Building Vision Foundation Models with Synthetic Data

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)

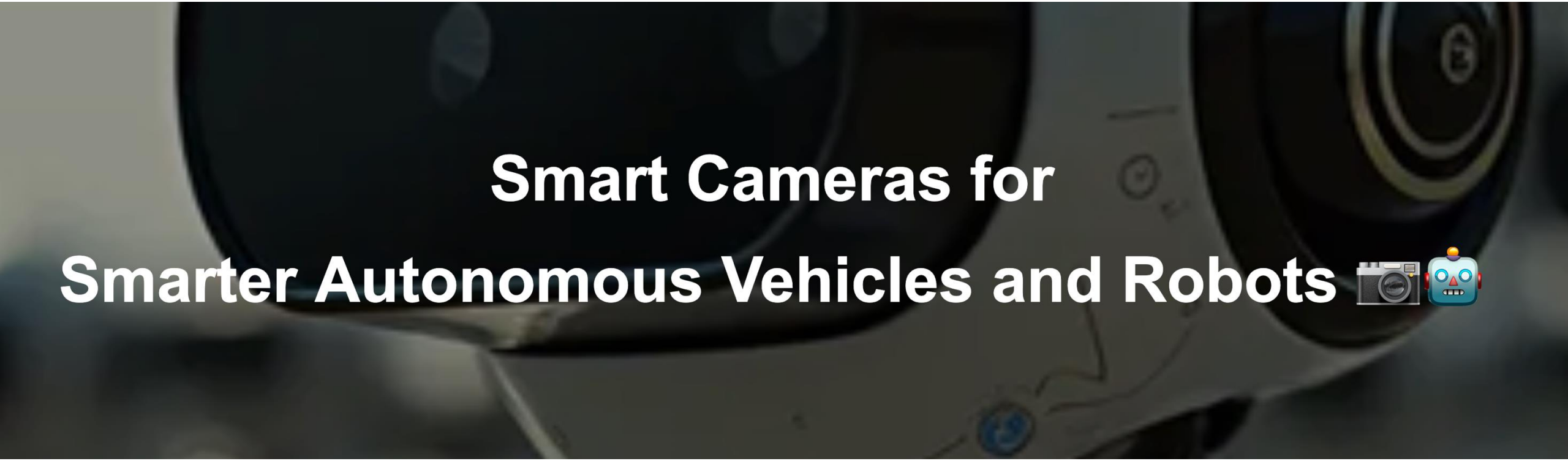
Recently Selected Projects (within 3 years) :

“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）

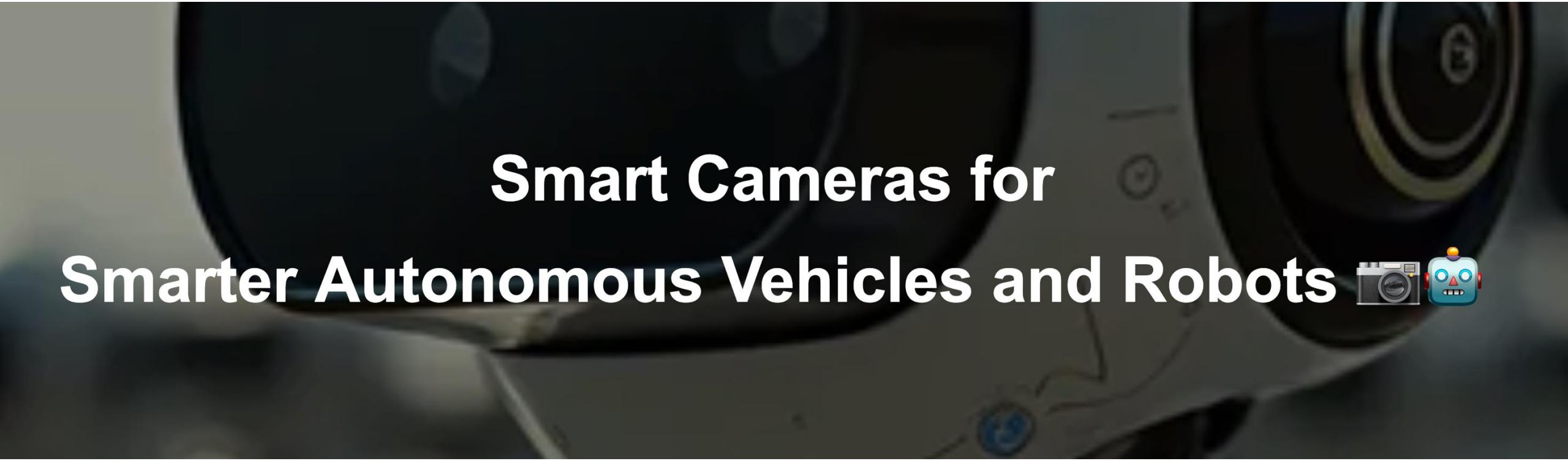
http:// hirokatsukataoka.net/

Today’s topic



https://supercamerai.github.io/

Today’s topic

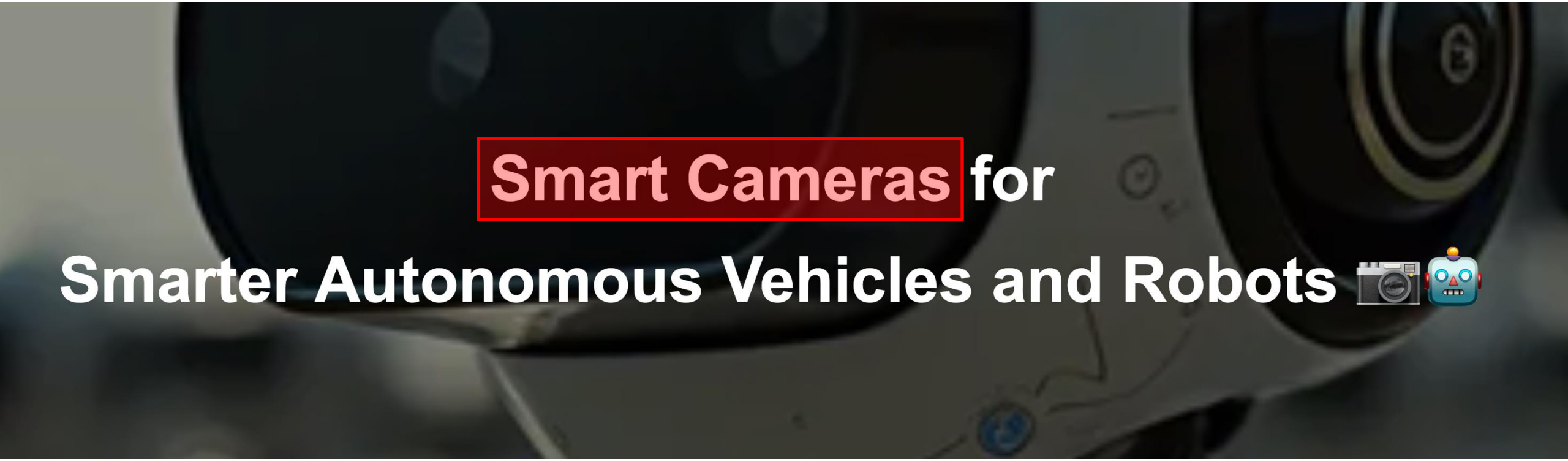


https://supercamerai.github.io/

OK, let me consider the workshop title 🤔

I don’t have any work at “pure camera”

Today's topic



https://supercamerai.github.io/

How about dividing this words into "imaging" and "understanding"??

Imaging

Discussion from GPT-5.1

RGB camera, image sensor, CMOS sensor, CCD sensor, camera lens, focal length, aperture, shutter speed, ISO sensitivity, exposure control, white balance, color calibration, color correction matrix, lens distortion correction, radial distortion, tangential distortion, vignetting correction, camera calibration, intrinsic parameters, extrinsic parameters, sensor noise, shot noise, read noise, dynamic range, HDR imaging, multi-exposure fusion, demosaicing, denoising, image sharpening, deblurring, motion deblurring, focus stacking, depth of field, aberration correction, chromatic aberration, spherical aberration, modulation transfer function (MTF), optical transfer function (OTF), image stabilization, optical image stabilization (OIS), electronic image stabilization (EIS), rolling shutter correction, global shutter, raw image processing, Bayer pattern, gamma correction, tone mapping, burst photography, multi-frame super-resolution, single-image super-resolution, image interpolation, frame interpolation, optical flow estimation, exposure bracketing, gradient-domain fusion, intrinsic image decomposition, shape-from-shading, photometric stereo, depth from defocus, depth from focus, coded aperture imaging, computational lens, synthetic aperture imaging, plenoptic imaging, light-field capture, image fusion, low-light imaging, dark frame subtraction, dehazing, deraining, color constancy, retinex-based enhancement, neural image signal processing (Neural ISP), learned demosaicing, learned denoising, neural HDR reconstruction, computational refocusing, digital zoom, sensor pixel binning, sensor readout pipeline, rolling shutter temporal modeling, radiometric calibration, geometric calibration, camera response function (CRF), tone reproduction, high-fidelity color imaging, lens flare correction, glare removal, reflection removal, illumination estimation.

