CraftFormer: Clustered jitter Reduced human pose with Adversarial Fusion Transformer

Esha Dasgupta[1], Bo Eun Kim[2], and Sang Hoon Yeo[1] and Hyung Jin Chang[1]

[1] University of Birmingham
[2] Dankook University

Abstract. We propose a multi-stream fusion framework, Clustered jitter Reduced human pose with an Adversarial Fusion Transformer, **CraftFormer**, for regressing a subject's 3D pose smoothly from sparse, "in-the-wild" video sequences rather than dense frame sequences. Traditional Human Pose Estimation, HPE, suffers from jitter, occlusion and a reliance on structured lab data and can result in unrealistic movement trends. These make it unsuitable for use in applications where the smoothness and naturalness of motion is critical. To overcome these limitations, CraftFormer utilises a token clustering strategy to reduce computational load and redundancy to cluster the input video into a compact set of informative keyframes, process them through separate spatial and temporal streams, then lift the fused embeddings to 3D joint positions and SMPL body parameters via a cross-stream fusion transformer. A linear regression head maps the lifted representations to SMPL pose and shape parameters, providing kinematically valid outputs. We train adversarially against motion capture sequences from AMASS, using the motion capture distribution as a biomechanical prior that penalises physically implausible motion dynamics.

1 INTRODUCTION

Human Pose Estimation is the task of regressing the shape and the pose of a person to best represent their positioning either from an image or a video. It tracks and predicts a person's location by their structural information and keypoints. HPE from monocular video is a subset of this problem with applications in healthcare [17], sport science [13] and graphical fields such as animation [21].

This field has seen significant progress, however several challenges still remain and are insufficiently explored by existing methods. For example, the most popular approach consist of processing every frame independently [12] or with light temporal information [7], which drives up the computation and attention costs without taking advantage of the redundancy offered by the human body in movement. Whereas, pipelines which lift from 2D to 3D poses decouple detection from reconstruction but this introduces a dependency on the lifter and ensures reliance on its accuracy, which can propagate errors from the model into downstream tasks [2]. In addition, the task of removing jitter and maintaining temporal smoothness is still an open problem and this restricts HPE's use in physiological based applications where smooth movement is critical.

There are two main approaches to the problem of lifting 2D poses to 3D. Regression-based methods predict 3D joint positions and/or SMPL parameters directly from image features [1], with recent work additionally including temporal context through recurrent networks [3, 7] or transformers [10, 18, 20]. And Optimisation-based methods fit parametric models to 2D detected poses in an iterative manner when it comes to inference time [8], as a trade off between flexibility and computational costs. Our proposed approach, CraftFormer, is regression-based but integrates the SMPL forward pass into the training graph, combining the speed of regression with the anatomical guarantees of parametric fitting and smooth motion.

To the best of our knowledge, the works most relevant to this paper are MotionBERT [20] and 4D-Humans [4]. MotionBERT lifts 2D sequences to 3D using a dual-stream transformer whereas 4D-Humans estimates SMPL parameters [11] from video using a transformer backbone. We differ from MotionBERT in performing lifting entirely within the model without pretraining with an external lifter, and in operating on sparse clustered keyframes rather than dense sequences. We differ from 4D-Humans in our dual-stream separation of spatial and temporal evidence, our adversarial AMASS prior operating directly on SMPL sequences, and our explicit uncertainty estimation propagated through the attention mechanism. Sparse sequence representations have been explored in video regression [9] but not, to our knowledge, applied to 3D pose estimation with a learned clustering objective that jointly considers pose diversity and detection confidence. The use of AMASS as a discriminator reference distribution, rather than as a training set with rendered video, is a novel framing that decouples the kinematic prior from the visual appearance domain entirely.

We propose an adversarial multi-stream fusion framework to estimate the 3D human pose from an in-the-wild video with a lower computation cost, called CraftFormer. The intuition of this paper is that much of human motion can be broken down into segments of motion and that between snapshots of changes in direction, a person moves between stages. Therefore, it should be possible to recreate a full sequence with just the points where end points change between body parts and discard the redundancies of when a person is shifting between them.

There are four main ways we address the limitations of existing work in our paper.

First, clustering is used to reduce the computation cost. Rather than processing all the frames in a video, we reduce each video sequence to a set of keyframes via a clustering mechanism that selects frames as cluster medoids to act as a representation. This jointly maximises the selected pose diversity and the confidence of the 2D detection, so keyframe tokens are chosen from the frames where the subject data is most reliable. The redundant frames are then discarded, reducing sequence length from hundreds of frames to a much lesser amount of tokens. This helps simulate occlusion during training and guides the model to infer the connecting motion from body-part relations and kinematic priors alone

without direct frame by frame supervision. This encourages the model to learn the structural constraints of the human body.

