Human Bas-Relief Generation from A Single Photograph

Zhenjie Yang*, Beijia Chen*, Youyi Zheng, Xiang Chen†, and Kun Zhou, *Fellow, IEEE*

Abstract—We present a semi-automatic method for producing human bas-relief from a single photograph. Given an input photo of one or multiple persons, our method first estimates a 3D skeleton for each person in the image. SMPL models are then fitted to the 3D skeletons to generate a 3D guide model. To align the 3D guide model with the image, we compute a 2D warping field to non-rigidly register the projected contours of the guide model with the body contours in the image. Then the normal map of the 3D guide model is warped by the 2D deformation field to reconstruct an overall base shape. Finally, the base shape is integrated with a fine-scale normal map to produce the final bas-relief. To tackle the complex intra- and inter-body interactions, we design an occlusion relationship resolution method that operates at the level of 3D skeletons with minimal user inputs. To tightly register the model contours to the image contours, we propose a non-rigid point matching algorithm harnessing user-specified sparse correspondences. Experiments demonstrate that our human bas-relief generation method is capable of producing perceptually realistic results on various single-person and multi-person images, on which the state-of-the-art depth and pose estimation methods often fail.

Index Terms—human bas-relief, multi-person, occlusion resolution, contour matching, single image.

1 Introduction

Bas-relief, a geometric abstraction of real-world objects, has been regarded as a respectable art form since the ancient times (Figure 1). Underlying its remarkable expressivity is a so-called phenomenon of *bas-relief ambiguity* [1], which states that given the right lighting conditions, the shading and shadowing effects of a Lambertian surface could be invariable under depth flattening when looking at the front view. To make its creation less laborious, research on bas-relief generation from various inputs has been active over the last decades [2], [3]. Among them, a single image is the most easy-to-acquire input form for bas-relief generation due to its ubiquity. However, for general objects, single-image based bas-relief generation is highly ill-posed due to the absence of 3D information and thus it often requires a large number of user interventions in different processing steps to guide the generation [4], [5].

Recent work succeeds in producing bas-reliefs for single images of specific objects, such as human face [6], [7], [8] and hair [9]. Prior knowledge of these objects is exploited to keep the amount of user interaction acceptable. In light of this, we seek to generate the bas-relief of humans from a single photograph. This task is, however, challenging due to its unique characteristics. First, although there exists extensive research on human shape estimation, accurately estimating the human shapes along with their dresses and accessories from a single image is difficult [10]. More importantly, people in a group photo tend to have complex body interactions and severe occlusions, which precisely reflects the emotional intimacy and communication between family members or close friends. Such complex intra- and inter-body interactions pose additional challenges to most existing 3D human shape estimation methods.

In this paper, we present a novel semi-automatic pipeline for generating a human bas-relief from a single photograph. The output bas-relief is a composition of an overall base shape and a detailed normal map where the detailed normal map is baked into the base shape via optimization [11]. To produce an accurate base shape, we first fit the SMPL human template [12] to multiple 3D skeletons estimated from the input image, forming a guide 3D model. In an essential stage, before we fit the SMPL models, we allow minimal user inputs to indicate depth ordering of skeletal bones to resolve the complex intra- and inter-body interactions, motivated by the opinion that the editing of the structure is easy for an average user [13]. We then compute a 2D deformation to non-rigidly register the projection contours of the guide 3D model to the body contours of the input image. Accordingly, we design strategies to incorporate minimal user interactions into the non-rigid contour registration to specify sparse correspondence points. Afterwards, we warp the normal map of the guide 3D model by the previously computed 2D deformation to obtain a base normal map tightly aligned with the image, from which we reconstruct

Fig. 1: A stone relief of Barabudur built in the 9th century.

- Z. Yang and B. Chen are with Zhejiang University.
- Y. Zheng, X. Chen and K. Zhou are with the State Key Lab of CAD&CG, Zijingang Campus, Zhejiang University, Hangzhou, China 310058.

*Joint first authors. †Corresponding author. E-mail: xchen.cs@gmail.com

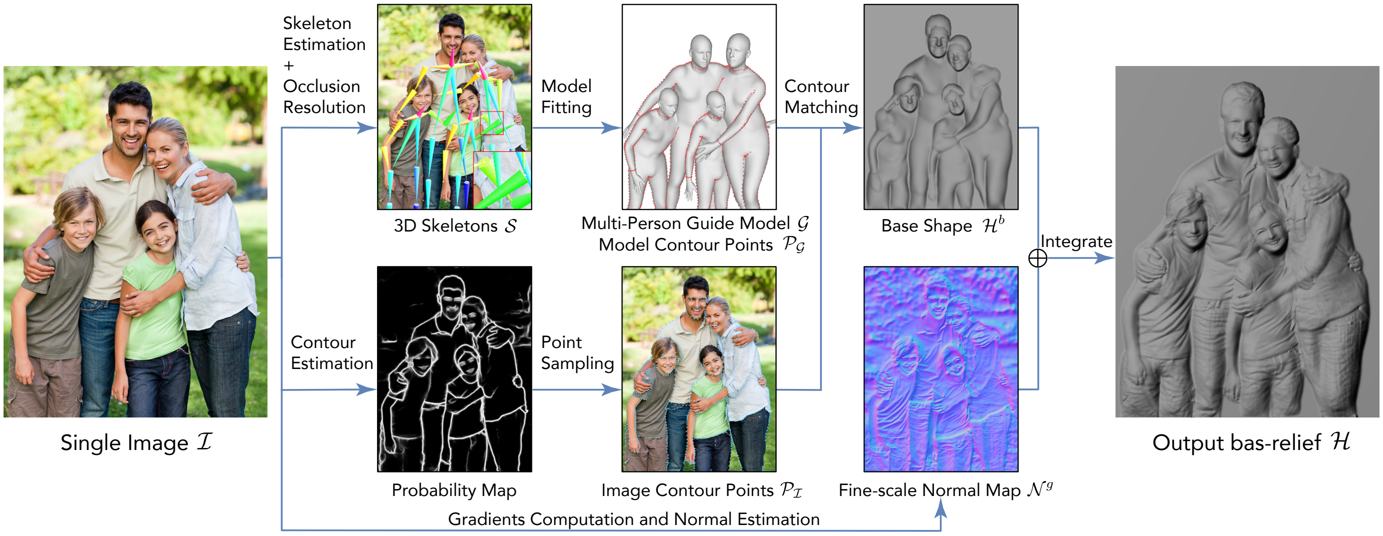

Skeleton Estimation + Occlusion Resolution
Model Fitting
Contour Matching
3D Skeletons $\mathcal{S}$
Multi-Person Guide Model $\mathcal{G}$
Model Contour Points $\mathcal{P}_{\mathcal{G}}$
Base Shape $\mathcal{H}^b$
Integrate
Single Image $\mathcal{I}$
Output bas-relief $\mathcal{H}$
Contour Estimation
Point Sampling
Probability Map
Image Contour Points $\mathcal{P}_{\mathcal{I}}$
Fine-scale Normal Map $\mathcal{N}^g$
Gradients Computation and Normal Estimation

Fig. 2: **Overview of the Pipeline.** Given a single group photo as input, our method produces a corresponding bas-relief by integrating an overall base shape and a fine-scale normal map. We first generate a multi-person guide model by fitting a template model to the estimated 3D skeletons. Then we extract image contours of the people in the photo from an nn-predicted probability map. Finally, we non-rigidly register the guide model to the image contours to generate the base shape. The normal map is obtained from the gradient information of the photo.

a compressed height field, *i.e.,* the overall base shape of the bas-relief.

We demonstrate both the efficacy and efficiency of the human bas-relief generation method on diverse sets of single-person and multi-person images. Experiments show that our method produces perceptually realistic results for single images with complex interactions and occlusions intra- and inter-bodies, on which the state-of-the-art methods for depth and pose estimation often fail.

