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/

コア技術

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.

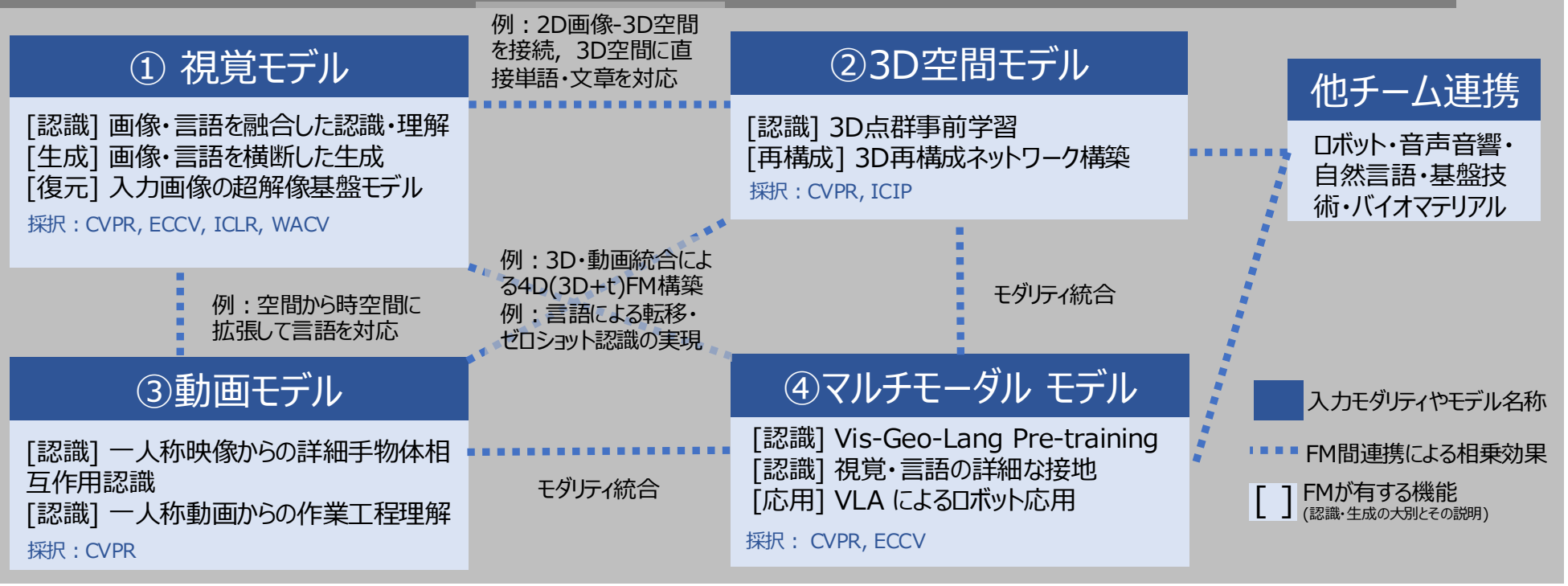
Ego4D(データ):
ウェアラブルセンサ等 一人称視点動画における大規模データ. 人間視点からの汎用的認識を目指して構築され, 自動車やロボットへの適用も可能.

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.

基盤モデル



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.

実世界適応

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】

コア技術

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.

基盤モデル

◆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
を実施してきたメンバーが参画

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

③動画モデル
[認識] 一人称動画からの作業工程理解
採択：CVPR

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)

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.

実世界適応

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】

コア技術

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.

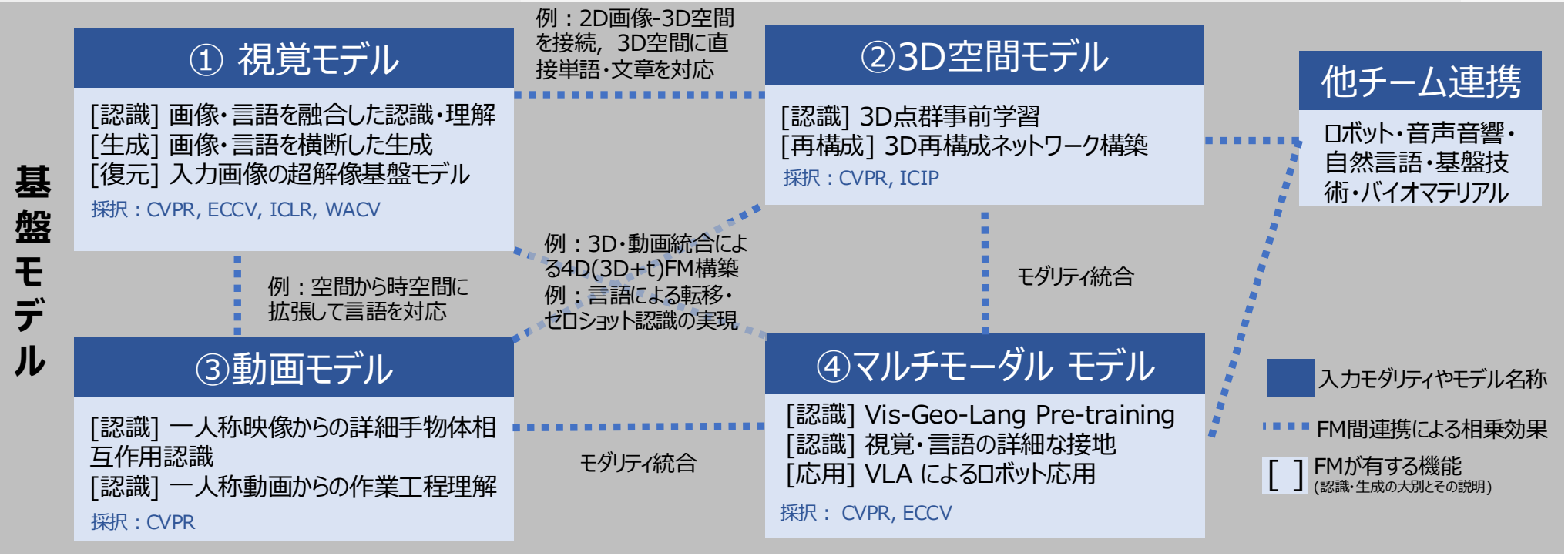
Ego4D(データ):
ウェアラブルセンサ等 一人称視点動画における大規模データ. 人間視点からの汎用的認識を目指して構築され, 自動車やロボットへの適用も可能.

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.

基盤モデル



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.

実世界適応

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:
- 内視鏡画像の自動診断
- 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

視覚+αのモダリティ①②③④ごとに基盤モデル構築

【政策予算・画像チーム】

コア技術

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.

基盤モデル

① 視覚モデル

[認識] 画像・言語を融合した認識・理解
[生成] 画像・言語を横断した生成
[復元] 入力画像の超解像基盤モデル
採択：CVPR, ECCV, ICLR, WACV

例：2D画像-3D空間を接続, 3D空間に直接単語・文章を対応

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

例：3D・動画統合による4D(3D+t)FM構築

モダリティ統合

視覚 / マルチモーダル基盤モデルの応用先も含めて検討

③動画モデル

[認識] 一人称映像からの詳細手物体相互作用認識
[認識] 一人称動画からの作業工程理解
採択：CVPR

モダリティ統合

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)

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.

