CVPR Nashville JUNE 11-15, 2025

Project Page

Robust 3D Shape Reconstruction in Zero-Shot from a Single Image in the Wild

Junhyeong Cho[1] Kim Youwang[2] Hunmin Yang[1,3] Tae-Hyun Oh[2,3]

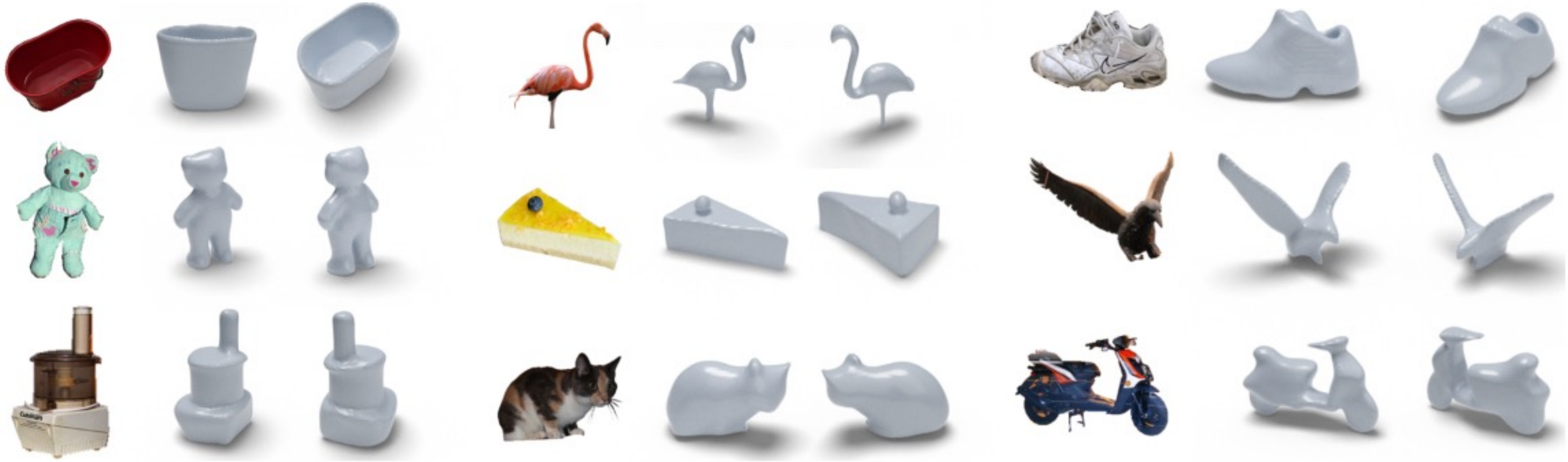
[1]ADD [2]POSTECH [3]KAIST

POSTECH
POHANG UNIVERSITY OF SCIENCE AND TECHNOLOGY

KAIST

Motivation

Recent monocular 3D shape reconstruction methods have shown promising zero-shot results on *object-segmented images without occlusions.*



Huang et al., “ZeroShape: Regression-based Zero-shot Shape Reconstruction”, CVPR, 2024.

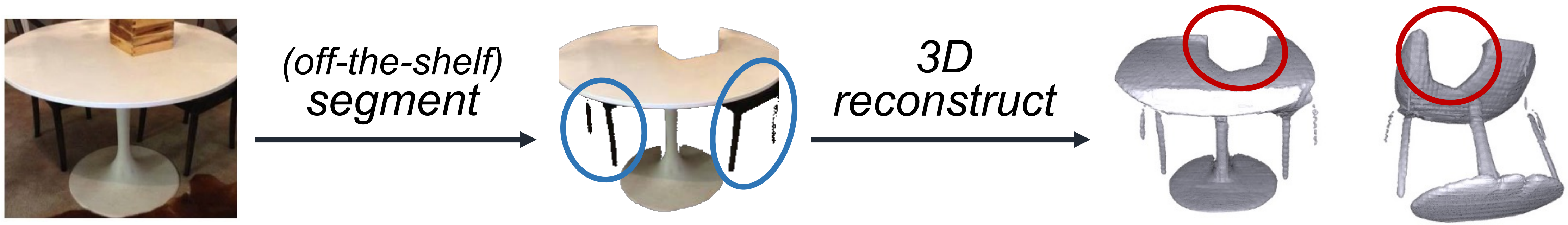
Motivation

What happens in the wild?

