A Robust Improvement of PoseFormer

Zhen Wang, Jipei Chen, Tony Chen
u6904458, u7546904, u7504537

Abstract

While transformer-based methods have shown promise in 2D-3D human pose estimation, a notable limitation lies in their neglect of the quality of 2D joint detection, leading to error accumulation over time. To address this, our work introduces an innovative input quality analysis through Discrete Cosine Transform and Discrete Wavelet Transform, effectively smoothing input data and enhancing noise robustness. Our method, termed Robust PoseFormer, selectively preprocesses input skeletons using DCT, chosen for its superior performance in preliminary comparisons. Additionally, we integrate a sparse attention mechanism with a deliberately designed loss function to further mitigate cumulative noise issues. The experimental results, validated using the Human3.6M dataset, confirm our model's efficacy, demonstrating competitive results while addressing the core challenge of noise accumulation in temporal-dependent models. Our method not only enhances the accuracy of HPE but also proposes a novel perspective for addressing noise-related challenges in transformer-based models. ***Code is available at Robust PoseFormer-COMP8539 Repository***

1. Introduction

The original paper has highlighted the recent success in using a transformer-based method to achieve an uplifting 2D-3D human pose estimation (HPE). It offers a new perspective distinct from the conventional CNN approach, by extending its ability to capture long-term global dependencies of the entire input pose [10]. This model also performs well against many cutting-edge models at that time. However, the model has a intricate problem is that it ignores the quality of 2D joint detection and may lead to a error accumulation for a temporal-related model[8, 9].

To this end, we analyse the input frames' quality based on an innovative perspective. We randomly plot the curve of the changes in x-coordinate of a joint in the input skeleton sequence over time, as shown in Figure 1 (black curve), and find that there are a lot of jitters in the curve. After we applying a Discrete Cosine Transform to the curve and reconstruct the curve with the first 3, 9, and 27 coefficients respectively,

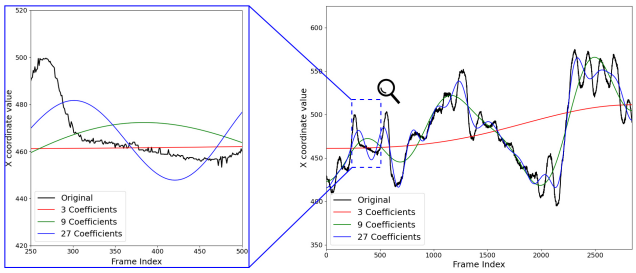

Figure 1: We randomly sample the x-axis of a joint from sequential frames in the input. In the figure, the black, red, green, and blue curves represent the changes in the x-coordinate of the same joint over time in the original input and its reconstructions using the first 3, 9, and 27 DCT coefficients, respectively.

we find that the curve becomes smoother, and as shown in Figure 1 (red, green, and blue curves), the reconstructed curve can capture the overall characteristics of the original video. This operation is equivalent to low-pass filtering the curve, and can filter out the high-frequency components, making subsequent processing more robust to noise.

In this paper, at the first stage, in order to improve the input data quality and make our processing more robust to noise, we deliver two methods to handle the problem in 2D-3D HPE. The two methods are Discrete Cosine Transform (DCT)[8, 9] and Discrete Wavelet Transform (DWT)[7], both of which are used to decompose the input signal or image into components of different frequency subbands. The advantage of the DWT is that it can provide a joint representation of time and frequency, while the advantage of the DCT is that it usually produces a sparser representation of the coefficients, which is an important advantage in many compression algorithms. After using two methods to preprocess the input skeleton sequence respectively, we compare and analyse the results, and finally select DCT as our data preprocessing method.

Moreover, we observed that introducing multiple frames with inherent noise can potentially result in error accumulation in time-dependent models, as highlighted by Mao et al.[8, 9]. Drawing inspiration from Child et al.[2], we strategically used a sparse attention mechanism as an output instead of using a linear regression head.

Consequently, our new coined model, Robust PoseFormer, integrated with denoised high-resolution frames, effectively mitigates the cumulative noise, achieve a competitive results compared to the original PoseFormer model.

