Temporally Consistent 3D Human Pose and Shape Estimation from Video

Aurel Shehaj[1,3], Nikolaos Vasilikopoulos[3,4], Iason Oikonomidis[3], Antonis A. Argyros[2,1,3]

[1]Graduate Program on Data Analysis and Machine-Statistical Learning (DAMSL), University of Crete, Greece
[2]Computer Science Department, University of Crete, Greece
[3]Institute of Computer Science, Foundation for Research and Technology–Hellas (FORTH), Greece
[4]VRULL GmbH, Austria

{shehaj, oikonom, argyros}@ics.forth.gr, vassilicopoulos@gmail.com

***Abstract*—Human pose and shape estimation (HPS) from monocular images has recently witnessed rapid progress through the use of parametric body models and deep networks. However, frame-wise formulations struggle with temporal jitter and inconsistent motion when applied to video. We extend the depth-guided D-PoSE framework to operate on short video segments and show that temporally-aware models improve sequence-level stability. Accurate, temporally coherent motion reconstruction is particularly important for immersive virtual reality (VR) applications, where jitter and inconsistent avatar behaviour can break presence and even induce visual discomfort in users. Our experiments demonstrate that recurrent and transformer-based temporal models reduce acceleration error by more than 60% on the challenging 3DPW benchmark while retaining or improving frame-wise reconstruction accuracy. The proposed pipeline thus moves D-PoSE towards motion-consistent video inference while retaining the advantages of its single-image spatial representation, making it particularly suited for VR and AR scenarios where temporal stability is critical.**

***Index Terms*—Human Pose and Shape Estimation; Temporal Modeling; 3D Human Reconstruction; Avatar Tracking; Human Motion Analysis; Transformers**

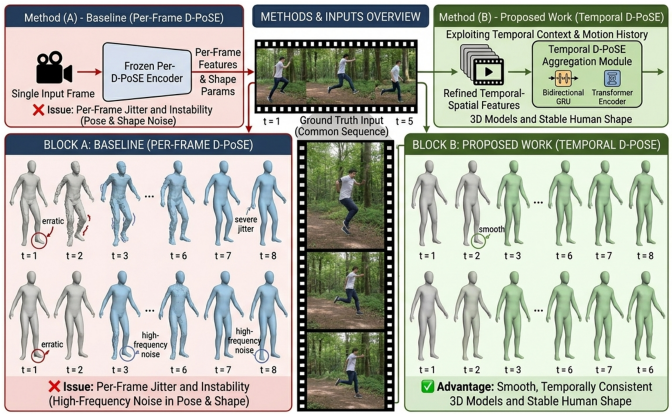

Fig. 1. D-PoSE vs the proposed temporally consistent D-PoSE: The proposed pipeline moves D-PoSE towards motion-consistent video inference while retaining the advantages of its single-image spatial representation, making it particularly suited for VR and AR scenarios where temporal stability is critical.

I. Introduction

Human pose and shape estimation (HPS) aims to reconstruct a complete 3D representation of the human body from visual observations. Modern approaches typically employ parametric body models such as SMPL [1] or SMPL-X [2], together with convolutional or transformer-based architectures, to regress pose and shape parameters directly from monocular images. Beyond its academic interest, accurate 3D human motion capture has broad applications in robotics, animation, healthcare, sports analytics, and immersive technologies.

In virtual and augmented reality (VR/AR) environments in particular, accurate reconstruction of full-body pose and shape is a key enabling technology for natural telepresence, social interaction, avatar embodiment, and training applications. However, consumer-grade VR systems typically provide only sparse tracking signals, mainly from the head and hands, making reliable full-body reconstruction a challenging task. Furthermore, tracking inaccuracies and motion jitter can negatively affect perceived realism, break the sense of presence, and even induce discomfort through visual–vestibular conflicts.

Despite achieving impressive frame-wise accuracy, existing monocular HPS methods often exhibit temporal instability when applied to video sequences. Because these approaches estimate pose independently for each frame, small estimation errors can accumulate into jittery and inconsistent motion trajectories. This limitation becomes particularly problematic in immersive VR applications, where fidelity and temporal consistency are critical for maintaining a convincing user experience. Consequently, improving the temporal coherence of human pose and shape estimation is not only important for animation and video editing, but also essential for the next generation of VR and AR systems.

Recently Vasilikopoulos et al. [3] introduced D-PoSE, a one-stage HPS framework that leverages depth estimation and body-part segmentation as representations that are beneficial for intermediate supervision. By jointly predicting depth maps, segmentation maps and SMPL-X parameters from monocular

This work was co-funded by the European Union (EUHE MAGICIAN - Grant Agreement 101120731).

RGB images, D-PoSE attains high frame-wise reconstruction accuracy. Nonetheless, it processes each frame independently and therefore does not explicitly exploit temporal continuity.

This paper investigates whether temporal awareness can improve the temporal consistency of D-PoSE without sacrificing reconstruction fidelity (Fig. 1). Overall, the proposed methodology makes the following contributions in HPS and the related field:

- We extend D-PoSE from a frame-based to a video-based pipeline that processes sliding windows of consecutive frames. The temporal extension does not change the spatial backbone or regression head, thereby preserving the strong spatial modeling capabilities of D-PoSE.
- We investigate recurrent (GRU) [4], [5] and attention-based (Transformer) architectures [6] for temporal aggregation, comparing bidirectional recurrence, purely temporal and spatio-temporal attention.
- We incorporate velocity and acceleration losses to encourage temporal smoothness and propose an online shape memory mechanism that reduces fluctuations in predicted body shape across frames.
- We conduct experiments on the BEDLAM [7] and 3DPW [8] datasets, evaluating reconstruction accuracy and temporal consistency. The results reveal favorable trade-offs between geometric fidelity and motion smoothness, demonstrating the suitability of our approach for VR applications that demand temporally stable full-body tracking.

Overall we examine the extent to which temporal modeling can mitigate the instability inherent in monocular HPS systems. The proposed extensions provide a systematic investigation of the accuracy–smoothness trade-off and demonstrate how temporal information can be leveraged to obtain more coherent human motion reconstructions.

