# モバイル端末向けリアルタイム3次元人体姿勢推定における時間軸情報の活用

電気通信大学 柳井研究室
中溝雄斗, 木村倫太郎, 柳井啓司

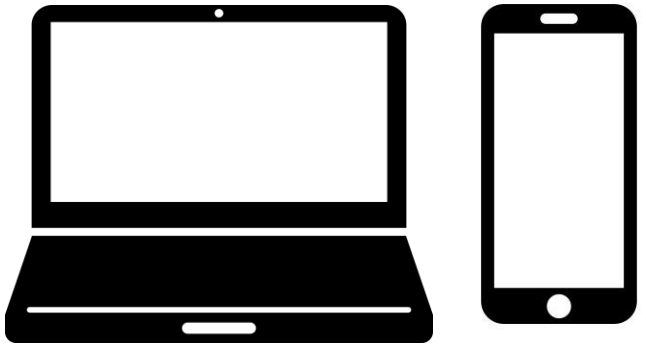
# はじめに

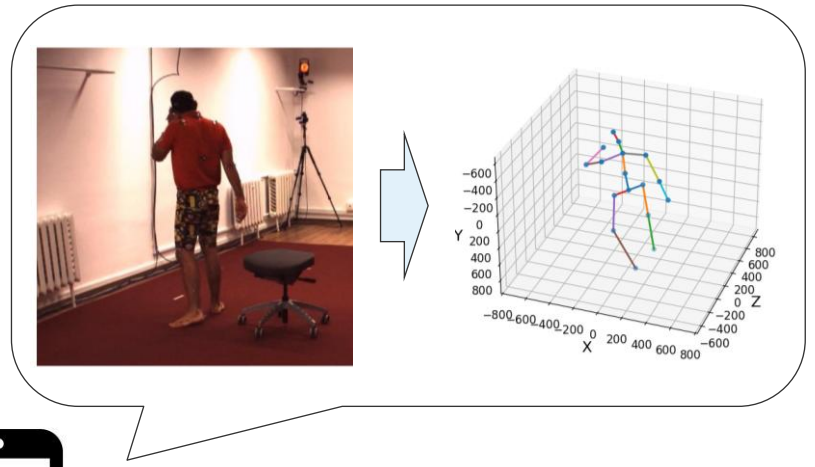
# 背景

## 3次元人体姿勢推定とは？

➡ 画像・映像・センサーデータなどから人物の関節座標を推定し、3次元の骨格モデルを復元すること



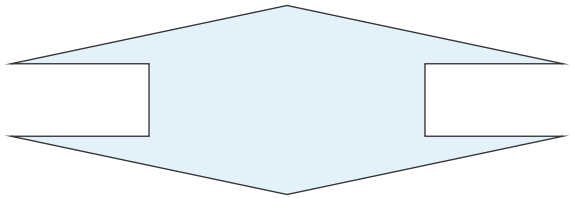




# 研究的背景

**基本的には画像・映像から姿勢を復元するタスクに注目**

- 昨今の研究は事前知識や時間軸情報の活用から高い精度を達成しているものの、高性能なGPUが前提



**モバイル端末でリアルタイムに推論することを目的とした研究・モデルは限定的**

- 事前知識や時間軸情報の活用不足から精度が不十分であるのが現状
- 近年の研究では現実空間の座標系での推定は行えない

# 課題と目標

既存のモバイル端末向けかつ、リアルタイム推論が可能な軽量モデルは

- 精度が不十分
- 現実空間の座標系での姿勢を推定できないものも

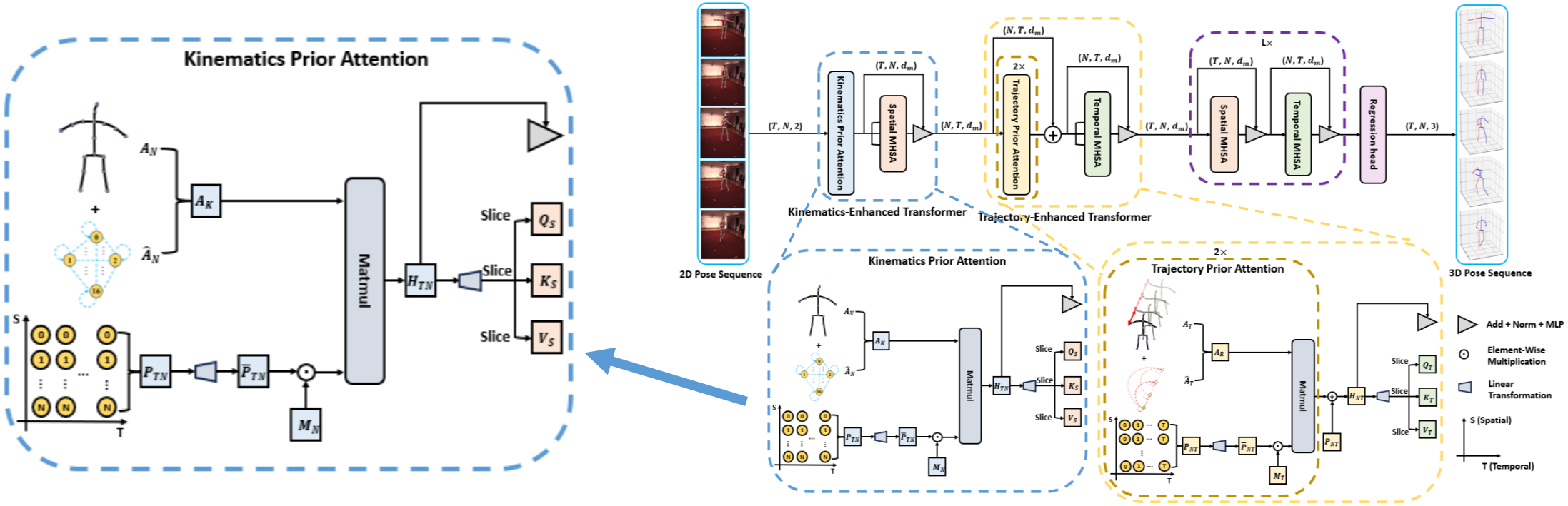
- **GTに依存しない現実空間の座標系での姿勢推定**
- **時系列処理による高精度性**

# 関連研究

# 3次元人体姿勢推定

## 長時間シーケンス活用と空間的事前知識の導入により高精度化

- KTPFormer [1]: 大規模構造と骨格情報による条件付け、
  長時間シーケンス処理を組み合わせることで高性能を実現




[1] Peng Jihua, et al. "Ktpformer: Kinematics and trajectory prior knowledge-enhanced transformer for 3d human pose estimation." CVPR, 2024.