2. Related Work

Our work is built upon the PoseFormer model [10]. Therefore, to begin, we will cover its general framework. The baseline method applies 2D-to-3D lifting HPE, treating each frame as a single token and concatenating all joint coordinates inside it. A variant of this approach introduces an additional dimension to denote the frame to which the token belongs. Recognizing that joint information is derived both spatially and temporally, PoseFormer directly models these aspects using two transformers in a spatial-temporal transformer architecture. This architecture comprises a sequence of spatial transformers followed by temporal transformer encoders, with the spatial transformers encoding spatial position embeddings using a linear encoder.

PoseFormer[10], conceptually, is the pioneering effort to use the ViTs as the primary framework for lifting-based 3D HPE, significantly surpassing earlier CNN-based approaches. Although Zhang $et\ al.$[13] acknowledge the competitiveness of the PoseFormer's performance, they argue that it might overlook unique temporal patterns associated with each joint. Consequently, they suggest employing alternating spatial-temporal transformer layers for more detailed joint-specific feature extraction. MHFormer[5] extends the structure of PoseFormer by integrating task-specific prior knowledge. They create several hypotheses to account for uncertain and ambiguous body parts to enhance performance in joint detection. Other works, like Strided Transformer[6], have identified the challenge of utilizing noise accumulated 2D pose sequences to produce a singular 3D pose. To this end, they introduced strided convolutions in the Vanilla Transformer Encoder to optimise accuracy and efficiency. Moreover, Einfalt $et\ al.$[4] also pointed to the same error accumulation problem, but they utilize masked token modelling for temporal upsampling of sparse 2D pose sequences, reducing computational complexity and enabling real-time 3D pose estimation.

Notably, the challenge of accrued noise in the temporal-dependent model has been a recurrent phenomenon. Therefore, besides changing the model's structure, the core problem naturally sheds light on the essential noise reduction task. However, using traditional image processing methods on the input pictures of 2D joints is not an effective way, since they are extremely different compared to conventional computer vision inputs, and it is necessary to select a new method [15]. To this end, Wang $et\ al.$[14], building upon the Transformer structure, present a noteworthy approach: they compress ViTs by emphasizing low-frequency components. Other work by Mao $et\ al.$[8, 9], using the DCT method with an RNN model in the HPE task, filters and refines the original input by selecting predominantly low-frequency data. Additionally, recent work by Magistris $et\ al.$[7] demonstrates the broad advantages of DWT in CNN-based pose estimation.

Inspired by the above works and the recent success of digital signal processing in the realm of computer vision, we adopt a frequency selection method tailored for an sparse-attention based model, aiming both to elevate accuracy and decrease input size. This technique, we believe, addresses the noise accumulation issue with a fresh and effective perspective.

3. Method

To lay the groundwork for our analysis, we will introduce the PoseFormer architecture. The spatial transformer feeds its results to the temporal transformer for temporal position embedding. The spatial transformer block receives a sequence of joint coordinates $X \in \mathrm{R}^{f\times(J\times 2)}$, where each frame's joint vector $\mathrm{x}^i \in \mathrm{R}^{1\times(J\times 2)}$ is treated as a 'pose token'. Through a linear projection layer, these tokens are transformed into higher-dimension features C, resulting in an embedding Z_0 that combines the positional embedding $E_{pos} \in \mathrm{R}^{J\times C}$ with the projected features $\left[x^1E, x^2E, \ldots, x^fE\right]$. This embedding is then passed to the transformer encoder consisting of MSA and MLP blocks, with the output for each frame being $Z_i \in \mathrm{R}^{J\times C}$.

Continuing to the temporal transformer block, the features for each frame $Z_i \in \mathrm{R}^{1\times(J\times C)}$ are concatenated across all frames, producing $Z \in \mathrm{R}^{f\times(J\times C)}$. After adding learnable temporal positional embedding and processing with MSA and MLP blocks, the output $\mathcal{R}^{f\times(J\times C)}$ is obtained. Finally, a weighted mean across the frame dimension yields the centered 3D pose $\mathcal{R}^{1\times(J\times C)}$, which is then condensed via an MLP block to the final output size $\mathcal{R}^{1\times(J\times 3)}$.

