

# **Building Vision Foundation Models with Very Limited Resources**

---

**Hirokatsu Kataoka**

National Institute of Advanced Industrial Science and Technology (AIST)  
Visual Geometry Group, University of Oxford (Oxford VGG)

<http://www.hirokatsukataoka.net/>

# Hirokatsu Kataoka

Ph.D. in Engineering (Keio University; Mar 2014)



## Profile :

- Chief Senior Researcher, AIST (Apr 2023 - Present)
- Academic Visitor, Visual Geometry Group, University of Oxford (Sep 2024 - Present)
- PI, LIMIT.Lab (Jun 2025 – Present; **Research initiative w/ VGG community**)
- Visiting Associate Professor, Keio University (Sep 2024 - Present)
- Adjunct Researcher, SB Intuitions (May 2024 - Present)
- Adjunct Associate Professor, Tokyo Denki University (Apr 2024 - Present)
- PI, cvpaper.challenge (May 2015 – Present; Community with 1,500+ collaborators)

## Recently Selected Projects (within 3 years):

- “Scaling Backwards: Minimal Synthetic Pre-training? (ECCV24)”
- “Rethinking Image Super-Resolution from Training Data Perspectives (ECCV24)”
- “Pre-training Vision Transformers with Very Limited Synthesized Images (ICCV23)”
- “Primitive Geometry Segment Pre-training (**BMVC23 Best Industry Paper Finalist**)”
- “SegRCDB: Semantic Segmentation via Formula-Driven Supervised Learning (ICCV23)”
- “Visual Atoms: Pre-training Vision Transformers with Sinusoidal Waves (CVPR23)”
- “Replacing Labeled Real-Image Datasets with Auto-Generated Contours (CVPR22)”
- “Point Cloud Pre-training with Natural 3D Structures (CVPR22)”
- “Pre-training without Natural Images (IJCV22; **ACCV Best Paper Honorable Mention**)”

片 かた  
岡 おか  
裕 ひろ  
雄 かつ

<http://hirokatsukataoka.net/>

1

## Visual Pre-training with Minimal Data & Supervision

Can a natural law train a visual model?

- ACCV 2020 Best Paper Honorable Mention Award
- Featured in MIT Technology Review (Feb. 4<sup>th</sup>, 2021)
- One single synthetic image enables to pre-train ViT (ECCV24)

2

## Visual Foundation Models without Real Data

Can synthetic pre-training make a vision foundation model?

- Industry-focused vision foundation models (arXiv 2025 / on-going work)
- Primitive Geometry Segmentation for medical 3D data (BMVC 2023 Best Industry Paper Finalist)

3

## Multimodal AI Models with Generative Models

Can generative models make next foundation models?

- Zero-shot 3D understanding (CVPRW25 / on-going work)
- Leading research initiative (LIMIT.Lab with VGG community)

1

## Visual Pre-training with Minimal Data & Supervision

Can a natural law train a visual model?

- ACCV 2020 Best Paper Honorable Mention Award
- Featured in MIT Technology Review (Feb. 4<sup>th</sup>, 2021)
- One single synthetic image enables to pre-train ViT (ECCV24)

2

## Visual Foundation Models without Real Data

Can synthetic pre-training make a vision foundation model?

- Industry-focused vision foundation models (arXiv 2025 / on-going work)
- Primitive Geometry Segmentation for medical 3D data (BMVC 2023 Best Industry Paper Finalist)

3

## Multimodal AI Models with Generative Models

Can generative models make next foundation models?

- Zero-shot 3D understanding (CVPRW25 / on-going work)
- Leading research initiative (LIMIT.Lab with VGG community)



Starting from the [Kataoka+, ACCV20/IJCV22], our team have proven...

- Visual pre-training can be done with mathematically generated images / without any real data
- Representations in video/3D/audio | Tasks in cls/det/seg are also learnable (Any Modality, Any Task)

# Fractal Database

**to make a pre-trained CNN model without any natural images.**

## Where did that idea come from?

How could we learn huge #parameters with only 1M images?

- Something like 'Natural Law' inside of the image dataset



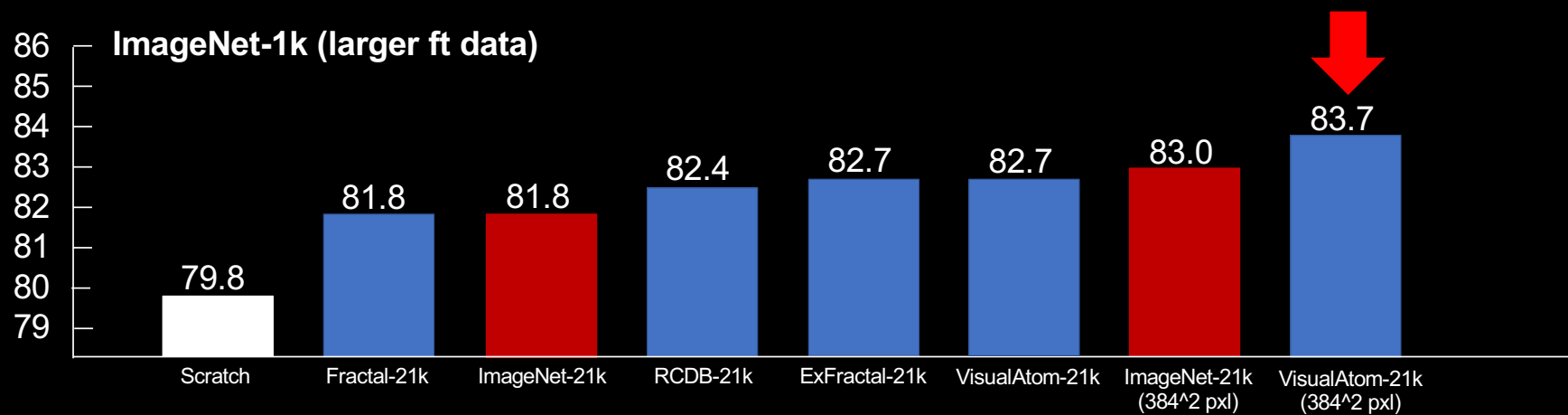
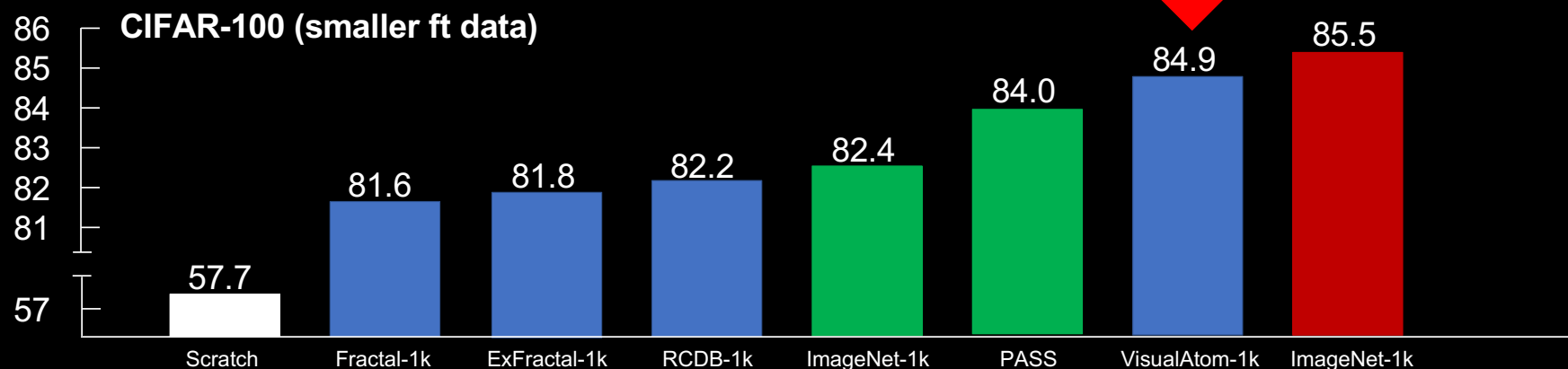
Observed fractal geometry on ImageNet dataset



We hypothesize DNN could learn  
'Natural Law' inside of the dataset

**Directly render and train primitives**

In image classification,



Synth pre-training is much closer to / even better than the real-image pre-training

FDSL

SSL

SL

# **Scaling Backwards: Minimal Synthetic Pre-training?**

**ECCV 2024**

(Collaborating with Oxford VGG & UTN FunAI Lab)

**Hirokatsu Kataoka**

National Institute of Advanced Industrial Science and Technology (AIST)

Visual Geometry Group, University of Oxford (Oxford VGG)

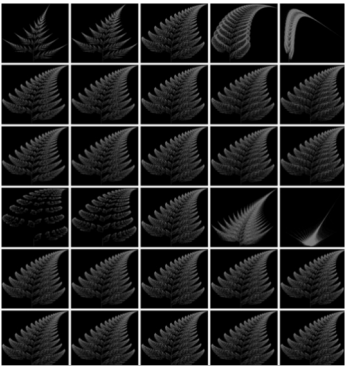
<http://www.hirokatsukataoka.net/>

Think backwards

How can we minimize the visual pre-training in the framework?

FractalDB:	1,000 categories	x 1,000 instances	1M images
OFDB:	1,000 categories	x 1 instance	1k images

The idea:      1 category                      x 1 instance                      | 1 image



Parameter set (x25)

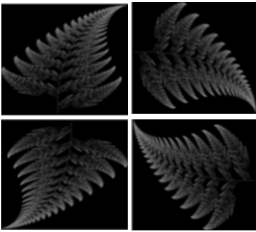
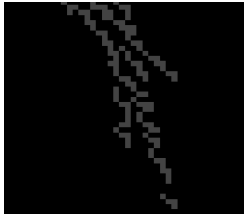


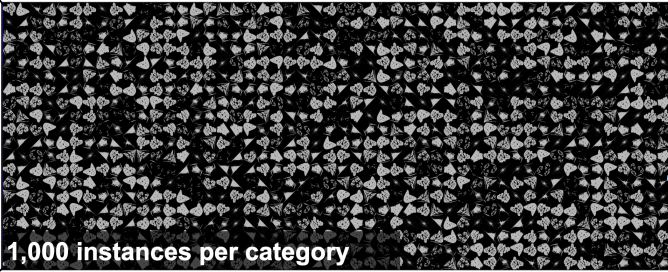
Image rotation  
(x4)



Patch pattern (x10)

Inside Fractal Category


FractalDB



1,000 instances per category

OFDB

One-instance FractalDB

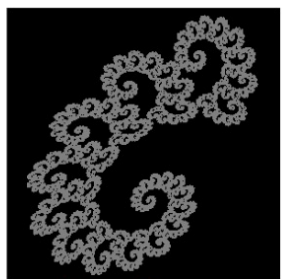


Only one Instance!!

