AR Poser: Automatically Augmenting Mobile Pictures with Digital Avatars Imitating Poses

Gokcen Cimen, Christoph Maurhofer, Robert W. Sumner, Martin Guay

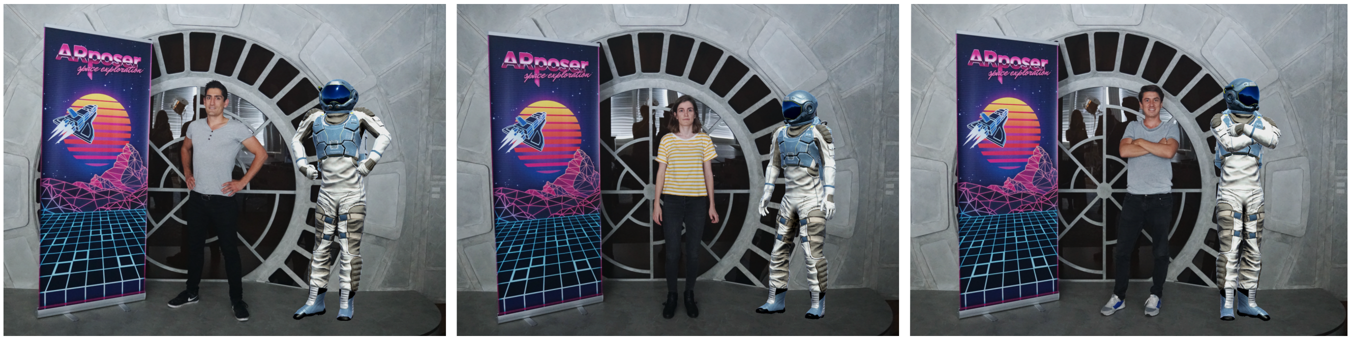


Figure 1: *Examples of poses automatically recovered and augmented with a digital character using our method.*

Abstract

We introduce *AR Poser*: a framework for posing *with, or as* a digital character. In this paper, we describe our first contribution to *AR Poser*: a technique for digital characters to recognize and automatically reproduce the same pose as a person in a picture (using only RGB information from a mobile device). 3D human pose estimation from RGB is an under-constrained and ambiguous problem that remains today an active field of study. Instead of addressing the general case of human pose estimation, we propose a solution that can be tailored to a specific scenario—such as entertainment poses for AR selfies. At the heart of our solution is a set of predefined poses (selfie poses) utilized to reduce ambiguities. In a nutshell, our method consists of two reliable steps: we first perform 2D pose estimation, and then perform a projection onto the 3D subspace to find the closest matching 3D pose. With our method, we are able to automatically create augmented reality selfies for a variety of different poses.

Keywords: Augmented Reality, Pose Estimation, Intelligent Virtual Characters.

Concepts:

1 Introduction

Digital augmentation of the real world opens new dimensions for ideation, communication and entertainment. For example, facial tracking combined with different mask overlays recently resulted in highly entertaining and popular mobile applications. In the future, we can imagine combining human shape estimation with digital character augmentation to unlock various entertaining selfie scenarios. Hence, we introduce *AR Poser*: a framework for posing with or as a digital character. In this paper, we describe our first contribution to *AR Poser*: a technique for digital characters to automatically reproduce the same pose as a person in a picture.

To automatically imitate the person's pose with a 3D digital character, we need to estimate the 3D pose of the person from a single monocular image (RGB). 3D human pose estimation from RGB is an under-constrained and ambiguous problem that remains an active field of study. Instead of addressing the general case of human pose estimation, we propose a solution that can be tailored to a specific scenario—such as poses AR selfies. At the heart of our solution is a set of predefined poses (selfie poses) utilized to reduce ambiguities associated with depth when estimating 3D poses. In a nutshell, our approach consists of breaking down the problem into two more reliable steps: first a 2D pose estimation, and then a projection onto our 3D subspace to find the closest matching 3D pose. With our method, we were able to automatically create augmented reality selfies for a variety of different poses.

