# Transformer-Based Framework for 3D Human Pose Estimation using YOLO Backbone

Miguel F. Lima[1,2][a], Ana Filipa Rodrigues Nogueira[1,2][b], Cláudia D. Rocha[1][c], Luís F. Teixeira[1,2][d] and Hélder Oliveira[1,3][e]

[1]*Instituto de Engenharia de Sistemas e Computadores, Tecnologia e Ciência (INESC TEC), Rua Dr. Roberto Frias, 4200-465, Porto, Portugal*

[2]*Faculdade de Engenharia da Universidade do Porto (FEUP), Rua Dr. Roberto Frias, 4200-465, Porto, Portugal*

[3]*Faculdade de Ciências da Universidade do Porto (FCUP), Rua do Campo Alegre, 1021-1055, Porto, Portugal*

{*miguel.f.brandao, ana.f.rodrigues, claudia.d.rocha, helder.f.oliveira*}*@inesctec.pt, luisft@fe.up.pt*

Keywords: 3D Human Pose Estimation, Transformer Network, YOLOv11 Backbone, Computer Vision, Deep Learning.

Abstract: Estimating 3D human poses from monocular videos is a crucial task for applications in healthcare, augmented reality, and robotics, yet it is challenged by occlusions and depth ambiguity. We introduce a new framework that utilises the YOLOv11 as a backbone for robust 2D keypoint detection and the TCPFormer, an innovative transformer-based architecture that leverages spatial and temporal transformers, to lift 2D poses to 3D. By integrating multi-scale attention, TCPFormer effectively captures local joint relationships and global sequence context, surpassing the accuracy of previous models. Evaluated on the MPI-INF-3DHP dataset, our approach presents an end-to-end pipeline for 3D pose estimation from image sequences, achieving superior performance compared to existing methods, with a mean per joint position error of 70.65 mm, an area under the curve of 59.05%, and 90.40% of keypoints correctly estimated within a 150 mm threshold.

## 1 INTRODUCTION

Estimating 3D human poses from monocular videos is a key task of computer vision, enabling applications in augmented reality, robotics, motion analysis (Udayan J et al., 2025) and healthcare (Zheng and Fang, 2024; Boudreault-Morales et al., 2025). By reconstructing the three-dimensional configuration of human joints from two-dimensional image sequences, this task enables advanced functionalities such as virtual character animation, action estimation, and human motion analysis. However, the problem is inherently challenging due to depth ambiguity, occlusions, and the complexity of the human motion across diverse scenarios (Guo et al., 2025).

Recent advances in deep learning have transformed 3D pose estimation, particularly for monocular videos, by shifting from traditional regression-based methods to more robust and versatile neural network architectures. Convolutional neural networks (CNNs) (Krizhevsky et al., 2012) excel at extracting spatial features but struggle with maintaining temporal coherence across frames in video sequences. Transformer-based architectures (Vaswani et al., 2017), such as PoseFormer (Zheng et al., 2021), introduced temporal and spatial attention, improving accuracy but usually at a higher computational cost. Meanwhile, efficient pose detection frameworks like YOLO (Maji et al., 2022) have enhanced 2D keypoint detection, yet their integration into 3D lifting pipelines remains somewhat underexplored.

Despite advancements in deep learning models, existing methods face limitations. CNN-based models lack the global context needed for robust 3D pose estimation under occlusions and suffer from depth ambiguity. While transformer-based approaches for 3D estimation from 2D images are, in general, more computationally intensive, and may still suffer to some extent from depth ambiguity. To address these challenges, we propose an end-to-end framework that leverages a transformer-based architecture for 3D pose lifting, TCPFormer (Liu et al., 2025), and uses the efficient YOLOv11 (Khanam and Hussain, 2024; Ultralytics, 2025b) as a backbone for 2D keypoint de-

[a] https://orcid.org/0009-0009-2253-2685
[b] https://orcid.org/0000-0002-9413-3300
[c] https://orcid.org/0000-0001-7254-0346
[d] https://orcid.org/0000-0002-4050-7880
[e] https://orcid.org/0000-0002-6193-8540

tection. TCPFormer leverages multi-scale attention to capture both local joint relationships and global sequence context, enabling better capture of temporal correlations, thereby mitigating the depth ambiguity problem.

