College of Engineering (CoE)
National Chung Cheng University

Report of 2025 CoE Intern Research Program

Student	Bandithvipho LANG	Faculty Mentor	Wen-Nung Lie
Department		Electrical Engineering	
Research Period 2025 03 03 to 2025 06 30		Date of Report Submission 2025 06 25	
Report Title	MHFusionNet: Multiple Hypotheses Fusion-Based Approach For 3D Human Pose Estimation		
Highlights of Report			
Signature of Student		Date	2025/06/25
Mentor Review Comment	□ Excellent □ Good □ Average □ Poor		
Signature of Mentor		Date	

College of Engineering

National Chung Cheng University

MHFusionNet: Multiple Hypotheses Fusion-Based Approach For 3D Human Pose Estimation

Student Bandithvipho LANG

Faculty Mentor Wen-Nung Lie

06/25, 2025

MHFusionNet: Multiple Hypotheses Fusion-Based Approach For 3D Human Pose Estimation

Bandithvipho LANG
Department of Electrical Engineering
National Chung Cheng University, Chia-Yi 621, Taiwan, R.O.C.

Abstract

In this project, presents the proposed Multiple Hypotheses Fusion-Based Approach known as a fusion-based for multiple possible 3D human pose hypotheses estimation. This project contributes a novel framework that addresses ambiguity and occluding problems in 3D human pose estimation. Most of the State-of-the-Art (SOTA) developed to the missing depth ambiguity and occlude but there are still limitations such as the ManiPose [3] produce the multiple hypotheses and used averaging to compute for final 3D human pose which makes the final 3D pose uncertainty and unreliable. Recently, D3DP [1] proposed the Joint-wise reprojection-based Multi-hypothesis Aggregation (JPMA) for probabilistic 3D human pose estimation by using diffusion-based which achieves exceptional performance. The proposed JPMA conducts joint-level aggregation based on reprojection errors by relying on intrinsic camera parameters for projecting 3D pose to the camera plane as a 2D coordinate (u, v), which is incompatible with real-world applications. Our proposed MHFusionNet, designed camera-parameter-free approach which is composed of two stages. For the first stage, leverages a pre-trained multiple hypotheses model to generate multiple 3D human pose. Second stage, the fusion network was designed based on two strategies feature fusion (FF) and early fusion (EF) techniques. This approach advances upon prior state-of-the-art (SOTA) methods by modeling uncertainty more effectively, rather than relying on simple assumptions like averaging.

1. Introduction

3D Human Pose Estimation (3D HPE) aims to localize joints and build a body representation (skeleton position) from input data such as RGB images and videos. The goal is to regress the 3D joint's locations of a human in the 3D space using the input of 2D pose. Human Pose Estimation (HPE) provides geometric and motion information of the human body and can be applied to a wide range of applications such as video animation, human-computer interactions (HCI) [5], action recognition [6], health care of elderly patients [7], gesture recognition [8], and video surveillance [9]. Generally,

the mainstream approach is to conduct 3D pose estimation in two stages, for the first stage the 2D pose is first obtained with a 2D pose detector which is estimating the 2D locations of human joints from RGB images. and then the second stage is performed to map these 2D locations to their corresponding 3D positions. In this work, we focus on the second stage, also known as the 2D-to-3D lifting process, inspiring from the recent state-of-the-art approaches [1, 2, 3, 4, 10, 11].

The 2D-to-3D lifting from monocular videos is an inverse problem, where multiple feasible solutions (i.e., hypotheses) exist due to its ill-posed nature given the missing depth ambiguity. Those approaches [10, 11] ignore this problem and only estimate a single solution (single hypothesis), which often leads to unsatisfactory results, especially when the person is severely occluded (right arm and elbow).

Since those multiple hypotheses methods developed to the missing depth ambiguity and occluded but there are still limitations such as the ManiPose [3] they produce the multiple hypotheses and used averaging to compute for final 3D human pose which is made the final 3D pose uncertainty and unreliable. Recently, D3DP [1] proposed the Joint-wise reprojection-based Multi-hypothesis Aggregation (JPMA) for probabilistic 3D human pose estimation by using diffusion-based which achieves exceptional performance. The proposed JPMA conducts joint-level aggregation based on reprojection errors by relying on intrinsic camera parameters for projecting 3D pose to the camera plane as a 2D coordinate (u, v), which is incompatible with real-world applications.

To address the problem that mentioned above, we developed the camera-parameter-free approach as they are more practical in real applications, this work proposed a fusion-based framework for 3D human pose estimation that integrates multiple hypotheses generated using diffusion-based models to improve the accuracy and robustness of final 3D skeleton predictions.

2. Proposed Method

Many existing Multiple Hypotheses approaches [1, 2, 3, 12, 13, 14] rely on simple averaging to calculate the 3D human pose from multiple predicted poses. However, averaging can limit the model's ability to fully capture the uncertainty and diversity in human pose predictions. In this work, we propose a more effective strategy by designing a dedicated fusion network, called MHFusionNet, that intelligently selects and integrates the most plausible information from multiple hypotheses to predict a more accurate final 3D human pose.

Our approach takes advantage of FusionFormer [2], While FusionFormer focuses on multiple views, our proposed MHFusionNet adapts its fusion mechanism to a different context by fusing multiple hypotheses generated by (SOTA) method.

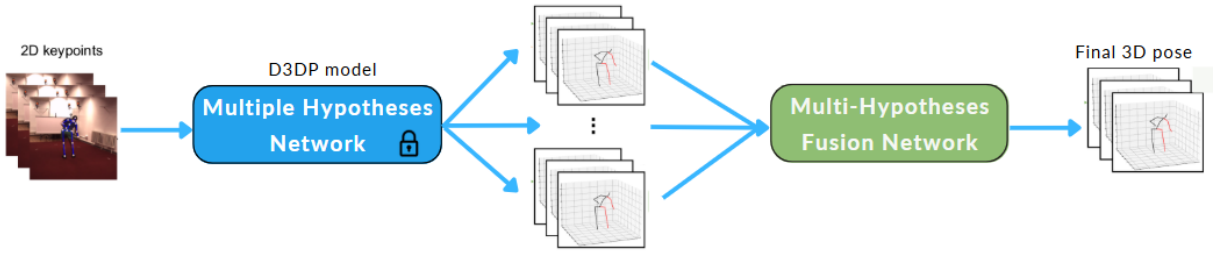

Figure 2.1 The proposed MHFusionNet Method

An overview of the proposed MHFusionNet is illustrated in **Figure 2.1**, the architecture consists of two stages: the Multiple Hypotheses Network and the Multiple Hypotheses Fusion Network.

