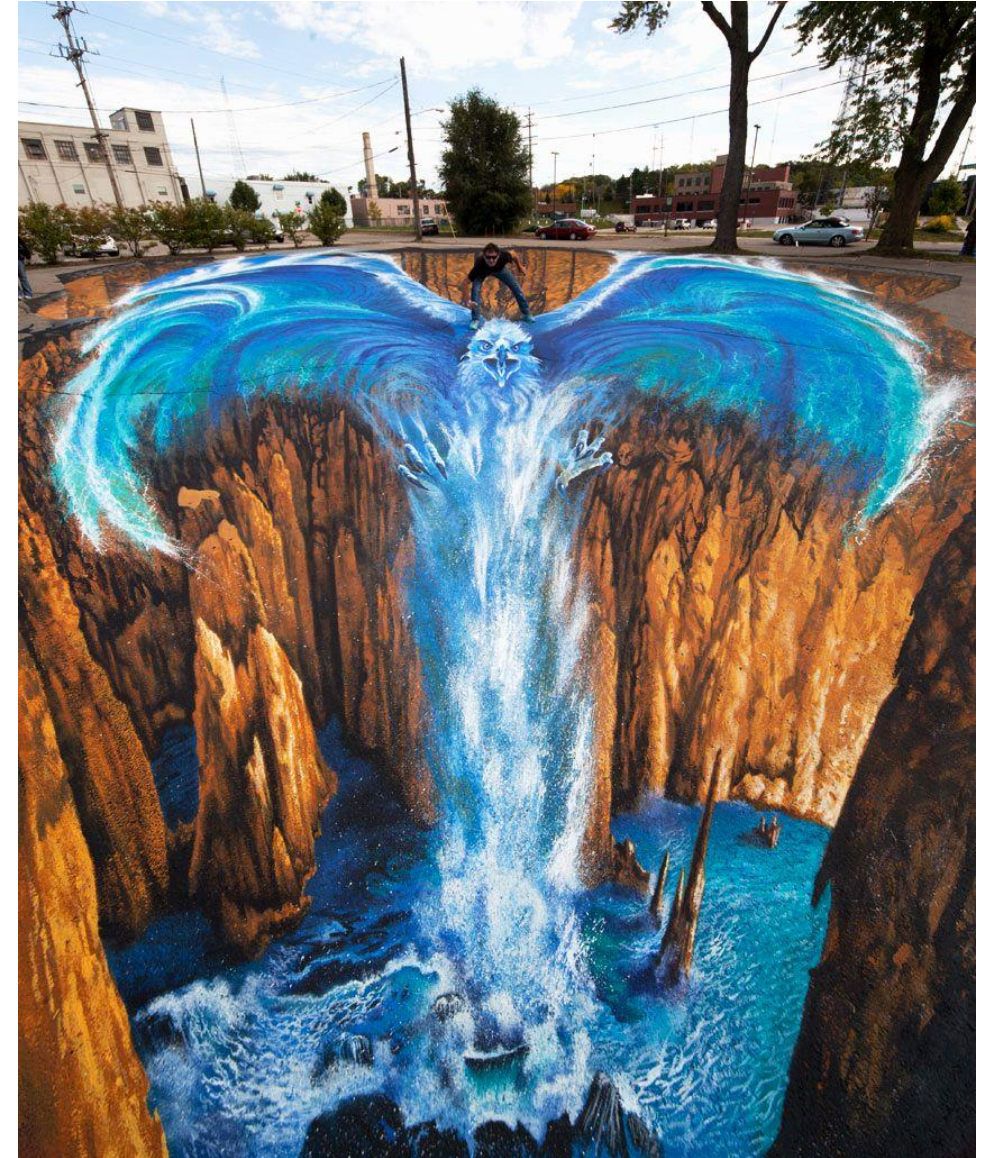


# Towards 3D-Aware 2D Models: Visual Understanding in the 3D World

presented by Boshen Xu

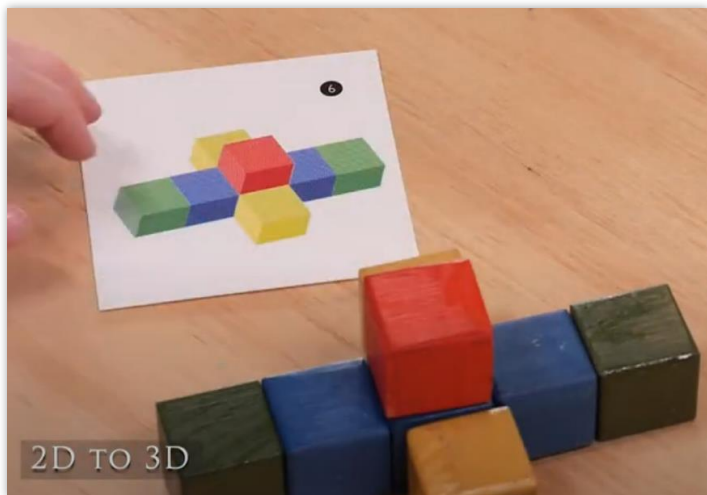
2024.12.18

# We **Hallucinate** 3D Information Due to Our **Knowledge**



# We Live in a 3D World

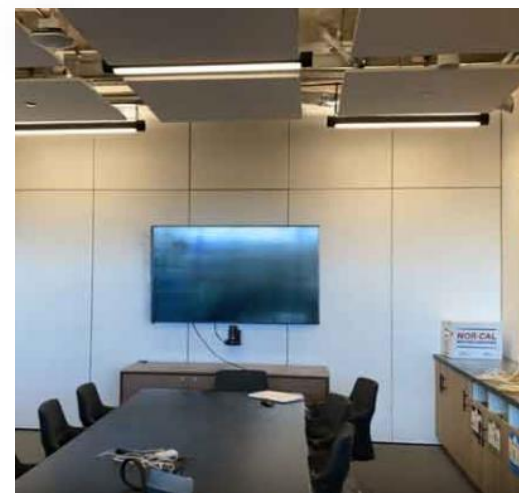
Through **embodied experiences**, we humans excel in understanding **3D spatial relations**, as well as **interacting with the world with 2D vision**.



