Leveraging Temporal Context in Human Pose Estimation: A Survey

Dana Skorvankova[a] and Martin Madaras[b]
Department of Applied Informatics, Comenius University, Bratislava, Slovakia

Keywords: Human Pose Estimation, Temporal Context, Point Clouds, Visual Transformer, Deep Learning.

Abstract: Human pose estimation, the task of localizing skeletal joint positions from visual data, has witnessed significant progress with the advent of machine learning techniques. In this paper, we explore the landscape of deep learning-based methods for human pose estimation and investigate the impact of integrating temporal information into the computational framework. Our comparison covers the evolution from methods based on Convolutional Neural Networks (CNNs) to recurrent architectures and visual transformers. While spatial information alone provides valuable insights, we delve into the benefits of incorporating temporal information, enhancing robustness and adaptability to dynamic human movements. The surveyed methods are adapted to fit the requirements of human pose estimation task, and are evaluated on a real large scale dataset, focusing on a single-person scenario, inferring from 3D point cloud inputs. We present results and insights, showcasing the trade-offs between accuracy, memory requirements, and training time for various approaches. Furthermore, our findings demonstrate that models relying on attention mechanisms can achieve competitive outcomes in the realm of human pose estimation within a limited number of trainable parameters. The survey aims to provide a comprehensive overview of machine learning-based human pose estimation techniques, emphasizing the evolution towards temporally-aware models and identifying challenges and opportunities in this rapidly evolving field.

1 INTRODUCTION

Human pose estimation is a task of localizing skeletal joints positions of a person's body from visual data. It has witnessed remarkable progress in recent years, primarily driven by the growth of machine learning techniques. In this paper, we explore the landscape of deep learning-based methods employed for human pose estimation and delve into the impact and potential benefits offered by the integration of temporal information into the computational framework.

The field of pose estimation has transitioned from traditional computer vision methods to more sophisticated approaches, with deep learning at its core. Convolutional Neural Networks (CNNs), recurrent architectures, and attention mechanisms have emerged as pivotal tools, demonstrating unprecedented capabilities in capturing intricate spatial relationships and contextual dependencies within visual data.

While spatial information alone provides valuable insights into the pose of an individual, the temporal dimension introduces a new layer of understanding. Human movements are inherently dynamic, and capturing the temporal evolution of poses adds crucial context to the analysis. In this context, we explore the benefits of incorporating temporal information into pose estimation models. Temporal integration not only improves the robustness of pose predictions but also facilitates the recognition of complex actions and behaviors, making these models more adaptable to real-world scenarios where human activities unfold dynamically.

This survey aims to provide a comprehensive overview of the recent developments in deep learning-based human pose estimation techniques and their evolution toward temporally-aware models. By examining the current state-of-the-art methods and the advantages gained through temporal integration, we aim to offer insights into the challenges and opportunities that lie ahead in this dynamic and rapidly evolving field. The main contributions of this paper are as follows:

(1) Our experimental findings hold significant practical implications. With our experiments, we fill the gap in existing research by identifying the direct impact of temporal context incorporation on the ac-

[a] https://orcid.org/0000-0003-3791-495X
[b] https://orcid.org/0000-0003-3917-4510

Skorvankova, D. and Madaras, M.
Leveraging Temporal Context in Human Pose Estimation: A Survey.
DOI: 10.5220/0012696800003720

In *Proceedings of the 4th International Conference on Image Processing and Vision Engineering (IMPROVE 2024)*, pages 83-90
ISBN: 978-989-758-693-4; ISSN: 2795-4943

curacy and robustness of pose estimation. Thus, our research provides valuable guidance for the development of more effective and reliable applications. These insights can inform the design and implementation of practical solutions, enhancing the real-world performance of pose estimation systems across various domains.

(2) We systematically optimized various existing approaches that leverage diverse techniques for processing sequential data. Focusing specifically on the task of human pose estimation, we fine-tuned and enhanced these methodologies to achieve superior performance, and evaluate them on real 3D human dataset.

(3) In our experiments, we demonstrate that visual transformers elevate the field of pose estimation and improve the accuracy of both single-frame and temporal predictions. Attention-based strategies emerge as the optimal type of deep learning tool for achieving precise, robust, and efficient human pose estimation applications.

