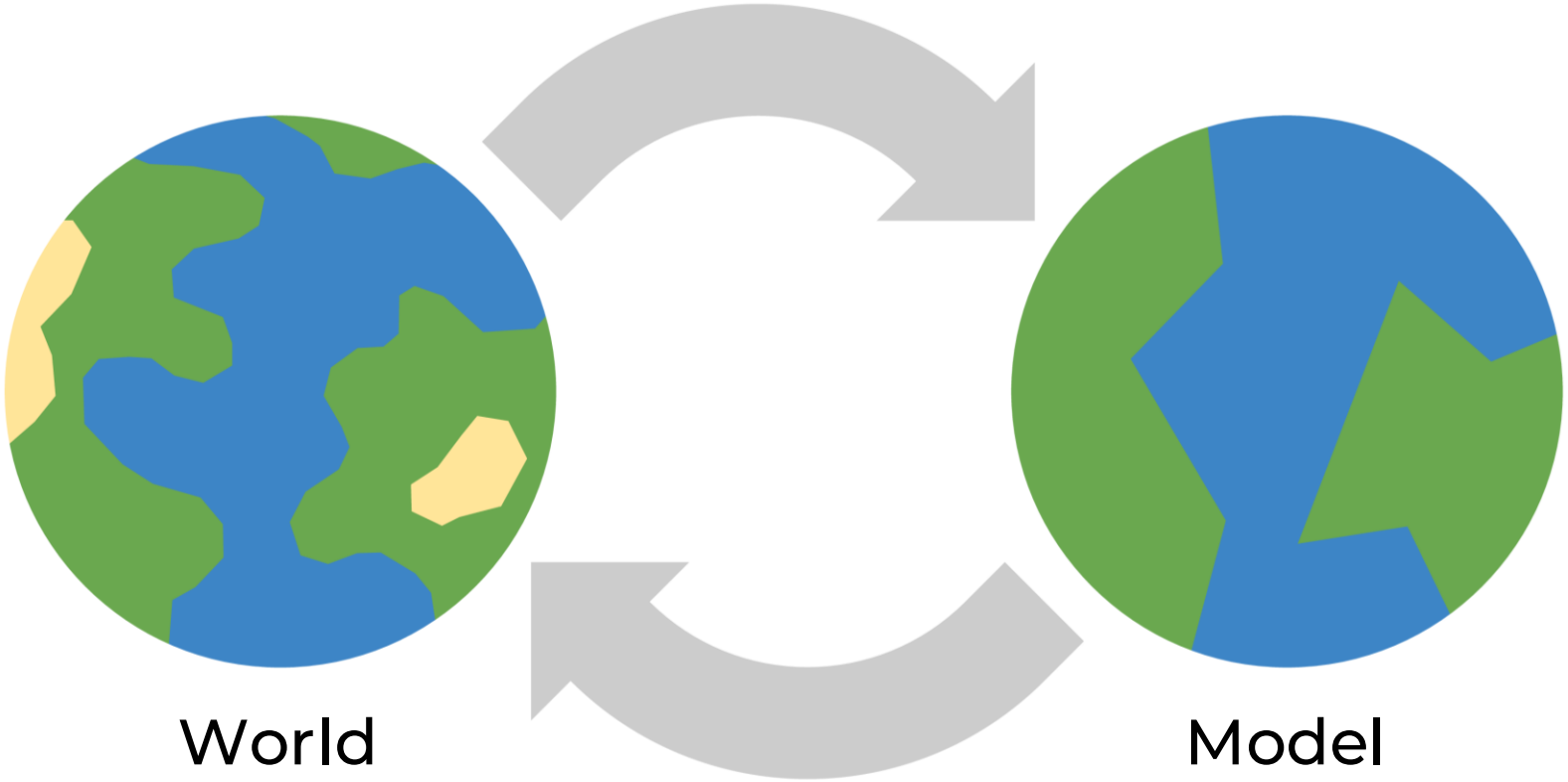
Revisiting the 3Rs in Computer Vision

Sangwoo Mo

Assistant Professor
POSTECH IME

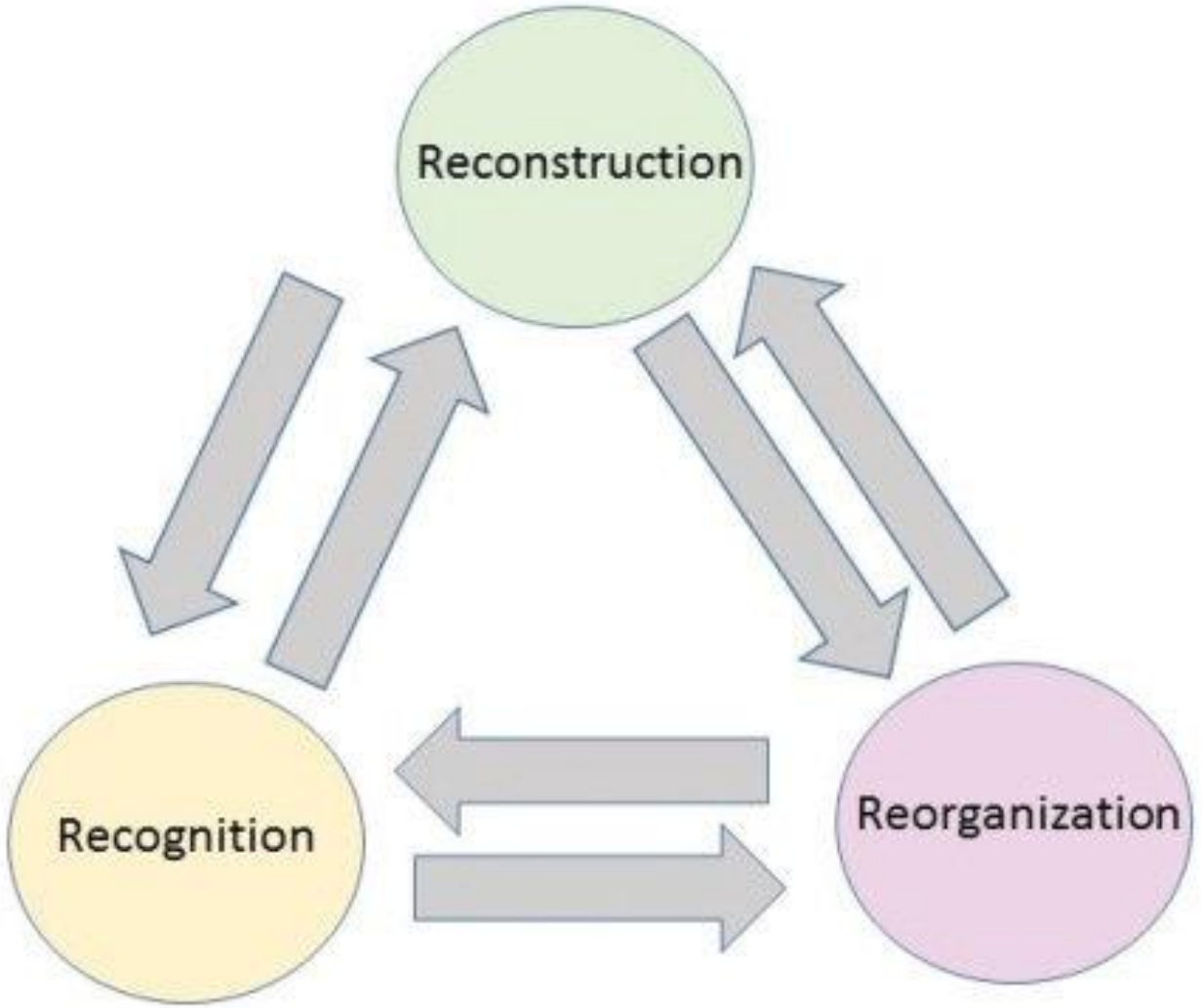
All Computer Vision Problems are Connected

All Computer Vision Problems are Connected



3Rs in Computer Vision





3Rs in Computer Vision

Is this a cat?

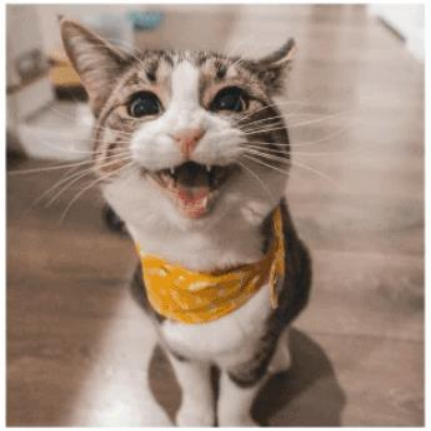
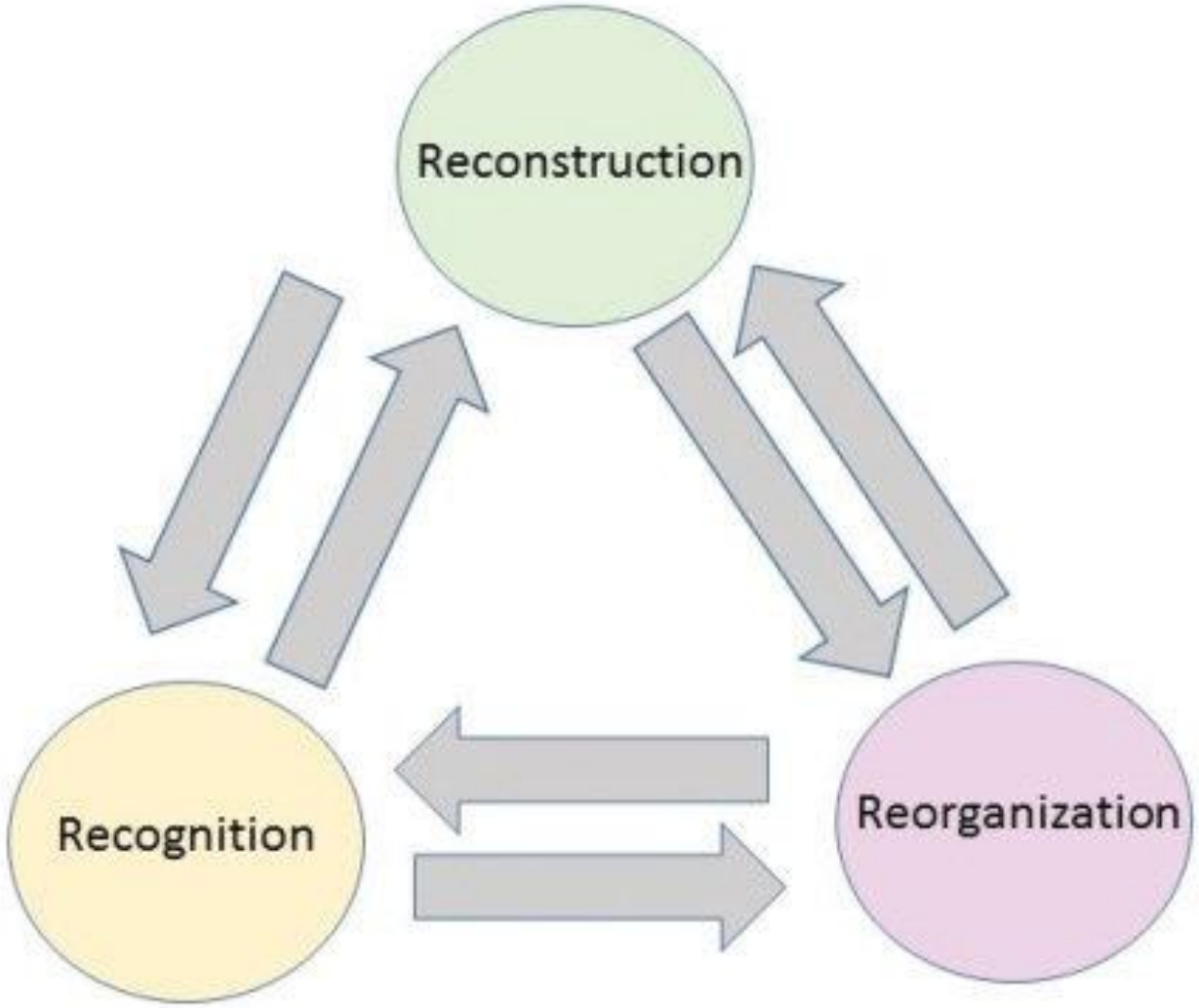
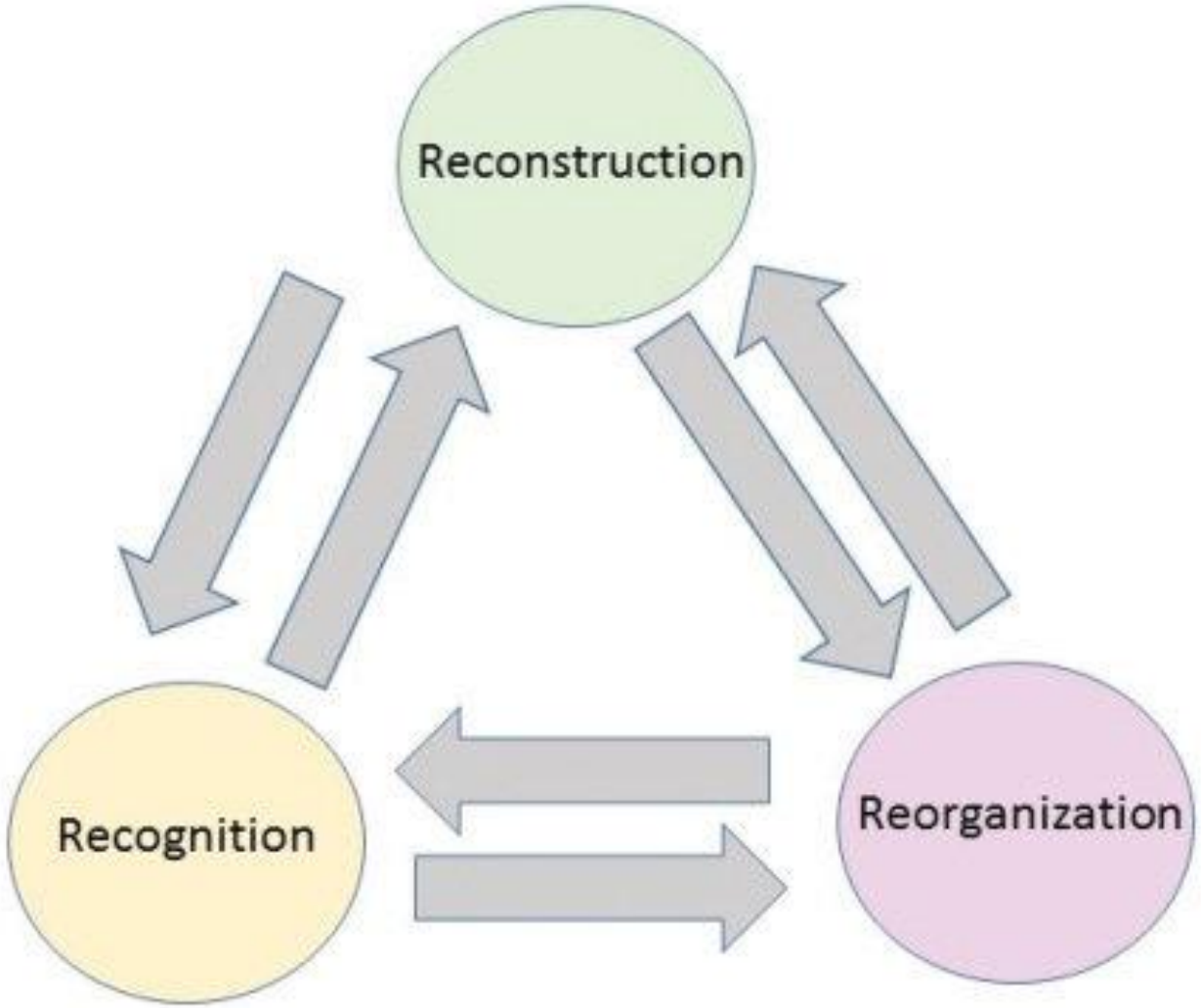


Image Classification



3Rs in Computer Vision



Is this a cat?

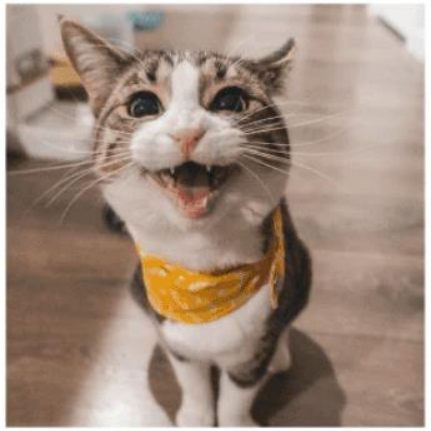


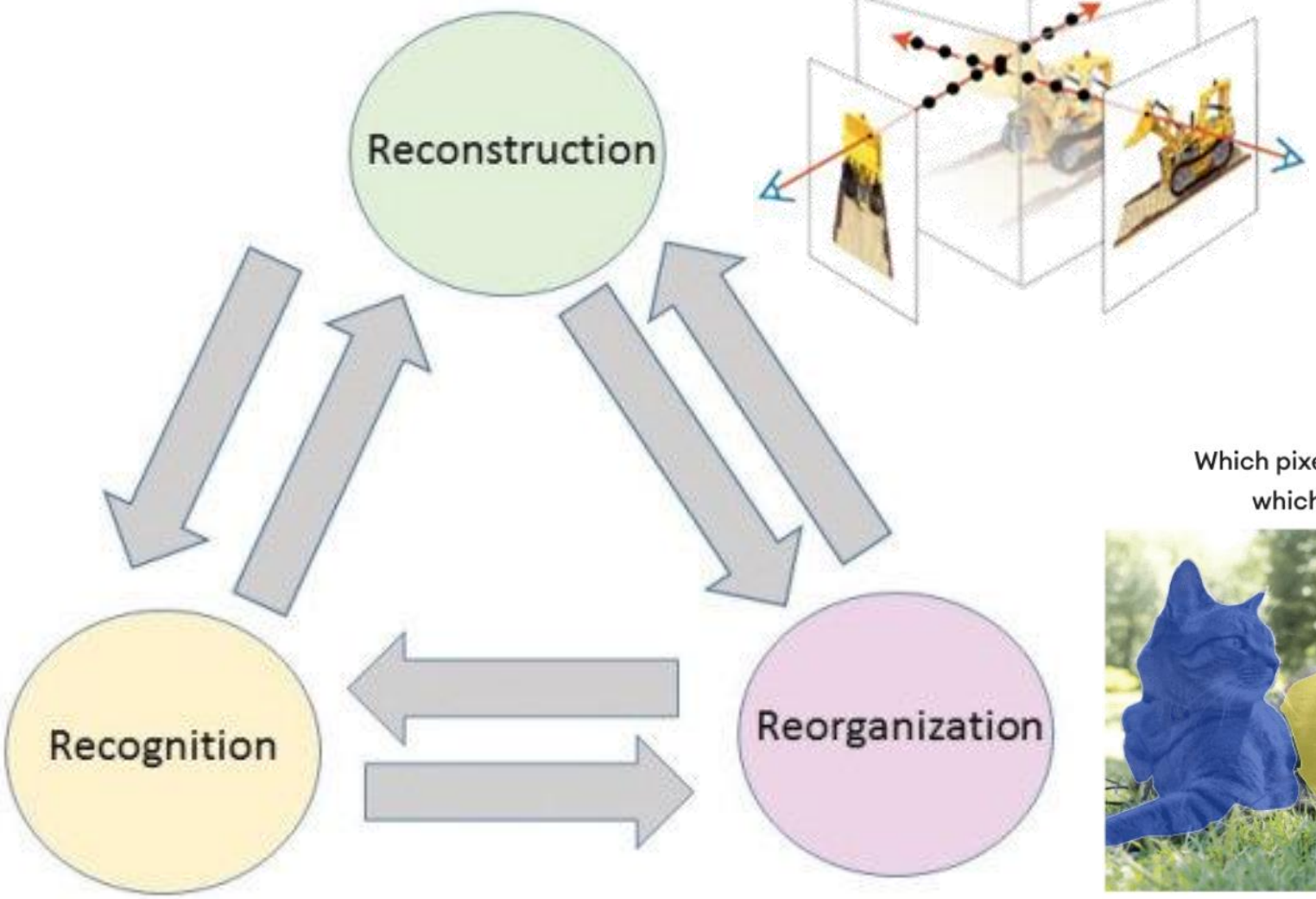
Image Classification

Which pixels belong to which object



Image Segmentation

3Rs in Computer Vision



Is this a cat?

