Topic: 3D Human Pose Estimation on Monocular Video by Transformer
Group 10: r11944014 戴靖婷、r11922096 張家誠、r11725002 鈕愷夏

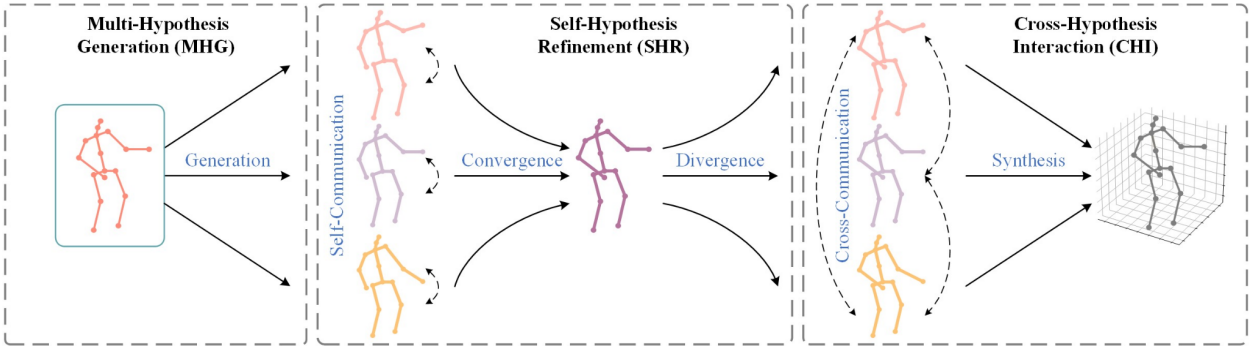
A. Introduction

透過單目相機生成的影片來估計 3D human poses 是一大挑戰，因為經常會有 depth ambiguity 以及 self-occlusion 的問題需要解決，透過 Transformer-based architecture 的架構更適合用於解決此類問題，同時處理 long-range dependency。

而我們的研究主要建立在 CVPR 2022 所提出來的 Multi-Hypothesis Transformer，透過更改 MHFormer 的架構及 loss function 以獲得更好的 3D human poses。