2 Related Work

There has been several entertaining applications of people augmentation in the recent years. Combining face tracking from RGB images with different 3D masks is a famous example popularized by the *Snapchat* app [Kazemi and Sullivan 2014]. Recently, rapid progress has been made in estimating human poses and shapes from RGB images. To our knowledge, no work has yet fully automatically augmented people's bodies. For example, clothes has been added to images and video, but semi-automatically [Rogge et al. 2011; Rogge et al. 2014]. There has been manual augmentation, where the application inserts a pre-defined character or object in the scene [Barakonyi and Schmalstieg 2006; Zünd et al. 2014; McIntosh et al.]. The seminal work of Barakonyi explores different interactions in AR with digital characters and objects [Barakonyi and Schmalstieg 2006]. More recently, Zund et al. [Zünd et al. 2014] evaluated different aspects of reality mixing techniques, and the McIntosh et al. [McIntosh et al.] created a magic bench where a character appeared on the bench on a display in front of the bench, and the character interacts with the person. Recently, the idea of tracking real world objects to interact with digital characters has been demonstrated using pre-defined marker-based tracking for rigid localization [Cimen et al. 2018]. To our knowledge, there has not yet been a case of a person's pose automatically estimated and augmented with a digital character.

Estimating a 3D human pose skeleton from an image is a challenging problem due to the ambiguities associated to the depth projection, as well as the variations in human shapes. Using a depth camera, methods have been proposed where a large data-set of 3D skeleton poses and depth image pairs are created, to then regress a model that maps depth images to 3D skeletons [Shotton et al. 2012; Buys et al. 2014; Zimmermann et al. 2018], resulting in popular products such as the *kinect*.

Recently, researchers have experimented with a similar approach but using only a single monocular RGB image [Mehta et al. 2017a; Tomè et al. 2017; Martinez et al. 2017; Mehta et al. 2017b]. At the heart of this approach is a human pose synthesis procedure that can creates a large data-set of human poses with various textures, from a which a model can be fitted. While it is showing promising results, it remains challenging to differentiate between different human shapes and cover a wide range of poses.

We are not the first to consider breaking down the problem into a first 2D pose estimation step, followed by a 3D re-construction step. A large set of human captured motion can be used to faciliate convergence to a natural 3D pose [Wang et al. 2014; Yasin et al. 2015; Chen and Ramanan 2016]. Wang et al. [Wang et al. 2014] represented 3D poses as a linear combination of a sparse set of bases learned from a large 3D pose dataset, and solve the 3D reconstruction as an optimization problem in the reduced space. Instead of online optimization, Yasin et al, [Yasin et al. 2015] learned a direct mapping off line from 2D to 3D poses, and applied it to the task of 3D pose retrieval. Our approach does not rely on a large data set. A direct mapping blends between 2D skeletons as input and might yield 3D poses with artiacts. Hence a direct optimization method guarantees the 3D pose preserves its initial details. One additional problem with large set of poses is that while they may contain a variety of activities, they are complicated to gather and might not contain the entertainment poses desired for the application. Our approach is technically most similar to [Chen and Ramanan 2016], but depends only on a small set of poses, geared towards the application.

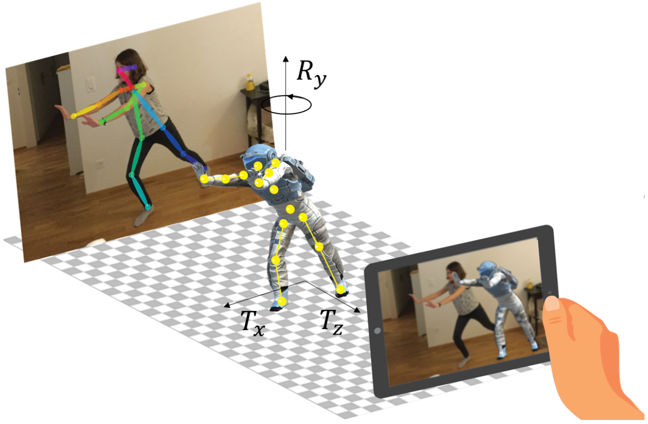


Figure 2: *From 2D pose estimation to 3D pose subspace and finding optimal character pose.*