Most real-world objects are *unsegmented* and ***partially occluded***

Motivation

In practice, existing methods suffer from *segmentation errors by off-the-shelf models* and *the prevalence of occlusions*!



(off-the-shelf) segment

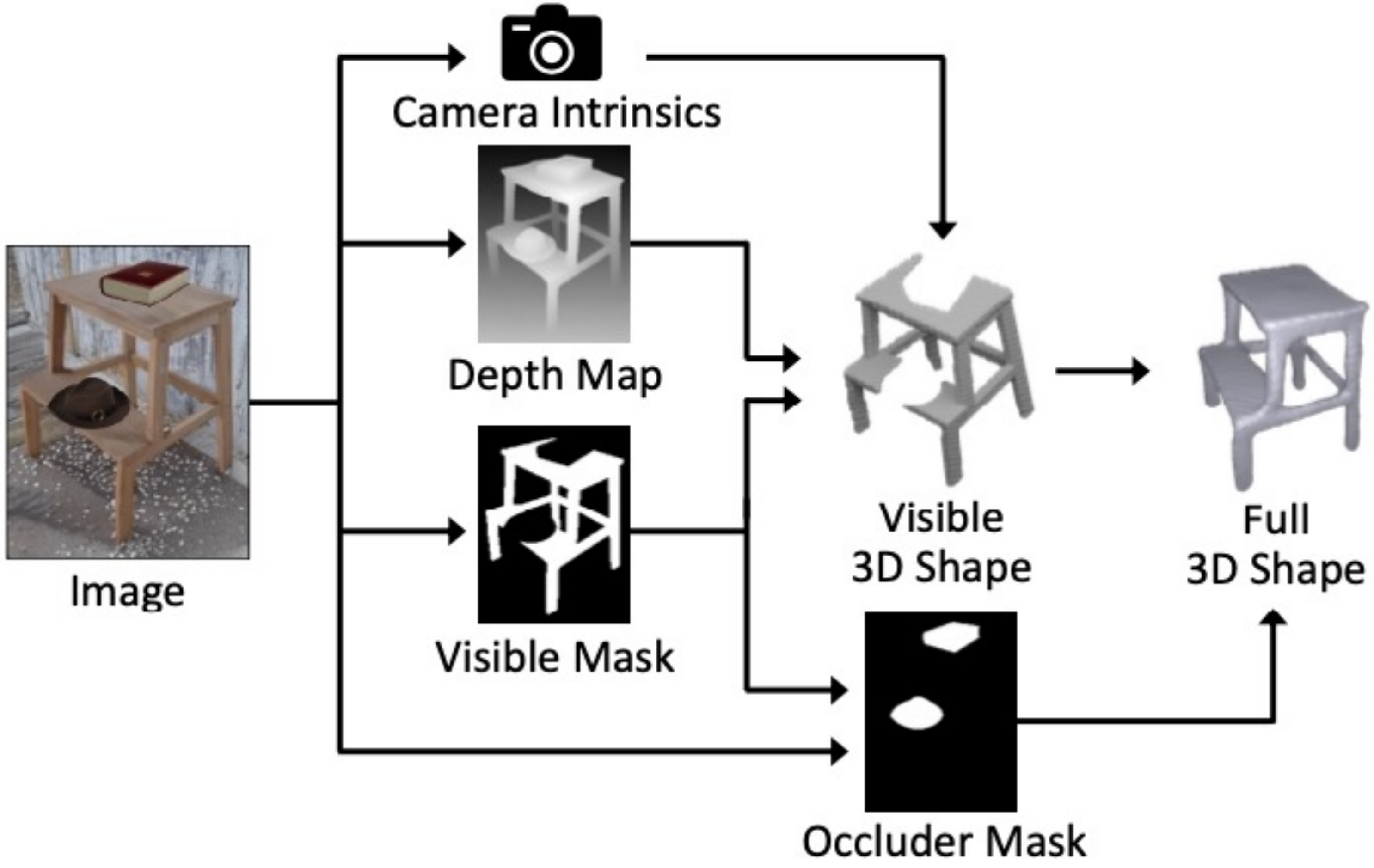