Think backwards

**Ultimately, is it possible to learn from just a single image?**

FractalDB:	1,000 categories	x 1,000 instances	1M images
OFDB:	1,000 categories	x 1 instance	1k images
The idea:	1 category	x 1 instance	1 image



**“Can a single synthetic image match a million real images?”**

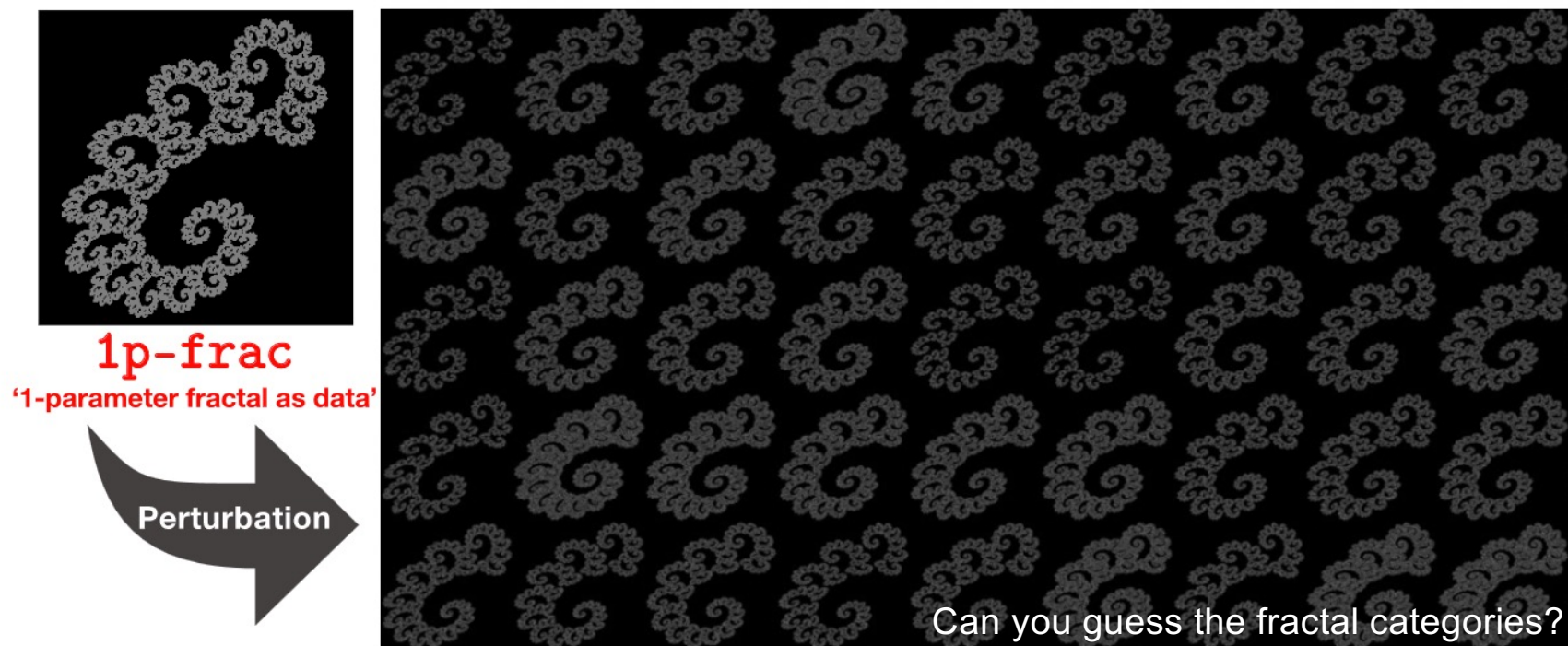
**Sounds like a crazy idea, but my intuition is ‘possible’**



What is the minimum requirement in visual pre-training?

**Essence is about classifying minute differences?**

Only a single fractal image, treating minute variations as pseudo-categories



Perturbations makes image categories / this enables to conduct a visual pre-training

## Scaling backwards?

### The comparisons between real and synthetic images

- Real images: ImageNet-1k
- Synth images: FractalDB, OFDB, and 1p-frac

**Table 1:** Scaling backwards in synthetic pre-training (Accuracies on CIFAR-100, Real: ImageNet, Synth: Fractal images).

Type\#Img	1	1k	1M
Real	N/A	76.9	85.5
Synth	84.2	84.0	81.6

#### Pre-training effectiveness

- Real img: #Images is important
- Synth img: Better performance for the task setting



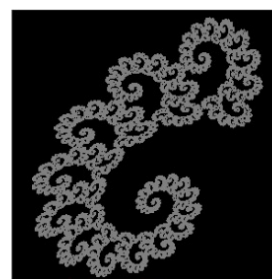
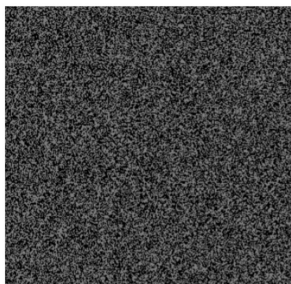
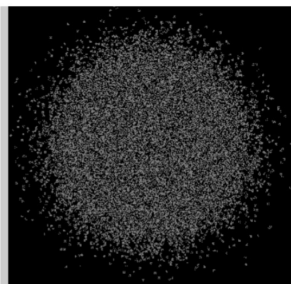
## Any random-image training possible?

### No pre-training effects in random / Gaussian images

- The result shows it is important to use a good formulation like fractal geometry

**Table 4:** Comparison with pre-training with a single noise image.

Method	C100 IN100	
Gaussian <sup>◇</sup>	1.1	5.7
Uniform <sup>◇</sup>	2.0	71.1
1p-frac <sup>◇</sup>	84.2	89.0



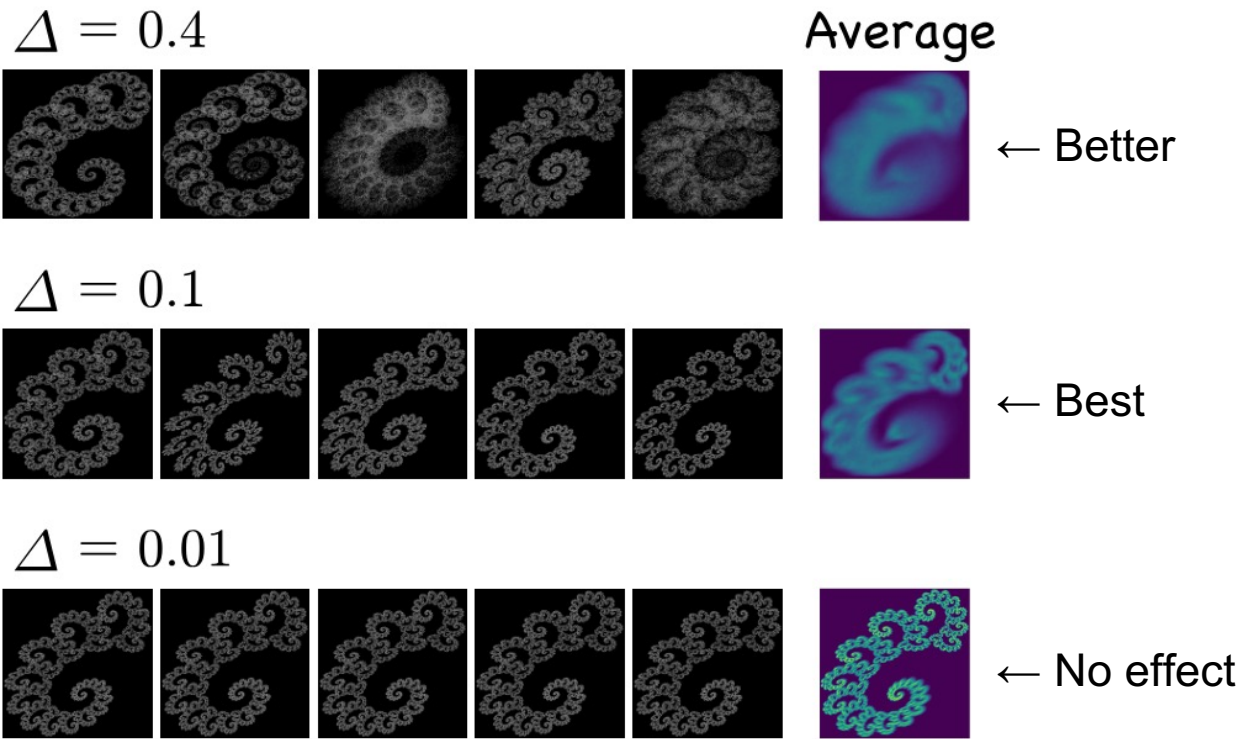
How fine does the shapes need to be?

