STRIDE: Single-video based Temporally Continuous Occlusion-Robust 3D Pose Estimation

Rohit Lal[‡1*] Saketh Bachu[1*] Yash Garg[1] Arindam Dutta[1] Calvin-Khang Ta[1] Hannah D. Cruz[1]
Dripta S. Raychaudhuri[†1] M. Salman Asif[1] Amit K. Roy-Chowdhury[1]
[1]University of California, Riverside
{rlal011,sbach008,ygarg002,adutt020,cta003,hdela004,drayc001,sasif,amitrc}@ucr.edu

Figure 1. **Effect of occlusions on pose estimation.** Image-based 3D pose estimators [3] 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 column. Notably, even state-of-the-art video approaches [36] fail on prolonged full occlusions spanning multiple frames, as in columns 4 and 5. This highlights yet 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.

Abstract

Accurately estimating 3D human poses is crucial for fields like action recognition, gait recognition, and virtual/augmented reality. However, predicting human poses under severe occlusion remains a persistent and significant challenge. Existing image-based estimators struggle with heavy occlusions due to a lack of temporal context, resulting in inconsistent predictions, while video-based models, despite benefiting from temporal data, face limitations with prolonged occlusions over multiple frames. Additionally, existing algorithms often struggle to generalize unseen videos. 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 estimating 3D human poses for each video. Our proposed approach handles occlusions not encountered during the model's training by refining a sequence of noisy initial pose estimates into accurate, temporally coherent poses at test time, effectively overcoming the limitations of existing methods. Our flexible, model-agnostic framework allows us to use any off-the-shelf 3D pose estimation method to improve robustness and temporal consis-*

* Equal contribution, ‡ Currently at NASA MSFC IMPACT. Work done while the author was at UCR., † Currently at AWS AI Labs. Work done while the author was at UCR.

2642-9381/25/$31.00 ©2025 IEEE
DOI 10.1109/WACV61041.2025.00087

2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV) | 979-8-3315-1083-1/25/$31.00 ©2025 IEEE | DOI: 10.1109/WACV61041.2025.00087

tency. We validate STRIDE*'s efficacy through comprehensive experiments on multiple challenging datasets where it not only outperforms existing single-image and video-based pose estimation models but also showcases superior handling of substantial occlusions, achieving fast, robust, accurate, and temporally consistent 3D pose estimates. Code is made publicly available at* https://github.com/take2rohit/stride

1. Introduction

Accurate 3D pose estimation [48] is an important problem in computer vision with a variety of real-world applications, including but not limited to action recognition [20], virtual and augmented reality [1], and gait recognition [10, 11, 50]. While the performance of 3D pose estimation algorithms has improved rapidly in recent years, the majority of these are image-based [29, 31, 33, 38], 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 [32, 41], 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 [3, 26, 36, 41] relies heavily on supervised training on large datasets captured in controlled settings [3]. This limits generalizability, as distribution shifts in uncontrolled environments 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. Image-based pose estimation methods [3] 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 [27, 28]. On the other hand, video-based approaches [26, 41, 47] also suffer from performance degradation, despite modeling temporal information, due to such prolonged occlusions being absent in the training data [5].

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 [12, 13, 26] fine-tune models using 2D cues like keypoints from test images. This approach has limitations as the 2D projection of 3D poses is ambiguous, as many 3D configurations can map to the same 2D pose. Also, 2D pose estimators can fail on unseen data distributions [18, 31], so fine-tuning on imperfect and ambiguous 2D poses can lead to incorrect model adaptation and degraded 3D pose predictions.

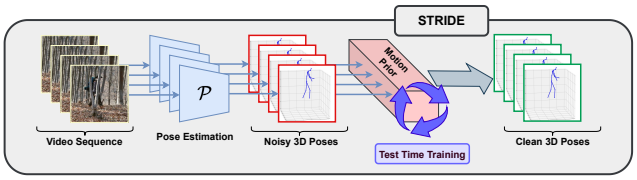

Figure 2. **Overview of our approach.** Our method enhances 3D pose estimation for occluded videos through test-time training of a motion prior model. We first extract initial 3D pose estimates from the test video using any 3D off-the-shelf pose estimator. To address occlusions and test distribution shifts, we then fine-tune the motion prior on that specific video by optimizing for smooth and continuous poses over the sequence.

