

「模倣 AI エージェントによる人物行動理解」

研究期間：2020 年 11 月～2023 年 3 月

研究者：大川 武彦

加速フェーズ期間：2023 年 4 月～2024 年 3 月

1. 研究のねらい

少子高齢化社会や後継者不足の問題は極めて深刻であり、国内では熟練ノウハウを持つ団塊世代の大量退職によるモノづくり産業の減衰が顕著となっている。持続的成長可能な社会の実現に向けて、デジタル技術を活用した熟練者技能の継承・伝承も検討されるべきである。熟練者の動作には、勘やコツなどの個人の感性に帰属する「暗黙知」の割合が多く、先端人工知能技術やロボティクスなどの最新研究を導入しても形式化することが難しい現状である。

この熟練者動作のモデリングとして、模倣アルゴリズムを利用して人の手動作系列の生成を行い、シミュレーション可能、またはロボット実行可能にする試みがなされている。しかし、このような模倣による人工知能やロボットの構成法は、このシステムの入力となる手操作情報の取得が困難であることから実験室環境等の制限された条件下でしか応用が進んでいない。

本研究では、調理、組み立て、生化学実験等の様々な人の作業映像から手動作系列を取得する技術を開発する。具体的には、実験室環境などの既存のラベル付けされた手操作認識データセットから学習したモデルを、それらの応用場面ごとに適応する戦略を考える。この新規映像における手操作認識の課題を解決することは、調理や製造、生化学実験を行うロボットの開発やその応用の促進へ貢献するものである。

加速フェーズでは、これまでの採択期間で開発した手操作認識モデルを時間情報を考慮するように拡張し、また下流タスクの行動認識との関係理解を目的として研究を行う。既存のフレーム単位の手操作認識では、作業映像から時間方向に一貫した手動作系列を取得することが難しく時間方向への拡張は必要である。また、時間を考慮した手操作認識からその行動についての認識は、様々な応用場面で手操作に着目して行動理解を行う際に有用な方法論である。さらに、認識した手作業の実行手続きの解析は、ロボットへその作業の実行手順を直接教示することを可能にする。

2. 研究成果

(1) 概要

模倣による人工知能やロボットの構成に向けて、様々な作業映像において人の手操作を認識することは重要な課題である。手操作に関する情報の中で手領域抽出、手姿勢推定、手物体検出等のタスクは、手がどのように物体を操作するかを詳細に表現することができる。

しかし、実世界の新しい映像データに手操作認識モデルを適用した場合のこれらのタスクの精度は未だ低く、どのようにこの新規映像にて高い性能の学習器を構成するかは非自明である。既存の手操作認識に関するデータセットは、手領域、手関節、手と物体位置に関するアノテーションの難易度の高さから環境の多様性が限定的な実験室で構築されている。このようなデータセットをもとに学習した手操作認識モデルは、新規映像において認識精度が

悪化し、手操作に関する情報の自動取得に利用できない。

上記の課題を解決するために、本研究では新規映像データが与えられた際に、自己教師あり学習やデータ拡張を活用して、手操作認識モデルを改善する方法を提案する。それらの詳細について後段にて記述する。

また、本 ACT-X 研究を通じて国内外の研究機関との共同研究や対外交流、国際的な会合への参加が実現した。アメリカのカーネギーメロン大学へ訪問研究者として滞在して研究実績を積むことができ、オムロンサイニックス社と産学連携して研究を実施した。さらにチューリッヒ工科大学、Meta 社、Microsoft 社などの有名研究機関へ訪問して、私の専門分野とその隣接分野の研究の状況、今後の社会実装への見通しなど議論することができた。

加速フェーズでは、第一に、前年度の研究を深化させ、国際学会 ICCV にて手操作認識に関するワークショップ、そして一人称視点手姿勢推定に関するコンペティションを開催した。第二に、手操作認識の時間拡張、行動理解との接続に関する研究を実施した。実際の調理映像から手物体の位置情報をトラッキングして、キャプションとしての作業動作記述する手法を提案した。また、本研究では既存の離散行動ラベルを出力するのではなく、動画キャプションという自然言語形式の出力に挑戦し、このタスクで利用可能な一人称視点調理映像データセットも構築した。