Perturbation should be adjusted

➤ The perturbation values are shown in the following figure

**Table 5:** Effects of perturbation degree  $\Delta$  ( $\sigma = 3.5$ ).

$\Delta$	C100	IN100
0.001	1.2	1.9
0.01	19.9	61.8
0.1	<b>84.2</b>	<b>89.0</b>
0.2	83.4	88.5
1.0	82.6	88.1



Vision transformer can classify even the case of ‘ $\Delta=0.1$ ’

## How about the scaling ViT?

**Its plausible 21k categories from a single fractal parameter**

Even better than that of the ImageNet-21k pre-training (see 82.1 vs 81.8)

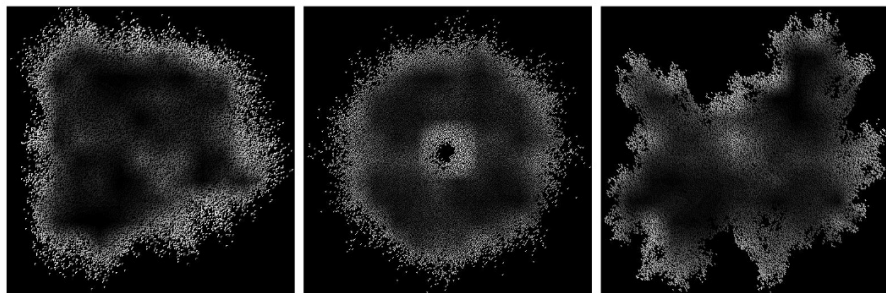
Pre-training	#Img	Type	ViT-B
Scratch	—	—	79.8
ImageNet-21k	14M	SL	81.8 ← IN-21k
FractalDB-21k	21M	FDSL	81.8
ExFractalDB-21k	21M	FDSL	82.7
RCDB-21k	21M	FDSL	82.4
VA-21k	21M	FDSL	82.7
OFDB-21k	21k	FDSL	82.2
3D-OFDB-21k	21k	FDSL	82.7
1p-frac (ours)	1	FDSL	82.1 ← The proposal

## Lessons from the synthetic training project

### What is the essence of visual pre-training?

- Previous: ViT tends to activate shape contours in visual tasks
- This work: ViT can classify pixel-level minute changes in categories

In synthetic data, a model indefinitely trains better performance, but learning efficiently with selected data yields better results



(d) Attention maps in fractal images with FractalDB-1k pre-trained DeiT. The brighter areas show more attentive areas.



1

## Visual Pre-training with Minimal Data & Supervision

Can a natural law train a visual model?

- ACCV 2020 Best Paper Honorable Mention Award
- Featured in MIT Technology Review (Feb. 4<sup>th</sup>, 2021)
- One single synthetic data enables to pre-train ViT (ECCV24)

2

## Visual Foundation Models without Real Data

Can synthetic pre-training make a vision foundation model?

- Industry-focused vision foundation models (arXiv 2025)
- Primitive Geometry Segmentation for medical 3D data (BMVC 2023 Best Industry Paper Finalist)

3

## Multimodal AI Models with Generative Models

Can generative models make next foundation models?

- Zero-shot 3D understanding (CVPRW25 / on-going work)
- Leading research initiative (LIMIT.Lab with VGG community)

# Industrial Synthetic Segment Pre-training

arXiv: 2505.13099

**Hirokatsu Kataoka**

National Institute of Advanced Industrial Science and Technology (AIST)

Visual Geometry Group, University of Oxford (Oxford VGG)

<http://www.hirokatsukataoka.net/>

## Towards the vision foundation models with synth train / limited resources

### Research questions:

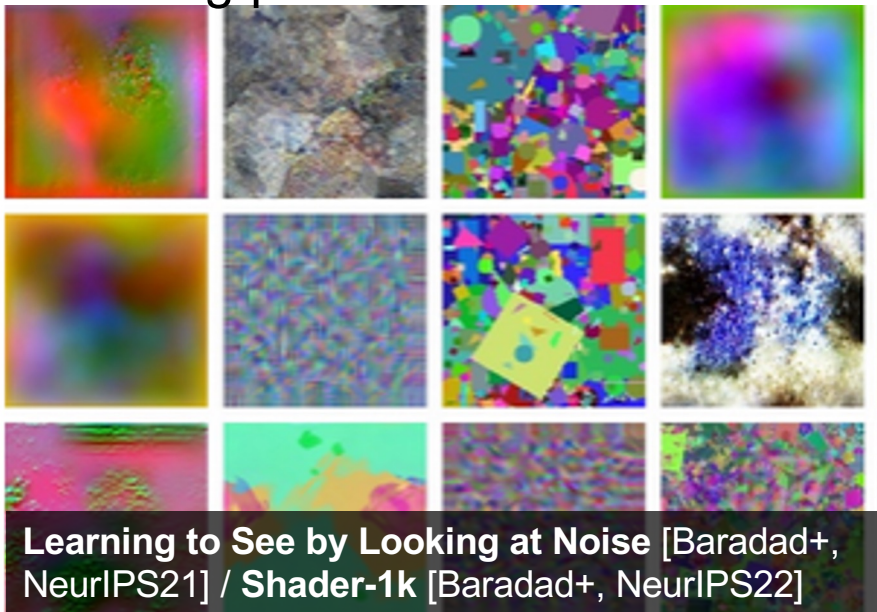
How can we beat SAM...

- with only synthetic data in pre-training phase?
- without any human supervision and real images?



## Insights from the prior art & industrial data

### Learning primitives



AT



PI

### Industrial data looks like...



<https://robot-mujinka.com/wp-content/uploads/2019/09/bara.jpg>

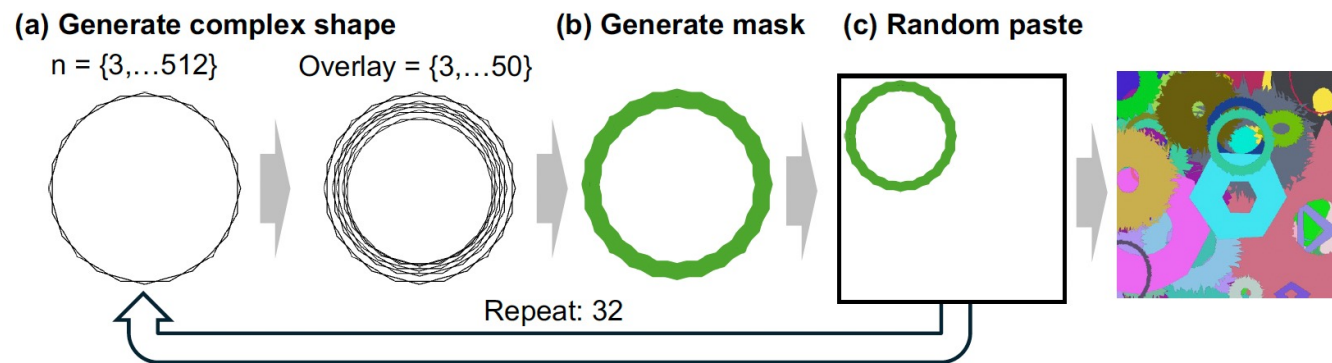
Complex, dense and hierarchical occlusion are the key in industrial & visual foundation models



## Visual pre-training for complex, dense and hierarchical occlusion handling

### Instance Core – Combination of many mathematically generated images

That is all, but much more difficult to separate each other comparing to the real world

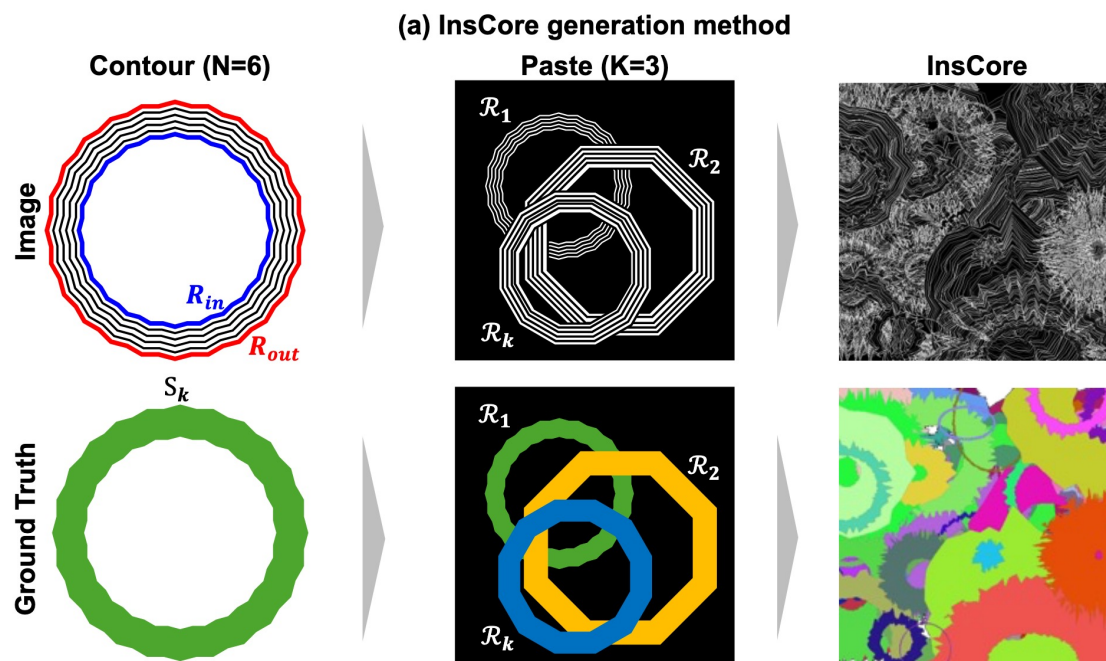


- Generate shape-oriented images
- Iterate putting pattern in 32 times

## Visual pre-training for complex, dense and hierarchical occlusion handling

### Instance Core – Combination of many mathematically generated images

That is all, but much more difficult to separate each other comparing to the real world



- A pair for image and semantic label
- Complex occlusions, dense and hierarchical masks

## Industrial image dataset

### Industrial data has different scenarios aren't covered by web data

- #Images and domain?
- We focused 'occlusion handling' visual pre-training

Evaluation dataset	Industrial domain	#Train		#Test		#Classes
		Image	MaskW	Image	Mask	
NuInsSeg [23]	Medical	532	23,127	133	7,571	6
LIVECell [10]	Biomedical	3,253	1,018,576	1,564	462,261	1
SpaceNet2 [32]	Remote sensing	3,080	8,7301	771	21,641	1
Industrial-iSeg [19]	Manufacturing	1,109	25,308	89	523	6
LogiSeg [22]	Logistics	1,384	10,018	300	2,093	7



## The impact of learning from the image primitives

### The InsCore pre-trained model is better effects on MS COCO segmentation

- Better than the ImageNet pre-trained model
- Only using x140 smaller pre-training dataset  
ImageNet 14M imgs vs. InsCore 0.1M synth imgs