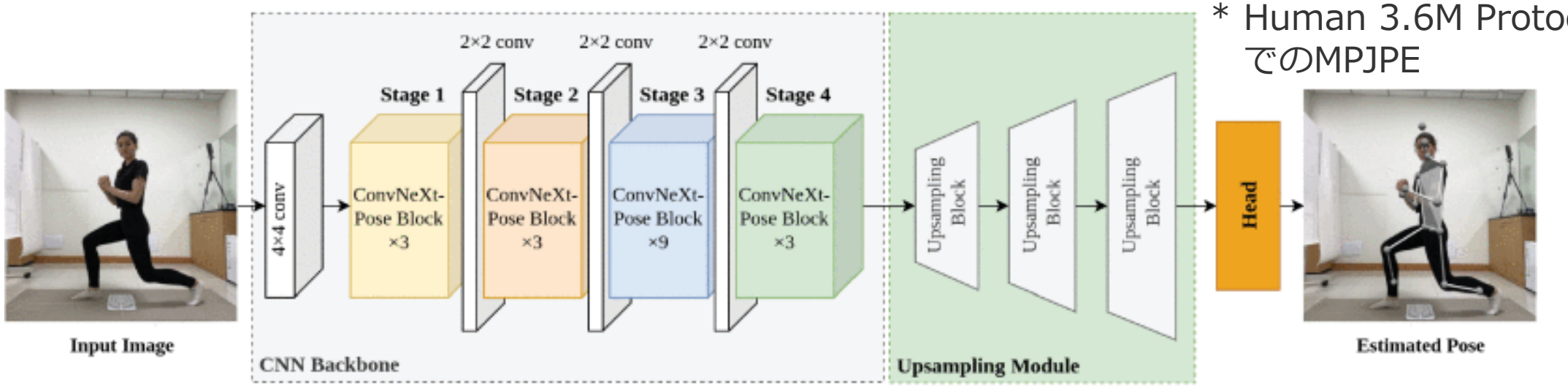
# 3次元人体姿勢推定

**モバイル端末向け軽量モデルでは時系列処理は不採用**

ConvNeXtPose [2]: ConvNeXt + 軽量アップサンプリングで高効率な推論を実現したが、精度は大規模なモデルに及ばない

- 代表的な指標*において、KTPFormer比で約71%の精度

※ ただし、現実空間の座標系での推定ではない




* Human 3.6M Protocol2上でのMPJPE

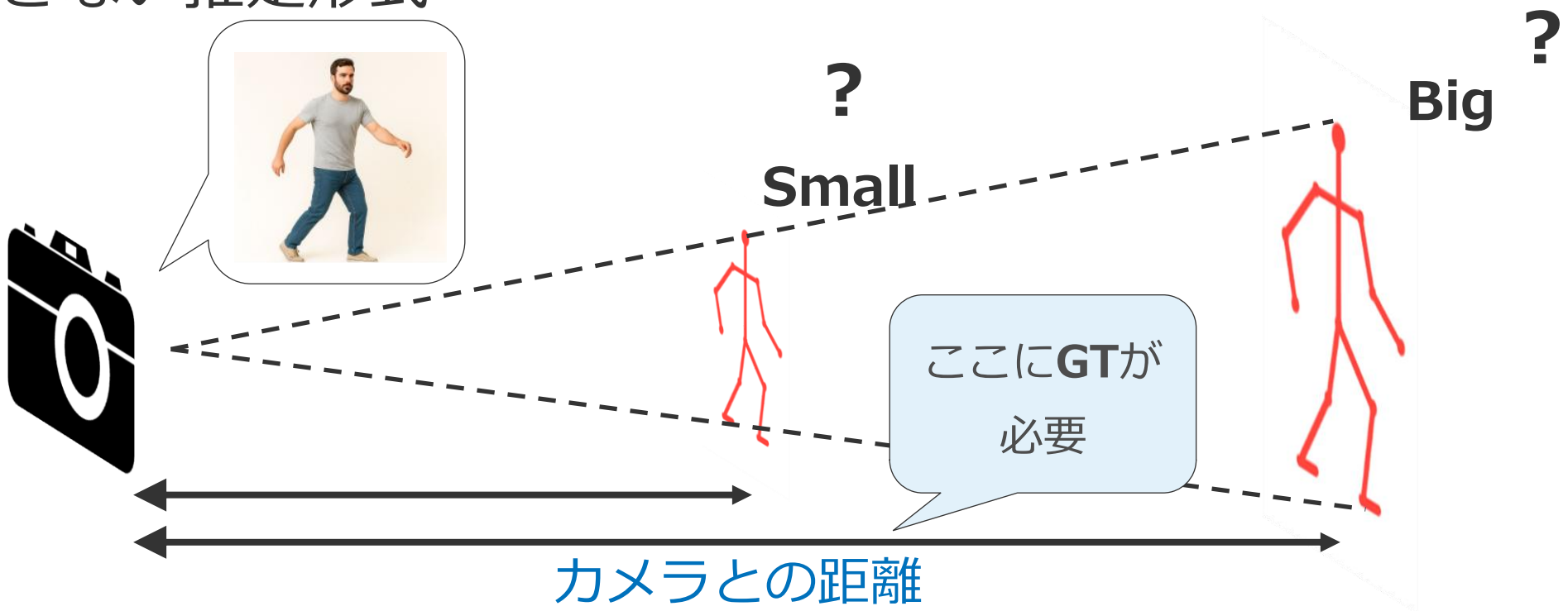
[2] Nguyen Hong Son, et al. "Convnextpose: A fast accurate method for 3d human pose estimation and its AR fitness application in mobile devices." IEEE Access, 2023.