To 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 [9, 51] 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. 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 [26, 32], our approach is *up to* $2\times$ *faster* than previous state-of-the-art method [26] and operates without accessing any labeled training data during inference time, making it privacy [34] and storage-friendly.

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

1. We propose a novel test-time training algorithm (STRIDE), for achieving temporally continuous 3D pose estimation under occlusions.

2. STRIDE is designed as a model-agnostic framework that leverages a human motion prior model to refine noisy 3D pose sequences from any off-the-shelf estimator into smooth and continuous predictions, highlighting efficiency and generalizability.

3. STRIDE achieves state-of-the-art results on multiple challenging benchmarks including OCMotion [15], Occluded Human3.6M and Human3.6M [16], thus demonstrating enhanced occlusion-robustness and temporal consistency. Additionally, STRIDE is computationally amicable, achieving a minimum of $2\times$ speed-up over existing analogous algorithms.

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. Recent deep learning-based methods for the problem have shown impressive performance on challenging academic datasets [3, 48]. [25] introduced the first CNN-based approach to regress 3D joints from a single image. Subsequent works [29, 33] improved upon this by incorporating multi-view constraints and depth information. Recent methods [39, 40] use kinematic and anatomical constraints along with data augmentation to achieve state-of-the-art results on academic datasets. However, these supervised methods often fail under distribution shifts. To address this, [21, 22] proposed self-supervised algorithms for 3D human pose estimation, which perform well on single images but struggle with occlusions and lack temporal continuity in video settings.

Video-based 3D pose estimation. Video-based 3D human pose estimation have shown impressive performance on challenging datasets. [49] directly regresses 3D poses using consistency between 3D joints and 2D keypoints. [30] utilized temporal convolutions for pose estimation in videos, while [2] exploited SMPL pose and shape parameters for fine-tuning HMR in the wild. [44] proposed a mixed spatio-temporal approach alternating between spatial and temporal consistency. HuMoR [32] maintained consistency across frames with weighted regularization using predicted contact probabilities. The state-of-the-art CycleAdapt [26] addresses domain shifts in 3D human mesh reconstruction by cyclically adapting HMRNet [17] and MDNet [26] 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 poses.

3D pose estimation under occlusion. Handling occlusions poses a significant challenge in both image-based and video-based 3D pose estimation. Approaches like 3DNBF [46] leverage generative models to estimate poses but do not account for any temporal continuity. To alleviate the problem in a video-based setting, [6] introduced data augmentation using occlusion labels with the Cylinder Man Model. Current methods address this by refining 3D poses for temporal consistency. Recent approaches such as GLAMR [41] recover human meshes globally from local motions and perform motion infilling based on visible motions. SmoothNet [42] uses a temporal refinement network to mitigate motion jitters from single image-based pose estimations. While effective for minor occlusions, these methods struggle with heavy occlusions. Also, these algorithms often fail to generalize under domain shifts. To improve on this, our approach adapts a motion prior from noisy 3D pose sequences to predict missing poses and maintain temporal consistency.

Test Time Optimization for 3D pose estimation. A major shortcoming of fully supervised learning is that it can only handle test cases that are similar to the ones seen during the training process. For example, a novel type of occlusion or a human pose during testing can confuse the model and reduce its performance. [43] suggested verifying the estimated 2D poses and additionally ensuring the consistency of the lifted 3D poses by using randomly projected 2D poses to enhance the 3D human pose estimation. Further, [35] introduced the idea of enforcing physical constraints on the estimated human poses to ensure they are physically plausible. Subsequently, [4] proposed combining top-down and bottom-up human pose estimation approaches to take advantage of their strengths and also perform test time optimization using a re-projection loss, and bone length regularizations. Although these methods handle unseen test cases well, they are not designed for handling heavy occlusions and fail to predict poses under complete occlusions. Our work focuses on these gaps by refining and filling in missing 3D poses to maintain temporal consistency.

3. Method

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, as illustrated in Fig. 3.

