Multi-View 3D Human Pose Estimation with Weakly Synchronized Images

Ling Li[1], Ruiwen Gu[1], Chongyang Wang[3], Junliang Xing[2], Xinchun Yu[1], Xiao-Ping Zhang[1] *

[1]Shenzhen Key Laboratory of Ubiquitous Data Enabling, Shenzhen International Graduate School, Tsinghua University, China, Shenzhen

[2]Department of Computer Science and Technology, Tsinghua University

[3] West China Hospital, Sichuan University

liling22@mails.tsinghua.edu.cn, grw23@mails.tsinghua.edu.cn, mvrjustid@gmail.com, jlxing@tsinghua.edu.cn, yuxinchun@sz.tsinghua.edu.cn, xpzhang@ieee.org

Abstract

Multi-view 3D human pose estimation (MHPE) is an important research task in computer vision. To maintain consistency during the data collection, hardware synchronization devices are commonly used to connect cameras, ensuring that images from different views are captured simultaneously. However, synchronizing with extra devices has two apparent limitations: the hardware is i) usually expensive and ii) less flexible for deployment in outdoor open scenarios. Suppose the model can improve its tolerance for the time differences in multi-view image capture. In that case, the difficulty and cost of deployment will be greatly reduced, and MHPE will become more widespread. In this paper, we try to answer how to build a model that performs pose estimation directly using "weakly synchronized images" from multiple views, where the captured images shift from each other within a frame. To this end, we introduce a new multi-view 3D human pose estimation task given weakly synchronized image inputs. Apart from existing well-synchronized datasets, we present the first weakly synchronized dataset comprising 800k images. Thereon, we propose SyncDiffPose, a novel model based on the diffusion method for pose estimation to denoise the error in such data. By combining simple synchronization strategies, e.g., the timer method, our approach can perform pose estimation without hardware calibration.

Introduction

In recent years, computer vision tasks have made rapid progress (Chen et al. 2024a; Qian et al. 2024; Chen et al. 2024b; Sun 2024; Sun et al. 2024; Hu et al. 2024a,b; Ling et al. 2024a). As an essential task, multi-view 3D human pose estimation has critical sports and art motion analysis applications (Li and Meng 2022; Li et al. 2023, 2024; Xu et al. 2024; Wang et al. 2022, 2024a; Jin et al. 2023; Wang et al. 2024b; Ling et al. 2024b). We use images from different views as input and predict 3D coordinates of human joints. Existing MHPE methods typically require camera synchronization during data collection to ensure temporal consistency between views. The most common method for synchronizing multi-view images is to use a unique device that connects the cameras for hardware-level synchronized capturing. However, such synchronization devices are usually expensive and unsuitable for outdoor use. Camera-based clock synchronization is an alternative solution. The challenge with this approach lies in the need to establish a unified protocol between cameras, leading to higher development costs. Additionally, camera clock synchronization relies on the stability of the WiFi signal, and fluctuations in the WiFi signal can significantly affect synchronization accuracy. Ideally, multi-view models should be able to eliminate the dependency on synchronization to promote broader application of MHPE methods.

Some recent studies have begun exploring ways to eliminate the need for hardware synchronization devices in MHPE (Wu et al. 2019; Yin et al. 2022; Liu et al. 2024). Earlier researchers explored a technical approach that first obtained time differences between different views based on synchronization algorithms to align inputs from different views for temporally estimated pose. These studies used synchronized MHPE datasets like Human3.6M (Ionescu et al. 2013) to generate desynchronized data between views, which were then used to train models to estimate the time differences between different views (Yin et al. 2022). However, this method can only achieve integer multiples of inter-frame time differences, meaning image frames can only be shifted by whole frames and cannot obtain fractional frame time differences. In practical applications, the errors caused by the intra-frame intervals are still significant.

Furthermore, the difficulty of eliminating inter-frame interval errors is not severe and can usually be easily eliminated using methods such as clapperboard or timer methods. The challenge lies in removing intra-frame errors, as synchronized data only allows for moving whole frames at a time, limiting the input's diversity. Liu *et al.* collected the IFID dataset containing inter-frame and intra-frame time differences to support research on this task (Liu et al. 2024). However, the IFID dataset lacks information on human pose. In human pose estimation, existing MHPE datasets are generally collected using hardware-level synchronization devices, resulting in a lack of weakly synchronized datasets and corresponding algorithms.

Increasing the frame rate is another simple method for addressing shifts between cameras. As the camera frame rate increases, the errors caused by intra-frame intervals are continuously reduced. However, the cost of cameras also increases significantly. Essentially, this strategy cannot resolve

* Corresponding author

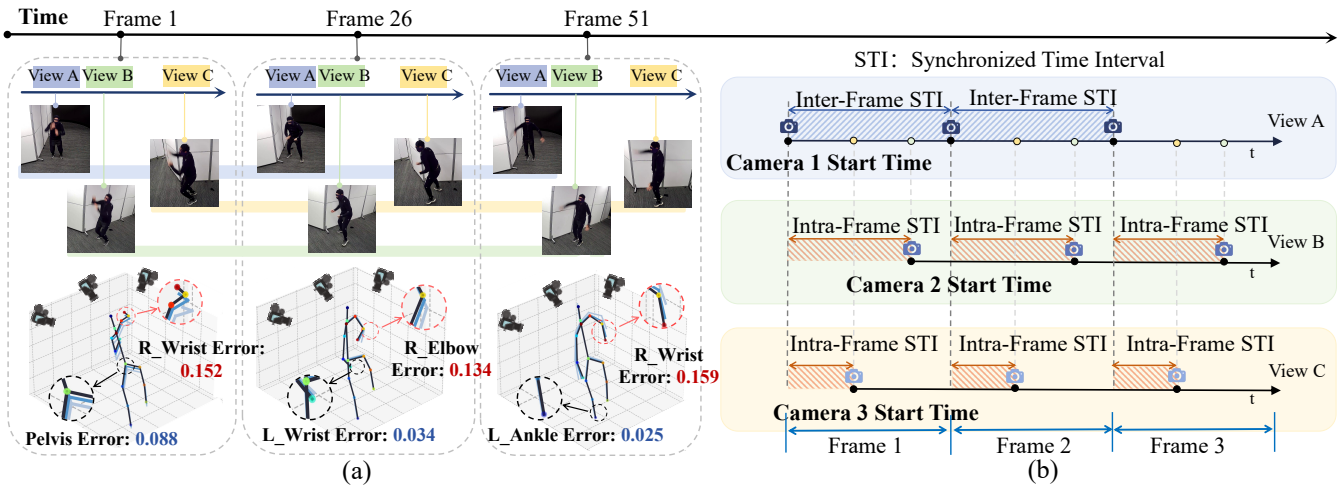

Figure 1: (a) The illustration of multi-view weakly synchronized images. These images are multi-view captures from different views with an intra-frame **Synchronization Time Interval (STI)**. Although the time intervals between images from different views are small, fast-moving key points, such as the ankle and wrist, can undergo rapid positional changes in scenarios involving high limb movement velocities. (b) Interpretation of intra-frame and inter-frame STI concepts. We refer to the STI duration between views within a single frame as intra-frame STI. STI duration between views that extends beyond the length of an entire frame is referred to as inter-frame STI.

the errors introduced by weak synchronization. Ideally, our model should be able to perform pose estimation directly based on **"weakly synchronized images"**.