2 RELATED WORK

The domain of pose estimation has experienced rapid evolution, transitioning from CNN-based techniques to the integration of vision transformers. Early efforts in human pose estimation concentrated on single-frame analysis utilizing CNNs (Mehta et al., 2020; Mehta et al., 2017; Sun et al., 2019). Techniques like OpenPose (Cao et al., 2021) and AlphaPose (Fang et al., 2017) extended CNN approaches to handle scenarios involving multiple individuals, marking the advent of multi-person pose estimation. Moreover, CNN-based methods have played a crucial role in advancing 3D pose estimation, facilitating the prediction of three-dimensional human poses from 2D keypoints through techniques such as lifting from 2D to 3D (Kang et al., 2023; Nie et al., 2023). To address the issue of temporally incoherent estimates when dealing with individual frames, in recent years, human pose estimation has undergone substantial progress, emphasizing the significance of incorporating temporal context for a more comprehensive understanding of human motion. The integration of temporal context into pose estimation has advanced notably, showcasing innovations in recurrent neural networks (Artacho and Savakis, 2020; Hossain and Little, 2018) and graph-based methods (Li et al., 2022; Wu and Shi, 2023; Yang et al., 2021) to enable accurate and robust tracking of human movements over time.

Recent strides in the field of temporal pose estimation involve the integration of attention mechanisms and transformers into pose estimation architectures (Liu et al., 2020; Tang et al., 2023). Vision transformers (Dosovitskiy et al., 2021; Liu et al., 2021) have gained prominence for their global context capturing capabilities, yielding improvements in both 2D and 3D human pose estimation (Zheng et al., 2021). The attention mechanisms in transformers facilitate the consideration of long-range dependencies among keypoints, enhancing accuracy. However, many existing transformer-based methods typically follow a two-stage process, involving intermediate 2D pose estimation that is subsequently lifted into 3D (Einfalt et al., 2023; Li et al., 2023; Zhao et al., 2023). These approaches are constrained not only by the precision of the initial 2D joint positions but also by challenges related to self-occlusions and ambiguities arising from the absence of depth information.

In this survey paper, we aim to explore numerous end-to-end strategies including recurrent, graph-based and attention-based methods, which eliminate the need for two separate networks for estimating 2D and 3D poses in distinct stages. We avoid the ambiguities related to 2D input representations in our research. Instead, the focus is on single-stage temporal pose estimation approaches directly estimating 3D poses from 3D input data. Specifically, we use unorganized point clouds, as it is the most widely used and straightforward 3D data format. Despite its relatively sparse structure, it allows us to extract all the important information without requiring an exhaustive number of model parameters.

3 EXAMINED METHODS

Within our research, we have implemented various pose estimation methods and refinement strategies incorporating temporal information, following the latest trends in the field. All of the models presented below were implemented by us, inspired by existing methodologies.

3.1 Single-Frame Methods

To adequately evaluate the impact of the temporal context, we also performed experiments with single-frame pose estimation networks. They represent baseline methods, which we aim to further improve using the spatio-temporal approaches.

3.1.1 Baseline Pose Estimation

As our baseline single-frame pose estimation method, we established a simple MLP-based network, with a PointNet(Qi et al., 2017)-like architecture. The model takes a set of unordered 3D points as input and applies a shared multi-layer perceptron (MLP) to each point independently, capturing local features. Then, the per-point features are aggregated using max pooling to obtain global features across the whole data sample. The model directly outputs the 3D joint coordinates of the human skeleton.

3.1.2 Segmentation-Guided Pose Estimation

We also include somewhat advanced single-frame pose estimation approach (Škorvánková and Madaras, 2021) for the comparison. It consists of a two-stage pipeline. The first stage involves an auxiliary segmentation network that classifies points of a point cloud representing a human pose into corresponding body regions. In the second stage, the original input point cloud is concatenated with the output regions from the segmentation network, forming a four-channel point cloud input. This data, preserving both local and global information, is then fed into the second model—the regression network, where joint coordinates are regressed. The second model is essentially the same as our baseline pose estimation network. In both stages of the approach, residual connections are used in shared multi-layer perceptron blocks to enhance feature propagation.

