AR Costumes: Automatically Augmenting Watertight Costumes from a single RGB Image

Christoph Maurhofer
ETH Zurich

Gokcen Cimen
ETH Zurich

Mattia Ryffel
Disney Research

Robert W. Sumner
Disney Research

Martin Guay
Disney Research

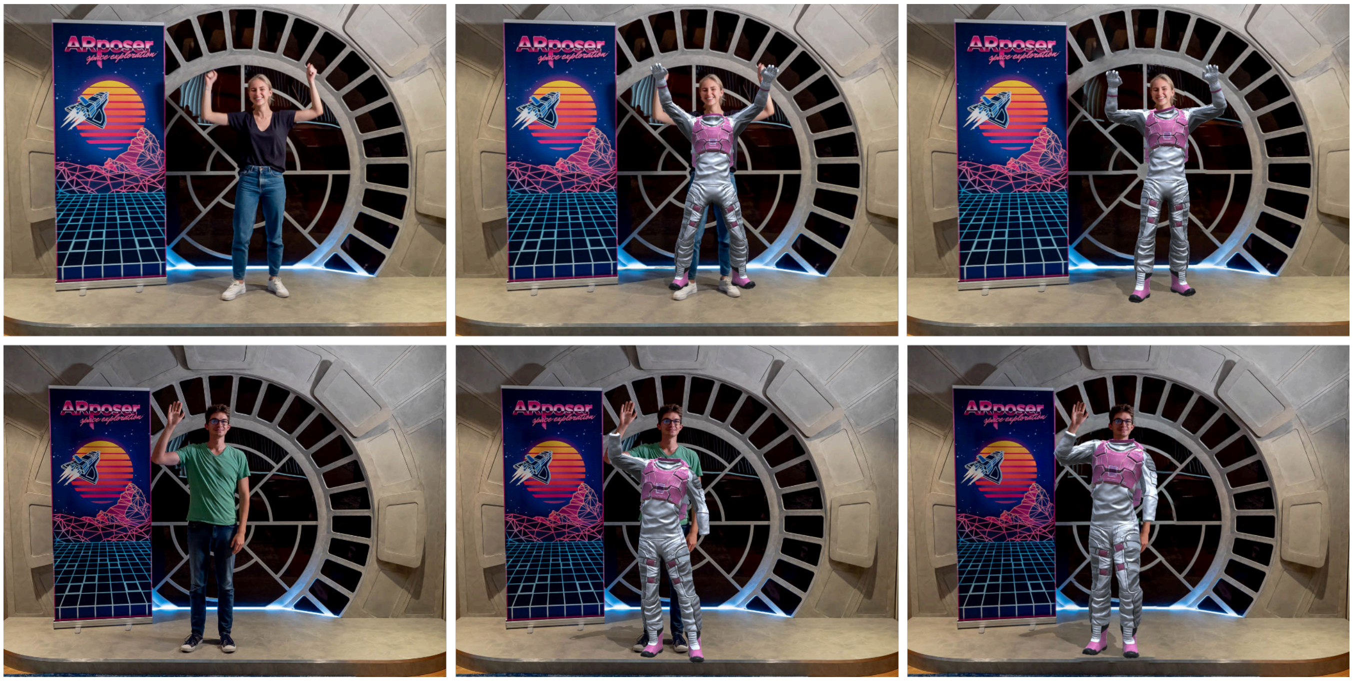


Figure 1: Naively matching a 3D costume pose to person's pose (middle column), results in several parts of the person visible. In this paper, we solve these problems with *shape* estimation of the costume, together with inpainting of the person's body.

ABSTRACT

We describe a method to automatically augment a watertight digital costume onto a person's body from a monocular RGB image. When overlaying a digital costume onto a body using pose matching, several parts of the person's cloth or skin remain visible due to differences in shape and proportions. In this paper, we present a practical solution to these artifacts which requires minimal costume parameterization work, and a straightforward inpainting approach. To our knowledge, our approach is the first to deliver plausible watertight costumes from RGB imagery only, and is compatible with mobile devices. We believe this can serve as a useful baseline for future improvements and comparisons.

CCS CONCEPTS

•**Computing methodologies →Mixed / augmented reality;** *Motion capture; Image manipulation;* Image processing;

KEYWORDS

Costume Fitting, Costume Augmentation, Augmented Reality, Shape Estimation, Inpainting

ACM Reference format:
Christoph Maurhofer, Gokcen Cimen, Mattia Ryffel, Robert W. Sumner, and Martin Guay. 2018. AR Costumes: Automatically Augmenting Watertight
Costumes from a single RGB Image. In *Proceedings of ACM conference on Virtual-Reality Continuum and its Applications in Industry, Tokyo, December 2018 (VRCAI'18),* 13 pages.
DOI: 10.475/123_4

1 INTRODUCTION

Imagine taking a selfie and magically wearing your favorite character or hero's suit. While we did see digital cloth added onto people in the past, it was often with a depth camera such as a *Kinect*, which is not always reliable in outdoor conditions, and is not as widespread as monocular cameras on mobile devices. In this paper, we carry out this concept from a single RGB image, in a manner compatible with mobile devices.

People come in different shapes and sizes, and estimating the best costume to fit their given pose and proportions is a challenge. While recent work supports estimating shapes, it might not be the desired solution to fully cover the body: artistic direction might require the shape to remain slender or muscular, for example. Hence, we approach this problem with a costume parametrization based on different skeleton *proportions* (variations in limb lengths such as legs and spine), and combine this with *inpainting* to remove the remaining visible parts, such as cloth or skin from the person behind, as shown in Fig.1.

Our solution is practical and requires minimal parametrization work. Given a 3D costume, we manually create different versions associated to a 3D skeleton of different proportions. Together with a data set of poses, we optimize for the best matching 3D costume to the person's 2D skeleton (estimated from the RGB image using a 2D pose tracker). Once the best matching shape (pose and proportions) is found, we need to remove the remaining visible regions of the person. To solve this, we estimate the person's body mask we want to remove (e.g. the body, but without the head or hands), and proceed with inpainting the masked region. To inpaint, we capture the background image without the person, and then compute a projective transform—or *homography*—from four feature points in the source image to the target image, followed by Poisson image editing to match the surrounding color and lighting.

The entire process runs in about ten seconds on a *Surface Book*, without an optimized solution, and can produce high resolution images if required. We show the results of our approach on various poses and people; some of which are systematically successful. To measure the quality of the results, we performed a qualitative study quantifying the average likability of a set of poses across multiple people. To evaluate the value of the different steps of our method, we performed an ablation study showing the effect of each step of our method.

2 RELATED WORK

The vision of digitally augmenting the real world was first introduced over fifty years ago [38], and has since been revisited countless times as progress in hardware, computer vision and synthetic imagery continues to be made—unlocking new possibilities for communication, education and entertainment.

A good example could be the different real-time interactions between digital characters and real objects possible using marker-based tracking, first explored in 2006 by Barakonyi's and his colleagues [3]. With recent progress in physics-based modeling of character motion, Cimen et al. demonstrated more realistic reactions and perturbations caused by real world objects onto the digital character [12], increasing the realism of such experiences.

With progress in human tracking, several augmentation concepts have been explored around body, face and hair—allowing people to try on virtual make-up and glasses using face tracking [22], hairs styles [24] using hair tracking, and clothing [15, 34, 35, 43]. In this work, we are focused on fitting a *watertight* costume to a body from a single RGB image. Most cloth augmentation systems such as Facecake's *Swivel* [15] utilizes depth imagery (e.g. a Kinect) (which is not reliable in outdoor conditions), constrained monocular settings with a tracking costume and a uniform background [35], or manual semi-automatic tracking [34]. While recent progress in pose tracking from RGB imagery [10, 40, 45] has made it possible to match the pose of a cloth automatically from RGB imagery, it still leaves parts of the body behind visible (as shown in Fig.1), and does not take into account differences in shape and proportions. In this paper, we combine not only pose, but also shape estimation together with *inpainting* to realise this concept of an automatic *watertight* costume. And we now describe in more detail the recent progress in these areas.