From preliminary experiments, we concluded that in the temporal transformer baseline, kinematic relationships are poorly captured due to the basic linear layer. This limitation arises from the overall model structure, where the spatial transformer is embedded into the temporal transformer module. Although this design choice aligns with human perception in the sense that each frame has its rich spatial information, we believe the model architecture sacrifices the global dependency between frames over the short-term kinematic relationship. Moreover, such an embedded model design will result in a high sparse encoding result inside the temporal transformer and is more likely to have quadratic computation growth with respect to the token number. As depicted in fig.2, we begin to elaborate possible improvements and forms the new proposed model, Robust PoseFormer.

MLP Attention Score in 3D HPE. Our goal is to diminish the input noise while maintaining the key features of the 3D HPE data. Therefore, we modify the original model's Attention structure from the original scaled dot

product Attention to MLP Attention to proficiently capture spatial-temporal relationships within pose sequences in our task. Given the input features $Z \in \mathrm{R}^{f\times(J\times C)}$, MLP Attention computes attention scores directly using a Multi-Layer Perceptron, eliminating the necessity for separate query, key, and value transformations. This produces attention weights $A \in \mathrm{R}^{f\times f}$, which are then applied to the value representations $V = ZW_V \in \mathrm{R}^{f\times C'}$, with $W_V \in \mathrm{R}^{(J\times C)\times C'}$. The final output is computed as $\text{Attention}(Z) = AV \in \mathrm{R}^{f\times C'}$. The robustness of MLP Attention to variations in input, coupled with its ability to apprehend complex relationships, renders it exceptionally apt for 3D HPE tasks, characterized by their highly variable input data.

Boost the short-term kinetic information. From the spatial transformer, we are interested to distinguish short-term kinetic over long-term dependency. The last step of original paper and other series of similar work in HPE domain all make the prediction based on the central frame from 2D output space. However, we think that there is potentially a lot sparsity in temporal attention and it cannot correctly penalize the frames that is not relevant to the prediction of the current central frame. On the other hand, they computed the weighted mean from $conv1d$, which is applying a learned filter to capture the dynamics over f dimension. It is a implicit form of weighted average without control of the fraction of importance from short-term kinetic information. To explicitly boost the short-term kinetic information and bring the most relevant frames to make the central prediction, we designed a self-attention mechanism over the frame's time dimension to compute the calculated attention score from other frames to this central frame. We refer it as the temporal attention and enable the dropout to insert some randomness increase the robustness.

With our capacity to pinpoint the most relevant temporal sequence $f^* = \text{argmax}_{f_i\in\mathcal{F}}\ \text{sparseAttention}(\mathcal{F})$, our next objective is to discount the influence of the least significant frame f_j in subsequent temporal block computations. We achieve this by establishing a robust update protocol that instructs the model to disregard updates from temporal positions receiving minimal attention during a given iteration. This is done by nullifying the gradient of the positional embedding tensor at those specific positions. By freezing the learning of the embedding of that frame, we assume that if this frame is relevant to the future batch sequence, they will be assigned greater weights through prolonged iteration. We also update the loss function to optimize the learning of temporal positional embedding to achieve selective sampling sequence with capturing the attention entropy loss for the attentions weights A as $L_{\text{penalty}} = -\sum_i A_i \log(A_i + \epsilon)$. So the total loss L is given by $L = L_{\text{MPJPE}} + \alpha \cdot L_{\text{penalty}}$.

Discrete Cosine Transform. Generally, the original skeleton sequence often contains jitters and noises, which tend to have higher frequencies in frequency domain. As

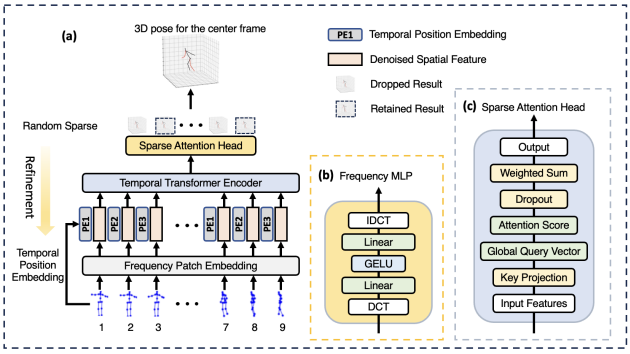

Figure 2: (a) Overview of our model. (b) FreqMLP (Frequency Multi-Layer Perceptron). (c) Sparse Attention Head.