3.1.3 Attention-Based Pose Estimation

In response to the current prominence of attention-based methods, we introduce an additional single-frame pose estimation approach denoted as Points in Transformer (PoinT). We incorporate both local and global feature processing for input point clouds in our architecture. This involves concatenating per-point features, initially extracted, with globally aggregated features spanning the entire point cloud. This can also be formulated as introducing the attention mechanism to the traditional PointNet, as we find it the most effective strategy to process point clouds. The diagram of our PoinT architecture is illustrated in Figure 1.

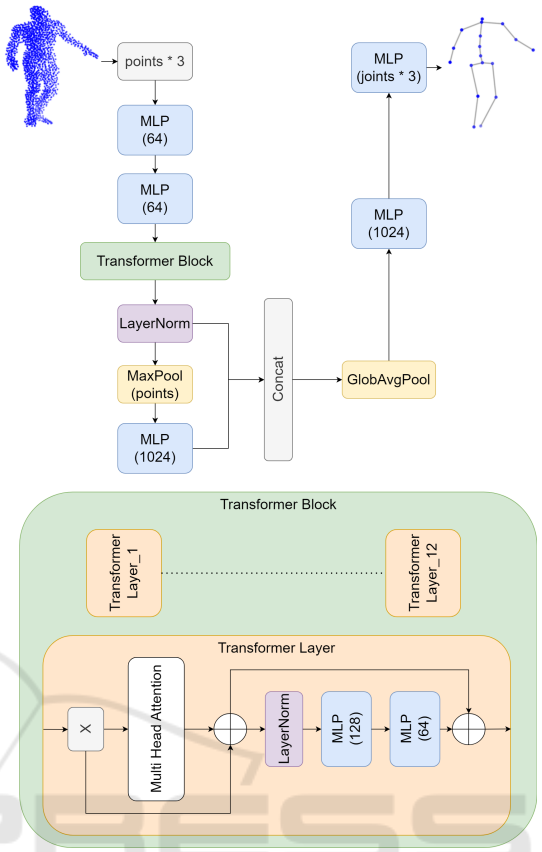

Figure 1: The architecture of PoinT model. Each MLP includes GELU activation and layer normalization. Numbers in brackets indicate number of units.

3.2 Temporal Methods

3.2.1 Pose Refinement Approaches

A portion of the spatio-temporal methods employs initially estimated human poses, inferred from a single frame. These poses are subsequently smoothed and refined by incorporating temporal context, involving the consideration of the sequence of surrounding frames during computation. Mainly, we used this strategy when the particular method required point-to-point correspondence within subsequent frames, a characteristic not inherent in unorganized point clouds.

Temporal Convolutions. First strategy we tested to refine initial single-frame pose predictions is using temporal convolutions (Lea et al., 2017; Pavllo et al., 2019; Chao et al., 2023). Unlike spatial convolutions that focus on spatial relationships within a single frame, temporal convolutions consider the temporal evolution of poses, recognizing the importance of motion dynamics for a comprehensive understanding of human actions. The size of the temporal kernel determines the extent of the temporal context taken into account. For our task, we also employ dilated temporal convolution kernels to extend the receptive field in time without increasing the number of model parameters. However, for the convolution across the temporal axis to be sensible, the point-to-point correspon-

dence between frames has to be maintained across the whole sequence of motion. In previous papers, temporal convolutions were applied either to 1D input representing joint locations, or 2D input images, both serving as organized data structures. Since we employ unorganized 3D point clouds as input, we use the temporal convolution approach only for fine-tuning the initially predicted single-frame poses.

Sequence-to-Sequence Modelling. Another approach we have included in our survey is using an architecture based on sequence-to-sequence modelling inspired by Hossain et al. (Hossain and Little, 2018). The model employs LSTM modules which are interconnected in an encoder-decoder fashion. We use this type of network, again, to refine the initially estimated 3D human poses using the preceding frames in the sequence. The technique could not be used directly on sequences of input point clouds, since the frame-to-frame correspondence is missing in the unordered data structure.

3.2.2 End-to-End Approaches

The other part of our experiments focus on direct approaches, which take a sequence of unordered point clouds as input, and learn to estimate 3D joint locations of the tracked person for the reference frame.