Dataset	#Data	mAP	mAP <sub>50</sub>	mAP <sub>75</sub>
From scratch	—	42.3	65.7	45.5
ImageNet-21k	14M	43.7	67.4	47.3
SegRCDB	0.1M	43.8	67.4	47.4
InsCore (Ours)	0.1M	<b>44.4</b>	<b>68.2</b>	<b>47.5</b>

## The impact of learning from the image primitives

### Analysis of #synth-images in visual pre-training

- Enough with 100k synthetic images

#Data	Fine-tuning (mAP)				
	ES	LC	SN2	liSeg	LS
20k	36.3	15.3	60.8	21.6	94.9
100k	37.1	<b>18.3</b>	<b>61.6</b>	<b>25.5</b>	<b>95.1</b>
200k	<b>37.2</b>	15.5	61.5	24.4	94.5
400k	37.1	16.6	61.4	25.3	95.0



## The impact of learning from the image primitives

### The InsCore pre-trained model surpassed the fine-tuned SAM!

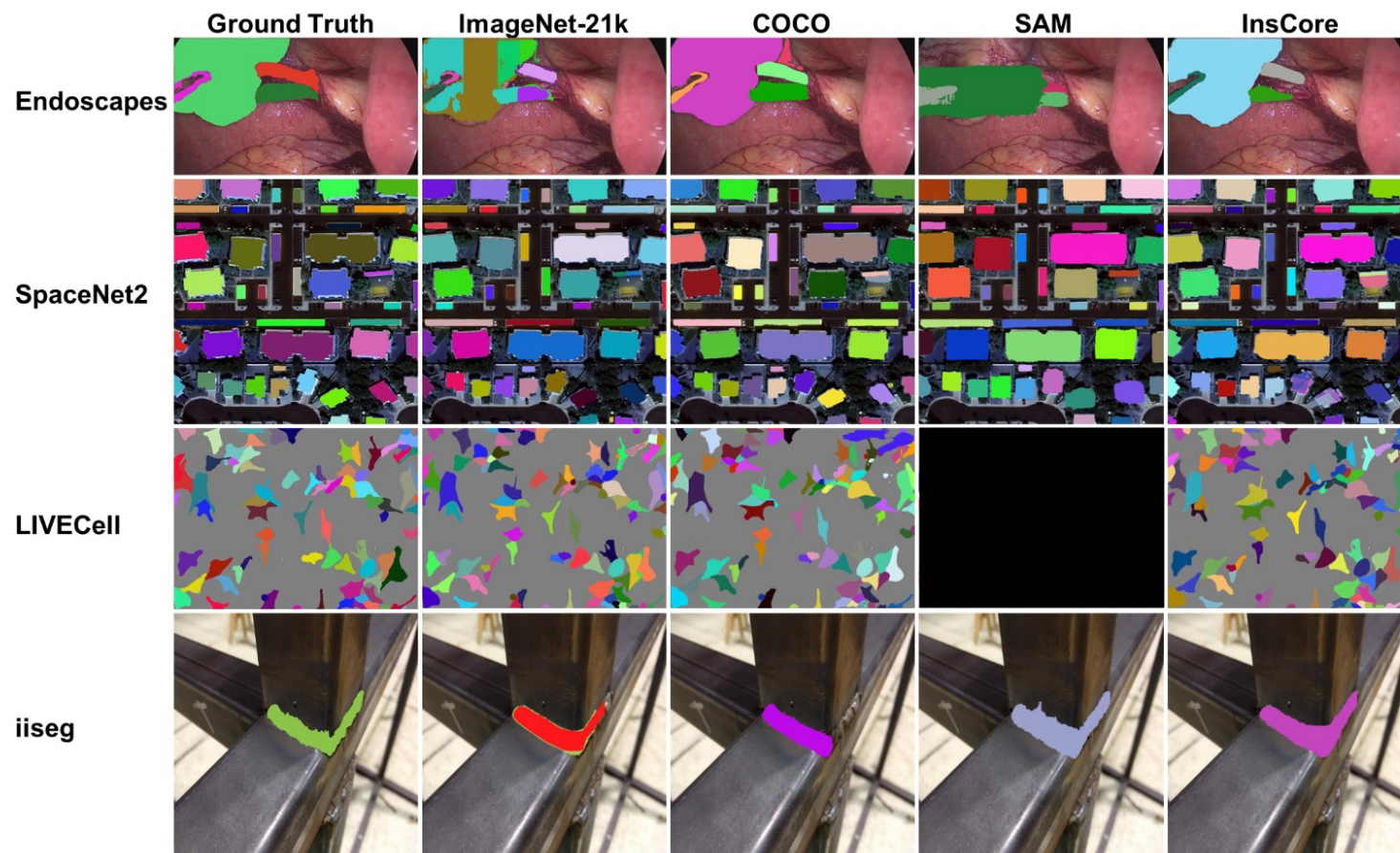
- Without real images nor human supervision in pre-training phase
- Only using x110 smaller pre-training dataset  
SA-1B 11M imgs, 1B segments vs. InsCore 0.1M synth imgs

Model	Backbone	Prompt	Pre-training		Fine-tuning (mIoU)			
			Dataset	Size	NuInsSeg	SpaceNet2	iiseg	LogiSeg
SAM (Zero-shot)	ViT-B	GT bbox	SA-1B	11M	40.1	56.0	46.4	91.1
SAM (Fine-tuning)					51.5	73.1	60.6	95.6
Mask R-CNN (Ours)	Swin-B	–	InsCore	0.1M	<b>66.0</b>	<b>76.9</b>	<b>60.8</b>	<b>96.4</b>

We can further improve the visual pre-training without real data, human supervision

## Visual results

### Qualitative results at each industrial dataset



The InsCore pre-trained segmentation surpassed SAM!

# Tech transfer with Toyota Industry Company (TICO)

産総研マガジン

🔍 記事検索

🔗 産総研マガジンとは

産総研の概要／研究データ／  
研究ユニットの紹介

産総研 オフィシャルサイト

産総研マガジン > LINK for Business > 「物流自動化の課題」に挑む豊田自動織機と産総研

📅 2025/02/05

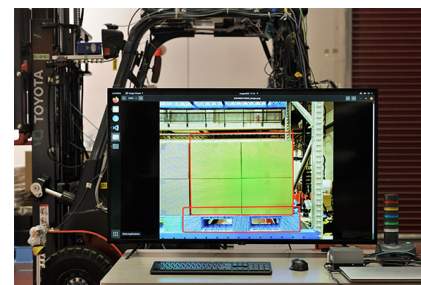
LINK for Business



「物流自動化の課題」に挑む豊田自動織機と産総研

現場の知見と先端技術を融合し、AIで荷姿の異常を検出

[https://www.aist.go.jp/aist\\_j/magazine/20250205.html](https://www.aist.go.jp/aist_j/magazine/20250205.html)



荷姿異常の例



現場による荷姿姿勢の違い



I led the successful tech transfer from academic research to industry



# **Primitive Geometry Segment Pre-training for 3D Medical Image Segmentation**

**BMVC 2023**

**Best Industry Paper Finalist**

**Hirokatsu Kataoka**

National Institute of Advanced Industrial Science and Technology (AIST)

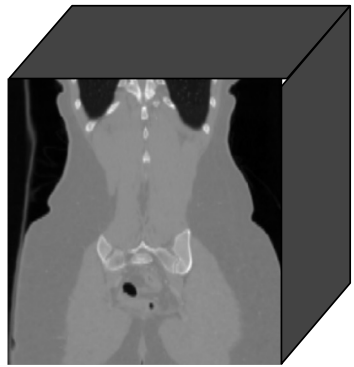
Visual Geometry Group, University of Oxford (Oxford VGG)

<http://www.hirokatsukataoka.net/>

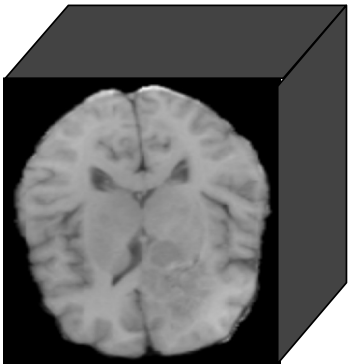
## 3D medical image segmentation

From the perspective, InsCore can be applied in the 3D x medical domain!

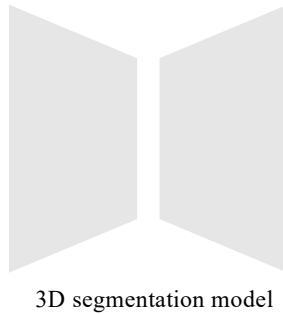
### Tumor/Organ Segmentation



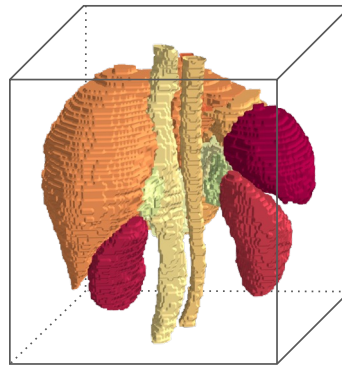
CT Scan (Input)



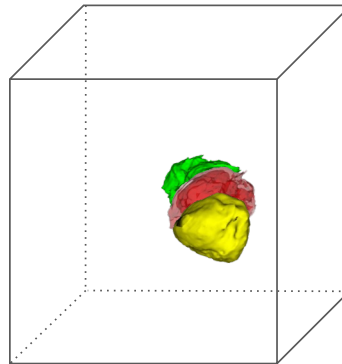
MRI Scan (Input)



3D segmentation model



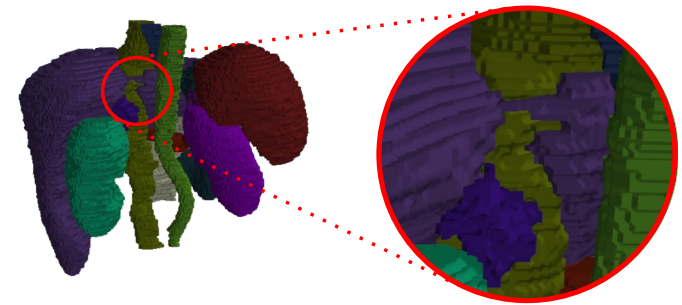
CT Scan (Output)