Our contributions are: *(1)* We propose an end-to-end framework that uses YOLOv11 for 2D keypoint detection from monocular video sequences, and incorporates TCPFormer's spatial and temporal attention mechanisms to enhance the accuracy of 2D-to-3D keypoint lifting. *(2)* We demonstrate that our TCPFormer-based framework, combined with YOLOv11, achieves state-of-the-art mean per joint position error (MPJPE) on benchmarks such as MPI-INF-3DHP (Mehta et al., 2017), surpassing existing methods.

The remainder of this paper is organised as follows: Section 2 reviews related work in 3D pose estimation and detection frameworks. Section 3 details the framework architecture consisting of TCPFormer and its integration with YOLOv11. Section 4 presents experimental results, including model comparisons. Section 5 concludes with insights and future work.

# 2 RELATED WORK

Recent advancements in 3D human pose estimation have increasingly relied on deep learning and transformers to address challenges such as depth ambiguity and occlusions in monocular video inputs. This section reviews prior work in three key areas: 3D human pose estimation methods, transformer-based architectures, and detection backbones like YOLO for 2D pose estimation. In this context, our approach integrates TCPFormer with a YOLOv11 backbone, aligning it with recent trends such as transformer-based models and recently proposed 2D pose estimation models in order to estimate 3D human pose.

## 2.1 Datasets in Context

The development of 3D human pose estimation methods relies on diverse datasets that capture a range of poses and environments. The MPI-INF-3DHP dataset consists of 8 subjects performing 2 sequences each, captured using multiple cameras, with annotated frames provided from 8 distinct camera views for training and evaluation (Mehta et al., 2017). It includes videos captured in indoor and outdoor settings, as well as green-screen environments, offering a challenging benchmark with varied conditions. Its test set comprises 6 videos. Another widely used dataset, Human3.6M, provides extensive indoor motion capture data but is less available due to restricted access and licensing constraints (Ionescu et al., 2014; Catalin Ionescu, 2011). The experiments conducted in this paper were performed on the MPI-INF-3DHP dataset, chosen for its open accessibility and diverse multi-view setup, which enable robust evaluation and comparison with existing methods (Nogueira et al., 2025).

## 2.2 3D Human Pose Estimation Methods

Recent advances in deep learning have transformed 3D pose estimation for monocular videos, shifting from traditional regression-based methods to end-to-end learning approaches that jointly optimize 2D keypoint detection and 3D pose estimation from raw images. Two-stage pipelines, which first detect 2D keypoints and then lift them to 3D, remain popular due to their modularity, allowing flexible integration of 2D detectors and 3D lifting models (Pavllo et al., 2019). For example, Martinez (Martinez et al., 2017) proposed a simple yet effective baseline that uses fully connected neural networks to predict 3D joint coordinates from pre-detected 2D keypoints, achieving robust single-frame performance. However, its frame-independent design lacks temporal coherence for video sequences, leading to inconsistent predictions. In contrast, VideoPose3D (Pavllo et al., 2019) employs temporal convolutions to model motion, improving temporal consistency but struggling with occlusions due to errors in 2D keypoint detection. Our work introduces a two-stage pipeline with TCPFormer for 3D lifting, effectively capturing both spatial and temporal features to address occlusions and enhance robustness.

## 2.3 Transformer-Based Architectures

Transformers have significantly advanced 3D human pose estimation by capturing temporal dependencies across frame sequences and spatial relationships among keypoints within each frame. PoseFormer (Zheng et al., 2021) introduced a spatio-temporal transformer to process 2D keypoint sequences, achieving state-of-the-art accuracy on datasets like Human3.6M. MotionAGFormer (Mehraban et al., 2023) extended this by combining transformers with graph convolutional networks in a hybrid architecture, using parallel transformer and graph-based modules to model both global and local joint relationships, achieving superior accuracy and efficiency on benchmarks like Human3.6M and MPI-INF-3DHP. Similarly, TCPFormer, which we adopt, employs multi-

scale attention to model local and global contexts across video frames, offering better accuracy than previous models. However, TCPFormer estimates the 3D pose using as input the ground-truth 2D pose estimates. Therefore, in order to make it an end-to-end approach, our method explores TCPFormer by pairing it with the Ultralytics YOLOv11 (Jocher et al., 2024) backbone, designed for efficient and accurate 2D pose detection.

