Omnidata: A Scalable Pipeline for Making Multi-Task Mid-Level Vision Datasets from 3D Scans

Group 6

Rishikesh Jadhav, Aniruddh Balram, Patrik Pordi, Jingxi Chen, Ji-Ze Jang, Quynh Thi Phung

Scribe : Rishikesh Jadhav

What problem are they trying to solve?

Data Scarcity: The datasets often focus on specific tasks like object recognition, leaving other vision tasks underrepresented.

Data Quality and Annotations: **G**enerating high-quality annotations for vision tasks like surface normal estimation and depth estimation can be labor-intensive and costly

Lack of Diverse Vision Tasks: Existing datasets are often biased towards recognition tasks and may not cover a wide range of vision tasks, such as depth estimation, surface normal estimation, and panoptic segmentation

Limited Control over Data Generation : The authors recognize the need for more control over the data generation process, including factors like camera intrinsics, scene lighting, and data domain

Authors Solution :

Data Scarcity: Generating multi-task mid-level vision datasets from 3D scans.

Data Quality and Annotations: Automating annotation generation for tasks like surface normal and depth estimation.

Lack of Diverse Vision Tasks: Creating datasets that cover a wide range of vision tasks beyond recognition.

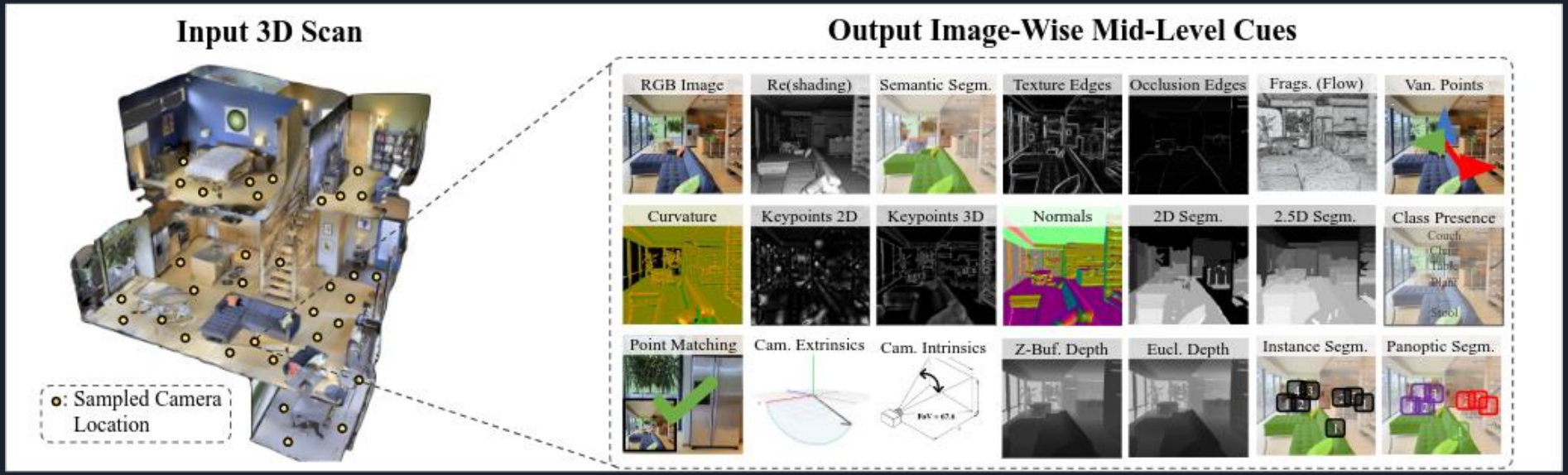
Limited Control over Data Generation : Providing control over data generation parameters to researchers for customization.

INTRODUCTION

The main concept is to create a pipeline for “steerable” (highly adaptable) computer vision datasets from 3D scans, leveraging mid-level cues for training vision models.

