UNIVERSITY OF CALIFORNIA
RIVERSIDE

Temporally Continuous 3D Pose Estimation under Occlusion

A Thesis submitted in partial satisfaction
of the requirements for the degree of

Master of Science

in

Electrical Engineering

by

Rohit Lal

June 2024

Thesis Committee:

Dr. Amit K. Roy-Chowdhury, Chairperson
Dr. Greg Ver Steeg
Dr. Jiachen Li

The Thesis of Rohit Lal is approved:

Committee Chairperson

University of California, Riverside

Acknowledgments

I wish to extend my deepest appreciation to my thesis advisor, Dr. Amit Roy-Chowdhury, whose unwavering support and insightful guidance have been the cornerstone of my research journey. The countless hours of fruitful discussions under your mentorship have not only enhanced my productivity but also transformed research into a passion. Your dedication and encouragement have provided me with invaluable opportunities for growth and development as a researcher, for which I am eternally grateful.

Special thanks to Dr. Greg and Dr. Jiachen, who graciously accepted the role of serving on my thesis committee. Their exceptional teaching prowess has profoundly impacted my academic development, leaving an indelible mark on my journey. Engaging in discussions with them has been illuminating, significantly enriching and shaping my research perspectives.

My heartfelt gratitude extends to my collaborators on the BRIAR project—Arindam, Calvin, Yash, Saketh, and Hannah. I also owe a debt of gratitude to Miraj, Sarosij, Rohit Kundu, and Udita for their unwavering support, which has made this endeavor fulfilling and memorable.

I am fortunate to have been part of an outstanding community of fellow students at UCR. To Sarah, Puneet, Manoj, Abhav, Gayatri, Anirudh, Tanmayee, and Rohan—your humour, and support have enriched my Master's experience making it a period to cherish.

A special acknowledgement goes to Michael, Kathy, and Kevin, who were instrumental in helping me navigate the initial challenges of settling in Riverside, a place far from home yet made familiar through their kindness.

I am grateful to my advisor at IISc, Dr. Anirban Chakraborty, for his encouragement to pursue further studies at UCR, and to my undergraduate friends, Khush, Himanshu, and

the IvLabs community at VNIT, for inspiring me towards a path of research.

Above all, my deepest gratitude is reserved for my parents. Their steadfast support, belief in my abilities, and unconditional love have been the bedrock of my resilience and success. This journey is as much a testament to their sacrifices as it is to my endeavors.

Acknowledgement of previously published materials. The text of this thesis, in part or in full, is a reprint of the material as appeared in my previously published/submitted papers for which I am the lead author.

1. Lal, R., Bachu, S., Garg, Y., Dutta, A., Ta, C. K., Raychaudhuri, D. S., Asif, M. S., & Roy-Chowdhury, A. K. (2023). STRIDE: Single-video based Temporally Continuous Occlusion Robust 3D Pose Estimation. arXiv preprint arXiv:2312.16221.

To my parents for all the support.

ABSTRACT OF THE THESIS

Temporally Continuous 3D Pose Estimation under Occlusion

by

Rohit Lal

Master of Science, Graduate Program in Electrical Engineering
University of California, Riverside, June 2024
Dr. Amit K. Roy-Chowdhury, Chairperson

The capability to accurately estimate 3D human poses is crucial for diverse fields such as action recognition and virtual/augmented reality. However, a persistent and significant challenge within this field is the accurate prediction of human poses under conditions of severe occlusion. Traditional image-based estimators struggle with heavy occlusions due to a lack of temporal context, resulting in inconsistent predictions. While video-based models benefit from processing temporal data, they encounter limitations when faced with prolonged occlusions that extend over multiple frames. Addressing these challenges, we propose `STRIDE` (**S**ingle-video based **T**empo**R**ally cont**I**nuous occlusion Robust 3**D** Pose **E**stimation), a novel Test-Time Training (TTT) approach to fit a human motion prior for each video. This approach specifically handles occlusions that were not encountered during the model's training. We validate `STRIDE` through comprehensive experiments on challenging datasets like Occluded Human3.6M, Human3.6M, and OCMotion.

Contents

List of Figures

List of Tables

Chapter 1

Introduction and Related Works

1.0.1 Introduction

Accurate 3D pose estimation [79, 59] is an important problem in computer vision with a variety of real-world applications, including but not limited to action recognition [31], virtual and augmented reality [2], and gait recognition [82, 15]. While the performance of 3D pose estimation algorithms has improved rapidly in recent years, the majority of these are image-based [61, 55, 59], estimating the pose from a single image. Consequently, these approaches still face inherent challenges in handling occluded subjects due to the limited visual information contained in individual images. To address these issues, recent efforts have explored video-based pose estimation algorithms [74, 60], leveraging temporal continuity across frames to resolve pose ambiguities from missing visual evidence.

Further, the success of both image and video-based state-of-the-art algorithms [4, 63, 74, 52] relies heavily on supervised training on large datasets captured in controlled settings [4]. This limits generalizability, as distribution shifts in uncontrolled environments

Figure 1.1: **Effect of occlusions on pose estimation.** Image-based 3D pose estimators [4] often struggle with heavy occlusions, as illustrated in this figure. Without temporal context, predictions on highly obscured frames are inconsistent with prior poses, like the erroneous pose in the third frame. Notably, even state-of-the-art video approaches [63] fail on prolonged full occlusions spanning multiple frames, as in frames 4-5. This highlights another critical limitation - models are brittle when deployed outside their training distributions. Without training examples of such long-duration occlusions, models fail to extrapolate reasonable poses. Our work addresses this through test-time training of a human motion prior. By fine-tuning on each new video, we tailor this parametric prior to handling sequence-specific occlusion patterns not observed during training. Given an initial noisy estimate, our approach refines the pose sequence into an accurate, temporally coherent output, as shown in the final row.

can significantly degrade performance. For example, consider a scenario of an individual walking through a forest, periodically becoming fully obscured by trees, as depicted in Fig. 1.1. Image-based pose estimation methods [4] struggle in such cases, as key spatial context is lost when the person is occluded. Without additional temporal cues, the model has insufficient visual evidence to accurately determine the 3D pose [54, 53]. On the other hand, video-based approaches [52, 74, 78] also suffer from performance degradation, despite modeling temporal information, due to such prolonged occlusions being absent in the training data [7].

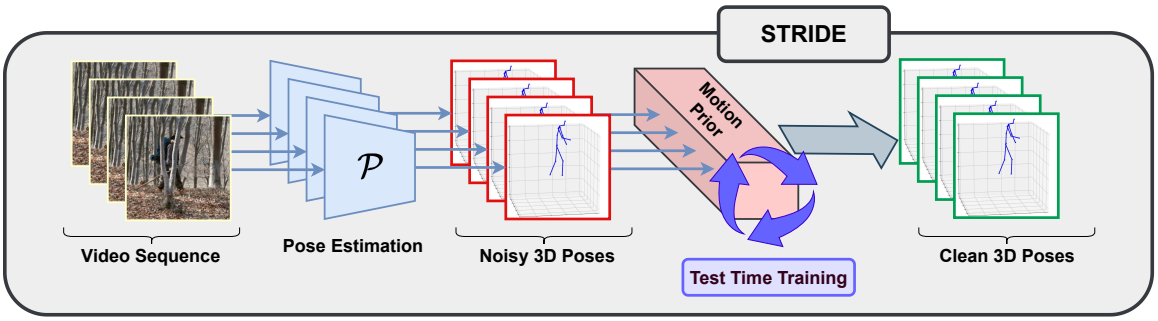


Figure 1.2: **Overview of our approach.** Our method enhances 3D pose estimation for occluded videos through test-time training of a motion prior model. We first extract initial 3D pose estimates from the test video using any 3D off-the-shelf pose estimator. To address occlusions and test distribution shifts, we then fine-tune the motion prior on that specific video by optimizing for smooth and continuous poses over the sequence.

To deal with this large diversity in contexts, occlusion patterns, and imaging conditions in real-world videos, we explore the Test-Time Training (TTT) paradigm for 3D pose estimation. TTT allows for efficient on-the-fly adaptation to the specific occlusion patterns and data distribution shifts present in each test video. This facilitates better generalization, improving the model's capability to handle even prolonged occlusions. Furthermore, this reduces reliance on large annotated datasets, which are costly to collect, especially for occluded motions.

Recent TTT approaches for 3D pose estimation [52, 51, 17, 16] fine-tune the model using 2D cues like keypoints or silhouettes extracted from the test images. However, this reliance on 2D cues has inherent limitations. Firstly, the 2D projection of 3D poses is ambiguous, as many plausible 3D configurations can map to the same 2D keypoints. Secondly, 2D pose estimators themselves are susceptible to errors on unseen data distributions [59, 27]. Thus, fine-tuning on potentially imperfect and ambiguous 2D poses can incorrectly update the model, leading to degraded 3D predictions.

To overcome the limitations of existing methods, we propose `STRIDE` (**S**ingle-video based **T**empo**R**ally cont**I**nuous occlusion Robust 3**D** Pose **E**stimation), a novel test-time training framework for 3D pose estimation under occlusion. The key component of our approach is a *parametric motion prior* that is capable of modeling the dynamics of natural human motions and poses. This motion prior is pre-trained using a BERT-style [14, 83] approach on 3D pose sequences, learning to reconstruct temporally coherent poses when given a series of noisy estimates as input. At test time, given a sequence of noisy 3D poses from any existing pose estimation algorithm, `STRIDE` leverages this pre-trained prior to produce a clean sequence by fine-tuning it on each new video. We use 3D kinematic losses for motion smoothing via adapting the model to the video-specific motion patterns. By leveraging the motion prior's inherent knowledge of natural human movement during test-time training, `STRIDE` avoids ambiguities of 2D pose information faced by existing approaches. An overview of our approach is shown in Fig. 1.2.

