UNIVERSIDADE ESTADUAL PAULISTA
"JÚLIO DE MESQUITA FILHO"

João Renato Ribeiro Manesco

3D Human Pose Estimation Based on Monocular RGB Images and Domain Adaptation

Bauru
2023

João Renato Ribeiro Manesco

3D Human Pose Estimation Based on Monocular RGB Images and Domain Adaptation

Dissertação apresentada como parte dos requisitos para obtenção do título de Mestre em Ciência da Computação, junto ao Programa de Pós-Graduação em Ciência da Computação, da Universidade Estadual Paulista "Júlio de Mesquita Filho".

Área de concentração: Computação Aplicada

Financiadora: FAPESP - Proc. 2021/02028-6 e 2022/07055-4

Orientador: Prof. Associado Aparecido Nilceu Marana
Co-Orientador: Prof. Titular Stefano Berretti

Bauru
2023

Manesco, João Renato Ribeiro.
3D Human Pose Estimation Based on Monocular RGB Images and Domain Adaptation / João Renato Ribeiro Manesco. – Bauru, 2023
74 f. : il., tabs.

Supervisor: Prof. Associado Aparecido Nilceu Marana
Dissertação (mestrado) - Universidade Estadual Paulista "Júlio de Mesquita Filho", Faculdade de Ciências, Bauru, 2023

1. Estimação de Pose Humana 3D. 2. Adaptação de Domínio. 3. Deep Learning. I. Marana, Aparecido Nilceu II. Universidade Estadual Paulista "Júlio de Mesquita Filho", Faculdade de Ciências. III. 3D Human Pose Estimation Based on Monocular RGB Images and Domain Adaptation.

CDU – 518.72:76

UNIVERSIDADE ESTADUAL PAULISTA

Câmpus de Bauru

ATA DA DEFESA PÚBLICA DA DISSERTAÇÃO DE MESTRADO DE JOÃO RENATO RIBEIRO MANESCO, DISCENTE DO PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO, DA FACULDADE DE CIÊNCIAS - CÂMPUS DE BAURU.

Aos 30 dias do mês de agosto do ano de 2023, às 09:00 horas, por meio de Videoconferência, realizou-se a defesa de DISSERTAÇÃO DE MESTRADO de JOÃO RENATO RIBEIRO MANESCO, intitulada **Estimação de Poses Humanas Tridimensionais Baseada em Imagens Monoculares Auxiliada por Adaptação de Domínio**. A Comissão Examinadora foi constituída pelos seguintes membros: Prof. Associado APARECIDO NILCEU MARANA (Orientador - Participação Virtual) da FC / UNESP Bauru SP, Prof. Associado JOAO PAULO PAPA (Participação Virtual) da FC / UNESP / Bauru (SP), Prof. Titular HELIO PEDRINI (Participação Virtual) do Instituto de Computação / Universidade Estadual de Campinas, Campinas (SP). Após a exposição pelo mestrando e arguição pelos membros da Comissão Examinadora que participaram do ato, de forma virtual, o discente recebeu o conceito final: **APROVADO**. Nada mais havendo, foi lavrada a presente ata, que após lida e aprovada, foi assinada pelo(a) Presidente(a) da Comissão Examinadora.

gov.br
Documento assinado digitalmente
APARECIDO NILCEU MARANA
Data: 30/08/2023 15:16:23-0300
Verifique em https://validar.iti.gov.br

Prof. Dr. APARECIDO NILCEU MARANA

João Renato Ribeiro Manesco

3D Human Pose Estimation Based on Monocular RGB Images and Domain Adaptation

Dissertação apresentada como parte dos requisitos para obtenção do título de Mestre em Ciência da Computação, junto ao Programa de Pós-Graduação em Ciência da Computação, da Universidade Estadual Paulista "Júlio de Mesquita Filho".

Área de concentração: Computação Aplicada

Financiadora: FAPESP - Proc. 2021/02028-6 e 2022/07055-4

Comissão Examinadora

Prof. Associado Aparecido Nilceu Marana
UNESP - Câmpus de Bauru - SP
Orientador

Prof. Titular Hélio Pedrini
UNICAMP - Campinas - SP

Prof. Associado João Paulo Papa
UNESP - Câmpus de Bauru - SP

Bauru
30 de agosto de 2023

I dedicate this dissertation to my cherished family, friends, and mentors, whose unwavering support has guided me to this milestone.

Acknowledgements

I extend my heartfelt gratitude first and foremost to my family whose enduring support and assistance have enabled me to conquer this dream.

Furthermore, my sincere appreciation goes to my professors and mentors who have played a pivotal role in my educational journey, particularly, Prof. Aparecido Nilceu Marana, my advisor, for his invaluable guidance, support and kindness throughout this period.

I'd also like to thank my colleagues at MICC, as well as my supervisor during my period abroad, Prof. Stefano Berretti, for all the help during my time in Italy.

I wish to express my thanks to my friends for their companionship and unwavering presence during both challenging and joyful moments.

Lastly, my gratitude extends to FAPESP for the sponsorship through Proc. 2021/02028-6 and 2022/07055-4.

Progress is made by trial and failure; the failures are generally a hundred times more numerous than the successes, yet they are usually left unchronicled.
William Ramsay

Resumo

Estimação de poses humanas em imagens monoculares é um importante e desafiador problema de Visão Computacional cujo objetivo é obter a forma do corpo de um indivíduo baseando-se em uma única imagem. Atualmente, métodos que empregam técnicas de *deep learning* destacam-se na tarefa de estimação de poses humanas 2D. Poses 2D podem ser utilizadas em um conjunto diverso e amplo de aplicações, de grande relevância para a sociedade. Entretanto, a utilização de poses 3D pode trazer resultados ainda mais precisos e robustos. Como rótulos referentes a poses 3D são difíceis de serem adquiridos e suas aquisições podem ser realizadas apenas em locais restritos, métodos totalmente convolucionais apresentaram desempenho insatisfatório para a tarefa. Uma estratégia para solucionar este problema consiste em utilizar estimadores de poses 2D, que já se encontram mais consolidados, para estimar poses 3D em duas etapas, a partir de poses 2D. Devido a restrições na aquisição das bases de dados, a melhora de performance desta estratégia só pode ser observada em ambientes controlados, desta forma, técnicas de adaptação de domínio podem ser aplicadas com o objetivo de melhorar a capacidade de generalização dos métodos por meio da inserção de novos ângulos de câmera e ações, advindos de domínios sintéticos. Neste trabalho, propomos um novo método, chamado de *Domain Unified Approach* (DUA), que visa resolver os problemas causados pela má representação de pose em cenários com domínios distintos, por meio da adição de três novos módulos ao estimador de poses: conversor de pose, estimador de incerteza e classificador de domínio. Treinado com um conjunto enorme de dados sintéticos (SURREAL) e aplicado a um conjunto de dados obtido de um cenário do mundo real (Human3.6M), nosso método DUA levou a uma redução de 44,1 mm no erro médio por posição de junta no espaço 3D, um resultado bastante competitivo com os resultados do estado da arte.

Palavras-chave: Estimação de Poses Humanas 3D, Poses 2D, Adaptação de Domínio.

Abstract

Human pose estimation in monocular images is an important and challenging problem in Computer Vision. Currently, methods that employ deep learning techniques excel in the task of 2D human pose estimation. 2D poses can be used in a diverse and broad set of applications, of great relevance to society. However, the use of 3D poses can bring even more accurate and robust results. Since labels referring to 3D poses are difficult to acquire and can only be obtained in restricted scenarios, fully convolutional methods tend to perform poorly on the task. One strategy to solve this problem is to use 2D pose estimators, already well established in the literature, to estimate 3D poses in two steps using 2D pose inputs. Due to database acquisition constraints, the performance improvement of this strategy can only be observed in controlled environments, therefore domain adaptation techniques can be used to increase the generalization capability of the system by inserting new actions and camera angles from synthetic domains. In this work, we propose a novel method called Domain Unified Approach (DUA), aimed at solving pose misalignment problems on a cross-dataset scenario, through a combination of three modules on top of the pose estimator: pose converter, uncertainty estimator, and domain classifier. Trained on a huge synthetic dataset (SURREAL) and applied to a dataset taken from a real-world scenario (Human3.6M), our DUA method led to a 44.1mm reduction in mean error per joint position in 3D space, a result quite competitive with state-of-the-art results.

Keywords: 3D Human Pose Estimation, 2D Poses, Domain Adaptation.

List of Figures

List of Tables

List of abbreviations and acronyms

AMASS	Archive of Motion Capture As Surface Shapes
CPN	Cascaded Pyramid Networks
CVAE	Conditional Variational Auto Encoder
DA	Domain Adaptation
DAN	Deep Adaptation Networks
DANN	Domain Adversarial Neural Networks
DDC	Deep Domain Confusion
DUA	Domain Unified Approach for 3D Human Pose Estimation
GCNs	Graph Convolutional Networks
GRL	Gradient Reversal Layer
GT	Ground Truth
H3.6M	Human3.6M
HG	Stacked Hourglass
HOG	Histograms of Oriented Gradient
LCNs	Localized Neural Networks
MK-MMD	Multi-Kernel Maximum Mean Discrepancy
MoSh	Motion and Shape capture
MPJPE	Mean Per-Joint Position Error
OOD	Out-of-Distribution
P-MPJPE	Procrustes-Aligned Mean Per-Joint Position Error
RegDA	Regressive Domain Adaptation for Unsupervised Keypoint Detection
RSD	Representation Subspace Distance
SemGCN	Semantic Graph Convolutional Networks
SMPL	Skinned Multi-Person Linear Model

Contents

1 Introduction

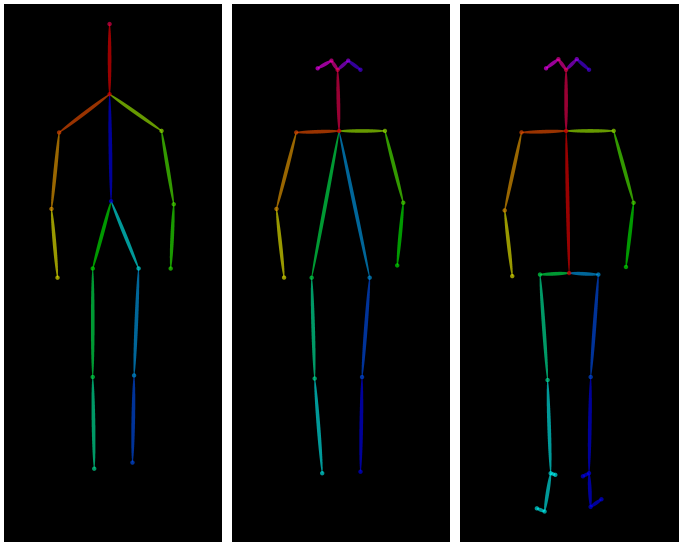
This chapter aims to introduce the task of 3D Human Pose Estimation, addressing current challenges and issues found in the existing literature, in order to properly describe the research that has been done and our contributions to the state of the art.

Human pose estimation is an essential and challenging computer vision problem. Its objective is to estimate the human body shape (pose) based on a single image, usually monocular. This shape can be inferred by the detection of joints in a skeleton, which are connected in such a way that each connection represents a part of the human body (ZHENG et al., 2020).

Currently, methods that use deep learning techniques using a bottom-up approach excel in the task of 2D pose estimation for images full of people, presenting good accuracy in real-time scenarios. Among these methods, we can cite OpenPose (CAO et al., 2019) and PifPaf (KREISS et al., 2019).

For the 2D pose representation, a few models have been proposed, among those we can find: MPI (INSAFUTDINOV et al., 2016), COCO (LIN et al., 2014) and BODY 25 (CAO et al., 2019). Figure 1 shows the main models used to represent 2D poses and Figure 2 shows examples of 2D human poses estimated in an image and represented using the COCO model (LIN et al., 2014).

Figure 1 – Main models used to represent 2D human poses.



(a) MPI (b) COCO (c) BODY 25

Source: Reprinted from Silva and Marana (2020), Copyright 2020, with permission from Elsevier.

2D poses can be employed in a diverse and vast set of applications, of major relevance to

Figure 2 – 2D human poses estimated in an image and represented through the COCO model (LIN et al., 2014).

Source: Cao et al. (2019), © 2019 IEEE.

society, among which we can mention: crowd control, action recognition, person identification, medical aid for therapies and sports analysis, human-computer interaction, augmented and virtual reality, and pedestrian location for autonomous cars (CHEN et al., 2020). However, the usage of 3D poses can bring even more accurate and robust results, as seen in 2D pose-based methods proposed by Silva and Marana (2020) and Jangua and Marana (2020), which, despite showing good results for action recognition and gait recognition, respectively, rely on the proper camera positioning to achieve success. If these methods used 3D poses on their pipeline, this restriction would be minimized and their performances would be even better.

There are a few ways to approach the 3D human pose estimation problem, for example by using depth sensors, infrared sensors, radio sensors, or even multiple camera perspectives by pose triangulation. However, these solutions end up being costly to implement or work in highly controlled environments (CHEN et al., 2020). Besides those restrictions, with the crescent growth of digital cameras shipped in mobile devices, like smartphones and webcams, a necessity to approach the 3D human pose estimation problem by using monocular RGB images emerges.

The usage of a single RGB camera introduces several challenges to the problem of human pose estimation, such as the occurrence of occlusions and the lack of full-body images of some individuals. Furthermore, variations in clothing, body type, and camera angle can have a negative impact on the performance of the methods (BARTOL et al., 2020). In Figure 3, it is possible to notice some of the challenges found in the problem of human pose estimation from images captured in real and non-controlled environments (domains). These challenges become even more significant when methods based on a single RGB image are used.

The referred challenges are further aggravated when the objective of the analysis is 3D poses since the majority of datasets used for the training of this task is obtained in controlled environments, through the usage of motion capture systems (DOERSCH; ZISSERMAN, 2019),

Figure 3 – Examples of challenges that can be found on the human pose estimation task in real environments (domains).

Source: Chen et al. (2020), Copyright 2020, with permission from Elsevier.

which decreases the amount of relevant data present, making it very difficult to apply the methods in real environments. In these cases, there is a significant disparity between the domain used for training and the real domain in which the methods are going to be applied.

The development of new methods and the improvement of techniques in the area of Computer Graphics and Computer Systems have enabled the emergence of animations and games with photorealistic features on environments and characters, as is the case of the game Grand Theft Auto V, from the company Rockstar North[1]. This kind of development enables the ability to create synthetic datasets through legal modifications in the game source code, such as the Joint Track Auto (JTA) dataset (FABBRI et al., 2018), allowing large sets of accurately labeled images to be obtained in a wide variety of environments, with different camera angles, and different times of the day. The usage of such diverse synthetic databases can contribute significantly to the improvement of 3D human pose estimation methods.

Despite the feasible contribution of synthetic data in the learning process of the task, its introduction can lead to a problem called domain shift, which occurs when the probability distributions from training and evaluation are different (KOUW; LOOG, 2018). In this case, distinctions between actions, environments, and camera angles between synthetic and real domains can directly impact in the efficacy of the method, thus, the usage of domain adaptation techniques as a way to mitigate the domain shift problem between datasets can be of great importance. Domain adaptation can be defined as the ability to apply an algorithm trained in one or more source domains to a different but related target domain (CSURKA, 2017; WANG; DENG, 2018; WEN et al., 2018).

[1] https://www.rockstargames.com/br/games/V

1.1 Problem

According to Martinez et al. (2017), state-of-the-art 3D two-step pose estimation techniques tend to perform better than their end-to-end counterparts. Even so, recent methods have reported high levels of overfitting regarding camera angles in frequently used databases, thereby impacting the performance of real applications (WEI et al., 2019; VÉGES et al., 2019; XU et al., 2020; XU et al., 2021; CHEN et al., 2021). Furthermore, evaluation protocols aimed to address these issues are rarely used in the literature, such that the most frequently used workaround is data augmentation across different camera angles.

The presence of overfitting in this type of scenario can be attributed to the fact that 3D pose dataset annotation needs to be captured in restricted and controlled environments, which implies a limitation in the set of actions and scenarios represented by the methods.

An alternative, and arguably more reliable solution to this problem, would be to include synthetic datasets during the training step, which expand the number of available poses, scenes, and camera angles, providing a way to deal with the overfitting problem caused by the limitations of a single database. As emphasized in Chapter 4, some few works seek to follow this strategy to assist the learning process. This type of solution, however, ends up requiring the manipulation of data from different domains (real and synthetic), which can be a challenging task due to possible differences in the distributions of both databases, which leads to a problem called domain shift (CSURKA, 2017).

Although domain adaptation methods are capable of handling the problem of domain shift in several types of situations, a major drawback can be noticed. The concept of domain adaptation, as well as the methods that arise from this concept, were all intended to work in a classification scenario (JIANG et al., 2021).

Therefore, in this work, we seek to address the problem of incorporating data from different domains by combining domain adaptation with techniques that address the representation problems found by the usage of different sensors in the 3D pose acquisition phase.

1.2 Objective

The primary objective of this research is to improve the existing methods for estimating 3D poses in monocular RGB images. To accomplish this, two properly labeled datasets composed of 3D human poses representing two distinct domains, real-world and synthetic, are used in conjunction with domain adaptation techniques in order to facilitate effective knowledge transfer between the two distinct domains.

Estimating 3D poses in monocular RGB images is a significant challenge as it requires inferring depth information from a single image. Various approaches have been proposed to address this issue, but in this study, our focus will be on methodologies that utilize 2D poses

as a foundation for 3D regression, as detailed in Section 2.3.3.

We found out, during the evaluation of the proposed methods, that treating the problem in a cross-domain scenario introduces many problems alongside the domain shift, such as the misrepresentation of poses between domains as well as the error propagation along the edge-joints, to deal with that, we proposed a domain unified approach for 3D human pose estimation, aimed at dealing with all the aforementioned problems in a cohesive way.

1.2.1 Specific Objectives

Besides the primary objective, this work has the following specific goals:

- Review and analyze the state-of-the-art techniques for estimating 3D poses in monocular RGB images;
- Assemble and fetch real-world and synthetic datasets to work in this cross-domain scenario;
- Investigate and implement domain adaptation techniques to enable the effective transfer of knowledge from synthetic to real domains;
- Propose a way to deal with the pose misrepresentation of joints between domains;
- Investigate the presence of domain shift in the aforementioned problem;
- Evaluate the performance of the proposed method on benchmark datasets, comparing them to existing methods, and quantifying the improvements achieved;
- Conduct experiments to assess the model efficacy and generalizability across diverse scenarios;
- Investigate the limitations of the proposed method and identify potential areas for further improvement and research.

1.3 Hypothesis

Currently, several challenges can be found in the field of 3D human pose estimation, such as pose misrepresentation, the propagation of errors on extreme kinematic joints, and the emergence of domain shift when integrating data from diverse sources.

Therefore, the core hypothesis of this work is that a unified approach, combining solutions related to pose conversion, uncertainty estimation, and domain adaptation, has the potential to achieve more robust pose estimation results, even in out-of-distribution scenarios.

1.4 Contributions

To date, few studies have investigated the direct use of domain adaptation methods to assist in the estimation of 3D poses from 2D poses in a monocular environment, however, the problem of working with multiple datasets introduces many details to the problem in such a way that the current solutions are unable to address all the nuances of cross-domain 3D human pose estimation. The following issues were identified as exacerbating the domain discrepancy in the task:

- The utilization of diverse body capture sensors across distinct datasets that leads to distinct pose representations, consequently resulting in misalignment between joints;
- Distinct domains frequently exhibit misalignment in their camera and action distributions, which can impact the accuracy and robustness of 3D human pose estimation;
- The propagation of errors on the edge kinematic groups, namely the arms, and legs, resulting in a substantial increase in the overall error.

In order to address the aforementioned problems, and subsequently mitigate the domain discrepancy, this study brings three main contributions to the pose estimation procedure:

- The introduction of a pose conversion technique aimed at achieving a Unified Pose Representation to overcome the observed differences between capture sensors;
- An enhancement and evaluation of the pose estimation training pipeline through the development of a novel uncertainty-based method;
- The creation of a domain adaptation model based on adversarial networks for 3D human pose estimation.