また、加速フェーズ期間の研究は、主にチューリッヒ工科大学の訪問研究員としてスイスへ滞在して研究を実施した。さらにシンガポール国立大学におけるセミナー講演や、国際会議 ICCV, ICVSS における発表など行い、自身の研究のアウトリーチの良い機会となった。一人称視点調理映像データセット構築のプロジェクトに関しては、オムロンサイニックス社と産学連携して、被験者を集めて映像収録、アノテーションを行い、作業内容の解析を行った。

(2) 詳細

A. 新規映像における手操作認識

様々な作業映像にて手操作認識を行うために、学習モデルが未だ経験していない新規の映像データにおける推論精度を向上させる研究を実施した。特に、手領域抽出、手姿勢推定、手物体検出のタスクは詳細なレベルでの手操作の状態を表現するため、ユーザの行動理解やロボットによる模倣に向けて重要な課題である。

1. 手領域抽出タスク (IEEE Access 2021 採択[1])

手領域抽出は、画像から手のピクセル単位のラベルを予測するタスクであり、手位置の同定、指先の状態の解析等に利用できる。研究[1]では、新規映像におけるユーザの手の見えと背景の環境の相違を明示的に解消し、信頼できる形式で自己教師あり学習を行う手法を提案する(図 1)。第一に、データ間の前景・背景に関する見えの相違を解消するために、画像スタイル変換ネットワークを提案し、前景同士、背景同士の画像スタイルの差異を解消した。第二に、ラベルのない新規映像に擬似的なラベルを付与してモデルを再学習する擬似ラベリング手法を提案した。ここでは、2 体のネットワークから予測の確信度を推定して、信頼できる擬似ラベルの

みを再学習に使用することで性能向上を図った。

本研究成果は、国際ジャーナル IEEE Access 2021 へ採択され、国内会議 MIRU2021 にてロングオーラル発表を行い、学生奨励賞を受賞した。さらに、本研究はオムロンサイニックス株式会社と共同研究を実施した成果である。

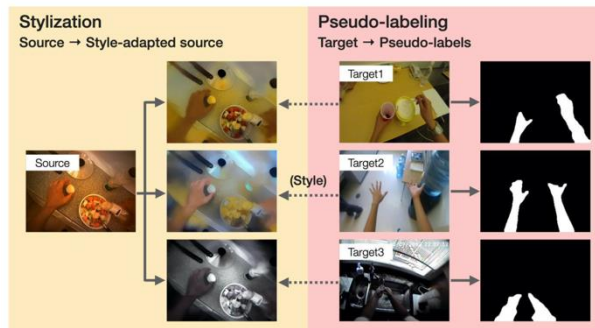


図 1: 新規映像における手領域抽出手法

2. 手領域抽出と手姿勢推定タスク (ECCV 2022 採択 & ワークショップ x2 発表[2])

[1]からさらに発展させ、新規映像において手姿勢推定と手領域抽出を同時に解く課題に取り組んだ。手姿勢推定は、手のキーポイントの座標を回帰するタスクであり、手領域より明示的に手の把持姿勢を推定するものである。しかし、手姿勢アノテーションは一層困難であり、利用可能なラベルありデータは実験室環境のものに限られる。研究[2]では、実験室データから構成された既存モデルを屋外などの非常に異なる撮影条件でうまく推論させることを目的とした (図2)。

本研究では、新規映像データ上で幾何的なデータ拡張前後における自己教師あり学習手法を提案した。データ拡張前後の一貫性学習に予測の信頼度推定、Teacher-Student 学習機構を組み合わせることで安定して性能向上することを発見した。

本研究は、カーネギーメロン大学へ訪問研究員として滞在していた期間の成果であり、一流国際会議 ECCV 2022 に採択された。さらに、同会議から手操作認識に関するワークショップ、複数視点からの人物行動理解に関するワークショップに招待され、ポスター発表を行なった。加速フェーズ期間では国内学会の MIRU2023 にて本研究の招待発表を行なった。

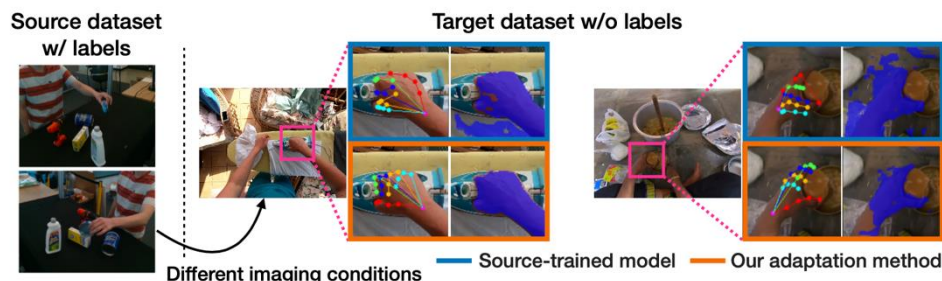


図 2: 新規映像における手姿勢推定と手領域抽出

3. 手物体検出タスク (ECCV 2022 ワークショップ発表[3])