The challenge of 3D human pose estimation based on weakly synchronized data is information fusion between views. Unlike synchronized data, weakly synchronized data from additional views introduces extra noise but also carries information. When the human body moves, the positions of joints that undergo rapid movement may change significantly in a very short time. Such varying positions of the same joint from different views could lead to significant errors using existing pose estimation models (Liu et al. 2024). We need to find a method to eliminate the synchronization discrepancies across different views, effectively utilizing the information from various views.

In this paper, we first define a novel task for Multi-view 3D Human Pose Estimation, focusing on data characterized by a Synchronization Time Interval (STI) confined to a single frame. We have termed this type of data "weakly synchronized images", a concept we propose for the first time, as illustrated in Figure 1. To mitigate the scarcity of weakly synchronized data, we have assembled a large-scale 3D human pose dataset consisting of multi-view images that are weakly synchronized. This dataset includes 800,000 frames and ground truth joint coordinates captured by OptiTrack devices, enabling comprehensive exploration of this task. To our knowledge, this is currently the first dataset targeting weakly synchronized data for MHPE. During the data collection, we ensure that one of the cameras and the Mocap system are well synchronized while the other cameras remain weakly synchronized. In pose estimation, the temporal weak synchronization between different views introduces noise. To address this issue, we introduce a novel model, SyncDiffPose, which employs a diffusion-based method to identify and eliminate noise in the pose estimation. This approach has demonstrated superior performance in MHPE tasks involving weakly synchronized images as inputs. In summary, our contributions are as follows:

- We introduce the concept of "weakly synchronized" for the first time and define a new task to extend the problem formulation of multi-view 3D human pose estimation.
- We have constructed the WeakSyncPose3D, the first large-scale multi-view weakly synchronized human pose dataset, which encompasses a broad spectrum of movements characterized by high motion frequencies.
- Considering pose estimation given weakly synchronized data faces the challenge of a mix of joints and noises, we designed our model based on the diffusion method, SyncDiffPose, which achieved promising performances.

Related Work

3D Human Pose Estimation

Monocular 3D human pose estimation is relatively prevalent, requiring pose estimation from a single-view image input (Moreno-Noguer 2017; Rhodin et al. 2018; Li et al. 2020). However, limited by the information that can be captured from a single camera, monocular 3D human pose estimation often encounters occlusion problems (Zhang, Chen, and Tu 2022; Zhang, Huang, and Wang 2020; Radwan, Dhall, and Goecke 2013). A natural solution is to estimate the pose from multiple views, known as multi-view 3D human pose estimation (Qiu et al. 2019; Jiang, Hu, and Xia 2023). In multi-view 3D human pose estimation, direct and voxel-based regression have gradually become two mainstream approaches. Typically, the performance of direct re-

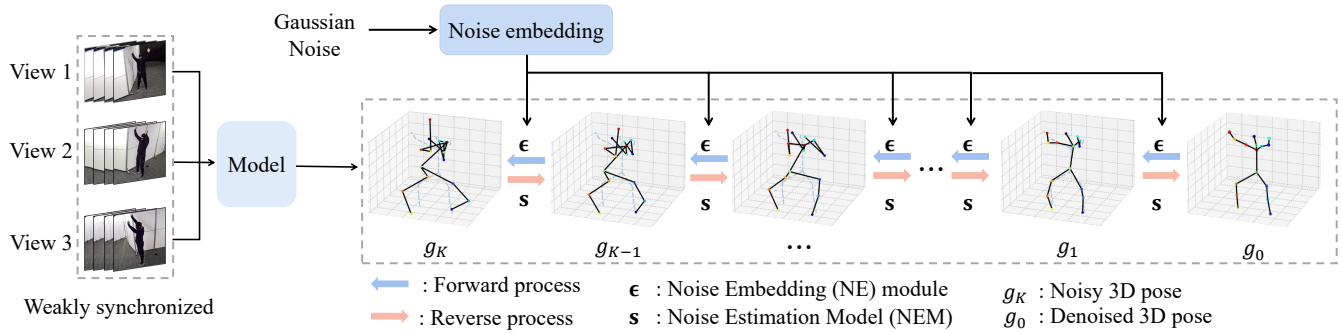

Figure 2: Overview of the architecture of SyncDiffPose. We incrementally embedded Gaussian noise, which is introduced from the Noise Embedding (NE) module ϵ into the "ground truth". After K rounds of noise addition, a noisy 3D pose g_K with high noise deviations is ultimately formed. Before initiating the reverse process, the intermediate 3D pose g_k, embedded with noise, is initialized using a standard modeling approach. The Noise Estimation Model (NEM) s is deployed to infer and systematically reduce the noise deviations throughout the reverse process. This iterative denoising over K steps culminates in retrieving the pristine, noise-free 3D pose g_0.

gression methods is slightly lower than voxel regression, but they often require less computational resources for training and inference. With the continuous development of convolution neural networks and Transformer models, recent methods (He et al. 2016; Vaswani et al. 2017; Liu et al. 2021) have performed well in multi-view 3D pose estimation. When sufficient view information is available, existing methods can obtain relatively more accurate results from images captured simultaneously from different views than what is estimated from the single view.

Diffusion Model for Human Pose Estimation

The emergence of Denoising Diffusion Probabilistic Models (DDPM) (Ho, Jain, and Abbeel 2020) has led to significant breakthroughs in generative models and is also widely applied across various tasks in computer vision (Lugmayr et al. 2022; Rombach et al. 2022; Hoogeboom et al. 2022; Tashiro et al. 2021). The diffusion method mainly consists of two phases: training and inference. During training, noise is continuously added to the information with a hidden Markov chain (Ephraim and Merhav 2002). Subsequently, a neural network is utilized to reverse the diffusion process and recover the data. In the inference phase, noise is continuously removed under the influence of conditions, ultimately achieving the generative target. Given that existing 2D pose estimation methods (Wang et al. 2020; Zhang, Zhu, and Ye 2019) have already achieved good performance, diffusion methods are primarily used in 3D human pose estimation (Feng et al. 2023). In this process, the 3D pose is typically considered as a pose carrying noise (Rommel et al. 2023) or originated directly from random noise (Zhou et al. 2023), using the 2D pose as a condition to denoise.

