CONFIDENTIAL EXTENDED ABSTRACT.
DO NOT DISTRIBUTE ANYWHERE.

モバイル端末向け 3 次元人体姿勢推定における時空間モデリングの活用

中溝 雄斗[1,a)]　木村 倫太郎[1,b)]　柳井 啓司[1,c)]

概要

既存のモバイル端末向け 3 次元人体姿勢推定手法は，処理時間の制約のため時空間モデリングの活用が限定的であり，推定精度や時間的一貫性に課題が残る。そこで本研究では，新たな軽量な単一フレーム推定と効率的な因果的時空間リファインメントからなる二段階構成を提案する。実験により，提案手法がモバイル GPU 上でリアルタイム動作を維持しつつ，ベースラインを上回る精度と時間的一貫性を示した。

1. はじめに

3 次元人体姿勢推定は，単眼画像または動画から人物の 3 次元関節座標を推定する技術であり，VR/AR，スポーツ解析，監視，リハビリテーションなど幅広い応用を支える基盤技術として注目されている [2], [3], [13]。近年では，Transformer [18]，State Space Model (SSM) [7], [8]，およびグラフ構造に基づく時空間統合により，3 次元人体姿勢推定の精度は大きく向上している [9], [15], [20]。特に，複数フレームを入力として高度な時空間モデリングを施す手法は，高い構造整合性と時間的一貫性を実現しており，時空間モデリングの有効性を示している。

一方で，これらの高精度手法の多くは高性能 GPU 上での固定長クリップ処理を前提としており，モバイル端末上でのストリーミング推論にはそのまま適用しにくい。

対称的に ConvNeXtPose [13] などの既存のモバイル端末向け軽量手法はシンプルな単一フレーム推定に依存しており，低レイテンシ動作を実現する一方で，モデル内部で関節間関係や時間的文脈を十分に統合できない。そのため，構造的な妥当性や時間的一貫性の観点では依然として改善の余地が残されている。

そこで本研究では，モバイル端末におけるリアルタイムストリーミング推論を前提として，3 次元人体姿勢推定の設計を再考する。特に厳しいレイテンシ制約の下では，単一フレームの空間バックボーンを単純に大型化するよりも，軽量な因果的時空間モデリングに計算資源を配分する方が有効であるという考えに基づき，軽量なフレーム単位空間推定器（Stage 1）により関節座標を推定し，続いて過去情報を状態として保持する因果的時空間リファインメント（Stage 2）によって逐次的に補正することで，将来フレームに依存しない低遅延なストリーミング推論を実現する。

実験の結果，提案手法は既存のモバイル端末向け手法と比較して低い計算量で高い精度を示し，単一フレームベースラインに対して時間的一貫性の改善も確認した。これにより，本手法は高精度な時系列処理とモバイル端末上での安定したリアルタイム動作の両立を可能にし，実環境で利用可能なオンデバイス 3 次元人体姿勢推定への有効な設計指針を示す。

2. 関連研究

2.1 3 次元人体姿勢推定

既存の 3 次元人体姿勢推定手法は，大きく直接回帰型の一段階手法と，2 次元関節表現を経由する二段階手法に分類される。一段階手法は，画像または動画から 3 次元関節座標を直接推定するものであり，Soft-argmax 系の手法 [17] などが代表例である。この系統は，中間表現を明示的に設計することなくエンドツーエンドに最適化できるため，画像中の意味的・幾何的情報を保持したまま 3 次元姿勢を推定できる利点を持つ。一方，二段階手法では，まず画像から 2 次元関節座標を推定し，その後その系列を 3 次元空間へリフティングする。VideoPose3D [14] 以降，多くの高性能手法がこの枠組みに属する。これらは低次元な関節表現に基づくため時系列モデリングと親和性が高く，特に Transformer ベースの手法は時空間文脈を効果的に統合することで高精度を達成している。しかしその一方で，入力を 2 次元関節列に圧縮する過程で画像空間の意味的手掛かりが失われ，これが性能上の制約となることも指摘されている [16]。