# 3次元姿勢推定

## 現実空間の座標系での推定ではないとは？

関節座標のXY座標を画像上のピクセル座標系で推定しており、

カメラとの距離のGT情報を参照しないと現実空間でのメートル座標系に変換できない推定形式

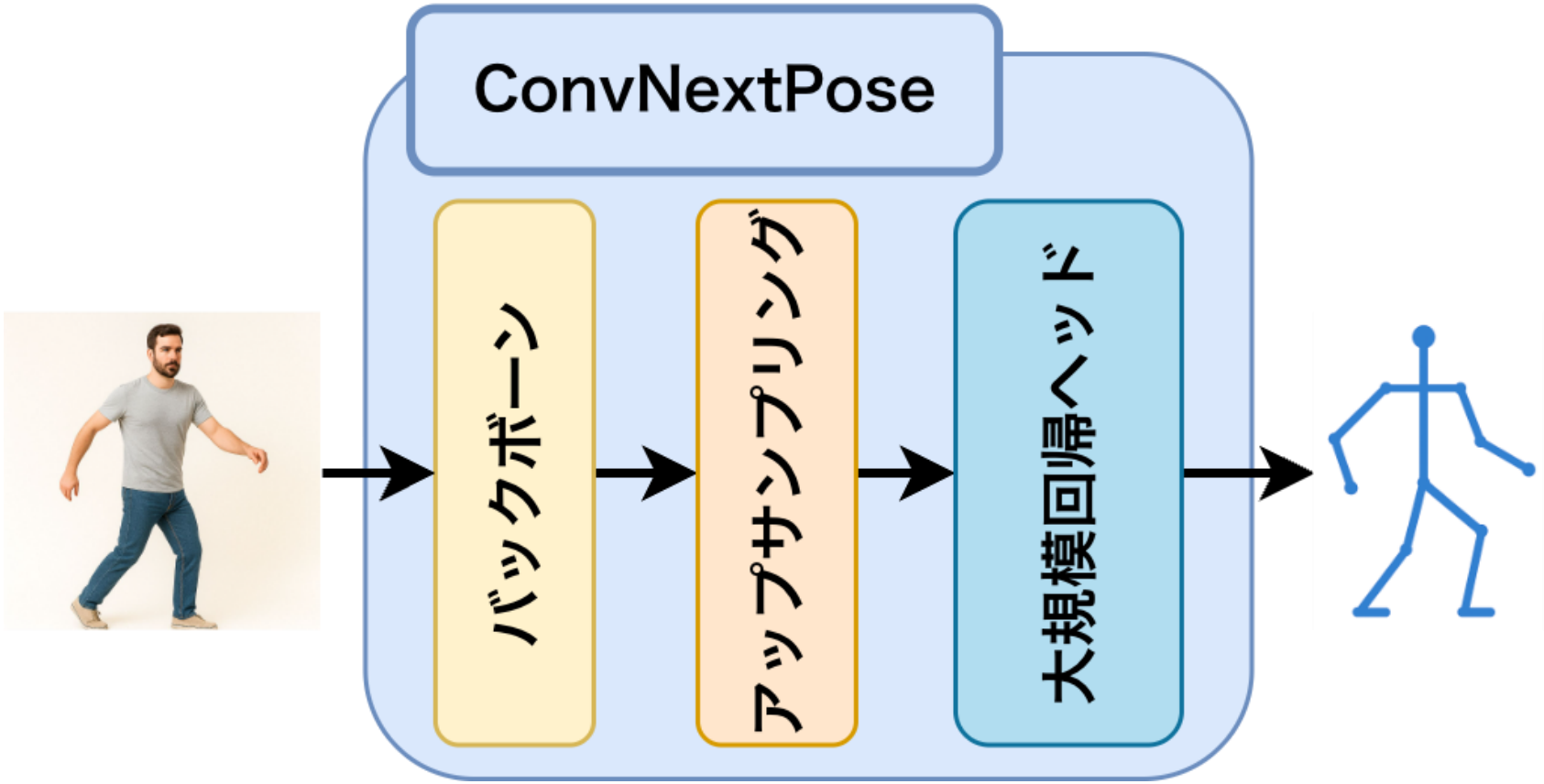


# 提案手法

# 提案手法の全体像

## ConvNeXtPoseをベースに設計する

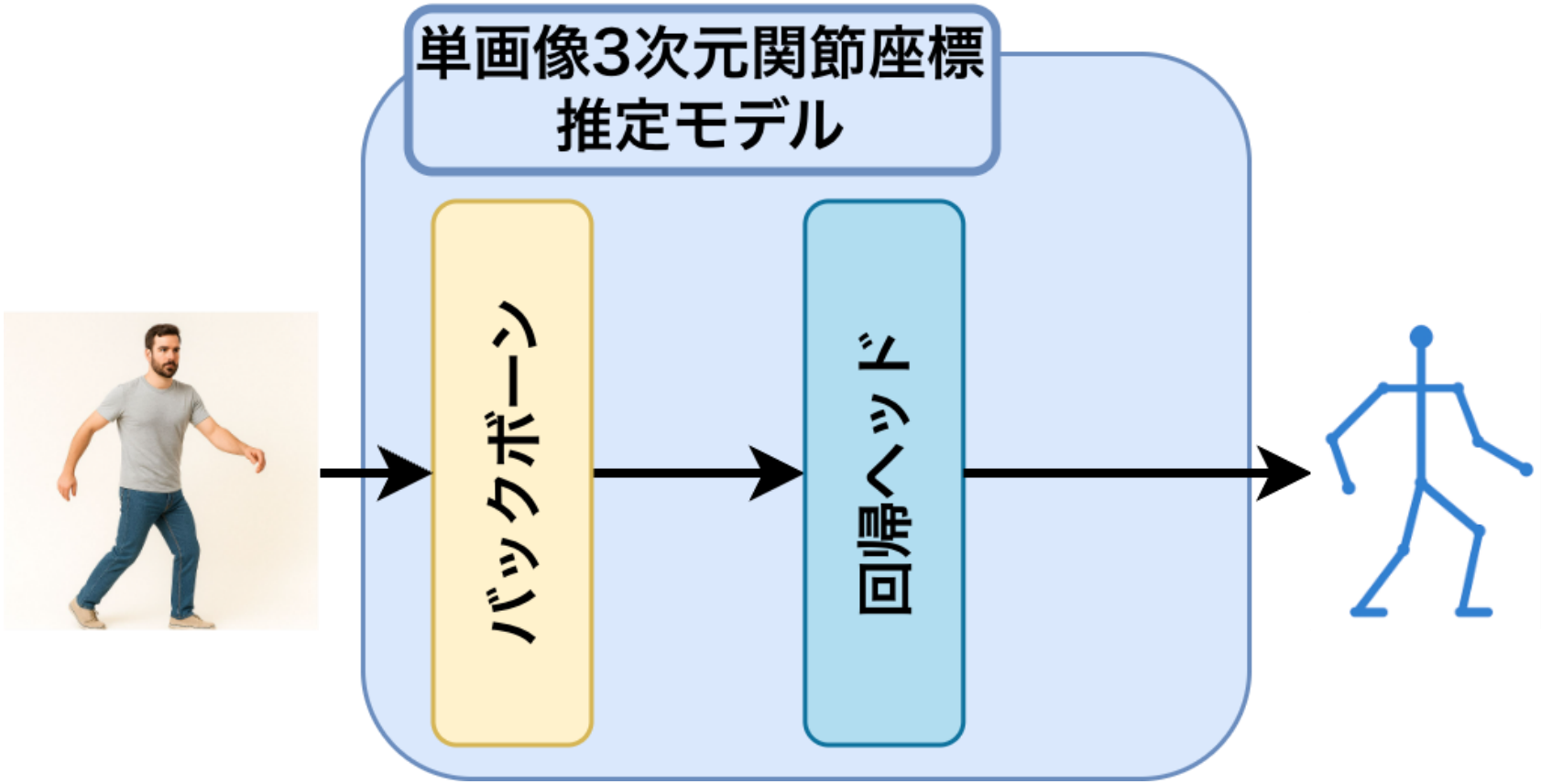
1. 軽量なヘッドへの換装 + アップサンプリングブロックの削除



# 提案手法の全体像

## ConvNeXtPoseをベースに設計する

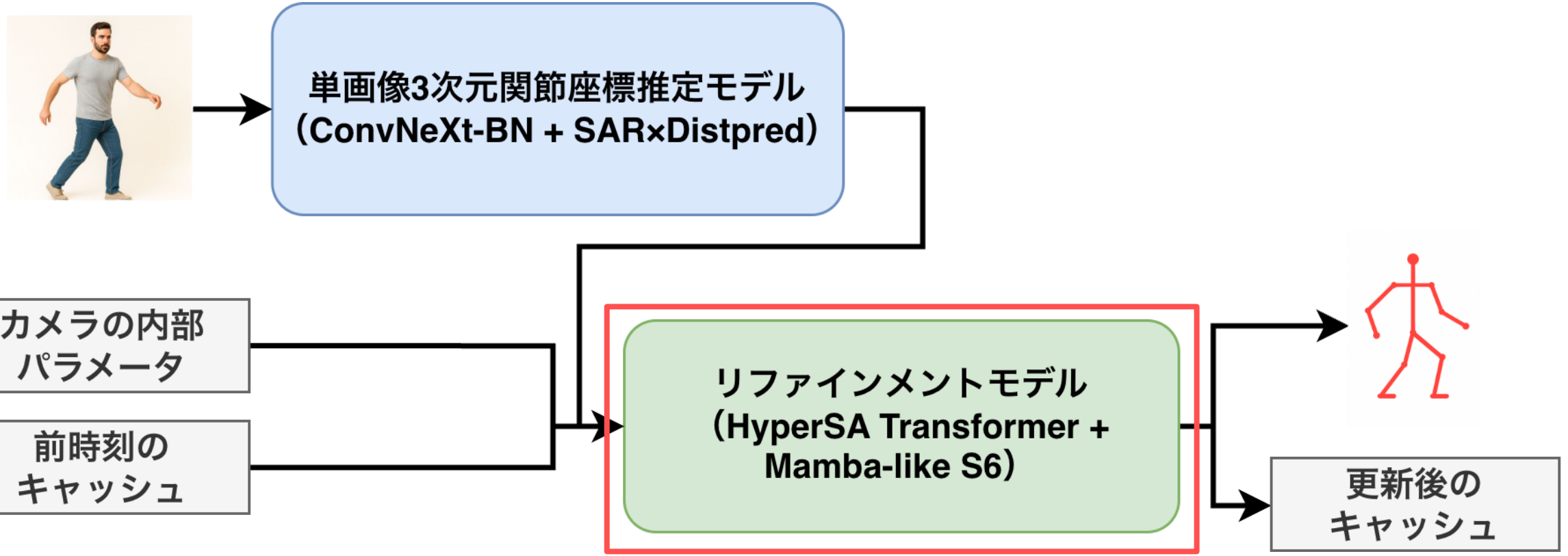
1. 軽量なヘッドへの換装 + アップサンプリングブロックの削除



# 提案手法の全体像

## ConvNeXtPoseをベースに設計する

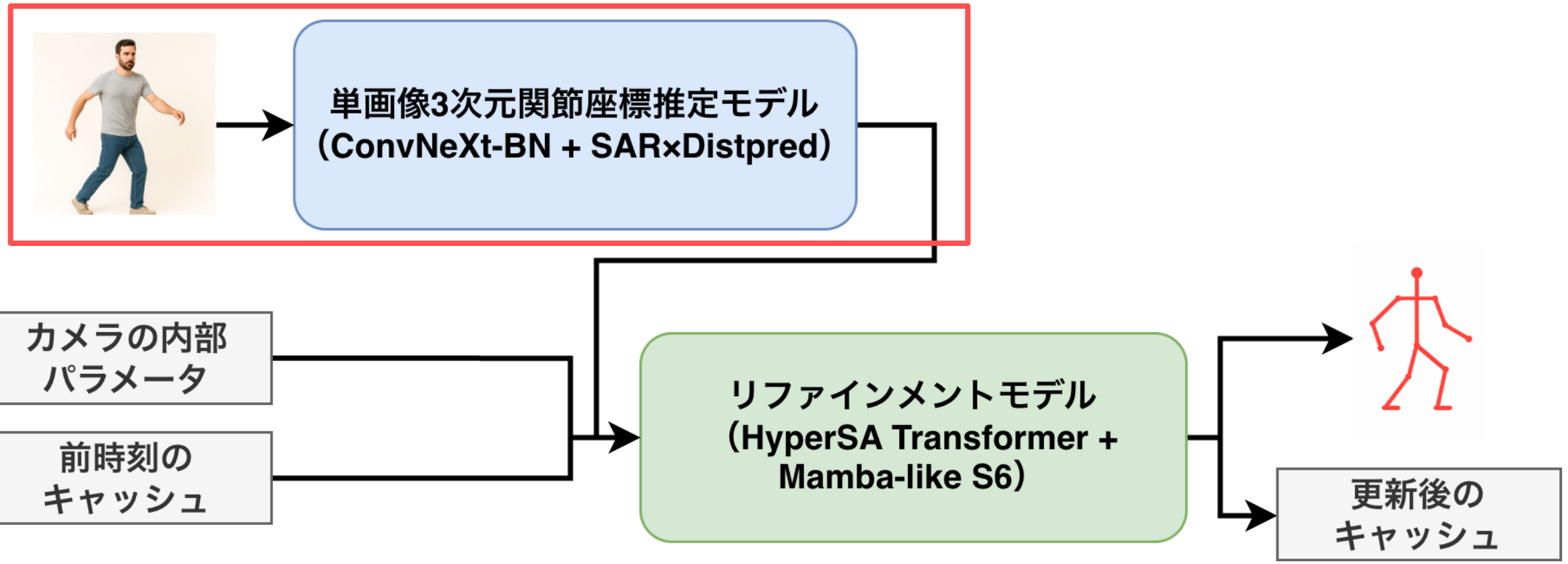
1. 軽量なヘッドへの換装 + アップサンプリングブロックの削除
2. 逐次的に時系列処理を行うリファインメントモデルを追加