In the first stage, we adopt the D3DP model as the baseline to generate multiple plausible 3D human pose hypotheses from a 2D input. These diverse hypotheses are then passed to the second stage, the Fusion Network (FN) which is specifically trained to identify and synthesize the most accurate final 3D human pose from the given set of the hypotheses.

The proposed MHFusionNet leverages a pre-trained multi-hypotheses model in the first stage to generate multiple 3D human pose. In this second stage, the FN was designed based on two strategies Feature Fusion (FF) and Early Fusion (EF) techniques. This approach advances upon prior state-of-the-art (SOTA) methods by modeling uncertainty more effectively, rather than relying on simple assumptions like averaging.

2.1. Multiple Hypotheses Generator

To generate diverse and plausible 3D human pose predictions, we utilize the D3DP model [1], a diffusion-based framework designed for 3D pose estimation. Unlike traditional models that produce a single deterministic output, D3DP can generate multiple hypotheses that reflect the inherent uncertainty and ambiguity in 2D-to-3D pose lifting, especially in cases of occlusion or visually similar joint configurations.

Diffusion-based 3D Human Pose Estimation is the method Diffusion-based with Joint-wise reprojection-based Multi-hypothesis Aggregation (JPMA). On the other hand, D3DP generates multiple possible 3D pose hypotheses for a single 2D observation. It gradually diffuses the ground truth 3D poses to a random distribution and learns a denoiser conditioned on 2D keypoints to recover the uncontaminated 3D poses. The proposed D3DP is compatible with existing 3D pose estimator. JPMA is proposed to assemble multiple hypotheses generated by D3DP into a single 3D pose for practical use. It reprojects 3D pose hypotheses to the 2D camera plane, selects the best hypotheses to the 2D camera plane, selects the best hypothesis joint-by-joint based on the reprojection errors, and combines the selected joints into the final pose. The

proposed JPMA conduct aggregation at the joint level and makes use of the 2D prior information, both of which have been overlooked by previous approaches. D3DP used a mixed spatial temporal Transformer-based method, as the backbone.

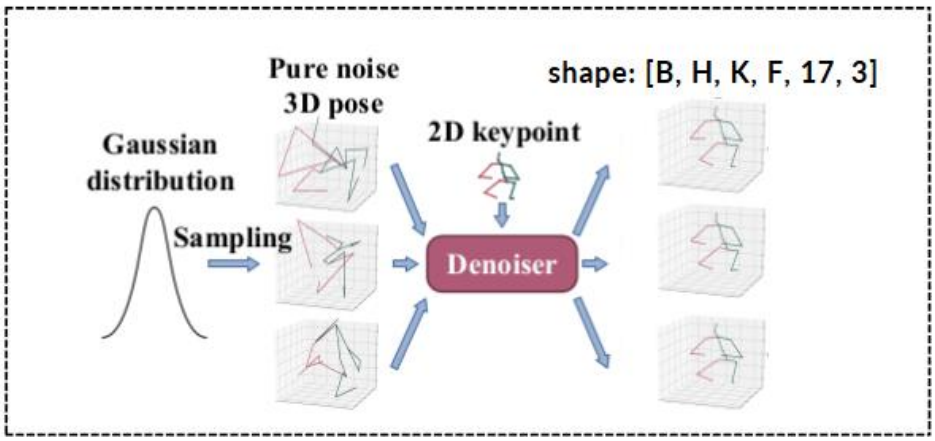

Figure 2.2 Multiple Hypotheses Generated [**1**]

The D3DP model uses a denoising diffusion process to generate multiple diverse 3D human pose hypotheses from a Gaussian distribution. The main idea is to start with random noise and progressively denoise it using a trained model conditioned on 2D keypoints. As we can see in **Figure 2.2**, D3DP begins by sampling the noise vectors from a standard Gaussian distribution. These noisy vectors are treated as corrupted 3D poses and are fed into a denoiser D, which is conditioned on the 2D keypoints x and timestep t. The denoiser learns to reconstruct the clean 3D pose hypotheses $\tilde{y}_0$, this process can be expressed as:

$$\tilde{y}_0 = D(y_t, x, t) \quad (1)$$

Where other processes will follow by formula **(Eq. 2, 3, 4, 5)** in D3DP [1], to generate multiple hypotheses, we repeat the sampling process H times from a Gaussian distribution $\mathcal{N}(0, I)$, giving us:

$$\tilde{Y} = \{\tilde{y}_0^{(1)}, \tilde{y}_0^{(2)}, \dots, \tilde{y}_0^{(H)}\} \quad (2)$$

Each hypothesis $\tilde{y}_0^{(0)} \in \mathbb{R}^{J\times 3}$, where J is the number of joints. Optionally, this process can be refined over K iteration (K: number of samplings timestep) using the DDIM strategy [15], which improves the accuracy of the generated hypotheses.

2.2. Feature Fusion

Feature Fusion is fusing the individual feature representations from each hypothesis, by taking the 3D input and mapping them into a high-dimensional space to obtain the feature.

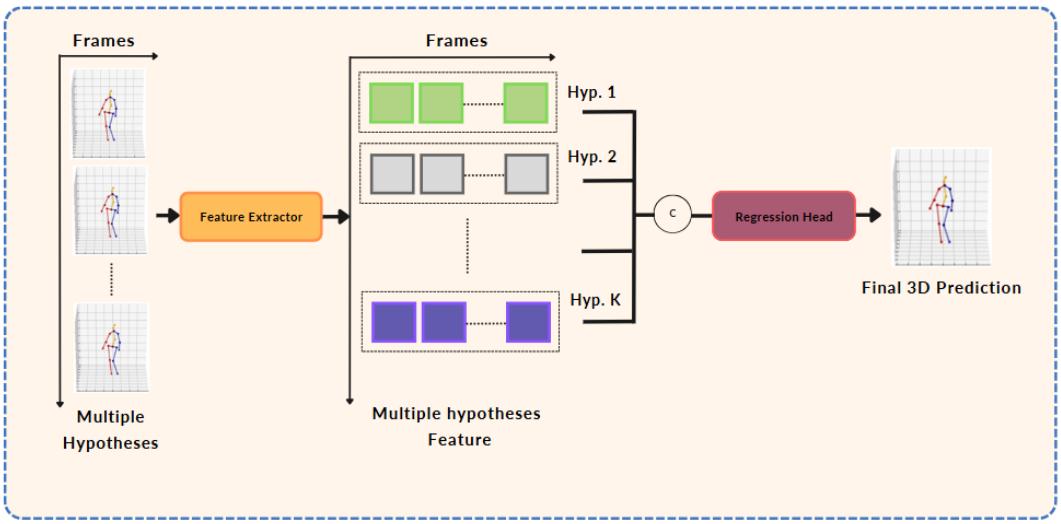

Figure 2.3 Overview of Feature Fusion (FF)