Secondly, we use a dual-stream projection to combine local and global information with a spatial and temporal stream. The resultant keyframes are processed by parallel encoders with complementary functions. The spatial stream applies windowed self-attention within anatomically defined joint groups to capture body-part relationships and joint configurations per frame. The temporal stream applies full self-attention across all the keyframes, for each joint, to encode the subject's trajectory, velocity and cross-frame consistency. On their own, the streams alone cannot recover the subject's 3D pose: the spatial stream has insufficient depth evidence from a single view, and the temporal stream has no structural body-part information. But combined, they can query each other to learn what they lack and result in a better representation than either alone.

Thirdly, we use end-to-end SMPL lifting. CraftFormer does the 3D lifting by a cross-stream fusion transformer that receives both the spatial and temporal embeddings to output 3D joint positions with their uncertainty estimates. A linear regression head maps these joint positions to their SMPL pose and shape parameters, which are decoded through a differentiable SMPL forward pass. Gradients flow end-to-end through the SMPL layer into the fusion transformer and back through both projection streams. This removes an external lifter dependency and allows the parametric body model to enforce kinematic plausibility during the training instead of as a correction after.

And lastly, we use an adversarial biomechanical prior. We train against the AMASS dataset using a discriminator that looks at overlapping windows of the generated SMPL sequence. AMASS provides motion capture information across a variety of subjects and activities which serves as a reference distribution of physically plausible human motion. The discriminator learns to detect the artifacts introduced by monocular video lifting eg jitter, foot skating, implausible joint movements; and provides a guidance to the generator to follow the motion consistency of the AMASS dataset. Because both the generated and real sequences are represented as SMPL joint positions, the discriminator operates in a consistent representation space without bridging a domain gap. SMPL pose sequences are smoothed in the DCT frequency domain before being passed to the discriminator to remove high frequency jitter. This allows the discriminator to focus on motion dynamics rather than noise.

Together these components form a pipeline in which clustered keyframes are used to represent geometric and trajectory information, a fusion transformer lifts this 2D view to a 3D projection via cross-stream deduction, an SMPL layer enforces anatomical plausibility and the adversarial prior shapes the distribution of generated pose sequences towards what real human bodies look like.

Our contribution can be summarized as follows:

- We introduce the adversarial multi-stream fusion framework CraftFormer which estimates the pose of a person from an in-the-wild video subject to occlusions and jitter.

– We utilise a clustering approach that reduces video sequences to sparse keyframes while simulating occlusion as a self-supervised training signal.
– We utilise an dual-stream Transformer based architecture in CraftFormer that separates spatial reasoning from temporal reasoning, with cross-stream fusion lifting 2D views to 3D joint positions with per-joint uncertainty estimates.
– We utilise an end-to-end differentiable SMPL integration in the training, removing the need for an pretrained lifter and providing kinematic plausibility for the poses.
– We utilise a DCT smoothed adversarial training approach with AMASS that works as a learnable biomechanical prior.

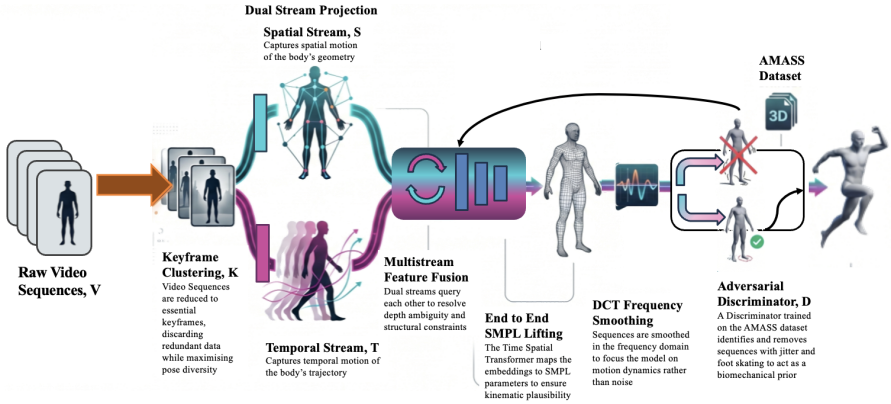

Fig. 1: Overall framework of CraftFormer. The multi-task framework takes a Video V as input and uses a clustering token strategy to reduce the sequence down to the essential keyframes, K_i. K_i are used to project temporal and spatial projections of the body over time to guide interpolation. Then, the Multi-stream Feature Fusion Encoder combines the two projections to recreate the pose sequence by querying information from each other. A Linear Regression Head projects them to SMPL parameters, then the sequence is smoothed in the frequency domain before a Discriminator, D ensures that the result is smooth and realistic.

2 RELATED WORKS

2.1 CNN Based Video Pose Estimation

Temporal Consistency has been a recurring issue tackled in the field. HMMR [6] learns the temporal representation of motion from a video by predicting the subject's motion in its past and future frames. By encoding the features, it produces smooth 3D mesh predictions and can hallucinate the temporal context.

