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

João Renato Ribeiro Manesco
School of Sciences
São Paulo State University (UNESP)
Bauru - SP, Brazil
joao.r.manesco@unesp.br

Stefano Berretti
Media Integration and Communication Center
University of Florence
Florence, Italy
stefano.berretti@unifi.it

Aparecido Nilceu Marana
School of Sciences
São Paulo State University (UNESP)
Bauru - SP, Brazil
nilceu.marana@unesp.br

***Abstract*—Human pose estimation in monocular images is a challenging problem in Computer Vision. Currently, while 2D poses find extensive applications, the use of 3D poses suffers from data scarcity due to the difficulty of acquisition. Therefore, fully convolutional approaches struggle due to limited 3D pose labels, prompting a two-step strategy leveraging 2D pose estimators, which does not generalize well to unseen poses, requiring the use of domain adaptation techniques. In this work, we introduce a novel Domain Unified Approach called DUA, which, through a unique combination of three modules on top of the pose estimator (pose converter, uncertainty estimator, and domain classifier), can improve the accuracy of 3D poses estimated from 2D poses. In the experiments carried out with SURREAL and Human3.6M datasets, our method reduced the mean per-joint position error (MPJPE) by 44.1 mm in the synthetic-to-real scenario, a quite significant result. Furthermore, our method outperformed all state-of-the-art methods in the real-to-synthetic scenario.**

I. Introduction

Human pose estimation is an important and challenging computer vision problem. It aims to estimate the human body shape (pose) based on a single image, usually monocular. This shape can be inferred by detecting joints in a skeleton, which are connected so that each connection represents a part of the human body [2]. 2D human pose estimation from monocular images has made significant strides in recent years, thanks to the development of well-consolidated, robust, and fast methods. These methods have found application in various fields, laying a solid foundation for exploring 2D pose estimation.

As using 3D poses rather than solely 2D poses has the potential to improve the precision and quality of applications, 3D pose estimation is a hot research topic in Computer Vision. The acquisition of 3D poses, however, presents a significant challenge. Unlike 2D poses, whose pose labels can be obtained, for instance, through crowdsourcing tools, obtaining 3D pose labels requires specialized solutions limited to very particular scenarios. This restriction reduces the diversity of the training data and makes using 3D pose estimation methods based on RGB monocular images difficult [3].

This paper summarizes the M.Sc. dissertation of João Renato Ribeiro Manesco [1]

A solution to this problem involves taking advantage of the maturity of 2D pose estimation methods to obtain 3D poses in two steps: a first step in which the 2D pose is obtained using consolidated methods, followed by a second step in which the 3D pose is estimated from the 2D pose. Methods based on this two-step strategy tend to perform better than the end-to-end ones [4]. However, recent two-step methods have reported high bias levels in frequently used databases, thereby impacting the performance on real applications [5].

Besides using the two-step strategy for 3D pose estimation, another objective of this work is to address the problem of incorporating data from domains other than the one used to train the model, for instance, synthetic image domains, to obtain better 3D human poses. The problem is that there is generally a shift between the real and synthetic image domains [6].

As one can observe, several factors influence the effectiveness of 3D human pose estimation, especially in scenarios that involve multiple datasets, such as the use of distinct body capture sensors across datasets, distinct image domains with misalignment in the camera and action distributions, and the error propagation on edge kinematic groups.

Although some methods in the literature address these problems, there is a lack of a comprehensive approach to deal with them in a unified way. Therefore, in this work, we propose a new method, called Domain-Unified Approach (DUA), for 3D human-pose estimation, based on a unified approach, that brings the following contributions:

- The creation of a novel and effective technique for converting pose models, achieving a Unified Pose Representation, to overcome the differences between capture sensors and pose misrepresentations;
- The enhancement 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 that unifies these solutions, mitigating the domain-shift problems and enhancing the overall effectiveness of the approach in cross-domain settings.