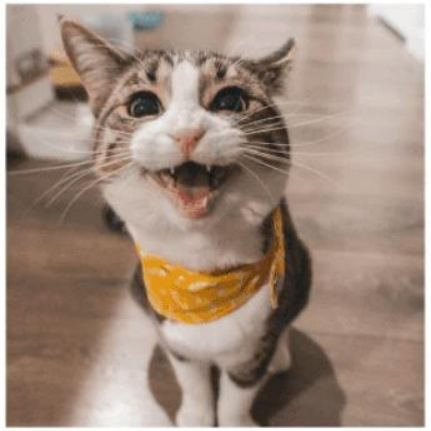


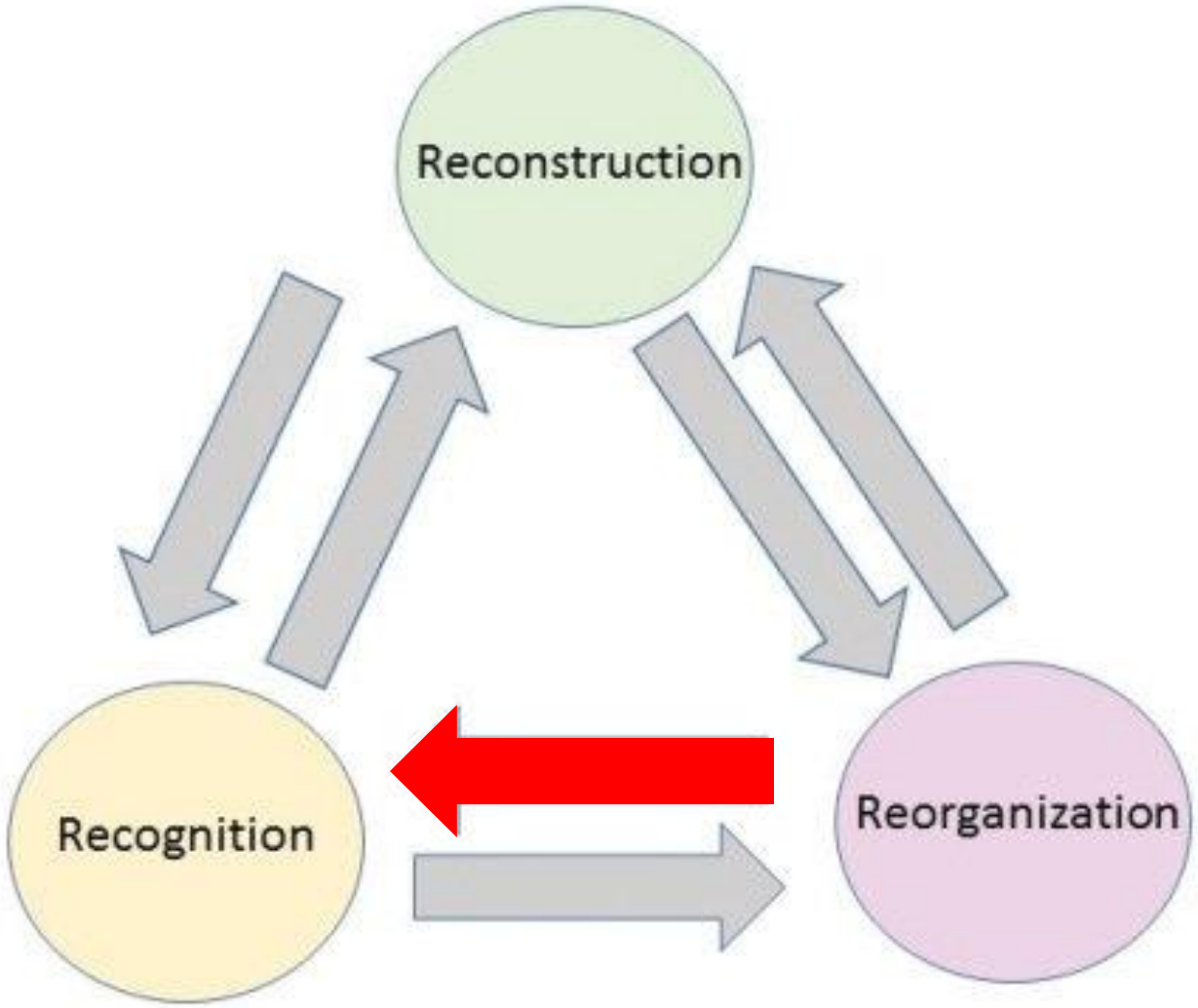
Image Classification

Which pixels belong to which object



Image Segmentation

1. Segment-grounded Recognition



Learning Hierarchical Image Segmentation For Recognition and By Recognition



Tsung-Wei Ke*



Sangwoo Mo*



Stella X. Yu

ICLR 2024

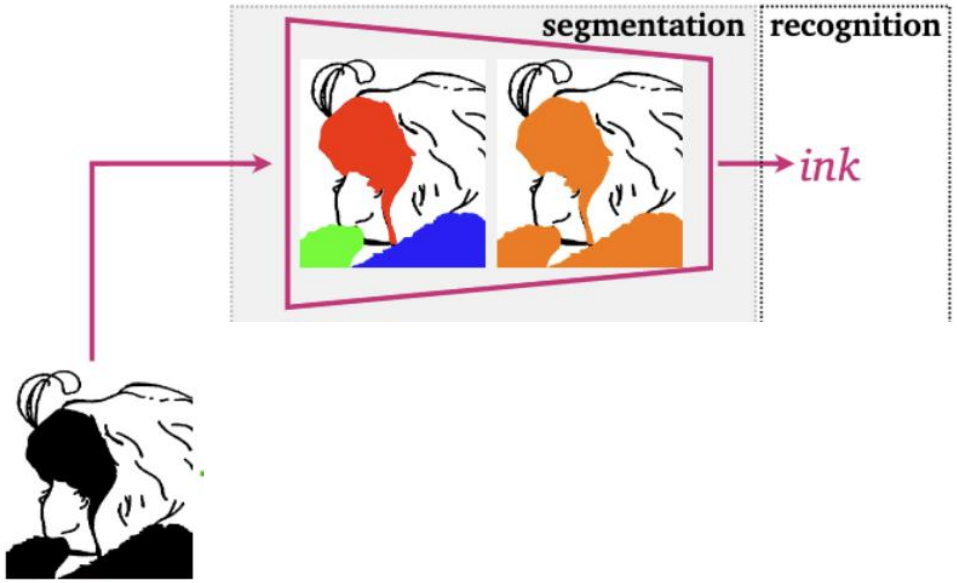
* Equal contribution

Segmentation & Recognition run **Concurrently**

Segmentation & Recognition run **Concurrently**

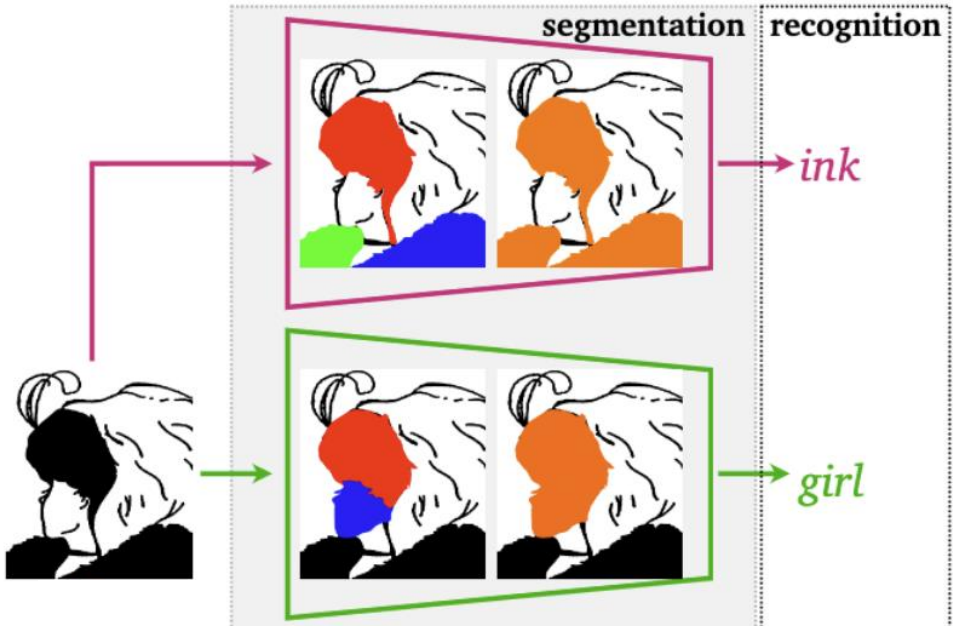


Segmentation & Recognition run **Concurrently**



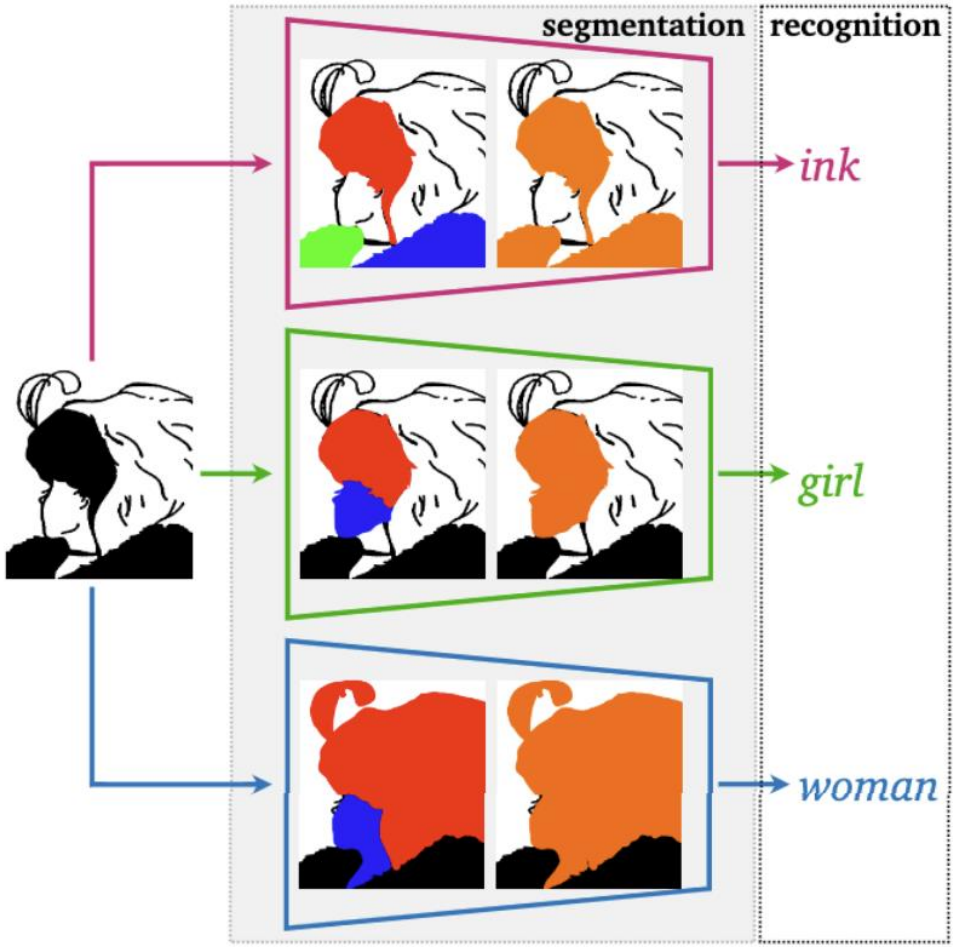


Segmentation & Recognition run **Concurrently**



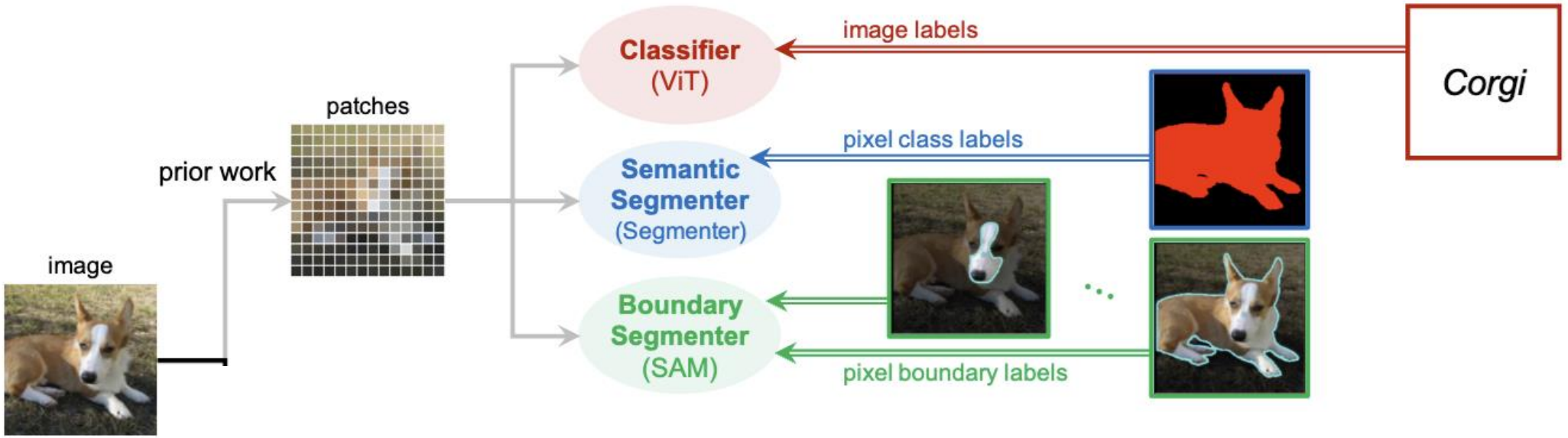


Segmentation & Recognition run **Concurrently**

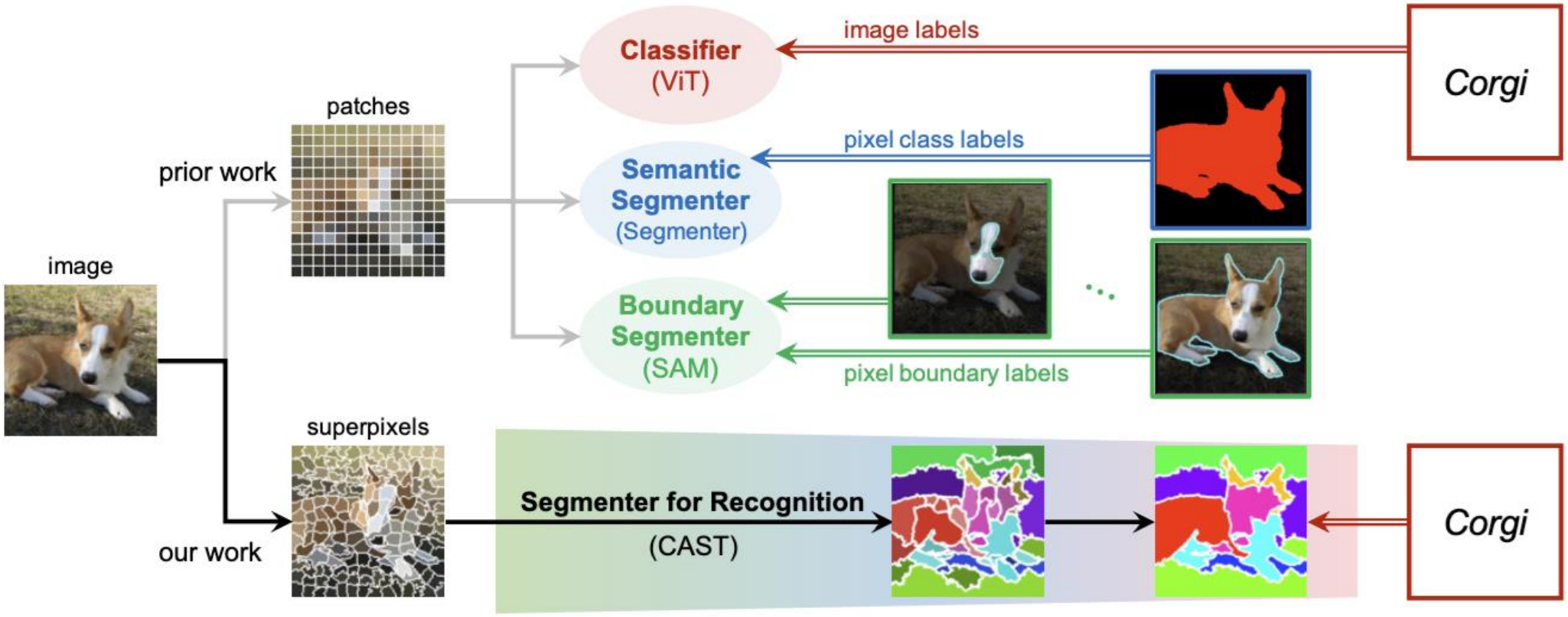




Continuum of Segmentation for Recognition

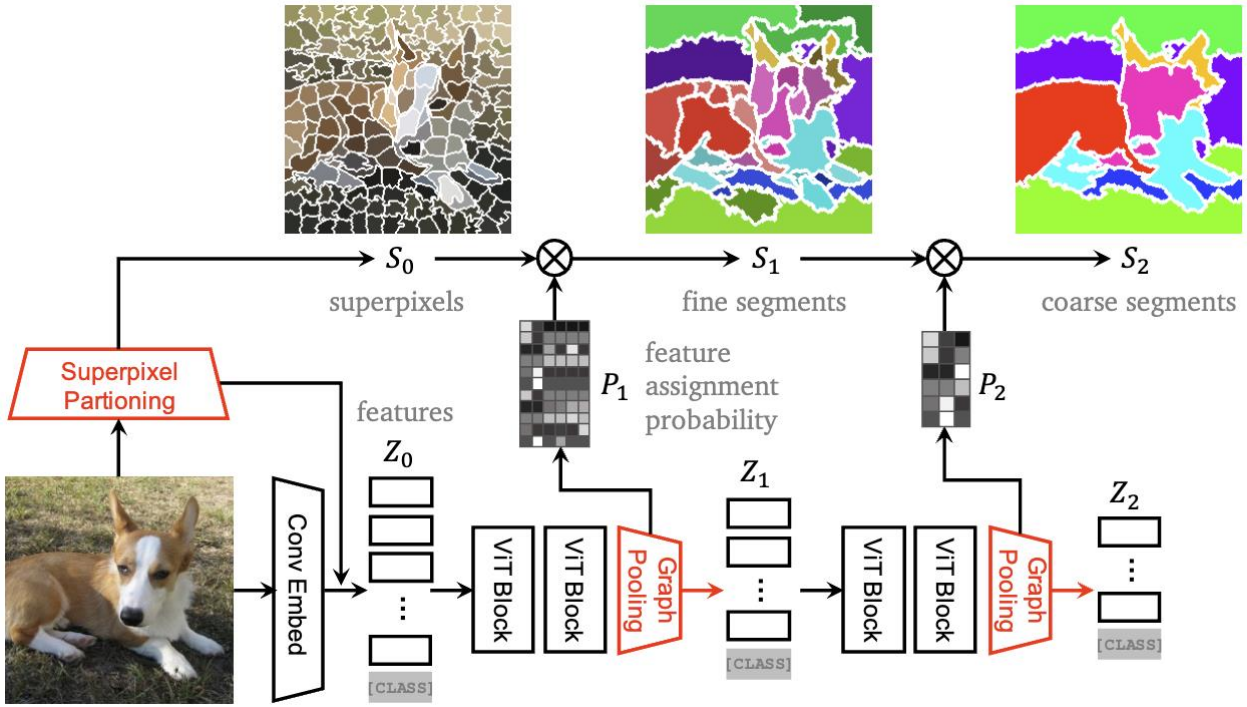


Continuum of Segmentation for Recognition



fine parts (e.g., ear) → coarse parts (e.g., head) → wholes (e.g., Corgi)