indicates papers from my publications/experience

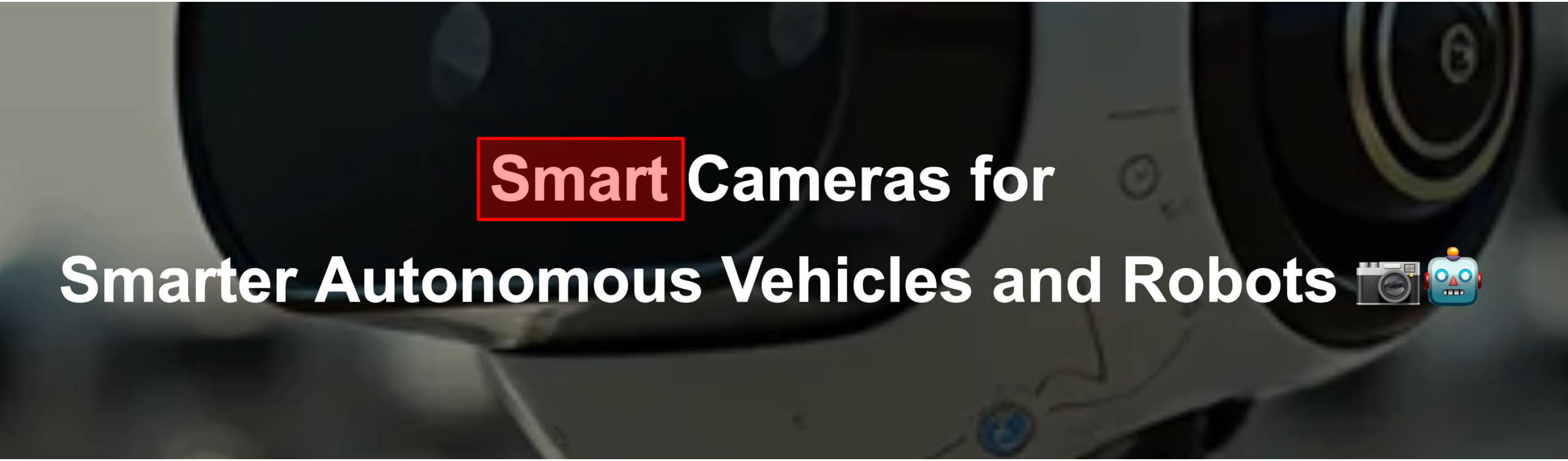
Understanding

Discussion from GPT-5.1

Image Classification, Object Detection, Instance Segmentation, Semantic Segmentation, Keypoint Detection, Human Pose Estimation, Facial Recognition, Fine-grained Classification, Part-based Recognition, Attribute Recognition, Salient Object Detection, Visual Relationship Detection, Scene Parsing, Scene Classification, Scene Graph Generation, Visual Grounding, Action Recognition, Temporal Action Segmentation, Video Classification, Spatiotemporal Understanding, Video Reasoning, Video Question Answering, Video Captioning, Event Detection, Action Anticipation, Future Prediction, Trajectory Prediction, Multi-view Understanding, 3D Object Detection, 3D Instance Segmentation, Point Cloud Classification, Shape Completion, Depth Estimation, Multi-view Stereo, Neural Radiance Fields, 3D Reconstruction, 4D Reconstruction, 4D Scene Understanding, Motion Understanding, Object Tracking, Multi-object Tracking, SLAM, Visual Navigation, Embodied AI, Manipulation Understanding,, Visual Question Answering (VQA), Visual Reasoning, Causal Reasoning, Cross-modal Reasoning, Vision-Language Understanding, Vision-Language Grounding, CLIP-based Understanding, Representation Learning, Self-supervised Learning, Contrastive Learning, Masked Image Modeling, Visual Similarity Learning, Metric Learning, Zero-shot Recognition, Few-shot Recognition, Domain Adaptation, Domain Generalization, Out-of-distribution Detection, Robust Recognition, Generative Understanding, Image Captioning, Dense Captioning, Image-to-Text Retrieval, Text-to-Image Retrieval, Multimodal Fusion, Multimodal Representation, Attention-based Understanding, World Model Learning, Spatial Understanding, Temporal Understanding, Part-based Understanding, Compositional Understanding, Causal Understanding, High-level Semantic Understanding.

indicates papers from my publications/experience

Today's topic



https://supercamerai.github.io/

Yes, consider how to make cameras smarter

Topics and research questions

1 Visual Foundation Model without Real Data

Can synthetic pre-training make a vision foundation model?

- Industrial synthetic segment pre-training (arXiv 2025)

2 Vision Foundation Model with Generative Models

Can generative models make next foundation models?

- S3OD: Towards Generalizable Salient Object Detection with Synthetic Data (arXiv 2025)

We’re using synthetic data for a smart camera

Industrial Synthetic Segment Pre-training

arXiv: 2505.13099

**Shinichi Mae*, Hirokatsu Kataoka*, Ryousuke Yamada,
Yoshihiro Fukuhara, Risa Shinoda, Christian Rupprecht**

AIST / Oxford VGG

* indicates equal contribution

VFM under synthetic data & limited resources

Research questions:

How can we surpass large-scale models such as SAM;

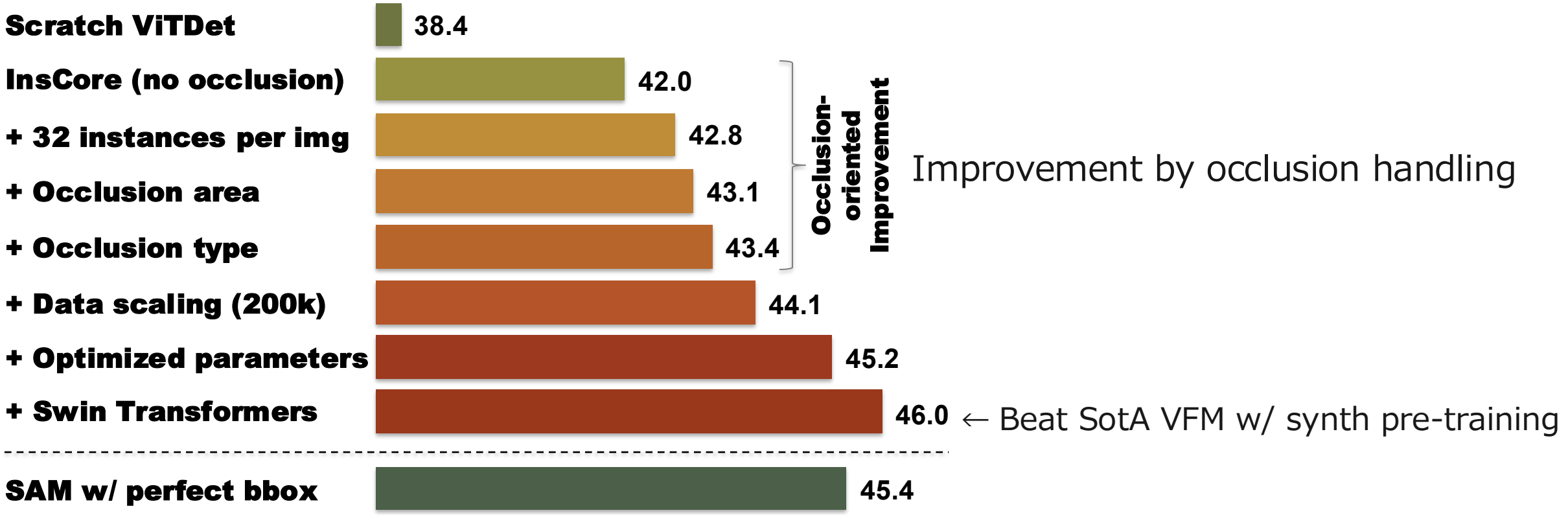
- Using only synthetic data for pre-training?
- Without human labels and real images?

Segment Anything Model (SAM) by Meta

→ General segmentation model with 11M images and 1B masks

[Kirillov et al. ICCV23]

From the results of InsCore vs SAM/SAM2: What we observed…

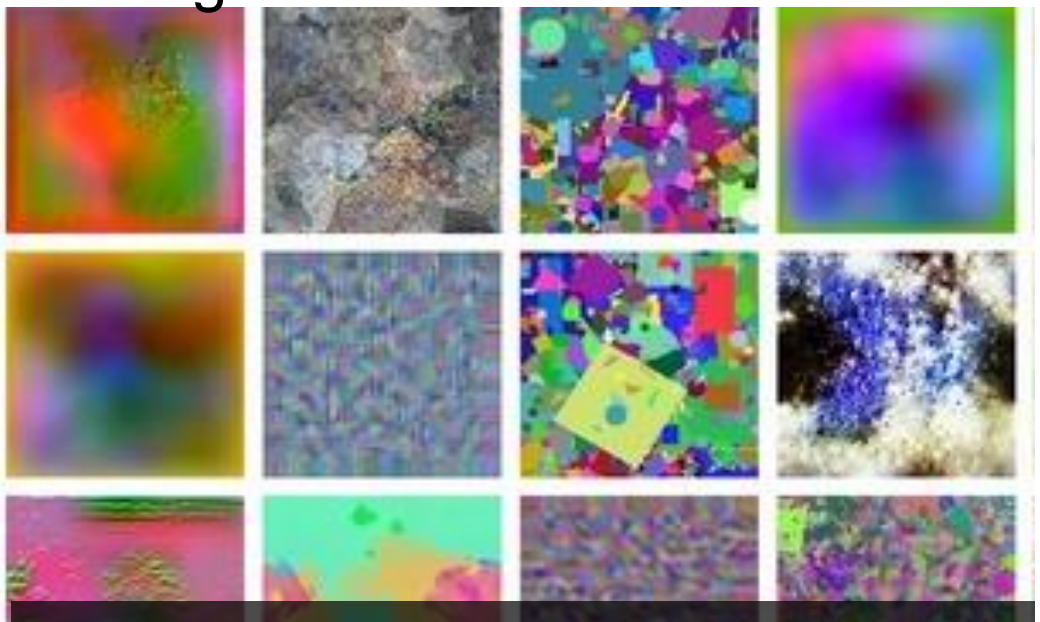


Improving InsCore led to performance surpassing SAM (and even SAM2)

→【Key Insight】Not appearance but occlusion handling is essential / Occlusion structures can be systematically generated in synthetic data

Inspired by related work & industrial data

Learning visual elements from noise:



Learning to See by Looking at Noise [Baradad+, NeurIPS21] / **Shader-1k** [Baradad+, NeurIPS22]

Antonio Torralba

Phillip Isola

Appearance of industrial data is more like:

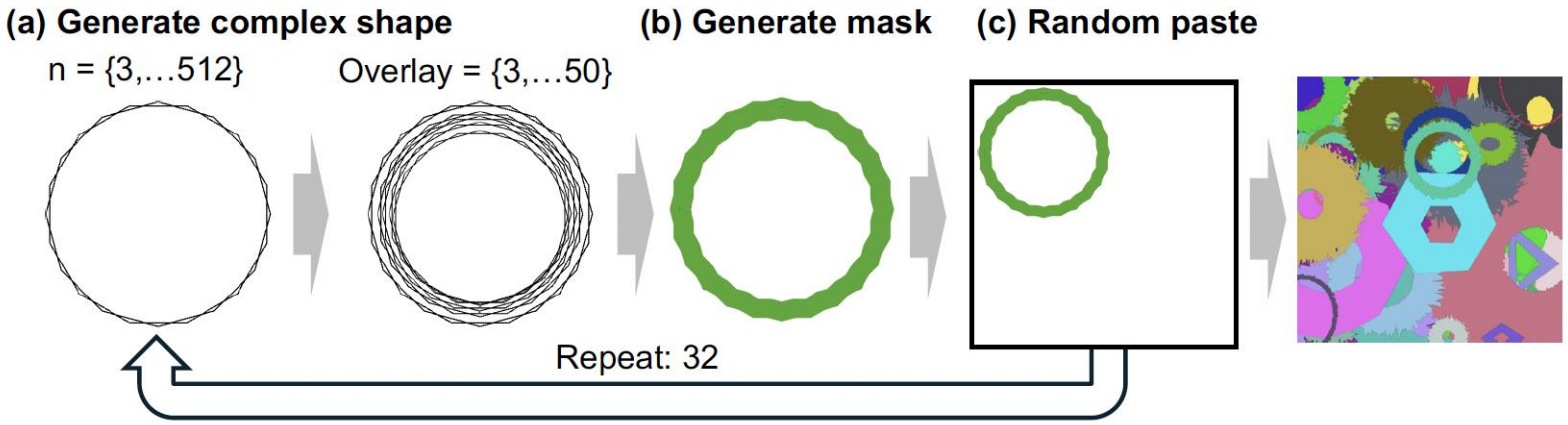
https://robot-mujinka.com/wp-content/uploads/2019/09/bara.jpg

Complex occlusions and dense configurations
→ Critical factors for VFM & industrial segmentation

Occlusion-aware synthetic pre-training

Instance Core: Combined shape contours with formula-generated images

Generates contour-based complex shapes, places many objects to induce heavy occlusion
Highly effective for learning segmentation in complex overlapping structures