B. Methodology

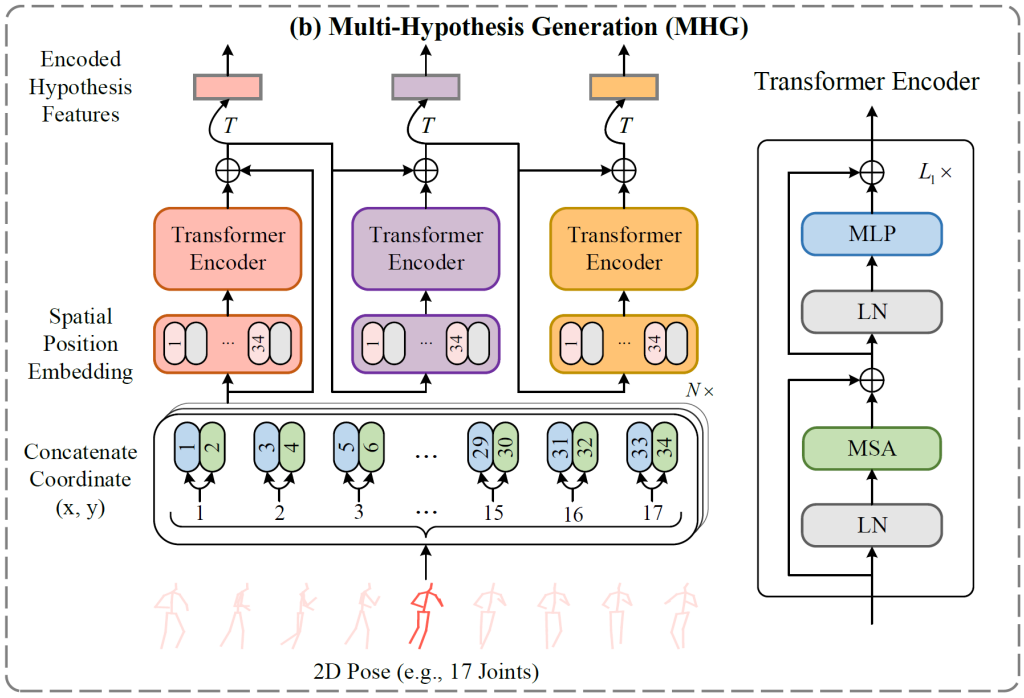
1. MHFormer



i. 輸入 2D pose sequence（由現成的 2D pose detector 而來）。

ii. 目標透過 spatial and temporal information in the multi-hypothesis feature，重建出當前 frame 的 3D poses。

iii. **Multi-Hypothesis Generation**



1. 在空間域中透過 cascaded Transformer-based architecture 生成不同深度的多種特徵。用於 model human joint relations 以及初始化 multi-hypothesis representations。同時應用 skip residual feature
2. input: 2D pose sequence / output: multiple hypotheses
3. output of MHG 包含了多種語意資訊的多層特徵，可以被視為不同 pose hypothese 的 init。

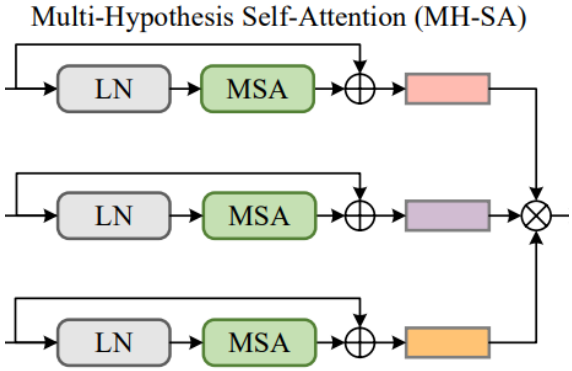
iv. Temporal Embedding

1. 得到 pose hypothese initation 後進一步建立跨 hypothesis 的 features，方便後續的方法（SHR、CHI）來捕捉時間依賴性。
2. 首先把 **spatial domain** 轉換至 **temporal domain**。加上可學習的 **temporal position embedding** 保存不同時間 frames。

v. **Self-Hypothesis Refinement**

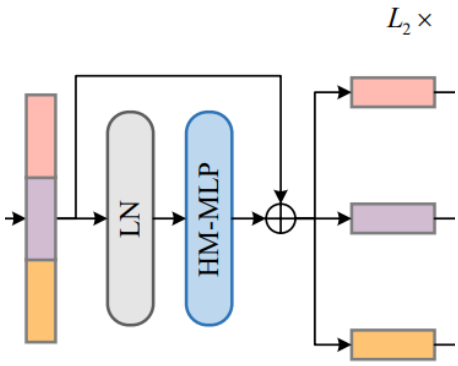
1. 轉變到時間域後，建構 SHR 來提煉 single-hypothesis features。
    a. 每一層的 SHR 是由一個 multi-hypothesis self-attention **(MH-SA) block**，以及 hypothesis-mixing **MLP block** 組成。

2. **multi-hypothesis self-attention (MH-SA) block**



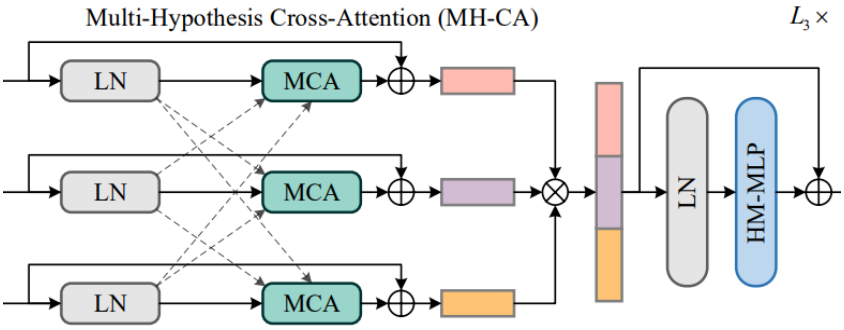
   a. MH-SA transformer model 的核心是 **MSA (Multi-head transformer)**，目標在於對每個假設獨立捕捉 single-hypothesis dependencies 以進行 communication。
   b. 最終 **self-hypothesis** 傳遞 **different hypothesis features** 訊息來達到 **feature enhancement**。