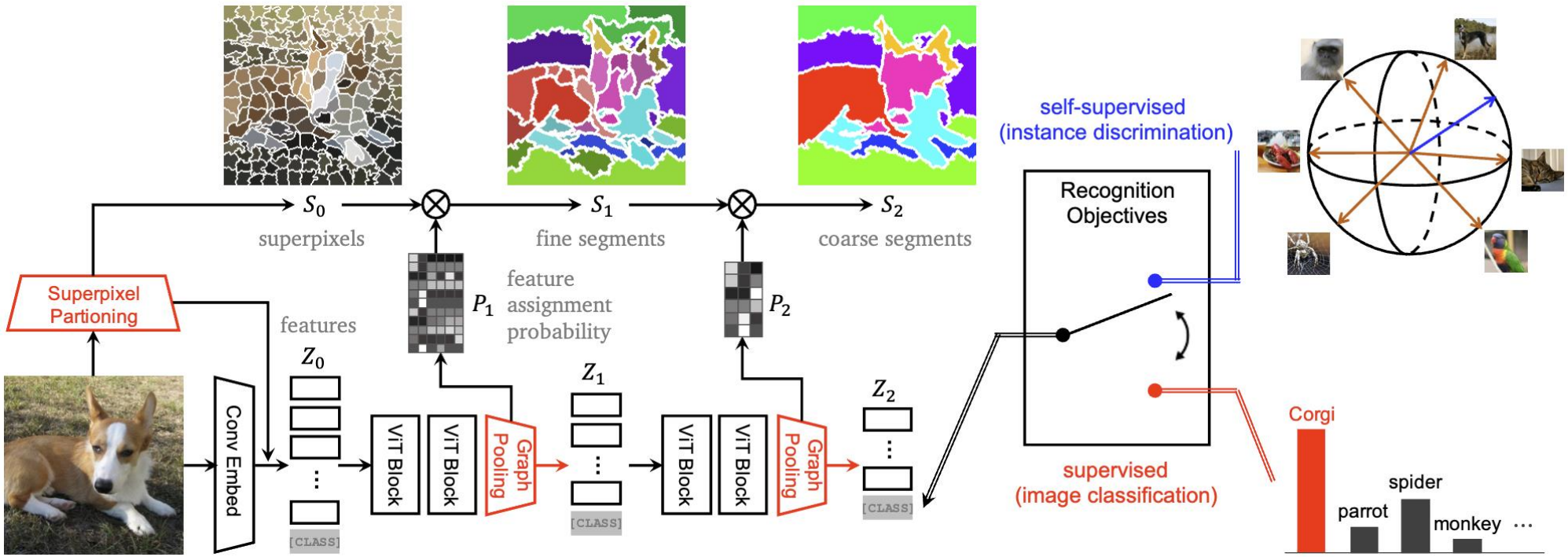
Our CAST Framework



Concurrent segmentation & recognition using

Adaptive **S**egment **T**okens

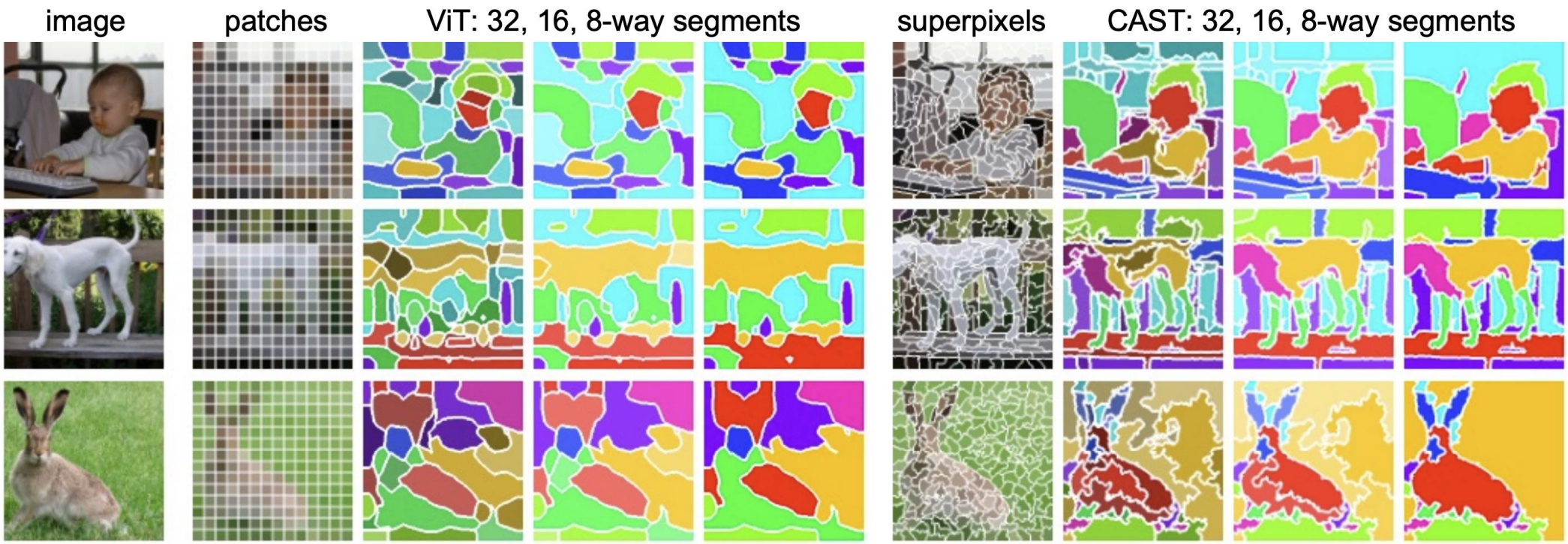
Our CAST Framework



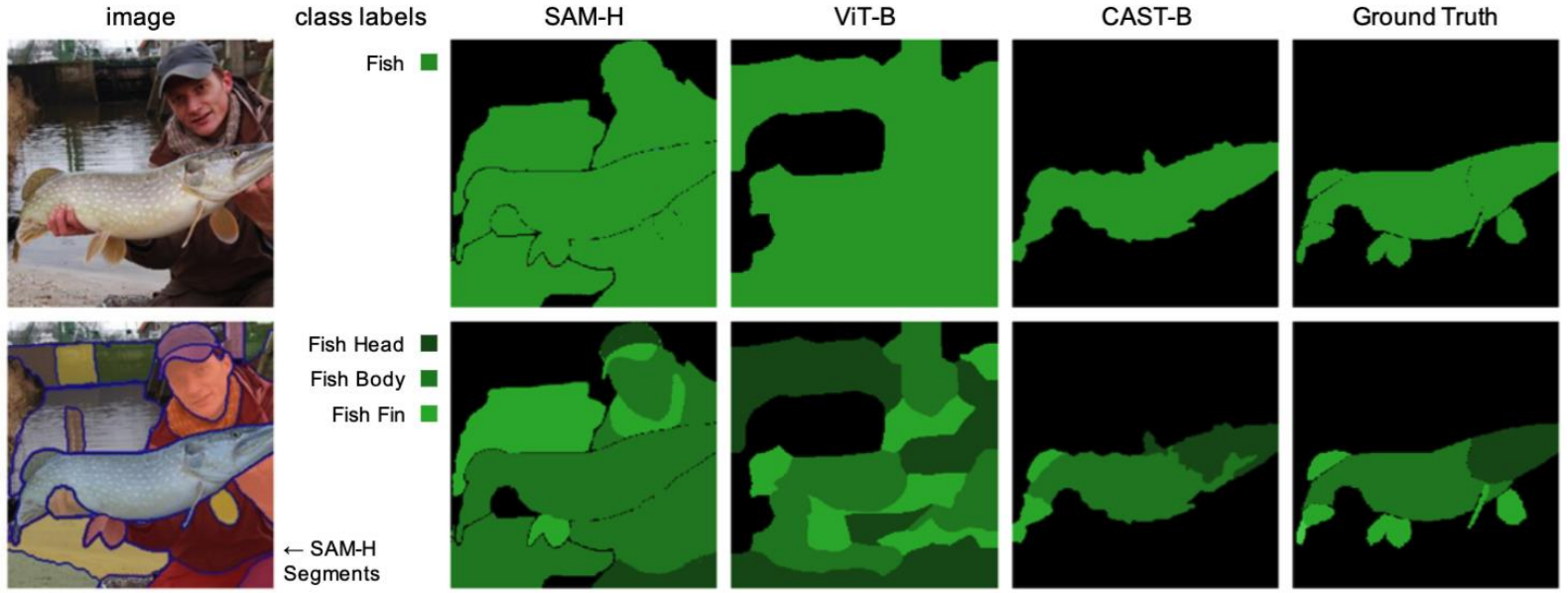
Hierarchical segmentation **for** recognition

learned unsupervisedly **by** recognition

Unsupervised Part-Whole Discovery



Unsupervised Part-Whole Discovery

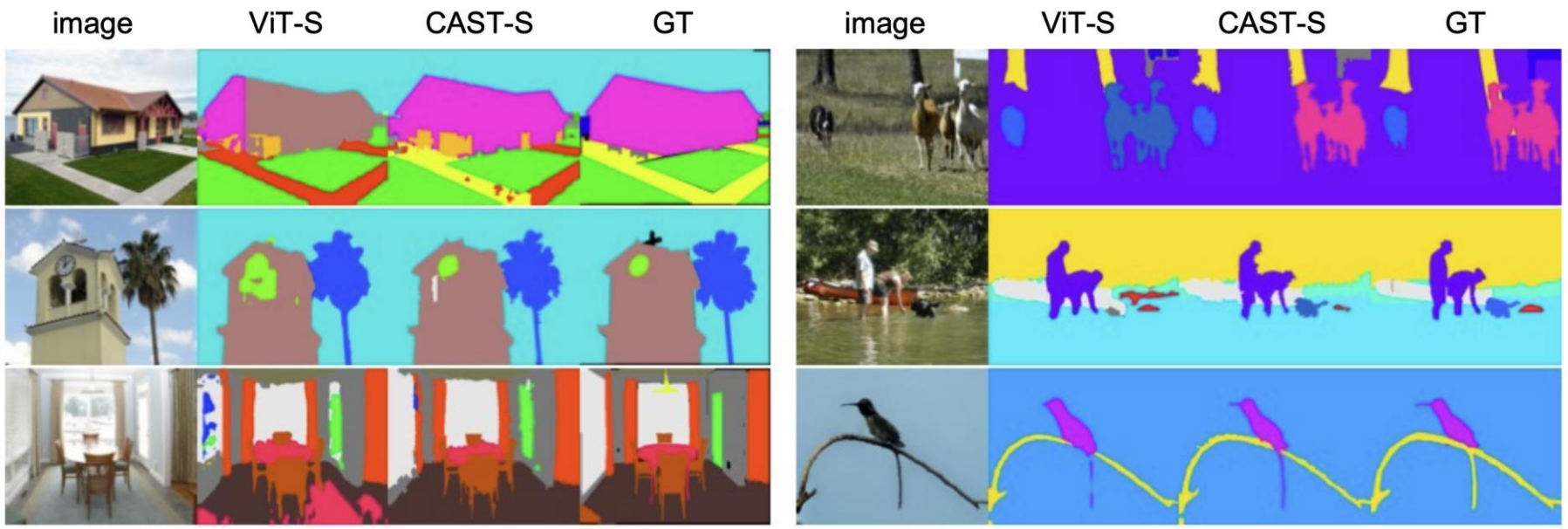


model	GFLOPS	region mIoU		boundary F-score	
SAM-B	677	18.03	10.15	20.71	7.25
SAM-H	3166	21.97	12.07	**32.66**	**11.82**
ViT-B	18	25.34	11.74	10.92	4.64
CAST-B	**13**	**29.66**	**13.20**	22.32	6.52

P R F	14 labels			4 labels			1 label		
HSG	20.7	18.6	19.6	24.1	30.6	26.9	20.5	36.1	26.2
CAST	**21.1**	**24.1**	**22.5**	**24.8**	**33.2**	**28.4**	**26.3**	**44.9**	**33.2**
gain	0.4	5.5	2.9	0.7	2.6	1.5	5.8	8.8	7.0

SOTA models (e.g., Segment Anything) lack this **consistency**

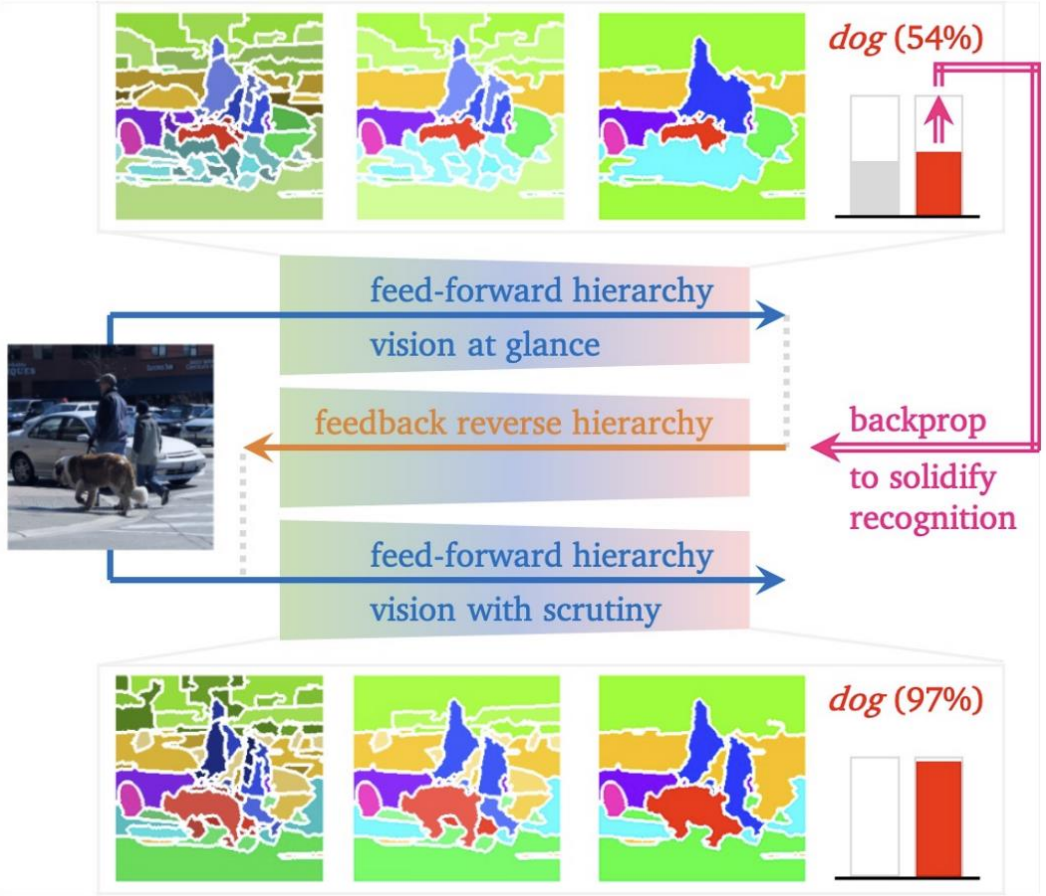
Better Segmentation & Recognition than ViT



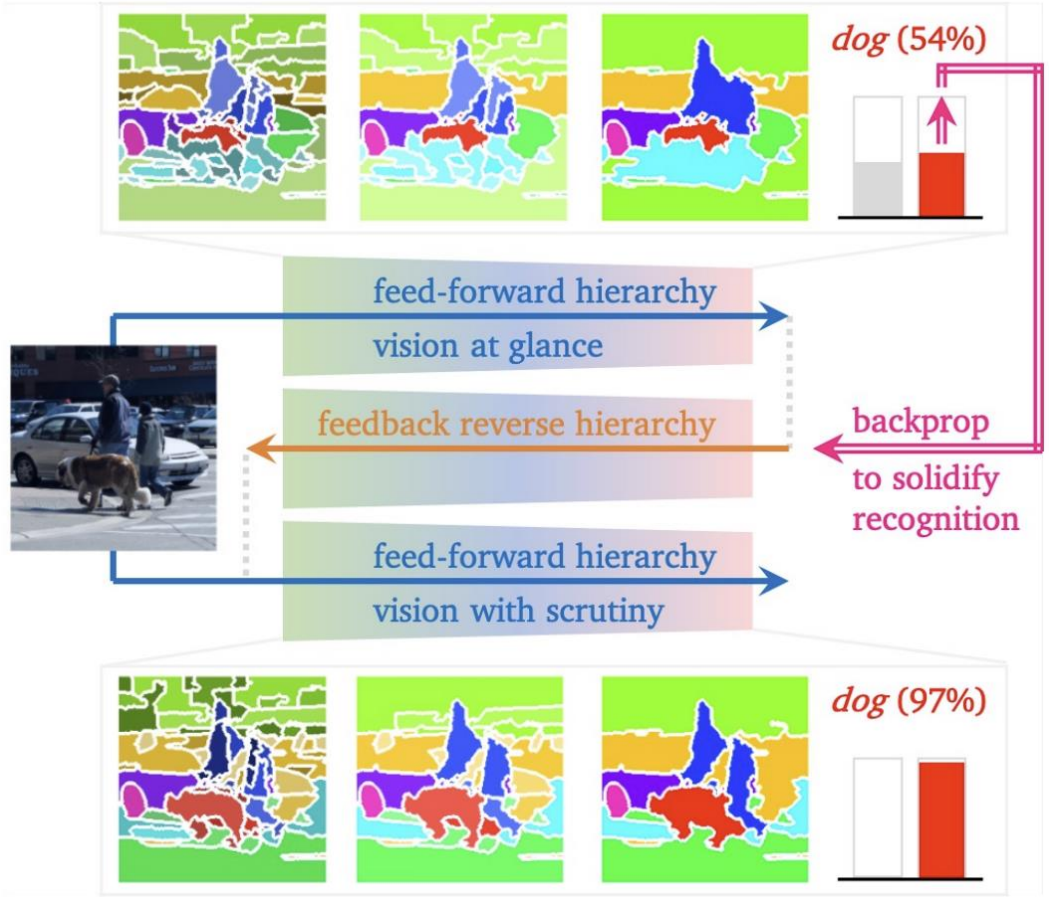
test on PASCAL-VOC	before tuning		after tuning	
ViT-S	30.9	16.1	65.8	40.7
ViT-S but with token pooling	34.5	19.8	67.2	41.9
ViT-S but with superpixels	32.2	21.2	66.5	46.7
CAST-S	**38.4**	**27.0**	**67.6**	**48.1**

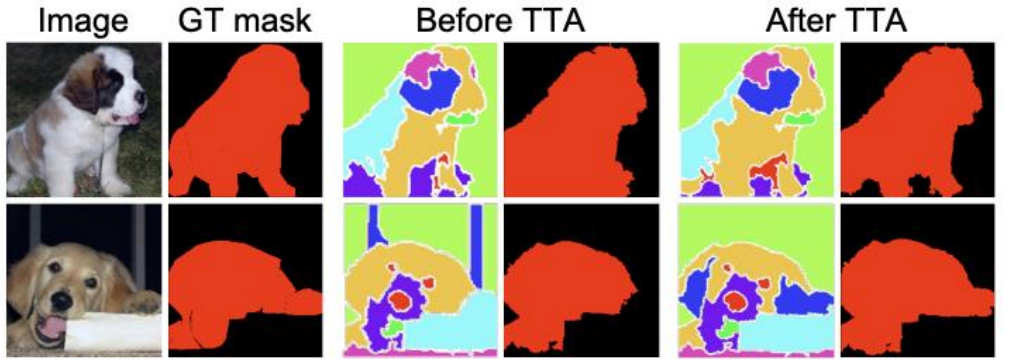
Model	GFLOPS	IN-100	IN-1K
ViT-S	4.7	78.1	67.9
Swin-T	4.5	78.3	63.0
CAST-S	**3.4**	**79.9**	**68.1**