Our main contributions are:

- The first semi-automatic pipeline for generating human bas-reliefs from a single image;
- An effective occlusions resolution method incorporating user-specified depth constraints between skeleton bones;
- A robust non-rigid contour registration method incorporating user-specified point correspondences.

2 Related Work

2.1 Bas-Relief Generation

Bas-relief generation has been an active research field of computer graphics for the last decades. Various methods have been introduced for producing bas-reliefs from a variety types of inputs, such as 3D shapes [14], [15], [16], [17], point clouds [18], photos [7], [13], [19], sketches [20], line-drawings [21], and paintings [4], to name a few. For comprehensive details of this research field, please refer to the recent surveys [2], [3].

Bas-relief is essentially a shallow height field that looks realistic only in the vicinity of the front view. When full 3D data is available, bas-relief generation boils down to a depth compression problem [15], [22], typically through nonlinear functions or adaptive histograms. A height field is then reconstructed to obey the compressed 3D data. During the reconstruction phase, different strategies are presented for controlling the resultant shapes. For example, the gradient-domain information is mostly preserved by solving a Poisson equation [15], [16]. The rolling guidance normal filter is employed to decouple the normal map into two layers for separate shapeup of base and details [17]. We also design our pipeline to handle the overall structure and shape details separately for further blending. We use the normal map to keep the gradient information during base shape generation. However, our input is merely a single photograph, while this line of work requires rich 3D information, which prevents the direct application of them to our problem.

For image-based bas-relief generation, the biggest challenge is the accurate extraction of visually significant and meaningful 3D information from the 2D input. Due to the ambiguities and complexities from its inverse nature, previous methods often involve user interventions to specify semantic properties of the underlying scene, *e.g.*, lighting conditions [19], region-based segmentations [4], and depth-orderings [5]. Such user guidances largely help resolve the ambiguity and make the 3D computation reliable. However, those methods target general objects and scenes, making their user interactions impractical and ineffective for the human body, which requires a lightweight design.

Recently, there are image-based bas-relief generation methods leveraging prior knowledge and focusing on particular objects, such as human faces [6], [8] and hairs [9]. The method presented by Zhang *et al.* [7] for portrait images is closest to our work. Their work first fits the input portrait image with a template 3D face to get an initial model, and then uses bi-Laplacian deformation to align the initial 3D face with the 2D feature points. Finally, they compress the depth of the aligned 3D face and enhance its fine-scale details via a Shape-from-Shading-based optimization. In their pipeline, user interactions are involved in marking feature points of the face and painting ambiguous regions of the normal map. We share with their work the fit-and-deform strategy at the high-level. However, compared with the face, humans in a multi-person image have a broader exploration space due to the non-rigid shapes and complex body language. We resort to 3D skeletons for resolving the occlusions among body parts. We also detect critical contour points to constrain the body shapes and adopt TPS rather than bi-Laplacian to handle the highly-nonlinear registration. User interaction strategies are correspondingly designed to ensure a minimal requirement of manual efforts.

2.2 3D Human Shape Estimation

As the 3D human shape estimation is a rather broad topic, we only discuss the estimation methods based on the input of a single image. The human bodies are prevalently represented as 3D templates [12], [23] with decoupled sets of parameters, *i.e.,* the pose and shape, statistically learned from high-precision scanning data. With the recent availability of large-scale human datasets [24], [25], [26], learning-based estimation methods obtain impressive performance. Some directly reconstruct the non-parametric human shapes [25], [27], while most others incorporate the parametric human bodies into the training loop [28], [29], [30], [31], [32].

Pose inherently describes the joint locations of human bodies. Recent advances in both 2D [33], [34] and 3D [35], [36], [37] estimation methods make the images or videos reliable sources of pose information. Thus, human shape estimation methods often take the pose as an intermediate representation in the network or for the joints regularization [28], [30], [31], [38]. For example, Kanazawa *et al.* [29] present a network to regress the parameters of the SMPL model, and supervise the training with a joint-reprojection error and an adversarial prior. Besides the parameters regression, Alldieck *et al.* propose to regress a per-vertex 3D offset field [39] or a UV-space displacement map [40] upon the SMPL model to improve the image-space alignment and fine details. However, these model-based methods only support the single-person scenario. They have difficulties guaranteeing a tight alignment with the body contours in the image space, even with the per-vertex optimization. Jiang *et al.* [41] adopt R-CNN to detect all people in a multi-person image and estimate their SMPL parameters. Interpenetration and depth-ordering losses are incorporated during training to encourage a coherent reconstruction. Since it is model-based, the image space alignment is not tight. Furthermore, intra- and inter-body interactions are insufficient in the training data, leading to unsatisfied predictions on family photos. In this work, we resort to the human poses as the guide for body regression, and we present a concise interaction model for effective occlusion resolution.

Annotations like the part segmentation [31] and silhouette [28], [42] are also used as intermediate predictions to help the supervision. However, high-quality part segmentation data are costly to collect and challenging to regress [25]. Moreover, the pixel-wise predictions are hard to be rectified interactively. On the other hand, the silhouette used in previous methods [28], [42] often ignores the internal contours, which are crucial for multi-person scenarios with complex self-occlusions and mutual occlusions. Also, the rectification mechanism is absent when the contour matching is not tight or even erroneous. Instead, we consider both the internal and external contours for non-rigid registration to generate the base shape, and we design an easy-to-use interaction for rectification. The recent work [32] also learns the image-to-surface correspondence [43] to regularize the shape estimation. While the method can work on the multi-person case to some extent, it does not explicitly tackle the occlusion relationships. In contrast, we design an interaction module on 3D skeletons to resolve all the incorrect occlusions intra- or inter-bodies.

Human depth estimation is most relevant to our work. Tang *et al.* [10] present a depth estimation method for human images. The end-to-end pipeline has intermediate modules for part segmentation, 3D joints, normals, and two-levels of depths. They compose the final depth from a base shape and a detailed shape and further refine it by the normal information. Smith *et al.* [44] and Gabeur *et al.* [45] propose to first regress the front and back depth, and then compose the full body through template fitting or Poisson reconstruction. The network training depends on a normal map loss or an adversarial loss. For these works, the accuracy and applicability of depth predictions are somewhat limited by the shape and pose variations and the effective resolutions of the synthetic training data. Moreover, multi-person images are not considered. In a method presented by Li *et al.* [46], they first obtain depth maps by running a multiview stereo method on real-world mannequin imitation videos and then train a depth estimation network using the obtained data for supervision. Their method can capture rough depth profiles on multi-person images. However, the body shapes, details, and occlusions are often missing for family photos, possibly due to the high dynamic range of depth data and the limited effective resolution of the acquired ground truth. To overcome the shortage of 3D human data, recent works [47], [48] exploit dance videos from social media and design warping-based strategies to enable self-supervision. In this work, we tackle the multi-person input by designing user-interaction incorporated mechanisms to handle the occlusion resolution and contour matching. The final bas-relief faithfully recovers both the large-scale shape variations and fine-scale geometric details in the image.

Recently, template-free methods obtain promising results on single-view 3D human reconstruction. For example, Zheng *et al.* [49] train a volume network to regress the spatial occupancy of 3D shapes on a voxel grid. While they also adopt a normal refinement network to augment the reconstructed shape, the fine details are still restricted by the voxel representation's resolution limitation. Among these works, PIFu [50], and its high-res successor PIFuHD [51] achieve state-of-the-art reconstruction quality. They propose to learn an implicit occupancy function in a pixel-wise way rather than relying on explicit volume representations or global latent features. Although the method often performs well on single-person images, it does not directly apply to multi-person input due to its data-driven nature. Moreover, the occlusions in such cases are not explicitly addressed, and multi-person data paired with 3D ground truth are still rare. In contrast, our method resorts to a lightweight framework to incorporate sparse user constraints on occlusion resolution and does not depend on a 3D multi-person dataset.