However it lacks the ability to reason over long range motion due to its reliance on a 1D temporal convolution window. CraftFormer avoids this with a full self attention in the temporal stream which does not limit it to local context. VIBE [7] improves upon this by using an adversarial motion prior. It uses a GRU based temporal encoder to encode input image features into temporal features. A motion discriminator trained on AMASS was used to encourage plausible motion, but it works on GRU hidden states limited to local context and not the joint sequences directly. Our Discriminator works directly on SMPL joint positions which gives it more global context. MEVA [12] approaches the problem of making the reconstructed motion smooth in a two stage approach, capturing coarse motion first, with a variational motion estimator, before person specific detail is added in as finer detail, with a motion residual regressor. This constrains the motion to lie on a plausible manifold of motion but decoupling the motion components constrains the model's ability to reason about the larger trajectory and the full body geometry, leading to the finer motion being dominated by current static features and leading to inconsistency across sequential frames. CraftFormer has a similar split- the temporal and spatial streams to capture trajectory and body dynamics but they are fused via cross-attention instead of sequential processing to let them update each other. TCMR [3] addressed the issue of overdependence on the current frame by focusing more heavily on the temporal information encoded in the past and future frames using attention based integration. This results in a much smoother output and lower acceleration loss. However, TCMR operates on every frame in its sequence and we operate on a sparser collection of keyframes.

2.2 Transformer Based Pose Estimation

PoseFormerV2 [19] introduces a frequency domain representation of joint sequences to extend PoseFormer. It uses the DCT to represent the motion's visual identity while removing the high frequency noise. This compact representation helps scale up the receptor field and increase robustness to noisy joints. Our DCT smoothing is conceptually related to this, but in a different stage of the pipeline. MotionBERT [20] tackles the problem using a dual stream spatio-temporal transformer, DSTformer, in an unified pretraining framework. It has a high accuracy on the task and performs better when a pretraining stage is included. It is the work most relevant to CraftFormer, however there are differences. CraftFormer lifts entirely in its fusion transformer, whereas MotionBERT does it in two steps with a pretrained motion encoder lifter. MotionBERT also processes dense frame sequences. 4D-Humans [4] uses a transformer based approach to reconstruct and track SMPL meshes for subjects directly from input frames and video. Its high accuracy and performance make it a common upstream dependency in this field. Hourglass Tokenizer [9] suggests an amendment to Transformer based approaches by including a pruning and recovering network to prune the network of redundant tokens and recover them after. This makes it both efficient and accurate. This provides the intuition for the clustering approach in CraftFormer.

3 METHODOLOGY

3.1 Overall Framework

CraftFormer, E_{CF}, is an novel multi-stream fusion human pose reconstructor which uses an adversarial approach to reconstruct pose sequences, $\texttt{Pose}_{\texttt{CF}}$, from a sparse "in-the-wild" video input, V. The SMPL model framework is utilised to represent the output pose of the framework in terms of its shape and pose parameters, $\texttt{SMPL}\ \{\ \beta,\ \rho\}$.

As seen in Figure 1, CraftFormer is built on a transformer-based backbone.

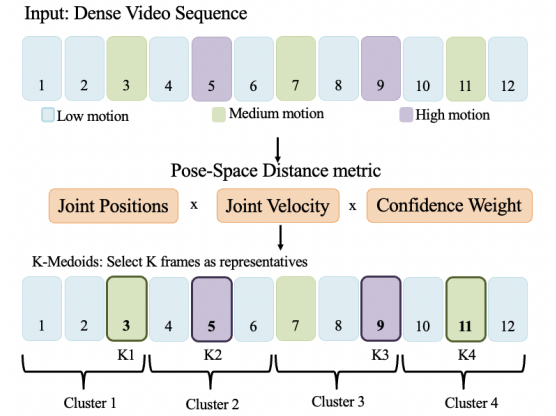

Fig. 2: This diagram shows how the token clustering is enacted on a frame sequence to select keyframe medoids and cluster the input into groups.

3.2 Video Input and a Token Clustering Strategy

Two of the main issues in Human Pose Estimation is that of jitter and occlusion affecting the reconstruction making it unrealistic and inaccurate. These are natural and standard in any in-the-wild videos and cannot be eliminated except in strictly controlled set-ups in vision laboratories which limits the usefulness of the videos taken there. In addition, processing the entire video sequence often leads to higher computational loads and redundancy as sequential frames are very similar to each other. Therefore, CraftFormer is designed to be able to mitigate some of the effects of both.

Rather than processing every frame in a video sequence, the model utilizes a clustering tokenizer to identify essential snapshots. This helps with reducing redundancy and minimizes the computational load by focusing on the most informative frames. In addition, by making the framework reason from incomplete sequences, the system simulates occlusion and forces the model to learn body-part relationships as a prior to accurately predict the pose. This approach helps

to deal with some of the common pitfalls, noise and gaps inherent for in-the-wild data in uncontrolled environments.