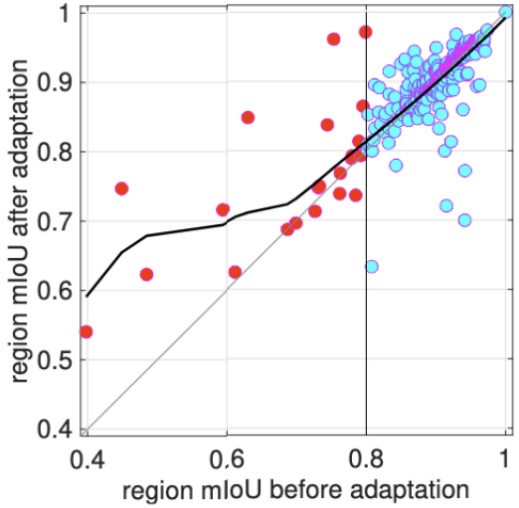
Recognition Also Guides Segmentation



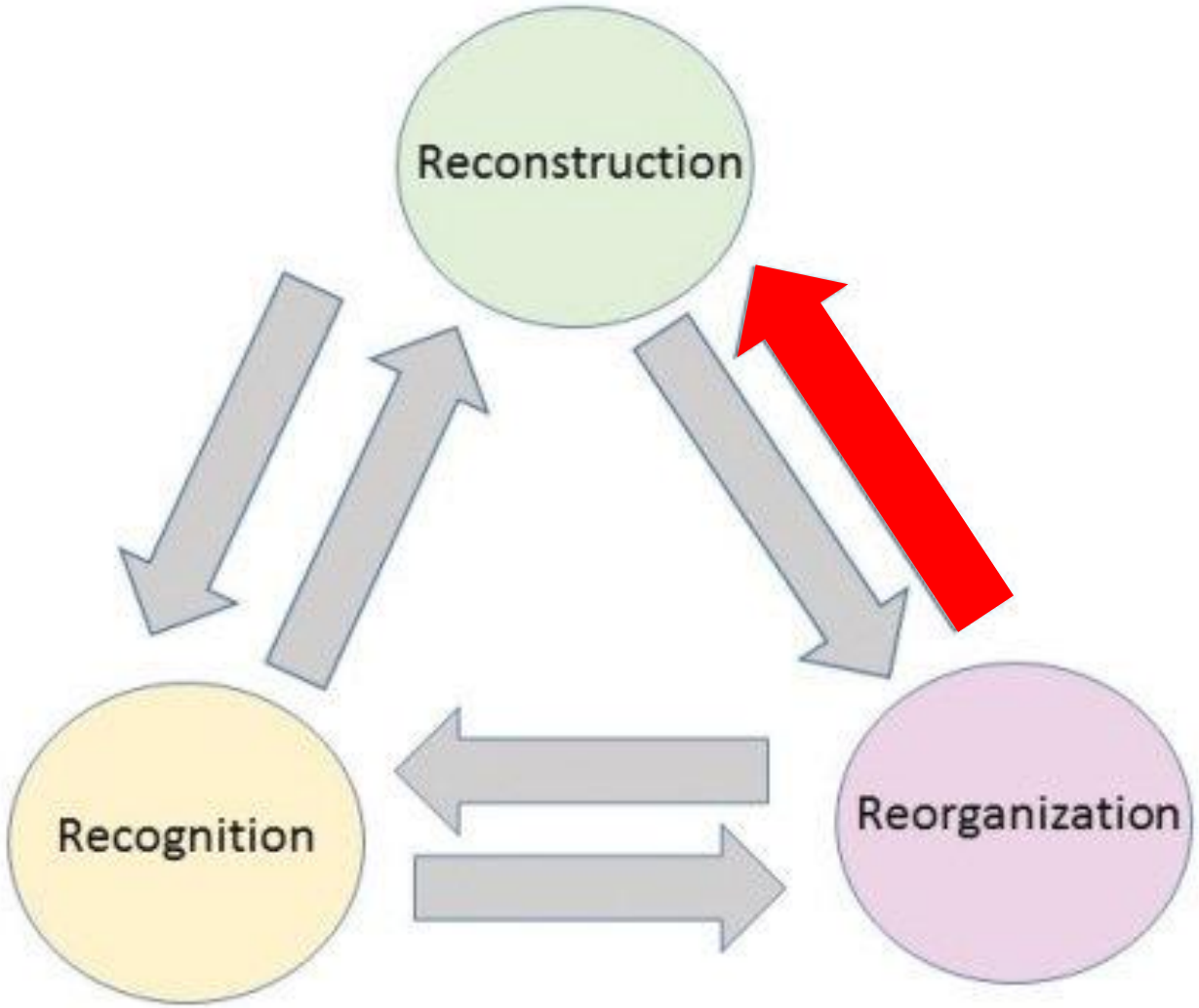
Recognition Also Guides Segmentation







2. Segment-grounded Dense Prediction



SHED Light on Segmentation for Depth Estimation



Seung-Hyun Lee*



Sangwoo Mo*



Stella X. Yu



Under Review

* Equal contribution

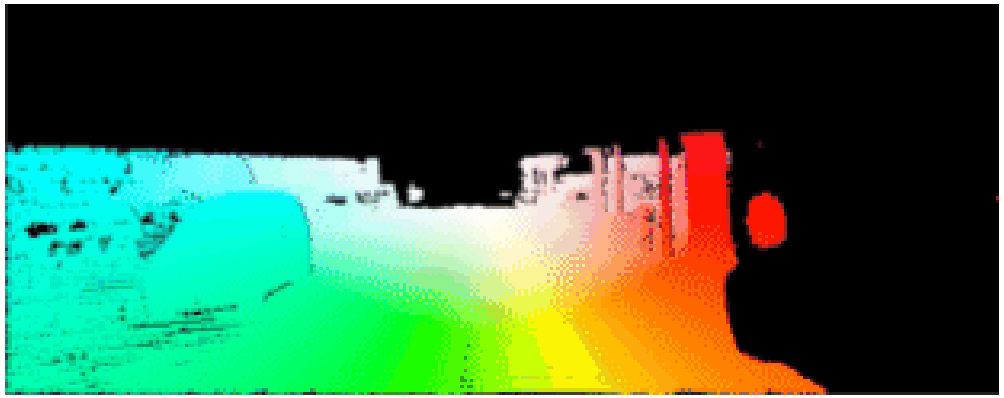


Dense Prediction is Also Segment-grounded

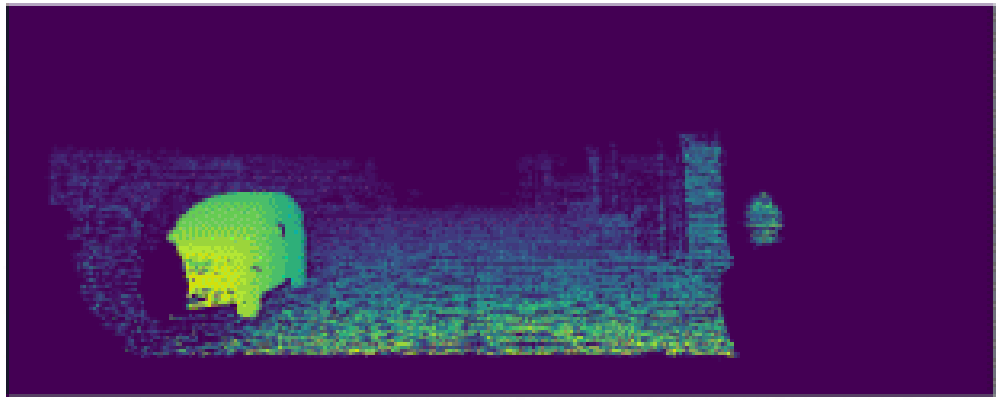
segmentation





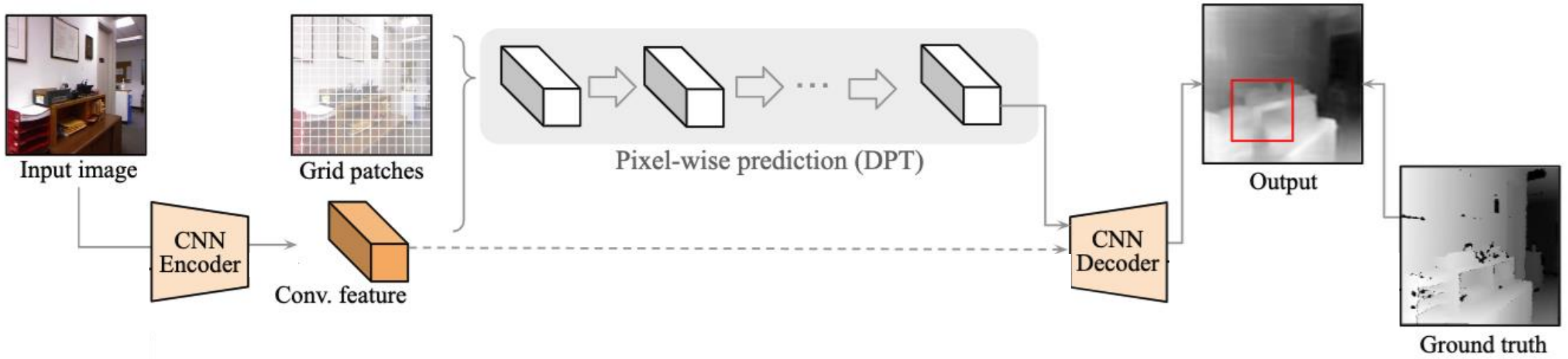


optical flow

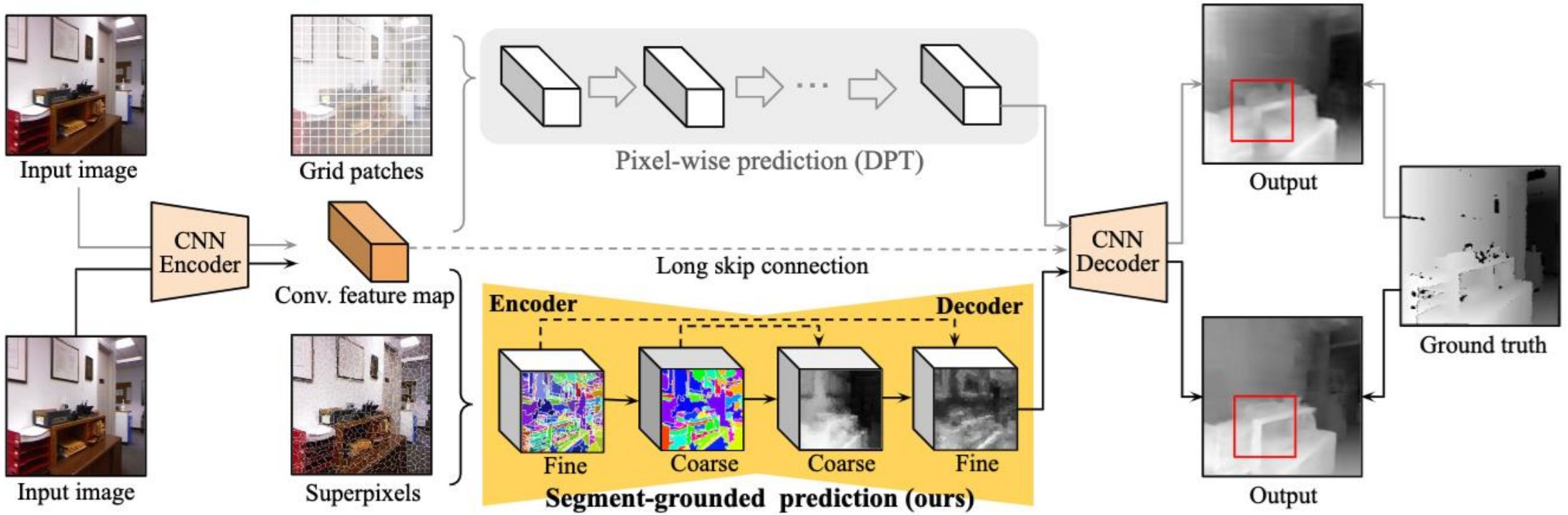


depth

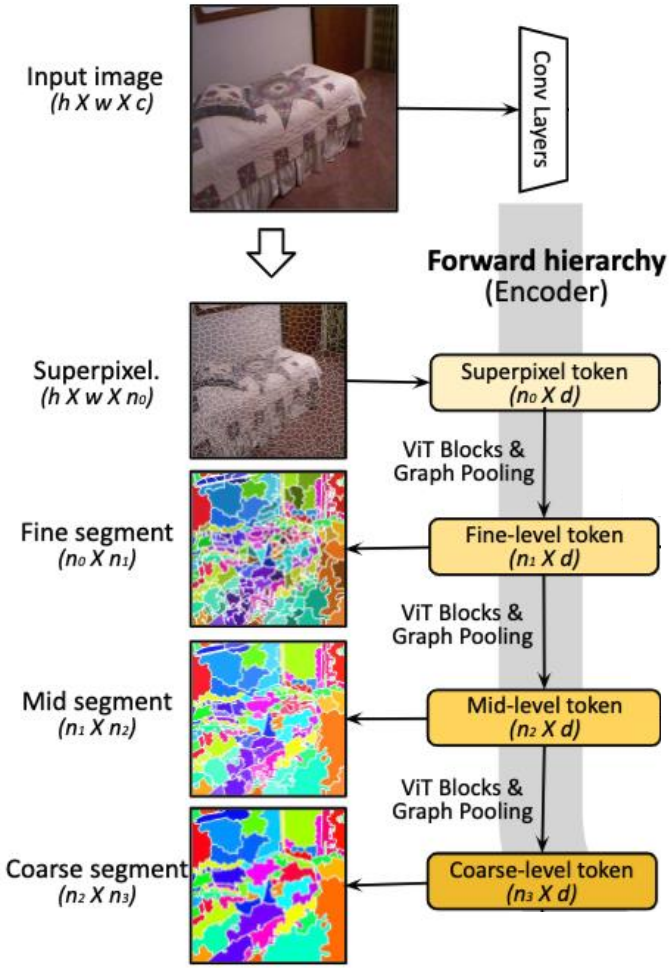
Dense Prediction is Also Segment-grounded



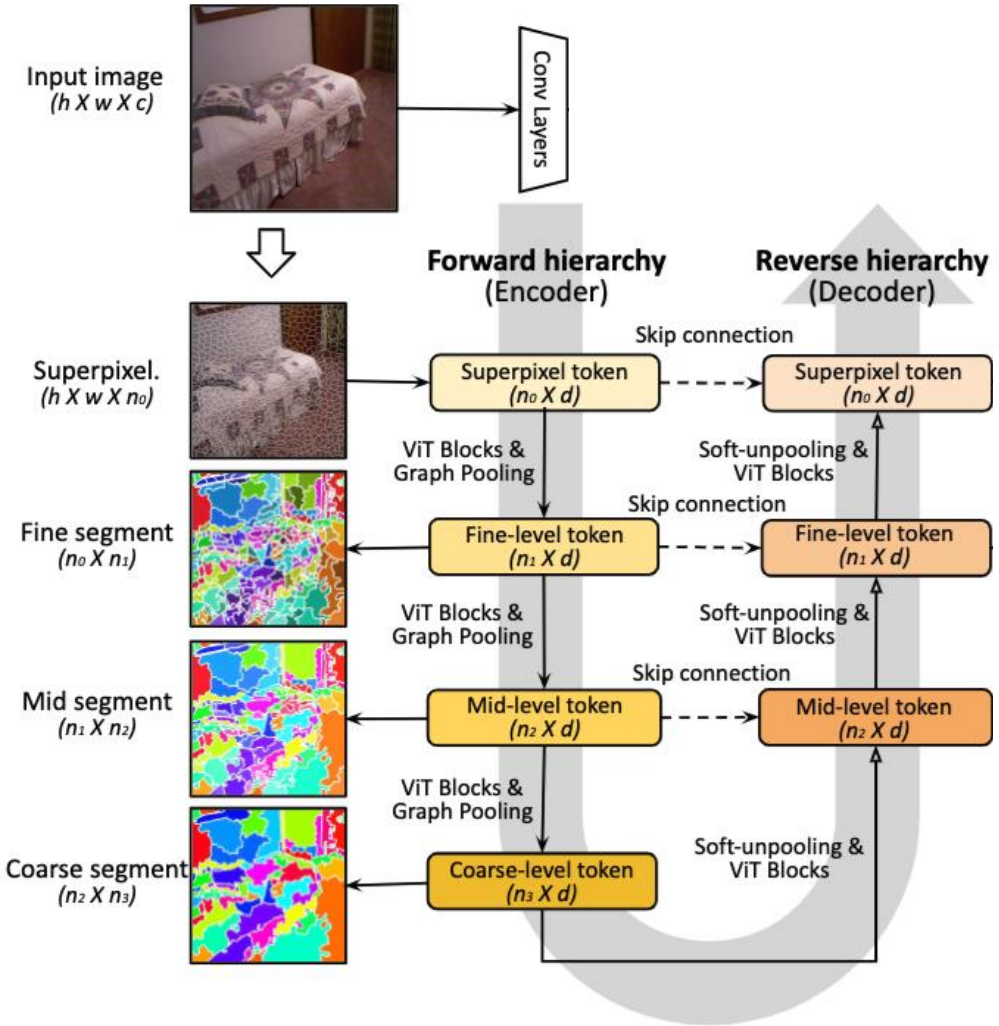
Dense Prediction is Also Segment-grounded