3 2D Pose Estimation

We have recently seen rapid progress in 2D pose estimation from monocular images using deep learning "in the wild". There are now packaged solutions that offer robust solutions for multiple subjects and occluded parts. In this work, we use the pre-trained network *OpenPose* [Cao et al. 2016], which was trained on the COCO [Lin et al. 2014] and MPII [Andriluka et al. 2014] datasets.

The pre-trained network takes as input the RGB image, and returns a list of joint positions y_i together with a confidence value c_i. For example, a partially visible body will result in a low confidence value for the joints outside the image. The neural network was trained over a large data-set of hand-annotated images—each with the skeleton of the people in the image.

The 2D skeleton has a set of joint names, that we associate to the 3D joints of our character, as shown in Fig.3. This map is defined manually in our case, but could be done automatically given corresponding T poses for example. With the 2D joint positions associated to 3D joints, we can proceed to the step of computing the best matching 3D pose.

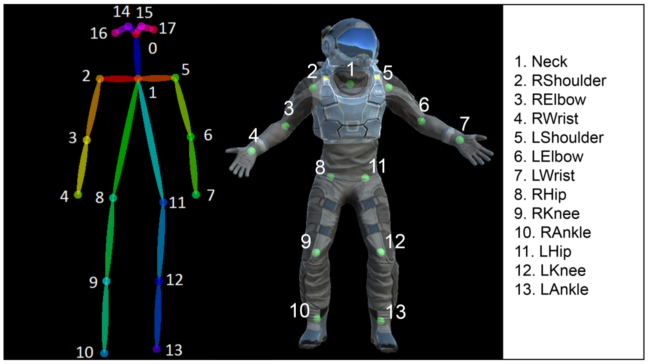


Figure 3: *The 2D skeleton on the left is obtained from OpenPose. It has 18 joints. On the right is the 3D character that we used in our experiment. A common subset of joints need to be mapped for the 3D pose matching process.*

4 3D Pose Projection

The way we project the 2D skeleton onto the 3D pose space is via local optimization. We assume a small set of 3D poses, in our case entertaining selfies, as shown in the results section. The 3D poses constrain the solution space to only plausible articulations of the character's body. Since possible selfie poses are symmetrical, we handle symmetries by mirroring the 3D poses in the dataset along the y-axis.

For each pose in the data base, we optimize for the rigid transformation that will bring the 3D pose, closest to the 2D projected skeleton, in terms for joint positions and bone direction similarity. The global transformation of a 3D pose is parameterized with 5 degrees of freedom, as we constrain the translation of the character along the y-axis, as shown in Fig. 2.

Formally, for each pose $X^k = \{x_i\}^k$ defined as a set of joint positions x_i, we optimize for a reduced rigid transformation M composed of a rotation around the y axis R_y, and translations along the x and z axises T_x, T_z—resulting in $M = T_y\, T_x\, R_y$ and shown in Fig.2—that minimizes the similarity cost between the 3D projected joint positions $P\ M\, x_i$ and the 2D joint positions y_i, where P is the view and projection transformation of the camera (see next section for how we estimate the mobile camera's parameters). Finally, we go through all the optimal transformations and poses pairs $< X^k, M >$, and pick the one that has the smallest cost value, resulting in the following optimization problem:

$$X^*, M^* = \underset{<X^k,M>}{\text{argmin}}\ \min_{M} \sum_i ||y_i - P\ M\, x_i||^2. \qquad (1)$$

We solve the internal optimization for the transformation M using gradient-based optimization along numerical derivatives. This requires initializing the 3D pose front facing the camera as to ensure convergence towards a sensible solution.

We described how to match the 3D pose to the 2D skeleton, but this depends on 3D camera parameters for the projection. Next we describe how we estimate these for the mobile device given a know *a priori* marker in the scene.