Simple steps:

1. Camera and Point Sampling
2. View Sampling (Wide-baseline multi-view, Smooth trajectory sampling)
3. Rendering Mid-level Cues



RELATED WORK

Static 3D Datasets : Many recent datasets rely on 3D scans, but they often focus on **specific domains**. **Omnidata** offers diversity across scenes and objects.

Vision-Focused Simulators: While simulators use 3D meshes, Omnidata bridges the gap between simulators and static vision datasets, prioritizing realism.

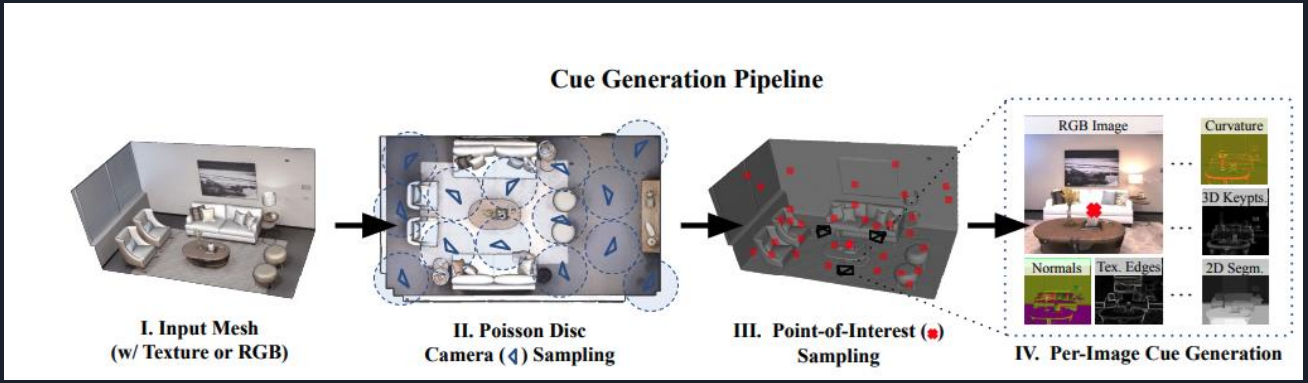
Linking 2D Images and 3D Meshes : Rely on hand-annotation, which is expensive and time-consuming.

Multi-Task Datasets: Existing multi-task learning (MTL) datasets often specialize in recognition tasks. Omnidata aims for a more comprehensive, real-world setting.

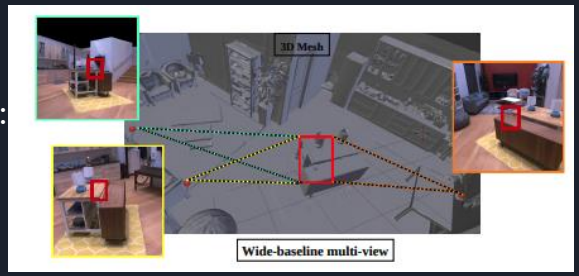
Data Augmentation + Domain Randomization(RL): Omnidata leverages a rich set of data augmentations, including including viewpoint consistency and Euclidean transforms, for static computer vision datasets, enhancing **model robustness**.

Auto Labeling: The pipeline harnesses the structure in 3D scans to efficiently compute and propagate labels, reducing annotation effort(basically the use pretrained models to reduce manual efforts).

