

# 修復 2D 偵測失效之姿態還原系統

## Pose Restoration System for 2D Detection Failures

指導老師：韓欽銓 教授

學生：吳云禾

國立聯合大學 資訊工程學系

苗栗市南勢里聯大 2 號

[cchan@nuu.edu.tw](mailto:cchan@nuu.edu.tw)

### 摘要

本研究針對 2D 姿態偵測器在遮擋與極端姿勢下產生之失準關節點，提出一套基於 SMPL-X 參數化人體模型之 3D 姿態還原系統。系統以單目影片為輸入，整合 2D 偵測、2D 至 3D 提升、姿態去噪與投影等模組，端到端產出 3D 姿態疊圖。去噪模型由圖卷積網路、Transformer 與 SMPL-X 解碼器構成，藉共用身材參數保證跨幀骨長一致。

**關鍵詞：**3D 人體姿態估計、SMPL-X、Transformer、時序去噪、深度學習

### Abstract

This study proposes a pose restoration system for 2D detection failures, aiming to recover inaccurate joint predictions caused by 2D pose detectors under occlusion and extreme poses. The system takes a monocular video as input and integrates 2D detection, 2D-to-3D lifting, pose denoising, and projection modules, producing 3D pose overlay videos end-to-end. The denoising network consists of a Graph Convolutional

Network for spatial features, a Transformer for temporal modeling, and an SMPL-X decoder; sharing body-shape parameters across the entire sequence guarantees consistent bone lengths across frames.

**Keywords:** 3D human pose estimation, SMPL-X, Transformer, Temporal denoising, Deep learning

### 一、前言

近年來，從單目影片估計 3D 人體姿態於動作捕捉、虛擬實境、運動分析等領域有廣泛應用。然而，現有 2D 姿態偵測器在面對遮擋、運動模糊或極端姿勢時，仍經常產生失準的關節點，這些誤差會直接傳遞到後續 3D 估計階段，導致整段影片出現「骨骼跳動」、「關節反折」等視覺異常。

本研究提出一套端到端 3D 姿態還原系統，以單目影片為輸入，整合 2D 姿態偵測、2D 至 3D 提升、時序去噪與投影等模組，輸出修正後之 3D 姿態。其中核心去噪模組將 2D 偵測失效之影響建模為 3D

姿態空間中之雜訊與遮罩，採用 SMPL-X 參數化人體模型作為解碼層，將輸出限制於合法的人體姿勢空間，藉此保證骨骼長度於整段序列中不變，並避免關節角度違反解剖學，期望解決下游 3D 姿態還原於真實場景中之穩定性與一致性問題。

## 二、系統架構與內容

### (一) 整體流程

本系統推論時依序由八個階段組成，整體流程如圖 1 所示。依序為：(1) 輸入影片；(2) 2D 姿態偵測：採用 MMPose 對每幀找出 17 個 COCO 標準關節之 2D 像素位置；(3) 2D 提升至 3D：以時序模型推斷各關節的深度資訊；(4) 座標系轉換；(5) 身材參數估計：以 SMPL-X 解碼器於整段序列預測共用身材參數；(6) 姿態去噪：對含雜訊之 3D 序列進行去噪，由 GCN、Transformer 與 SMPL-X 解碼器組成，逐幀預測姿態參數並沿用 (5) 之身材參數確保跨幀骨長一致，此階段為本研究之核心；(7) 投影回 2D；(8) 輸出疊圖影片。

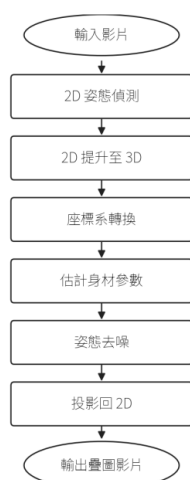


圖 1 專題架構圖

### (二) 去噪模型架構

姿態去噪階段由三個模組組成。空間編碼器使用圖卷積網路 (GCN) 學習關節間結構關係；時序編碼器使用 Transformer [2] 整合時序資訊；SMPL-X [1] 解碼器逐幀預測姿態參數與整段序列共用之身材參數，再透過 SMPL-X 之正向運動學運算，得到最終 17 個關節位置。