Besides dealing with all the highlighted problems, our novel method reduced the mean per-joint position error (MPJPE)

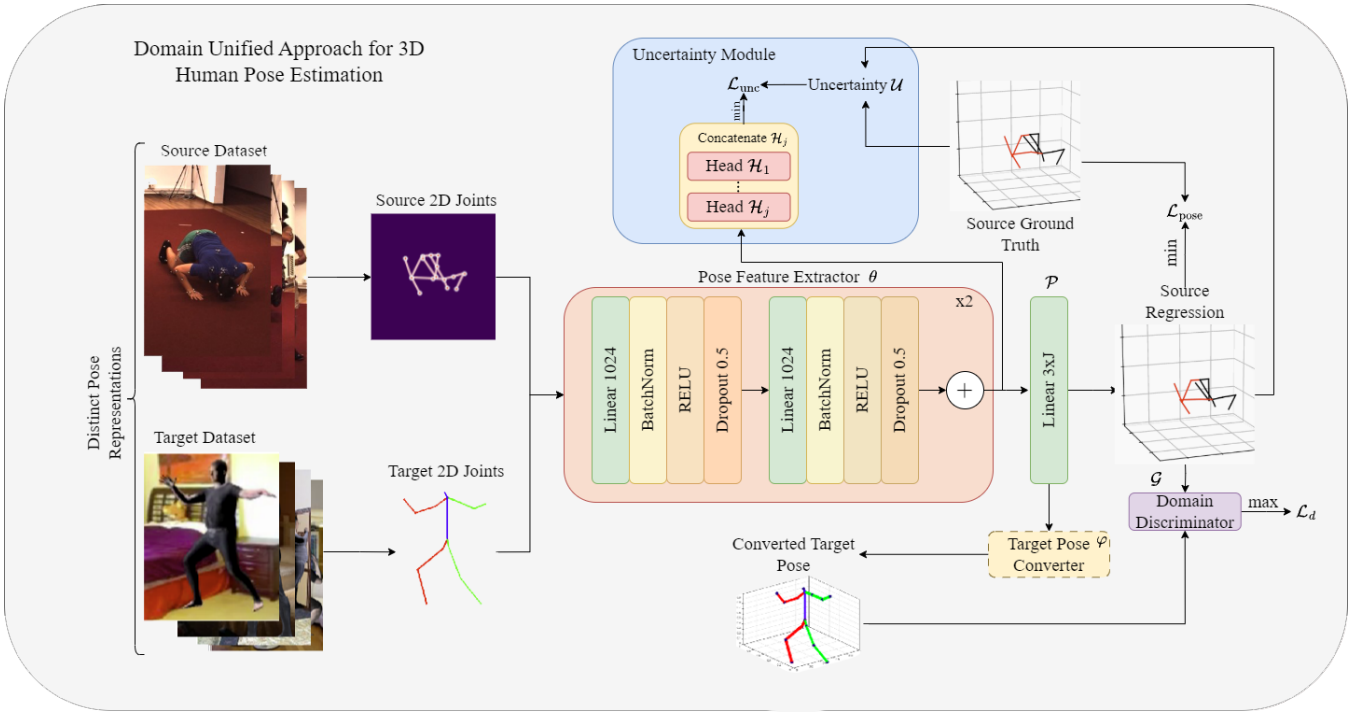


Fig. 1: Diagram of Domain Unified Approach (DUA). The method is composed of three main modules on top of the 3D pose estimator: the unified pose representation, the uncertainty estimation, and the domain discriminator. The dashed lines on the pose converter represent frozen weights.

in the synthetic-to-real scenario (from SURREAL to Human3.6M datasets) by 44.1mm, outperforming all state-of-the-art methods in the real-to-synthetic scenario (from SURREAL to Human3.6M datasets).

One of the most innovative aspects of our method is the proposed pose conversion mechanism. This mechanism introduces a novel approach to pose representation, offering a unified pose representation that effectively addresses the problem caused by multiple pose models.

II. Proposed Method

Aiming to tackle the problem of 3D human pose estimation based on monocular RGB images and domain adaptation, this work introduces a novel solution called Domain-Unified Approach (DUA). This novel method combines domain adaptation, a unified pose representation, and a unique training technique to mitigate error propagation at extreme joints. DUA demonstrates its prowess by making pose estimations from two domains with a low error rate. The key concept is to maximize the distance between 3D human poses using a domain discriminator optimized simultaneously with a feature extractor, as illustrated in Figure 1.

DUA has an architecture inspired by the DANN (Domain Adversarial Neural Networks) method [7]. To find the desired pose, given a pose estimator Π, the following pose loss is used:

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

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.

In the DUA method, the unified pose representation obtained by the converter undergoes a strategic pre-training phase, and its weights remain fixed during the training process. Conversely, the other components of the method are trained in an online mode, ensuring a robust and comprehensive training approach.

