

A review paper *

A REVIEW OF DEEP LEARNING AND SYNTHETIC DATA GENERATION FOR HUMAN POSE ESTIMATION IN AUTONOMOUS ROBOTIC SYSTEMS: METHODS, CHALLENGES, AND FUTURE DIRECTIONS

Mladen Dinčić, Ivan Ćirić, Emina Petrović, Miloš Simonović, Mladen
Kuzev, Nikola Ivačko

Department of Mechatronics and Control, Faculty of Mechanical Engineering, University
of Niš, Serbia

Abstract. *Reliable perception of human posture and movement is essential for autonomous robotic systems operating in shared or dynamic environments. Human pose estimation provides this capability by detecting body keypoints and reconstructing human motion from visual data, but its practical use in robotics is still affected by challenges such as occlusion, viewpoint changes, limited real-time performance, and insufficient annotated training data. This paper analyzes recent developments in 2D and 3D human pose estimation from the perspective of autonomous robotic perception. The review covers fundamental body representation models, single-person and multi-person estimation strategies, and the main differences between skeleton-based and model-based 3D approaches. Special attention is given to deep learning architectures, including convolutional neural networks, vision transformers, and emerging state-space models, with emphasis on their accuracy, computational complexity, and suitability for real-time robotic applications. In addition, the paper discusses synthetic data generation as a practical solution for improving dataset diversity and annotation quality. Rendering-based simulation, motion-capture data, generative methods, and fusion-based techniques are discussed as tools for producing controllable training samples. A small proof-of-concept example based on a game-based 3D environment is also included to demonstrate how camera and character positions can be extracted and used to automatically calculate camera-to-human distance as a spatial annotation. The main limitation of this approach is the synthetic-to-real domain gap, which can reduce model reliability when synthetic datasets are applied to real-world robotic scenarios. Future research should therefore focus on robot-oriented synthetic datasets, improved synthetic-to-real transfer, generative artificial intelligence, and multimodal sensing to support more robust human perception in autonomous robotic systems.*

Key words: Human Pose Estimation, Deep Learning, Autonomous Robotic Systems, Synthetic Data, Synthetic-to-Real Transfer

*Received: July 02, 2026 / Accepted: September 18, 2026

Corresponding author: Mladen Dinčić

Institution: Faculty of Mechanical Engineering, University of Niš

E-mail: mladen.dincic@masfak.ni.ac.rs

1. INTRODUCTION

Human pose estimation (HPE) represents an important perception task in autonomous robotic systems, since it enables robots and intelligent machines to interpret human body posture and motion from visual input in real time. In general, HPE refers to the detection and localization of anatomical keypoints from images or video data [1, 2]. In recent years, the development of deep learning methods has significantly improved the performance of pose estimation systems, leading to higher accuracy and greater robustness in complex visual conditions [3]. As a result, HPE has become relevant for a wide range of applications, including human-aware navigation of autonomous drones [4], pedestrian motion understanding in autonomous vehicles, and medical applications such as rehabilitation, movement evaluation, and physical performance monitoring [5].

Although substantial progress has been achieved, the lack of large, diverse, and accurately annotated datasets remains one of the main obstacles, especially in the case of 3D human pose estimation. Compared with 2D pose datasets, which have become increasingly larger and more varied, 3D datasets are still limited by the requirements of 3D data acquisition [1, 6–8]. Accurate 3D pose annotation usually depends on dedicated equipment, such as motion capture systems, depth cameras, or controlled laboratory setups. For this reason, many existing 3D pose benchmarks are recorded in indoor environments with restricted variation in subjects, backgrounds, camera positions, and movement patterns. These limitations reduce the generalization ability of pose estimation models when they are deployed in realistic and dynamic robotic environments.

Synthetic data generation offers a practical way to reduce these limitations by creating artificial training samples with automatically available annotations. Through controllable 3D scenes, synthetic datasets can be generated at large scale while systematically varying important factors such as illumination, camera viewpoint, human appearance, clothing, body shape, scene background, and motion type. Previous studies have demonstrated that synthetic data can improve generalization to unseen viewpoints by increasing variation in visual appearance, motion, and camera conditions [9]. Another important advantage of 3D synthetic environments is the possibility of obtaining precise ground-truth information without manual annotation. This includes 2D and 3D joint locations, body orientation, camera parameters, spatial relations between humans, robots, and surrounding objects, as well as angular information within the scene. These properties are particularly important for robotic perception, where pose estimation models must function reliably under changing viewpoints, partial occlusions, and safety-critical interaction scenarios. In this context, simulation and game-engine environments are useful not only for rendering visual samples, but also for extracting additional scene-level annotations, such as the relative position or distance between the camera and the human subject. Nevertheless, the effective use of synthetic data for HPE depends on motion realism, rendering quality, sample diversity, careful control of scene parameters, and the remaining synthetic-to-real domain gap [8]. Modern 3D rendering environments and generative AI-based methods provide promising tools for improving these aspects. Although synthetic data can be applied to many visual perception tasks, this review focuses on its role in 2D and 3D human pose estimation for autonomous robotic systems.

2. HUMAN POSE ESTIMATION PRINCIPLES AND REPRESENTATIONS

A fundamental aspect of every human pose estimation framework is the selection of an appropriate body representation model. Typically, the human body is represented by a set of anatomical keypoints associated with major joints, including the shoulders, elbows, wrists, hips, knees, and ankles. The geometric configuration and relative positions of these keypoints provide the basis for describing human posture. Existing studies generally classify body representations into three principal categories: skeleton-based models, surface-based models, and volumetric models [10].