手を対象とするだけでなく操作している物体も同時に検出することは、手と物体の作用関係を明らかにする役割を持つ。新規映像において、2枚の学習画像を重ね合わせてデータの偏りを緩和する Mixup と呼ばれるデータ拡張が有効であることが示されている。しかし、手・物体検出においては、2枚の手作業中の画像を混合すると、特定の領域に手や物体が集中して、手・物体境界の識別能力が低下するなど、意図しないバイアスが発生する。研究[3]は、手・物体検出におけるこの意図しないバイアスを軽減しつつ、データ混合の効果を活用する Background Mixup と呼ばれるデータ拡張手法を提案する。図 3 のように、生化学実験等の小規模の学習データに、既存の大規模な背景データを混合することで、データの多様性を向上させ、手・物体検出器をより頑健にする。

本研究成果は、ECCV 2022 の手操作認識に関するワークショップにて採択され、ポスター発表を行った。

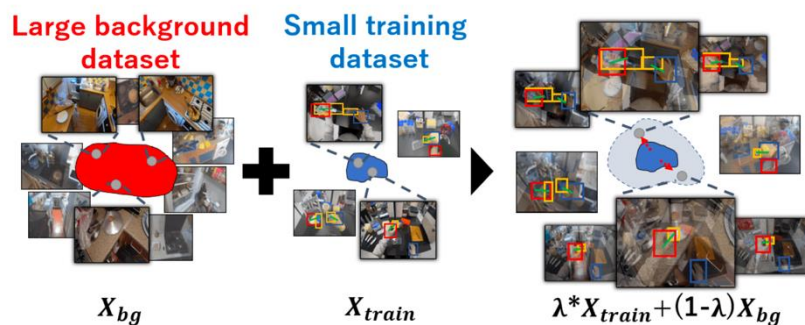


図 3: 手・物体検出における背景画像混合データ拡張

B. ICCV ワークショップとコンペティションの主催（加速フェーズ）

2023 年 10 月に手操作認識に関するワークショップ HANDS をパリで開催された国際会議 ICCV にて主催した[b]。これは前年度成果 A.2-3 の研究を ECCV2022 にて発表した際に、HANDS ワークショップの主催者らと知り合うことができ、ICCV2023 では主催者側として参加するに至った。2 月からシンガポール国立大学、東京大学、チューリッヒ工科大学、バーミンガム大学のチームで開催を計画し、私の担当する一人称視点手姿勢推定のコンペティションも開催した。ワークショップでは、著名なスピーカーを招待して手に関するコンピュータビジョン研究の最前線を学ぶ機会となった。本会議にて Best Paper Finalist となった物体把持動作生成 UniDexGraps++ の講演や、Meta 社のマルチカメラ手動作認識の研究者にも招待講演を依頼し、最先端の知見交換の良い機会となった（図4）。また、本会議からワークショップテーマに関連する13件の研究を招待し、その他 5 件のワークショップ論文を採択してポスター発表も行った。



図 4: ワークショップ当日の様子

一人称視点手姿勢推定のコンペティションは、ワークショップの草案を提出した 2 月から公開の 6 月まで学習データとコード、投稿の評価サーバーの準備を行い、そこから 10 月までコンペティションの運営を行なった。タスクは、一人称視点からの 3 次元手姿勢推定を設定して、私の主著研究である AssemblyHands [a] データセットから学習・評価セットを構成した。シングルビューの学習画像から参加者は推論器を学習して、テスト時にはマルチビューの画像が与えられ、マルチビューからの予測の集約を行う。18 チームが参加し、トップチームは我々のベースライン結果から 41%改善をした。また、トップ3のチームにはテクニカルレポートの提出を依頼して、それらの知見も一般公開するに至った。このワークショップチャレンジの全てのタスクをまとめたテクニカルレポートも公開した [b]。