# 提案手法の全体像

## ConvNeXtPoseをベースに設計する

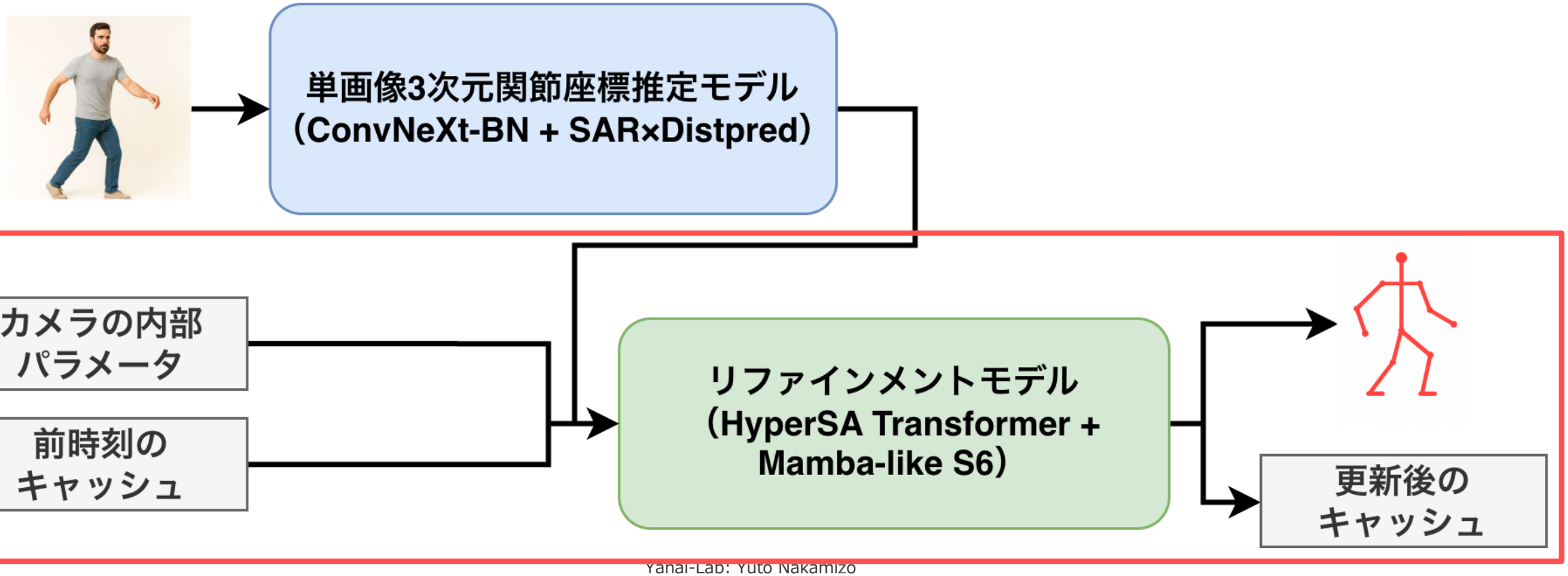
1. 軽量なヘッドへの換装 + アップサンプリングブロックの削除
2. 逐次的に時系列処理を行うリファインメントモデルを追加