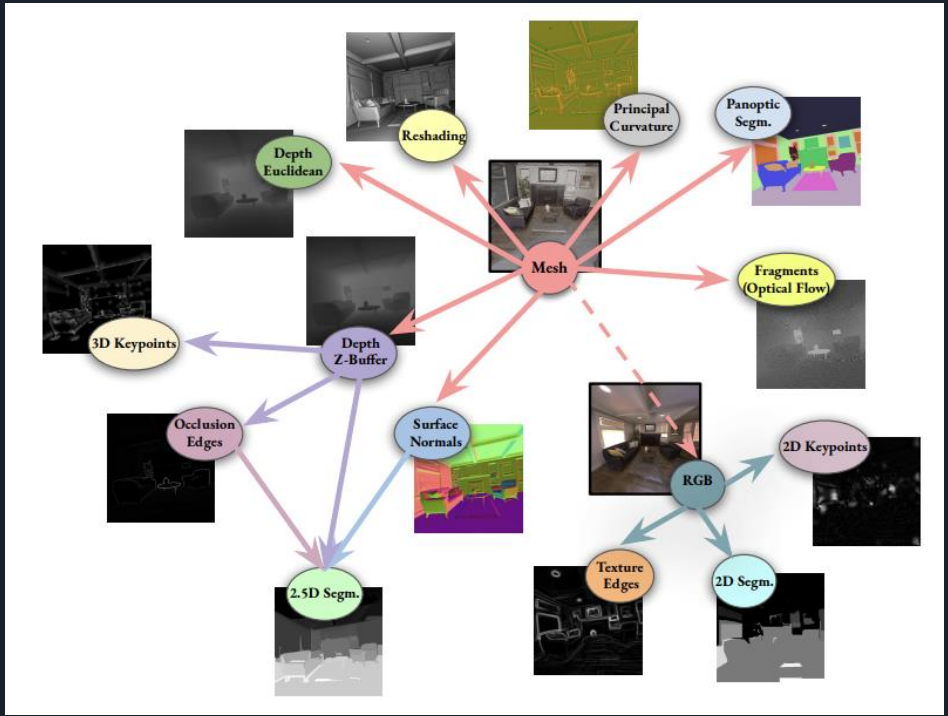
PIPELINE OVERVIEW



1. The annotator generates camera locations and points-of-interest along the mesh. (Filtering- each camera sees at least one point, and each point is seen by a minimum number of cameras (default: 3)
2. For each camera and each point-of-interest, it creates a view from that camera fixated on the point.
3. For each space-point-view triplet, the annotator renders all the mid-level cues



How are all of these cues generated?



The pipeline uses some of the mid-level cues to produce others.

Mid-level cues provided for the starter set.

STARTER DATASET OVERVIEW

The starter data is generated from seven mesh-based datasets, categorized as follows:

Indoor scene datasets: Replica, HyperSim, Taskonomy, Habitat-Matterport (HM3D).

Aerial/outdoor datasets: BlendedMVG.

Diagnostic/Structured datasets: CLEVR.

Object-centric datasets: generated by placing Google Scanned Objects around buildings in the Replica dataset, following the approach used by ObjectNet for image diversity in classification(simulation).

The Starter Dataset is a large dataset annotated with Omnidata's annotator, containing a diverse range of images from various scenes

RESULTS

Dataset statistics

Dataset	Images			Spaces			Points
	Train	Val	Test	Train	Val	Test	
CLEVR	60,000	6,000	6,000	1	0	0	72,000
Replica	56,783	23,725	23,889	10	4	4	4,150
Replica + GSO	107,404	43,450	42,665	10	4	4	31,167
Hypersim	59,543	7,386	7,690	365	46	46	74,619
Taskonomy	3,416,314	538,567	629,581	379	75	79	684,052
BlendedMVG	79,023	16,787	16,766	341	74	73	112,576
Habitat-Matterport	8,470,855	1,061,021	-	800	100	-	564,328
Total (no CLEVR)	12,189,922	1,690,936	720,591	1,905	303	206	1,434,892