Available Datasets for MHPE

Human3.6M (Ionescu et al. 2013), CMU-panoptic (Joo et al. 2015), MPI-INF-3DHP (Mehta et al. 2017), among others, are the benchmark datasets for multi-view 3D human pose. However, these datasets use hardware synchronization devices to connect cameras during collection. This results in their inability to obtain weakly synchronized data, making the consequent models lack versatility in real-world application scenarios. The IFID dataset (Liu et al. 2024) is the first naturally collected multi-view desynchronized dataset with intra-frame interval annotations for the task of temporal synchronization. Unfortunately, the IFID dataset lacks annotations for 3D human poses. To fill this gap, we spot the need for a dataset that utilizes multi-view weakly synchronized images for human pose estimation.

Method

Problem Formulation

Weakly Synchronized Images. In this study, we introduce the concept of "weakly synchronized images" to characterize collections of images gathered from multiple views within a specific environment that are not perfectly synchronized. More precisely, weakly synchronized images refer to a series of image sequences $\mathcal{D} = \{\mathcal{D}_j\}_{j=1}^{M}$, collected from M distinct views $V = \{v_1, v_2, \ldots, v_M\}$ within the same environment, where $\mathcal{D}_j = (\mathbf{I_{1,j}}, \mathbf{I_{2,j}}, \ldots, \mathbf{I_{N,j}}) \in \mathbb{R}^{N\times H\times W\times 3}$, each $\mathbf{I_{i,j}} \in \mathbb{R}^{H\times W\times 3}$ represents the i-th sampling instance of an image from view v_j. A key attribute of these images is that, although the sampling instances from different views are not perfectly synchronized in time, the discrepancy in the timestamps $t_{i,j}$ and $t_{i,k}$ for the corresponding samples $\mathbf{I_{i,j}}$ and $\mathbf{I_{i,k}}$ from any two views v_j and v_k is constrained within a predefined maximum synchronization error Δt_{frame}. Formally, weakly synchronized images satisfy the following conditions

$$|t_{i,j} - t_{i,k}| < \Delta t_{frame}, \tag{1}$$

where $i \in \{1, 2, \ldots, N\}$, $j, k \in \{1, \ldots, M\}$ and $j \neq k$. $\Delta t_{frame} = 1/f$, where f represents the camera frame rate.

Target for Modeling. The traditional multi-view 3D human pose estimation task uses images captured simultaneously by multi-view cameras as input. Here, we extend the input

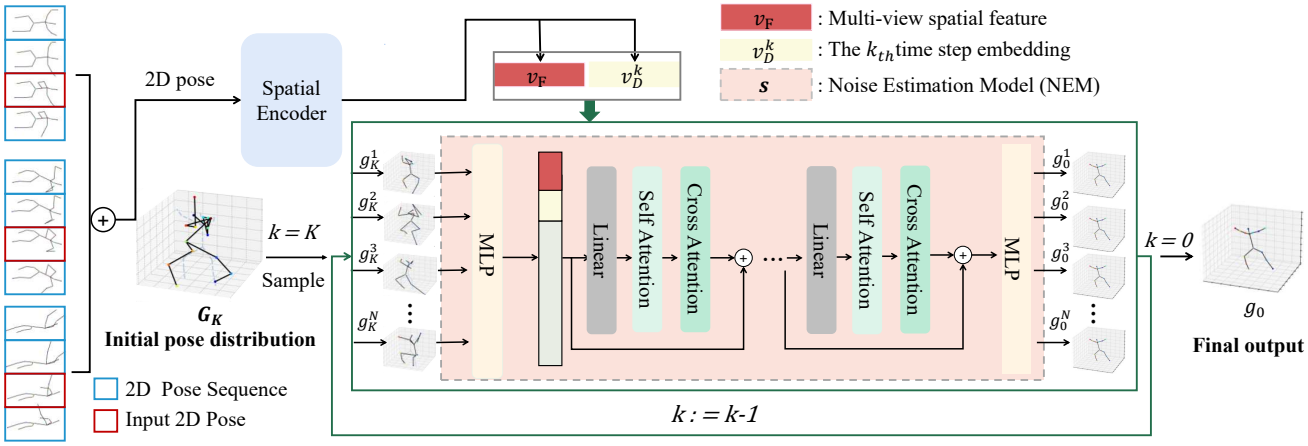

Figure 3: The pipeline of reverse process. The encoder extracts spatial information from weakly synchronized 2D poses, generating the feature vector v_F. Temporal encoding is applied to produce a step feature v_D^k for each k-th step. We then initialize the noisy 3D pose distribution G_K and sample N noisy pose samples $\{g_k^i\}_{i=1}^N$ from this distribution, which is fed into the reverse process model. These samples undergo K iterations of noise estimation and removal within the Noise Estimation Module (NEM) $\boldsymbol{s}$. The process yields N denoised pose samples $\{g_0^i\}_{i=1}^N$, from which the average $g_0 = \frac{1}{N}\sum_{i=1}^N g_0^i$ is computed, forming our final posture output.

to include a set of weakly synchronized image sequences $\mathcal{D} \in \mathbb{R}^{N\times M\times H\times W\times 3}$. We define it as follows

$$\mathbf{f} : \mathcal{T}(\mathcal{D}, \Phi, \phi) = \mathcal{C}. \tag{2}$$

Here, Φ is the 3D human pose estimation model based on weakly synchronized images, ϕ represents the camera parameters, $\mathcal{C} = (C_1, C_2, \ldots, C_N) \in \mathbb{R}^{N\times J\times 3}$ denotes the 3D human pose coordinates, and $\mathbf{f}$ maps the model's prediction to keypoint coordinates. In this paper, we process frames individually and define the problem as

$$\mathbf{f} : \mathcal{T}(\{\mathbf{I_{i,j}}\}_{j=1}^M, \Phi, \phi) = C_i. \tag{3}$$

WeakSyncPose3D Dataset

We deployed 3 optical and 4 motion capture cameras within a $4m \times 3m$ area. We synchronize one of the optical cameras with the 4 motion capture cameras. Performances carry out various sports on the site to highlight the importance of intra-frame STI and increase the motion frequency. The wide variety of poses and high-frequency movements also increase the difficulty of pose estimation. The supplementary materials provided detailed introductions to the optical cameras, mocap cameras, and the action capture setup we designed.