Pose Estimation Estimating a 3D pose from a single monocular image is an inherently ambiguous problem due to depth and self-occlusions. By synthesizing large data-sets of depth image and pose pairs using a data-set of poses and cloth textures, it is possible to regress models to classify pixels as belonging to different body parts, or to directly output skeleton poses [7, 37, 49]. With deep learning, this approach has also been shown to work with monocular RGB images only [30, 31, 39, 44]. The main drawback of this approach is the lack of available dataset, the effort required to synthesize all the pose-image pairs. A practical approach is to utilize the data-sets that have been labeled with 2D joint positions, train a 2D pose tracker such as the one proposed by Cao et al. [9], and to then lift the pose in 3D using different approaches [10, 11, 40, 45]. While pose estimation is sufficient for many applications, it falls short when fitting cloth onto a person, which requires a good estimate of shape.

Shape Estimation Estimating shapes from a single RGB image is often approached using a template, or parametric body model such as SCAPE [2] or SMPL [29], which is optimized to match image features such as silhouettes and 2D joints [5, 16]. Automatic methods which yield accurate 3D shapes are restricted to naked humans or tight clothing. A user can be involved to help the fitting process [48]. Instead of this traditional multi-step approach where a 2D skeleton is first estimated and then a 3D shape fitted, recent work has trained 3D shape inference from a single RGB through 2D landmark correspondence loss [23] adversely with a 3D shape consistency discriminator. It is worth taking a look at *DensePose* [18], which trained a deep net on a newly annotated dataset of 2D body part patches. While these works are exciting, only a few of the above mentioned data-sets are available publicly and for commercial purposes. Hence in this paper, we used a two step approach which first estimates a 2D skeleton, then optimizations for the best shape—similarly to Keep it SIMPL [5], but with a simpler, hand-crafted parameterization of the digital costume.

Inpainting To remove artifacts such as cloth and skin when overlaying a digital costume, we need to fill the pixels with plausible information such as the background behind the person. Inpainting is the name of the trade for filling or restoring image gaps, and is a well studied problem in image processing [17]. There is a dichotomy between methods that are aware of the background [20, 25], and methods that "do their best" without [4, 13, 26, 42]. These later methods seek to propagate the information in the best possible way according to boundaries [4], or to texturize the missing information in a statistically consistent manner with respect to local and global structures [13, 26, 42]. Best results are nowadays obtained with deep nets, which can learn higher order statics on images by training on very large data sets [21, 28, 32, 46]. Without knowledge of the background, these methods do not yield consistent nor exact results. Using a pre-scanned background, Hays and Efros [20] search for the closest image in a large database, and fill the missing parts by cutting & pasting. Whyte et al. [41] also focused on other images, but of the same scenes, using a homography transformation between 4 points in the scene to map source and target images. Copy & pasting often holds artifacts with edge disconnects. Darabi et al. [14] proposed an advanced blending scheme that mixes different sources for a smoother fill. More recently, Klose et al. [25] use the captured images along camera parameters to estimate the scene information (depth and color), and to then re-project the final visible color. In our work, we choose the practical approach of using a homography (as in [41]), followed by Poisson image editing to restore lighting color and ensure smooth edge continuity.

Photoreal Digital Try On Recent progress in deep learning applied to images has enabled virtual *try-on* of specific pieces of cloth [19], or of full body appearance [47]. These methods utilize large datasets of photoreal images. In fact, some recent work such as [19] utilize no 3D information at all and are thus quite different from our approach. We believe it is an exciting direction and it would be interesting to evaluate these approaches on stylized characters and depictions, or study how our synthetic compositions could enhance the learning process.

3 3D COSTUME SHAPE

To summarize, our approach breaks down the problem of costume fitting from a single RGB image into two main parts: a shape estimation described in this Section, followed by *mask* estimation and *inpainting* for the *remaining* visible parts, described in Section 4.

Hence given an RGB image containing a single person, our goal is to find a costume shape which best fits the pose *and* proportions of the person. Our 3D shape estimation follows a 2D inference plus 3D matching type of approach as in [10, 11], but extended with estimating proportions (Section 3.2) followed by refinement (Section 3.3).

We first estimate the 2D skeleton with joint positions y_i of the person using a deep neural network [9] trained on labelled 2D joint data COCO [27] and MPII [1]. We then parameterized the 3D costume mesh with different 3D skeleton poses k *and* proportions c (variations in limb lengths), resulting in $p = c \times k$ shapes in our data set.

From the 2D skeleton, we optimize for the closest 3D pose k^* in Section 3.1, then search for the optimal proportions c^* using a heuristic that favors shoulders and hips for closer perceptual similarity in Section 3.2. The final pose is close to the 2D skeleton, but could still be refined. Hence we perform a final full space refinement optimization to match more exactly the limb directions and joint positions of the 2D skeleton, as described in Section 3.3.

3.1 3D Pose Match

For each pose k in the data set, we optimize for the rigid transformation that will bring the 3D pose, closest to the 2D projected skeleton, in terms of joint positions similarity. The global transformation of a 3D pose is parameterized with 4 degrees of freedom: one rotation around the y axis, together with three global translations.

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 three translations T—resulting in $M = T\,R_y$—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. Finally, we go through all the optimal transformations and pose pairs k, M, and pick the one that has the smallest cost value, resulting in the following optimization problem:

$$k^*, M^* = \underset{k}{\operatorname{argmin}} \min_{M}\ E_p = \sum_{i}^{|X^k|} ||y_i - P\ M\,x_i||^2. \qquad (1)$$

We optimize 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.

3.2 Proportions Estimate

Given our closest pose k^*, we seek to choose the closest matching proportions c^* to better fit the 2D skeleton. In our experiments, we found that comparing the sum of all joint positions, such as in the previous section, did not lead to perceptually similar proportions, or resulted in confusing the optimization (1) into the wrong pose. We found that focusing on the shoulders and hips, which are visually more prominent, yielded better results perceptually, and more robust pose and proportions pairs.

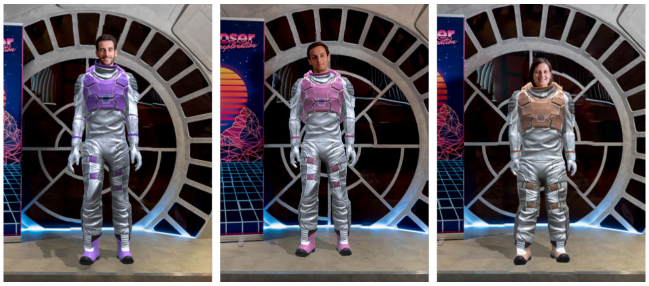


Figure 2: Our three shapes with variations in limb lenghts. The arms and legs are longer on the left, and shorter on the right. A better estimate of the proportions helps the refinement of the pose converge to a better solution.

Our selection criteria is based on two features $f = \left[f_{s/w}, f_{h/w}\right]$ measuring the shoulder-to-waist ratio $f_{s/w}$, and the shoulder width versus average upper body height ratio $f_{s/h}$, defined as:

$$f_{s/w} = \frac{|\ S_L - S_R\ |}{|\ H_L - H_R\ |}$$
$$f_{s/h} = \frac{2 \cdot |\ S_L - S_R\ |}{|\ S_L - H_L\ | + |\ S_R - H_R\ |},$$

where S_L and S_R are the left and right shoulders, and H_L, H_R the hips of the skeleton in 3D.

We select the 3D shape c which has the closest feature vector to the *target* 2D skeleton features f_t when inverse projected onto a plane centered on the 3D costume. Specifically, we pick the shape c that minimizes the weighted sum at the L2 norm:

$$c^* = \underset{c}{\operatorname{argmin}} \|w \;\; [f_t - f_c]^T \;\|^2.$$

where $w = [w_0 \; w_1]$ are both equal to 1 in our implementation.

While there is a variety of different proportions in people, we found that three main modes ($| \; c \; |= 3$) was sufficient to span most of our subjects, and represented a satisfactory compromise between speed, set-up complexity and quality. Additionally, the refinement step discussed next can contribute to fixing slight proportion mismatches in 2D, as we optimize in 3D allowing the limbs to visual shorten when projected onto the screen.