II. Related Work

A. Single-Image Human Pose and Shape Estimation

Early human pose estimation methods represented the body as an articulated set of parts and inferred configurations by matching image evidence to kinematic or pictorial-structure models [9]–[11]. Such methods established important ideas about part relations and structural constraints, but they were sensitive to appearance variation, strong occlusions, and the depth ambiguities inherent in monocular images. Deep learning methods later improved 2D pose estimation by regressing joint heatmaps or coordinates from images, and improved 3D pose estimation either directly from images or by lifting detected 2D joints to 3D coordinates [12]–[16]. However, keypoint-based representations remain sparse and do not explicitly model body surface geometry or person-specific shape.

Human mesh recovery methods address this limitation by estimating the parameters of a statistical body model. Optimization-based approaches such as SMPLify fit a body model to detected image evidence [17], while regression-based methods predict body-model parameters directly from RGB images. HMR introduced end-to-end monocular recovery of SMPL pose and shape parameters, using an adversarial prior to discourage implausible bodies [18]. Subsequent work improved single-image HPS through model-fitting-in-the-loop supervision [19], part-guided robustness to occlusion [20], kinematics-aware regression [21], global-location-aware reasoning [22], and transformer-based mesh reconstruction [23], [24]. Intermediate geometric or semantic cues have also been explored: Neural Body Fitting uses body-part segmentation as an explicit intermediate representation [25], while DSR exploits semantic clothing information through differentiable semantic rendering [26]. D-PoSE follows this line by using estimated depth maps and body-part segmentation as intermediate cues for single-image HPS, and serves as the frame-wise baseline for our temporal extension [3].

B. Parametric Human Body Models

Parametric body models map low-dimensional pose and shape parameters to articulated 3D human meshes. SCAPE [27] and the widely adopted SMPL model [1] introduced a differentiable mesh representation based on pose and shape parameters. SMPL-X further extends this formulation with articulated hands, facial expression, and additional joints for expressive whole-body reconstruction [2]. These models underpin most modern HPS methods, including D-PoSE, by providing a compact and consistent body representation.

C. Sequence Models for Temporal Aggregation

Video-based HPS can exploit the fact that human motion is continuous and temporally correlated. Recurrent architectures such as LSTMs and GRUs maintain hidden states over time and have been widely used for sequence modeling [4], [5], [28]. In human pose and shape estimation, recurrent temporal encoders have been used to aggregate information across frames and reduce frame-wise jitter [29]–[31]. Temporal convolutional networks provide another efficient way to model motion history; VideoPose3D, for example, uses dilated temporal convolutions over 2D keypoint sequences for 3D pose estimation in video [32]. Transformers, utilizing self-attention, allow each frame to attend to other frames in a temporal window [6]. PoseFormer [33], 4DHumans [34] and MotionBERT [35], demonstrate the effectiveness of transformer-based temporal modeling for pose sequence and human motion representation learning.

Although focused on pose rather than full body-mesh recovery, these works highlight the value of temporal context and have influenced video-based HPS. Accordingly, we evaluate both GRU- and transformer-based temporal aggregation, comparing recurrent memory with attention-based context modeling.

D. Temporal Human Pose and Shape Estimation

Frame-wise HPS methods can produce accurate reconstructions on individual images, but applying them independently to video often leads to temporal inconsistency, jitter, and unstable

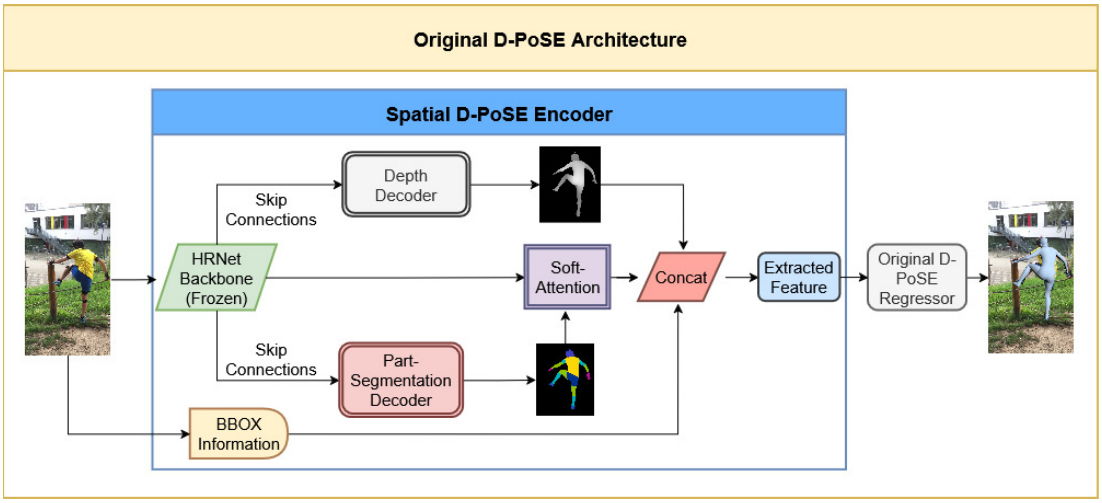

Fig. 2. D-PoSE processes frames independently and therefore does not explicitly model temporal continuity. The proposed extension incorporates temporal context to improve consistency across consecutive frames.

shape estimates. Early video-based methods therefore introduced temporal constraints or learned dynamics to improve coherence. HMMR [36] learns 3D human dynamics from video using a temporal encoding of image features. VIBE [29] uses a temporal encoder together with an adversarial motion discriminator trained on motion-capture data to encourage realistic motion sequences. TCMR [30] further revisits the use of temporal features, showing that past and future context should be exploited without letting the current frame dominate the representation. MPS-Net [37] introduces temporal-attentive mechanisms for monocular video HPS, while WHAM [38] extends video-based recovery toward world-grounded human motion estimation.

Temporal coherence can also be improved as a post-processing step. SmoothNet [39] is a plug-and-play temporal refinement network that reduces jitter in existing pose or body-estimation outputs without changing the underlying per-frame estimator. Such approaches are attractive because they can be applied to many backbones, but they do not directly integrate temporal reasoning into the image-to-body reconstruction pipeline. Our work instead builds on the D-PoSE single-image formulation and introduces explicit temporal aggregation over its learned representations, aiming to preserve the geometric benefits of depth and segmentation cues while improving consistency across video frames.