実世界適応

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】

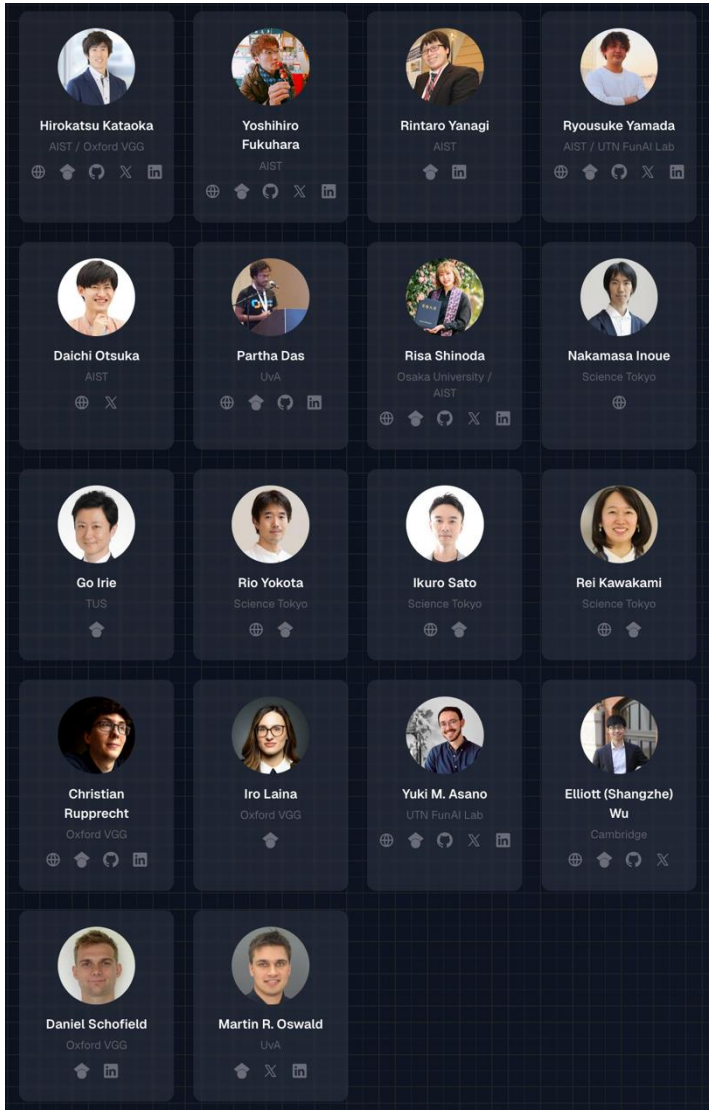
LIMIT.LAB

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

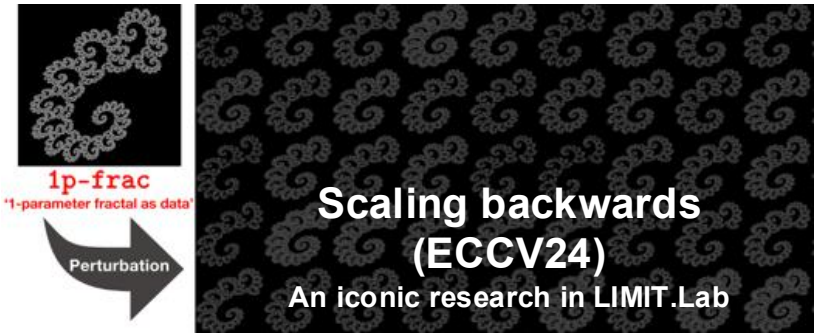


https://limitlab.xyz/

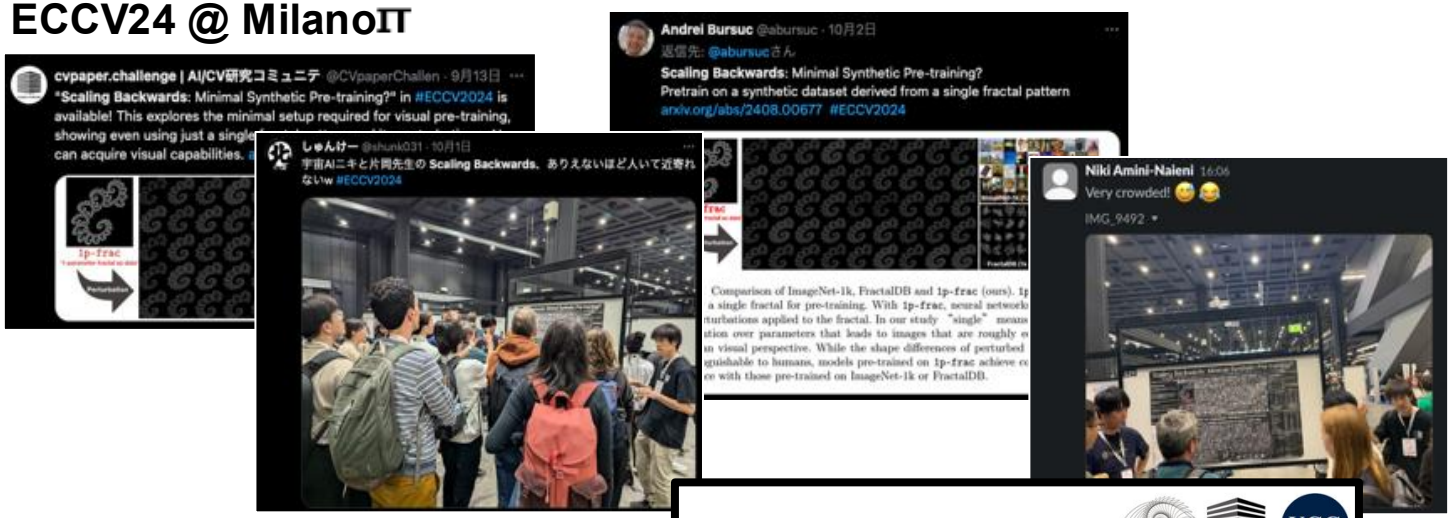
AIST 主導の研究イニシアティブとして機能

直近数年の主な国際活動・連携・研究

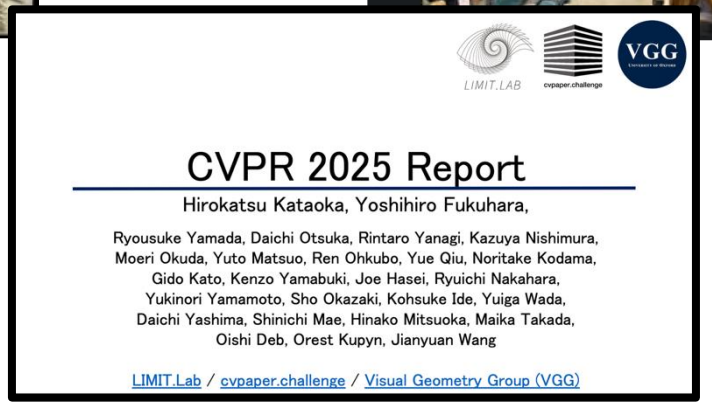
- 2023年10月 ICCV 2023 / 2024年6月 CVPR 2024 / 2025年10月 ICCV 2025 にて LIMIT Workshop 開催
- 2025年6月 国際研究コミュニティ LIMIT.Lab 設立
- 2025年 国内外コミュニティ, AIST PJから合計70件超の主著・共著論文投稿
- その他 招待講演・国際ワークショップ・国際交流会など定期開催



ECCV24 @ Milano IT



LIMIT Workshop Organizers



トップ国際会議 ICCV 2025 @ Hawaii 開催

2025年10月 ICCV 2025 Workshop 開催

トップ国際会議 ICCV 2025 @ Hawaii 開催


Opening remarks


Photo by Seokeon Choi (Best Paper at LIMIT)

Francis Engelmann (Stanford Univ)

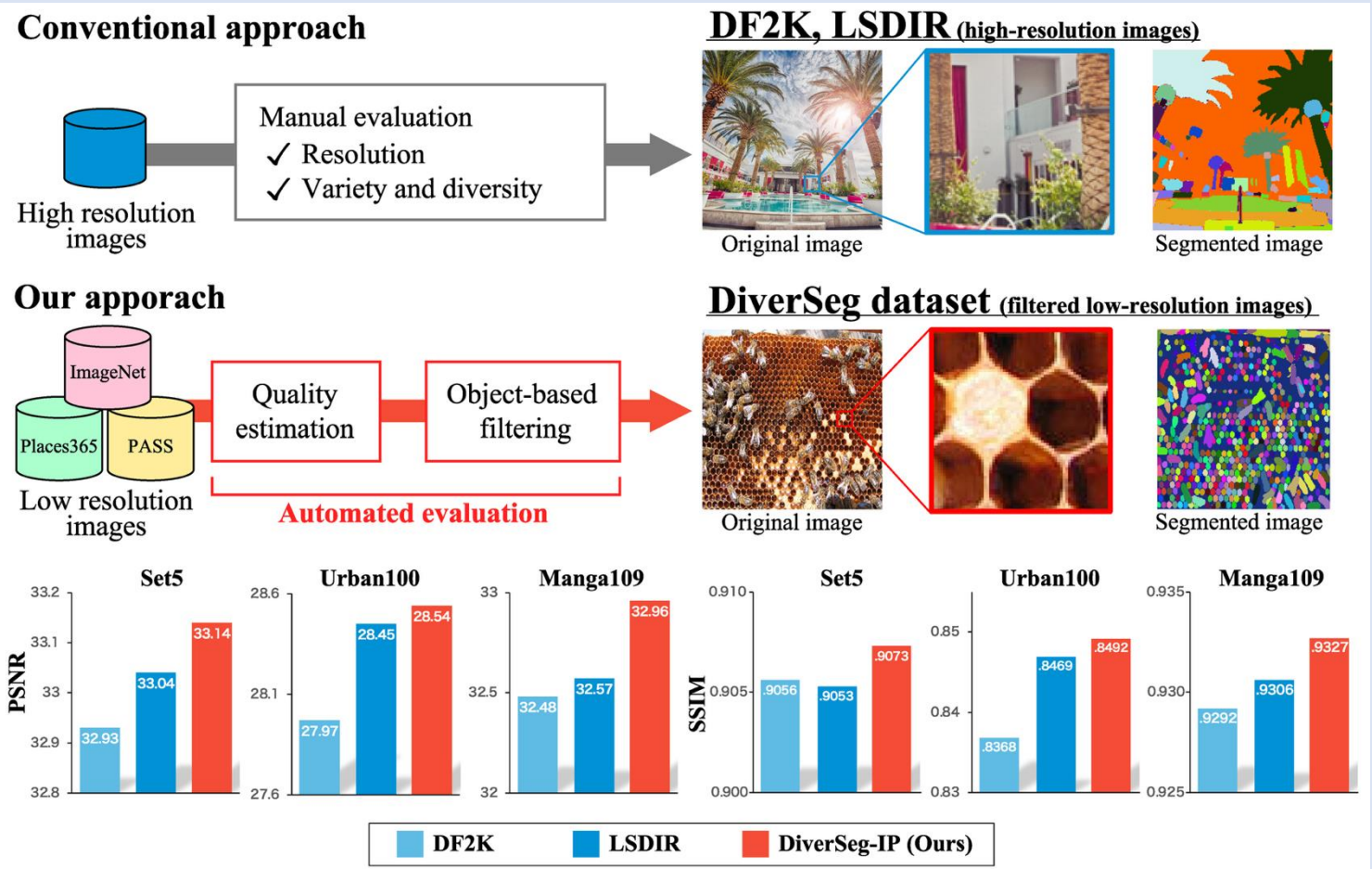
国際活動を経るごとに
人脈形成 → 研究コミュニティ →
研究チーム → 研究自体
が強化される枠組みとなっている