また，一段階手法における回帰戦略としては，ヒートマップベース回帰と直接回帰が広く用いられる。ヒートマップ

[1] 電気通信大学
[a)] nakamizo-y@mm.inf.uec.ac.jp
[b)] kimura-r@mm.inf.uec.ac.jp
[c)] yanai@cs.uec.ac.jp

手法は空間的な事前知識を保持しやすい反面，高解像度特徴マップやそれに対する大規模な Softmax 演算を必要とするため，計算量とメモリ帯域の負荷が大きい。これに対し直接回帰は効率的であるが，局所的な空間情報を失いやすく，精度面で課題を持つ。SAR [19] はこの両者の中間的な設計として，局所特徴に対するオフセット回帰を導入し，効率性と空間的事前知識の活用の両立を図っている。

さらに近年では，時空間統合のための構造として Transformer に加え，State Space Model (SSM) [7], [8] も注目されている。SSM は系列長に対して線形計算量で長距離依存を扱えるうえ，逐次更新に適した構造を持つ。特に Mamba [8] に代表される入力依存型の状態更新は，3 次元人体姿勢推定にも応用されつつある [5], [6], [9]。ただし，既存手法の多くはオフライン処理や双方向文脈を前提としており，モバイル環境下での因果的ストリーミング推論を主目的とした設計は十分ではない。

2.2 モバイル端末向け 3 次元人体姿勢推定

モバイル端末上でのリアルタイム 3 次元人体姿勢推定を対象とした研究としては，MobileHumanPose [4]，ConvNeXtPose [13] などが挙げられる。MobileHumanPose は MobileNet 系の軽量バックボーンを採用することで，実機上での低レイテンシ推論を実現した。ConvNeXtPose もまた，軽量な CNN バックボーンと簡潔な回帰ヘッドにより，AR フィットネスなどの応用を意識した実用的な速度とより優れた精度を示している。

これらの手法はいずれも，単一フレームから直接 3 次元姿勢を推定するシンプルな一段階構造を基本としており，リアルタイム動作を達成している点で重要である。一方で，主な焦点は空間特徴抽出と単一画像推定の効率化にあり，逐次入力されるフレーム列をモデル内部で時空間的に統合する仕組みは限定的である。加えて，特に近年のモバイル端末向けの手法 [4], [13] では精度面の恩恵から integral regression [17] が採用されることが多いが，これは大規模な Softmax 正規化とそれに伴う遅延を招き，単一フレーム処理だけで 30 fps 弱の処理速度となり，追加の時空間モデリングを組み込む余地が小さい。

したがって，モバイル端末上の厳しい計算制約の下で，リアルタイム性を維持しつつストリーミング入力に対する時空間モデリングを行う枠組みは，依然として十分に確立されていない。本研究ではこのギャップに着目し，単一画像推定の計算コストを抑えてレイテンシ予算を確保し，その余裕を軽量な時空間モデリングへ再配分する統合的なアーキテクチャを提案する。これにより，モバイル端末上での高精度かつ低遅延な 3 次元人体姿勢推定を目指す。

