

視覚基盤モデルの構築

片岡 裕雄

産総研 人工知能研究センター

<https://hirokatsukataoka.net/jp/>

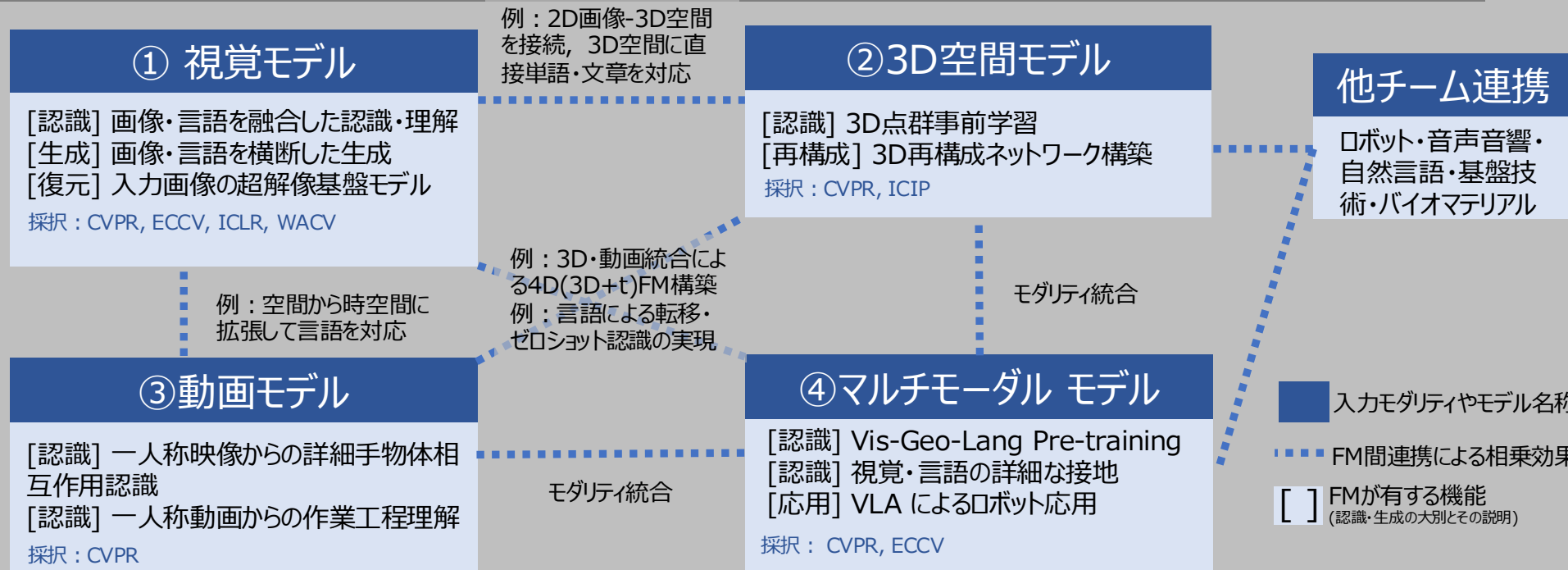
FDSL(データ):
数式からの教師付データ生成により権利・倫理問題を根本解消. 画像・言語・音響など任意の教師・データペア構築.

Ego4D(データ):
ウェアラブルセンサ等 一人称視点動画における大規模データ. 人間視点からの汎用的認識を目指して構築され, 自動車やロボットへの適用も可能.

3D ResNet(モデル):
時空間動画画像認識の世界的ベースライン. CNNの知見をTransformerや自己教師あり学習に応用.

LIMIT(モデル&データ):
計算機・データ・教師ラベル等が限られていてもAI基盤モデル学習する技術. 人工生成データと少量実データの組み合わせにより基盤モデル構築を実現.

※上記は画像チームのメンバーが保有するコア技術. 最先端技術を常時キャッチアップしつつ実装.



ロボット実装への戦略

【マルチモーダル基盤モデル】①～③ 2D/3D/言語/動画 基盤モデルと統合による④マルチモーダル基盤への拡張. まずは外界を的確に捉える視覚機能を獲得し, ロボット操作により実世界へ影響.

【生成的/マルチモーダル事前学習】 生成規則・生成AIから人工的に生成されたデータを再利用, ロボットの動作が求められる環境にて柔軟に学習. 視覚・聴覚・言語を同時学習することも実験的に成功.

ロボット:

- 工業製品ロボット操作
- 店舗の商品陳列ロボット
- 想定モデル: ①③④

リモートセンシング:

- 国土利用地図の高頻度更新・災害即応
- 都市開発・環境変化モニタリング
- 想定モデル: ①②③

物流:

- フォークリフト自動運転
- 物体検知-障害物自動回避
- 荷役異常自動判定
- 想定モデル: ①③④

医療:

- 内視鏡画像の自動診断
- 3D-CT による臓器復元
- 想定モデル: ①②③

ビジュアル素材:

- ユーザ指定画像生成
- 権利の所在を明らかにした画像入手
- 想定モデル: ①②④

採択: CVPR, ICCV, BMVC talk, Journal, 産総研マガジン掲載

【政策予算・画像チーム】

FDSL(データ):

数式からの教師付データ生成により権利・倫理問題を根本解消. 画像・言語・音響など任意の教師・データペア構築.

Ego4D(データ):

ウェアラブルセンサ等 一人称視点動画における大規模データ. 人間視点からの汎用的認識を目指して構築され, 自動車やロボットへの適用も可能.

3D ResNet(モデル):

時空間動画像認識の世界的ベースライン. CNNの知見をTransformerや自己教師あり学習に応用.

LIMIT(モデル&データ):

計算機・データ・教師ラベル等が限られていてもAI基盤モデル学習する技術. 人工生成データと少量実データの組み合わせにより基盤モデル構築を実現.

※上記は画像チームのメンバーが保有するコア技術. 最先端技術を常時キャッチアップしつつ実装.

◆視覚・マルチモーダル事前学習

◆一人称動画認識

◆動画認識モデルベースライン

◆限定資源下における視覚モデル構築

◆リモートセンシング

例: 2D画像-3D空間を接続し3D空間に直交するFM構築

例: 5D(3D+2t)FM構築
例: 言語による転移・ゼロショット認識の実現

②3D空間モデル

[認識] 3D点群事前学習
[再構成] 3D再構成ネットワーク構築
採択: CVPR, ICIP

他チーム連携

ロボット・音声音響・自然言語・基盤技術・バイオマテリアル

モデル統合

④マルチモーダルモデル

[認識] Vis-Geo-Lang Pre-training
[認識] 視覚・言語の詳細な接地
[応用] VLA によるロボット応用
採択: CVPR, ECCV

入力モダリティやモデル名称

FM間連携による相乗効果

FMが有する機能 (認識・生成の大別とその説明)

③動画モデル

を実施してきたメンバーが参画

相互作用認識

[認識] 一人称動画からの作業工程理解

採択: CVPR

モデル統合

ロボット:

- 工業製品ロボット操作
- 店舗の商品陳列ロボット
- 想定モデル: ①③④

リモートセンシング:

- 国土利用地図の高頻度更新・災害即応
- 都市開発・環境変化モニタリング
- 想定モデル: ①②③

物流:

- フォークリフト自動運転
- 物体検知-障害物自動回避
- 荷役異常自動判定
- 想定モデル: ①③④

医療:

- 内視鏡画像の自動診断
- 3D-CT による臓器復元
- 想定モデル: ①②③

ビジュアル素材:

- ユーザ指定画像生成
- 権利の所在を明らかにした画像入手
- 想定モデル: ①②④

採択: CVPR, ICCV, BMVC talk, Journal, 産総研マガジン掲載

【政策予算・画像チーム】

ロボット実装への戦略

【マルチモーダル基盤モデル】①～③2D/3D/言語/動画 基盤モデルと統合による④マルチモーダル基盤への拡張. まずは外界を的確に捉える視覚機能を獲得し, ロボット操作により実世界へ影響.

【生成的/マルチモーダル事前学習】生成規則・生成AIから人工的に生成されたデータを再利用, ロボットの動作が求められる環境にて柔軟に学習. 視覚・聴覚・言語を同時学習することも実験的に成功.

FDSL(データ):

数式からの教師付データ生成により権利・倫理問題を根本解消. 画像・言語・音響など任意の教師・データペア構築.

Ego4D(データ):

ウェアラブルセンサ等 一人称視点動画における大規模データ. 人間視点からの汎用的認識を目指して構築され, 自動車やロボットへの適用も可能.