G. ICCV コンペティションタスクに関連する研究（加速フェーズ）

ICCV2023 では一人称視点からの 3 次元手姿勢推定のタスクを構築してコンペティションを実施した。このタスクに関する調査として既存の手姿勢推定データセット・ベンチマークとその学習方法に関するサーベイ [4] を実施した。

1. 手姿勢推定に関するサーベイ (IJCV2023 採択 [4])

本サーベイでは、効率的なアノテーションと学習の観点から 3 次元手姿勢推定に関する包括的な分析を行なった。3 次元手姿勢推定では、3 次元姿勢アノテーションの収集が精密な認識のために必要であり、映像理解、AR/VR、ロボティクスなどの応用において重要なステップである。しかし、3 次元情報へのアクセスの難しさやオクルージョンのため、よくアノテーションされた 3 次元手姿勢を取得することは困難である。最近の研究がこのアノテーションの問題にどのように対処しているかを明らかにするため、手動、生成モデルベース、手センサーベース、計算手法に分類されるアノテーション方法を調査した。これらのアノテーション方法は常に大規模に利用可能ではないため、十分なアノテーションデータがない場合の学習方法、特に自己教師あり学習、半教師あり学習、ドメイン適応を調査した。効率的なアノテーションと学習の分析に基づき、この分野の限界と

今後の可能性についてさらに議論した。

本研究成果は、コンピュータビジョンに関するトップジャーナルである IJCV にて出版された。

2. 柔軟にマルチビュー画像へ適応した一人称視点姿勢推定 (CVPR2024 採択[5])

成果 C のコンペティションでは、テスト時に与えられるマルチビュー画像間のカメラパラメーターを既知としていたが、このようなマルチビュー推論とその学習は以下の制約がある。1) 訓練のためにマルチビュー画像へのアノテーションが必要であり、このデータ構築は高価である。2) テスト中に訓練時と同じカメラのパラメーター/レイアウトでない場合、学習したモデルを適用できない。そこで、新たにマルチビュー画像へのアノテーションを必要としない、かつ、カメラパラメーター未知の任意の 2 視点画像ペアへ学習器を適応する問題設定を提案する (図5)。提案手法では、各視点の予測からの擬似ラベルを統合する方法と、視点間のカメラ変換が不変である制約から構成される。これにより、信頼できる擬似ラベルを 2 視点画像へ生成して、学習器を任意のマルチカメラへ適応することができる。本研究成果は、コンピュータビジョンに関するトップ国際会議である CVPR にて発表された。

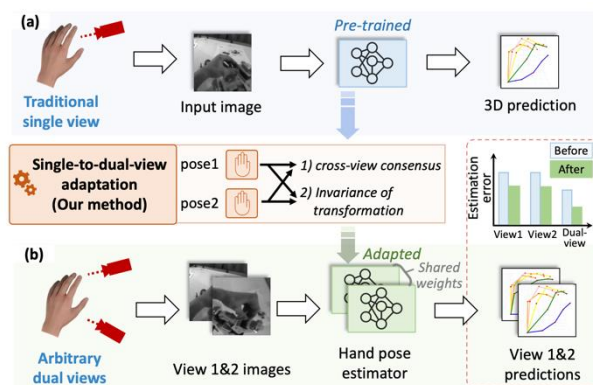
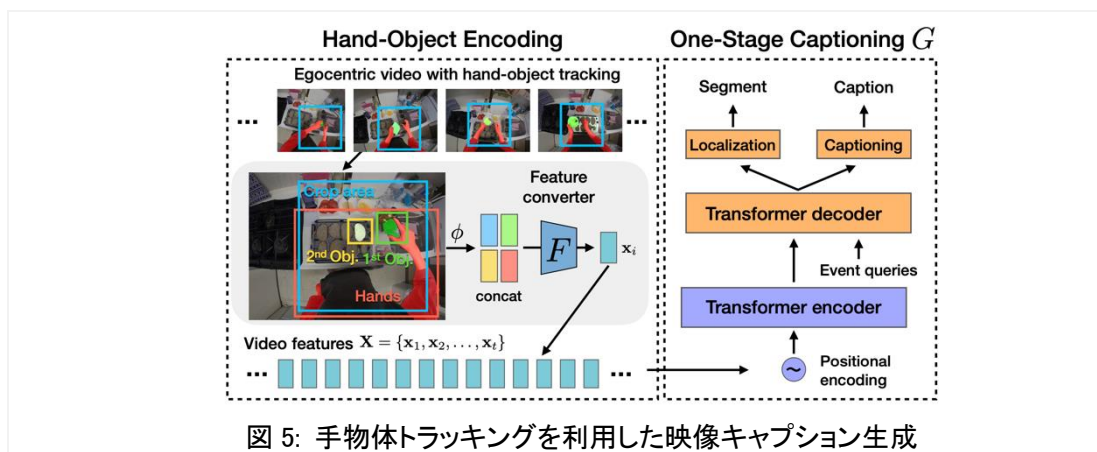