A key advantage of our algorithm is that it can work alongside any off-the-shelf pose estimator to improve temporal consistency, providing model-agnostic pose enhancements. This allows `STRIDE` to not only surpass image-based pose estimators that lack contextual cues to resolve occlusions, but also outperform video-based methods. Notably, `STRIDE` can handle situations with up to 100% occlusion of the human body over many consecutive frames. In comparison to existing test-time video based pose estimation method [60, 52], our approach is **up to 2 times faster** than previous state-of-the-art method [52] and operates without accessing any labeled training data during inference time, making it privacy [62] and storage-friendly.

Contributions. In summary, we make the following key contributions:

1. We propose a novel test-time training method, `STRIDE`, for achieving temporally continuous 3D pose estimation under occlusion.

2. A motion prior model that refines noisy 3D pose sequences into smooth and continuous predictions.

3. A model-agnostic framework that can refine poses from any off-the-shelf estimator, highlighting efficiency and generalizability.

4. State-of-the-art results on challenging benchmarks including Occluded Human3.6M, Human3.6M [21], and OCMotion [20]. We demonstrate enhanced occlusion robustness and temporal consistency.

1.0.2 Related Works

Monocular 3D pose estimation. Monocular 3D pose estimation is a fundamental and challenging problem in computer vision which involves the localisation of 3D spatial pose coordinates from just a single image. The problem is inherently complex due to the diversity of body shapes, clothing, self-occlusions, etc. Despite these bottlenecks, recent deep learning-based methods have shown impressive performance on challenging academic datasets [4, 79]. [46] proposed the first CNN-based approach to regress 3D joints from a single image in an end-to-end fashion. Since then, numerous works [61, 55] have improved upon these ideas by using additional information such as multi-view constraints and depth information. Recent works [71] employ kinematic constraints for improved pose estimation, while [72] uses anatomical constraints and data augmentation for obtaining state-of-the-art

results on academic datasets. However, it is worth noting that these methods are trained under supervised settings and often fail to generalize under distribution shifts. Owing to these weaknesses, [35, 37] proposed self-supervised algorithms for 3D human pose estimation. Although these works perform well in single image-based settings, they failed to generalise under occlusions and also lack temporal continuity when extended to video-based settings.

2D-3D human pose lifting. Modern 3D human pose estimation encounters significant challenges in generalization due to limited labeled data for real-world applications. [45] addressed this issue by breaking down the problem into 2D pose estimation and 2D to 3D lifting. Subsequently, [6] improved on this by including self-supervised geometric regularization, by synthetic data usage [84], spatio-temporal transformers [80], and frequency domain analysis [78]. [83] achieved state-of-the-art results by modelling motion priors from a sequence of 2D poses. Although these works perform well up to a certain degree, they suffer from two problems: 1) depth ambiguity of 2D human poses, 2) inaccurate 3D human poses if the initial 2D human poses are noisy. In contrast, we focus on 3D pose estimation in a video-based setting and does not involve any 2D-3D pose lifting.

Video-based 3D pose estimation. Video based 3D human pose estimation have demonstrated impressive performance gains on challenging datasets. Work proposed in [81] performs direct regression to 3D human poses by employing consistency between 3D joints and 2D keypoints. [56] utilized temporal convolutions for 3D human pose estimation in videos. Works like [3] exploited SMPL pose and shape parameters from videos and used it for fine tuning HMR for improved human pose estimation in the wild. Further, [76] proposed a mixed spatio-temporal approach for 3D human pose regression which alternated between

spatial consistency and temporal consistency. A recent method, HuMoR [60] performed a weighted regularization using predicted contact probabilities to maintain consistency among joint positions and joint heights across frames. The current state-of-the-art method CycleAdapt[52] handled the domain shift [32] between training and testing phases in 3D human mesh reconstruction by cyclically adapting a human mesh reconstruction network (HMRNet [22]) and a human motion denoising network (MDNet [52]) during test time. Despite the success of the above methods in maintaining temporal consistency, they are extremely slow due to an external optimization step and do not generalise well under distribution shifts. Severe occlusions often degrade the performance of these methods due to missing poses. Our work emphasizes these shortcomings and brings temporal continuity under severe occlusions by leveraging a motion-prior model that seamlessly handles missing 3D human pose estimates.

3D pose estimation under occlusion. Handling occlusions is a challenging problem, especially in video-based 3D pose estimation settings. [8] performed data augmentation using occlusion labels for 3D data using a novel Cylinder Man Model. Current methods solve this problem by refining the 3D poses to maintain temporal consistency. Recent methods like GLAMR [74] performed human mesh recovery in the global coordinate system from extracted motions in the local coordinate system and performed motion infilling for missing poses based on visible motions. SmoothNet [75] uses a temporal refinement network that takes poses from existing single image based pose estimation methods for alleviating motion jitters. Although these methods handle minor occlusions that infrequently occur in the scene, they do not perform well under heavy occlusions.

Chapter 2

Methodology

We address the problem of extracting temporally continuous 3D pose estimates from a monocular video that may contain heavy occlusions. Given an off-the-shelf monocular 3D pose estimator $\mathcal{P}$ (either image or video-based) that produces temporally inconsistent poses due to occlusions or domain gaps, our goal is to output clean, temporally coherent 3D pose sequences that better match natural human motion dynamics. To achieve this, we propose a two-stage approach, illustrated in Fig. 2.1:

1. **Learning a motion prior:** We first pre-train a self-attention-based motion prior model $\mathcal{M}$ on labeled 3D pose datasets in a BERT-style manner [14, 83]. During pre-training, we synthetically corrupt the 3D joint inputs with noise to simulate occlusions and other errors. $\mathcal{M}$ is then trained to denoise these inputs and reconstruct a sequence of temporally coherent 3D pose estimation. This allows $\mathcal{M}$ to learn very strong general priors of natural human motion dynamics.

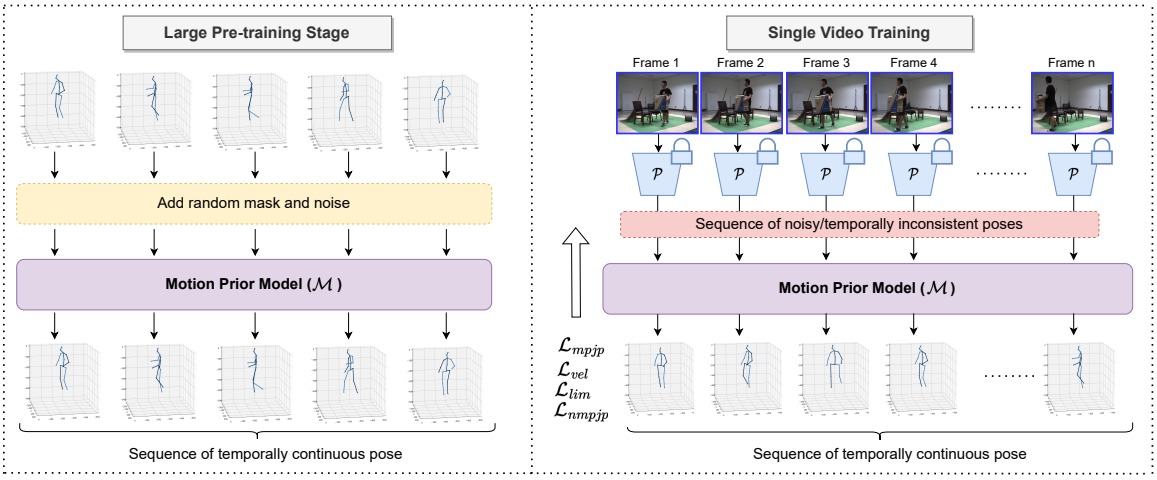


Figure 2.1: The presented figure illustrates the pipeline for our temporally continuous pose estimation, `STRIDE`. Initially, we pre-train a motion prior model, denoted as $\mathcal{M}$, using a diverse set of 3D pose data sourced from various public datasets. The primary objective of this motion prior model is to generate a sequence of poses that exhibit temporal continuity when provided with a sequence of initially noisy poses. Moving into the single video training stage, we acquire a sequence of noisy poses using a 3D pose estimation model, $\mathcal{P}$. The weights of $\mathcal{P}$ are held constant during this phase. Subsequently, we pass this noisy pose sequence through the motion prior model $\mathcal{M}$ and retrain it using various supervised losses, as outlined in Eq. 2.5. The end result of this training process is a model capable of producing temporally continuous 3D poses for that specific video.

2. **Test-time alignment:** For a given test video, we obtain potentially noisy per-frame poses using $\mathcal{P}$ [4] and fine-tune the motion prior model $\mathcal{M}$ in an unsupervised manner to align it to the specific motion exhibited in the video. This adaptation step allows us to obtain temporally continuous pose estimates for the given video.