Yossi Gandelsman (Reve)

Michael Ryoo (Stonybrook / Salesforce)

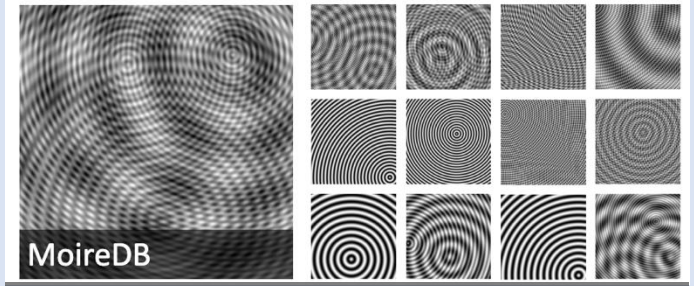

Photo by Jack Langerman
https://x.com/jacklangerman/status/1970111913037299992

Yonglong Tian (OpenAI)



超解像基盤モデル

[Ohtani+, ECCV 2024]



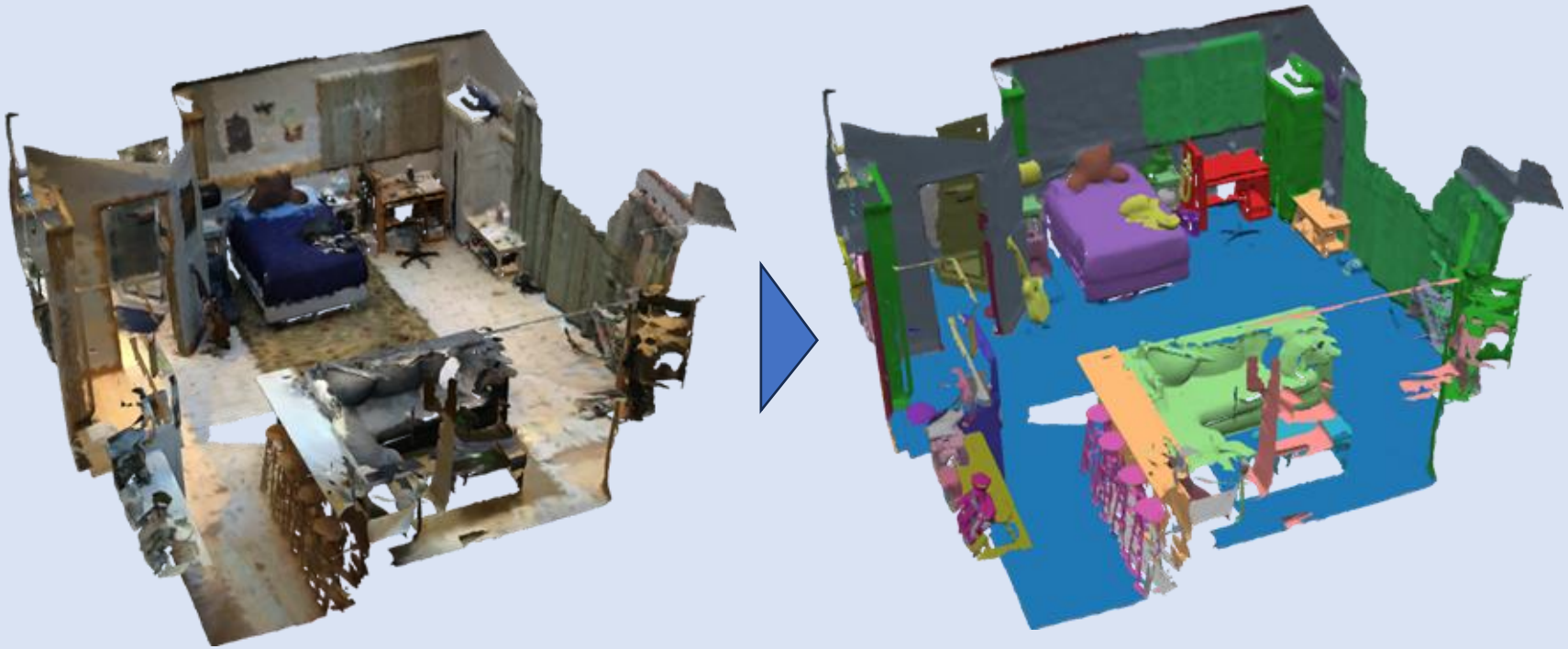
モアレ画像による ロバスト性向上

[Matsuo+, CVPRW 2025]

生成データによる 超解像学習

[Kodama+, CVPRW 2025]