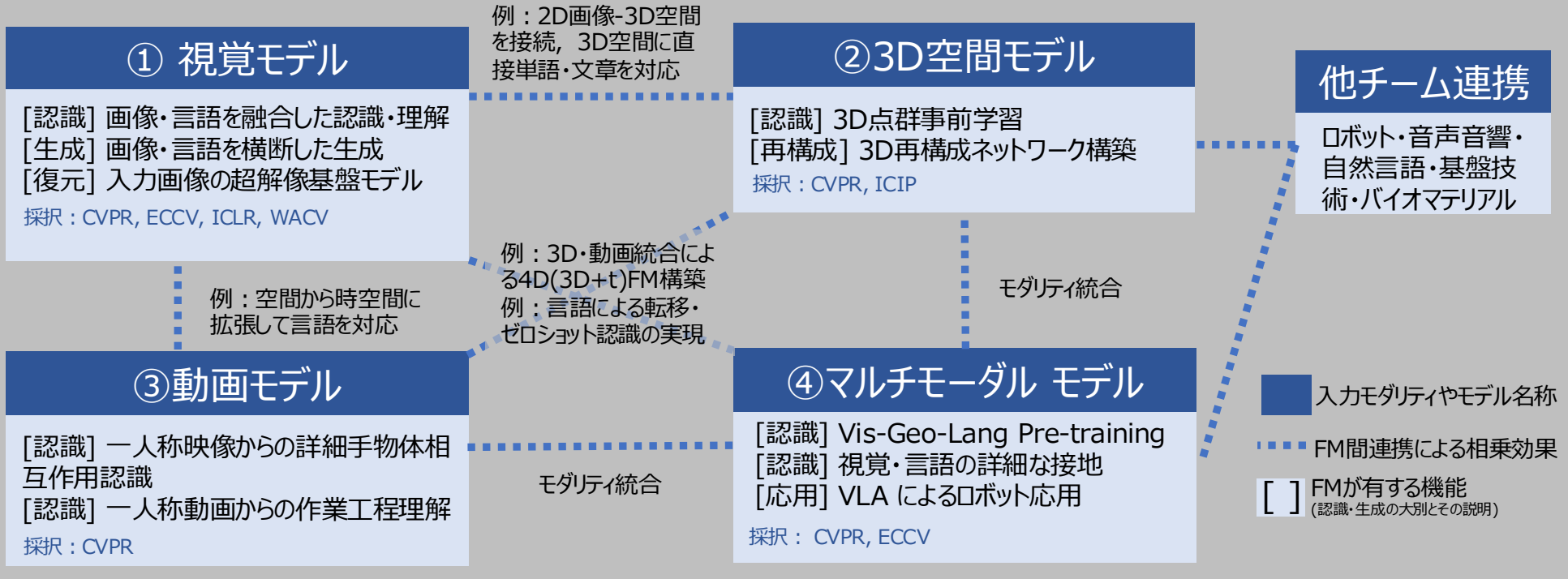
3D ResNet(モデル):

時空間動画画像認識の世界的ベースライン. CNNの知見をTransformerや自己教師あり学習に応用.

LIMIT(モデル&データ):

計算機・データ・教師ラベル等が限られていてもAI基盤モデル学習する技術. 人工生成データと少量実データの組み合わせにより基盤モデル構築を実現.

※上記は画像チームのメンバーが保有するコア技術. 最先端技術を常時キャッチアップしつつ実装.



ロボット実装への戦略

【マルチモーダル基盤モデル】①～③2D/3D/言語/動画 基盤モデルと統合による④マルチモーダル基盤への拡張. まずは外界を的確に捉える視覚機能を獲得し, ロボット操作により実世界へ影響.

【生成的/マルチモーダル事前学習】生成規則・生成AIから人工的に生成されたデータを再利用, ロボットの動作が求められる環境にて柔軟に学習. 視覚・聴覚・言語を同時学習することも実験的に成功.

ロボット:

- 工業製品ロボット操作
- 想定モデル: ①③④

リモートセンシング:

- 国土利用地図の高頻度
- 都市開発・環境変化モニタリング
- 想定モデル: ①②③

物流:

- フォークリフト自動運転
- 荷役異常自動判定
- 想定モデル: ①③④

医療:

- 中視鏡画像の自動診断
- 臓器復元
- 想定モデル: ①②③

ビジュアル素材:

- ユーザ指定画像生成
- 権利の所在を明らかにした画像入手
- 想定モデル: ①②④

視覚+αのモダリティ①②③④ごとに基盤モデル構築

採択: CVPR, ICCV, BMVC talk, Journal, 産総研マガジン掲載

【政策予算・画像チーム】

FDSL(データ):

数式からの教師付データ生成により権利・倫理問題を根本解消. 画像・言語・音響など任意の教師・データペア構築.

Ego4D(データ):

ウェアラブルセンサ等 一人称視点動画における大規模データ. 人間視点からの汎用的認識を目指して構築され, 自動車やロボットへの適用も可能.

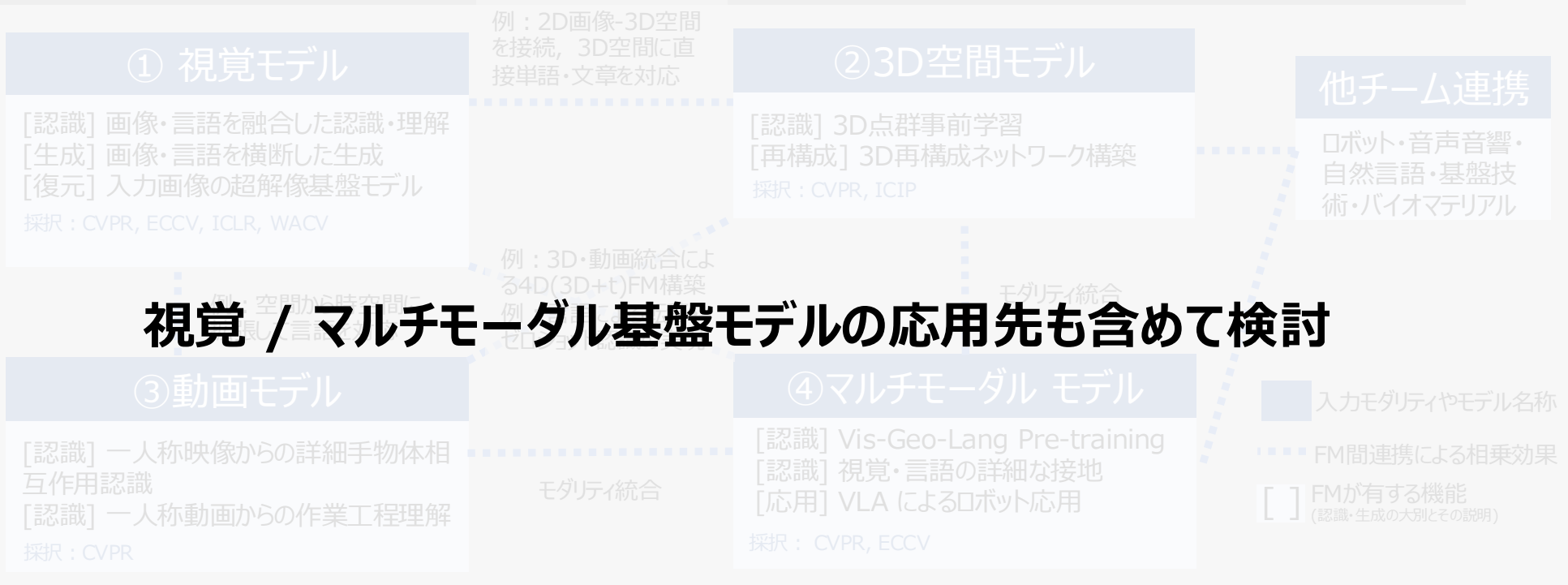
3D ResNet(モデル):

時空間動画画像認識の世界的ベースライン. CNNの知見をTransformerや自己教師あり学習に応用.

LIMIT(モデル&データ):

計算機・データ・教師ラベル等が限られていてもAI基盤モデル学習する技術. 人工生成データと少量実データの組み合わせにより基盤モデル構築を実現.

※上記は画像チームのメンバーが保有するコア技術. 最先端技術を常時キャッチアップしつつ実装.



視覚 / マルチモーダル基盤モデルの応用先も含めて検討

ロボット実装への戦略

【マルチモーダル基盤モデル】①～③2D/3D/言語/動画 基盤モデルと統合による④マルチモーダル基盤への拡張. まずは外界を的確に捉える視覚機能を獲得し, ロボット操作により実世界へ影響.

【生成的/マルチモーダル事前学習】生成規則・生成AIから人工的に生成されたデータを再利用, ロボットの動作が求められる環境にて柔軟に学習. 視覚・聴覚・言語を同時学習することも実験的に成功.