In Section 2.0.1, we describe the architecture of the motion prior model $\mathcal{M}$. Next, in Section 2.0.2, we detail the masked sequence modelling approach used for pre-training the motion prior $\mathcal{M}$ on synthetically corrupted pose sequences. Finally, in Section 2.0.3, we introduce the self-supervised losses used for fine-tuning $\mathcal{M}$ at test time on each video.

2.0.1 Network Architecture

We base our motion prior model $\mathcal{M}$ on the DSTFormer architecture [83], originally proposed for lifting 2D poses to 3D. Here, we modify and adapt DSTFormer for the sequence-to-sequence task of denoising and smoothing noisy 3D pose sequence inputs. Specifically, the motion prior $\mathcal{M}$ takes in a sequence of 3D body poses represented as $\mathbf{X} \in \mathbb{R}^{T\times J\times 3}$, where T is the number of frames, J is the number of joints, and each pose consists of $J \times 3$ coordinate values. It then denoises the input sequence to produce refined temporally coherent 3D poses $\bar{\mathbf{X}} \in \mathbb{R}^{T\times J\times 3}$. $\mathcal{M}$ contains two key components: 1) a *spatial block* to capture the orientation of joints, and 2) a *temporal block* to model the temporal dynamics of a joint. The spatial block refines poses in each frame, while the temporal block smooths the transitions between frames. We describe these components below:

Spatial block. This block utilizes *Spatial Multi-Head Self-Attention* (S-MHSA) to model relationships among joints within each pose in the input sequence. Mathematically, the S-MHSA operation is defined as:

$$\text{S-MHSA}(\mathbf{Q}_\text{S}, \mathbf{K}_\text{S}, \mathbf{V}_\text{S}) = [\texttt{head}_1; ...; \texttt{head}_h]\mathbf{W}^P_\text{S};\ \text{head}_i = \text{softmax}(\frac{\mathbf{Q}^i_\text{S}(\mathbf{K}^i_\text{S})^\intercal}{\sqrt{d_K}})\mathbf{V}^i_\text{S}$$

Here, $\mathbf{Q}^i_\text{S}, \mathbf{K}^i_\text{S}, \mathbf{V}^i_\text{S}$ denote the query, key, and value projections for the i^{th} attention head, d_k is the key dimension, and $\mathbf{W}^P_\text{S}$ is the projection parameter matrix. We apply S-MHSA to features of different time steps in parallel. The output undergoes further processing, including residual connection and layer normalization (LayerNorm), followed by a multi-layer perceptron (MLP).

Temporal block. This block utilizes *Temporal Multi-Head Self-Attention* (T-MHSA) to model the relationships between poses across time steps, thereby enabling the smoothing of

the pose trajectories over the sequence. It operates similarly to S-MHSA but is applied to per-joint temporal features parallelized over the spatial dimension:

$$\text{T-MHSA}(\mathbf{Q}_\text{T}, \mathbf{K}_\text{T}, \mathbf{V}_\text{T}) = [\texttt{head}_1; ...; \texttt{head}_h]\mathbf{W}^P_\text{T};\ \texttt{head}_i = \text{softmax}(\frac{\mathbf{Q}^i_\text{T}(\mathbf{K}^i_\text{T})^\intercal}{\sqrt{d_K}})\mathbf{V}^i_\text{T}$$

By attending to temporal relationships, T-MHSA produces smooth pose transitions over time.

Dual-Stream Spatio-temporal Transformer. We then use the dual-stream architecture which employs spatial and temporal Multi-Head Self-Attention mechanisms. These mechanisms capture intra-frame and inter-frame body joint interactions, necessitating careful consideration of three key assumptions: both streams model comprehensive spatio-temporal contexts, each stream specializes in distinct spatio-temporal aspects, and the fusion dynamically balances weights based on input characteristics.

2.0.2 Learning a Motion Prior

To build a strong prior for human motion dynamics, we draw inspiration from the success of large language models like BERT [14] that leverage large-scale self-supervised pre-training. Here, we extend this paradigm to 3D human pose estimation. Specifically, given a dataset of 3D pose sequences, we synthetically mask these sequences to simulate occlusions and other errors. Similar to [12, 83], the prior $\mathcal{M}$ is trained to denoise these noisy inputs to reconstruct a sequence of temporally coherent 3D poses.

During pre-training, we apply both joint-level and frame-level masking to a 3D pose sequence $\mathbf{X}$ to obtain a corrupted sequence $\texttt{mask}(\mathbf{X})$ which mimics realistic scenarios of imperfect predictions and occlusions. The prior $\mathcal{M}$ is trained to reconstruct the complete 3D

motion sequence $\bar{\mathbf{X}}$ from this corrupted input $\mathbf{X}$ by minimizing losses on 3D joint positions $\mathcal{L}_{\text{3D}}$ between the reconstruction and the ground-truth pose. Additionally, we incorporate a velocity loss $\mathcal{L}_{\text{O}}$ following [56, 76].

$$\mathcal{L}_{\text{3D}} = \sum_{t=1}^{T} \sum_{j=1}^{J} \parallel \bar{\mathbf{X}}_{t,j} - \mathbf{X}_{t,j} \parallel_2 \qquad \mathcal{L}_{\text{O}} = \sum_{t=2}^{T} \sum_{j=1}^{J} \parallel \bar{\mathbf{O}}_{t,j} - \mathbf{O}_{t,j} \parallel_2$$

where $\bar{\mathbf{O}}_t = \bar{\mathbf{X}}_t - \bar{\mathbf{X}}_{t-1}$, $\mathbf{O}_t = \mathbf{X}_t - \mathbf{X}_{t-1}$.

2.0.3 Test-Time Alignment

Given the pre-trained motion prior model $\mathcal{M}$ that takes in noisy 3D poses and outputs temporally coherent predictions, our goal is to leverage this for pose estimation on new test videos. We first obtain an initial noisy estimate of the 3D pose sequence using any off-the-shelf pose detector $\mathcal{P}$ [4]. As these models struggle on occlusions and distribution shifts, their outputs lack temporal consistency. To address this, we pass the noisy poses through $\mathcal{M}$ to achieve a refined estimate.

Although the prior refines pose, some inconsistencies like domain shift and novel human motion may be present in the videos. Hence, we propose additional test-time training of $\mathcal{M}$ using geometric and physics-based constraints to adapt to such situations. Similar to internal learning approaches like Deep Video Prior [41], our proposed self-supervision strategy fine-tunes the motion prior to the specifics of each test video for enhanced outputs. Specifically, we utilize four different losses that regularize (1) the velocity of joints, (2) scale variations in predictions (3) the size of limbs, and (4) the smoothness of poses (5) in missing frames. Crucially, only $\mathcal{M}$ is updated during test-time training while $\mathcal{P}$ remains fixed to preserve the pose estimation capabilities of off-the-shelf models.

Limb loss: Limb length consistency is an important aspect of anatomically plausible 3D human pose predictions. This loss encourages the model to produce temporally stable limb lengths, contributing to more realistic and physically plausible pose estimations. The idea is to penalize variability in limb lengths across frames. If the limb lengths exhibit large variations, it may indicate inconsistency or instability in the predicted poses. The limb loss function $\mathcal{L}_{\text{lim}}$ is defined as follows,

$$\mathcal{L}_{\text{lim}} = \frac{1}{J}\sum_{j=1}^{J-1} \underbrace{\frac{1}{T}\sum_{t=1}^{T}\left(\boldsymbol{\mathcal{J}}_{t,j} - \frac{1}{T}\sum_{t'=1}^{T}\boldsymbol{\mathcal{J}}_{t',j}\right)^2}_{\text{Variance of Joint Lengths Across Time}} . \tag{2.1}$$

Here $\boldsymbol{\mathcal{J}} \in \mathbb{R}^{T\times(J-1)}$ represents a matrix of the normalised length of limb $j < (J-1)$ at any time $t < T$. By calculating the variance of limb lengths and taking the mean, the loss encourages the model to produce more consistent and stable limb lengths across the entire sequence. This can be beneficial in applications where it is crucial to maintain anatomical consistency in the predicted 3D poses.

To further regularize for the cases where the 3D pose estimation model $\mathcal{P}$ fails to detect any pose, we use linear interpolation between joints. Consider that the video consists of N frames, out of which the model fails to predict anything for q frames. The linear extrapolation and interpolation function $L : \mathbb{R}^{(N-q)\times J\times 3} \rightarrow \mathbb{R}^{N\times J\times 3}$ fills in the missing inputs. This provides pseudo-labels during training for two of our loss functions. These pseudo-labels also help to ensure temporal continuity in the predicted poses.