3 Overview

Our pipeline requires a single image $\mathcal{I}$, of one or a group of persons, as the only input (see Figure 2). The output bas-relief $\mathcal{H}$ is a shape approximation of the people with their clothes in $\mathcal{I}$. We construct $\mathcal{H}$ by merging two levels of ingredients:

$$\mathcal{H} = \mathsf{merge}(\mathcal{H}^b, \mathcal{N}^g), \tag{1}$$

where $\mathcal{H}^b$ is a base shape containing the global shape and layout of human bodies compressed in depth, and $\mathcal{N}^g$ is a normal map capturing fine-grained geometric details such as the wrinkles of clothes and facial details in the image.

We are not trying to recover the exact 3D human body from $\mathcal{I}$ as single-image based reconstruction is generally ill-posed. Our goal is to generate a compressed height field, *i.e.*, the bas-relief of the people. Therefore, the first task is to compute a smooth and rough human model as a guidance of overall shape and layout. However, directly regressing such a model with accurate spatial relationships from the input image is challenging, even with state-of-the-art methods like [32], due to the abundant occlusions and complex interactions among body parts. To this end, we choose to extract the skeletons $\mathcal{S}$ of the people by off-the-shelf neural networks, and resolve in 3D the occlusive relationships among the skeletons. After that, we fit a parametric human template to each skeleton in $\mathcal{S}$ to generate a multi-person guide model $\mathcal{G}$. Compared with the end-to-end body regression, explicitly estimating the 3D skeletons enables an easy incorporation of user corrections for the depth ordering between skeleton bones. Meanwhile, such corrections are much more easier to interact and more accurate than segmenting regions and specifying their relative depth ordering [5], [20].

The guide model $\mathcal{G}$ only contains approximate human shapes which do not tightly match the body profiles in the image. Therefore, we first extract projected model contours from $\mathcal{G}$ and body contours from $\mathcal{I}$, and then match the two sets of points uniformly sampled on those contours in the 2D space. Based on the 2D deformation computed from the non-rigid point matching, we warp the normal map of $\mathcal{G}$ to reconstruct a height field $\mathcal{H}^b$, *i.e.*, the base shape, whose shape profiles are consistent with the body profiles in $\mathcal{I}$.

Finally, we compute a detailed normal map $\mathcal{N}^g$ from the input image, and produce a bas-relief model $\mathcal{H}$ through an optimization that balances the large-scale shape variations in $\mathcal{H}^b$ and the fine-scale geometric details in $\mathcal{N}^g$.

Our method is semi-automatic, and it requires minimal user interaction to help with the occlusion resolution, contour matching, and special region masking when necessary.

4 Multi-Person Guide Model

Constructing a bas-relief of high fidelity relies on a stable capturing of the global human structures in the image. Therefore, we estimate the overall body shape of each person and resolve the spatial relationships among them. The resultant multi-person model provides a reliable reference guide for the following base shape generation.

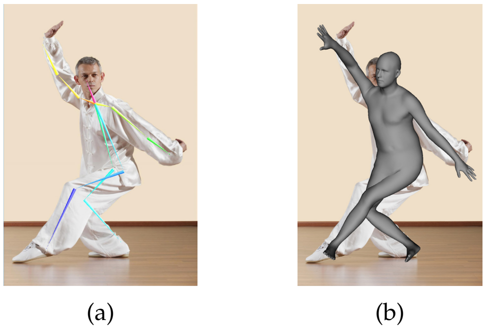

(a) (b)

Fig. 3: **Guide Model Generation.** (a) The 3D skeleton estimated from the image. (b) The guide model computed by fitting SMPL to the skeleton.

4.1 Pose and Camera Estimation

To obtain a 3D shape approximation of the human bodies in the image, we first estimate each person's skeleton model $\mathcal{S}_i,\ i = 1, \ldots, N$. Considering the robustness and generality of deep learning methods, we choose OpenPose [33] for 2D human pose estimation and adopt a lightweight network [52] to generate the 3D skeletons $\mathcal{S}$ from the 2D pose.

The estimated 3D skeletons are represented in their own coordinate systems. We need to transform them into a unified camera coordinate system while ensuring the 3D joints are projected to their 2D counterparts in the image space. Consequently, we optimize the following energy to obtain the intrinsic parameters $\mathbf{K}$ of a unique pinhole camera and the similarity transformation $\mathbf{T_i}$ of each skeleton $\mathcal{S}_i$

$$E(\mathbf{K}, \mathbf{T_i}) = \sum_{i=1}^{N} \sum_{\mathbf{v} \in \mathcal{S}_i} \|\mathbf{p} - \pi_{\mathbf{K}}(\mathbf{T_i v})\|^2 + \sum_{i=1}^{N} (t_z^i - \bar{t}_z)^2, \tag{2}$$

where the first term penalizes the re-projection error, and the second term regularizes the 3D skeletons to have similar depths. In Equation 2, $\mathbf{v}$ is a keypoint of the 3D skeleton and $\mathbf{p}$ is its corresponding 2D keypoint in the image space. The projection function $\pi_{\mathbf{K}}$ is dependent on the intrinsic matrix $\mathbf{K}$, in which we assume the focal lengths of two axes are the same. For the extrinsic matrix $\mathbf{T} = [s\mathbf{R}|\mathbf{t}]$, $\mathbf{R}$ is the rotation matrix, $\mathbf{t} = [t_x, t_y, t_z]^\top$ is the translation vector, and s is a scalar value for handling inconsistent scales of different skeletons. For the regularization, $\bar{t}_z = \frac{1}{N}\sum_{i=1}^{N} t_z^i$ is the mean translation in the z coordinate. The optimization is nonlinear, in which we set the initial value of focal length to 500, $\mathbf{t}$ to $[0, 0, 400]^\top$, and s to 1. These values are chosen empirically and work well for all our examples in practice. To prevent the collapse of s and $\mathbf{t}$ to infinitesimal values, we also constrain s to have a lower bound of 0.3.

4.2 Occlusion Resolution

After pose and camera estimation (section 4.1), we obtain an initial 3D skeleton for each person whose projection matches its 2D counterpart. However, the spatial relationships among these 3D skeletons are not necessarily correct. Consequently, the bodies computed by fitting SMPL to these skeletons (section 4.3) often exhibit occlusions inconsistent with the given image, either intra or inter bodies (Figure 4b).

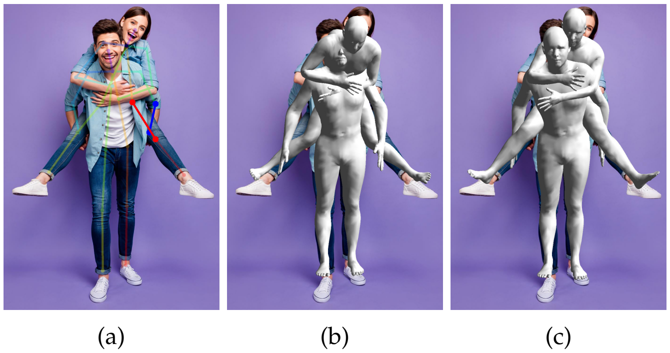
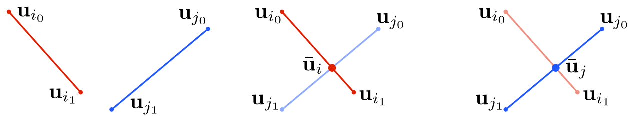


Fig. 4: **Occlusion Resolution.** (a) The user interaction of the depth order, of the red-blue pair, for bone-level occlusion resolution. (b-c) The generated bodies without and with occlusion resolution.



(a) no occlusion (b) l_1 upon l_2 (c) l_2 upon l_1

Fig. 5: Occlusion relationship between two bone segments.

Thus, we present a semi-automatic method to recover the correct occlusion relationships before fitting the 3D bodies.