RESULTS

Zero-shot depth estimation

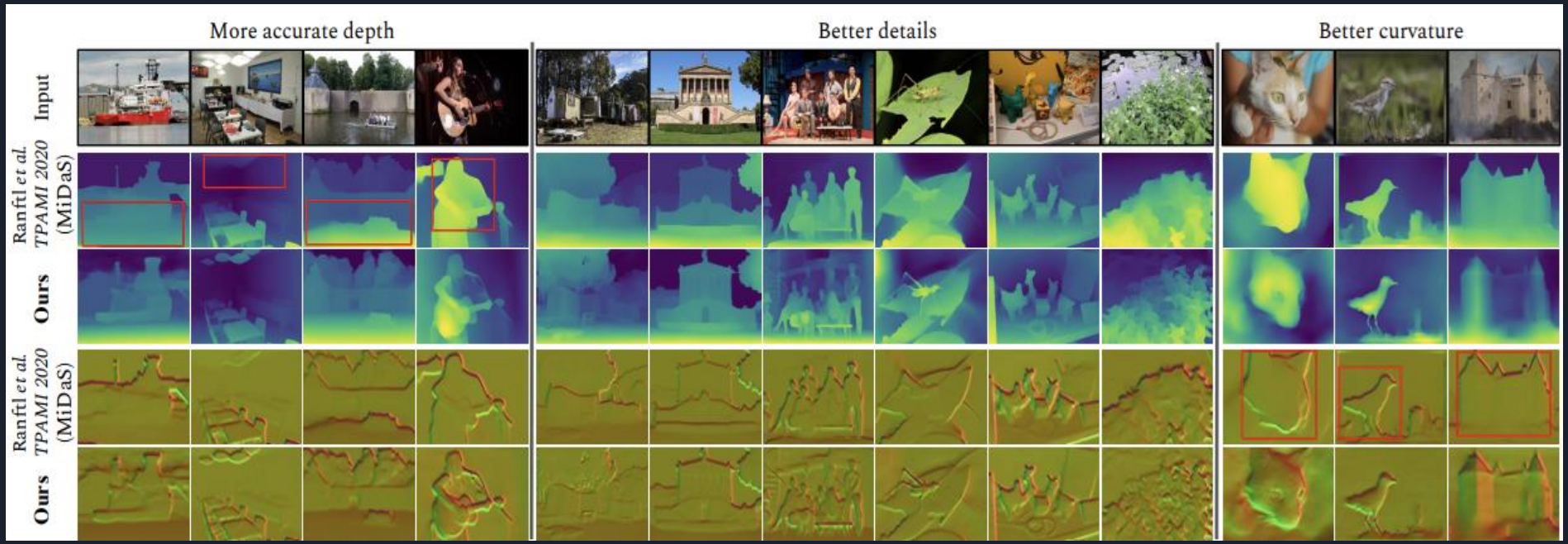
Method	Test Data	L1 Error (↓)	$\delta > 1.25$ (↓)	$\delta > 1.25^2$ (↓)	$\delta > 1.25^3$ (↓)
XTC [71]	OASIS [13]	1.180	85.28	71.86	60.22
MiDaSv3 [46]		0.8057	82.03	67.25	55.35
Omnidata		**0.7901**	**81.00**	**65.22**	**52.93**
XTC [71]	NYU [53]	0.5279	70.41	49.90	36.28
MiDaSv3 [46]		0.3838	63.84	41.65	28.97
Omnidata		**0.2878**	**51.73**	**30.98**	**20.86**

Zero-shot surface normal estimation (check AUC for Unet + Omnidata)