Mean Per Joint Position (MPJP) loss: This loss focuses on the accuracy of the pose estimation by penalizing deviations in the spatial position of individual joints. It computes the mean Euclidean distance between the predicted $\hat{\mathbf{X}}$ poses and pseudo-poses $\tilde{\mathbf{X}} = L(\hat{\mathbf{X}})$

where $\hat{\mathbf{X}}$ is the noisy sequence of poses obtained from $\mathcal{P}$. It measures the average distance between corresponding joints in the predicted and pseudo labels. It is defined as follows,

$$\mathcal{L}_{\text{MPJP}} = \frac{1}{T \cdot J \cdot 3} \sum_{t=1}^{T} \sum_{j=1}^{J} \sum_{d=1}^{3} \|\hat{\mathbf{X}}_{t,j,d} - \tilde{\mathbf{X}}_{t,j,d}\|_2 \tag{2.2}$$

Normalized MPJP (N-MPJP) loss: This loss function introduces a normalization step to address scale variations between the predicted and target poses. It calculates the scale factor based on the norms of the predicted and target poses and then applies this scale factor to the predicted poses before computing the MPJPE. The normalization in $\mathcal{L}_{\text{N-MPJP}}$ aims to make the model more robust to variations in absolute pose values. It is particularly useful when the scale of the poses in the training and testing data may differ. By incorporating scale information, $\mathcal{L}_{\text{N-MPJP}}$ addresses scale-related issues during training, potentially improving the model's generalization to different scenarios.

$$\mathcal{L}_{\text{NMPJP}} = \mathcal{L}_{\text{MPJP}}(s\hat{\mathbf{X}}, \tilde{\mathbf{X}}); \text{ where } s = \frac{\sum_{t=1}^{T} \sum_{j=1}^{J} \sum_{d=1}^{3} \left\| \tilde{\mathbf{X}}_{t,j,d} \cdot \hat{\mathbf{X}}_{t,j,d} \right\|_2}{\sum_{t=1}^{T} \sum_{j=1}^{J} \sum_{d=1}^{3} \left\| \hat{\mathbf{X}}_{t,j,d} \right\|_2^2} \tag{2.3}$$

In Equation 2.3, s represents the scale. The combination of both $\mathcal{L}_{\text{NMPJP}}$ and $\mathcal{L}_{\text{MPJP}}$ losses allows the model to simultaneously optimize for accurate joint positions ($\mathcal{L}_{\text{MPJP}}$) and address scale variations ($\mathcal{L}_{\text{NMPJP}}$). The incorporation of $\mathcal{L}_{\text{NMPJP}}$ allows the model to learn to handle scenarios where the pose scale may differ between training and testing data.

Velocity loss: We optimize velocity loss similar to Eq. 2.4, but instead of ground truth, we use pseudo-labels.The velocity loss helps in smoothing the movement and removing unwanted jittering across frames.

$$\mathcal{L}_{\text{vel}} = \frac{1}{N \cdot (J-1)} \sum_{t=1}^{T-1} \sum_{j=1}^{J} \sum_{d=1}^{3} \left\| \hat{\mathbf{V}} - \tilde{\mathbf{V}} \right\|_2 \tag{2.4}$$

where $\hat{\mathbf{V}} = \hat{\mathbf{X}}_{t+1,j,d} - \hat{\mathbf{X}}_{t,j,d}$ and $\tilde{\mathbf{V}} = \tilde{\mathbf{X}}_{t+1,j,d} - \tilde{\mathbf{X}}_{t,j,d}$ represent velocities of predicted poses and pseudo label poses respectively.

Overall Loss. In summary, by combining all the above-mentioned losses into one final loss function as shown in Eq. 2.5, $\mathcal{M}$ is trained to produce accurate joint positions, maintain anatomical consistency, and handle scale variations,

$$\mathcal{L}_{total} = \lambda_1 \mathcal{L}_{mpjp} + \lambda_2 \mathcal{L}_{vel} + \lambda_3 \mathcal{L}_{lim} + \lambda_4 \mathcal{L}_{nmpjp} \tag{2.5}$$

Here, λ_i, where $i \in 1, 2, 3, 4$, refers to loss-weighing hyper-parameters which remain constant for all evaluations.

Chapter 3

Experimentation

In this section, our primary objective is to provide a comprehensive understanding of our approach. We elaborate on the datasets employed and conduct a thorough comparison with state-of-the-art methodologies. Furthermore, we analyze the qualitative results, pinpointing areas where existing methods may falter. As a conclusive step, we perform an ablation study to assess the impact of pre-training and different loss functions, shedding light on their contributions to our experimental framework.

We conduct evaluations on three datasets with varying levels of occlusion: Human3.6M, representing scenarios without occlusion; OCMotion, moderate occlusion; and Occluded Human3.6M, representing heavy occlusion. The metrics assessed include Procrustes-aligned mean per joint position error (PA-MPJPE), mean per joint position error (MPJPE), and acceleration error (Accel), measured as the disparity in acceleration between ground-truth and predicted 3D joints. We report the metrics in (mm). We use BEDLAM-CLIFF [4] as the off-the-shelf pose estimation method. We compare the error rates of `STRIDE` and the

baseline methods in Table 3.1,3.2 and 3.3. The best results are in **bold** and green arrows indicate the percentage improvement over the previous state-of-the-art method.

3.0.1 Datasets

Human3.6M [21]. An indoor-scene dataset, Human3.6M is a pivotal benchmark for 3D human pose estimation from 2D images. Captured with a 4-camera setup, it includes 11 subjects, each with 15 different actions, annotated in the 17 keypoints format. Following [4], we retain every 1 in 5 frames in the test split comprising the S9 and S11 sequence. We perform experiments on the original publically available Human3.6M dataset to show that our method achieves comparable performance with other state-of-the-art methods.

OCMotion [20]. OCMotion is a video dataset that extends the 3DOH50K image dataset [77], incorporating natural occlusions. The dataset comprises 300K images captured at 10 FPS, featuring 43 sequences observed from 6 viewpoints. Its annotations for 3D motion include SMPL, 2D poses, and camera parameters. The sequences {0013, 0015, 0017, 0019} are designated for testing. Our method does not require supervised training, so we have only used the test split when performing all experiments.

Occluded Human3.6M. We curate the Occluded Human3.6M dataset to evaluate our method, specifically designed for assessing human pose estimation under significant occlusion, unlike existing datasets such as Human3.6M, MPI-INF-3DHP[47], and 3DPW[67]. To accomplish this, we use random erase occlusions on Human3.6M videos, completely covering a person up to 100%. These occlusions persist spatially and temporally for 1.6 seconds within 3.2 seconds of the video.

BRIAR [13]. BRIAR is a large-scale biometric dataset featuring videos of human subjects captured in extremely challenging conditions. These videos are recorded at varying distances *i.e* close range, 100m, 200m, 400m, 500m, and unmanned aerial vehicles (UAV), with each video lasting around 90 seconds. Most of the pose estimation methods fail on this dataset due to the extreme amount of domain shifts. Additionally, BRIAR lacks ground truth data for poses, which means evaluations of pose estimation methods on this dataset can only be qualitative, relying on visual assessments rather than quantitative metrics.

3.0.2 Quantitative Results

Table 3.1: 3D Pose estimation results on **Occluded Human3.6M**. This dataset is crucial as it is the only dataset that has significant occlusion. The results underscore that our method surpasses all state-of-the-art with substantial percentage improvements, affirming its robustness in handling occlusions.

	Method	PA-MPJPE	MPJPE	Accel
Image	CLIFF [42]	183.5	100.5	38.4
	BEDLAM [4]	179.5	98.9	39.1
Video	GLAMR [74]	213.9	380.3	42.3
	PoseFormerV2 [78]	193.9	260.2	38.7
	CycleAdapt [52]	77.6	132.6	48.7
	MotionBERT [83]	76.1	112.8	28.7
	Our Method	**59.0** (57%↓)	**80.7** (18%↓)	**26.6** (7%↓)

Our method is most effective under heavy occlusions. We significantly outperform other state-of-the-art methods on the Occluded Human3.6M dataset as shown in Table 3.1. Notably, `STRIDE` performs significantly better than BEDLAM despite using pseudo-labels

Table 3.2: **3D pose estimation results on OCMotion [20]**. Our method outperforms other image and video-based pose estimation methods. While PoseFormerV2 has the lowest accel., it also exhibits the highest PA-MPJPE error. This is due to oversmoothing and inaccurate interpolation between poses which compromises the pose estimation accuracy.

	Method	PA-MPJPE	Accel	Avg
Image	OOH [77]	55.0	48.6	51.8
	PARE [29]	52.0	43.6	47.8
	BEDLAM [4]	47.1	49.0	48.0
Video	PoseFormerV2 [78]	126.3	**28.5**	77.4
	GLAMR [74]	89.9	51.3	70.6
	CycleAdapt [52]	74.6	57.5	66.0
	ROMP [64]	48.1	57.2	52.6
	Our Method	**46.2** (2%↓)	47.8	**47.0** (2%↓)

from BEDLAM. BEDLAM fails to produce poses under heavy occlusion; hence, the evaluation results drop significantly. However, since `STRIDE` incorporates temporal information to address these gaps in the video, we predict reasonable poses even in case of heavy occlusions and improve the result of BEDLAM by a significant margin. It is important to note that by using `STRIDE` we do not only outperform BEDLAM, but we also outperform all the other existing video- and image-based state-of-the-art methods. This is mainly because existing methods do not incorporate human motion prior and hence results in temporally implausible poses.

Since Occluded Human3.6M contains artificial occlusions, we also evaluated on the OCMotion dataset, which contains real-world, natural occlusions. Table 3.2 shows that

Table 3.3: 3D pose estimation results on **Human3.6M**. Our evaluation demonstrates that our results are comparable to the BEDLAM-CLIFF baseline. This is due to the occlusion-free nature of the Human3.6M, which yields already refined and consistent poses with limited room for improvement.