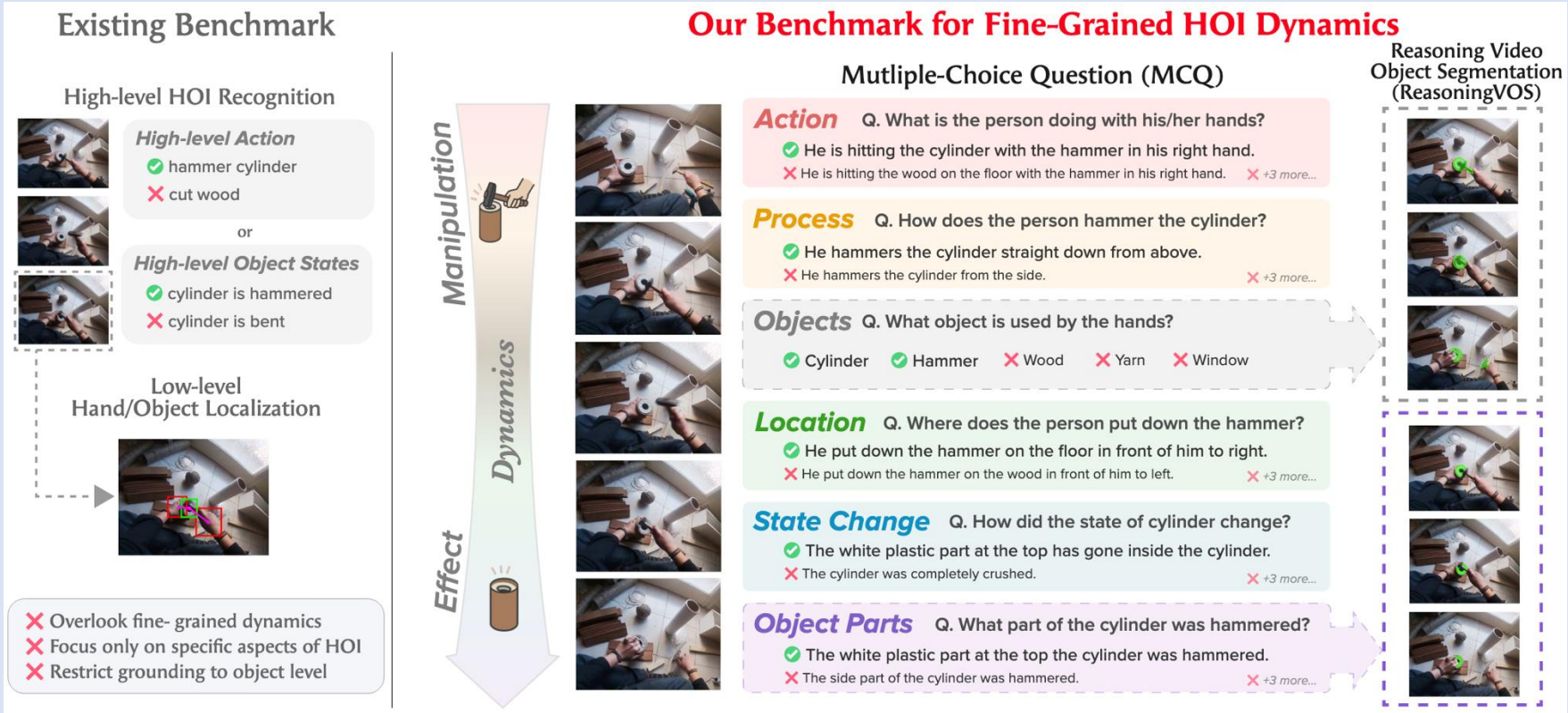
② 3D空間基盤モデル



3D点群基盤モデル

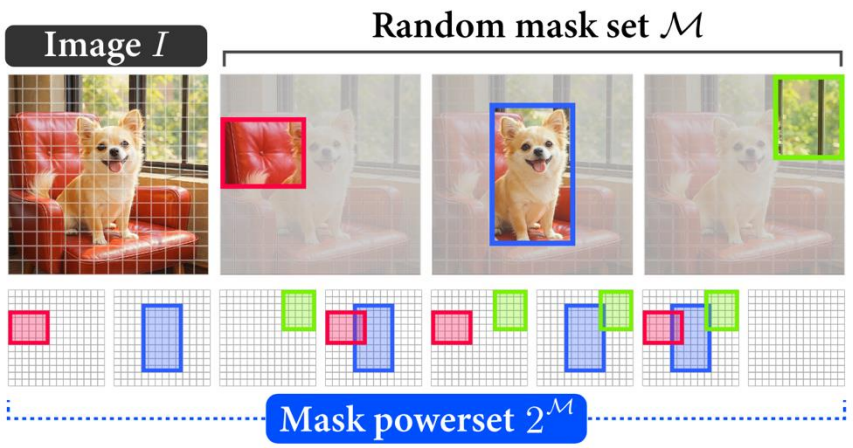
[Yamada+, CVPR 2026]

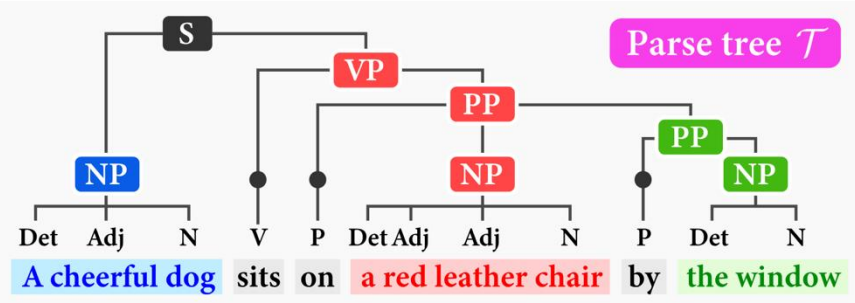
後で詳しく紹介

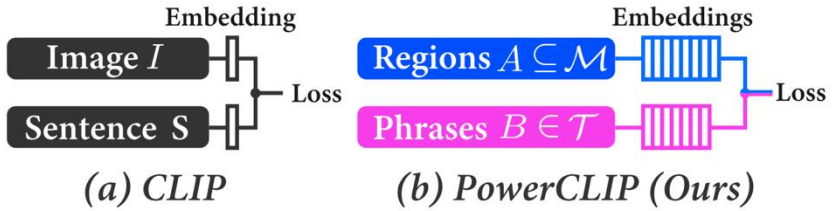


動画基盤モデルベンチマーキング

[Tateno+, CVPR 2026]