1.5 Document Organization

Besides this introductory chapter, this document also feature the following chapters:

Chapter 2: This chapter provides a literature review on 3D Human Pose Estimation. Additionally, a taxonomy is proposed to facilitate the systematic discussion and definition of current works.

Chapter 3: Provides a concise literature review on domain adaptation, laying the foundation to connect domain adaptation and 3D human pose estimation, and providing a formal definition of the domain-adaptative 3D human pose estimation problem.

Chapter 4: Introduces relevant works that directly address our research problem, specifically focusing on cross-domain evaluation in 3D human pose estimation and issues related to pose misrepresentation.

Chapter 5: Presents the Domain-Unified Approach (DUA) method proposed to solve the 3D human pose estimation task in a cross-domain scenario.

Chapter 6: Introduces the evaluation protocol as well as the databases and metrics used for evaluation;

Chapter 7: This chapter is dedicated to presenting the experiments and their corresponding results related to the proposed method.

Chapter 8: Concludes the work and introduces directions for future works.

2 Human Pose

This chapter provides an introduction to the human pose research problem that includes the definition of human pose, the ways human pose have been represented, and some applications of human pose. Additionally, it presents a literature review on 3D Human Pose Estimation methods and the taxonomy proposed, in this dissertation, to facilitate the systematic discussion and definition of current works.

2.1 Human Pose Definition and Representation

According to Stamou et al. (2005), a human pose can be described as an articulated body, that is, an object composed of a set of rigid parts connected through joints, which allow the execution of translational and rotational movement in six degrees of freedom.

Therefore, human pose estimation is a problem that aims to find a particular pose P in a space Π that contains all possible articulated poses. In the context of RGB images, the objective is to extract a set of features from the image that represent each joint of the human body.

There are several ways to represent a human pose, as shown in Figure 4, each one with a specific purpose. The most commonly used model in tasks of 2D and 3D pose estimation is the kinematic model, as it is a flexible and intuitive model to use (ZUFFI et al., 2012), however, this model lacks shape and texture information. Planar models are used when the human body silhouette and its respective deformations are relevant, enabling the acquisition of shape data. Finally, volumetric models are generally used when the reconstruction of three-dimensional body models is desired (ZHENG et al., 2020).

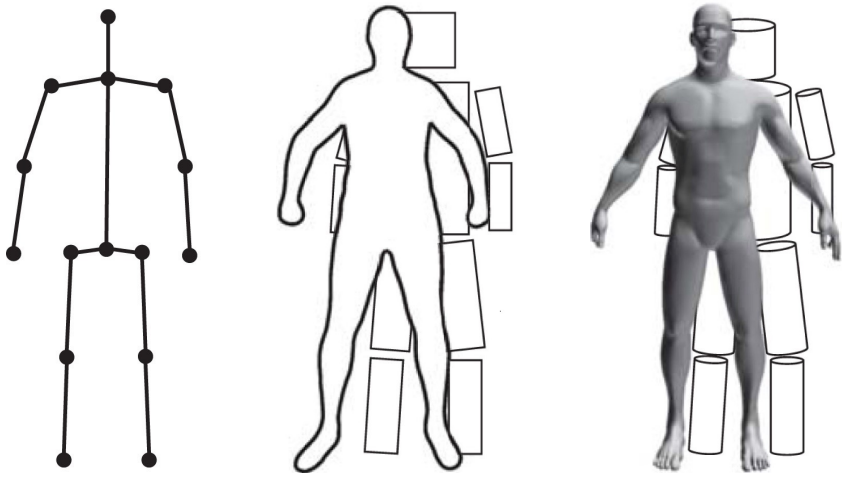
2.2 Applications of Human Pose Estimation

Human pose estimation is a fundamental and traditional computer vision problem, with the potential to serve as a crucial foundation for solving many challenges in diverse domains. In this section, we aim to illustrate the practical applications of human poses across a wide spectrum of fields. By showcasing the versatility and relevance of this task, we intend to provide an understanding of how this task is important in various domains.

2.2.1 Action Recognition

Pose information has been traditionally used as input data for training action recognition models, due to the movement information it carries, as well as, for being a low-dimensional piece

Figure 4 – Different models used to represent the human pose.

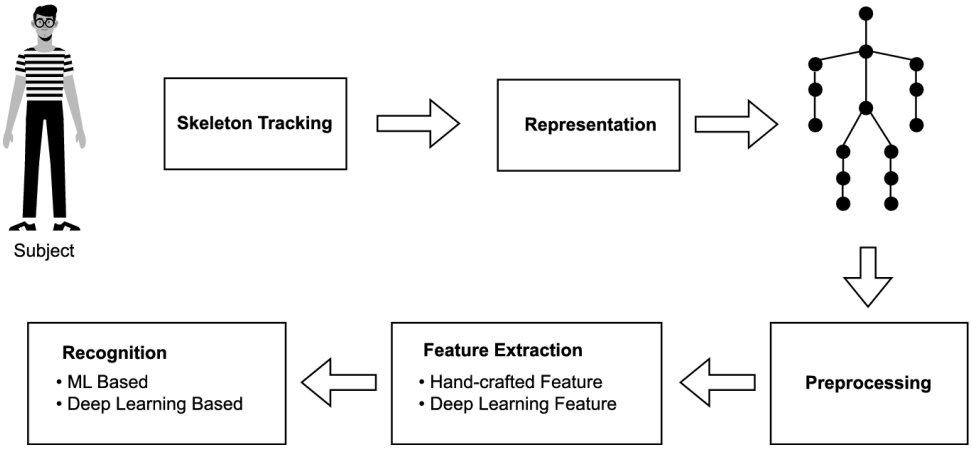


(a) Kinematic (b) Planar (c) Volumetric

Source: Chen et al. (2020), Copyright 2020, with permission from Elsevier.

of data. Earlier works used 3D skeleton information to map 3D information into Lie algebra vector spaces in order to be able to classify actions through this new representation (VEMULAPALLI et al., 2014). By using 3D poses, Devanne et al. (2014) represent actions as shape trajectories in a Riemannian manifold so that kNN can be performed on this manifold to perform action classification. A general pipeline of action recognition can be seen in Figure 5.

Figure 5 – Usual pipeline of a skeleton-based action recognition method.



Source: Sarker et al. (2021), reproduced with permission from Springer Nature.

Kim and Reiter (2017) employ skeleton-based action recognition through a Temporal Convolutional Neural Network. Angelini et al. (2018), on the other hand, do action recognition by the usage of an LSTM on 2D poses via OpenPose. Another usage of skeleton-based action recognition is given by Belluzzo and Marana (2022), who use a method where the skeleton joints projected on a black background were used as a basis for action recognition.

Silva and Marana (2020) map the bone segments of the pose to points in the parameter

space and encodes this set of points in a bag of poses classifier, used to perform action recognition.

Some authors try to employ traditional human pose estimation techniques to perform feature extraction for the action recognition task, such as Duan et al. (2022), in which the authors represent the skeletons as 3D volume heatmaps, instead of graphs, and perform convolutions upon these heatmaps.

As the extracted poses can be represented through a graph between the joints, the usage of pose information also enables the development of models based on graph neural networks as a way to model the neighboring relationship between joints. One work that exploits this is the one proposed by Shi et al. (2022), in which the authors address the problem using a pose-based graph convolutional network used to encode the body part features.

Action recognition by itself can be applied in many applications, however, the reliance of certain action recognition methods on 3D poses, or the decrease in efficacy observed in 2D-based methods when changes in the orientation are observed, increases the necessity of a 3D pose estimator capable of working in a variety of scenarios.

2.2.2 Healthcare

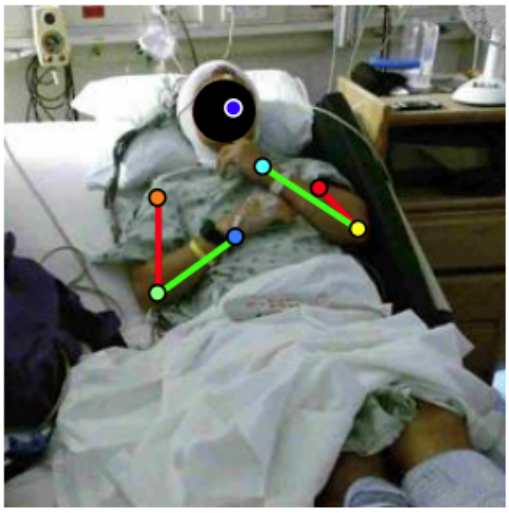
The usage of human poses in healthcare can provide several quantitative and qualitative information regarding human motion in certain areas of public health evaluation and treatment. An example of that is the work proposed by Lu et al. (2020), in which several 3D body skeletons are obtained and tracked in order to investigate the motor severity of Parkinson's Disease based on their gait.

Not only movement can be captured, but pose estimation can also help in finding reliable posture for patients in clinical hospital environments, by monitoring the patient activity in the hospital bed (CHEN et al., 2018). An example of this case can be seen in Figure 6, where the upper body pose of a patient was extracted for further evaluation.

Another application of human poses in this setting is provided by Gu et al. (2019), where a system based on computer vision is proposed as a way to perform physical therapy at home. This system, called ExerciseCheck, makes it possible for the person performing the exercise to compare their own movement with the desired movement recorded in the software.

Lastly, some works employ human poses on fall detection monitoring of elderly people, in order to provide immediate assistance. Chen et al. (2020) does that by analyzing information on the external rectangle of the predicted pose. Meanwhile, Alaoui et al. (2021) map the skeletons to a Riemannian manifold of semi-definite matrices in order to establish a dissimilarity measure between fallen skeletons so that an SVM could classify the fall.

Figure 6 – Upper body pose information of a patient positioned in a hospital bed. This information can be used to monitor proper posture and patient activity in the hospital.



Source: Chen et al. (2018), © 2018 IEEE.

2.2.3 Augmented and Virtual Reality

Human pose information can be used as a way to enhance the immersion of the interaction between real and digital objects in augmented and virtual reality scenarios. Stübl et al. (2023), for example, create an industrial augmented reality application in the domain of furniture production in order to assist in the quality inspection of the factory worker.

Weng et al. (2019) allow for the character animation obtained from a single photo on the palm of your hand by performing pose estimation and motion transfer on the animated character, in such a way that, by introducing a pose and a rigid body in the character, it is able to move in an augmented reality scenario, as illustrated in Figure 7, where a 3D model of the character in the painting starts to run in the direction of the person.

Figure 7 – Augmented reality application, in which a pose is applied to a 3D model of a character visualized in a photo.

Source: Weng et al. (2019), © 2019 IEEE.

Lastly, another augmented reality application was proposed by Zhang et al. (2021a),

in which, by retrieving information from professional tennis matches, a set of video sprites containing the players and ball trajectories is generated. The correct player actions are generated based on a pose retrieval system, in which, the most adequate pose for a certain scenario is used to select the sprites.

2.3 Pose Estimation Approaches

Early approaches tried to solve the pose recognition problem by employing traditional image processing techniques, such as the Yang and Ramanan (2013) approach, which uses histograms of oriented gradient (HOG) to create a set of features that identify each body part. However, this kind of approach proved to be inadequate to solve the problem accurately in real environments. The development of deep learning techniques and the emergence of convolutional neural networks have pushed the area of pose detection based on monocular images so that new methods started to show excellent results and overcome the traditional techniques (DANG et al., 2019).

There are various ways to approach the pose estimation problem through deep neural networks. Chen et al. (2020) classify those approaches according to the variations present when modeling the problem, these variations can appear in the form of the use or not of a pose to obtain new poses, the organization of the neural network, or the number of people involved in the scene.

Regarding the usage of poses, methods can be classified as generative or discriminative. Generative methods use the joint information on the labels to find a new set of poses and, during training, use this information to find viable joint positions in images. Discriminative methods, on the other hand, aim to learn a function that maps a person to a given pose space without knowledge of the pose models, and, based on that function, select a pose or a set of poses from a dictionary to represent the novel pose.

The neural network architecture is also a criterion for classifying pose estimation techniques, which are divided into single-stage (methods that use a single end-to-end architecture) and multiple-stage (methods that split the task between multiple networks). Estimation of 3D poses with multiple individuals is an example of a multi-stage method, where up to three networks can be defined, one to detect the individuals, one to find their 2D poses, and finally, one to project the 2D poses into the 3D space. In general, it is necessary to distinguish the approaches into two categories: those that deal with a single person in the image and those that deal with multiple people.

2.3.1 Single Person Approach

In this case, the problem refers to the situation where there is only one person in the scene, in a well-defined region, simplifying the job. If there are more people in the scene, it is

necessary to perform a preprocessing step to isolate only one person in the image. In this case, there are two types of methods commonly employed to solve the problem: methods based on regression and methods based on detection.

2.3.1.1 Regression-based Methods

Regression-based methods seek to find all the joints of the human body in an image through end-to-end networks. Toshev and Szegedy (2014) were the first to approach this problem using an AlexNet.

A problem found in this approach was that the usage of only raw information about the joint positions was not interesting, since this did not take into account the information in the neighbourhood of the joint, therefore, the supervision was converted to use heat maps representing the probable neighbourhoods of the joints.

The problem of using only heat maps is that depending on their resolution, the information regarding joints may end up being inaccurate in the decision process. To deal with this problem, Nibali et al. (2018) propose a numerical transform to calculate joint coordinates using heat maps obtained from a neural network.

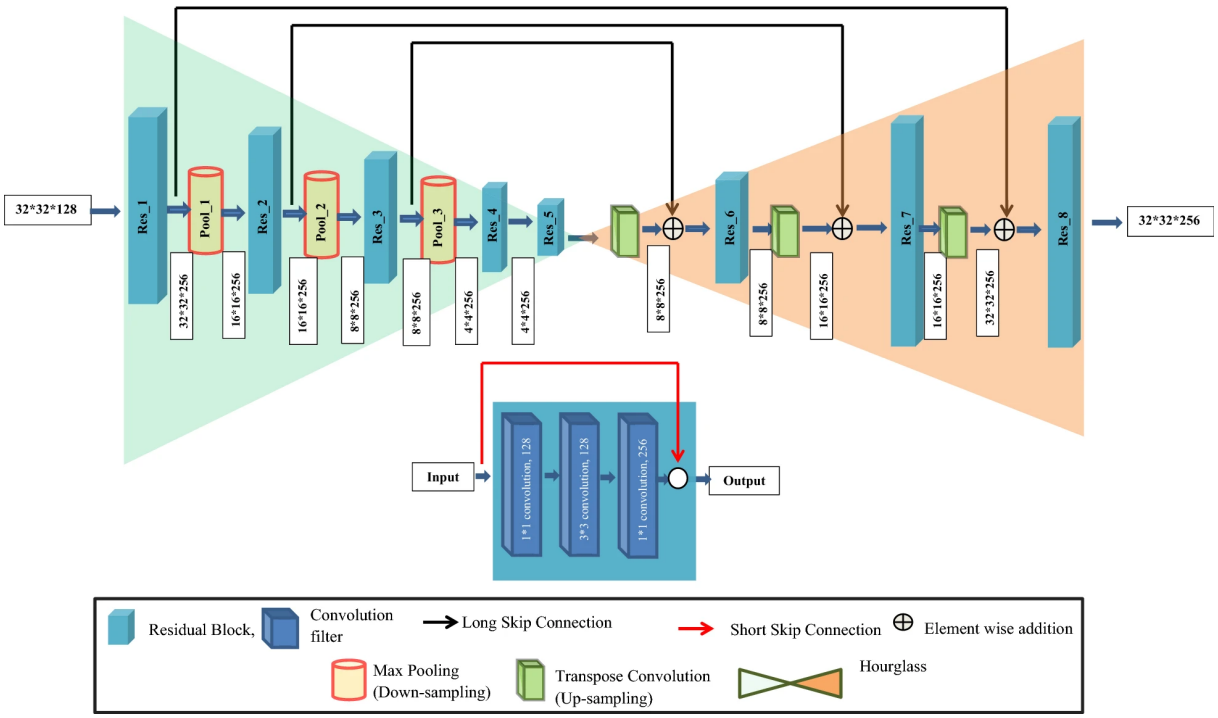
2.3.1.2 Detection-based Methods

Detection-based methods aim to perform the detection of the body segments separately so that the found body parts found are later organized to represent the human body. These kinds of methods are usually more susceptible to variations in background complexity and occlusion.

In order to improve the training process of convolutional neural networks, Jain et al. (2015) propose the usage of a heat map channel related to a joint-centered Gaussian distribution, this way, each joint will have its own heat map. Since then, most detection methods end up working with heap maps. Furthermore, Luvizon et al. (2019) propose a soft-argmax function with the goal of converting detection-based networks into regression-based networks, making an end-to-end network suitable for use in detection environments.

Some of the detection-based approaches still use traditional architectures, like GoogLeNet (RAFI et al., 2016). Another architecture that has become prominent recently, exhibiting good results in the pose estimation task, is the stacked hourglass architecture, proposed by Newell et al. (2016). This kind of architecture works by employing a set of cascaded pooling and upsampling layers in order to extract information at various scale ranges, thus obtaining more accurate information about the orientation and position of each joint. An example of an hourglass model can be seen in Figure 8.

Figure 8 – Hourglass module. Each one of the blue boxes is a residual module presented below the network.



Source: Misra et al. (2020), licensed under CC by 4.0.

2.3.2 Multiple People Approach

The presence of multiple people in an image can significantly increase the complexity of the problem, requiring either preprocessing to isolate people in the image or architectures that are able to identify individuals in images, even then, problems involving interaction between individuals can still occur.

One approach to the detection of subjects in images is called the Top-down approach. In this case, high-level abstractions are used first to detect the individuals and from that point, pose estimation is applied to each one of the subjects.

Another approach is called Bottom-up, in which, instead of detecting the subjects directly, the objective is to find the joints and limbs of all the subjects and then sort them by clustering the joints. This kind of approach can have major problems when there are people interacting very close together in an image.

2.3.3 3D Pose Estimation

Although there are commercial solutions that address the 3D pose estimation problem, these solutions work mostly in restricted environments, as is the case of those based on Kinect which has a depth sensor, or use markers for body detection (CHEN et al., 2020), which ends up being quite restrictive. Therefore, there is a need to propose more flexible solutions to 3D pose estimation, which can be used in uncontrolled scenarios, preferably using a low-cost easily

accessible monocular RGB camera.

With the emergence of deep learning and convolutional neural networks, the performance of the methods began to improve. An initial solution involving end-to-end networks for 3D pose regression was proposed, however, this type of solution is difficult to be employed in scenarios different from those used in training, as the databases for 3D pose estimation need to be captured in controlled environments, decreasing the diversity of the data and impacting the generalization ability of the networks (LI; CHAN, 2015).

With the development and popularization of 2D pose estimation techniques achieving surprising results, such as the OpenPose (CAO et al., 2019), PifPaf (KREISS et al., 2019) and *Stacked Hourglass* (NEWELL et al., 2016), approaches that seek to take advantage of the maturity of 2D pose estimators in 3D pose estimation have been gaining popularity. This is done by performing a two-step pose estimation process, a first step that is responsible for obtaining valid 2D poses and a second step in which, from the 2D pose, a 3D pose is retrieved.