shown in fig. 1, the input sequence of this paper also contains jitters and noises. In order to improve the input data quality, we introduce a low-pass filtering in this paper based on the DCT to suppress the high-frequency components of the input sequence. DCT is a method that can transform a signal or image containing limited data points from the temporal domain (or spatial domain) into the frequency domain. It is closely related to the Discrete Fourier Transform but uses only real numbers, making it more efficient for certain applications, such as image and video compression algorithms [1]. There are 4 common DCT types in total. In this paper, we use Type-II, which is the most commonly used form of DCT and is defined mathematically as follows.

$$X[k] = \sqrt{\frac{2}{N}} \cdot C(k) \sum_{n=0}^{N-1} x[n] \cos\left(\frac{\pi}{N}\left(n+\frac{1}{2}\right)k\right) \tag{1}$$

Here, $X[k]$ are the DCT coefficients, $x[n]$ are the samples of the original signal, n is the temporal (or spatial) index of the sequence, N is the length of the signal, and k is the index of the DCT coefficient. C(k) is a normalization coefficient defined as follows.

$$C(k) = \sqrt{\frac{2}{N}} \times \begin{cases} \frac{1}{\sqrt{2}} & \text{if } k = 0 \\ 1 & \text{if } k = 1, 2, \ldots, N-1 \end{cases} \tag{2}$$

To apply DCT to the project, a frequency multi-layer perceptron (FreqMLP) is designed in the feed-forward networks for features in temporal domain in this paper. In the FreqMLP, we apply DCT and Inverse Discrete Cosine Transform (IDCT) before and after the vanilla MLP. In the DCT part, we retain only the first finite number of DCT coefficients. This will remove the high-frequency components in the image. In the IDCT part, it reconstructs the low-pass filtered images from frequency domain to the temporal domain. The following is the mathematical expression of IDCT.

$$x[n] = \sqrt{\frac{2}{N}} \cdot C(k) \sum_{k=0}^{N-1} X[k] \cos\left(\frac{\pi}{N}\left(n+\frac{1}{2}\right)k\right) \tag{3}$$

After the processing above, most of the jitters and noises can be filtered out, but some normal motion details will be

mistakenly filtered out too. To address this issue, FreqMLP serves as an adaptable filter in the frequency domain. This enables the dynamic modification of the weight of each frequency component within the embedding of 2D joint coordinates, also known as temporal-domain features. In doing so, FreqMLP complements existing frequency-based features [12].

Discrete Wavelet Transform. In the same spirit of addressing the issue of jitters and noises in the input sequence, we also explore the use of DWT. DWT serves as another method for transforming a signal from the temporal or spatial domain into the frequency domain. Unlike DCT, which focuses on cosine bases, DWT employs both approximation and detail coefficients through the use of wavelets. This provides a more granular analysis of the input sequence [7].

To implement DWT in our framework, we introduce a wavelet-based multi-layer perceptron module. In this model, DWT is initially applied to the input sequence to compute both approximation and detail coefficients. We selectively keep only the lower-frequency approximation coefficients to suppress the high-frequency noises. Subsequently, a vanilla MLP is used to further refine the selected coefficients. Finally, the Inverse Discrete Wavelet Transform is applied to convert these coefficients back into the temporal domain, and reconstruct the output sequence.

4. Experiments

In this section, we meticulously evaluate our proposed method, detailing our experimental setup, including dataset, metrics, and implementation. While leveraging the comprehensive Human3.6M dataset and robust evaluation protocols ensures a thorough assessment, we acknowledge potential limitations due to severe computational constraints on the 3D HPE task and the use of a single dataset. We strive to balance extensive experimentation with available resource, and we discuss these content in the "Limitations" section, ensuring transparency in our results interpretation.

4.1. Dataset and Evaluation Metrics

Dataset. We utilized the Human3.6M dataset [16], renowned as the most comprehensive publicly available dataset for 3D HPE. It encompasses 3.6 million video frames, each meticulously annotated with accurate 3D ground truth positions. The dataset features 11 professional actors performing a variety of actions, captured from four unique camera perspectives. This diversity in activity types and settings provides a robust framework for evaluating the efficacy of our proposed methods across different scenarios. In alignment with the protocol established in the original PoseFormer work [10], our experimental configuration included using all 15 actions for both training and testing phases, with subjects S1, S5, S6, S7, and S8 designated for training, and subjects S9 and S11 reserved for testing.