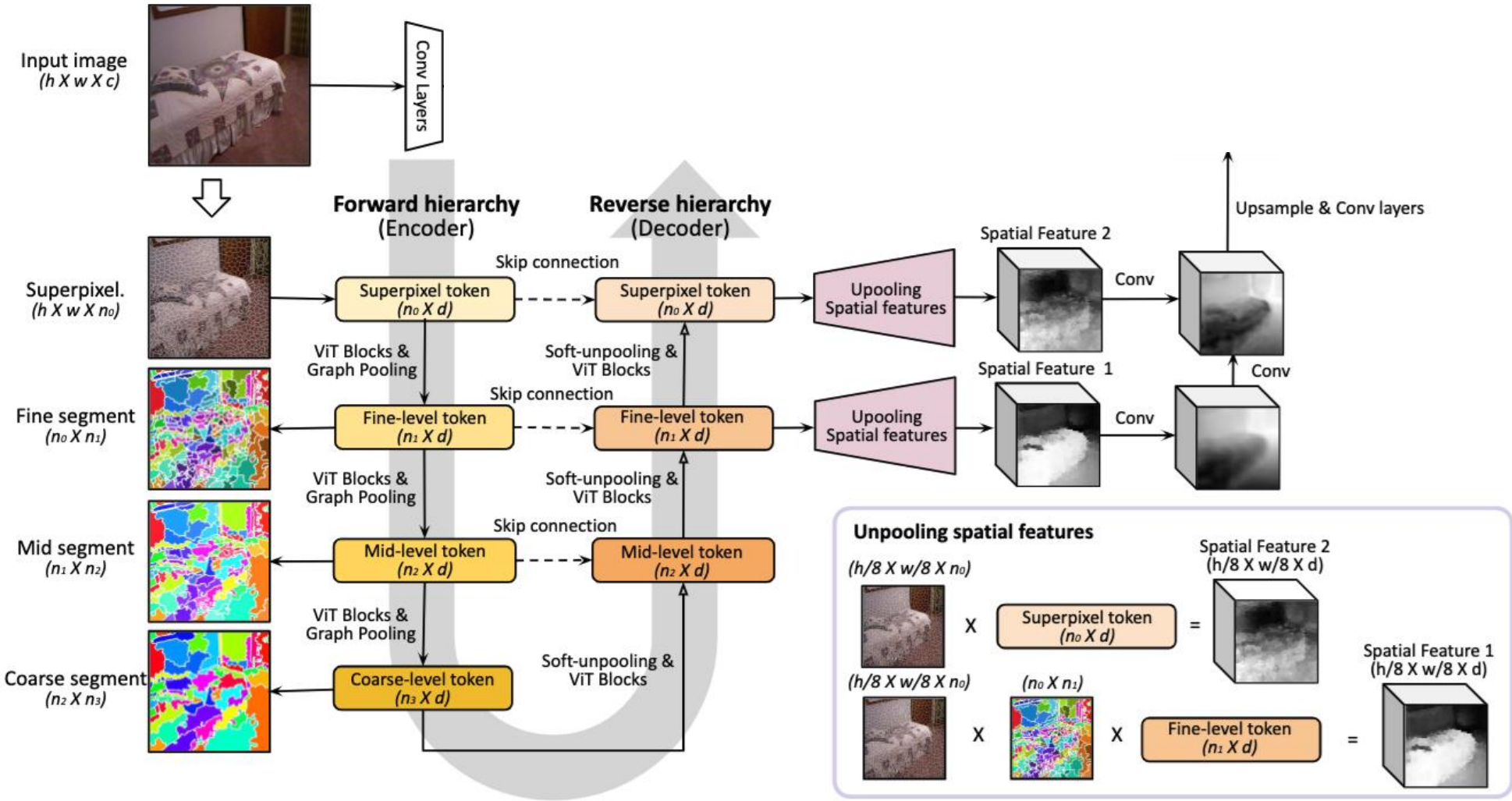
Our SHED Framework



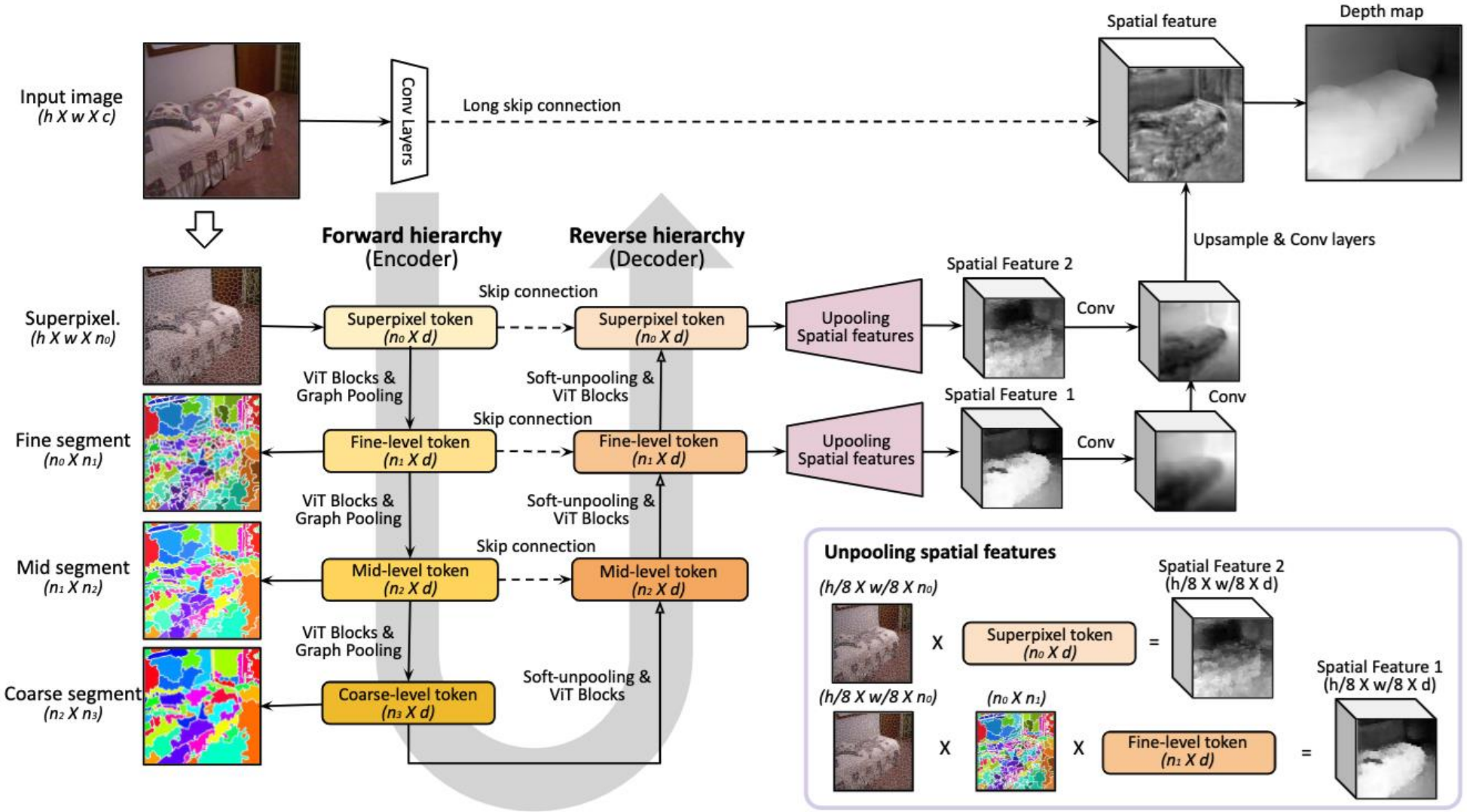
Our SHED Framework



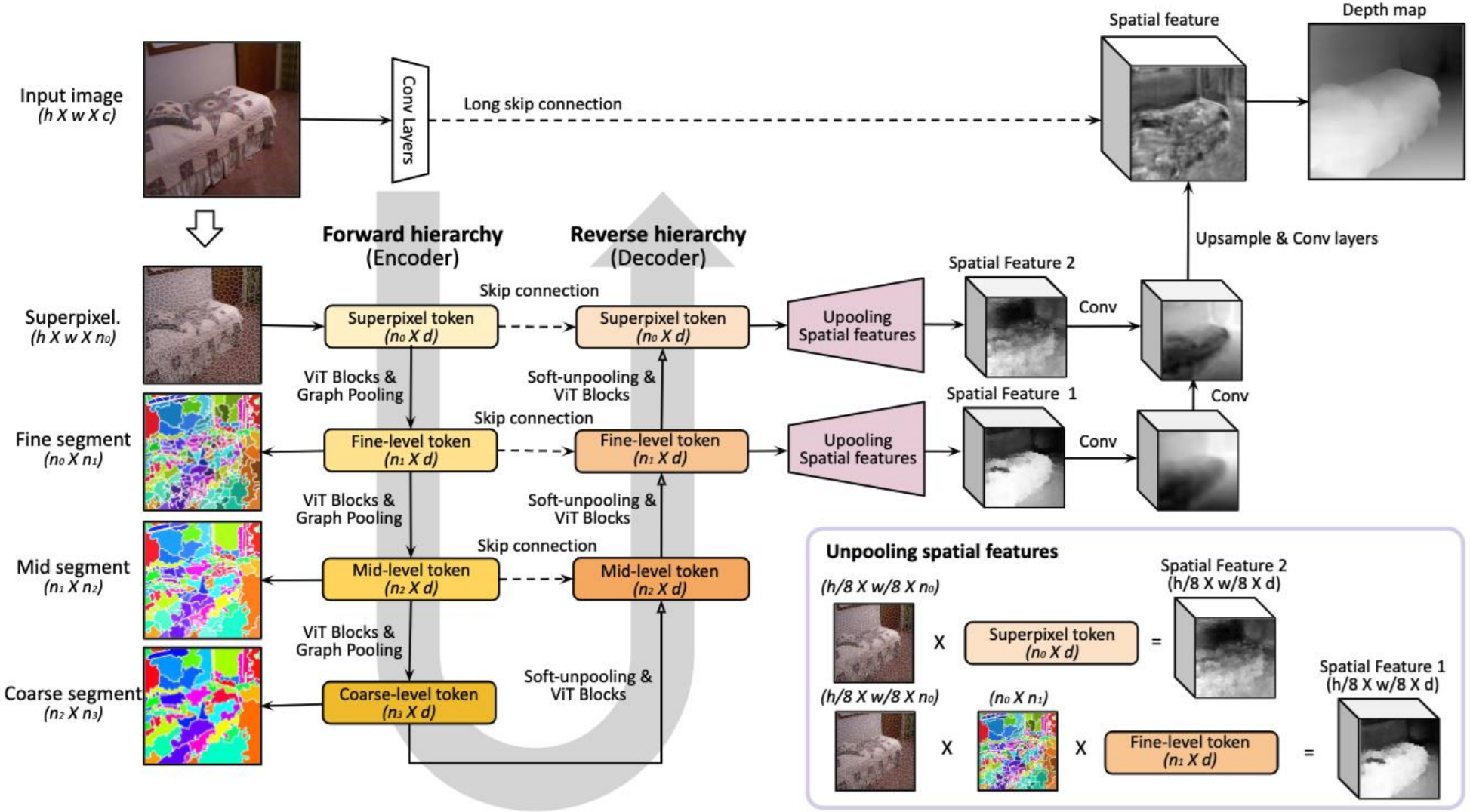
Our SHED Framework



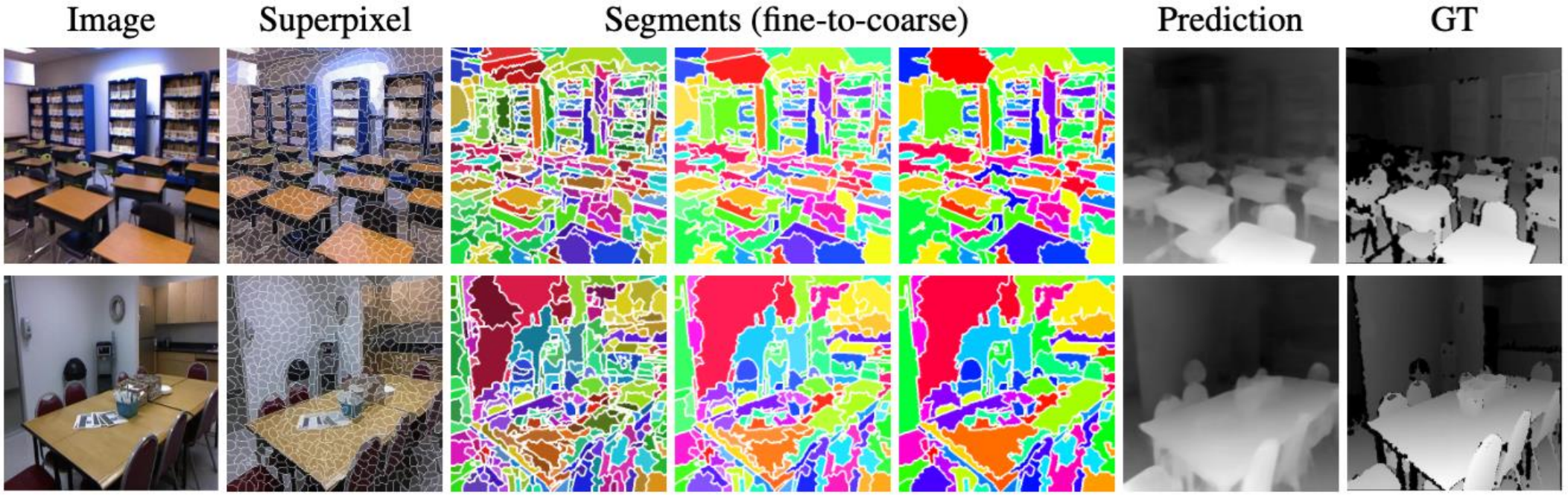
Our SHED Framework



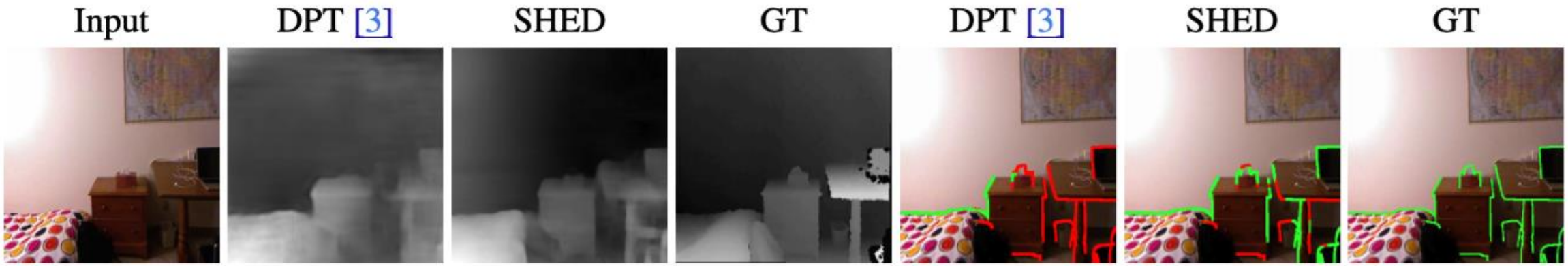
Segment Hierarchy for Estimating Depth



Segment-consistent Dense Prediction



Segment-consistent Dense Prediction



Method	Boundary Accuracy (Precision / Recall) ↓	Intra-segment coherence (Precision / Recall) ↓
DPT [3]	6.429 / 1.637	0.631 / 0.664
SHED (ours)	**5.875 / 0.732**	**0.614 / 0.614**

Layout-aware Representation Learning

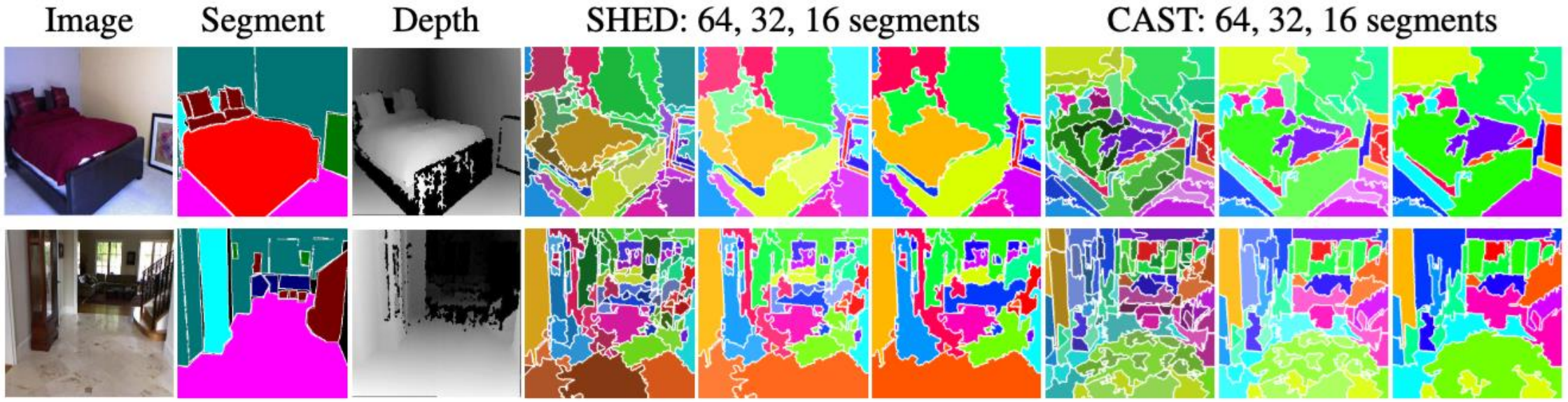
a) Top-5 retrieval examples



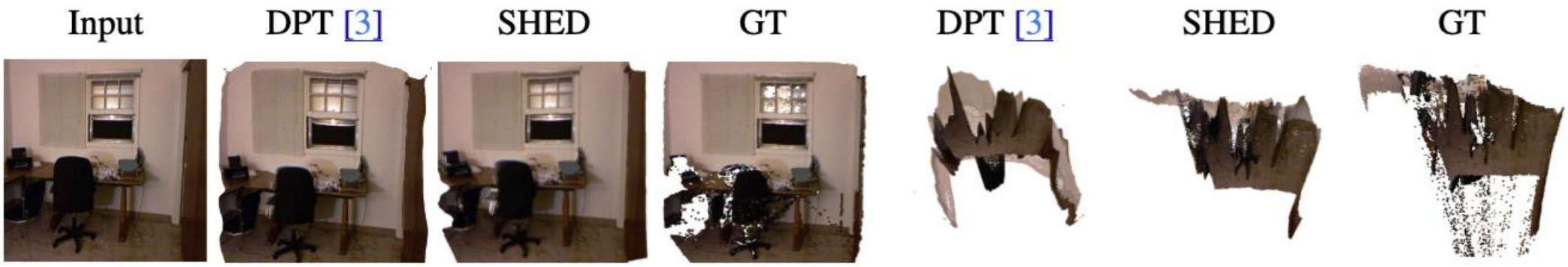
b) Top-K retrieval accuracy (%)

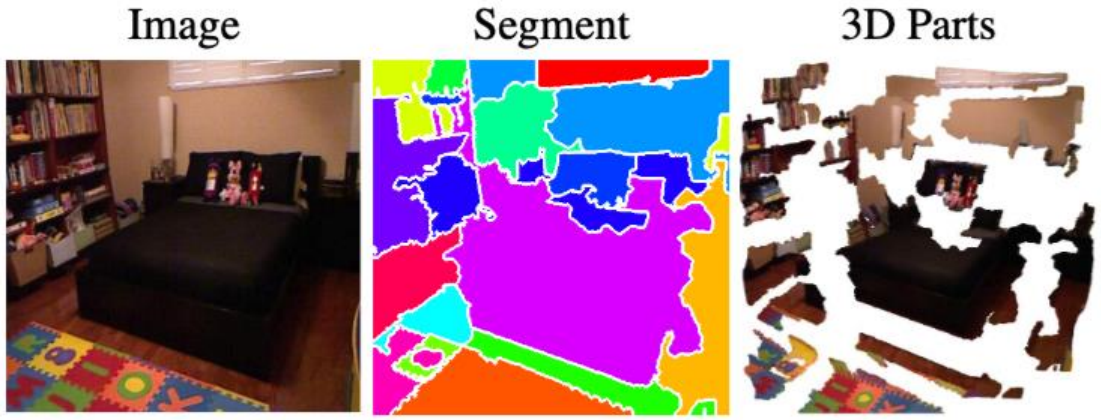
Method	Top-1	Top-3	Top-5
Scene retrieval			
DPT [3]	45.2	69.7	77.2
SHED (ours)	**60.5**	**78.7**	**87.0**
Frame retrieval (k=5)			
DPT [3]	18.5	31.0	38.3
SHED (ours)	**30.5**	**42.3**	**48.3**