3. Hypothesis-Mixing MLP



   a. 前面所說的多種 hypotheses 在 MH-SA 中獨立處理，沒有做到跨 hypotheses 的資訊交換，因此我們進一步連接 enhanced features of multiple hypotheses 再丟入 hypothesis-mixing MLP 進行資料交換。

vi. **Cross-Hypothesis Interaction**



1. 透過 CHI 建立 interactions among multi-hypothesis features，CHI 由兩個 block 組成（multi-hypothesis cross-attention (MH-CA) and hypothesis-mixing MLP）。
    a. MH-CA
        i. MH-SA 缺少跨 hypotheses 的 connection，因此這邊利用 MHCA 用於捕捉 multihypothesis correlations，由多個 multi-head cross-attention (MCA) elements 平行組成。
    b. Hypothesis-Mixing MLP
        i. 用於聚合 features of all hypotheses 來生成 single hypothesis representation 。

vii. Regression Head
1. linear transformation layer 應用再最終輸出透過回歸生成唯一的 3D pose。

viii. Loss Function
1. Mean Squared Error:

$$\mathcal{L} = \sum_{n=1}^{N} \sum_{i=1}^{J} \left\| Y_i^n - \widetilde{X}_i^n \right\|_2$$

2. **Our changes in MHFormer**:
    a. Model architecture:
        i. 經過觀察所有模型架構之後，發現 SHR block 為兩層，其中第一層的重點在於特徵強化，而第二層主要是做資訊交換。
        ii. 第一層的結構和 MHFormer 中的第一步 MHG 十分相像，且第二層的資訊交換與最後一步的 CHI 的 Cross attention 概念類似，因此我們決定更改 MHG 和 CHI 的架構與層數來代替移除的 SHR block。

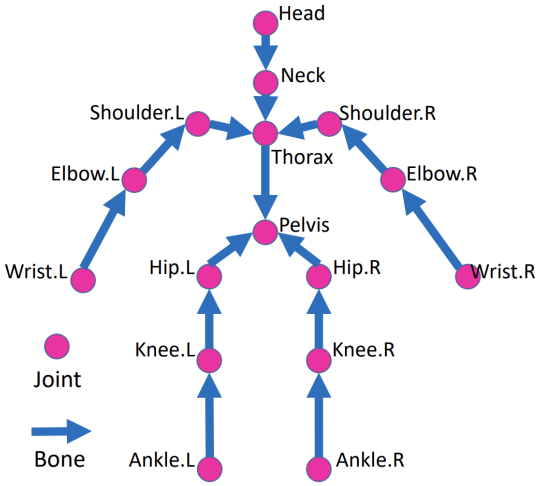
b. Loss function:

i. $$\mathcal{L} = \sum_{n=1}^{N} \sum_{i=1}^{J} \left\| Y_i^n - \widetilde{X}_i^n \right\|$$ norm = 1, 2

ii. Bone Loss [2]:

1. Bone definition: 考量連接關節之間的關係；以骨盆為 root，第 k 個 bone 的定義是第 k 個關節的 parent joint 減掉第 k 個關節。

$$\mathbf{B}_k = \mathbf{J}_{parent(k)} - \mathbf{J}_k$$



2. **Joint Pair Loss**:

$$\Delta \mathbf{J}_{u,v} = \sum_{m=1}^{M-1} \mathbf{J}_{I(m+1)} - \mathbf{J}_{I(m)}$$
$$= \sum_{m=1}^{M-1} sgn(parent(I(m)), I(m+1)) \cdot N^{-1}(\tilde{\mathbf{B}}_{I(m)})$$