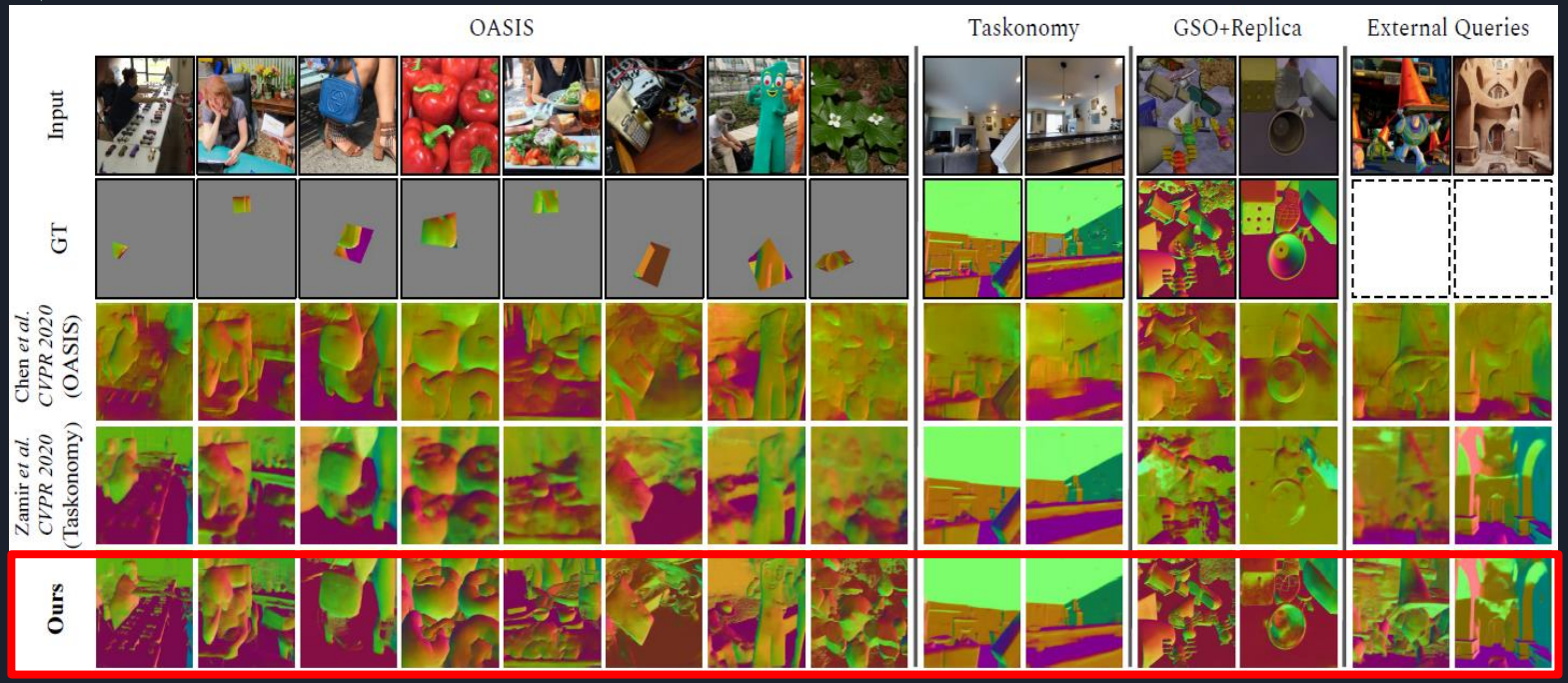
Method	Training Data	Anglular Error°		% Within t°			Relative Normal	
		Mean	Median	11.25°	22.5°	30°	AUC_o	AUC_p
Hourglass [12]	OASIS [13]	**23.91**	18.16	**31.23**	59.45	**71.77**	0.5913	0.5786
Hourglass [12]	SNOW [14]	31.35	26.97	13.98	40.20	56.03	0.5329	0.5016
Hourglass [12]	NYU [54]	35.32	29.21	14.23	37.72	51.31	0.5467	0.5132
PBRS [74]	NYU [54]	38.29	33.16	11.59	32.14	45.00	0.5669	0.5253
UNet [50]	SunCG [55]	35.42	28.70	12.31	38.51	52.15	0.5871	0.5318
UNet [50]	**Omnidata**	24.87	**18.04**	31.02	**59.53**	71.37	**0.6692**	**0.6758**
Human (Approx.)	-	17.27	12.92	44.36	76.16	85.24	0.8826	0.6514

Qualitative comparison with MiDaS on zero-shot OASIS **depth estimation**

Rectangular boxes show regions useful for comparison.



Qualitative results of zero-shot **surface normal estimation**



Qualitative results of **panoptic segmentation** with PanopticFPNs trained on COCO and Omnidata