3D reconstruct

by segmentation errors

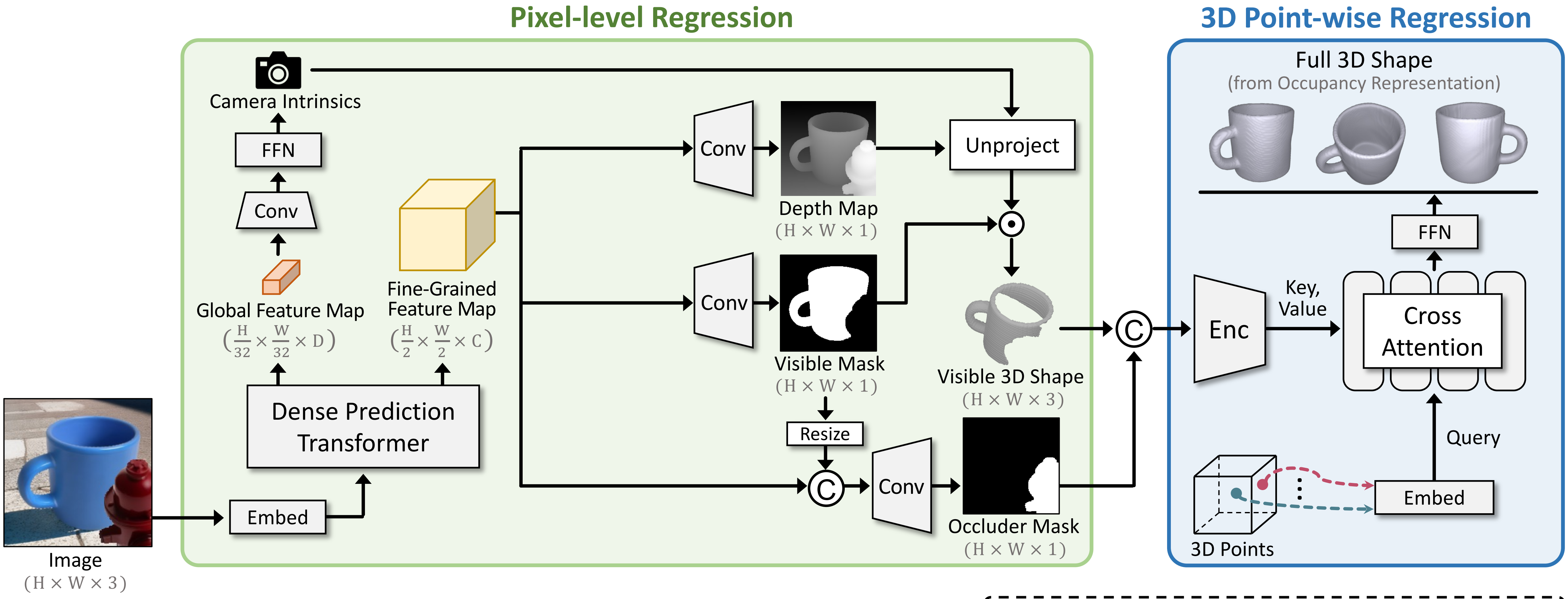
by occlusions

Our Approach: ZeroShape-W

- Occlusion-Aware 3D Shape Reconstruction Model

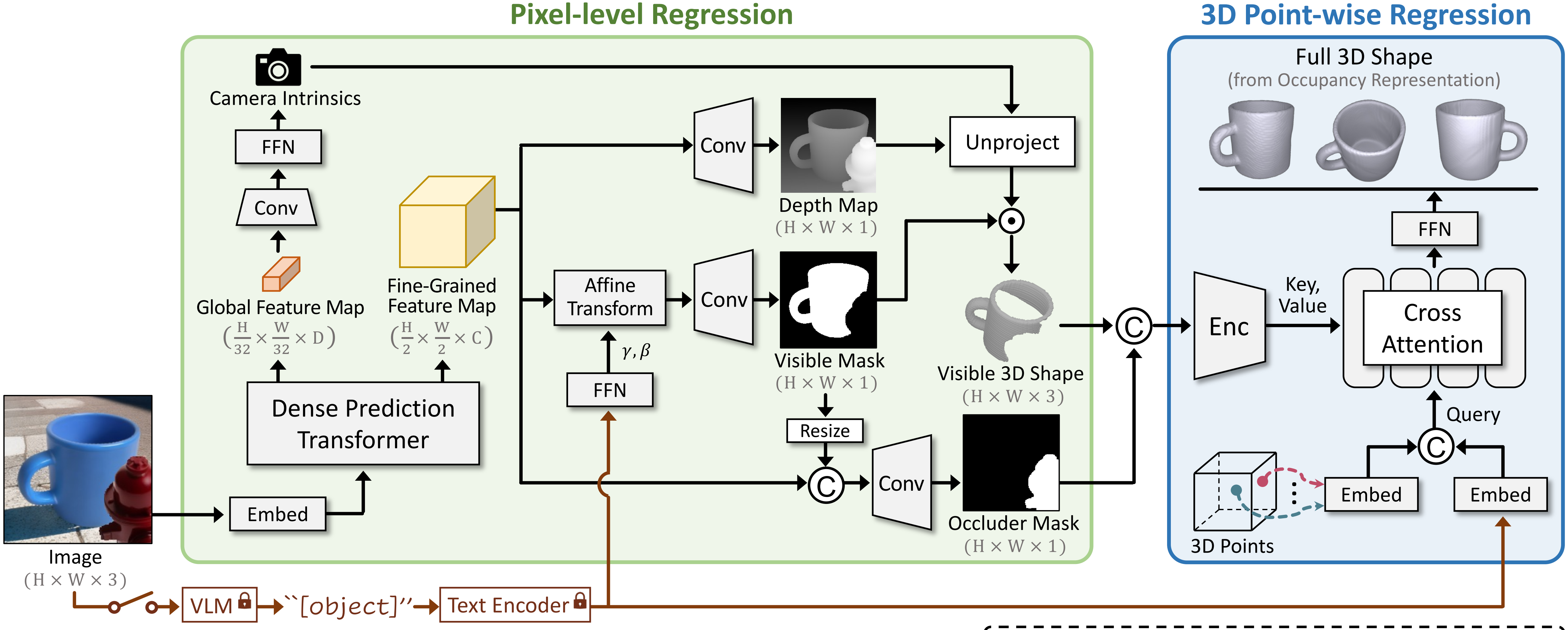


Model Architecture



🔒 Frozen Weights ⊙ Hadamard Product Ⓒ Concatenation