Annotation Details. To obtain weakly synchronized data with intra-frame STI for multi-view images sampled in the same environment, we first use a manual calibration technique known as the "timer method" to eliminate inter-frame STI. At the beginning of the recording, we start a timer on the field and calibrate the time based on the timer's display. The precision of the timer we use is 1 millisecond. During the calibration process, we manually synchronize the images from different views and remove the video segments that include the timer at the start of the recording. To acquire human keypoint coordinates, we wear motion capture suits and attach markers to the corresponding joints on the body. OptiTrack devices can accurately and in real-time capture the 3D coordinates of human body keypoints.

Data generalizability. It is noteworthy that we collected several segments of sports data under natural conditions, ensuring random initialization times across different cameras. This randomness in startup times guarantees that the collected weakly synchronized data encompasses a wide range of intra-frame synchronization intervals, thereby enhancing the generalizability of the dataset.

Diffusion Formulation for Pose Estimation

Diffusion models are a type of probabilistic generative models (Lugmayr et al. 2022) which aim to transition from the noisy sample, denoted as g_K, into a task-specific data sample g_0 through a systematic denoising process across K steps, formulated as $g_K \to g_{K-1} \to \ldots \to g_0$. This procedure, known as reverse diffusion, counteracts the forward diffusion path $g_0 \to g_1 \to \ldots \to g_K$.

For the model to learn this reverse diffusion, it necessitates a series of intermediate noisy samples $\{g_k\}_{k=1}^{K-1}$ to facilitate the transition from the original sample g_0 to the latent noisy sample g_K. The forward diffusion is executed to generate intermediate samples, which can be viewed as a Markov process and defined as

$$\begin{aligned} q(g_k|g_{k-1}) &= \mathcal{N}\left(g_k; \sqrt{1-\beta_k}g_{k-1}, \beta_k\mathbf{I}\right), \\ q(g_{1:K}|g_0) &= \prod_{k=1}^{K} q(g_k|g_{k-1}), \end{aligned} \tag{4}$$

where $\mathcal{N}(g_k;\cdot)$ denotes a normal distribution with a mean vector of $\sqrt{1-\beta_k}g_{k-1}$ and covariance matrix of $\beta_k\mathbf{I}$, β_k

Method	Head	L-Ankle	L-Elbow	Hip	L-Knee	Shoulder	L-Wrist	Pelvis	R-Ankle	R-Elbow	R-Knee	R-Wrist	Spine	**Avg.**
Denis *et al.*	34.42	56.06	182.40	60.76	87.62	130.97	190.02	70.74	96.20	145.59	184.52	184.64	94.88	112.26
Karim *et al.*	37.61	54.41	182.54	63.50	85.27	129.54	178.16	68.09	94.06	138.65	178.47	179.84	98.81	110.82
MvP	35.50	49.56	168.05	58.18	81.42	125.85	171.33	65.40	92.51	135.47	167.52	163.31	94.51	104.70
Ours	**23.10**	**28.45**	**127.98**	**44.24**	**63.83**	**99.65**	**137.27**	**48.95**	**74.26**	**110.32**	**112.79**	**117.97**	**85.84**	**84.86**
Denis *et al.*	28.80	49.65	178.91	55.80	78.65	125.78	177.41	64.85	94.22	135.09	176.71	173.60	86.49	105.57
Karim *et al.*	24.12	33.39	153.41	51.91	70.27	116.83	157.68	54.25	79.62	123.71	152.31	151.25	80.31	93.29
MvP	32.99	40.71	152.63	46.04	67.49	114.81	159.17	57.84	91.67	131.27	148.43	145.34	81.12	94.29
Ours	**18.75**	**22.39**	**111.06**	**35.66**	**61.22**	**87.08**	**117.49**	**43.55**	**74.02**	**102.18**	**95.71**	**100.66**	**78.16**	**70.24**

Table 1: Comparison with existing methods. In the upper part of the table, the results are acquired using 2D pose estimations from the ShuffleNet V2 (Ma et al. 2018) model. In the lower part, the ground truth 2D poses are used. The positions of human body joints are consistent with the positions of joints in the MPII dataset (Andriluka et al. 2014).

	Method	Head	L-Ankle	L-Elbow	Hip	L-Knee	Shoulder	L-Wrist	Pelvis	R-Ankle	R-Elbow	R-Knee	R-Wrist	Spine	**Avg.**
S	Denis *et al.*	66.32	61.04	68.29	66.30	50.87	68.75	73.85	72.61	55.88	57.81	51.19	54.39	50.69	63.08
y	Karim *et al.*	60.14	61.07	55.40	48.74	66.08	56.05	62.79	54.49	51.66	69.69	51.53	63.48	66.17	57.78
n	MvP	**49.61**	**21.36**	26.76	**24.55**	**19.58**	**41.69**	28.65	**21.89**	30.28	44.45	53.73	46.40	**32.88**	32.97
c.	**Ours**	56.95	32.62	**10.04**	44.11	19.29	51.98	**10.87**	26.38	**15.46**	**32.50**	**28.09**	**16.11**	42.05	**30.85**
U	Denis *et al.*	77.42	122.86	**43.09**	95.65	**65.74**	76.96	80.20	141.90	107.58	131.69	119.05	99.90	106.04	99.70
n	Karim *et al.*	78.01	85.13	139.80	82.40	71.20	92.21	80.26	66.94	122.85	139.48	71.30	107.45	51.99	90.91
S	MvP	56.31	116.38	127.70	77.03	124.33	112.42	86.07	**65.22**	51.70	**69.99**	71.17	129.71	117.44	92.98
y.	**Ours**	**50.18**	**55.41**	58.38	**74.84**	85.35	**75.13**	**69.71**	75.84	**49.85**	82.50	**65.39**	**81.33**	**51.91**	**69.17**

Table 2: Comparison of Results for Different Methods on Human3.6M. The upper section of the table displays the training and testing results obtained using the original Human3.6M dataset, assessing the performance of our model on synchronized data. The lower section presents the results where a random frameshift of 0 or 1 frame was applied across different views.

is a predetermined noise variance schedule modulating the noise at arbitrary sampling step.