3.3 Global-local Refinement

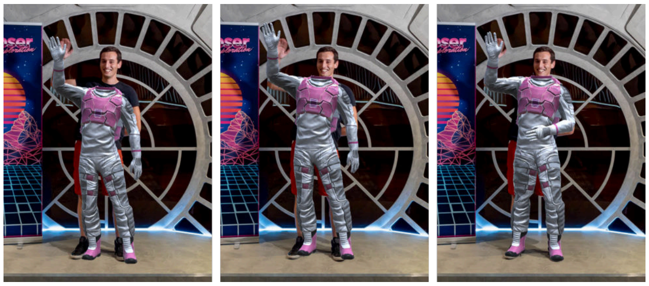


Figure 3: We optimize globally for the root position to adjust the scale, and alternate with local optimization of the joint angles in a back-and-forth manner, to finally converge to a well matching pose.

At this point, we have a 3D shape (pose k^* and proportions c^*) which is close to the person's shape, but is still different in the exact bone orientation and joint position, as shown on the left in Fig.3. To remove these differences, we perform an additional refinement step with respect to the full degrees of freedom of the 3D character: the joint orientations $Q = q_i$ and the root position x_0 of the character. Because bone positions may not match exactly, we weight down this objective in optimization (1), and add an additional objective function which seeks to match the bone *directions*, resulting in the following optimization:

$$Q^*, x_0^* = \min_{Q, x_0} w_p \; E_p + w_{dir} \; E_{dir},$$

$$E_{dir} = \sum_i \|(y_i - y_{p(i)}) - (PM^* x_i - PM^* x_{p(i)})\|^2,$$

where $p(i)$ is the parent of i. We solve this problem in a global / local fashion where we optimize for the global position while keeping the orientation fixed, and solve for the individual joint orientations while keeping the position fixed. Both of these steps are performed using local gradient descent along numerical derivatives.

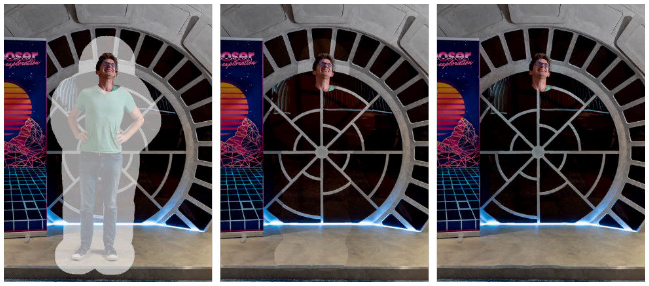


Figure 4: We first estimate the person's mask using the estimated 2D skeleton and *Grabcut.* Then we define a Homography transformation from target image coordinates to source (background) coordinates in order to color the masked pixels. Finally we apply Poisson image editing to fix the remaining color discrepencies.

We now have a costume that matches closely in pose and proportions, but when overlayed over the person, leaves cloth and skin from the person visible, as shown on the right in Fig.3. We remove these in the next section by estimating a 2D mask and inpainting.

4 INPAINTING AND COMPOSITION

The costume shape overlayed on the person at this point still has cloth or skin visible, as shown in Fig. 4. To remove these artifacts, we estimate the 2D mask of the person's body and head, and then inpaint the body area using background information. When rendering the 3D costume, we can obtain an odd look when the lighting and shadows differ from the real world, and when the costume appears plastic or unnatural. Hence we estimate the lighting direction by sampling the picture, and filter the final render to produce a more natural look for the costume.

4.1 Masking

To compute the 2D mask, we use *Grabcut* [36], which requires an initial labelling of the *foreground, probably foreground* and *background* pixels. We use the estimated 2D skeleton, and set *foreground* pixels that are within a distance r of a few pixels of the joint positions, and within $2r$ of the skeleton bones—defined as lines between joints. For the head specifically, we set a slightly larger ellipse to indicate the facial pixels to obtain a more precise boundary. Pixels within a larger radius are marked as *probably foreground*, while the rest remains assumed *background.* We run the algorithm for 5 iterations which yields reasonable results in most cases. With complex backgrounds, it sometimes misclassifies pixels. To circumvent this problem we simply inflate the mask to be inpainted. The final result can be seen in Fig.4.

4.2 Inpainting

Our goal is to color the masked pixels with plausible underlying scene values. Hence we capture the environment (with a video) and seek to find the pixel colors that best match the structure of the captured background, while resembling the colorization of the target picture.

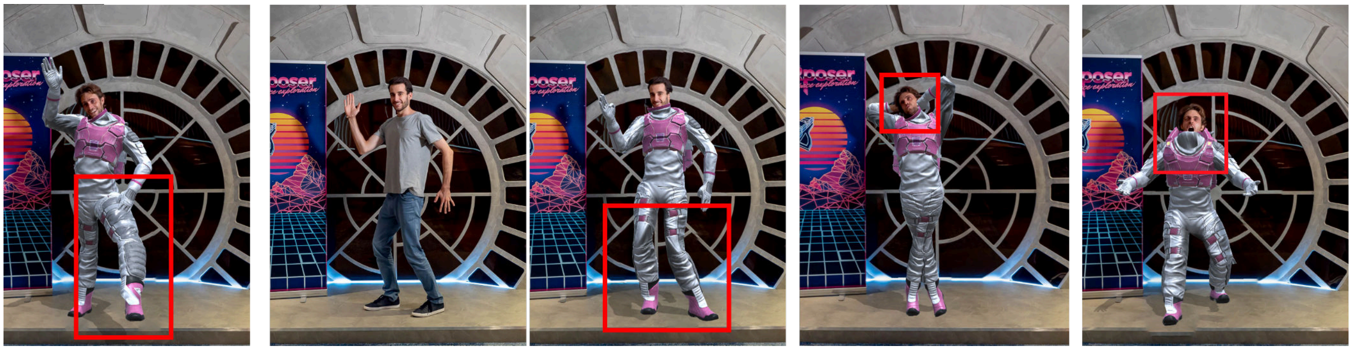
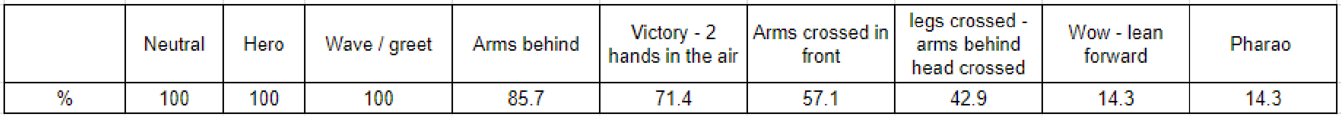


Figure 5: Our optimzation may result in large deformations when misclassifying the person's proportions (left). Another issue is we do not track the 2D feet orientation at the moment, and cannot reproduce this pharao pose at the moment (middle). Similarly limgs crossing are not prevented for the moment in our optimization. Estimating the mask area of the face in the legs crossing pose, without the hands, is challenging. Finally, poses that expose the inner area of the mesh are not taken into account at the moment, and methods to adress this are discussed in our results section.



	Neutral	Hero	Wave / greet	Arms behind	Victory - 2 hands in the air	Arms crossed in front	legs crossed - arms behind head crossed	Wow - lean forward	Pharao
%	100	100	100	85.7	71.4	57.1	42.9	14.3	14.3

Figure 6: Average likeability score for the 9 poses, performed by 7 subjects. Some of the poses are well handled accross people, others yield mitigated likebility, while others are not well handled by our current method.

Our solution consists in computing a *projective* transformation (a.k.a Homography) from the closest matching background with respect to camera parameters, to the new target image, using 4 corresponding points in the images: in our case, the 4 corners of the *AR Poser* poster. When capturing the background, we record the camera position $x^{'}$ and orientation $q^{'}$. Given a new camera position and orientation x and q (at runtime), we search our dataset for the nearest background image. Note that for speed we used a KD-tree.