### 2.4 Detection Backbones for 2D Pose Estimation

Efficient 2D keypoint detection is critical for two-stage 3D pose estimation. Traditional backbones like ResNet (He et al., 2016) provide robust feature extraction but are computationally heavy and may not generalize well for pose estimation. The YOLO framework, originally designed for object detection, has been adapted for pose estimation. YOLO-pose (Maji et al., 2022) introduced a heatmap-free, anchor-based approach for multi-person 2D pose estimation, achieving high accuracy on COCO's dataset benchmarks. Recent versions, such as YOLOv8, incorporate enhancements like efficient multi-resolution feature (EMRF) modules, improving speed and precision. YOLOv11, used in our work, further refines these capabilities of the architecture for accurate and real-time estimations. Unlike prior methods that use YOLO for 2D pose detection alone, we integrate YOLOv11 directly into the pose estimation pipeline, feeding its keypoints to TCPFormer for 3D lifting.

Our work distinguishes itself by combining two recently proposed models, pairing TCPFormer's advanced transformer-based 3D lifting capabilities with YOLOv11's state-of-the-art 2D keypoint detection, addressing accuracy limitations of prior approaches while maintaining the possible suitability for real-time applications.

## 3 PROPOSED METHOD

In this section, we describe the method explored for 3D human pose estimation for single view, which integrates a fine-tuned YOLOv11 backbone for 2D keypoint detection with the TCPFormer architecture for lifting to 3D poses. This two-stage pipeline takes advantage of the efficiency of YOLOv11 for real-time 2D detection and the temporal modelling capabilities of TCPFormer to capture complex motion dynamics in video sequences.

### 3.1 Overview

The proposed pipeline processes monocular video sequences to estimate 3D human poses. First, we load the default weights of YOLOv11x-pose model (Ultralytics, 2025a) pre-trained on the COCO dataset and finetune it to the MPI-INF-3DHP dataset to detect 2D keypoints from input frames. Then we use the TCPFormer trained to detect 3D keypoints from 2D keypoints on the MPI-INF-3DHP. TCPFormer lifts 2D keypoints to 3D by modelling temporal correlations across 27 frames (T = 27). This integration combines the strengths of efficient and accurate 2D keypoint estimation with advanced transformer-based temporal processing, potentially reducing error accumulation in traditional two-stage methods. Figure 1 illustrates the proposed pipeline.

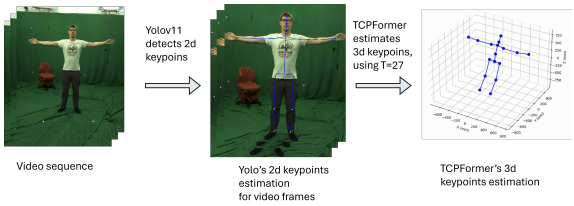


Figure 1: Overview of the proposed pipeline: YOLOv11 for 2D keypoint detection followed by TCPFormer for 3D lifting.

### 3.2 YOLOv11 Backbone

The YOLOv11-pose model detects 17 keypoints per person instance, in the same keypoints format defined in the COCO dataset.

The YOLOv11 architecture inherits an efficient feature extraction backbone, optimised for real-time performance, with specialised output heads for keypoint prediction. YOLOv11-pose outputs keypoint coordinates in the form of [x, y, visibility] tensors, along with confidence scores for each keypoint. The model processes input images resized to 640x640 pixels by default and supports person detection in a single pass, making it suitable for video sequences.

To align with the distinct keypoint annotation structure of our dataset, we fine-tuned the model to estimate keypoints in accordance with the MPI-INF-3DHP dataset's structure. The keypoints and their correspondence between YOLOv11's default predictions and the MPI-INF-3DHP dataset format are presented in Table 1.

Figure 2 shows a representation with the keypoints and indexes, following the MPI-INF-3DHP dataset format.

Table 1: Comparison of keypoints and their indices between YOLOv11 (COCO) and MPI-INF-3DHP dataset.

| Index | YOLOv11 (COCO) | MPI-INF-3DHP |
|---|---|---|
| 0 | Nose | Head |
| 1 | Left Eye | Spine Shoulder |
| 2 | Right Eye | Right Shoulder |
| 3 | Left Ear | Right Elbow |
| 4 | Right Ear | Right Hand |
| 5 | Left Shoulder | Left Shoulder |
| 6 | Right Shoulder | Left Elbow |
| 7 | Left Elbow | Left Hand |
| 8 | Right Elbow | Right Hip |
| 9 | Left Wrist | Right Knee |
| 10 | Right Wrist | Right Ankle |
| 11 | Left Hip | Left Hip |
| 12 | Right Hip | Left Knee |
| 13 | Left Knee | Left Ankle |
| 14 | Right Knee | Root/Pelvis |
| 15 | Left Ankle | Spine (Center torso) |
| 16 | Right Ankle | Neck |