Suppose $\overline{\alpha}_k = \prod_{i=1}^{k} \alpha_i = \prod_{i=1}^{k}(1-\beta_i)$, we can accordingly articulate g_k as a linear amalgamation of the original sample g_0 and a noise term ϵ, g_k can be defined as

$$g_k = \sqrt{\overline{\alpha}_k} g_0 + \sqrt{(1-\overline{\alpha}_k)}\epsilon. \tag{5}$$

where each element of noise term ϵ is drawn from $\mathcal{N}(0,1)$. Due to the additive property of independent Gaussian distributions, we can denote the conditional probability of g_k given g_0 as

$$q(g_k|g_0) = \mathcal{N}\left(g_k; \sqrt{\overline{\alpha}_k} g_0, (1-\overline{\alpha}_k)\mathbf{I}\right). \tag{6}$$

As the maximum step $K \to \infty$, the distribution of g_K converges to a standard Gaussian distribution, indicating the complete transformation of g_0 into Gaussian noise.

Utilizing the initial sample g_0 and intermediate noisy samples $\{g_k\}_{k=1}^{K}$ produced via forward process, the model s is optimized to predict the reverse distribution $p(g_{k-1}|g_k)$. Each reverse diffusion step is typically modeled as a function f, which can be denoted as

$$g_{k-1} = f(g_k, s). \tag{7}$$

SyncDiffPose Model Architecture

As illustrated in Figure 2, the process of 3D pose estimation from weakly synchronized images is divided into two stages: (i) using a basic model for 3D pose estimation to obtain an initial uncertain pose distribution G_k with noise deviations, and (ii) executing the reverse diffusion process facilitates noise elimination. Utilizing a noise estimation model s, we iteratively mitigate the noise embedded within the initial pose distribution G_k, culminating in the attainment of a high-quality, definitive pose $g_0 \in \mathbb{R}^{J\times 3}$.

SyncDiffPose Forward Diffusion

To effectively learn progressive denoising, intermediate "ground truth" distributions are necessary as training supervisory signals. Specifically, with a defined target pose g_0, the forward diffusion sequence ($g_0 \to g_1 \to \ldots \to g_K$), where K denotes the maximal diffusion steps, is designed to systematically escalate the uncertainty in g_0, converging towards the pose uncertainty g_K.

The final pose g_K is not just random noise. It arises from the 3D pose distribution of an existing model, displaying complex characteristics. Based on this analysis, we adopt the forward noise model shown in Figure 2. Differently, we emphasize the use of Noise Embedding (NE) to simulate the noise generated by weak temporal synchronization, thereby replicating the noisy and uncertain 3D pose distribution G_k. Specifically, the forward process based on NE is as follows

$$\min_{n} \sum_{i=1}^{n} \mathcal{L}\left(g_K^i, \hat{g_K^i} \mid \Phi_{NE}\right). \tag{8}$$

$g_K^1, \ldots, g_K^n$ are n poses with noise sampled from the noisy pose distribution G_K. Here, $\mathcal{L}$ denotes the loss function. We use Mean Square Error (MSE) loss here.Φ_{NE} represents the NE module, which consists of several simple linear layers. $\hat{g_K}$ is determined by the NE module

$$\hat{g_K} = \mu^{NE} + \sqrt{\overline{\alpha}_k}(g_0 - \mu^{NE}) + \sqrt{(1-\overline{\alpha}_k)} \cdot \epsilon^{NE}. \tag{9}$$

Here, μ^{NE} is the mean determined by the NE module, $\epsilon^{NE} \sim \mathcal{N}(0, \Sigma^{NE})$, where Σ^{NE} is the covariance matrix. To be more specific, $\mu^{NE} = \mathcal{L}(\mu_{g_0}|\Phi_{NE})$, where Φ_{NE} represents the NE model, consisting of linear layers. $\mathcal{L}(\mu_{g_0}|\Phi_{NE})$ represents the feature transformation based on

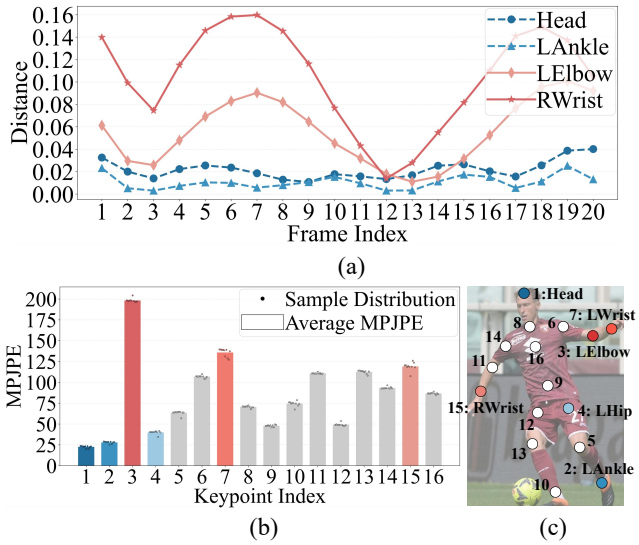

Figure 4: (a) The Euclidean distances between the positions of specific joints across consecutive frames in a sequence. (b) When using traditional models for pose estimation with weakly synchronized multi-view data, the weak synchronization has a more significant impact on the accuracy of limb extremities. (c) Mapping between the keypoint index and joint locations on the human body.

the NE model, which serves as the input for g_0. $\mu_{g_0} = \mu(g_0)$ is the mean value of the determined pose g_0. In this manner, when $\alpha_K \approx 0$, $\hat{g_K}$ is drawn from the fitted NE module, i.e., $\hat{g_K} = \mu^{NE} + \epsilon^{NE}$. This enables us to generate samples as supervisory signals for training the NE module.

Reverse Diffusion Process

The reverse diffusion process aims to eliminate the noise in the uncertain 3D pose distribution G_K, thereby obtaining a noise-free and accurate 3D pose g_0. The pipeline of the reverse process is illustrated in Figure 3.

Multi-View Spatial Encoder. We deploy the Spatial Encoder to extract comprehensive spatial information v_F from the multi-view 2D poses $\mathcal{S} = \{S_1, S_2, \ldots, S_M\}, S_M \in \mathbb{R}^{J\times 2}$. Where M represents the number of views, and J denotes the number of keypoints. Here, we denote v_F as

$$v_F = \mathcal{F}(\text{concat}(S_1, S_2, \ldots, S_M)), v_F \in \mathbb{R}^{J\times 64}. \quad (10)$$

Here, we set $M = 3, J = 16$. $\mathcal{F}$ is a linear layer that transforms the dimension from $\mathbb{R}^{M\times J\times 2}$ to $\mathbb{R}^{J\times 64}$.