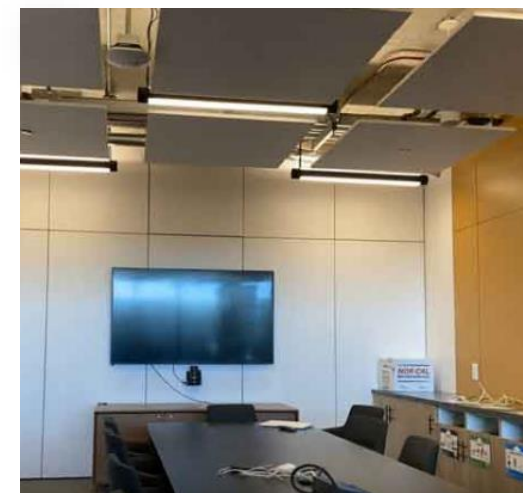
2D-to-3D Imagination



Spatial Sensing



Multi-view Consistency



**Mimicing this paradigm, how to develop 3D-aware 2D models?**

What is 3D data?

What type of 3D-awareness is desirable?

What problem can 3D-awareness solve?

# 3D Data Types

## Raw Data

(From Camera / Stereo / Laser...)

Point Cloud:  $N \times 3$

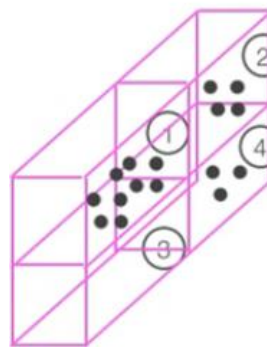


RGBD Img



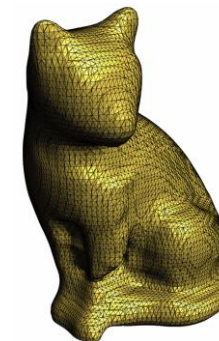
## 3D Data Representations

Voxels:  $D \times H \times W$



Mesh:

(Vertices, Edges, Faces)



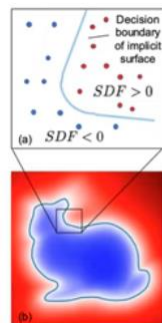
Multi-View Imgs



Video



SDF:  $f: (x, y, z) \rightarrow \mathbb{R}$     NeRF:  $f: (x, y, z, d) \rightarrow c, \sigma$



# 3D Awareness in 2D Representation Learning

Raw Data

(From Camera / Stereo / Laser...)

Point Cloud:  $N \times 3$



RGBD Img



Multi-View Imgs

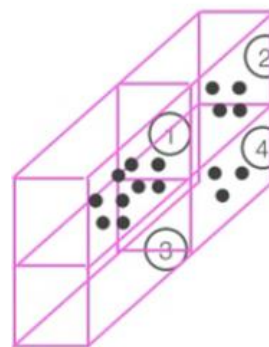


Video



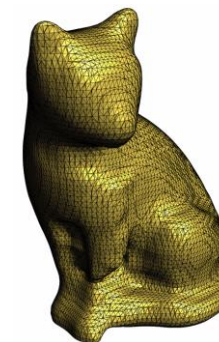
3D Data Representations

Voxels:  $D \times H \times W$

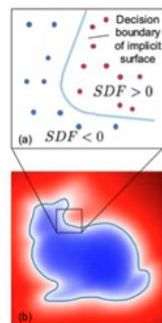


Mesh:

(Vertices, Edges, Faces)



SDF:  $f: (x, y, z) \rightarrow \mathbb{R}$     NeRF:  $f: (x, y, z, d) \rightarrow c, \sigma$

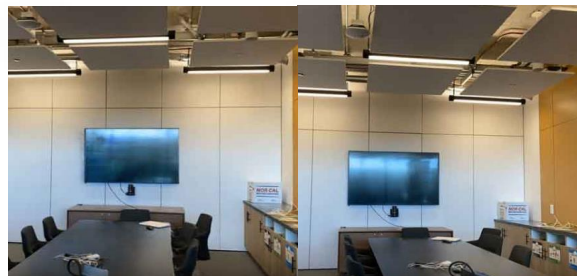


# What Problems Can 3D-Awareness Solve?

- 3D-aware **image models**: Know better about the "Real 3D World" and the "Camera Pose".



Imagination/3D from single-view  
(CAT3D, NeurIPS23 oral,  
DepthAnything, CVPR24)



Multi-view consistency  
(Lift3D, CVPR24,  
CONDENSE, ECCV24 oral)

- 3D-aware **video models**: Know better about the "Motion", and "Physics"



Better video understanding  
(SpatialTracker, CVPR24 Highlight)



Better camera & content movement control  
(3DTrajMaster, arxiv24)



# Recent Advances in 3D-Aware 2D Models

## Text

Q: How far is [A] from [B]?

A: It's [Distance].

SpatialVLM, CVPR24  
SpatialRGPT, NeurIPS24

## RGBD Img



DepthAnything, CVPR24  
DepthAnythingv2, NeurIPS24  
Metric3D, ICCV2023  
Metric3Dv2, T-PAMI24

## Multi-View Imgs



FiT3D, ECCV24  
ConDense, ECCV24 oral  
3DCorrEnhance, arxiv24  
CAT3D, NeurIPS24 oral

## Video



See3D, arxiv24  
CAT4D, arxiv24

# Recent Advances in 3D-Aware 2D Models

## Text

Q: How far is [A] from [B]?

A: It's [Distance].

SpatialVLM, CVPR24  
SpatialRGPT, NeurIPS24

## RGBD Img



DepthAnything, CVPR24  
DepthAnythingv2, NeurIPS24  
Metric3D, ICCV2023  
Metric3Dv2, T-PAMI24

## Multi-View Imgs



FiT3D, ECCV24  
ConDense, ECCV24 oral  
3DCorrEnhance, arxiv24  
CAT3D, NeurIPS24 oral

## Video



See3D, arxiv24  
CAT4D, arxiv24

# SpatialVLM: Endowing V-LLM with Spatial Reasoning

User



You are a cleaning robot that is 1 meter wide, now you are standing in a living room and see the image, you want to move from here to the door that leads to the backyard, do you think I can go through the path between the sofa and the table and chairs.



GPT-4V\*



As an AI, I'm unable to physically interact with environments, but I can provide some insights based on the image provided. ... From the image, we can make some visual estimations. Judging from the image, it appears that the sofa and the table are quite close to each other, and there may not be enough space to accommodate a 1-meter wide robot....

Spatial-VLM



[VLM Reasoning] How wide is the path between the sofa and the table and chairs? [VLM Answer] 1.56m [Answer] Yes, the robot can go through the path between the sofa and the table and chairs since it is wider than the robot's width.

User



Which can is the closest to the yellow finger that is on the right most side of the picture, the blue, orange or the silver one?