依序從第 u 個關節到第 v 個關節，加總路徑上所有經過的關節與前一個關節之間的 bone loss，進而計算出 u 和 v 之間的 joint pair loss。而使用這個方法的目的是要盡可能地消除累積誤差。

C. Encountered Problem

1. 原本想要透過 3D human pose estimation 去估測行人是否要穿越馬路，然而行人穿越馬路僅有少量的 github 開源代碼，且難以重新設計 human poses 當作 model

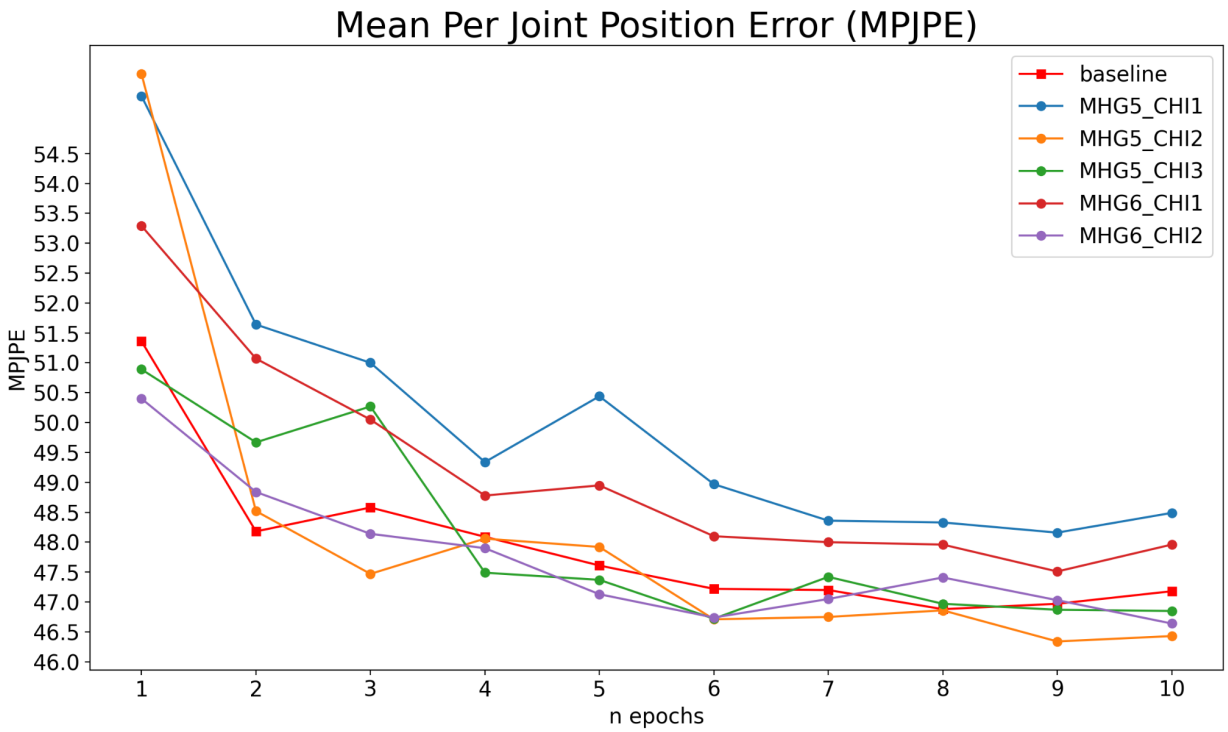
的 input 去重新訓練，因此我們最後只好限縮範圍在改良 3D human pose estimation 上面。