Temporal Dynamic Graph CNN. One of the end-to-end strategies we have experimented with is using a dynamic graph convolutional neural network (DGCNN) inspired by Wang et al. (Wang et al., 2019). The original model proposed in the paper was employed to address high-level tasks on single-frame point clouds. The network is based on graph convolutions, hence representing the point cloud as a graph structure, dynamically updating the graph in-between layers. The so-called *EdgeConv* operation consists of computing per-point features by applying a multi-layer perceptron (MLP), constructing a graph based on nearest neighbors in the feature space, and pooling among the neighboring edge features. The main contribution of the approach is the suggestion to recompute the graph after each MLP, based on inter-point distances in feature space.

We adopted the idea of dynamic graph convolutions and took it further by designing a Temporal DGCNN, incorporating the dynamic graph topology into our pose estimation (SGPE) regression network. The proposed architecture is depicted in Figure 2. As illustrated in the figure, we feed a sequence of point clouds into the model and concatenate the global per-frame features before feeding them to the bottleneck to regress the 3D joint coordinates of the last (reference) frame.

PointLSTM. Another strategy we examined and implemented is LSTM model directly processing unordered point clouds, following the research of Min et al. (Min et al., 2020). Originally, this strategy was applied to solve the task of gesture recognition, however we aim to utilize the approach to track human body. We acquired the per-point internal states within the LSTM. For each point of the point cloud, the hidden states are updated by aggregating relative features of its K nearest neighboring points in the previous frame.

Following the original paper, we integrated PointLSTM into a modified FlickerNet architecture (Min et al., 2019), replacing one of the network intermediate layers by the PointLSTM module. The architecture consists of five subsequent modules. In the first stage, intra-frame features are extracted using spatial neighborhood grouping. In the second to fourth stages, inter-frame features are extracted with spatial-temporal grouping, and the point clouds are sub-sampled using density-based sampling between two neighboring inter-frame layers. We are experimenting with three distinct models based on which layer is replaced by PointLSTM: (1) PointLSTM-early, (2) PointLSTM-middle, and (3) PointLSTM-late. The three inter-frame layers in the FlickerNet are replaced, respectively, to examine how well the LSTM can extract important features at various stages.

4 EVALUATION

4.1 Benchmark Data

Within our experiments, we use the CMU Panoptic dataset (Joo et al., 2017) to train and test the models described above. It is currently the only large-scale dataset containing multi-modal data capturing real human subjects interacting in various scenarios. For the sake of our research, we employ the portion of the dataset that focuses on a single person in the scene, marked as *Range of motion*. It includes over 2 hours of recordings, which yields over 141,000 frames in total. Since prior to our work, there was no protocol established for the utilized section of the dataset and the stated task, and considering the amount of data present in the selected part of the dataset, we split the data to train and test set with 70:30 ratio.In the pre-processing steps, the sequences are further sliced to generate input sequences for the particular methods. We initially sub-sampled all the point clouds to 512

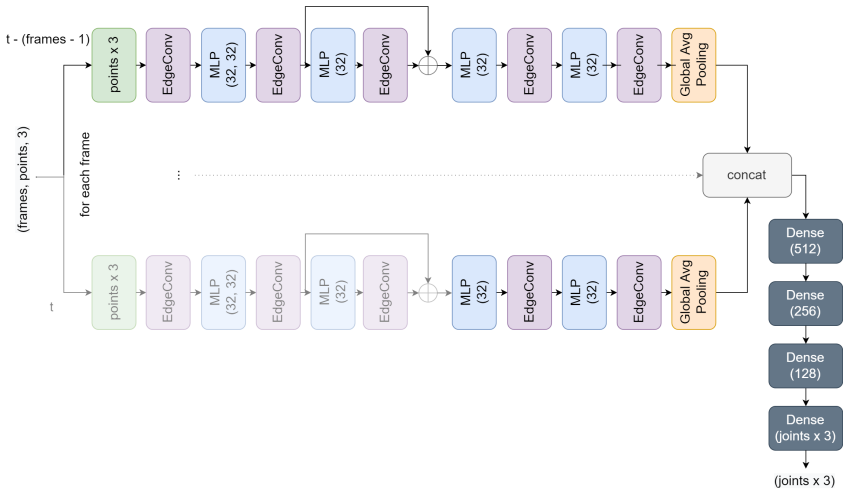

Figure 2: The proposed architecture of Temporal DGCNN. EdgeConv layer composes of the graph recomputation and max aggregation of the neighboring features.

points using farthest point sampling (FPS), and then decreased this size even further in some of the methods, as indicated in the next section.