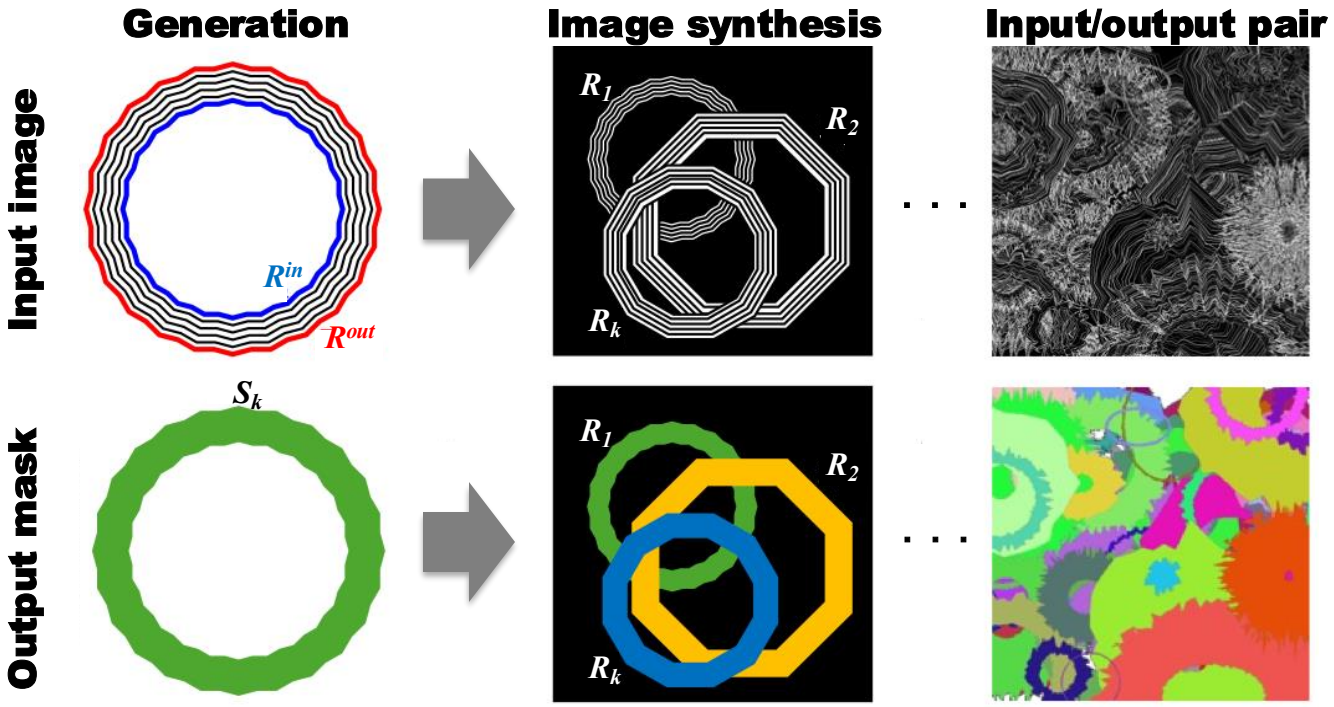


- Formula-generated shape contours (32 per image)
- Occlusion-oriented locations

Occlusion-aware synthetic pre-training

Instance Core: Combined shape contours with formula-generated images

Generates contour-based complex shapes, places many objects to induce heavy occlusion
Highly effective for learning segmentation in complex overlapping structures



- Input/output pairs with complex occlusions
- Transformer learns to map contour images to GT masks

Human annotator is nearly impossible for this complexity

Qualitative evaluation: Industrial datasets

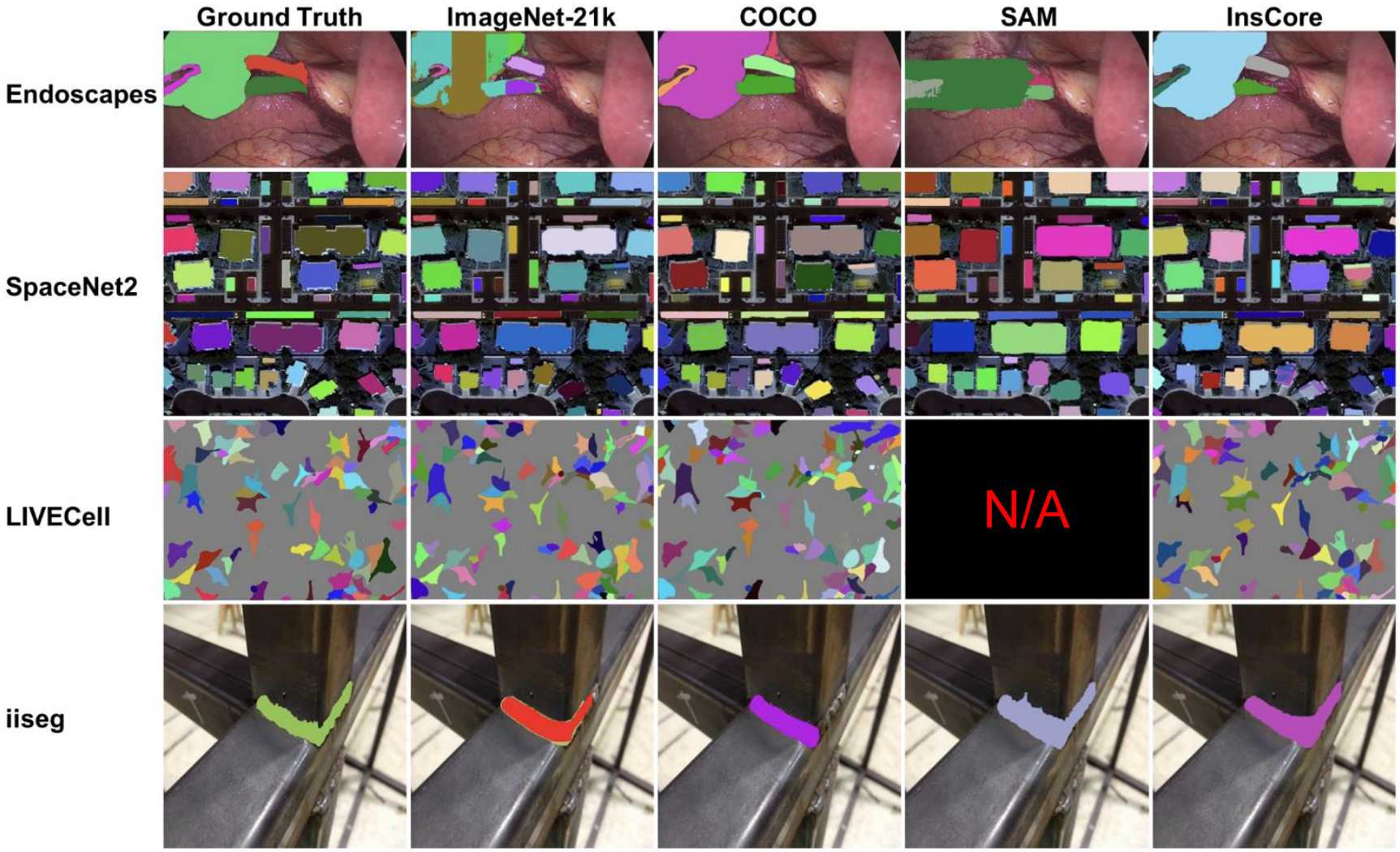
Five industrial scenarios selected:

- Medical · biomedical · remote sensing · manufacturing · logistics
- Datasets chosen for high occlusion and dense object layouts

Evaluation dataset	Industrial domain	#Train set		#Test images		#Categories
		#Images	#Masks	#Images	#Masks	
Endoscapes (ES) [27]	Medical	343	1,615	74	270	6
LIVECell (LC) [10]	Biomedical	3,253	1,018,576	1,564	462,261	1
SpaceNet2 (SN) [40]	Remote sensing	3,080	87,301	771	21,641	1
Industrial-iSeg (IS) [20]	Manufacturing	1,109	25,308	89	523	6
LogiSeg (LS) [24]	Logistics	1,384	10,018	300	2,093	7

Qualitative evaluation

Visual comparisons across datasets: SAM (previous) & InsCore (proposal)



The InsCore pre-trained segmentation surpassed SAM!

Baseline construction for SAM / SAM2

SAM / SAM2 receive complete bbox visual prompts

SAM / SAM2

Starting segmentation with image + bbox prompts

InsCore(ours)

Starting segmentation with image only

SAM / SAM2 start with a favorable advantage, InsCore starts without any boxes

Baseline construction for SAM / SAM2

SAM / SAM2 receive complete bbox visual prompts

SAM / SAM2

InsCore(ours)

Model	PT	Backbone	Head	ES	LC	SN	II	LS	Average	Desc.
SAM	SA-1B	ViT-B	MRCNN	3.1	0.6	59.7	30.5	14.3	21.6	No prompting
SAM	SA-1B	ViT-B	GT Bbox	48.7	9.3	52.3	36.4	80.3	45.4	w/ perfect bbox & noise (1 - 3 px)
SAM	SA-1B	ViT-B	GT Bbox	48.2	7.6	50.0	30.5	79.8	43.2	w/ perfect bbox & noise (1 - 5 px)
SAM	SA-1B	ViT-B	GT Bbox	44.1	2.2	34.4	10.3	75.7	33.4	w/ perfect bbox & noise (5 - 10 px)
SAM	SA-1B	ViT-B	GT Bbox	27.5	0.3	3.4	3.6	56.7	18.3	w/ perfect bbox & noise (10 - 30 px)
SAM2	SA-1B+V	ViT-B	GT Bbox	56.4	9.9	52.2	32.8	71.6	44.5	w/ perfect bbox & noise (1 - 3 px)
SAM2	SA-1B+V	ViT-B	GT Bbox	56.3	8.0	50.0	27.0	71.3	42.5	w/ perfect bbox & noise (1 - 5 px)

SAM / SAM2 start with a favorable advantage, InsCore starts without any boxes

Baseline construction for SAM / SAM2

SAM / SAM2 receive complete bbox visual prompts

SAM / SAM2

InsCore(ours)

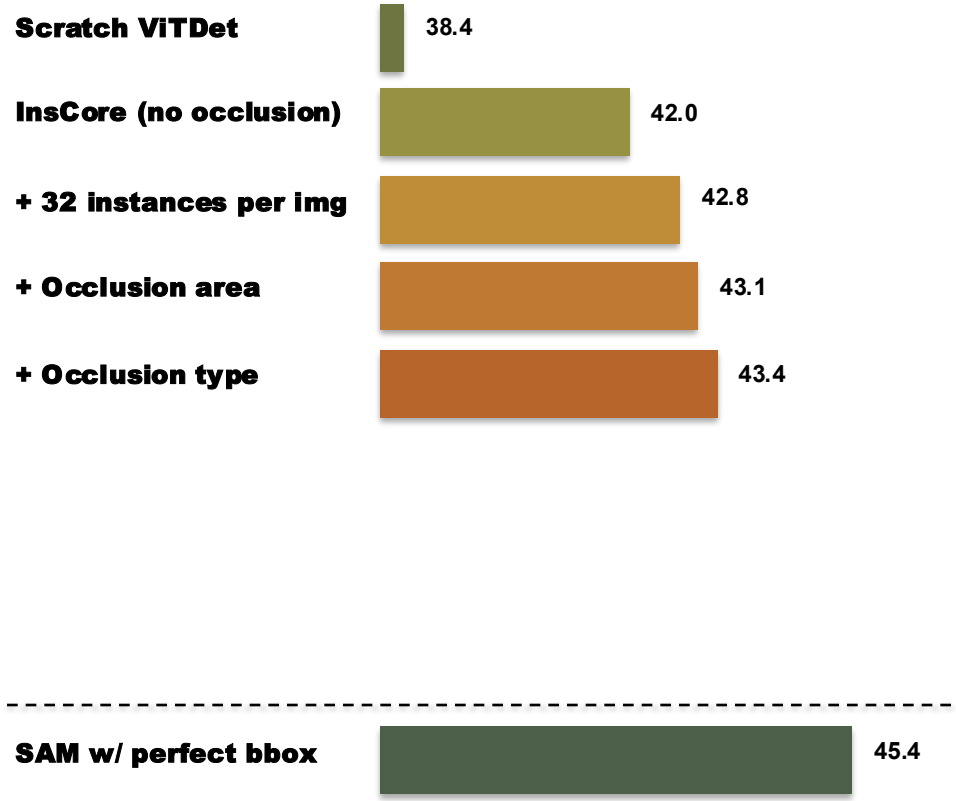
SAM / SAM2 start with a favorable advantage, InsCore starts without any boxes