ロボット:

- 工業製品ロボット操作
- 店舗の商品陳列ロボット
- 想定モデル: ①③④

リモートセンシング:

- 国土利用地図の高頻度更新・災害即応
- 都市開発・環境変化モニタリング
- 想定モデル: ①②③

物流:

- フォークリフト自動運転
- 物体検知-障害物自動回避
- 荷役異常自動判定
- 想定モデル: ①③④

医療:

- 内視鏡画像の自動診断
- 3D-CT による臓器復元
- 想定モデル: ①②③

ビジュアル素材:

- ユーザ指定画像生成
- 権利の所在を明らかにした画像入手
- 想定モデル: ①②④

採択: CVPR, ICCV, BMVC talk, Journal, 産総研マガジン掲載

【政策予算・画像チーム】

限定資源下におけるマルチモーダル / 視覚基盤モデル構築

✕ 単一の大規模基盤モデル構築 ○ 効率よく多様なモダリティの基盤モデル構築

→ 音声・言語のように日本語学習ができないので、視覚において全掛けはリスクと判断

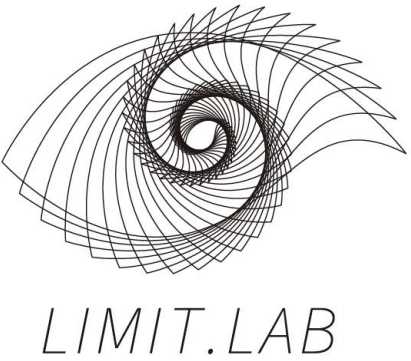
→ 複数モダリティ間のデータ収集・モデル構築・横展開/社会実装を経て知見獲得と学習改善

政策予算プロジェクトの基本方針：LIMIT

限定資源下におけるマルチモーダル / 視覚基盤モデル構築

→ 同トピックにおいて英独蘭との国際連携を展開

【LIMIT.Community / LIMIT.Lab】



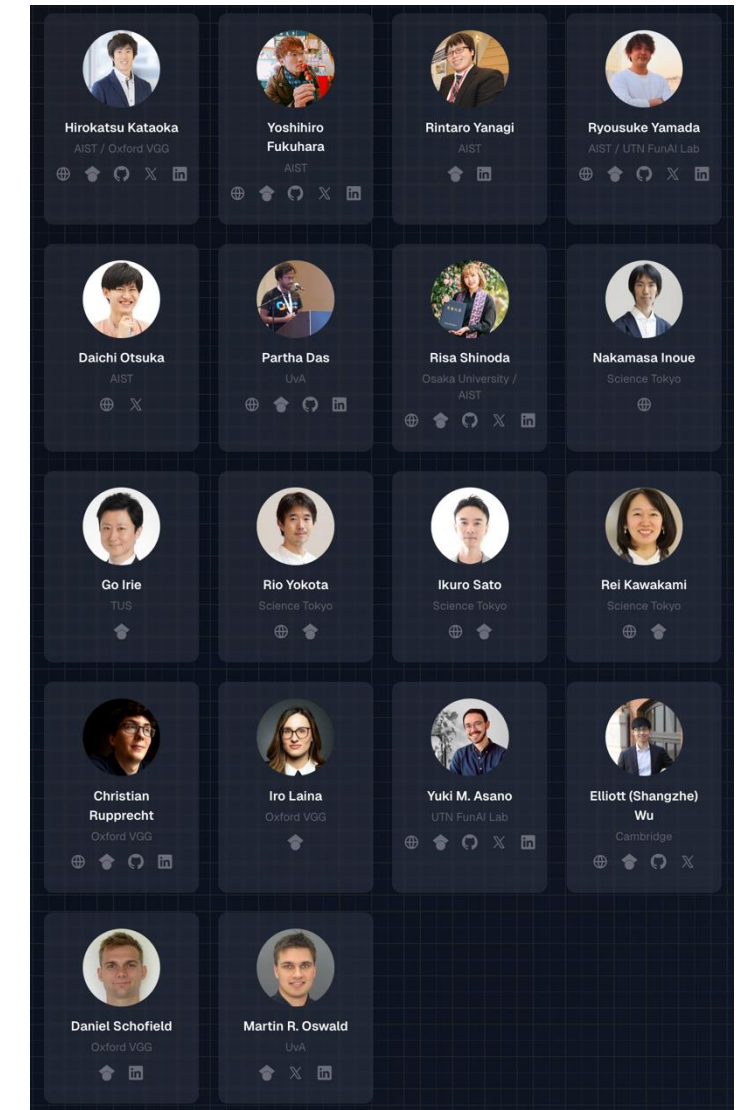
Community => LIMIT.Community

- 国際メンバー 200名超
- LIMIT Workshops 開催@ ICCV23, 25 & CVPR24

LIMIT.Labに研究メンバーが集結

- JP AIST
- GB Oxford VGG, Cambridge VSL
- DE UTN FunAI Lab
- NL UvA VISLab, CVLab

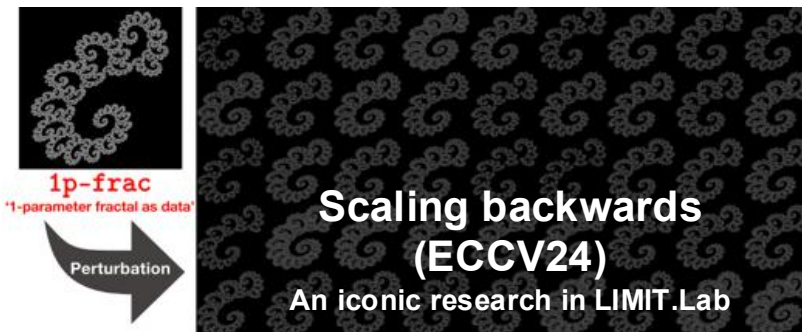
AIST 主導の研究イニシアティブとして機能



<https://limitlab.xyz/>

直近数年の主な国際活動・連携・研究

- 2023年10月 ICCV 2023 / 2024年6月 CVPR 2024 / 2025年10月 ICCV 2025にて LIMIT Workshop 開催
- 2025年6月 国際研究コミュニティ LIMIT.Lab 設立
- 2025年 国内外コミュニティ, AIST PJから合計70件超の主著・共著論文投稿
- その他 招待講演・国際ワークショップ・国際交流会など定期開催



ECCV24 @ MilanoIT



LIMIT Workshop Organizers



FunAI seminar @ NurembergDE



VGG seminar @ OxfordGB

CVPR 2025 Report

Hirokatsu Kataoka, Yoshihiro Fukuhara,

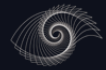
Ryousuke Yamada, Daichi Otsuka, Rintaro Yanagi, Kazuya Nishimura, Moeri Okuda, Yuto Matsuo, Ren Ohkubo, Yue Qiu, Noritake Kodama, Gido Kato, Kenzo Yamabuki, Joe Hasei, Ryuichi Nakahara, Yukinori Yamamoto, Sho Okazaki, Kohsuke Ide, Yuiga Wada, Daichi Yashima, Shinichi Mae, Hinako Mitsuoka, Maika Takada, Oishi Deb, Orest Kupyn, Jianyuan Wang

[LIMIT.Lab](https://limit.lab) / cvpaper.challenge / [Visual Geometry Group \(VGG\)](https://visualgeometrygroup.com)

2025年10月 ICCV 2025 Workshop 開催

トップ国際会議 ICCV 2025 @ Hawaii 開催

HELD AS PART OF



LIMIT.Workshop

AT

ICCV
OCT 19-23, 2025



HONOLULU
HAWAII

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

ICCV 2025 Workshop

📅 October 19, 2025, 1:00 PM – 6:00 PM (HST) 📍 Hawaii Convention Center 306 A

Organizers

Hirokatsu Kataoka
AIST / Oxford VGG



Website [🔗](#)

Yuki M. Asano
University of Technology Nuremberg (UTN)



Website [🔗](#)

Iro Laina
Oxford VGG



Website [🔗](#)

Rio Yokota
Institute of Science Tokyo



Website [🔗](#)

Nakama Inoue
Institute of Science Tokyo



Website [🔗](#)

Rintaro Yanagi
AIST



Website [🔗](#)

Partha Das
University of Amsterdam



Website [🔗](#)

Connor Anderson
Kitware



Website [🔗](#)

Ryosuke Yamada
AIST



Website [🔗](#)

Dalchi Otsuka
AIST



Website [🔗](#)

Yoshihiro Fukuhara
AIST / Waseda University



Website [🔗](#)