5 Augmentation and Mobile Setup

To incorporate a 3D character into the real world picture using a mobile device, we need to estimate the camera parameters: a view and perspective matrix. The perspective matrix is given by the device, while we use marker-base technology (*Vuforia*) [Vuforia 2017] to recognize and track the camera's transformations. We print a real world marker that is about the size of a person, and process the texture for visual features. When the mobile device takes a picture, it contains the marker, which is then used to estimate the orientation and position of the camera.

The 3D character pose used in the optimization (section 4), is initialized to roughly fit inside the bounding box of the marker. The optimization adjusts the character's depth translation to match the same size as the person's 2D skeleton. If the character is to be smaller, (e.g. a dwarf) we wait until the end of the optimization, to scale the final 3D pose back to its original size.

Finally, the neural network we use in section 3 (OpenPose) to estimate the 2D pose of a person is sizable and runs optimally on a graphics card. Deploying such a system on a mobile device represents a significant integration effort, and will suffer from a loss of performance due to the difference in hardware. Our solution is to place the 2D pose estimation “in the cloud”, and send messages between the mobile device taking pictures, and the 2D pose estimation running on a server.

6 Results and Discussion

We designed a creative concept around Space Exploration that resulted in 12 relevant poses. We started with a set of 10 poses, and invited people to experience the appplication. The subjects performed poses we did not have, which we then crafted and included in the dataset, removing the ones that were not relevant. After two such iterations, we 12 relevant poses shown below.

The pictures were taken from a mobile device, sent to a server for the 2D skeleton estimation (running OpenPose), and then the 3D pose matching was performed on the mobile device. The whole process took about 2 seconds.

6.1 Limitations

The sum of joint positions that we minimize is successful at matching the shape of the character, but does not always succeed at finding a perceptually similar size for the character. It can be seen in our results that sometimes the character is larger than others. We could fix this with a final pass that adjusts the size based on the shoulder and feet proportions, which seem to be visually important.

Naturally, poses not present in the database fail to be discovered. This is a limitation by design. Also, at the moment we only tackled and demonstrated pose similarity for body joints—excluding the face and the hands. In consequence, similar body poses that have different hand gestures will fail to be discriminated. We think this could be tackled with a 2 step matching where first the full body is matched, then the different hand poses are considered.

7 Conclusion and Future Work

We proposed a practical approach to produce augmented reality selfies with digital characters. It relies on a set of predefined poses that are automatically selected and adjusted based on a 2D pose estimate of the character. While a few minor improvements are required to match people of different sizes, it unlocks possibilities to investigate new interactions not yet described in this paper. For example, we have the digital character be worn as a suit by the person in the picture—similar to augmenting clothes. We could estimate the shape of the person from a humanoid and utilize this 3D geometry estimation to support partial occlusions as well as casting approximate shadows from the subject to the character.

References

ANDRILUKA, M., PISHCHULIN, L., GEHLER, P., AND SCHIELE, B. 2014. 2d human pose estimation: New benchmark and state of the art analysis. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

BARAKONYI, I., AND SCHMALSTIEG, D. 2006. Ubiquitous animated agents for augmented reality. In *Proceedings of the 5th IEEE and ACM International Symposium on Mixed and Augmented Reality*, IEEE Computer Society, Washington, DC, USA, ISMAR '06, 145–154.

BUYS, K., CAGNIART, C., BAKSHEEV, A., LAET, T. D., SCHUTTER, J. D., AND PANTOFARU, C. 2014. An adaptable system for rgb-d based human body detection and pose estimation. *Journal of Visual Communication and Image Representation 25*, 1, 39 – 52. Visual Understanding and Applications with RGB-D Cameras.

CAO, Z., SIMON, T., WEI, S., AND SHEIKH, Y. 2016. Realtime multi-person 2d pose estimation using part affinity fields. *CoRR abs/1611.08050*.

CHEN, C., AND RAMANAN, D. 2016. 3d human pose estimation = 2d pose estimation + matching. *CoRR abs/1612.06524*.