GPT-4V



After estimating the distances by measuring the Euclidean distances between the yellow finger and each of the cans, it is clear that the silver can is the closest to the yellow finger with an approximate distance of 110 units. In comparison, the blue can is about 262 units away, and the orange can is approximately 233 units away from the yellow finger.

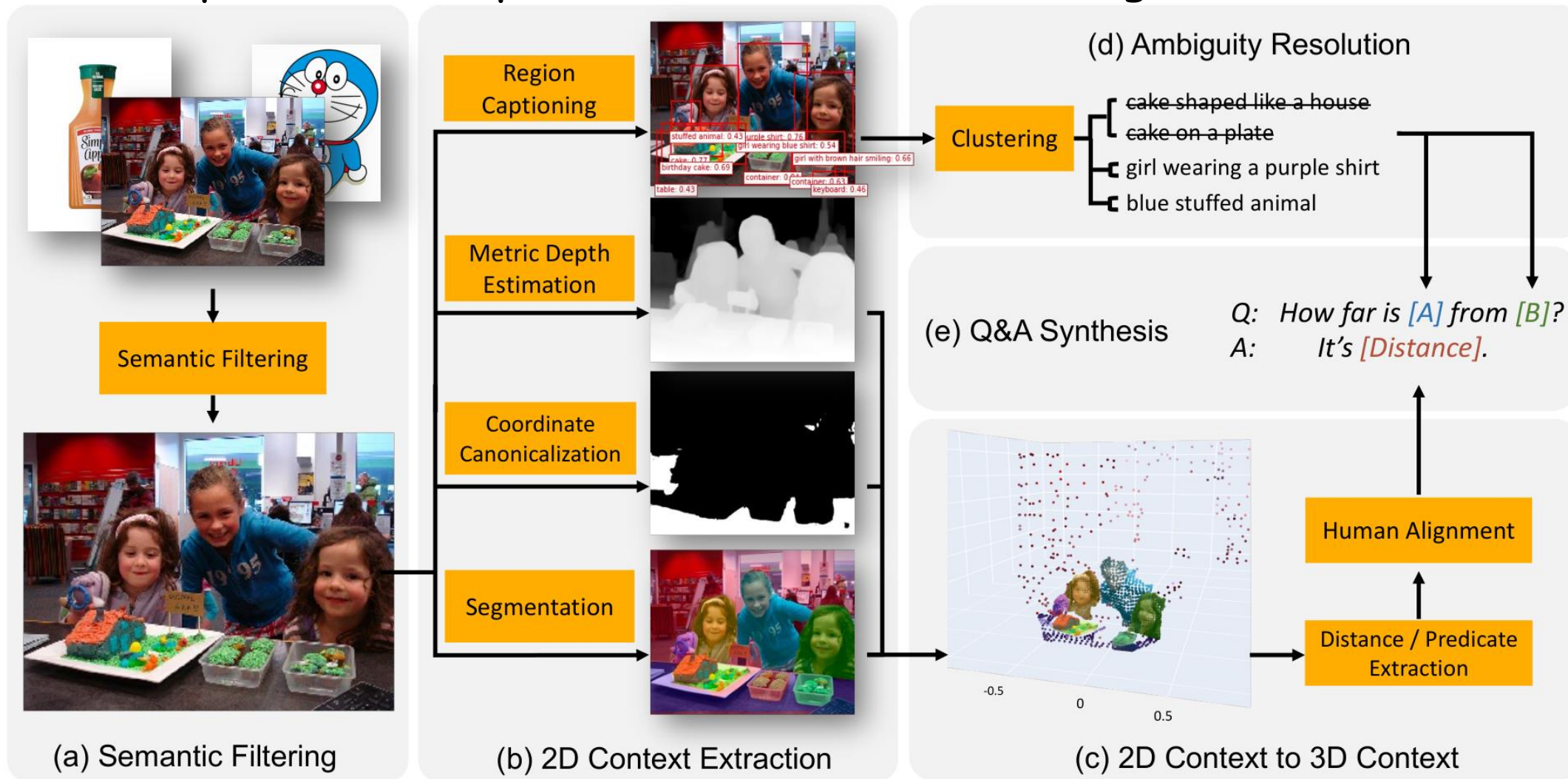
Spatial-VLM



[VLM Reasoning] How far is the yellow finger from the blue can? [VLM Answer] 0.5m [VLM Reasoning] How far is the yellow finger from the orange can? [VLM Answer] 0.59m [VLM Reasoning] How far is the yellow finger from the silver can? [VLM Answer] 0.21m [Answer] The silver can is the closest to the yellow finger that is on the right most side of the picture.

# SpatialVLM: Endowing V-LLM with Spatial Reasoning

- 2B template-based spatial VQA on 10M indoor imgs



# SpatialVLM: Endowing V-LLM with Spatial Reasoning



## Qualitative Q & A

Find out if the girl wearing a purple shirt is situated **behind or in front of** the girl wearing a blue sweatshirt.      It's the woman wearing white shirt.

## Quantitative Q & A

Measure the distance from the cake shaped like a house to the girl wearing a purple shirt?      Around **half a meter**

- Accuracy of VLMs on binary predicate prediction tasks (Qualitative)

Method	GPT-4V	LLaVA-1.5	InstructBLIP	PaLI	PaLM-E	PaLM 2-E	Ours
Accuracy	68.0%	71.3%	60.4%	60.7%	50.2%	50.4%	<b>75.2%</b>

- Accuracy of VLMs on quantitative questions about spatial relationship (Quantitative)

	GPT-4V	LLaVA-1.5	InstructBLIP	PaLI	PaLM-E	PaLM 2-E	Ours
Output numbers %	1.0%	20.9%	26.0%	52.0%	83.2%	88.8%	<b>99.0%</b>
In range [50, 200]%	0.0%	13.0%	7.9%	5.3%	23.7%	33.9%	<b>37.2%</b>

# Recent Advances in 3D-Aware 2D Models

## Text

Q: How far is [A] from [B]?

A: It's [Distance].