2. 由於我們的 GPU 資源不夠充足（所有結果都是由一個 3090 所得到的），因此我們透過把 reception filed 設小一些、epochs 的數量也調到 10，將每一次的實驗都壓縮到五個小時左右，而最後仍然有許多想要嘗試的組合沒辦法嘗試。
3. 想要將其他篇論文 [3] 的理念應用在這篇論文上，但架構銜接比較複雜，並未在期限內成功完成。

D. Experiments

1. Settings:
    a. Evaluation metric: Mean per joint position error (MPJPE)
    b. Dataset: Human 3.6M
    c. Parameters: 為了在計算效率和成果間取平衡。
        i. Receptive Field: 27 frames
        ii. Number of epochs: 10 epochs
        iii. Batch size: 256

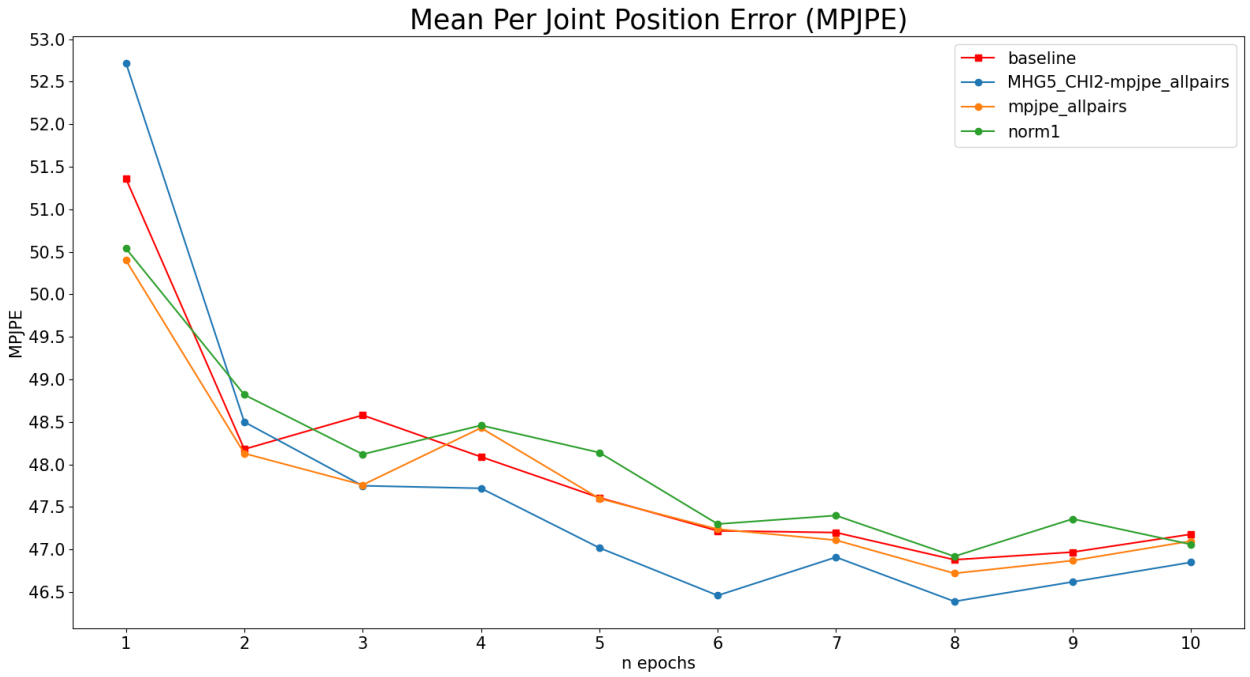
2. Model architecture:



	MPJPE (mm)	*Param (M)*	*Training time (mins per epoch)*
MHFormer (baseline)	46.88	18.92	22
5-layer MHG + 1-layer CHI	48.15	**6.54**	**13.5**
6-layer MHG + 1-layer CHI	47.35	6.56	15
6-layer MHG + 2-layer CHI	46.64	12.88	20
5-layer MHG + 3-layer CHI	46.85	19.17	25
5-layer MHG + 2-layer CHI	**46.33**	12.86	19

a. 第一個是原本的 MHFormer，下面全部都是我們改過的 model 所得到的結果，可以看到移除 SHR block 之後，越少層的 MHG 使用的參數量越少，training 的時間也明顯縮短許多。
b. 而 MPJPE 的部分是 5 層 MHG 再加上 2 層 CHI 最優，而參數的使用量也比原本的 MHFormer 還要更少。因此，我們最後以 5-layer MHG + 2-layer CHI 為主要模型架構。
c. 該論文原本是以 4 層 MHG、2 層 SHR 和 1 層 CHI 組成 MHFormer，我們移除了中間的那 2 層 SHR，分別由 MHG 和 CHI 取代，確實也是這樣的搭配獲得了最好的結果，因此最終的實驗結果符合我們的預期，

3. Loss function:



	MPJPE (mm)	*Training time (mins per epoch)*
MHFormer (baseline)	46.88	**22**
norm 1	46.92	**22**
mpjpe with allpairs	<u>46.72</u>	25
5-layer MHG + 2-layer CHI / mpjpe with allpairs	**46.39**	**22**

a. 首先使用 norm 1，MPJPE 與原本的 norm 2 沒有差非常多。
b. 原本預期加上 All joint pair loss 之後總體的 MPJPE 會大幅下降，但最後結果只有些許的變好（從 46.88 下降至 46.72），然而速度的部分反而因為計算稍微變複雜而變慢了一些。

c. 使用上面較佳的模型搭配上新的 loss function 依然可以獲得相對不錯的結果，但與單純使用原本的 loss function 相去不遠，因此我們最後是使用改動過後的模型作為我們的 best model。

4. Our final model: 5-layer MHG + 2-layer CHI

E. Discussion

1. 觀察 SHR block 的架構後，將 SHR 移除並改變 MHG 和 CHI 的架構和層數以還原 SHR 的功能，實驗證明能大幅減少參數使用量和訓練時間。除此之外，我們也實驗了繼續增加 MHG 和 CHI 的層數，發現剛好多加一層的 MHG 和多加一層的 CHI 正好能取代原本的兩層 SHR 功能，獲得最好的結果。
2. 另外，透過使用 All joint pair loss 當作 loss function，可以稍微提升 MHFormer 在 3D human poses 的結果，然而效果不甚顯著且需要多花費一些額外的訓練時間，因此我們最後並未使用新的 loss function 去做比較。
3. 在另一篇論文中提出了 strided transformer [3]，此篇論文提及一般 attention block 在 attention 結束後都會接上連接層，而 strided transformer 將其換成 CNN 用來聚合結果，效果也非常不錯，且可以降低計算時間，或許可以用來將 CHI 最後一步換成 strided transformer 的架構來減少參數量並提升精度。

F. Division of work

r11944014 戴靖婷:
paper survey, loss function and result plotting implementation, experiment, presentation

r11922096 張家誠:
paper survey, model architecture and loss function implementation, experiment

11725002 鈕愷夏:
paper survey, idea of improving model architecture

G. Reference

[1] Wenhao Li, Hong Liu, Hao Tang, Pichao Wang, Luc Van Gool. “MHFORMER: Multi-hypothesis transformer for 3D human pose estimation. “ In *Computer Vision and Pattern Recognition (CVPR 2022)*.

[2] Xiao Sun, Jiaxiang Shang , Shuang Liang , Yichen Wei. “Compositional human pose regression. “ In the *International Conference on Computer Vision (ICCV 2017)*.

[3] Wenhao Li, Hong Liu, Runwei Ding, Mengyuan Liu, Pichao Wang, Wenming Yang. “Exploiting Temporal Contexts with Strided Transformer for 3D Human Pose Estimation. " In *IEEE International Conference on Multimedia and Expo Workshops (ICMEW 2022)*.