

三次元人物姿勢推定のためのテスト時個人適応

代表研究者

温 依林

東京大学生産技術研究所 特任助教

1 はじめに

三次元人物姿勢推定は、スポーツトレーニングにおける動作評価や人間とロボットの協調作業など、幅広い応用分野で人間行動の理解を支える基盤技術である。近年の学習ベースの手法は大きな性能向上を遂げているが[1, 2, 3]、事前学習時のドメインと実環境のギャップによる精度低下は依然として課題である。また、人間の動作は個人の身体形状や習慣に応じた固有の特性を持ち、同一人物が同じ動作を繰り返す場合でも環境の変化や習熟の進展によって自然なばらつきが生じる。しかし、既存研究ではこうした個人特性が十分に考慮されてこなかった。本研究では、フィジカルスキル練習における姿勢の個人性の活用を中心的な課題として設定し、手法開発とデータセット構築の両面から既存研究を補完することを目指した。

手法開発では、テスト時のオンライン個人適応の観点からアプローチを検討した。テストドメインを事前学習には含まれない同一人物のビデオストリームとして定義し、新しいビデオクリップが到着するたびにモデルの重みを更新することで個人特性を継続的に捉えることを目指した。既存のテスト時適応手法(BOA[4]、DynaBOA[5]、CycleAdapt[6])を分析した結果、自己教師あり学習に誤差を含む推定結果を用いるため誤差が蓄積しやすく、長時間の継続適応では精度が低下するという根本的な課題が明らかになった。この問題に対処するため、Motion Discretization と Soft Reset を組み合わせた新たな手法(OTTPA[7])を提案した。Motion Discretization では、事前学習時に教師なしクラスタリングによりアンカーモーションのコードブックを構築し、テスト時の正規化信号として活用することで適応を安定させるとともに、Self-replay により潜在空間の representation drift を防ぐ。Soft Reset では、各クリップへの適応後に姿勢推定器を指数移動平均に差し戻すことで、ノイズの多い更新の影響を抑えつつ個人特性を保持する。以上の手法により、長期にわたる継続個人適応を通じた推定精度の向上を実現した。

データセット構築では、卓球練習に焦点を当てた新たな三次元人物姿勢データセットを構築し、個人化姿勢推定における既存データセットの限界を補完した(図1)。既存の Human3.6M[8]、3DPW[9]、EgoHumans[10]などのデータセットは多様な環境・動作の網羅を重視しており、各参加者が同一動作を行う回数は数回程度にとどまることが多く、個人ごとの動作特性を捉えることが難しい。これに対し、構築したデータセットはマシン練習(制御された球出しによる反復練習)と他の参加者との練習(ラリー・試合)の二部構成からなり、安定した条件下での動作の繰り返しとばらつきの両方を収録することで、個人特性の観察に適した構成となっている。さらに、このデータセットを用いてドメイン汎化手法(SAM3D-Body[11])とテスト時個人適応手法(OTTPA[7])をベースラインとして評価し、個人化三次元人物姿勢推定における知見と今後の課題を示した。

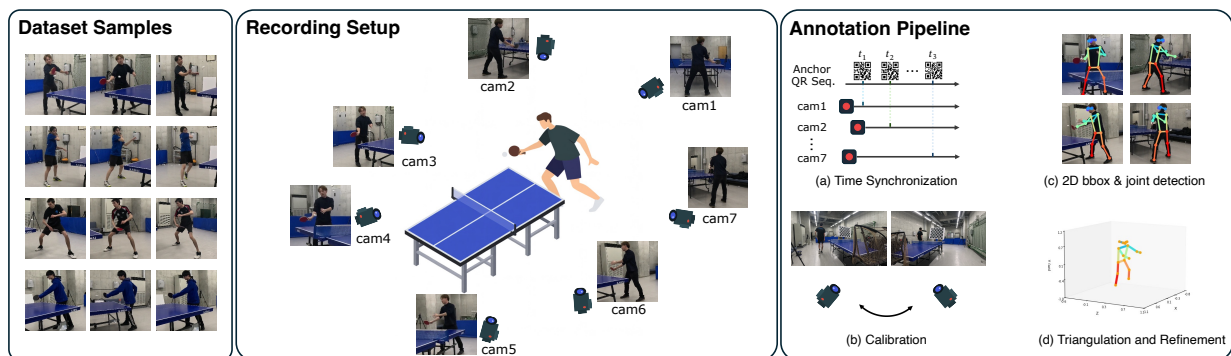


図1：構築されたデータセットのサンプル、収録環境、およびアノテーションの取得パイプライン

2 テスト時個人適応

2-1 問題設定と概要

提案手法（OTTPA）は、オンラインテスト時個人適応という問題に着目する。テストドメインを、事前学習時に未観測の同一人物が映るビデオストリームとして定義し、到着した各ビデオクリップに対してモデルをオンラインで適応させながら三次元関節座標を即時に予測する。

図2に示した通り、提案フレームワークは、姿勢推定器FとモーションデノイズングネットワークMの二つのコンポーネントから構成される。Fは入力画像から SMPL[12]モデルの姿勢・体型・位置パラメータを推定し、三次元関節座標を出力する。Mは姿勢推定器の出力から動きの連続性を捉え、適応を支援する教師信号を生成する。両コンポーネントは交互に適応され、テスト時のアピランス・動作特性をサイクル的に捉える。

既存手法では誤差蓄積と表現ドリフトが長期適応の障害となっていた。この問題に対処するため、本手法は二つの核心的な設計を導入する。第一に、Motion Discretization では事前学習時に教師なしクラスタリングによりアンカーモーションのコードブックを構築し、テスト時の正則化と self-replay に活用することで誤差蓄積と表現ドリフトを抑制する。第二に、Soft Reset では各クリップへの適応後に姿勢推定器を指数移動平均に差し戻すことで、ノイズの多い更新の影響を抑えつつ個人特性を保持する。