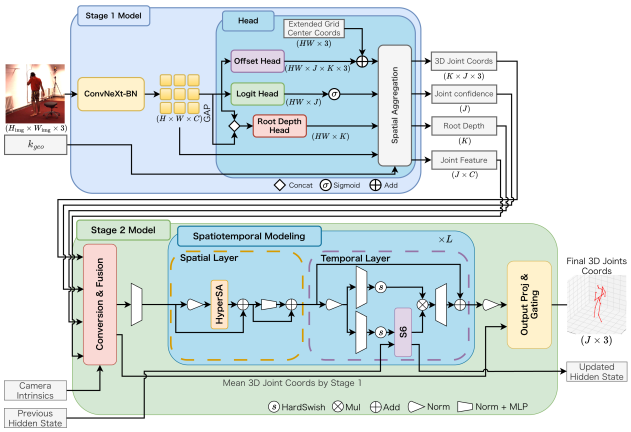


図 1 提案手法の概要図（拡大版を補足資料に掲載）

3. 手法

3.1 概要

本研究では，モバイル端末上で関節間および時間的関係をモデリングする二段階アーキテクチャを提案する。提案手法は，画像情報を関節トークンへ射影するフレーム単位の空間推定（Stage 1）と，その関節トークン上で因果的に情報統合を行う時空間モデリング（Stage 2）から構成される。概要を図 1 に示す。

Stage 1 では，入力フレームから各関節の 3 次元座標，不確実性，および集約特徴を推定し，その根拠となる画像の特徴量を幾何学的意味を持つ関節トークンへ射影する。Stage 2 では，Stage 1 の出力（関節推定値，信頼度，不確実性表現，画像特徴）を入力とし，因果制約下で時空間統合を行うことで，単一画像推定に内在するジッタや遮蔽による失敗を補正する。さらに，近年の軽量なモバイル端末向け手法 [4], [13] で省略されてきたルート関節の絶対深度推定を導入し，カメラ座標系と整合したルート相対 3 次元関節座標を出力する。

3.2 Stage 1: フレーム単位の空間推定

Stage 1 では，高解像度ヒートマップを生成しない直接回帰設計を採用し，不確実性を保持する多仮説 3 次元人体姿勢推定として空間推定を行う。入力画像 $\mathbf{I}_t$ に対し，3 次元関節座標の K 個の仮説，ルート関節の絶対深度の K 個の仮説，および後段で用いる信頼度・特徴を同時に推定する。

3.2.1 3 次元姿勢座標推定

バックボーンには ConvNeXtPose [13] に基づく BatchNorm 版 ConvNeXt（ConvNeXt-BN）を用い，アップサンプリング層を除去して低解像度特徴マップ $\mathbf{F} \in \mathbb{R}^{H \times W \times C}$ を直接利用する。このとき，単一画像からの 3 次元推定は本質的に不確実であるため，Stage 1 では各関節座標を点推定ではなく，各関節について K 個の仮説 $\{\hat{\mathbf{P}}_j'^{(k)}\}_{k=1}^{K}$ として保持する。学習には DistPred [11] に基づく CRPS 近

似損失を用い，精度項と仮説間多様性項を同時に最適化する。特に断りのない限り，学習時・推論時ともに同一の仮説数 K を用いる。

また，高密度ヒートマップ推定の代わりに，低解像度特徴上でも効果的に動作する SAR [19] に基づく直接回帰を採用する。これにより，モバイル GPU で典型的にボトルネックとなる正規化処理を回避し，Stage 2 の時空間モデリングに必要なレイテンシを確保する。さらにモバイル実装を考慮し，SAR の定式化を離散的最大信頼度選択から連続的な信頼度重み付き集約へ置き換え，特徴量のボクセル化も省略する。加えて，回帰ヘッドはマルチタスクへの対応のため MLP で実装する。

3.2.2 ルート関節の絶対深度の推定

ルート関節の絶対深度 d_{root} についても，各グリッドから専用 MLP により K 個の仮説 $d'^{(k)}_{\mathrm{root},g}$ を推定する。深度推定は画像全体の大域情報に依存しやすいため，Global Average Pooling で得たグローバル特徴 $\mathbf{f}_{\mathrm{global}}$ を結合して推定を行う。さらに RootNet [12] に従い正規化係数 k_{geo} を導入する。なお，本推定にも SAR の概念を導入するが，深度専用の信頼度推定は実施せず，関節座標推定時のルート関節の信頼度 $s_{\mathrm{root},g}$ を再利用する。加えて，タスク間干渉を避けるため，深度ヘッドおよび対応するロジット経路から共有部への勾配伝播を停止する。