A. Unified Pose Representation

The incompatibility between the pose representations is a common problem in 3D human pose estimation when dealing with different datasets. Previous works have already discussed this issue in the literature. One such work aims to learn unified representations by utilizing different data sources concurrently [8]. This problem arises from various body capture sensors and different 3D pose representations used by works in the literature, leading each dataset to have its own representation. Figure 2 shows the difference in the pose representations used by common 3D human pose datasets in the literature, SMPL, and H3.6M formats.

This conversion problem was already discussed in the literature [9]. However, previous approaches tried to find a normalization technique through handcrafted features. This

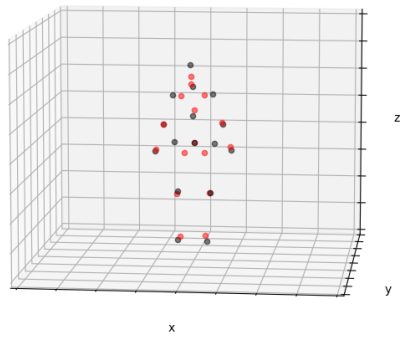


Fig. 2: Overlapped joints of the Human3.6M dataset coming from two distinct pose representations, SMPL (red) and the original H3.6M format (black).

approach sometimes works well 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 $\mathcal{B}$, 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 $\varphi : \mathcal{A} \mapsto (\mathcal{B} - \mathcal{A})$. The weights of the φ can be found using a single-layer residual neural network using gradient descent, as illustrated by Figure 3.

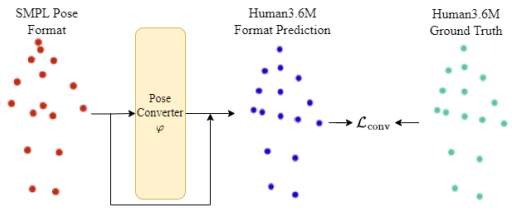


Fig. 3: Pose conversion method used to find a unified pose representation.

B. Uncertainty Estimation

The issue of 3D human pose estimation presents a challenge in error propagation within the most extreme kinematic group, which is aggravated by the ill-defined monocular estimation resulting from self-occlusion during varying camera perspectives. Therefore, 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. In contrast, aleatoric uncertainty aims to model unexplained uncertainties within the current training data [10]. Previous works have explored uncertainty modeling through Bayesian networks for 3D human pose estimation using different approaches [11], [12]. In this work, we propose a method based on a naive definition of uncertainty.

As such, 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, in which 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.

III. Material and Methodology

All experiments in this work were conducted using a computer with two Intel Xeon E5620 CPUs, 48GB of RAM, and an NVIDIA TitanXP GPU with 12GB of VRAM. 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 mechanism was pre-trained and its weights were frozen on the DUA method. Before training, a preprocessing step was applied to center all the poses with their hip joint on the origin of the coordinate system. In addition, all the poses were represented in the camera coordinate system by multiplying the pose coordinates by the inverse of their respective extrinsic camera matrix.

A. Datasets and Metrics

We evaluated our method on two datasets for cross-domain experiments: SURREAL [13] and Human3.6M [14]. 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. Examples of images from both datasets can be seen in Figure 4.

The method proposed and developed in this work was evaluated by using two standard metrics applied in the literature in the 3D human pose estimation problem: (i) MPJPE (Mean Per-Joint Position Error), the mean error in millimeters between the estimated points, and the real points after the root joint alignment; (ii) P-MPJPE (Procrustes-Aligned Mean Per-Joint Position Error), the mean error, in millimeters, between the estimated points and the real points after Procrustes alignment.

IV. Results and Discussion

DUA aims to unify the solution to distinct 3D human pose estimation problems by combining the pre-trained pose converter (with frozen weights), the pose uncertainty module,

TABLE I: 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 + DA	108.9 mm	70.1 mm	96.1 mm	57.9 mm	74.0 mm	52.2 mm
SURREAL + DA	-	-	55.8 mm	37.7 mm	106.7 mm	65.6 mm

*Denotes No Domain Adaptation.

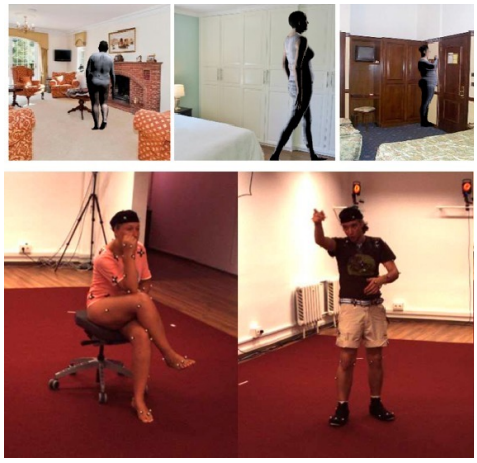


Fig. 4: Examples of images from SURREAL (first row) and Human3.6M (second row) datasets.