# 提案手法の全体像

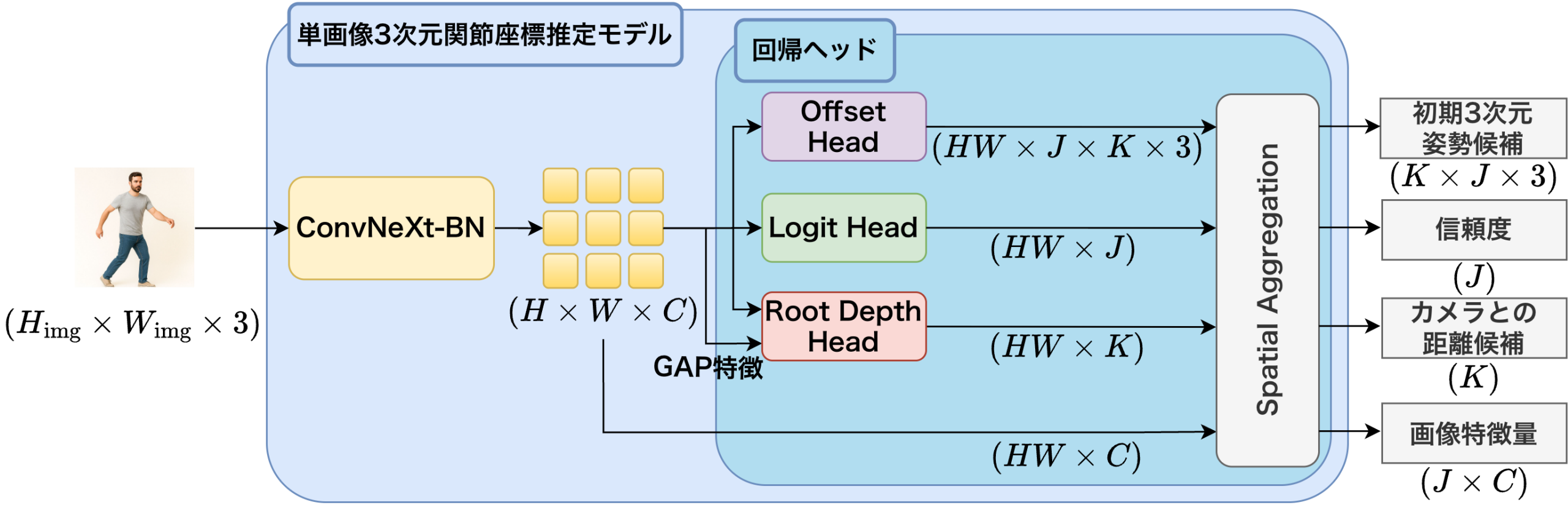
## ConvNeXtPoseをベースに設計する

1. 軽量なヘッドへの換装 + アップサンプリングブロックの削除
2. 逐次的に時系列処理を行うリファインメントモデルを追加



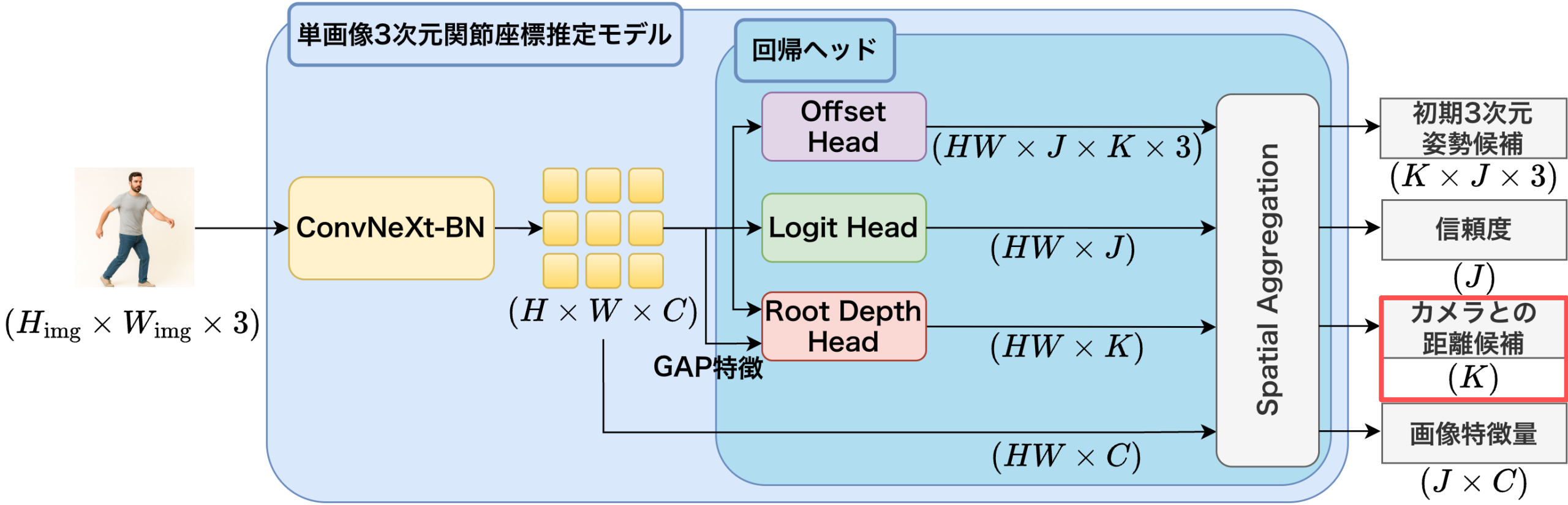
# 単画像3次元関節座標推定モデル

**フレームから初期3次元姿勢候補、カメラとの距離候補、信頼度を推定（画像特徴量も合わせて抽出）**



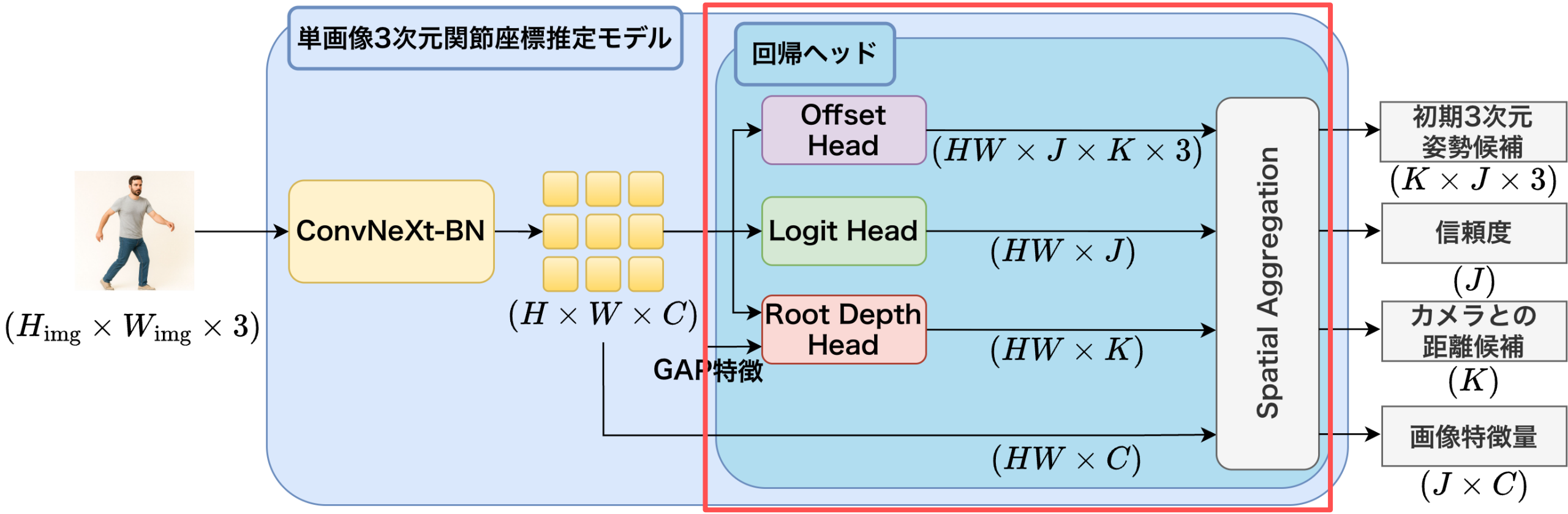
# 単画像3次元関節座標推定モデル

**フレームから初期3次元姿勢候補、カメラとの距離候補、**
**信頼度を推定（画像特徴量も合わせて抽出）**



# 単画像3次元関節座標推定モデル

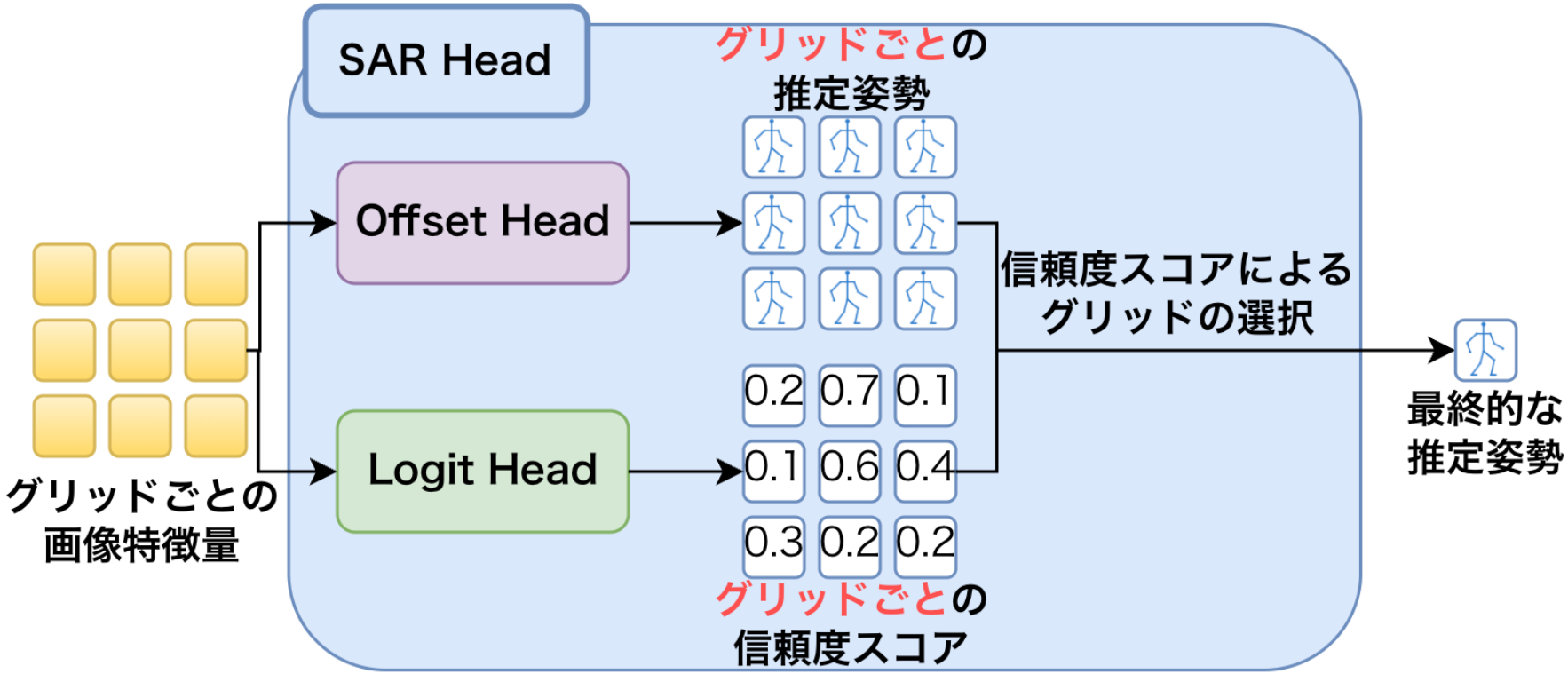
**フレームから初期3次元姿勢候補、カメラとの距離候補、**
**信頼度を推定（画像特徴量も合わせて抽出）**



# 単画像3次元関節座標推定モデル：回帰ヘッド

## SAR [3]を回帰ヘッドとして導入

- 各グリッドの画像特徴量に基づいて座標を直接推定する回帰手法
- 大規模な正規化処理が不要なため、モバイル適性あり



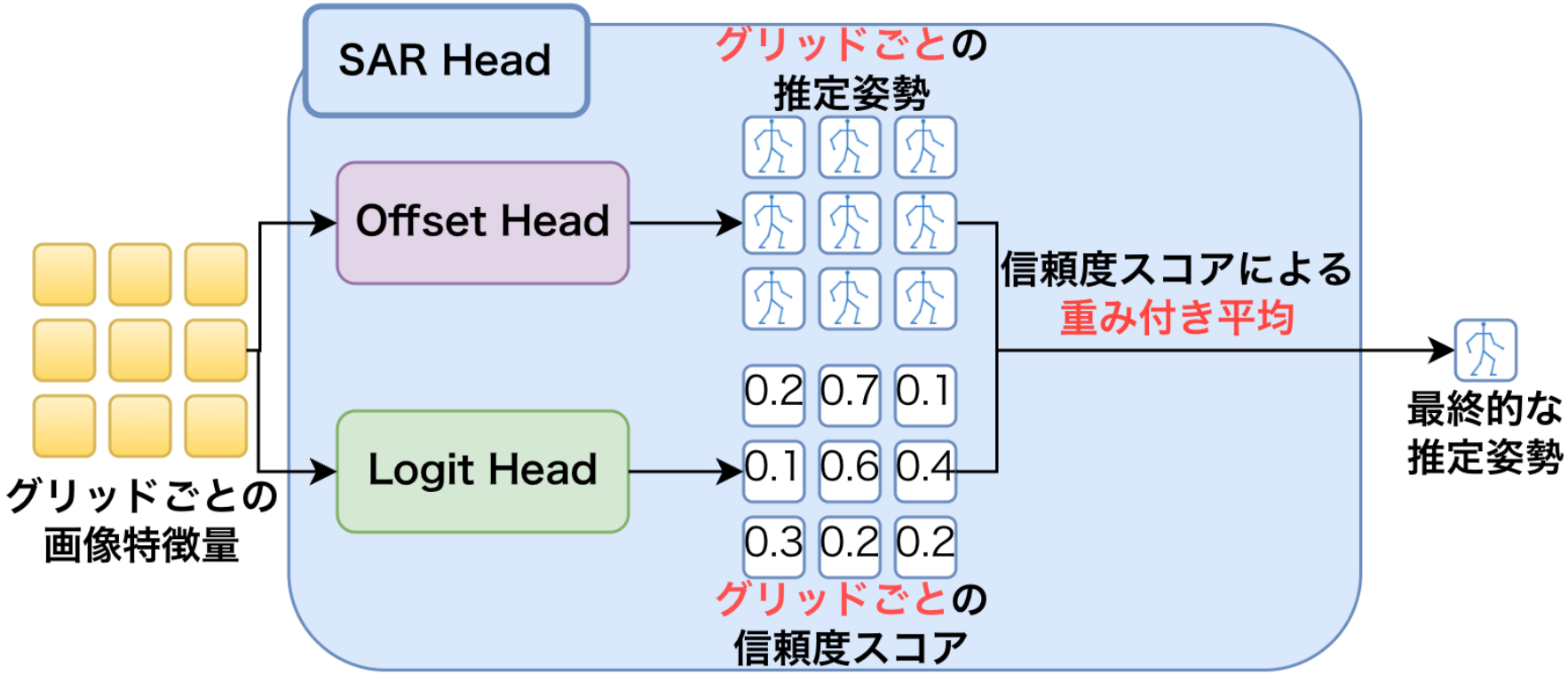

※ カメラからの距離候補や信頼度、画像特徴量の集約も同様に実施

[3] Dongkai Wang, et al. "Spatial-aware regression for keypoint localization." CVPR, 2024.