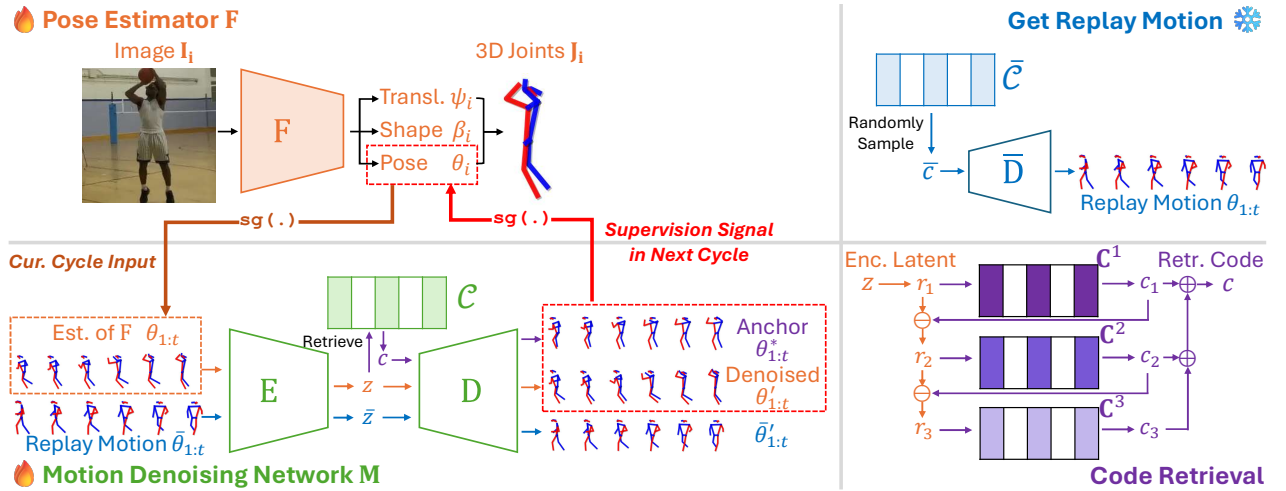


図2：テスト時には、姿勢推定器FとモーションデノイズングネットワークMを交互に更新する。Motion Discretization を用いてFの適応を正則化するとともに、Mの適応における Self-replay を実現する。

2-2 Motion Discretization のためのコードブックとアンカーモーションセット構築

事前学習時に、モーションデノイズングネットワークの潜在空間に対して教師なしクラスタリングを行い、離散的なモーション表現のコードブックCを構築する。コードブックは残差設計を採用した多層構造であり、コードの更新には指数移動平均（EMA）を用いる。

具体的には、図2に示すように、入力モーション系列をエンコーダに通して潜在表現 $\mathbf{z} \in \mathbb{R}^d$ を得る。次に残差 $\mathbf{r}_1 = \mathbf{z}$ を初期化し、各層 l において、現在の残差 $\mathbf{r}_l$ に最も近いコード $\mathbf{c}_l \in \mathbf{C}^l$ を検索して、残差を更新する処理を繰り返す（式1）。選択されたコードは、対応する残差を用いた指数移動平均により逐次更新される。

$$(式1) \quad \mathbf{c}_l = \operatorname{argmin}_{\mathbf{c} \in \mathbf{C}^l} \|\mathbf{c} - \mathbf{r}_l\|_2, \quad \mathbf{r}_{l+1} = \mathbf{r}_l - \mathbf{c}_l$$

このようにして構築したコードブックのコードをデコードすることで、人間の動作空間を広くカバーする多様かつ規則的なアンカーモーションの集合を得る。このアンカーモーション集合は、テスト時適応における正則化信号として機能し、長期適応の安定化に中心的な役割を果たす。

2-3 Motion Discretization による姿勢推定器の適応の正則化

テスト時には、姿勢推定器Fの出力モーション系列 $\theta_{1:t}$ をエンコーダEで潜在表現 $\mathbf{z} = E(\theta_{1:t})$ に変換し、コードブックを用いて離散化した $\mathbf{c} = \sum_{i=1}^k \mathbf{c}_i$ を得る。離散的なコード $\mathbf{c}$ と元の潜在表現 $\mathbf{z}$ をデコーダDにそれぞれ入力し、デノイズドモーション $\theta'_{1:t} = D(\mathbf{z})$ とアンカーモーション $\theta^*_{1:t} = D(\mathbf{c})$ をそれぞれ生成する。次のサイクルでは、アンカーモーションは姿勢推定器Fの更新を監督するアンカー損失として活用される（式2）。

$$(式2) \quad L_{ach} = \|\theta - sg(\theta^*)\|$$

$sg(\cdot)$ はバックプロパゲーションを停止する演算子である。自己教師あり学習で得られるデノイズドモーション θ' は、継続適応中の推定誤差の影響を受けやすい。これに対し、コードブックに基づくアンカーモーション θ^* は高周波ノイズを除去しながら一貫した動作パターンを保持するため、長期適応の安定化に寄与する。

Fの更新に用いる全体的な損失関数は、アンカー損失 L_{ach} に加え、CycleAdapt[6]で導入された二次元再投影誤差、デノイズドモーションとの一致損失、および SMPL 形状の一貫性損失を組み合わせたものである（式3）。

$$(式3) \quad L_F = L_p + \lambda_1 L_s + \lambda_2 L_{2D} + \lambda_3 L_{ach}$$

ここで $L_p = \|\theta - sg(\theta')\|$ はデノイズドモーションとの一致損失、 $L_s = \|\beta - sg(\beta')\|$ は SMPL 形状の一貫性損失、 L_{2D} はFの推定結果の二次元投影と検出器から得た二次元キーポイントとの再投影誤差である。 λ_1 、 λ_2 、 λ_3 はそれぞれの損失項の重みパラメータであり、本研究では $\lambda_1 = 0.001$ 、 $\lambda_2 = 0.1$ 、 $\lambda_3 = 0.3$ に設定した。

2-4 Self-Replay によるモーションデノイズングネットワークの適応

モーションデノイズングネットワークのオンライン適応では、コードブックから一貫したアンカーモーションをデコードする能力を徐々に失う representation drift が生じる恐れがある。この問題に対処するため、コードブックを活用した Self-replay を導入する。

具体的には、事前学習済みのコードブックからランダムにコードをサンプリングし、事前学習済みデコーダでデコードすることでリプレイモーションを生成する。

テスト時の推定結果 $\theta_{1:t}$ とリプレイモーション $\bar{\theta}_{1:t}$ を並行してMに入力し、Mは以下の損失（式4）により自己教師的に更新される。

$$(式4) \quad L_M = \|\bar{\theta}'_{1:t} - sg(\bar{\theta}_{1:t})\| + \|\theta'_{1:t} - sg(\theta_{1:t})\|$$

ここで $\bar{\theta}'_{1:t}$ と $\theta'_{1:t}$ は、リプレイモーションとテスト時のモーションの潜在表現をデコードした出力である。これにより、元の学習データへのアクセスなしに事前学習時の規則性を蒸留し、適応を通じて一貫したアンカーモーション表現を維持する。

2-5 Soft Reset による姿勢推定器の安定した適応

各ビデオクリップへの適応後、姿勢推定器Fに対して Soft Reset を適用する。具体的には、クリップ適応前の重み F_{pre} と適応後の重みFを指数移動平均により合成し、次のクリップの適応に使用するFの重みとして設定する（式5）。

$$(式5) \quad F \leftarrow \mu_F F_{pre} + (1 - \mu_F) F$$

ここで $\mu_F = 0.95$ は減衰係数である。Soft Reset はノイズの多い更新の影響を抑制しつつ、過去の適応で獲得した個人特性（体型・動作習慣）を保持することで、長期にわたる安定した個人適応を実現する。