Depth-aware Segment Hierarchy

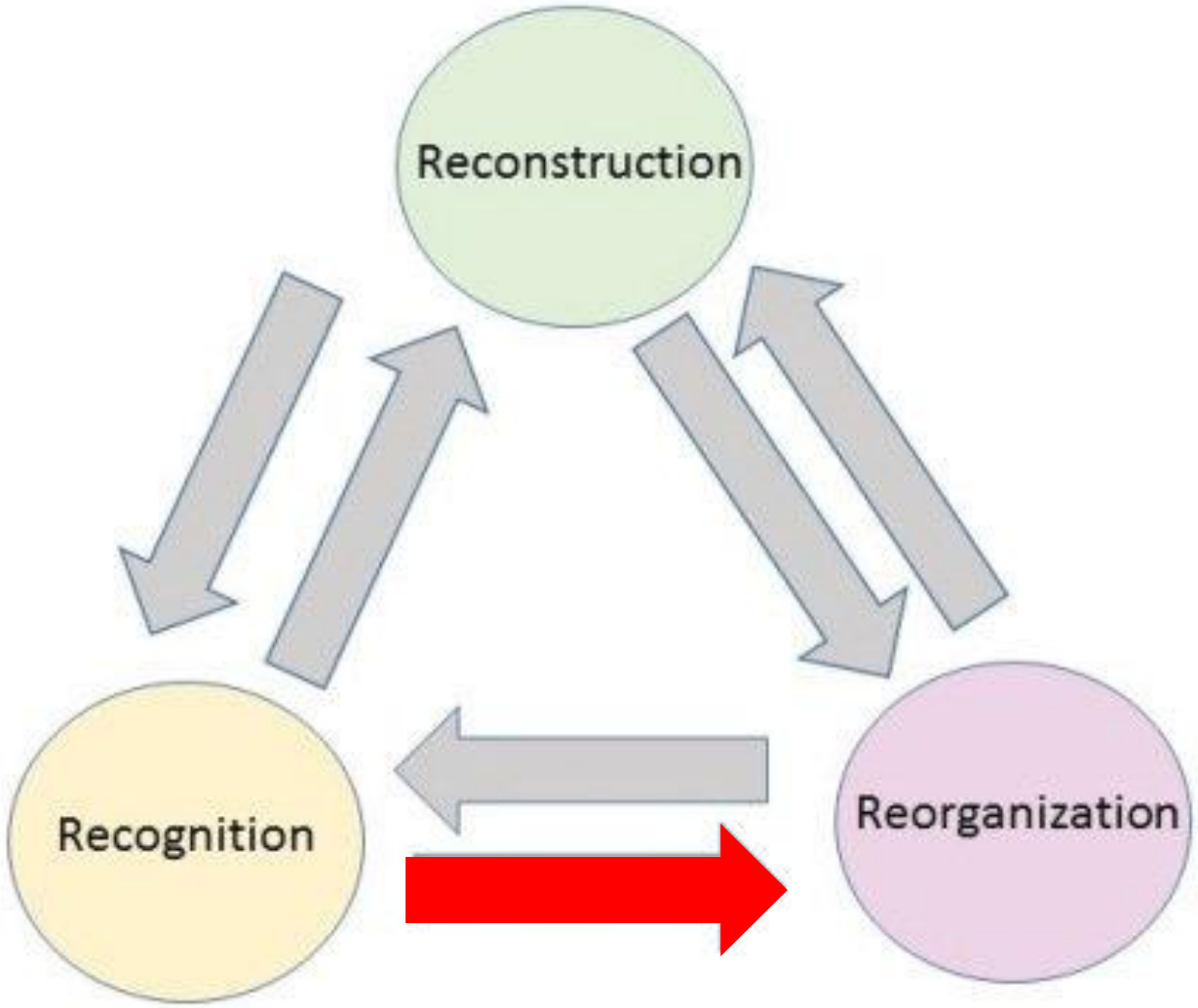


Accurate and Structured 3D Reconstruction





3. Recognition-guided Visual Grounding



Open Ad-hoc Categorization with Contextualized Feature Learning



Zilin Wang*



Sangwoo Mo*



Stella X. Yu



Sima Behpour



Liu Ren

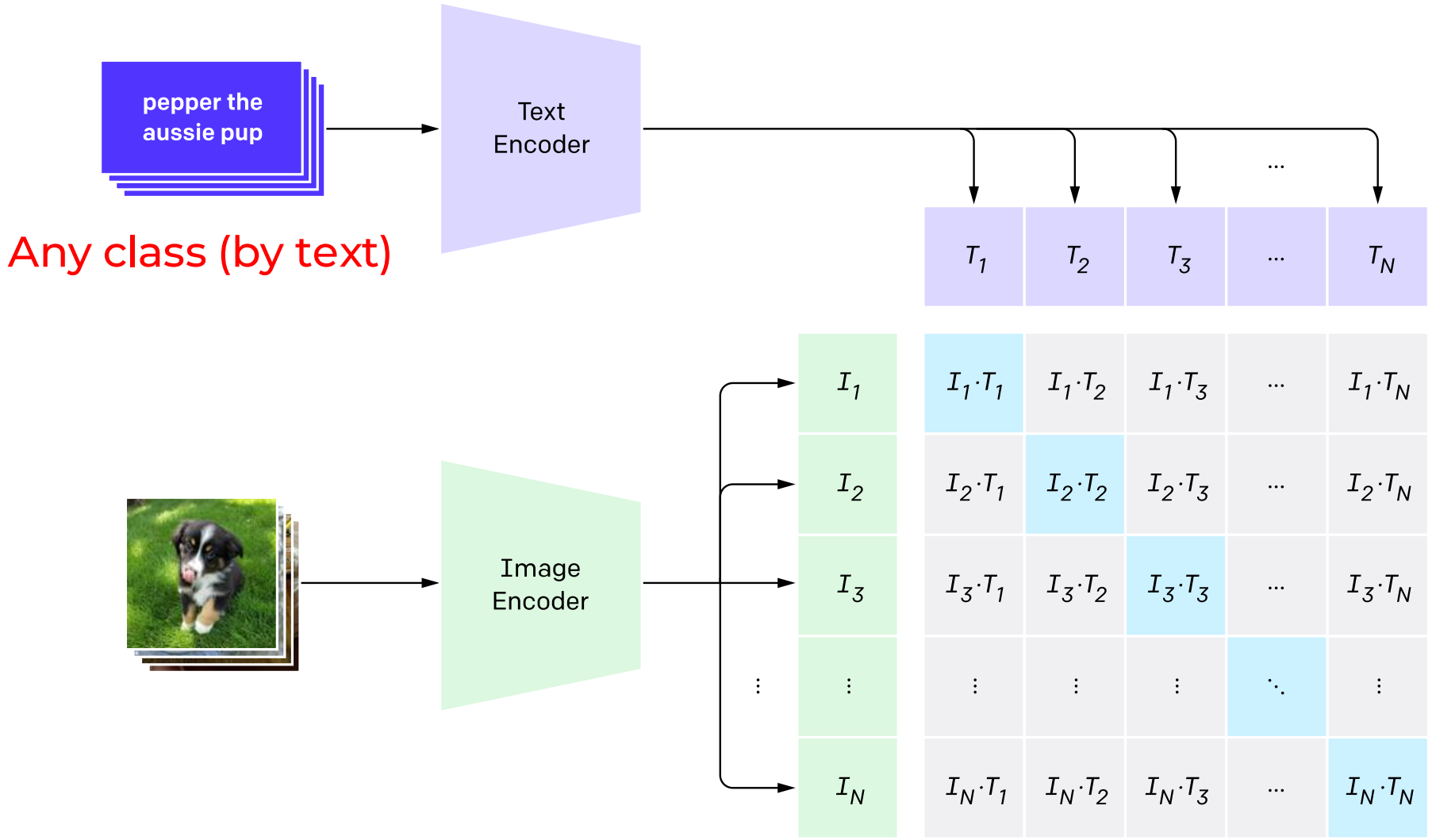


CVPR 2025

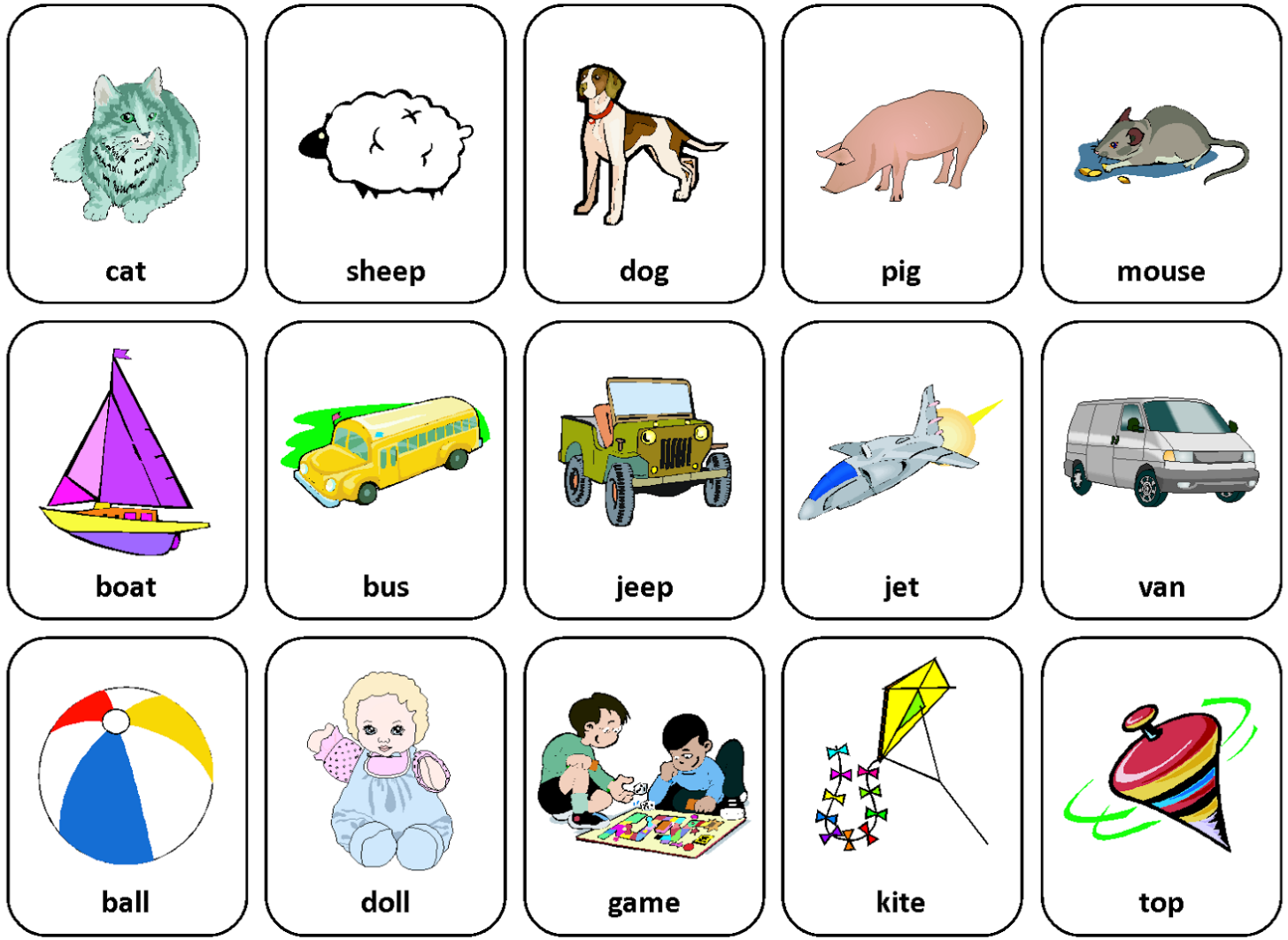
* Equal contribution



Is Categorization **Solved**?



CLIP works well for **Basic Categories**



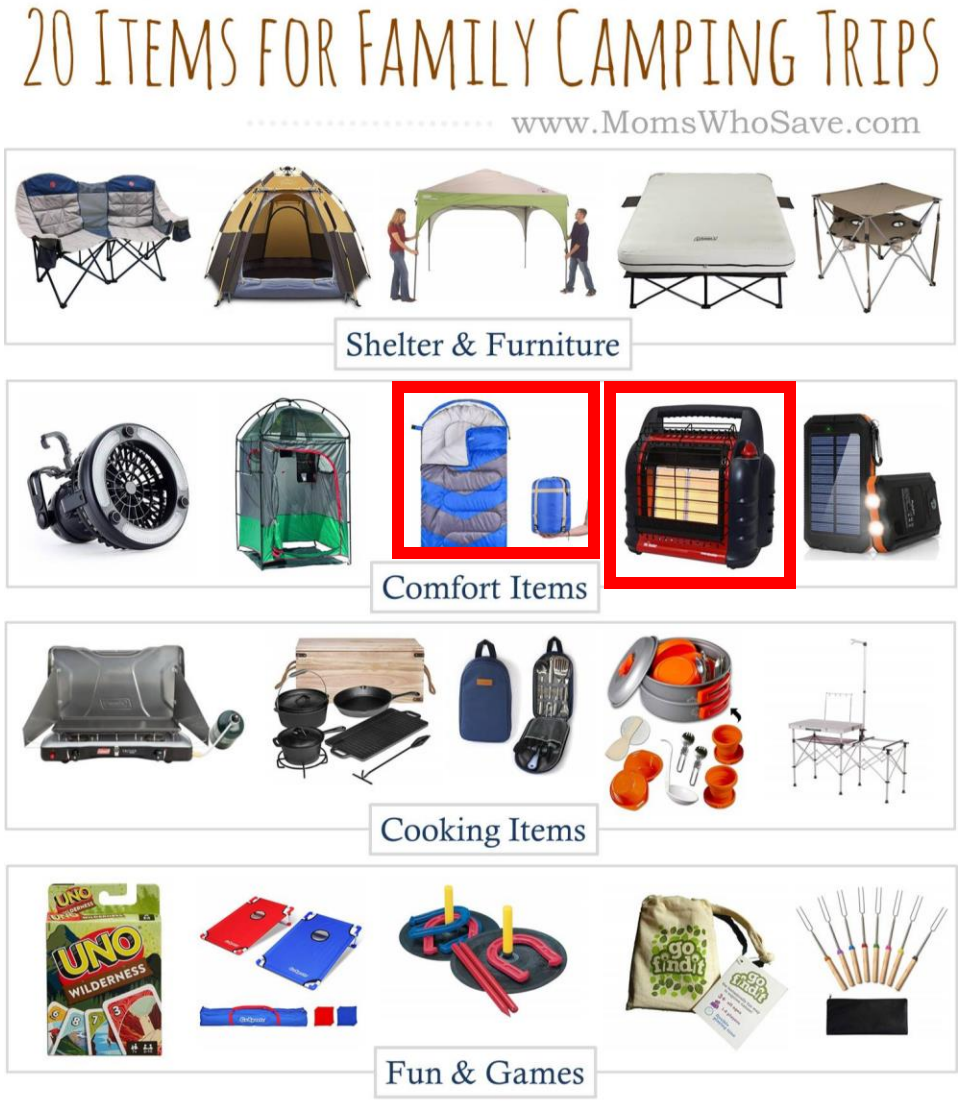
Source: https://testyyettrying.blogspot.com/2013/07/sorting-by-category-free-speech-therapy.html

However, Categorization is often **Ad-hoc**



Source: https://momswhosave.com/family-camping-trip-checklist/

However, Categorization is often **Ad-hoc**



Source: https://momswhosave.com/family-camping-trip-checklist/

Categorization is often **Open**



What else?

?

Source: https://momswhosave.com/family-camping-trip-checklist/

Open Ad-hoc Categorization



What else?

?

Source: https://momswhosave.com/family-camping-trip-checklist/

Open Ad-hoc Categorization

Semantics



What else?

?

Source: https://momswhosave.com/family-camping-trip-checklist/

Open Ad-hoc Categorization

Semantics



Visuals

What else?

?

Source: https://momswhosave.com/family-camping-trip-checklist/

Open Ad-hoc Categorization

Semantics (CLIP)



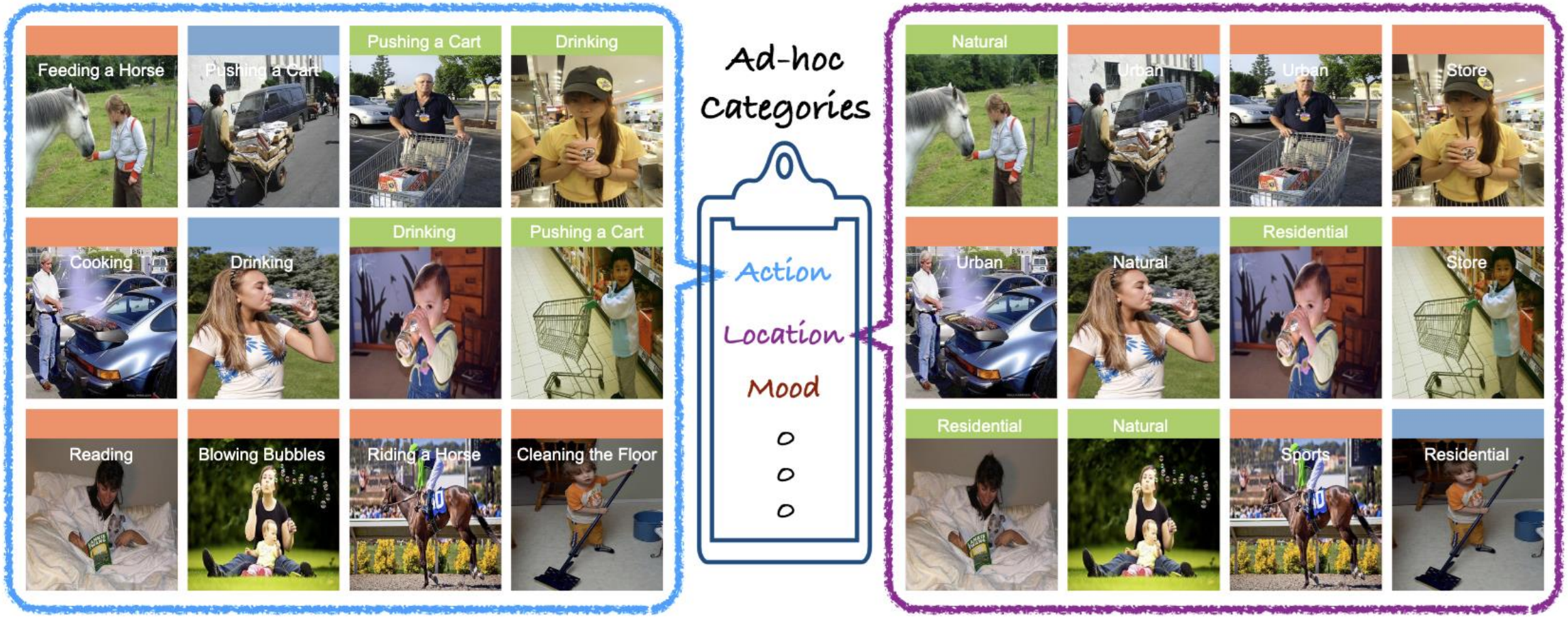
Visuals (GCD)

What else?

?

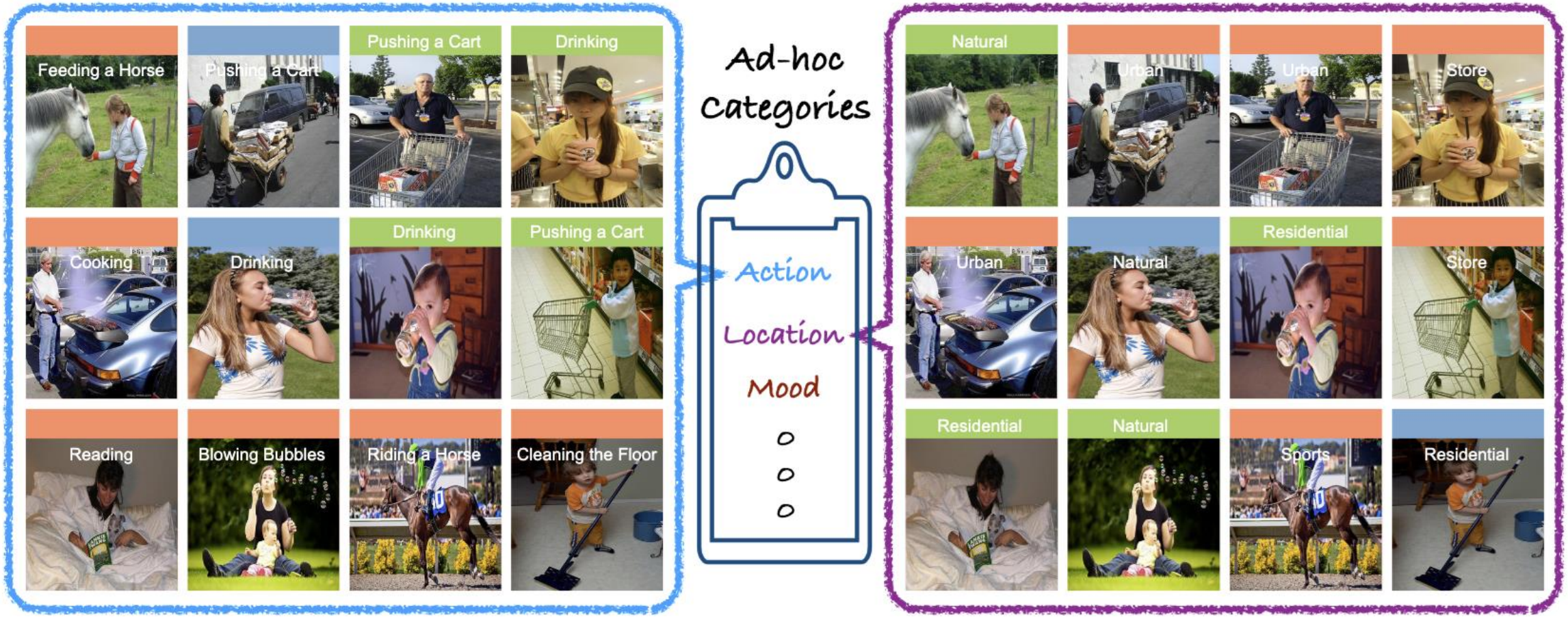
Source: https://momswhosave.com/family-camping-trip-checklist/