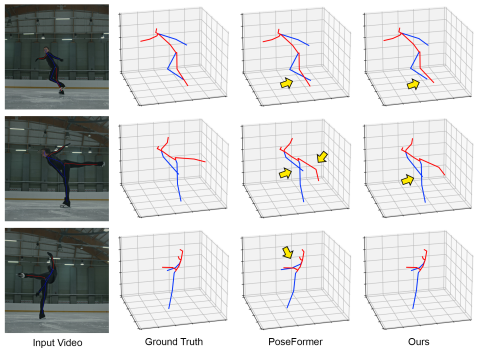

Figure 3: Qualitative comparisons of our model with PoseFormer and the ground truth.

Evaluation Metrics. To deliver a comprehensive evaluation of our proposed model, we employed various evaluation metrics, tailored to different input scenarios and protocols. Our analysis was executed under two distinct input conditions: 1) utilizing 2D poses detected by the Cascaded Pyramid Network (CPN) [3], and 2) leveraging ground truth 2D poses. The performance was gauged using three principal protocols: MPJPE, P-MPJPE, and N-MPJPE [11]. Specifically, the Mean Per Joint Position Error (MPJPE) calculates the average Euclidean distance between the predicted and ground truth joint positions, directly reflecting the prediction accuracy. The Procrustes-Aligned Mean Per Joint Position Error (P-MPJPE) conducts a Procrustes analysis to align the predicted and ground truth poses prior to computing the MPJPE, yielding a measure resilient to rigid body transformations. Lastly, the Normalized Mean Per Joint Position Error (N-MPJPE) normalizes the MPJPE by the torso size, accommodating variations in the subjects' body dimensions for a size-independent evaluation.

4.2. Implementation Details

Our proposed methodology was developed utilizing PyTorch. The computational workload for both training and testing phases was managed by two NVIDIA V100 GPUs. For the frame sequence lengths, a standard size of $frame = 9$ was consistently applied. To enhance the robustness and generalizability of our model, horizontal pose flipping was incorporated as a data augmentation technique during both training and testing stages. The optimization of our model was carried out using the Adam optimizer over the course of 130 epochs, with an applied weight decay set at 0.1. To systematically adjust the learning rate, we implemented an exponential decay schedule, starting from an initial learning rate of 2×10^{-4}, and subsequently reducing it by a factor of 0.98 after each epoch. The batch size for our training process was configured to 1024. Additionally, stochastic depth regularization was employed for the transformer encoder layers, using a rate of 0.1 to prevent over-fitting.

Table 1: Comparison of Original and Modified Models on Ground Truth Data

Protocol	Model	Dir.	Photo	Disc.	Eat.	WalkD.	Purch.	Pose	Walk	Greet	Phone	Wait	Sit	Smoke	WalkT.	SitD.	Average
MPJPE	Baseline	43.40	52.15	47.68	39.76	48.41	43.25	46.59	38.27	46.04	46.13	46.12	52.92	48.08	38.18	56.43	46.20
	Ours	39.27	48.44	44.73	35.64	45.27	39.82	44.68	35.58	42.30	44.01	43.68	51.35	43.77	35.98	54.56	43.30
P-MPJPE	Baseline	32.68	41.10	36.50	32.06	37.62	32.92	33.06	29.05	35.84	35.17	34.18	41.85	38.10	30.33	45.78	35.70
	Ours	30.28	38.76	34.28	28.49	35.84	30.41	31.96	28.00	33.30	34.01	33.42	39.82	35.26	28.26	43.40	33.70
N-MPJPE	Baseline	41.30	50.13	45.39	38.78	45.87	41.66	45.34	37.13	45.22	43.62	44.40	51.17	45.99	37.11	54.25	44.50
	Ours	38.20	46.71	43.27	35.03	43.39	38.63	43.73	35.06	41.71	42.32	42.67	49.67	42.52	34.80	52.56	42.00

Table 2: Comparison of Original and Modified Models on Cascaded Pyramid Network Data