Given the nearest background image, we want a warping function that maps coordinates x, y in the target image, to coordinates x', y' in the source (background) image. This requires a projective transformation, which is possible in higher dimensions (i.e. 3) using four points that correspond in both images, by using the *homogenous coordinates trick*. We first define a transformation that will map the first 3 canonical coordinates (e.g. $\begin{bmatrix}1 & 0 & 0\end{bmatrix}^T$) to the first three coordinates of the target image:

$$H = \begin{bmatrix}x_1 & x_2 & x_3\end{bmatrix} \begin{bmatrix}\lambda & 0 & 0\\ 0 & \mu & 0\\ 0 & 0 & \tau\end{bmatrix}$$

after dehomonization. For example, the trivial case $H \begin{bmatrix}1 & 0 & 0\end{bmatrix}^T = \begin{bmatrix}\lambda x_1 & \lambda y_1 & \lambda\end{bmatrix}^T$, which after de-homogenization (dividing by λ), results in $\begin{bmatrix}x_1 & y_1 & 1\end{bmatrix}^T$, thus mapping correctly to x_1. To define λ, μ, τ, and thus H, we leverage x_4 and map $\begin{bmatrix}1 & 1 & 1\end{bmatrix}^T$ to the last point x_4, resulting in

$$H \begin{bmatrix}1\\1\\1\end{bmatrix} = x_4 \Leftrightarrow X \begin{bmatrix}\lambda & 0 & 0\\ 0 & \mu & 0\\ 0 & 0 & \tau\end{bmatrix} \begin{bmatrix}1\\1\\1\end{bmatrix} = x_4 \Leftrightarrow \begin{bmatrix}\lambda\\ \mu\\ \tau\end{bmatrix} = X^{-1} x_4.$$

Now that we have H, mapping cannonical points to target points X, we can compute a similar mapping H' from cannonical points to source X'. Hence, we are able to close the loop by inverting H and mapping target points to source points in order to find the corresponding colors, resulting in:

$$\begin{bmatrix}x' & y' & w'\end{bmatrix} = H' H^{-1} \begin{bmatrix}x & y & 1\end{bmatrix} \tag{2}$$

which needs to be de-homogenized to obtain the final coordinates in the source image.

Now sampling pixels from this function yields similar color and structure, but does not ensure boundary smoothness and color consistency, as can be seen in Fig.4. Hence we further optimize the pixel values to blend with the target image by minimizing the target color gradient while preserving the source color gradient—a method known as Poisson image editing [33]. We solve this using an existing packaged solution in OpenCV [6].

4.3 Composition

We render the 3D costume using rasterization rendering in *Unity*. Simply overlaying the inpainted picture with the rendered costume might hide parts of the head of the target person. To avoid this, we attach a simple, transparent 3D object approximating a generic human head to the neck-bone of the character's rig, which acts as a depth mask during the render pass and occludes the relevant parts of the costume.

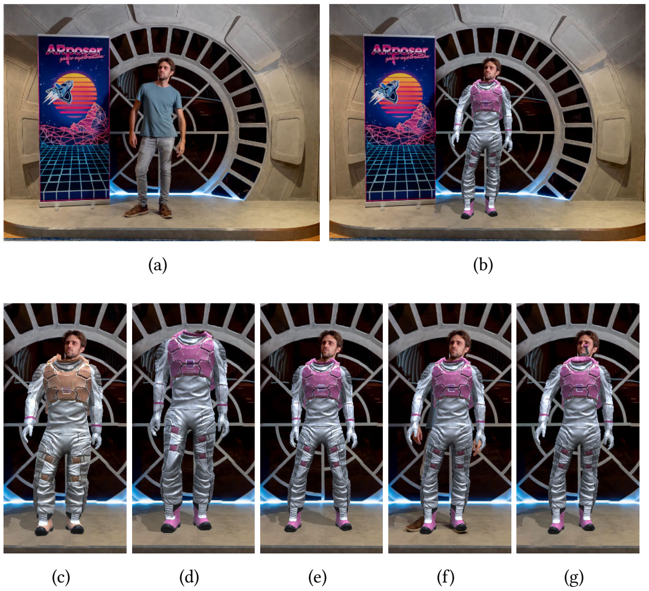


Figure 7: To judge the importance of each step in our method, we performed an ablation study by computing the results with the full pipeline, each time leaving out on step. Image (b) shows the result with all steps applied to source image (a). The bottom row shows the partial results with: (c) no proportion estimate, (d) no size refinement, (e) no bone direction refinement, (f) no inpainting and (g) no approximate head masking.

As for the rendering, we use a single directional light to approximate the scene lighting. When the light direction is different from the one in the picture, the rendering looks odd. Hence we need to find an appropriate lighting direction, which we do by sampling the face of the person in the image. Additionally, Phong shading tends to yield plastic-looking materials, which differs from the overall feel of the picture. Our quick fix is to add noise to the costume's rendering.

To estimate the lighting direction, we use the 2D face landmarks from the 2D pose estimation to sample different points in the source picture. We then sample their HSV values by averaging the neighboring pixels. In particular, we use points around the cheeks and forehead since they tend to have less unwanted noise in comparison to glasses or hair. Thanks to the face joints we can also align a 3D mesh of a face to match the joint positions.

By sampling the same set of points over the 3D mesh, we can read the normal direction of that vertex, and by a weighted average of the normals multiplied by the value of the pixels, we can infer a rough approximation of the direction of the light source. We use the resulting vector to set the new rotation of a directional light that illuminates the virtual costume and creates shadows in the ground. A more accurate approach is described in [8], but an implementation in this context is left for future work.

5 RESULTS AND DISCUSSION

We created a system to accurately overlay a person in a monocular RGB image with a watertight 3D costume matching in proportions and pose. It furthermore improves the quality of the result by removing visible artifacts of the source picture by inpainting the relevant areas, but keeping specific body parts of the target person, resulting in a realistic image composition, at arbitrary resolutions. We performed both an *ablation* study of the different steps of our approach (5.2, demonstrating their effect on the final outcome, as well as a qualitative study (5.1) assessing the quality of the results for different poses (and body proportions).

All results were generated using our in-house implementation: the pictures are taken from a surface book, sent to a server for the 2D skeleton estimation (running a pose tracker on the GPU). The skeleton is sent back to the device which processes the skeleton and image to match the shape and perform inpainting. The whole process takes about 10 seconds, from which two thirds is used by computing the segmentation with *Grabcut* [36] and the inpainting using Poisson image editing [33]. Our code was not optimized for speed.

5.1 Qualitative Study

Our data set holds 12 poses and we performed a qualitative user study of 9 poses, similar to the most recurrent ones people do. We had 7 different person perform the 9 poses. We then showed the results different people and asked to rank the likability of the results as binary value: 1 for like, and 0 for do not like. The average of the evaluations shown in table 6 resulted in 4 of the poses with a success rate above 80 %, with 3 having 100%, 3 having mitigated likability, and 3 being systematically unconvincing (bellow 20%.

The mitigated likeability we believe are due to two main artifacts. We sometimes obtain large deformations when our proportions classification fails, which causes the subsequent refinement stage to over-compensate resulting in la large deformations, as shown in on the left in Fig.5. The second artifact is the collar, which sometimes overlap with the mouth, which changes the nature of the costume. We think this could be addressed by fitting a 3D head model to the person's face, and avoiding interpenetration of the costume with the head.

The systematically unconvincing results we believe are due to poses are method cannot handle properly at the moment. Our optimization (sections 3.1 and 3.2) does not hold 2D feet markers, and so we fixe the orientation the feet—preventing from matching the sideways pose of the *pharao*, as shown in Fig.5. Similarly, we do not avoid intersections between limbs, which can cause the legs crossing pose to fail in most cases. This could be improved with a subspace optimization of the costume shape, or similarly by restricting the bones to anatomically plausible angles. Finally, the "wow" pose which leans forward exposes the inner area of the mesh, which our method does not handle automatically. We would need a 3D model of the person's head to cull the back side of the mesh from being visible after rendering.

5.2 Ablation Study

To evaluate the effect of the different steps as well as their necessity, we generated the results by iteratively leaving one out. Figure 7 shows the results. Refinement and inpainting have the most dramatic effect and leaving them out results in unconvincing compositions. A lesser impactful step is our proportions estimation which selects amongst a few discrete costumes (3 shapes). We observed that an ill-matched character can be adjusted by the refinement process. However, it can be observed that the visual quality of the result is generally better when a costume with a similar body type is selected. The same holds for the method used to color correct the inpainted image material. In many cases, histogram matching is enough to get a convincing result, but the Poisson energy minimization compensates for much more differences in color and can make the difference in more extreme cases.