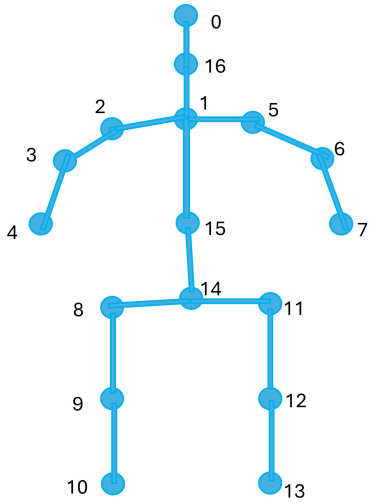


Figure 2: 2D keypoints and indexes.

#### 3.2.1 YOLOv11 Adaptation for Training

To adapt YOLOv11-pose for our task, we start by loading pretrained weights from the YOLOv11x-pose.pt model, which is pretrained on the COCO keypoints dataset. We then fine-tune the model on the MPI-INF-3DHP dataset, which provides annotated 2D and 3D poses in diverse scenarios.

In the hyperparameter configuration file, we specified 100 epochs, a batch size of 8, an image size of 640 pixels, and an automatic learning rate scheduler provided by Ultralytics. Additionally, we set a patience of 20 epochs for early stopping. This process optimises the model for accurate 2D keypoint detection under diverse conditions, such as different subjects, sequences, camera views and occlusions, all scenarios present in the MPI-INF-3DHP dataset.

The key innovation in YOLOv11 compared to prior versions lies in its refined architecture for higher Mean Average Precision (mAP) in pose tasks while maintaining low latency, achieved through optimisations in feature extraction and output regression layers (Khanam and Hussain, 2024).

### 3.3 TCPFormer for 3D Pose Lifting

For lifting the detected 2D keypoints to 3D poses, we explore TCPFormer (Liu et al., 2025), a transformer-based architecture that models temporal correlations using an implicit pose proxy. TCPFormer processes sequences of 2D keypoints extracted from video frames and outputs corresponding 3D joint positions.

The main difference between TCPFormer and previous methods is that TCPFormer introduces an implicit pose proxy to enhance temporal correlation modelling in 3D human pose estimation. In prior approaches, one pose in a sequence of length (T) establishes temporal relationships through a single 1-to-(T) mapping, limiting the depth of temporal interactions captured. In contrast, TCPFormer leverages a proxy of length (L), where each proxy element facilitates an independent 1-to-(T) mapping. This allows the model to aggregate multiple temporal perspectives, resulting in a more robust temporal correlation learning. Figure 3 shows an illustration of the model architecture.

The core components of TCPFormer include:

#### 3.3.1 Proxy Update Module (PUM)

The Proxy Update Module (PUM) refines the implicit pose proxy, initialised via a Gaussian distribution. It integrates spatio-temporal features from the encoder to update the proxy, aligning it with the input 2D pose sequence's dynamics. This ensures robustness for challenging MPI-INF-3DHP scenarios like occlusions.

#### 3.3.2 Proxy Invocation Module (PIM)

The Proxy Invocation Module (PIM) enhances the pose sequence's features using the updated implicit pose proxy. Through cross-attention, it emphasises key joints and frames, improving temporal and spatial coherence for accurate 3D pose estimation on MPI-INF-3DHP.

#### 3.3.3 Proxy Attention Module (PAM)

The Proxy Attention Module (PAM) models temporal correlations using an aggregation attention matrix,