2025年10月 ICCV 2025 Workshop 開催

トップ国際会議 ICCV 2025 @ Hawaii 開催

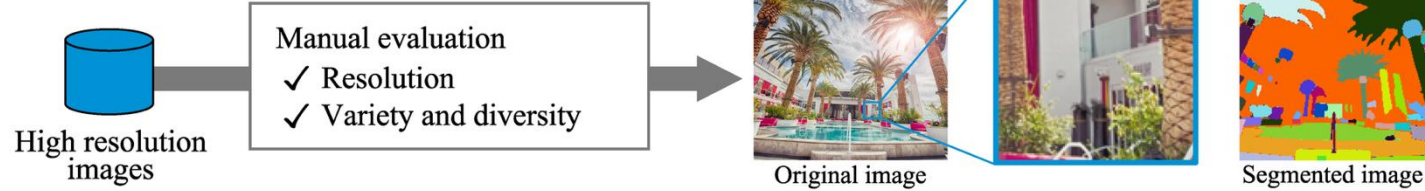


国際活動を経るごとに
人脈形成 → 研究コミュニティ →
研究チーム → 研究自体
が強化される枠組みとなっている

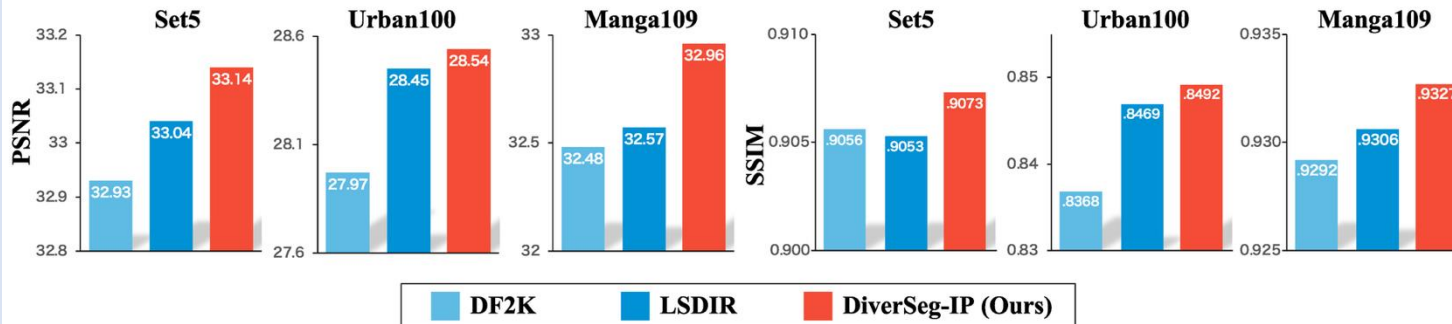
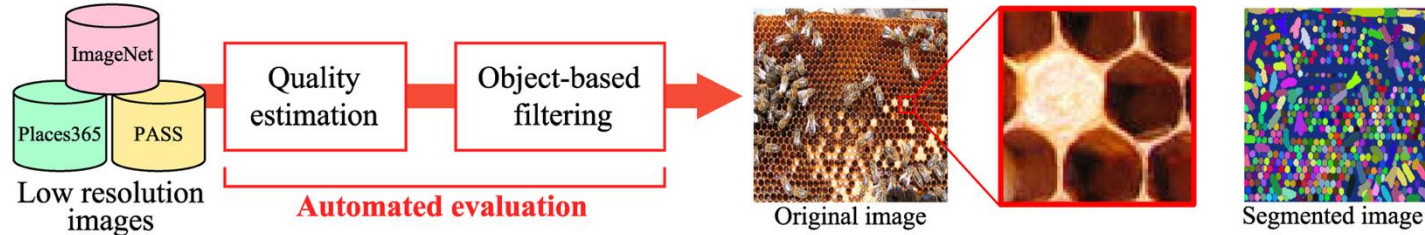


① 視覚基盤モデル

Conventional approach

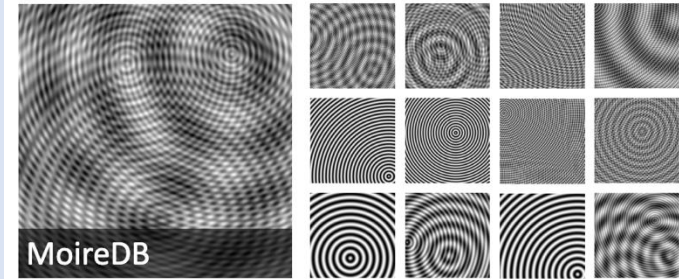


Our approach



超解像基盤モデル

[Ohtani+, ECCV 2024]



モアレ画像による ロバスト性向上

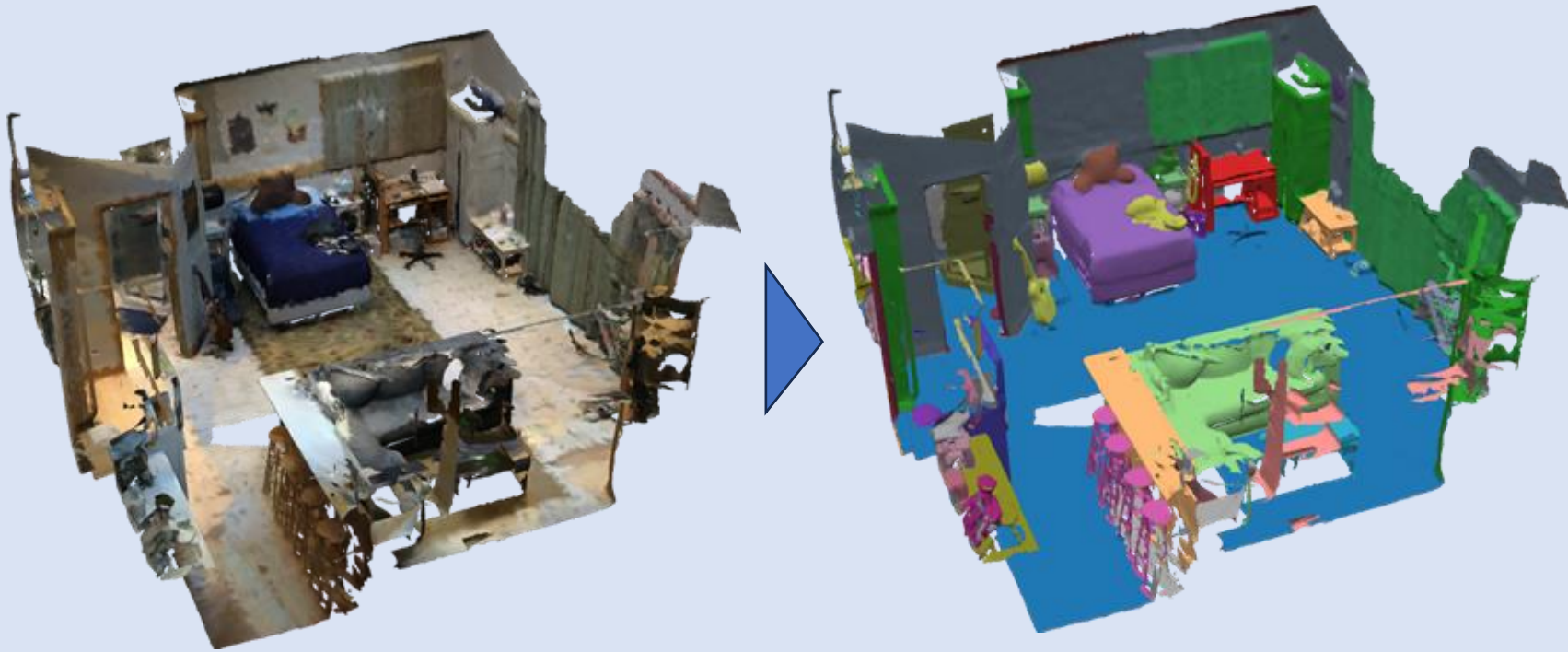
[Matsuo+, CVPRW 2025]



生成データによる 超解像学習

[Kodama+, CVPRW 2025]

② 3D空間基盤モデル



3D点群基盤モデル

[Yamada+, CVPR 2026]

後で詳しく紹介

③ 動画モデル

Existing Benchmark

High-level HOI Recognition



High-level Action

- ✓ hammer cylinder
- ✗ cut wood

or

High-level Object States

- ✓ cylinder is hammered
- ✗ cylinder is bent



Low-level Hand/Object Localization



- ✗ Overlook fine-grained dynamics
- ✗ Focus only on specific aspects of HOI
- ✗ Restrict grounding to object level

Our Benchmark for Fine-Grained HOI Dynamics

Mutiple-Choice Question (MCQ)



Action Q. What is the person doing with his/her hands?

- ✓ He is hitting the cylinder with the hammer in his right hand.
- ✗ He is hitting the wood on the floor with the hammer in his right hand. ✗ +3 more...

Process Q. How does the person hammer the cylinder?

- ✓ He hammers the cylinder straight down from above.
- ✗ He hammers the cylinder from the side. ✗ +3 more...

Objects Q. What object is used by the hands?

- ✓ Cylinder ✓ Hammer ✗ Wood ✗ Yarn ✗ Window

Location Q. Where does the person put down the hammer?

- ✓ He put down the hammer on the floor in front of him to right.
- ✗ He put down the hammer on the wood in front of him to left. ✗ +3 more...

State Change Q. How did the state of cylinder change?

- ✓ The white plastic part at the top has gone inside the cylinder.
- ✗ The cylinder was completely crushed. ✗ +3 more...

Object Parts Q. What part of the cylinder was hammered?

- ✓ The white plastic part at the top the cylinder was hammered.
- ✗ The side part of the cylinder was hammered. ✗ +3 more...

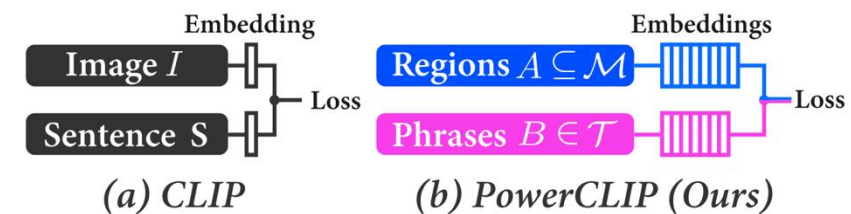
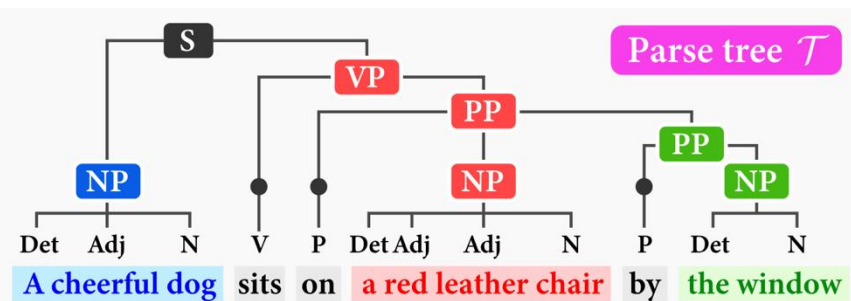
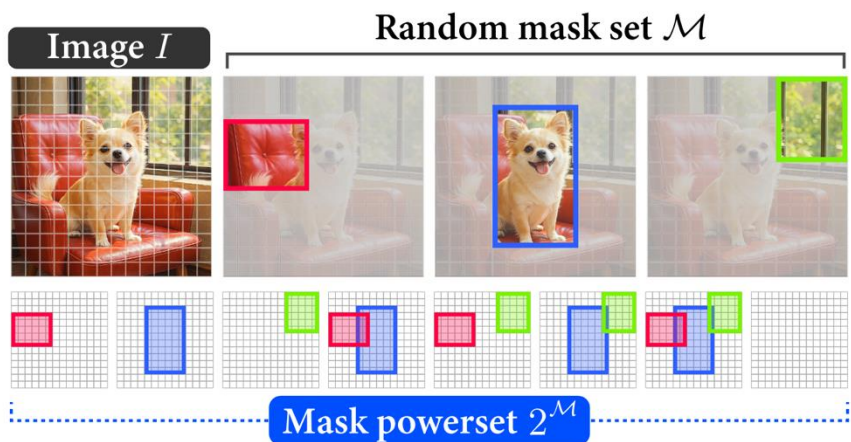
Reasoning Video Object Segmentation (ReasoningVOS)