Resolving body occlusions is challenging. As shown in Figure 4a, the occlusion relationship between two people is often ambiguous and not a single boolean value, since the bodies are non-rigid and the mutual interactions are complex. A key observation is the body parts are precisely rigid, and the occlusion relationship between two parts is definite. Therefore, we resolve the occlusions at the bone-level of 3D skeletons. For two arbitrary bones in $\mathcal{S}$, we check if their projections in 2D, *i.e.*, two line segments, have an intersection point (see Figure 5). If they do, the occlusion relationship has to be explicitly specified by the user. Each time the system prompts a pair of intersecting segments on the GUI for resolution, the user can push a button to switch their depth order (Figure 4a). Such a user-interaction is converted into a spatial constraint, *i.e.*, there has to be a signed depth gap between the two specified bones. Our goal is to adjust the keypoint depth of $\mathcal{S}$ to satisfy the user-provided constraints while preserving the overall shape of $\mathcal{S}$. Thus, we build a graph structure by taking the keypoints as nodes and the bones as edges. Then we formulate the occlusion resolution as a graph Laplacian minimization [53], [54] subject to a small set of depth constraints

$$E(\mathbf{z}) = \left\|\mathbf{L}\mathbf{z} - \mathbf{L}\mathbf{z}^{(0)}\right\|^2 + \omega \sum_{f,b \in \mathsf{occpairs}} \left(\bar{z}_f - \bar{z}_b + d_{gap}\right)^2 , \tag{3}$$

where $\mathbf{L}$ is the matrix of graph Laplacian operator, $\mathbf{z}$ and $\mathbf{z}^{(0)}$ stack all the z-coordinates of the deformed and undeformed keypoints, $\bar{z}_i = (1-\alpha)z_{i_0} + \alpha z_{i_1}$ is the z-coordinate of the intersection point on the i-th bone. We set the weight ω to 0.1 and the depth gap d_{gap} to 15 for compensation of bone thickness.

As shown in Figure 4c, the bodies computed by fitting SMPL to the occlusion resolved 3D skeletons have mutual interactions consistent with the image. With such an ease of partial order specification on skeleton bones, we let the users help constrain the occlusions while keeping the interactions minimal. This is a significant reason we explicitly estimate the 3D skeletons rather than generate the bodies directly from the image like the recent end-to-end method [32].

4.3 3D Body Fitting

For each 3D skeleton $\mathcal{S}_i$ estimated in section 4.1, we generate a body $\mathcal{G}_i = \mathsf{body}^{\mathsf{smpl}}(\theta^*, \beta^*)$ by fitting the SMPL template [12] to $\mathcal{S}_i$

$$\theta^*, \beta^* = \arg\min_{\theta,\beta} \sum_{\mathbf{v} \in \mathcal{S}_i} \left\|\mathbf{v} - \mathsf{joint}^{\mathsf{smpl}}_{\mathbf{v}}(\theta, \beta)\right\|^2 , \tag{4}$$

where $\mathbf{v}$ is a keypoint of $\mathcal{S}_i$ and $\mathsf{joint}^{\mathsf{smpl}}_{\mathbf{v}}$ is its corresponding point in the template skeleton, θ and β are the pose and shape parameters of the template model, respectively. Figure 3 shows a result of the body fitting.

Since the energy function in Equation 4 is highly nonlinear, we use the estimated skeleton $\mathcal{S}_i$ to compute θ_{init}, the initial values of the pose parameters, for better convergence. Due to the structural difference between $\mathcal{S}_i$ and the skeleton of the template, we only constrain the joints that have direct correspondences.

5 Contour Matching

The multi-person guide model $\mathcal{G}$ has poses and occlusion relationships consistent with the input image $\mathcal{I}$. However, the projection contours of $\mathcal{G}$ in the image space are not yet tightly matched with the body contours in $\mathcal{I}$ (Figure 4c). Though we could further optimize the shape parameters of the SMPL models to put the contours closer, an exactly tight match is unreachable due to the existence of clothes and the reduced space of the SMPL model. Fortunately, our goal here is to produce a compressed height field with accurate profiles as the base shape $\mathcal{H}^b$ of bas-relief. To produce such a base shape, we first register the two sets of contours in the 2D space by computing a non-rigid planar deformation and then warp the normal map of $\mathcal{G}$ based on the computed planar deformation to generate $\mathcal{H}^b$.

5.1 Contour Extraction

The contours describe depth discontinuities in the image, which reflect the overall shapes and the spatial relationships between different body parts. However, multi-person images often exhibit complex intra- and inter-body interactions, making accurate prediction of the segmentation boundaries a challenging task, even for advanced instance segmentation methods (*e.g.*, [55]). Furthermore, these methods can only find the outside boundaries of the silhouettes, but not the shape edges inside the masks caused by self-occlusions. To address these issues, we adopt a learning-based edge detection method [56] to generate a probability map of contours, which is clipped with a fixed threshold of 120 to retain the main features (Figure 6a). The nonzero pixels in this map are candidate contour points. We then perform the Fisher-Yates shuffle algorithm [57] to sample a subset of points uniformly from all the candidates as the

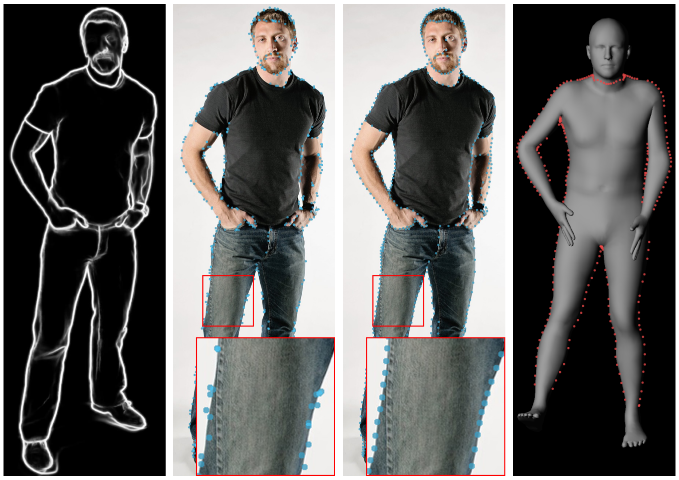
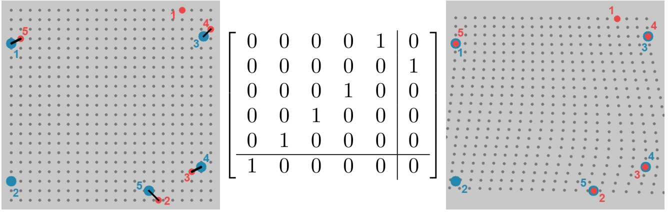


Fig. 6: **Contour Extraction.** (a) The probability map of image contours. (b) The initial centers sampled from the probability map. (c) The centers after k-means clustering. (d) The contour extracted from the guide 3D model.



$$\left[\begin{array}{ccccc|c} 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ \hline 1 & 0 & 0 & 0 & 0 & 0 \end{array}\right]$$

(a) initial state (b) optimal $\mathbf{Z}$ (c) registration

Fig. 7: **Non-rigid Registration Example**. (a) The initial states of the source points (red) and the target points (blue), and the optimal correspondences (black lines) between them. (b) The matrix for the optimal correspondences. The last row and column are extra markings for outliers (no correspondences). (c) The result of applying the optimal non-rigid transformation f.

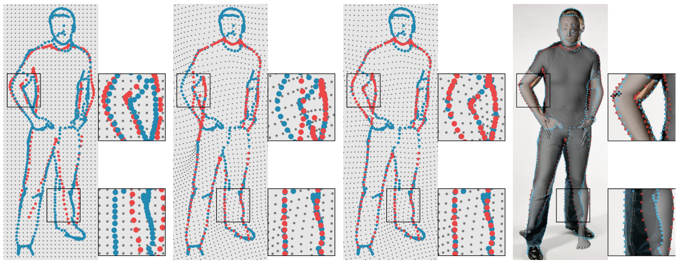


Fig. 8: **Non-rigid Contour Matching.** (a) The initial state of the model contour (red) and the image contour (blue). (b-c) The contour matching result without and with user specified point pairs. (d) The user selection of the correspondent points (black) in the GUI.

initial clustering centers (Figure 6b). Next, we perform k-means on the candidates to distribute the centers evenly (Figure 6c). Finally, we regard the resultant centers as the image contour points $\mathcal{P}_{\mathcal{I}}$ for matching.