3 卓球練習を対象とした三次元人物姿勢データセットの構築

3-1 データセット構築の背景と目的

表1に示す通り、既存の三次元人物姿勢データセットは、屋内マーカーベースの収録から屋外マーカーレス設定へと発展を遂げてきた。しかし、これらのデータセットは多様な環境・動作の網羅を重視しており、

各参加者が特定の動作を行う回数は数回程度にとどまることが多い。そのため、個人ごとの動作の類似性とその自然なばらつきを観察することが難しく、個人特性に着目した姿勢推定研究に必要なデータ資源が不足していた。

この課題を補完するため、本研究では卓球練習に着目した新たなデータセットを構築した。卓球のスキル練習では、同一ショット種を繰り返す構造化された練習形式が自然に採用される。また、マシン練習（制御された条件下での反復）と他の参加者との練習（自然に変化する条件下での実戦）の二部構成により、安定した動作の繰り返しと自然なばらつきの両方を網羅的に収録できる。さらに、スポーツ動作は速い動きと頻繁な遮蔽を伴うため、姿勢推定手法の評価に適した実応用シナリオを提供する。各参加者は1人あたり20分以上の練習映像を提供し、各ショット種で数十～数百回の繰り返しが自然に蓄積される。

既存データセットとの比較を表1に示す。本データセットは13名の参加者から1.1Mポーズ・8Mフレームを収録しており、構造化された練習形式を備えた唯一のデータセットである。

データセット	参加者数	ポーズ数	画像数	マーカーレス	構造化練習
Human3.6M [8]	9	900K	3.6M	×	×
MPI-INF-3DHP [13]	8	93K	1.3M	✓	×
3DPW [9]	7	49K	51K	×	×
SportsPose [14]	24	177K	1.5M	✓	×
EgoHumans [10]	20	232K	125K	✓	×
本データセット	13	1.1M	8M	✓	✓

表1：既存データセットとの比較

3-2 データ収集

図1に示した通り、屋内卓球室に7台のGoProカメラ（4K・60fps）を参加者の周囲に配置し、ビデオを収録した。カメラ同期を実現するため、タイムスタンプを0.033秒精度でエンコードしたQRコードシーケンス動画（30fps）を事前に作成し、各撮影セッション開始時にこのQRコードシーケンス動画を各カメラに順番に提示した。各カメラの映像からQRコードを検出してシーケンス動画と照合することで、カメラ間の時刻同期を行った。カメラの内部・外部パラメータのキャリブレーションにはChArUcoボードを用い、2台以上のカメラに同時提示することで各カメラの相対位置・姿勢を取得した。

3-3 アノテーションパイプライン

本研究では、三次元関節座標の付与にCOCO[15]形式を採用し、半自動パイプラインを実装した。まず各カメラからの撮影において、YOLO11[16]とByteTrack[17]を用いて参加者の二次元バウンディングボックスを検出・追跡し、追跡結果を手動で確認した。その後、ViTPose++[18]により各フレームの二次元関節座標を検出した。得られた多視点二次元関節座標に対してRANSACを用いた三角測量を適用し、三次元関節座標の初期値を取得した。さらにPost Refinementを適用することで、三次元関節座標の品質を向上させた。

アノテーション品質の評価として、一部の画像に対してクラウドソーシングによる二次元関節座標の手動アノテーションを実施し、三次元アノテーションの再投影結果と比較した。再投影誤差は13ピクセルであり、4K解像度の1%未満に収まっており、十分な精度が確保されている。

3-4 練習セッションの構成

撮影は二つのセッションから構成される。

マシン練習では、卓球マシン（Pongbot Nova S Pro）を用いて球の速度・回転・コースを制御した反復練習を実施した。フォアハンドとバックハンドに対して、まずドライブ・ループ・プッシュの単発ショット練習を行い、続いてこれらの組み合わせ練習およびランダムドリルへと移行した。単発ショットは1種類あたり3セット、組み合わせドリルは2セットを収録した。各ドリルは緩い速度から開始し、その後速度を上げたセットを収録した。また、参加者の習熟状況に応じてさらに追加セットを収録した。この構成により、安定した条件下で同一ショットの繰り返しを大量に収録し、個人の動作習慣を観察しやすいデータを確保した。

他の参加者との練習では、参加者が上級者とのラリーおよび11点制の試合（2ゲーム）を実施した。ラリーはフォアハンドおよびバックハンドの打ち合いで構成され、ドライブとループの両方をカバーした。この

セッションでは、上級者は参加者の状況に応じて球の速度と回転を適宜調整した。マシン練習とは異なり、球のコース・速度・回転がショットごとに自然に変化するため、より実戦的で多様な動作条件下でのデータが得られる。また、二つのセッションを異なる日に収録した参加者も存在しており、参加者のアピアランスとカメラ配置の変化を含む多様なデータを提供する。

3-5 データセット統計

参加者は男性 9 名・女性 4 名の計 13 名であり、初心者から地域リーグで競技する上級者まで多様な技能レベルを網羅している。このうち 3 名は左利きである。なお、2 名の参加者（P7、P9）はスケジュールの都合により他の参加者との練習セッションに参加していない。

参加者・ドリル別の収録時間の詳細は図 3 に示す。本データセットは多様な条件下でのショットの繰り返しを十分に収録しており、個人化姿勢推定の研究に向けた有用なデータ資源を提供する。既存データセットと比較して、本データセットは同一参加者による長時間の継続練習映像を収録している点が特徴であり、個人ごとの動作の一貫性と自然なばらつきの両方を観察できるリソースとなっている。