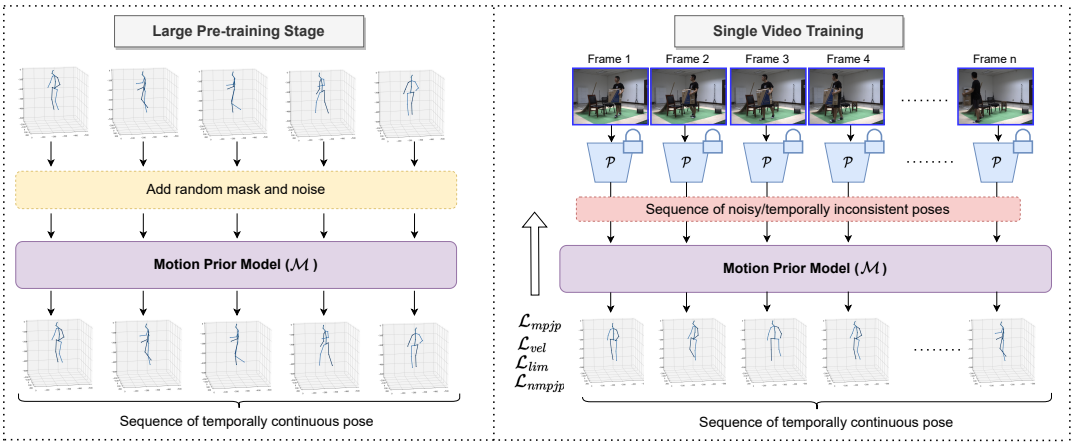

Figure 3. The presented figure illustrates the pipeline for our temporally continuous pose estimation, STRIDE. Initially, we pre-train a motion prior model, denoted as $\mathcal{M}$, using a diverse set of 3D pose data sourced from various public datasets. The primary objective of this motion prior model 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 Equation 6. The end result of this training process is a model capable of producing temporally continuous 3D poses for that specific video.

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 [9, 51]. 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 poses. This allows $\mathcal{M}$ to learn strong general priors of natural human motion dynamics.

2. **Test-time alignment:** For a given test video, we obtain noisy per-frame poses using $\mathcal{P}$ [3] and adapt 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 optimal pose estimates for the given video.

Section 3.1 outlines the architecture of the motion prior model $\mathcal{M}$. Section 3.2 discusses the masked sequence modelling approach for pre-training $\mathcal{M}$ on synthetically corrupted pose sequences. Lastly, Section 3.3 introduces the self-supervised losses for fine-tuning $\mathcal{M}$ during test time on individual videos.

3.1. Network Architecture

We base our motion prior model $\mathcal{M}$ on the DSTFormer architecture [51], 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. $\mathcal{M}$ denoises the input sequence to produce refined temporally coherent 3D poses $\bar{\mathbf{X}} \in \mathbb{R}^{T \times J \times 3}$. Additional details on the architecture and implementation method is provided in the Section 1 and Section 2 of supplementary material respectively.

3.2. 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 [9] 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 [7, 51], the prior $\mathcal{M}$ is trained to denoise these noisy inputs to reconstruct a sequence of temporally coherent 3D poses. We selected BERT-style training for STRIDE's motion prior because it is highly effective at capturing bidirectional dependencies in human motion data and excels in scenarios where pose information may be incomplete or corrupted. BERT-style training leverages a bidirectional approach, allowing the model to consider both preceding and succeeding context simultaneously. This is particularly advantageous for understanding human motion, as it mirrors real-world situations where poses might be partially missing or noisy.

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 $\mathcal{L}_{3D}$ between the reconstruction and the ground-truth pose. Additionally, we incorporate a velocity loss following [30, 44].