Problem Setup



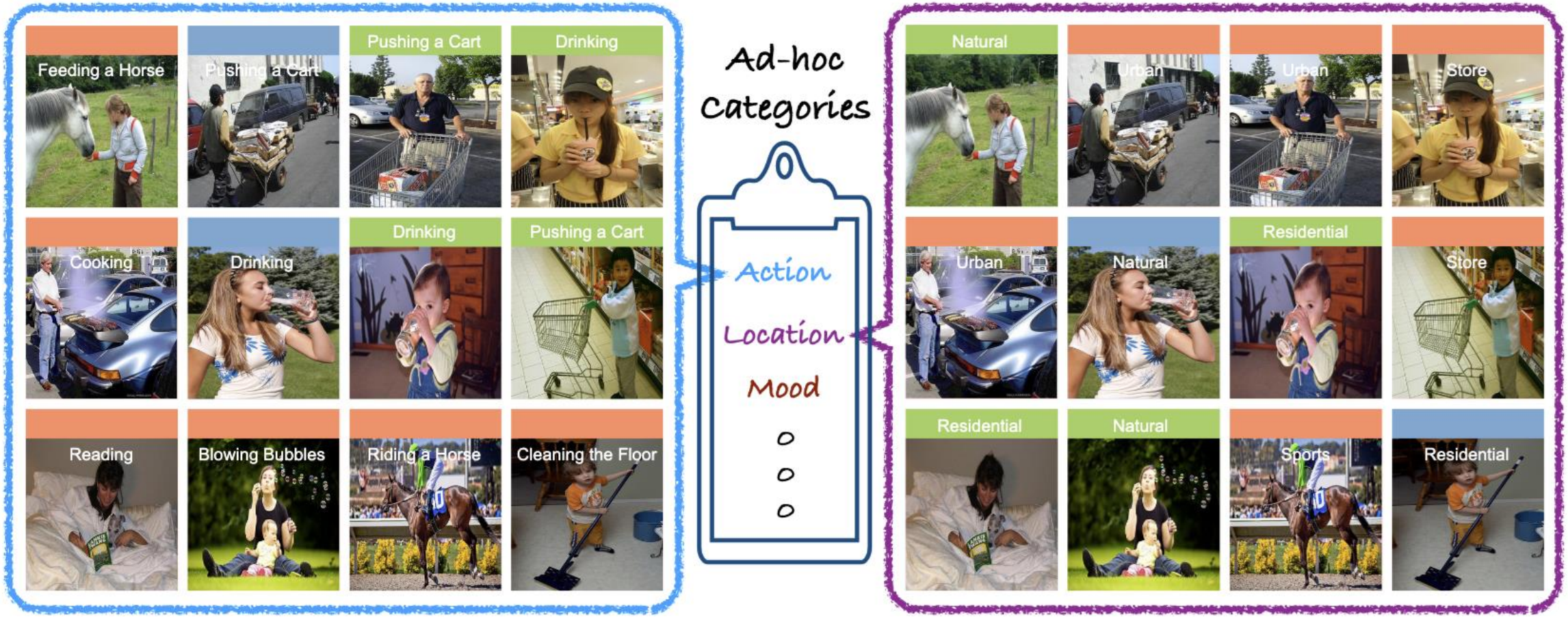
Problem Setup

Labeled



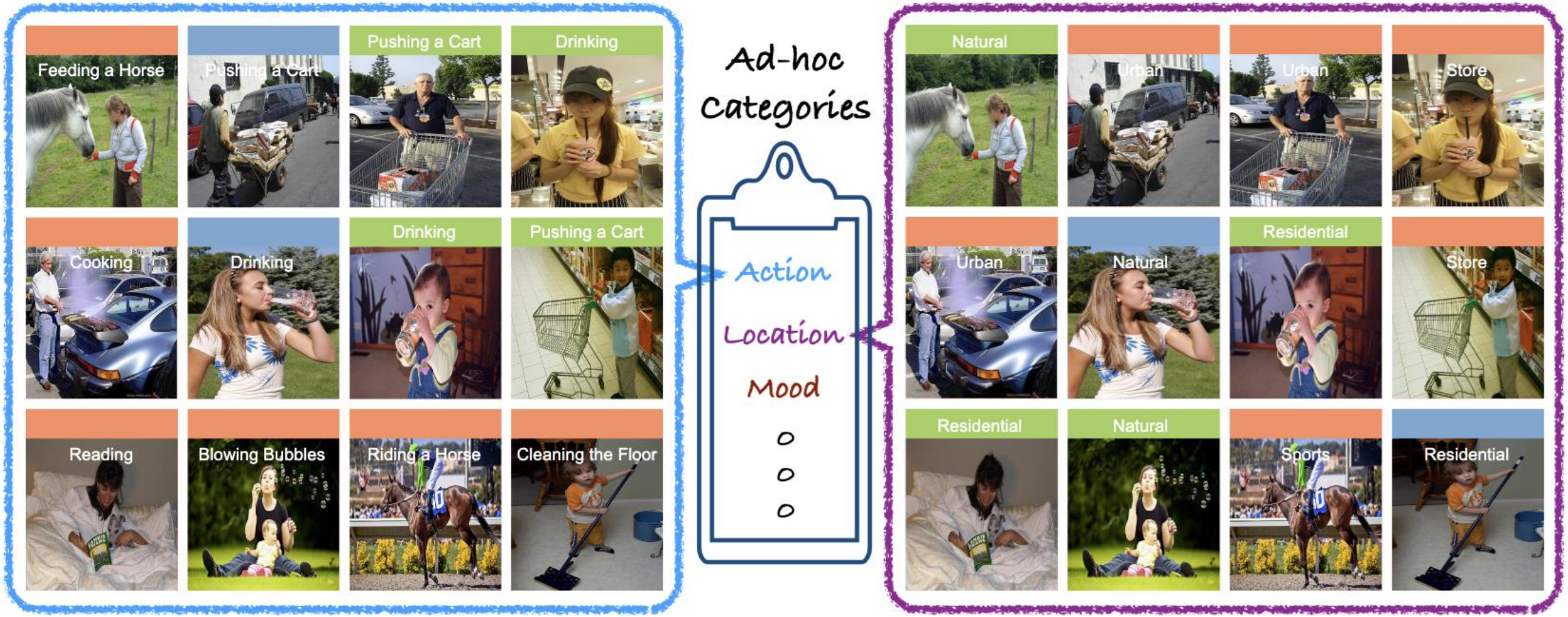
Problem Setup

Labeled
Unlabeled known

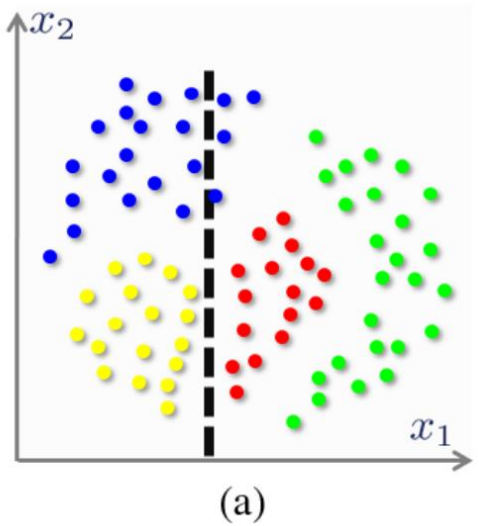


Problem Setup

Labeled
Unlabeled known
Unlabeled unknown

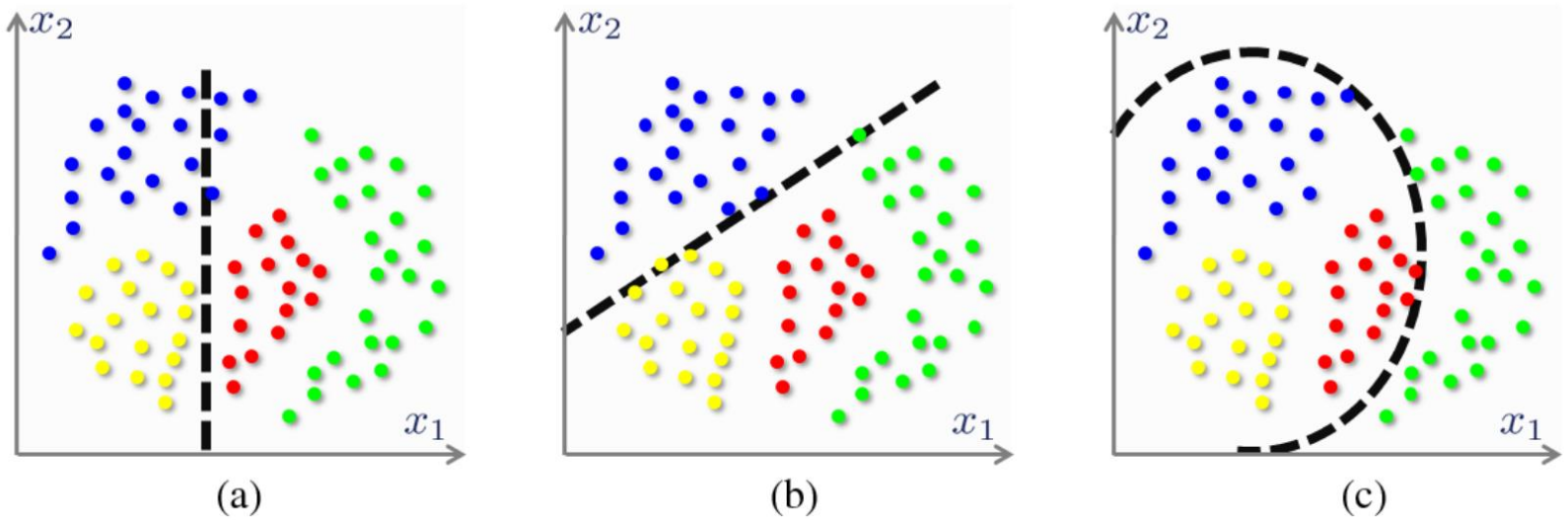


No Single Representation Rules Them All



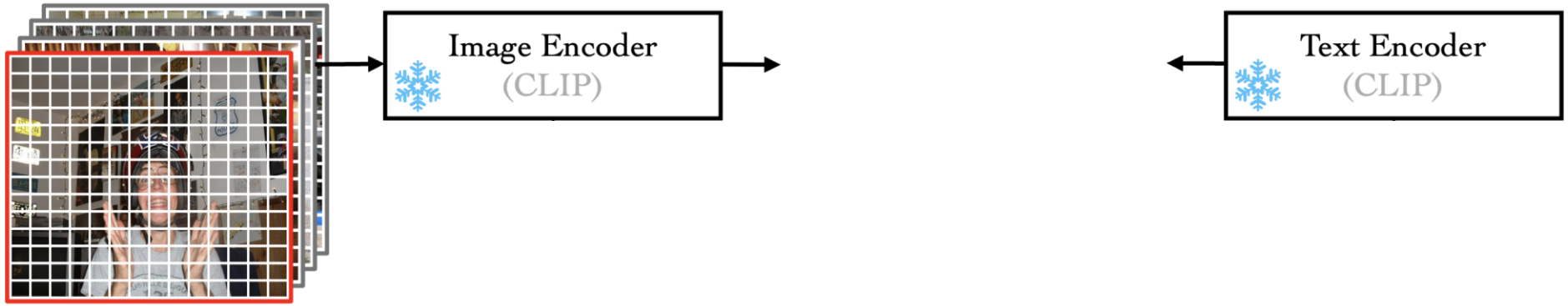
Source: https://datascience.stackexchange.com/questions/100015/orthogonal-decision-boundary-in-decision-tree

No Single Representation Rules Them All



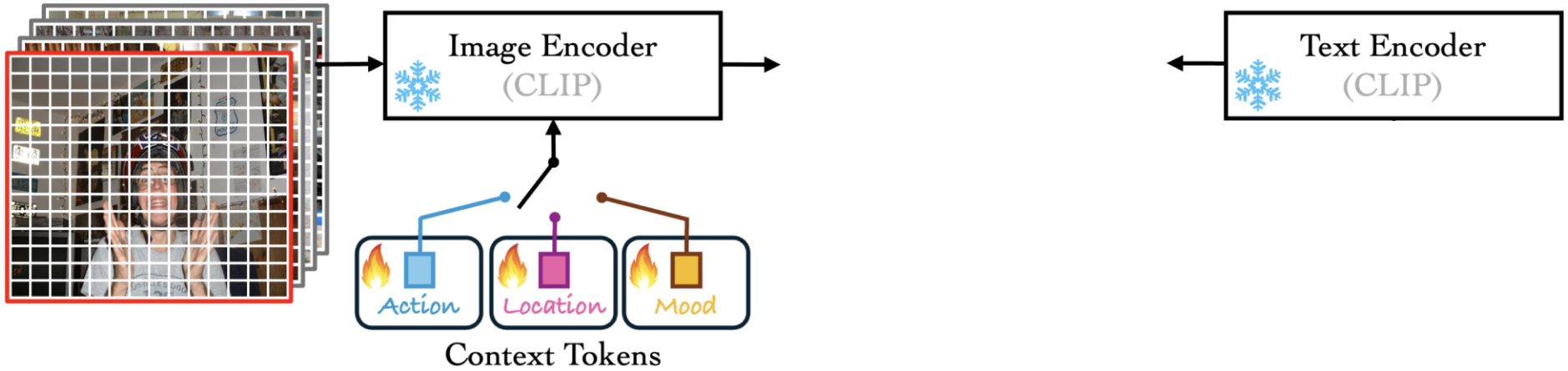
Source: https://datascience.stackexchange.com/questions/100015/orthogonal-decision-boundary-in-decision-tree

Our OAK Framework



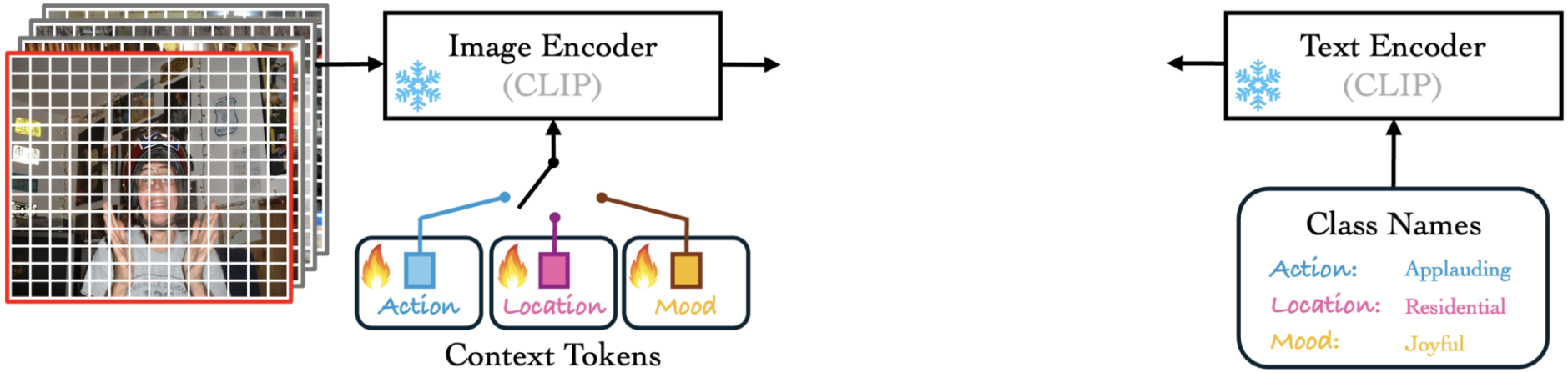
1. **Build on CLIP** to leverage its general knowledge

Our OAK Framework



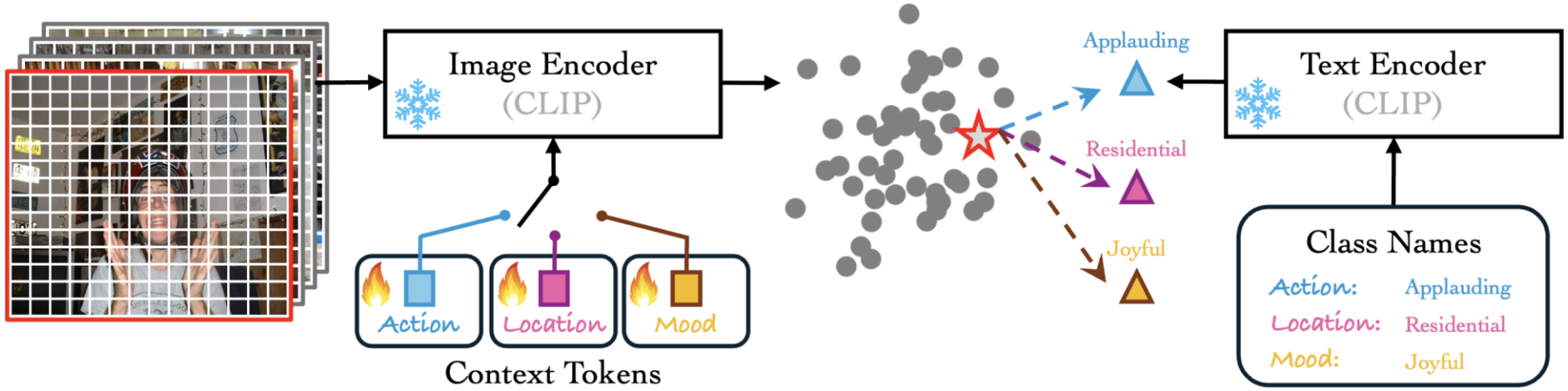
1. Build on CLIP to leverage its general knowledge
2. Tune only the **context token** to adapt visual features

Our OAK Framework



1. Build on CLIP to leverage its general knowledge
2. Tune only the **context token** to adapt visual features
3. Top-down guidance from **semantic cues**

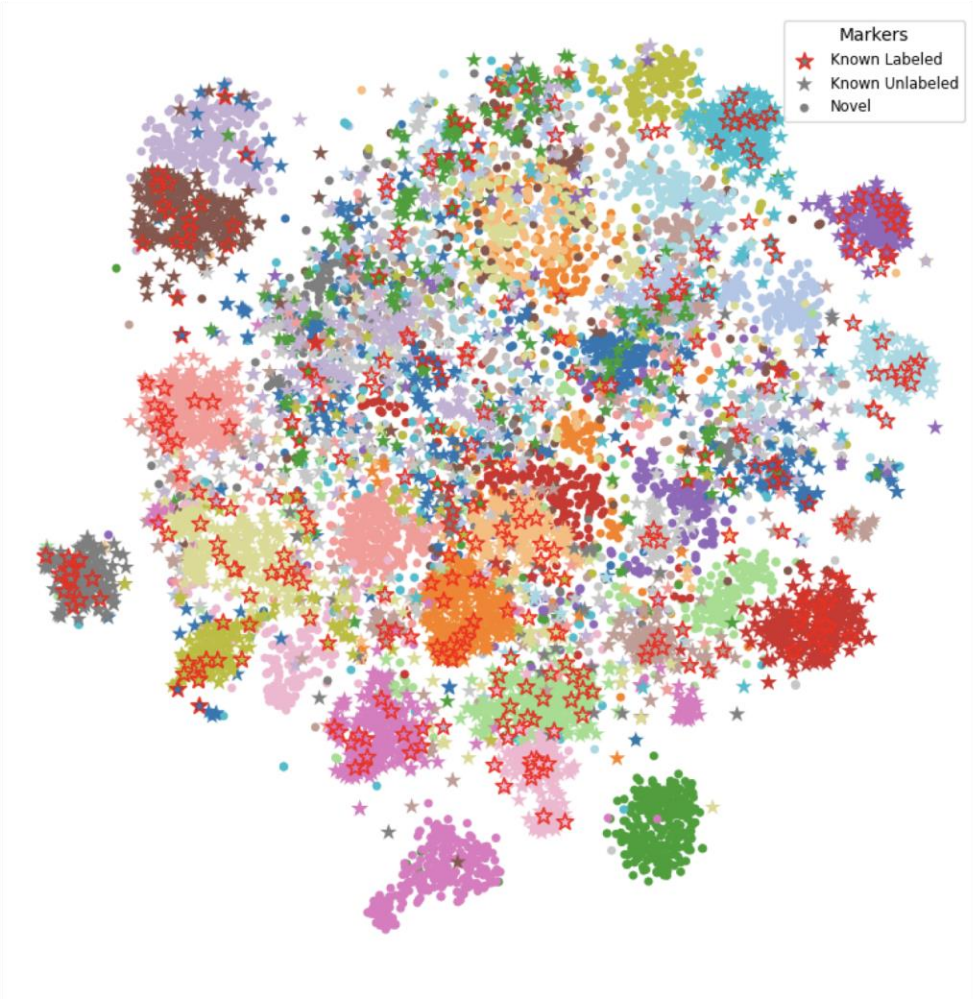
Our OAK Framework



1. Build on CLIP to leverage its general knowledge
2. Tune only the context token to adapt visual features
3. Top-down guidance from semantic cues
4. Bottom-up clustering of **visual cues** → novel clusters

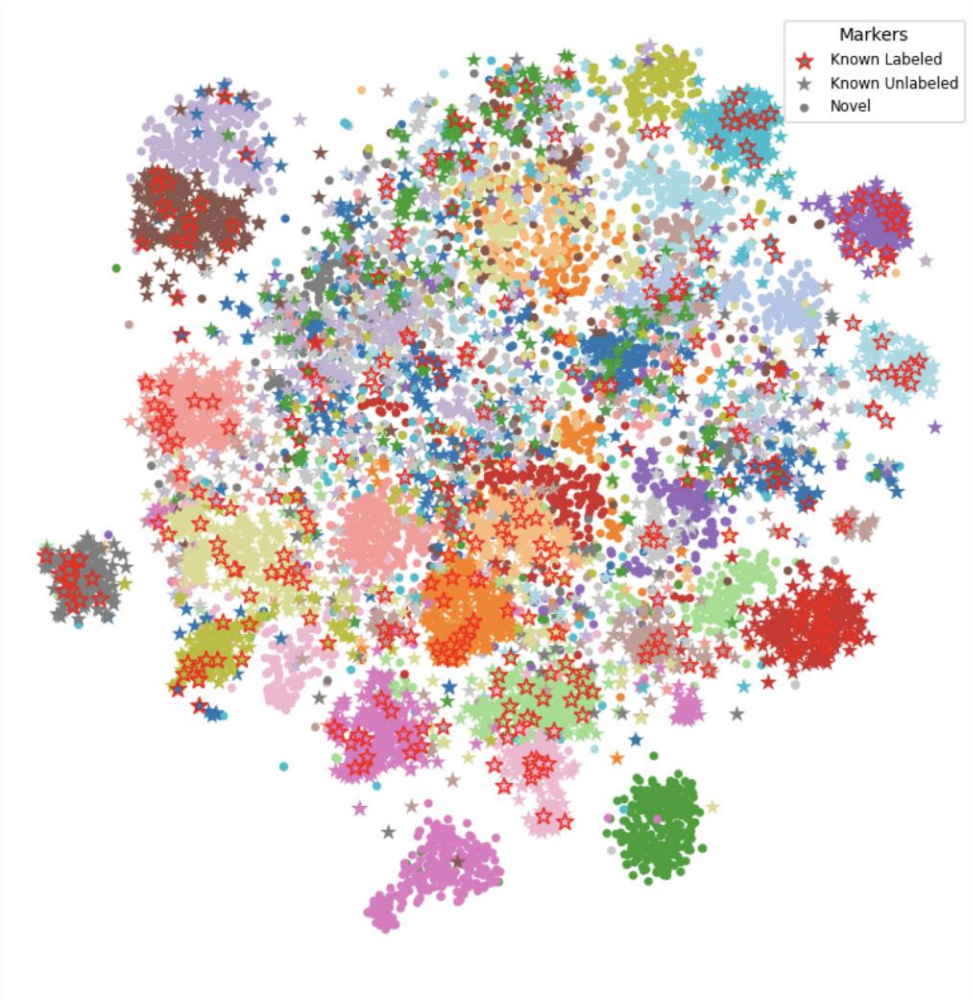
OAK **Contextualizes** Visual Features

Before (CLIP)