YOLO26 [16] is utilised to obtain the 2D keypoints for each frame across V along with the joint confidence weight scores. This allows the framework to cluster via similarity of only pose feature vectors. A pure appearance based clustering approach is avoided to avoid reliance on the subject's background and overlook sections of fast motion. We follow the intuition of the Token Pruning Cluster from [9] to cluster tokens but we utilize K-medoids instead of a K-Means based algorithm to cluster pose tokens so our selected cluster representatives correspond to a real frame in the sequence. The clustering criterion has to consider two properties, the confidence of the 2D pose token and how fast the motion is in the frames with sequences in the video of higher velocity movements having more essential frames to capture it. The distance, d, between two points, i and j, can be expressed as follows:

$$d(i,j) = \alpha * ||pose_i - pose_j|| * confidence\_weight(i) \tag{1}$$

Figure 2 shows the process in a visual format.

$$K_i = Cluster(V) \tag{2}$$

This results in the keyframes, K_i being extracted from the Video, V. The joint confidence scores allows for uncertainty to be propagated through the framework and let the later components weigh the data based on the likelihood of the pose rather than treat high and low confidence predictions equally. This is represented as a 3D uncertainty vector, the diagonal of the covariance of the different axes to allow for the different levels of uncertainty between the more straightforward 'flat' placement and the more ambiguous depth placement.

3.3 Temporal and Spatial Projections of the Data

The essential keyframes that define V, called K_i, obtained from the Clustering Strategy, are now split into two streams to simultaneously extract information from them and are passed through CNNs to obtain the features for each frame. The Temporal Projection, $Proj_{Temp}$, captures global information where it represents the full pose over time. This is the main indicator for depth and it serves to constrain a joint with low confidence in one frame by its velocity and movement in neighbouring keyframes. Per joint, the global stream encodes the velocity, acceleration and the temporal position per joint.

The Spatial Projection, $Proj_{Spat}$, captures local information where it focuses on the relations between the body joints within specific keyframes. The simulated occlusion from the clustering is training the model to hallucinate the unseen frames from the chosen keyframes by focusing on the body features such as joint angle limits and left-right symmetry. Interpolation across the snapshots is guided by these projections which consist of what a pose looks like in a given time and how the joints are changing given time. This provides context for the person's

movements. The stream encodes the joints, the confidence and the part group attention masks. Anatomical groups are defined for the local stream to restrict local attention to in-group tokens.

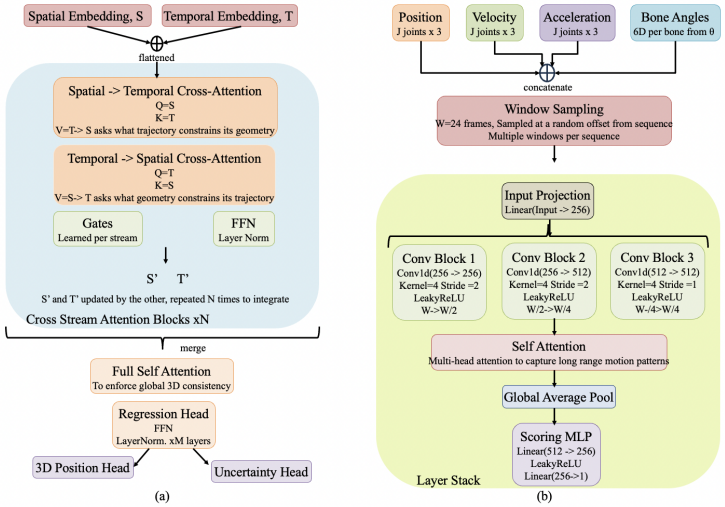

Fig. 3: a) This diagram shows the flow of data through the layers of the Time Spatial Transformer to lift a 2D pose to a 3D pose by utilising the information inherent in two different streams. b) This diagram shows the structure of the Discriminator in CraftFormer to guide the pose estimation towards plausible results.

3.4 Multi-stream Feature Fusion Encoder

Once CraftFormer has the temporal and spatial projections it fuses the two streams together in a Time-Spatial Transformer, TST, to take advantage of both local and global data to lift the 2D pose to a 3D SMPL pose. Figure 4 (a) shows the structure of the TST is more detail.