3.3. 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 $\mathcal{P}$ [3]. 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 [23], our proposed self-supervision strategy fine-tunes the motion prior to the specifics of each test video for enhanced outputs. We use four loss regularizers targeting different aspects of human motion: (1) Limb Loss, (2) Mean Per Joint Position (MPJP) Loss, (3) Normalized MPJP (N-MPJP) Loss, and (4) Velocity Loss. 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,

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

Here $\mathcal{J} \in \mathbb{R}^{T\times(J-1)}$ 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,

$$\mathcal{L}_{\text{MPJP}} = \frac{1}{T\cdot J\cdot 3}\sum_{t=1}^{T}\sum_{j=1}^{J}\sum_{d=1}^{3}\|\hat{\mathbf{X}}_{t,j,d} - \tilde{\mathbf{X}}_{t,j,d}\|_2 \quad (2)$$

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.

$$\mathcal{L}_{\text{NMPJP}} = \mathcal{L}_{\text{MPJP}}(s\hat{\mathbf{X}}, \tilde{\mathbf{X}}) \quad (3)$$

$$\text{where } s = \frac{\sum_{t=1}^{T}\sum_{j=1}^{J}\sum_{d=1}^{3}\left\|\tilde{\mathbf{X}}_{t,j,d}\cdot\hat{\mathbf{X}}_{t,j,d}\right\|_2}{\sum_{t=1}^{T}\sum_{j=1}^{J}\sum_{d=1}^{3}\left\|\hat{\mathbf{X}}_{t,j,d}\right\|_2^2} \quad (4)$$

In Equation 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 Equation 5, but instead of ground truth, we use pseudo-labels,

$$\mathcal{L}_{\text{vel}} = \frac{1}{N\cdot(J-1)}\sum_{t=1}^{T-1}\sum_{j=1}^{J}\sum_{d=1}^{3}\left\|\hat{\mathbf{V}} - \tilde{\mathbf{V}}\right\|_2 \quad (5)$$

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. The velocity loss helps in smoothing the movement and removing unwanted jittering across frames.

Overall Loss. In summary, by combining all the abovementioned losses into one final loss function as shown in

Equation 6, $\mathcal{M}$ is trained to produce accurate joint positions, maintain anatomical consistency, and handle scale variations,

$$\mathcal{L}_{total} = \lambda_1 \mathcal{L}_{\text{lim}} + \lambda_2 \mathcal{L}_{\text{MPJP}} + \lambda_3 \mathcal{L}_{\text{NMPJP}} + \lambda_4 \mathcal{L}_{\text{vel}} \quad (6)$$

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

4. Experiments and Results

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 [3] as the off-the-shelf pose estimation method. We compare the error rates of `STRIDE` and the baseline methods in Tables 1, 2 and 3. The best results are in **bold** and arrows indicate the percentage improvement over the best existing algorithm. Qualitative video results can be found on our shared GitHub repository.

4.1. Datasets

Human3.6M [16]: This indoor-scene dataset is crucial for 3D human pose estimation from 2D images. We use every 1 in 5 frames in the test split and achieve comparable performance to state-of-the-art methods.

OCMotion [15]: This video dataset extends the 3DOH50K image dataset with natural occlusions, comprising 300K images at 10 FPS. We use only the test split since our method does not require supervised training.

Occluded Human3.6M: We prepare a new dataset by modifying the Human3.6M [16]. It is curated to evaluate pose estimation under significant occlusion. This dataset uses random erase occlusions covering up to 100% of a person for 1.6 seconds within 3.2-second videos.

BRIAR [8]: Features videos of human subjects in extremely challenging conditions, recorded at varying distances and from UAVs. Additional details on datasets and implementation specifics, are provided in Section 7 of supplementary material.

4.2. Quantitative Results

	Method	PA-MPJPE	MPJPE	Accel
Image	CLIFF [24]	183.5	100.5	38.4
	BEDLAM [3]	179.5	98.9	39.1
Video	GLAMR [41]	213.9	380.3	42.3
	PoseFormerV2 [47]	193.9	260.2	38.7
	CycleAdapt [26]	77.6	132.6	48.7
	MotionBERT [51]	76.1	112.8	28.7
	`STRIDE` (ours)	**59.0** (57%↓)	**80.7** (18%↓)	**26.6** (7%↓)

Table 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 `STRIDE` surpasses all state-of-the-art with substantial percentage improvements, affirming its robustness in handling occlusions.