4.2 Results

Following the temporal convolution strategy, after an extensive number of experiments, we obtained the best results with a simple model, which convolves across sequences of 9 frames at a time. The input sequences of initial 3D pose estimations in our experiments are produced by our baseline model, SGPE, and PoinT network, as described in Section 3.1. Furthermore, we validated both symmetric and causal temporal convolution settings. Symmetric convolution means the reference frame is located in the centre of the input sequence, while causal convolution only has access to past frames. We can conclude from the results shown in Table 1, as well as in Figure 4, that fine-tuning the single-frame pose estimations using temporal convolutions increases the accuracy of all of the single-frame models we have experimented with.

Regarding the sequence-to-sequence LSTM network, we preserved the original number of 1024 units inside the LSTM cell on both encoder and decoder side. We have also tested multi-layered decoder consisting of multiple sequentially chained LSTM cells, however our best results were achieved with just one layer for both encoder and decoder. Based on the results (as shown in Table 1), we may conclude the accuracy of our single-frame pose estimation is already high, and may not be largely affected by outliers caused by time-inconsistent predictions. Hence, the sequence-to-sequence refinement does not lower

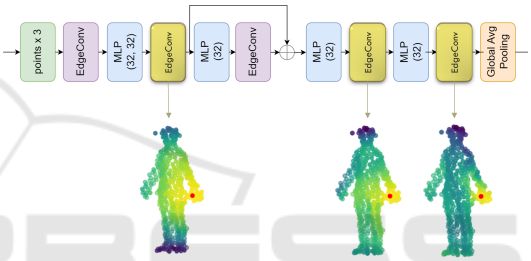

Figure 3: Structure of the feature space produced at different layers of the Temporal DGCNN. The distance in feature space from the red point to all the other points in the point cloud is visualized.

the mean error, however slightly increases the mean average precision of the estimations. The reported results were achieved using the input sequence length of 5 frames, same as in (Hossain and Little, 2018); with the temporal loss incorporated into the training process. Temporally computed loss means the error is calculated not only against the reference frame ground truth, but also against the previous frames from the sequence. The further from the reference frame it is located in the sequence, the lower weight is assigned to the loss computed from that ground truth pose. Using a simple mean absolute error as a loss function yielded the mean per joint position error of approximately 2.43 cm, whilst incorporating the temporal loss it has slightly decreased to 2.38 cm.

During the experiments, we have also validated the hyper-parameters of the Temporal DGCNN, such as the number of nearest neighbors used while constructing the graph, the input sequence length, and the input point clouds resolution. We obtained the best results using 20 neighboring points in EdgeConv, sequence length of 5 frames with stride 2 (taking every

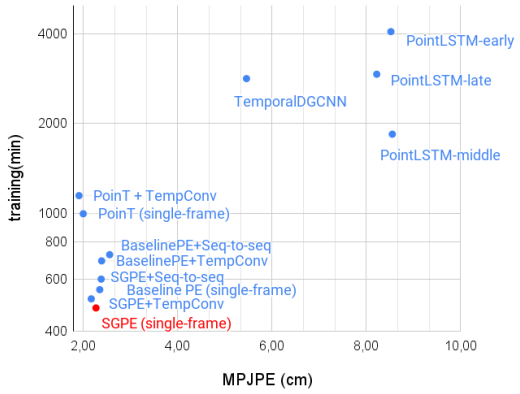

Figure 4: Comparison of mean per joint position error (MPJPE) and training time of the evaluated approaches. The method with the most favorable trade-off is the one located closest to the bottom-left corner.