### (三) 訓練方法

訓練過程中，對乾淨之 3D 姿態套用四種隨機擾動策略以模擬偵測器失效之多樣型態：(一)「高斯雜訊」對全部關節加入標準差 5 公分之隨機位移、(二)「關節鏈遮罩」整條肢體一併遮蔽（模擬被物件擋住）、(三)「時序遮罩」連續多幀同時遮蔽（模擬整段偵測失敗）、(四)「片段遮罩」隨機散點遮蔽（模擬零星關節錯誤）。需特別說明，本研究採用之合成雜訊與真實偵測雜訊之分布存在差距，此差距亦為後續端到端實驗中 MPJPE 改善幅度受限之主因。訓練輸入示意如圖 2 所示。

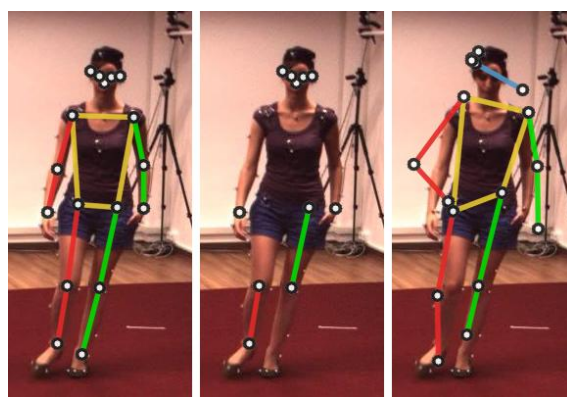


圖 2 訓練輸入示意：(左) 乾淨 GT、(中) GT 加隨機遮罩、(右) GT 加高斯雜訊

#### (四) 為何採用 SMPL-X

若直接迴歸 17 個關節之座標，逐幀獨立的輸出沒有任何結構機制強制跨幀的骨骼長度一致——同一人的手臂可能某幀 35 公分、下一幀 32 公分，造成視覺上明顯的骨頭抖動。改採 SMPL-X 後，整段序列共用同一組身材參數，骨頭長度於結構上不可能改變，此為本設計帶來的最關鍵效益。模型解碼後所得之人體網格如圖 3 所示。



圖 3 SMPL-X 解碼後重建之人體網格

### 三、系統實作成果

#### (一) 評估指標

本研究採用平均關節位置誤差作為評估指標。對每一幀的每一個關節，計算模型預測位置與標準答案位置之間的 3D 空間直線距離，再對全部 17 個關節與全部幀取平均，定義如式 (1)：

$$MPJPE = (1/TJ) \sum_{\{t,j\}} \left\| \hat{p} - p \right\|_2 \quad (1)$$

其中 T 為總幀數，J 為關節數（本研

究  $J = 17$ ）， $\hat{p}_{\{t,j\}}$  與  $p_{\{t,j\}}$  分別表示第 t 幀第 j 個關節之模型預測位置與標準答案位置； $\|\cdot\|_2$  為 3D 歐式範數，計算兩點間之直線距離。單位為公分，數字越小代表輸出越接近標準答案。

#### (二) 端到端實驗

對 H3WB [3] 中之走路驗證影片完整跑一遍流程（2D 偵測 → 3D 估計 → 本模型 → 投影），再以標準答案為基準計算 MPJPE 誤差，完整結果如表 1 所示，模型最終輸出投影回原始影片之效果如圖 4 所示。MPJPE 對絕對深度估計極為敏感，而本系統並未介入 2D→3D 之深度推斷階段；於真實偵測雜訊分布下達成此改善幅度，已可驗證去噪模組於真實場景之實用性。

表 1 端到端實驗之 MPJPE 結果

| 衡量指標  | 輸入      | 輸出      | 改善 |
|-------|---------|---------|----|
| MPJPE | 22.9 cm | 21.9 cm | 4% |