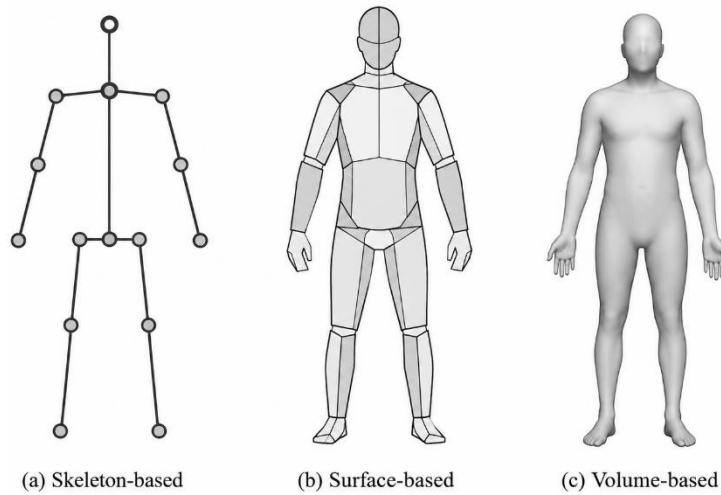


Fig. 1 Human body representation models

Among the available body representation approaches, the skeleton-based model shown in Fig. 1(a), often referred to as a kinematic or articulated body model, remains the predominant choice for both 2D and 3D human pose estimation tasks [10]. In this representation, the human body is modeled as a set of interconnected joints linked by body segments, providing a compact and computationally efficient representation of human posture and movement [11]. Several standardized keypoint configurations have been established in the literature, including the 16-joint annotation scheme introduced by MPII [12], the 17-keypoint format adopted by COCO [13], and the more detailed 25-keypoint framework employed by OpenPose [14].

Surface-based representations, shown in Fig. 1(b), on the other hand, characterize the human body using contours, silhouettes, or surface patches that approximate the shape of individual body parts. These models provide a denser and more detailed geometric description of the human body compared with skeletal representations [15]. However, their ability to accurately determine joint positions is limited. For this reason, they are used less frequently in contemporary deep learning-based human pose estimation frameworks, where precise localization of anatomical keypoints is typically the primary objective [10].

Volume-based representations, shown in Fig. 1(c), model the human body as a complete three-dimensional structure, allowing the simultaneous description of skeletal

configuration, body morphology, and surface variations. Unlike skeleton-based approaches that focus primarily on joint coordinates, these models provide a richer depiction of human anatomy by encoding the full spatial extent of the body. Early volumetric methods often approximate body parts using simple geometric elements, such as cylinders or ellipsoids, to construct a coarse body model. More recent approaches employ parametric mesh representations, most notably the Skinned Multi-Person Linear (SMPL) model, which generates diverse body shapes and poses through a compact set of learned parameters [10]. Owing to their ability to represent detailed body geometry and shape characteristics, parametric mesh models are particularly well suited for advanced 3D human pose and body shape estimation frameworks [16, 17].

2.1 2D Human Pose Estimation

Human pose estimation techniques are commonly categorized into single-person and multi-person frameworks, depending on the number of individuals present in the scene. Single-person pose estimation focuses on images containing a single subject, enabling the model to dedicate its full capacity to the precise identification of anatomical keypoints without the additional complexity of distinguishing between multiple individuals. Contemporary deep learning solutions typically rely on encoder-decoder architectures. In these architectures, the encoder is responsible for extracting hierarchical visual features from the input image, while the decoder transforms these features into predictions of body joint locations. Keypoint estimation can be performed either by directly regressing joint coordinates or by generating confidence maps that indicate the probability of a keypoint occurring at each image location [18].

Among the available strategies, heatmap-based approaches have emerged as the prevailing paradigm in 2D human pose estimation. Instead of predicting coordinates directly, these methods estimate spatial probability distributions for each body joint, resulting in improved localization accuracy and greater resilience to challenging conditions such as self-occlusion, background clutter, and highly articulated human poses [3]. Multi-person pose estimation presents substantially greater complexity than its single-person counterpart, as the number of subjects within an image is not predetermined and individuals frequently exhibit mutual occlusions or spatial overlap. To address these challenges, existing methods are generally classified into two main categories: top-down and bottom-up approaches [18].

Top-down frameworks operate by first identifying human instances through a person detection stage, after which a dedicated pose estimation model is applied independently to each detected individual. This strategy typically delivers high localization accuracy because the pose estimation task is performed on cropped regions containing a single person. However, its computational requirements increase proportionally with the number of detected subjects, which may limit efficiency in densely populated scenes. By contrast, bottom-up approaches begin by locating all body keypoints across the entire image without prior person detection. Then, the detected joints are associated and assembled into separate human skeletons. By processing the scene globally, these methods exhibit improved scalability and are generally more suitable for real-time applications and environments containing large numbers of individuals [18]. In autonomous robotic systems, this trade-off is particularly important because the pose estimation strategy must balance localization accuracy with computational efficiency and real-time operation. For example, top-down

approaches may be suitable for collaborative robots or human-robot interaction scenarios involving a limited number of people, while bottom-up methods can be advantageous for mobile robots, autonomous vehicles, or safety-monitoring systems operating in dynamic multi-person environments.

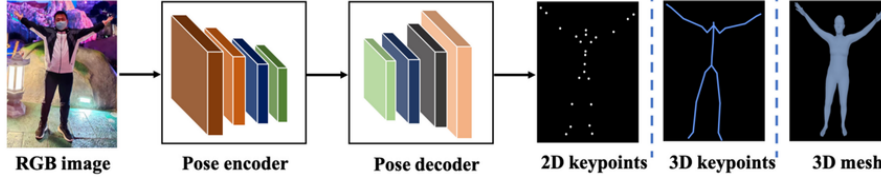


Fig. 2 Typical pipeline for single-person pose estimation [18]

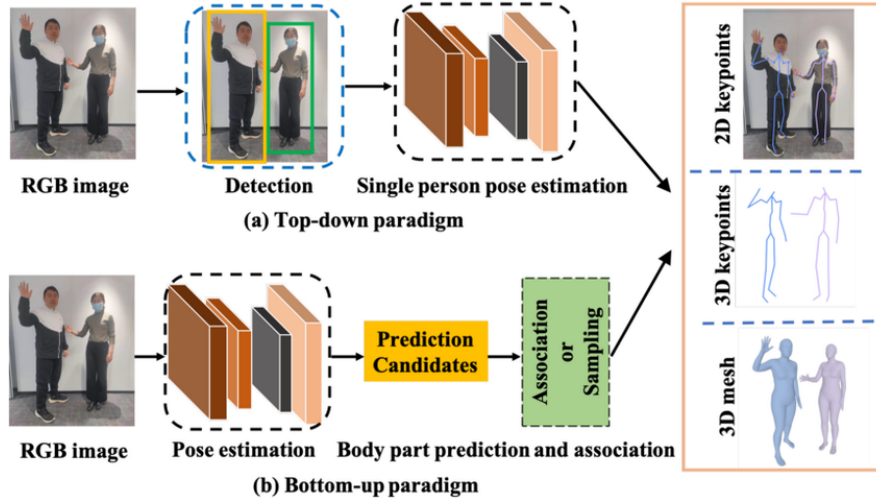


Fig. 3 Typical pipeline for multi-person pose estimation [18]