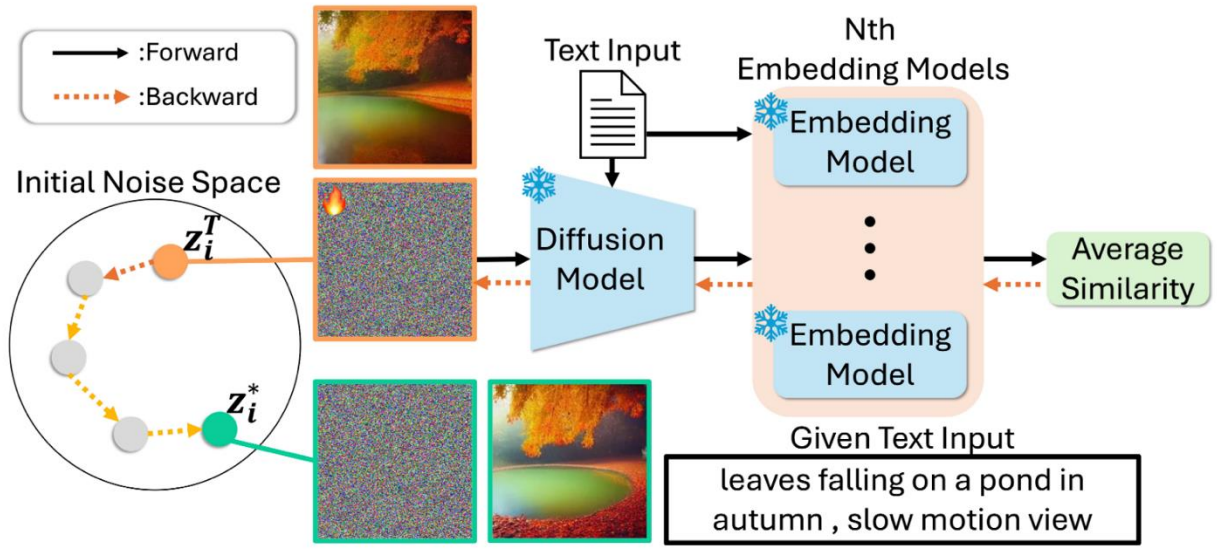






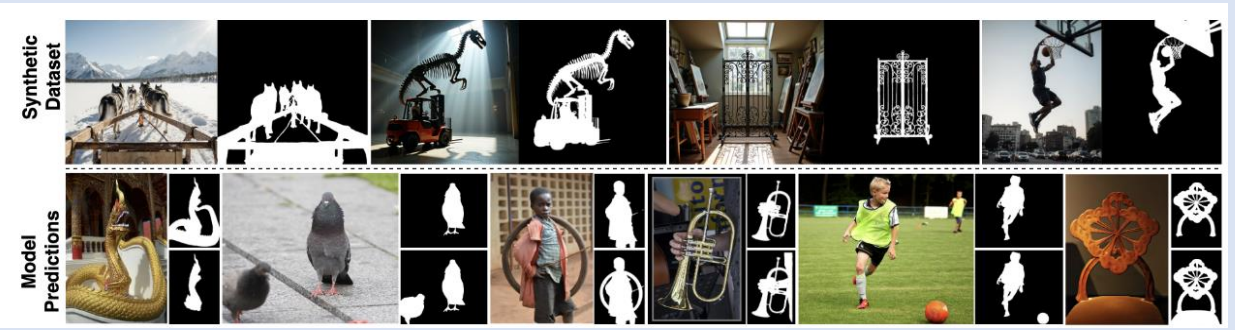
視覚言語モデル

[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

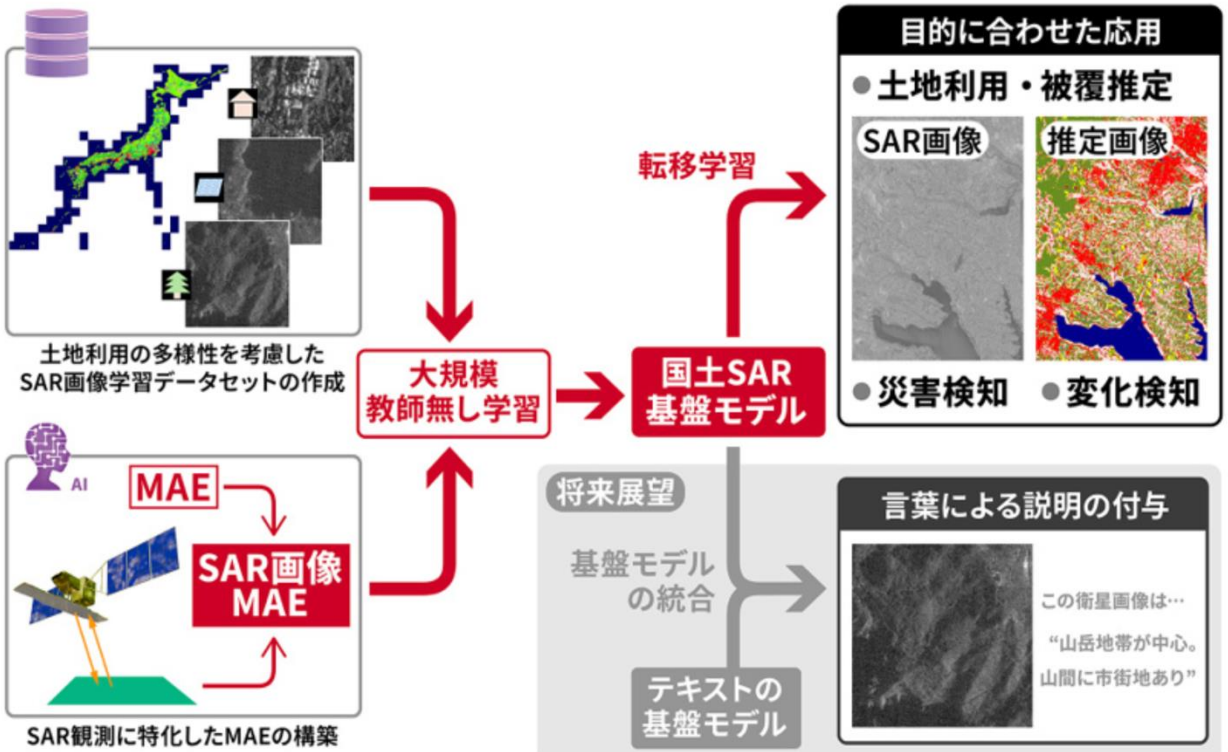
[Kupyn+, ICLR 2026]
後で詳しく紹介

実世界適応

衛星画像基盤モデル

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－



野生動物保全ベンチマーク

[Shinoda+, ICCV 2025]

自動運転フォークリフト

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

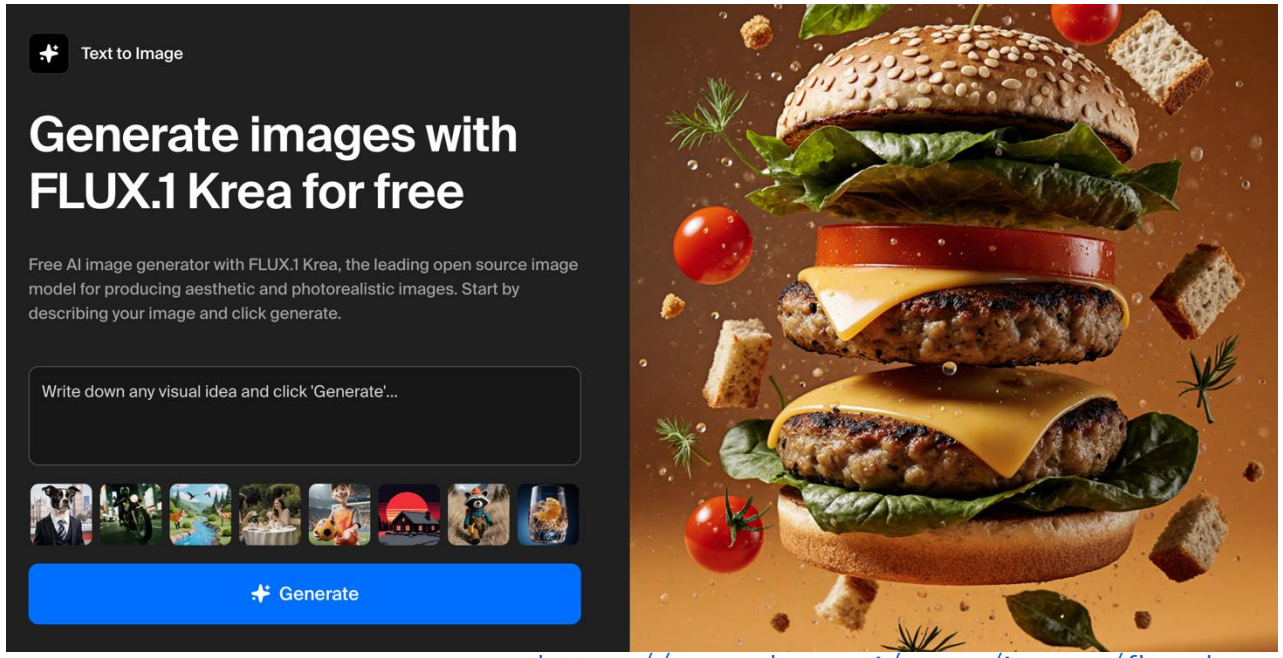
Oxford VGG / AIST

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)



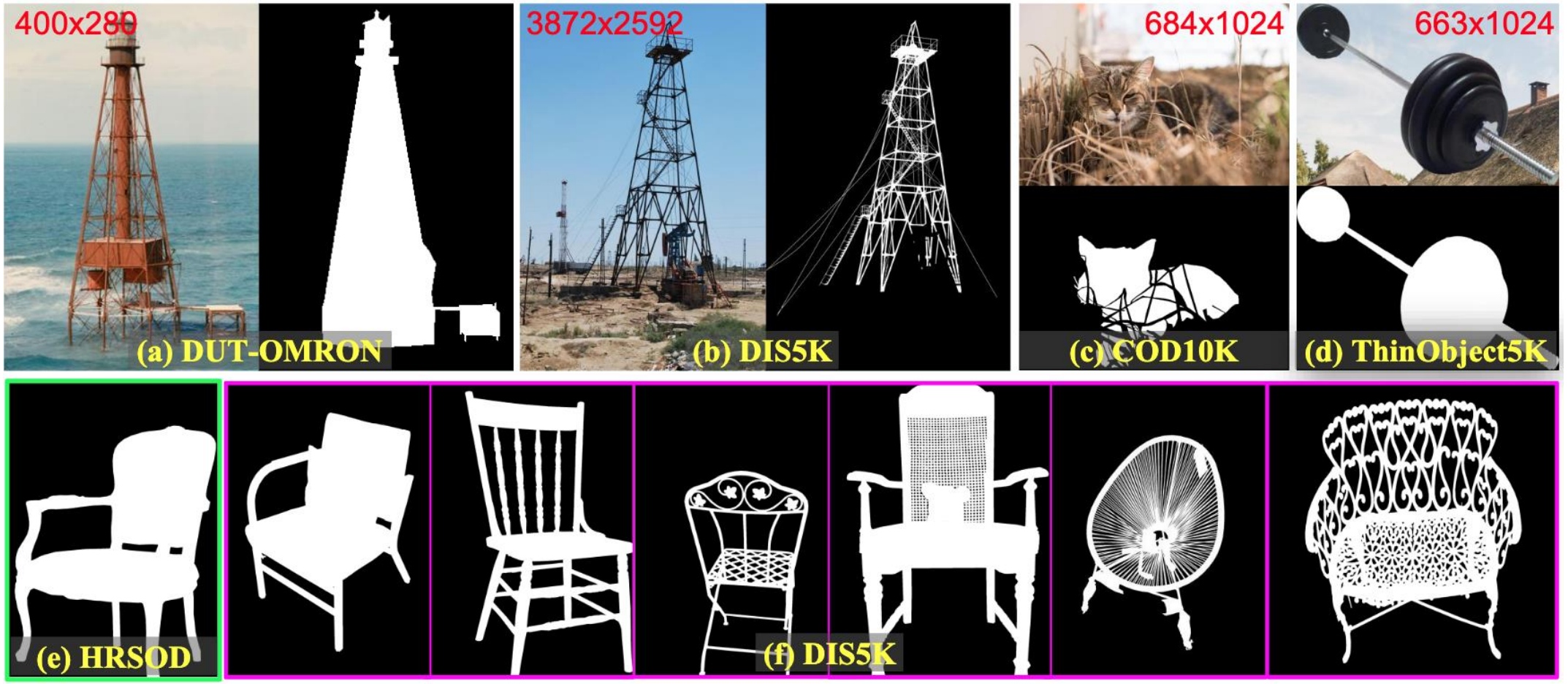
https://www.krea.ai/apps/image/flux-krea