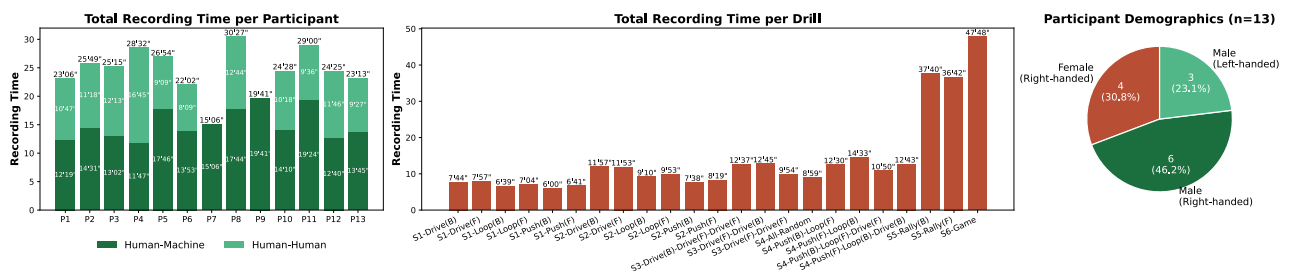


図 3：参加者・ドリル別の収録時間の分析および参加者の統計。

4 既存データセットにおける評価

4-1 データセットと評価指標

提案手法は、Human3. 6M[8]でモデルを事前学習し、Ego-Exo4D[19]と 3DPW[9]でテストおよび評価を行う。Human3. 6M [8] は、制御された屋内環境で収録された大規模データセットである。本手法の事前学習には、既存のテスト時適応手法に倣い、このデータセットを使用した。

Ego-Exo4D [19]は、世界各地でスキル動作を収録したデータセットである。本研究では、多様な身体動作と動作の繰り返しを含み、個人特性の捕捉に適したバスケットボール・サッカー・ダンスの三シナリオを対象とした。魚眼カメラによる収録と多様な動作が引き起こすドメインギャップが、このデータセットへの適応をより困難にする。評価には 30 名の参加者の歪み補正済み外部視点映像を使用した。評価指標として、17 関節の root-aligned の関節位置誤差 (MPJPE) とプロクラステス分析後関節位置誤差 (MPJPE-PA) をミリメートル単位で報告する。

3DPW [9]は、手持ちの移動カメラで屋外収録したデータセットであり、各参加者について多様なアピアランス・背景を含む映像が収録されている。既存のテスト時適応研究に倣い、テストセット (5 名・計 35, 515 フレーム) を評価に使用し、個人化テストドメインとして設定する。推定した関節を MPJPE および MPJPE-PA で評価し、さらに出力メッシュを正解と比較することで頂点位置誤差 (MPVPE) をミリメートル単位で報告する。

4-2 関連手法との比較

まず、公開コードのある既存のオンラインテスト時適応手法との比較を行う。ベースラインは BOA [4]、DynaBOA [5]、CycleAdapt [6]である。これらの手法はすべて Human3. 6M で事前学習した ResNet-50 ベースの姿勢推定器を採用し、同一のビデオストリームに適応させた。二次元検出には OpenPose[20]または ViTPose

[18, 21]を使用し、公平な比較を確保した。また、事前学習済みの重みをそのままテストデータに適用した姿勢推定器もベースラインとして含めた。

Ego-Exo4D および 3DPW における定量的比較を表 2 に示す。これらの結果は、既存のオンラインテスト時適応手法に対する提案手法の優位性を示している。本手法の設計が長期適応における課題を緩和し、個人特性を効果的に活用する頑健な推定を可能にする。また、ViTPose による二次元検出を使用することで Ego-Exo4D における性能が向上する一方、3DPW では限定的な改善にとどまった。これは、3DPW に含まれる参加者の重なりを ViTPose が適切に分離できない場合があるためと考えられる。提案手法の定性的な結果例は図 4 で示す。

さらに、既存手法については、DynaBOA と CycleAdapt は不完全な推定値に基づく自己教師あり学習による誤差蓄積が顕著であり、継続適応を通じて精度が低下する傾向が確認された。また、BOA と DynaBOA は MPJPE および MPVPE において特に大きな誤差を示した。これは、三次元教師信号が不十分なまま誤差を含む二次元教師信号に依存することで、奥行き推定が不正確になるためである。さらに、BOA と DynaBOA はテスト時に事前学習画像を保持する必要があるため、ストレージ使用量と検索処理による実行時間が増大する。

	Ego-Exo4D								3DPW		
	Basketball		Soccer		Dance		All Scenarios				
	MPJPE	MPJPE-PA	MPJPE	MPJPE-PA	MPJPE	MPJPE-PA	MPJPE	MPJPE-PA	MPJPE	MPJPE-PA	MPVPE
Pre-trained F	215.8	122.7	198.6	114.7	198.1	111.0	205.8	116.5	230.3	123.4	253.4
w/ OpenPose (Cao et al. 2017)											
BOA (Guan et al. 2021) [†]	158.4	72.8	115.3	63.3	117.8	69.3	135.1	70.0	98.2	55.8	114.2
DynaBOA (Guan et al. 2022) [†]	173.0	73.4	129.1	68.2	141.0	70.9	153.3	71.6	139.7	63.8	155.1
CycleAdapt (Nam et al. 2023)	160.4	80.8	126.3	78.5	135.3	81.0	145.0	80.5	141.0	79.6	155.6
Ours	141.6	71.3	105.3	61.3	106.4	67.0	121.5	68.1	83.9	51.6	100.3
w/ ViTPose (Xu et al. 2022b,a)											
BOA (Guan et al. 2021) [†]	149.6	65.3	108.9	58.2	101.9	61.9	123.5	62.9	91.5	53.9	105.5
DynaBOA (Guan et al. 2022) [†]	154.8	64.3	114.8	58.6	108.5	61.3	129.4	62.2	118.4	56.1	134.5
CycleAdapt (Nam et al. 2023)	144.8	69.6	112.1	62.9	108.3	67.1	124.6	67.6	109.6	66.6	124.8
Ours	137.4	63.4	102.2	56.8	99.8	59.5	116.4	60.8	85.0	53.3	100.4

表 2 : Ego-Exo4D および 3DPW において、事前学習済み姿勢推定器および既存のオンラインテスト時適応手法との比較。†は適応時に元の事前学習データを使用することを示す。