SpatialVLM, CVPR24  
SpatialRGPT, NeurIPS24

## RGBD Img



DepthAnything, CVPR24  
DepthAnythingv2, NeurIPS24  
Metric3D, ICCV2023  
Metric3Dv2, T-PAMI24

## Multi-View Imgs



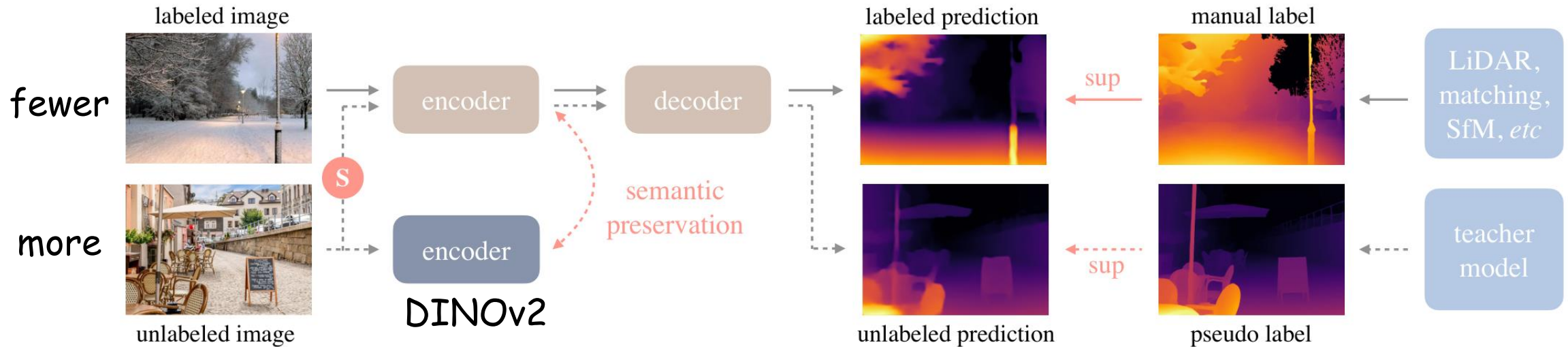
FiT3D, ECCV24  
ConDense, ECCV24 oral  
3DCorrEnhance, arxiv24  
CAT3D, NeurIPS24 oral