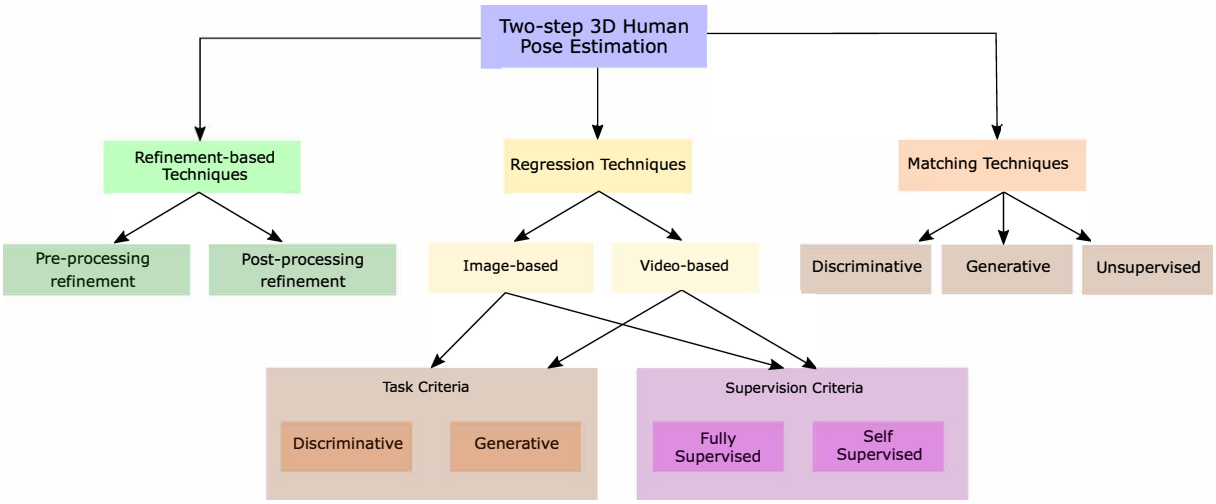
The idea is that 2D pose estimation techniques, whose labels are easier to retrieve in a diverse range of situations, can help to ease the labor of obtaining 3D data, and can provide accurate enough information for the 3D pose lifting (MARTINEZ et al., 2017). This is, however, a difficult problem, since depth information, which is already scarce in images, is lost with the 2D pose, and the problem itself is non-invertible, as a single 3D pose may have more than one 2D pose projection representing itself.

Despite being a difficult problem, two-step 3D human pose estimation has recently been accomplished in the literature. Moreno-Noguer (2017) was one of the first to achieve that, by creating Euclidean distance matrices representing the relationship between different joints in a spatial context, making it possible to achieve results comparable to those of end-to-end networks using only information found in 2D poses. Meanwhile, Martinez et al. (2017) goes further and show that, through proper preprocessing of the poses, a simple residual architecture is able to estimate 3D poses with higher accuracy than end-to-end methods using only a 2D pose as input.

2.4 Proposed Taxonomy for 3D Human Pose Estimation Methods

Aiming at a better understanding of 3D pose estimation techniques based on 2D poses, we proposed, through an analysis of the literature, a taxonomy in order to categorize the different methods according to key characteristics shared among them. This taxonomy is presented in Figure 9. In general, three major types of 3D pose estimation techniques based on 2D poses have been identified in our analysis through clustering similar techniques: pose refinement-based techniques, pose regression techniques, and matching techniques. After analyzing each method's key features, we can formulate proper definitions for them. These definitions will be described in the following paragraphs.

Figure 9 – Taxonomy proposed in this dissertation to categorize the different two-step 3D pose estimation approaches found in the literature.



Source: Manesco and Marana (2022). Reproduced with permission from Springer Nature.

Refinement techniques, as established by Wang et al. (2019a) are defined by a pre-processing or post-processing step in the pose dataset, in order to find a common space or perspective for the poses, or, through a refining step applied on the 3D poses already predicted, based on joint error, a frame sequence, or bone consistency, improving the accuracy of an already established 3D pose estimator.

Matching techniques, such as the one proposed by Chen and Ramanan (2017), are characterized by the creation of a dictionary with poses from the training set, in such a way that the 2D pose input is matched with the most similar pose that can be found in the dictionary, and the corresponding 3D pose is chosen as a pose candidate. Matching techniques can operate in a partial context, that divides the poses dictionaries by kinematic groups, or in a complete context, where the whole body is used as input. In addition, they can work in a discriminative context whose objective is to only obtain the desired pose, or in a generative context, which aims to create an artificially generated dictionary to help.

Lastly, regression-based techniques, such as Martinez et al. (2017), consist of using mechanisms, such as neural networks, to recover a 3D pose from the input 2D pose, in a regression learning context. Regression techniques can operate in an image-based analysis, i.e. a single frame is considered for the evaluation, or in a video-based analysis, where a sequence of frames is observed. Finally, regression-based techniques are divided between both task and supervision criteria.

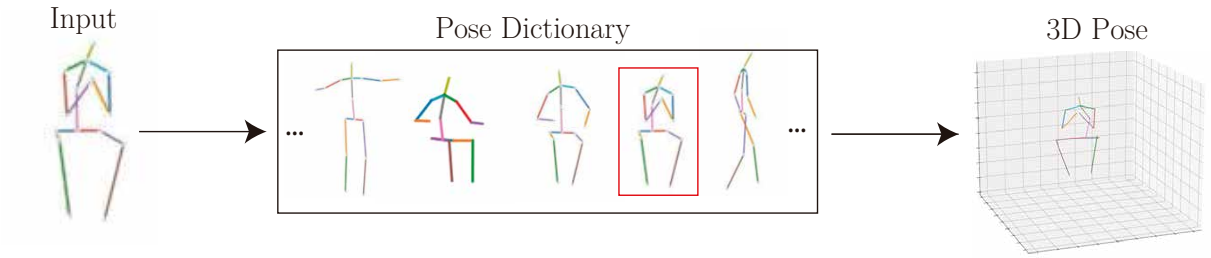
Regarding the task criteria, regression-based techniques resemble matching techniques, in which there is a discriminative context, whose goal is to regress the 3D pose, and a generative context, which can operate through data augmentation mechanisms or in adversarial generative scenarios aimed to generate new poses and increase the generalization capacity of the system. As for the supervision criterion, we can classify the techniques into fully supervised, whose 3D

ground truth data is used during the regression step, or self-supervised, where only the 2D pose information is considered, making use of other types of information as the label, such as the projections of the predicted 3D poses into a 2D camera.

2.4.1 Matching Techniques

Matching techniques work on top of a dictionary created from the training dataset. One of the pioneers in this approach were Chen and Ramanan (2017), proposing a discriminative method that consists of creating a dictionary of 2D projections from the training set. The method operates by searching for the 2D projection with the highest probability of representing the input 2D pose and relating it to the target 3D pose. Figure 10 illustrates the operation of a generic matching technique.

Figure 10 – Illustration of the operation of a generic matching technique.



Source: The Author.

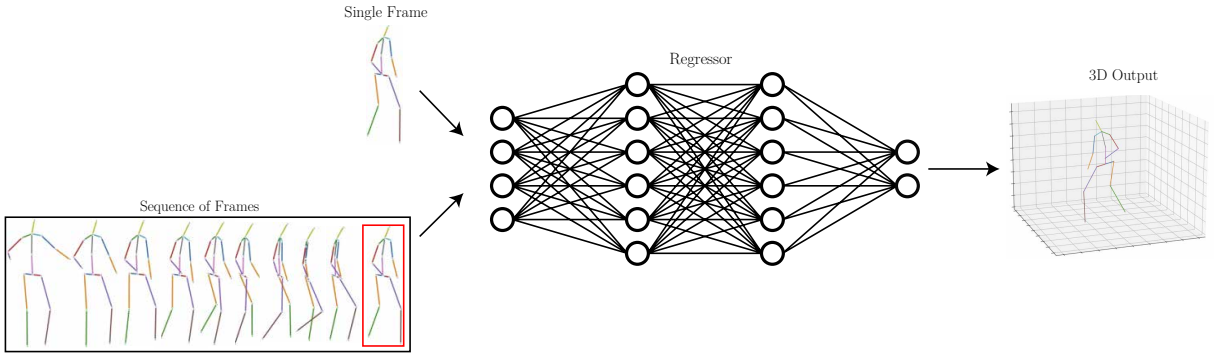
An alternative approach works in a generic generative scheme that can accommodate various types of matching techniques. The approach consists of augmenting the dataset based on a set of anatomical constraints, upon which, traditional matching techniques can be employed (JAHANGIRI; YUILLE, 2017).

Localized approaches, such as the one proposed by Yang et al. (2019), work by dividing the pose matching into upper and lower kinematic groups, which are then used to populate the 2D pose dictionary through several camera perspectives, by combining different groups. Other locality-based methods further spread the local kinematic groups, aiming to divide the task into different local-matching problems (ZHOU et al., 2020). A different approach works on an unsupervised scenario, where a sparse representation in a linear combination of pose basis is learned before the matching occurs (JIANG et al., 2019).

2.4.2 Regression Techniques

Regression techniques, on the other hand, seek to estimate three-dimensional coordinates from techniques such as neural networks, using a 2D pose as input. There are several approaches to regression methods, whose temporal information or 3D labels may or may not be used during learning. Figure 11 illustrates the general processing approach of regression methods.

Figure 11 – General scheme of operation of regression methods.



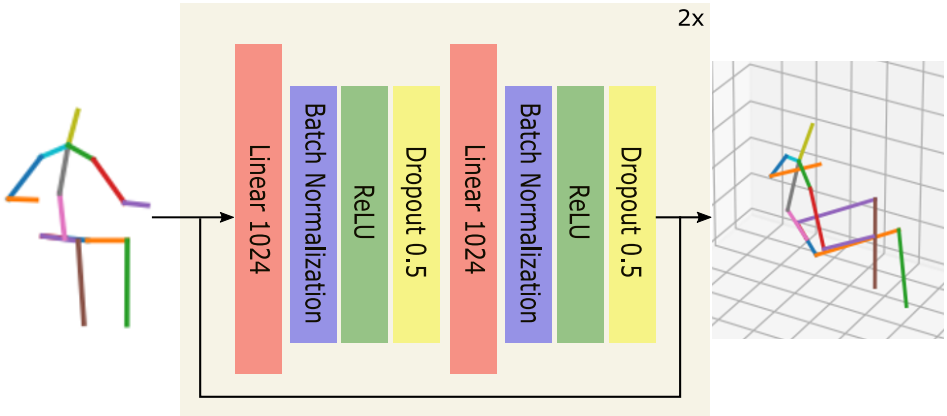
Source: The Author.

2.4.2.1 Image-based Regression Techniques

Image-based methods aim to perform human pose estimation based on a single frame or 2D input pose. Moreno-Noguer (2017) was one of the pioneers in this area through Euclidean-distance matrices used as input to neural networks. The method was distinguished by its comparable performance to the end-to-end methods observed at the time.

Another relevant work is the one proposed by Martinez et al. (2017), in which the authors create a baseline for pose estimation through proper pose processing and a simple residual neural network. The proposed method consists of a simple residual neural network combined with a suitable processing of the poses, achieving powerful results, and outperforming the end-to-end approaches of the time. This processing, considered standard for several subsequent methods, consists in multiplying the 3D poses by the inverse of the extrinsic camera parameter matrix, aiming at representing the pose in a canonical space. Figure 12 shows the residual neural network architecture proposed to solve the two-step 3D pose estimation problem.

Figure 12 – Residual neural network architecture proposed to solve the two-step 3D pose estimation problem.



Source: Adapted from Martinez et al. (2017), © 2017 IEEE.

Other methods are also inspired by this architecture, such as Véges et al. (2019), which

uses it in a siamese neural network scheme aiming to find rotation-equivalent representations, i.e. where the rotation projection matrix has the same values, such that the poses are represented in a universal space before being estimated.

Pavlakos et al. (2018) treat the lack of depth information through an ordinal relationship, modeled through a neural network. Another method aims to diversify the set of predicted poses, through the usage of a Gaussian mixture model, used to predict more than one valid pose. Xu et al. (2021) utilizes pose grammar and data augmentation to deal with the problem. Another fresh perspective to the problem deals with a bone representation, instead of 2D joints, to estimate the 3D poses (WEI et al., 2021). Li and Lee (2019), on the other hand, employ a Gaussian mixture model, aiming to find the parameters of the distribution of the model in M Gaussian kernels, this is done through a deep neural network and allows the generation of multiple hypotheses of poses that satisfy the problem, enabling the choice of the best pose among those available.

2.4.2.1.1 Graph Neural Networks

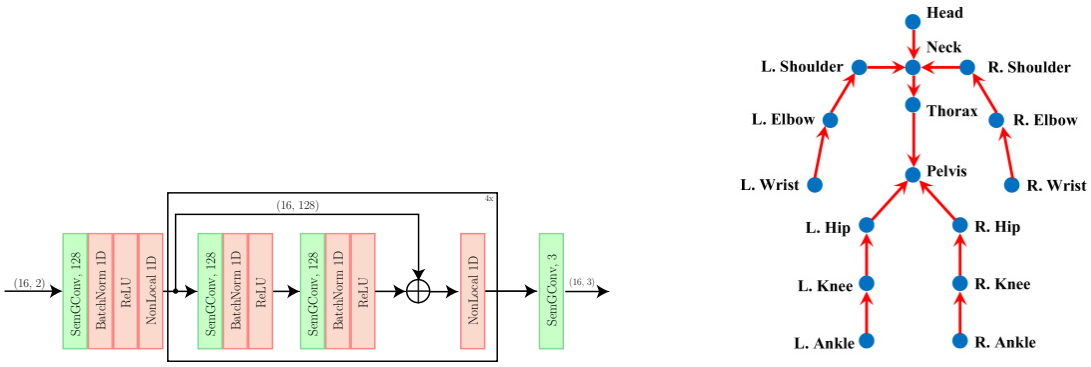
Due to the recent success of graph networks in tasks that employ poses, such as action recognition, several methods attempting to employ graph networks in the task of 3D pose estimation have been proposed. Simple approaches to the problem include replacing the linear layers in the model in Figure 12 with convolutional graph layers (BANIK et al., 2021).

Early uses of this approach involved employing SemGCNs to solve the problem, a kind of graph neural network that works with a semantic representational model for 2D poses, starting from a set of operations called semantic convolutions. This set of semantic convolutions works in such a way that each of the nodes of a graph, which already semantically defines the problem, has its own convolution matrix, solving a problem in which traditional graph neural networks do not perform well in this task, as they share the same convolution filter between all nodes. Aspects of the SemGCN network are exposed in Figure 13.

Ci et al. (2020), in contrast, propose the creation of Localized Neural Networks (LCNs) to solve the problem, combining concepts of Graph Convolutional Networks (GCNs), which have a problem in the representation of convolution filters, with concepts of fully connected networks, which do not exploit the connections between vertices directly during learning. To accomplish that, each of the nodes in an LCN has its own weight matrix with its own set of filters.

SemGCNs have been explored in several other scenarios, such as density mixture models to generate multiple pose hypotheses (ZOU et al., 2021), or even in a generative-adversarial context, in which one SemGCN acts as a generator of 3D poses, whereas the other SemGCN acts as a discriminator to differentiate real poses from artificially generated poses (XIA; XIAO, 2020).

Figure 13 – SemGCN network architecture (a), accompanied by the representation of the semantic graph of a pose used as a basis for the development of the task (b).



(a) SemGCN architecture. (b) Semantic Graph Pose Representation.

Source: Adapted from Zhao et al. (2019), © 2019 IEEE.

An interesting use of the SemGCN has been proposed by Sun et al. (2020), who employ them in a generative context, through a stereo pose generator module to obtain new viewpoints for an input pose. In this way, the problem is transformed into a multi-view 3D pose acquisition problem.

Some graph-based methods started expanding on the idea of the SemGCN through the usage of attention and transformers. Yin et al. (2023), for example, employ an attention mechanism aiming to extract global joint features from the input skeleton without neglecting local and neighboring information.

Another such work is the one proposed by Zhao et al. (2022), which introduces the GraFormer architecture, that works by embedding graph convolutional layers working together with an attention block capable of learning implicit relationships with a dynamic adjacency matrix.

2.4.2.2 Video-based Regression Techniques

In contrast to image-based techniques, video-based techniques employ a collection of frames in order to learn. One of the first approaches working with videos employed LSTMs to perform pose prediction (HOSSAIN; LITTLE, 2018). This type of approach has continued to be explored, such as the method of He et al. (2019), which employs a BiLSTM after splitting the poses into kinematic groups, with a regression head for each group.

Based on the idea of dividing poses into local groups, Zeng et al. (2020) split a single pose into local poses, focusing on the arms, legs, and torso, and proposed the use of an SRNet, composed of recurrent neural networks, to add global context information to local connections.

Other approaches seek to use temporal convolution networks for learning, either by creating memory banks (DENG et al., 2020) or by analyzing bone consistency, such as the method of Chen et al. (2021), which explores anatomical constraints through a network used to

predict bone size by analyzing random frames in conjunction with an analysis of bone direction consistency on consecutive frames.

Taking advantage of the trend of graph methods in image-based 3D pose estimation, Zhang et al. (2021b) propose a dynamic spatial convolutional graph method, whose goal is to generalize the spatial relationship between vertices over time. Assuming that the local analysis of traditional graph techniques always takes into account the same vertices, even if there is no interaction between them, the authors propose a dynamic graph convolution, which updates the joint neighborhood according to the Euclidean distance between the other vertices in a given frame.

Liu et al. (2021), in turn, use a set of attention mechanisms combined with temporal dilated convolutions, trained with the full frame sequence in order to estimate 3D poses in a temporal context. Meanwhile, Li et al. (2022) employ a pose-based transformer to learn a proper spatiotemporal representation of different poses, generating multiple initial pose hypotheses and learning the temporal relationship between them in order to deal with the non-invertible pose representation problem.

The concept is also examined in the work of Zhao et al. (2023), which utilizes a spatial encoder to detect the correlation between joints in a single frame. The authors of this work also employ a temporal encoder to acquire spatiotemporal data from the joints. In addition, they incorporate a Discrete Cosine Transform to represent the overall movement of the pose through a low-frequency representation of the skeleton.

To handle incomplete input sequences, Einfalt et al. (2023) replaced the missing poses with position-aware upsampling tokens. These tokens were then transformed into 3D pose estimates through self-attention of the entire sequence.

2.4.2.3 Self-Supervised Methods

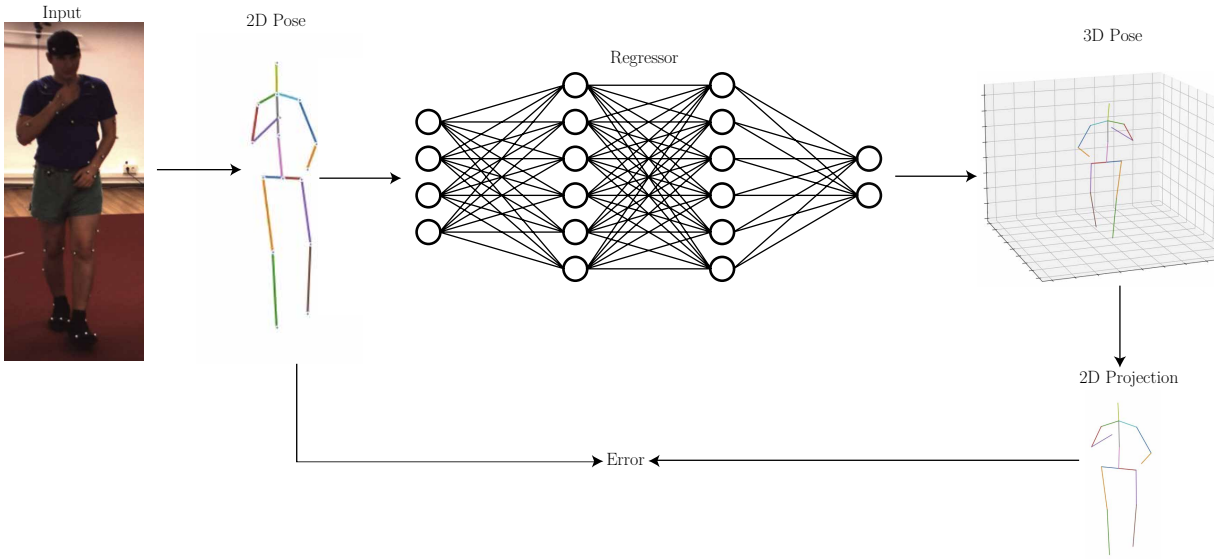
Self-supervised methods work by using the information present in the extracted 3D pose itself as a label, without the need for direct information from the 3D ground truths. Figure 14 illustrates the general behavior of self-supervised 3D human pose estimation techniques.

Dabral et al. (2018) work on this type of method by ensuring that the estimated 3D pose satisfies a set of anatomical constraints, such as the truth of an angle, or valid symmetry for twin limbs. Following a similar line, Xie et al. (2019) start from geometric constraints based on symmetry, in a context of graph networks, to perform the method supervision.