	Method	PA-MPJPE	MPJPE	Accel
Image	CLIFF [42]	56.1	89.6	-
	BEDLAM-HMR [4]	51.7	81.6	-
	BEDLAM-CLIFF [4]	50.9	70.9	39.14
Video	GLAMR [74]	-	-	-
	CycleAdapt [52]	64.5	106.3	57.25
	MotionBERT* [83]	64.15	95.8	**14.8**
	Our Method	**50.4** (1%↓)	**69.7** (2%↓)	37.1

our approach `STRIDE` attains state-of-the-art results on the OCMotion dataset [20]. Since we obtained good pseudo-labels from BEDLAM under partial occlusions, we observe the proximity of our results to BEDLAM. *It is important to highlight that methods such as [64, 29] are supervised and trained on the training split of OCMotion. In contrast, our approach does not assume access to any labeled training dataset.*

Our method demonstrates minor improvement over BEDLAM-CLIFF on the original Human3.6M dataset, as evidenced in Table 3.3. The marginal enhancement is primarily due to the nature of the Human3.6M dataset, which lacks occlusions, thereby limiting the potential for improvement beyond the baseline.

Inference speed: Table 3.4 compares the inference times of various 3D pose estimation methods on a 243-frame OCMotion video using an RTX 3090 GPU. HuMor and GLAMR

Table 3.4: Inference time for various 3D pose estimation methods.

Method	Time (sec)
HuMor [60]	> 600
GLAMR [74]	> 600
PoseFormerV2 [78]	129
CycleAdapt [52]	126
Our Method	**68** (46%↓)

are notably slower, exceeding 10 minutes due to their intensive pose optimization phase. In contrast, PoseFormerV2 and CycleAdapt show efficiency improvements with inference times of 129 and 126 seconds, respectively. `STRIDE` outperforms these, achieving a significant reduction to 68 seconds, making it 46% faster and highlighting its suitability for real-time applications without sacrificing accuracy.

3.0.3 Qualitative Results

Our evaluation juxtaposes `STRIDE` against leading state-of-the-art techniques like CycleAdapt [52]. Key insights from our comparison include:

Occluded Human3.6M: Traditional approaches often fall short in accurately predicting missing 3D poses, struggling with high levels of occlusion. In contrast, our method utilizes the dynamics of human motion to precisely infill missing poses, leading to a 57% error improvement in performance compared to the former state-of-the-art method. These improvements can be visualized in Fig. 3.1.

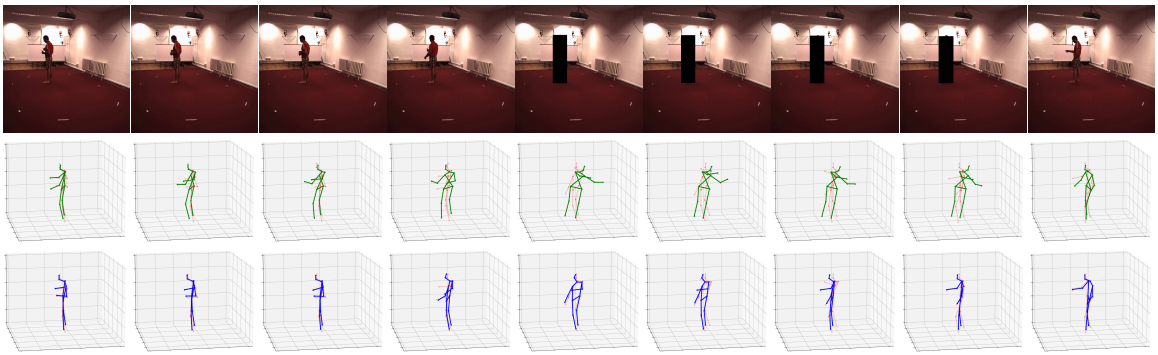


Figure 3.1: **3D pose estimation results on Occluded Human3.6M**. CycleAdapt (*second* row) fails to generalize in cases when there is complete occlusion. `STRIDE` (*third* row) produces temporally coherent pose infilling due to test time training. Note that the translucent red color represents the ground truth poses.

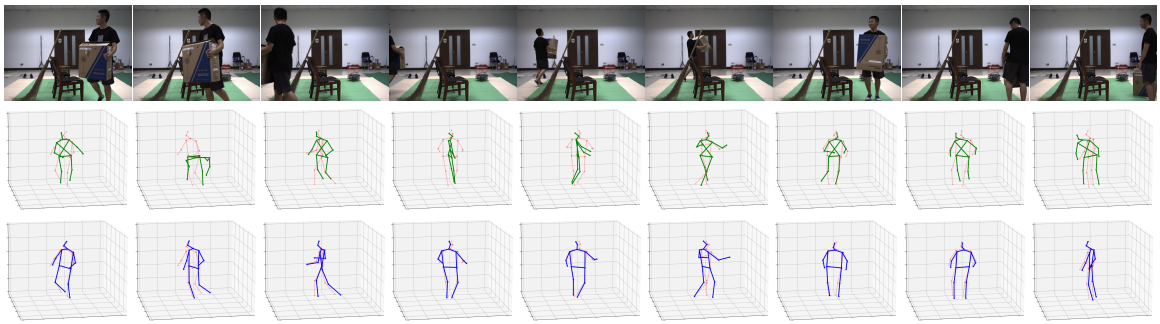


Figure 3.2: **3D pose estimation results on OCMotion (*0013, Camera01*)**. This figure demonstrates how our method incorporates temporal continuity into video sequences under occlusion. The *second* row represents 3D poses predicted by CycleAdapt [52]. The *third* row represents 3D poses predicted by `STRIDE`. Note: The 3D poses shown in translucent red color in the *second* and *third* row represent the ground truths.

BRIAR [13]: The videos within the BRIAR dataset present a substantial domain shift, a scenario not previously encountered by existing methodologies. Our algorithm distinguishes itself by mitigating these distribution shifts, resulting in markedly superior performance. While other techniques yield almost random predictions under these conditions, our method dynamically adapts to this domain shift during test time. Although direct quantitative comparisons are impossible due to the absence of ground truth, the visual comparisons provided through our videos demonstrate our method's enhanced efficacy.

OCMotion [19]: In Fig. 3.5, we compare our method against an existing state-of-the-art pose estimation method CycleAdapt [52]. In Frame 5, we can observe that CycleAdapt fails to perform well in cases when there is self-occlusion. We observe that `STRIDE`'s predictions are best aligned with the ground truth poses, even under significant occlusions or when the person goes out of the frame.

3.0.4 Ablation Study

An ablation study conducted in Table 3.5 provides quantitative insights into the significance of each component in `STRIDE`. Starting from a baseline with substantial errors, the introduction of a motion prior alone drastically improves performance, underscoring its effectiveness in driving the model toward realistic human pose dynamics. The addition of L_{mpjp} enhances spatial accuracy, further lowering MPJPE to 82.1 and PA-MPJPE to 60.4. The improvement with L_{vel} suggests its role in smoothing motion. The best results are observed when L_{nmpjp} is also included, indicating its critical function in accounting for scale variations.

In conclusion, the ablation study reveals that each component contributes to improving the accuracy and temporal consistency of the pose estimations, with the full combination of components yielding the state-of-the-art results. This indicates that while the motion prior sets a strong foundation for plausible poses, the various loss functions refine and stabilize the pose predictions to align closely with natural human movement dynamics and unseen poses. We find that using any off-the-shelf pose estimation method yields similar improvements, thereby making `STRIDE` agnostic to any specific 3D pose estimation method.

Table 3.5: **Ablation study** This table demonstrates how the inclusion of a pre-trained motion prior and various losses collectively contributes to the model's accuracy on Occluded Human3.6M dataset.

Prior	$\mathcal{L}_{mpjp}$	$\mathcal{L}_{vel}$	$\mathcal{L}_{lim}$	$\mathcal{L}_{nmpjp}$	MPJPE	PA-MPJPE
✗	✗	✗	✗	✗	179.5	98.9
✓	✗	✗	✗	✗	106.5	80.2
✓	✓	✗	✗	✗	82.1	60.4
✓	✓	✓	✗	✗	81.4	59.6
✓	✓	✓	✓	✗	81.1	59.6
✓	✓	✓	✓	✓	**80.7**	**59.0**

3.1 Implementation Details

We implement the proposed motion encoder DSTformer with depth N = 5, number of heads h = 8, feature size = 512, embedding size = 512. For pretraining, we use sequence length T = 243. The pretrained model could handle different input lengths thanks to the transformer-based backbone. During finetuning, we set the backbone learning rate to be 0.1× of the new layer learning rate.

Setup. We have implemented the proposed model using PyTorch. For our experiments, we utilized a CentOS machine equipped with 4 NVIDIA 3090 GPUs, specifically designed for accelerating pretraining tasks. It's worth noting that for finetuning and inference processes, a single GPU typically proves to be more than adequate.

Pretraining. We do large scale pertaining using AMASS and Training split of Human3.6M. For the implementation of AMASS [44], we initiate the process by rendering the parameterized human model SMPL+H. Subsequently, we extract 3D keypoints using a

predefined regression matrix. The extraction of 3D keypoints from the Human3.6M dataset is accomplished through camera projection. Motion clips with a length of $T = 243$ are sampled for the 3D mocap data. The input channels are set to $C_{\text{in}} = 3$, representing the x and y coordinates along with confidence values. Data augmentation is applied through random horizontal flipping.

The entire network undergoes training for a total of 90 epochs, employing a learning rate of 0.0005 and a batch size of 64, facilitated by the Adam optimizer. The weights assigned to the loss terms are parameterized by $\lambda_{\text{O}} = 20$. Additionally, we set the 3D skeleton masking ratio to 15%, aligning with BERT's configuration. This involves using 10% frame-level masks and 5% joint-level masks. Despite variations in the proportion of these mask types, only marginal differences are observed.