Despite substantial progress in video-based HPS, it remains unclear how temporal aggregation interacts with depth- and segmentation-guided human pose and shape estimation. The present work addresses this question by extending D-PoSE with temporal modeling and systematically comparing recurrent and transformer-based aggregation strategies, with the goal of improving temporal coherence while preserving reconstruction accuracy.

III. Method

A. D-PoSE Architecture

Our method extends the depth-guided D-PoSE framework [3], part of which serves as the spatial backbone throughout this work. D-PoSE estimates SMPL-X pose, shape, and camera parameters from a single RGB image by combining intermediate geometric and semantic representations with a part-aware regression framework.

Given an input image, an HRNet-W48 [40] backbone extracts multi-scale visual features. Two auxiliary decoder branches predict a dense depth map and a body-part segmentation map, which provide geometric and semantic supervision during training. The segmentation predictions additionally guide a part-aware attention mechanism that aggregates the spatial features into 22 body-part embeddings of dimension $C_f = 720$. These attention-derived representations, together with the global camera/shape features, are subsequently used by the D-PoSE regression head to estimate SMPL-X pose, shape, and camera parameters.

Unlike the original D-PoSE, which processes each frame independently, we retain the entire spatial pipeline unchanged and introduce temporal aggregation between the attention feature extraction stage and the regression head (Fig. 3). The temporal module refines the intermediate feature representations before they are passed to the original regressor, allowing temporal consistency to be learned while preserving the strong spatial modelling capabilities of D-PoSE.

B. Temporal Window Generation

Our temporal extension builds directly on top of the D-PoSE spatial architecture described above. During temporal training, the HRNet backbone remains unchanged and its weights are frozen. We insert an additional temporal aggregation module between feature extraction and regression: for each frame, we extract the attention-refined feature streams and reshape them into temporal sequences. These sequences are processed by a

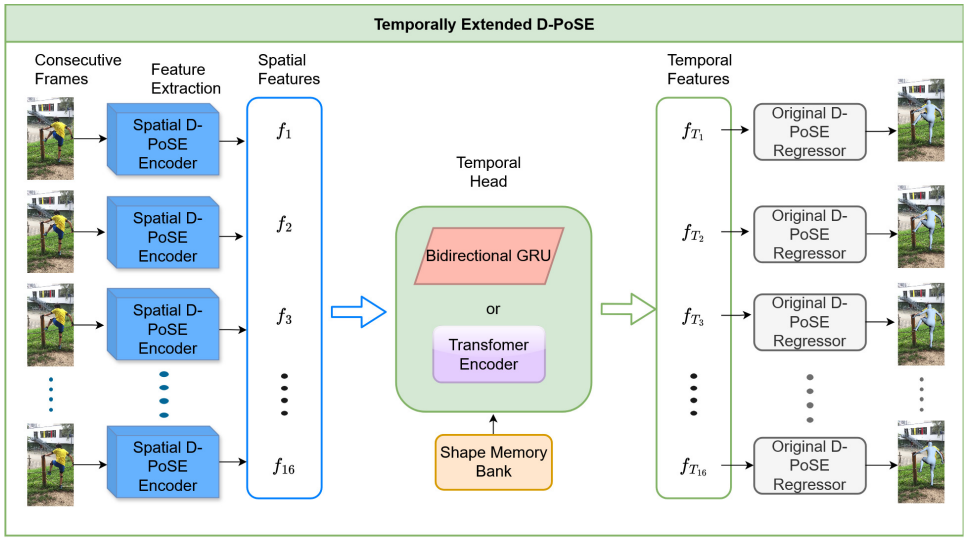

Fig. 3. Overview of the proposed framework. Frames are processed independently by the frozen D-PoSE spatial pipeline (Fig. 2). The resulting attention-based pose and camera/shape feature streams are grouped into temporal windows and refined using a temporal aggregation module (GRU or Transformer). An optional shape memory bank further encourages identity consistency before the original D-PoSE regression head predicts SMPL-X parameters.

recurrent or attention-based module to exploit motion continuity. The refined temporal features are then fed into the original regression head to produce pose and shape predictions. By preserving the spatial components of D-PoSE, our approach retains strong frame-wise reconstruction performance while adding temporal coherence.

More specifically, our temporally aware model starts by dividing each sequence into short overlapping windows. Let $V = \{I_1, I_2, \ldots, I_N\}$ denote a video with N frames. We sample windows of fixed length T using a stride s; the i-th window is $W_i = \{I_i, I_{i+1}, \ldots, I_{i+T-1}\}$. Sliding the window by stride s provides overlapping temporal context while increasing the number of training samples. Short windows keep memory requirements manageable and allow the framework to handle sequences of arbitrary length: regardless of how long V is, it is decomposed into independent windows that can be processed in batches.

In practice, video sequences rarely have a length divisible by T. When the final window contains fewer than T valid frames, we pad the remainder of the window by repeating the last valid observation. For example, if the final segment is $\{I_K, I_{K+1}, \ldots, I_N\}$ with $N - K + 1 < T$, then the padded window becomes $\{I_K, \ldots, I_N, I_N, \ldots, I_N\}$. Along with each window we store a binary validity mask $M \in \{0, 1\}^T$ indicating which positions correspond to real frames and which are padded. During training, temporal consistency losses are evaluated only over valid frame pairs or triplets, ensuring that artificial repetitions do not contribute to the gradients. The use of padding and validity masks thus permits fixed-length temporal input with a small memory penatly, without altering the effective supervision.

C. Temporal Feature Streams

Unlike post-processing approaches that smooth final pose estimates, our temporal module operates on intermediate attention-derived feature representations extracted by the D-PoSE spatial pipeline, thus enabling the network to refine the latent representation used for pose estimation. Specifically, two feature streams are used: (i) pose-related attention features and (ii) camera/shape-related attention features. These correspond to the outputs of the part-aware attention module before they are passed to the regression head.

For each frame t, the frozen D-PoSE network produces attention-based feature tensors $x_t \in \mathbb{R}^{C_f \times S}$, where C_f denotes the feature dimensionality and $S = 22$ corresponds to the body-part tokens produced by the attention mechanism. Each token represents a semantically meaningful body region obtained through segmentation-guided feature aggregation.

For a temporal window of length T, the per-frame features are stacked into a sequence $X = \{x_1, x_2, \ldots, x_T\}$, which is represented as $X \in \mathbb{R}^{B \times T \times C_f \times S}$ for a batch of B windows.