The TST is the component that resolves the depth ambiguities and constraining trajectories from the two streams. It serves to predict the 3D joint positions and their uncertainties, σ, after they've been lifted from 2D. In the fusion, each stream serves as a query with the key and value obtained from the other stream in order to triangulate in a bidirectional grounding. In the spatial-to-temporal direction, the query from $Proj_{Spat}$, and the key and value from $Proj_{Temp}$ work together to reason what the trajectory informs the framework given the body's geometry that cannot be determined from a single frame. In the temporal-to-spatial direction, the query from $Proj_{Temp}$, and the key and value from $Proj_{Spat}$ work together to reason where the body geometry tells the model where a joint could be given its trajectory across frames. Gates across each stream control how

much they update, then they are concatenated and projected through a linear layer to combine the data into a single unified space.

Following this unification, full self-attention is applied on the data to enforce full 3D consistency ie the body must be physically consistent across all of the frames. Finally, this learned representation is given to a fully connected layer in the TST to output the 3D joint positions and their uncertainties. After the TST, a linear regression layer takes this in as input and predicts the SMPL parameters of θ, the pose per frame, and β, which is shared across the frames. A larger set of layers are not necessary for this step and would add unnecessary parameters because the joint positions are a near linear function of θ.

3.5 Pose Cleaning

Once CraftFormer has the SMPL pose sequences predicted for the input, the pose sequence is cleaned to remove any residual jitter and sharp movement still remaining. This is done by applying a DCT to the sequence to transform it from the time to the frequency domain. A low pass filter is applied over the new frequency representation to keep the slower moving body movements and remove artefacts and higher frequency jitter. For CraftFormer we use a per-joint cut-off where the joints further out from the body eg wrists and ankles have a higher cut off threshold than the joints closer to the body centre eg the hips. This correlates with the natural motion of a body. Then the IDCT is applied to transform it back into the time range. This lets the Discriminator learn more smooth motions.

3.6 Discriminator

Finally, a Discriminator, D, is utilised as an adversarial approach to make the interpolation appear realistic. The function of this module is to distinguish real motion from CraftFormer's predicted motion to drive the reconstructed motion to avoid artefacts such as jitter, foot skating and penetration, unnatural movements and velocities and more. It fulfills this by being trained on the AMASS dataset, which is a set of clean motion capture sequences with SMPL data. Figure 4 (b) shows the structure of the Discriminator. Its input includes the details that are the most useful for it to understand what it has to guard against, velocity and acceleration teach it to look out for jitter and sharp movement, bone angles teach it to look for unnatural twisting and the position of the joints is useful for floor penetration.

3.7 Training

The training for the framework uses the loss terms as follows. The total loss is defined by:

$$L_{Total} = L_{joint} + L_{smooth} + L_{adv} + L_{shape} + L_{unc} \tag{3}$$

where L_{joint} controls the reconstructed pose taking the confidence into account so it learns more from the poses it is more confident about. Heteroscedastic loss is used to allow the framework to take advantage of the varying confidence of its predictions over the joints and the varying data quality.

$$L_{joint} = \sum_{j}(1/\sigma_j^2).\|J(\beta,\theta)_j - y_j\|^2 + log\sigma_j^2 \tag{4}$$

where σ_j is the per joint uncertainty, $J(\beta,\theta)$ are the predicted SMPL joints and y is the ground truth. L_{smooth} has the task of keeping the predicted motion smooth and stable between the frames with an L2 loss where θ_t is the pose at time t.

L_{unc} is a term to regularise σ in the heteroscedastic loss to avoid the framework from calling every predicted joint uncertain and having them contribute negligibly to the loss term. This avoids the confidence predictions from becoming useless. We do this by comparing the predicted 3D joint uncertainty to the 2D joint uncertainty that it was lifted from as a proxy.

$$L_{unc} = \sum_{j} \|\sigma\_pred_j - \sigma\_proxy_j\|^2 \tag{5}$$

The model is trained in three stages. The first stage involves the framework training with a subset of the loss equation; L_{joint}, L_{smooth} and L_{unc}; without predicting SMPL parameters and without the discriminator. The poses are predicted in 3D joint space with a bone length regulariser term to keep the skeleton consistent.

The function of L_{shape} in the loss term is to regularise the shape of the SMPL mesh and prevent extreme predictions with an L2 loss.

In the second stage, an SMPL parameter prediction head is added, L_{shape} is added to the equation, and fine tuned on the mapping between the predicted 3D joints and their ground truth SMPL poses. This avoids nonplausible predictions of SMPL parameters from the start.

L_{adv} shows the loss for the generator part of the framework which guides it to fool the discriminator whereas L_D shows the loss for the discriminator itself which guides it to distinguish between real and fake data. y is the real data in this circumstance and $\hat{y}$ is the predicted.

$$L_{adv} = -E[logD(\hat{y})] \tag{6}$$

$$L_D = -E[logD(y)] - E[log(1 - D(\hat{y}))] \tag{7}$$

The adversarial loss is not switched on from the start of the training. It is ramped up from nothing to a full component in the third stage of training. This is to let the Generator framework learn how to produce plausible poses first.