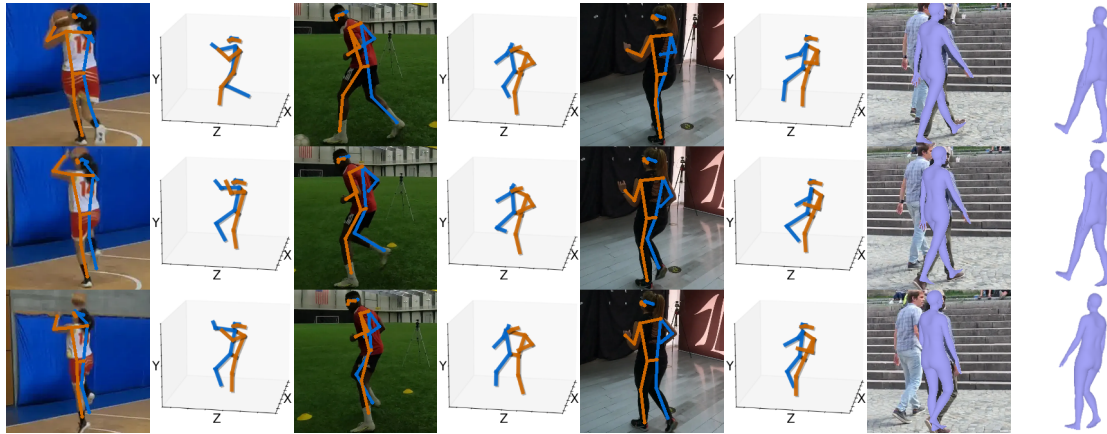


図 4 : Ego-Exo4D および 3DPW における提案手法の結果例。

また、表 3 に示す通り、事前学習済みモデルをテストドメインに直接適用するドメイン汎化手法との評価も実施した。これらの手法は Human3.6M を超える大規模データセットで事前学習されており、HMR-2.0b[3]はさらに強力な ViT バックボーンを採用している。ResNet-50 バックボーンを使用する既存手法との比較では、提案手法は ViTPose 使用時により優れた性能を示した。さらに、OpenPose 使用時には、3DPW においてこれらのベースラインを上回り、Ego-Exo4D でも MPJPE で競争力のある結果を達成した。HMR-2.0b との比較では、ViTPose 使用時に Ego-Exo4D で優れた精度を示し、3DPW においても同等の精度を得た。これらの比較は、事前学習ドメインと大きく異なるテストドメインへの適用が困難である一方、提案手法がこの学習・テストドメイン間のギャップを効果的に埋めることを示している。

	Ego-Exo4D		3DPW		
	MPJPE	MPJPE-PA	MPJPE	MPJPE-PA	MPVPE
Pre-trained F	205.8	116.5	230.3	123.4	253.4
SPIN (Kolotouros et al. 2019)	137.3	71.5	96.9	59.2	116.4
I2L-Mesh (Moon and Lee 2020)	136.6	73.8	93.2	57.7	110.1
PyMAF (Zhang et al. 2021)	135.5	70.2	92.8	58.9	110.1
HMR-2.0b (Goel et al. 2023) †	125.2	65.6	81.3	54.3	94.4
VIBE (Kocabas, Athanasiou, and Black 2020)	130.6	70.1	93.5	56.5	113.4
TCMR (Choi et al. 2021)	126.4	66.9	95.0	55.8	111.3
GLoT (Shen et al. 2023)	127.1	66.7	89.9	53.5	107.8
Ours (w/ OpenPose)	121.5	68.1	83.9	51.6	100.3
Ours (w/ ViTPose)	116.4	60.8	85.0	53.3	100.4

表 3：より大規模な学習データを活用するドメイン汎化手法との比較。比較対象の手法はいずれも Ego-Exo4D および 3DPW を事前学習に使用していない。†は ViT-H/16 バックボーンを使用することを示す。

4-3 アブレーション考察

まず、Motion Discretization の効果を考察する。その長期適応における効果を図 6 および表 4 に示す。Self-replay とアンカー損失の両方を活用することで相乗効果が生まれ、最良の性能が得られる。これは、Self-replay が representation drift を緩和することで、適応を通じて規則的かつ現実的なアンカーモーションの生成を維持するためである（図 5）。これらのアンカーモーションが信頼性の高い参照信号として機能し、不完全な推定値に基づく自己教師あり学習の悪影響を緩和する。一方、アンカー損失のみを使用した場合は、ノイズの多いアンカーモーションにより誤差が増大する。

w/ Soft-Reset on F	w/ Motion Discretization		MPJPE	MPJPE-PA
	w/ $L_{ach}(\theta^*)$	w/ Self-Replay		
✓	✓	✓	144.3	80.6
			122.9	69.2
			138.2	74.8
			121.5	68.1
✓	✓	✓	122.9	69.2
			123.2	69.2
			122.9	69.2
			121.5	68.1

表 4：Ego-Exo4D におけるアブレーション考察。さらに、図 5 は Motion Discretization が長期適応に寄与することを示す。

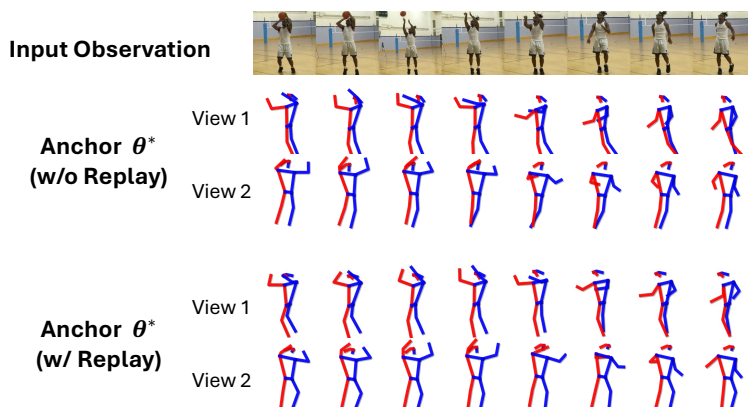


図 5：バスケットボールにおいて、40 分の観測・適応後に取得されたアンカーモーション。提案する Self-replay により、適応を通じて現実的かつ規則的なアンカーモーションのデコードを維持できる。

また、表5の左段で示す通り、姿勢推定器の適応における三次元モーション監督信号を考察する。コードブックから得られるアンカーモーション θ^* に加え、潜在表現をデコードして得るデノイズドモーション θ' も有効であることが確認された。アンカーモーションは長期的な自己教師あり学習のもとで規則的なパターンをより安定して保持する一方、デノイズドモーションは離散化によって失われる fine-grained motion を捉える。両者の相補的な役割により、頑健かつ高精度な推定が実現する。

w/ L_p (θ')	w/ L_{ach} (θ^*)	MPJPE	MPJPE-PA	Soft-Reset Decay μ_F	MPJPE	MPJPE-PA	w/ Cont. Adapt.	MPJPE	MPJPE-PA
✓		122.9	69.2	0	138.2	74.8			
	✓	123.1	69.8	0.9	122.7	69.2		125.3	70.2
✓	✓	121.5	68.1	0.95	121.5	68.1	✓	121.5	68.1
				1.00	125.3	70.2			

表5：Ego-Exo4Dにおける三次元モーション教師信号・Soft Reset 減衰係数・継続適応の分析。

次に、Soft Reset を考察する。表4上部と表5中段に示す通り、Soft Reset は、過去の適応を通じて蓄積した個人特性を活用しながら、個々のバッチにおける適応から生じた誤差の影響を抑制することで性能を向上させる。