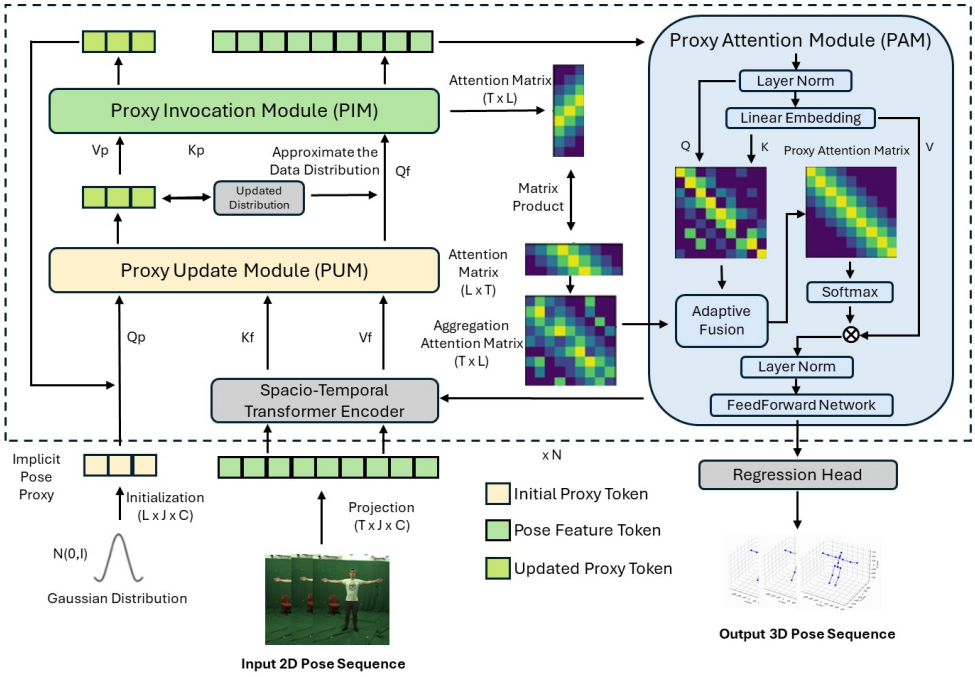


Figure 3: TCPFormer architecture. A spatio-temporal encoder extracts features from the 2D pose sequence. Then, an implicit pose proxy is introduced, initialised using a Gaussian distribution. These features, along with the proxy, are passed to the proxy update module to refine the implicit pose proxy. Subsequently, the proxy invocation module uses the updated proxy to enhance the pose sequence features. An aggregation attention matrix is derived from the two cross-attention matrices and, together with the pose sequence features, is fed into the proxy attention module to capture comprehensive temporal correlations. This process is repeated N times, followed by a regression head to produce the final 3D pose sequence. Image based on TCPFormer's paper (Liu et al., 2025)

derived from cross-attention between features and the proxy. It captures both local and global dependencies.

TCPFormer processes the sequences, lifting 2D keypoints to 3D by leveraging these modules to build comprehensive temporal correlations.

When trained on the MPI-INF-3DHP dataset using ground-truth (GT) 2D pose inputs for 2D-to-3D lifting, TCPFormer achieves a Mean Per Joint Position Error (MPJPE) of 17.8 mm, as reported in the original study. Under identical training conditions, our run of TCPFormer yields a comparable MPJPE of 17.3 mm on the GT in the same dataset as shown in Table 2. This close alignment in performance validates the robustness and reproducibility of TCPFormer's approach for 3D human pose estimation.

Table 2: Comparison of MPJPE performance on the MPI-INF-3DHP dataset between the reported TCPFormer's result and our run using the 2D ground-truth poses as input.

| In Paper (mm) | Our Run (mm) | Difference (%) |
|---|---|---|
| 17.8 | 17.3 | 2.8 |

In our methodology, we train TCPFormer to predict 3D keypoints from 2D keypoints estimated by YOLOv11 model for pose estimation, fine-tuned on the MPI-INF-3DHP dataset. YOLOv11 processes input images to generate 2D keypoint coordinates, which TCPFormer then uses as input to estimate 3D joint positions. This two-stage pipeline evaluates TCPFormer's ability to model temporal dynamics, leveraging its transformer-based architecture to capture long-range dependencies across video frames, even when input keypoints contain noise or occlusions typical of real-world scenarios.

## 3.4 Training and Implementation Details

The pipeline is implemented in two stages. For YOLOv11x-pose fine-tuning, we used the Ultralytics framework with 100 epochs, a batch size of 8 and an automatic learning rate scheduler from Ultralytics on MPI-INF-3DHP images.

For the TCPFormer training, we used 2D keypoints estimated by the YOLO-Pose model as input, with sequences of 27 frames from the MPI-INF-3DHP dataset.