L_{smooth} and L_{adv} together tackle the task of keeping the motion smooth from jitter but realistic which often can have sharp angles and turns. To handle not

overlooking genuine sharp motion, L_{smooth} is relaxed as the training continues and the responsibility of training the motion is handed to L_{adv} when it comes into play in the total loss equation.

4 EXPERIMENTS

4.1 Evaluation Metrics, Datasets and Implementation Details

Evaluation Metrics: We report the Mean Per Joint Position Error (MPJPE), the Mean Per Vertex Position Error (MPVPE)and the Acceleration Error (Accel) [3] for our method.

Datasets: There are two datasets used for training CraftFormer; 3DPW [15] and Human3.6M [5]. 3DPW is an in-the-wild dataset with accurate 3D human poses. It consists of 60 video sequences with both 2D and 3D annotations, taken with a moving hand-held camera which introduces artefacts such as motion blur, occlusion and camera motion into the data which more controlled datasets lack. We also use the Human3.6M dataset. This is a large dataset consisting of 3.6 million images their subsequent 3D poses in a fully controlled laboratory space. It has 11 actors, 6 male and 5 female, who were photographed over 17 scenarios using 4 50Hz calibrated cameras, 1 time of flight sensor ad 10 motion cameras over an effective coverage of 4m x 3m. The recorded poses were calculated using time of flight data and have 24 labels for various parts of the body.

Meanwhile AMASS [14] is used to train the Discriminator to recognize plausible motion sequences. It is a large scale dataset of human motion that contains 40 hours of motion data over 300 subjects with a common parametarised body model, an extended SMPL-H and DMPLs, which is compatible with SMPL and SMPL-X.

4.2 Comparison with State-of-the-Art Methods

	3DPW			Human3.6M			
Method	*MPJPE* ↓	*MPVPE* ↓	*Accel* ↓	*MPJPE* ↓	*Accel* ↓	*Params*	Notes
HMMR [6]	116.5	139.3	15.2	-	-	37M*	ResNet50 backbone
VIBE [7]	91.9	-	25.4	65.6	27.3	48M*	ResNet50+GRU
MEVA [12]	86.9	-	11.6	76.0	15.3	53M*	ResNet-50 backbone
TCMR [3]	86.5	102.9	6.8	62.3	3.9	28M*	ResNet50+GRU
4D-Humans [4]	81.3	-	-	50.0	-	670.5M	ViT-H backbone
MotionBERT [20]	76.9	88.1	-	53.8	-	42.5M	dense 243 frames
CraftFormer	87.4	128.3	10.4	60.3	7.7	25M	sparse 16 frames

Table 1: Table to compare the evaluation metrics of CraftFormer against the state of the art models on the 3DPW and the Human3.6M Dataset. Parameter counts are given for the methods. Numbers with an * are calculated from architecture.

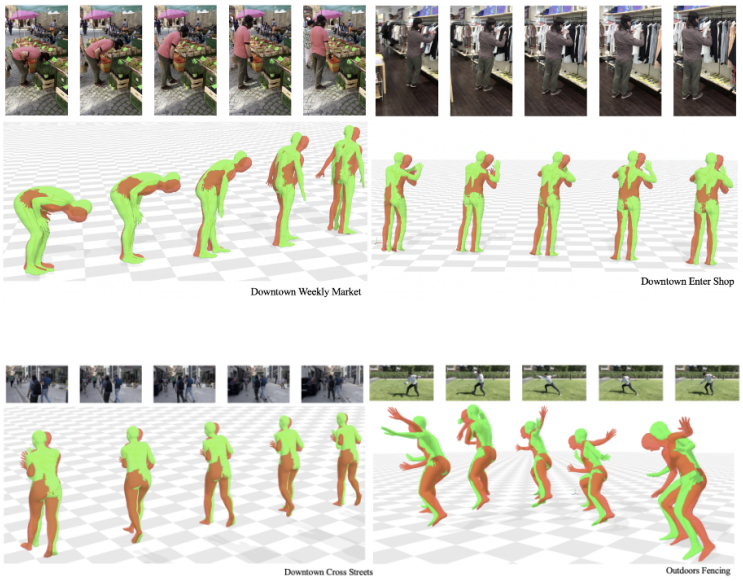

Fig. 4: A Figure to show the qualitative results of CraftFormer on the 3DPW Dataset. A sequence of images from the dataset videos are shown on the first row with the second row displaying the ground truth SMPL pose in green with the CraftFormer predictions in red overlaid on top.

We compare CraftFormer against other relevant works in Table 1. For evaluation we use the evaluation metrics and datasets outlined above. We note that while CraftFormer does not achieve state of the art, it achieves a respectable MPJPE and the second best acceleration result while using less frames to process with from the input videos and less parameters than the other methods. This is useful because current pose estimation is not smooth enough for use in fields such as robotics and physiological analysis. CraftFormer manages to remain competitive with the smoothness of its result while optimised for in the wild data and less expensive computation.