2.2 3D Human Pose Estimation

Unlike 2D pose estimation, which determines the locations of body joints within the image coordinate system, 3D human pose estimation estimates their positions in three-dimensional space. This makes the task more challenging because depth information is ambiguous in monocular images, and large-scale annotated 3D datasets are still limited. As a result, developing robust and generalizable 3D pose estimation models remains an active area of research.

Current 3D HPE techniques can be broadly categorized into skeleton-based and model-based approaches. Skeleton-based methods focus on predicting the three-dimensional coordinates of body joints directly. To achieve this, they commonly employ strategies such as volumetric heatmap prediction, 2D-to-3D pose lifting, or the fusion of image-derived features with detected 2D keypoints to enhance geometric consistency and spatial accuracy [19–21]. Model-based approaches, on the other hand, incorporate parametric human body representations, with the Skinned Multi-Person Linear (SMPL) model being among the

most widely adopted frameworks [22]. Rather than estimating only joint locations, these methods simultaneously recover body pose, shape, and surface structure [23]. By encoding the human body through a compact set of pose and shape parameters, model-based techniques facilitate detailed three-dimensional human reconstruction from monocular imagery. Consequently, they have become increasingly important in applications including virtual and augmented reality, motion analysis, human-computer interaction, and autonomous robotic systems [18].

For autonomous robotic systems, 3D human pose estimation provides richer spatial information about human posture and body configuration relative to the robot and its surroundings. Such information is valuable for applications including human-aware navigation, collision avoidance, collaborative robotics, and UAV-based human monitoring, where three-dimensional understanding of human position and movement is essential.

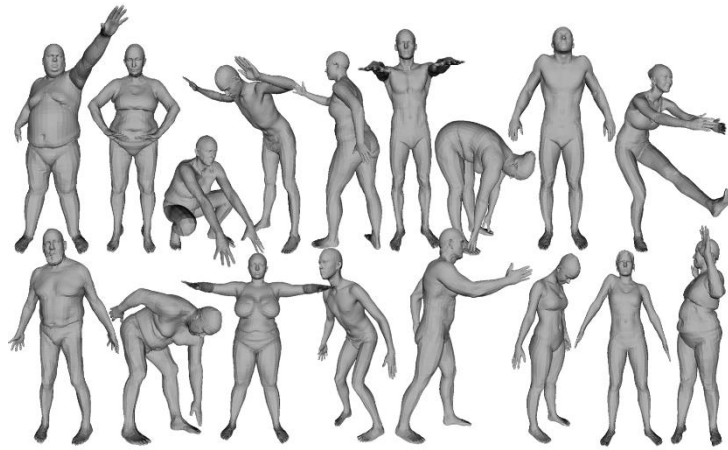


Fig. 4 Sample SMPL model registrations from a multi-pose dataset [22]

3. DEEP LEARNING APPROACHES FOR HUMAN POSE ESTIMATION

Deep learning represents a subset of machine learning techniques that utilize multi-layer artificial neural networks to automatically extract and learn hierarchical patterns from raw input data. By progressively transforming information through successive processing layers, these models can capture increasingly abstract feature representations. Among the various deep learning architectures, convolutional neural networks (CNNs) have become particularly influential in computer vision applications. CNNs employ trainable convolutional kernels that systematically learn visual patterns, ranging from low-level features such as edges and textures to more complex semantic structures in deeper network layers [24, 25].

The adoption of deep learning, especially CNN-based architectures, has revolutionized human pose estimation by enabling end-to-end learning frameworks capable of directly mapping visual inputs to body joint predictions. In recent years, transformer-based models have emerged as a powerful alternative to conventional CNNs. Originally introduced for

natural language processing tasks, transformers have been successfully adapted to computer vision through self-attention mechanisms that capture long-range relationships and global contextual information across an entire image.

The following sections provide a review of representative CNN-based and transformer-based approaches for human pose estimation, together with an overview of their reported performance on widely used benchmark datasets.

3.1 CNN-Based Methods

One of the earliest breakthroughs in deep learning-based human pose estimation was achieved by DeepPose [26], which demonstrated that body joint locations could be estimated directly from input images through a cascade of convolutional neural networks. By treating pose estimation as a coordinate regression problem, DeepPose enabled end-to-end learning of human body configurations. However, directly predicting joint coordinates proved challenging, often resulting in reduced localization accuracy due to the highly nonlinear relationship between image content and spatial coordinates. To overcome these limitations, subsequent research shifted toward heatmap-based formulations, where the network predicts a probability map for each anatomical keypoint instead of regressing its coordinates directly. A notable contribution in this direction was the Stacked Hourglass architecture [27], which employed a multi-stage encoder-decoder design with repeated bottom-up and top-down processing. This structure allowed the network to integrate fine-grained spatial information with broader contextual cues, thereby improving keypoint localization across multiple image scales.

For multi-person pose estimation, OpenPose [14] introduced an efficient bottom-up strategy based on Part Affinity Fields (PAFs). These vector fields encode spatial and directional relationships between connected body parts, enabling simultaneous detection of body joints and their association into individual human skeletons within a single network inference step. This approach significantly improved real-time performance in crowded scenes containing multiple subjects.

Another major advancement was the introduction of High-Resolution Network (HRNet) [28]. Unlike conventional architectures that progressively reduce spatial resolution and later recover it through upsampling, HRNet preserves high-resolution feature maps throughout the entire network while continuously exchanging information across parallel multi-resolution branches. This capability results in highly accurate keypoint localization and has established HRNet as one of the most influential backbone architectures for both 2D and 3D human pose estimation systems [25].