Noise Estimation Model $\boldsymbol{s}$. As illustrated in Figure 3, our Noise Estimation Model $\boldsymbol{s}$ comprises lightweight sequence-to-sequence modules, each containing a linear layer, a self-attention, and a cross-attention structure. The modules are interconnected through residual connections. In the k-th step, we take $g_k \in \mathbb{R}^{N\times J\times 3}$ as the input and predict the noise distribution. The specific dimensional information is provided in the supplementary materials.

Reverse Process. The diffusion process focuses on refining an uncertain pose distribution G_K with noise during the training or evaluation phases into a precise pose g_0. Given a noisy pose $\hat{g}_k$, the Noise Embedding Model $\boldsymbol{\epsilon}$ conditioned upon the diffusion step k and the extracted spatial-temporal context v_F methodically deduces $\hat{g}_{k-1}$ from $\hat{g}_k$. The noise model, conditioned on the diffusion step k and the extracted spatial feature v_F, estimates the current noise, thereby continuously aiding in the denoising of the pose. This process can be denoted as:

$$\hat{g}_{k-1} = f(s_\theta\left(\hat{g}_k, v_F, v_D^k\right)), \quad k \in \{1, \ldots, K\}. \quad (11)$$

Method	Diff	NE	Upper	Lower	**Avg.**
MvP	✗	✗	159.54	84.56	104.70
	✓	✗	157.93	83.44	103.41
SyncPose	✗	✗	177.87	93.46	115.98
	✓	✗	154.44	83.93	101.84
	✓	✓	**123.39**	**61.30**	**84.86**

Table 3: Performance comparison between methods that particularly adopted diffusion and NE modules. The input 2D poses are predicted by the ShuffleNet V2 (Ma et al. 2018).

Experiments

Datasets and Evaluation Metrics

Datasets. The WeakSyncPose3D dataset is the first multi-view weakly synchronized 3D human pose estimation dataset. It comprises a total of $835,272$ frames of action, with $624,996$ frames for training, $93,816$ frames for validation, and $116,460$ frames for testing.

Human3.6M (Ionescu et al. 2013) is a widely used multi-view 3D human pose dataset. During data collection, synchronization devices were used to ensure that the input images from different views were fully synchronized in time. The dataset includes 11 action subjects. We trained on 5 subjects (S1, S5, S6, S7, S8) and tested on 2 subjects (S9, S11).

Evaluation Metrics. We use common evaluation metrics, such as mean per-joint position error (MPJPE), to measure the accuracy of denoised 3D pose against ground truth data. MPJPE measures the average Euclidean distance between the denoised 3D pose and the ground truth in 3D space. For a denoised 3D pose J and its corresponding ground truth pose J^*, both composed of N joints with their 3D coordinates, we calculate MPJPE as follows

$$\textbf{MPJPE} = \frac{1}{N}\sum_{i=1}^{N}\|J_i - J_i^*\|_2, \quad (12)$$

where J_i and J_i^* represent the 3D coordinates of the i-th joint, and $\|\cdot\|_2$ denotes the L2 norm.

Experiment Results

We compared our model with several advanced 3D human pose estimation models (Tome et al. 2018; Iskakov et al. 2019; Zhang et al. 2021) on the WeakSyncPose3D and Human3.6M (Ionescu et al. 2013) datasets, with the results shown in Table 1 and Table 2, respectively. Among the various methods, Denis *et al.* (Tome et al. 2018), and Karim

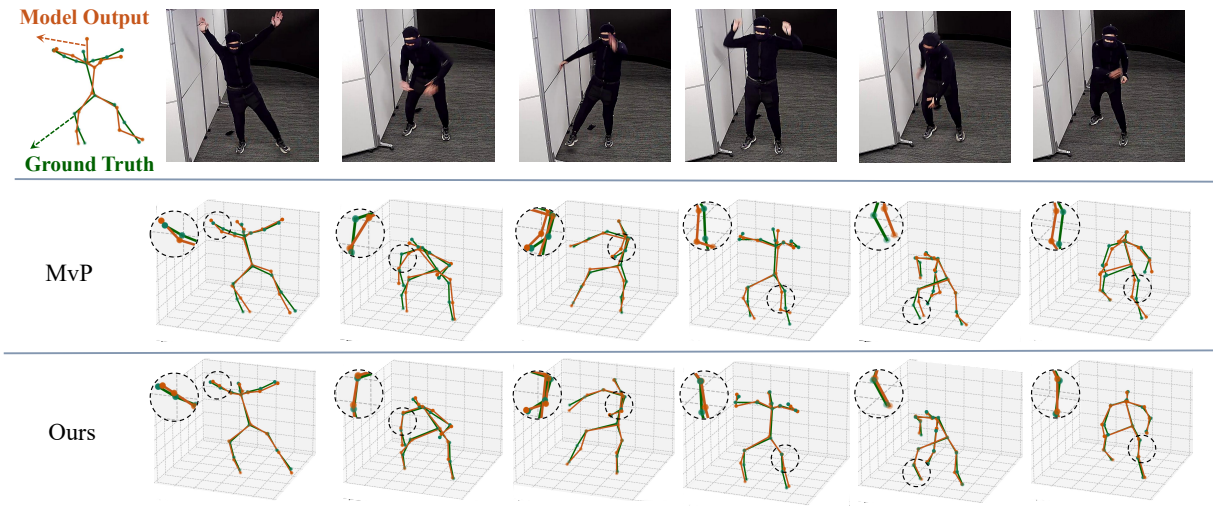

Figure 5: Qualitative comparisons of 3D human poses estimated by different methods. The green-colored skeleton represents the ground truth, and the significant improvements achieved by our method compared to other single-view and multi-view methods are highlighted with a black dotted circle and magnified.

et al. (Iskakov et al. 2019) represent high-performing CNN-based approaches. MvP (Zhang et al. 2021) is an attention mechanism-based method with excellent performance.

We separately trained and tested our model on the WeakSyncPose3D and Human3.6M datasets. Table 1 shows the test results of our model after training on the WeakSyncPose3D dataset. Table 2 presents the performance of our model on the Human3.6M dataset, including both the original data and the one-frame frameshift weakly synchronized data. Compared to the improvements on the synchronized dataset, our model shows even more significant improvements on the weakly synchronized dataset.

In Figure 4 (a), we calculate the Euclidean distances between the positions of specific points and analyze the fluctuation of positions of various keypoints in successive frames within weakly synchronized data. Our analysis indicates significant positional fluctuations between successive frames at limb extremities, such as the left elbow and right wrist. Given that the maximum STI for weakly synchronized data corresponds to a one-frame interval, these fluctuations demonstrate the errors resulting from weak synchronization. Consequently, intra-frame STI significantly impacts these points. The results of MPJPE for each point in Figure 4 (b) also validate this observation. From Figure 5, we can intuitively observe the excellent performance of our method.