Wang et al. (2019b) utilize a mechanism for generating 3D poses without the need for 3D labels, for this purpose, the output of a 3D pose estimator module based on LSTMs serves as input to a 3D pose refiner module, which, in its intermediate layers, compares the projection of the estimated 3D pose with the 2D pose input in order to define the error function.

Another work, proposed by Klug et al. (2020) aims to define theoretical limits for

Figure 14 – General scheme of a self-supervised method based on the error obtained between the 3D pose projection and the input 2D pose.



Source: The Author.

self-supervised approaches, through a minimum threshold error for pose estimation techniques that use weak projections. The threshold is based on the error propagated by the distortions caused by the projection, thus, even if a technique can estimate almost perfectly a 3D pose, the distortion caused by the projection of the 3D pose on the camera, used for error propagation, will be taken into account and a minimum threshold can be defined for each of the analyzed databases.

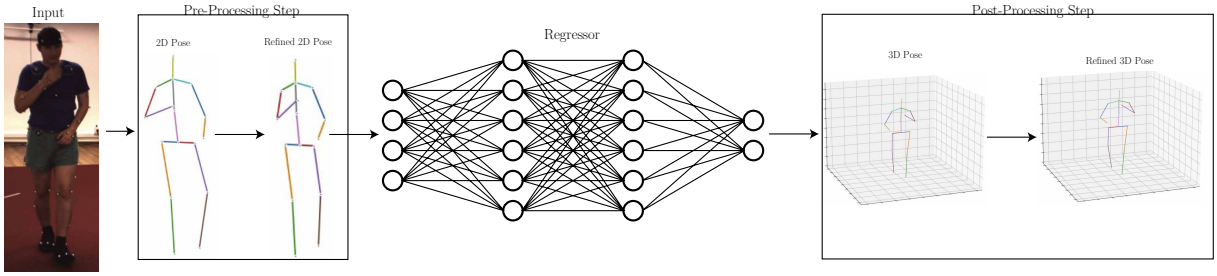
2.4.3 Refinement-based techniques

Refinement-based techniques work on improving the pose representation, so that other pose estimation techniques can achieve better results. As illustrated in Figure 15. Wang et al. (WANG et al., 2019a), for example, carry out the post-processing of the poses by learning a distance matrix, upon which k pose-basis are retrieved and used to reposition the 3D pose in a new space.

Guo et al. (2019), on the other hand, perform the refinement by grouping joints into different categories: easy, medium, and hard, referring to the error magnitude at the joints. In this way, each type of joint undergoes another regression step in a neural network, seeking to adjust the poses in groups, without the group with greater error influencing the other groups.

Regarding the pre-processing of the poses, one approach aims to fix the poses by maintaining the consistency of the gravity center along the frames (XU; WU, 2020). Liang et al. (2020) employ a neural network on the 2D poses, in such a way that all of the 2D poses are represented in the same perspective, dealing with camera angles overfitting. Lastly, following a similar idea, Wei et al. (2019) employ a hierarchical correction network in a generative

Figure 15 – Overview of refinement methods, illustrating a situation where both pre-processing and post-processing techniques are applied.



Source: The Author.

adversarial context, to find a new perspective for the 2D poses.

Xu et al. (2020) create a video processing technique through a cinematic analysis based on the bone's length and angle that projects 2D poses to a common perspective, and at the end of the prediction, refine the predicted poses with a low 2D confidence score based on their trajectory.

3 Domain Adaptation

This chapter aims to present the motivation and the definition of domain adaptation, and to provide a short literature review on domain adaptation, with the intent of orienting the reader within this subject matter and drawing the necessary connections between domain adaptation and 3D human estimation.

3.1 Motivation and Definition

When images from the training dataset and the test dataset show differences between their data distributions, a problem called domain shift arises. This problem can cause a negative impact on the accuracy of classifiers, leading images to be misclassified (KOUW; LOOG, 2018). One way to deal with this problem is to use domain adaptation techniques (PATEL et al., 2015).

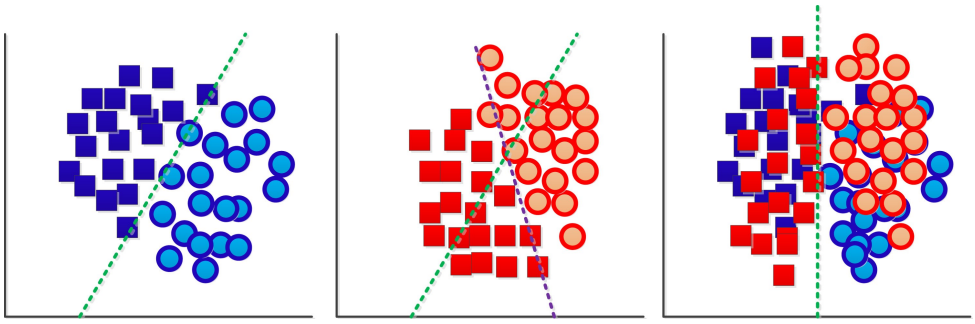
Some authors define domain adaptation as a sub-area of transfer learning, whose goal is to use data from a domain other than the one used for training, in order to improve the accuracy of the classifier when applied to an alternative dataset (CSURKA, 2017). Recent studies in the literature, instead, define domain adaptation as a subarea of domain generalization. In both approaches, the common goal is to address the domain shift problem in unsupervised target distributions, the difference being that domain adaptation techniques typically focus on addressing the domain shift within a well-defined target domain, leveraging accessible data to assist in the distribution learning process. Domain generalization, on the other hand, covers a broader scope, emphasizing generalization to out-of-distribution (OOD) unseen domains solely based on the available source data (ZHOU et al., 2022).

Figure 16 shows the importance of applying domain adaptation techniques when dealing with the domain shift problem. In the middle graph, when comparing the source domain classifier (highlighted in blue), with the target domain classifier (highlighted in red), one can observe the negative impact of using a different distribution to evaluate a classifier trained in a distinct domain. The right graph shows a scenario in which domain adaptation techniques were applied and a common domain was found between both datasets, solving the domain shift problem.

The formal definition of the concepts that compose the basis of domain adaptation theory, according to Pan and Yang (2010) are exposed hereafter. Given a Dataset composed of

Definition 1 (Domain)**.** *A domain D is composed of a feature space $\mathcal{F}$ with d dimensions and a marginal probability function $P(x)$, which means that, given a data point x, we have $D = \{\mathcal{F}, P(x)\}$, with $x \in \mathcal{F}$.*

Figure 16 – Example of a domain shift problem. The first graph (left) shows source domain data with a trained classifier highlighted in blue. The second graph (middle), shows the target domain with a trained classifier highlighted in red. In the third graph (right), data from both domains are shown after domain adaptation, with a classifier trained on the common domain.



Source: Reprinted from Chai et al. (2016), Copyright 2016, with permission from Elsevier.

Definition 2 (Task). *Given a domain D composed from the set of training data $\{x, y\}$, a task $\mathcal{T}$ consists of a set of labels $\mathcal{Y}$ and an objective predictive function $f(x)$, that can be learned from the training data, which means that $\mathcal{T} = \{\mathcal{Y}, f(x)\}$, with $y \in \mathcal{Y}$ and $f(x) = P(y|x)$.*

Definition 3 (Domain Adaptation). *Given a source domain $\mathcal{D}_\mathcal{S} = \{X^\mathcal{S}, Y^\mathcal{S}\}$ and a target domain $\mathcal{D}_\mathcal{T} = \{X^\mathcal{T}, Y^\mathcal{T}\}$, and assuming that $\mathcal{D}_\mathcal{S} \neq \mathcal{D}_\mathcal{T}$ regarding their marginal probabilities $P(X^\mathcal{S}) \neq P(X^\mathcal{T})$, and two tasks $\mathcal{T}_\mathcal{S} \approx \mathcal{T}_\mathcal{T}$, with conditional distribution $P(Y^\mathcal{S}|X^\mathcal{S}) \approx P(Y^\mathcal{T}|X^\mathcal{T})$; The goal of domain adaptation is to improve the prediction $f_\mathcal{T}(\cdot)$ in the target domain $\mathcal{D}_\mathcal{T}$, using the source domain $\mathcal{D}_\mathcal{S}$ data.*

3.2 Deep Domain Adaptation

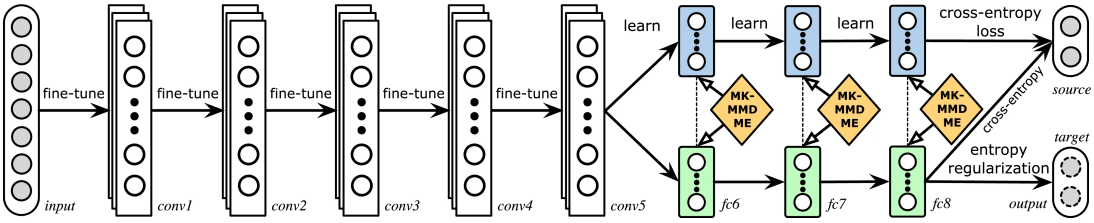
The subject of domain adaptation has also benefited from recent developments in the field of deep learning. Initially, deep neural networks were used only as feature extractors for later application of traditional domain adaptation techniques, however, the development of the area enabled the establishment of architectures and training protocols focused on domain adaptation in the deep learning scenario (CSURKA, 2017). These architectures come in different forms, with methods based on the discrepancy between domains, methods that use adversarial training, methods based on autoencoders, and methods that take advantage of the spatial relationship between the data. Two architectures that are fairly simple to apply and show excellent results are the Deep Adaptation Networks (DAN) architecture (LONG et al., 2018) and the Domain Adversarial Neural Networks (DANN) architecture (GANIN et al., 2017).

3.2.1 Deep Adaptation Network

The Deep Adaptation Network (LONG et al., 2018) architecture is based on the principle that the last layers of a convolutional neural network have more specific features, thus the first

layers have their weights frozen, while the last fully connected layers of a pre-trained network are fine-tuned in such a way that the Multi-Kernel Maximum Mean Discrepancy (MK-MMD) metric, used to calculate the distance between distributions, from the last layers are minimized by being part of a domain error function that composes the fine-tuning part of the training. Figure 17 shows one example of a DAN network.

Figure 17 – The Deep Adaptation Network architecture.



Source: Long et al. (2018), © 2018 IEEE.

The MK-MMD distance between two probability distributions p and q is defined by the distance between the average representation of their distributions in a reproducing kernel Hilbert Space, from a kernel k, from which, the following expression can be obtained:

$$d_k^2(p, q) \triangleq ||E_p[\phi_k(x)] - E_q[\phi_k(x)]||_{\mathcal{H}_k}^2. \tag{1}$$

Thus, the error is minimized by:

$$\min_{\Theta} \frac{1}{n} \sum_{i=1}^{n} J(\theta(x_i), y_i) + \lambda \sum_{l=4}^{5} d_k^2(\mathcal{D}_{\mathcal{S}}^l, \mathcal{D}_{\mathcal{T}}^l), \tag{2}$$

where J is the error function applied on the neural network output $\theta(x_i)$ for an input data object x_i, in comparison to a true label y_i, $d_k^2(\mathcal{D}_{\mathcal{S}}^l, \mathcal{D}_{\mathcal{T}}^l)$ denotes the MK-MMD between the source domain ($\mathcal{S}$) and target domain ($\mathcal{T}$) at layer l.

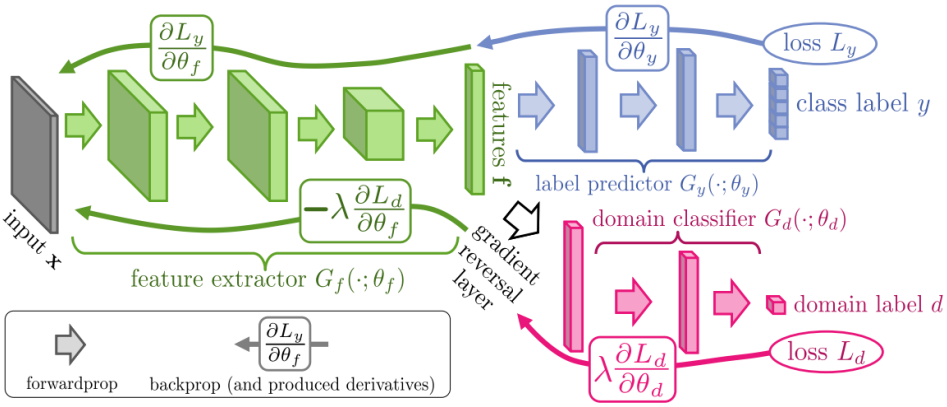
3.2.2 Domain Adversarial Neural Network

The Domain Adversarial Neural Network architecture (GANIN et al., 2017) follows an adversarial training protocol through a domain-classifier model, that is used to assure that the feature extraction module finds a uniform representation of the features in a common domain.

The method combines a feature generator G_f, with a domain discriminator G_d and a label predictor G_y. The main idea of the method is to maximize the confusion of the domain classifier in a way that does not incur performance loss on the label predictor. Figure 18 shows how the DANN architecture works.

Initially, the method obtains the feature representation $G_f(X)$ of a data input X. These features serve two distinct purposes: predicting class labels $G_y(G_f(X))$, and domain labels $G_d(G_f(X))$.

Figure 18 – The Domain Adversarial Neural Network architecture.



Source: Ganin et al. (2017), reproduced with permission from Springer Nature.

After properly obtaining the desired labels, the method proceeds into maximizing the domain classifier confusion through the usage of a Gradient Reversal Layer (GRL). The idea is that the gradient of the domain classifier is reversed with respect to the feature extractor during backpropagation, this is achieved by multiplying its gradients by a negative scalar $-\Lambda$.

By reversing the gradient and maximizing domain confusion, the training process compels the feature extractor to learn domain-invariant features. This ensures that the feature distributions over both domains are as indistinguishable as possible when passed to the domain classifier. The overall loss of the method is given by Equation 3, where L_y represents the loss of the label predictor, L_d the loss of the domain classifier, and y the ground-truth label information.

$$L_{\text{DANN}} = L_y(G_y(G_f(X)), y) - \Lambda L_d(G_d(G_f(X))). \tag{3}$$

3.3 Regressive Domain Adaptation

Although domain adaptation methods are able to deal with the domain shift problem in several types of situations, a fundamental issue can be observed: the concept of domain adaptation, as well as the methods that arise from this concept, were envisaged to operate in a classification scenario.

Jiang et al. (2021) show that there are few methods focused on addressing the particularities of regression methods and that, in certain scenarios, traditional Domain Adaptation techniques, are not able to adequately adapt to the task. This fact occurs as a consequence of the behavior of the border that separates the predictions in the different tasks. In a classification scenario, the border between classes is usually well-defined in both domains and, when applying domain adaptation techniques, the margins of the border that defines the classes in the source domain can be enlarged, increasing the generalization capacity of the classifier for the new

domain.

In regression problems, on the other hand, due to the problem being located in a continuous space, the decision margins are not as clear as in classification problems. This problem can be aggravated when dealing with keypoint detection, whose output also involves a high-dimensional discrete space.

Because of this particularity of domain adaptation methods in regression tasks, such as the estimation of 3D poses from 2D poses, some techniques that are specific to the regression task have been proposed or adapted for the application of domain adaptation in this kind of scenario.

Although regression-based domain adaptation methods have been developed to deal with problems that traditional domain adaptation methods could not solve, one problem still persists: current methods aimed at solving this problem work only on very specific regression problems and do not deal well with other types of tasks.

Some regression-based techniques were considered during the development of this work, among them, we can cite Representation Subspace Distance (RSD) (CHEN et al., 2021) and Regressive Domain Adaptation for Unsupervised Keypoint Detection (RegDA) (JIANG et al., 2021), however, none of them turned out to be suitable for the problem of 3D pose estimation based on 2D poses. The RSD method encounters a numerical problem during the optimization of this specific problem, meanwhile, the RegDA technique requires the use of ground-false data, which is hard to obtain efficiently in this particular case.

Therefore, our challenge is to propose a method capable of operating with regression to solve the following problem:

Problem 1 (Domain Adaptative 3D Human Pose Estimation)**.** *Given a source domain $\mathcal{D}_\mathcal{S}$ composed of a set of poses X, Y and an unsupervised target domain $\mathcal{D}_\mathcal{T}$ consisting of a set of pose annotations X, with distinct marginal probabilities $\mathcal{D}_\mathcal{S} \neq \mathcal{D}_\mathcal{T}$, the goal is to find a feature map θ and a pose estimator head $\mathcal{P}$ such that the conditional probability $\mathbb{P}(\theta(X_\mathcal{S})|X_\mathcal{S}) \approx \mathbb{P}(\theta(X_\mathcal{T})|X_\mathcal{T})$ without negatively affecting the efficacy of the pose regressor head $\mathcal{P}$.*

4 Related Work

In the previous chapters, we presented a literature review and discussed earlier works that are related to the human pose estimation and domain adaptation in a more general sense. In this chapter, we present works that are more closely related to the focus of this dissertation, since they either deal with pose representation issues or are used in a cross-domain evaluation scenario. We also discuss some works that inspired us while solving the problem.

4.1 3D Human Pose Estimation in Cross-Domain Scenarios

The idea of using domain adaptation to 3D Human Pose Estimation has been discussed before. Zhang et al. (2019), for example, proposed a method in which a synthetic depth-based dataset is used for domain adaptation during the learning step. However, the idea of evaluating 3D Human Pose Estimation in a Cross-Domain scenario is still not discussed by them.

Recent works started to notice the discrepancy in performance between data obtained from distinct distributions. To deal with this issue, several approaches have been proposed, such as that of Dabral et al. (2018), in which synthetic pose datasets artificially generated were used to enhance the amount of data available during the training. Other authors also followed this data augmentation paradigm by using Generative Adversarial frameworks (YANG et al., 2018) or a Conditional Variational Auto Encoder (CVAE), aiming to generate poses from another dataset distribution (JIANG et al., 2021).

The expansion of the training set through data augmentation is further discussed by recent works aimed at working directly in cross-dataset scenarios, in which the discrepancy in performance is even more noticeable. One such work introduced augmentation by adjustment of distinct geometric factors through a joint optimization algorithm trained online (GONG et al., 2021).

Gholami et al. (2022) addressed the domain gap caused by the cross-dataset evaluation through the weakly-supervised generation of synthetic 3D motions. In this way, the target distribution could be represented only by looking at the 2D poses, working both as a pose estimation technique and as a synthetic pose generator. A distinct approach that also employs synthetically generated poses, focused on alleviating the domain shift jointly through feature spaces and pose spaces using semantic awareness and skeletal pose adaptation.

The idea of directly using domain adaptation techniques to approach this problem has been discussed in previous works. One such work (GUAN et al., 2021) utilized the Skinned Multi-Person Linear (SMPL) model and proposed a method called Bilevel Online Adaptation to reconstruct mesh and pose, through a multi-objective optimization problem using temporal

constraints to deal with the domain discrepancy.

Chai et al. (2023), on the other hand, observed that most of the distribution discrepancy of cross-dataset evaluation stems from camera parameters and the diversity of local structures during training. Thus, they employed domain adaptation by combining a global position alignment mechanism, aiming to eliminate the viewpoint inconsistency, and a local pose augmentation used to enhance the diversity of the available poses.

The approach proposed by Kundu et al. (2022) introduced the usage of uncertainty mechanisms to work with self-supervised 3D human pose estimation. This operated in such a way that by minimizing the uncertainty for the unsupervised real dataset alongside a supervised synthetic dataset, it is possible to perform cross-dataset pose adaptation. Zhang et al. (2021), on the other hand, proposed a method for learning casual representations in order to generate out-of-distribution features that can properly generalize to unseen domains, and in order to compare the efficacy of their method, they compare it to previously established domain adaptation techniques, such as DDC (TZENG et al., 2014), DAN (LONG et al., 2018), and DANN (GANIN et al., 2017).