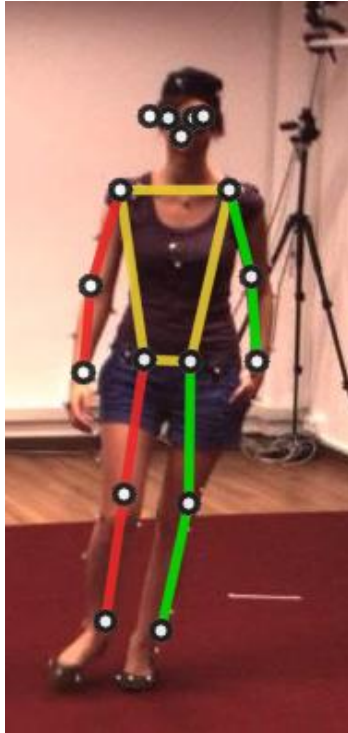


圖 4 模型輸出之 3D 姿態投影回原始影片

### (三) 網頁互動系統

為驗證本系統之實用性，本研究實作一套基於 Flask 後端與 Three.js 前端之網頁互動系統，使用者可透過瀏覽器即時操作。整體前後端架構如圖 5 所示，系統主介面如圖 6 所示。

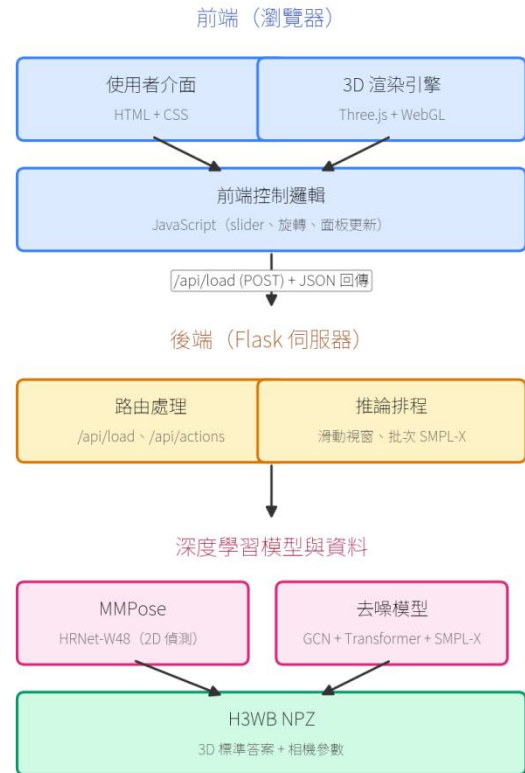


圖 5 網頁互動系統之前後端架構

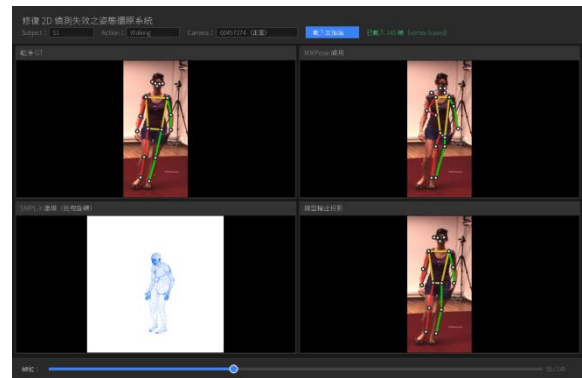


圖 6 網頁主介面：四面板比對與幀號滑桿

系統採用 WebGL 之 Three.js 函式庫實作 3D 視角控制：使用者可以滑鼠拖曳旋轉 SMPL-X 重建之人體網格，從任意角度檢視模型輸出之解剖細節。圖 7 為同一姿態於三個不同視角下之展示。



圖 7 SMPL-X 重建人體網格之三個視角

#### (四) 多動作展示

為驗證系統並非僅針對單一動作而設計，本研究於驗證資料中之多種動作上進行測試。圖 8 呈現三種代表性動作（走路、用餐、擺姿勢）之 MMPose 偵測結果與模型輸出投影之對照。可觀察到於不同姿態下，2D 偵測器於部分關節皆出現誤差，而本模型透過 SMPL-X 之結構先驗有效還原為合理之人體骨架。本章節之多動作測試以視覺定性比較為主，量化結果以走路動作之表 1 為代表。