Ablation Study

Comparison of results regarding using the diffusion model. As seen from Table 3, applying the diffusion method improves model performance. In our view, diffusion has the potential to partially mitigate coordinate deviations arising from temporal discrepancies between different views or other contributing factors.

Comparison of results regarding using the NE module. As seen from Table 3, the diffusion method combined with the NE module significantly improves. Considering the upper and lower error limits associated with weakly synchronized data, we posit that the NE module maps the noise into the corresponding error space, thereby aligning the noise distribution effectively.

Conclusion

In this study, we further expanded the multi-view 3D human pose estimation task based on "weakly synchronized images". Introducing this concept significantly reduces the deployment complexity and cost of MHPE, greatly facilitating the broader application of MHPE methods. To confront this newly identified challenge, we have compiled a comprehensive dataset comprising 800k frames of weakly synchronized 3D human poses from multiple views, establishing the first dataset dedicated to this specific task. We also proposed a novel framework based on the diffusion method for 3D human pose estimation, SyncDiffPose, specifically designed to mitigate the inaccuracies introduced by the weak synchronization. Our experimental results demonstrate the outstanding performance of this approach.

Limitations. Despite achieving commendable results and effectively mitigating noise introduced by weak synchronization, our model's performance is constrained by insufficient synchronization capabilities, preventing it from reaching the optimal levels observed in models that utilize synchronized images as input.

Acknowledgements

This research is supported by the Shenzhen Ubiquitous Data Enabling Key Lab under grant ZDSYS20220527171406015 and by the Tsinghua Shenzhen International Graduate School-Shenzhen Pengrui Endowed Professorship Scheme of the Shenzhen Pengrui Foundation.

References

Andriluka, M.; Pishchulin, L.; Gehler, P.; and Schiele, B. 2014. 2D Human Pose Estimation: New Benchmark and State of the Art Analysis. In *Proceedings of the IEEE Conference on computer Vision and Pattern Recognition*, 3686–3693.

Chen, Y.; Huang, W.; Liu, X.; Deng, S.; Chen, Q.; and Xiong, Z. 2024a. Learning Multiscale Consistency for Self-Supervised Electron Microscopy Instance Segmentation. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, 1566–1570.

Chen, Y.; Shi, H.; Liu, X.; Shi, T.; Zhang, R.; Liu, D.; Xiong, Z.; and Wu, F. 2024b. TokenUnify: Scalable Autoregressive Visual Pre-training with Mixture Token Prediction. *arXiv preprint arXiv:2405.16847*, 1–28.

Ephraim, Y.; and Merhav, N. 2002. Hidden Markov processes. *IEEE Transactions on Information Theory*, 48: 1518–1569.

Feng, R.; Gao, Y.; Tse, T. H. E.; Ma, X.; and Chang, H. J. 2023. DiffPose: SpatioTemporal Diffusion Model for Video-Based Human Pose Estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 14861–14872.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep Residual Learning for Image Recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 770–778.

Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising Diffusion Probabilistic Models. In *Advances in Neural Information Processing Systems*, 6840–6851.

Hoogeboom, E.; Satorras, V. G.; Vignac, C.; and Welling, M. 2022. Equivariant Diffusion for Molecule Generation in 3D. In *Proceedings of the International Conference on Machine Learning*, 8867–8887.

Hu, D.; Guan, R.; Liang, K.; Yu, H.; Quan, H.; Zhao, Y.; Liu, X.; and He, K. 2024a. ScEGG: An Exogenous Gene-guided Clustering Method for Single-cell Transcriptomic Data. *Briefings in Bioinformatics*, 25: bbae483.

Hu, D.; Liu, S.; Wang, J.; Zhang, J.; Wang, S.; Hu, X.; Zhu, X.; Tang, C.; and Liu, X. 2024b. Reliable Attribute-missing Multi-view Clustering with Instance-level and feature-level Cooperative Imputation. In *Proceedings of the ACM International Conference on Multimedia*, 1456–1466.

Ionescu, C.; Papava, D.; Olaru, V.; and Sminchisescu, C. 2013. Human3.6M: Large Scale Datasets and Predictive Methods for 3D Human Sensing in Natural Environments. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36: 1325–1339.

Iskakov, K.; Burkov, E.; Lempitsky, V.; and Malkov, Y. 2019. Learnable Triangulation of Human Pose. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 7718–7727.

Jiang, B.; Hu, L.; and Xia, S. 2023. Probabilistic Triangulation for Uncalibrated Multi-View 3D Human Pose Estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 14850–14860.

Jin, Z.; Wang, Z.; Wang, Q.; Jia, J.; Bai, Y.; Zhao, Y.; Li, H.; and Wang, X. 2023. HoloSinger: Semantics and Music Driven Motion Generation with Octahedral Holographic Projection. In *Proceedings of the ACM International Conference on Multimedia*, 9393–9395.

Joo, H.; Liu, H.; Tan, L.; Gui, L.; Nabbe, B.; Matthews, I.; Kanade, T.; Nobuhara, S.; and Sheikh, Y. 2015. Panoptic Studio: A Massively Multiview System for Social Motion Capture. In *Proceedings of the IEEE International Conference on Computer Vision*, 3334–3342.

Li, R.; and Meng, L. 2022. Multi-View Spatial-Temporal Network for Continuous Sign Language Recognition. *arXiv preprint arXiv:2204.08747*.

Li, R.; Zhang, Y.; Zhang, Y.; Zhang, H.; Guo, J.; Zhang, Y.; Liu, Y.; and Li, X. 2024. Lodge: A Coarse to Fine Diffusion Network for Long Dance Generation Guided by the Characteristic Dance Primitives. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1524–1534.

Li, R.; Zhao, J.; Zhang, Y.; Su, M.; Ren, Z.; Zhang, H.; Tang, Y.; and Li, X. 2023. FineDance: A Fine-grained Choreography Dataset for 3D Full Body Dance Generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 10234–10243.

Li, S.; Ke, L.; Pratama, K.; Tai, Y.-W.; Tang, C.-K.; and Cheng, K.-T. 2020. Cascaded Deep Monocular 3D Human Pose Estimation With Evolutionary Training Data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6173–6183.

Ling, L.; Xing, J.; Yu, X.; and Zhang, X.-P. 2024a. Deviation Wing Loss for High-Performance 2D Pose Estimation. In *Proceedings of the IEEE International Conference on Multimedia and Expo*, 1–6.