Some works tried to solve the problem of pose misrepresentation which is also found in the literature regarding cross-dataset evaluation. The work of Rapczyński et al. (2021) aimed to solve this issue through virtual camera augmentation, and a joint-harmonization mechanism, supported by scale normalizatoin. The harmonization mechanism was used to ensure that joints were represented in the same position across all datasets, meanwhile, normalizing the scale ensures that all the limbs of the subjects have the same proportion. The approach of Sárándi et al. (2023), instead, involved using an autoencoder to learn a set of latent key points that can properly represent all of the distinct datasets across the same embedding.

5 Proposed Method

This Chapter, introduces the Domain-Unified Approach method, a novel solution we propose for addressing the challenges discussed in previous Chapters.

Even with various methods being proposed to tackle specific aspects of the problem of 3D pose estimation from monocular RGB images, a comprehensive approach for addressing cross-dataset human pose estimation remains lacking. Considering the limitations of existing methods in addressing the multifaceted challenges posed by domain discrepancy, often focusing on addressing specific aspects of the domain discrepancy problem, in this work we introduce a novel method, called Domain-Unified Approach (DUA), which tackles this issue from a unified perspective combining domain adaptation techniques with a universal pose representation and a specialized training technique to mitigate error propagation at the edges. This chapter presents the DUA method.

5.1 DUA - Domain Unified Approach

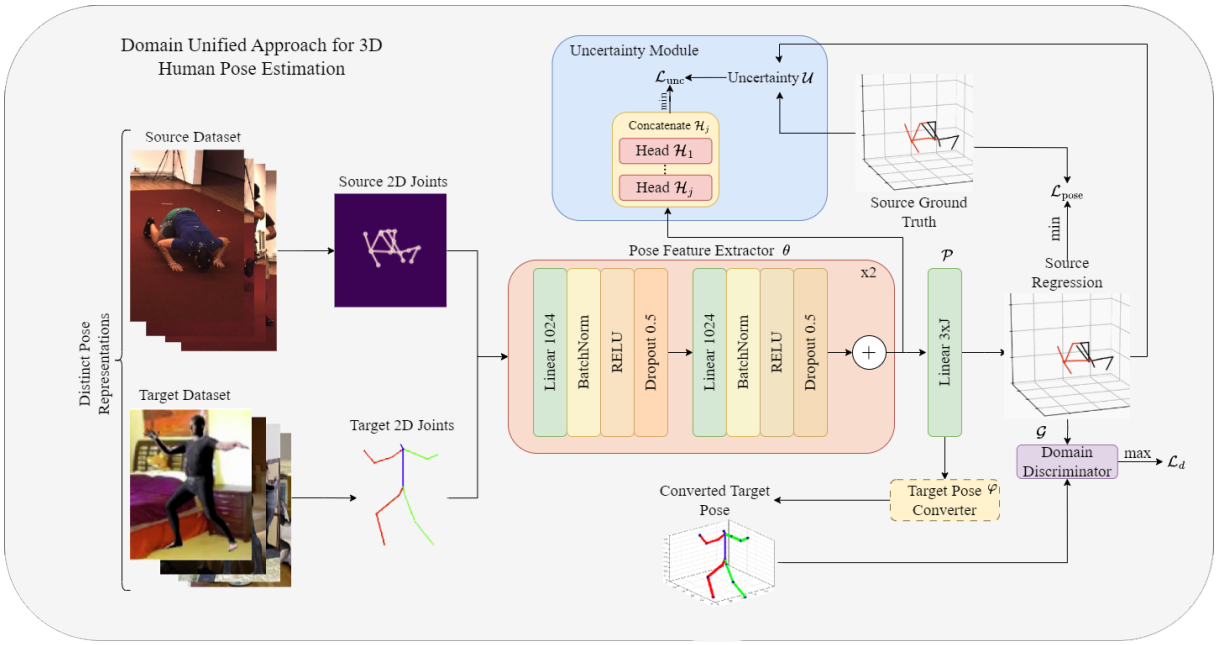
In order to address the **Problem 1**, discussed in Section 3.3, we proposed DUA (Domain Unified Approach), a method capable of accurately inferring poses from source and target domains with minimal error, by combining a pose conversion unit, presented in Section 5.2, and the uncertainty loss mechanism, described in Section 5.3, with a domain adaptation module. The general idea of DUA is to maximize the distance between 3D human poses using a domain discriminator that is jointly optimized with the entire deep learning system. Based on the proposed taxonomy discussed in Section 2.4, this method can be categorized as a hybrid approach that combines image-based regression and pre-processing refinement.

Figure 19 depicts the DUA method, which is structured around three main modules, all operating on top of a backbone pose estimator. Initially, the pose estimator serves as a feature extractor, from which one can obtain poses from a dedicated pose head $\mathcal{P}$. These extracted features are fused with pose predictions to generate an uncertainty estimate, aiding the training process. Furthermore, the predicted target-domain poses undergo transformation into a unified pose representation, harmonizing joint distribution with the source domain. Lastly, a domain discriminator is employed, tasked with distinguishing between source and target poses, in order to establish a consistent feature representation within a common domain.

The DUA method has an architecture inspired by the DANN architecture. To find the desired pose, given a pose estimator Π, the following pose loss is used:

$$\mathcal{L}_{\text{pose}}(x) = \beta(y - \Pi(x))^2 + (1 - \beta)\|y - \Pi(x)\|, \tag{4}$$

Figure 19 – Proposed Domain Unified approach for 3D human pose estimation. The method is composed of three main modules on top of the 3D pose estimator: the unified pose representation module, the uncertainty estimation module, and the domain discriminator. The dashed lines on the pose converter represent frozen weights.



Source: The Author.

where $0 \leq \beta \leq 1$ is a hyperparameter that controls the importance of each part of the pose loss.

The pose estimator is engaged in a minimax game, aiming to minimize $\mathcal{L}_{\text{pose}}$, while simultaneously maximizing the domain discrepancy of the joints to find the optimal representation from the pose feature extractor θ. This is achieved using a domain classifier G, trained with the loss:

$$\mathcal{L}_d(x) = G(\theta(x)) \log(G(\theta(x))) + (1 - G(\theta(x)))log(1 - G(\theta(x))). \tag{5}$$

In DUA method, the unified pose representation obtained by the converter is pre-trained, and its weights remain frozen during training. On the other hand, the other components of the method are trained in an online mode. The overall training loss is given by:

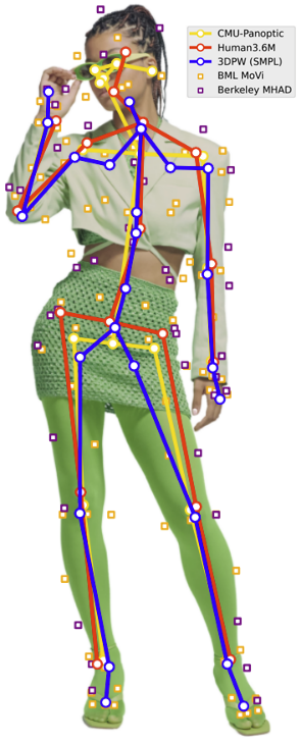
$$\mathcal{L} = \lambda\mathcal{L}_d + \gamma\mathcal{L}_{\text{unc}} + \mathcal{L}_{\text{pose}}, \tag{6}$$

where $0 < \lambda < 1$ and $0 < \gamma < 1$ are regularization parameters.

5.2 Unified Pose Representation

The incompatibility between the pose representations is a commonly observed problem in 3D human pose estimation when dealing with different datasets. Previous works have already

Figure 20 – Five distinct pose representations used by common 3D Human Pose datasets found in the literature.



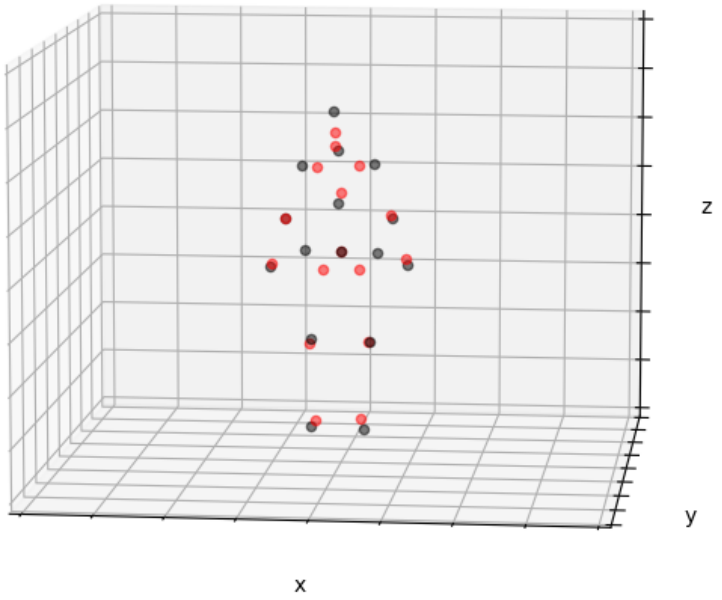
Source: Sárándi et al. (2023), © 2023 IEEE.

discussed this issue in the literature. One such work aims to learn unified representations by utilizing different data sources concurrently (SÁRÁNDI et al., 2023). This problem arises from the existence of various body capture sensors and different 3D pose representations being used in the literature leading each dataset to have its own representation. Figure 20 shows five distinct 3D pose representations found in the literature.

This problem was previously addressed in the task of pose shape estimation, by the creation of the Archive of Motion Capture As Surface Shapes (AMASS) (MAHMOOD et al., 2019). This represents a large and varied database of human motion that unifies 15 different optical marker-based datasets through the lenses of the SMPL (Skinned Multi-Person Linear) representation. This is done through the usage of the Motion and Shape capture (MoSh) technique, aimed at estimating body shape SMPL parameters given the 3D pose data (LOPER et al., 2014).

An example of the moshed SMPL representation of the 3D pose found in the Human3.6M dataset is found in Figure 21. In this figure, the SMPL representation (red) is juxtaposed with the H3.6M representation (black) to show their differences, with the biggest impact being on the hips and the head. Therefore, to address this problem in the pose representations, our approach aims to train a pose converter to transform the 3D human pose to the singular pose representation using data obtained from both the SMPL and original Human3.6M representations.

Figure 21 – Overlapped joints of the Human3.6M dataset coming from two distinct pose representations, SMPL (red) and the original H3.6M format (black). This makes explicit the difference in the pose representations being used by common 3D human pose datasets in the literature.



Source: The Author.

This conversion problem was already discussed in the literature (RAPCZYŃSKI et al., 2021). However, previous approaches tried to find a harmonization and normalization technique through handcrafted features. This approach works well in some cases but does not preserve the body proportions after normalization. Thus, we developed a pose converter to dynamically learn how to convert from one pose representation to another. The idea of our converter network is to dynamically find an array, based on the network weights and the 3D pose input, upon which when adding this array to a pose in representation $\mathcal{A}$, a pose in an arbitrary format $\mathcal{B}$ is found.

In mathematical terms, the mapping function $\Phi : \mathcal{A} \mapsto \mathcal{B}$ takes a set of joints $X_{\mathcal{A}}$ represented in pose format $\mathcal{A}$ and calculates weights to map $X_{\mathcal{A}}$ to a representation $X_{\mathcal{B}}$ in the pose format $\mathcal{B}$.

Instead of directly mapping $\mathcal{A}$ to $\mathcal{B}$, the task of converting between representation spaces of the same semantic skeleton graph involves finding trajectory vectors that describe the new joint positions and their trajectories in the new pose space. To simplify this process, we work directly with the joint trajectory vectors by introducing a mapping function $\varphi : \mathcal{A} \mapsto (\mathcal{B} - \mathcal{A})$, in such a way that:

$$X_{\mathcal{B}} = \varphi(X_{\mathcal{A}}) + X_{\mathcal{A}}. \tag{7}$$

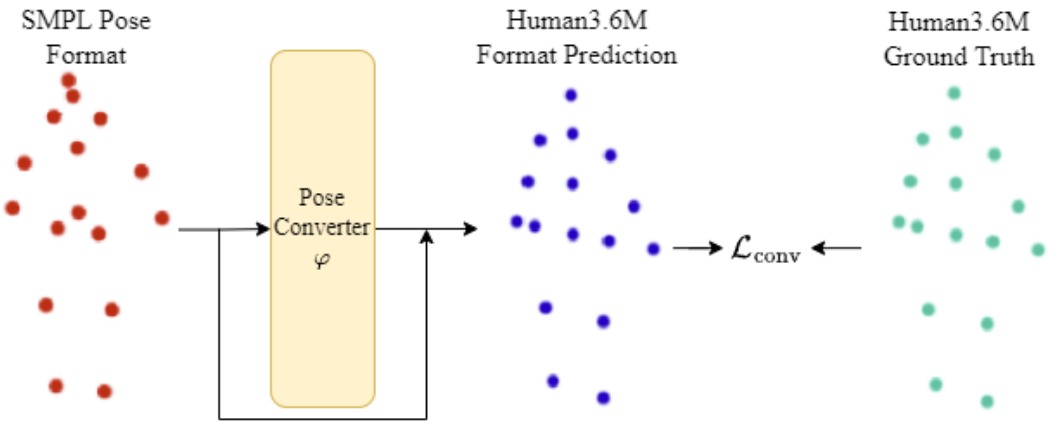
The weights of the mapping function φ are obtained through a single-layer residual neural network using gradient descent. To train this network, a loss function combining mean

squared error and mean average error is employed. The loss is given by:

$$\mathcal{L}_{\text{conv}} = \alpha(X_{\mathcal{B}} - X_{\mathcal{A}})^2 + (1-\alpha)\|X_{\mathcal{B}} - X_{\mathcal{A}}\|, \tag{8}$$

where $0 \leq \alpha \leq 1$ is a hyperparameter used to impose the importance of each loss term. Figure 22 illustrates the learning process of the proposed method.

Figure 22 – Pose conversion method used to find a unified pose representation.



Source: The Author.

5.3 Pose Uncertainty

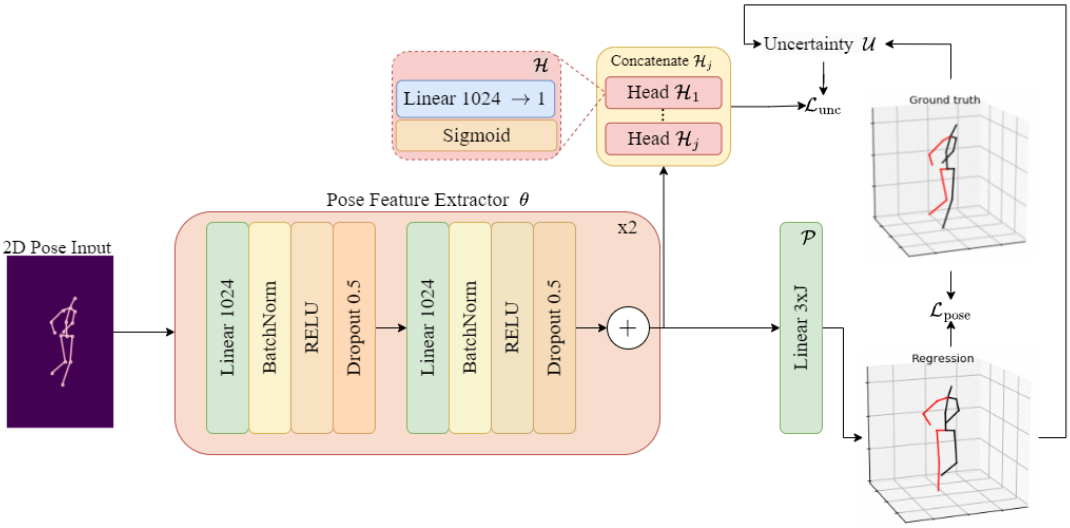
The issue of 3D human pose estimation presents a challenge in the form of error propagation within the most extreme kinematic group, compounded by the ill-defined monocular estimation resulting from self-occlusion during varying camera perspectives. In order to mitigate this problem, an approach has been devised to quantify and reduce the uncertainties arising from such scenarios.

Uncertainty in Bayesian networks has been defined in two forms: epistemic uncertainty captures the model's ignorance despite sufficient training data with well-defined data distributions, while aleatory uncertainty aims to model unexplained uncertainties within the current training data (KENDALL et al., 2018). Previous works have explored uncertainty modeling through Bayesian networks for 3D human pose estimation using different approaches (LI et al., 2023; KUNDU et al., 2022). In this work, we propose a method based on a naive definition of uncertainty.

To quantify uncertainty, our method utilizes the features extracted from the pose estimator to predict the probability of a joint being incorrect. A random variable $\mathcal{U}$ is generated by mapping the normalized Euclidean distance of the joint difference, where joints with small distances are mapped near zero and those with significant distances are mapped to one. This mapping allows for improved assessment and quantification of uncertainty associated with

individual joints. Figure 23 illustrates our devised approach using the method proposed by Martinez et al. (2017) as the backbone.

Figure 23 – Uncertainty-based 3D human pose estimation method devised using Martinez et al. (2017) as backbone.



Source: The Author.

Our method consists of J heads, each representing one joint in the pose representation. After passing through the sigmoid activation function, each head provides the probability of a specific joint being incorrect. This probability is learned through supervised training by comparing the output to the normalized Euclidean distance. The Uncertainty Error is calculated as the L1 distance between the array composed of the heads and the normalized distance.

Given the outputs of a pose feature extractor θ for an input x, inserted into each of the J heads $\mathcal{H}_j$, and using a pose estimator head $\mathcal{P}$ to obtain the output 3D pose, we define $\mathcal{U}(x)$ as the desired pose uncertainty obtained by concatenating each head $\mathcal{H}_j$. In other words, using the concatenation operation denoted by $||$ and the sigmoid function σ, we have:

$$\mathcal{U}(x) = ||\sigma(\mathcal{H}_j(\theta(x))). \tag{9}$$

By combining the pose feature extractor and the pose estimator head through the function $\Pi(x) = P(\theta(x))$, where P represents the pose estimator, the uncertainty $\mathcal{U}(x)$ of a given pose x can be learned. This is achieved by comparing the normalized Euclidean distance of each joint in the predicted pose $\Pi(x)$ to the ground-truth 3D pose y, and comparing it to the output uncertainty using L1 distance. The loss function for training is defined as:

$$\mathcal{L}_{\text{unc}}(x) = \left\| \frac{(\sqrt{\Pi(x) - y})^2}{\|\Pi(x)\|_2 \|y\|_2} - \mathcal{U}(x) \right\|. \tag{10}$$

6 Experimental Settings

This chapter aims to present the experimental settings used in this work in order to conduct the experiments, including the hardware configuration, the hyperparameters values, and the training and evaluation protocols. This chapter also presents the metrics and datasets utilized to assess the DUA method, proposed in this dissertation for 3D human pose estimation based on monocular RGB images and domain adaptation.

6.1 Hardware Configuration

All experiments were conducted using a computer with two Intel Xeon E5620 CPUs, 48GB of RAM, and an NVIDIA TitanXP GPU with 12GB of VRAM.

6.2 Hyperparameter Values

During training, a batch size of 2048 was employed, with a learning rate of $1e^{-3}$ paired with the Adam optimizer. For hyperparameters, $\alpha = 0.5$ were employed in the pose conversion scenario, for the pose estimator, $\lambda = 0.01$, $\gamma = 0.1$, and $\beta = 0.4$ were chosen via empiric evaluation. The pose conversion was pre-trained and its weights were frozen on the DUA method.

6.3 Preprocessing

Prior to training, a preprocessing step was applied to center all the poses with their hip joint on the origin of the coordinate system. Additionally, the 2D joint coordinates were normalized to fit within a $[-1, 1]$ coordinate system by scaling the image input space. The 3D joint coordinates followed the standard protocol established by Martinez et al. (2017), where they were transformed into the camera coordinate system. This transformation involves centering the pose around the camera's optical center, and is achieved by applying the inverse rotation and translation of the camera position to the pose joints, using the extrinsic camera matrixes from the training data.

6.4 Evaluation Protocol

Cross-dataset evaluation in 3D human pose estimation shows a significant challenge due to the inherent misalignment of target distributions, especially when synthetic data is involved. The scarcity of literature addressing this specific scenario has motivated only a few authors to

explore the evaluation protocol for assessing cross-domain generalization when synthetic data is involved (KUNDU et al., 2022; ZHANG et al., 2021).