図 5: 単一視点画像モデルの任意の 2 視点画像ペアへの適応

D. 手操作認識の時間拡張と行動理解への接続 (加速フェーズ)

1. 手物体トラッキングから手作業映像キャプション生成 (CVPRW2024 採択[6])

研究成果 A.3 では実映像における手物体検出の改善を検討したが、本研究では詳細な手物体の検出とそのトラッキングを利用して、映像中の行動理解を行うことを目指した (図 5)。調理などの手物体インタラクションを含む一人称視点映像から行動理解を試みる場合には、詳細な手物体状態の認識とそれらの時間方向のトラッキングが重要である。我々は手検出ベースのトラッキングと手物体セグメンテーションのハイブリッド手法を利用して、一人称視点映像の効果的な特徴表現を提案した。さらに、行動理解のタスクとして、従来の行動分類の枠組みではなく、自然言語を用いたキャプション生成を提案した。自然言語を利用することで、行動の離散ラベルの出力と異なって、作業動作に関する言語的な解釈を可能にする。また、このベンチマーク構成のために、キャプション付きの 226 調理映像を 44 人のユーザから集め、データセットも提供した。



3. 今後の展開

A. コンテキスト情報を利用した手操作認識

本 ACT-X 研究で開発した手操作認識モデルは画像フレーム毎に推論するものであったが、特に一人称視点映像ではモーションブラーや手が視野外に消えてしまうなど単一フレームからの推論では、手操作の状態推定が難しいケースが多い。これを解決するための今後の展開として、映像のコンテキスト情報を利用して動作の意味や時間的な整合性を考慮した手操作認識が考えられる。また、これらの補助情報を元にした認識モデルの改善は、新規映像に対する適応においても重要な手がかりとなると考える。

将来的に一人称視点からの手操作認識の実用に向けて、上記のようなモデリングの可能性についてコミュニティを上げて議論する必要がある。そのため、これまでのように手操作認識にのみ注視するのではなく、今後数年間で手操作認識とコンテキストの関連に関する学会のワークショップやチャレンジを構築して、当テーマの探索を加速させる必要がある。

B. コンテキスト情報を利用した手動作軌跡の生成（加速フェーズ実施後追記）

本 ACT-X 研究期間では、認識の枠組み（動画像からある情報を抽出する）内で取り組んできた。しかし、近年では生成 AI に代表されるような動画像生成、自然言語生成に基づく事前学習など顕著な発展を遂げている。これらの進歩から手動作軌跡の生成技術も今後研究が進むと考えられる。例えば、あるシーンに対する人の物体把持の動作軌道を生成することでロボット把持タスクへ転用できる可能性がある。また、ニュースに対して自動で手話を生成する支援技術や、対話型のバーチャルアシスタントが発話内容に対して非言語の自然な手ジェスチャーを行う設計にするなど、様々な応用が考えられる。これらの実用には、条件となるコンテキスト情報（生成する行動内容や文章による動作の説明文）を入力として、動作軌道を生成する研究が必要である。

4. 自己評価

模倣人工知能やロボット構成に関わる重要な人物行動理解のタスクとして、手操作認識の理解に焦点を当て、国内外の著名な研究機関と共同で研究を進めて、一流国際会議やジャーナルへの採択等の業績を上げることができた。研究費については、これらの共同研究を行

う際の旅費・滞在費に主に充てたため、本 ACT-X の支援なしではこれらの共同研究を行うことが出来なかった。本研究の社会実装を進めるために、企業訪問や産学連携により現在の実用状況と今後の展開について議論ができた。さらに、ACT-X 領域会議や国際学会、現地の研究所訪問を通して、隣接するロボット、AR/VR、グラフィクス、自然言語処理等の研究者と交流を深めて、自身の専門とその立ち位置を明確にすることが出来た。