InsCore performance improvement

Occlusion-aware enhancements boost scores across datasets

#ins	ES	LC	SN	IS	LS	Ave
1	22.1	11.5	61.9	21.3	93.2	42.0
2	22.1	11.7	62.2	22.3	93.8	42.4
4	22.9	12.3	62.0	21.9	93.9	42.6
8	23.5	11.7	**62.3**	22.3	**94.5**	42.8
16	22.5	12.7	61.9	21.2	93.8	42.4
32	22.1	13.4	**62.3**	**22.4**	93.8	42.8
64	**24.1**	**13.6**	**62.3**	21.5	**94.4**	**43.1**

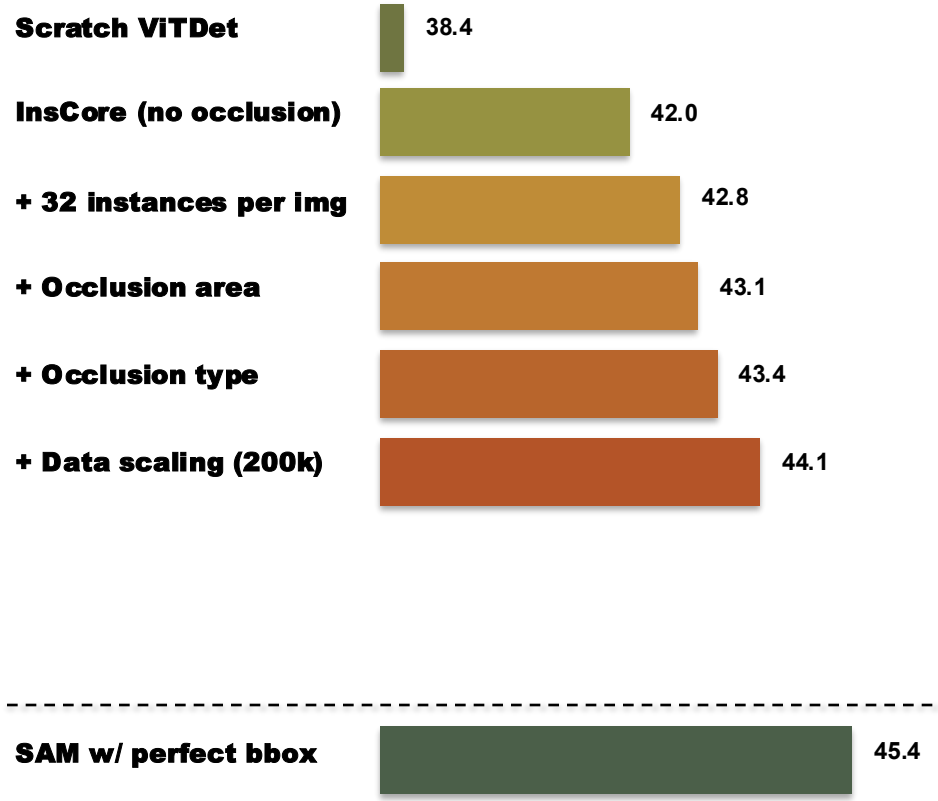
Type	ES	LC	SN	IS	LS	Ave
Random	**24.8**	**13.8**	62.3	22.7	93.6	**43.4**
Grid	22.9	12.4	62.3	22.3	93.6	42.7
Gaussian	23.4	12.3	62.1	22.7	94.2	42.9
Poisson	23.0	12.0	**62.4**	**23.6**	93.7	42.9
Cluster	22.9	11.8	**62.4**	23.3	93.5	42.7
Spiral	23.5	12.4	62.1	21.9	94.0	42.7
Mixed all	23.1	12.3	62.3	23.2	**94.4**	43.1



InsCore performance improvement

Data scaling effects: Larger synthetic datasets yield higher scores

#Images	ES	LC	SN	IS	LS	Ave
12.5k	**24.9**	**13.8**	63.0	23.3	94.3	43.8
25k	22.4	12.8	63.2	**25.2**	**94.9**	43.7
50k	24.4	13.3	63.2	24.4	94.2	43.9
100k	24.8	**13.8**	62.3	22.7	93.6	43.4
200k	24.4	12.7	**63.4**	**25.2**	**94.9**	**44.1**
400k	24.2	12.6	63.2	24.0	94.2	43.6



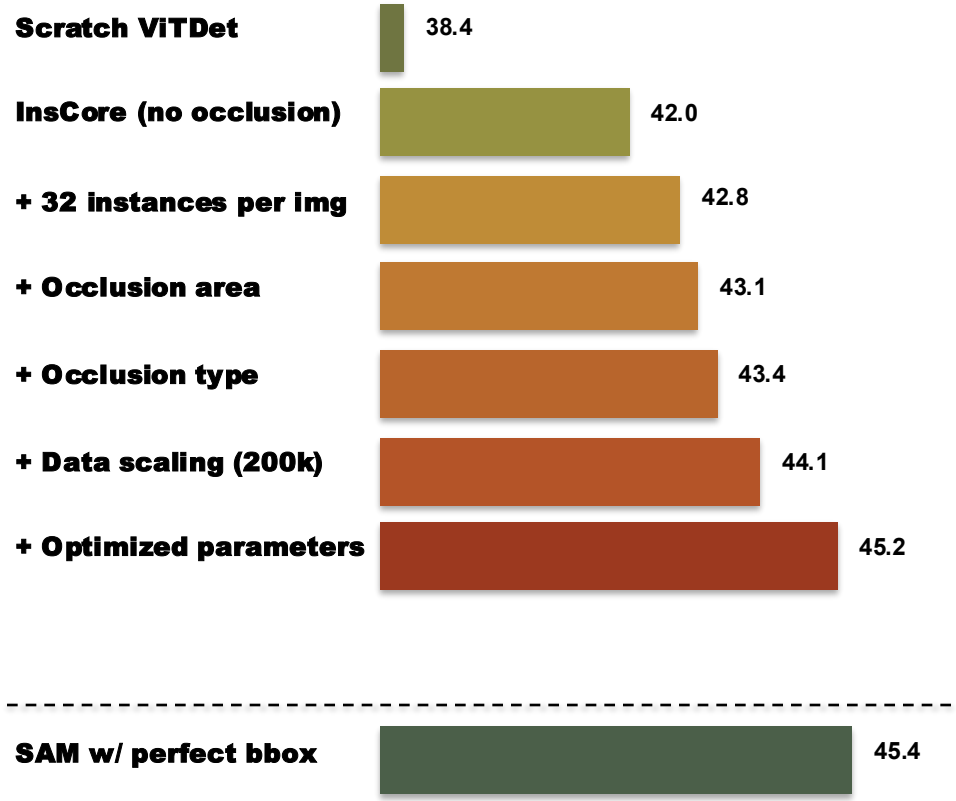
InsCore performance improvement

Parameter tuning effects: Batch size x epochs adjustments further improve scores

#Batch	ES	LC	SN	IS	LS	Average
8	**25.6**	13.1	**63.1**	**25.5**	94.6	**44.3**
16	24.3	12.5	**63.1**	23.4	94.5	43.5
32	24.8	13.1	**63.1**	24.8	**94.8**	44.1
64	23.2	**13.2**	62.6	23.4	94.2	43.3

#Epochs	ES	LC	SN	IS	LS	Ave
0	16.6	11.9	59.5	12.1	92.2	16.6
50	24.1	13.6	62.3	21.5	94.4	43.1
100	23.2	13.2	62.6	23.4	94.2	43.3
200	**24.8**	**13.4**	**63.2**	**26.1**	**94.9**	**44.5**

#Epochs	ES	LC	SN	IS	LS	Ave
60	**24.8**	13.4	63.2	**26.1**	94.9	**44.5**
100	24.1	**14.8**	**63.5**	24.4	95.1	44.3
200	24.0	**14.8**	63.2	24.6	**95.2**	44.3



InsCore performance improvement

Comparison with other pre-training methods on ViT-B backbone:

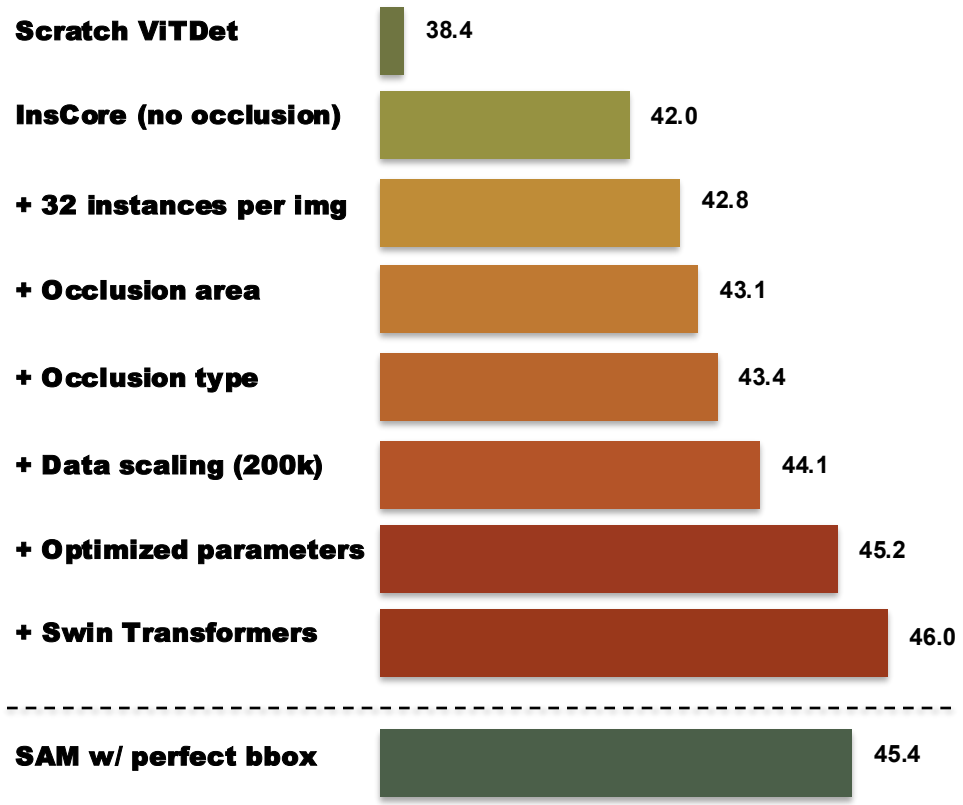
InsCore outperforms other pre-training methods with real images / synthetic images