The model is optimised with two objectives: a 3D

pose regression loss ($\mathcal{L}_{3D}$) (Equation 1) and a temporal consistency loss ($\mathcal{L}_T$) (Equation 2), like proposed in the TCPFormer paper (Liu et al., 2025):

$$\mathcal{L}_{3D} = \frac{1}{JT} \sum_{j=1}^{J} \sum_{t=1}^{T} \|\hat{Y}_{j,t} - Y_{j,t}\|_2 \quad (1)$$

where $J$ is the number of joints, $T$ is the sequence length (T=27 in this experiment), $\hat{Y}_{j,t}$ is the predicted 3D pose, and $Y_{j,t}$ is the ground-truth 3D pose for the $j$-th joint in the $t$-th frame. The temporal consistency loss, designed to ensure smooth motion, is given by:

$$\mathcal{L}_T = \frac{1}{J(T-1)} \sum_{j=1}^{J} \sum_{t=2}^{T} \|\Delta\hat{Y}_{j,t} - \Delta Y_{j,t}\|_2 \quad (2)$$

where $\Delta\hat{Y}_{j,t} = \hat{Y}_{j,t} - \hat{Y}_{j,t-1}$ and $\Delta Y_{j,t} = Y_{j,t} - Y_{j,t-1}$ represent the differences between consecutive frames for predicted and ground-truth poses, respectively. The final loss is the combination of the 3D pose regression loss (Equation 1) and the temporal consistency loss (Equation 2).

Training utilises a batch size of 16, a learning rate of 0.0005, a learning rate decay of 0.99, and 90 epochs on a NVIDIA V100-SXM2-32GB GPU. This setup evaluates TCPFormer's ability to model temporal dynamics from YOLO-Pose estimated keypoints across the diverse MPI-INF-3DHP dataset.

# 4 EXPERIEMENTAL RESULTS

In this section, we present the experimental results of our TCPFormer-based approach for end-to-end 3D human pose estimation on the MPI-INF-3DHP dataset. Our model was trained using 2D keypoints estimated from the YOLO-Pose model on sequences of 27 frames. The evaluation is done on the test set, which includes diverse indoor, green-screen, and outdoor scenes in six video sequences. The primary metric is the MPJPE, which measures the average Euclidean distance between predicted and ground-truth 3D joint positions after root alignment. Our method achieves an MPJPE of 70.65 mm, demonstrating reasonable accuracy in challenging monocular settings. To provide a more comprehensive assessment, we also report Percentage of Correct Keypoints (PCK) metrics, normalised relative to the torso diameter (PCK@X%_torso) and absolute threshold of 150 mm (PCK@X%_150mm), as well as the Area Under the Curve (AUC) for PCK across thresholds from 0 to 150 mm. The results are shown in Table 3. The comparison with other models is shown in Table 4.

Table 3: Performance metrics of TCPFormer on MPI-INF-3DHP dataset with 2D keypoints from YOLO.

| Metric | Value |
|---|---|
| MPJPE (mm) | 70.65 |
| PCK@10%_torso (%) | 24.09 |
| PCK@20%_torso (%) | 50.07 |
| PCK@30%_torso (%) | 68.70 |
| PCK@100%_torso (%) | 96.94 |
| PCK@10%_150mm (%) | 13.23 |
| PCK@20%_150mm (%) | 29.03 |
| PCK@30%_150mm (%) | 44.89 |
| PCK@100%_150mm (%) | 90.40 |
| AUC (%) | 59.05 |

These results indicate strong performance at higher thresholds (e.g., 96.94% at PCK@100%_torso), reflecting the model's ability to accurately localize most joints in less restrictive conditions. However, lower PCK values at stricter thresholds highlight challenges with fine-grained accuracy, particularly in outdoor scenes with occlusions or viewpoint variations. The AUC of 59.05% further underscores balanced performance across error thresholds, showing its suitability for real-world applications with partial occlusions where some inaccuracies are acceptable. Figures 4, 5, and 6 show how the model performs across different sequences of the test set, demonstrating reliable predictions.

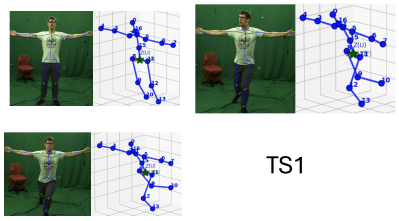


Figure 4: Detected 2D and 3D keypoints using our framework on 3 example frames from testset 1.