To ensure the smoothness of the noise and prevent severe jittering, we initially sample noise $\boldsymbol{z} \in \mathbb{R}^{T_K \times J}$ for $T_K = 27$ keyframes. Subsequently, we upsample it to $\boldsymbol{z}' \in \mathbb{R}^{T \times J}$ and introduce a small Gaussian noise $\mathcal{N}(0, 0.002^2)$.

3D Pose Estimation. We conduct training during the inference stage for a duration of 30 epochs, employing the following hyperparameters: Batch size: 1, Learning rate: 0.0002, Weight decay: 0.01, Learning rate decay: 0.99

The total loss, denoted as

$$\mathcal{L}_{\text{total}} = \lambda_1 \mathcal{L}_{\text{mpjp}} + \lambda_2 \mathcal{L}_{\text{vel}} + \lambda_3 \mathcal{L}_{\text{lim}} + \lambda_4 \mathcal{L}_{\text{nmpjp}},$$

is comprised of multiple components, each weighted by specific coefficients. For this configuration, we set the weights as follows: $\lambda_1 = 1, \quad \lambda_2 = 20, \quad \lambda_3 = 200, \quad \lambda_4 = 0.5.$

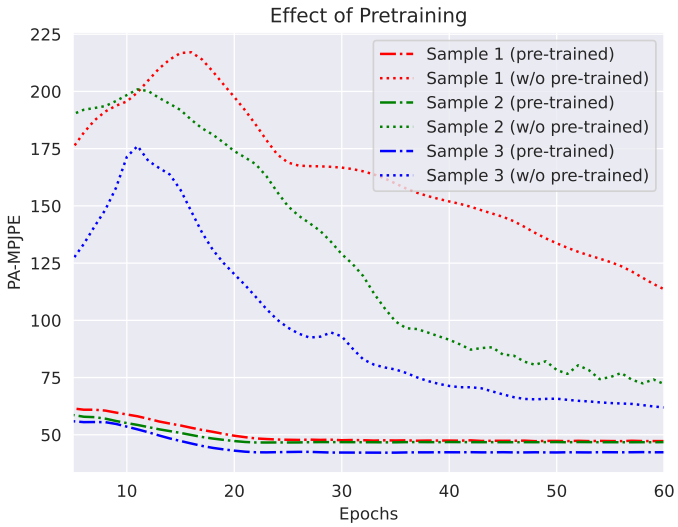


Figure 3.3: **Effect of large-scale pre-training**. We take 5 random samples from Occluded Human3.6M and try to align DSTFormer architecture. We find that when DSTFormner is initialised with motion-prior weights it converges faster.

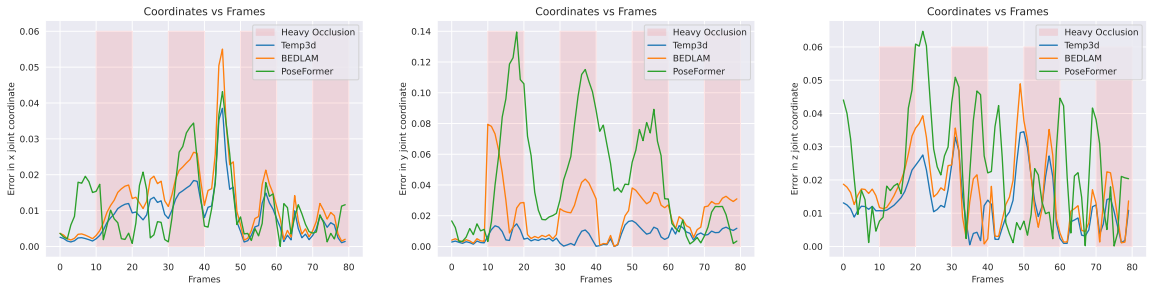


Figure 3.4: The figure above, from left to right, illustrates the variation in error values across the x, y, and z coordinates within a single video. Notably, `STRIDE` exhibits relatively lower error, particularly in scenarios involving occlusion. Furthermore, for y-coordinate, it is evident that the error demonstrates a remarkable level of smoothness.

3.2 Temporal Smoothness

The existing metric falls short in capturing temporal smoothness or assessing errors during occlusion. Additionally, there's a likelihood that a model excelling in occluded scenarios might not significantly impact overall performance if non-occluded cases dominate the results. This becomes particularly apparent in cases of sporadic temporal occlusion.

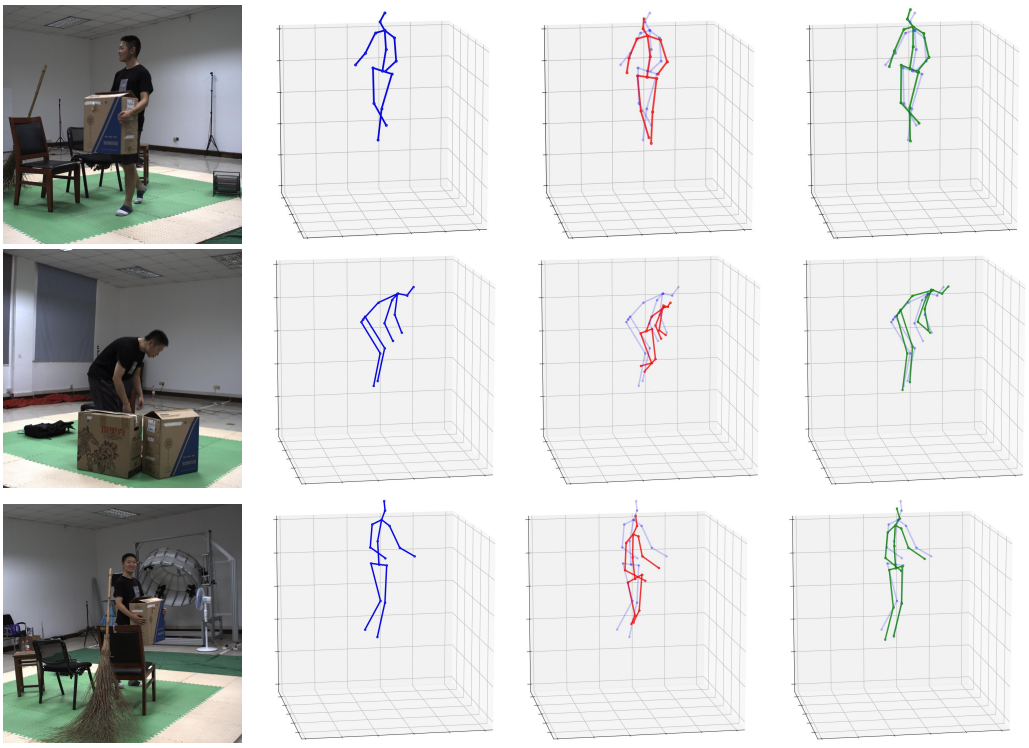
To address this issue and gain deeper insights into predictions during occlusions, we visualize various errors in Fig. 3.4. This plot illustrates how the error in the x, y, and z coordinates evolves in a video featuring occlusions. Notably, other methods demonstrate subpar performance during occlusions, with the error in the x and z coordinates being relatively minimal, exerting less influence on the final error. In contrast, the y-coordinate error predominantly contributes to the overall error, where `STRIDE` stands out by consistently having the least amount of error. The noteworthy aspect is the sustained and consistent performance throughout the occluded duration.

Method	PA-MPJPE	MPJPE
PoseFormerV2 [78]	193.9	260.2
MotionBERT [83]	76.1	112.8
BEDLAM [4]	179.5	98.9
BEDLAM Interpolation [4]	64.1	83.3
`STRIDE` (ours)	**59.0**	**80.7**

Table 3.6: Quantitative comparison of 3D Pose estimation methods on **Occluded Human3.6M**.

3.3 Additional Qualitative Comparisons

In Fig. 3.5 we compare our method against a different state-of-the-art 3D pose estimation method named PoseFormerV2 [78]. One trivial way to improve the results of BEDLAM is by linear interpolation between frames. However, we found that just interpolation was not enough as it may miss the results. Our loss optimization during



(a) OCMotion Image (b) Ground Truth (c) PoseFormerV2 (d) `STRIDE` (ours)

Figure 3.5: This figure shows how our method works when tested in natural occlusion cases. The translucent blue color in the *second column*, *third column*, and *fourth column* represents the ground truth. Blue, red, and green similarly represent Ground Truth, PoseformerV2 and `STRIDE` results, respectively.

inference helps to achieve the best results. Interpolation results are shown in Table 3.6.

We provide videos where we compare `STRIDE` against CycleAdapt and GLAMR here

`https://bit.ly/stride_results`

Chapter 4

Conclusion and Future Works

In conclusion, while existing 3D human pose estimation methods excel in various scenarios, they struggle with handling significant occlusions. In this work, we introduce `STRIDE`, an unsupervised approach which utilizes large-scale pre-training, self-supervised learning, and temporal context to enhance 3D pose estimation for a single video containing occlusions during the test time. `STRIDE` achieves state-of-the-art results on datasets that contain significant human body occlusions such as Occluded Human3.6M and OCMotion thus demonstrating improved occlusion robustness. Currently, a limitation of `STRIDE` is that it can only extract temporally continuous 3D poses when there are no human-to-human occlusions. Future work will focus on adapting `STRIDE` for multi-person occlusion scenarios. Future work can also involve the using temporally continuous pose estimates to enhance downstream tasks such as action recognition, mesh recovery, and gait recognition.