PT data	Backbone	Image type	#PT images	ES	LC	SN	IS	LS	Ave
Scratch	ViT-B	–	–	16.6	11.9	59.5	12.1	92.2	38.4
ImageNet-21k [8]	ViT-B	Real	14M	**29.1**	**16.4**	61.1	24.1	94.2	**45.0**
SegRCDB [33]	ViT-B	Synthetic	0.1M	23.0	13.0	61.3	21.9	94.1	42.6
Hypersim [32]	ViT-B	Synthetic	0.07M	22.0	**15.7**	60.2	20.9	94.3	42.6
Virtual KITTI [4]	ViT-B	Synthetic	0.02M	19.1	14.7	59.9	17.3	94.3	41.1
SAIL-VOS [13]	ViT-B	Synthetic	0.1M	23.0	14.2	59.9	18.4	**95.2**	42.1
InsCore	ViT-B	Synthetic	0.1M	24.8	13.4	**63.2**	**26.1**	94.9	44.5
InsCore*	ViT-B	Synthetic	0.2M	**25.6**	15.6	**63.5**	**26.1**	**95.6**	**45.2**

InsCore performance improvement

Enhancement with SwinTransformer

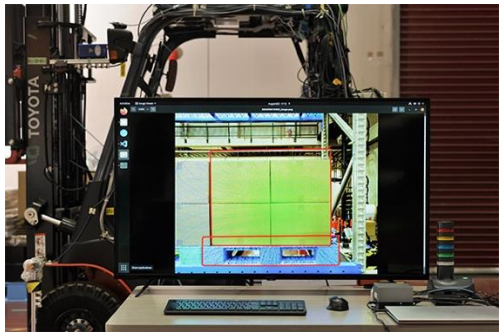
Method	Backbone	Fine-tuning (mAP)					Ave
		ES	LC	SN	IS	LS	
SAM [19]	ViT-B	48.7	9.3	52.3	**36.4**	80.3	45.4
SAM2 [31]	ViT-B	**56.4**	9.9	52.2	32.8	71.6	44.5
InsCore	ViT-B	25.6	15.6	**63.5**	26.1	**95.6**	45.2
InsCore	Swin-B	29.7	**18.3**	61.6	25.5	95.1	**46.0**



Synthetic pre-training exceeds scaled training with SAM / SAM2

Collaboration outcomes

https://www.aist.go.jp/aist_j/magazine/20250205.html



Deployment of InsCore-trained models for industrial applications

S3OD: Towards Generalizable Salient Object Detection with Synthetic Data

arXiv: 2510.21605

Orest Kupyn, Hirokatsu Kataoka, Christian Rupprecht

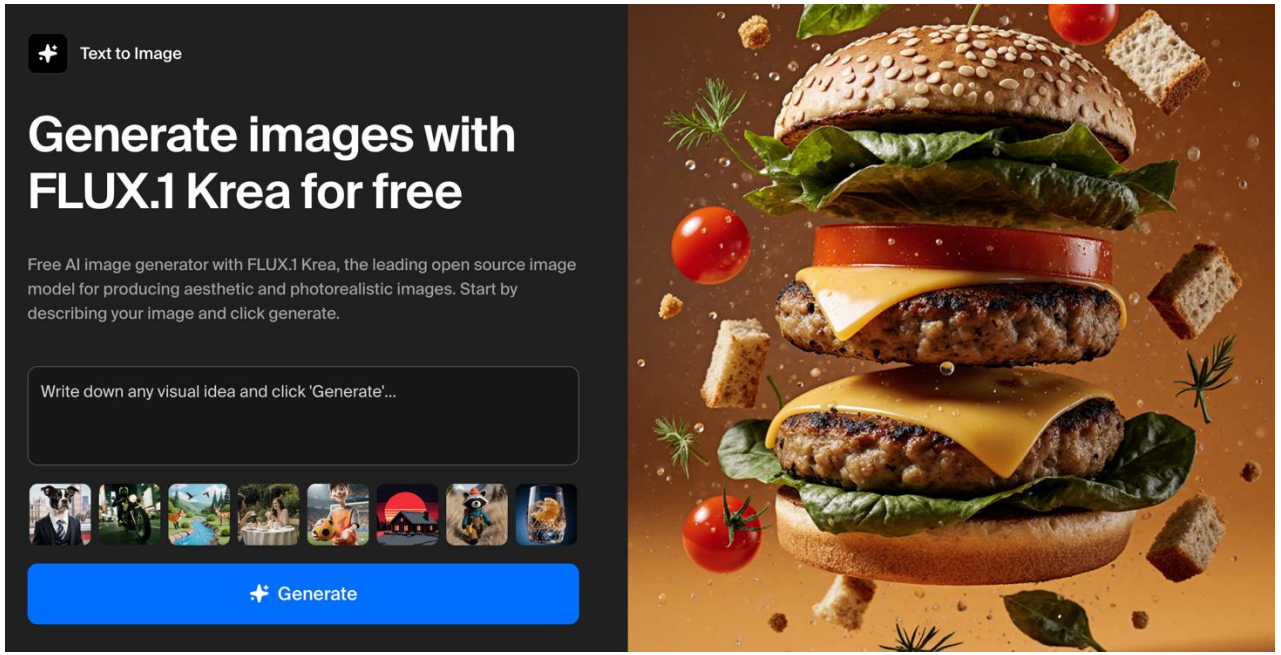
Oxford VGG / AIST

VFM under synthetic data from latest generative models

Research questions:

Can generative models make next vision foundation models?

- With a text-to-image model (FLUX DiT)
- With a strong feature representation (DINOv3)



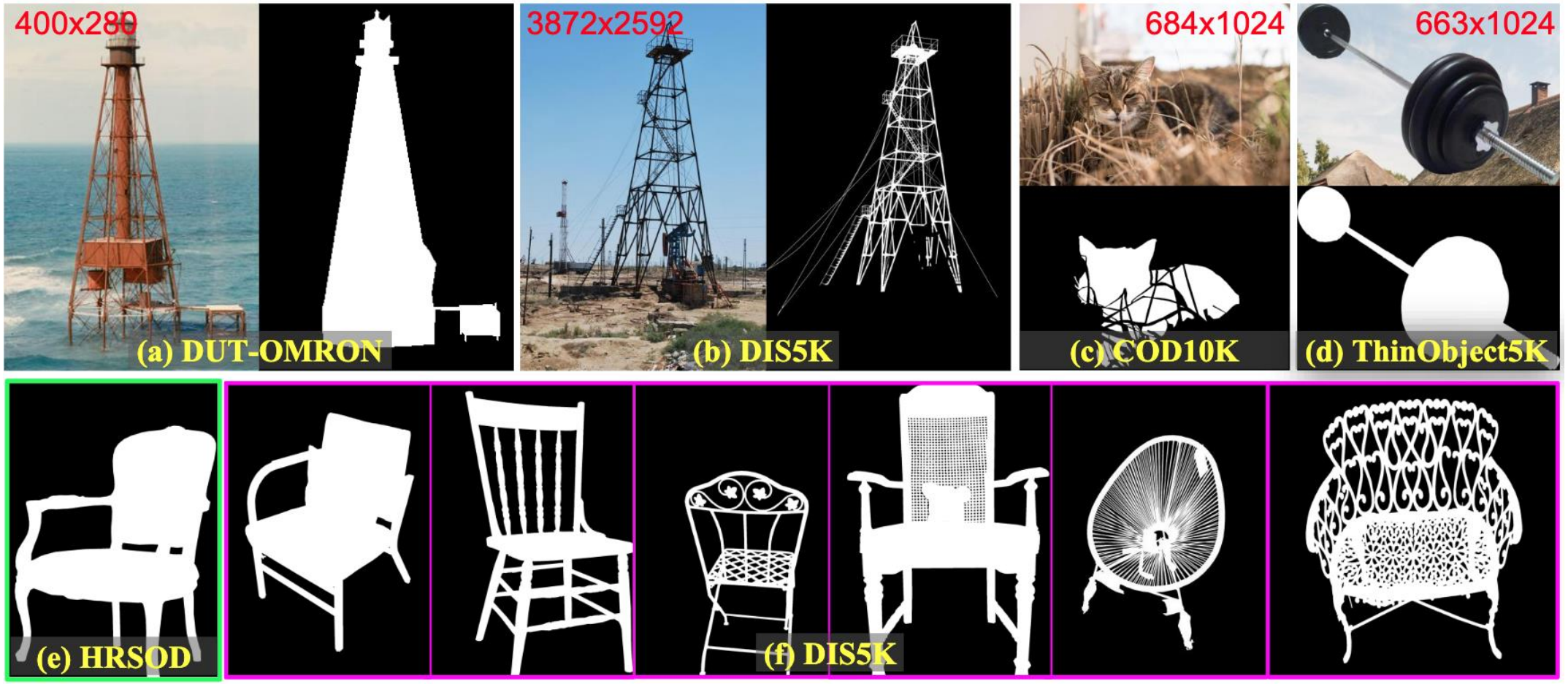
https://www.krea.ai/apps/image/flux-krea

https://ai.meta.com/research/publications/dinov3/

We try to create a VFM without any real images / human annotations

Challenges in high-resolution salient object detection (SOD)

- High-resolution SOD / DIS datasets are relatively small
- Pixel-accurate masks can take hours per image and include ambiguity



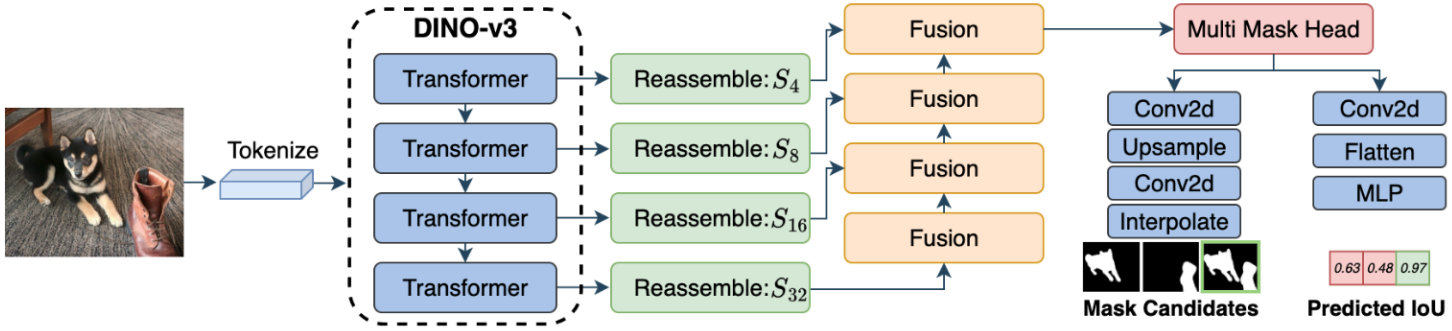
https://arxiv.org/abs/2203.03041

Overview & contributions

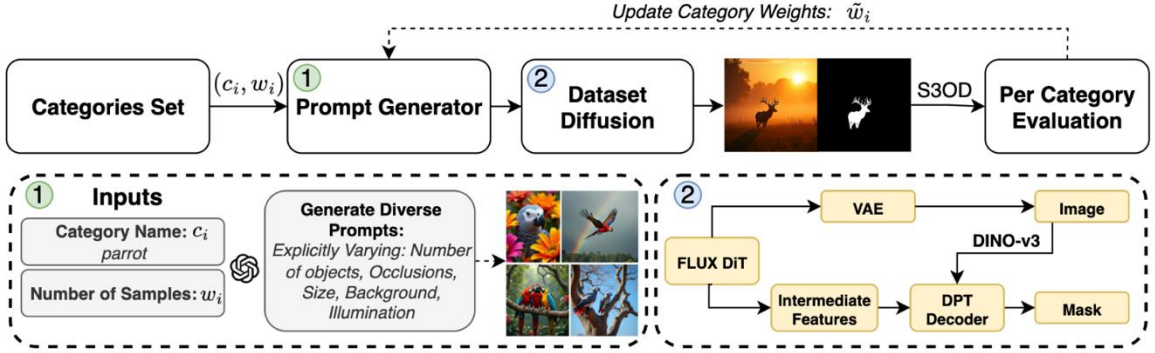
- Unify DIS & HR-SOD as high-fidelity salient segmentation
- Train a single, simple model that generalizes across datasets with synthetic data