4.3 Qualitative Results

The qualitative results in Figure 4 highlight some of the results of CraftFormer. The top row visualises the input frames from 3DPW and the second row shows the predicted body pose in an SMPL representation in red, with the ground truths shown in green from the Dataset. Ait-Viewer was used to visualise the comparison. There are some interesting things to note. The regressed human poses are smooth across the frames with an internal consistency to the movement even if there is a visual gap between the predicted pose and the ground truth. This shows that the smoothing and the Discriminator are favouring enforcing physical plausibility. The greatest errors come when the subject is moving

Approach	MPJPE ↓	MPVPE ↓	Accel ↓
CraftFormer Base	87.4	128.3	10.4
Without Uncertainty	116.7	175.9	19.3
Without clustering	107.2	173.2	15.9
Without Temporal Projection	98.1	144.8	21.5
Without Spatial Projection	105.6	159.5	16.2
Without Fusion Transformer	102.5	142.0	17.8
Without Freq Filtering	92.8	136.4	17.1
Without Discriminator	109.3	169.6	19.6

Table 2: An Ablation study carried out on the 3DPW dataset to compare the effectiveness of the chosen pipeline for CraftFormer. Each row removes one component of the full model and compares against standard evaluation metrics.

quickly, as is natural, such as lunging or jumping with results more accurate when the subject is walking, standing still or taking part in movements that conform more to the distribution spread of AMASS. These errors are concentrated in the legs and the arms with a higher accuracy centred around the body's core.

4.4 Ablation Study

An Ablation Study is conducted to validate the structure of the model and pipeline and the results shown in Table 2. Seven tests are carried out on 3DPW to analyze the effect of removing a key feature of the framework to quantify its contribution. All ablations use the same training procedure and hyperparameters, varying one component at a time from our full model.

The first removes the Uncertainty Propagation from the pipeline which removes the ability of CraftFormer to be confident in ascertaining which frames and pose estimates it has confidence in to be reliable and robust. As shown the accuracy of the recreated pose, mesh and acceleration drops because the model relies more on data that is depth ambiguous and this affects the gradients in training.

The second study removes the Clustering of dense video into keyframes and uses the whole sequence as its input. This increases the sequence length that the architecture sees but the MPJPE performance worsens indicating that the occlusion simulating that the clustering was doing by discarding between frames is a helpful addition to the pipeline.

The third and fourth studies remove the Temporal and Spatial Projections respectively, which removes a large component of the information CraftFormer is able to process and gain context from. This is reflected in the higher errors confirming that they provide useful and non redundant information. Removing the temporal stream leads to a higher acceleration loss due to the removed constraint in trajectory but suggests that the spatial stream on its own provides a better depth clue than the temporal when it must. Relatedly, the fifth study explores the effect of removing fusing the dual streams together to integrate the local and global context and does worse which suggests that the two streams must collaborate with each other.

The sixth study looks at the resultant when the fused stream of data is not smoothed in the frequency domain. This has the least impact on MPJPE and MPVPE but has a significant impact on the acceleration error because this step allows the discriminator to focus on the motion dynamics that contribute to realistic motion and not per frame noise.

And lastly, the seventh study seeks to capture the effect of the Discriminator on the results by removing it. This results in a large spike in acceleration error and a significant spike in MPVPE showing that the Discriminator is providing an useful function to the pipeline by forcing CraftFormer to generate motion in the manner of the AMASS dataset.

5 CONCLUSION

This paper proposes CraftFormer, a transformer-based architecture for monocular 3D human pose estimation. Rather than processing dense frame sequences, we reduce video to a sparse set of informative keyframes via a clustering approach which reduces the computational cost while simultaneously creating implicit occlusion conditions that strengthen the model's learned body-part priors. Rather than relying on a pretrained lifter, we perform 3D lifting entirely within a cross-stream fusion transformer that combines complementary spatial and temporal embeddings through bidirectional cross-attention. Rather than applying SMPL as a post correction, we integrate the parametric body model end-to-end in the training graph, allowing kinematic constraints to shape learned representations. And rather than using AMASS as a source of rendered training video, we use it as a reference distribution for a temporal discriminator that functions as a learnable biomechanical prior, penalising the artifacts of monocular lifting.

Limitations: CraftFormer assumes a fixed number of keyframes per sequence, which does not adapt to the complexity of individual clips. Sequences with fast or variable motion may be undersampled at K=16, while static or slow sequences may retain redundant tokens which add computation costs without benefit. A dynamic K selected per video based on its motion would address this at the cost of variable-length batching. The discriminator is also trained on AMASS which has distribution gaps; walking is represented proportionally higher compared to poses such as sports, dancing or manipulation. This may limit the effectivity of the Discriminator on sequences from underrepresented action tasks. Finally, we cannot recover the absolute depth value without additional camera parameter supervision, and our performance on non-standard viewpoints or extreme body orientations reflect the ambiguity inherent to single view fitting.