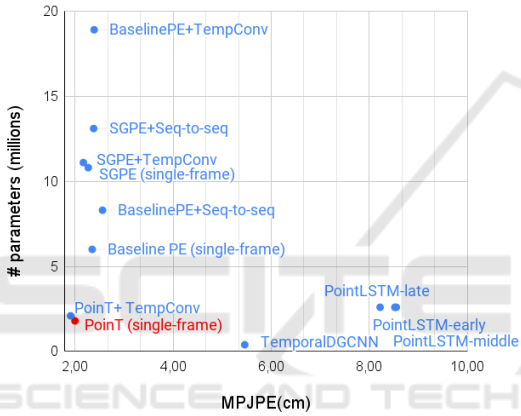

Figure 5: Comparison of mean per joint position error (MPJPE) and overall number of parameters of the evaluated models. The method with the most favorable trade-off is the one located closest to the bottom-left corner.

second frame from a sequence of 10 frames), and the point clouds initially down-sampled to 256 points using FPS. Moreover, we visualize the feature spaces produced at different stages of the network on a sample human body point cloud (Figure 3). We can observe, that in the particular case depicted in the figure, the point within the left hand is learnt to be gradually distinguished from the rest of the body, since the hand tends to move somewhat independently from the body core and the other limbs.

We can infer from the results that the Temporal DGCNN does not reach the accuracy of the single-frame pose estimation. Despite the small number of trainable parameters within the model, the training procedure is rather time consuming, mainly due to the graph re-computations after each layer. Also, since the original DGCNN was proposed to process generic objects, certain symmetry was usually present in the point cloud structure; whereas the complex structure of human poses is often asymmetrical and might pose a more complicated problem.

Next part of our experiments was focused on point cloud-processing LSTM model. We report pose estimation results of the PointLSTM-early, middle and late, following the original paper (Min et al., 2020), replacing different layers of the modified FlickerNet by the PointLSTM layer. After validation, we fixed the number of nearest neighbors for each point in the network to 16. To control the computational costs, we perform random sub-sampling of the point clouds to 256 points before feeding it to the model during training, and uniform sampling is applied when testing the model (same as in the original paper). We use the input sequence length of 8 frames, mostly due to limited computational resources. Also, we maintain the approach from the original paper, and assign the number of each frame within the input sequence as a fourth feature channel of each point of a point cloud. We trained all PointLSTM models for 50 epochs, while one epoch takes approximately 1 to 1.5 hour on a single Quadro RTX 4000. Lowest errors reached for the PointLSTM-early, middle and late are listed in Table 1. In spite of PointLSTM architectures keeping a relatively small amount of model parameters, the overall training time significantly exceeds that of the two-stage refinement approaches. Furthermore, for the accuracy to be competitive compared to the other tested methods, the PointLSTM model would likely need further adjustments to capture the complexity of human poses in motion, or the hardware resources for the experiments would need to be much larger.

We visualize the trade-off between mean per joint position error (MPJPE) and overall training time of the methods in Figure 4. Depending on the specific application environment, different approaches might be considered optimal. While the temporal convolution refinement yields the best test accuracy when applied to the transformer model, it also slightly increases the time requirements of the learning process. All in all, the models inferring pose from a single frame are deemed the most universal, as they exhibit the most favorable trade-off between accuracy and memory or time requirements. However, in specific scenarios where precision is considered the highest priority, temporal convolutions should be used to fine-tune the initial single-frame estimations. On the other hand, if the highest priority is given to memory requirements or computational complexity, the single frame transformer model, or even the temporal dynamic graph CNN represents a well-designed solution as it achieves sufficient accuracy with a small number of parameters in the model.

Table 1: Quantitative results of implemented methods. Mean per joint position error (MPJPE) and mean average precision at 10 cm threshold (mAP@10) are reported as evaluation metrics. *Symmetric* indicates the reference frame is in the middle of the input sequence. Whole training time of all models within the particular method is shown (in minutes). The total number of trainable parameters is in millions.