3.2 Transformer-Based Methods

The introduction of transformer-based architecture has substantially enhanced the ability of human pose estimation models to capture long-range dependencies and structural relationships among body joints. One of the pioneering transformer approaches, TokenPose [29], represented each body keypoint as an independent learnable token. Through self-attention mechanisms, these keypoint tokens interact with visual tokens derived from image feature maps, enabling the network to jointly learn appearance information and the underlying anatomical relationships between body parts. While TokenPose still employs a convolutional backbone, typically HRNet, for low-level feature extraction, it demonstrated that transformer-driven modeling of joint interactions can

achieve performance comparable to leading CNN-based approaches while requiring fewer parameters.

Further progress was achieved with ViTPose [30], which showed that a standard Vision Transformer can be used as a unified framework for both feature extraction and pose estimation, removing the dependency on convolutional backbones altogether. Built upon a Vision Transformer pre-trained using masked image modeling, ViTPose combines a simple transformer encoder with a lightweight decoder composed of deconvolution layers for keypoint prediction. Experimental evaluations on the COCO benchmark demonstrated state-of-the-art accuracy, while also highlighting strong scalability across architectures ranging from approximately 100 million to over 1 billion parameters. In addition, ViTPose exhibited robust performance in challenging scenarios involving severe occlusions, pose variability, and significant appearance changes, confirming the effectiveness of transformer-based representations for large-scale human pose estimation tasks.

3.3 Emerging Architectures and Performance Overview

Despite the remarkable progress achieved by transformer-based architectures in human pose estimation, their dependence on self-attention operations results in computational costs that grow quadratically with the number of input tokens. This characteristic presents a significant obstacle for processing extended temporal sequences and for applications requiring real-time performance. To address these limitations, state space models (SSMs) have recently gained attention as a computationally efficient alternative, enabling long-range dependency modeling with linear complexity.

Among these approaches, PoseMamba [31] represents a fully SSM-driven framework for monocular video-based 3D human pose estimation. The method integrates bidirectional spatio-temporal processing with both global and local feature interactions. Furthermore, it incorporates a skeleton-aware joint reordering mechanism that exploits the geometric structure of the human body to improve local relationship modeling while preserving global spatio-temporal coherence. As a result, PoseMamba achieves competitive performance with substantially reduced parameter counts and computational overhead relative to transformer-based solutions.

Such computational efficiency is particularly relevant in autonomous robotic platforms, where pose estimation algorithms are frequently deployed on resource-constrained hardware and are expected to operate under strict real-time requirements [32]. Consequently, the suitability of an HPE model should be evaluated not only through accuracy-oriented benchmarks but also through practical deployment criteria, including inference speed, memory consumption, scalability in multi-person environments, and resilience to viewpoint variations and partial occlusions. These factors directly affect the reliability of human perception in applications such as collaborative robotics, human-aware navigation, autonomous vehicles, and UAV-based monitoring, where delayed or inaccurate pose estimates may reduce the effectiveness of downstream decision-making and safety-related functions.

Table 1 summarizes representative performance results reported for several widely used deep learning approaches in human pose estimation. Since the methods are evaluated on different datasets, protocols, and metrics, the reported values are not directly comparable. They are included only to summarize commonly used benchmarks and evaluation criteria in HPE research [25].

Table 1 Overview of representative deep learning architectures for human pose estimation

Method	Year	Arch.	Dataset	Metric	Perf.
DeepPose [26]	2014	CNN	FLIC	PCP	0.69
Stacked Hourglass [27]	2016	CNN	MPII	PCKh@0.5	90.9
OpenPose [14]	2017	CNN	MPII	mAP	75.6
HRNet [28]	2021	CNN	COCO	AP	75.5
TokenPose-L [29]	2021	CNN + Transformer	COCO	AP	75.8
ViTPose-G [30]	2022	Transformer	COCO	AP	80.9
PoseMamba-L [31]	2025	SSM	Human3.6M	MPJPE (mm)	38.1

The reported metrics are dataset-dependent and should not be directly compared. AP, mAP, PCKh, and PCP are accuracy-based metrics, where higher values indicate better performance, while MPJPE measures 3D joint error and is therefore better when lower.

4. SYNTHETIC DATA GENERATION APPROACHES FOR HUMAN POSE ESTIMATION

Compared with conventional real-world data acquisition, simulation-driven data generation frameworks provide substantially greater control over the conditions under which training samples are produced. Parameters including camera placement, lighting conditions, environmental context, human appearance, body morphology, and movement patterns can be systematically manipulated while maintaining access to precise ground-truth annotations. This capability is particularly important in human pose estimation, where synthetic data generation pipelines can automatically generate 2D keypoint labels, 3D joint coordinates, segmentation masks, depth maps, and camera calibration information without the need for labor-intensive manual annotation. Current synthetic data generation strategies can generally be categorized into several complementary groups [33]. Rendering-based approaches rely on computer graphics and simulation engines to create data from explicitly modeled virtual environments in which scene geometry, illumination, materials, camera configurations, and motion parameters are fully controllable. In contrast, generative techniques, including those based on generative adversarial networks (GANs), learn statistical characteristics from real-world datasets and subsequently produce visually similar synthetic samples. Another category consists of fusion-based methods, which integrate synthetic humans or objects into real image backgrounds to enhance visual variability while retaining accurate annotations [34].