Ling, L.; Yang, W.; Yu, X.; Xing, J.; and Zhang, X.-P. 2024b. Translating Motion to Notation: Hand Labanotation for Intuitive and Comprehensive Hand Movement Documentation. In *Proceedings of the ACM International Conference on Multimedia*, 4092–4100.

Liu, Y.; Ai, H.; Xing, J.; Li, X.; Wang, X.; and Tao, P. 2024. Advancing Video Synchronization with Fractional Frame Analysis: Introducing a Novel Dataset and Model. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 3828–3836.

Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; and Guo, B. 2021. Swin Transformer: Hierarchical Vision Transformer using Shifted Windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 10012–10022.

Lugmayr, A.; Danelljan, M.; Romero, A.; Yu, F.; Timofte, R.; and Van Gool, L. 2022. RePaint: Inpainting Using Denoising Diffusion Probabilistic Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11461–11471.

Ma, N.; Zhang, X.; Zheng, H.-T.; and Sun, J. 2018. ShuffleNet V2: Practical Guidelines for Efficient CNN Architecture Design. In *Proceedings of the European Conference on Computer Vision*, 116–131.

Mehta, D.; Rhodin, H.; Casas, D.; Fua, P.; Sotnychenko, O.; Xu, W.; and Theobalt, C. 2017. Monocular 3D Human Pose Estimation in the Wild Using Improved CNN Supervision. In *Proceedings of the IEEE International Conference on 3D vision*, 506–516.

Moreno-Noguer, F. 2017. 3D Human Pose Estimation From a Single Image via Distance Matrix Regression. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2823–2832.

Qian, H.; Chen, Y.; Lou, S.; Khan, F.; Jin, X.; and Fan, D.-P. 2024. MaskFactory: Towards High-quality Synthetic Data Generation for Dichotomous Image Segmentation. In *Proceedings of the Neural Information Processing Systems*, 1–24.

Qiu, H.; Wang, C.; Wang, J.; Wang, N.; and Zeng, W. 2019. Cross View Fusion for 3D Human Pose Estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 4342–4351.

Radwan, I.; Dhall, A.; and Goecke, R. 2013. Monocular Image 3D Human Pose Estimation under Self-Occlusion. In *Proceedings of the IEEE International Conference on Computer Vision*, 1888–1895.

Rhodin, H.; Spörri, J.; Katircioglu, I.; Constantin, V.; Meyer, F.; Müller, E.; Salzmann, M.; and Fua, P. 2018. Learning Monocular 3D Human Pose Estimation From Multi-View Images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 8437–8446.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-Resolution Image Synthesis With Latent Diffusion Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10684–10695.

Rommel, C.; Valle, E.; Chen, M.; Khalfaoui, S.; Marlet, R.; Cord, M.; and Pérez, P. 2023. DiffHPE: Robust, Coherent 3D Human Pose Lifting with Diffusion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 3220–3229.

Sun, H. 2024. Ultra-High Resolution Segmentation via Boundary-Enhanced Patch-Merging Transformer. arXiv:2412.10181.

Sun, H.; Xu, L.; Jin, S.; Luo, P.; Qian, C.; and Liu, W. 2024. PROGRAM: PROtotype GRAph Model based Pseudo-Label Learning for Test-Time Adaptation. In *The Twelfth International Conference on Learning Representations*, 1–31.

Tashiro, Y.; Song, J.; Song, Y.; and Ermon, S. 2021. CSDI: Conditional Score-based Diffusion Models for Probabilistic Time Series Imputation. In *Advances in Neural Information Processing Systems*, 24804–24816.

Tome, D.; Toso, M.; Agapito, L.; and Russell, C. 2018. Rethinking Pose in 3D: Multi-stage Refinement and Recovery for Markerless Motion Capture. In *Proceedings of the International Conference on 3D Vision*, 474–483.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is All you Need. In *Advances in Neural Information Processing Systems*, 1–11.

Wang, J.; Long, X.; Gao, Y.; Ding, E.; and Wen, S. 2020. Graph-PCNN: Two Stage Human Pose Estimation with Graph Pose Refinement. In *Proceedings of the European Conference on Computer Vision*, 492–508.

Wang, Z.; Jia, J.; Sun, S.; Wu, H.; Han, R.; Li, Z.; Tang, D.; Zhou, J.; and Luo, J. 2024a. DanceCamera3D: 3D Camera Movement Synthesis with Music and Dance. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7892–7901.

Wang, Z.; Jia, J.; Wu, H.; Xing, J.; Cai, J.; Meng, F.; Chen, G.; and Wang, Y. 2022. GroupDancer: Music to Multi-People Dance Synthesis with Style Collaboration. In *Proceedings of the ACM International Conference on Multimedia*, 1138–1146.

Wang, Z.; Li, J.; Qin, X.; Sun, S.; Zhou, S.; Jia, J.; and Luo, J. 2024b. DanceCamAnimator: Keyframe-Based Controllable 3D Dance Camera Synthesis. In *Proceedings of the ACM International Conference on Multimedia*, 10200–10209.

Wu, X.; Wu, Z.; Zhang, Y.; Ju, L.; and Wang, S. 2019. Multi-Video Temporal Synchronization by Matching Pose Features of Shared Moving Subjects. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2729–2738.

Xu, Z.; Zhang, Y.; Yang, S.; Li, R.; and Li, X. 2024. Chain of Generation: Multi-Modal Gesture Synthesis via Cascaded Conditional Control. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 6387–6395.

Yin, L.; Han, R.; Feng, W.; and Wang, S. 2022. Self-Supervised Human Pose based Multi-Camera Video Synchronization. In *Proceedings of the ACM International Conference on Multimedia*, 1739–1748.

Zhang, F.; Zhu, X.; and Ye, M. 2019. Fast Human Pose Estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3517–3526.

Zhang, J.; Cai, Y.; Yan, S.; Feng, J.; et al. 2021. Direct Multi-view Multi-person 3D Pose Estimation. In *Advances in Neural Information Processing Systems*, volume 34, 13153–13164.

Zhang, J.; Chen, Y.; and Tu, Z. 2022. Uncertainty-Aware 3D Human Pose Estimation from Monocular Video. In *Proceedings of the ACM International Conference on Multimedia*, 5102–5113.

Zhang, T.; Huang, B.; and Wang, Y. 2020. Object-Occluded Human Shape and Pose Estimation From a Single Color Image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7376–7385.

Zhou, J.; Zhang, T.; Hayder, Z.; Petersson, L.; and Harandi, M. 2023. Diff3DHPE: A Diffusion Model for 3D Human Pose Estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2092–2102.