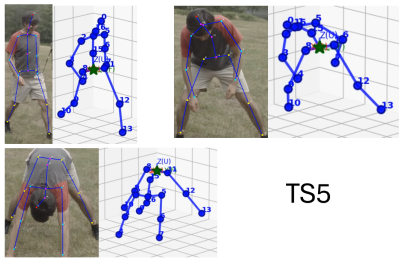


Figure 5: Detected 2D and 3D keypoints using our framework on 3 example frames from testset 5.

This evaluation reveals that the model excels in capturing temporal dynamics during the activities, thanks to the implicit pose proxy mechanism, but still

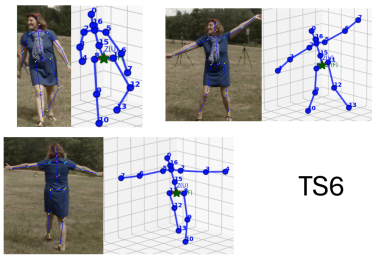


Figure 6: Detected 2D and 3D keypoints using our framework on 3 example frames from testset 6.

struggles with some motions in MPI-INF-3DHP's diverse environments. Compared to other 3D estimation methods (Table 4), our approach outperforms recent models in metrics like MPJPE, PCK and AUC. Our method, using a temporal window of 27 frames (T=27), achieves an MPJPE of 70.7 mm, a PCK of 90.4%, and an AUC of 59.1%. In comparison with other recently proposed methods evaluated on this dataset, our model consistently outperforms them across the MPJPE, PCK, and AUC metrics. For instance, relative to Anatomy-Aware 3D (Chen et al., 2022), which employs a longer temporal window of 81 frames, our approach achieves a superior MPJPE of 70.7mm compared to 78.8mm and a PCK of 90.4% compared to 87.9%, while utilizing substantially fewer frames for temporal context (27 vs. 81). Similarly, our model surpasses RepNet (Wandt and Rosenhahn, 2019) in AUC, attaining 59.1% against 58.5%. These results, summarized in Table 4, highlight the efficiency and effectiveness of the implicit pose proxy mechanism in leveraging shorter temporal sequences to achieve better performance.

# 5 CONCLUSIONS AND FUTURE WORK

In this paper, we presented a novel transformer-based framework for 3D human pose estimation, integrating the TCPFormer architecture with the YOLOv11 backbone for robust 2D keypoint detection. Evaluated on the MPI-INF-3DHP dataset, our approach achieved a MPJPE of 70.65 mm, demonstrating competitive performance in challenging monocular settings with diverse indoor, green-screen, and outdoor scenes. The combination of TCPFormer's multi-scale attention mechanism, which effectively captures local joint relationships and global temporal context, with YOLOv11's efficient and accurate 2D keypoint estimation, addressed key limitations of prior methods, such as depth ambiguity and occlusion handling. Our results, including low MPJPE, high PCK values (up to 96.94% at PCK@100%_torso) and an AUC of 59.05%, indicate robust performance across various thresholds, outperforming recent methods on image inputs while using significantly fewer temporal frames.

This work contributes to the field by showcasing the potential of a tightly integrated two-stage pipeline, leveraging state-of-the-art detection and lifting techniques to enhance accuracy while possibly maintaining suitability for real-time applications. The implicit pose proxy in TCPFormer proved particularly effective in modeling temporal dynamics, as evidenced in the obtained results, despite challenges with rapid motions or occlusions.

Looking ahead, future work will be evaluating the feasibility of integrating this framework into real-time applications, such as healthcare, augmented reality, and robotics.

# ACKNOWLEDGEMENTS

This work is co-financed by Component 5- Capitalization and Business Innovation of core funding for Technology and Innovation Centres (CTI), integrated in the Resilience Dimension of the Recovery and Resilience Plan within the scope of the Recovery and Resilience Mechanism (MRR) of the European Union (EU), framed in the Next Generation EU, for the period 2021- 2026, with reference 21.

# REFERENCES

Boudreault-Morales, G.-E. et al. (2025). The effect of depth data and upper limb impairment on lightweight monocular rgb human pose estimation models. *BioMedical Engineering OnLine*, 24.

Catalin Ionescu, Fuxin Li, C. S. (2011). Latent structured models for human pose estimation. In *International Conference on Computer Vision*.

Chen, T., Fang, C., Shen, X., Zhu, Y., Chen, Z., and Luo, J. (2022). 3d human pose estimation in videos via bone direction and length prediction. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(1):198–209. Date of Publication: 04 February 2021.