As shown in **Figure 2.3**, the Feature Fusion consists of two modules, Feature Extractor and Regression Head. The process begins with the multiple 3D human pose as the input passed through the Feature Extractor that converts the raw data from the 3D pose into high dimensional feature. These features encode spatial-temporal information of the body joints and the dynamics across frames.

2.2.1. Feature Extractor

Feature Fusion extracts high-dimensional features representations from each hypothesis independently, then integrates them through fusion mechanism. By motivating from the FusionFormer [3], in the Feature Extractor We adopt PoseFormer [13] as the feature extractor in the main experiments.

Our Feature extractor is built upon a Poseformer-based architecture, which was originally designed to operate on 2D input keypoints in spatial-temporal. In our MHFusionNet, we modified the Poseformer [13] to accept 3D input, allowing it to extract high-dimensional feature embeddings from each individual 3D pose hypothesis. The overall procedure of our feature extractor is illustrated in **Figure 2.4**.

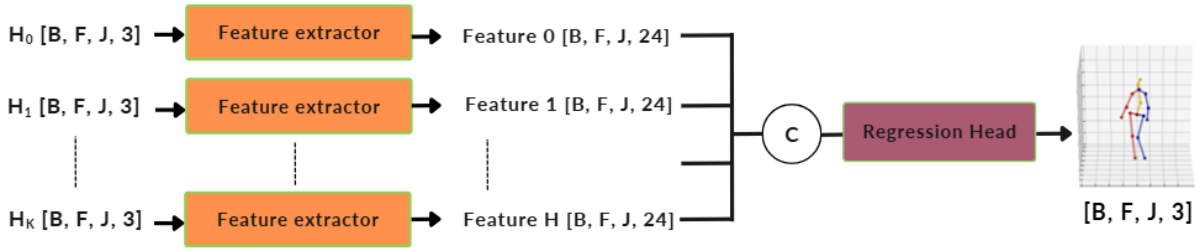

Figure 2.4 Overview of Feature Extractor

After obtaining the set of multiple 3D human pose from the pre-trained model D3DP model, we have an input tensor H_i of shape [B, F, J, 3], where 3 denotes the (x, y, z) coordinates:

$$F_{3D}: H_i \in \mathbb{R}^{B\times F\times J\times 3} \quad (3)$$

Each input H_i contains H hypotheses across F frames, yielding an overall set, that denotes the estimation result as $\mathcal{P}_{3D}$:

$$\mathcal{P}_{3D} = \mathcal{F}_{3D}(I) \in \mathbb{R}^{B\times F\times H\times J\times 3} \quad (4)$$

The D3DP model takes an input I (2D observations), $\mathcal{F}_{3D}(I)$ denoted the D3DP model's inference yielding H hypotheses of 3D pose across F frames as shown in **(Eq. 4)**. At this point, $\mathcal{P}_{3D}$ contains a set of hypotheses 3D poses for each frane across the batch represented by B.

To extract high-dimensional feature embeddings, $\mathcal{P}_{3D}$ is passed through the Feature Extractor, which projects each 3D point into latent space of high dimensionality. This can be represented:

$$\mathcal{F}_{embed} = Embed(\mathcal{P}_{3D})\mathbb{R}^{B\times F\times H\times J\times C_H} \quad (5)$$

Where C_H is the number of channels of each keypoint. Subsequently, the feature extractor employs several layers to extract the relationship between keypoint. Each joint of every hypothesis and frame is embedded into this latent space, allowing the model to learn spatial and temporal relationships across the pose sequence.

To embedded features F_{embed} are further processed by the Pose Feature Encoder E_{pose}, which applied a series of attention mechanisms as in Poseformer-based, extracting the relationship between joints, frame, and hypotheses:

$$\mathcal{F}^0_{pose} = E_{pose}(F_{embed}) \in \mathbb{R}^{B\times F\times H\times J\times C_H} \quad (6)$$

Where:

- $\mathcal{P}_{3D}$: Output of the D3DP model across F frames and H hypotheses
- F_{embed}: High-dimensional embeddings of the 3D input $\mathcal{P}_{3D}$
- F_{pose}: Final feature representation learned by the pose feature encoder

F_{pose}, capture both the spatial structure within each hypothesis and the temporal consistency across frames. They are then passed to the regression head to produce the final refined 3D pose prediction.

2.2.2. Regression Head

To maximize the capacity of our fusion network, we employ a simple 3D pose regression head to map the fused hypotheses features into final 3D human pose predictions. Once we have the encoded feature map F_{pose} from the pose feature encoder:

$$F_{pose} \in \mathbb{R}^{B\times F\times H\times J\times C_H} \quad (7)$$

The goal of the regression head is to aggregate information across the H hypotheses and regress to a single final 3D human pose prediction. Before we pass the F_{pose} to the regression head, first we concatenate the feature embeddings across the hypothesis dimension H. This creates a combined feature for each joint across all hypotheses:

$$F_{concat} = concat(F_{pose}) \in \mathbb{R}^{B\times F\times J\times (H\times C_H)} \quad (8)$$

As we can see in **(Eq. 8)** above, we concatenated the F_{pose} across hypotheses on C_H channels. This concatenation allows the model to access information from all hypotheses simultaneously, making it possible to learn inter-hypotheses relationships and select the best parts of each hypothesis.