## Video



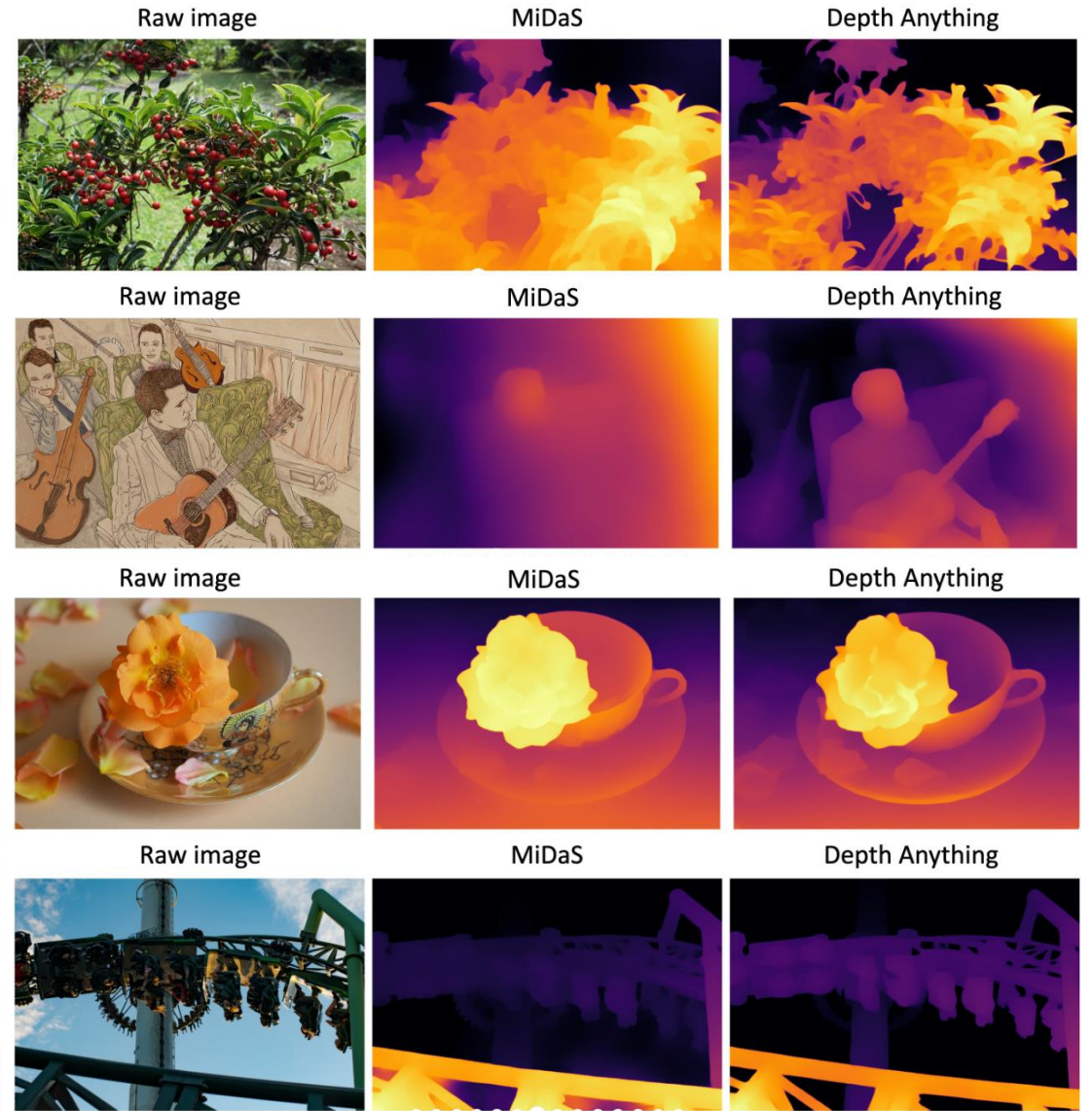
See3D, arxiv24  
CAT4D, arxiv24

# DepthAnything: the Power of Large-Scale Unlabeled Data

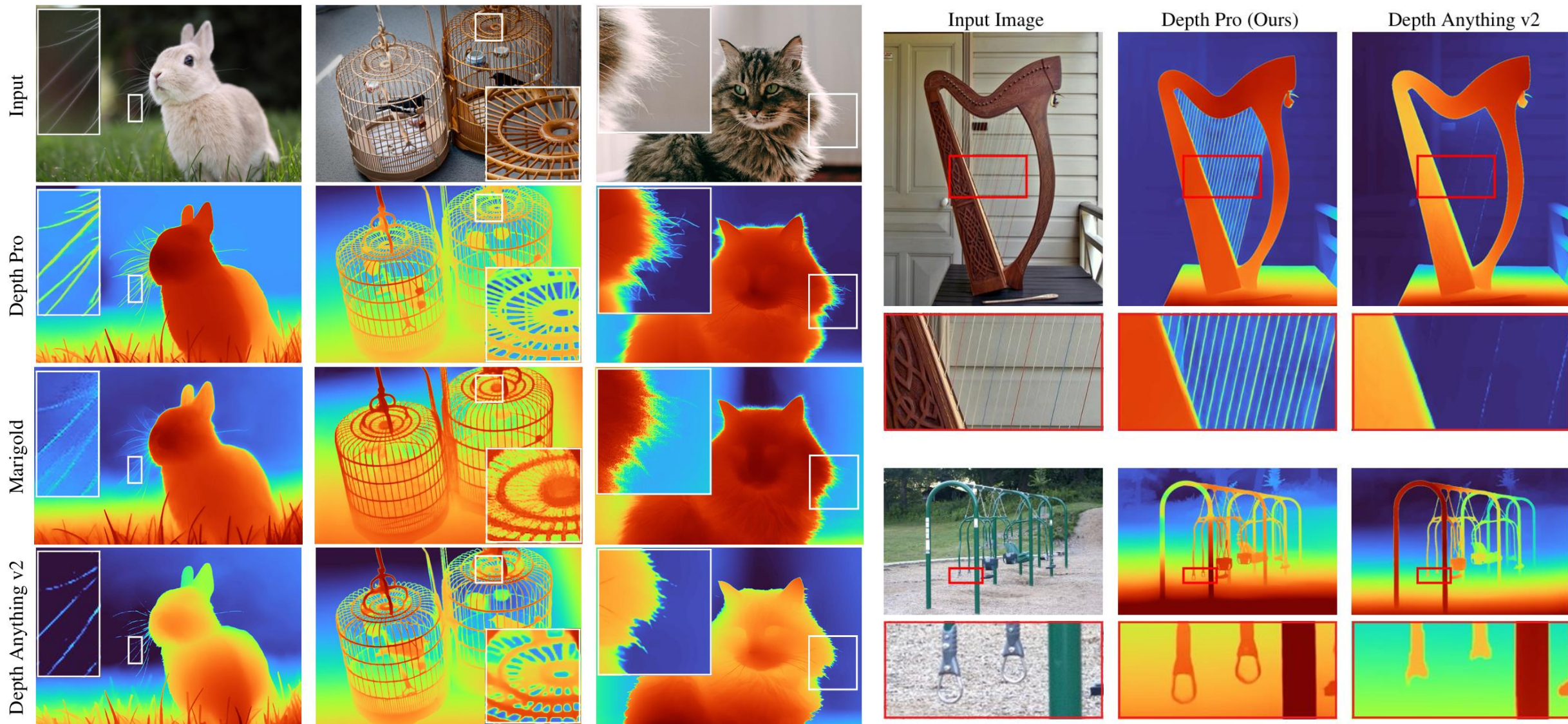


- Labeled image: (1) train a teacher model; (2) basic ability for encoder
- Unlabeled image: (1) providing semantic perception for encoder; (2) Distilling knowledge from DINOv2

# DepthAnything: the Power of Large-Scale Unlabeled Data



# Some State-of-the-art Depth Estimation Demo



# Recent Advances in 3D-Aware 2D Models

## Text

Q: How far is [A] from [B]?  
A: It's [Distance].

SpatialVLM, CVPR24  
SpatialRGPT, NeurIPS24

## RGBD Img



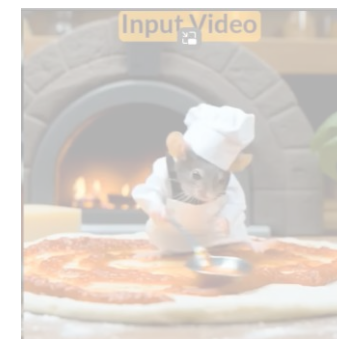
DepthAnything, CVPR24  
DepthAnythingv2, NeurIPS24  
Metric3D, ICCV2023  
Metric3Dv2, T-PAMI24

## Multi-View Imgs



FiT3D, ECCV24  
ConDense, ECCV24 oral  
3DCorrEnhance, arxiv24  
CAT3D, NeurIPS24 oral

## Video



See3D, arxiv24  
CAT4D, arxiv24

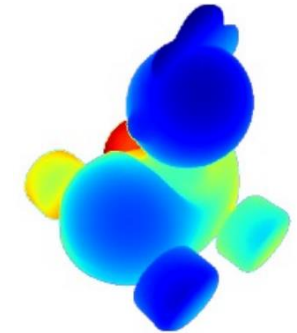
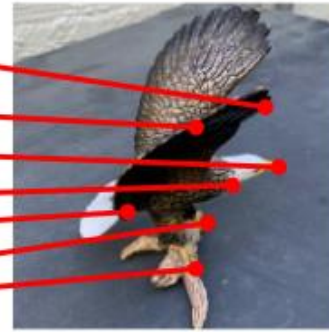
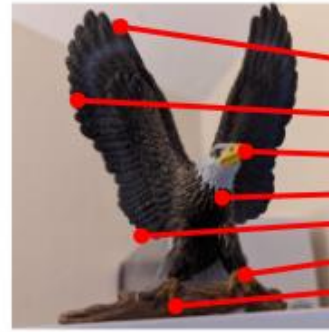
# Probe3D: 3D Perception Evaluation of Foundation Models

- Single-view (depth & surface normal)

- Multi-view (consistency)

SV Evaluation: Probing

Object



decoder

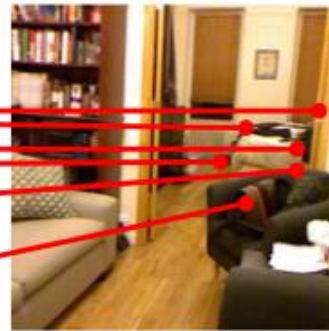
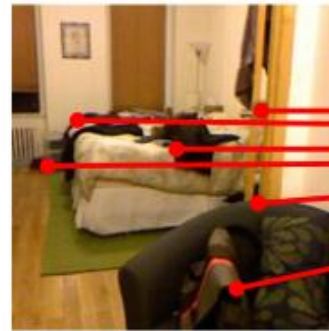
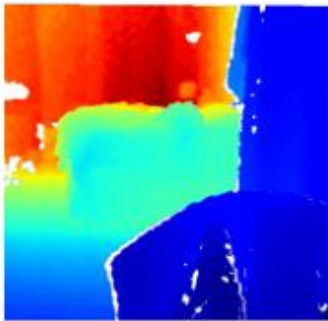
multi-scale