最後に、継続適応を考察する。図6および表5右部に示す通り、継続適応を通じた個人特性の活用により、平均的な性能向上が確認された。ここでのベースラインは、各ビデオクリップの適応前にモデルを事前学習済み重みにリセットする設定である。特に、継続適応による性能向上は、サッカーおよびダンスのシナリオで顕著である。一方、バスケットボールシナリオでは、カメラ配置の影響で一部参加者の映像が大きく歪んでいるため、ドメインギャップが拡大してバッチごとの誤差が増大し、継続適応の効果が限定的となった。

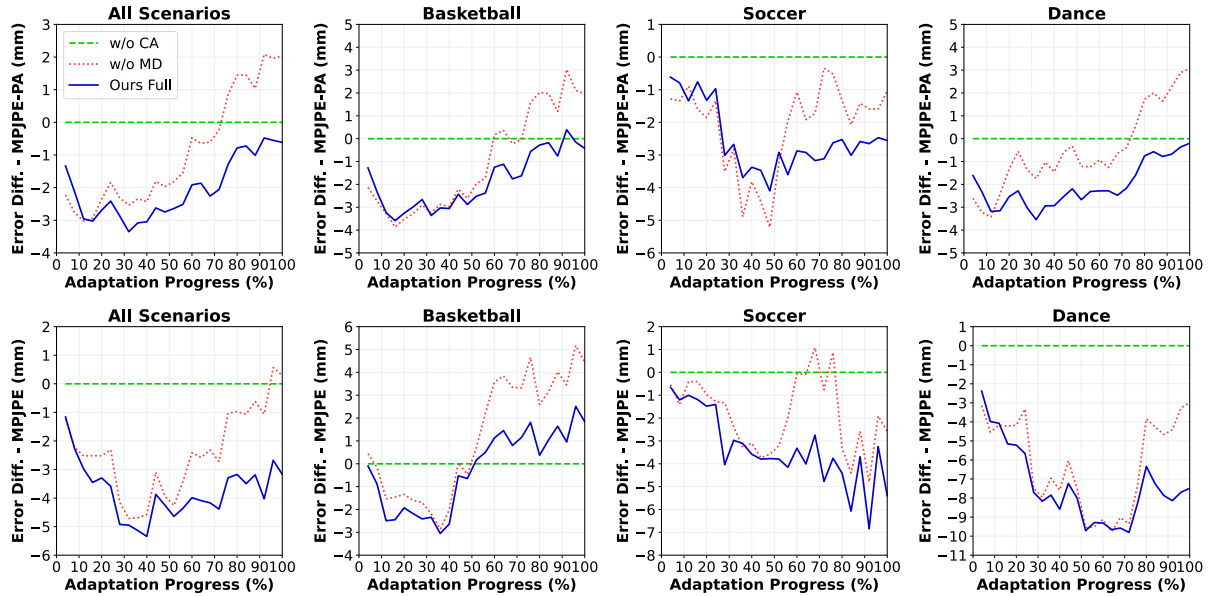


図6：各参加者の適応過程における誤差差分の推移。縦軸は、各ビデオクリップの適応を常に事前学習済み重みから開始する「継続適応なし」のベースライン（破線）に対する誤差差分を示す。提案手法（実線）を、アンカー損失と Self-replay の両方を外した「Motion Discretizationなし」の変種（点線）と比較する。縦軸の値が低いほど性能が優れていることを示す。

5 構築されたデータセットにおける評価とベースライン設定

5-1 評価設定

右利きの参加者には図 1 に示した cam3、左利きの参加者には cam6 の映像を使用して全収録データで評価を行った。これらのカメラ視点からの撮影は遮蔽が最も少ないためである。評価指標として、root-aligned な関節位置誤差 (MPJPE) およびプロクラステス分析後関節位置誤差 (MPJPE-PA) をミリメートル単位で報告する。

5-2 ベースライン設定

SAM3D-Body[11]は多様な実環境条件における高い汎化性能を示す最先端の人物姿勢推定モデルである。自動データエンジンにより収集された豊やかで多様なデータで学習された大規模モデルを採用しており、汎化力に優れる。本研究では、ゼロショット設定で本データセットに適用し、卓球練習における高速な動作と頻繁な自己遮蔽という課題のもとでの汎化性能を検証した。

本研究で提案された OTTPA[7]は、同一人物のラベルなし映像ストリームに対して自己教師あり学習によりモデルを継続的に更新するオンラインテスト時個人適応手法である。Motion Discretization と Soft Reset により適応を正則化し、長期適応における誤差蓄積を緩和する。各参加者のマシン練習セッションと他の参加者との練習セッションの映像を順次連結してテストビデオストリームを構成した。また、事前学習済み重みをそのまま適用した設定 (OTTPA-Pretrained) と、継続適応を行わず各ビデオクリップの開始時に事前学習済み重みにリセットする設定 (OTTPA-Reset) も合わせて評価する。これらのベースラインにより、誤差蓄積への対処と個人の動作パターン捕捉に関する OTTPA の能力を多角的に検証する。

Unit: mm	SAM3D-Body [11]	OTTPA-Pretrained [7]	OTTPA-Reset [7]	OTTPA [7]
MPJPE	56.7	179.3	94.1	92.3
MPJPE-PA	30.3	108.5	58.7	57.8

表 6：構築されたデータセットにおける定量的な結果

5-3 SAM3D-Body の考察

表 6 に示した通り、SAM3D-Body は大規模データによる強力な事前学習により、全体的に最良の性能を示した。これは特に、Human3.6M のみで事前学習した ResNet-50 バックボーンを使用する OTTPA-Pretrained との比較において顕著である。しかし、SAM3D-Body は遮蔽時および奥行き推定において課題が残る。SAM3D-Body は単一画像のみを処理するため、フレーム間の時間的一貫性を活用できない。その結果、スポーツシナリオにおける頻繁な遮蔽下での奥行き推定が困難となり、個人の体型一貫性や動作パターンを捉えることができない (図 7)。これは、特定の参加者の体型・姿勢パターンへの適応ではなく、多様な個人への汎化を目的とするドメイン汎化手法の本質的な限界である。

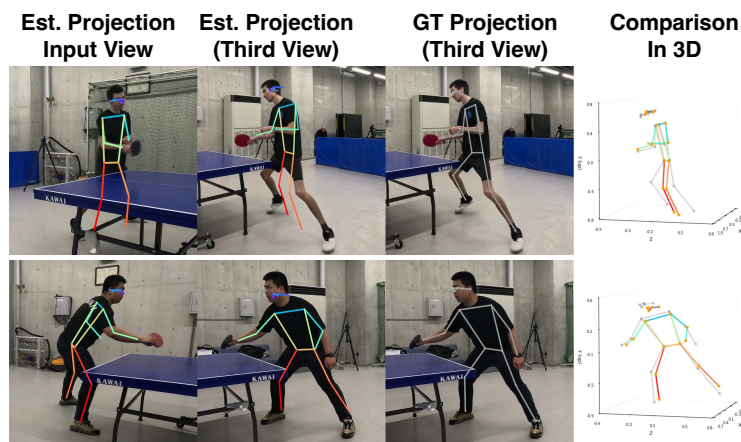


図 7： SAM3D-Body の定性的結果例。SAM3D-Body は遮蔽されない関節の推定において高い精度を示す一方、遮蔽された関節の推定と奥行き推定に課題が残る。