The temporal module operates directly on these body-part representations. By applying temporal aggregation before pose regression, the model can exploit motion continuity while preserving the rich spatial and semantic information extracted by the D-PoSE attention mechanism.

D. GRU-Based Temporal Modeling

We investigate gated recurrent units (GRUs) as our first temporal aggregation mechanism. GRUs are recurrent neural networks that maintain a hidden state summarizing past observations and use gating mechanisms to control the flow of information. In our setup, the sequence of attention-derived body-part embeddings $\{x_1, x_2, \ldots, x_T\}$ is processed by a bidirectional GRU layer. Each embedding contains information

from the 22 body-part tokens generated by the D-PoSE attention mechanism. Temporal modeling is therefore performed directly on semantically structured body-part features rather than on raw image features or predicted poses. The forward GRU reads the sequence from left to right, updating its hidden state $h_t^{\rightarrow}$ at each step; the backward GRU reads the sequence from right to left, producing states $h_t^{\leftarrow}$. The two states are concatenated to obtain the bidirectional representation $h_t = [h_t^{\rightarrow}; h_t^{\leftarrow}]$. Because our windows are processed offline during training, bidirectionality allows the model to exploit context from both past and future frames. In practice, we set the hidden dimensionality of each direction to 512 so that the concatenated state has size 1024. The output of the GRU is projected back to the original feature dimension via a linear layer and added to the input embedding via a residual connection. These residuals preserve frame-level information and facilitate optimization. We experiment with both a single shared GRU and a token-wise formulation. In the shared configuration, all body-part tokens are processed jointly within a common recurrent model. In the token-wise configuration, inspired by ReFit [31], each of the 22 body-part tokens is processed independently across time using a dedicated recurrent pathway . This allows different body regions to learn specialized temporal dynamics while maintaining the same overall architecture. Once temporal refinement is applied, the updated embeddings $\hat{x}_t$ are flattened along the temporal dimension and passed to the original regression head. This integration requires no changes to the SMPL-X regressor and thus preserves the baseline's spatial modeling capacity.

E. Transformer-Based Temporal Modeling

Transformers provide an alternative to recurrent networks for modeling temporal dependencies. Given a sequence of attention-derived body-part embeddings produced by the D-PoSE attention mechanism, a transformer encoder applies self-attention to model relationships across time. Unlike recurrent architectures, which process frames sequentially, self-attention allows each token to directly access information from all other frames within the temporal window. This enables the model to capture both short-term and long-range motion dependencies.

In our implementation, the transformer operates on the same attention-based feature streams used by the GRU variants. Specifically, the temporal module receives the $attention_pose$ and $attention_cam_shape$ representations produced by the frozen D-PoSE spatial pipeline. These features contain semantically structured body-part information that is subsequently refined through self-attention.

We employ a transformer encoder consisting of four stacked encoder layers with eight attention heads per layer. Sinusoidal positional encodings are added to the input embeddings to preserve temporal ordering, allowing the model to distinguish between otherwise identical feature vectors occurring at different positions within the sequence.

Two attention formulations are investigated. In the temporal formulation, attention is applied independently across time for each body-part token, enabling the model to capture motion dynamics for individual body regions. In the spatio-temporal formulation, body-part tokens from all frames are flattened into a single space-time sequence, allowing attention to be computed jointly across both temporal positions and anatomical regions. This enables the model to capture interactions between different body parts while simultaneously modeling their temporal evolution.

Each encoder layer contains a multi-head self-attention block followed by a position-wise feed-forward network with GELU activation. Residual connections are employed throughout the architecture to preserve the original feature information and facilitate optimization. The temporally refined embeddings are projected back to the original feature dimensionality and subsequently passed to the unchanged D-PoSE regression head for pose, shape, and camera estimation.

All transformer experiments operate on temporal windows of length T=16 frames. Larger temporal windows can provide additional context but increase computational complexity due to the quadratic scaling of self-attention with sequence length.

F. Shape Consistency Mechanism

To encourage stable body shape estimates, we introduce an optional shape memory mechanism. For each tracked individual, a running average of the shape-related feature representations is maintained and can be used to replace the current estimate. This encourages consistency of shape predictions while preserving pose dynamics. The mechanism operates online during both training and inference and can be integrated seamlessly with any temporal architecture, but may be extended with an offline averaging step similar to VIBE [29].

G. Training Objective

The total loss is a sum of frame-wise D-PoSE losses and temporal consistency losses. The original D-PoSE losses include 2D keypoint reprojection, 3D joint position, pose regression, shape regression, mesh vertex, depth, and segmentation terms [3]. Temporal consistency is encouraged through first- and second-order differences on both joint positions and pose rotations. The velocity loss is

$$\mathcal{L}_{\text{vel}} = \frac{1}{\sum_t M_t} \sum_{t=1}^{T-1} M_t \left\| \Delta x_t^{gt} - \Delta x_t^{pred} \right\|_1, \tag{1}$$

and the acceleration loss is

$$\mathcal{L}_{\text{acc}} = \frac{1}{\sum_t M'_t} \sum_{t=1}^{T-2} M'_t \left\| \Delta^2 x_t^{gt} - \Delta^2 x_t^{pred} \right\|_1, \tag{2}$$

where M_t and M'_t mask valid frame pairs and triplets. We optimize $\mathcal{L}_{\text{total}} = \mathcal{L}_{D\text{-PoSE}} + \lambda_{\text{vel}}\mathcal{L}_{\text{vel}} + \lambda_{\text{acc}}\mathcal{L}_{\text{acc}}$.

In our experiments we use two temporal loss settings. The moderate temporal loss (TL) employs moderate temporal regularization with $(\lambda^j_{vel}, \lambda^p_{vel}, \lambda^j_{acc}, \lambda^p_{acc}) = (0.5, 0.2, 0.2, 0.1)$ for joint velocities, pose velocities, joint accelerations, and pose accelerations, respectively. The second configuration, denoted as STL (Strong Temporal Loss), substantially increases the

influence of temporal supervision by setting the corresponding weights to $(5.0, 5.0, 2.0, 2.0)$. All results labeled TL or STL correspond to these settings.