S3OD dataset: Full-synthetic 139k+ images

S3OD architecture: DINOv3 backbone and DPT multi-mask prediction



Iterative generation pipeline: Improving masks from feedbacks



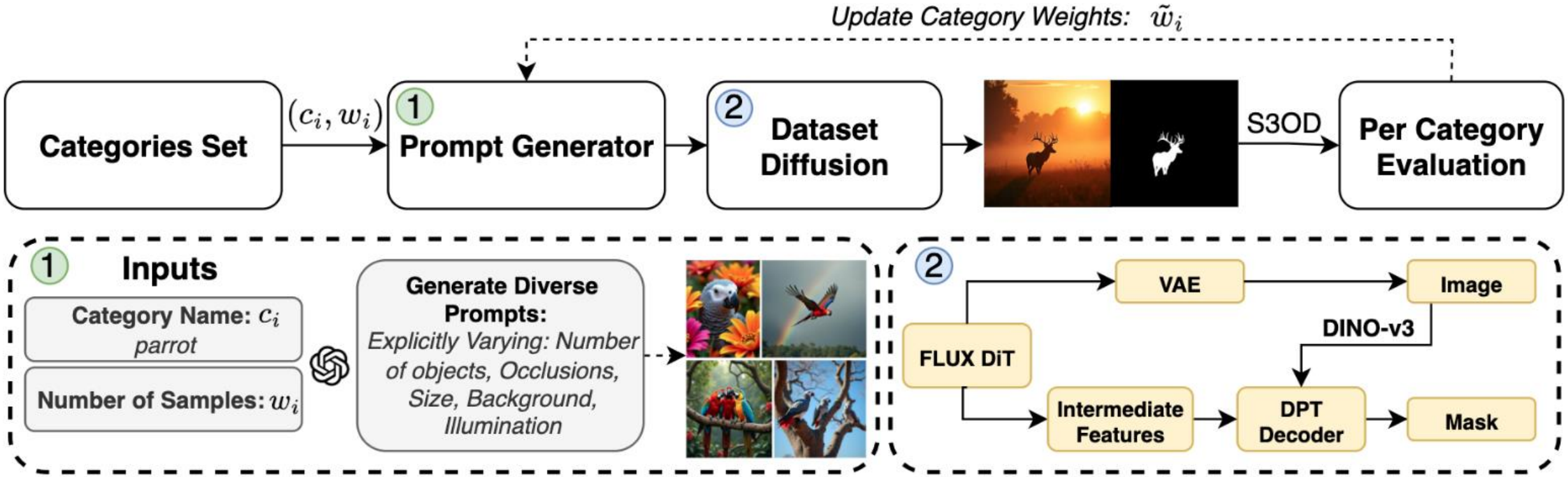
SotA results: Synth-only data can reach to the highest scores

Method	Data	Overall $F_m \uparrow$	$S_\alpha \uparrow$	$E_M^\Phi \uparrow$	MAE$\downarrow$
InSpyreNet	DUTS	.811	.830	.864	.065
BiRefNet	SOD	.825	.839	.861	.058
S3OD	SOD	.863	.856	.906	.049
S3OD	S3OD	**.881**	**.884**	**.925**	**.039**

S3OD dataset

Dataset for Scaling, Synthetic, and Salient Object Dataset

- Built on a high-resolution T2I model with FLUX DiT
- 1. Prompt generator & 2. dataset diffusion from a category set

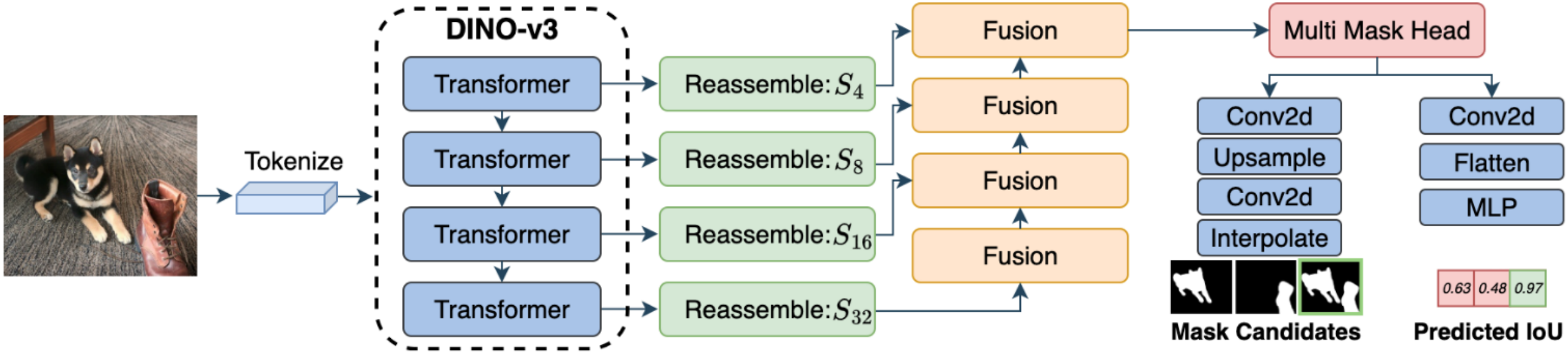


The dataset contains 139k images & 1,676 unique categories

Model: DINOv3 backbone & multimask decoder

Model architecture

- DINOv3 backbone with ViT
- Multi-mask decoder with DPT prediction head



Experiments: cross-dataset generalization & SotA

Synth-only S3OD training outperforms prior methods across all datasets

- DIS-5k, DAVIS-S, HRSOD-TE, UHRSOD-TE, DUTS-TE, DUT-OMRON
- 20-50% improvement in MAE

Method	Data	Overall			
		$F_m \uparrow$	$S_\alpha \uparrow$	$E_M^\Phi \uparrow$	MAE↓
InSpyreNet	DUTS	.811	.830	.864	.065
BiRefNet	SOD	.825	.839	.861	.058
S3OD	SOD	.863	.856	.906	.049
S3OD	S3OD	**.881**	**.884**	**.925**	**.039**

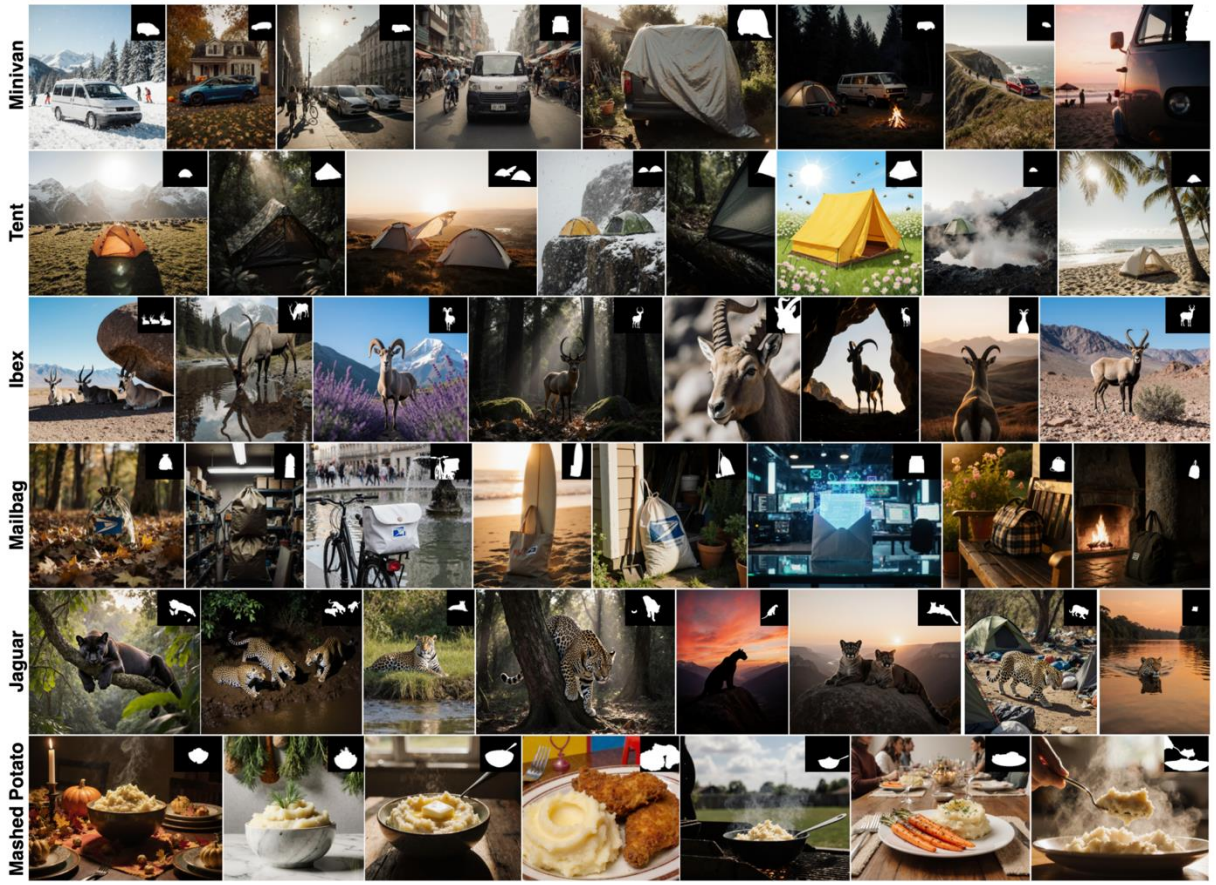
Method	Data	DAVIS-S				HRSOD-TE				UHRSOD-TE				DUTS-TE				DUT-OMRON			
		$F_m \uparrow$	$S_\alpha \uparrow$	$E_M^\Phi \uparrow$	MAE↓	$F_m \uparrow$	$S_\alpha \uparrow$	$E_M^\Phi \uparrow$	MAE↓	$F_m \uparrow$	$S_\alpha \uparrow$	$E_M^\Phi \uparrow$	MAE↓	$F_m \uparrow$	$S_\alpha \uparrow$	$E_M^\Phi \uparrow$	MAE↓	$F_m \uparrow$	$S_\alpha \uparrow$	$E_M^\Phi \uparrow$	MAE↓
InSpyreNet	DIS	.921	.937	.966	.015	.891	.912	.923	.038	.914	.922	.932	.033	.845	.880	.895	.046	.713	.801	.812	.071
BiRefNet	DIS	.919	.936	.961	.014	.887	.915	.926	.031	.922	.924	.937	.032	.860	.886	.910	.036	.744	.819	.835	.054
MVANet	DIS	.907	.929	.959	.016	.902	.919	.930	.033	.922	.926	.941	.032	.852	.877	.893	.042	.711	.792	.838	.072
S3OD	DIS	.951	.950	.973	.010	.923	.913	.932	.030	.946	.927	.947	.029	.902	.901	.926	.035	.808	.830	.858	.061
S3OD	S3OD	**.970**	**.967**	**.988**	**.005**	**.954**	**.955**	**.972**	**.016**	**.954**	**.944**	**.961**	**.023**	**.937**	**.938**	**.962**	**.020**	**.860**	**.887**	**.911**	**.040**