Bibliography

[1] Khush Agrawal, Rohit Lal, Himanshu Patil, Surender Kannaiyan, and Deep Gupta. Deepsct: Deep learning based self correcting object tracking mechanism. In *2021 National Conference on Communications (NCC)*, pages 1–6. IEEE, 2021.

[2] Taravat Anvari and Kyoungju Park. 3d human body pose estimation in virtual reality: A survey. In *2022 13th International Conference on Information and Communication Technology Convergence (ICTC)*, pages 624–628, 2022.

[3] Anurag Arnab, Carl Doersch, and Andrew Zisserman. Exploiting temporal context for 3d human pose estimation in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3395–3404, 2019.

[4] Michael J. Black, Priyanka Patel, Joachim Tesch, and Jinlong Yang. Bedlam: A synthetic dataset of bodies exhibiting detailed lifelike animated motion, 2023.

[5] Ju Yong Chang, Gyeongsik Moon, and Kyoung Mu Lee. Poselifter: Absolute 3d human pose lifting network from a single noisy 2d human pose, 2020.

[6] Ching-Hang Chen, Ambrish Tyagi, Amit Agrawal, Dylan Drover, Rohith Mv, Stefan Stojanov, and James M Rehg. Unsupervised 3d pose estimation with geometric self-supervision. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5714–5724, 2019.

[7] Yu Cheng, Bo Yang, Bo Wang, Wending Yan, and Robby T Tan. Occlusion-aware networks for 3d human pose estimation in video. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 723–732, 2019.

[8] Yu Cheng, Bo Yang, Bo Wang, Wending Yan, and Robby T. Tan. Occlusion-aware networks for 3d human pose estimation in video. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019.

[9] Hongsuk Choi, Gyeongsik Moon, Ju Yong Chang, and Kyoung Mu Lee. Beyond static features for temporally consistent 3d human pose and shape from a video. In *CVPR*, 2021.

[10] Vasileios Choutas, Lea Muller, Chun-Hao P. Huang, Siyu Tang, Dimitrios Tzionas, and Michael J. Black. Accurate 3d body shape regression using metric and semantic attributes, 2022.

[11] Xiao Chu, Wei Yang, Wanli Ouyang, Cheng Ma, Alan L. Yuille, and Xiaogang Wang. Multi-context attention for human pose estimation, 2017.

[12] Hai Ci, Mingdong Wu, Wentao Zhu, Xiaoxuan Ma, Hao Dong, Fangwei Zhong, and Yizhou Wang. Gfpose: Learning 3d human pose prior with gradient fields, 2022.

[13] David Cornett, Joel Brogan, Nell Barber, Deniz Aykac, Seth Baird, Nicholas Burchfield, Carl Dukes, Andrew Duncan, Regina Ferrell, Jim Goddard, et al. Expanding accurate person recognition to new altitudes and ranges: The briar dataset. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 593–602, 2023.

[14] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding, 2019.

[15] Arindam Dutta, Rohit Lal, Dripta S Raychaudhuri, Calvin-Khang Ta, and Amit K Roy-Chowdhury. Poise: Pose guided human silhouette extraction under occlusions. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 6153–6163, 2024.

[16] Shanyan Guan, Jingwei Xu, Michelle Zhang He, Yunbo Wang, Bingbing Ni, and Xiaokang Yang. Out-of-domain human mesh reconstruction via dynamic bilevel online adaptation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(4):5070–5086, 2022.

[17] Shanyan Guan, Jingwei Xu, Yunbo Wang, Bingbing Ni, and Xiaokang Yang. Bilevel online adaptation for out-of-domain human mesh reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10472–10481, 2021.

[18] Mir Rayat Imtiaz Hossain and James J. Little. *Exploiting Temporal Information for 3D Human Pose Estimation*, page 69–86. Springer International Publishing, 2018.

[19] Buzhen Huang, Yuan Shu, Jingyi Ju, and Yangang Wang. Occluded human body capture with self-supervised spatial-temporal motion prior. *arXiv preprint arXiv:2207.05375*, 2022.

[20] Buzhen Huang, Tianshu Zhang, and Yangang Wang. Object-occluded human shape and pose estimation with probabilistic latent consistency. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.

[21] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. Human3.6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(7):1325–1339, jul 2014.

[22] Angjoo Kanazawa, Michael J. Black, David W. Jacobs, and Jitendra Malik. End-to-end recovery of human shape and pose. In *Computer Vision and Pattern Recognition (CVPR)*, 2018.

[23] Angjoo Kanazawa, Michael J. Black, David W. Jacobs, and Jitendra Malik. End-to-end recovery of human shape and pose, 2018.

[24] Angjoo Kanazawa, Jason Y. Zhang, Panna Felsen, and Jitendra Malik. Learning 3d human dynamics from video. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5607–5616, 2019.

[25] Yangyuxuan Kang, Yuyang Liu, Anbang Yao, Shandong Wang, and Enhua Wu. 3d human pose lifting with grid convolution. *arXiv preprint arXiv:2302.08760*, 2023.

[26] Isinsu Katircioglu, Bugra Tekin, Mathieu Salzmann, Vincent Lepetit, and Pascal Fua. Learning latent representations of 3d human pose with deep neural networks. *International Journal of Computer Vision*, 126, 12 2018.

[27] Donghyun Kim, Kaihong Wang, Kate Saenko, Margrit Betke, and Stan Sclaroff. A unified framework for domain adaptive pose estimation. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXIII*, pages 603–620. Springer, 2022.

[28] Muhammed Kocabas, Nikos Athanasiou, and Michael J Black. Vibe: Video inference for human body pose and shape estimation. In *CVPR*, 2020.

[29] Muhammed Kocabas, Chun-Hao P Huang, Otmar Hilliges, and Michael J Black. Pare: Part attention regressor for 3d human body estimation. In *ICCV*, 2021.

[30] Nikos Kolotouros, Georgios Pavlakos, Michael J Black, and Kostas Daniilidis. Learning to reconstruct 3d human pose and shape via model-fitting in the loop. In *ICCV*, 2019.

[31] Yu Kong and Yun Fu. Human action recognition and prediction: A survey. *International Journal of Computer Vision*, 130(5):1366–1401, 2022.

[32] Vikash Kumar, Rohit Lal, Himanshu Patil, and Anirban Chakraborty. Conmix for source-free single and multi-target domain adaptation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 4178–4188, 2023.

[33] Vikash Kumar, Himanshu Patil, Rohit Lal, and Anirban Chakraborty. Improving domain adaptation through class aware frequency transformation. *International Journal of Computer Vision*, 131(11):2888–2907, 2023.

[34] Vikash Kumar, Sarthak Srivastava, Rohit Lal, and Anirban Chakraborty. Caft: Class aware frequency transform for reducing domain gap. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2525–2534, 2021.

[35] Jogendra Nath Kundu, Siddharth Seth, Anirudh Jamkhandi, Pradyumna YM, Varun Jampani, Anirban Chakraborty, et al. Non-local latent relation distillation for self-adaptive 3d human pose estimation. *Advances in Neural Information Processing Systems*, 34:158–171, 2021.

[36] Jogendra Nath Kundu, Siddharth Seth, MV Rahul, Mugalodi Rakesh, Venkatesh Babu Radhakrishnan, and Anirban Chakraborty. Kinematic-structure-preserved representation for unsupervised 3d human pose estimation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 11312–11319, 2020.

[37] Jogendra Nath Kundu, Siddharth Seth, Pradyumna YM, Varun Jampani, Anirban Chakraborty, and R Venkatesh Babu. Uncertainty-aware adaptation for self-supervised 3d human pose estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 20448–20459, 2022.

[38] Rohit Lal, Yash Garg, Arindam Dutta, Calvin-Khang Ta, Dripta S Raychaudhuri, M Salman Asif, and Amit K Roy-Chowdhury. Temp3d: Temporally continuous 3d human pose estimation under occlusions. *arXiv preprint arXiv:2312.16221*, 2023.

[39] Rohit Lal, Arihant Gaur, Aadhithya Iyer, Muhammed Abdullah Shaikh, and Ritik Agrawal. Open-set multi-source multi-target domain adaptation. *arXiv preprint arXiv:2302.00995*, 2023.

[40] Kyoungoh Lee, Inwoong Lee, and Sanghoon Lee. Propagating lstm: 3d pose estimation based on joint interdependency. In Vittorio Ferrari, Martial Hebert, Cristian Sminchisescu, and Yair Weiss, editors, *Computer Vision – ECCV 2018*, pages 123–141, Cham, 2018. Springer International Publishing.

[41] Chenyang Lei, Yazhou Xing, Hao Ouyang, and Qifeng Chen. Deep video prior for video consistency and propagation, 2022.

[42] Zhihao Li, Jianzhuang Liu, Zhensong Zhang, Songcen Xu, and Youliang Yan. Cliff: Carrying location information in full frames into human pose and shape estimation, 2022.

[43] Zhengyi Luo, S Alireza Golestaneh, and Kris M Kitani. 3d human motion estimation via motion compression and refinement. In *ACCV*, 2020.