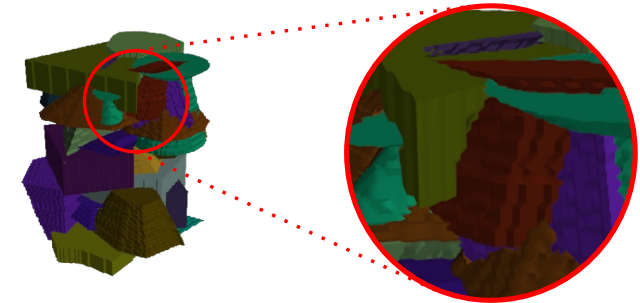
MRI Scan (Output)

### Deep analysis and their synthetic data

Real Organ



Synth Organ  
(PrimGeoSeg; Ours)



# Construction of PrimGeoSeg dataset

1. **Primitive object generation** : Combining simple **xy-plane** and **z-axis** rule
2. **Assembled object generation** : Randomly arranging primitive objects in 3D voxel

## xy-plane rule :

2D primitive shapes

Ellipse, 3-poly, 4-poly, 5-poly, 6-poly, 7-poly, **8-poly**, 9-poly

## z-axis rule :

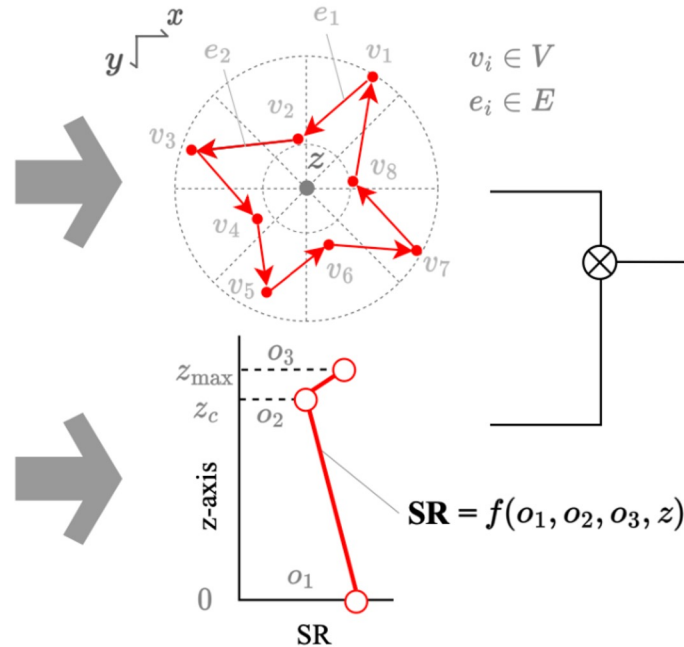
Similarity ratio along z-axis

1. Pillar

2. Cone

**3. Concave**

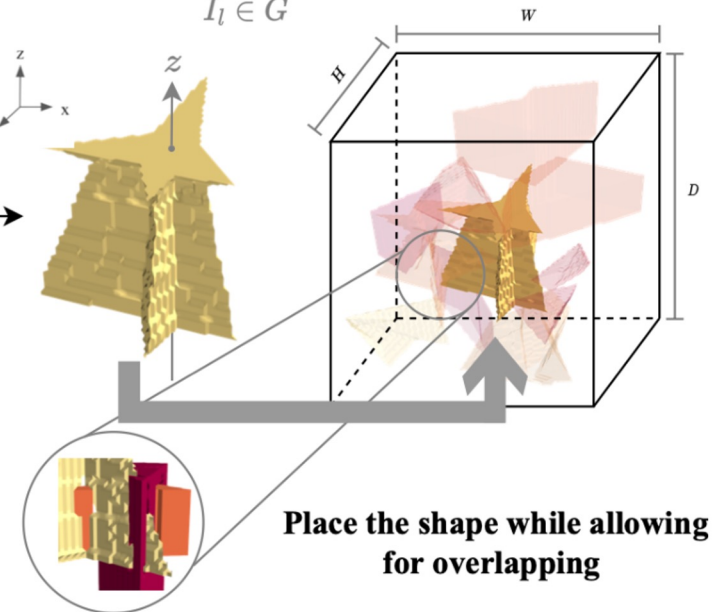
4. Convex



**Primitive object generation**

Primitive object  $I_l$   
 $I_l \in G$

Assembled object  $S_i$



**Assembled object generation**

## Experimental results: comparison with prior arts

### 👉 BTCV : 13 organs in abdomen

Pre-training	PT Num	Type	Avg.	Spl	RKid	LKid	Gall	Eso	Liv	Sto	Aor	IVC	Veins	Pan	rad	lad
<i>UNETR</i>																
Scratch	0	–	73.0	90.2	91.1	90.7	47.0	63.8	95.3	76.5	85.1	82.1	67.9	72.3	46.1	40.8
Chen <i>et al.</i> [6]	0.8K	SSL	75.8	95.2	<b>95.5</b>	93.8	51.9	52.3	<b>98.8</b>	80.0	87.8	82.7	66.1	68.9	60.8	51.3
PrimGeoSeg	0.8K	FDSL	77.4	88.9	94.0	93.8	59.8	65.7	95.4	79.3	88.3	82.6	69.9	76.8	58.5	53.3
PrimGeoSeg	50K	FDSL	<b>80.9</b>	<b>95.7</b>	94.2	<b>94.1</b>	<b>61.9</b>	<b>69.6</b>	96.7	<b>85.5</b>	<b>89.5</b>	<b>84.4</b>	<b>74.7</b>	<b>81.9</b>	<b>64.3</b>	<b>58.7</b>
<i>SwinUNETR</i>																
Scratch	0	–	78.3	92.3	93.2	93.8	55.9	61.3	94.0	77.0	87.5	80.4	74.2	76.1	68.8	63.6
Tang <i>et al.</i> [25]	5K	SSL	81.6	95.3	93.2	93.0	<b>63.6</b>	74.0	96.2	79.3	<b>90.0</b>	83.3	<b>76.1</b>	82.3	<b>69.0</b>	<b>65.1</b>
PrimGeoSeg	5K	FDSL	<b>82.0</b>	<b>95.7</b>	<b>94.4</b>	<b>94.4</b>	61.0	<b>75.5</b>	<b>96.7</b>	<b>83.3</b>	89.1	<b>85.6</b>	75.2	<b>84.3</b>	67.9	62.4

### 👉 MSD : Lung tumor / spleen

Pre-training	Type	UNETR		SwinUNETR	
		Lung	Spleen	Lung	Spleen
Scratch	–	52.5	95.0	63.5	96.3
Tang <i>et al.</i> [25]	SSL	–	–	65.2	96.5
PrimGeoSeg	FDSL	<b>62.2</b>	<b>96.3</b>	<b>67.9</b>	<b>96.6</b>

### 👉 BraTS : Brain tumor

Pre-training	Type	UNETR				SwinUNETR			
		Avg.	ET	WT	TC	Avg.	ET	WT	TC
Scratch	–	88.1	84.8	91.3	88.1	90.0	86.8	<b>92.9</b>	90.3
PrimGeoSeg	FDSL	<b>88.7</b>	<b>85.6</b>	<b>91.8</b>	<b>88.9</b>	<b>90.3</b>	<b>87.0</b>	<b>92.9</b>	<b>91.0</b>

🔪 For further comparison, please see supplementary materials

## Experimental results: comparison with prior arts

### 👉 BTCV : 13 organs in abdomen

Pre-training	PT Num	Type	Avg.	Spl	RKid	LKid	Gall	Eso	Liv	Sto	Aor	IVC	Veins	Pan	rad	lad
<i>UNETR</i>																
Scratch	0	–	73.0	90.2	91.1	90.7	47.0	63.8	95.3	76.5	85.1	82.1	67.9	72.3	46.1	40.8
Chen <i>et al.</i> [6]	0.8K	SSL	75.8	95.2	<b>95.5</b>	93.8	51.9	52.3	<b>98.8</b>	80.0	87.8	82.7	66.1	68.9	60.8	51.3
PrimGeoSeg	0.8K	FDSL	77.4	88.9	94.0	93.8	59.8	65.7	95.4	79.3	88.3	82.6	69.9	76.8	58.5	53.3
PrimGeoSeg	50K	FDSL	<b>80.9</b>	<b>95.7</b>	94.2	<b>94.1</b>	<b>61.9</b>	<b>69.6</b>	96.7	<b>85.5</b>	<b>89.5</b>	<b>84.4</b>	<b>74.7</b>	<b>81.9</b>	<b>64.3</b>	<b>58.7</b>
<i>SwinUNETR</i>																
Scratch	0	–	78.3	92.3	93.2	93.8	55.9	61.3	94.0	77.0	87.5	80.4	74.2	76.1	68.8	63.6
Tang <i>et al.</i> [25]	5K	SSL	81.6	95.3	93.2	93.0	<b>63.6</b>	74.0	96.2	79.3	<b>90.0</b>	83.3	<b>76.1</b>	82.3	<b>69.0</b>	<b>65.1</b>
PrimGeoSeg	5K	FDSL	<b>82.0</b>	<b>95.7</b>	<b>94.4</b>	<b>94.4</b>	61.0	<b>75.5</b>	<b>96.7</b>	<b>83.3</b>	89.1	<b>85.6</b>	75.2	<b>84.3</b>	67.9	62.4

### 👉 MSD : Lung tumor / spleen

Pre-training	Type	UNETR		SwinUNETR	
		Lung	Spleen	Lung	Spleen
Scratch	–	52.5	95.0	63.5	96.3
Tang <i>et al.</i> [25]	SSL	–	–	65.2	96.5
PrimGeoSeg	FDSL	<b>62.2</b>	<b>96.3</b>	<b>67.9</b>	<b>96.6</b>