After obtaining F_{concat}, we pass it to the Regression Head (RH) to produce the final 3D human pose prediction. The regression head (RH) acts as a mapping network that takes the combined feature space and learns to regress it to valid 3D pose:

$$\tilde{\mathcal{P}}_{3D} = \mathcal{R}_\theta(F_{concat}) \in \mathbb{R}^{B\times F\times J\times 3} \quad (9)$$

Where $\mathcal{R}_\theta(.)$ is the regression head implemented with learnable parameters θ and $\tilde{\mathcal{P}}_{3D}$ is the Final 3D human pose prediction, matching the ground truth shape $(B, F, J, 3)$.

To investigate the best way to map from the high-dimensional F_{concat} space to the final 3D pose, we implemented and experimented with different network structures for our regression head, including MLP, ResidualFC, and DenseFC. Each approach operates on the input feature of shape $(H \times C)$. The details of each design below.

a). Multi-layer Perceptron (MLP)

MLP is used as a neural network to model the relationship between input features and a continuous output variable. MLP is a feadforward network comprised of a sequence of Linear and activation functions. In this network, we constructed three layers by using Linear module and ReLU activation functions as shown in **Figure 2.5**.

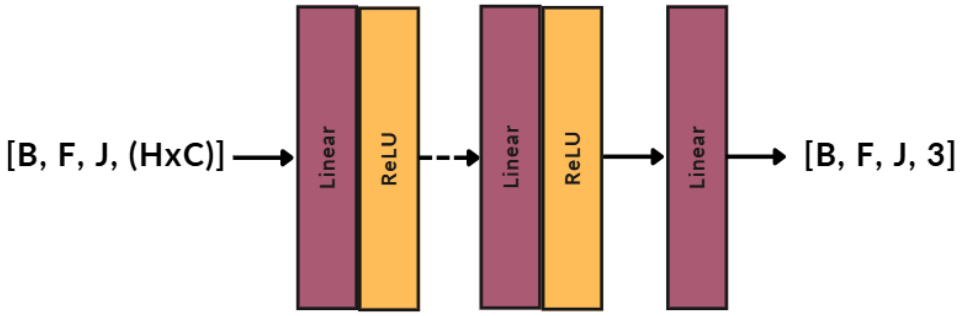

Figure 2.5 MLP Network

Once F_{concat} is computed, it is passed to a Regression Head $\mathcal{R}_\theta(.)$, which transforms the feature embeddings into the final 3D pose prediction. As illustrated in **Figure 2.5**, since we concatenate hypotheses feature on C-dimensional we get the input feature $(H \times C)$ that first pass through a Linear Layer that maps it into a high-dimensional, denoted as:

$$\tilde{\mathcal{P}}_{3D} = Linear(ReLU(Linear(F_{concat}))) \in \mathbb{R}^{B\times F\times J\times 3} \quad (10)$$

The final Linear Layer maps the feature down from high-dimensional to the desired output $\tilde{\mathcal{P}}_{3D} \in \mathbb{R}^{B\times F\times J\times 3}$. This design provides a simple regression head that captures the global relationship across hypotheses and joint features.

b). Residual Fully Connected Layer (ResidualFC)

The ResidualFC block is designed to process the combined input features of $(H \times C)$ and regress it into the final 3D human pose $[B, F, J, 3]$. Its design is inspired by the concept of residual connections, allowing the network to learn deeper features representation while preserving information flow across layers.

In **Figure 2.6**, illustrates the process of the residualFC block used in our Fusion Network. The process can be broken down in the description below.

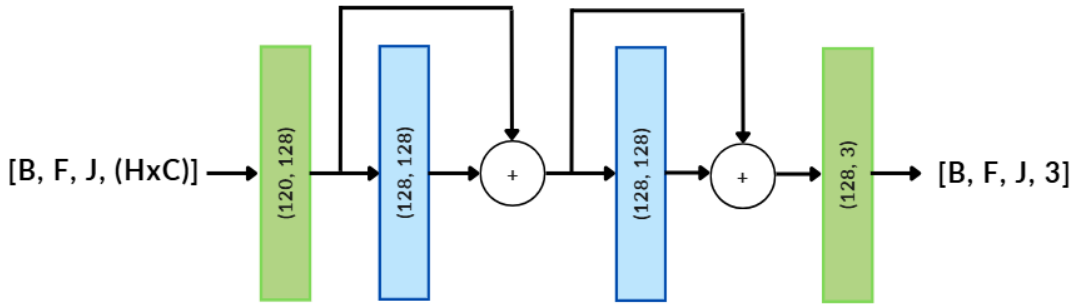

Figure 2.6 ResidualFC Network

After we obtained F_{concat} the output from the feature extractor $(H \times C)$ first passed through the Linear module or MLP block as shown in **Figure 2.6,** this maps the input into higher-dimensional latent space, which is:

$$F_0 = W_{in}(H \times C) \in \mathbb{R}^{120\times 128} \quad (11)$$

Each residual block comprises a linear module followed by Layer Normalization and a LeakyReLU activation. A dropout layer also includes regularization. The output of these layers, $g(F_i)$, is then added to the input F_i following the residual concept, which denoted as:

$$F_{i+1} = F_i + g(F_i) \in \mathbb{R}^{120\times 128} \quad (12)$$

After passing through N_{stage} residual blocks, the final feature F_{final} is connected to the final linear module (MLP) to predict the final 3D human pose $[B, F, J, 3]$.

$$\tilde{\mathcal{P}}_{3D} = W_{out}(F_{Final}) \in \mathbb{R}^{120\times 3} \quad (13)$$

This design allows the Fusion Network to encode feature relationships across hypotheses and be able to capture intricate spatial and temporal dynamics across multiple hypotheses, making it highly robust for 3D human pose estimation.

c). Dense Fully Connected Layer (DenseFC)

In our denseFC network in **Figure 2.7**, we used with two numbers (Blue Block), each layer progressively refines the features, capturing intricate relationships between joints while mitigating overfitting.