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 1. Notably, `STRIDE` performs significantly better than BEDLAM despite using pseudo-labels 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 result 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 2 shows that our approach `STRIDE` attains state-of-the-art results on the OCMotion dataset [15]. 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 [19, 37] are supervised and trained on the training split of OCMotion. In contrast to these algorithms, our proposed approach (`STRIDE`) does not assume access to any labeled training dataset.*

Our method demonstrates minor improvement over BEDLAM-CLIFF [3] on the original Human3.6M dataset, as evidenced in Table 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. A thorough analysis of our findings, including the observed enhancement in temporal

	Method	PA-MPJPE	Accel	Avg
Image	OOH [45]	55.0	48.6	51.8
	PARE [19]	52.0	43.6	47.8
	BEDLAM [3]	47.1	49.0	48.0
Video	PoseFormerV2 [47]	126.3	**28.5**	77.4
	GLAMR [41]	89.9	51.3	70.6
	CycleAdapt [26]	74.6	57.5	66.0
	ROMP [37]	48.1	57.2	52.6
	STRIDE (ours)	**46.2** (2%↓)	47.8	**47.0** (2%↓)

Table 2. **3D pose estimation results on OCMotion [15].** STRIDE 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	MPJPE	Accel
Image	CLIFF [24]	56.1	89.6	-
	BEDLAM-HMR [3]	51.7	81.6	-
	BEDLAM-CLIFF [3]	50.9	70.9	39.14
Video	GLAMR [41]	-	-	-
	CycleAdapt [26]	64.5	106.3	57.25
	MotionBERT* [51]	64.15	95.8	**14.8**
	STRIDE (ours)	**50.4** (1%↓)	**69.7** (2%↓)	37.1

Table 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.

smoothness, is provided in the Section 3 of supplementary.

Inference speed: Table 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 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. We have ensured that the testing is consistent across all algorithms, and results are reported *including the TTT phase with 30 epochs*. Our method takes less time because we train only on a single video at test time, which contains significantly fewer frames compared to the entire dataset.

Method	Time (sec)
HuMor [32]	> 600
GLAMR [41]	> 600
PoseFormerV2 [47]	129
CycleAdapt [26]	126
STRIDE (ours)	**68** (46%↓)

Table 4. Inference time for various 3D pose estimation methods.

4.3. Qualitative Results

To provide a comprehensive analysis and comparison of STRIDE against other methods, we have compiled and shared several qualitative video results in the supplementary material. Our evaluation juxtaposes STRIDE against leading state-of-the-art techniques like CycleAdapt [26]. Key insights from our comparison include:

Occluded Human3.6M: In this proposed dataset, 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 compared to the former methods.

BRIAR [8]: 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 3D pose data on BRIAR, the visual comparisons provided through our videos convincingly demonstrate our method's enhanced adaptability and efficacy.

OCMotion [14]: In Fig. 4, we compare our method against an existing state-of-the-art pose estimation method CycleAdapt [26]. 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.

Please refer to the Section 4 supplementary material for additional qualitative 3D pose estimation results and Section 6 of supplementary material for details on mesh generation in videos and mesh recovery results.

4.4. Ablation Study

An ablation study conducted in Table 5 provides insights into the significance of each component in STRIDE. Start-