動画基盤モデルベンチマーキング

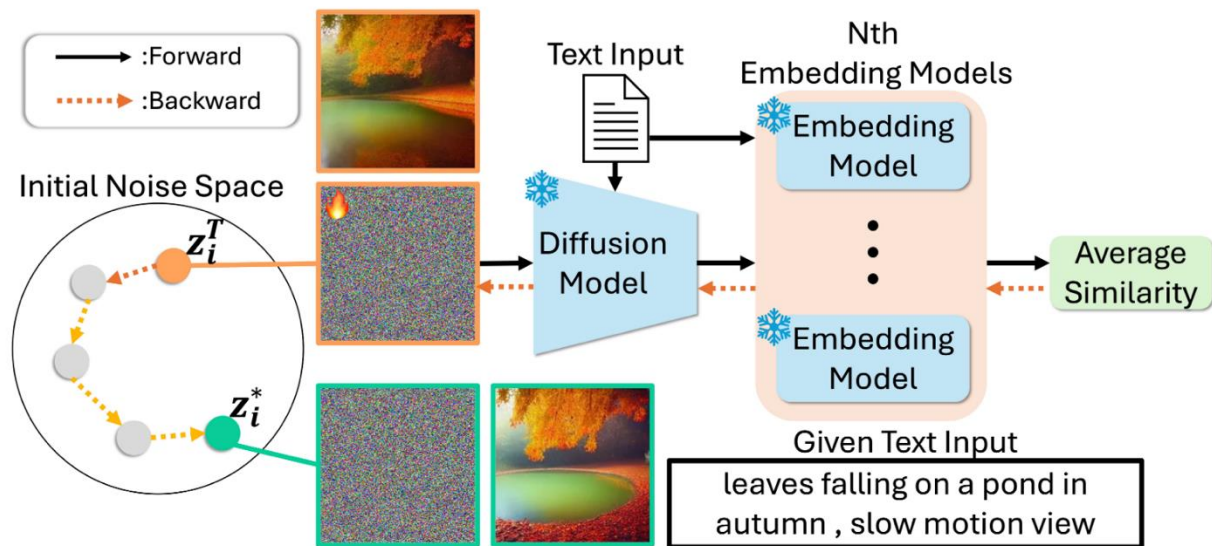
[Tateno+, CVPR 2026]

④ マルチモーダルモデル



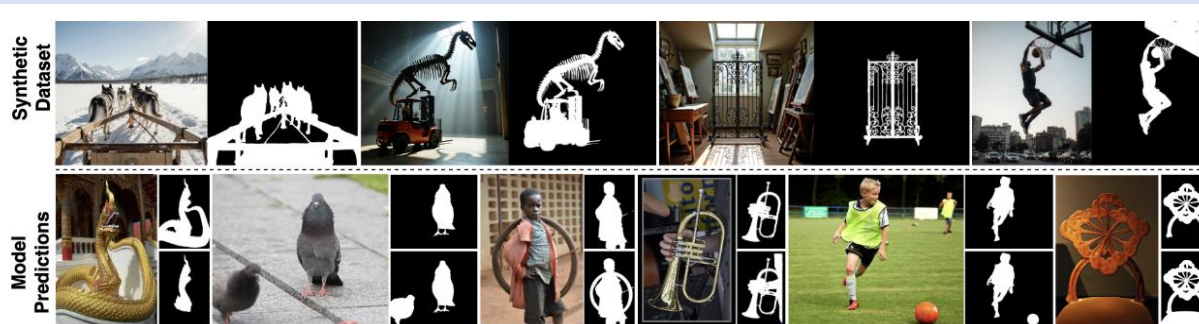
視覚言語モデル

[Kawamura+, CVPR 2026]



より良い生成学習データの探索

[Ohkubo+, WACV 2026]



基盤モデルが次世代の基盤モデルを構築

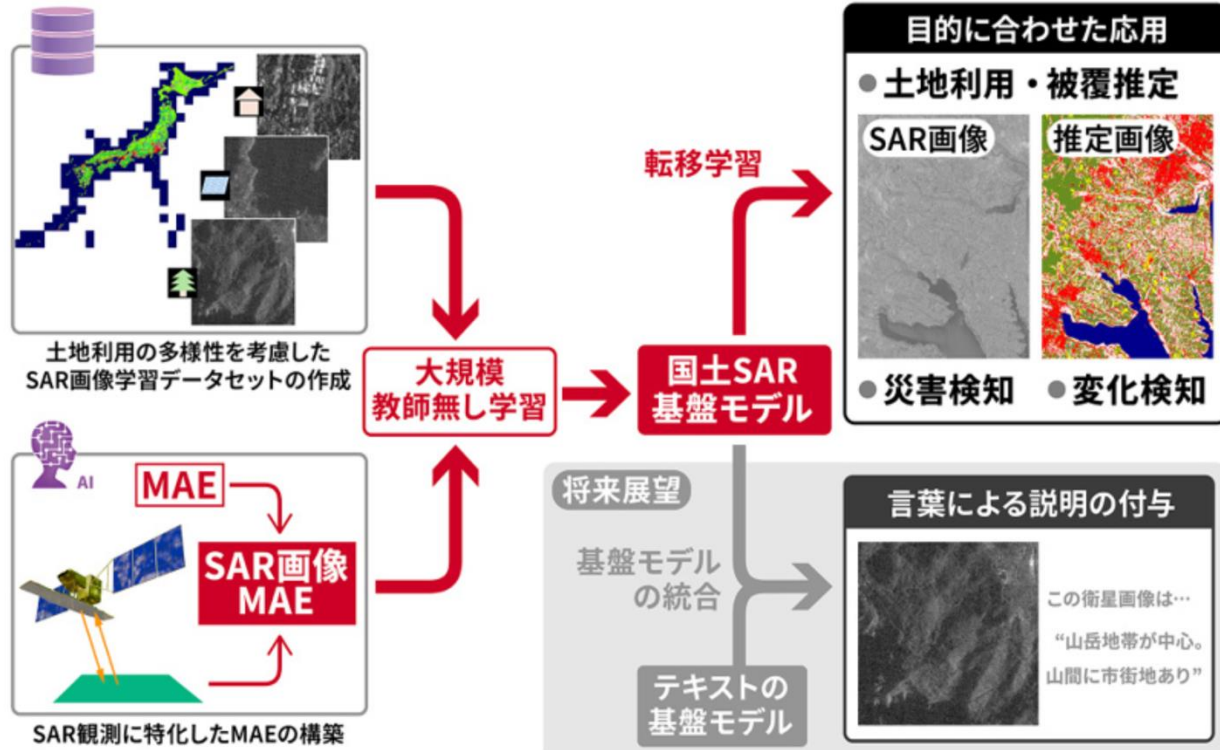
[Kupyn+, ICLR 2026]

後で詳しく紹介

実世界適応

人工衛星「だいち2号」の観測データを活用して国土に特化したSAR基盤モデルを構築

—SAR観測データへのAI利用をより手軽に—



衛星画像基盤モデル



野生動物保全ベンチマーク

[Shinoda+, ICCV 2025]