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



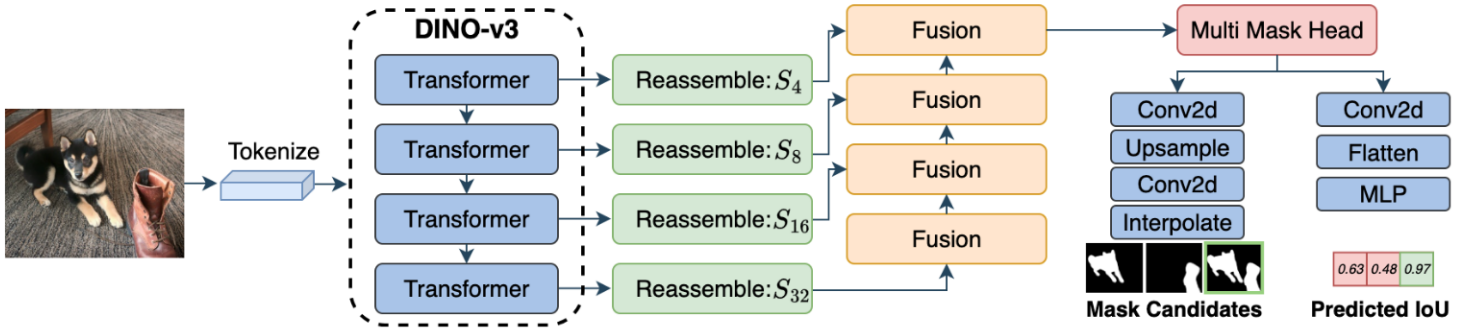
https://arxiv.org/abs/2203.03041

概要 & 貢献

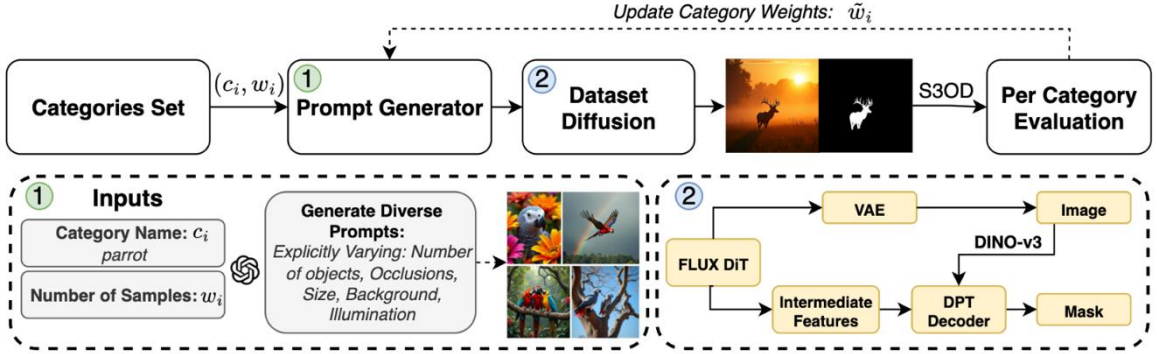
- 高精細セグメンテーションのデータを自動生成
- 合成データにより単一・単純モデルでも汎化性を担保
- Text prompting, DINOv3特徴, 高精度生成マスクにより最高水準のモデル

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	.863	.856	.906	.049
S3OD	S3OD	**.881**	**.884**	**.925**	**.039**

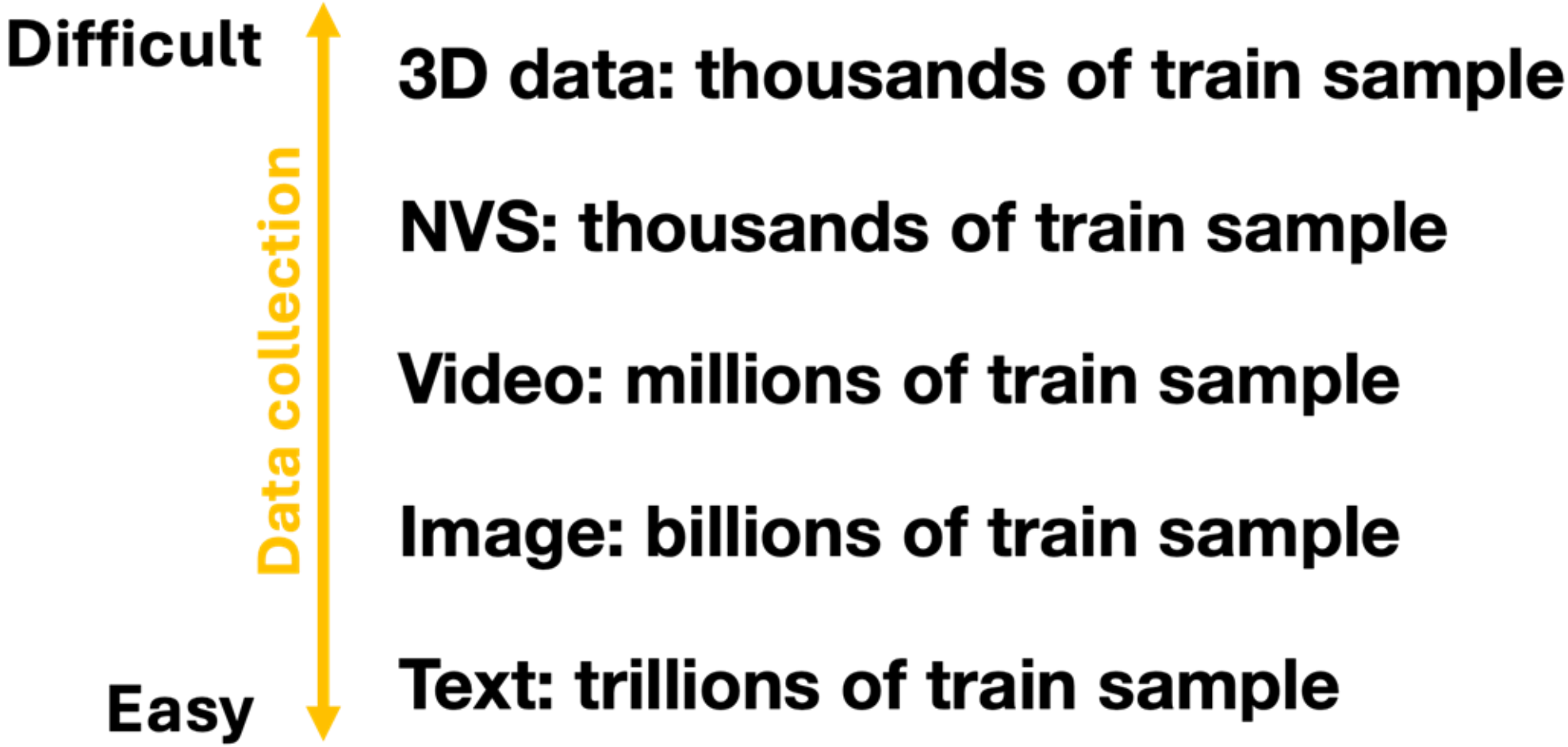
3D sans 3D Scans:
Scalable Pre-training from Video-Generated Point Clouds

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



Video x Reconstruction x SSL

Research questions:

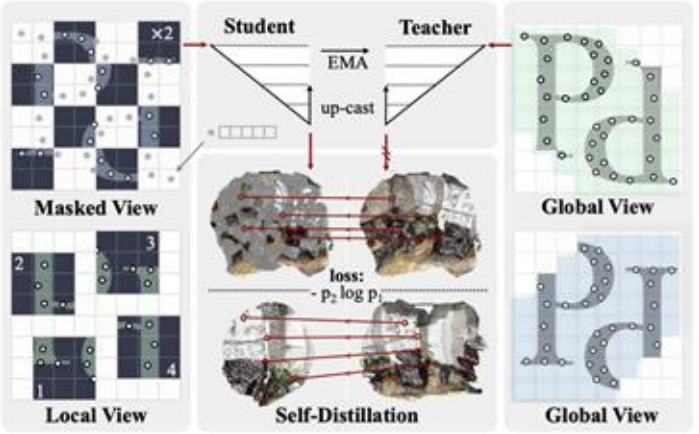
- ラベルなしの動画から3D表現を獲得できるのか？
  + 再構成データ: 効率的な3D 事前学習データが作成できる?
  + 点群モデルのスケーリング: 大規模なデータは学習可能か?
  + 統一的なモデル: 屋内外のデータに対して適応可能なのか？

Unlabeled video

[Shashanka Venkataramanan, ICLR2024]

Reconstruction

[Jianyuan Wang, CVPR2025]