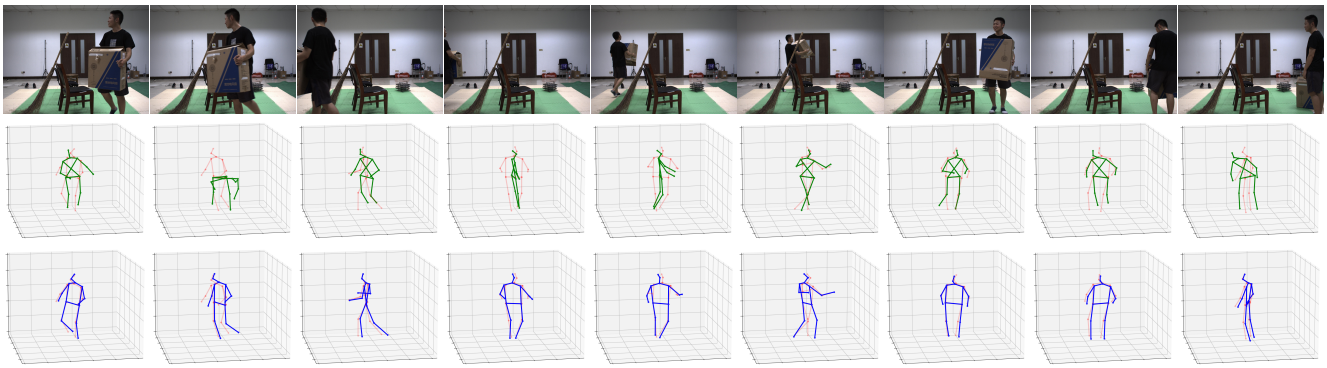

Figure 4. **3D pose estimation results on OCMotion (*0013, Camera01*)**. This figure demonstrates how our method incorporates temporal continuity into video sequences under occlusion. The *second* row represents 3D poses predicted by CycleAdapt [26]. The *third* row represents 3D poses predicted by STRIDE. Note: The 3D poses shown in translucent red color in the *second* and *third* row represent the ground truths.

$\mathcal{M}$	$\mathcal{L}_{mpjp}$	$\mathcal{L}_{vel}$	$\mathcal{L}_{lim}$	$\mathcal{L}_{nmpjp}$	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**

Table 5. **Ablation study.** This table illustrates how integrating a pre-trained motion prior and various losses collectively contribute to STRIDE's final accuracy on the Occluded Human3.6M dataset.

ing from a baseline with substantial errors, introducing a motion prior alone drastically improves performance, underscoring its effectiveness in driving the model toward realistic human pose dynamics. Adding L_{mpjp} enhances spatial accuracy, further lowering MPJPE to 82.1 and PA-MPJPE to 60.4. 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. Thus, 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 state-of-the-art results. This shows 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. In Section 5 of supplementary material, we show how varying off-the-shelf pose estimation methods within the backbone of STRIDE affects its performance. 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.

5. Conclusion

We introduce STRIDE, a novel algorithm for self-supervised test-time training aimed at improving 3D human pose estimation in individual video frames. STRIDE utilizes extensive self-supervised pre-training to develop a robust model of human motion priors. It integrates self-supervised optimization with temporal regularization, achieving state-of-the-art performance in terms of both pose accuracy and computational efficiency across diverse challenging datasets, even those with significant occlusions. A limitation of STRIDE is its ability to extract temporally continuous 3D poses only in scenarios where there are no human-to-human occlusions. Future efforts can concentrate on adapting STRIDE for situations involving multi-person occlusions. Handling such situations requires modelling of complex human-to-human interaction alongside external identification and tracking framework. In conclusion, STRIDE sets a new benchmark for 3D human pose estimation in occluded environments and introduces promising directions for enhancing related downstream applications.

Acknowledgment

The work was partially supported by NSF grants 2326309 and 2312395, DURIP grant N000141812252, USDA grant 2021-67022-33453, and the Office of the Director of National Intelligence (ODNI), specifically through the Intelligence Advanced Research Projects Activity (IARPA), under contract number [2022- 21102100007]. The views and conclusions in this research reflect those of the authors and should not be construed as officially representing the policies, whether explicitly or implicitly, of ODNI, IARPA, or the U.S. Government. Nevertheless, the U.S. Government retains the authorization to reproduce and distribute reprints for official government purposes, regardless of any copyright notices included.

References

[1] 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. 2

[2] 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. 3

[3] Michael J. Black, Priyanka Patel, Joachim Tesch, and Jinlong Yang. Bedlam: A synthetic dataset of bodies exhibiting detailed lifelike animated motion, 2023. 1, 2, 3, 4, 5, 6, 7

[4] Yu Cheng, Bo Wang, and Robby T. Tan. Dual networks based 3d multi-person pose estimation from monocular video. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(2):1636–1651, 2023. 3