Model Architecture



🔒 Frozen Weights ⊙ Hadamard Product Ⓒ Concatenation

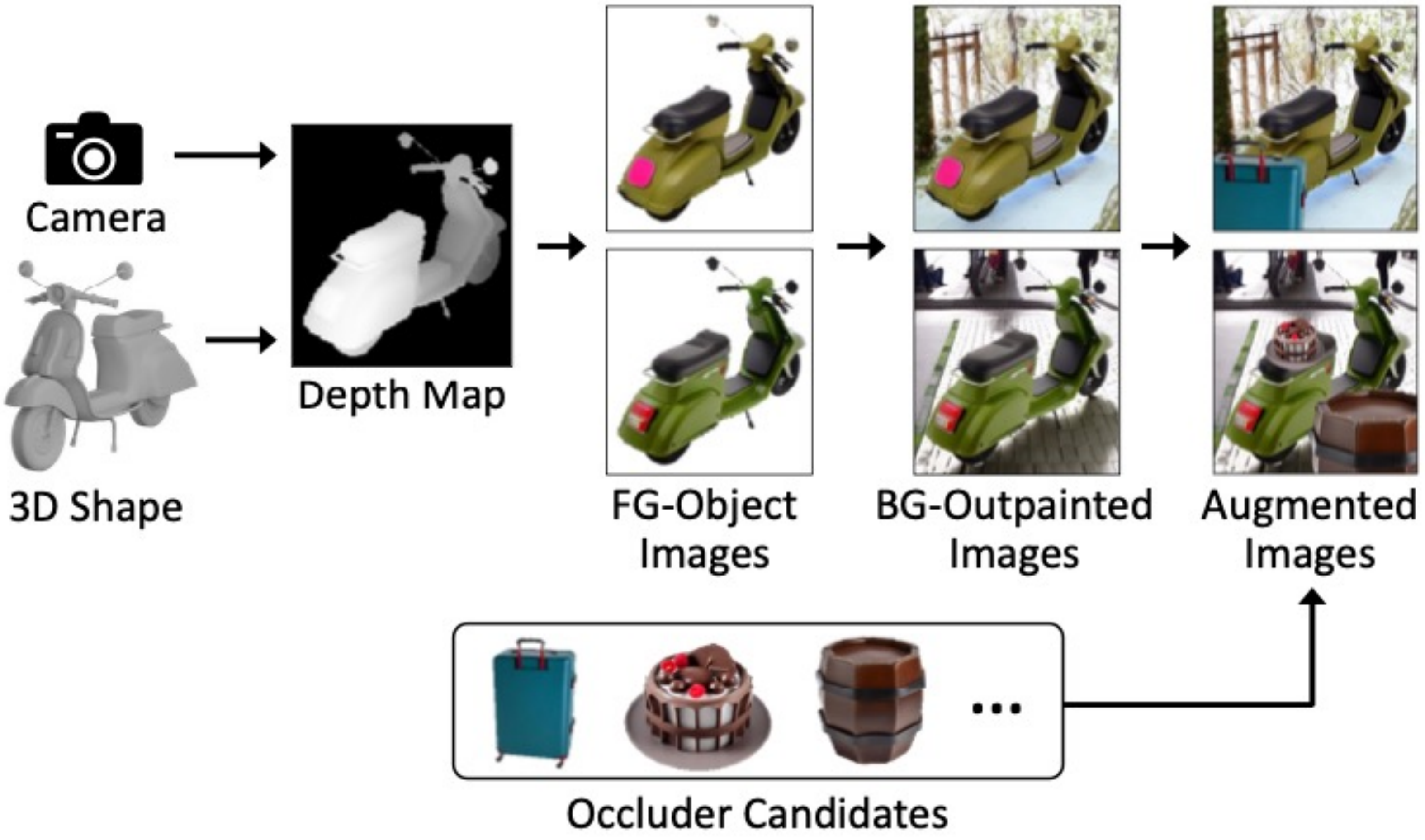
Challenge

How can we train our model?

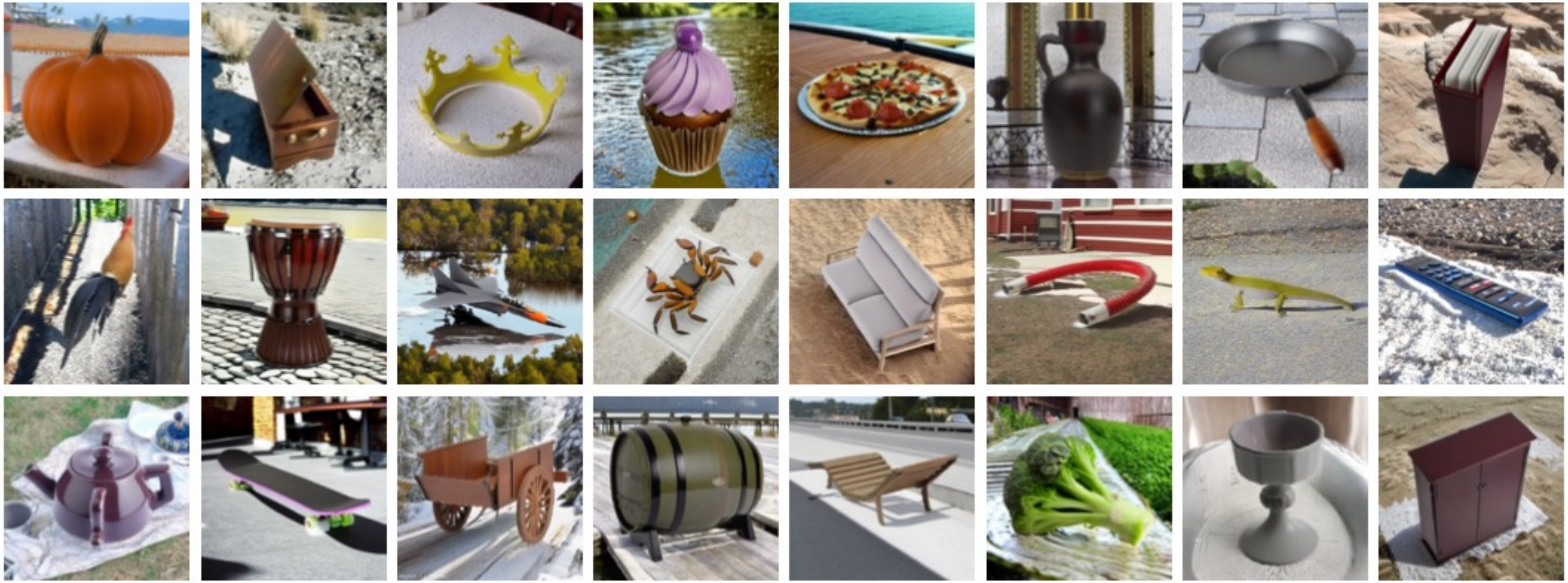
- The availability of real-world images annotated with precise 3D shapes is limited.
- An alternative approach is to synthesize realistic renderings from 3D shape collections (e.g., ShapeNet).
  - However, this is also limited by the availability of high-quality synthetic assets (e.g., 4K-res texture maps, HDR environment maps).