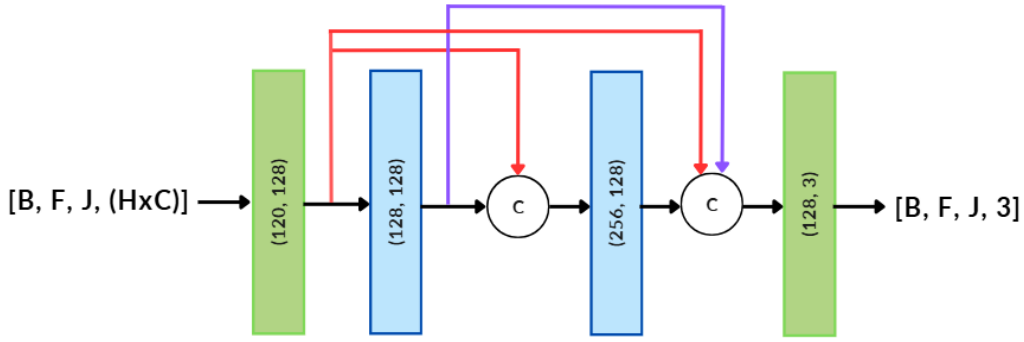

Figure 2.7 DenseFC Network

Dense connections ensure efficient feature reuse and mitigate vanishing gradient issues, facilitating deeper learning. Let begin with the $(H \times C)$ passed through the Linear module, which maps the data into a higher-dimensional feature of shape (120, 128). This projection allows the network to encode the input in a higher-dimensional:

$$X_0 = W_o(H \times C) \Longrightarrow \mathbb{R}^{120\times 128} \quad (14)$$

The output of the first block, $Y_1 \in \mathbb{R}^{128}$, is concatenated with the original in put $X_0 \in \mathbb{R}^{120\times 128}$, denoted as:

$$X_1 = [X_0, Y_1] \in \mathbb{R}^{120\times(128+128)} \quad (15)$$

Then, the combined feature X_1 is passsed through another stage (second stage), yielding $Y_2\mathbb{R}^{128}$, then they concatenated with X_1 across the feature channel:

$$X_2 = [X_1, Y_2] \in \mathbb{R}^{120\times(128+128+128)} \quad (16)$$

After going through the $N_{stage} = 2$ (Number of stages), the final concatenation X_2 is passed through a concluding Linear Layer:

$$\tilde{\mathcal{P}}_{3D} = W_{out}(X_2) \in \mathbb{R}^{120\times 3} \quad (17)$$

Where W_{out}, is a fully connected linear layer that maps the combined features down to the original shape [B, F, J, 3].

2.3. Early Fusion

Early Fusion combines the raw hypothesis in the early stage in the network before passing to any deep feature fusion, allowing the network to learn dependencies directly from the joint coordinates.

In **Figure 2.8**, begins with H different 3D pose hypotheses, each containing joint coordinates in 3D space (x, y, z), which are concatenated along the channel dimensions to form the input of size $[B, F, J, (C \times H)]$, where B is batch, F is represented Frames, J is number of joints, and H is number of hypotheses. This concatenation strategy allows the fusion network to simultaneously process all pose hypotheses without losing spatial and temporal relationships between joints across different hypotheses.

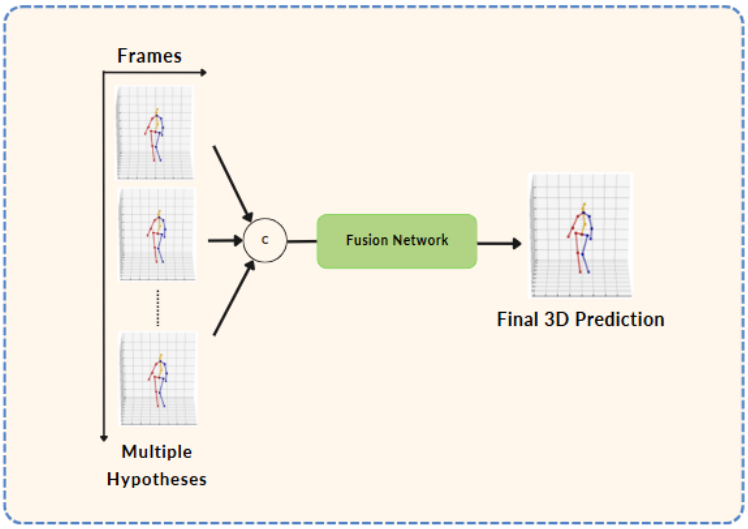

Figure 2.8 Overview of Early Fusion

After concatenating each hypothesis on C-channels, then it's fed into our fusion network that implements two distinct architectural approaches: ResidualFC and DenseFC networks as the same as we implemented in regression head for Feature Fusion.

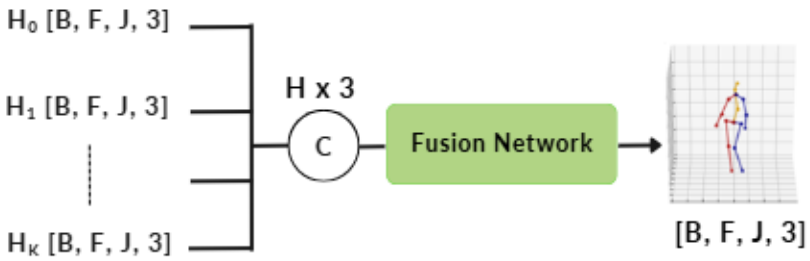

Figure 2.9 Detail architecture of Early Fusion

After obtaining the set of multiple 3D human pose hypotheses $H = \{H_0, H_1, H_2, \dots, H_k\}$ from the pre-trained D3DP model, which is mentioned in hypotheses generator section above, we have a set of hypotheses, each shaped as:

$$H_k \in \mathbb{R}^{B \times F \times J \times 3}$$

Where $k = 0, 1, 2, \dots, K$, each H_k represents a 3D pose hypothesis predicted from the same 2D input or 2D observation, across all frames. To perform early fusion, all hypotheses are concatenated along the last axis (C-channels), so that every joint at each frame holds information from all hypotheses:

$$H_{fused} = Concat(H_0, H_1, H_2, \dots, H_k) \in \mathbb{R}^{B \times F \times J \times (H \times 3)} \quad (18)$$

This allows the network to jointly reason over all hypotheses poses simultaneously, embedding the fused hypotheses into a shared representation. After that, the fused H_{fused} is then passed to the Fusion Network:

$$\tilde{P}_{3D} = E_{fused}(H_{fused}) \in \mathbb{R}^{B \times F \times J \times 3} \quad (19)$$

Where:

- $\tilde{\mathcal{P}}_{3D}$ is the Final 3D human pose prediction
- E_{fused} is the fusion network that consists of residualFC and denseFC
- H_{fused} is the output after we concatenated $(H \times C)$

a). Residual Fully Connected Layer (ResidualFC)

For the ResidualFC in the Early Fusion architecture, we adopt the same core structure as used in the Feature Fusion (FF) approach. However, unlike FF, the Early Fusion strategy does not incorporate a separate feature extractor. Instead, it relies solely on the fusion network to learn meaningful representations directly from the raw concatenated.