5-4 OTTPA の考察

次に、本研究で提案された OTTPA を詳細に検証する。OTTPA-Pretrained との比較から、テスト時適応による明確な性能向上が確認された。これは、事前学習ドメインが限られている場合に、自己教師ありモデル適応による個人特性の捕捉が有効であることを示している。事前学習で使った Human3.6M は、本データセットの卓球映像とは環境・参加者・動作・マーカーベースのアピアランスのいずれの面でも大きく異なる。しかし、SAM3D-Body が提案手法より優れた性能を示すことは、より強力な事前学習モデルとテスト時適応の組み合わせが今後の重要な研究方向であることを示唆している。

さらに、継続適応の有効性を検証するため、OTTPA-Reset との比較をセッション別に行った。表 7 に示す通り、マシン練習セッション内では継続適応の明確な効果が見られ、S1～S4 全体を通じた性能向上が確認された。しかし、他の参加者との練習セッションに移行すると効果が薄れ、試合セッションでは継続適応による性能が OTTPA-Reset を下回った。

	Human-Machine Session								Human-Human Session			
	S1		S2		S3		S4		S5		S6	
	MPJPE	MPJPE-PA	MPJPE	MPJPE-PA	MPJPE	MPJPE-PA	MPJPE	MPJPE-PA	MPJPE	MPJPE-PA	MPJPE	MPJPE-PA
OTTPA-Reset	89.9	54.5	97.6	56.0	100.9	61.6	94.4	58.9	88.4	57.8	96.9	59.5
OTTPA	86.6	54.4	92.9	55.7	95.0	60.7	88.1	56.4	88.7	55.1	101.1	62.6

表 7：OTTPA における継続適応のセッション別評価。適応は S1 から S6 へと順次進行し、S1～S4 がマシン練習、S5～S6 が他の参加者との練習（ラリーおよび試合）である。

これは以下の要因によるものと考えられる。第一に、他の参加者との練習はマシン練習とは異なる日に収録されている場合が多く、参加者のアピアランスやカメラ配置に差異が生じる。第二に、他の参加者との練習では球のコース・速度・回転がショットごとに自然に変化するため、動作パターンがマシン練習と比べて多様かつ予測困難になる。一貫したアピアランス・練習条件下では継続適応が個人特性の活用にも有効である一方、こうした新たな環境では誤差蓄積と破滅的忘却が適応済みモデルに悪影響を及ぼす。これらの結果は、多様かつ変化する観測条件にわたる頑健な長期適応が未解決の課題であることを示している。

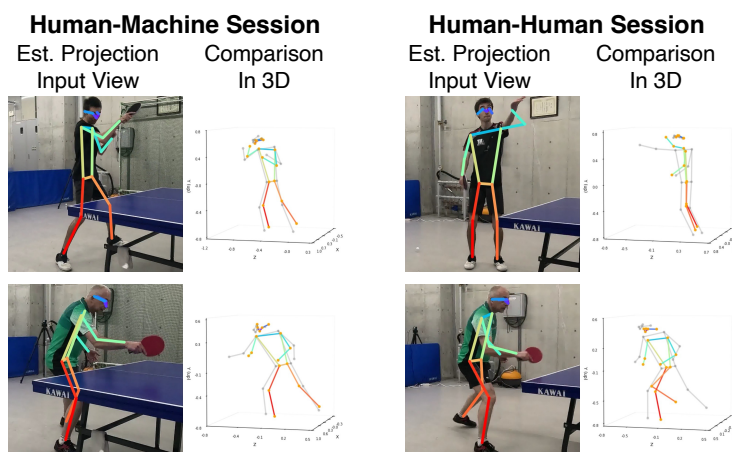


図 8：提案された OTTPA の定性的結果例。マシン練習セッション内では継続適応が個人特性を効果的に捉える一方、マシン練習から他の参加者との練習へのドメインシフトに伴い、破滅的忘却と誤差蓄積という課題がある。

参加者 13 名それぞれについて継続適応の有効性を検証した結果、継続適応は 13 名中 9 名で MPJPE において OTTPA-Reset を上回った（表 8）。継続適応が有効でない典型的なケースとして二種類が確認された。一方は、P6・P7 のような参加者で、単一クリップの適応でも誤差が大きく、継続適応によりさらに誤差が累積するケースである。他方は、左利き選手において継続適応の効果が限定的となるケースであり、事前学習データに左利き選手のサンプルが少ないことが一因と考えられる。これらの知見は、より強力な事前学習モデルを用いた適応と、頑健な長期適応のための誤差蓄積・破滅的忘却への対処という二つの重要な今後の方向性を示している。

	P1		P2		P3		P4		P5		P6		P7	
	MPJPE	PA	MPJPE	PA	MPJPE	PA	MPJPE	PA	MPJPE	PA	MPJPE	PA	MPJPE	PA
OTTPA-Reset	80.4	56.4	103.5	67.8	91.8	62.7	90.9	67.9	93.3	57.1	121.4	84.4	115.3	66.3
OTTPA	79.5	55.7	94.4	59.1	100.3	73.6	88.1	57.7	96.8	61.3	122.8	68.6	121.7	71.6
	P8		P9		P10		P11		P12		P13			
	MPJPE	PA	MPJPE	PA	MPJPE	PA	MPJPE	PA	MPJPE	PA	MPJPE	PA		
OTTPA-Reset	99.1	56.6	96.0	45.3	91.8	55.7	72.5	41.6	92.8	56.6	87.0	48.5		
OTTPA	94.4	63.9	80.7	45.1	90.6	54.6	72.3	45.7	89.7	52.8	82.1	44.7		