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



S3OD: Towards Generalizable Salient Object Detection with Synthetic Data

ICLR 2026 accepted paper

Orest Kupyn, Hirokatsu Kataoka, Christian Rupprecht

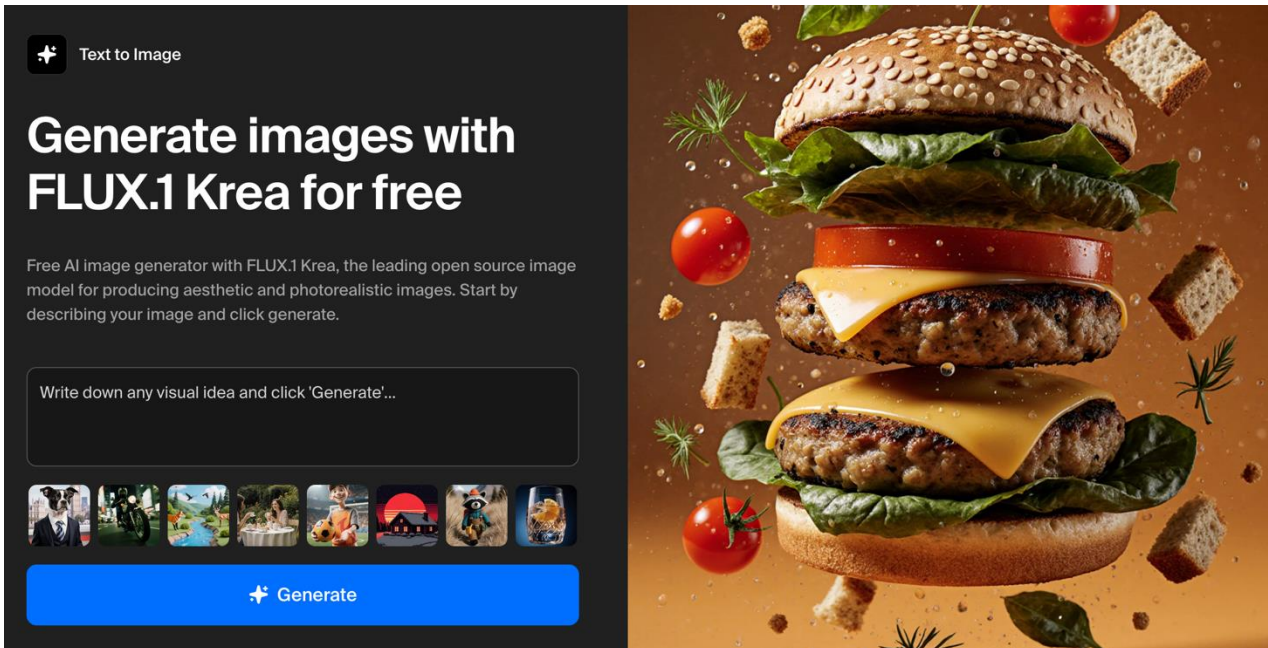
Oxford VGG / AIST

生成モデル / 合成データによるVFM構築

Research questions:

生成モデルは **次の** 視覚基盤モデル (VFM) を構築できるのか?

- Text-to-Image モデル (Flux DiT)
- 汎用視覚特徴抽出 (DINOv3)



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

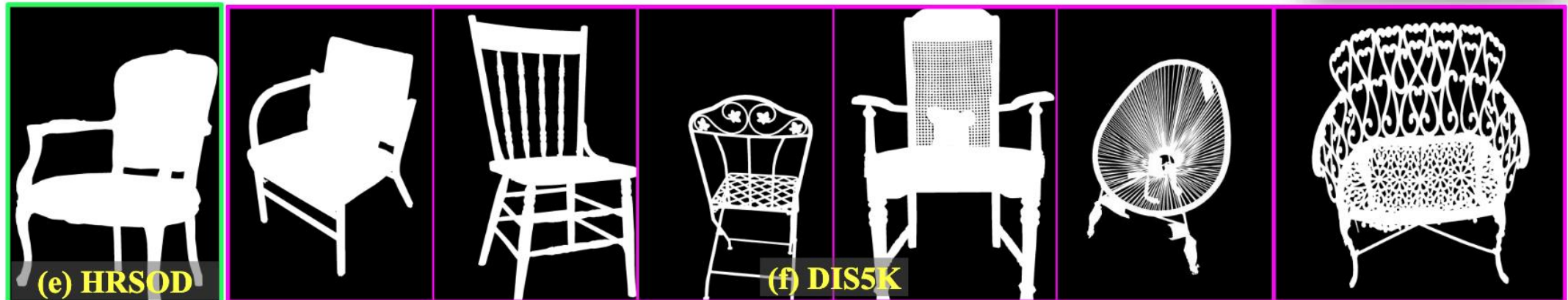
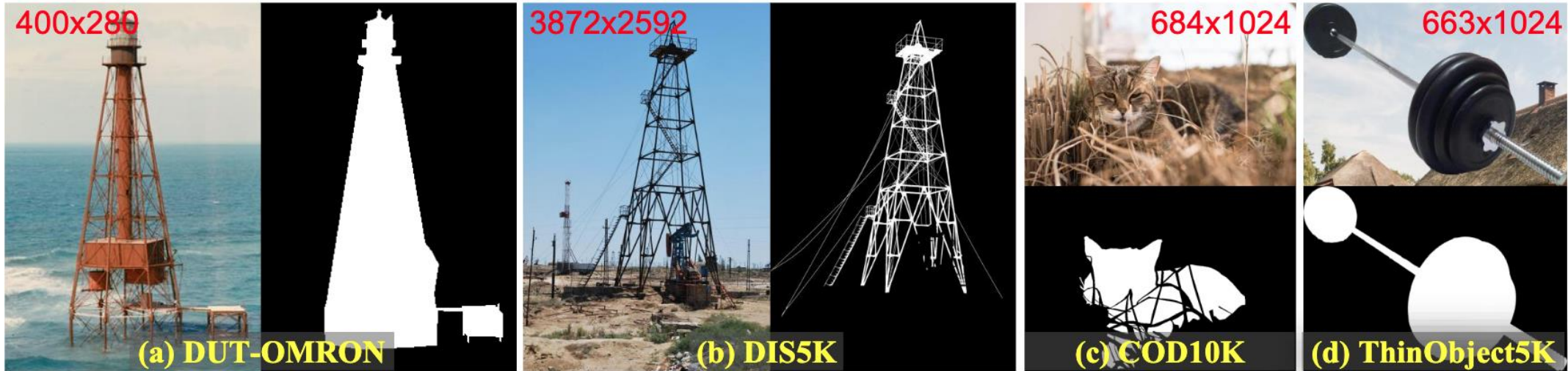


<https://ai.meta.com/research/publications/dinov3/>

実画像やアノテーションなしに VFM の構築に挑戦する

顕著性抽出 / 高精細画像セグメンテーション

- 顕著性抽出 / 高精細画像セグメンテーションのデータセットは比較的小規模
- 画素の精密性が重要なので人為的なラベル付に多大なる時間を要する



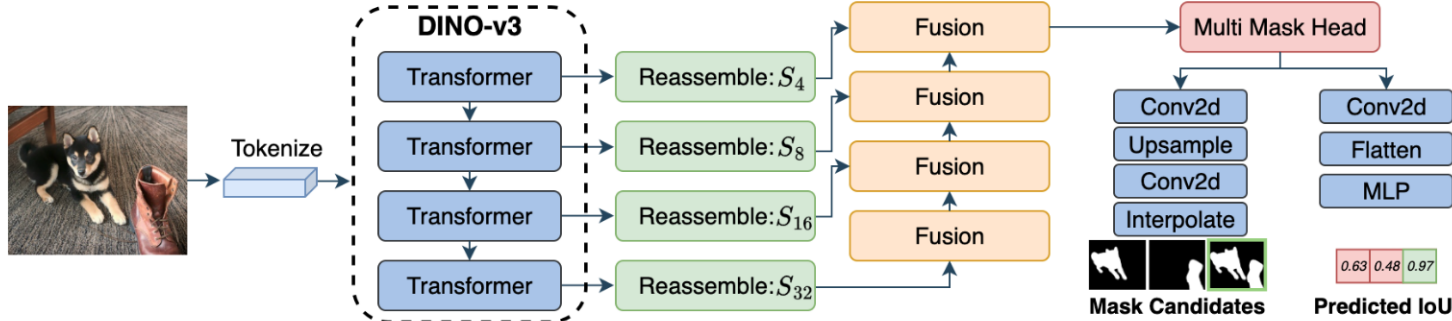
概要 & 貢献

- 高精細セグメンテーションのデータを自動生成
- 合成データにより単一・単純モデルでも汎化性を担保
- Text prompting, DINOv3特徴, 高精度生成マスクにより最高水準のモデル

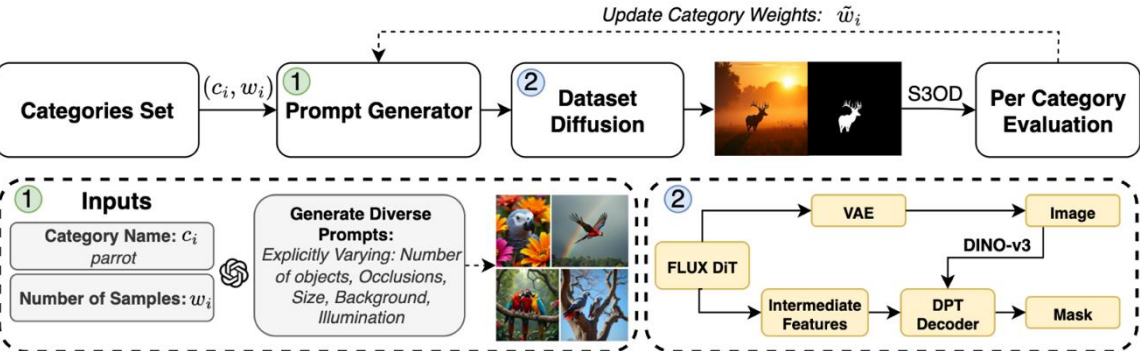
データセット: 合成された13.9万画像



モデル: DINOv3 backbone & DPT 多タスク学習



繰り返しパイプライン: フィードバックによるマスク精度向上



最高水準の性能: 実データを使わずともベースよりも高いスコア

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	<u>.863</u>	<u>.856</u>	<u>.906</u>	<u>.049</u>
S3OD	S3OD	.881	.884	.925	.039