For the guide model, we first divide its mesh faces into front-oriented and back-oriented groups according to their normals in the camera coordinate system. Next, we locate the vertices on the common boundary of the two groups and take their projections in 2D as the model contour points $\mathcal{P}_{\mathcal{G}}$ for matching (Figure 6d). For all our examples, we empirically constrain $|\mathcal{P}_{\mathcal{I}}| = 1.2\,|\mathcal{P}_{\mathcal{G}}|$ to ensure a sufficient number of image points for matching.

5.2 Non-rigid Point Matching

Given the point sets $\mathcal{P}_{\mathcal{I}}$ and $\mathcal{P}_{\mathcal{G}}$, our goal is to find the optimal correspondences $\mathbf{Z}$ between them and the non-rigid planar transformation f represented as a warping function that aligns them in 2D. We conduct the simultaneous minimization

$$\min_{\mathbf{Z},f} \sum_{i=1}^{N}\sum_{j=1}^{K} Z_{ij}\,\|\mathbf{y}_j - f(\mathbf{x}_i)\|^2 + \lambda\,\|\mathsf{L}(f)\|^2 - \xi \sum_{i=1}^{N}\sum_{j=1}^{K} Z_{ij} \tag{5}$$

subject to $Z_{ij} \in \{0,1\}$, $\sum_{j=1}^{K+1} Z_{ij} = 1$, and $\sum_{i=1}^{N+1} Z_{ij} = 1$. Here we have $N = |\mathcal{P}_{\mathcal{G}}|, K = |\mathcal{P}_{\mathcal{I}}|$ and $\mathbf{x}_i \in \mathcal{P}_{\mathcal{G}}, \mathbf{y}_j \in \mathcal{P}_{\mathcal{I}}$. The first term measures the approximation fidelity, the second term represents a constraint on the smoothness of f, and the third term penalized the number of outliers. Figure 7 shows an example of the non-rigid matching formulation.

In practice, we adopt the TPS-RPM [58] to process Equation 5. The method parameterizes the non-rigid transformation f and the smoothness measure L as the thin-plate spline formulation [59] and uses softassign [60] to relax the binary correspondence $\mathbf{Z}$. The variables f and $\mathbf{Z}$ are alternatively optimized via an annealing schedule. We set the weight λ to be linearly dependent on the temperature for annealing, and set the weight ξ to 1e-2 for tolerating a large number of outliers. See the supplemental for more details.

The fully-automatic optimization cannot ensure a perfect matching due to the complexity of body parts and the existence of outliers. Once again, we leverage user interactions to guide the optimization process. Interacting with the transform function f is impractical. Thus we let the users pick several pairs of correspondent points. Each point pair specified by the user is converted to a hard constraint for $\mathbf{Z}$. For a specified point pair (i, j), we fix Z_{ij} to 1 and all the other entries in the i-th row and j-th column to 0. We found such interactions convenient, and on average, 4 such point pairs are sufficient for a tight contour alignment. Figure 8 shows an interaction example.

6 Shape Integration

6.1 Base Shape

Given the multi-person guide model $\mathcal{G}$ (section 4) and the non-rigid transformation f_{tps} (section 5), we generate the base shape $\mathcal{H}^b$, which is a height field accurate in overall layout but without fine-scale details.

To achieve this, we first render a normal map (Figure 9b) of $\mathcal{G}$ and then warp it by f_{tps} to obtain the base normal map $\mathcal{N}^b$ (Figure 9c). After that, we generate the base shape $\mathcal{H}^b$ (Figure 9d) from $\mathcal{N}^b$ by using the modeling method of Ji *et*

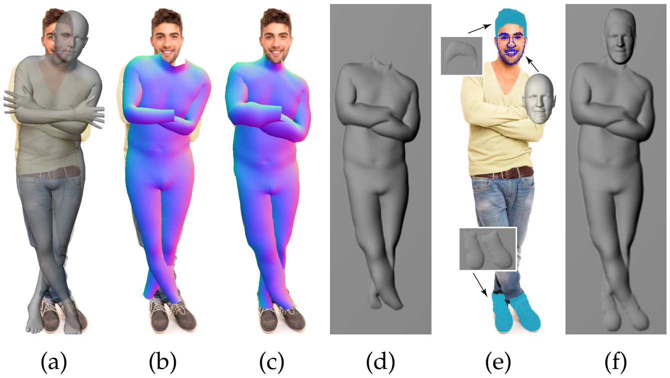


Fig. 9: **Base Shape Generation.** (a) The guide model. (b) The normal map rendered from the guide model. (c) The normal map warped by the non-rigid transformation. (d) The base shape reconstructed from the warped normal map. (e) The special processing of head, hairs, hands, and feet. (f) The full base shape.

al. [16], which minimizes the squared difference between the Laplacian of $\mathcal{H}^b$ and $\mathcal{N}^b$ with intuitive height control.

Since the head of the SMPL model lacks facial characteristics and the hands and feet are hard to fit robustly, in practice, we render $\mathcal{N}^b$ without these parts. For the head, we instead extract facial landmarks from the image [61] and fit them with a 3D facial expression model [62] to obtain a high-quality head (Figure 9e) tightly aligned with the image. Then we render the head's depth map and convert it to a height map, which is superimposed onto the base shape $\mathcal{H}^b$ to update it. For the hairs, hands, and feet, we first let the users paint a small set of strokes in the image to mask out these regions [63]. Then we smooth the mask to estimate the gradients on its boundary. Next, we use the gradients as the boundary conditions for solving a Laplacian problem to get an approximate height field (Figure 9e). Finally, we update $\mathcal{H}^b$ by composing the height field with Possion Editing [64] (Figure 9f). See the supplemental for more details.

6.2 Fine-scale Normal Map

The base shape misses fine-grained geometric details from the input image such as the wrinkles of the clothes and the faces. To enhance these features, we compute a deceptive but robust normal map to approximate the fine-scale details. We design a multi-level extension to the method in [16]. First, we convert the input image $\mathcal{I}$ to a grayscale image $\mathcal{I}_{\text{gray}}$ and filter it with Gaussian kernels of different sizes. Then we extract the per-pixel normals $\mathbf{n}^k(u,v) = \text{normalize}\left([-\nabla\mathcal{I}^k_{\text{gray}}(u,v), 1]^\mathsf{T}\right)$ at each level k. Finally, the normals at different levels are averaged to obtain the fine-scale normal map $\mathcal{N}^g$. We found such a method is more robust on a single image than photometric stereo methods like the shape-from-shading. We show a comparison with an SFS-based method and an intrinsic image optimization method in Figure 15.

6.3 Shape Sharpening

To produce the final bas-relief, we employ the method of Nehab et al. [11] with orthographic projection to compose the overall shape $\mathcal{H}^b$ and the fine-scale details $\mathcal{N}^g$ together by a least-squares optimization

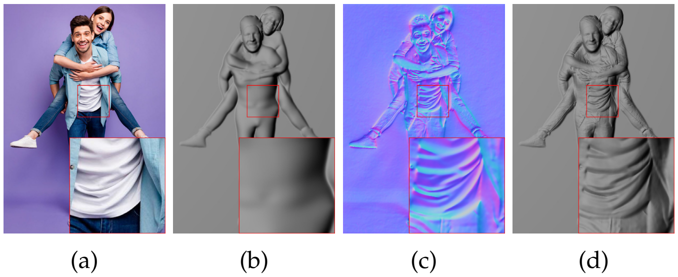


Fig. 10: **Shape Integration.** (a) The input image. (b) The base shape. (c) The fine-scale normal map. (d) The final bas-relief.

$$E(\mathcal{H}) = \alpha\sum_i(\mathcal{H}_i-\mathcal{H}^b_i)^2+(1-\alpha)\left[(T^u_i\cdot\mathcal{N}^g_i)^2+(T^v_i\cdot\mathcal{N}^g_i)^2\right], \tag{6}$$

where i is the pixel index, and T^u and T^v are linearly constructed from $\mathcal{H}$ representing two tangents to the optimized surface. We set α to 0.1 for the body region and 0.4 for the head region. To discard the background information in $\mathcal{N}^g$, we use the non-zero heights in $\mathcal{H}^b$ as a binary mask. Figure 10 shows an example of shape integration. We refer to the supplemental for more details.