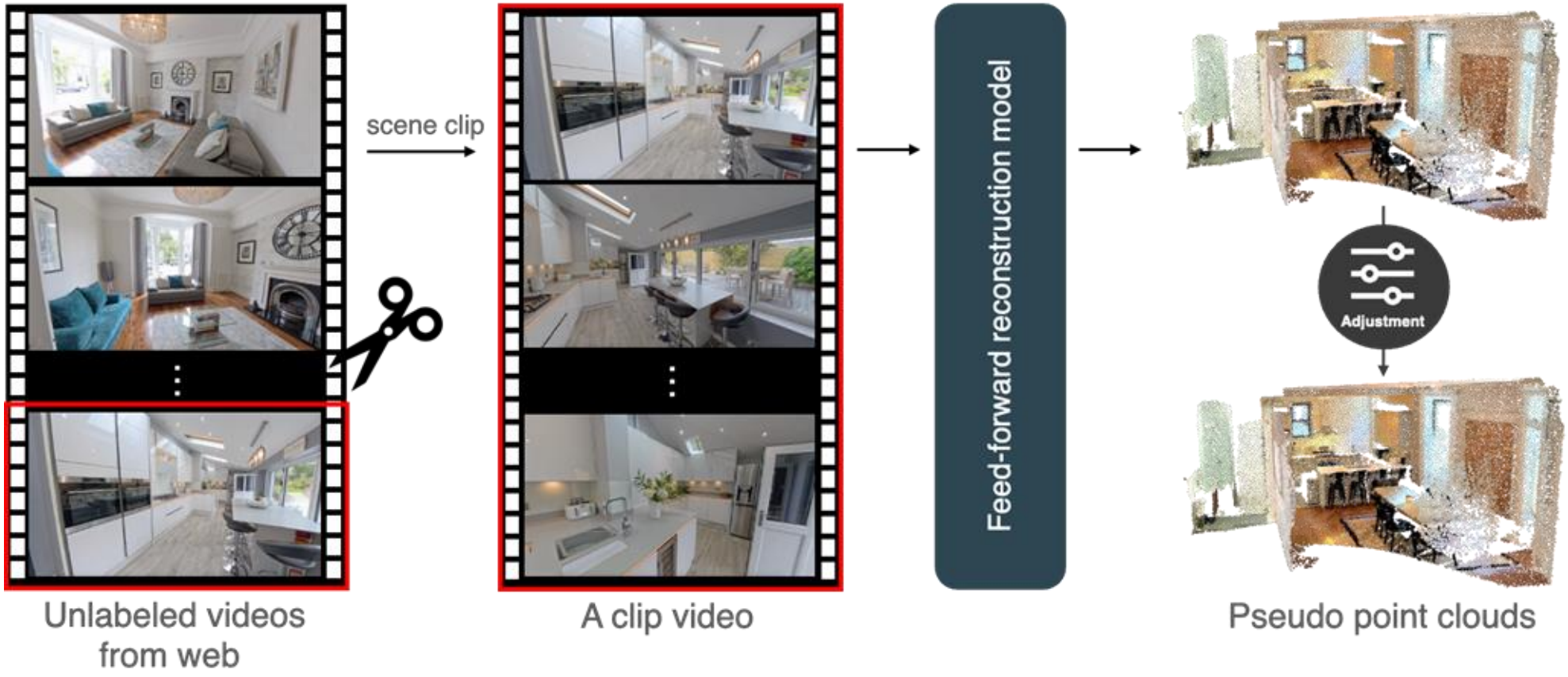


SSL

[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



学習設定

1. 事前学習

Sonata on fake point clouds (RoomTours)