3D sans 3D Scans: Scalable Pre-training from Video-Generated Point Clouds

CVPR 2026 accepted paper

**Ryousuke Yamada, Kohsuke Ide, Yoshihiro Fukuhara,
Hirokatsu Kataoka, Gilles Puy, Andrei Bursuc, Yuki M. Asano**

AIST / UTN

3Dデータ: 3D 自己教師あり学習のボトルネック

Difficult

Data collection

Easy

3D data: thousands of train sample

NVS: thousands of train sample

Video: millions of train sample

Image: billions of train sample

Text: trillions of train sample

Video x Reconstruction x SSL

Research questions:

- ラベルなしの動画から3D表現を獲得できるのか？
 - + 再構成データ: 効率的な3D 事前学習データが作成できる？
 - + 点群モデルのスケーリング: 大規模なデータは学習可能か？
 - + 統一的なモデル: 屋内外のデータに対して適応可能なのか？



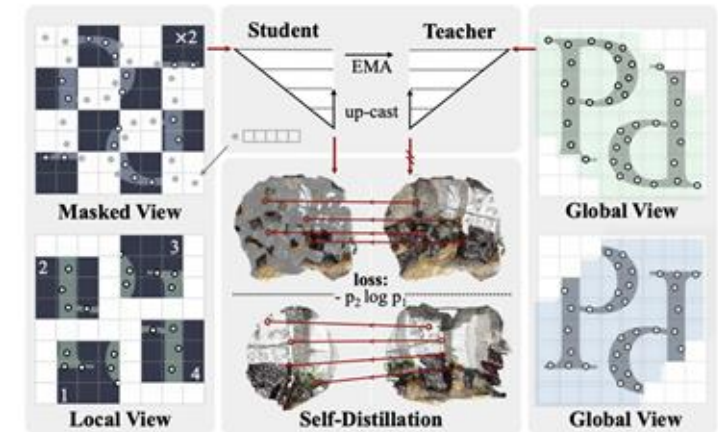
Unlabeled video

[Shashanka Venkataramanan, ICLR2024]



Reconstruction

[Jianyuan Wang, CVPR2025]

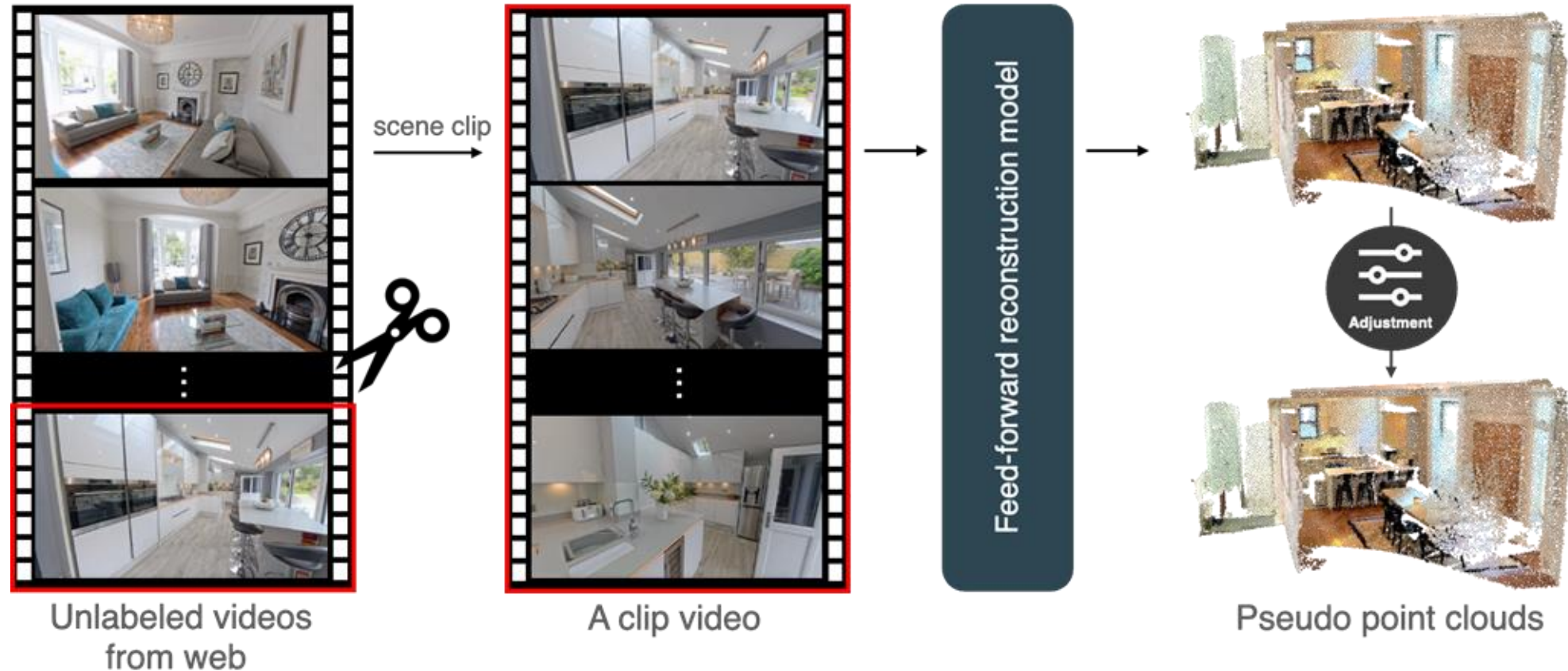


SSL

[Xiaoyang Wu, CVPR2025]

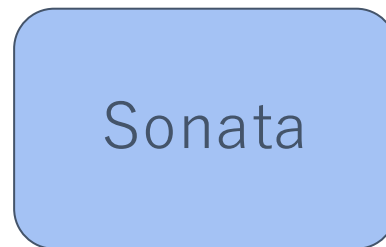
動画データセット：RoomTours

- 動画シェアサイトから 5,082 動画を記録
 - 動画あたり平均 4.92 シーン（部屋）が記録
 - 三次元再構成モデル π^3 により動画→3D点群データに変換



1. 事前学習

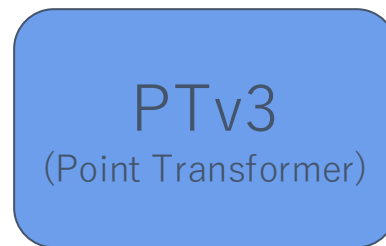
Sonata on fake point clouds (RoomTours)