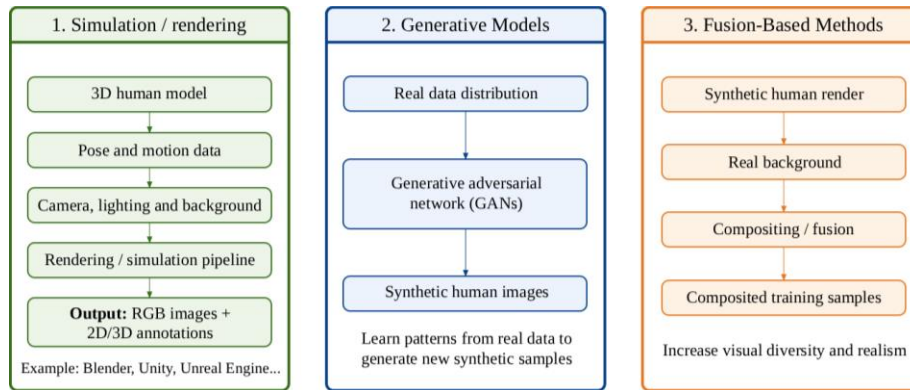


Fig. 5 Overview of synthetic data generation approaches for human pose estimation

Among these alternatives, simulation and rendering frameworks are especially well suited for HPE applications because they preserve complete access to the underlying three-dimensional scene representation and can automatically generate task-specific annotations. This characteristic is particularly valuable in robotic systems, where humans, robots, cameras, and surrounding objects can all be modeled within a common virtual 3D environment. Therefore, simulation-based pipelines allow the extraction of not only body joint locations but also depth measurements, camera calibration parameters, body orientation estimates, viewing perspectives, and detailed spatial relationships between individuals and environmental objects [34, 35]. Such information is often difficult or expensive to obtain with high precision in real-world datasets, but it can be directly used for the training, validation, and optimization of perception systems intended for autonomous robotic operation. In the context of 3D HPE, rendering pipelines commonly combine parametric human body representations with motion capture recordings to generate realistic movement sequences under controlled conditions. Other approaches extend this strategy through broader pose-space coverage and texture diversity, generating articulated 3D human models from sampled body poses, applying clothing textures extracted from real images, and varying viewpoints, illumination conditions, and backgrounds. Experimental results indicate that broad pose-space coverage and texture diversity are particularly important for improving the effectiveness of synthetic training data [36]. A representative example employs the SMPL body model [22] together with motion trajectories from the Carnegie Mellon University Motion Capture (CMU MoCap) dataset, rendering animated virtual humans in Unity from the perspective of an egocentric fisheye camera positioned on the chest. By introducing random variations in textures, illumination settings, and background content, the framework produces diverse RGB imagery accompanied by corresponding 2D and 3D pose labels as well as localization heatmaps [35]. This concept has also been demonstrated at substantially larger scale, with synthetic human datasets comprising more than six million rendered frames and providing additional pixel-level annotations such as depth maps, body-part segmentation masks, optical flow, and surface normals [37]. More recent work has also demonstrated the value of photorealistic synthetic data under highly challenging viewpoints. The xR-EgoPose dataset was designed for egocentric 3D human pose estimation and incorporates extensive variation in body shape, skin tone, clothing, background, lighting, and motion. Its fisheye

camera, mounted on a head-mounted display (HMD), introduces severe self-occlusion and strong perspective distortion, providing a particularly challenging setting for evaluating and training 3D pose estimation methods [38]. Together, these examples illustrate the ability of rendering-based systems to create large-scale pose estimation datasets with highly accurate annotations while substantially reducing the need for extensive real-world image collection and manual labeling procedures.

Modern game engines provide another important class of synthetic data generation environments by combining interactive scene control with real-time rendering and direct access to internal scene information. UnrealCV, for example, extends Unreal Engine with programmatic control over virtual cameras and scene properties such as lighting, materials, and object positions, while enabling the automatic extraction of RGB images, depth maps, surface normals, and object-level ground truth [39]. Game-based environments have also been used directly for large-scale human pose estimation and tracking datasets. By exploiting photorealistic virtual urban scenes, such frameworks can provide dense annotations for multiple people, including 2D and 3D joint coordinates, visibility and occlusion information, and persistent tracking identities across video sequences [40]. Together, these examples further illustrate that interactive game environments can serve not only as image generators, but also as structured sources of automatically available ground-truth information for pose-estimation and tracking applications.

Despite these advantages, the effectiveness of synthetic datasets is often constrained by the synthetic-to-real domain gap. Models trained exclusively on rendered imagery may exhibit lower performance when deployed in real-world environments due to differences in appearance, illumination, texture characteristics, and scene complexity. One strategy for reducing this discrepancy is domain randomization, in which visual and geometric properties of the simulated environment are deliberately varied over a broad range during training. By randomizing factors such as textures, object placement, camera parameters, illumination, and scene composition, models trained exclusively on simulated imagery can learn representations that transfer more effectively to real-world robotic perception tasks [41]. Rather than relying solely on photorealistic rendering or extensive domain randomization, another strategy is to design representations that are inherently less sensitive to synthetic-to-real differences. Silhouette-based synthetic data generation provides one such example by emphasizing body shape information rather than full RGB appearance, thereby reducing the impact of texture, lighting, and background variability [42]. This observation suggests that successful synthetic data generation for HPE depends not only on the realism and variability of simulated data, but also on the selection of feature representations that facilitate effective transfer between simulated and real-world domains.

Recent developments in 3D-aware generative modeling further highlight the growing role of generative artificial intelligence in the creation of controllable synthetic human data. Although AvatarGen was not specifically developed for human pose estimation, it illustrates the capability of modern generative frameworks to produce realistic clothed human avatars with adjustable poses, diverse appearances, and flexible camera viewpoints using only 2D image supervision. The approach leverages SMPL-guided pose conditioning, canonical avatar construction, and deformation-based modeling to separate appearance attributes from body pose, allowing generated avatars to be reconfigured and observed from arbitrary viewpoints [43]. From the perspective of synthetic data generation for HPE, such methods are significant because they demonstrate how generative AI can support the production of realistic human avatars with controllable pose, clothing,

viewpoint, and environmental conditions. These capabilities may ultimately complement traditional rendering pipelines and contribute to the development of more diverse and representative training and evaluation datasets.

Complementing these 3D-aware approaches, diffusion-based generative models provide a more direct example of how generative AI can support synthetic human-data generation for pose-related tasks. Diffusion model with Human Pose Correction (HPC) combines text-conditioned image generation with human body priors to produce photorealistic images of people while reducing anatomically implausible body configurations that commonly occur in general-purpose diffusion models. By incorporating SMPL-based pose information into the generation process, the method produces synthetic images paired with corresponding 3D body meshes that can serve as ground-truth supervision for Human Mesh Recovery models. Experimental results show that such synthetic data can improve performance in challenging domains where real 3D annotations are limited, while also enabling substantial variation in pose, appearance, and scene context. However, limitations remain in preserving fine anatomical details and complex human-object or human-scene interactions, indicating that additional geometric and structural constraints may still be required when generative models are used for pose-oriented synthetic data generation [44].