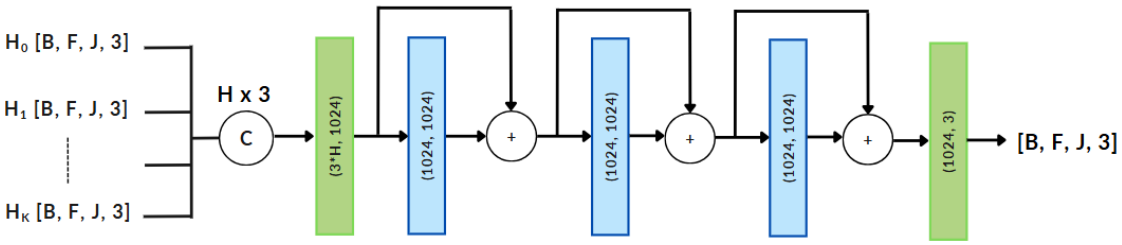

Figure 2.10 ResidualFC for Early Fusion

As illustrated in **Figure 2.10**, after we obtain this concatenated tensor of size $H \times 3$, we pass it through the fusion network, which is implemented using a residual fully connected network. This network first processes the input by mapping the concatenated dimension from $H \times 3$ into a higher latent space of size 1024, effectively

increasing the representational capacity and enabling the model to learn more complex feature interactions as described in **(Eq. 18)**. By following this **(Eq. 18)**, the output passed through the three residual stages as represented by **Figure 2.10**. stage consists of a fully connected layer followed by a non-linear activation function (e.g., PReLU), and includes a skip connection to preserve learned information in process of the residual stage as the same as **(Eq. 11, 12)** in Feature Fusion (FF). In each residual stage, we increase the number of stages to 3 and maintain the hidden dimensionality at **1024** to improve the network's learning capacity, since the Early Fusion framework does not include a separate feature extractor. This increased complexity allows the model to better capture relationships across hypotheses and joints. Finally, the output from the last residual stage is mapped back from 1024 to 3 through a linear projection layer to the original shaped of the 3D human pose as represented in (**Eq. 13**).

b). Dense Fully Connected Layer (DenseFC)

For the DenseFC in the Early Fusion (EF) architecture, we also follow the core principle from the Feature Fusion (FF) approach. However, the same as in ResidualFC, the Early Fusion framework does not incorporate a separate feature extractor. Instead, it relies entirely on the fusion network to learn meaningful representations directly from the raw concatenated 3D pose hypotheses.

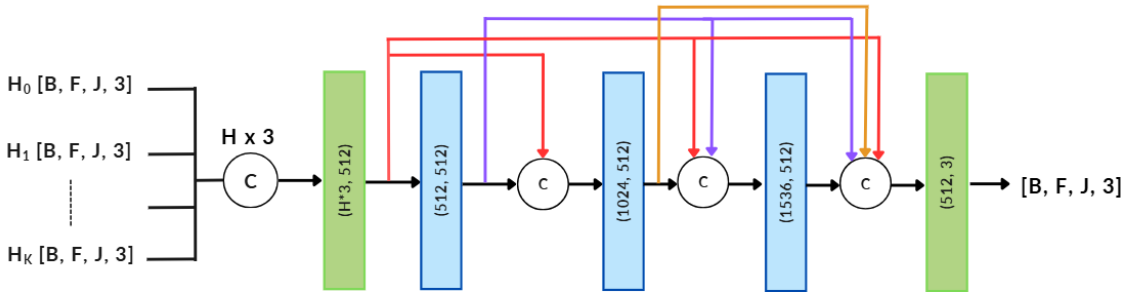

Figure 2.11 DenseFC for Early Fusion

In **Figure 3.14**, illustrates the DenseFC architecture for Early Fusion. After we obtained the concatenated tensor of size $H \times 3$, we pass it through the fusion network, first maps the input feature dimention $H \times 3$ into a latent space of size 512, as described in (**Eq. 14, 15**). This projection increases the feature representation capabilities and serves as the input for the subsequent dense stages. Each dense stage adjusts its weight dimensions begin with dense layer receives input of size 512 and output 512. Then, the second dense layer receives concatenated input size 1024 and output 512. After that, the dense layer receives concatenated input by following the concept of denseFC it increases to 1536. Finally, the output from the last dense stage is mapped back to original shaped from 512 to 3 as represented in **(Eq. 17**). This final

mapping enables the model to output accurately and refined 3D coordinates for each joint across all frames.

3. Experiment results

In this section, we will introduce experiments with each network and do the comparison and evaluation of the performance.

3.1. Experiment Settings

We implemented our proposed method on the PyTorch platform on an Intel Core i9, with RAM 64GB and NVIDIA GeForce RTX Dual GPU 2080Ti, VRAM 24GB. Our MHFusionNet was trained for 200 epochs by using an initial learning rate of 0.001 with the AdamW optimizer. Our experiment used Human3.6M as the dataset and evaluated with MPJPE.

Human3.6M is one of the most popular and largest datasets for 3D human pose estimation which contains 3.6 million human poses (2D/3D skeleton key points), which is captured by high-speed motion recording the dataset, and they produced high-quality images of a variety of human postures and activities. There are 15 daily activities performed by 11 human subjects in indoor environments. We followed the standard protocol by using 5 subjects (S1, S5, S6, S7, and S8) for training and 2 subjects (S9, S11) for testing. In this experiment, we use Human3.6M as the training datasets.

The performance of our proposed method is evaluated by using the standard evaluation metric. Protocol 1 is the mean per-joint position error (MPJPE) in (**Eq. 20**), which computes the average of the Euclidean distance error in millimeters (mm) between the predicted skeleton and the ground truth.

$$MPJPE = \frac{1}{F.J}\sum_{f=1}^{F}\sum_{j=1}^{J}\left\|\tilde{P}_f^j - P_f^j\right\|_2 \qquad (20)$$

Where f is the frame index, $\tilde{P}^j = (\tilde{X}^j, \tilde{Y}^j, \tilde{Z}^j)$ is the $j\text{-}th$ predicted 3D skeletal joint, $S^j = (X^j, Y^j, Z^j)$ is the ground truth, J is the number of joints ($J = 17$), and F is the number of frames in testing.