encoder

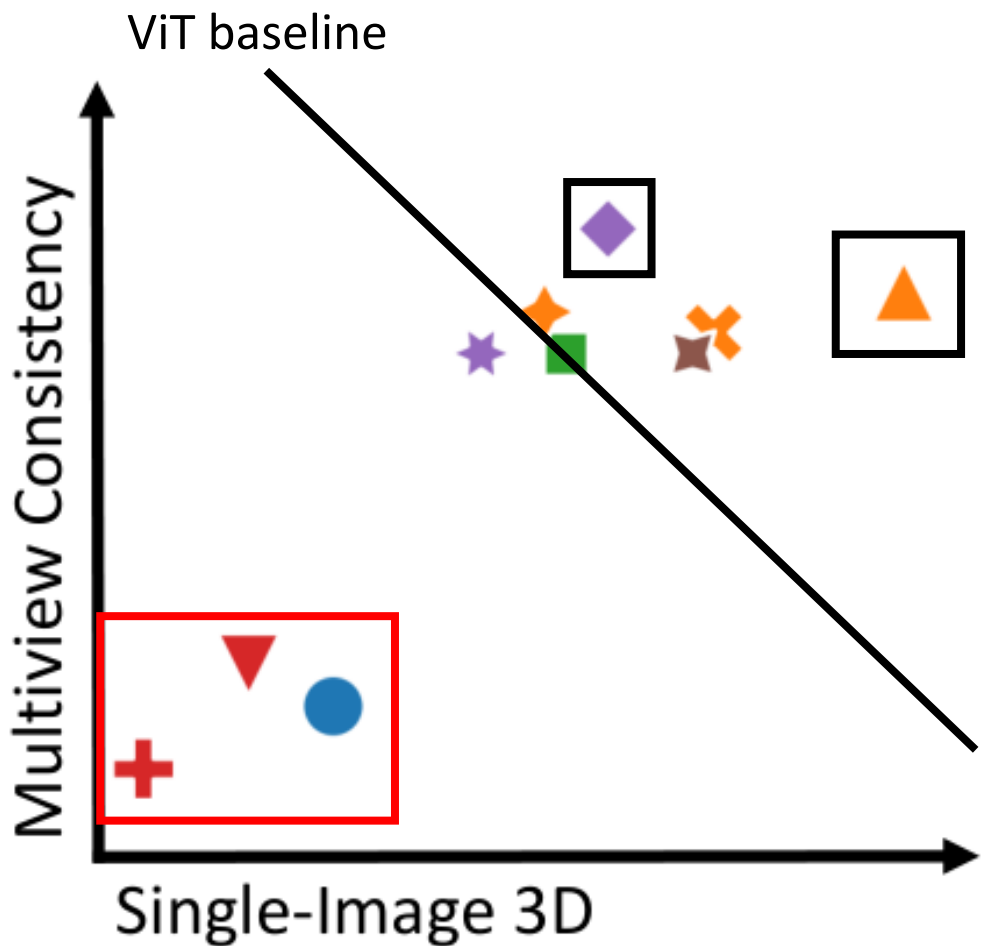
Image



Scene



# Probe3D: 3D Perception Evaluation of Foundation Models



## Evaluated Models

● MAE

◆ DINO

✕ iBOT

▲ DINOv2

■ DeiT III

✖ StableDiffusion

⊕ CLIP

▼ SigLIP

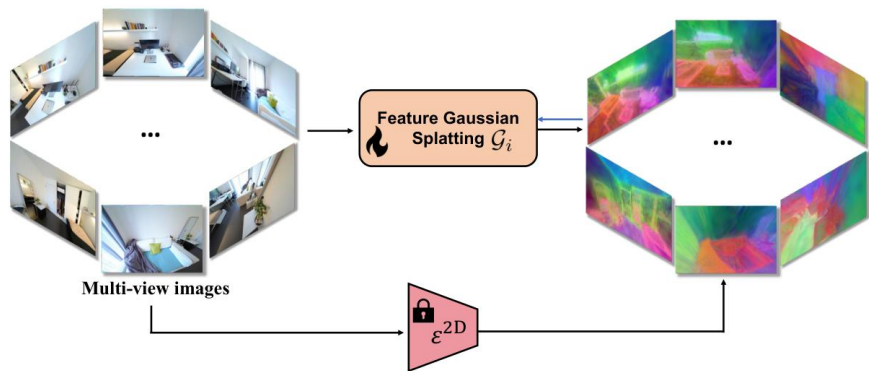
◆ MiDaS

★ SAM

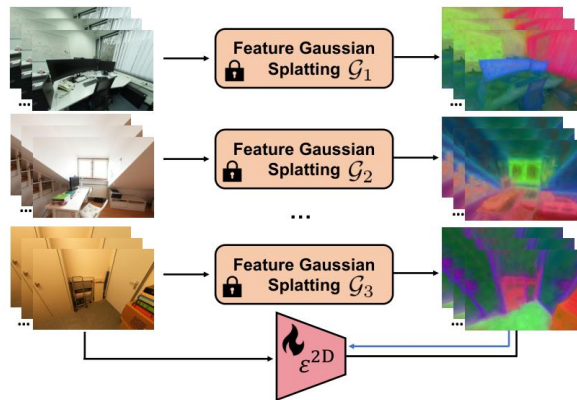
**How to develop 3D-awareness for 2D foundation models?**

# FiT3D: Finetuning DINOv2 by Multi-View Images

Stage I: Lifting Features to 3D (Per-scene)