Future directions: Future directions include extending this work is to anatomical muscle activation estimation. The SMPL output from our pipeline provides a vertex representation of the body, and the per-joint uncertainty estimates produced by the fusion transformer provide a natural confidence signal for downstream tasks.

References

1. Bogo, F., Kanazawa, A., Lassner, C., Gehler, P., Romero, J., Black, M.J.: Keep it SMPL: Automatic estimation of 3D human pose and shape from a single image. In: Computer Vision – ECCV 2016. Lecture Notes in Computer Science, Springer International Publishing (Oct 2016)
2. Chang, J.Y., Moon, G., Lee, K.M.: Absposelifter: Absolute 3d human pose lifting network from a single noisy 2d human pose. CoRR **abs/1910.12029** (2019), http://arxiv.org/abs/1910.12029
3. Choi, H., Moon, G., Chang, J.Y., Lee, K.M.: Beyond static features for temporally consistent 3d human pose and shape from a video. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2021)
4. Goel, S., Pavlakos, G., Rajasegaran, J., Kanazawa, A., Malik, J.: Humans in 4D: Reconstructing and tracking humans with transformers. In: ICCV (2023)
5. Ionescu, C., Papava, D., Olaru, V., Sminchisescu, C.: Human3.6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. IEEE Transactions on Pattern Analysis and Machine Intelligence (2014)
6. Kanazawa, A., Zhang, J.Y., Felsen, P., Malik, J.: Learning 3d human dynamics from video. CoRR **abs/1812.01601** (2018), http://arxiv.org/abs/1812.01601
7. Kocabas, M., Athanasiou, N., Black, M.J.: Vibe: Video inference for human body pose and shape estimation. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2020)
8. Kolotouros, N., Pavlakos, G., Black, M.J., Daniilidis, K.: Learning to reconstruct 3d human pose and shape via model-fitting in the loop. CoRR **abs/1909.12828** (2019), http://arxiv.org/abs/1909.12828
9. Li, W., Liu, M., Liu, H., Wang, P., Cai, J., Sebe, N.: Hourglass tokenizer for efficient transformer-based 3d human pose estimation (2024), https://arxiv.org/abs/2311.12028
10. Liu, J., Liu, M., Liu, H., Li, W.: Tcpformer: Learning temporal correlation with implicit pose proxy for 3d human pose estimation (2025), https://arxiv.org/abs/2501.01770
11. Loper, M., Mahmood, N., Romero, J., Pons-Moll, G., Black, M.J.: SMPL: A Skinned Multi-Person Linear Model. Association for Computing Machinery, New York, NY, USA, 1 edn. (2023), https://doi.org/10.1145/3596711.3596800
12. Luo, Z., Golestaneh, S.A., Kitani, K.M.: 3d human motion estimation via motion compression and refinement. In: Proceedings of the Asian Conference on Computer Vision (ACCV) (November 2020)
13. Luvizon, D.C., Picard, D., Tabia, H.: 2d/3d pose estimation and action recognition using multitask deep learning. CoRR **abs/1802.09232** (2018), http://arxiv.org/abs/1802.09232
14. Mahmood, N., Ghorbani, N., Troje, N.F., Pons-Moll, G., Black, M.J.: AMASS: Archive of motion capture as surface shapes. In: International Conference on Computer Vision. pp. 5442–5451 (Oct 2019)
15. von Marcard, T., Henschel, R., Black, M., Rosenhahn, B., Pons-Moll, G.: Recovering accurate 3d human pose in the wild using imus and a moving camera. In: European Conference on Computer Vision (ECCV) (sep 2018)
16. Sapkota, R., Cheppally, R.H., Sharda, A., Karkee, M.: Yolo26: Key architectural enhancements and performance benchmarking for real-time object detection (2026), https://arxiv.org/abs/2509.25164

17. Selim Gilon*, Emily Y. Miller, S.D.U.: Opencap: Human movement dynamics from smartphone videos (2023), https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1011462
18. Wu, C., Tan, G., Li, C.: Hevitpose: High-efficiency vision transformer for human pose estimation (2023), https://arxiv.org/abs/2311.13615
19. Zhao, Q., Zheng, C., Liu, M., Wang, P., Chen, C.: Poseformerv2: Exploring frequency domain for efficient and robust 3d human pose estimation (2023), https://arxiv.org/abs/2303.17472
20. Zhu, W., Ma, X., Liu, Z., Liu, L., Wu, W., Wang, Y.: Motionbert: A unified perspective on learning human motion representations. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (2023)
21. Çimen, G., Maurhofer, C., Sumner, B., Guay, M.: Ar poser: Automatically augmenting mobile pictures with digital avatars imitating poses. CGVCVIP (2018)