Overall, these approaches demonstrate that synthetic data generation for HPE can benefit from combining controllability and annotation accuracy of simulation-based methods with the increased visual diversity offered by generative models.

4.1 Spatial annotation from a game-based 3D environment

To illustrate the practical extraction of scene-level annotations from a virtual environment, a small proof-of-concept example was performed using a game-based 3D environment from Red Dead Redemption 2 (RDR2). The aim of this example was not to create a complete synthetic dataset, but to demonstrate that spatial metadata can be extracted from an existing interactive 3D scene and used as annotation information for future pose-estimation-oriented datasets.

The spatial annotations were extracted using a custom C# script executed through ScriptHookRDR2DotNet V2, a modding framework that allows custom C# scripts to run within the RDR2 game environment and access selected native game-engine functions. A free-camera modification was used to control the virtual viewpoint during image capture. During gameplay, the script periodically accesses native functions to obtain the final rendered camera position and rotation, as well as the world position of the playable character entity. These values were stored as three-dimensional vectors in the game world coordinate system and continuously updated in a text log file. For each selected screenshot, the current log output was saved as the corresponding annotation file, containing the camera position, character position, and camera-to-character distance.

The camera-to-character distance was calculated using the Euclidean distance defined in Eq. (1):

$$d = \sqrt{(x_p - x_c)^2 + (y_p - y_c)^2 + (z_p - z_c)^2} \quad (1)$$

where x_c, y_c and z_c denote the camera coordinates, while x_p, y_p and z_p denote the coordinates of the playable character in the RDR2 game-world coordinate system. The X

and Y components define the horizontal position within the virtual environment, while Z represents elevation. The value d denotes the Euclidean distance between the camera and the playable-character entity. The calculated distance d is reported in game units. Although RAGE-engine documentation describes the spatial units as SI-based and indicates that positions are expressed in meters, no independent calibration was performed in this proof-of-concept to verify the exact metric conversion for the specific RDR2/free-camera setup. Therefore, the reported distances are conservatively presented as internal game-world units rather than experimentally validated metric distances.

The character coordinates used in this example correspond to the world position of the playable-character entity returned by the native `GET_ENTITY_COORDS` function. They should therefore be interpreted as the entity reference position in the game world, rather than as the position of a specific anatomical keypoint or body landmark such as the head, pelvis, or body center.

The three samples were intentionally selected to illustrate different scene conditions relevant to human pose estimation: a close-view sample, a long-distance sample, and a partially occluded sample. The visual samples are shown in Figs. 6, 7 and 8, while the extracted spatial annotations and calculated distances are summarized in Tables 2, 3 and 4. The same extraction procedure can be repeated for additional camera positions, character poses, and environmental configurations, since the script continuously records the current camera and character coordinates during scene setup.

The screenshots shown in Figs. 6-8 were captured in-game from a commercially available copy of RDR2 during the proof-of-concept experiment. They are used only as illustrative material to demonstrate the proposed annotation procedure. RDR2 is the property of Rockstar Games/Take-Two Interactive.



Fig. 6 Close-view sample used for camera-to-human distance annotation. Screenshot captured by the authors in Red Dead Redemption 2.

Table 2 Extracted spatial annotations for the close-view sample

Parameter	X	Y	Z
Camera position	2461.296	-1392.571	45.541
Character position	2468.351	-1390.423	45.448
Camera-to-human distance (game units)	7.375		



Fig. 7 Long-distance sample used for camera-to-human distance annotation. Screenshot captured by the authors in Red Dead Redemption 2.

Table 3 Extracted spatial annotations for the long-distance sample

Parameter	X	Y	Z
Camera position	2381.733	-1363.494	46.352
Character position	2401.429	-1376.097	45.499
Camera-to-human distance (game units)	23.399		



Fig. 8 Partially occluded sample used for camera-to-human distance annotation. Screenshot captured by the authors in Red Dead Redemption 2.

Table 4 Extracted spatial annotations for the partial-occlusion sample

Parameter	X	Y	Z
Camera position	2432.593	-1351.548	45.568
Character position	2425.076	-1358.583	45.390
Camera-to-human distance (game units)	10.297		

These samples show that an existing game-based 3D environment can be used to obtain basic spatial annotations without manual distance measurement. Although limited in scope, the example illustrates how similar environments could support future synthetic datasets for human pose estimation.

5. CONCLUSION AND FUTURE WORK

The development of human pose estimation has moved from early convolutional regression-based solutions toward more advanced architectures capable of preserving spatial detail, modeling relationships between distant body keypoints, and processing temporal information in video sequences. These improvements are important for autonomous robotic systems, where human perception must be both accurate and computationally efficient. In practical robotic applications, however, high benchmark accuracy alone is not sufficient. Pose estimation models must also remain reliable under occlusions, changing viewpoints, complex motion patterns, and limited hardware resources.

A major requirement for further progress is the availability of diverse datasets with reliable annotations. Synthetic data generation can help address this requirement by enabling controlled variation of scene parameters while preserving access to ground-truth

information. This paper reviewed several approaches for generating synthetic data for human pose estimation, including rendering-based simulation, motion-capture-driven pipelines, generative models, and fusion-based techniques. In addition, the presented proof-of-concept example showed that an existing game-based 3D environment can be used to extract basic spatial metadata from an interactive virtual scene. The calculated camera-to-human distance illustrates how scene-level annotations may complement conventional pose labels and support future dataset generation workflows.

One limitation of this work is that the paper primarily provides a literature-based review, while the proof-of-concept example is limited in scale and scope. The RDR2-based example demonstrates only the extraction of basic spatial metadata from three selected scenes and does not constitute a complete synthetic dataset. In addition, the extracted character position corresponds to the entity reference position rather than a full anatomical pose representation, and the calculated distances are reported in game units without independent metric calibration. Therefore, the presented example should be interpreted as an illustrative demonstration of the annotation principle, not as a validated dataset-generation pipeline or a direct evaluation of human pose estimation performance in real robotic conditions.