[5] 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. 2

[6] 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. 3

[7] 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. 4

[8] 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. 6, 7

[9] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding, 2019. 2, 4

[10] Arindam Dutta, Rohit Lal, Yash Garg, Calvin-Khang Ta, Dripta S Raychaudhuri, Hannah Dela Cruz, and Amit K Roy-Chowdhury. Posture: Pose guided unsupervised domain adaptation for human body part segmentation. *arXiv preprint arXiv:2407.03549*, 2024. 2

[11] 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. 2

[12] 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. 2

[13] 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. 2

[14] 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. 7

[15] 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. 3, 6, 7

[16] 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. 3, 6

[17] 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. 3

[18] 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. 2

[19] Muhammed Kocabas, Chun-Hao P Huang, Otmar Hilliges, and Michael J Black. Pare: Part attention regressor for 3d human body estimation. In *ICCV*, 2021. 6, 7

[20] Yu Kong and Yun Fu. Human action recognition and prediction: A survey. *International Journal of Computer Vision*, 130(5):1366–1401, 2022. 2

[21] 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. 3

[22] 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. 3

[23] Chenyang Lei, Yazhou Xing, Hao Ouyang, and Qifeng Chen. Deep video prior for video consistency and propagation, 2022. 5

[24] 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. 6, 7

[25] 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. 3

[26] 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. 2, 3, 6, 7, 8

[27] Alejandro Newell, Zhiao Huang, and Jia Deng. Associative embedding: End-to-end learning for joint detection and grouping, 2017. 2

[28] Alejandro Newell, Kaiyu Yang, and Jia Deng. Stacked hourglass networks for human pose estimation, 2016. 2

[29] 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. 2, 3

[30] 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. 3, 5

[31] 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. 2

[32] Davis Rempe, Tolga Birdal, Aaron Hertzmann, Jimei Yang, Srinath Sridhar, and Leonidas J. Guibas. Humor: 3d human motion model for robust pose estimation, 2021. 2, 3, 7

[33] 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. 2, 3

[34] Paul M Schwartz and Daniel J Solove. The pii problem: Privacy and a new concept of personally identifiable information. *NYUL rev.*, 86:1814, 2011. 2

[35] Soshi Shimada, Vladislav Golyanik, Weipeng Xu, and Christian Theobalt. Physcap: physically plausible monocular 3d motion capture in real time. *ACM Trans. Graph.*, 39(6), nov 2020. 3

[36] Soyong Shin, Juyong Kim, Eni Halilaj, and Michael J. Black. Wham: Reconstructing world-grounded humans with accurate 3d motion, 2023. 1, 2

[37] 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. 6, 7

[38] Calvin-Khang Ta, Arindam Dutta, Rohit Kundu, Rohit Lal, Hannah Dela Cruz, Dripta S Raychaudhuri, and Amit Roy-Chowdhury. Multi-modal pose diffuser: A multimodal generative conditional pose prior. *arXiv preprint arXiv:2410.14540*, 2024. 2

[39] 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. 3

[40] 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. 3

[41] Ye Yuan, Umar Iqbal, Pavlo Molchanov, Kris Kitani, and Jan Kautz. Glamr: Global occlusion-aware human mesh recovery with dynamic cameras, 2022. 2, 3, 6, 7

[42] 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. 3

[43] Jianfeng Zhang, Xuecheng Nie, and Jiashi Feng. Inference stage optimization for cross-scenario 3d human pose estimation. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 2408–2419. Curran Associates, Inc., 2020. 3

[44] 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. 3, 5

[45] Tianshu Zhang, Buzhen Huang, and Yangang Wang. Object-occluded human shape and pose estimation from a single color image. In *CVPR*, 2020. 7

[46] Yi Zhang, Pengliang Ji, Adam Kortylewski, Angtian Wang, Jieru Mei, and Alan L Yuille. 3D-Aware Neural Body Fitting for Occlusion Robust 3D Human Pose Estimation. In *The IEEE/CVF International Conference on Computer Vision*, 2023. 3

[47] 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. 2, 6, 7

[48] 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. 2, 3

[49] 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. 3

[50] Haidong Zhu, Wanrong Zheng, Zhaoheng Zheng, and Ram Nevatia. Sharc: Shape and appearance recognition for person identification in-the-wild, 2023. 2

[51] 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. 2, 4, 6, 7