Can generative models make next vision foundation models? → Yes

Discussion & future direction

A single model trained on synth-only dataset can surpass the prev methods

- S3OD dataset & model provide scalable training for HR-SOD



More samples from S3OD dataset

Can generative models make next vision foundation models? → Yes

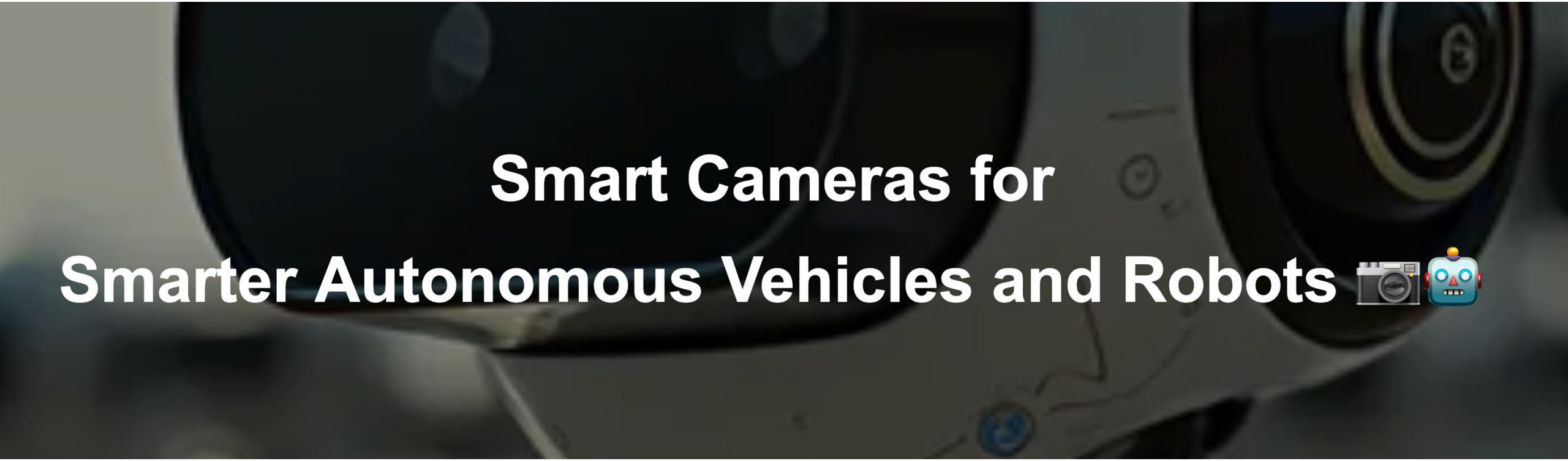
CV community requires Pure Vision Research

Hirokatsu Kataoka

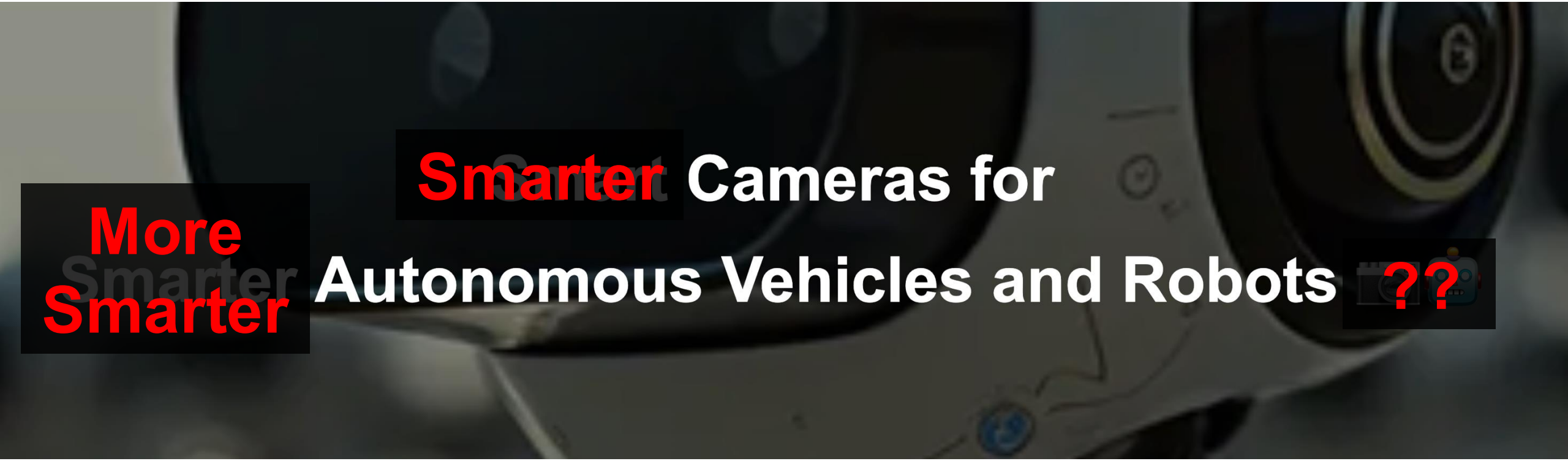
National Institute of Advanced Industrial Science and Technology (AIST)
Visual Geometry Group, University of Oxford (Oxford VGG)

http://www.hirokatsukataoka.net/

Today’s topic (once more consider!)



Today's topic (once more consider!)

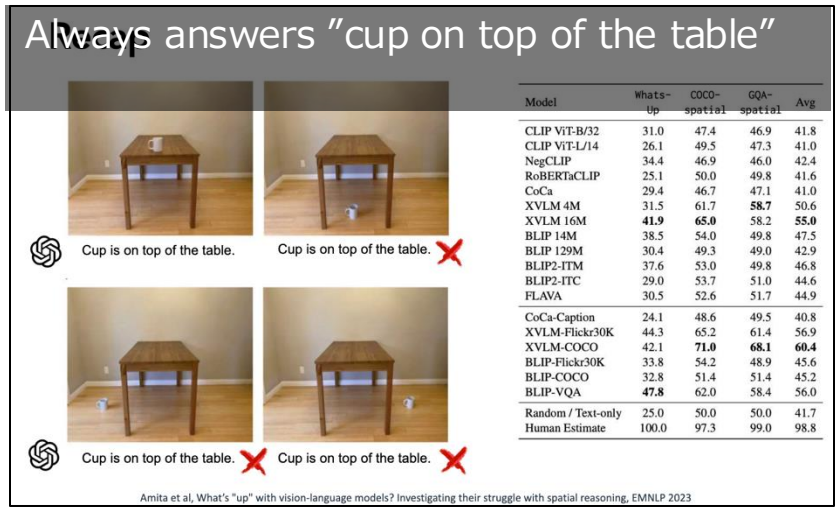


- For a smarter camera, I believe we should further improve visual learning

Future vision foundation models

Strong language bias with LLM / VLLM?
Less researches in the pure vision researches

Always answers "cup on top of the table"



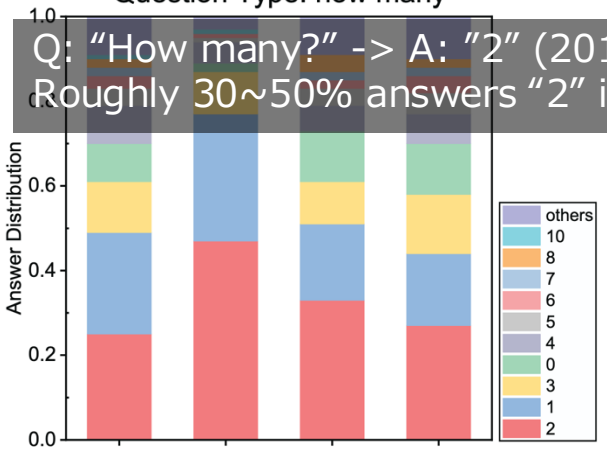
Model	What's Up	COCO-spatial	GQA-spatial	Avg
CLIP ViT-B/32	31.0	47.4	46.9	41.8
CLIP ViT-L/14	26.1	49.5	47.3	41.0
NegCLIP	34.4	46.9	46.0	42.4
RoBERTaCLIP	25.1	50.0	49.8	41.6
CoCa	29.4	46.7	47.1	41.0
XVLM 4M	31.5	61.7	**58.7**	50.6
XVLM 16M	**41.9**	**65.0**	58.2	**55.0**
BLIP 14M	38.5	54.0	49.8	47.5
BLIP 129M	30.4	49.3	49.0	42.9
BLIP2-ITM	37.6	53.0	49.8	46.8
BLIP2-ITC	29.0	53.7	51.0	44.6
FLAVA	30.5	52.6	51.7	44.9
CoCa-Caption	24.1	48.6	49.5	40.8
XVLM-Flickr30K	44.3	65.2	61.4	56.9
XVLM-COCO	42.1	**71.0**	**68.1**	**60.4**
BLIP-Flickr30K	33.8	54.2	48.9	45.6
BLIP-COCO	32.8	51.4	51.4	45.2
BLIP-VQA	**47.8**	62.0	58.4	56.0
Random / Text-only	25.0	50.0	50.0	41.7
Human Estimate	100.0	97.3	99.0	98.8

Amita et al, What's "up" with vision-language models? Investigating their struggle with spatial reasoning, EMNLP 2023

Slide from: https://musi-workshop.github.io/files/talk_manling_musi_2025.pdf

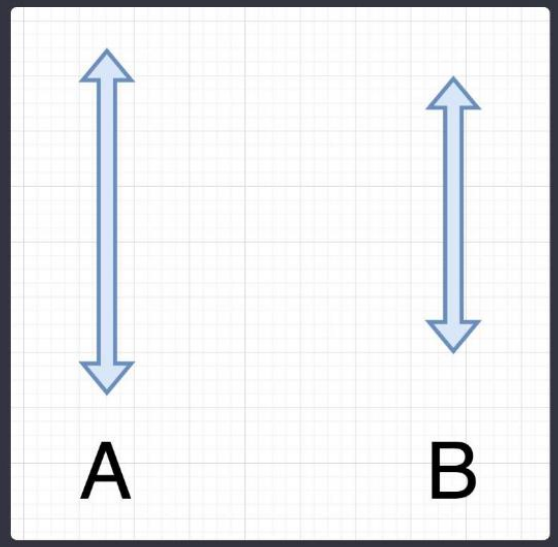
Question Type: how many

Q: "How many?" -> A: "2" (2019)
Roughly 30~50% answers "2" in the Q.



[Guo et al. SIGIR19]

Current VLMs have Poor **Geometric** Understanding [Wang et al., 2024]



GPT4-V

Both arrows, A and B, appear to be of the same length.