3.3 Stage 2: 時空間リファインメント

Stage 2 では，ハイパーグラフに基づく高次関節関係モデリングと，SSM に基づく因果的時間モデリングを，ストリーミング推論と整合する形で統合し，カメラ座標系におけるルート相対 3 次元関節座標 $\mathbf{p}_t^{\mathrm{ref}} \in \mathbb{R}^{J\times 3}$ を推定する。

3.3.1 トークン化処理

幾何的一貫性を保った時空間モデリングのため，Stage 1 の推定結果をカメラ座標系へ変換する。具体的には，得られた $\hat{\mathbf{P}}'^{(k)}_{t,j}$ および $\hat{d}^{(k)}_{t,\mathrm{root}}$ に対して，カメラ内部パラメータに基づくバックプロジェクション Π^{-1} を適用し，絶対座標 $\mathbf{P}^{(k)}_{t,j}$ を得る。つぎに，ルート関節を減算してルート相対座標 $\mathbf{p}^{(k)}_{t,j}$ へ変換し，仮説平均 $\bar{\mathbf{p}}_{t,j}$ と平均からの偏差 $\mathbf{\Delta}^{(k)}_{t,j}$ を算出する。これらに推定信頼度 $\hat{c}_{t,j}$ と集約特徴 $\hat{\mathbf{f}}_{t,j}$ を結合し，Stage 2 の入力トークンを構成する。これにより，単一画像推定の不確実性，信頼度，および画像由来の意味情報を後段へ伝達する。

3.3.2 時空間モデリング

続いて，MixSTE [20] に従い，各トークンに対し，空間モデリングと時間モデリングを L 層交互に適用する。空間層では，フレーム内関節トークン $\{\mathbf{v}_{t,j}\}_{j=1}^{J}$ を Transformer 型アテンションと MLP [18] によりモデリングする。アテンション機構には，ハイパーエッジに基づく高次骨格関係を符号化する HyperSA [21] を用いる。本研究では，計算効率とストリーミング推論時の安定性を優先し，学習可能ではなく固定ハイパーグラフを採用する。時間層では，関節ごとの時系列トークン $\{\mathbf{v}_{\tau,j}\}_{\tau\le t}$ に対し，Mamba [8] に従う S6 機構を用いた SSM ベース状態更新モジュールにより因果的時間モデリングを行う。さらに，ストリーミング推論時のキャッシュサイズ削減のため，Mamba の因果 1D 畳み込みおよびチャネル拡張を除去し，状態更新のみで時間モデリングを実現する。

3.3.3 出力射影

時空間モデリングの最終特徴から線形層により残差補正 $\Delta\mathbf{p}^{\mathrm{res}}_{t,j}$，再構成座標 $\mathbf{p}^{\mathrm{new}}_{t,j}$，およびゲートロジット $g^{\mathrm{logit}}_{t,j}$ を推定する。シグモイド関数により得たゲート係数 $\alpha_{t,j}$ を用いて，関節ごとに残差経路と再構成経路を混合する。これにより，Stage 1 の初期推定に対する局所的補正だけでなく，大きな再配置も可能にする。

3.4 学習戦略と損失設定

Stage 2 からの勾配が Stage 1 表現を不安定化させることを防ぐため，二段階で学習する。まず Stage 1 を画像データで学習し，その後 Stage 1 を固定して Stage 2 を動画データで学習する。なお，Stage 1 の学習では深度関連損失を除外することで 2 次元ラベルのみ付与されたデータも利用するが，Stage 2 の学習では 3 次元アノテーションを持つデータのみを使用する。

Stage 1 の学習では，関節座標についてはグリッドごとに計算した平均推定に対する回帰損失と DistPred [11] に基づく CRPS 近似損失をを SAR [19] に準拠して空間的に集約を行ったものを損失とする。なおこの損失は X, Y を正規化画像座標，Z を正規化相対深度として表現した座標空間で計算する。ルート関節の絶対深度についても同様にグリッドごとに計算した平均損失と分布損失を空間的に集約したものを損失とする。したがって，最終的な Stage 1 の損失はこれらの損失の和として定義される。