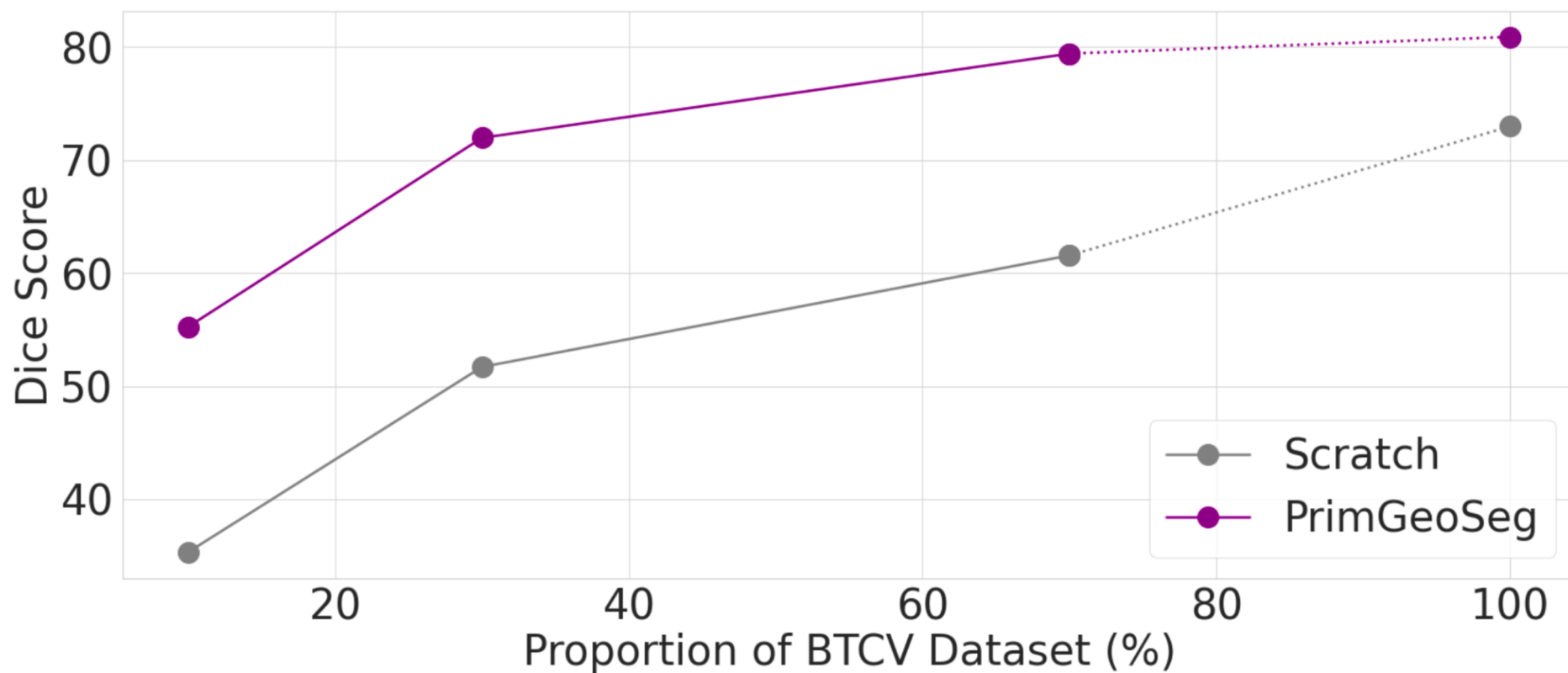
### 👉 BraTS : Brain tumor

Pre-training	Type	UNETR				SwinUNETR			
		Avg.	ET	WT	TC	Avg.	ET	WT	TC
Scratch	–	88.1	84.8	91.3	88.1	90.0	86.8	<b>92.9</b>	90.3
PrimGeoSeg	FDSL	<b>88.7</b>	<b>85.6</b>	<b>91.8</b>	<b>88.9</b>	<b>90.3</b>	<b>87.0</b>	<b>92.9</b>	<b>91.0</b>

🔪 For further comparison, please see supplementary materials

## Experimental results: comparison with prior arts

👉 On BTCV dataset : 13 organs in abdomen



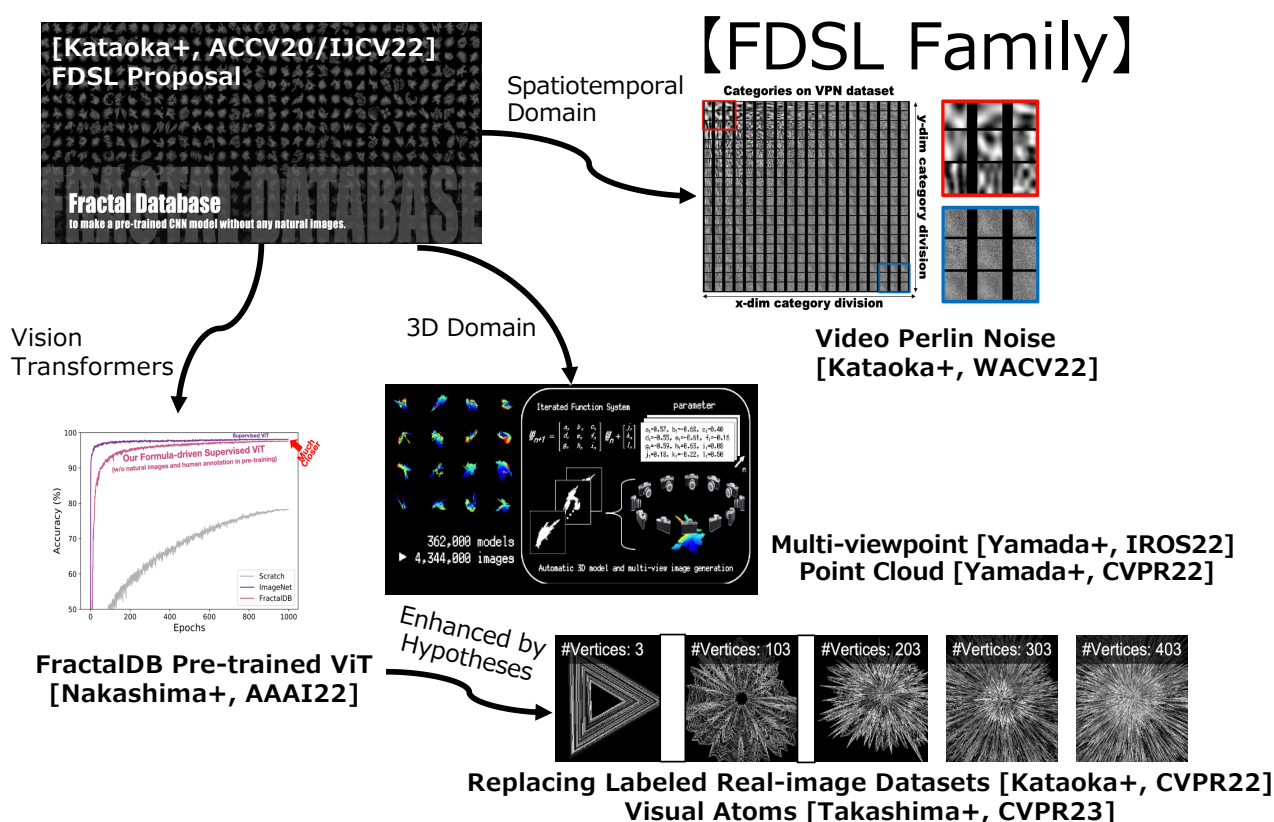
✓ PrimGeoSeg is more effective under limited data settings



# Lessons from the synthetic training projects

## Any task, any modality in limited synthetic data?

- Flexibly define the modality and its labels with generative models / mathematical formula
- Its possible to construct foundation models with very limited data



## Now added:

- Audio training [Shibata+, ICASSP25]
- Text-to-Image (On-going)
- Super resolution [Kodama+, CVPRW25]
- Multimodal - 2D image + 3D point cloud [Yamada+, ECCV24]
- Multimodal – 2D, 3D, and text (next slides)
- and other modalities / tasks

1

## Visual Pre-training with Minimal Data & Supervision

Can a natural law train a visual model?

- ACCV 2020 Best Paper Honorable Mention Award
- Featured in MIT Technology Review (Feb. 4<sup>th</sup>, 2021)
- One single synthetic data enables to pre-train ViT (ECCV24)

2

## Visual Foundation Models without Real Data

Can synthetic pre-training make a vision foundation model?

- Industry-focused vision foundation models (arXiv 2025)
- Primitive Geometry Segmentation for medical 3D data (BMVC 2023 Best Industry Paper Finalist)

3

## Multimodal AI Models with Generative Models

Can generative models make next foundation models?

- Zero-shot 3D understanding (CVPRW25 / on-going work)
- Leading research initiative (LIMIT.Lab with VGG community)



# **Text-guided Synthetic Geometric Augmentation for Zero-shot 3D Understanding**

**CVPRW 2025 / On-going work**

**Hirokatsu Kataoka**

National Institute of Advanced Industrial Science and Technology (AIST)

Visual Geometry Group, University of Oxford (Oxford VGG)

<http://www.hirokatsukataoka.net/>

## Towards multimodal AI models

We've proved that “primitive patterns” can make a vision foundation model

- Can generative models flexibly make next foundation models?
- How about a multimodal AI model? Zero-shot recognition in the wild?