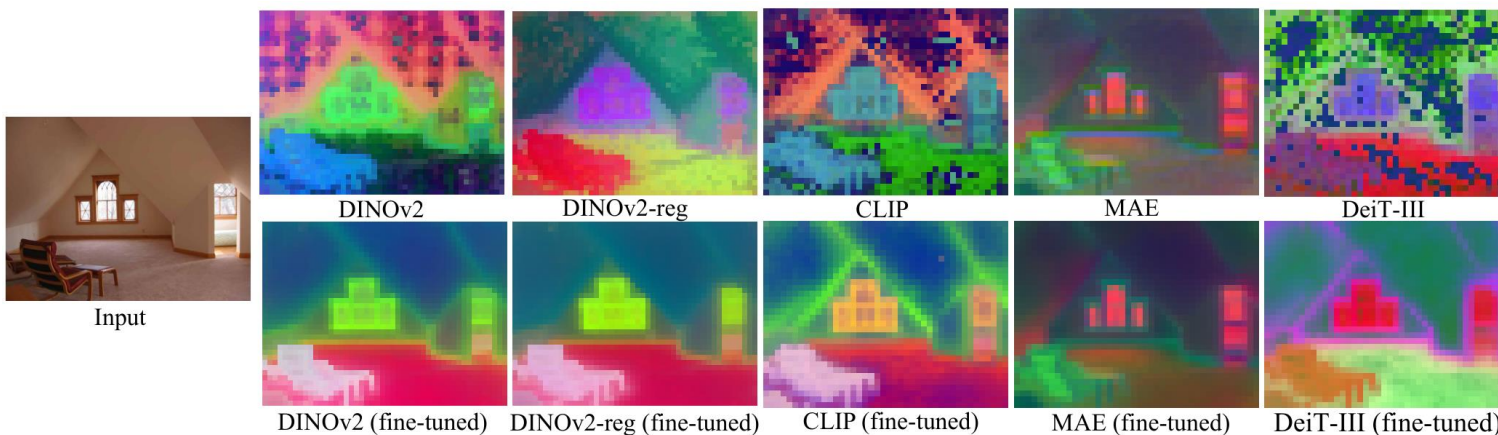


Stage II: 3D-Aware Fine-Tuning (Multi-scene)



More efficient & effective

- Model: DINOv2-small
- Training: ScanNet++, 230 scenes w/ 140K multi-view images
- Time Cost: full-finetuning 8.5h on one A100



# FiT3D: Generalization on Segmentation & Depth Estimation

- Indoor scene-level semantic segmentation

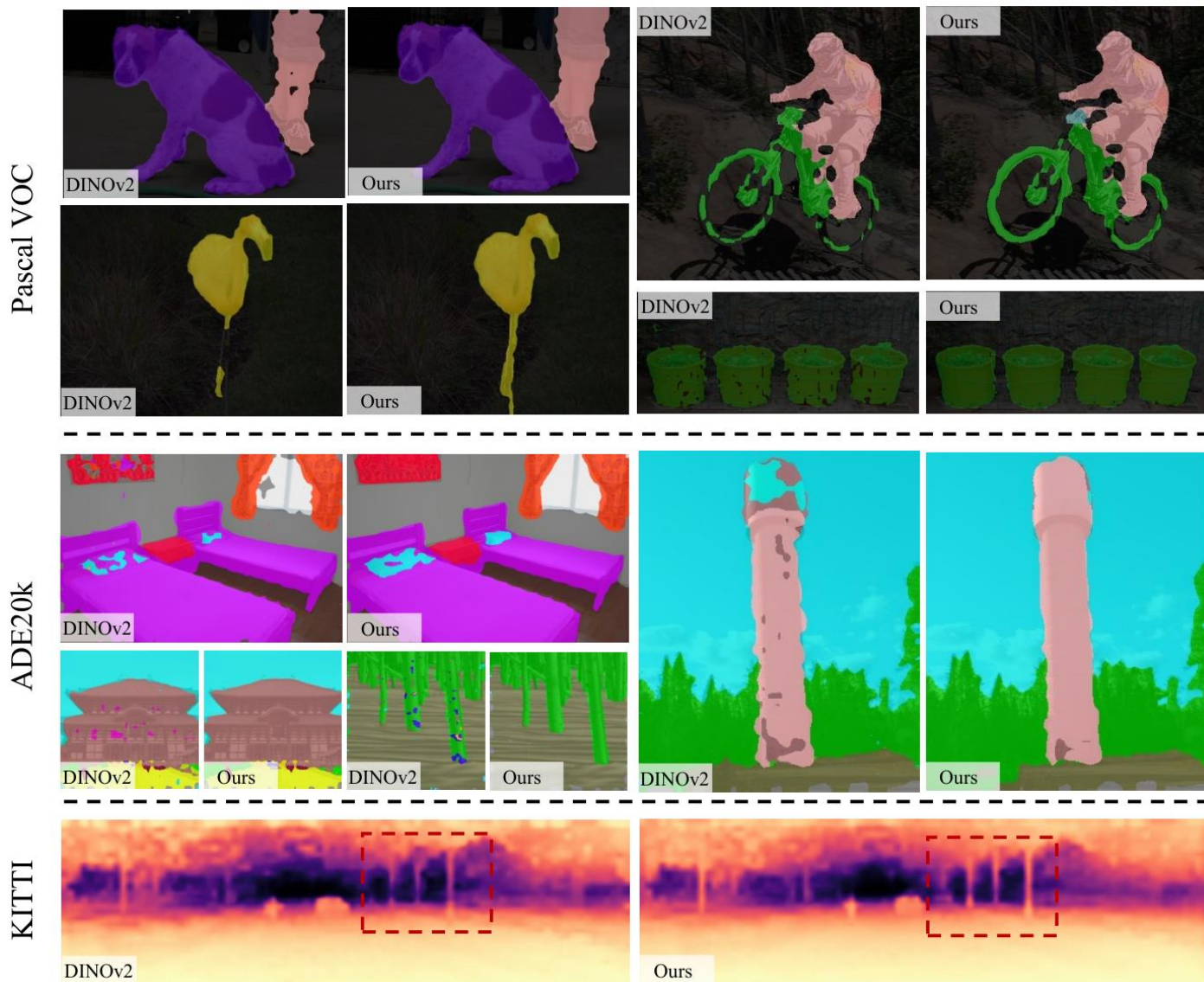
Method	ScanNet++ [65]			NYUv2 [55]			ScanNet [9]		
	mAcc (↑)	mIoU (↑)	aAcc (↑)	mAcc (↑)	mIoU (↑)	aAcc (↑)	mAcc (↑)	mIoU (↑)	aAcc (↑)
DINOv2 [44]	40.84	30.19	80.25	76.88	65.55	82.43	55.86	43.6	73.54
+ Ours	<b>43.4</b>	<b>32.76</b>	<b>83.54</b>	<b>80.52</b>	<b>67.5</b>	<b>83.37</b>	<b>58.32</b>	<b>44.84</b>	<b>74.37</b>

- (OOD dataset) semantic segmentation

Method	ADE20k [67]			Pascal VOC [17]			KITTI [19]	
	mAcc (↑)	mIoU (↑)	aAcc (↑)	mAcc (↑)	mIoU (↑)	aAcc (↑)	RMSE (↓)	Rel (↓)
DINOv2 [44]	56.74	44.28	79.73	90.61	81.14	95.72	3.03	0.10
+ Ours	<b>58.71</b>	<b>45.93</b>	<b>81.05</b>	<b>91.04</b>	<b>82.35</b>	<b>96.14</b>	<b>2.91</b>	<b>0.09</b>