7 Experimental Results

7.1 Hardware and Software

We build our system on a desktop PC with an Intel Core i5-4590 CPU, an NVIDIA GTX 1080Ti GPU, and 32G memory. We implement the numerical computations with Eigen [65] and Ceres Solver [66], the image processing with OpenCV [67], and the GUI with Qt. For the experiments, we downsample each photo to have a height of 1000 pixels, which achieves a balance between the runtime efficiency and the quality of details.

7.2 Evaluation and Comparison

To show the efficacy and robustness of our bas-relief generation method, we apply our approach to a variety of group photos with diverse body postures, ages, and genders (Figure 11, the 2nd row). The results demonstrate that the method generates faithful shapes recovering correct occlusions between limbs and trunks, and also produces the fine details of bodies and clothes.

The final bas-relief is essentially a continuous and compressed depth-field. Therefore we compare our method with state-of-the-art human depth estimation methods on single images (Figure 11, rows 3 to 5). All the methods are based on neural networks and fully automatic. In particular, the method of Li *et al.* [46] produces reasonable depth-orderings and partially captures large-scale body profiles in the depth map, but the shapes are rough and flat, where the salient shading effects and fine-scale features are lost. The method of Tang *et al.* [10] cannot handle multi-person images, on which the method produces incomplete predications and confuses body parts belonging to different people. For single-person images, the output is rather coarse, and

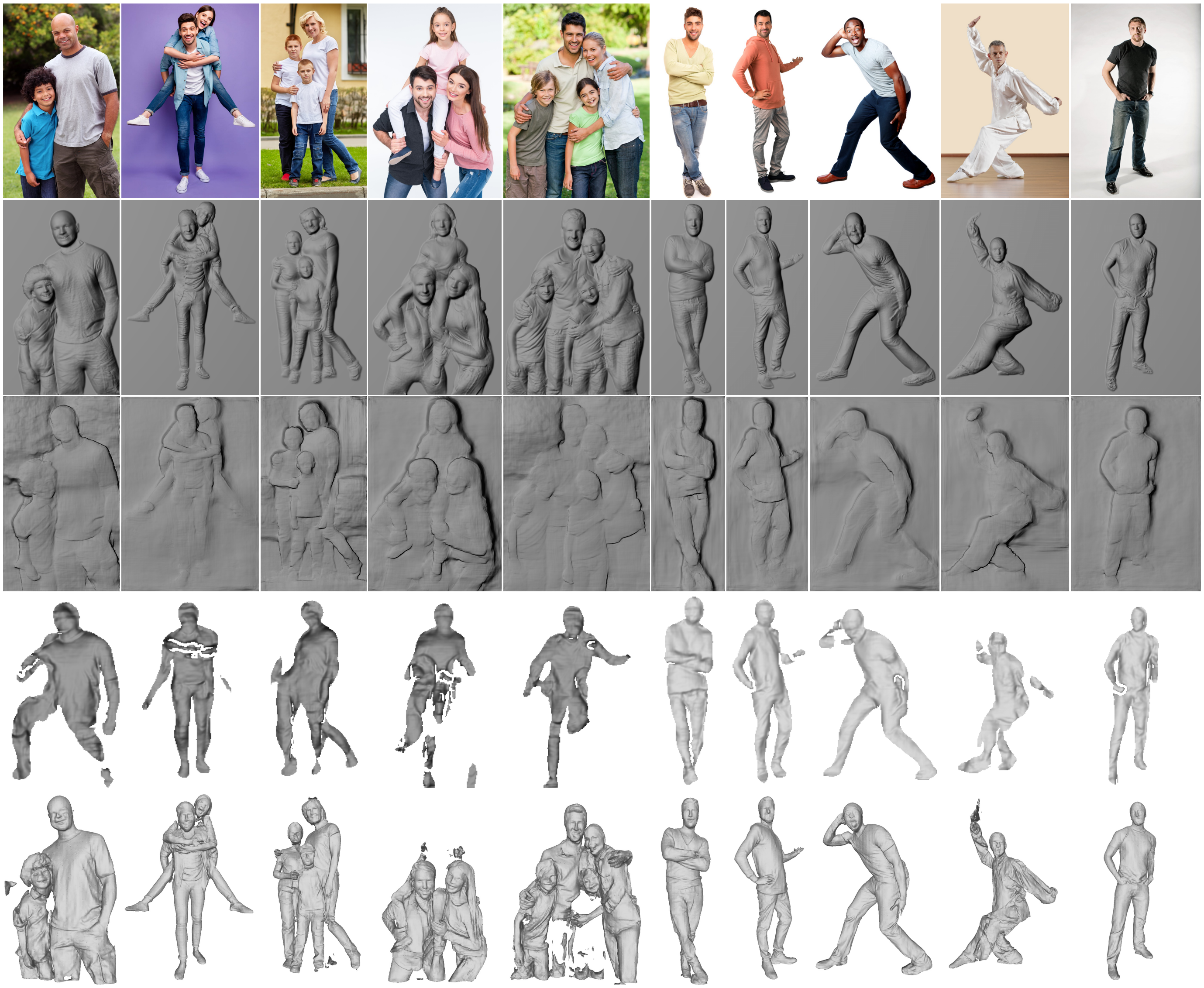


Fig. 11: **Comparison to NN-based Depth Estimation.** From top to bottom: the input images, the bas-reliefs produced by our method, the depth estimation results of Li *et al.* [46], the depth estimation results of Tang *et al.* [10], and the full body reconstruction results of PIFuHD [51].

the mask prediction is inaccurate. To be rigorous, PIFuHD [51] is not a depth estimation method but for full-body reconstruction. When looking at the front view, it produces high-quality shapes with abundant fine details, at least for single-person input. The method does not directly support multi-person images, where body parts are often missed. The results also exhibit depth artifacts and fine-scale noise (see the supplemental). In contrast, our method produces an intact and detailed bas-relief of all the subjects and succeeds in resolving the complex body interactions with minimal user interventions. To evaluate the qualities of the results, we recruit 134 volunteers for the user study. Specifically, for each image, we show them the results of the four methods in random order and ask them to select the one that looks most similar to the input or choose none. Figure 12 shows that, on average, our method obtains 63% of the votes, a significant preference over the alternatives, while the second one PIFuHD gets 23%. Note that to be fair, we only show the front view of the results to minimize the interference of depth artifacts of PIFuHD.

The base shape $\mathcal{H}^b$ is fundamental to the quality of the produced bas-relief. To demonstrate its efficacy, we conduct a naive estimation of the base shape and use it to generate the bas-relief according to Equation 6. Precisely, we first manually mask the person in the input image. Then we smooth the mask and estimate approximate gradients at the boundary. Finally, we solve a Laplacian problem by taking the gradients as boundary conditions to produce a naive base shape and add details onto it. Figure 13 shows that the bas-reliefs constructed from naive base shapes are rather flat and lose important depth clues between interacting body parts. In contrast, our results embody genuine depth relationships and realistic 3D-perceptions.