Method	symmetric	MPJPE (cm)	mAP@10 (%)	training (min)	# params
Baseline PE (single-frame)	-	2.35	97.80	553	6.0M
SGPE (single-frame)	-	2.27	98.40	**480**	10.8M
PoinT (single-frame)	-	2.00	98.65	995	1.8M
Baseline PE + TempConv	no	2.39	98.13	691	18.9M
Baseline PE + TempConv	yes	2.39	98.16	691	18.9M
SGPE + TempConv	no	2.17	98.47	563	23.7M
SGPE + TempConv	yes	2.17	98.48	515	11.1M
PoinT + TempConv	no	1.95	98.68	1140	2.1M
PoinT + TempConv	yes	**1.91**	**98.71**	1145	2.1M
Temporal DGCNN	no	5.64	90.56	2832	**0.4M**
Temporal DGCNN	yes	5.46	91.19	2832	**0.4M**
Baseline PE + Seq-to-seq	no	2.56	98.08	725	8.3M
SGPE + Seq-to-seq	no	2.38	98.53	600	13.1M
PointLSTM-early	no	8.52	76.27	4074	2.6M
PointLSTM-middle	yes	8.55	76.59	1842	2.6M
PointLSTM-late	no	8.49	76.73	2676	2.6M
PointLSTM-late	yes	8.22	77.91	2930	2.6M

5 CONCLUSIONS

This paper comprehensively explores the landscape of deep learning-based methods for human pose estimation, with a specific focus on the integration of temporal information. The survey covers the evolution from traditional methods to advanced techniques based on convolutional neural networks, recurrent architectures, and attention mechanisms. The incorporation of temporal context is investigated for its impact on robustness and adaptability to dynamic human movements. The experimental findings provide valuable insights into the performance of various models. The single-frame pose estimation models, including the baseline model, SGPE, and PoinT, demonstrate high accuracy with competitive evaluation metrics. The introduction of temporal convolutions for refinement further enhances the accuracy of these models, with the PoinT + TempConv approach achieving the lowest mean per joint position error. Even though the single-frame methods have the best trade-off between accuracy and computational requirements, it seems that in specific environments where the highest possible accuracy is needed, it may be more convenient to incorporate temporal information for fine-tuning the single-frame estimations. Furthermore, the paper explores spatio-temporal methods, such as sequence-to-sequence modeling using LSTMs and end-to-end approaches like the Temporal DGCNN. While the sequence-to-sequence LSTM refinement does not significantly affect MPJPE, it contributes to an increase in mean average precision. The Temporal DGCNN, despite its lower accuracy compared to single-frame models, presents a satisfactory trade-off between memory requirements and achieved accuracy, making it a viable option in scenarios where computational complexity is a priority or limited resources are provided. Our research contributes valuable insights into the strengths and weaknesses of different methods, offering guidance for the development of effective and reliable human pose estimation applications. Our survey also underscores the importance of temporal information and its role in enhancing the robustness of pose prediction models. As the field continues to evolve, addressing challenges and leveraging opportunities in this dynamic domain remains a key focus for future research.

REFERENCES

Artacho, B. and Savakis, A. (2020). Unipose: Unified human pose estimation in single images and videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Cao, Z., Hidalgo, G., Simon, T., Wei, S., and Sheikh, Y. (2021). Openpose: Realtime multi-person 2d pose estimation using part affinity fields. *IEEE Transactions on Pattern Analysis Machine Intelligence*, 43(01):172–186.

Chao, X., Ge, Z., and Leung, H. (2023). Video2mesh: 3d

human pose and shape recovery by a temporal convolutional transformer network. *IET Computer Vision*, 17(4):379–388.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., and Houlsby, N. (2021). An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*.

Einfalt, M., Ludwig, K., and Lienhart, R. (2023). Uplift and upsample: Efficient 3d human pose estimation with uplifting transformers. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 2903–2913.

Fang, H.-S., Xie, S., Tai, Y.-W., and Lu, C. (2017). Rmpe: Regional multi-person pose estimation. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 2353–2362.

Hossain, M. R. I. and Little, J. J. (2018). Exploiting temporal information for 3d human pose estimation. In Ferrari, V., Hebert, M., Sminchisescu, C., and Weiss, Y., editors, *Computer Vision – ECCV 2018*, pages 69–86, Cham. Springer International Publishing.