Our Training Data

- Data Synthesis & Occlusion Augmentation

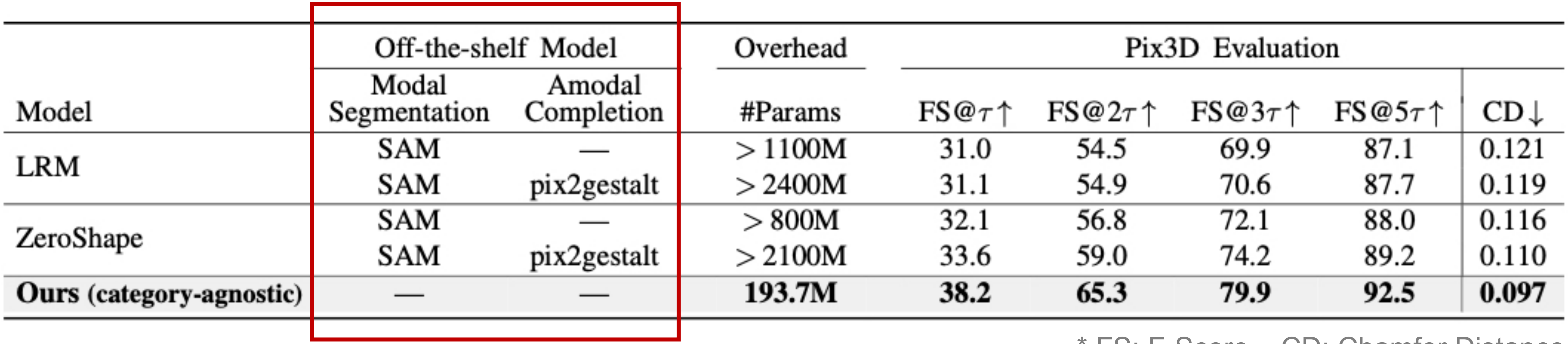


Our Training Data



- > 50K 3D Shapes from ShapeNetCore.v2
- > 40K 3D Shapes from Objaverse-LVIS
- > 1,000 Object Categories
- > 1 Million Synthetic Images