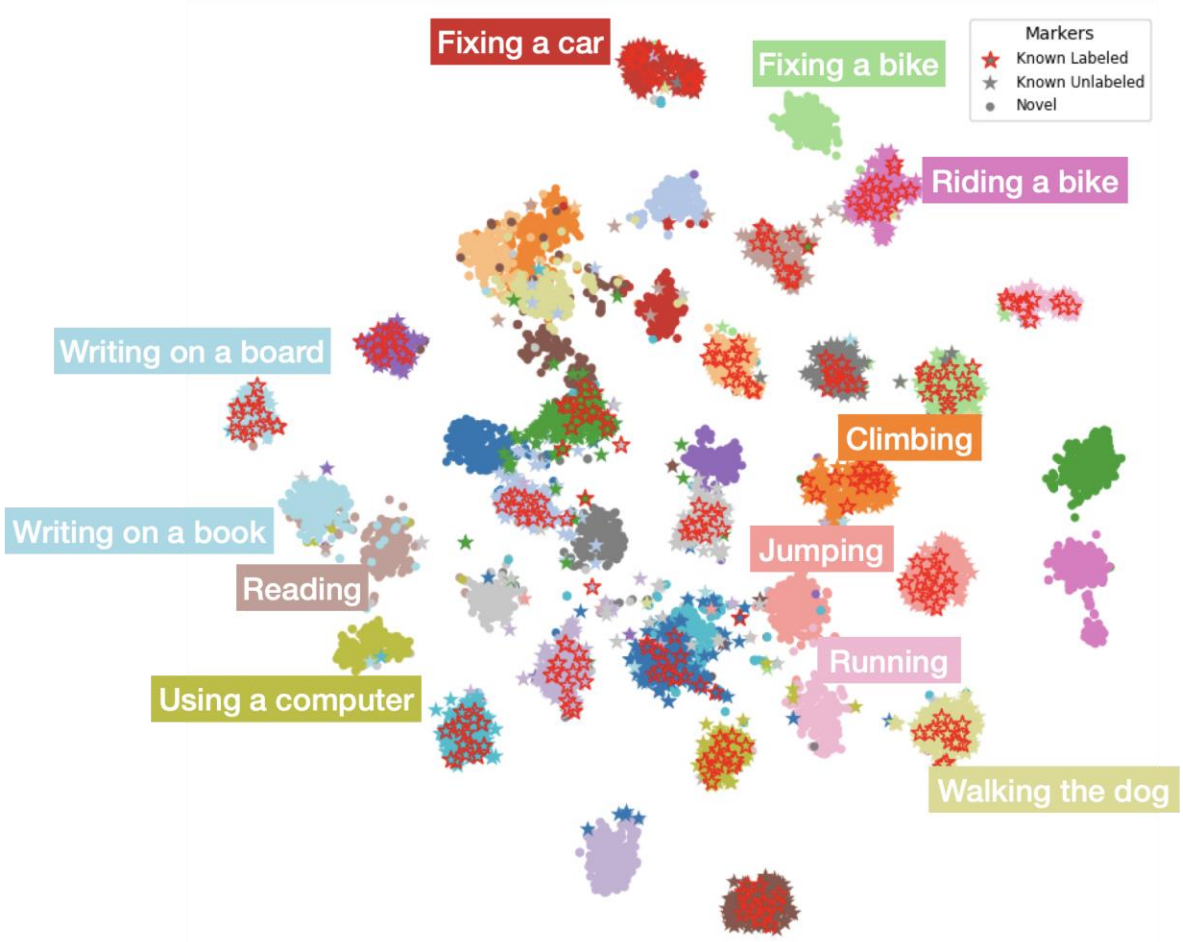


OAK **Contextualizes** Visual Features

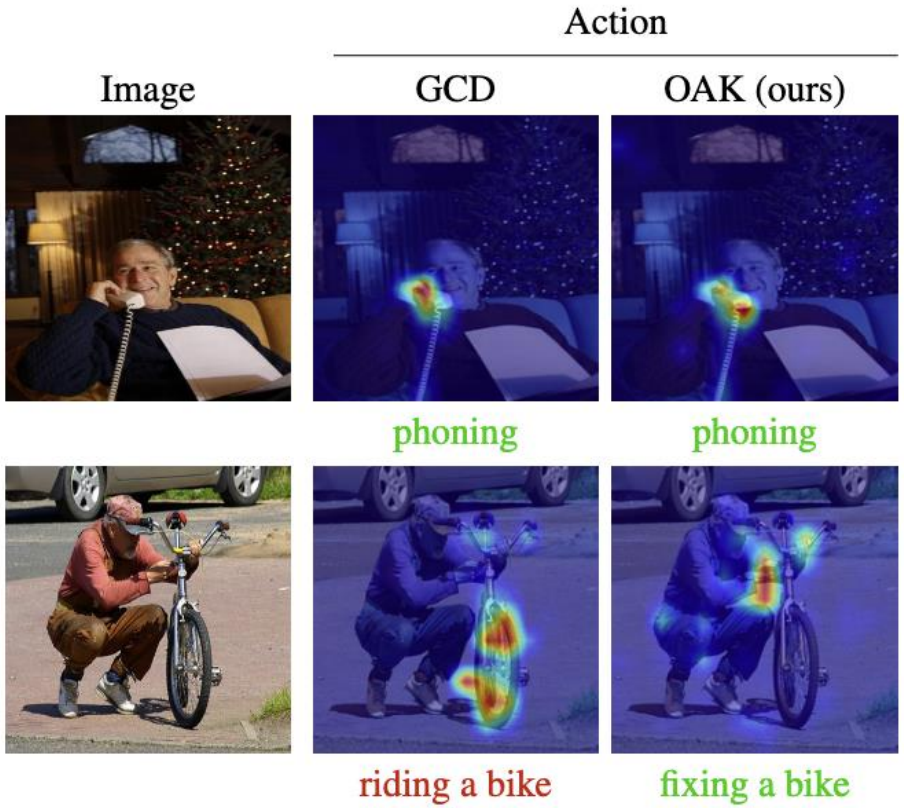
Before (CLIP)



After (OAK)

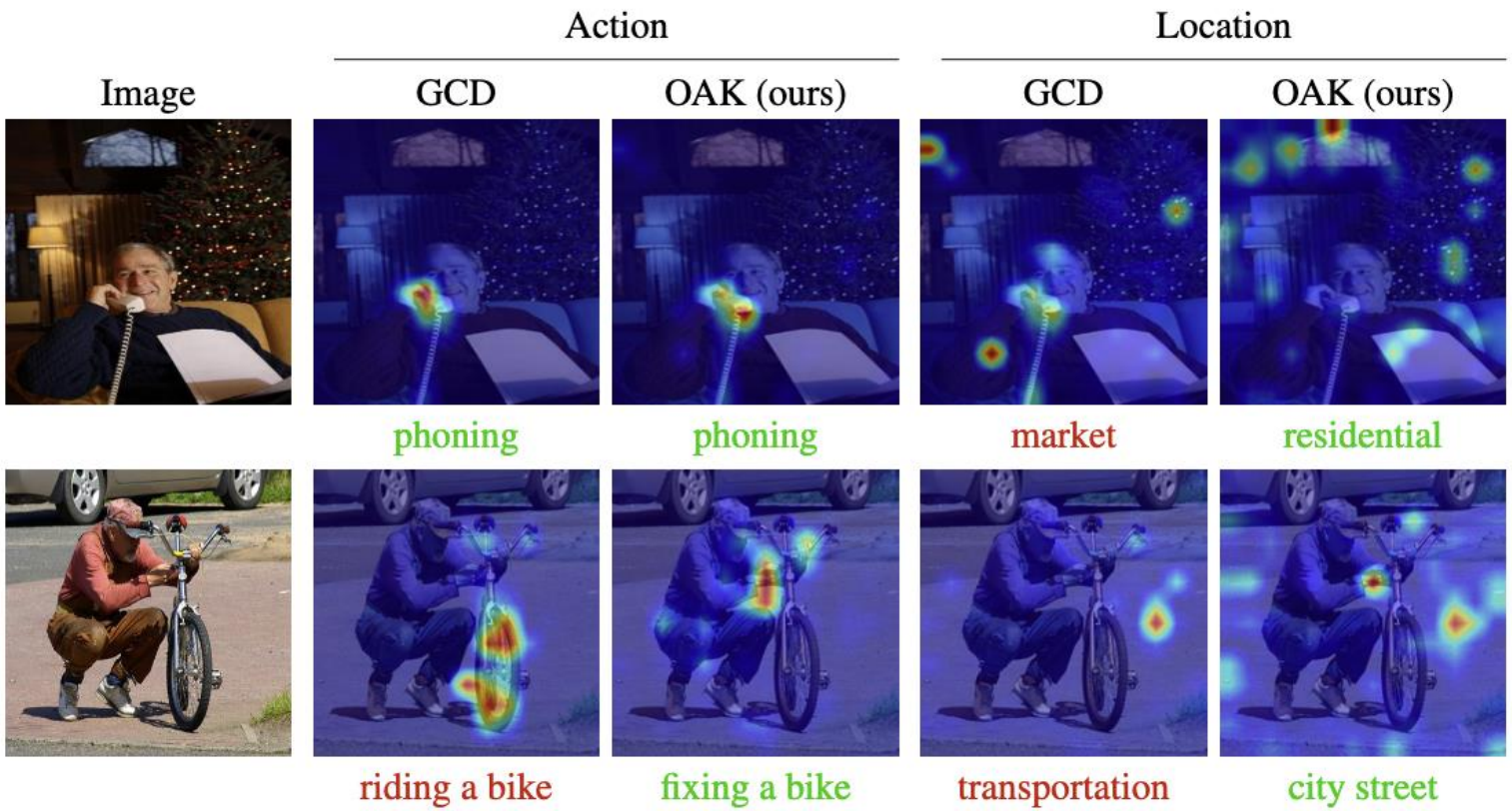


OAK Contextualizes **Visual Attention**



Focus on
hands

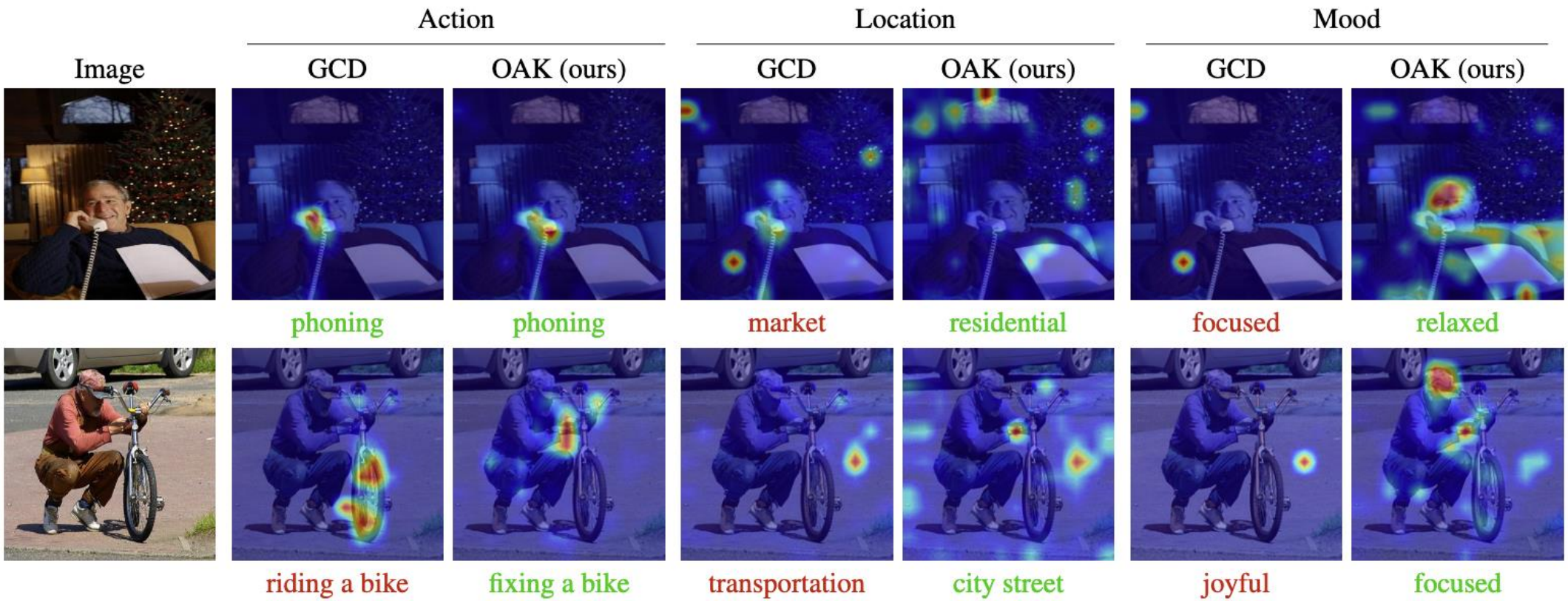
OAK Contextualizes **Visual Attention**



Focus on hands

Focus on **background**

OAK Contextualizes **Visual Attention**



Semantic & Visual Cues are Both Necessary

Semantic only

Method	Known				Novel				Overall			
	Act	Loc	Mood	Omni[a]	Act	Loc	Mood	Omni	Act	Loc	Mood	Omni
CLIP-ZS	**96.2**	60.4	**87.4**	75.0	-	-	-	-	-	-	-	-
CLIP-ZS + LLM vocab	93.5	58.3	75.9	75.0	38.6	34.2	35.4	0.0	65.2	47.5	55.0	43.0
CLIP-ZS + GT vocab	93.6	59.6	78.8	75.0	80.3	59.7	65.8	29.4	86.7	59.7	72.1	38.3
SS-KMeans	67.0	56.3	43.2	0.0	53.9	71.1	78.2	35.3	60.2	62.9	61.3	47.7
GCD	89.4	75.4	64.3	75.0	67.8	80.8	40.6	0.0	78.3	77.8	52.1	52.3
OAK (ours)	88.9	**83.9**	68.8	0.0	**85.1**	**88.4**	**87.4**	**47.1**	**86.9**	**85.9**	**78.4**	**70.3**

[a] With only 8 images overlapping across all contexts, text is toned down to light gray due to unreliable results.

Semantic & Visual Cues are Both Necessary

Semantic only

Visual only

Method	Known				Novel				Overall			
	Act	Loc	Mood	Omni[a]	Act	Loc	Mood	Omni	Act	Loc	Mood	Omni
CLIP-ZS	**96.2**	60.4	**87.4**	75.0	-	-	-	-	-	-	-	-
CLIP-ZS + LLM vocab	93.5	58.3	75.9	75.0	38.6	34.2	35.4	0.0	65.2	47.5	55.0	43.0
CLIP-ZS + GT vocab	93.6	59.6	78.8	75.0	80.3	59.7	65.8	29.4	86.7	59.7	72.1	38.3
SS-KMeans	67.0	56.3	43.2	0.0	53.9	71.1	78.2	35.3	60.2	62.9	61.3	47.7
GCD	89.4	75.4	64.3	75.0	67.8	80.8	40.6	0.0	78.3	77.8	52.1	52.3
OAK (ours)	88.9	**83.9**	68.8	0.0	**85.1**	**88.4**	**87.4**	**47.1**	**86.9**	**85.9**	**78.4**	**70.3**

[a] With only 8 images overlapping across all contexts, text is toned down to light gray due to unreliable results.

Semantic & Visual Cues are Both Necessary

	Known				Novel				Overall			
Method	Act	Loc	Mood	Omni[a]	Act	Loc	Mood	Omni	Act	Loc	Mood	Omni
Semantic only												
CLIP-ZS	**96.2**	60.4	**87.4**	75.0	-	-	-	-	-	-	-	-
CLIP-ZS + LLM vocab	93.5	58.3	75.9	75.0	38.6	34.2	35.4	0.0	65.2	47.5	55.0	43.0
CLIP-ZS + GT vocab	93.6	59.6	78.8	75.0	80.3	59.7	65.8	29.4	86.7	59.7	72.1	38.3
Visual only												
SS-KMeans	67.0	56.3	43.2	0.0	53.9	71.1	78.2	35.3	60.2	62.9	61.3	47.7
GCD	89.4	75.4	64.3	75.0	67.8	80.8	40.6	0.0	78.3	77.8	52.1	52.3
Use both												
OAK (ours)	88.9	**83.9**	68.8	0.0	**85.1**	**88.4**	**87.4**	**47.1**	**86.9**	**85.9**	**78.4**	**70.3**

[a] With only 8 images overlapping across all contexts, text is toned down to light gray due to unreliable results.

OAK Can **Name** the Discovered Clusters

a) Stanford	Action		Location		Mood	
GT Label	cooking	reading	restaurant or dining area	urban area or city street	focused	relaxed
Prediction	preparing a meal	reading a book	commercial kitchen	suburban street	explorative	admiring

b) Clevr-4	Texture		Color		Shape		Count	
GT Label	zigzag	circles	pink	cyan	star	diamond	4	9
Prediction	wavy lines	checkerboard	pink	turquoise	star shape	diamond shape	4	17

OAK Can **Name** the Discovered Clusters

a) Stanford	Action		Location		Mood	
GT Label	cooking	reading	restaurant or dining area	urban area or city street	focused	relaxed
Prediction	preparing a meal	reading a book	commercial kitchen	suburban street	explorative	admiring

b) Clevr-4	Texture		Color		Shape		Count	
GT Label	zigzag	circles	pink	cyan	star	diamond	4	9
Prediction	wavy lines	checkerboard	pink	turquoise	star shape	diamond shape	4	17

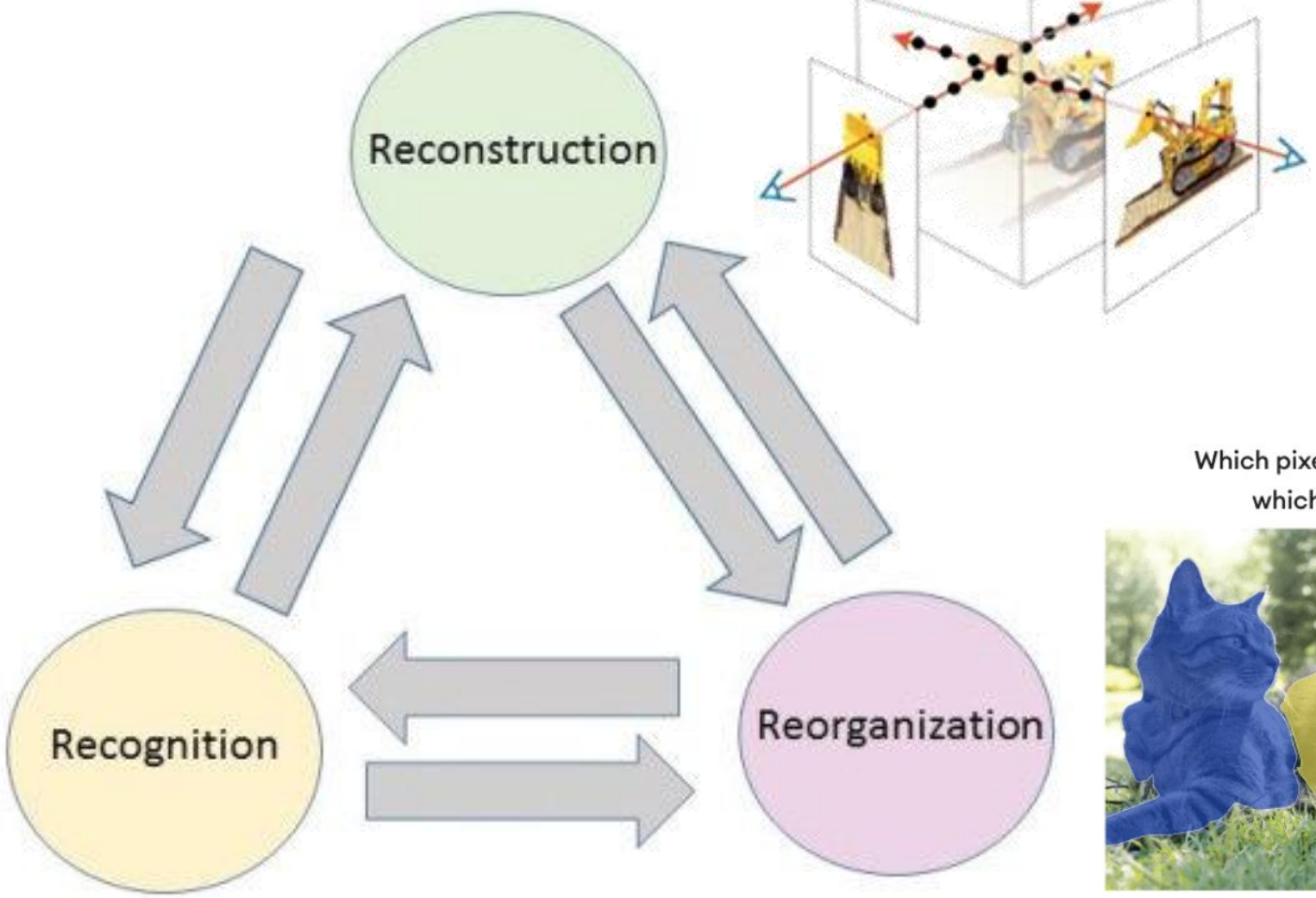
jumping named as *dancing*



pouring liquid named as *carrying a box*



3Rs in Computer Vision



Is this a cat?

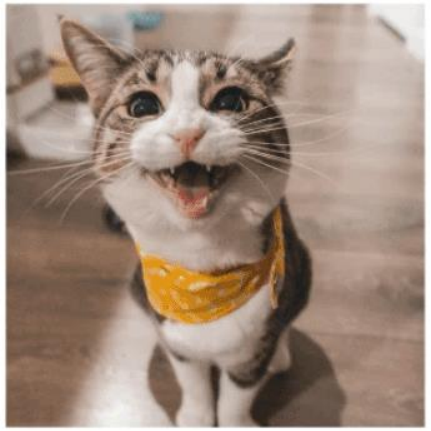


Image Classification

Which pixels belong to which object



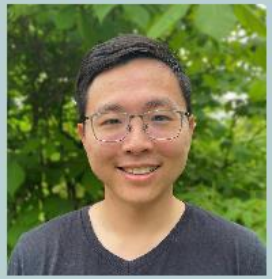
Image Segmentation

Q. Do Foundation Models Capture the 3Rs?



ICCV 2025 SP4V Workshop

Keynote Speakers



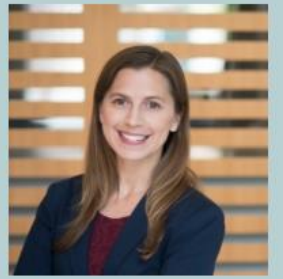
Danfei Xu
Georgia Tech & NVIDIA



João Carreira
Google DeepMind



Jiajun Wu
Stanford



Kristen Grauman
UT Austin



Saining Xie
NYU



Vincent Sitzmann
MIT

Organizers



Sangwoo Mo
UMich



Congyue Deng
Stanford



Hila Chefer
Black Forest Labs



Daniel Zoran
Google DeepMind



Kaichun Mo
NVIDIA



Leonidas Guibas
Stanford



Stella Yu
UMich

Top