CIMEN, G., YUAN, Y., SUMNER, R., COROS, S., AND GUAY, M. 2018. Interacting with intelligent characters in ar. *International SERIES on Information Systems and Management in Creative eMedia (CreMedia)*, 2017/2.

KAZEMI, V., AND SULLIVAN, J. 2014. One millisecond face alignment with an ensemble of regression trees. In *Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition*, IEEE Computer Society, Washington, DC, USA, CVPR '14, 1867–1874.

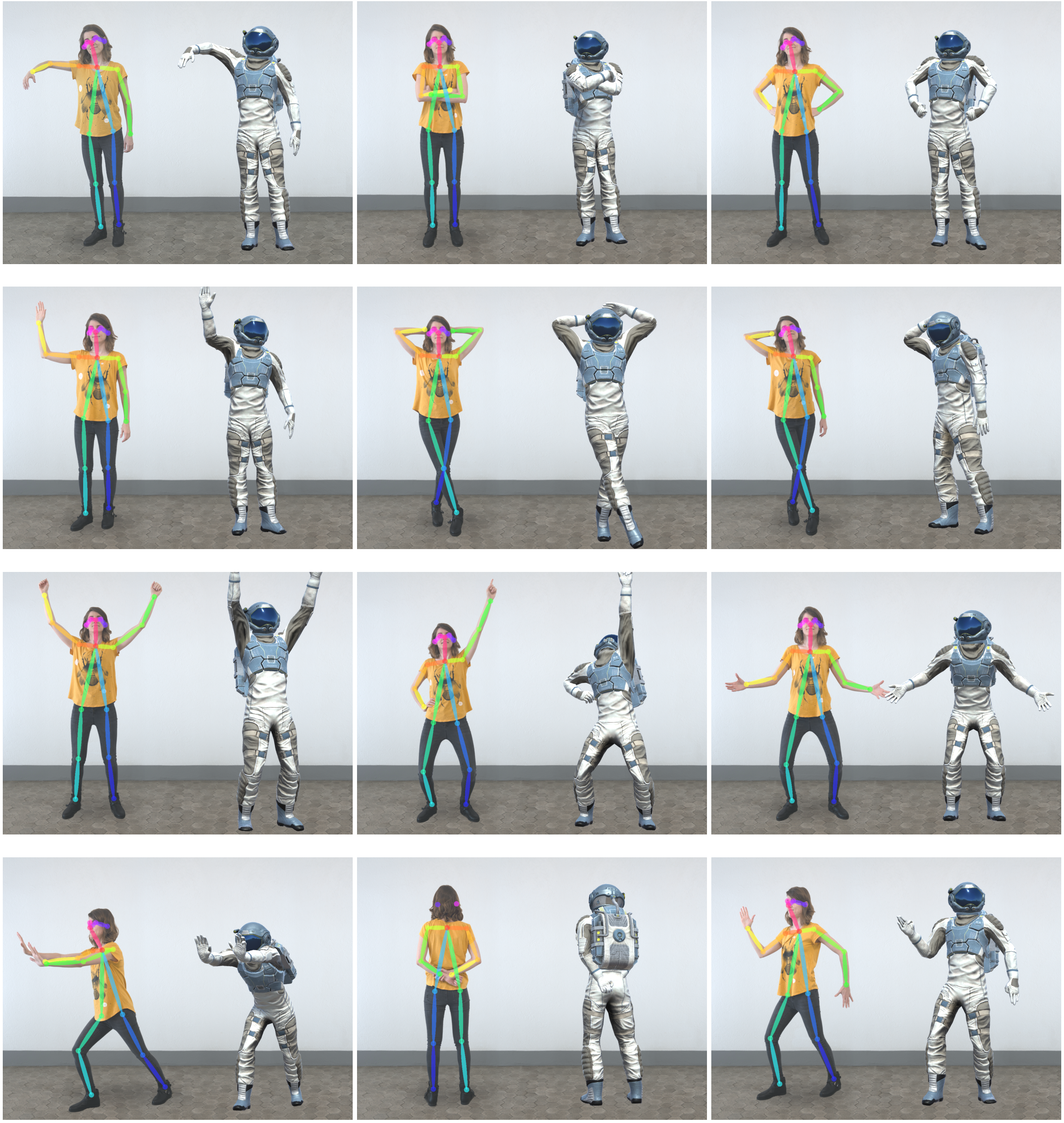
LIN, T., MAIRE, M., BELONGIE, S. J., BOURDEV, L. D., GIRSHICK, R. B., HAYS, J., PERONA, P., RAMANAN, D., DOLLÁR, P., AND ZITNICK, C. L. 2014. Microsoft COCO: common objects in context. *CoRR abs/1405.0312*.