3.2. Ablation Study

To verify the impact of each proposed design of the individual components of MHFusionNet, we conduct extensive ablation experiments on Human 3.6M dataset under the evaluation matric protocol #1 (MPJPE).

To study the effectiveness and performance of fusion networks in our proposed MHFusionNet, we were designed based on two fusion strategies which are Feature Fusion (FF) and Early Fusion (EF) techniques. In Feature Fusion (FF) techniques were designed by following the concept of neural networks, including an MLP, ResidualFC,

DenseFC for regression head (RH), and used Poseformer-based for Feature Extractor.

3.2.1. Feature Fusion (FF)

In Feature Fusion, the architecture comprises two main modules: the feature extractor and the regression head. In this experiment, the Poseformer model is used as the feature extractor due to its proven effectiveness in capturing temporal dependencies in 3D human pose estimation tasks. To investigate the influence of the regression head we designed the different networks including MLP, ResidualFC, and DenseFC.

Table 3.1 Results Comparison of Different Regression Head designed

Feature Extractor	Regression Head	MPJPE (mm)	Model Size (MB)
Poseformer	MLP	40.09	119.0
	ResidualFC	40.02	183.9
	DenseFC	40.48	282.0

As shown in **Table 3.1**, the ResidualFC regression head achieved the lowest Mean Per Joint Position Error (MPJPE) of 40.02 mm, indicating the highest accuracy among the three designs. In contrast, while the MLP design yielded a comparable MPJPE of 40.09 mm, it had the smallest model size (119.0MB) than residualFC, making it more suitable for deployment in resource-constrained environments where computational cost. The DenseFC design resulted in the highest MPJPE of 40.48 mm and also largest model size (282.0 MB) since it was designed to connect each layer to every other layer in a feed-forward it's also caused the memory size. Overall, the ResidualFC presents a balance of trade-off, offering the best performance among other designs.

3.2.2. Early Fusion (EF)

In the Early Fusion (EF), we concatenated all hypotheses along the feature dimensional (C-dimensional) to form a single, unified representation. This concatenated feature is then passed through the Network to predict the final 3D human pose. Unlike feature fusion, early fusion does not employ a separate feature extractor, it relies entirely on the regression network to learn meaningful representations directly from the raw concatenated input. To investigate the learning capacity of the EF network, we increase the number of linear size and the number of stages in both residualFC and denseFC as we can see the detail in the above description.

Table 3.2 Results Comparison of different Early Fusion designed

Early Fusion	MPJPE (mm)	Model Size (MB)
ResidualFC	40.19	75.9
DenseFC	40.31	105.9

As shown in **Table 3.2**, the ResidualFC design once again achieved better performance with a lower MPJPE of 40.19 mm, compared to 40.31 mm for DenseFC. Furthermore, ResidualFC also has a smaller model size (75.9 MB) compared to DenseFC (105.9 MB), making it more computationally efficient while still achieving superior accuracy.

Table 3.3 Results Comparison between FF and EF

Fusion Network	MPJPE (mm)	Model Size (MB)
Feature Fusion (FF)	40.02	183.9
Early Fusion (EF)	40.19	75.9

As presented in **Table 3.3**, the Feature Fusion approach achieves the lowest MPJPE of 40.02 mm, slightly outperforming Early Fusion, which records an MPJPE of 40.19 mm. Although the difference in accuracy is relatively small (only 0.17 mm), it suggests that FF benefits from the separation of feature extraction and regression stages, allowing the model to extract richer and more abstract representations before final pose estimation.

In contrast, Early Fusion (EF) integrates all information in the early stage by directly concatenating pose hypotheses. While this reduces model complexity and leads to a significantly smaller model size (75.9 MB) almost 60% smaller than Feature Fusion (FF), the slight drop in accuracy but it's also acceptable since EF got small model size.

3.2.3. Performance comparison with State-of-the-Art (SOTA) Methods

We compared our proposed fusion-based methods with a baseline approach derived from the original D3DP (Direct 3D Pose) method, which generates multiple 3D pose hypotheses. In the baseline, the final pose is computed by averaging all predicted hypotheses. While averaging is a simple and efficient strategy, it often comes with notable trade-offs, particularly in scenarios involving uncertain or noisy predictions.

Table 3.4 Results Comparison of Our proposed method with averaging

	Number of Hypotheses	**MPJPE (mm)**
D3DP (Averaging)	5	40.41
Feature Fusion (Ours)	5	40.02
Early Fusion (Ours)	5	40.19

As illustrated in **Table 3.4**, both of our proposed fusion approaches—Feature Fusion (40.02 mm) and Early Fusion (40.19 mm)—outperform the D3DP averaging baseline (40.41 mm) in terms of Mean Per Joint Position Error (MPJPE). This demonstrates that the fusion network more effectively than Average. Because the averaging operates by computing the mean position of each joint across all hypotheses.

In scenarios where one or more hypotheses include significant errors (e.g., due to occlusion or incorrect joint placement), these outliers can disproportionately affect the result. Even if the majority of hypotheses are accurate, the averaging process will still blend in the poor-quality estimates, potentially degrading the overall prediction.

Example, if we generated five multiple hypotheses of 3D human poses and four out of five hypotheses correctly predict the wrist position, but one predicts it far off due to a noisy input or occluded, the averaged wrist location will be skewed away from the true position then it can be led to a larger MPJPE.

In this **table 3.5**, we present a comparison between our proposed fusion-based methods and the JPMA from D3DP model. JPMA includes two types of aggregation strategies, J-Agg (Joint Aggregation): Aggregates joint predictions across hypotheses, and P-Agg (Pose Aggregation)**:** Aggregates entire pose predictions. They also report **oracle-based results** (J-Best and P-Best), which represent the theoretical upper-bound performance when selecting the best joint or pose from the set of hypotheses, assuming access to ground-truth data.

Table 3.5 Results Comparison of our method with JPMA from D3DP model

	Number of Hypotheses	**MPJPE (mm)**
J-Best	5	37.70
P-Best	5	39.58
J-Agg	5	39.66
P-Agg	5	39.80
Feature Fusion (Our)	5	40.02
Early Fusion (Our)	5	40.19

While our proposed Feature Fusion and Early Fusion methods achieve slightly higher MPJPEs (40.02 mm and 40.19 mm, respectively) than JPMA's J-Agg (39.66 mm) and P-Agg (39.80 mm), the performance gap is relatively small between 0.3 to 0.5 mm and is offset by key practical advantages of our approach.