5.3 Limitations and Future Work

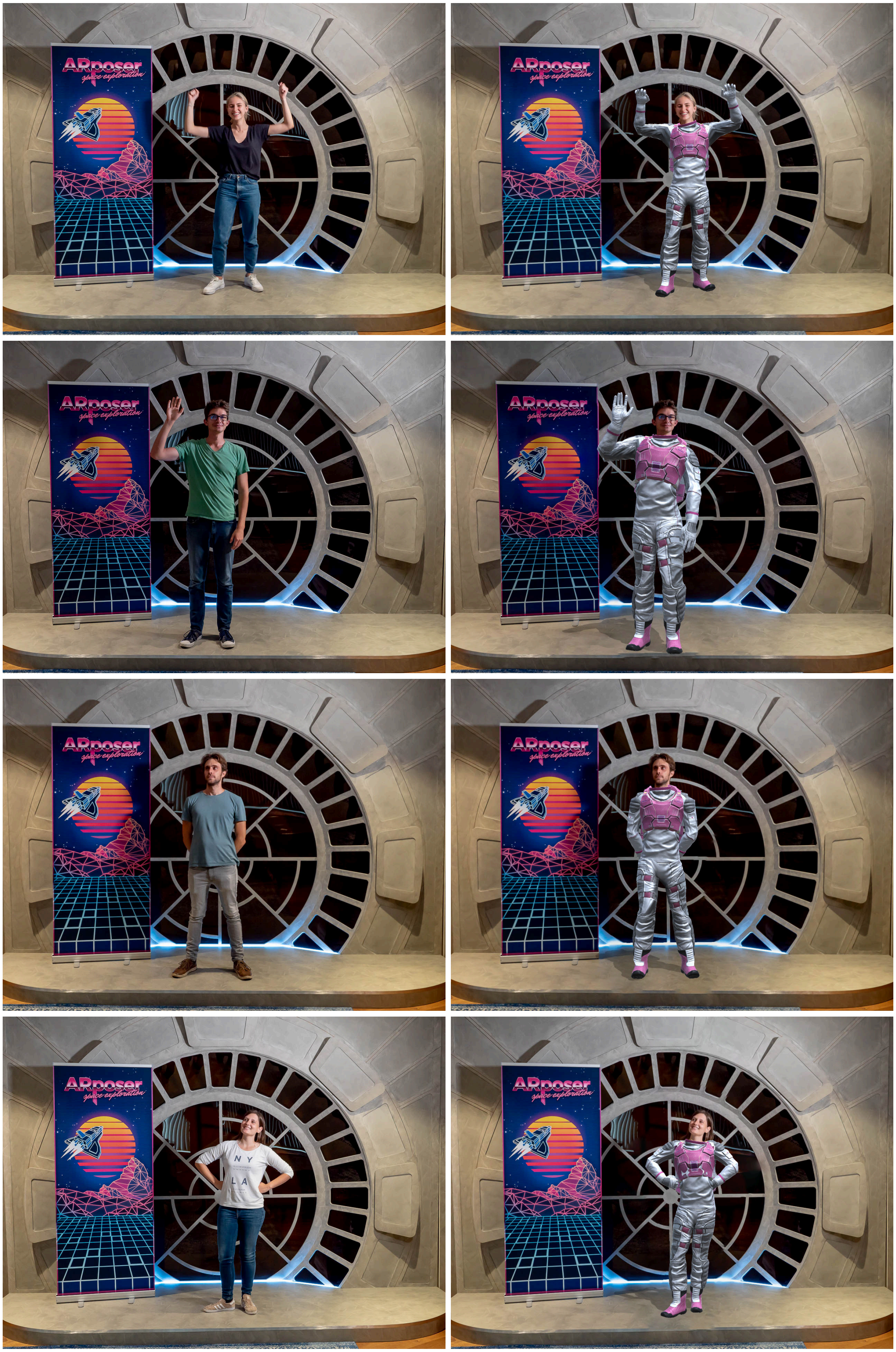
Trying with new characters, different than the astronauts, requires fine tuning parameters in our optimization (sections 3.1 and 3.2). We also observed that we could obtain better results for certain people and poses by tweaking the parameters. Note that we kept them fixed for our evaluation, but it could be interesting to classify the optimization parameters based on the person's picture. Additionally, Our method remains to be tested with children, who have more variation in limb lengths, compared to adults. Finally, we think that accommodating characters that have significantly different or exaggerated limb proportions, such as a cartoon character with tiny legs, would require changing the head position, and thus revisiting our design.

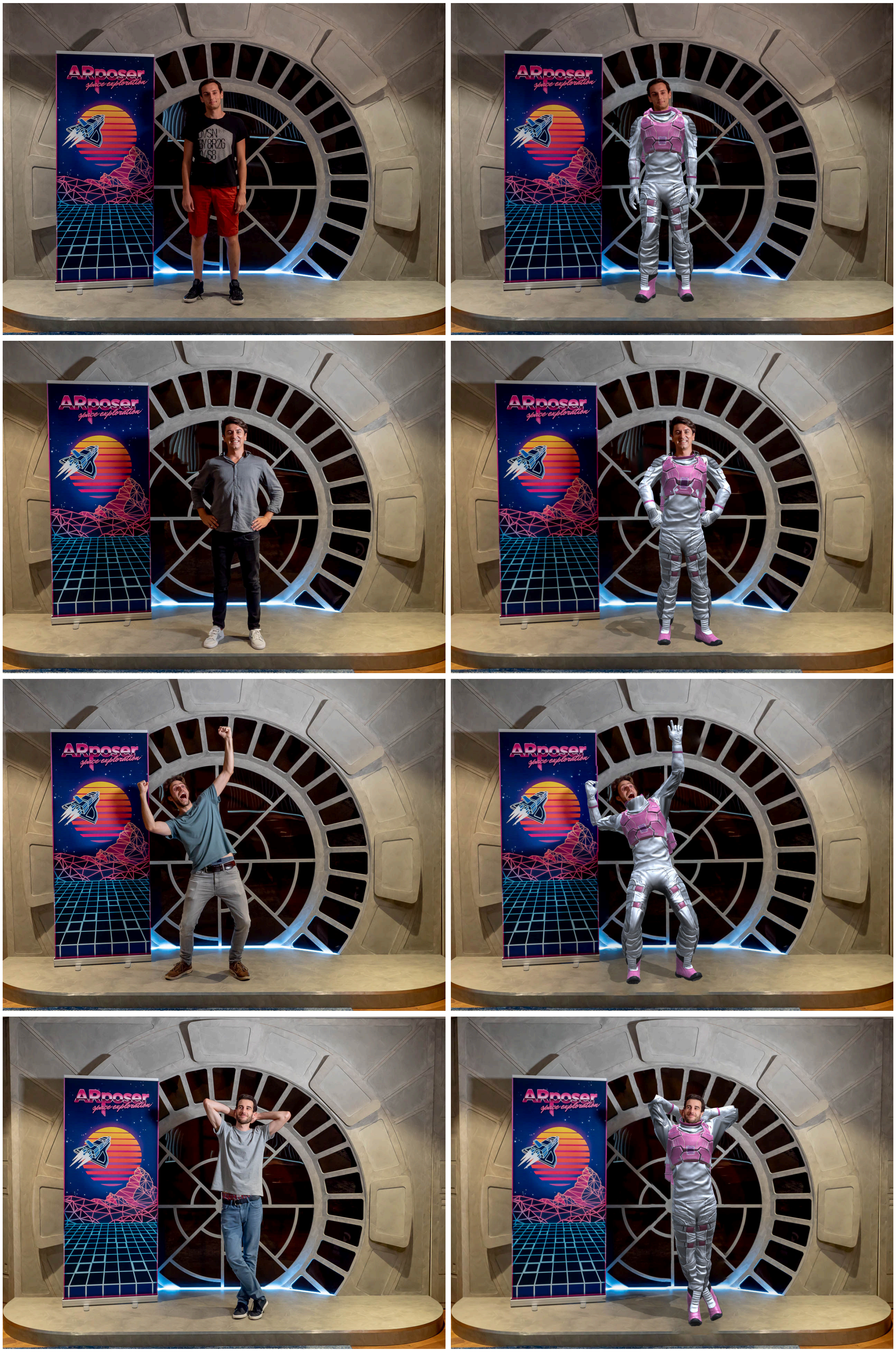
Our inpainting requires scanning the area before hand, which requires starting over when the environment changes. Also, we inpaint using a projection transformation derived from a known marker in the scene (the AR Poser poster), and in the event it changes location, we must rescan the environment once again.