[44] Naureen Mahmood, Nima Ghorbani, Nikolaus F. Troje, Gerard Pons-Moll, and Michael J. Black. Amass: Archive of motion capture as surface shapes, 2019.

[45] Julieta Martinez, Rayat Hossain, Javier Romero, and James J Little. A simple yet effective baseline for 3d human pose estimation. In *Proceedings of the IEEE international conference on computer vision*, pages 2640–2649, 2017.

[46] Dushyant Mehta, Helge Rhodin, Dan Casas, Pascal Fua, Oleksandr Sotnychenko, Weipeng Xu, and Christian Theobalt. Monocular 3d human pose estimation in the wild using improved cnn supervision. In *2017 international conference on 3D vision (3DV)*, pages 506–516. IEEE, 2017.

[47] Dushyant Mehta, Helge Rhodin, Dan Casas, Pascal Fua, Oleksandr Sotnychenko, Weipeng Xu, and Christian Theobalt. Monocular 3d human pose estimation in the wild using improved cnn supervision. In *3D Vision (3DV), 2017 Fifth International Conference on*. IEEE, 2017.

[48] Dushyant Mehta, Oleksandr Sotnychenko, Franziska Mueller, Weipeng Xu, Srinath Sridhar, Gerard Pons-Moll, and Christian Theobalt. Single-shot multi-person 3d pose estimation from monocular rgb. In *2018 International Conference on 3D Vision (3DV)*, pages 120–130. IEEE, 2018.

[49] Dushyant Mehta, Srinath Sridhar, Oleksandr Sotnychenko, Helge Rhodin, Mohammad Shafiei, Hans-Peter Seidel, Weipeng Xu, Dan Casas, and Christian Theobalt. Vnect: real-time 3d human pose estimation with a single rgb camera. *ACM Transactions on Graphics*, 36(4):1–14, July 2017.

[50] Francesc Moreno-Noguer. 3d human pose estimation from a single image via distance matrix regression, 2016.

[51] Ramesha Rakesh Mugaludi, Jogendra Nath Kundu, Varun Jampani, et al. Aligning silhouette topology for self-adaptive 3d human pose recovery. *Advances in Neural Information Processing Systems*, 34:4582–4593, 2021.

[52] Hyeongjin Nam, Daniel Sungho Jung, Yeonguk Oh, and Kyoung Mu Lee. Cyclic test-time adaptation on monocular video for 3d human mesh reconstruction. In *International Conference on Computer Vision (ICCV)*, 2023.

[53] Alejandro Newell, Zhiao Huang, and Jia Deng. Associative embedding: End-to-end learning for joint detection and grouping, 2017.

[54] Alejandro Newell, Kaiyu Yang, and Jia Deng. Stacked hourglass networks for human pose estimation, 2016.

[55] Bruce Xiaohan Nie, Ping Wei, and Song-Chun Zhu. Monocular 3d human pose estimation by predicting depth on joints. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 3467–3475. IEEE, 2017.

[56] Dario Pavllo, Christoph Feichtenhofer, David Grangier, and Michael Auli. 3d human pose estimation in video with temporal convolutions and semi-supervised training. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7753–7762, 2019.

[57] Qucheng Peng, Ce Zheng, and Chen Chen. Source-free domain adaptive human pose estimation, 2023.

[58] Ibrahim Radwan, Abhinav Dhall, and Roland Goecke. Monocular image 3d human pose estimation under self-occlusion. In *2013 IEEE International Conference on Computer Vision*, pages 1888–1895, 2013.

[59] Dripta S Raychaudhuri, Calvin-Khang Ta, Arindam Dutta, Rohit Lal, and Amit K Roy-Chowdhury. Prior-guided source-free domain adaptation for human pose estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14996–15006, 2023.

[60] Davis Rempe, Tolga Birdal, Aaron Hertzmann, Jimei Yang, Srinath Sridhar, and Leonidas J. Guibas. Humor: 3d human motion model for robust pose estimation, 2021.

[61] Helge Rhodin, Jörg Spörri, Isinsu Katircioglu, Victor Constantin, Frédéric Meyer, Erich Müller, Mathieu Salzmann, and Pascal Fua. Learning monocular 3d human pose estimation from multi-view images. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8437–8446, 2018.

[62] Paul M Schwartz and Daniel J Solove. The pii problem: Privacy and a new concept of personally identifiable information. *NYUL rev.*, 86:1814, 2011.

[63] Soyong Shin, Juyong Kim, Eni Halilaj, and Michael J. Black. Wham: Reconstructing world-grounded humans with accurate 3d motion, 2023.

[64] Yu Sun, Qian Bao, Wu Liu, Yili Fu, Michael J Black, and Tao Mei. Monocular, one-stage, regression of multiple 3d people. In *ICCV*, 2021.

[65] Matthew Tancik, Pratul Srinivasan, Ben Mildenhall, Sara Fridovich-Keil, Nithin Raghavan, Utkarsh Singhal, Ravi Ramamoorthi, Jonathan Barron, and Ren Ng. Fourier features let networks learn high frequency functions in low dimensional domains. *Advances in Neural Information Processing Systems*, 33:7537–7547, 2020.

[66] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need, 2023.

[67] Timo von Marcard, Roberto Henschel, Michael Black, Bodo Rosenhahn, and Gerard Pons-Moll. Recovering accurate 3d human pose in the wild using imus and a moving camera. In *European Conference on Computer Vision (ECCV)*, sep 2018.

[68] Timo von Marcard, Roberto Henschel, Michael J. Black, Bodo Rosenhahn, and Gerard Pons-Moll. Recovering accurate 3d human pose in the wild using imus and a moving camera. In *Proceedings of the European Conference on Computer Vision (ECCV)*, September 2018.

[69] Jacob Walker, Kenneth Marino, Abhinav Gupta, and Martial Hebert. The pose knows: Video forecasting by generating pose futures, 2017.

[70] Jingdong Wang, Ke Sun, Tianheng Cheng, Borui Jiang, Chaorui Deng, Yang Zhao, Dong Liu, Yadong Mu, Mingkui Tan, Xinggang Wang, Wenyu Liu, and Bin Xiao. Deep high-resolution representation learning for visual recognition, 2020.

[71] Jingwei Xu, Zhenbo Yu, Bingbing Ni, Jiancheng Yang, Xiaokang Yang, and Wenjun Zhang. Deep kinematics analysis for monocular 3d human pose estimation. In *Proceedings of the IEEE/CVF Conference on computer vision and Pattern recognition*, pages 899–908, 2020.

[72] Yuanlu Xu, Wenguan Wang, Tengyu Liu, Xiaobai Liu, Jianwen Xie, and Song-Chun Zhu. Monocular 3d pose estimation via pose grammar and data augmentation. *IEEE transactions on pattern analysis and machine intelligence*, 44(10):6327–6344, 2021.

[73] Wei Yang, Shuang Li, Wanli Ouyang, Hongsheng Li, and Xiaogang Wang. Learning feature pyramids for human pose estimation, 2017.

[74] Ye Yuan, Umar Iqbal, Pavlo Molchanov, Kris Kitani, and Jan Kautz. Glamr: Global occlusion-aware human mesh recovery with dynamic cameras, 2022.

[75] Ailing Zeng, Lei Yang, Xuan Ju, Jiefeng Li, Jianyi Wang, and Qiang Xu. Smoothnet: A plug-and-play network for refining human poses in videos. In *European Conference on Computer Vision*. Springer, 2022.

[76] Jinlu Zhang, Zhigang Tu, Jianyu Yang, Yujin Chen, and Junsong Yuan. Mixste: Seq2seq mixed spatio-temporal encoder for 3d human pose estimation in video. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13232–13242, 2022.

[77] Tianshu Zhang, Buzhen Huang, and Yangang Wang. Object-occluded human shape and pose estimation from a single color image. In *CVPR*, 2020.

[78] Qitao Zhao, Ce Zheng, Mengyuan Liu, Pichao Wang, and Chen Chen. Poseformerv2: Exploring frequency domain for efficient and robust 3d human pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8877–8886, 2023.

[79] Ce Zheng, Wenhan Wu, Chen Chen, Taojiannan Yang, Sijie Zhu, Ju Shen, Nasser Kehtarnavaz, and Mubarak Shah. Deep learning-based human pose estimation: A survey. *ACM Comput. Surv.*, 56(1), aug 2023.

[80] Ce Zheng, Sijie Zhu, Matias Mendieta, Taojiannan Yang, Chen Chen, and Zhengming Ding. 3d human pose estimation with spatial and temporal transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11656–11665, 2021.

[81] Xiaowei Zhou, Menglong Zhu, Spyridon Leonardos, Konstantinos G Derpanis, and Kostas Daniilidis. Sparseness meets deepness: 3d human pose estimation from monocular video. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4966–4975, 2016.

[82] Haidong Zhu, Wanrong Zheng, Zhaoheng Zheng, and Ram Nevatia. Sharc: Shape and appearance recognition for person identification in-the-wild, 2023.

[83] Wentao Zhu, Xiaoxuan Ma, Zhaoyang Liu, Libin Liu, Wayne Wu, and Yizhou Wang. Motionbert: A unified perspective on learning human motion representations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15085–15099, 2023.

[84] Yue Zhu and David Picard. Decanus to legatus: Synthetic training for 2d-3d human pose lifting. In *Proceedings of the Asian Conference on Computer Vision*, pages 2848–2865, 2022.