Protocol	Model	Dir.	Photo	Disc.	Eat.	WalkD.	Purch.	Pose	Walk	Greet	Phone	Wait	Sit	Smoke	WalkT.	SitD.	Average
MPJPE	Baseline	47.17	59.40	50.76	47.88	53.65	47.77	48.49	39.27	50.42	52.57	48.29	59.87	52.42	41.12	67.86	51.10
	Ours	44.36	56.71	50.01	45.29	51.47	45.60	46.90	36.40	47.97	51.53	46.59	58.63	50.34	38.55	66.52	49.10
P-MPJPE	Baseline	36.24	45.33	39.02	38.03	41.44	36.05	36.63	29.25	40.61	39.68	36.32	48.90	41.80	32.89	54.12	39.80
	Ours	34.64	44.39	38.49	35.58	40.86	35.05	35.97	28.63	39.19	40.00	35.94	47.50	40.94	32.12	52.66	38.80
N-MPJPE	Baseline	44.61	57.37	49.13	46.31	51.38	45.05	46.89	37.28	49.05	50.31	46.74	58.40	50.54	38.89	65.24	49.10
	Ours	43.25	55.24	48.90	44.28	49.77	44.13	46.00	35.58	47.18	50.23	45.64	57.08	49.37	37.69	64.09	47.90

4.3. Comparisons with Original model

This section systematically compares our modified model's performance against the original using ground truth data and data from the Cascaded Pyramid Network. The results are detailed in Table 1 and Table 2. Lower values in these tables mean better performance, with red highlights indicating significant improvements by our model.

Evaluation on Ground Truth Data Our model surpasses the original in all aspects of ground truth data, showing improved pose estimation and adaptability to different body sizes. Figure 3 demonstrates our model's robustness to rigid body transformations.

Evaluation on CPN Data Moreover, our model consistently excels over the original with CPN data, effectively handling noisy 2D detections, alignment errors, and variations in body sizes, even in real-world scenarios.

These results confirm the superiority of our modifications across all evaluation protocols and data types. We recognize these advancements and suggest further validations with more datasets and real-world tests to confirm our model's effectiveness and broader applicability.

4.4. Ablation Study

In our comprehensive ablation study, we evaluated the impact of different model structures on our pose estimation performance, both on ground truth data and 2D poses detected by the CPN. Across the most evaluation protocols, our final model, integrating DCT, Sparse attention and new gradient method, consistently outperformed other variants, emphasizing its effectiveness in capturing pose information accurately. Notably on GT data, our final model achieved the best results, showcasing its robustness across different evaluation aspects (Table 3). The DCT plus Sparse attention Model also demonstrated competitive performance, especially in alignment and body size normalized metrics, signifying the potential of DCT in pose estimation tasks. Nevertheless, for the final protocol, our model exhibited a slight decrease in performance. It might be because our boosted method increases the torse size by decreasing some relative frames.

When evaluated on CPN data, our final model still held its ground, tying for the best MPJPE (Table 4). Interestingly, the DCT Model also performed exceptionally well, highlighting its capacity to handle real-world, noisy 2D detections. In contrast, the MLP Attention Model and the DWT Model, while showing some improvements in alignment-resistant performance, didn't manage to outshine the Original Model in other metrics. These results collectively underscore the superior capability of our final model, affirming our final strategic as a powerful combination for pose estimation tasks. This not only fortifies our model's performance under varied evaluation protocols but also establishes its adaptability to different data sources, paving the way for robust and reliable pose estimation in practical applications.

Table 3: Comparison across different models on GT data. The red font means the best, and the blue font means the second best. Note that the Baseline is our reproduced PoseFormer model.

Model Structure	MPJPE (mm)	P-MPJPE (mm)	N-MPJPE (mm)
Baseline*	46.2	35.7	44.5
MLP Attention Model	46.4	35.5	44.4
DWT Model	44.9	34.5	42.8
DCT Model	44.5	34.3	42.8
DCT + Sparse Attention	43.5	34.1	41.8
Our Final Model*	43.3	33.7	42.0

Table 4: Comparison across different models on CPN data. The red font means the best, and the blue font means the second best. The green font means that our model with only 40 epoch outperforms the baseline. Note that the Baseline is our reproduced PoseFormer model.