Stage 2 の学習では，カメラ座標系におけるルート相対 3 次元関節座標を用いて最適化する。KTPFormer [15] に倣った回帰誤差（重み付き MPJPE）と時間平滑化正則化に加え，正解骨長への一致および入力系列内骨長一貫性を導入し，これらの和を損失として定義する。

4. 実験

4.1 実験設定

学習および評価には Human3.6M [10] を主に用い, Stage 1 の汎化性能向上のため MPII [1] も併用する。すべてのデータセットで 2 次元関節座標から人物バウンディングボックスを生成し，人物領域を切り出して入力とする。学習・推論では 18 関節表現（thorax 関節を追加）を用い，評価時には 17 関節で評価する。MPII には Human3.6M の関節定義へ合わせる joint-mapping を適用する。評価は ConvNeXtPose [13] における Human3.6M Protocol 2 に従

表 1　ConvNeXtPose との定量比較

Model	P2 MPJPE (mm) ↓	MACs (G) ↓	Params (M) ↓	GPU (FP16) (ms) ↓
w/ GT absolute root joint depth				
ConvNeXtPose-XS [13]	56.6	**0.82**	**3.53**	25.99 ± 0.39
ConvNeXtPose-S [13]	51.8	1.76	7.44	31.73 ± 0.41
ConvNeXtPose-M [13]	51.05	2.82	7.59	2104.88 ± 10.93
ConvNeXtPose-L [13]	**49.75**	4.30	8.38	2151.13 ± 8.76
PocketSTPose (Ours)	49.93	1.02	10.69	$\mathbf{13.49 \pm 0.02}$
w/o GT absolute root joint depth				
PocketSTPose (Ours)	50.39	1.09	11.64	14.57 ± 0.02

い，S1, S5, S6, S7, S8 で学習し，S9, S11 で評価する。なお，このとき ConvNeXtPose に準じ，評価対象フレームは 64 フレーム間隔でサンプリングされる。主評価指標には MPJPE を用い，全評価対象における平均を報告する。加えて，時間的一貫性の指標として動画単位での骨長の分散と変動係数（BL_{var}, BL_{cv}）もそれぞれその平均を報告する。実機レイテンシは Samsung Galaxy S20（Snapdragon 865）上で測定し，すべてのモデルを TensorFlow Lite 形式へ変換して GPU delegate（FP16 を許容する設定）で評価する。Stage 1 では ConvNeXtPose-XS のバックボーン構成を用い，入力解像度は 256×256，各グリッドあたりのサンプル数は $K = 100$ とする。Stage 2 は層数 $L = 4$，埋め込み次元 256，アテンションヘッド数 8 で構成する。なお，評価では，評価対象フレームに対して先行フレームを 10 fps でサンプリングした $T = 30$ フレームを入力し，レイテンシはキャッシュを用いた 1 フレームごとのストリーミング推論で測定する。

4.2　先行研究との比較

提案手法を，精度と実機レイテンシのバランスに優れるモバイル端末向けの先行研究である ConvNeXtPose [13] と比較する。表 1 に，Human3.6M Protocol 2（P2）における精度，計算コスト，および実機上で測定した GPU 推論レイテンシを示す。なお，レイテンシについては，すべて前述の環境で実測した値を報告する。また，ConvNeXtPose は評価時にルート関節の絶対深度の GT を用いるため，公平な比較のために，我々の手法についてもルート関節の絶対深度の GT を用いた結果（w/ GT absolute root joint depth）を併せて報告する。

表 1 に示すように，提案手法は，大型の ConvNeXtPose-M/L に近い精度を，より少ない計算量で実現している。ルート関節の絶対深度の GT を用いる同条件の比較では，ConvNeXtPose-L の約 24%の MACs で，同等の MPJPE を達成した。さらに，より実用的なルート関節の絶対深度の GT を用いない設定でも，ConvNeXtPose-M より高い精度を，約 38%の MACs で実現している。