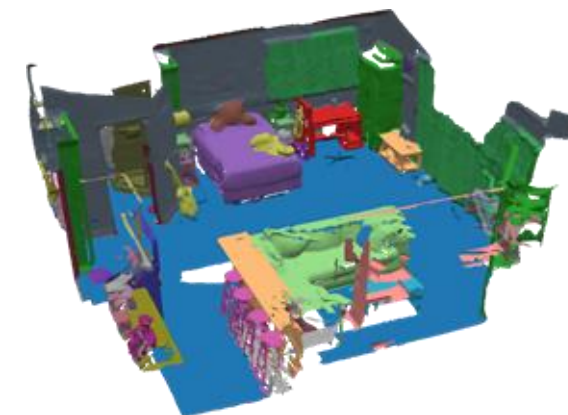
自己蒸留
Self-distillation

2. Linear Probing / Full Fine-tuning

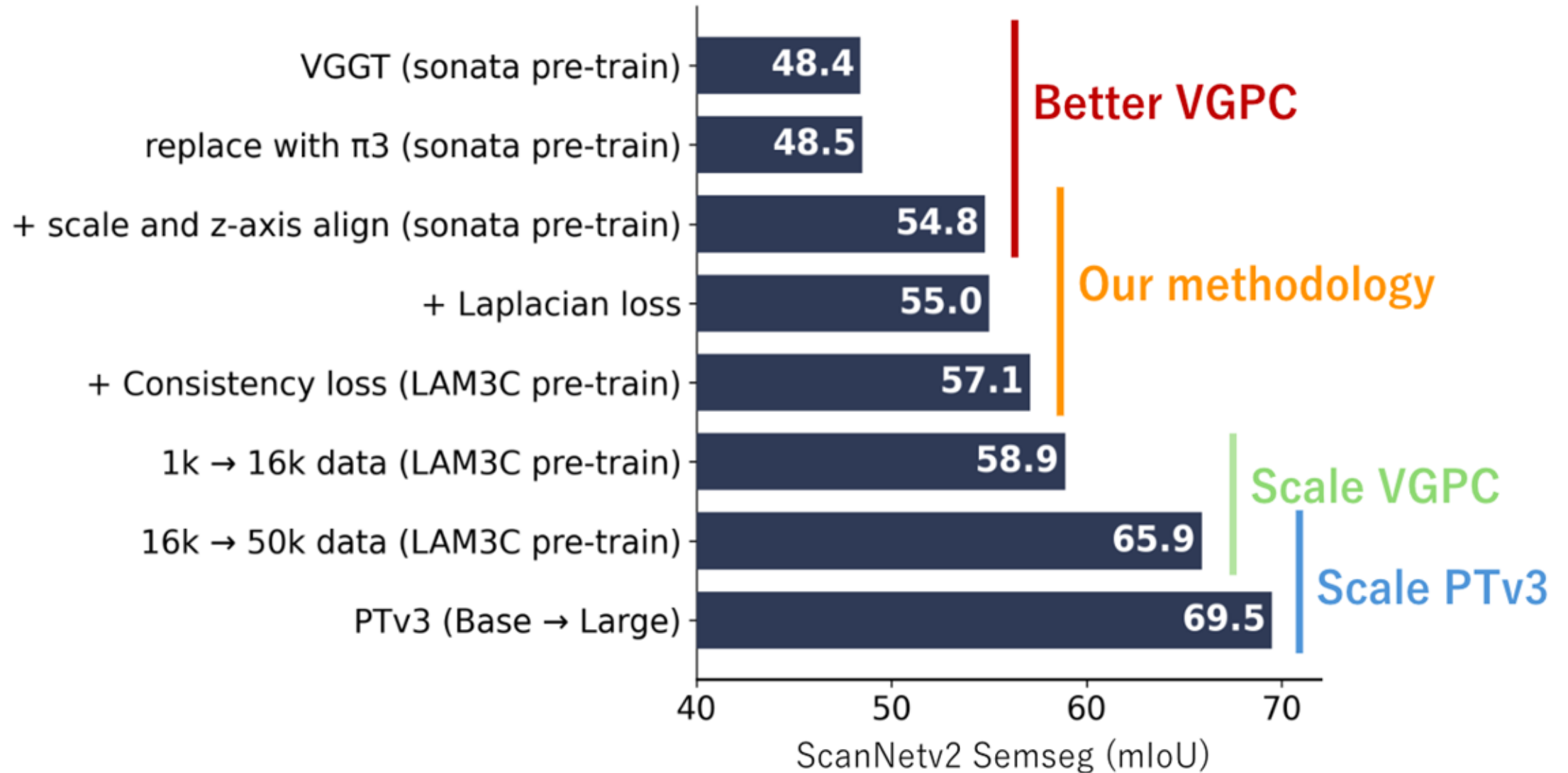
Semantic Segmentation on ScanNet



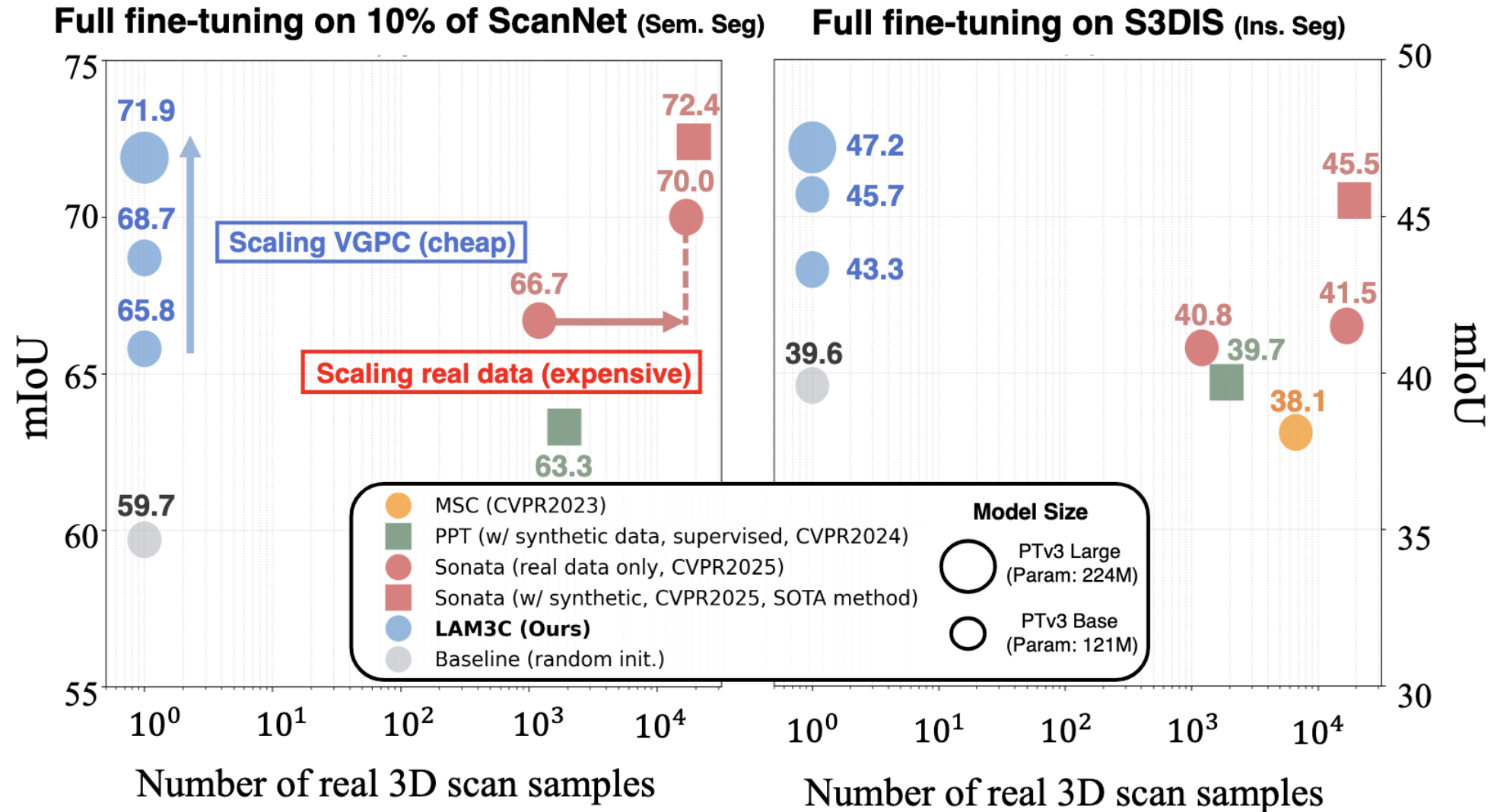
学習済みパラメータ
により初期化



スコア評価



スコア評価



限定資源下におけるマルチモーダル / 視覚基盤モデル構築 → 2D / 3D / 動画 / マルチモーダル から 実世界適応まで

① 視覚基盤モデル

Conventional approach

High resolution images → Manual evaluation (Resolution, Variety and diversity) → Original image / Segmented image

Our approach

Low resolution images (ImageNet, Places365, PASS) → Quality estimation → Object-based filtering → DiverSeg dataset (filtered low-resolution images) → Original image / Segmented image

Automated evaluation

超解像基盤モデル

[Ohtani+, ECCV 2024]

MoireDB

モアレ画像によるロバスト性向上

[Matsu+, CVPRW 2025]

生成データによる超解像学習

[Kodama+, CVPRW 2025]

④ マルチモーダルモデル

Random mask set \mathcal{M}

Image I → Mask powerset $2^{\mathcal{M}}$

Parse tree T

Image I / Sentence S → Embedding → Regions $A \subset \mathcal{M}$ / Phrases $B \in T$ → Loss

(a) CLIP (b) PowerCLIP (Ours)

視覚言語モデル

[Kawamura+, CVPR 2026]

Nth Embedding Models

Text Input → Diffusion Model → Average Similarity

Given Text Input: leaves falling on a pond in autumn, slow motion view

より良い生成学習データの探索

[Ohkubo+, WACV 2026]

基盤モデルが次世代の基盤モデルを構築

[Kupyn+, ICLR 2026]

② 3D空間基盤モデル

3D点群基盤モデル

[Yamada+, CVPR 2026]

③ 動画モデル

Existing Benchmark vs. Our Benchmark for Fine-Grained HOI Dynamics

Multiple-Choice Question (MCQ)

Action: Is the person sitting with their back to the camera?

Process: How does the person hammer the nail?

Location: Where does the person put down the hammer?

State Change: How did the state of the cylinder change?

Object Parts: What part of the cylinder was hammered?

動画基盤モデルベンチマーキング

[Tateno+, CVPR 2026]

衛星画像基盤モデル

実世界適応

人工衛星「だいち2号」の観測データを活用して国土に特化したSAR基盤モデルを構築

— SAR観測データへのAI利用をより手軽に —

土地用途の多様性を考慮したSAR画像学習データセットの作成

大規模教師無し学習 → 国土SAR基盤モデル

目的に合わせた応用

- 土地利用・被覆推定
- SAR画像 推定画像
- 災害検知
- 変化検知

将来展望

MAE → SAR画像 MAE → SAR観測に特化したMAEの構築

基盤モデルの統合 → 言葉による説明の付与

テキストの基盤モデル

この衛星画像は... "山岳地帯が中心、山間に市街地あり"

野生動物保全ベンチマーク

AnimalClue

Logos from the wildlife tracking and monitoring

自動運転フォークリフト

「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で高度の異常を検出

https://www.aist.go.jp/aist_1/magazine/20250205.html

多数の基盤モデル開発 - データ収集・モデル構築 - やそれら横展開・社会実装を経て今後さらに視覚基盤モデル構築を改善する