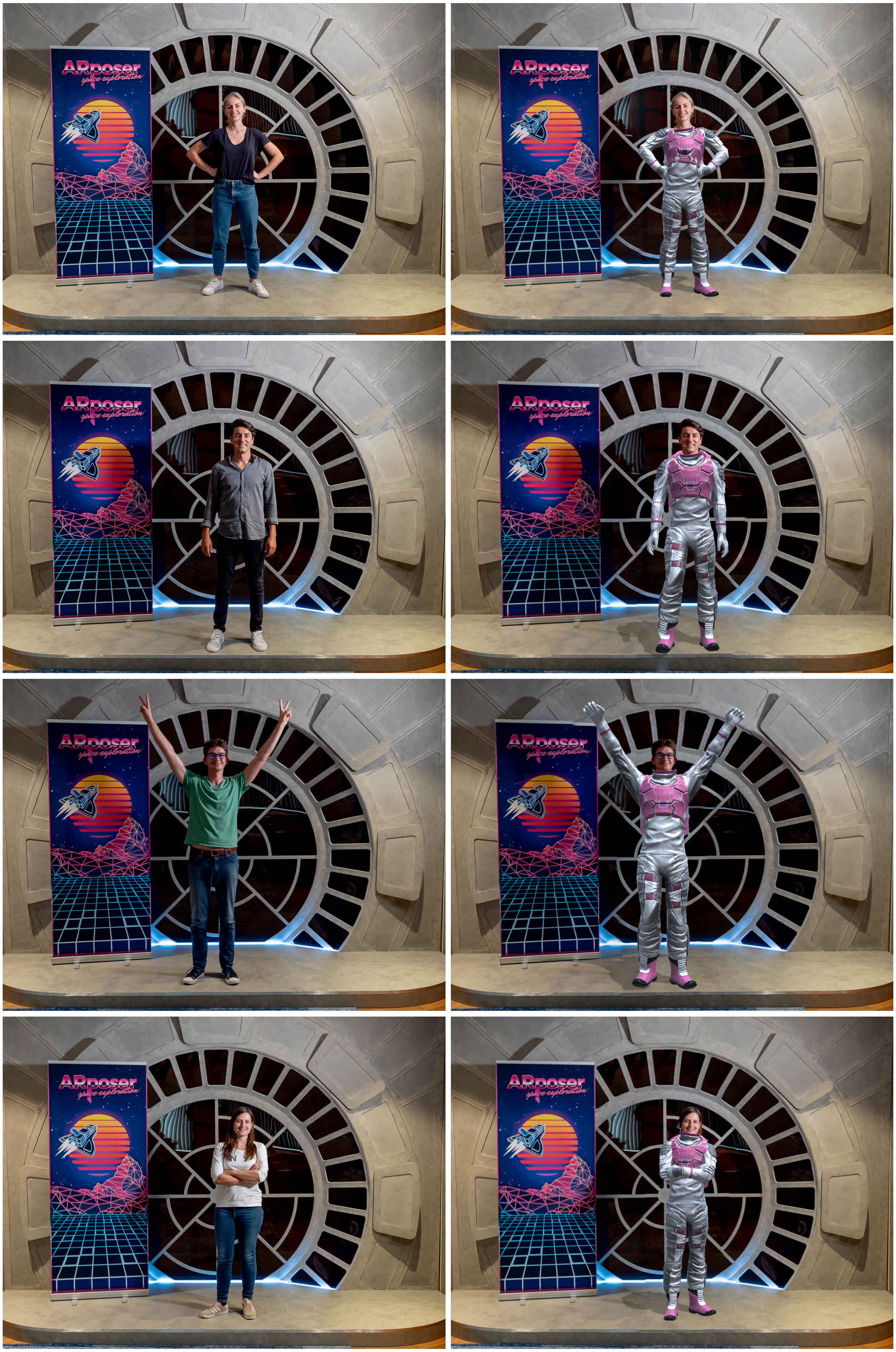
Finally, when masking the target person using *Grabcut* [36], we don't always get a segmentation that is precise enough. This results in body and background parts that are still visible after the inpainting, or it may hide parts of the head. Additionally, the approximated head model used to hide parts of the 3D costume may not be accurate enough (see figure 5). This could be improved by using a more detailed, dynamically adjustable model, to estimate the shape of the person.