To generate the multi-person guide model $\mathcal{G}$, we choose to estimate skeletons as intermediates and fit the SMPL template to them. There could be alternatives for this step. We

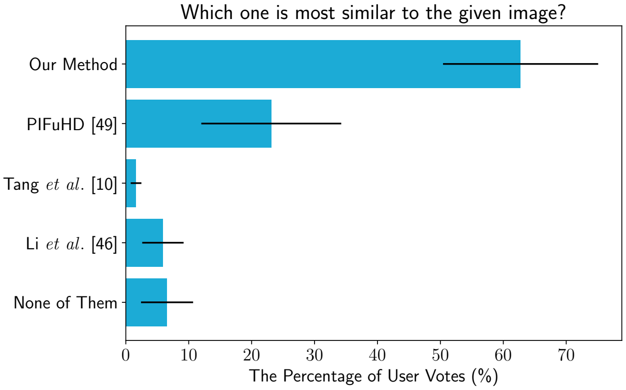


Fig. 12: The user study shows that our results obtain significant preference over alternative ones. The error bar indicates the standard deviation of the per-image vote percentages.

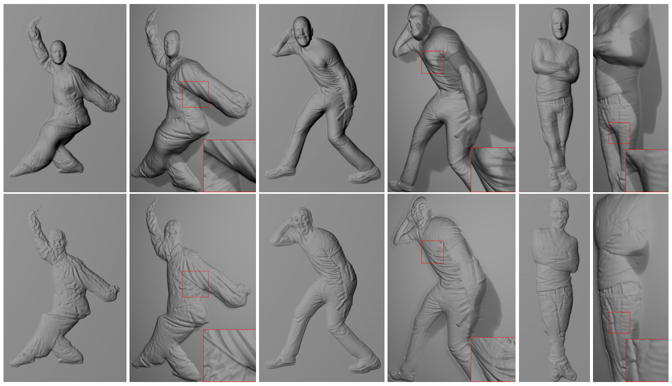


Fig. 13: **Comparison to Alternative Base Shapes.** The bas-reliefs constructed from our full base shape (top row) and the naive base shape (bottom row). We show both the front views and 45-degree side views of the results.

thus compare our strategy with the latest NN-based body reconstruction methods [29], [32], [41]. Figure 14 shows that we achieve similar fitting qualities like them in the case of single-person input. However, they do not work well on multi-person input and produce the wrong number of bodies. The method of Kanazawa *et al.* [29] always outputs one regardless of how many people appear. The method of Guler *et al.* [32] produces all the appeared bodies only in some cases, *e.g.*, only one is obtained while there are two (2nd row). The method of Jiang *et al.* [41] tends to generate more bodies than exists, *e.g.*, three bodies are predicted but only two in the image (2nd row). Moreover, the occlusion relationships produced by these methods are often wrong, *e.g.*, the 3rd row. Although the method explicitly deals with occlusions in the training loss, it still fails on family photos manifesting intertwined body parts rather than holistic body orderings. In contrast, our method generates all the appeared bodies and resolves all the wrong occlusions with 3D skeletons at part-level.

To produce the final bas-relief model, we integrate an overall base shape with a fine-scale normal map to enhance the details. Shape-from-Shading based methods are often adopted to extract geometric details from the input image of objects like the human face [7]. However, in our case the wide variety of materials and textures of faces, hairs and clothes breaks the basic assumptions, *e.g.*, the spatially invariant reflectance and constant albedo, of SFS-based methods. Compared with our normal estimation (Figure 15b), the SFS-based method [68], even with an adaptive albedo model [9], generates unreliable normal maps (Figure 15c,d). Unlike the SFS-based method, intrinsic image optimization jointly recovers the shape, illumination, and reflectance. We compare our method with a data-driven method, SIRFS [69]. Though the normals produced are visually more plausible (Figure 15e) than the SFS-based method, the result is still spatially inaccurate and misses fine-scale details, which our normal estimation can robustly approximate. Note that we use the image pixels of the facial skin to estimate a lighting model (Figure 15a) for both [68] and [69].

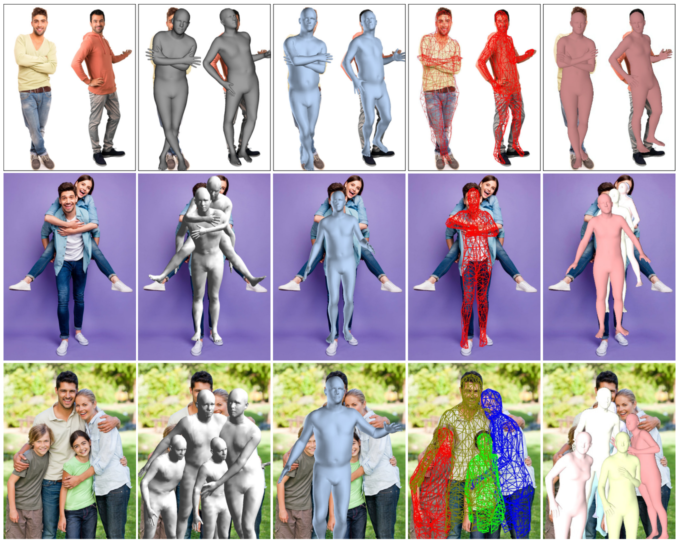


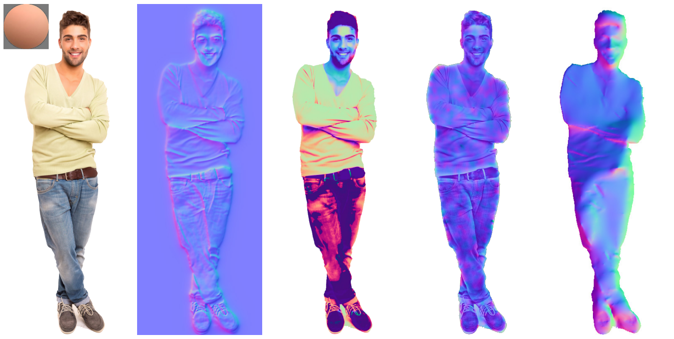
Fig. 14: **Comparison of Guide Models.** From left to right: the input image, our guide models, the result of Kanazawa et al. [29], the result of Guler et al. [32], and the result of Jiang *et al.* [41]. Note that the correct occlusion resolution, rather than tight image alignment, is the key to success for the guide models. We ensure the latter by contour matching.

7.3 User interaction

We carefully design our pipeline and algorithms to minimize the user-interaction, such that it is affordable by an average user. Our system requires the user to specify the occlusion relationships of the intersecting skeleton bones, pick key point pairs in contours for non-rigid registration, and paint strokes for segmentation of the hairs, hands, and feet when they are visible. Table 1 collects the number of interactions for the examples shown in Figure 11. For the interaction time of these examples, on average, we take less than 2 minutes for point pair selections, 1 minute for bone occlusions, and 1 minute for region segmentations. For the most complex one, it takes about 8 minutes in total.

7.4 Running Time

For all the examples shown in Figure 11, the total computation time without taking account of the user interactions is less than 4 minutes. The most time-consuming step is the



(a) (b) (c) (d) (e)

Fig. 15: **Comparison to Shape-from-Shading methods.** (a) The input image. (b) The fine-scale normal map from our method. (c) The fine-scale normal map from an SFS-based method [68]. Their method assumes the model has constant albedo, which does not hold for human bodies with textured clothes, shoes, and hairs, leading to incorrect normal estimations. (d) The fine-scale normal map from [68] with an adaptive albedo model [9]. (e) The fine-scale normal map from SIRFS [69].

TABLE 1: Statistics of the user interactions for specifying the order of intersecting bones, the matched contour point pairs, and the painting strokes for masks of hands/feet/hairs.

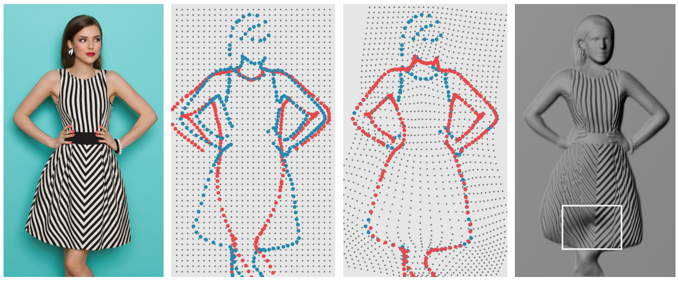
ID in Figure 11 (left to right)	1	2	3	4	5	6	7	8	9	10
#Bone Pairs	3	11	10	10	14	5	0	0	2	0
#Point Pairs	2	3	8	7	6	3	2	4	6	3
#Paint Strokes	3	2	12	8	10	3	5	4	4	4

generation of the multi-person guide model, which takes about 2 minutes for the most complex example (the one with four people). The contour matching takes less than 20 seconds, and the shape integration takes about 1.5 minutes.