レイテンシの観点では，表 1 に示すように，ConvNeXtPose-XS/S は GPU 上で数十 ms で動作する一方，より高解像度ヒートマップを用いる M/L では 2000 ms を超える極端な遅延が生じる。これに対し，提案手法は高解像度ヒートマップやその正規化を必要としない。その結果，GPU（FP16）設定で約 14 ms の推論時間を達成し，ルート関節の絶対深度の GT を用いない場合でも増加はわずかである。したがって，提案手法は実用条件下でも安定したリアルタイム性能を維持することが可能である。

4.3　時空間モデリングの有効性

時空間モデリング自体の有効性を検証するため，バックボーンを拡張した Stage 1 と比較した。ルート関節の絶対深度の GT を用いた設定では，提案手法は拡張した Stage 1 に対して，MPJPE を 51.20 mm から 49.93 mm へ，BL_{var} を $196.27\,\mathrm{mm}^2$ から $41.97\,\mathrm{mm}^2$ へ，BL_{cv} を 4.49 から 2.31 へ改善した。なお，この際の GPU レイテンシはそれぞれ 13.49 ± 0.02 ms と 15.33 ± 0.04 ms である。また，ルート関節の絶対深度の GT を用いない実用設定でも，同様に GPU レイテンシでは 14.57 ± 0.02 ms と 16.40 ± 0.06 ms と，提案手法のほうが良好であるにも関わらず，MPJPE を 55.87 mm から 50.39 mm へ，BL_{var} を $209.17\,\mathrm{mm}^2$ から $25.81\,\mathrm{mm}^2$ へ，BL_{cv} を 4.77 から 1.87 へ改善した。これらの結果は，単一フレームモデルの容量増加だけではなく，高度な時空間モデリングそのものが関節座標精度と時間的一貫性の両方に寄与することを示している。

5.　まとめ

本研究では，モバイル端末上でのリアルタイムストリーミング 3 次元人体姿勢推定のための新たなフレームワークを提案した。軽量な単一フレーム推定と状態更新に基づく時空間リファインメントを分離することで，厳しいレイテンシ制約下でも時空間依存関係を統合できる。実機上で約 15 ms のリアルタイム推論を実現しつつ，Human3.6M Protocol 2 で 50.39 mm の MPJPE を達成した。今後は，リアルタイム性能を維持したまま非ルート相対な絶対座標系での時空間リファインメントへの拡張が期待される。

参考文献

[1] Andriluka, M., Pishchulin, L., Gehler, P. and Schiele, B.: 2D Human Pose Estimation: New Benchmark and State of the Art Analysis, *Proc. of IEEE Computer Vision and Pattern Recognition*, pp. 3686–3693 (2014).

[2] Bazarevsky, V., Ivan, G., Karthik, R., Tyler, Z., Fan, Z. and Matthias, G.: BlazePose: On-device Real-time Body Pose Tracking, *arXiv:2006.10204* (2020).

[3] Camillo, L., Jiuqiang, T., Hadon, N., Chris, M., Esha, U., Michael, H., Fan, Z., Chuo-Ling, C., Ming, Y., Juhyun, L., Wan-Teh, C., Wei, H., Manfred, G. and Matthias, G.: MediaPipe: A Framework for Perceiving and Processing Reality, *Third Workshop on Computer Vision for AR/VR at IEEE Computer Vision and Pattern Recognition* (2019).

[4] Choi, S., Choi, S. and Kim, C.: MobileHumanPose: Toward real-time 3D human pose estimation in mobile devices, *Proc. of IEEE Computer Vision and Pattern Recognition*, pp. 2328–2338 (2021).

[5] Cui, H. and Hayama, T.: HGMamba: Enhancing 3D Human Pose Estimation with a [H]yperGCN-Mamba Network, *Proc. of International Joint Conference on Neural Networks* (2025).