# 単画像3次元関節座標推定モデル：回帰ヘッド

## SAR [3]を回帰ヘッドとして導入

- 各グリッドの画像特徴量に基づいて座標を直接推定する回帰手法
- 大規模な正規化処理が不要なため、モバイル適性あり




※ カメラからの距離候補や信頼度、画像特徴量の集約も同様に実施

[3] Dongkai Wang, et al. "Spatial-aware regression for keypoint localization." CVPR, 2024.

# 単画像3次元関節座標推定モデル：回帰ヘッド

## DistPred [4]を導入し、姿勢推定と絶対深度推定を拡張

- 単一値ではなく複数のサンプルを出力し、それらによる分布として回帰する手法
- 推定結果の不確実性を後段の処理に伝えるために利用

首の関節座標は
**(x, y, z) = (1.0, 2.0, 3.0)** か
**(1.1, 1.9, 2.5)** か…



モデル

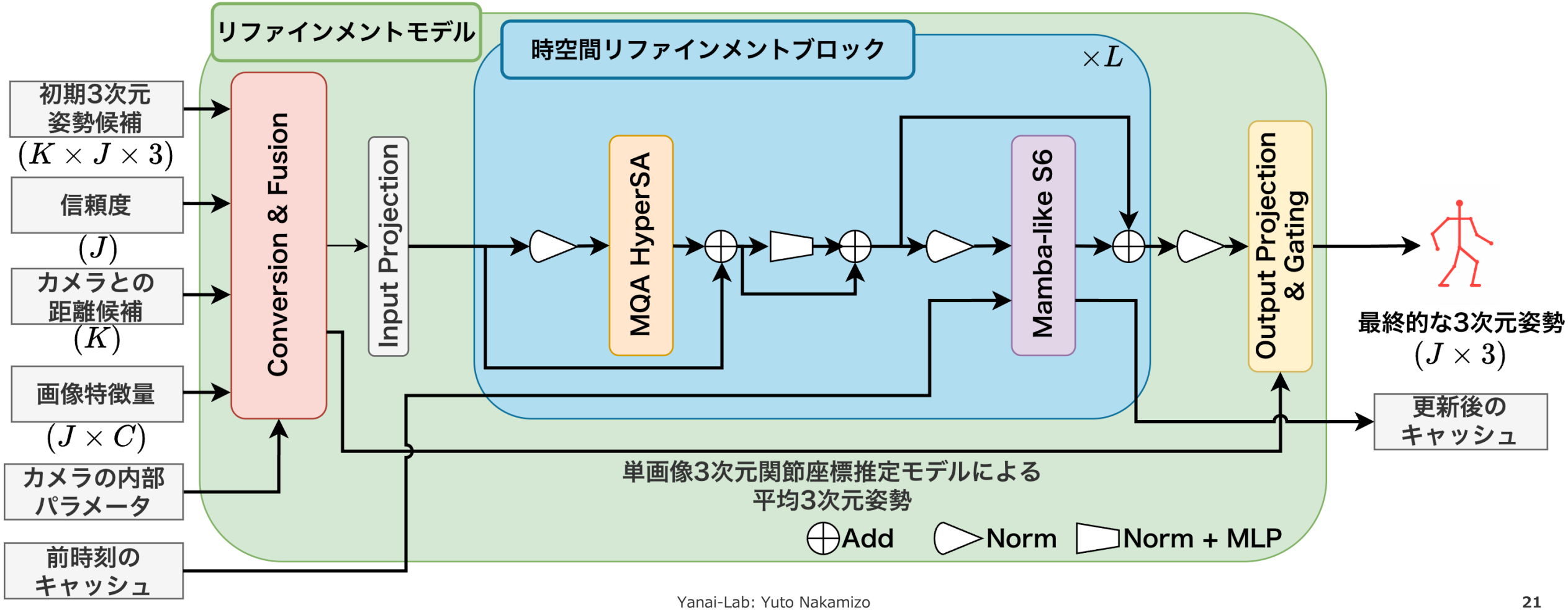
※ カメラからの距離候補の推定でも同様に実施

[4] Daojun Liang. "DistPred: A Distribution-Free Probabilistic Inference Method for Regression and Forecasting.", ACM SIGKDD, 2025.