MARTINEZ, J., HOSSAIN, R., ROMERO, J., AND LITTLE, J. J. 2017. A simple yet effective baseline for 3d human pose estimation. *CoRR abs/1705.03098*.

MCINTOSH, K., MARS, J., KRAHE, J., MCCANN, J., RIVERA, A., MARSICO, J., ISRAR, A., LAWSON, S., AND MAHLER, M. Magic bench: A multi-user, multi-sensory ar/mr platform. In *ACM SIGGRAPH 2017 VR Village*.

MEHTA, D., RHODIN, H., CASAS, D., FUA, P., SOTNYCHENKO, O., XU, W., AND THEOBALT, C. 2017. Monocular 3d human pose estimation in the wild using improved cnn supervision. In *2017 Fifth International Conference on 3D Vision (3DV)*.

MEHTA, D., SRIDHAR, S., SOTNYCHENKO, O., RHODIN, H., SHAFIEI, M., SEIDEL, H.-P., XU, W., CASAS, D., AND



THEOBALT, C. 2017. Vnect: Real-time 3d human pose estimation with a single rgb camera. *ACM Trans. Graph. 36*, 4, 44:1–44:14.

ROGGE, L., NEUMANN, T., WACKER, M., AND MAGNOR, M. 2011. Monocular pose reconstruction for an augmented reality clothing system. In *Proc. Vision, Modeling and Visualization (VMV)*, Eurographics, 339–346.

ROGGE, L., KLOSE, F., STENGEL, M., EISEMANN, M., AND MAGNOR, M. 2014. Garment replacement in monocular video sequences. *ACM Trans. Graph. 34*, 1 (Dec.), 6:1–6:10.

SHOTTON, J., GIRSHICK, R., FITZGIBBON, A., SHARP, T., COOK, M., FINOCCHIO, M., MOORE, R., KOHLI, P., CRIMINISI, A., KIPMAN, A., AND BLAKE, A. 2012. Efficient human pose estimation from single depth images. IEEE.

TOMÈ, D., RUSSELL, C., AND AGAPITO, L. 2017. Lifting from the deep: Convolutional 3d pose estimation from a single image. *CoRR abs/1701.00295*.

VUFORIA, 2017. Qualcomm, http://www.qualcomm.com/vuforia.

WANG, C., WANG, Y., LIN, Z., YUILLE, A. L., AND GAO, W. 2014. Robust estimation of 3d human poses from a single image. *CoRR abs/1406.2282*.

YASIN, H., IQBAL, U., KRÜGER, B., WEBER, A., AND GALL, J. 2015. 3d pose estimation from a single monocular image. *CoRR abs/1509.06720*.

ZIMMERMANN, C., WELSCHEHOLD, T., DORNHEGE, C., BURGARD, W., AND BROX, T. 2018. 3d human pose estimation in RGBD images for robotic task learning. *CoRR abs/1803.02622*.

ZÜND, F., LANCELLE, M., RYFFEL, M., SUMNER, R. W., MITCHELL, K., AND GROSS, M. 2014. Influence of animated reality mixing techniques on user experience. In *Proceedings of the Seventh International Conference on Motion in Games*, ACM, MIG '14, 125–132.