Date: 2023.10

ChatGPT 5

The two arrows **A** and **B** might *look* different in size because of their width and position, but if you compare carefully, they are actually the **same length**.

This is a variation of a common visual illusion — our brains interpret the thicker arrow (A) as longer, even though both are equal.

Would you like me to overlay measurement lines on your image to prove they are identical?

Date: 2025.10

GPT-5 still struggles with spatial understanding ? (2024 - 2025)

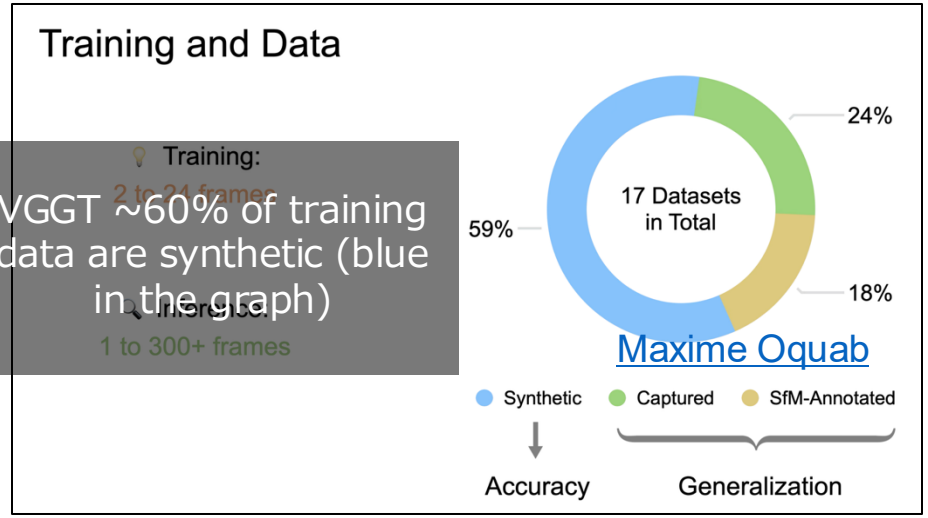
Slide from: https://musi-workshop.github.io/files/talk_manling_musi_2025.pdf

Huge room for improvement in pure vision abilities

Recent pure vision researches

Potential of synthetic data + single Transformer integration

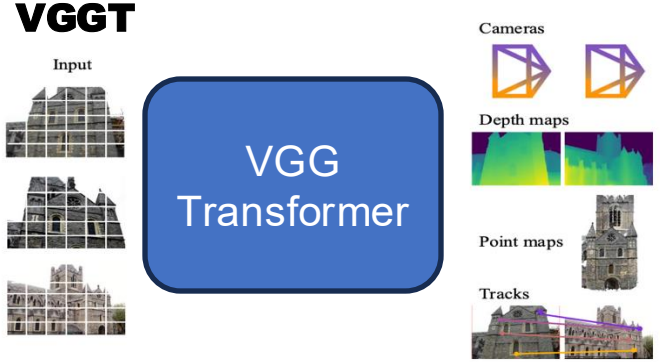
VGGT



VGGT ~60% of training data are synthetic (blue in the graph)

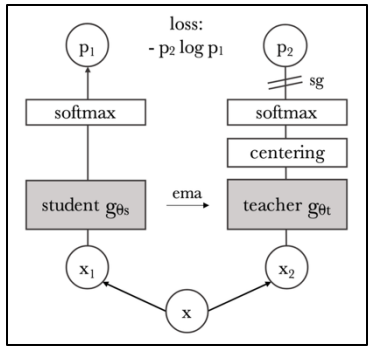
Slide from: https://docs.google.com/presentation/d/1JVuPnuZx6RgAy-U5Ezobg73XpBi7FrOh/edit

VGGT



[Wang et al. CVPR25]

DINO

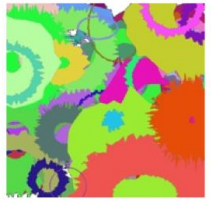


[Caron et al. ICCV21]
[Oquab et al. arXiv23]
[Simeoni et al. arXiv25]

SAM/SAM2

[Kirillov et al. ICCV23]

InsCore



InsCore demonstrates synthetic pre-training can surpass large-scale models pre-trained with real images

Trained with a single Transformer

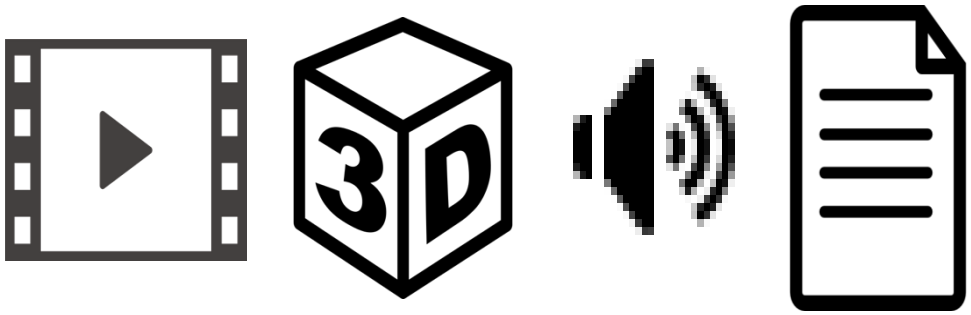
Synthetic data can flexibly arrange the teacher labels & vision tasks

For smarter cameras & more smarter applications?

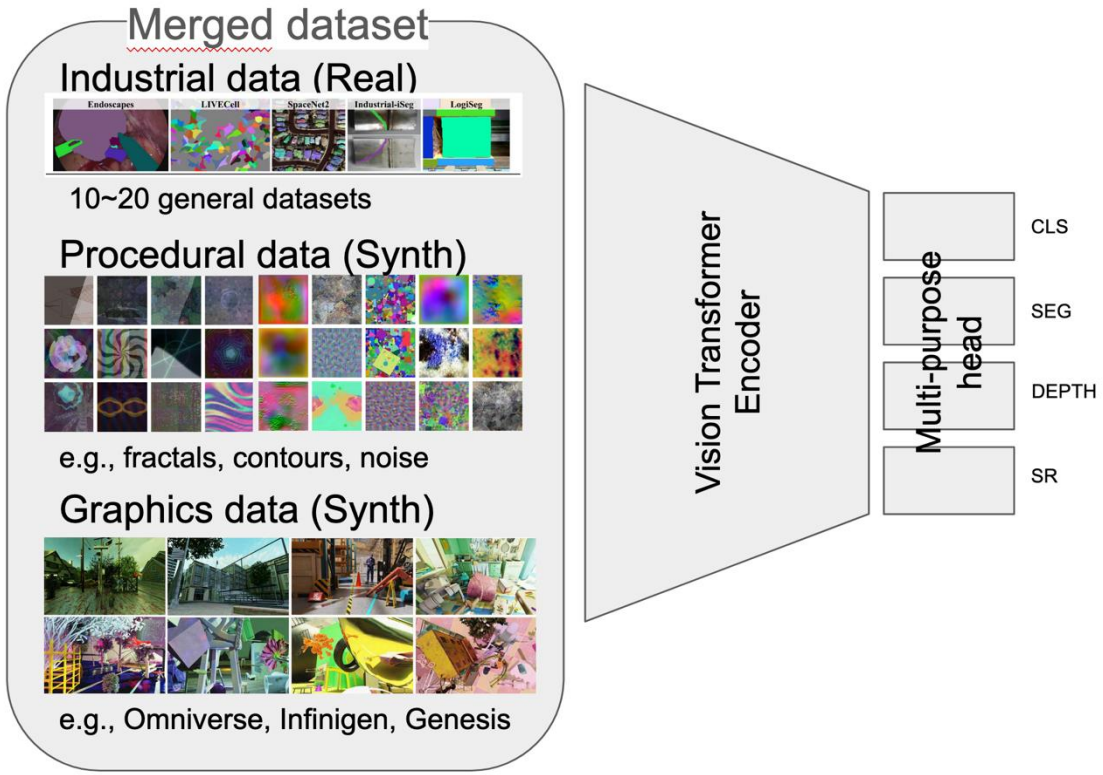
With further improved visual models through synthetic data

e.g., our CVPR22, WACV22, ECCV24, ICASSP25…

Video 3D Audio Text



Any modality



Any task

Any modality, any task, with unified Transformer

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

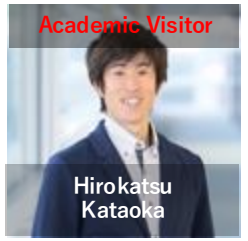
【LIMIT.Community / LIMIT.Lab】

LIMIT.LAB

Community => LIMIT.Community
- 150+ researchers / students
- LIMIT Workshops @ ICCV25&CVPR24

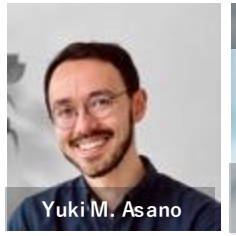
Research Lab => LIMIT.Lab
- JP AIST
- GB Oxford VGG, Cambridge VSL
- DE UTN FunAI Lab
- NL UvA VISLab

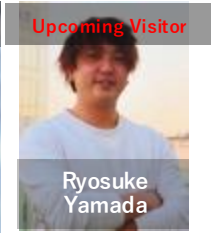
【Oxford VGGGB】



+Postdocs, Ph.D. students at VGG

【UTN FunAILabDE / former UvA VISLabNL】





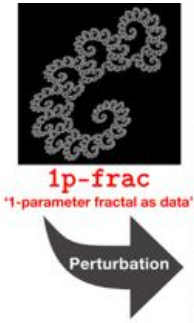
+Postdocs, Ph.D. students at FunAILab

【Cambridge VSLGB】

+Upcoming Ph.D. students at VSL

I'm leading this community-driven research!

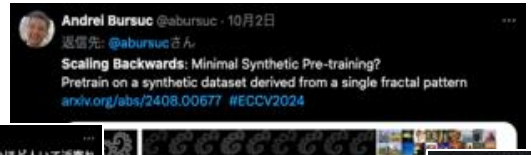
We will conduct the researches to improve pure vision foundation model

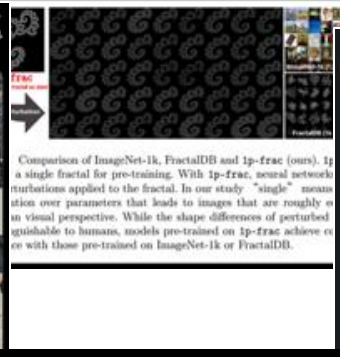


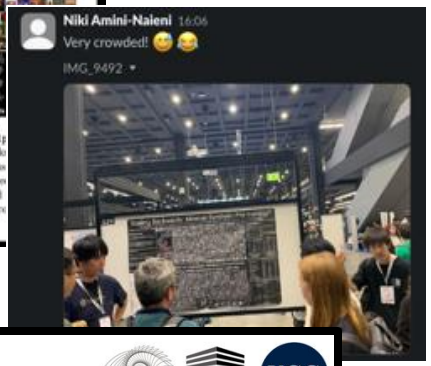
Scaling backwards (ECCV24)
An iconic research in LIMIT.Lab

LIMIT Workshop 23 @ Paris FR
LIMIT Workshop 24 @ Seattle US
LIMIT Workshop 25 @ Hawaii US

ECCV24 @ Milano IT







LIMIT Workshop Organizers