**Photo Wake-Up**

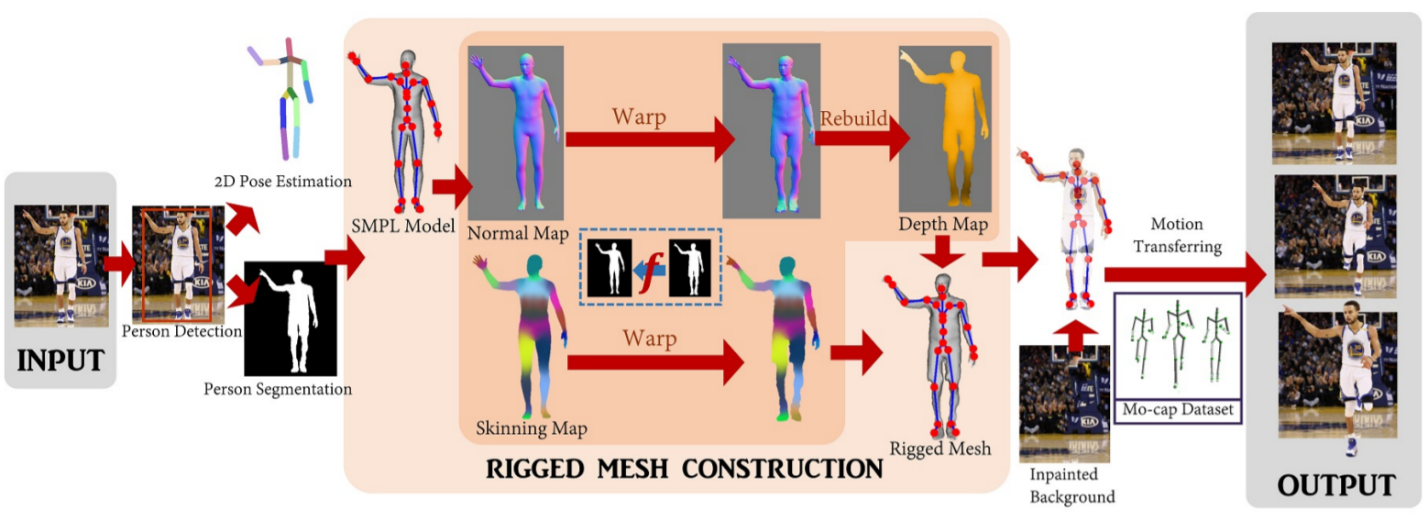
**3D Character Animation from a Single Photo**

In recent years, tremendous amount of progress is being made in the field of 3D Machine Learning, which is an interdisciplinary field that fuses computer vision, computer graphics and machine learning. Photo wake-up belongs to that field. Researchers at the University of Washington have developed this technique for animating a human subject (such as walking toward the screen, running, sitting, and jumping) from a single photo. The technique is demonstrated by using a variety of single-image inputs, including photographs, realistic illustrations, cartoon drawings, and abstracted human forms. The output animation can be played as a video, viewed interactively on a monitor, and as an augmented or virtual reality experience, where a user with headset can enjoy the central figure of a photo coming out into the real world.



*Figure 1: Given a single photo as input (far left), model creates a 3D animatable version of the subject, which can now walk towards the viewer (middle). The 3D result can be experienced in augmented reality (right); in the result above the user has virtually hung the artwork with a HoloLens headset and can watch the character run out of the painting from different views.*

The overall system works as follows (Fig. 2): First apply state-of-the-art algorithms to perform person detection, segmentation, and 2D pose estimation. From the results, devise a method to construct a rigged mesh. Any 3D motion sequence can then be used to animate the rigged mesh. More specifically, “Mask R-CNN” is used for person detection and segmentation. 2D body pose is estimated using “Convolutional pose machines", and person segmentation is refined using “Dense CRF". Once the person is segmented out of the photo, “PatchMatch” (a randomized correspondence algorithm for structural image editing) is applied to fill in the regions where the person used to be.



*Figure 2: Overview of our method. Given a photo, person detection, 2D pose estimation, and person segmentation, is performed using off-the-shelf algorithms. Then, A SMPL template model is fit to the 2D pose and projected into the image as a normal map and a skinning map. The core of our system is: find a mapping between person’s silhouette and the SMPL silhouette, warp the SMPL normal/skinning maps to the output, and build a depth map by integrating the warped normal map. This process is repeated to simulate the model’s back view and combine depth and skinning maps to create a complete, rigged 3D mesh. The mesh is further textured, and animated using motion capture sequences on an unpainted background.*

**Mesh Construction and Rigging**

The key technical idea of this method is how to recover an animatable, textured 3D mesh from a single photo to fit the proposed application. Mesh construction and rigging block is responsible for that. It begins by fitting the SMPL (skinned multi-person linear) morphable body model to a photo, including the follow-on method for fitting a shape in 3D to the 2D skeleton (method was published in 2016 and called “automatic estimation of 3D human pose and shape from a single image”). The recovered SMPL model provides an excellent starting point, but it is semi-nude, does not conform to the underlying body shape of the person and, importantly, does not match the clothed silhouette of the person. To that end 2D approach was taken: warp the SMPL silhouette to match the person silhouette in the original image and then apply that warp to projected SMPL normal maps and skinning maps. The resulting normal and skinning maps can be constructed for both front and (imputed) back views and then lifted into 3D, along with the fitted 3D skeleton, to recover a rigged body mesh that exactly agrees with the silhouettes, ready for animation. The center box in Figure 2 illustrates the approach.

**Mesh Warping, Rigging, & Skinning:** Here we describe the process for constructing a rigged mesh for a subject without self-occlusion. Let’s start with the 2D pose of the person and the person’s silhouette mask S. For simplicity, we refer to S both as a set and as a function, i.e., as the set of all pixels within the silhouette, and as a binary function S(x) = 1 for pixel x inside the silhouette or S(x) = 0 for x outside the silhouette. To construct a 3D mesh with skeletal rigging, we first fit a SMPL model to the 2D input pose using the “automatic estimation of 3D human pose and shape from a single image” method, which additionally recovers camera parameters. We then project this mesh into the camera view to form a silhouette mask . The projection additionally gives us a depth map (x), a normal map (x) and a skinning map (x) for pixels x ∈ . The skinning map is derived from the per-vertex skinning weights in the SMPL model and is thus vector-valued at each pixel (one skinning weight per bone).