In this work, our focus lies on evaluating the performance of synthetic to real cross-domain pose estimation. Building upon previous works, we adopt a widely used general domain adaptation evaluation protocol to assess the effectiveness of domain generalization. Specifically, we employ an unsupervised training approach using the target dataset training split, while utilizing the supervised source data for training. The evaluation is conducted using both synthetic and real datasets as source data.

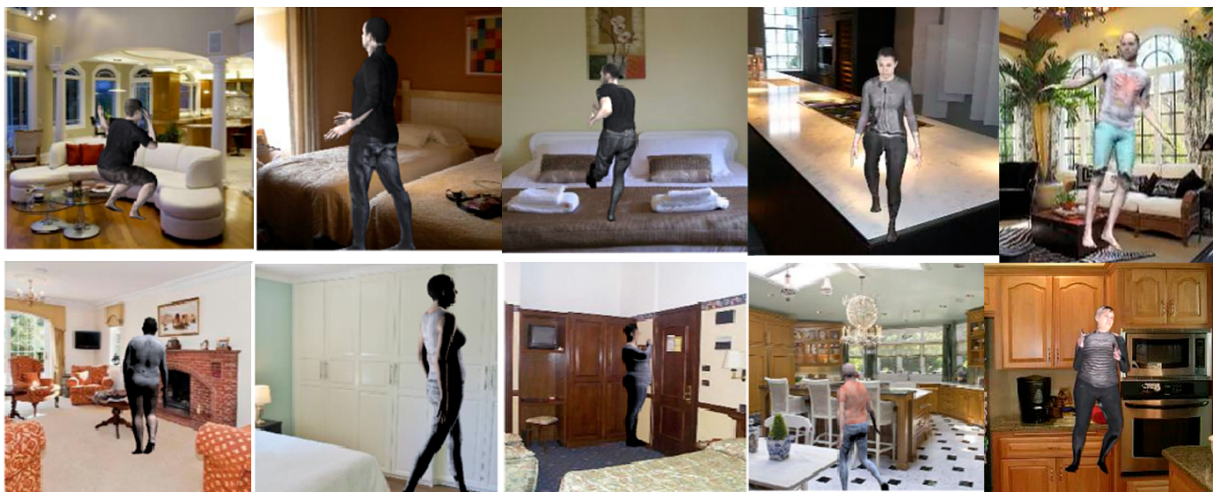
For the purpose of comparison, we adopt the unified pose representation of the Human3.6M model as our baseline, with the pose converter trained to transform the SMPL pose representation of the dataset into the Human3.6M representation. By employing this unified pose representation, we aim to facilitate meaningful comparisons with existing approaches.

6.5 Datasets

We employed two datasets in our cross-domain experiments: SURREAL (VAROL et al., 2017) and Human3.6M (MARCARD et al., 2018). In particular, the SURREAL dataset was used to represent the synthetic image domain, while the Human3.6M dataset was used to represent the real people image domain.

The SURREAL is a large-scale dataset containing more than 6 million photorealistic synthetic image frames found in real environments with large variations in texture, body time, camera positioning, and pose actions. The database contains information about the depth map, body parts, optical flow, and 2D and 3D joints. Figure 24 shows examples of images, environments, and actions found in the SURREAL dataset

Figure 24 – Example images found in the SURREAL dataset.



Source: The Author.

The Human3.6M database is composed of real people images obtained from a Motion Capture system based on markers, it contains scenes of 11 professional actors obtained in a controlled environment. The database has about 3.6 million annotations of 3D poses, considering four different angles. This dataset also has three evaluation protocols with different data for training and testing. Figure 25 shows some examples of images present in the Human3.6M dataset, showcasing the environment, some of the actions, and some actors present in the dataset.

Figure 25 – Example images found in the Human3.6M dataset.

Source: The Author.

6.6 Metrics

The methods proposed and developed in this work were evaluated by the standard metrics applied in the literature in the 3D human pose estimation problem, in order to provide a comparison with previous works dealing with the same task the following metrics were used:

MPJPE: The Mean Per-Joint Position Error (MPJPE) represents the mean error, in millimeters, between the estimated points and the real points after the root joint alignment;

P-MPJPE: The Procrustes-Aligned Mean Per-Joint Position Error (P-MPJPE) represents the mean error, in millimeters, between the estimated points and the real points after Procrustes alignment.

7 Results and Discussion

In this chapter, we present the results obtained by DUA (Domain Unified Approach) method, in the experiments conducted according to the settings detailed in Chapter 6.

DUA consists of three essential modules: pose conversion, pose uncertainty, and a domain-unified approach, as established in Section 5.1. In order to provide a comprehensive analysis, we show the results of each module in its respective section. By evaluating the performance of each module individually, we gain insights into its effectiveness and contribution toward accurate and robust pose estimation.

7.1 Unified Pose Representation

The pose conversion method underwent training for 100 epochs, and the converted poses were evaluated using the MPJPE (Protocol 1) and P-MPJPE (Protocol 2) metrics on the Human3.6M dataset. We conducted evaluations in two scenarios: by action and by joints. The results, shown in Tables 1 and 2, respectively, clearly demonstrate the significant error reduction achieved by the conversion method when compared to SMPL poses without conversion. This reduction in error is particularly impactful when predicting poses from different domains.

Table 1 – MPJPE and P-MPJPE measures, by groups of actions, of the SMPL model without pose conversion (right) and with pose conversion (left), when overlapped to the original Human3.6M pose. Error-values in millimetres (mm) - the lower the better.

Action	**MPJPE Without Conversion**	**P-MPJPE Without Conversion**	**MPJPE With Conversion**	**P-MPJPE With Conversion**
Direct.	63.84	60.66	28.95	21.81
Discuss	69.74	60.04	35.19	26.59
Eating	70.95	60.04	33.72	26.07
Greet	63.18	59.99	28.90	23.13
Phone	71.20	61.81	36.77	29.83
Photo	68.89	62.11	32.77	26.06
Pose	63.29	60.25	29.68	23.77
Purch.	72.37	60.47	34.97	24.33
Sitting	77.79	62.63	38.93	31.20
SittingD.	81.35	64.39	42.91	32.94
Smoke	71.38	62.59	34.29	27.42
Wait	66.40	60.06	33.23	26.12
WalkD.	69.82	60.01	35.47	26.16
Walk	62.24	59.10	30.51	24.24
WalkT.	61.74	59.15	29.05	23.86
Avg.	68.95	60.89	33.70	26.24

The results presented in Table 1 demonstrate a notable reduction in the mean per-joint position error (MPJPE) across all action groups when comparing the converted poses with

their original counterparts. This significant decrease highlights the detrimental effects of pose misrepresentation and the positive impact of utilizing a properly converted pose through our method. By effectively converting the poses to a more appropriate representation, our method successfully mitigates the negative effects of misrepresentation, leading to improved pose estimation performance.

Table 2 – MPJPE and P-MPJPE measures, per joints, of the SMPL model without pose conversion (right) and with pose conversion (left), when overlapped to the original Human3.6M pose. Error values in millimetres (mm) - the lower the better.

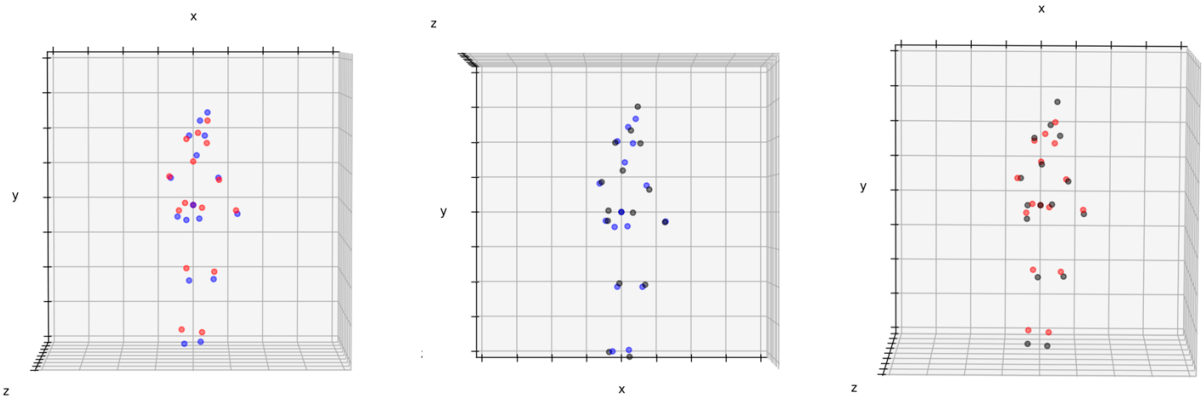
Joint	MPJPE Without Conversion	P-MPJPE Without Conversion	MPJPE With Conversion	P-MPJPE With Conversion
Hip	00.00	20.89	00.00	18.09
Right Hip	125.03	122.87	14.81	26.84
Right Knee	44.73	44.61	34.72	29.37
Right Foot	47.09	29.11	41.28	28.71
Left Hip	104.66	104.26	18.62	19.88
Left Knee	49.04	47.62	39.96	34.10
Left Foot	63.97	46.92	49.60	31.48
Spine	67.05	89.22	28.48	16.83
Thorax	71.94	62.84	35.18	24.20
Head	118.46	99.80	54.31	44.69
Left Shoulder	106.59	88.95	39.91	31.98
Left Elbow	60.27	35.68	41.93	23.66
Left Wrist	44.09	28.21	27.07	15.80
Right Shoulder	105.49	90.97	43.43	31.86
Right Elbow	55.11	37.31	42.07	26.76
Right Wrist	39.19	28.52	29.89	17.91

Furthermore, the results presented in Table 2 demonstrate the effectiveness of our method in achieving a proper pose representation across all individual joints. The observed reduction in error for each joint indicates the successful correction of mispositioning issues and highlights the ability of our approach to mitigate error propagation from previous representations. The largest errors per joint, previously contained within misrepresented poses are now dislocated to edge joints, which aligns with the expected behavior in normal pose estimation.

In Figure 26, we provide visual comparisons to illustrate the effectiveness of our pose conversion method. On the left, a Human3.6M pose (depicted by red dots) is superimposed on the SMPL pose without conversion (depicted by blue dots). In the middle, the resulting conversion (depicted by black dots) is superimposed on the original SMPL pose (blue dots). On the right, we compare the Human3.6M pose (red dots) to the SMPL pose converted to the Human3.6M format using our method (black dots). It can be noted that the conversion of the SMPL pose to the Human3.6M pose model helps to reduce errors, particularly in the hip area, thus improving the overall pose estimation accuracy and alignment with the target pose format.

It is evident from the visual comparisons that the pose conversion successfully aligns the SMPL poses with the target format, effectively capturing the key joint positions and

Figure 26 – Overlapping of the different pose representations available, Human3.6M (red), SMPL ground-truth (blue), converted pose (black). Item (a) shows a Human3.6M (red dots) pose superimposed on the original SMPL pose (blue dots). Item (b) shows the resulting pose after conversion (black dots) superimposed to the original SMPL pose (blue dots). Item (c) shows the converted pose (black dots) superimposed to the original Human3.6m pose (red dots).



(a) (b) (c)

Source: The Author.

maintaining the overall pose structure. The converted poses exhibit improved similarity to the ground truth poses, indicating the accuracy and fidelity of the conversion process. This qualitative analysis further corroborates the quantitative results, demonstrating the efficacy of the pose conversion method in achieving a more accurate and compatible representation for 3D human pose estimation tasks.

7.2 Pose Uncertainty

The proposed approach was trained for 200 epochs for each representation. The results obtained are presented in Table 3, where a double horizontal line separates the results obtained from the 2D ground truth and the 2D detections. Results were evaluated in two scenarios, using a linear backbone and a graph-based backbone, and with three types of 2D input information: ground truth, stacked hourglass, and a more robust 2D pose extractor based on cascaded pyramid networks.

The experimental results demonstrate that our approach surpasses the performance of their respective backbones, providing better results in almost all the action groups, and effectively mitigating the performance gap between graph-based and linear methods. However, it is important to note that our method is still dependent on the choice of a backbone architecture, as it inherits and improves upon the errors already present in certain action groups from the backbone. In some cases, the original backbones still outperformed our method, but on average, our approach resulted in lower errors. Moreover, we observed that selecting a more accurate 2D

Table 3 – MPJPE measures obtained from the 3D Human Pose considering all of the established scenarios with distinct 2D Pose sources: Stacked Hourglass (HG) (denoted by $*$), Cascaded Pyramid Networks (CPN) (denoted by $\dagger$), and Ground Truth (GT) (denoted by $\ddagger$). Experiments were performed using a linear backbone (indicated by $\star$) and a graph-based backbone (indicated by $\S$). The best results are presented in bold. Error-values in millimeters (mm) - the lower the better.

	Direct.	**Discuss**	**Eating**	**Greet**	**Phone**	**Photo**	**Pose**	**Purch.**
Martinez et al. (2017)*	51.8	56.2	58.1	59.0	69.5	78.4	55.2	58.1
Zhao et al. (2019)*	48.2	60.8	51.8	64.0	64.6	**53.6**	51.1	67.4
Ours*	50.9	54.5	56.0	57.0	65.7	72.6	52.6	54.5
Ours*	50.1	56.1	58.0	56.7	63.5	73.1	52.4	54.4
Ours†*	**46.2**	**50.4**	**51.1**	52.7	57.3	67.8	**50.5**	**48.2**
Ours†§	46.4	52.1	51.8	**52.2**	**56.5**	67.0	51.5	49.0
Zhao et al. (2019)‡	37.8	49.4	37.6	40.9	45.1	**41.4**	**40.1**	48.3
Ours‡*	**34.8**	**42.4**	35.6	**39.9**	42.2	50.3	42.6	**37.0**
Ours‡§	41.0	46.2	**34.4**	**39.9**	**38.1**	46.6	44.3	39.1

	Sitting	**SittingD.**	**Smoke**	**Wait**	**WalkD.**	**Walk**	**WalkT.**	**Avg.**
Martinez et al. (2017)*	74.0	94.6	62.3	59.1	65.1	49.5	52.4	62.9
Zhao et al. (2019)*	88.7	**57.7**	73.2	65.6	48.9	64.8	51.9	60.8
Ours*	70.2	89.4	59.7	57.4	62.9	47.3	51.5	60.2
Ours*	71.1	93.5	60.2	57.8	61.7	47.8	51.9	60.6
Ours†*	**64.1**	72.6	53.7	**51.3**	57.2	41.8	46.7	**54.1**
Ours†§	64.7	73.7	**53.3**	51.4	**55.1**	**41.5**	**45.2**	**54.1**
Zhao et al. (2019)‡	50.1	**42.2**	53.5	44.3	40.5	47.3	39.0	43.8
Ours‡*	49.5	54.0	**40.3**	**41.6**	42.6	32.1	35.0	41.3
Ours‡§	**46.0**	52.4	39.5	43.5	**40.4**	**31.5**	**34.2**	**41.1**

Human Pose estimator significantly enhanced the performance of our method, indicating the importance of leveraging reliable 2D pose information in the overall pose estimation process.

Overall, our results validate the effectiveness of incorporating pose uncertainty estimation into the 3D human pose estimation pipeline, leading to enhanced accuracy and robustness in capturing human poses across different action categories.

7.3 Domain Unified Approach

DUA aims to unify previously proposed approaches by combining the pre-trained pose converter (with frozen weights), the pose uncertainty module, and the domain adaptation protocol. The domain adaptation method was trained online for 300 epochs, and an evaluation was conducted using both the SURREAL and Human3.6M datasets as source data. Results of the domain adaptation method are presented in Table 4.

Notably, the utilization of domain adaptation significantly mitigates the problem caused by domain discrepancy, when evaluating the most practical scenario, of training with a huge synthetic dataset and applying it to a real-world scenario (SURREAL $\rightarrow$ Human3.6M), our method leads to a reduction of 44.1 mm in the mean per-joint position error (MPJPE). This

improvement brings our results closer to the state of the art in the field. Additionally, we included SURREAL data without conversion only in the domain adaptation scenario to compare and evaluate the effectiveness of the converted pose in the overall method results.

Table 4 – Results from the MPJPE metric (mm – the lower the better) obtained from different domain adaptation scenarios.

Source Dataset	**Target Dataset**					
	SURREAL		**Converted SURREAL**		**Human3.6M**	
	MPJPE	P-MPJPE	MPJPE	P-MPJPE	MPJPE	P-MPJPE
Human3.6M (No DA)	107.6 mm	65.7 mm	100.3 mm	62.6 mm	41.3 mm	32.7 mm
SURREAL (No DA)	-	-	40.4 mm	29.0 mm	150.8 mm	88.2 mm
Human3.6M + DUA	108.9 mm	70.1 mm	96.1 mm	57.9 mm	74.0 mm	52.2 mm
SURREAL + DUA	-	-	55.8 mm	37.7 mm	106.7 mm	65.6 mm

The impact of employing pose conversions is evident in the Human3.6M $\rightarrow$ SURREAL scenario, where the evaluation is conducted on a larger dataset comprising a distinct subset of actions. In this challenging scenario, the performance of our method with domain adaptation experiences a slight decrease with domain adaptation. This can be attributed to the misrepresentation of poses, which poses a difficulty for the optimization of our domain discriminator. However, by incorporating pose conversion into our method, we are able to effectively mitigate this discrepancy and observe a performance improvement.

Furthermore, when compared to other state of the art methods employed in the real-to-synthetic cross-dataset scenario (as shown in Table 5), our method outperforms all previously proposed techniques. One significant advantage of our approach is the use of a universal pose representation, wherein the conversion step mitigates issues arising from variations in body capture sensors across different datasets.

Table 5 – Quantitative results obtained on the H3.6M $\rightarrow$ SURREAL evaluation. Table results and layout are obtained from experiments conducted by Zhang et al. (2021) and Kundu et al. (2022), bold indicates the best result.

Method	**H3.6M $\rightarrow$ SURREAL**	
	MPJPE	**P-MPJPE**
DDC (TZENG et al., 2014)	117.5	80.1
DAN (LONG et al., 2018)	114.2	78.4
DANN (GANIN et al., 2017)	113.6	77.2
Zhang et al. (2021)	103.3	69.1
Kundu et al. (2022)	99.6	67.2
Kundu et al. (2022)*	99.4	65.1
DUA (Ours)	**96.1**	**57.9**

* Denotes an alternative approach for cross-domain evaluation conducted using test-time adaptation.

We also observed that the pose conversion step substantially impacts our method's overall performance. This finding highlights the importance of finding a unified pose representation to mitigate domain discrepancy. In order to make that clear, ablation results are

made explicit in Table 6. This table effectively illustrates the importance of each component of our method. Specifically, the Unified Pose Representation emerges as the primary contributor to addressing the problem. Conversely, adopting domain adaptation in isolation, without a common representation causes a negative impact on the results. Interestingly, no such adverse influence is observed when solely employing the shared pose representation. However, this negative impact is not seen when both methods are combined, with a clear enhancement in results becoming evident.

Table 6 – Explicit ablation results of our method evaluated on the Human3.6M → SURREAL evaluation setting.