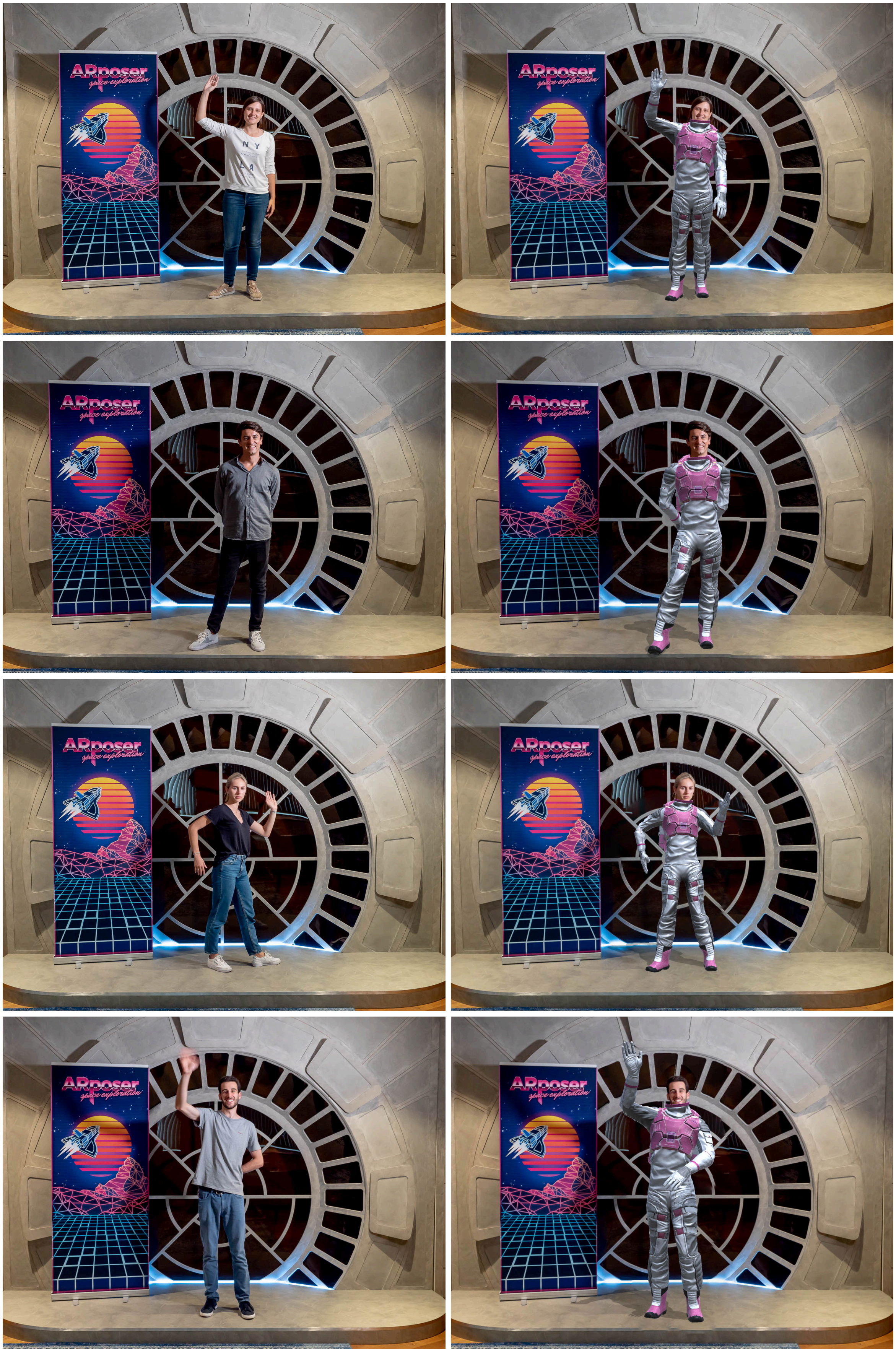
6 CONCLUSION AND FUTURE WORK

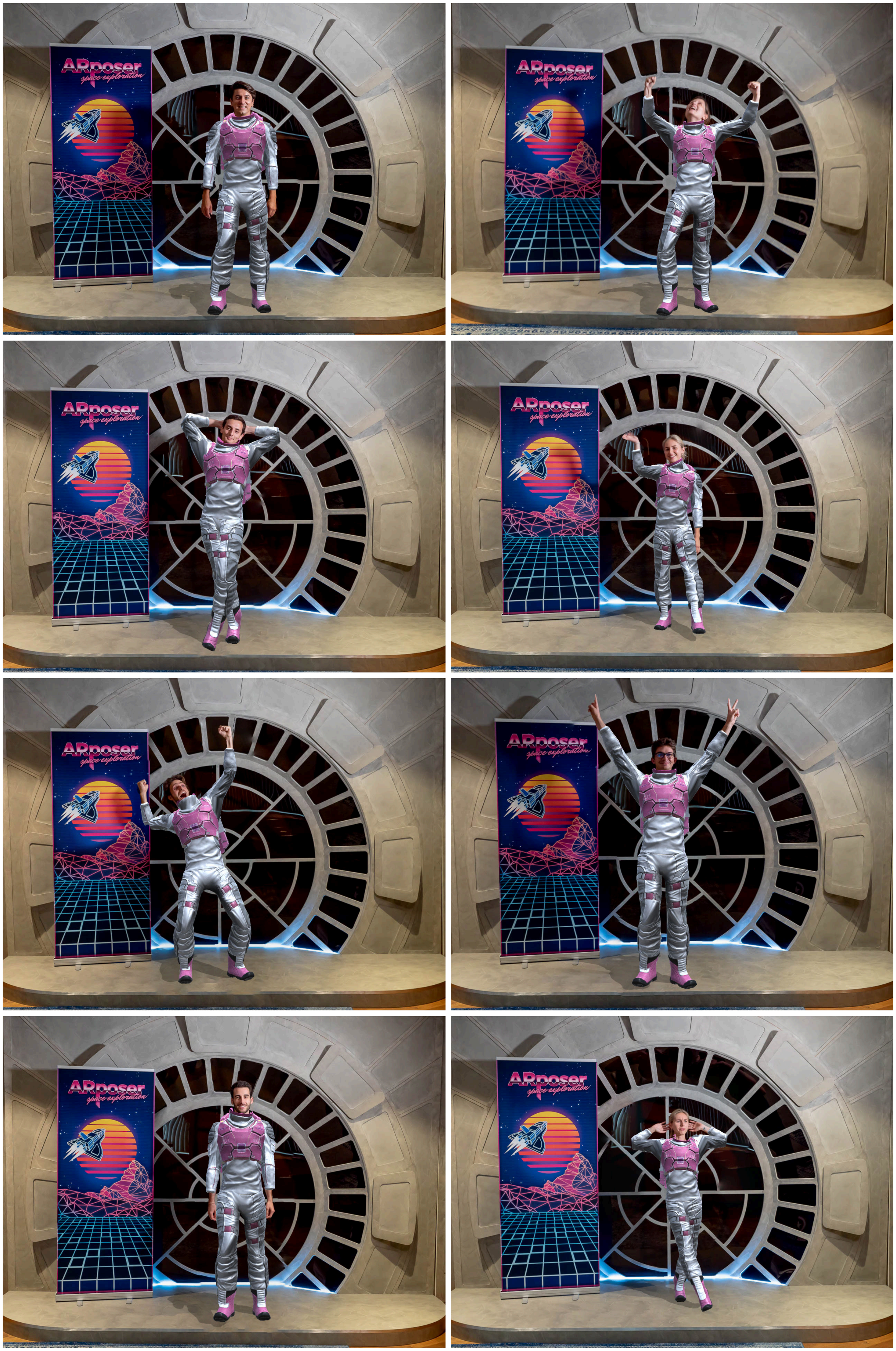
We described a method to automatically fit a watertight costume onto a person from a single RGB image, at arbitrary resolutions. Our approach is based on a 3D matching type of approach, and handles the inpainting problems associated with seamlessly compositing the costume onto the person. With this approach, we were able to handle a variety of poses, and people of different proportions. Our approach is not perfect and we provide examples of the failure cases as well. We believe this will help future comparisons and motivate improvements. For example, our approach could be improved with more robust pose and shape estimation—perhaps with deep learning—and eventually run in real-time. In the near future, we would like to augment cartoon costumes with exaggerated proportions, which cannot be handled by our current method.


ARposer
space exploration


ARposer
space exploration


ARposer
space exploration


ARposer
space exploration


ARposer
space exploration

REFERENCES

[1] Mykhaylo Andriluka, Leonid Pishchulin, Peter V. Gehler, and Bernt Schiele. 2014. 2D Human Pose Estimation: New Benchmark and State of the Art Analysis. *2014 IEEE Conference on Computer Vision and Pattern Recognition* (2014), 3686–3693.

[2] Dragomir Anguelov, Praveen Srinivasan, Daphne Koller, Sebastian Thrun, Jim Rodgers, and James Davis. 2005. SCAPE: shape completion and animation of people. *ACM Trans. Graph.* 24 (2005), 408–416.

[3] Istvan Barakonyi and Dieter Schmalstieg. 2006. Ubiquitous Animated Agents for Augmented Reality. In *Proceedings of the 5th IEEE and ACM International Symposium on Mixed and Augmented Reality (ISMAR '06)*. 145–154.

[4] Marcelo Bertalmío, Guillermo Sapiro, Vicent Caselles, and Coloma Ballester. 2000. Image inpainting. In *Proceedings of the 27th Annual Conference on Computer Graphics and Interactive Techniques (SIGGRAPH '00)*. 417–424.

[5] Federica Bogo, Angjoo Kanazawa, Christoph Lassner, Peter V. Gehler, Javier Romero, and Michael J. Black. 2016. Keep it SMPL: Automatic Estimation of 3D Human Pose and Shape from a Single Image. In *Computer Vision (ECCV 2016)*. 561–578.

[6] G. Bradski. 2000. The OpenCV Library. *Dr. Dobb's Journal of Software Tools* (2000).

[7] Koen Buys, Cedric Cagniart, Anatoly Baksheev, Tinne De Laet, Joris De Schutter, and Caroline Pantofaru. 2014. An adaptable system for RGB-D based human body detection and pose estimation. *J. Visual Communication and Image Representation* 25 (2014), 39–52.

[8] Dan A. Calian, Lalonde Jean-Francois, Paulo Gotardo, Tomas Simon, Matthews Iain, and Kenny Mitchell. 2018. From Faces to Outdoor Light Probes. In *Eurographics*.

[9] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. 2017. Realtime Multi-person 2D Pose Estimation Using Part Affinity Fields. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2017), 1302–1310.

[10] Ching-Hang Chen and Deva Ramanan. 2017. 3D Human Pose Estimation = 2D Pose Estimation + Matching. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2017), 5759–5767.

[11] Gokcen Cimen, Christoph Maurhofer, Robert W. Sumner, and Martin Guay. 2018. AR Poser: Automatically Augmenting Mobile Pictures with Digital Avatars Imitating Poses. *12th International Conference on Computer Graphics, Visualization, Computer Vision and Image Processing* (2018).

[12] Gokcen Cimen, Ye Yuan, Robert Sumner, Stelian Coros, and Martin Guay. 2018. Interacting with Intelligent Characters in AR. *International SERIES on Information Systems and Management in Creative eMedia (CreMedia)* 2017/2 (2018), 24–29.

[13] Antonio Criminisi, Patrick Pérez, and Kentaro Toyama. 2004. Region filling and object removal by exemplar-based image inpainting. *IEEE Transactions on Image Processing* 13, 9 (2004), 1200–1212.