## Generated 3D dataset for zero-shot understanding

“Wine bottle”  
text



2D image



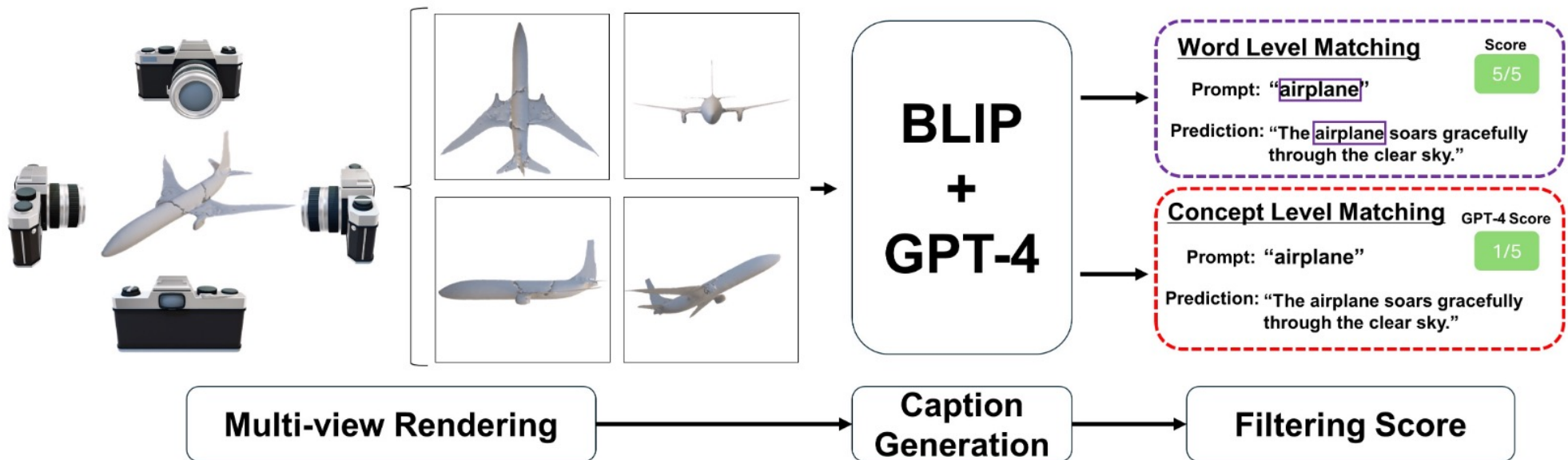
3D point cloud

It's not a positive way in the limited-resource concept, but this is the first step...

## Detailed pipeline

### Text-to-3D-to-Image

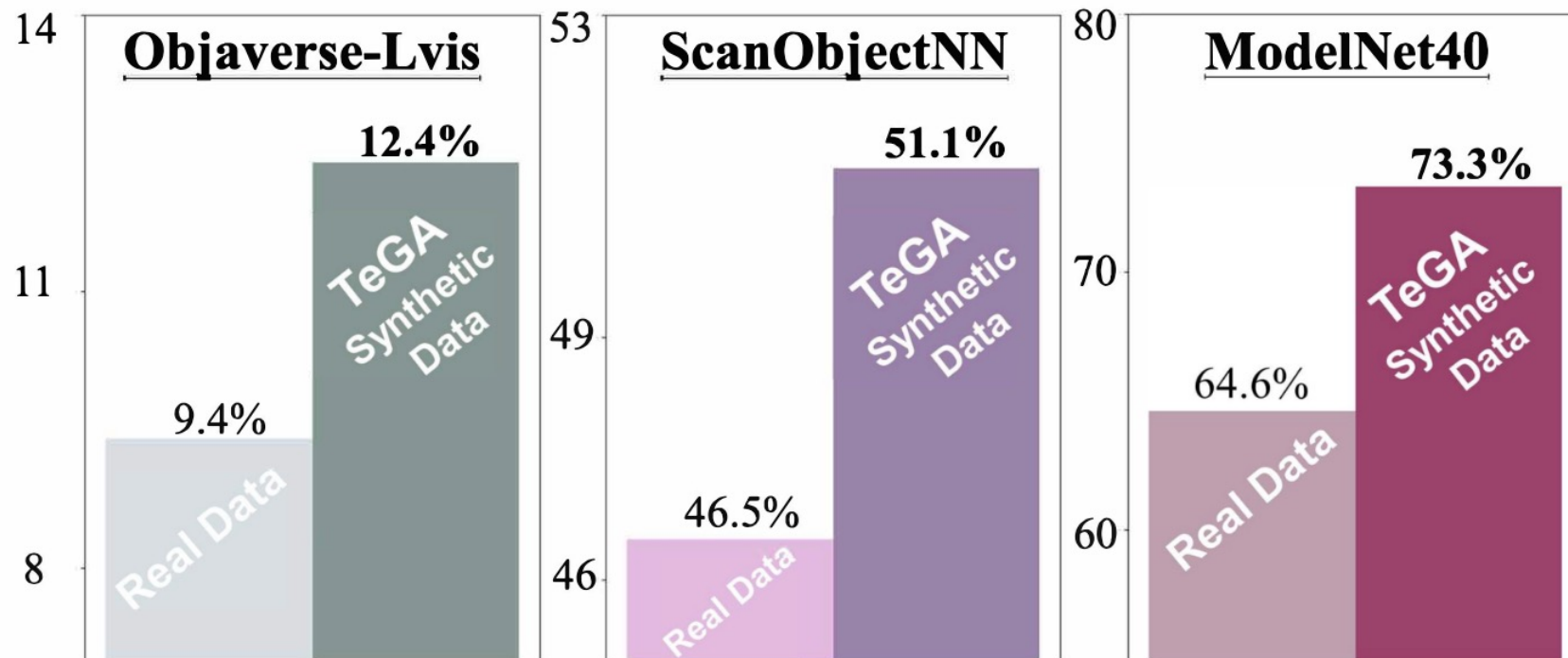
- Text-to-3D model: Point-E
- 3D dataset: ShapeNet
- 2D images: multi-view rendering
- Text: BLIP + GPT-4



## Main results

### With and without synthetic 3D data (TeGA)

- Models: 2D-3D-Text models (MixCon3D)
- Zero-shot 3D understanding on three datasets (Objaverse, ScanObjectNN, ModelNet)

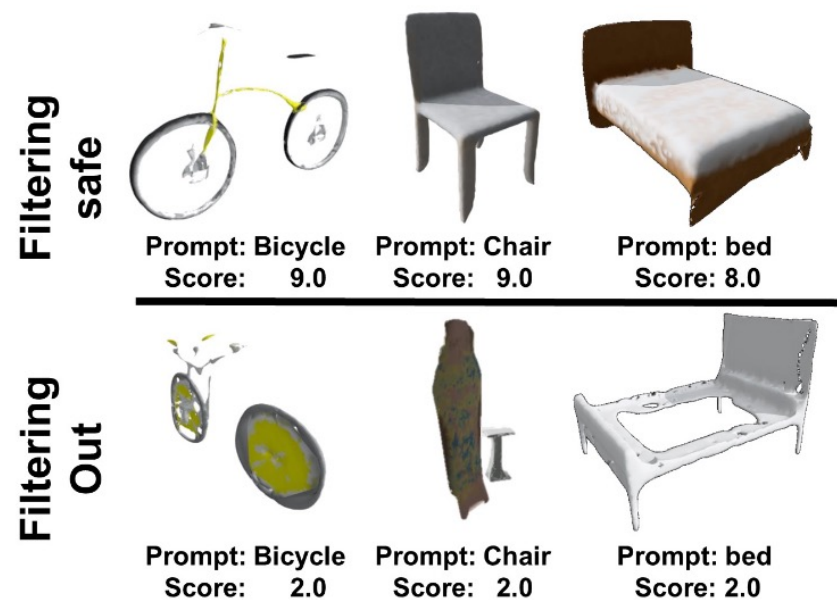
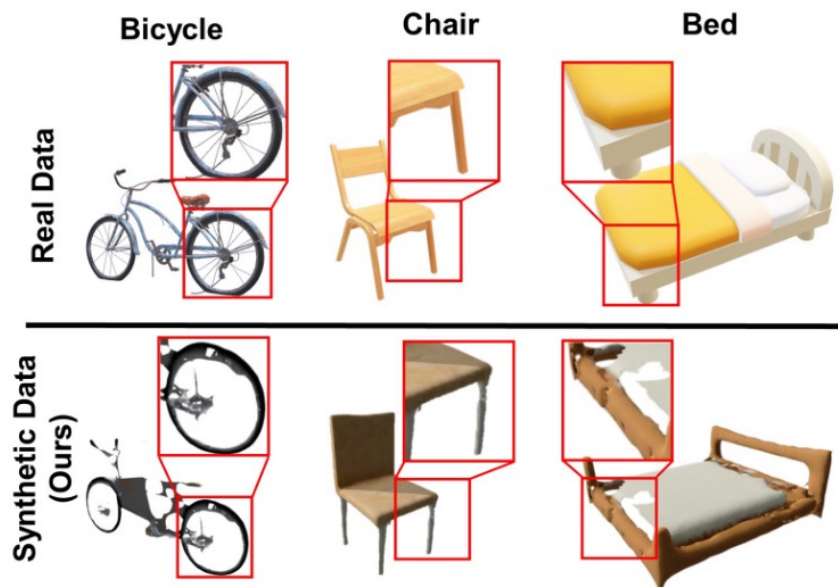


Not a perfect method for now, but its reaching latest levels in zero-shot 3D understanding

## Dataset analysis

### Generated dataset by Point-E

- ShapeNet vs. Generated data (left fig)
- Data filtering: IN & OUT (right fig)



# **LIMIT.Lab**

**-Research initiative-**

**Multimodal AI Foundation Models with Very Limited Resources**

**Hirokatsu Kataoka**

National Institute of Advanced Industrial Science and Technology (AIST)

Visual Geometry Group, University of Oxford (Oxford VGG)

<http://www.hirokatsukataoka.net/>

# LIMIT.Lab

## Research initiative w/ VGG community

【LIMIT.Community / LIMIT.Lab】



Community => LIMIT.Community

- 100+ researchers / students
- LIMIT Workshops @ ICCV25&CVPR24

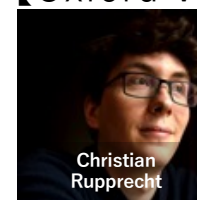
Research Lab => LIMIT.Lab

- 🇯🇵 AIST
- 🇬🇧 Oxford VGG, Cambridge VSL
- 🇩🇪 UTN FunAI Lab
- 🇳🇱 UvA VISLab

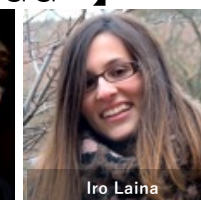
## I'm leading this community-driven research!

- 10-year experience in Japanese community-driven research
- We enhance the CV researches w/ world-wide talent starting from VGG community?