Method	MPJPE
w/o Unified Pose Representation and DA	107.6 mm
w/o Unified Pose Representation	108.9 mm
w/o DA	100.3 mm
DUA	96.1 mm

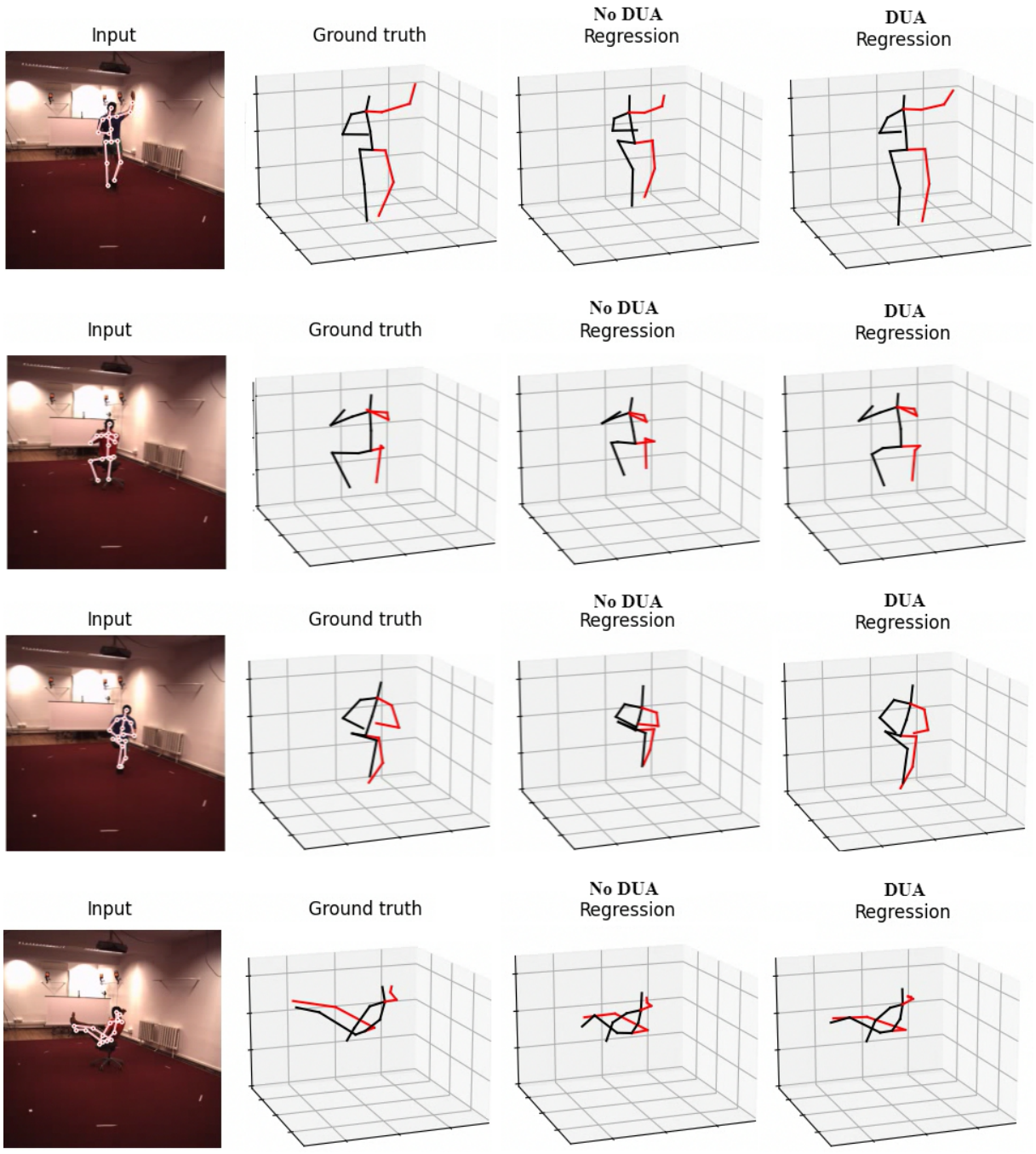
Qualitative results presented in Figure 27, clearly demonstrate the consequences of not applying domain adaptation, resulting in significant distortions in certain pose regions, including body proportion loss and mispositioning of joints, such as the hips.

It is worth noting that the performance of the same-domain scenario exhibits a slight decrease in efficacy when employing domain adaptation techniques. This can be attributed to the objective of our approach, which aims to obtain a more generalized set of pose features through domain adaptation. By doing so, our method mitigates the risk of overfitting to specific actions or camera angles, as previously discussed in the literature (CHEN et al., 2021).

Although there is a small decrease in performance in the same-domain scenario, the overall benefit of achieving improved generalization and robustness across different domains outweighs this slight trade-off.

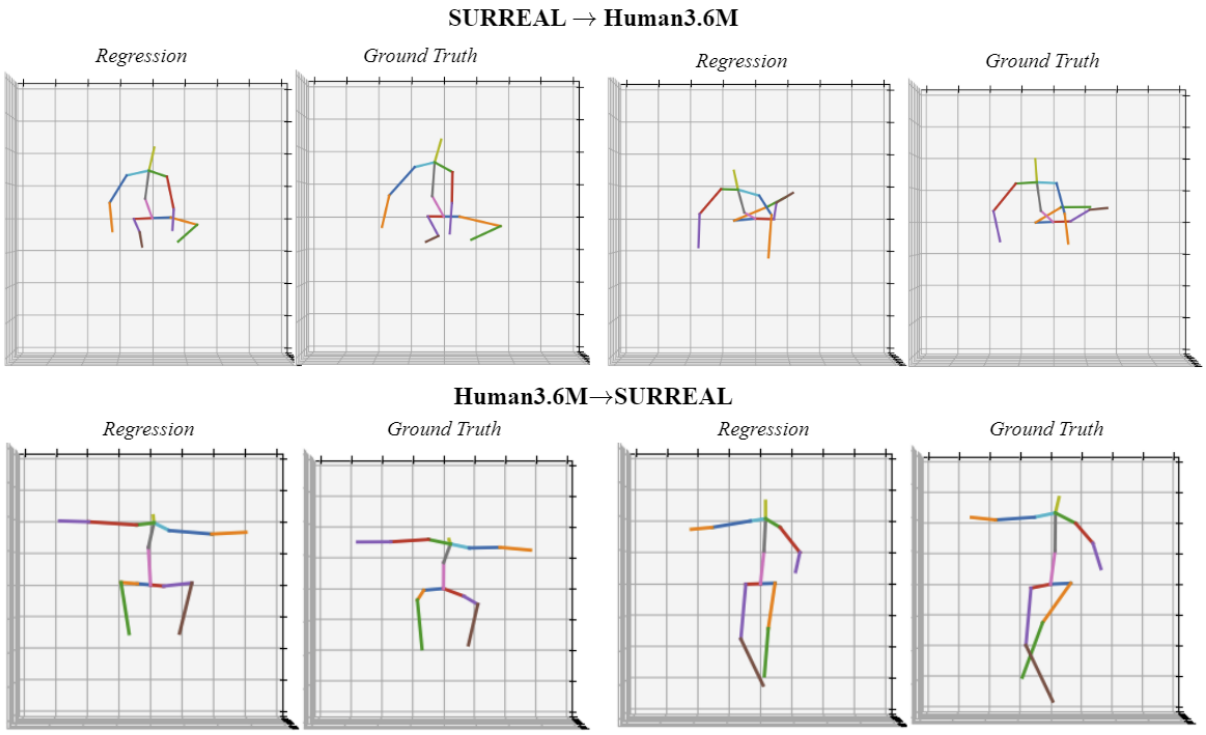
Finally, Figure 28 presents a qualitative assessment of the poses where the method exhibited the least favorable overall performance. As it is possible to notice, our method finds it difficult to deal with specific instances of leg self-occlusion, as well as how to deal with poses found on the target dataset without a proper correspondence on its source counterpart. Given the dissimilarity in pose space for actions between the two datasets, except for walking, it is difficult to learn certain aspects of unseen poses and this complexity carries over to the final outcome.

Figure 27 – Qualitative results obtained from our proposed approach on the Human3.6M dataset.



Source: The Author.

Figure 28 – Poses in which the method achieved the worst performance per scenario.



Source: The Author.

8 Conclusion and Future Work

In conclusion, our proposed technique has successfully addressed key challenges in 3D human pose estimation by integrating the pre-trained pose converter, the pose uncertainty module, and the domain adaptation protocol. Through the application of domain adaptation, we have effectively tackled the issue of domain discrepancy, leading to a remarkable reduction of 44.1 mm in mean per-joint position error MPJPE, when training with the synthetic dataset SURREAL and evaluating with Human3.6M. This substantial improvement achieves state-of-the-art performance in the field, highlighting the efficacy of our domain adaptation method.

By utilizing a universal pose representation and incorporating the pose conversion step, we effectively addressed challenges arising from variations in body capture sensors across different datasets. This capability enhances the adaptability and generalization of our approach, validating our proposed hypothesis and providing robust and accurate pose estimation results.

Although there is still room for further advancements in the field of 3D human pose estimation, such as improving the backbone estimation or enhancing the quality of uncertainty estimation, our method represents a significant step forward in improving the accuracy and robustness of pose estimation by the combination of our pose conversion, pose uncertainty estimation, and domain adaptation modules.

8.1 Contributions

In this work, we have proposed a new method, called DUA, to address the challenge of cross-domain 3D Human Pose Estimation. Our method adopts a Domain Unified Approach that comprehensively addresses multiple aspects of the problem, which albeit discussed in the literature, have been overlooked altogether. By integrating these perspectives, our method offers a cohesive solution to tackle three critical issues in 3D Human Pose Estimation: pose misrepresentation, propagation of errors at joint edges, and domain misalignment in cross-domain scenarios. Consequently, our contributions are as follows:

- The creation of a novel and effective technique for converting poses between two datasets, that achieves a Unified Pose Representation to overcome the observed differences between capture sensors, associated with the misrepresentation challenge;
- An enhancement and evaluation of the pose estimation training pipeline through a novel uncertainty-based method used to mitigate the propagation of errors at joint edges;
- The creation of a domain adaptation model based on adversarial networks for 3D human pose estimation, that unifies these solutions in order to mitigate domain-shift problems

and enhance the overall effectiveness of the approach in cross-domain settings.

In this work, we have also proposed a taxonomy to categorize the different two-step 3D pose estimation approaches found in the literature.

As far as we know, all these contributions are original.

8.2 Future Work

In this work, we propose a novel method for dealing with cross-domain 3D Human Pose estimation. While our method shows promising results, there are opportunities for future enhancements. One possible improvement is transitioning to a graph-based neural network as the backbone, since this neural network model has the potential to significantly boost the method's performance. Moreover, an area for investigation lies in adapting these methods to work with video-based approaches, exploiting temporal information to aid in the domain adaptation process. Additionally, adopting self-supervision could eliminate the need for correcting the misalignment in cross-dataset poses.

To further refine the uncertainty mechanism, we recommend exploring alternative formal approaches that enable more robust handling of uncertainty. Another avenue worth investigating is feature extraction of depth information, using the synthetic data acquired from the AMASS dataset SMPL models, with this, a shared space between keypoints and 3D depth information could be found, offering a more reliable solution to the non-invertibility challenge in the projection process. These advancements have the potential to elevate the overall performance and effectiveness of our proposed method.

8.3 Published Articles

The following papers, related to this work, have been published:

- MANESCO, João Renato Ribeiro; MARANA, Aparecido Nilceu. A Survey of Recent Advances on Two-Step 3D Human Pose Estimation. In: **11th Brazilian Conference on Intelligent Systems (BRACIS 2022)**. Cham: Springer International Publishing, 2022. p. 266-281.
- MANESCO, João Renato Ribeiro; BERRETTI, Stefano; MARANA, Aparecido Nilceu. DUA: A Domain-Unified Approach for Cross-Dataset 3D Human Pose Estimation. *Sensors*, v. 23, n. 17, p. 7312, 2023.

The following papers, not strictly related to this work, have been published during the Master's period:

- MANESCO, João Renato Ribeiro; MARANA, Aparecido Nilceu. Combining ArcFace and Visual Transformer Mechanisms for Biometric Periocular Recognition. **IEEE Latin America Transactions**, v. 21, n. 7, p. 814-820, 2023.
- CANTO, Victor Hugo Braguim; MANESCO, João Renato Ribeiro; SOUZA, Gustavo Botelho de; MARANA, Aparecido Nilceu. Dog Face Recognition Using Vision Transformer. In: **12th Brazilian Conference on Intelligent Systems (BRACIS 2023)**. Cham: Springer Nature Switzerland, 2023. p. 33-47.

Bibliography

ALAOUI, A. Y.; TABII, Y.; THAMI, R. O. H.; DAOUDI, M.; BERRETTI, S.; PALA, P. Fall detection of elderly people using the manifold of positive semidefinite matrices. *Journal of Imaging*, MDPI, v. 7, n. 7, p. 109, 2021.

ANGELINI, F.; FU, Z.; LONG, Y.; SHAO, L.; NAQVI, S. M. Actionxpose: A novel 2d multi-view pose-based algorithm for real-time human action recognition. *arXiv preprint arXiv:1810.12126*, 2018.

BANIK, S.; GARCÍA, A. M.; KNOLL, A. 3d human pose regression using graph convolutional network. In: *2021 IEEE International Conference on Image Processing (ICIP)*. [S.l.: s.n.], 2021. p. 924–928.

BARTOL, K.; BOJANIC, D.; PETKOVIC, T.; D'APUZZO, N.; PRIBANIC, T. A Review of 3D Human Pose Estimation from 2D Images. In: *Proceedings of 3DBODY.TECH 2020 - 11th International Conference and Exhibition on 3D Body Scanning and Processing Technologies, Online/Virtual, 17-18 November 2020*. Online: Hometrica Consulting - Dr. Nicola D'Apuzzo, 2020. ISBN 978-3-033-08209-0.

BELLUZZO, B.; MARANA, A. N. Human action recognition based on 2d poses and skeleton joints. In: SPRINGER. *Brazilian Conference on Intelligent Systems*. [S.l.], 2022. p. 71–83.

CAO, Z.; HIDALGO, G.; SIMON, T.; WEI, S.-E.; SHEIKH, Y. Openpose: realtime multi-person 2d pose estimation using part affinity fields. *IEEE transactions on pattern analysis and machine intelligence*, IEEE, v. 43, n. 1, p. 172–186, 2019.

CHAI, W.; JIANG, Z.; HWANG, J.-N.; WANG, G. Global adaptation meets local generalization: Unsupervised domain adaptation for 3d human pose estimation. *arXiv preprint arXiv:2303.16456*, 2023.

CHAI, X.; WANG, Q.; ZHAO, Y.; LIU, X.; BAI, O.; LI, Y. Unsupervised domain adaptation techniques based on auto-encoder for non-stationary eeg-based emotion recognition. *Computers in Biology and Medicine*, v. 79, p. 205–214, 2016. ISSN 0010-4825. Available on: <https://www.sciencedirect.com/science/article/pii/S0010482516302797>.

CHEN, C.; RAMANAN, D. 3d human pose estimation = 2d pose estimation + matching. In: *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Los Alamitos, CA, USA: IEEE Computer Society, 2017. p. 5759–5767. ISSN 1063-6919.

CHEN, K.; GABRIEL, P.; ALASFOUR, A.; GONG, C.; DOYLE, W. K.; DEVINSKY, O.; FRIEDMAN, D.; DUGAN, P.; MELLONI, L.; THESEN, T. et al. Patient-specific pose estimation in clinical environments. *IEEE journal of translational engineering in health and medicine*, IEEE, v. 6, p. 1–11, 2018.

CHEN, T.; FANG, C.; SHEN, X.; ZHU, Y.; CHEN, Z.; LUO, J. Anatomy-aware 3d human pose estimation with bone-based pose decomposition. *IEEE Transactions on Circuits and Systems for Video Technology*, IEEE, 2021.

CHEN, W.; JIANG, Z.; GUO, H.; NI, X. Fall detection based on key points of human-skeleton using openpose. *Symmetry*, MDPI, v. 12, n. 5, p. 744, 2020.

CHEN, X.; WANG, S.; WANG, J.; LONG, M. Representation subspace distance for domain adaptation regression. In: PMLR. *International Conference on Machine Learning*. [S.l.], 2021. p. 1749–1759.

CHEN, Y.; TIAN, Y.; HE, M. Monocular human pose estimation: A survey of deep learning-based methods. *Computer Vision and Image Understanding*, v. 192, p. 102897, mar. 2020. ISSN 10773142.

CI, H.; MA, X.; WANG, C.; WANG, Y. Locally connected network for monocular 3d human pose estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, IEEE, 2020.

CSURKA, G. *Domain Adaptation in Computer Vision Applications*. 1st. ed. [S.l.]: Springer Publishing Company, Incorporated, 2017. ISBN 3319583468.

DABRAL, R.; MUNDHADA, A.; KUSUPATI, U.; AFAQUE, S.; SHARMA, A.; JAIN, A. Learning 3d human pose from structure and motion. In: *Proceedings of the European Conference on Computer Vision (ECCV)*. [S.l.: s.n.], 2018. p. 668–683.

DANG, Q.; YIN, J.; WANG, B.; ZHENG, W. Deep learning based 2D human pose estimation: A survey. *Tsinghua Science and Technology*, v. 24, n. 6, p. 663–676, 2019. ISSN 18787606.

DENG, W.; ZHENG, Y.; LI, H.; WANG, X.; WU, Z.; ZENG, M. Vh3d-lsfm: Video-based human 3d pose estimation with long-term and short-term pose fusion mechanism. In: SPRINGER. *Chinese Conference on Pattern Recognition and Computer Vision (PRCV)*. [S.l.], 2020. p. 589–601.

DEVANNE, M.; WANNOUS, H.; BERRETTI, S.; PALA, P.; DAOUDI, M.; BIMBO, A. D. 3-d human action recognition by shape analysis of motion trajectories on riemannian manifold. *IEEE transactions on cybernetics*, IEEE, v. 45, n. 7, p. 1340–1352, 2014.

DOERSCH, C.; ZISSERMAN, A. Sim2real transfer learning for 3D human pose estimation: Motion to the rescue. *CoRR*, abs/1907.02499, 2019.

DUAN, H.; ZHAO, Y.; CHEN, K.; LIN, D.; DAI, B. Revisiting skeleton-based action recognition. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. [S.l.: s.n.], 2022. p. 2969–2978.

EINFALT, M.; LUDWIG, K.; LIENHART, R. Uplift and upsample: Efficient 3d human pose estimation with uplifting transformers. In: *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. [S.l.: s.n.], 2023. p. 2903–2913.

FABBRI, M.; LANZI, F.; CALDERARA, S.; PALAZZI, A.; VEZZANI, R.; CUCCHIARA, R. Learning to detect and track visible and occluded body joints in a virtual world. In: *Proceedings of the European Conference on Computer Vision (ECCV)*. [S.l.: s.n.], 2018. p. 430–446.

GANIN, Y.; USTINOVA, E.; AJAKAN, H.; GERMAIN, P.; LAROCHELLE, H.; LAVIOLETTE, F.; MARCHAND, M.; LEMPITSKY, V. Domain-adversarial training of neural networks. In: _____. *Domain Adaptation in Computer Vision Applications*. Cham: Springer International Publishing, 2017. p. 189–209. ISBN 978-3-319-58347-1. Available on: <https://doi.org/10.1007/978-3-319-58347-1_10>.

GHOLAMI, M.; WANDT, B.; RHODIN, H.; WARD, R.; WANG, Z. J. Adaptpose: Cross-dataset adaptation for 3d human pose estimation by learnable motion generation. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. [S.l.: s.n.], 2022. p. 13075–13085.

GONG, K.; ZHANG, J.; FENG, J. Poseaug: A differentiable pose augmentation framework for 3d human pose estimation. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. [S.l.: s.n.], 2021. p. 8575–8584.

GU, Y.; PANDIT, S.; SARAEE, E.; NORDAHL, T.; ELLIS, T.; BETKE, M. Home-based physical therapy with an interactive computer vision system. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*. [S.l.: s.n.], 2019. p. 0–0.

GUAN, S.; XU, J.; WANG, Y.; NI, B.; YANG, X. Bilevel online adaptation for out-of-domain human mesh reconstruction. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. [S.l.: s.n.], 2021. p. 10472–10481.

GUO, Y.; ZHAO, L.; ZHANG, S.; YANG, J. Coarse-to-fine 3d human pose estimation. In: SPRINGER. *International Conference on Image and Graphics*. [S.l.], 2019. p. 579–592.

HE, X.; WANG, H.; QIN, Y.; TAO, L. 3d human pose estimation with grouping regression. In: SPRINGER. *Chinese Conference on Image and Graphics Technologies*. [S.l.], 2019. p. 138–149.

HOSSAIN, M. R. I.; LITTLE, J. J. Exploiting temporal information for 3d human pose estimation. In: *Proceedings of the European Conference on Computer Vision (ECCV)*. [S.l.: s.n.], 2018. p. 68–84.