Guo, Y. et al. (2025). A survey of the state of the art in monocular 3d human pose estimation: Methods, benchmarks, and challenges. *Sensors*, 25(8):2409.

He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778.

Ionescu, C., Papava, D., Olaru, V., and Sminchisescu, C. (2014). Human3.6m: Large scale datasets and predictive methods for 3d human sensing in natural envi-

Table 4: Comparison of our method and other 3D pose estimation models on MPI-INF-3DHP. Best results are in **bold**, and second-best results are underlined.

| Model | MPJPE (mm) ↓ | PCK (%) ↑ | AUC (%) ↑ |
|---|---|---|---|
| PnPproj (Zhang et al., 2025) | 119.4 | 73.1 | - |
| RepNet (Wandt and Rosenhahn, 2019) | 97.8 | 82.5 | 58.5 |
| Anatomy-Aware 3D (81 frames) (Chen et al., 2022) | 78.8 | 87.9 | 54.0 |
| Anatomy-Aware 3D (243 frames) (Chen et al., 2022) | 79.1 | 87.8 | 53.8 |
| **Ours (27 frames)** | **70.7** | **90.4** | **59.1** |


ronments. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Jocher, G., Chaurasia, A., and Qiu, J. (2024). Ultralytics yolov11: Real-time object detection, segmentation, pose estimation, and tracking. https://github.com/ultralytics/ultralytics Version 8.3.0.

Khanam, R. and Hussain, M. (2024). Yolov11: An overview of the key architectural enhancements. *arXiv preprint arXiv:2410.17725*. Submitted on 23 Oct 2024.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 25.

Liu, J., Liu, M., Liu, H., and Li, W. (2025). Tcpformer: Learning temporal correlation with implicit pose proxy for 3d human pose estimation. In *Proceedings of the 39th AAAI Conference on Artificial Intelligence*.

Maji, D., Nagori, S., Mathew, M., and Poddar, D. (2022). Yolo-pose: Enhancing yolo for multi person pose estimation using object keypoint similarity loss. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 2637–2646.

Martinez, J., Hossain, R., Romero, J., and Little, J. J. (2017). A simple yet effective baseline for 3d human pose estimation. In *IEEE International Conference on Computer Vision (ICCV)*, pages 2640–2649.

Mehraban, S., Adeli, V., and Taati, B. (2023). Motionagformer: Enhancing 3d human pose estimation with a transformer-gcnformer network. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 2926–2935.

Mehta, D., Rhodin, H., Casas, D., Fua, P., Sotnychenko, O., Xu, W., and Theobalt, C. (2017). Monocular 3d human pose estimation in the wild using improved cnn supervision. In *3D Vision (3DV), 2017 Fifth International Conference on*. IEEE.

Nogueira, A. F. R., Oliveira, H. P., and Teixeira, L. F. (2025). Markerless multi-view 3d human pose estimation: A survey. *Image and Vision Computing*, 155:105437.

Pavllo, D., Feichtenhofer, C., Grangier, D., and Auli, M. (2019). 3d human pose estimation in video with temporal convolutions and semi-supervised training. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7753–7762.

Udayan J, D. et al. (2025). Deep learning in monocular 3d human pose estimation: Systematic review of contemporary techniques and applications. *Multimedia Tools and Applications*.

Ultralytics (2025a). Pose estimation - ultralytics yolo documentation. https://docs.ultralytics.com/pt/tasks/pose/. Accessed: July 17, 2025.

Ultralytics (2025b). Yolov11: Real-time object detection, instance segmentation, and pose estimation. https://docs.ultralytics.com/models/yolov11/. Accessed: July 17, 2025.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems (NeurIPS)*, 30.

Wandt, B. and Rosenhahn, B. (2019). Repnet: Weakly supervised training of an adversarial reprojection network for 3d human pose estimation. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7782–7791.

Zhang, X., Chen, Y., Lai, H., and Zhang, H. (2025). Weakly supervised 3d human pose estimation based on pnp projection model. *Pattern Recognition*, 163:111464.

Zheng, C., Zhu, S., Mendieta, M., Yang, T., Chen, C., and Ding, Z. (2021). 3d human pose estimation with spatial and temporal transformers. *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 11606–11615.

Zheng, S. and Fang, Q. (2024). A real-time 3d motion detection system based on a monocular camera. *2024 IEEE-EMBS Conference on Biomedical Engineering and Sciences (IECBES)*, pages 244–248.