# リファインメントモデル

**前段の出力と過去の状態を保存したキャッシュに基づき、**

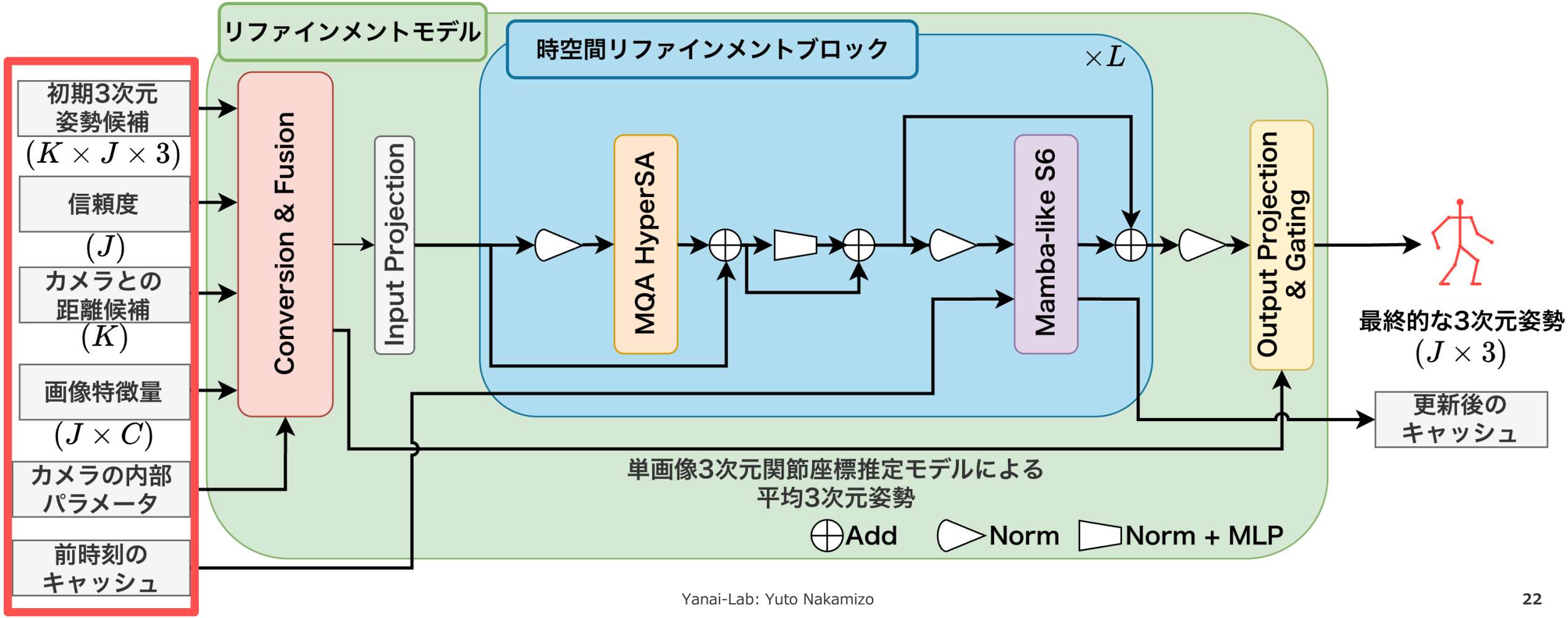
**前段の3次元姿勢を調整する**



# リファインメントモデル

**前段の出力と過去の状態を保存したキャッシュに基づき、**

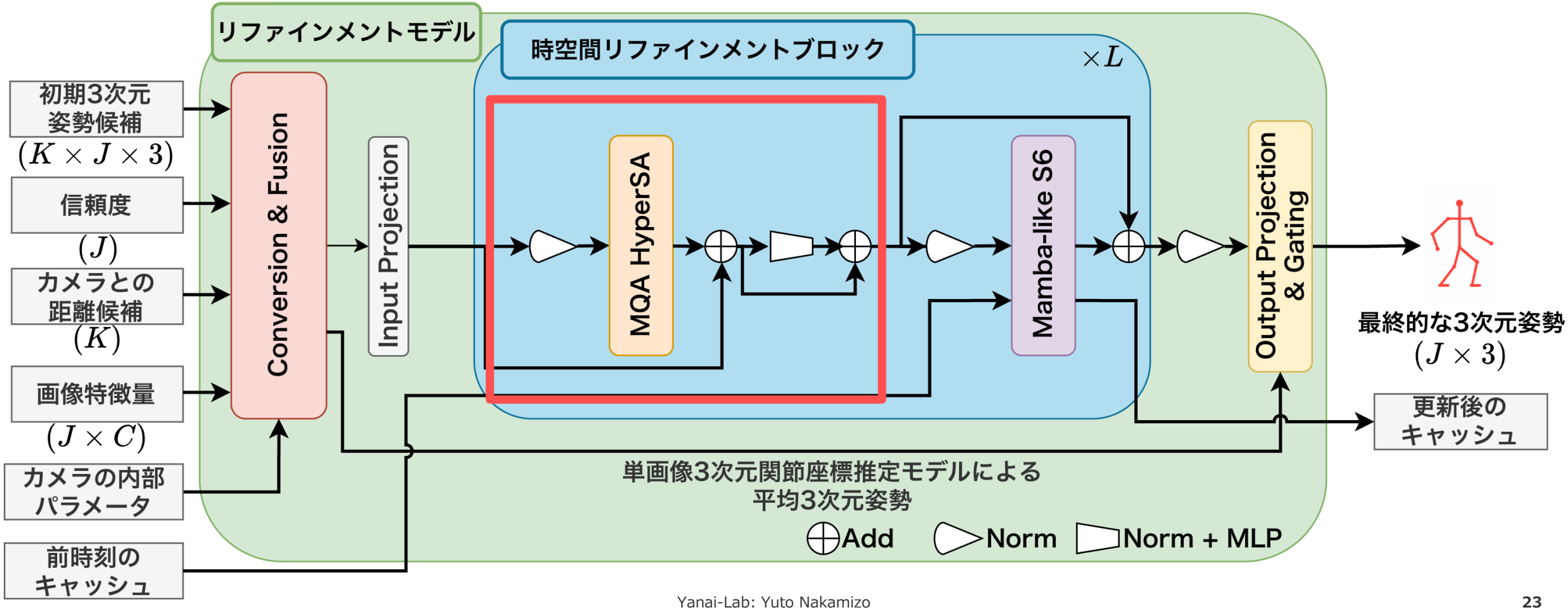
**前段の3次元姿勢を調整する**



# リファインメントモデル

**前段の出力と過去の状態を保存したキャッシュに基づき、**

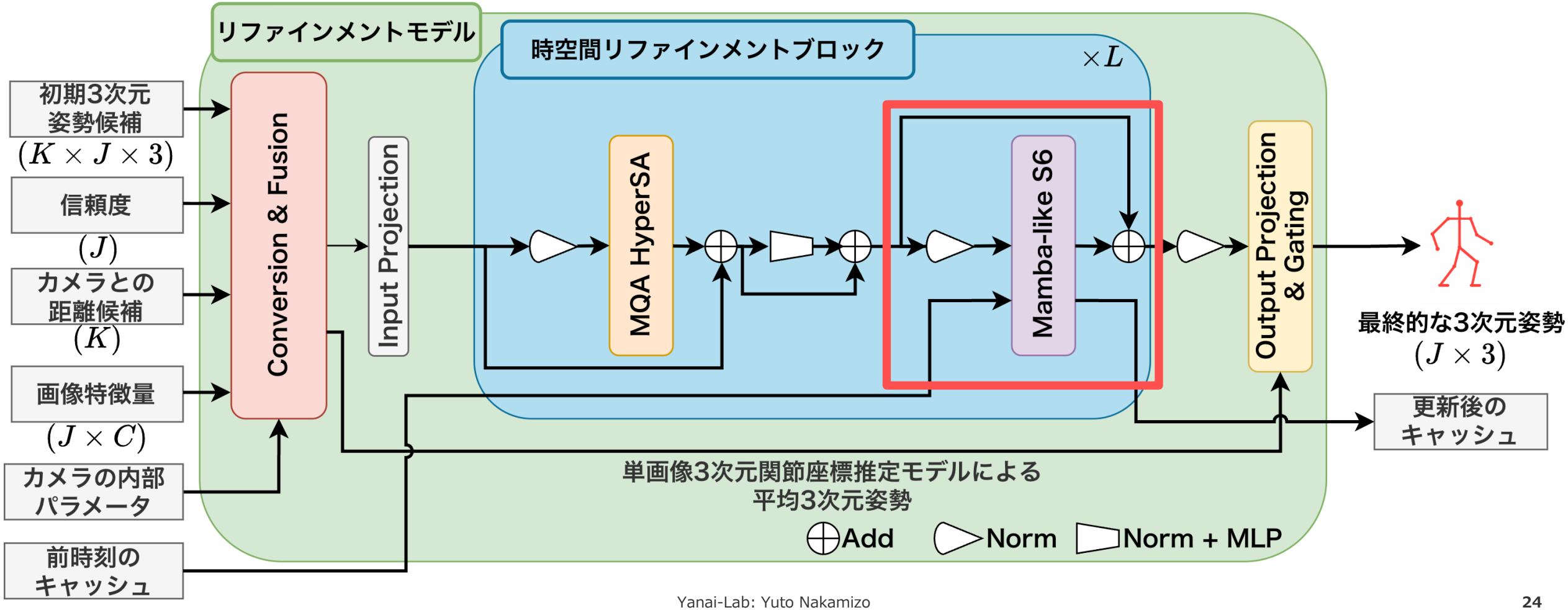
**前段の3次元姿勢を調整する**



# リファインメントモデル

**前段の出力と過去の状態を保存したキャッシュに基づき、**

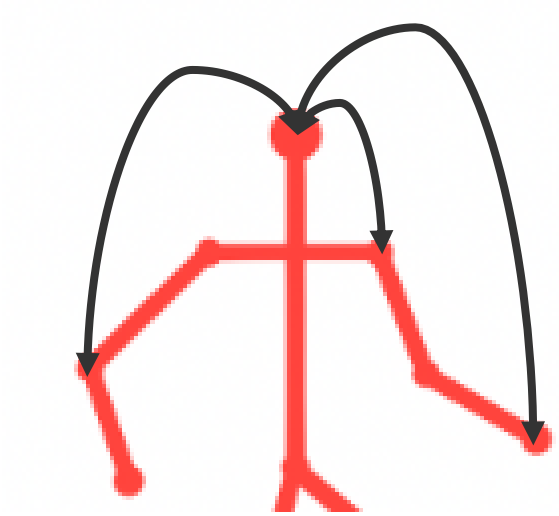
**前段の3次元姿勢を調整する**



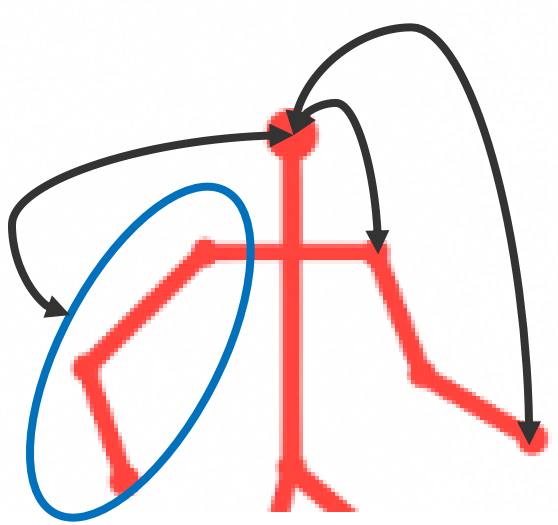
# リファインメントモデル：HyperSA [5]

## ハイパーグラフという概念を導入したSelf-Attention

- 通常のSelf-Attentionでは関節ペア同士の関係性しか見れなかったが、関節と関節グループ同士の関係性も考慮可能に



**Self-Attention**
1対1の関係のみ



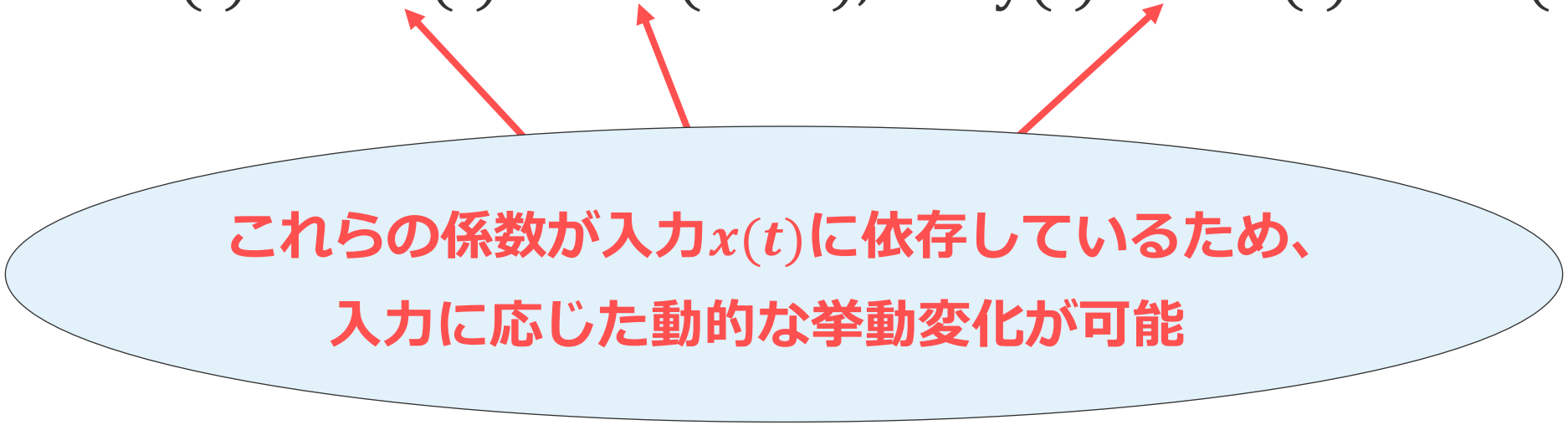
**HyperSA**
1対グループの関係も

[5] Yuxuan Zhou, et al. "Hypergraph Transformer for Skeleton-based Action Recognition." arXiv, 2022.