【Oxford VGG 🇬🇧】



Christian  
Rupprecht



Iro Laina

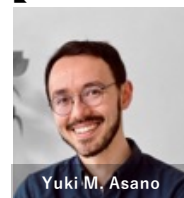


Academic Visitor

Hirokatsu  
Kataoka

+ Postdocs, Ph.D. students at VGG

【UTN FunAIlab 🇩🇪 / former UvA VISLab 🇳🇱】

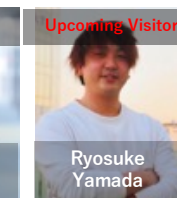


Yuki M. Asano



Co-supervisor

Hirokatsu  
Kataoka



Upcoming Visitor

Ryosuke  
Yamada

+ Postdocs, Ph.D. students at FunAIlab

【Cambridge VSL 🇬🇧】



Elliott Wu

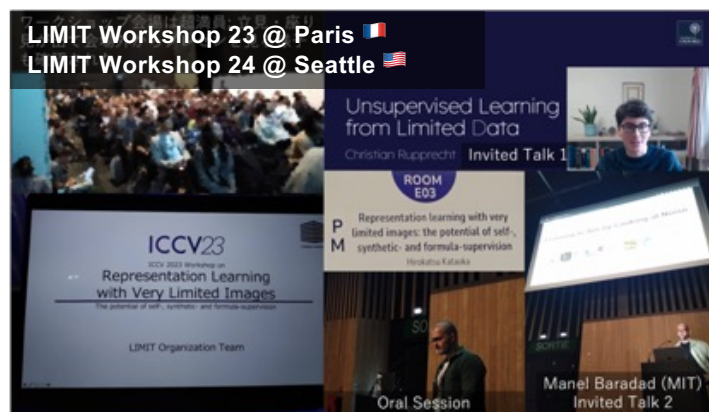
+ Upcoming Ph.D. students at VSL



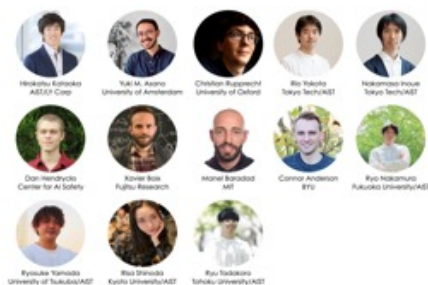
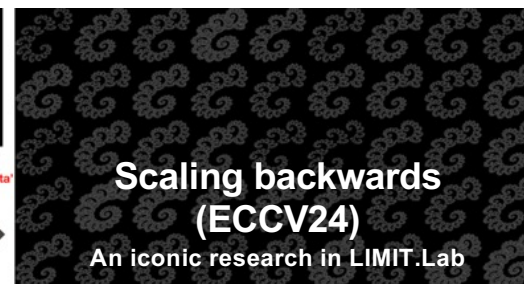
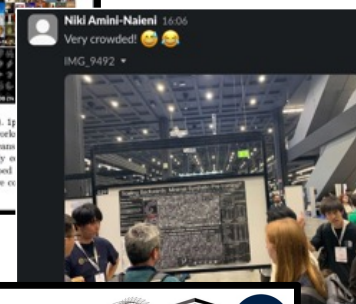
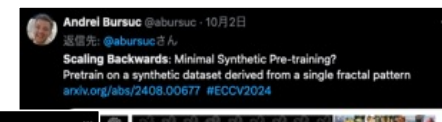
# LIMIT.Lab, so far

## As a stealth mode,

- Organizing several workshops / networking
- Conceptual researches in limited resources
- CVPR 2025 report



## ECCV24 @ Milano



LIMIT Workshop Organizers



FunAI seminar @ Nuremberg



VGG seminar @ Oxford



LIMIT.Lab, from now!

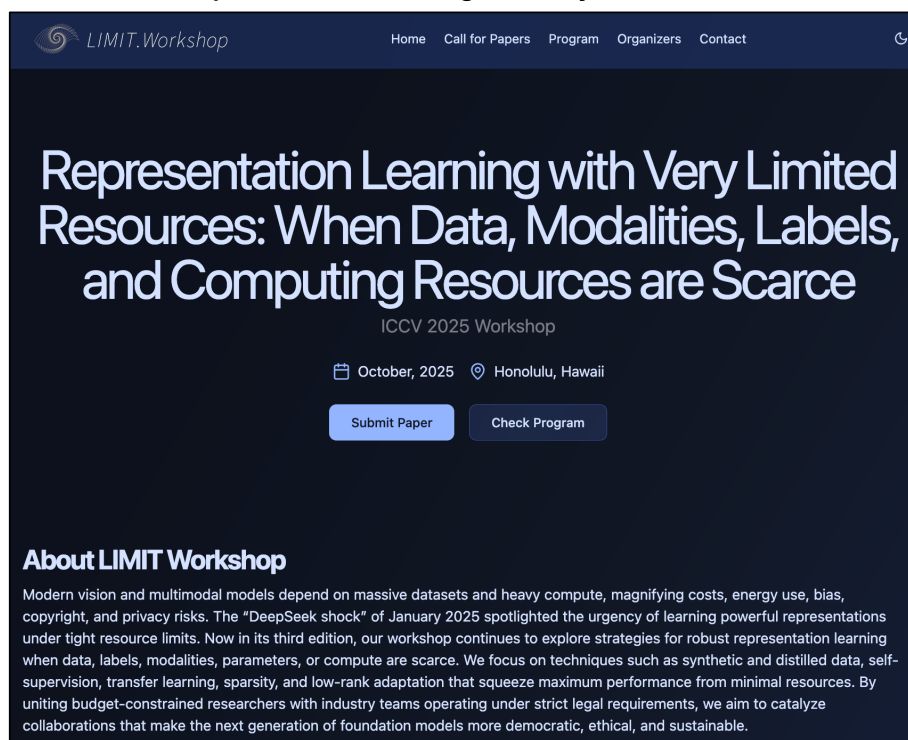
## LIMIT researches in multiple aspects

- 1. Minimalist 3D foundation models (like VGGT) for minimal representation?
- 2. Synthetic data for zero-shot understanding in the wild?
- 3. Multipurpose Transformer pre-training?

## Concept switching, from “LIMIT” to “FOUND” FOUNDation Data

### Two accepted organizing workshops at ICCV 2025

#### LIMIT: Representation Learning with Very Limited Resources



The screenshot shows the LIMIT Workshop website. The header includes the LIMIT Workshop logo and navigation links: Home, Call for Papers, Program, Organizers, and Contact. The main content area features the title "Representation Learning with Very Limited Resources: When Data, Modalities, Labels, and Computing Resources are Scarce" in large white text on a dark blue background. Below the title, it says "ICCV 2025 Workshop" and provides the date "October, 2025" and location "Honolulu, Hawaii". There are two buttons: "Submit Paper" and "Check Program". At the bottom, there is an "About LIMIT Workshop" section with a paragraph of text.

Representation Learning with Very Limited Resources: When Data, Modalities, Labels, and Computing Resources are Scarce

ICCV 2025 Workshop

October, 2025 Honolulu, Hawaii

Submit Paper Check Program

**About LIMIT Workshop**

Modern vision and multimodal models depend on massive datasets and heavy compute, magnifying costs, energy use, bias, copyright, and privacy risks. The “DeepSeek shock” of January 2025 spotlighted the urgency of learning powerful representations under tight resource limits. Now in its third edition, our workshop continues to explore strategies for robust representation learning when data, labels, modalities, parameters, or compute are scarce. We focus on techniques such as synthetic and distilled data, self-supervision, transfer learning, sparsity, and low-rank adaptation that squeeze maximum performance from minimal resources. By uniting budget-constrained researchers with industry teams operating under strict legal requirements, we aim to catalyze collaborations that make the next generation of foundation models more democratic, ethical, and sustainable.

<https://iccv2025-limit-workshop.limitlab.xyz/>

#### FOUND: Foundation Data for Industrial Tech Transfer



The screenshot shows the FOUND Workshop website. The header includes the FOUND 2025 logo and navigation links: Home, Call for Papers, Program, Organizers, and Contact. The main content area features the title "Foundation Data for Industrial Tech Transfer" in large blue text on a light blue background. Below the title, it says "ICCV 2025 Workshop" and provides the date "October, 2025" and location "Honolulu, Hawaii". There are two buttons: "Submit Paper" and "Check Program". At the bottom, there is an "About FOUND Workshop" section with a paragraph of text.

Foundation Data for Industrial Tech Transfer

ICCV 2025 Workshop

October, 2025 Honolulu, Hawaii

Submit Paper Check Program

**About FOUND Workshop**

Recently, Transformer-based foundation models have achieved outstanding performance across a broad spectrum of benchmarks spanning recognition and generation tasks, and their versatility has fueled rapid advances in both AI research and industrial deployment. To seamlessly adapt these models to downstream tasks in diverse real-world domains-including medicine, manufacturing, robotics, and the creative industries-and thereby deliver tangible impact on human life, it is indispensable to develop Tech Transfer technologies that encompass domain-specific fine-tuning and robust MLOps pipelines, where the decisive factor is the breadth and quality of data available for those tasks. At the same time, model evaluation is approaching saturation on conventional IID benchmarks, prompting growing calls to redesign evaluation metrics and tasks that dispense with the IID assumption and explicitly capture out-of-distribution (OOD) and long-tail phenomena. Advancing both (i) Tech Transfer to heterogeneous downstream tasks and (ii) the definition of next-generation evaluation criteria therefore hinges on curating and exploiting broader and deeper data resources-Foundation Data, as we term them. Against this backdrop, the ICCV 2025 workshop “FOUND: Foundation Data for Industrial Tech Transfer” will convene researchers from industry and academia to share techniques for realizing Foundation Data and to engage in comprehensive discussions on model adaptation and the design of novel evaluation tasks grounded in such data, with the ultimate aim of opening new horizons for AI research and application.

<https://iccv2025-found-workshop.limitlab.xyz/>

- Good connection from LIMIT to FOUND in terms of academic/industry, research community, & tech transfer with AI models
- LIMIT: Building multimodal AI foundation models with very limited resources
- FOUND: Foundation data for the last-mile industrial tech transfer

**A better academic research, a better tech transfer in the concept of LIMIT**