[6] Cui, H., Hua, W., Huang, R., Jia, S. and Hayama, T.: SasMamba: A Lightweight Structure-Aware Stride State Space Model for 3D Human Pose Estimation, *Proc. of IEEE Winter Conference on Applications of Computer Vision* (2026).

[7] Dao, T., Fu, D. Y., Saab, K. K., Thomas, A. W., Rudra, A. and Re, C.: Hungry Hungry Hippos: Towards Language Modeling with State Space Models, *Proc. of International Conference on Learning Representations* (2023).

[8] Gu, A. and Dao, T.: Mamba: Linear-Time Sequence Modeling with Selective State Spaces, *arXiv:2312.00752* (2023).

[9] Huang, Y., Liu, J., Xian, K. and Qiu, R. C.: PoseMamba: monocular 3D human pose estimation with bidirectional global-local spatio-temporal state space model, *Proc. of AAAI Conference on Artificial Intelligence*, pp. 3842–3850 (2025).

[10] Ionescu, C., Papava, D., Olaru, V. and Sminchisescu, C.: Human3.6M: Large Scale Datasets and Predictive Methods for 3D Human Sensing in Natural Environments, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 36, No. 7, pp. 1325–1339 (2014).

[11] Liang, D.: DistPred: A Distribution-Free Probabilistic Inference Method for Regression and Forecasting, *Proc. of ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 753–764 (2025).

[12] Moon, G., Chang, J. Y. and Lee, K. M.: Camera Distance-Aware Top-Down Approach for 3D Multi-Person Pose Estimation From a Single RGB Image, *Proc. of IEEE International Conference on Computer Vision*, pp. 10132–10141 (2019).

[13] Nguyen, H. S., Kim, M., Im, C., Han, S. and Han, J.: ConvNeXtPose: a fast accurate method for 3D human pose estimation and its AR fitness application in mobile devices, *IEEE Access*, Vol. 11, pp. 117393–117402 (2023).

[14] Pavllo, D., Feichtenhofer, C., Grangier, D. and Auli, M.: 3D Human Pose Estimation in Video With Temporal Convolutions and Semi-Supervised Training, *Proc. of IEEE Computer Vision and Pattern Recognition*, pp. 7745–7754 (2019).

[15] Peng, J., Zhou, Y. and Mok, P.: KTPFormer: Kinematics and Trajectory Prior Knowledge-Enhanced Transformer for 3D Human Pose Estimation, *Proc. of IEEE Computer Vision and Pattern Recognition*, pp. 1123–1132 (2024).

[16] Qiu, Z., Qiansheng, Y., Jian, W. and Dongmei, F.: IVT: An End-to-End Instance-guided Video Transformer for 3D Pose Estimation, *Proc. of ACM International Conference Multimedia*, pp. 6174–6182 (2022).

[17] Sun, X., Xiao, B., Wei, F., Liang, S. and Wei, Y.: Integral Human Pose Regression, *Proc. of European Conference on Computer Vision*, pp. 536–553 (2018).

[18] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L. and Polosukhin, I.: Attention is all you need, *Proc. of Neural Information Processing Systems*, p. 6000–6010 (2017).

[19] Wang, D. and Zhang, S.: Spatial-Aware Regression for Keypoint Localization, *Proc. of IEEE Computer Vision and Pattern Recognition*, pp. 624–633 (2024).

[20] Zhang, J., Tu, Z., Yang, J., Chen, Y. and Yuan, J.: MixSTE: Seq2seq Mixed Spatio-Temporal Encoder for 3D Human Pose Estimation in Video, *Proc. of IEEE Computer Vision and Pattern Recognition*, pp. 13232–13242 (2022).

[21] Zhou, Y., Cheng, Z.-Q., Li, C., Fang, Y., Geng, Y., Xie, X. and Keuper, M.: Hypergraph Transformer for Skeleton-based Action Recognition, *arXiv:2211.09590* (2022).