IV. Experimental Evaluation

A. Datasets

a) BEDLAM: We train all temporal models on the BEDLAM dataset [7], a recent large-scale synthetic motion corpus designed for human mesh recovery. Each BEDLAM sequence is generated by animating a high-fidelity SMPL-X body with realistic motion capture data and rendering the result from multiple viewpoints. The dataset provides per-frame annotations for 3D joint locations, SMPL-X pose and shape parameters, dense mesh vertices, depth images and semantic body-part segmentations. With tens of thousands of frames encompassing a wide range of body shapes, motions and backgrounds, BEDLAM offers rich supervision for training. In our experiments we use BEDLAM exclusively for training and the test-split of 3DPW for validation.

b) 3DPW: Evaluation is performed on the 3D Poses in the Wild (3DPW) dataset [8], which contains real outdoor videos of everyday activities captured with on-body IMUs and handheld cameras. Each sequence includes synchronized IMU and monocular video recordings; the authors provide accurate SMPL annotations obtained through multi-view markerless motion capture. 3DPW sequences cover walking, running, sitting, carrying objects and other challenging actions in unconstrained outdoor environments. We use the official validation split for evaluation. Because our models are trained solely on synthetic BEDLAM data, evaluation on 3DPW measures cross-dataset generalization to real-world scenes.

B. Implementation Details

Our implementation is based on PyTorch and runs on a single NVIDIA RTX-A100 GPU. We initialize the HRNet-W48 backbone from the best frame-wise D-PoSE checkpoint and keep it frozen throughout temporal training to reduce memory consumption and isolate the effect of temporal modeling. Only the temporal modules, depth and segmentation decoders, feature projection layers, keypoint-attention layers, and the regression head remain trainable. As described earlier, sequences are divided into fixed-length windows. Unless otherwise noted we use windows of $T = 16$ frames with stride $s = 16$ during training; in ablation experiments we also explore a shorter stride of 8 frames to increase the number of training samples.

a) Training schedule: We employ the Adam optimizer for all our experiments, with an initial learning rate of 10^{-4} and zero weight decay. A "ReduceLROnPlateau" scheduler monitors the validation acceleration error every 500 iterations and reduces the learning rate by a factor of 0.5 when the error plateaus, down to a minimum of 10^{-5}. Gradient clipping with a maximum norm of 1.5 is applied to stabilize training. Batch sizes depend on the temporal module and GPU memory budget; for GRU models we use batches of 12-14 windows, whereas transformer models process 10-12 windows per batch. We train all temporal variants for up to two epochs on BEDLAM, selecting the checkpoint with the lowest validation acceleration error.

b) Augmentations: To preserve the natural temporal structure of the sequences, we disable all random image augmentations used in the original D-PoSE training. No random crop, flip, color jitter or scaling is applied. This ensures that consecutive frames correspond to physically consistent motion and that temporal losses are computed over meaningful trajectories. Similarly, we do not employ temporal augmentations such as time reversal or frame dropping.

C. Evaluation Metrics

We evaluate our models using standard metrics for HPS alongside a temporal smoothness measure.

a) Mean Per-Joint Position Error (MPJPE): MPJPE computes the average Euclidean distance between predicted and ground-truth 3D joint locations: $\frac{1}{J}\sum_{j=1}^{J}\|\hat{p}_j - p_j\|_2$, where J is the number of joints, p_j denotes the ground truth and $\hat{p}_j$ the prediction. We report MPJPE in millimetres.

b) Procrustes-Aligned MPJPE (PA-MPJPE): PA-MPJPE aligns the predicted and ground-truth skeletons with a rigid Procrustes transformation (rotation, translation and scale) before computing the average joint error. This removes global misalignment and focuses on articulated pose accuracy.

c) Mean Vertex Error (MVE): MVE measures the mean Euclidean distance between predicted and ground-truth mesh vertices, providing a dense assessment of the reconstructed surface. It is computed as $\frac{1}{V}\sum_{i=1}^{V}\|\hat{v}_i - v_i\|_2$, where V is the number of mesh vertices, and v_i, $\hat{v}_i$ the respective vertices.

d) Acceleration Error (AccErr): Because temporal coherence is our primary focus, we compute acceleration error following Kanazawa et al. [36]. For a sequence of predicted joint positions $\{\hat{p}_1, \ldots, \hat{p}_T\}$, we approximate the discrete acceleration as $\hat{a}_t = \hat{p}_{t+2} - 2\hat{p}_{t+1} + \hat{p}_t$. The error is then $\frac{1}{T-2}\sum_{t=1}^{T-2}\|\hat{a}_t - a_t\|_2$, where a_t is the ground-truth acceleration. Lower acceleration error indicates smoother and more temporally consistent motion and is widely used in video-based HPS because high-frequency prediction jitter appears as large second-order motion differences. All metrics are averaged over sequences in the validation set; lower is better for all metrics.

D. Overall Comparison

Table I reports quantitative results on 3DPW for the D-PoSE baseline and our temporal variants. All temporal models significantly reduce acceleration error compared to the baseline. GRU-based models achieve the strongest balance between accuracy and smoothness, with the per-joint GRU guided by Strong Temporal Loss (PJ-GRU + STL) reducing acceleration error by over 60% while also lowering MPJPE and MVE. Transformer-based models achieve even lower acceleration error (down to 11.4) but at some cost to frame-wise accuracy.

The results in Table I underscore the importance of modeling joint-specific motion. A shared GRU processes all joint features together, capturing global temporal patterns but failing

TABLE I
QUANTITATIVE COMPARISON ON 3DPW. LOWER IS BETTER. GRU: GATED RECURRENT UNIT; PJ: PER-JOINT; TL: TEMPORAL LOSS; STL: STRONG TEMPORAL LOSS; T-TRANS: TEMPORAL TRANSFORMER; SM: SHAPE MEMORY; AVG: TEMPORAL AVERAGING.

Method	MPJPE (mm)	PA-MPJPE (mm)	MVE (mm)	AccErr (mm)
D-PoSE Baseline	80.8	46.9	93.8	34.7
GRU-based				
Shared-GRU	82.1	48.2	95.5	29.7
PJ-GRU	81.9	48.0	95.4	25.2
PJ-GRU + TL	81.9	**46.5**	94.6	15.6
PJ-GRU + STL	**76.8**	46.7	**90.2**	12.8
Transformer-based				
T-Trans	87.9	49.7	104.6	12.7
T-Trans + SM	84.0	48.5	99.0	11.9
ST-Trans + SM	83.5	48.9	98.3	12.4
T-Trans + SM + Avg	83.8	50.2	99.2	**11.4**

to account for the heterogeneous dynamics of individual limbs. In contrast, the per-joint GRU maintains separate hidden states for each joint token, enabling it to learn distinct motion filters for arms, legs and torso. This flexibility reduces acceleration error from 29.7 mm to 25.2 mm while marginally improving PA-MPJPE. We conclude that allocating separate recurrent capacity to each joint is beneficial for articulated human motion, even when the overall parameter count remains similar.