- Other foundation models also benefits

	DINOv2-reg		CLIP		MAE		DeiT-III	
	mIoU (↑)	RMSE (↓)	mIoU (↑)	RMSE (↓)	mIoU (↑)	RMSE (↓)	mIoU (↑)	RMSE (↓)
Original	30.92	0.4190	25.61	0.4324	17.19	0.4855	18.62	0.4350
+ Ours	<b>33.39</b>	<b>0.3824</b>	<b>28.82</b>	<b>0.3960</b>	<b>20.27</b>	<b>0.4795</b>	<b>22.98</b>	<b>0.3820</b>



# Recent Advances in 3D-Aware 2D Models

## Text

Q: How far is [A] from [B]?  
A: It's [Distance].

SpatialVLM, CVPR24  
SpatialRGPT, NeurIPS24

## RGBD Img



DepthAnything, CVPR24  
DepthAnythingv2, NeurIPS24  
Metric3D, ICCV2023  
Metric3Dv2, T-PAMI24

## Multi-View Imgs



FiT3D, ECCV24  
ConDense, ECCV24 oral  
3DCorrEnhance, arxiv24  
CAT3D, NeurIPS24 oral

## Video

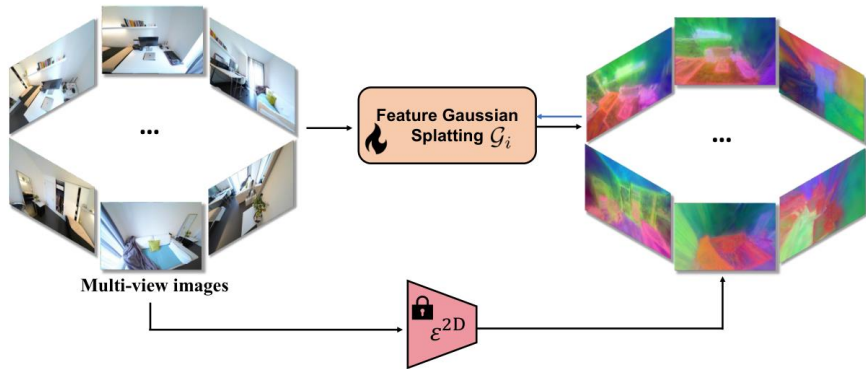


See3D, arxiv24  
CAT4D, arxiv24

# Other Efforts To Develop 3D-Aware Model

## Perception Models

2D Perceptions



3D Reconstruction

## Generative Models

3D Native Generation



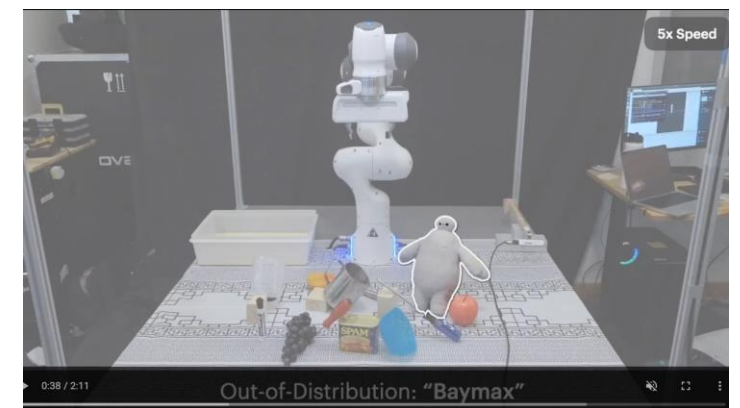
2D-to-3D Generation

## Robotics

3D Pretraining



3D Reconstruction



# Summary of 3D-Awareness

- Ways to Express 3D data & Learn 3D-Awareness
  - Text: simple, but inefficient, low-performance.
  - RGBD: surface representation, developed as *useful foundation models*
  - Multi-View: providing *sufficient information* to the whole 3D scene
  - Video: *natural, easy to scale up*
- Potential Impact
  - Image Models: Know better about the "Real 3D World" and the "Camera Pose"
  - *video models: Know better about the "Motion" and "Physics"*
  - *Robotics: Know better about the "Real 3D World", develop "Embodied Intelligence"*

Towards Learning the Better Real World

# Increasing Attention on 3D Vision



MENU	Select Primary Subject Area	# papers	Oral	Poster
1	Recognition: detection, categorization, retrieval	177	25	152
2	Image and video synthesis and generation	157	26	131
3	3D from multi-view and sensors	137	26	111
4	Low-level vision	110	19	91
5	Vision + language	105	20	85
6	Segmentation, grouping and shape analysis	99	16	83
7	Transfer/ low-shot/ long-tail learning	86	15	71
8	Deep learning architectures and techniques	85	20	65
9	Self- & semi- & meta- & unsupervised learning	84	7	77
10	Video analysis and understanding	77	15	62
11	Pose estimation and tracking	62	14	48
12	Representation learning	61	11	50
13	3D from single images	60	10	50

## CVPR 2023 by the Numbers



Selecting a category below changes the paper list on the right.

**SELECT** ↓ Top 10 overall by number of authors

	AUTHORS	PAPERS
1	3D from multi-view and sensors	1,090   246
2	Image and video synthesis and generation	889   185
3	Humans: Face, body, pose, gesture, movement	813   166
4	Transfer, meta, low-shot, continual, or long-tail learning	688   153
5	Recognition: Categorization, detection, retrieval	673   139
6	Vision, language, and reasoning	631   118
7	Low-level vision	553   126

## CVPR 2024

Computer Vision and Pattern Recognition Conference  
Seattle | June 17-21, 2024

### Paper Topics and Totals

1	Image and video synthesis and generation	329
2	3D from multi-view and sensors	276
3	Humans: Face, body, pose, gesture, movement	202
4	Vision, language, and reasoning	152
5	Low-level vision	131
6	Recognition: Categorization, detection, retrieval	127
7	Transfer, meta, low-shot, continual, or long-tail learning	123
8	Multi-modal learning	110
9	Segmentation, grouping and shape analysis	107
10	3D from single images	106
11	Datasets and evaluation	95
12	Autonomous driving	87
13	Video: Action and event understanding	78

### Oral (top 5% papers)



### Highlight (top 10% papers)