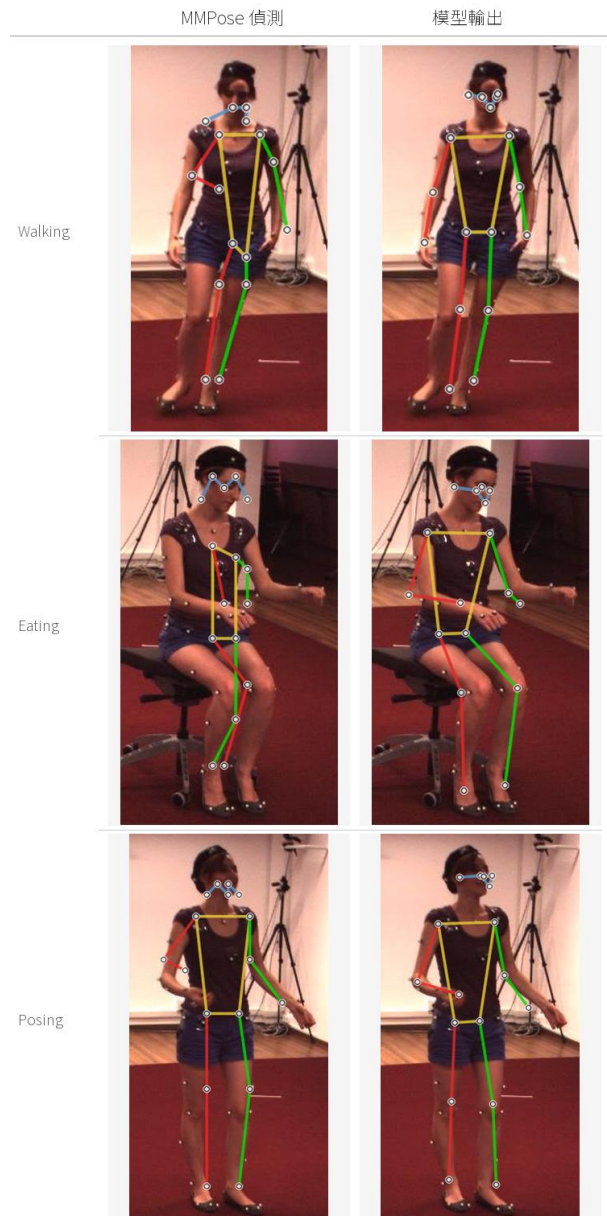


圖 8 多動作 demo：MPose vs 模型輸出  
(上：走路、中：用餐、下：擺姿勢)

#### 四、結論

本專題針對 2D 姿態偵測於遮擋與極端姿勢下之失準問題，提出基於 SMPL-X 之 3D 姿態還原系統，並完成端到端 pipeline 與網頁互動展示；端到端實驗驗證本系統可有效降低關節位置誤差並提升時序穩定度。本研究之已知限制有三：  
(一) 訓練資料以日常室內動作為主，當輸

入影片動作與訓練分布差異過大時，模型輸出表現會明顯衰減；(二) 平均關節位置誤差僅衡量點距，無法反映姿勢的視覺合理性，且 SMPL-X 之關節角度合法屬於有助於收斂之歸納偏置，並非數學硬約束，模型仍可能輸出視覺上不自然之姿勢；(三) 滑動視窗推論時之跨窗座標平均會略微違反 SMPL-X 之骨長保證。

未來工作可考慮：(一) 引入更廣泛之動作資料集（如 AMASS）以提升泛化能力；(二) 加入骨骼角度合理範圍與地面接觸約束，改善視覺品質；(三) 於姿態參數空間進行跨窗平均後再前向，以維持骨長嚴格一致。

## 五、參考文獻

- [1] G. Pavlakos, V. Choutas, N. Ghorbani, T. Bolkart, A. Osman, D. Tzionas, and M. J. Black, "Expressive body capture: 3D hands, face, and body from a single image", Proc. IEEE/CVF CVPR, pp.10975–10985, 2019.
- [2] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need", Advances in Neural Information Processing Systems (NeurIPS), Vol. 30, pp.5998–6008, 2017.
- [3] Y. Zhu, N. Samet, and D. Picard, "H3WB: Human3.6M 3D whole-body dataset and benchmark", Proc. IEEE/CVF Int. Conf. on Computer Vision (ICCV), pp.20166–20177, 2023.