INSAFUTDINOV, E.; PISHCHULIN, L.; ANDRES, B.; ANDRILUKA, M.; SCHIELE, B. Deepercut: A deeper, stronger, and faster multi-person pose estimation model. In: LEIBE, B.; MATAS, J.; SEBE, N.; WELLING, M. (Ed.). *Computer Vision – ECCV 2016*. Cham: Springer International Publishing, 2016. p. 34–50. ISBN 978-3-319-46466-4.

JAHANGIRI, E.; YUILLE, A. L. Generating multiple diverse hypotheses for human 3d pose consistent with 2d joint detections. In: *Proceedings of the IEEE International Conference on Computer Vision Workshops*. [S.l.: s.n.], 2017. p. 805–814.

JAIN, A.; TOMPSON, J.; LECUN, Y.; BREGLER, C. Modeep: A deep learning framework using motion features for human pose estimation. In: CREMERS, D.; REID, I.; SAITO, H.; YANG, M.-H. (Ed.). *Computer Vision – ACCV 2014*. Cham: Springer International Publishing, 2015. p. 302–315. ISBN 978-3-319-16808-1.

JANGUA, D.; MARANA, A. A new method for gait recognition using 2d poses. In: SBC. *Anais do XVI Workshop de Visão Computacional*. [S.l.], 2020. p. 69–74.

JIANG, J.; JI, Y.; WANG, X.; LIU, Y.; WANG, J.; LONG, M. Regressive domain adaptation for unsupervised keypoint detection. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. [S.l.: s.n.], 2021. p. 6780–6789.

JIANG, M.; YU, Z.; ZHANG, Y.; WANG, Q.; LI, C.; LEI, Y. Reweighted sparse representation with residual compensation for 3d human pose estimation from a single rgb image. *Neurocomputing*, Elsevier, v. 358, p. 332–343, 2019.

JIANG, Y.; LIU, X.; WU, D.; ZHAO, P. Residual deep monocular 3d human pose estimation using cvae synthetic data. In: IOP PUBLISHING. *Journal of Physics: Conference Series*. [S.l.], 2021. v. 1873, p. 012003.

KENDALL, A.; GAL, Y.; CIPOLLA, R. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. [S.l.: s.n.], 2018. p. 7482–7491.

KIM, T. S.; REITER, A. Interpretable 3d human action analysis with temporal convolutional networks. In: *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*. [S.l.: s.n.], 2017. p. 20–28.

KLUG, N.; EINFALT, M.; BREHM, S.; LIENHART, R. Error bounds of projection models in weakly supervised 3d human pose estimation. In: IEEE. *2020 International Conference on 3D Vision (3DV)*. [S.l.], 2020. p. 898–907.

KOUW, W. M.; LOOG, M. An introduction to domain adaptation and transfer learning. *arXiv preprint arXiv:1812.11806*, 2018.

KREISS, S.; BERTONI, L.; ALAHI, A. Pifpaf: Composite fields for human pose estimation. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. [S.l.: s.n.], 2019. p. 11977–11986.

KUNDU, J. N.; SETH, S.; YM, P.; JAMPANI, V.; CHAKRABORTY, A.; BABU, R. V. Uncertainty-aware adaptation for self-supervised 3d human pose estimation. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. [S.l.: s.n.], 2022. p. 20448–20459.

LI, C.; LEE, G. H. Generating multiple hypotheses for 3d human pose estimation with mixture density network. In: *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. [S.l.: s.n.], 2019. p. 9879–9887.

LI, H.; SHI, B.; DAI, W.; ZHENG, H.; WANG, B.; SUN, Y.; GUO, M.; LI, C.; ZOU, J.; XIONG, H. Pose-oriented transformer with uncertainty-guided refinement for 2d-to-3d human pose estimation. *arXiv preprint arXiv:2302.07408*, 2023.

LI, S.; CHAN, A. B. 3D Human Pose Estimation from Monocular Images with Deep Convolutional Neural Network. In: CREMERS, D.; REID, I.; SAITO, H.; YANG, M.-H. (Ed.). *Computer Vision – ACCV 2014*. Cham: Springer International Publishing, 2015. v. 9004, p. 332–347. ISBN 978-3-319-16808-1. Series Title: Lecture Notes in Computer Science.

LI, W.; LIU, H.; TANG, H.; WANG, P.; GOOL, L. V. Mhformer: Multi-hypothesis transformer for 3d human pose estimation. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. [S.l.: s.n.], 2022. p. 13147–13156.

LIANG, G.; ZHONG, X.; RAN, L.; ZHANG, Y. An adaptive viewpoint transformation network for 3d human pose estimation. *IEEE Access*, IEEE, v. 8, p. 143076–143084, 2020.

LIN, T.-Y.; MAIRE, M.; BELONGIE, S.; HAYS, J.; PERONA, P.; RAMANAN, D.; DOLLÁR, P.; ZITNICK, C. L. Microsoft coco: Common objects in context. In: FLEET, D.; PAJDLA, T.; SCHIELE, B.; TUYTELAARS, T. (Ed.). *Computer Vision – ECCV 2014*. Cham: Springer International Publishing, 2014. p. 740–755.

LIU, R.; SHEN, J.; WANG, H.; CHEN, C.; CHEUNG, S.-c.; ASARI, V. K. Enhanced 3d human pose estimation from videos by using attention-based neural network with dilated convolutions. *International Journal of Computer Vision*, Springer, v. 129, n. 5, p. 1596–1615, 2021.

LONG, M.; CAO, Y.; CAO, Z.; WANG, J.; JORDAN, M. I. Transferable representation learning with deep adaptation networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, IEEE, v. 41, n. 12, p. 3071–3085, 2018.

LOPER, M. M.; MAHMOOD, N.; BLACK, M. J. MoSh: Motion and shape capture from sparse markers. *ACM Transactions on Graphics, (Proc. SIGGRAPH Asia)*, ACM, New York, NY, USA, v. 33, n. 6, p. 220:1–220:13, nov. 2014. Available on: <http://doi.acm.org/10.1145/2661229.2661273>.

LU, M.; POSTON, K.; PFEFFERBAUM, A.; SULLIVAN, E. V.; FEI-FEI, L.; POHL, K. M.; NIEBLES, J. C.; ADELI, E. Vision-based estimation of mds-updrs gait scores for assessing parkinson's disease motor severity. In: SPRINGER. *Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part III 23*. [S.l.], 2020. p. 637–647.

LUVIZON, D. C.; TABIA, H.; PICARD, D. Human pose regression by combining indirect part detection and contextual information. *Computers & Graphics*, v. 85, p. 15–22, dez. 2019. ISSN 00978493.

MAHMOOD, N.; GHORBANI, N.; TROJE, N. F.; PONS-MOLL, G.; BLACK, M. J. AMASS: Archive of motion capture as surface shapes. In: *International Conference on Computer Vision*. [S.l.: s.n.], 2019. p. 5442–5451.

MANESCO, J. R. R.; MARANA, A. N. A survey of recent advances on two-step 3d human pose estimation. In: SPRINGER. *Brazilian Conference on Intelligent Systems*. [S.l.], 2022. p. 266–281.

MARCARD, T. von; HENSCHEL, R.; BLACK, M. J.; ROSENHAHN, B.; PONS-MOLL, G. Recovering Accurate 3D Human Pose in the Wild Using IMUs and a Moving Camera. In: FERRARI, V.; HEBERT, M.; SMINCHISESCU, C.; WEISS, Y. (Ed.). *Computer Vision – ECCV 2018*. Cham: Springer International Publishing, 2018. v. 11214, p. 614–631. ISBN 978-3-030-01248-9 978-3-030-01249-6. Series Title: Lecture Notes in Computer Science.

MARTINEZ, J.; HOSSAIN, R.; ROMERO, J.; LITTLE, J. J. A simple yet effective baseline for 3d human pose estimation. In: *2017 IEEE International Conference on Computer Vision (ICCV)*. [S.l.: s.n.], 2017. p. 2659–2668.

MISRA, T.; ARORA, A.; MARWAHA, S.; CHINNUSAMY, V.; RAO, A. R.; JAIN, R.; SAHOO, R. N.; RAY, M.; KUMAR, S.; RAJU, D.; JHA, R. R.; NIGAM, A.; GOEL, S. SpikeSegNet-a deep learning approach utilizing encoder-decoder network with hourglass for spike segmentation and counting in wheat plant from visual imaging. *Plant Methods*, Springer Science and Business Media LLC, v. 16, n. 1, mar. 2020.

MORENO-NOGUER, F. 3d human pose estimation from a single image via distance matrix regression. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. [S.l.: s.n.], 2017. p. 2823–2832.

NEWELL, A.; YANG, K.; DENG, J. Stacked Hourglass Networks for Human Pose Estimation. In: LEIBE, B.; MATAS, J.; SEBE, N.; WELLING, M. (Ed.). *Computer Vision – ECCV 2016*. Cham: Springer International Publishing, 2016. v. 9912, p. 483–499. ISBN 978-3-319-46484-8. Lecture Notes in Computer Science.

NIBALI, A.; HE, Z.; MORGAN, S.; PRENDERGAST, L. Numerical Coordinate Regression with Convolutional Neural Networks. *arXiv:1801.07372 [cs]*, maio 2018. ArXiv: 1801.07372.

PAN, S. J.; YANG, Q. A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, v. 22, n. 10, p. 1345–1359, Oct 2010. ISSN 1041-4347.

PATEL, V. M.; GOPALAN, R.; LI, R.; CHELLAPPA, R. Visual domain adaptation: A survey of recent advances. *IEEE signal processing magazine*, IEEE, v. 32, n. 3, p. 53–69, 2015.

PAVLAKOS, G.; ZHOU, X.; DANIILIDIS, K. Ordinal depth supervision for 3d human pose estimation. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. [S.l.: s.n.], 2018. p. 7307–7316.

RAFI, U.; LEIBE, B.; GALL, J.; KOSTRIKOV, I. An Efficient Convolutional Network for Human Pose Estimation. In: *Procedings of the British Machine Vision Conference 2016*. York, UK: British Machine Vision Association, 2016. p. 109.1–109.11. ISBN 978-1-901725-59-9.

RAPCZYŃSKI, M.; WERNER, P.; HANDRICH, S.; AL-HAMADI, A. A baseline for cross-database 3d human pose estimation. *Sensors*, MDPI, v. 21, n. 11, p. 3769, 2021.

SÁRÁNDI, I.; HERMANS, A.; LEIBE, B. Learning 3d human pose estimation from dozens of datasets using a geometry-aware autoencoder to bridge between skeleton formats. In: *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. [S.l.: s.n.], 2023. p. 2956–2966.

SARKER, S.; RAHMAN, S.; HOSSAIN, T.; AHMED, S. F.; JAMAL, L.; AHAD, M. A. R. Skeleton-based activity recognition: preprocessing and approaches. *Contactless Human Activity Analysis*, Springer, p. 43–81, 2021.

SHI, L.; ZHANG, Y.; CHENG, J.; LU, H. Action recognition via pose-based graph convolutional networks with intermediate dense supervision. *Pattern Recognition*, Elsevier, v. 121, p. 108170, 2022.

SILVA, M. V. da; MARANA, A. N. Human action recognition in videos based on spatiotemporal features and bag-of-poses. *Applied Soft Computing*, Elsevier, v. 95, p. 106513, 2020.

STAMOU, G.; KRINIDIS, M.; LOUTAS, E.; NIKOLAIDIS, N.; PITAS, I. 2D and 3D Motion Tracking in Digital Video. In: *Handbook of Image and Video Processing*. [S.l.]: Elsevier, 2005. p. 491–XVIII. ISBN 978-0-12-119792-6.

STÜBL, G.; HEINDL, C.; EBENHOFER, G.; BAUER, H.; PICHLER, A. Lessons learned from human pose interaction in an industrial spatial augmented reality application. *Procedia Computer Science*, Elsevier, v. 217, p. 912–917, 2023.

SUN, J.; WANG, M.; ZHAO, X.; ZHANG, D. Multi-view pose generator based on deep learning for monocular 3d human pose estimation. *Symmetry*, v. 12, n. 7, 2020. ISSN 2073-8994. Available on: <https://www.mdpi.com/2073-8994/12/7/1116>.

TOSHEV, A.; SZEGEDY, C. Deep pose: Human pose estimation via deep neural networks. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. [S.l.]: IEEE, 2014. p. 1653–1660.

TZENG, E.; HOFFMAN, J.; ZHANG, N.; SAENKO, K.; DARRELL, T. Deep domain confusion: Maximizing for domain invariance. *arXiv preprint arXiv:1412.3474*, 2014.

VAROL, G.; ROMERO, J.; MARTIN, X.; MAHMOOD, N.; BLACK, M. J.; LAPTEV, I.; SCHMID, C. Learning from synthetic humans. In: *CVPR*. [S.l.: s.n.], 2017.

VÉGES, M.; VARGA, V.; LŐRINCZ, A. 3d human pose estimation with siamese equivariant embedding. *Neurocomputing*, Elsevier, v. 339, p. 194–201, 2019.

VEMULAPALLI, R.; ARRATE, F.; CHELLAPPA, R. Human action recognition by representing 3d skeletons as points in a lie group. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. [S.l.: s.n.], 2014. p. 588–595.

WANG, C.; QIU, H.; YUILLE, A. L.; ZENG, W. Learning basis representation to refine 3d human pose estimations. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. [S.l.: s.n.], 2019. v. 33, n. 01, p. 8925–8932.

WANG, K.; LIN, L.; JIANG, C.; QIAN, C.; WEI, P. 3d human pose machines with self-supervised learning. *IEEE transactions on pattern analysis and machine intelligence*, IEEE, v. 42, n. 5, p. 1069–1082, 2019.

WANG, M.; DENG, W. Deep visual domain adaptation: A survey. *Neurocomputing*, Elsevier BV, v. 312, p. 135–153, Oct 2018. ISSN 0925-2312.

WEI, G.; LAN, C.; ZENG, W.; CHEN, Z. View invariant 3d human pose estimation. *IEEE Transactions on Circuits and Systems for Video Technology*, IEEE, v. 30, n. 12, p. 4601–4610, 2019.

WEI, G.; WU, S.; TANG, K.; LI, G. BoneNet: Real-time 3d human pose estimation by generating multiple hypotheses with bone-map representation. *Computer-Aided Design and Applications*, CAD Solutions, LLC, v. 18, n. 6, p. 1448–1465, fev. 2021. Available on: <https://doi.org/10.14733/cadaps.2021.1448-1465>.

WEN, G.; CHEN, H.; CAI, D.; HE, X. Improving face recognition with domain adaptation. *Neurocomput.*, Elsevier Science Publishers B. V., Amsterdam, The Netherlands, The Netherlands, v. 287, n. C, p. 45–51, abr. 2018. ISSN 0925-2312.

WENG, C.-Y.; CURLESS, B.; KEMELMACHER-SHLIZERMAN, I. Photo wake-up: 3d character animation from a single photo. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. [S.l.: s.n.], 2019. p. 5908–5917.

XIA, H.; XIAO, M. 3d human pose estimation with generative adversarial networks. *IEEE Access*, IEEE, v. 8, p. 206198–206206, 2020.

XIE, Z.; XIA, H.; FENG, C. A multi-scale recalibrated approach for 3d human pose estimation. In: SPRINGER. *Pacific-Asia Conference on Knowledge Discovery and Data Mining*. [S.l.], 2019. p. 400–411.

XU, H.; WU, S. 3d human pose estimation based on center of gravity. In: IEEE. *2020 International Joint Conference on Neural Networks (IJCNN)*. [S.l.], 2020. p. 1–7.

XU, J.; YU, Z.; NI, B.; YANG, J.; YANG, X.; ZHANG, W. Deep kinematics analysis for monocular 3d human pose estimation. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. [S.l.: s.n.], 2020. p. 899–908.

XU, Y.; WANG, W.; LIU, T.; LIU, X.; XIE, J.; ZHU, S.-C. Monocular 3d pose estimation via pose grammar and data augmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, IEEE, 2021.

YANG, J.; WAN, L.; XU, W.; WANG, S. 3d human pose estimation from a single image via exemplar augmentation. *Journal of Visual Communication and Image Representation*, Elsevier, v. 59, p. 371–379, 2019.

YANG, W.; OUYANG, W.; WANG, X.; REN, J.; LI, H.; WANG, X. 3d human pose estimation in the wild by adversarial learning. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. [S.l.: s.n.], 2018. p. 5255–5264.

YANG, Y.; RAMANAN, D. Articulated Human Detection with Flexible Mixtures of Parts. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, v. 35, n. 12, p. 2878–2890, dez. 2013. ISSN 0162-8828, 2160-9292.

YIN, Y.; LIU, M.; ZHU, Q.; ZHANG, S.; HUSSIEN, N. A.; FAN, Y. Multi-branch attention graph convolutional networks for 3d human pose estimation. *IEEE Transactions on Instrumentation and Measurement*, IEEE, 2023.

ZENG, A.; SUN, X.; HUANG, F.; LIU, M.; XU, Q.; LIN, S. Srnet: Improving generalization in 3d human pose estimation with a split-and-recombine approach. In: SPRINGER. *European Conference on Computer Vision*. [S.l.], 2020. p. 507–523.

ZHANG, H.; SCIUTTO, C.; AGRAWALA, M.; FATAHALIAN, K. Vid2player: Controllable video sprites that behave and appear like professional tennis players. *ACM Transactions on Graphics (TOG)*, ACM New York, NY, v. 40, n. 3, p. 1–16, 2021.

ZHANG, J.; WANG, Y.; ZHOU, Z.; LUAN, T.; WANG, Z.; QIAO, Y. Learning dynamical human-joint affinity for 3d pose estimation in videos. *IEEE Transactions on Image Processing*, IEEE, v. 30, p. 7914–7925, 2021.

ZHANG, X.; WONG, Y.; KANKANHALLI, M. S.; GENG, W. Unsupervised domain adaptation for 3d human pose estimation. In: *Proceedings of the 27th ACM International Conference on Multimedia*. [S.l.: s.n.], 2019. p. 926–934.

ZHANG, X.; WONG, Y.; WU, X.; LU, J.; KANKANHALLI, M.; LI, X.; GENG, W. Learning causal representation for training cross-domain pose estimator via generative interventions. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. [S.l.: s.n.], 2021. p. 11270–11280.

ZHAO, L.; PENG, X.; TIAN, Y.; KAPADIA, M.; METAXAS, D. N. Semantic graph convolutional networks for 3d human pose regression. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. [S.l.: s.n.], 2019. p. 3425–3435.

ZHAO, Q.; ZHENG, C.; LIU, M.; WANG, P.; CHEN, C. 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*. [S.l.: s.n.], 2023. p. 8877–8886.

ZHAO, W.; WANG, W.; TIAN, Y. Graformer: Graph-oriented transformer for 3d pose estimation. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. [S.l.: s.n.], 2022. p. 20438–20447.

ZHENG, C.; WU, W.; YANG, T.; ZHU, S.; CHEN, C.; LIU, R.; SHEN, J.; KEHTARNAVAZ, N.; SHAH, M. Deep learning-based human pose estimation: A survey. *CoRR*, abs/2012.13392, 2020. Available on: <https://arxiv.org/abs/2012.13392>.

ZHOU, K.; LIU, Z.; QIAO, Y.; XIANG, T.; LOY, C. C. Domain generalization: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, IEEE, 2022.

ZHOU, S.; JIANG, M.; WANG, Q.; LEI, Y. Towards locality similarity preserving to 3d human pose estimation. In: *ACCV Workshops*. [S.l.: s.n.], 2020. p. 136–153.

ZOU, L.; HUANG, Z.; GU, N.; WANG, F.; YANG, Z.; WANG, G. Gmdn: A lightweight graph-based mixture density network for 3d human pose regression. *Computers & Graphics*, Elsevier, v. 95, p. 115–122, 2021.

ZUFFI, S.; FREIFELD, O.; BLACK, M. J. From Pictorial Structures to deformable structures. In: *2012 IEEE Conference on Computer Vision and Pattern Recognition*. Providence, RI: IEEE, 2012. p. 3546–3553.