[14] Soheil Darabi, Eli Shechtman, Connelly Barnes, Dan B Goldman, and Pradeep Sen. 2012. Image Melding: Combining Inconsistent Images using Patch-based Synthesis. *ACM Transactions on Graphics (TOG) (Proceedings of SIGGRAPH 2012)* 31, 4 (2012).

[15] Facecake. 2015. Swivel. https://www.facecake.com/

[16] Peng Guan, A. Weiss, A. O. Bălan, and M. J. Black. 2009. Estimating human shape and pose from a single image. In *2009 IEEE 12th International Conference on Computer Vision*. 1381–1388.

[17] Christine Guillemot and Olivier Le Meur. 2014. Image Inpainting : Overview and Recent Advances. *IEEE Signal Processing Magazine* 31 (2014), 127–144.

[18] Riza Alp Güler, Natalia Neverova, and Iasonas Kokkinos. 2018. DensePose: Dense Human Pose Estimation In The Wild. *CoRR* abs/1802.00434 (2018). arXiv:1802.00434 http://arxiv.org/abs/1802.00434

[19] Xintong Han, Zuxuan Wu, Zhe Wu, Ruichi Yu, and Larry S. Davis. 2017. VITON: An Image-based Virtual Try-on Network. *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2017).

[20] James Hays and Alexei A Efros. 2007. Scene Completion Using Millions of Photographs. *ACM Transactions on Graphics (SIGGRAPH 2007)* 26, 3 (2007).

[21] Satoshi Iizuka, Edgar Simo-Serra, and Hiroshi Ishikawa. 2017. Globally and locally consistent image completion. *ACM Trans. Graph.* 36 (2017), 107:1–107:14.

[22] Ana Javornik, Yvonne Rogers, Delia Gander, and Ana Maria Moutinho. 2017. MagicFace: Stepping into Character Through an Augmented Reality Mirror. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems (CHI '17)*. 4838–4849.

[23] Angjoo Kanazawa, Michael J. Black, David W. Jacobs, and Jitendra Malik. 2017. End-to-end Recovery of Human Shape and Pose. *CoRR* abs/1712.06584 (2017).

[24] Ira Kemelmacher-Shlizerman. 2016. Transfiguring portraits. *ACM Trans. Graph.* 35 (2016), 94:1–94:8.

[25] Felix Klose, Oliver Wang, Jean-Charles Bazin, Marcus Magnor, and Alexander Sorkine-Hornung. 2015. Sampling Based Scene-space Video Processing. *ACM Trans. Graph.* 34, 4, Article 67 (July 2015), 11 pages.

[26] Nikos Komodakis. 2006. Image Completion Using Global Optimization. *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)* 1 (2006), 442–452.

[27] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, Lubomir D. Bourdev, Ross B. Girshick, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. Microsoft COCO: Common Objects in Context. In *ECCV*.

[28] Guilin Liu, Fitsum A. Reda, Kevin J. Shih, Ting-Chun Wang, Andrew Tao, and Bryan Catanzaro. 2018. Image Inpainting for Irregular Holes Using Partial Convolutions. *CoRR* abs/1804.07723 (2018).

[29] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. 2015. SMPL: A Skinned Multi-person Linear Model. *ACM Trans. Graph.* 34, 6 (2015), 248:1–248:16.

[30] Julieta Martinez, Rayat Hossain, Javier Romero, and James J. Little. 2017. A Simple Yet Effective Baseline for 3d Human Pose Estimation. *2017 IEEE International Conference on Computer Vision (ICCV)* (2017), 2659–2668.

[31] Dushyant Mehta, Helge Rhodin, Dan Casas, Pascal Fua, Oleksandr Sotnychenko, Weipeng Xu, and Christian Theobalt. 2017. Monocular 3D Human Pose Estimation in the Wild Using Improved CNN Supervision. *2017 International Conference on 3D Vision (3DV)* (2017), 506–516.

[32] Deepak Pathak, Philipp Krähenbühl, Jeff Donahue, Trevor Darrell, and Alexei A. Efros. 2016. Context Encoders: Feature Learning by Inpainting. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2016), 2536–2544.

[33] Patrick Pérez, Michel Gangnet, and Andrew Blake. 2003. Poisson Image Editing. *ACM Trans. Graph.* 22, 3 (2003), 313–318.

[34] Lorenz Rogge, Felix Klose, Michael Stengel, Martin Eisemann, and Marcus Magnor. 2014. Garment Replacement in Monocular Video Sequences. *ACM Trans. Graph.* 34, 1, Article 6 (2014), 10 pages.

[35] Lorenz Rogge, Thomas Neumann, Markus Wacker, and Marcus Magnor. 2011. Monocular Pose Reconstruction for an Augmented Reality Clothing System. In *Proc. Vision, Modeling and Visualization (VMV)*. Eurographics, 339–346.

[36] Carsten Rother, Vladimir Kolmogorov, and Andrew Blake. 2004. "GrabCut": interactive foreground extraction using iterated graph cuts. *ACM Trans. Graph.* 23 (2004), 309–314.

[37] Jamie Shotton, Ross B. Girshick, Andrew W. Fitzgibbon, Toby Sharp, Mat Cook, Mark Finocchio, Richard Moore, Pushmeet Kohli, Antonio Criminisi, Alex Kipman, and Andrew Blake. 2012. Efficient Human Pose Estimation from Single Depth Images. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 35 (2012), 2821–2840.

[38] Ivan E. Sutherland. 1968. A Head-mounted Three Dimensional Display. In *AFIPS Fall Joint Computing Conference*. 757–764.

[39] Denis Tomè, Chris Russell, and Lourdes Agapito. 2017. Lifting from the Deep: Convolutional 3D Pose Estimation from a Single Image. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2017), 5689–5698.

[40] Chunyu Wang, Yizhou Wang, Zhouchen Lin, Alan L. Yuille, and Wen Gao. 2014. Robust Estimation of 3D Human Poses from a Single Image. *2014 IEEE Conference on Computer Vision and Pattern Recognition* (2014), 2369–2376.

[41] Oliver Whyte, Josef Sivic, and Andrew Zisserman. 2009. Get Out of my Picture! Internet-based Inpainting. In *BMVC*.

[42] Zongben Xu and Jian Sun. 2010. Image Inpainting by Patch Propagation Using Patch Sparsity. *IEEE Transactions on Image Processing* 19 (2010), 1153–1165.

[43] Shan Yang, Tanya Ambert, Zherong Pan, Ke Wang, Licheng Yu, Tamara L. Berg, and Ming C. Lin. 2016. Detailed Garment Recovery from a Single-View Image. *CoRR* abs/1608.01250 (2016).

[44] Wei Yang, Wanli Ouyang, Xiaolong Wang, Jimmy S. J. Ren, Hongsheng Li, and Xiaogang Wang. 2018. 3D Human Pose Estimation in the Wild by Adversarial Learning. *CoRR* abs/1803.09722 (2018).

[45] Hashim Yasin, Umar Iqbal, Björn Krüger, Andreas Weber, and Juergen Gall. 2015. 3D Pose Estimation from a Single Monocular Image. *CoRR* abs/1509.06720 (2015).

[46] Jiahui Yu, Zhe Lin, Jimei Yang, Xiaohui Shen, Xin Lu, and Thomas S. Huang. 2018. Generative Image Inpainting with Contextual Attention. *CoRR* abs/1801.07892 (2018).

[47] Mihai Zanfir, Alin-Ionut Popa, Andrei Zanfir, and Cristian Sminchisescu. 2018. Human Appearance Transfer. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[48] Shizhe Zhou, Hongbo Fu, Ligang Liu, Daniel Cohen-Or, and Xiaoguang Han. 2010. Parametric Reshaping of Human Bodies in Images. In *ACM SIGGRAPH 2010 Papers (SIGGRAPH '10)*. ACM, New York, NY, USA, Article 126, 10 pages.

[49] Christian Zimmermann, Tim Welschehold, Christian Dornhege, Wolfram Burgard, and Thomas Brox. 2018. 3D Human Pose Estimation in RGBD Images for Robotic Task Learning. *CoRR* abs/1803.02622 (2018).