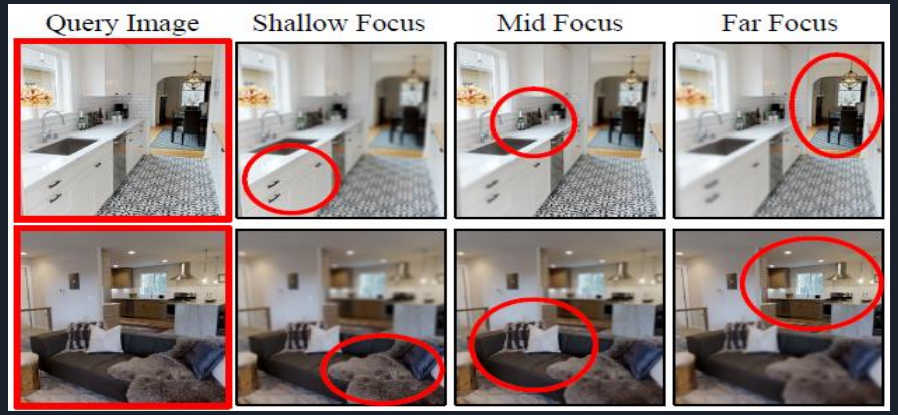
The Omnidata model trained jointly on Taskonomy, Replica, and Hypersim shows good performance on indoor scenes without people.

ILLUSTRATIVE DATA-FOCUSED ANALYSES (One For All)

Inter-dataset domain transfer performance for **surface normal estimation** and **panoptic Segmentation**(starter dataset test).

Image refocusing augmentation on Taskonomy(for better generalization performance).

	Surface normal estimation: L1 Error (↓)						Panoptic Quality (PQ) (↑)			
Train/Test	Taskonomy	Replica	Hypersim	Replica+GSO	BlendedMVG	h. mean	Taskonomy*	Replica	Hypersim	h. mean
Taskonomy*	**4.85**	7.76	8.69	13.89	15.55	8.53	8.39	3.95	11.67	6.55
Replica	9.36	**3.98**	11.78	10.28	15.02	8.24	1.01	**41.97**	4.50	2.43
Hypersim	7.28	7.57	**6.72**	11.34	12.94	8.56	**9.35**	14.08	25.39	13.80
Replica+GSO	13.88	4.94	15.05	**5.17**	14.03	8.26	-	-	-	-
BlendedMVG	17.1	14.23	16.93	14.87	**8.85**	13.58	-	-	-	-
Omnidata	5.32	4.24	6.53	6.45	11.53	**6.11**	9.14	41.24	**30.16**	**17.98**



Are **Mid level Cues** as inputs useful???

Semantic segmentation results using models trained on Replica and tested on Replica, Hypersim and Taskonomy.

These models received mid-level cues as input(except RGB).

Input/Supervised Domains	GT Mid-Level			Predicted Mid-Level		
	Cross-Entropy (↓)			Cross-Entropy (↓)		
	Repl.	H.Sim	Task.	Repl.	H.Sim	Task.
RGB	0.61	5.87	7.55	0.61	5.87	7.55
(All Above) + Normals	0.47	4.47	6.12	0.61	5.44	7.12
(All Above) + 3D Edges	**0.46**	4.47	6.75	0.54	5.06	6.49
(All Above) + (2D Edges, Z-Depth, 3D Keypts)	**0.46**	**3.86**	**6.04**	**0.53**	**4.9**	**6.13**

As we can see the models notably benefited from the mid-level cues.

CONCLUSION AND LIMITATION

Summary:

- This paper presented a way to generate better quality data for training visual learning tasks from 3D scans
- The authors show the improvement/effect by their “**camera and point sampling**” and “view sampling” on better FOV, viewing angle and viewing distance distribution for generated data .

Limitation:

- The “camera and point sampling” in the paper is to designed to densely sampling the 3D space, however this does not lead to better sampling in “**semantic objects**”
- For the **generated RGB** images from mesh, the quality may not be that high (or requires super fine-grained mesh -> huge storage space)

Resources and References

Resources:

Pipeline Code and Docs: Simplifies technology adoption.

Docker Container: Includes required software for easy setup.

PyTorch Dataloaders: Efficiently loads dataset.

Starter Dataset: Contains 14.5M labeled images.

Convenience Utilities: Aid data management and mesh filtering.

Pretrained Models: Includes MiDaS implementation.

References:

Link to the paper : https://arxiv.org/abs/2110.04994

Link to the website : https://omnidata.vision/

Link to the Video : https://www.youtube.com/watch?v=jAXaASBB5N0&t=70s