Future work should focus on improving the transfer of synthetic data to real-world robotic conditions and on developing datasets that are better aligned with autonomous robotic perception. Generative artificial intelligence, advanced 3D rendering pipelines, and game-based annotation methods could support the creation of more diverse training and evaluation scenarios, including variations in pose, appearance, viewpoint, illumination, occlusion, and spatial relations. Further practical studies are also needed to evaluate HPE models beyond standard accuracy metrics, especially in terms of inference time, computational requirements, robustness, and real-time deployment feasibility. Multimodal sensing may provide an additional direction for improving reliability when camera-based perception is affected by occlusion, poor lighting, privacy limitations, or restricted viewpoints [45].

Acknowledgement:

This research was financially supported by the Ministry of Science, Technological Development and Innovation of the Republic of Serbia (Contract No. 451-03-33/2026-03/200109).

REFERENCES

1. Song, L., Yu, G., Yuan, J., Liu, Z., 2021, *Human pose estimation and its application to action recognition: A survey*, Journal of Visual Communication and Image Representation, 76, pp. 103055.
2. Pfister, T., Charles, J., Zisserman, A., 2015, *Flowing ConvNets for Human Pose Estimation in Videos*, Proc. 2015 IEEE International Conference on Computer Vision (ICCV), Santiago, Chile, pp. 1913–1921.
3. Wei, S.-E., Ramakrishna, V., Kanade, T., Sheikh, Y., 2016, *Convolutional Pose Machines*, Proc. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, pp. 4724–4732.
4. Chen, J., He, B., Singh, C.D., Fermüller, C., Aloimonos, Y., 2024, *Active Human Pose Estimation via an Autonomous UAV Agent*, Proc. 2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Abu Dhabi, United Arab Emirates, pp. 7801–7808.
5. Stenum, J., Cherry-Allen, K.M., Pyles, C.O., Reetzke, R.D., Vignos, M.F., Roemmich, R.T., 2021, *Applications of Pose Estimation in Human Health and Performance across the Lifespan*, Sensors, 21(21), pp. 7315.

6. Sigal, L., Balan, A.O., Black, M.J., 2010, *HumanEva: Synchronized Video and Motion Capture Dataset and Baseline Algorithm for Evaluation of Articulated Human Motion*, International Journal of Computer Vision, 87(1–2), pp. 4–27.
7. Ionescu, C., Papava, D., Olaru, V., Sminchisescu, C., 2014, *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), pp. 1325–1339.
8. Mehta, D., Rhodin, H., Casas, D., Fua, P., Sotnychenko, O., Xu, W., Theobalt, C., 2017, *Monocular 3D Human Pose Estimation in the Wild Using Improved CNN Supervision*, Proc. 2017 International Conference on 3D Vision (3DV), Qingdao, pp. 506–516.
9. Varol, G., Laptev, I., Schmid, C., Zisserman, A., 2021, *Synthetic Humans for Action Recognition from Unseen Viewpoints*, International Journal of Computer Vision, 129(7), pp. 2264–2287.
10. Dibenedetto, G., Sotiropoulos, S., Polignano, M., Cavallo, G., Lops, P., 2025, *Comparing Human Pose Estimation through deep learning approaches: An overview*, Computer Vision and Image Understanding, 252, pp. 104297.
11. Liang, S., Sun, X., Wei, Y., 2018, *Compositional Human Pose Regression*, Computer Vision and Image Understanding, 176–177, pp. 1–8.
12. Andriluka, M., Pishchulin, L., Gehler, P., Schiele, B., 2014, *2D Human Pose Estimation: New Benchmark and State of the Art Analysis*, Proc. 2014 IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, pp. 3686–3693.
13. Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L., 2014, *Microsoft COCO: Common Objects in Context*, Computer Vision – ECCV 2014, Lecture Notes in Computer Science, 8693, pp. 740–755, Springer International Publishing, Cham.
14. Cao, Z., Simon, T., Wei, S.-E., Sheikh, Y., 2017, *Realtime Multi-person 2D Pose Estimation Using Part Affinity Fields*, Proc. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, pp. 1302–1310.
15. Guler, R.A., Neverova, N., Kokkinos, I., 2018, *DensePose: Dense Human Pose Estimation in the Wild*, Proc. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, pp. 7297–7306.
16. Kanazawa, A., Black, M.J., Jacobs, D.W., Malik, J., 2018, *End-to-End Recovery of Human Shape and Pose*, Proc. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, pp. 7122–7131.
17. Pavlakos, G., Zhu, L., Zhou, X., Daniilidis, K., 2018, *Learning to Estimate 3D Human Pose and Shape from a Single Color Image*, Proc. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, pp. 459–468.
18. Liu, W., Bao, Q., Sun, Y., Mei, T., 2023, *Recent Advances of Monocular 2D and 3D Human Pose Estimation: A Deep Learning Perspective*, ACM Computing Surveys, 55(4), pp. 1–41.
19. Pavlakos, G., Zhou, X., Derpanis, K.G., Daniilidis, K., 2017, *Coarse-to-Fine Volumetric Prediction for Single-Image 3D Human Pose*, Proc. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, pp. 1263–1272.
20. Martinez, J., Hossain, R., Romero, J., Little, J.J., 2017, *A Simple Yet Effective Baseline for 3d Human Pose Estimation*, Proc. 2017 IEEE International Conference on Computer Vision (ICCV), Venice, pp. 2659–2668.
21. Cai, Y., Ge, L., Liu, J., Cai, J., Cham, T.-J., Yuan, J., Thalmann, N.M., 2019, *Exploiting Spatial-Temporal Relationships for 3D Pose Estimation via Graph Convolutional Networks*, Proc. 2019 IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Korea (South), pp. 2272–2281.
22. Loper, M., Mahmood, N., Romero, J., Pons-Moll, G., Black, M.J., 2015, *SMPL: a skinned multi-person linear model*, ACM Transactions on Graphics, 34(6), pp. 1–16.
23. Omran, M., Lassner, C., Pons-Moll, G., Gehler, P., Schiele, B., 2018, *Neural Body Fitting: Unifying Deep Learning and Model Based Human Pose and Shape Estimation*, Proc. 2018 International Conference on 3D Vision (3DV), Verona, pp. 484–494.
24. Ćirić, I., Simonović, M., Pavlović, M.G., Cvetković, S., 2025, *Mašinska vizija u mehatronici*, 1st ed., Mašinski fakultet Univerziteta u Nišu, Niš., 405 p.
25. Kappan, M.M., Sandoval, E.B., Meijering, E., Cruz, F., 2025, *A survey on deep learning for 2D and 3D human pose estimation*, Artificial Intelligence Review, 59(1), pp. 32.
26. Toshev, A., Szegedy, C., 2014, *DeepPose: Human Pose Estimation via Deep Neural Networks*, Proc. 2014 IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, pp. 1653–1660.
27. Newell, A., Yang, K., Deng, J., 2016, *Stacked Hourglass Networks for Human Pose Estimation*, Computer Vision – ECCV 2016, Lecture Notes in Computer Science, 9912, pp. 483–499, Springer International Publishing, Cham.