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. The domain adaptation method results are presented in Table I. More details on each module of our method with further qualitative results can be seen in Chapter 7 of the dissertation [1] and in the paper [15].

Notably, using 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 led to a reduction of 44.1 mm (from 150.8 mm without Domain Adaptation to 106.7 mm with DUA) in the mean per-joint position error (MPJPE).

The impact of our innovative approach, employing pose conversions, is evident in the Human3.6M $\rightarrow$ SURREAL scenario. Here, we evaluate a larger dataset, which includes a distinct subset of actions. In this challenging scenario, the effectiveness of our method experiences a slight decrease when using domain adaptation. This decrease can be observed by analyzing the domain adaptation method without pose conversion in Table I, where misrepresenting poses causes difficulties in optimizing our domain discriminator. However, by incorporating our dynamic pose conversion mechanism, we effectively mitigate this sudden decrease found in the domain adaptation scenario, as inferred from the 'Converted SURREAL' column of the Table I, which indicates the complete results of the DUA method with proper pose representation.

Furthermore, our method surpasses other state-of-the-art methods in the real-to-synthetic cross-dataset scenario, as shown in Table II.

One significant advantage of our approach is the use of a unified pose representation, wherein the conversion step mitigates issues arising from variations in body capture sensors across different datasets.

TABLE II: Quantitative results obtained on the H3.6M $\rightarrow$ SURREAL evaluation. Table results and layout are obtained from experiments conducted by [12] and [16]. Bold indicates the best result. More details on that can be found in the dissertation [1].

Method	H3.6M $\rightarrow$ SURREAL	
	MPJPE $\downarrow$	P-MPJPE $\downarrow$
DDC	117.5	80.1
DAN	114.2	78.4
DANN	113.6	77.2
[16]	103.3	69.1
[12]	99.6	67.2
[12]*	99.4	65.1
DUA (Ours)	**96.1**	**57.9**

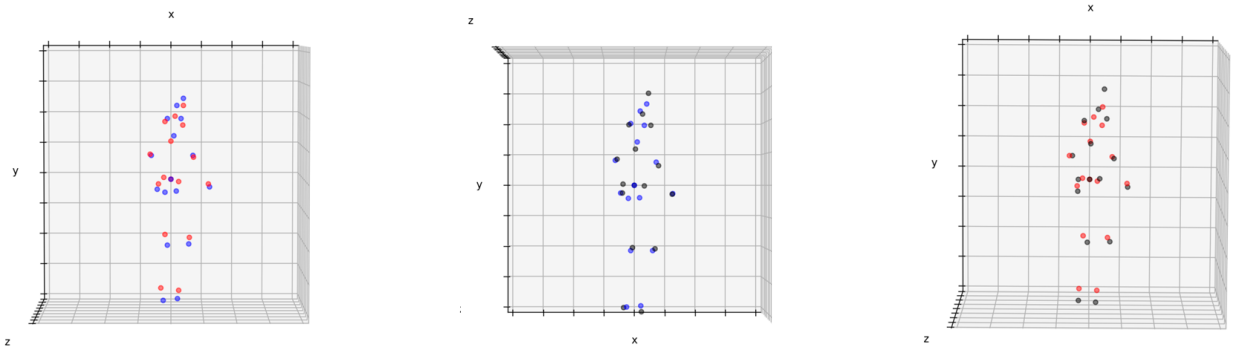
*Denotes an alternative approach for cross-domain evaluation conducted using test-time adaptation.

The use of a pose conversion mechanism stands as a significant contribution to the field, emphasizing the importance of finding a unified pose representation to mitigate domain discrepancy effectively. To further illustrate this point, we have provided ablation results in Table III.

TABLE III: Ablation results obtained on the Human3.6M $\rightarrow$ SURREAL scenario.