表 8： OTTPA における継続適応の参加者別評価。P2・P11・P12・P13 は女性参加者、P5・P7・P10 は左利き選手である。

6 まとめと今後の展望

本研究では、三次元人物姿勢推定における個人特性の活用を中心的な課題として、手法開発とデータセット構築の両面から取り組んだ。

手法開発では、継続適応における誤差蓄積という根本的な課題に対し、Motion Discretization と Soft Reset を組み合わせたテスト時個人適応手法 (OTTPA) を提案した。Ego-Exo4D および 3DPW における評価実験により、既存のオンラインテスト時適応手法を上回る性能を達成した。データセット構築では、卓球練習を対象とした三次元人物姿勢データセットを構築した。構造化された練習形式により、13 名の参加者の個人ごとの動作の繰り返しとばらつきを網羅的に収録した。SAM3D-Body と提案された OTTPA をベースラインとして評価した結果、ドメイン汎化手法は遮蔽・奥行き推定に限界があり、テスト時適応手法は観測条件が変化する場合には誤差蓄積・破滅的忘却が課題として残ることが明らかになった。

今後の課題として、以下の二点が挙げられる。第一に、ViT ベースのより強力な事前学習モデルとテスト時適応の統合による性能向上である。第二に、アピアランス・動作条件が変化する長期シナリオでの頑健な適応の実現であり、誤差蓄積と破滅的忘却への対処が鍵となる。本研究で構築したデータセットは、これらの課題に取り組む今後の研究における有用な評価基盤となることが期待される。

【参考文献】

- [1] N. Kolotouros, G. Pavlakos, M. J. Black, K. Daniilidis, "Learning to Reconstruct 3D Human Pose and Shape via Model-fitting in the Loop," ICCV, 2019.
- [2] M. Kocabas, N. Athanasiou, M. J. Black, "VIBE: Video Inference for Human Body Pose and Shape Estimation," CVPR, 2020.
- [3] S. Goel, G. Pavlakos, J. Rajasegaran, A. Kanazawa, J. Malik, "Humans in 4D: Reconstructing and Tracking Humans with Transformers," ICCV, 2023.
- [4] S. Guan, J. Xu, Y. Wang, B. Ni, X. Yang, "Bilevel Online Adaptation for Out-of-Domain Human Mesh Reconstruction," CVPR, 2021.
- [5] S. Guan, J. Xu, M. Z. He, Y. Wang, B. Ni, X. Yang, "Out-of-Domain Human Mesh Reconstruction via Dynamic Bilevel Online Adaptation," IEEE TPAMI, 2022.
- [6] H. Nam, D. S. Jung, Y. Oh, K. M. Lee, "Cyclic Test-Time Adaptation on Monocular Video for 3D Human Mesh Reconstruction," ICCV, 2023.
- [7] Y. Wen, K. Dong, Y. Sugano, "Robust Long-term Test-Time Adaptation for 3D Human Pose Estimation through Motion Discretization," AAAI, 2026.
- [8] C. Ionescu, D. Papava, V. Olaru, C. Sminchisescu, "Human3.6M: Large Scale Datasets and Predictive Methods for 3D Human Sensing in Natural Environments," IEEE TPAMI, 2014.
- [9] T. von Marcard, R. Henschel, M. J. Black, B. Rosenhahn, G. Pons-Moll, "Recovering Accurate 3D Human Pose in the Wild using IMUs and a Moving Camera," ECCV, 2018.
- [10] R. Khrodar, A. Bansal, L. Ma, R. Newcombe, M. Vo, K. Kitani, "Ego-Humans: An Ego-Centric 3D Multi-Human Benchmark," ICCV, 2023.

- [11] X. Yang, D. Kukreja, D. Pinkus, A. Sagar, T. Fan, J. Park, S. Shin, J. Cao, J. Liu, N. Ugrinovic, M. Feiszli, J. Malik, P. Dollar, K. Kitani, "SAM 3D Body: Robust Full-Body Human Mesh Recovery," CVPR, 2026.
- [12] M. Loper, N. Mahmood, J. Romero, G. Pons-Moll, M. J. Black, "SMPL: A Skinned Multi-Person Linear Model," ACM Transactions on Graphics, vol. 34, no. 6, 2015.
- [13] D. Mehta, H. Rhodin, D. Casas, P. Fua, O. Sotnychenko, W. Xu, C. Theobalt, "Monocular 3D Human Pose Estimation in the Wild using Improved CNN Supervision," 3DV, 2017.
- [14] C. K. Ingwersen, C. M. Mikkelsen, J. N. Jensen, M. R. Hannemose, A. B. Dahl, "SportsPose - A Dynamic 3D Sports Pose Dataset," CVPR, 2023.
- [15] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, C. L. Zitnick, "Microsoft COCO: Common Objects in Context," ECCV, 2014.
- [16] G. Jocher, J. Qiu, "Ultralytics YOLO11," version 11.0.0, 2024.
- [17] Y. Zhang, P. Sun, Y. Jiang, D. Yu, F. Weng, Z. Yuan, P. Luo, W. Liu, X. Wang, "ByteTrack: Multi-Object Tracking by Associating Every Detection Box," ECCV, 2022.
- [18] Y. Xu, J. Zhang, Q. Zhang, D. Tao, "ViTPose++: Vision Transformer Foundation Model for Generic Body Pose Estimation," IEEE TPAMI, vol. 46, pp. 1212–1230, 2024.
- [19] K. Grauman, A. Westbury, L. Torresani, K. Kitani, J. Malik, T. Afouras, K. Ashutosh, V. Baiyya, S. Bansal, B. Boote, et al., "Ego-Exo4D: Understanding Skilled Human Activity from First- and Third-Person Perspectives," CVPR, 2024.
- [20] Z. Cao, T. Simon, S.-E. Wei, Y. Sheikh, "Realtime Multi-Person 2D Pose Estimation Using Part Affinity Fields," CVPR, 2017.
- [21] Y. Xu, J. Zhang, Q. Zhang, D. Tao, "ViTPose: Simple Vision Transformer Baselines for Human Pose Estimation," NeurIPS, 2022.

〈 発 表 資 料 〉

題 名	掲載誌・学会名等	発表年月
Robust Long-term Test-Time Adaptation for 3D Human Pose Estimation through Motion Discretization	Proceedings of the AAAI Conference on Artificial Intelligence	2026 年 3 月
Prosthesis-Aware 3D Human Pose Estimation: A Dataset and Benchmark for RSP Users	European Conference on Computer Vision	2026 年 9 月
Continuous Self-Supervised Adaptation for Personalized Human Pose Estimation	第 28 回 画像の認識・理解シンポジウム (MIRU 2025) 口頭発表	2025 年 7 月
A Preliminary Study on 3D Pose Estimation for Blade Prosthesis Users	第 28 回 画像の認識・理解シンポジウム (MIRU 2025)	2025 年 7 月
Evaluating 3D Body Pose Estimation in Continuous Sports Practice	第 29 回 画像の認識・理解シンポジウム (MIRU 2026)	2026 年 8 月
3D Human Pose Estimation for Running-Specific Prosthesis Users	第 29 回 画像の認識・理解シンポジウム (MIRU 2026)	2026 年 8 月