Although JPMA may show slightly better MPJPE, these methods rely heavily on the availability of intrinsic camera parameters, which are essential for transforming joint coordinates between image and camera spaces. To use J-Agg or P-Agg effectively, accurate **camera calibration** must be performed to estimate the intrinsic parameters (focal length, principal point, etc.). This process can be time-consuming, environment-specific, and requires specialized tools or calibration patterns.

Overall, our proposed methods Feature Fusion and Early Fusion don’t require intrinsic parameters, nor do they perform explicit coordinate transformations. Instead, the network learns to reason over multiple hypotheses directly, making it much more robust and deployment friendly.

4. Discussion

Our proposed MHFusionNet, which includes Feature Fusion (FF) and Early Fusion (EF), achieves competitive performance in Multiple 3D human pose estimation compared to SOTA methods like D3DP which is proposed JPMA (J-Agg, P-Agg) and averaging techniques. While JPMA methods show slightly better MPJPE, they require intrinsic camera parameters, limiting their use in real-world applications. Averaging, though simple, is sensitive to outliers and often degrades accuracy when poor hypotheses are present.

5. Conclusion

In this work, we proposed MHFusionNet, a multi-hypothesis fusion framework for 3D human pose estimation that effectively integrates multiple 3D pose hypotheses using deep learning. The network was implemented in two architectures: Feature Fusion (FF), which leverages a dedicated feature extractor (Poseformer), and Early Fusion (EF), which performs direct concatenation and regression. Through extensive experiments on the Human3.6M dataset, we demonstrated that both fusion strategies achieve competitive accuracy compared to state-of-the-art methods such as D3DP with JPMA (J-Agg, P-Agg) and averaging-based techniques. While JPMA shows slightly better accuracy, its dependence on camera calibration and intrinsic parameters makes it less practical for real-world deployment.

Acknowledgements

The completion of this project would not have been possible without the assistance and support, direction, and inspiration in various ways around me. This report was completed in 2025 at National Chung Cheng University under supervision of Professor Wen-Nung Lie. I would like to express my heartfelt appreciation to Professor Wen-Nung Lie, whose teaching and guidance enabled me to better understand my thesis project which led to a successful outcome. I sincerely thank him for his support in solving the difficult challenge of this project.

References

[1] W. Shan *et al.*, "Diffusion-Based 3D Human Pose Estimation with Multi-Hypothesis Aggregation," *Proc. IEEE Int. Conf. Comput. Vis.*, pp. 14715–14725, 2023, doi: 10.1109/ICCV51070.2023.01356.

[2] Y. Cai, W. Zhang, Y. Wu, and C. Jin, "FusionFormer: A Concise Unified Feature Fusion Transformer for 3D Pose Estimation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, 2024

[3] C. Rommel, V. Letzelter, N. Samet, R. Marlet, M. Cord, P. Pérez, and E. Valle, "ManiPose: Manifold-Constrained Multi-Hypothesis 3D Human Pose Estimation," *arXiv preprint arXiv:*, 2023. [Online]. Available: *arXiv URL*

[4] X. Liang, A. Angelopoulou, E. Kapetanios, B. Woll, R. Al Batat, and T. Woolfe, "A multi-modal machine learning approach and toolkit to automate recognition of early stages of dementia among British Sign Language users," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2020, pp. 278–293

[5] K. Peppas, K. Tsiolis, I. Mariolis, A. Topalidou-Kyniazopoulou, and D. Tzovaras, "Multi-modal 3D human pose estimation for human-robot collaborative applications," in *Proc. Struct., Syntactic, Stat. Pattern Recognit. (S+SSPR)*, Padua, Italy, 2021, pp. 355–364.

[6] P. Zhang, C. Lan, J. Xing, W. Zeng, J. Xue, and N. Zheng, "View adaptive neural networks for high performance skeleton-based human action recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 8, pp. 1963–1978, 2019

[7] J. C. Chiang et al., "Posture monitoring for health care of bedridden elderly patients using 3D human skeleton analysis via machine learning approach," *Appl. Sci.*, vol. 12, no. 6, p. 3087, 2022

[8] N.-H. Nguyen, T.-D.-T. Phan, G.-S. Lee, S.-H. Kim, and H.-J. Yang, "Gesture recognition based on 3D human pose estimation and body part segmentation for RGB data input," *Sensors*, vol. 20, no. 18, p. 5045, Sep. 2020

[9] K. Boekhoudt, A. Matei, M. Aghaei, and E. Talavera, "HR-Crime: Human-related anomaly detection in surveillance videos," in *Proc. Int. Conf. Comput. Anal. Images Patterns (CAIP)*, Virtual Event, 2021, pp. 164-174.

[10] C. Zheng, S. Zhu, M. Mendieta, T. Yang, C. Chen, and Z. Ding, "3D Human Pose Estimation with Spatial and Temporal Transformers," in *Proc. IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021, pp. 11656–11665.

[11] J. Zhang, Z. Tu, J. Yang, Y. Chen, and J. Yuan, "MixSTE: Seq2seq Mixed Spatio-Temporal Encoder for 3D Human Pose Estimation in Video," in *Proc. IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023, pp. 13232–13242.

[12] W. Li, H. Liu, H. Tang, P. Wang, and L. Van Gool, "MHFormer: Multi-Hypothesis Transformer for 3D Human Pose Estimation," in *Proc. IEEE/CVF*

Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 13147–13156.

[13] J. Gong, L. G. Foo, Z. Fan, Q. Ke, H. Rahmani, and J. Liu, "DiffPose: Toward More Reliable 3D Pose Estimation," in *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023, pp. 13041-13051.

[14] Q. Cai, X. Hu, S. Hou, L. Yao, and Y. Huang, "Disentangled Diffusion-Based 3D Human Pose Estimation with Hierarchical Spatial and Temporal Denoiser," in *Proc. IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023, pp. 22580-22590.

[15] J. Song, C. Meng, and S. Ermon, "Denoising Diffusion Implicit Models," *arXiv preprint arXiv:2010.02502*, 2020.