Quantitative Comparison

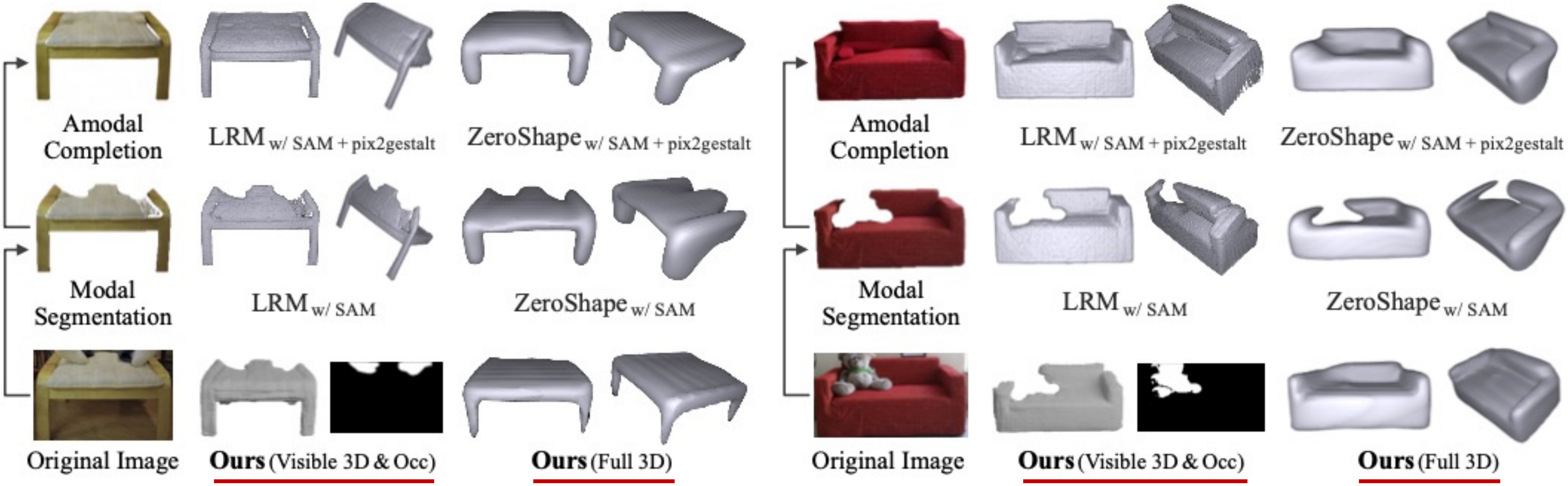


	Off-the-shelf Model		Overhead	Pix3D Evaluation				
Model	Modal Segmentation	Amodal Completion	#Params	FS@τ↑	FS@2τ↑	FS@3τ↑	FS@5τ↑	CD↓
LRM	SAM	—	$>$1100M	31.0	54.5	69.9	87.1	0.121
	SAM	pix2gestalt	$>$2400M	31.1	54.9	70.6	87.7	0.119
ZeroShape	SAM	—	$>$800M	32.1	56.8	72.1	88.0	0.116
	SAM	pix2gestalt	$>$2100M	33.6	59.0	74.2	89.2	0.110
Ours (category-agnostic)	—	—	**193.7M**	**38.2**	**65.3**	**79.9**	**92.5**	**0.097**

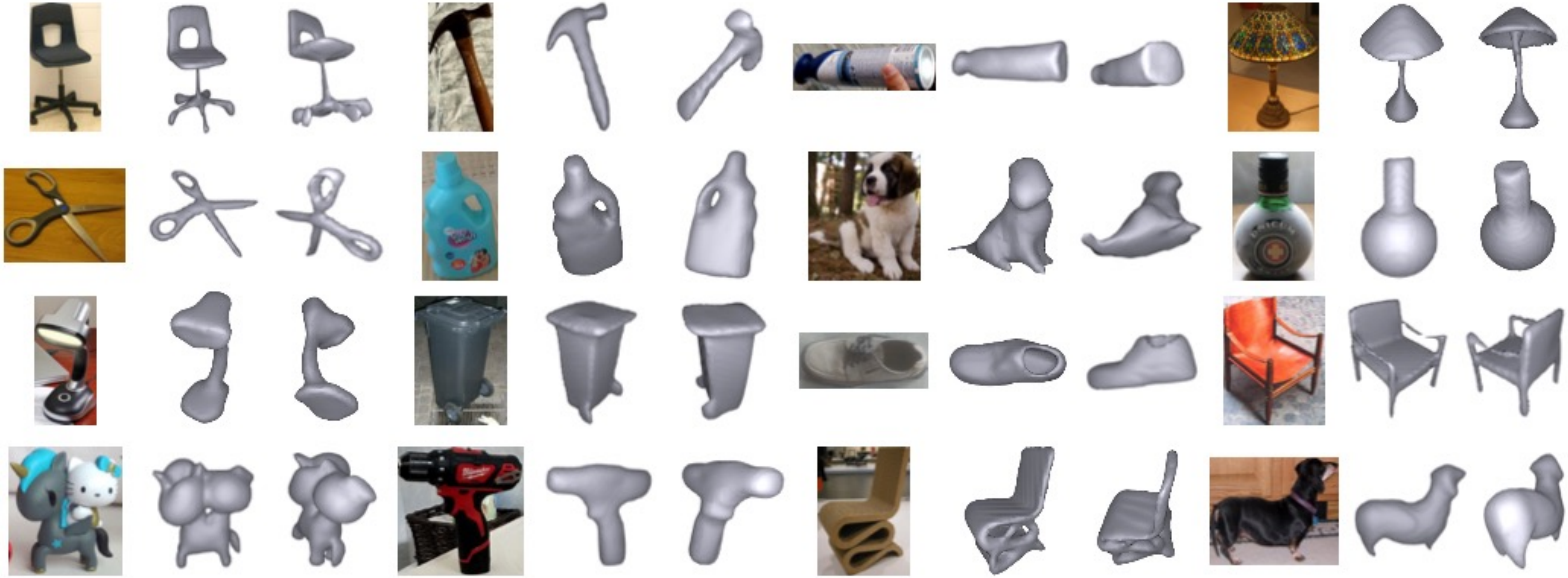
* FS: F-Score CD: Chamfer Distance

- In terms of FS@τ and CD, our model outperforms the strongest baseline ZeroShape$_{\text{w/SAM+pix2gestalt}}$ by a large margin of 13%
- The number of parameters used by our model is less than 1/12 of the parameters used by LRM$_{\text{w/SAM+pix2gestalt}}$