Method	MPJPE
without Unified Pose Representation and without Domain Adaptation	107.6 mm
without Unified Pose Representation and with Domain Adaptation	108.9 mm
with Unified Pose Representation and without Domain Adaptation	100.3 mm
DUA - with Unified Pose Representation and with Domain Adaptation	**96.1 mm**

In Figure 5, 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



(a) (b) (c)

Fig. 5: Overlapping of the different pose representations available, Human3.6M (red), SMPL ground-truth (blue), converted pose (black). (a) Shows a Human3.6M (red dots) pose superimposed on the original SMPL pose (blue dots). (b) Shows the resulting pose after conversion (black dots) superimposed to the original SMPL pose (blue dots). (c) Shows the converted pose (black dots) superimposed to the original Human3.6m pose (red dots).

our method (black dots). It can be noted that conversion of the SMPL pose to the Human3.6M pose model helps to reduce errors, particularly in the hip area, thus improving pose estimation accuracy and alignment with the target pose format.

Regarding the domain adaptation step of our method, Figure 6 shows some results obtained with and without domain adaptation. One can observe that when domain adaptation is not applied, there are 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 decrease 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 [17].

V. Conclusion

Experimental results show that our method (DUA) 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 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 improvement achieves state-of-the-art performance, highlighting the success of our method. By utilizing a unified pose representation and incorporating the pose conversion step, we have effectively addressed challenges arising from variations in body capture sensors across different datasets and enhanced the adaptability and generalization of our approach.

VI. Main Contributions

The method proposed and developed in this M.Sc dissertation, called DUA, provides a comprehensive solution to solve critical issues in 3D human pose estimation: i) Pose misrepresentation; (ii) Propagation of errors at joint edges; and (iii) Domain misalignment in cross-domain scenarios. In the experiments carried out with SURREAL and Human3.6M datasets, DUA reduced the MPJPE by 44.1 mm in the synthetic-to-real scenario, a quite significant result. Furthermore, DUA outperformed all state-of-the-art methods in the real-to-synthetic scenario.

This Master's thesis has been supported by a FAPESP scholarship (Proc. 2021/02028-6) and benefited, through a FAPESP Research Internship Abroad Scholarship (Proc. 2022/07055-4), by a fruitful international collaboration with Prof. Stefano Berretti, from MICC (The Media Integration and Communication Center), University of Florence, Italy.

During this Master's course, two papers related to the topic of the dissertation were 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. Available at: https://link.springer.com/chapter/10.1007/978-3-031-21689-3_20.
- **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. Available at: https://www.mdpi.com/1424-8220/23/17/7312. The paper's source code is available on Github: https://github.com/jrjoaorenato/DUA.

The following paper, not strictly related to the dissertation, was published during the Master's period as the first author:

- **MANESCO, João Renato Ribeiro**; MARANA, Aparecido Nilceu. Combining ArcFace and Visual Trans-

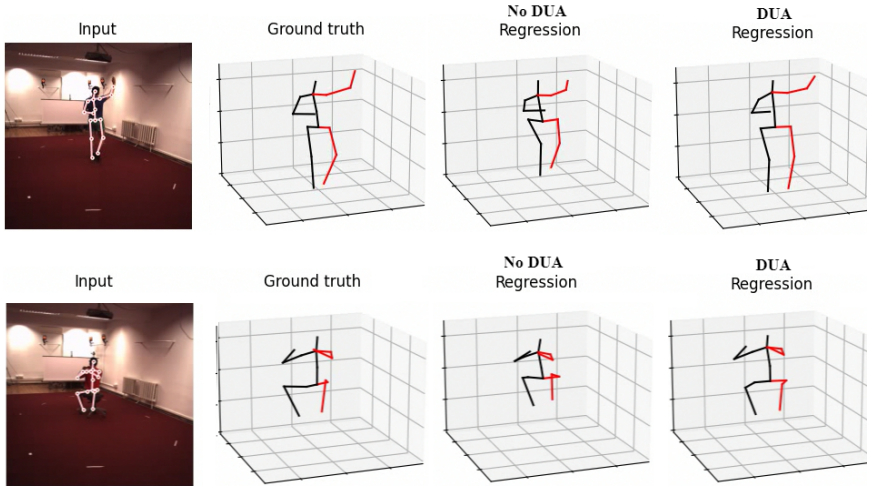


Fig. 6: Results obtained by using our proposed approach with and without the domain adaptation (DA) step, on two samples of Human3.6M dataset. Left: ground truth pose; Center: pose without DA; Right: pose with DA.

former Mechanisms for Biometric Periocular Recognition. IEEE Latin America Transactions, v. 21, n. 7, p. 814-820, 2023. Available at https://ieeexplore.ieee.org/abstract/document/10244180. The paper's source code is available on Gitlab: https://gitlab.com/jrjoaorenato/perioculartransformersarcface.