Model Structure	MPJPE (mm)	P-MPJPE (mm)	N-MPJPE (mm)
Baseline*	51.1	39.8	49.1
MLP Attention Model	51.4	39.4	49.4
DWT Model	50.5	39.2	48.9
DCT Model	49.9	39.3	48.5
DCT + Sparse Attention	49.9	39.2	48.3
Our Final Model*(only 40 epoch)	50.6	39.0	48.7
Our Final Model*	49.1	38.8	47.9

4.5. Limitation

The performance after introducing temporal attention showed moderate improvement over the originally reproduced PoseFormer, which is strong side effects that the temporal transformer can bring a negative influence to the prediction by introducing distinct long-term frames. It also shows the potential that such an embedded transformer inside transformer architecture could waste complexity over small benefits. There should be more space for improvement regarding using the spatial encoded features in the temporal dimension. To completely verify our finding with emphasis on short-term kinetic movements, we need to compare the current temporal attention with the temporal attention removing drop-out. Moreover, We use the baseline from our reproducible trials of the original PoseFormer implementation with 9 frames rather than the 81 frames recorded in their paper. We noticed that there were inconsistencies between the evaluation results of the publicly released code parameter and the original paper's result.

As we mentioned before the evaluation section, we use Human3.6M as the sole source for training and testing. The size of Human3.6M requires a full run-time of 2 days on the GPU cluster resource, and each run costs us considerable money. We have leveraged the computational resources we have at hand to deliver the analysis. We acknowledge that more datasets are publicly available to improve the training phase. For example, the MPI-INF-3DHP dataset contains many outdoor scenes. To enable more reasoning for the experiment, we should include a more comprehensive ablation study. While we recognize the merit of validating our approach on multiple datasets, the exhaustive nature of Human3.6M facilitated a thorough and meaningful evaluation, revealing both the strengths and areas for potential enhancement in our model.

5. Future Work

In this paper, we introduce DCT and DWT to complete the low-pass filtering of input data and improve the data quality. However, in this process, we did not optimise the selection of cutoff frequency. This means that we use a static passband threshold to adapt to any input data. This may lead to unsatisfactory data preprocessing results for some input sequences. In future work, we will explore the optimization of cutoff frequency so that it has better results for most input sequences.

Furthermore, we plan to introduce Kalman filter to the data preprocessing of this project, which is a dynamic adaptive filter. Kalman filter uses a series of measurements observed over time, containing statistical noise and other inaccuracies, and producing estimates of unknown variables that tend to be more accurate than those based on a single measurement alone. The dynamic adaptive characteristics of Kalman filter make it more ideal for most input sequences.

6. Conclusions

In the scope of this project, we have successfully reengineered the operational framework of the Spatial-temporal transformer. A thorough analysis of the PoseFormer architecture was conducted, leading to substantiated enhancements. By adhering to the configurations detailed in the original publication, we have managed to replicate the reported findings. Despite computational constraints, we executed a comprehensive evaluation using the complete sequence of the Human3.6M dataset, incorporating ground truth data for validation.

7. Confidential Peer Review

Zhen Wang. Propose ideas and coding to verify the feasibility. Perform evaluations using the improved model. Investigate the reasons behind the model's achievements.

Jipei Chen. Propose ideas and coding to verify the feasibility. Incorporate all the team's findings, evaluations, and related works into a comprehensive report.

Tony Chen. Preprocessing of input data, the implementation of DCT and DWT. Responsible for experiments setup on cloud GPU server. Provide suggestions to experiments.

Overall, all team member contributes equally to this project.

8. Personal Reflection

In summary, we worked well as a team, and everyone's contribution was key to our success. Our exploration into the PoseFormer architecture revealed inherent redundancies, leading to the implementation of several enhancements that enabled us to surpass the CPN MPJPE benchmark as reported by [10]. However, we were limited by our computational resources, affecting the depth of our experiments. Looking back, I could have discussed these limitations earlier with our lecturer or chosen a project that required fewer computational resources.