Qualitative Comparison

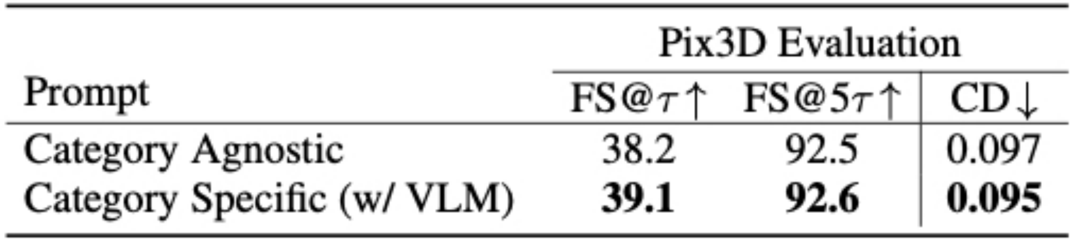


Reconstruction of Diverse Objects



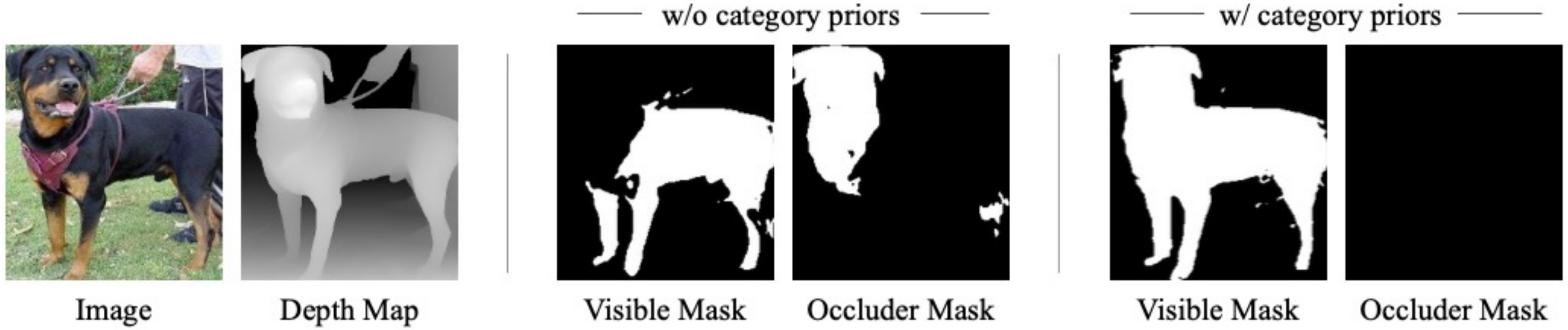
Our regression-based model has learned generalizable 3D shape priors!

Effect of Category Priors

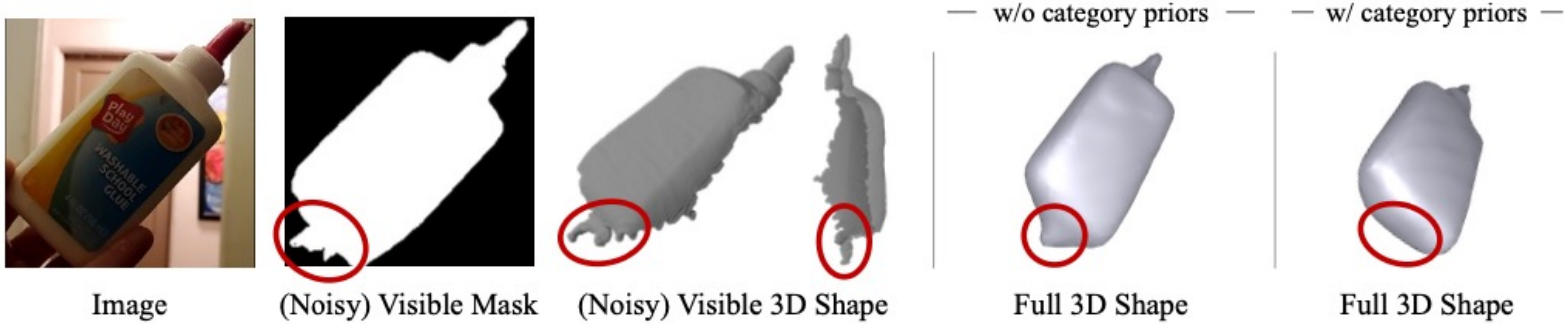


	Pix3D Evaluation		
Prompt	FS@$\tau\uparrow$	FS@$5\tau\uparrow$	CD$\downarrow$
Category Agnostic	38.2	92.5	0.097
Category Specific (w/ VLM)	**39.1**	**92.6**	**0.095**

- Mask Regression



- Occupancy Regression



Thank You

Project Page is here!