The following paper, not strictly related to the dissertation, was published during the Master's period as a co-author:

- 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. Available at: https://link.springer.com/chapter/10.1007/978-3-031-45389-2_3.

Apart from the publications already listed, partial results of this dissertation were presented at the XI Workshop of the Computer Science Graduate Program - UNESP (XI WPPGCC) in 2022. These results were published as an extended abstract.

The source code of the DUA method is available on Github:https://github.com/jrjoaorenato/DUA.

VII. Acknowledgements

This work has been supported by the São Paulo Research Foundation (FAPESP) (grants 2022/07055-4, 2021/02028-6, 2024/00789-8, 2013/07375-0) and Petrobras/Fundunesp (Process 2662/2017).

References

[1] J. R. R. Manesco, "3D Human Pose Estimation Based on Monocular RGB Images and Domain Adaptation," Master's thesis, São Paulo State University, School of Sciences, 2023.

[2] C. Zheng, W. Wu, T. Yang, S. Zhu, C. Chen, R. Liu, J. Shen, N. Kehtarnavaz, and M. Shah, "Deep learning-based human pose estimation: A survey," *arXiv preprint arXiv:2012.13392*, 2020.

[3] Y. Chen, Y. Tian, and M. He, "Monocular human pose estimation: A survey of deep learning-based methods," *Computer Vision and Image Understanding*, vol. 192, p. 102897, Mar. 2020.

[4] J. Martinez, R. Hossain, J. Romero, and J. J. Little, "A simple yet effective baseline for 3d human pose estimation," in *2017 IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 2659–2668.

[5] G. Wei, C. Lan, W. Zeng, and Z. Chen, "View invariant 3d human pose estimation," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 30, no. 12, pp. 4601–4610, 2019.

[6] G. Csurka, *Domain Adaptation in Computer Vision Applications*, 1st ed. Springer Publishing Company, Incorporated, 2017.

[7] Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. Marchand, and V. Lempitsky, "Domain-adversarial training of neural networks," *J. Mach. Learn. Res.*, vol. 17, no. 1, p. 2096–2030, Jan. 2016.

[8] I. Sárándi, A. Hermans, and B. Leibe, "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*, 2023, pp. 2956–2966.

[9] M. Rapczyński, P. Werner, S. Handrich, and A. Al-Hamadi, "A baseline for cross-database 3d human pose estimation," *Sensors*, vol. 21, no. 11, p. 3769, 2021.

[10] A. Kendall, Y. Gal, and R. Cipolla, "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*, 2018, pp. 7482–7491.

[11] H. Li, B. Shi, W. Dai, H. Zheng, B. Wang, Y. Sun, M. Guo, C. Li, J. Zou, and H. Xiong, "Pose-oriented transformer with uncertainty-guided refinement for 2d-to-3d human pose estimation," *arXiv preprint arXiv:2302.07408*, 2023.

[12] J. N. Kundu, S. Seth, P. YM, V. Jampani, A. Chakraborty, and R. V. Babu, "Uncertainty-aware adaptation for self-supervised 3d human pose estimation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 20 448–20 459.

[13] G. Varol, J. Romero, X. Martin, N. Mahmood, M. J. Black, I. Laptev, and C. Schmid, "Learning from synthetic humans," in *CVPR*, 2017.

[14] T. von Marcard, R. Henschel, M. J. Black, B. Rosenhahn, and G. Pons-Moll, "Recovering accurate 3d human pose in the wild using imus and a moving camera," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 601–617.

[15] J. R. R. Manesco, S. Berretti, and A. N. Marana, "Dua: A domain-unified approach for cross-dataset 3d human pose estimation," *Sensors*, vol. 23, no. 17, p. 7312, 2023.

[16] X. Zhang, Y. Wong, X. Wu, J. Lu, M. Kankanhalli, X. Li, and W. Geng, "Learning causal representation for training cross-domain pose estimator via generative interventions," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 11 270–11 280.

[17] T. Chen, C. Fang, X. Shen, Y. Zhu, Z. Chen, and J. Luo, "Anatomy-aware 3d human pose estimation with bone-based pose decomposition," *IEEE Transactions on Circuits and Systems for Video Technology*, 2021.