# リファインメントモデル：S6 [6]

**入力$x(t)$を隠れ状態$h(t)$を介して、出力$y(t)$に変換する**

$$h(t) = \bar{A}x(t) + \bar{B}h(t-1), \qquad y(t) = Ch(t) + Dx(t)$$




本研究ではS6の代表的なアーキテクチャであるMamba [6]から逐次処理に不向きな要素を一部削除した形式で利用

[6] Albert Gu, et al. "Mamba: Linear-Time Sequence Modeling with Selective State Spaces." arXiv, 2023.

# 実験

# 実験設定

## 単画像3次元関節座標推定モデル

- バックボーンにはConvNeXtPose-XSのバックボーンを使用
- グリッドごとに推定するサンプル数は$K = 100$に設定

## リファインメントモデル

- 時空間の処理数は$\mathrm{L} = 4$に設定

## 学習データ

- 入力は256x256、動画入力時は30フレーム（10fps）
- 1段階目はMPII [7] + Human 3.6M [8]、2段階目はHuman 3.6M

[7] Andriluka Mykhaylo, et al. "2D Human Pose Estimation: New Benchmark and State of the Art Analysis." CVPR, 2014.
[8] Ionescu Catalin, et al. "Human3.6m: Large scale datasets and predictive methods for 3d human sensing in natural environments." IEEE TPAMI, 2014.

# 評価項目

## 3次元姿勢推定精度

- Human 3.6M Protocol2により評価
- mean per-joint position error (MPJPE）を報告
  - 各関節がどの程度、GTとずれているのかを評価

## レイテンシ

- Samsung Galaxy S20のGPU上で計測
- LiteRT Benchmark Tools*を利用
- 平均と標準偏差を報告

* https://www.tensorflow.org/lite/performance/measurement

# 定量評価：モバイル端末向け軽量モデルとの比較

**既存手法よりもレイテンシを低下させながら、精度を改善**

| | MPJPE (mm)↓ | レイテンシ (ms)↓ |
|---|---|---|
| カメラからの距離にGTを利用する前提の場合 | | |
| ConvNeXtPose-XS | 56.6 | 25.99 ± 0.39 |
| ConvNeXtPose-M | 51.05 | 2104.88 ± 10.93 |
| ConvNeXtPose-L | **49.75** | 2151.13 ± 8.76 |
| Ours | 49.93 | **13.49 ± 0.02** |
| カメラからの距離もモデル自身で推定する前提の場合 | | |
| Ours | 50.39 | 14.57 ± 0.02 |

# 定量評価：モバイル端末向け軽量モデルとの比較

**既存手法よりもレイテンシを低下させながら、精度を改善**

| | MPJPE (mm)↓ | レイテンシ (ms)↓ |
|---|---|---|
| カメラからの距離にGTを利用する前提の場合 | | |
| ConvNeXtPose-XS | 56.6 | 25.99 ± 0.39 |
| ConvNeXtPose-M | 51.05 | 2104.88 ± 10.93 |
| ConvNeXtPose-L | **49.75** | 2151.13 ± 8.76 |
| Ours | 49.93 | **13.49 ± 0.02** |
| カメラからの距離もモデル自身で推定する前提の場合 | | |
| Ours | 50.39 | 14.57 ± 0.02 |

**-12%**

**-48%**

# 定量評価：モバイル端末向け軽量モデルとの比較

**既存手法よりもレイテンシを低下させながら、精度を改善**

| | MPJPE (mm)↓ | レイテンシ (ms)↓ |
|---|---|---|
| カメラからの距離にGTを利用する前提の場合 | | |
| ConvNeXtPose-XS | 56.6 | 25.99 ± 0.39 |
| ConvNeXtPose-M | 51.05 | 2104.88 ± 10.93 |
| ConvNeXtPose-L | **49.75** | 2151.13 ± 8.76 |
| Ours | 49.93 | **13.49 ± 0.02** |
| カメラからの距離もモデル自身で推定する前提の場合 | | |
| Ours | 50.39 | 14.57 ± 0.02 |

# 定量評価：時間軸考慮の有効性

**同レイテンシーまで拡大した1段階目モデルと比較しても、**
**5mm以上の精度改善を達成**

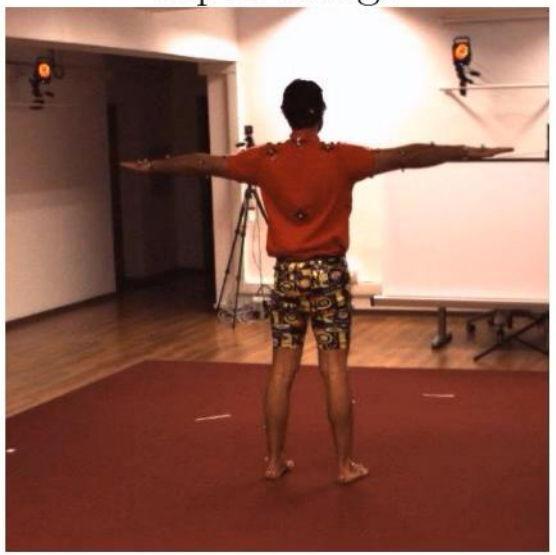
| | MPJPE (mm)↓ | レイテンシ (ms)↓ |
|---|---|---|
| 拡張1段階目モデル | 55.87 | 16.40 ± 0.06 |
| Ours | **50.39** | **14.57** ± **0.02** |

※カメラからの距離もモデル自身で推定する設定で評価

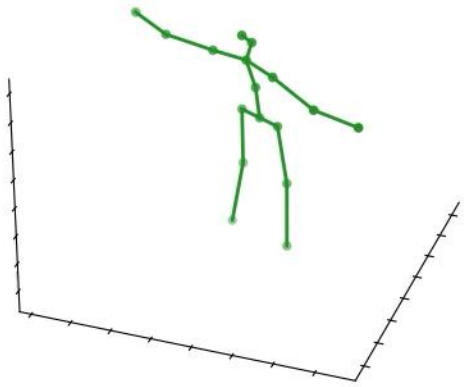
# 定性評価

**特に遮蔽が激しく、可視性が悪い環境での推定精度を改善**

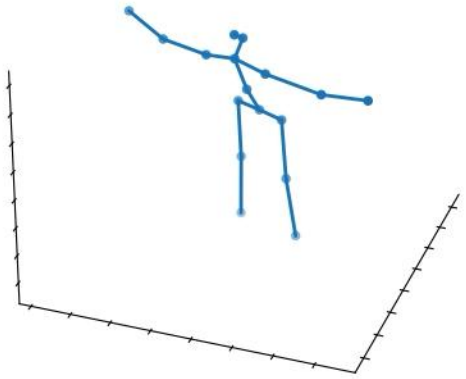
Input Image



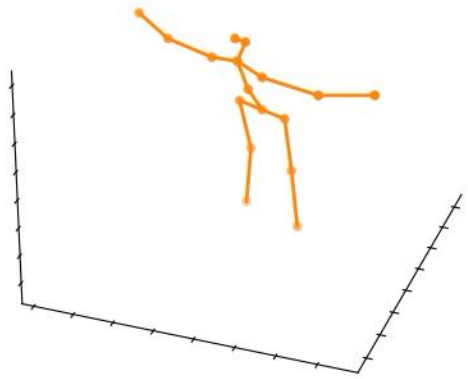
Ground Truth



ConvNeXtPose-XS



Ours



# まとめ

# まとめ

**高精度かつリアルタイム推論が可能な、GTに依存しない3次元人体姿勢推定モデルを提案**

- ヘッドの換装などによるフレーム処理の軽量化
- カメラとの距離の推定の導入によるGT依存の解消
- 逐次処理型のモデルによる効率的な時間軸情報の考慮

**今後の課題**

- ルート関節の絶対座標も含めた、絶対座標系でのリファインメントへの拡張