28. Wang, J., Sun, K., Cheng, T., Jiang, B., Deng, C., Zhao, Y., Liu, D., Mu, Y., Tan, M., Wang, X., Liu, W., Xiao, B., 2021, *Deep High-Resolution Representation Learning for Visual Recognition*, IEEE Transactions on Pattern Analysis and Machine Intelligence, 43(10), pp. 3349–3364.
29. Li, Y., Zhang, S., Wang, Z., Yang, S., Yang, W., Xia, S.-T., Zhou, E., 2021, *TokenPose: Learning Keypoint Tokens for Human Pose Estimation*, Proc. 2021 IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, pp. 11293–11302.
30. Xu, Y., Zhang, J., Zhang, Q., Tao, D., 2022, *ViTPose: Simple Vision Transformer Baselines for Human Pose Estimation*, Proc. Advances in Neural Information Processing Systems 35, New Orleans, Louisiana, USA, pp. 38571–38584.
31. Huang, Y., Liu, J., Xian, K., Qiu, R.C., 2025, *PoseMamba: Monocular 3D Human Pose Estimation with Bidirectional Global-Local Spatio-Temporal State Space Model*, Proceedings of the AAAI Conference on Artificial Intelligence, 39(4), pp. 3842–3850.
32. Bulat, A., Tzimiropoulos, G., 2017, *Binarized Convolutional Landmark Localizers for Human Pose Estimation and Face Alignment with Limited Resources*, Proc. 2017 IEEE International Conference on Computer Vision (ICCV), Venice, pp. 3726–3734.
33. Tsirikoglou, A., Eilertsen, G., Unger, J., 2020, *A Survey of Image Synthesis Methods for Visual Machine Learning*, Computer Graphics Forum, 39(6), pp. 426–451.
34. De Melo, C.M., Torralba, A., Guibas, L., DiCarlo, J., Chellappa, R., Hodgins, J., 2022, *Next-generation deep learning based on simulators and synthetic data*, Trends in Cognitive Sciences, 26(2), pp. 174–187.
35. Aso, K., Hwang, D.-H., Koike, H., 2019, *Generating Synthetic Humans for Learning 3D Pose Estimation*, Proc. 2019 IEEE Conference on Virtual Reality and 3D User Interfaces (VR), Osaka, Japan, pp. 1519–1520.
36. Chen, W., Wang, H., Li, Y., Su, H., Wang, Z., Tu, C., Lischinski, D., Cohen-Or, D., Chen, B., 2016, *Synthesizing Training Images for Boosting Human 3D Pose Estimation*, Proc. 2016 Fourth International Conference on 3D Vision (3DV), Stanford, CA, USA, pp. 479–488.
37. Varol, G., Romero, J., Martin, X., Mahmood, N., Black, M.J., Laptev, I., Schmid, C., 2017, *Learning from Synthetic Humans*, Proc. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, pp. 4627–4635.
38. Tome, D., Peluse, P., Agapito, L., Badino, H., 2019, *xR-EgoPose: Egocentric 3D Human Pose From an HMD Camera*, Proc. 2019 IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Korea (South), pp. 7727–7737.
39. Qiu, W., Yuille, A., 2016, *UnrealCV: Connecting Computer Vision to Unreal Engine*, Computer Vision – ECCV 2016 Workshops, Lecture Notes in Computer Science, 9915, pp. 909–916, Springer International Publishing, Cham.
40. Fabbri, M., Lanzi, F., Calderara, S., Palazzi, A., Vezzani, R., Cucchiara, R., 2018, *Learning to Detect and Track Visible and Occluded Body Joints in a Virtual World*, Computer Vision – ECCV 2018, Lecture Notes in Computer Science, 11208, pp. 450–466, Springer International Publishing, Cham.
41. Tobin, J., Fong, R., Ray, A., Schneider, J., Zaremba, W., Abbeel, P., 2017, *Domain randomization for transferring deep neural networks from simulation to the real world*, Proc. 2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Vancouver, BC, pp. 23–30.
42. Hori, R., Hachiuma, R., Saito, H., Isogawa, M., Mikami, D., 2021, *Silhouette-Based Synthetic Data Generation For 3D Human Pose Estimation With A Single Wrist-Mounted 360° Camera*, Proc. 2021 IEEE International Conference on Image Processing (ICIP), Anchorage, AK, USA, pp. 1304–1308.
43. Zhang, J., Jiang, Z., Yang, D., Xu, H., Shi, Y., Song, G., Xu, Z., Wang, X., Feng, J., 2023, *AvatarGen: A 3D Generative Model for Animatable Human Avatars*, Computer Vision – ECCV 2022 Workshops, Lecture Notes in Computer Science, 13803, pp. 668–685, Springer Nature Switzerland, Cham.
44. Weng, Z., Bravo-Sánchez, L., Yeung-Levy, S., 2024, *Diffusion-HPC: Synthetic Data Generation for Human Mesh Recovery in Challenging Domains*, Proc. 2024 International Conference on 3D Vision (3DV), Davos, Switzerland, pp. 257–267.
45. Ren, Y., Wang, Z., Wang, Y., Tan, S., Chen, Y., Yang, J., 2022, *GoPose: 3D Human Pose Estimation Using WiFi*, Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies, 6(2), pp. 1–25.