Joo, H., Simon, T., Li, X., Liu, H., Tan, L., Gui, L., Banerjee, S., Godisart, T. S., Nabbe, B., Matthews, I., Kanade, T., Nobuhara, S., and Sheikh, Y. (2017). Panoptic studio: A massively multiview system for social interaction capture. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Kang, Y., Liu, Y., Yao, A., Wang, S., and Wu, E. (2023). 3d human pose lifting with grid convolution. In *Proceedings of the Thirty-Seventh AAAI Conference on Artificial Intelligence and Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence and Thirteenth Symposium on Educational Advances in Artificial Intelligence*, AAAI'23/IAAI'23/EAAI'23. AAAI Press.

Lea, C., Flynn, M. D., Vidal, R., Reiter, A., and Hager, G. D. (2017). Temporal convolutional networks for action segmentation and detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Li, W., Du, R., and Chen, S. (2022). Skeleton-based spatio-temporal u-network for 3d human pose estimation in video. *Sensors*, 22(7).

Li, W., Liu, H., Ding, R., Liu, M., Wang, P., and Yang, W. (2023). Exploiting temporal contexts with strided transformer for 3d human pose estimation. *IEEE Transactions on Multimedia*, 25:1282–1293.

Liu, J., Guang, Y., and Rojas, J. (2020). Gast-net: Graph attention spatio-temporal convolutional networks for 3d human pose estimation in video. *CoRR*, abs/2003.14179.

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., and Guo, B. (2021). Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 10012–10022.

Mehta, D., Sotnychenko, O., Mueller, F., Xu, W., Elgharib, M., Fua, P., Seidel, H.-P., Rhodin, H., Pons-Moll, G., and Theobalt, C. (2020). XNect: Real-time multi-person 3D motion capture with a single RGB camera. volume 39.

Mehta, D., Sridhar, S., Sotnychenko, O., Rhodin, H., Shafiei, M., Seidel, H.-P., Xu, W., Casas, D., and Theobalt, C. (2017). Vnect: Real-time 3d human pose estimation with a single rgb camera. volume 36.

Min, Y., Chai, X., Zhao, L., and Chen, X. (2019). Flickernet: Adaptive 3d gesture recognition from sparse point clouds. In *BMVC*, page 105.

Min, Y., Zhang, Y., Chai, X., and Chen, X. (2020). An efficient pointlstm for point clouds based gesture recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Nie, Q., Liu, Z., and Liu, Y. (2023). Lifting 2d human pose to 3d with domain adapted 3d body concept. *Int. J. Comput. Vision*, 131(5):1250–1268.

Pavllo, D., Feichtenhofer, C., Grangier, D., and Auli, M. (2019). 3d human pose estimation in video with temporal convolutions and semi-supervised training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Qi, C. R., Su, H., Mo, K., and Guibas, L. J. (2017). Pointnet: Deep learning on point sets for 3d classification and segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Sun, K., Xiao, B., Liu, D., and Wang, J. (2019). Deep high-resolution representation learning for human pose estimation. In *Proceedings Of The IEEE Conference On Computer Vision And Pattern Recognition (CVPR)*.

Tang, Z., Qiu, Z., Hao, Y., Hong, R., and Yao, T. (2023). 3d human pose estimation with spatio-temporal criss-cross attention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4790–4799.

Wang, Y., Sun, Y., Liu, Z., Sarma, S. E., Bronstein, M. M., and Solomon, J. M. (2019). Dynamic graph cnn for learning on point clouds. *ACM Trans. Graph.*, 38(5).

Wu, M. and Shi, P. (2023). Human pose estimation based on a spatial temporal graph convolutional network. *Applied Sciences*, 13(5).

Yang, Y., Ren, Z., Li, H., Zhou, C., Wang, X., and Hua, G. (2021). Learning dynamics via graph neural networks for human pose estimation and tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8074–8084.

Zhao, Q., Zheng, C., Liu, M., Wang, P., and Chen, C. (2023). Poseformerv2: Exploring frequency domain for efficient and robust 3d human pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8877–8886.

Zheng, C., Zhu, S., Mendieta, M., Yang, T., Chen, C., and Ding, Z. (2021). 3d human pose estimation with spatial and temporal transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 11656–11665.

Škorvánková, D. and Madaras, M. (2021). Human pose estimation using per-point body region assignment. *COMPUTING AND INFORMATICS*, 40(2):387–407.