Figures 4 and 5 provide qualitative examples of the effect of temporal modeling. Compared to the frame-wise D-PoSE baseline, the temporal extension produces substantially smoother predictions across consecutive frames while preserving pose accuracy. The reduction in frame-to-frame jitter is particularly evident in challenging articulated regions such as the hands, especially under occlusion, consistent with the improvements in acceleration error reported in Table I.

E. Ablation Studies

We analyze the impact of architectural choices, temporal loss weights, and shape memory. Table I additionally shows a comparison between shared GRU to a per-joint GRU. The per-joint GRU yields lower acceleration error and slightly better accuracy, suggesting that joint-specific dynamics are beneficial.

As seen in Table I, increasing the velocity and acceleration weights (STL) significantly reduces acceleration error while also improving frame-wise accuracy. Moderate regularization (TL) reduces jitter with minimal effect on accuracy. Comparing the last two rows reveals that strengthening temporal supervision has a doubly beneficial effect: the strong temporal loss (STL) not only reduces acceleration error by nearly 50% relative to the base PJ-GRU but also lowers MPJPE and MVE. The derivative-based terms act as a regularizer that suppresses high-frequency prediction noise while encouraging coherent joint trajectories. Excessively large weights could risk oversmoothing rapid motions; however, within the tested range, increasing the loss weights improved both temporal stability and frame-wise accuracy.

a) Shape consistency and Transformer variants: The shape consistency mechanism plays a crucial role in stabilizing transformer-based models. Without it, the plain temporal transformer (T-Trans) exhibits low acceleration error but comparatively higher MPJPE and MVE, indicating that the network has difficulty preserving identity across frames. Adding shape memory (T-Trans + SM) reduces MPJPE by 3.9 mm and MVE by 5.6 mm, while further decreasing acceleration error. Flattening the spatio-temporal dimension (ST-Trans + SM) enables the transformer to jointly attend over space and time, yielding a modest additional improvement in accuracy. A post-hoc prediction averaging step (T-Trans + SM + Avg) yields the smoothest motion but slightly increases MPJPE and MVE. These variants illustrate that transformer encoders can produce smooth trajectories when equipped with auxiliary mechanisms such as shape memory and prediction averaging, but they tend to lag behind recurrent models in per-frame accuracy.

b) Temporal consistency improvement: To quantify the benefits of temporal modeling, we compute the relative reduction in acceleration error compared to the frame-wise baseline, presented in Table II. A shared GRU reduces acceleration error by 14%, a per-joint GRU by 27%, and adding moderate temporal losses achieves an improvement of 55%. The strongest GRU variant (PJ-GRU + STL) achieves a 63% reduction while also improving MPJPE. Transformer variants provide even larger relative gains: T-Trans + SM reduces acceleration error by 65%, and combining shape memory with prediction averaging reaches a 67% improvement. These statistics underscore how temporal aggregation and derivative-based losses dramatically suppress motion jitter, and they highlight the trade-off between smoothness and frame-wise fidelity among different temporal architectures.

V. CONCLUSION

We presented temporal extensions of the depth-guided D-PoSE framework for monocular 3D human pose and shape estimation. Rather than introducing a new temporal architecture, this work systematically investigated how lightweight recurrent and attention-based temporal aggregation, derivative-

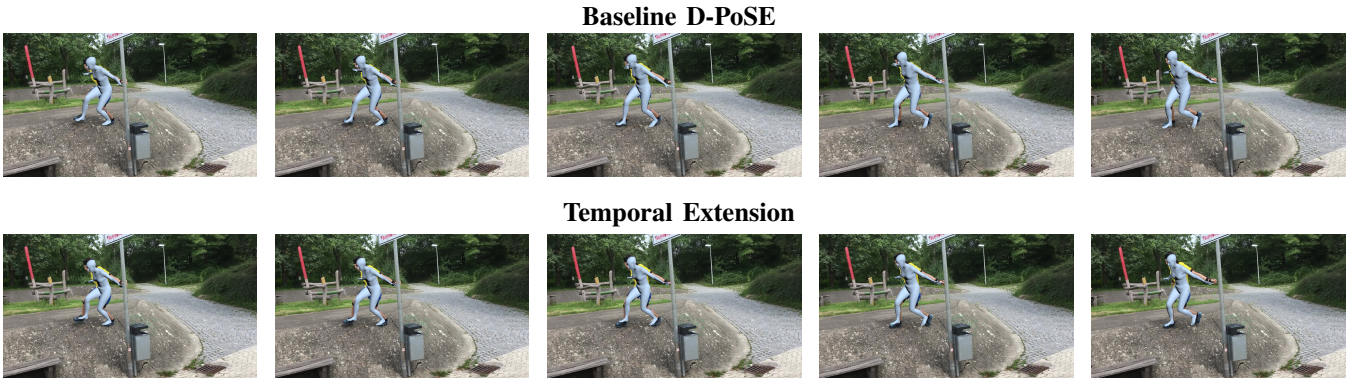

Fig. 4. Qualitative comparison over consecutive frames 135–139. The temporal model reduces prediction jitter, and maintains estimation accuracy.