References

[1] N. Ahmed, T. Natarajan, and K.R. Rao. *Discrete Cosine Transform*. IEEE Transactions on Computers, vol. C-23, no. 1, pp. 90-93, 1974. DOI: 10.1109/T-C.1974.223784 3

[2] Rewon Child, Scott Gray, Alec Radford, and Ilya Sutskever. *Generating Long Sequences with Sparse Transformers.* arXiv:1904.10509 [cs.LG], 2019. https://arxiv.org/abs/1904.10509 1

[3] Yilun Chen, Zhicheng Wang, Yuxiang Peng, Zhiqiang Zhang, Gang Yu, and Jian Sun. *Cascaded Pyramid Network for Multi-Person Pose Estimation.* arXiv:1711.07319 [cs.CV], 2018. https://arxiv.org/abs/1711.07319 4

[4] Moritz Einfalt, Katja Ludwig, and Rainer Lienhart. *Uplift and Upsample: Efficient 3D Human Pose Estimation with Uplifting Transformers.* arXiv preprint arXiv:2210.06110, 2022. [Online]. Available: https://arxiv.org/abs/2210.06110 2

[5] Wenhao Li, Hong Liu, Hao Tang, Pichao Wang, and Luc Van Gool. *MHFormer: Multi-Hypothesis Transformer for 3D Human Pose Estimation.* In 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages 13137-13146, 2022. doi: https://doi.org/10.1109/CVPR52688.2022.01280 2

[6] Wenhao Li, Hong Liu, Runwei Ding, Mengyuan Liu, Pichao Wang, and Wenming Yang. *Exploiting Temporal Contexts with Strided Transformer for 3D Human Pose Estimation.* arXiv preprint arXiv:2103.14304, 2022. [Online]. Available: https://arxiv.org/abs/2103.14304 2

[7] Giorgio De Magistris, Matteo Romano, Janusz T. Starczewski, and Christian Napoli. *A Novel DWT-based Encoder for Human Pose Estimation.* In: Proceedings of the Scholar's Yearly Symposium of Technology, Engineering and Mathematics, Bruneck, Italy, July 23, 2022, CEUR Workshop Proceedings, Vol. 3360, pp. 33–40. CEUR-WS.org, 2022. https://ceur-ws.org/Vol-3360/p05.pdf 1, 2, 4

[8] Wei Mao, Miaomiao Liu, and Mathieu Salzmann. *History Repeats Itself: Human Motion Prediction via Motion Attention.* 2020. arXiv:2007.11755 [cs.CV]. 1, 2

[9] Wei Mao, Miaomiao Liu, Mathieu Salzmann, and Hongdong Li. *Learning Trajectory Dependencies for Human Motion Prediction.* 2020. arXiv:1908.05436 [cs.CV]. 1, 2

[10] Ce Zheng, Sijie Zhu, Matias Mendieta, et al. *3D Human Pose Estimation with Spatial and Temporal Transformers.* In IEEE CVPR 2021. 1, 2, 4, 6

[11] 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.* arXiv:2012.13392 [cs.CV], 2023. https://arxiv.org/abs/2012.13392 4

[12] Zhao, Qitao and Zheng, Ce and Liu, Mengyuan and Wang, Pichao and Chen, Chen. *PoseFormerV2: Exploring Frequency Domain for Efficient and Robust 3D Human Pose Estimation*. arXiv preprint arXiv:2303.17472, 2023. 4

[13] Jinlu Zhang, Zhigang Tu, Jianyu Yang, Yujin Chen, and Junsong Yuan. *MixSTE: Seq2seq Mixed Spatio-Temporal Encoder for 3D Human Pose Estimation in Video.* arXiv preprint arXiv:2203.00859, 2022. [Online]. Available: https://arxiv.org/abs/2203.00859 2

[14] Zhenyu Wang, Hao Luo, Pichao WANG, Feng Ding, Fan Wang, and Hao Li. *VTC-LFC: Vision Transformer Compression with Low-Frequency Components.* In Advances in Neural Information Processing Systems, edited by Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, 2022. [Online]. Available: https://openreview.net/forum?id=HuiLIB6EaOk 2

[15] Kai Xu, Minghai Qin, Fei Sun, Yuhao Wang, Yen-Kuang Chen, and Fengbo Ren. *Learning in the Frequency Domain.* arXiv preprint arXiv:2002.12416, 2020. [Online]. Available: https://arxiv.org/abs/2002.12416 2

[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, 2014. https://doi.org/10.1109/TPAMI.2013.248 4