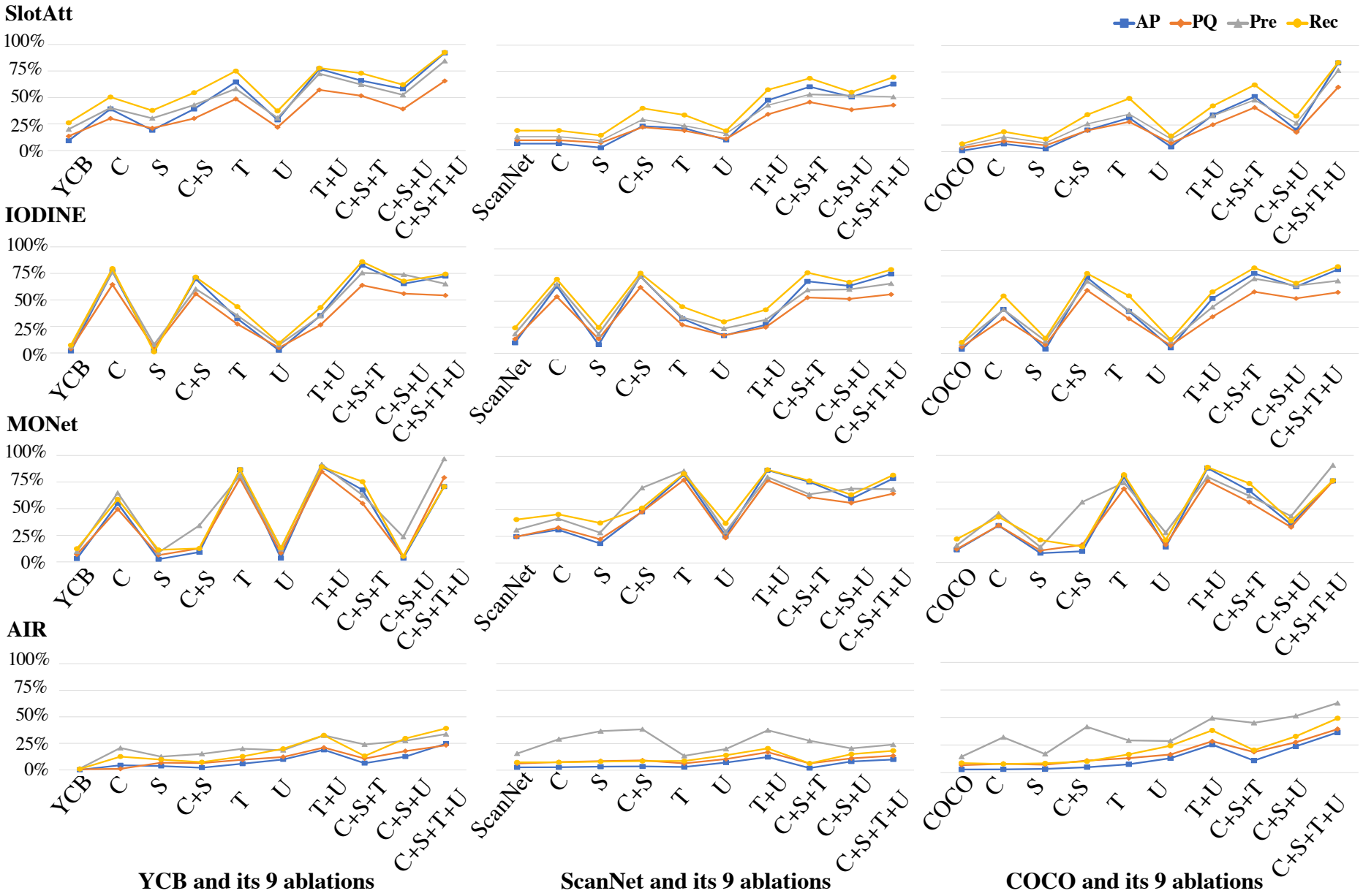
U – **U**niformed scale ablation

Change objects scale to be uniform

YCB ScanNet COCO YCB-U ScanNet-U COCO-U

CSTU – fully ablated





Quantitative results of ablation experiments

Why do unsupervised models fail on real-world datasets?

	object-level		scene-level	
	Object Color Gradient	Object Shape Concavity	Inter-object Color Similarity	Inter-object Shape Variation
AIR				
MONet				
IODINE				
Slot Attention				

Finding 1
Different models favor different objectness bias;

Finding 2
None of the model can fully capture the true objectness biases in real-world images.

Can unsupervised methods segment objects from single images?

Success on **synthetic** datasets vs. Failure on **real-world** datasets

4 + 1 methods

unsupervised segmentation methods

6 dataset

synthetic datasets

real-world datasets

15 ablation settings

- effectiveness of complexity factors;
- objectness biases in different models.

4 complexity factors

- Quantitatively evaluate the objectness biases in datasets

Future direction

- more discriminative objectness biases (e.g. motions);
- learn from single-object-dominant dataset.

Thanks

Project page: https://vlar-group.github.io/UnsupObjSeg.html
GitHub: https://github.com/vLAR-group/UnsupObjSeg
Arxiv: https://arxiv.org/abs/2210.02324