Sonata

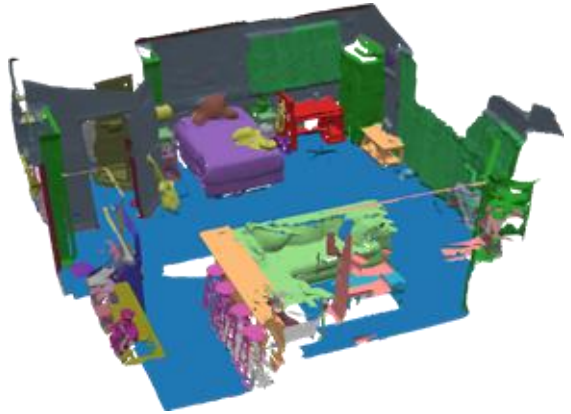
自己蒸留
Self-distillation

2. Linear Probing / Full Fine-tuning

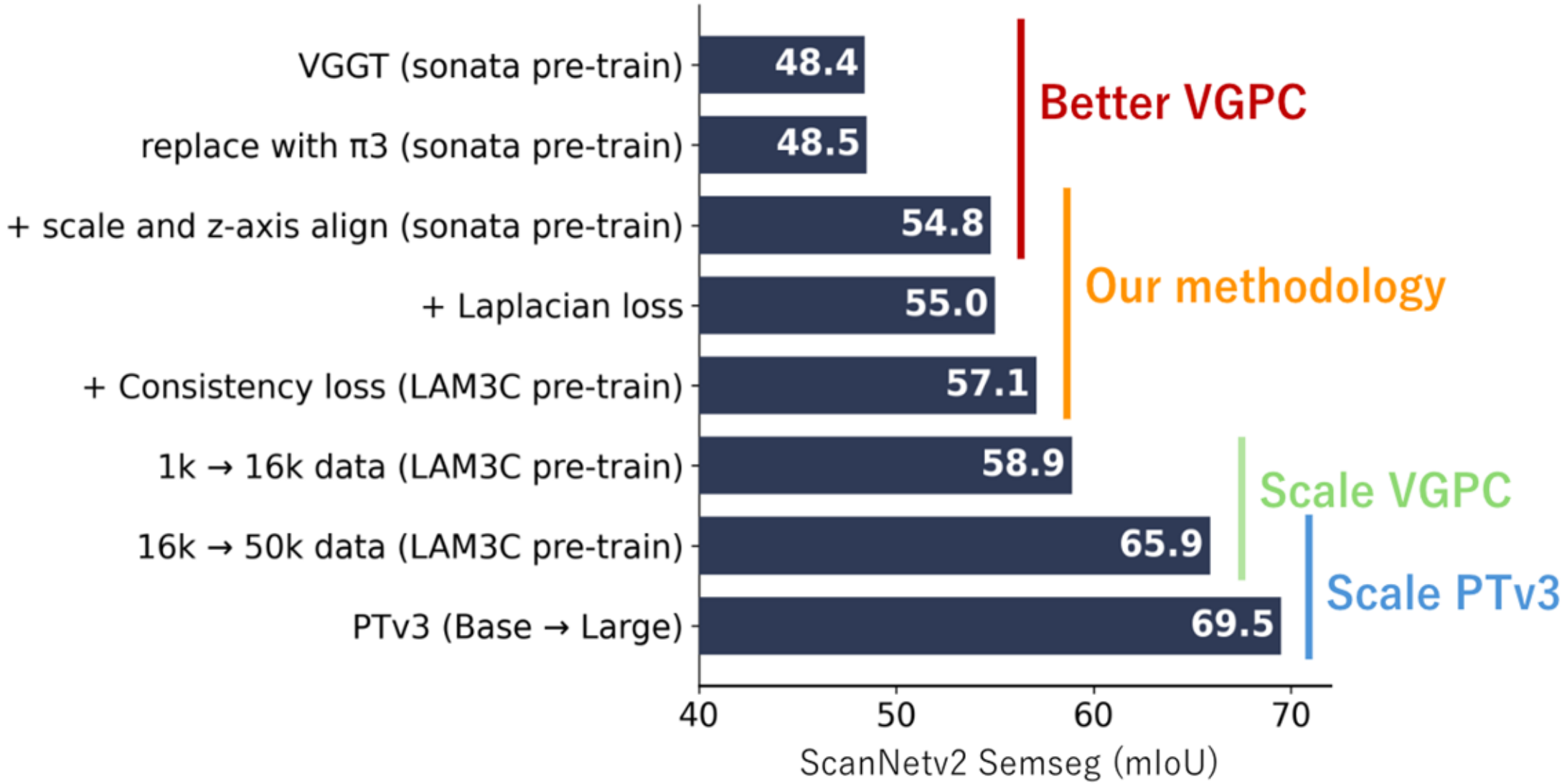
Semantic Segmentation on ScanNet

学習済みパラメータ
により初期化

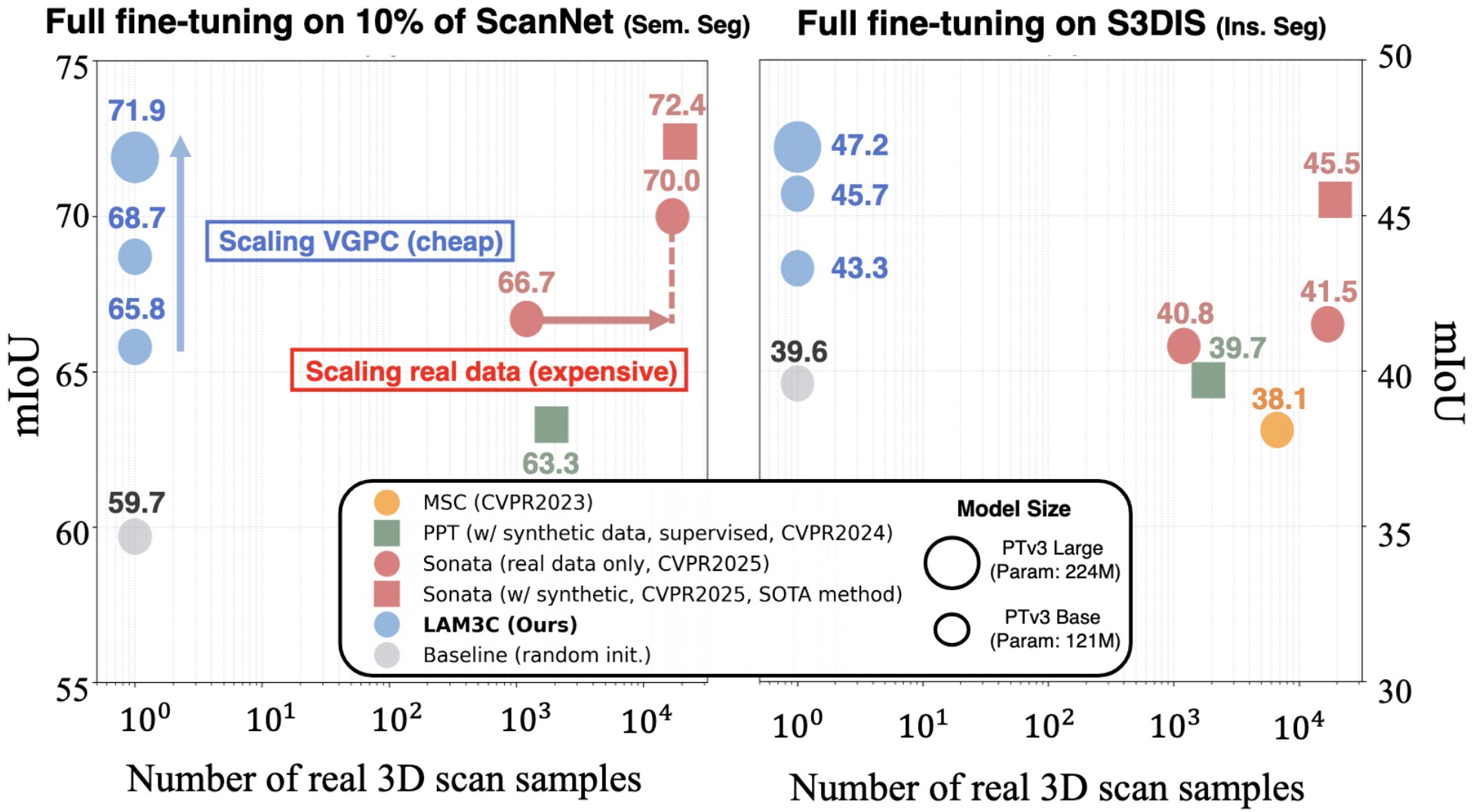
PTv3
(Point Transformer)



スコア評価


VGGT (sonata pre-train)
48.4
replace with π3 (sonata pre-train)
48.5
+ scale and z-axis align (sonata pre-train)
54.8
+ Laplacian loss
55.0
+ Consistency loss (LAM3C pre-train)
57.1
1k → 16k data (LAM3C pre-train)
58.9
16k → 50k data (LAM3C pre-train)
65.9
PTv3 (Base → Large)
69.5
40
50
60
70
ScanNetv2 Semseg (mIoU)
Better VGPC
Our methodology
Scale VGPC
Scale PTv3

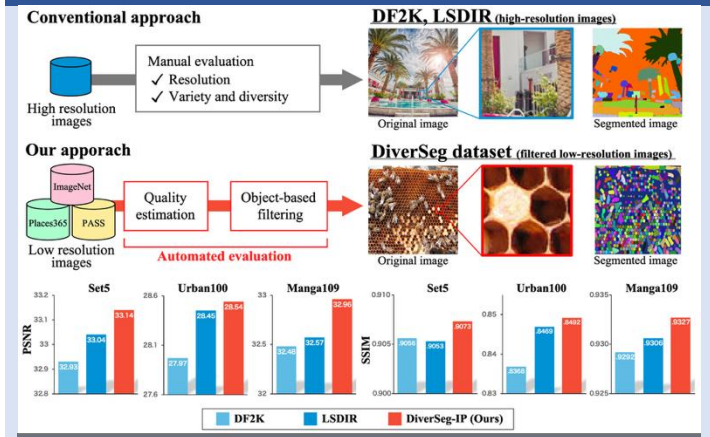
スコア評価


Full fine-tuning on 10% of ScanNet (Sem. Seg)
Full fine-tuning on S3DIS (Ins. Seg)
mIoU
Number of real 3D scan samples
Number of real 3D scan samples
71.9
68.7
65.8
59.7
72.4
70.0
66.7
63.3
47.2
45.7
43.3
39.6
45.5
41.5
40.8
39.7
38.1
Scaling VGPC (cheap)
Scaling real data (expensive)
MSC (CVPR2023)
PPT (w/ synthetic data, supervised, CVPR2024)
Sonata (real data only, CVPR2025)
Sonata (w/ synthetic, CVPR2025, SOTA method)
LAM3C (Ours)
Baseline (random init.)
Model Size
PTv3 Large (Param: 224M)
PTv3 Base (Param: 121M)

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

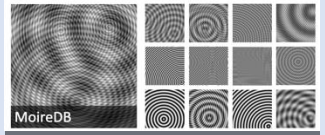
AIST
Create the Future, Collaborate Together

① 視覚基盤モデル



超解像基盤モデル

[Ohtani+, ECCV 2024]



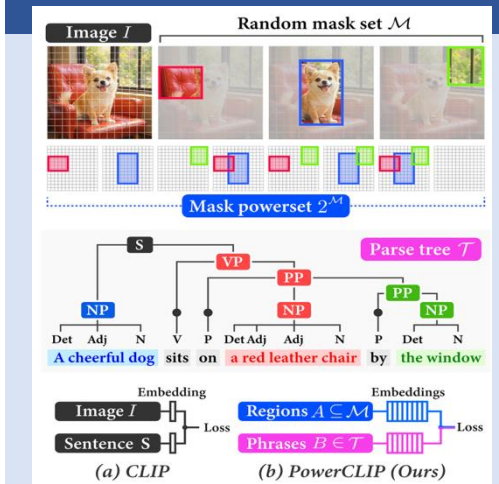
モアレ画像による
ロバスト性向上

[Matsuo+, CVPRW 2025]

生成データによる
超解像学習

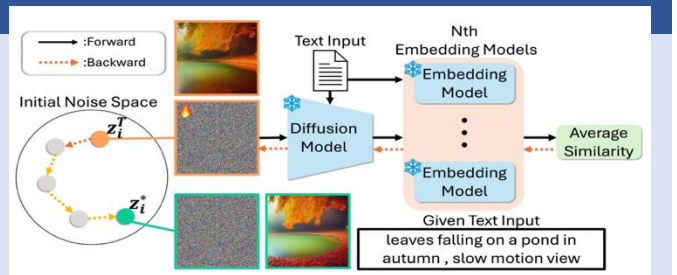
[Kodama+, CVPRW 2025]

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

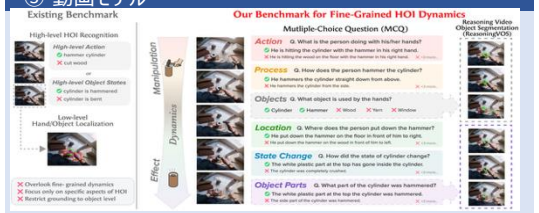
[Kupyn+, ICLR 2026]

② 3D空間基盤モデル

3D点群基盤モデル

[Yamada+, CVPR 2026]

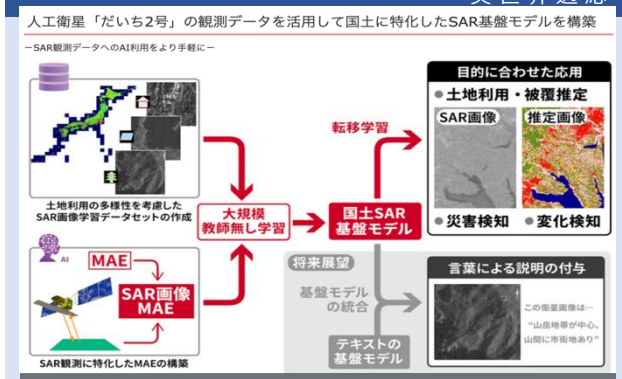
③ 動画モデル



動画基盤モデルベンチマーキング

[Tateno+, CVPR 2026]

実世界適応



衛星画像基盤モデル

野生動物保全ベンチマーク

[Shinoda+, ICCV 2025]

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