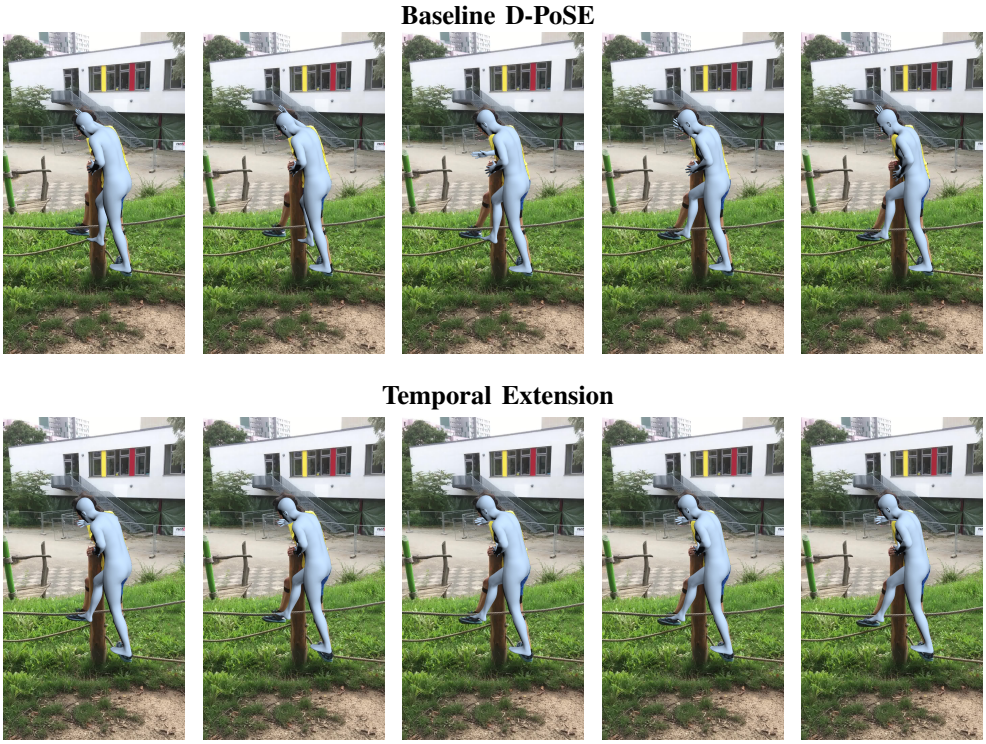

Fig. 5. Qualitative comparison over consecutive frames 401–405. The top row shows the frame-wise D-PoSE baseline, while the bottom row shows the proposed temporal extension. Although D-PoSE recovers plausible poses on a per-frame basis, small temporal fluctuations are visible across consecutive frames. The proposed temporal extension leverages motion continuity to produce smoother and more temporally coherent predictions.

TABLE II
ACCELERATION ERROR IMPROVEMENT RELATIVE TO THE BASELINE.

Model	AccErr	Improvement (%)
Baseline	34.7	–
Shared-GRU	29.7	14.4
PJ-GRU	25.2	27.4
PJ-GRU + TL	15.6	55.0
PJ-GRU + STL	12.8	63.1
T-Trans	12.7	63.4
T-Trans + SM	11.9	65.7
ST-Trans + SM	12.4	64.3
T-Trans + SM + Avg	11.4	67.1

based supervision, and an online shape memory mechanism interact with a depth-guided HPS pipeline. Across all evaluated variants, temporal modeling substantially improved sequence-level stability while maintaining competitive frame-wise reconstruction accuracy.

Our experiments demonstrate that recurrent and transformer-based temporal models provide different trade-offs between reconstruction fidelity and temporal smoothness. The best recurrent model (PJ-GRU + STL) reduced acceleration error by more than 60% relative to the frame-wise baseline while simultaneously improving MPJPE and MVE, while transformer-based models achieved the smoothest motion with acceleration error reductions of up to 67%. Overall, the results indicate that modeling temporal

context before pose regression is an effective strategy for reducing frame-to-frame prediction jitter while preserving the geometric benefits of the underlying D-PoSE representation. Beyond improving conventional HPS metrics, these gains are particularly relevant for monocular avatar reconstruction in VR and AR, where prediction jitter directly affects avatar stability, embodiment and telepresence. Although evaluated on standard HPS benchmarks, the observed reductions in temporal instability target artifacts that degrade immersive experiences.

One of the limimations of the current framework is that it operates on fixed-length 16-frame windows and relies on a frozen spatial encoder, preventing joint optimization of spatial and temporal representations. Furthermore, the experimental evaluation is limited to training on BEDLAM and evaluation on 3DPW. Finally, although the proposed architecture is compatible with sliding-window inference, we do not evaluate runtime latency or end-to-end deployment within a VR system. Future work will investigate causal temporal aggregation, joint end-to-end optimization, broader cross-dataset validation, and real-time evaluation in immersive VR applications.

REFERENCES

[1] M. Loper, N. Mahmood, J. Romero, G. Pons-Moll, and M. J. Black, "SMPL: A skinned multi-person linear model," *SIGGRAPH Asia*, vol. 34, no. 6, pp. 248:1–248:16, 2015.

[2] G. Pavlakos, V. Choutas, N. Ghorbani, T. Bolkart, A. A. A. Osman, D. Tzionas, and M. J. Black, "Expressive body capture: 3d hands, face, and body from a single image," in *IEEE/CVF CVPR*, 2019, pp. 10 975–10 985.

[3] N. Vasilikopoulos, D. Drosakis, and A. Argyros, "D-pose: Depth as an intermediate representation for 3d human pose and shape estimation," *arXiv preprint arXiv:2410.04889*, 2024.

[4] K. Cho, B. van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using rnn encoder–decoder for statistical machine translation," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, 2014, pp. 1724–1734.

[5] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, "Empirical evaluation of gated recurrent neural networks on sequence modeling," *arXiv preprint arXiv:1412.3555*, 2014.

[6] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *NeurIPS*, vol. 30, 2017.

[7] M. J. Black, P. Patel, J. Tesch, and J. Yang, "Bedlam: A synthetic dataset of bodies exhibiting detailed lifelike animated motion," in *IEEE/CVF CVPR*, 2023, pp. 8726–8737.

[8] T. von Marcard, R. Henschel, M. J. Black, B. Rosenhahn, and G. Pons-Moll, "Recovering accurate 3d human pose in the wild using imus and a moving camera," in *ECCV*, 2018.

[9] M. A. Fischler and R. A. Elschlager, "The representation and matching of pictorial structures," *IEEE Transactions on Computers*, vol. C-22, no. 1, pp. 67–92, 1973.

[10] P. F. Felzenszwalb and D. P. Huttenlocher, "Pictorial structures for object recognition," *IJCV*, vol. 61, no. 1, pp. 55–79, 2005.

[11] M. Andriluka, S. Roth, and B. Schiele, "Pictorial structures revisited: People detection and articulated pose estimation," in *IEEE/CVF CVPR*, 2009, pp. 1014–1021.