(加速フェーズ実施後追記) 加速フェーズでは、手操作認識の時間発展と行動理解への継続に関する領域へ研究を進めた。この期間にヨーロッパで著名な研究機関であるチューリッヒ工科大にて訪問研究を実施した。さらに、一流国際会議のワークショップ主催やトップジャーナルの採択など業績を上げることができた。また、シンガポール国立大学や国際集会での発表など数多くのアウトリーチ機会を設けることができた。

5. 主な研究成果リスト

(1) 代表的な論文(原著論文)発表

研究期間累積件数: 3件

研究期間累積件数: 3件 (加速フェーズ実施後更新)

[1] Takehiko Ohkawa, Takuma Yagi, Atsushi Hashimoto, Yoshitaka Ushiku, and Yoichi Sato. Foreground-Aware Stylization and Consensus Pseudo-Labeling for Domain Adaptation of First-Person Hand Segmentation. IEEE Access, vol. 9, pp. 94644–94655, 2021.

新規映像における手領域抽出に関する研究. 内容は 2.研究成果(2)詳細 A.1 に記載.

[2] Takehiko Ohkawa, Yu-Jhe Li, Qichen Fu, Ryosuke Furuta, Kris M. Kitani, and Yoichi Sato. Domain Adaptive Hand Keypoint and Pixel Localization in the Wild. In Proceedings of European Conference on Computer Vision (ECCV), 2022 (in press) and Two Invited Posters in European Conference on Computer Vision Workshops (ECCVW), 2022.

新規映像における手姿勢推定と手領域抽出に関する研究. 内容は 2.研究成果(2)詳細 A.2 に記載.

[3] Koya Tango, Takehiko Ohkawa, Ryosuke Furuta, and Yoichi Sato. Background Mixup Data Augmentation for Hand and Object-in-Contact Detection. In European Conference on Computer Vision Workshops (ECCVW), 2022 (in press).

手物体検出における画像混合データ拡張に関する研究. 内容は 2.研究成果(2)詳細 A.3 に記載.

[4] Takehiko Ohkawa, Ryosuke Furuta, and Yoichi Sato. Efficient Annotation and Learning for 3D Hand Pose Estimation: A Survey. International Journal of Computer Vision (IJCV), vol. 131, pp. 3193–3206, 2023.

3次元手姿勢推定に関するサーベイ. 内容は 2.研究成果(2)詳細 C.1 に記載. ※加速フェーズ実施の成果

[5] Ruicong Liu, Takehiko Ohkawa, Mingfang Zhang, Yoichi Sato. Single-to-Dual-View Adaptation for Egocentric 3D Hand Pose Estimation. To appear in IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024.



マルチビュー画像へ適応した 3 次元手姿勢推定. 内容は 2.研究成果(2)詳細 C.2 に記載. ※加速フェーズ実施の成果

[6] Takehiko Ohkawa, Takuma Yagi, Taichi Nishimura, Ryosuke Furuta, Atsushi Hashimoto, Yoshitaka Ushiku, and Yoichi Sato. Dense Video Captioning of Egocentric Human Activities Using Web Instructional Videos. To appear in IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 2024.

一人称視点映像における手作業映像キャプション生成の研究. 内容は 2.研究成果(2)詳細 D.1 に記載. ※加速フェーズ実施の成果

(2)特許出願

研究期間全出願件数:0 件(特許公開前のものも含む)

(3)その他の成果(主要な学会発表、受賞、著作物、プレスリリース等)

- 2021 年 7 月 MIRU 学生奨励賞(研究[1]に関する受賞)
- 2022 年 2 月 UTokyo-IIS Research Collaboration Initiative Award
- 2023 年 7 月 ICVSS ポスター発表 ※加速フェーズ 実施の成果
 - [a] Takehiko Ohkawa, Kun He, Fadime Sener, Tomas Hodan, Luan Tran, and Cem Keskin. AssemblyHands: Towards Egocentric Activity Understanding via 3D Hand Pose Estimation. Poster Presentation at International Computer Vision Summer School (ICVSS), 2023.
- 2023 年 10 月 ICCV HANDS ワークショップ開催 ※加速フェーズ 実施の成果
 - ホームページ: <https://sites.google.com/view/hands2023/home>
 - [b] チャレンジレポート: Zicong Fan*, Takehiko Ohkawa*, Linlin Yang*, et al. Benchmarks and Challenges in Pose Estimation for Egocentric Hand Interactions with Objects. arXiv:2403.16428 (*equal contribution).