[12] A. Toshev and C. Szegedy, "Deeppose: Human pose estimation via deep neural networks," in *IEEE/CVF CVPR*, 2014, pp. 1653–1660.

[13] J. Tompson, A. Jain, Y. LeCun, and C. Bregler, "Joint training of a convolutional network and a graphical model for human pose estimation," in *NeurIPS*, vol. 27, 2014, pp. 1799–1807.

[14] A. Newell, K. Yang, and J. Deng, "Stacked hourglass networks for human pose estimation," in *ECCV*. Springer, 2016, pp. 483–499.

[15] J. Martinez, R. Hossain, J. Romero, and J. J. Little, "A simple yet effective baseline for 3d human pose estimation," in *IEEE/CVF ICCV*, 2017, pp. 2640–2649.

[16] K. Sun, B. Xiao, D. Liu, and J. Wang, "Deep high-resolution representation learning for human pose estimation," in *IEEE/CVF CVPR*, 2019, pp. 5693–5703.

[17] F. Bogo, A. Kanazawa, C. Lassner, P. Gehler, J. Romero, and M. J. Black, "Keep it smpl: Automatic estimation of 3d human pose and shape from a single image," in *ECCV*. Springer, 2016, pp. 561–578.

[18] A. Kanazawa, M. J. Black, D. W. Jacobs, and J. Malik, "End-to-end recovery of human shape and pose," in *IEEE/CVF CVPR*, 2018, pp. 5674–5683.

[19] N. Kolotouros, G. Pavlakos, M. J. Black, and K. Daniilidis, "Learning to reconstruct 3d human pose and shape via model-fitting in the loop," in *IEEE/CVF ICCV*, 2019.

[20] M. Kocabas, C.-H. P. Huang, O. Hilliges, and M. J. Black, "Pare: Part attention regressor for 3d human body estimation," in *IEEE/CVF ICCV*, 2021, pp. 11 127–11 137.

[21] J. Li, C. Xu, Z. Chen, S. Bian, L. Yang, and C. Lu, "Hybrik: A hybrid analytical-neural inverse kinematics solution for 3d human pose and shape estimation," in *IEEE/CVF CVPR*, 2021, pp. 3383–3393.

[22] Z. Li, J. Liu, Z. Zhang, S. Xu, and Y. Yan, "Cliff: Carrying location information in full-image fomo for human mesh recovery," in *ECCV*, 2022, pp. 590–606.

[23] K. Lin, L. Wang, and Z. Liu, "End-to-end human pose and mesh reconstruction with transformers," in *IEEE/CVF CVPR*, 2021, pp. 1954–1963.

[24] ——, "Mesh graphormer," in *IEEE/CVF ICCV*, 2021, pp. 12 939–12 948.

[25] M. Omran, C. Lassner, G. Pons-Moll, P. Gehler, and B. Schiele, "Neural body fitting: Unifying deep learning and model-based human pose and shape estimation," in *3DV*, 2018.

[26] S. K. Dwivedi, N. Athanasiou, M. Kocabas, and M. J. Black, "Learning to regress bodies from images using differentiable semantic rendering," in *IEEE/CVF ICCV*, 2021, pp. 11 250–11 259.

[27] D. Anguelov, P. Srinivasan, D. Koller, S. Thrun, J. Rodgers, and J. Davis, "Scape: Shape completion and animation of people," in *ACM SIGGRAPH*. ACM, 2005, pp. 408–416.

[28] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[29] M. Kocabas, N. Athanasiou, and M. J. Black, "Vibe: Video inference for human body pose and shape estimation," in *IEEE/CVF CVPR*, 2020.

[30] H. Choi, G. Moon, J. Y. Chang, and K. M. Lee, "Beyond static features for temporally consistent 3d human pose and shape from a video," in *IEEE/CVF CVPR*, 2021, pp. 1964–1973.

[31] Y. Wang and K. Daniilidis, "Refit: Recurrent fitting network for 3d human recovery," in *IEEE/CVF ICCV*, 2023, pp. 14 696–14 706.

[32] D. Pavllo, C. Feichtenhofer, D. Grangier, and M. Auli, "3d human pose estimation in video with temporal convolutions and semi-supervised training," in *IEEE/CVF CVPR*, 2019, pp. 7753–7762.

[33] C. Zheng, S. Zhu, M. Mendieta, T. Yang, C. Chen, and Z. Ding, "3d human pose estimation with spatial and temporal transformers," in *IEEE/CVF ICCV*, 2021, pp. 11 656–11 665.

[34] S. Goel, G. Pavlakos, J. Rajasegaran, A. Kanazawa, and J. Malik, "Humans in 4d: Reconstructing and tracking humans with transformers," in *IEEE/CVF ICCV*, 2023, pp. 14 783–14 794.

[35] W. Zhu, X. Ma, Z. Liu, L. Liu, W. Wu, and Y. Wang, "Motionbert: A unified perspective on learning human motion representations," in *IEEE/CVF ICCV*, 2023, pp. 15 085–15 099.

[36] A. Kanazawa, J. Y. Zhang, P. Felsen, and J. Malik, "Learning 3d human dynamics from video," in *IEEE/CVF CVPR*, 2019.

[37] W.-L. Wei, J.-C. Lin, T.-L. Liu, and H.-Y. M. Liao, "Capturing humans in motion: Temporal-attentive 3d human pose and shape estimation from monocular video," in *IEEE/CVF CVPR*, 2022, pp. 13 211–13 220.

[38] S. Shin, J. Kim, E. Halilaj, and M. J. Black, "Wham: Reconstructing world-grounded humans with accurate 3d motion," in *IEEE/CVF CVPR*, 2024, pp. 2070–2080.

[39] A. Zeng, L. Yang, X. Ju, J. Li, J. Wang, and Q. Xu, "Smoothnet: A plug-and-play network for refining human poses in videos," in *ECCV*, 2022, pp. 625–642.

[40] J. Wang, K. Sun, T. Cheng, B. Jiang, C. Deng, Y. Zhao, D. Liu, Y. Mu, M. Tan, X. Wang, W. Liu, and B. Xiao, "Deep high-resolution representation learning for visual recognition," *IEEE Trans. on PAMI*, vol. 43, no. 10, pp. 3349–3364, 2021.