8 Conclusion, Limitation and Future Work

We have introduced an easy-to-use semi-automatic pipeline for human bas-relief generation from a single photograph. It produces a high-fidelity height field of the human bodies with an accurate spatial layout and realistic fine-scale details. The skeletons-based occlusion resolution and non-rigid contours registration, together with the reliable user-interactions, ensure the robust inference and extraction of global shape and structure information from the input image.

Our method has a few limitations. The computation of fine-scale normal map is prone to inaccurate results under strong global-illumination effects like shadows and reflections. For complex texture, the color variation is converted to the high-frequency change of the height field (Figure 16), and whether it is a desired feature or not depends on the user. Loose-fitting clothing like the skirt is challenging for base shape generation due to the vast difference in shape contours, even with the non-rigid registration method (see the white frame in Figure 16d).



(a) (b) (c) (d)

Fig. 16: Bas-relief for a loose cloth with strong texture.

We design the current pipeline to keep it lightweight and away from the necessity of expensive multi-person training data. Learning-based methods and data-driven priors have the potential to promote the automation level of base shape generation. Given sufficient training data paired with 3D ground-truth and explicit occlusion models like [70], it is promising to enable the non-template depth estimation networks to resolve the partial-orderings of occlusion regions in multi-person images. Building accurate and dense image-to-surface correspondences [43] has the potential to further decrease the manual efforts for contours matching. An enhanced template fitting method [71] can also help remove the segmentation needs for particular regions like the hands.

Acknowledgments

The authors would like to thank the anonymous reviewers for their constructive comments. This research is partially supported by the National Key Research & Development Program of China (2018YFE0100900), the National Natural Science Foundation of China (61772024, 61732016, 61890954), and the Fundamental Research Funds for the Central Universities.

References

[1] P. N. Belhumeur, D. J. Kriegman, and A. L. Yuille, "The bas-relief ambiguity," *International journal of computer vision*, vol. 35, no. 1, pp. 33–44, 1999.

[2] J. Kerber, M. Wang, J. Chang, J. J. Zhang, A. Belyaev, and H.-P. Seidel, "Computer assisted relief generation—a survey," vol. 31, no. 8, pp. 2363–2377, 2012.

[3] Y.-W. Zhang, J. Wu, Z. Ji, M. Wei, and C. Zhang, "Computer-assisted relief modelling: A comprehensive survey," vol. 38, no. 2, pp. 521–534, 2019.

[4] A. Reichinger, S. Maierhofer, and W. Purgathofer, "High-quality tactile paintings," *Journal on Computing and Cultural Heritage (JOCCH)*, vol. 4, no. 2, pp. 1–13, 2011.

[5] Q. Zeng, R. R. Martin, L. Wang, J. A. Quinn, Y. Sun, and C. Tu, "Region-based bas-relief generation from a single image," *Graphical models*, vol. 76, no. 3, pp. 140–151, 2014.

[6] J. Wu, R. R. Martin, P. L. Rosin, X.-F. Sun, F. C. Langbein, Y.-K. Lai, A. D. Marshall, and Y.-H. Liu, "Making bas-reliefs from photographs of human faces," *Computer-Aided Design*, vol. 45, no. 3, pp. 671–682, 2013.

[7] Y.-W. Zhang, C. Zhang, W. Wang, Y. Chen, Z. Ji, and L. Hui, "Portrait relief modeling from a single image," *IEEE Transactions on Visualization and Computer Graphics*, 2019.

[8] Y. Liu, Z. Ji, Y.-W. Zhang, and G. Xu, "Example-driven modeling of portrait bas-relief," *Computer Aided Geometric Design*, p. 101860, 2020.

[9] M. Chai, L. Luo, K. Sunkavalli, N. Carr, S. Hadap, and K. Zhou, "High-quality hair modeling from a single portrait photo," *ACM Transactions on Graphics (TOG)*, vol. 34, no. 6, pp. 1–10, 2015.

[10] S. Tang, F. Tan, K. Cheng, Z. Li, S. Zhu, and P. Tan, "A neural network for detailed human depth estimation from a single image," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 7750–7759.

[11] D. Nehab, S. Rusinkiewicz, J. Davis, and R. Ramamoorthi, "Efficiently combining positions and normals for precise 3d geometry," *ACM transactions on graphics (TOG)*, vol. 24, no. 3, pp. 536–543, 2005.

[12] M. Loper, N. Mahmood, J. Romero, G. Pons-Moll, and M. J. Black, "Smpl: A skinned multi-person linear model," *ACM transactions on graphics (TOG)*, vol. 34, no. 6, p. 248, 2015.

[13] T.-P. Wu, J. Sun, C.-K. Tang, and H.-Y. Shum, "Interactive normal reconstruction from a single image," *ACM Transactions on Graphics (TOG)*, vol. 27, no. 5, pp. 1–9, 2008.

[14] P. Cignoni, C. Montani, and R. Scopigno, "Computer-assisted generation of bas-and high-reliefs," *Journal of graphics tools*, vol. 2, no. 3, pp. 15–28, 1997.

[15] T. Weyrich, J. Deng, C. Barnes, S. Rusinkiewicz, and A. Finkelstein, "Digital bas-relief from 3d scenes," *ACM transactions on graphics (TOG)*, vol. 26, no. 3, pp. 32–es, 2007.

[16] Z. Ji, W. Ma, and X. Sun, "Bas-relief modeling from normal images with intuitive styles," *IEEE transactions on visualization and computer graphics*, vol. 20, no. 5, pp. 675–685, 2013.

[17] M. Wei, Y. Tian, W.-M. Pang, C. C. Wang, M.-Y. Pang, J. Wang, J. Qin, and P.-A. Heng, "Bas-relief modeling from normal layers," *IEEE transactions on visualization and computer graphics*, vol. 25, no. 4, pp. 1651–1665, 2018.

[18] J. Nie, W. Shi, Y. Liu, H. Gao, F. Xu, Z. Zhang, and G. Jiang, "Bas-relief generation from point clouds based on normal space compression with real-time adjustment on cpu," *arXiv preprint arXiv:1912.13140*, 2019.

[19] M. Alexa and W. Matusik, "Reliefs as images." *ACM Trans. Graph.*, vol. 29, no. 4, pp. 60–1, 2010.

[20] D. Sỳkora, L. Kavan, M. Čadík, O. Jamriška, A. Jacobson, B. Whited, M. Simmons, and O. Sorkine-Hornung, "Ink-and-ray: Bas-relief meshes for adding global illumination effects to hand-drawn characters," *ACM Transactions on Graphics (TOG)*, vol. 33, no. 2, pp. 1–15, 2014.

[21] M. Kolomenkin, G. Leifman, I. Shimshoni, and A. Tal, "Reconstruction of relief objects from line drawings," in *CVPR 2011*. IEEE, 2011, pp. 993–1000.

[22] X. Sun, P. L. Rosin, R. R. Martin, and F. C. Langbein, "Bas-relief generation using adaptive histogram equalization," *IEEE transactions on visualization and computer graphics*, vol. 15, no. 4, pp. 642–653, 2009.

[23] D. Anguelov, P. Srinivasan, D. Koller, S. Thrun, J. Rodgers, and J. Davis, "Scape: shape completion and animation of people," in *ACM SIGGRAPH 2005 Papers*, 2005, pp. 408–416.

[24] C. Ionescu, D. Papava, V. Olaru, and C. Sminchisescu, "Human3. 6m: Large scale datasets and predictive methods for 3d human sensing in natural environments," *IEEE transactions on pattern analysis and machine intelligence*, vol. 36, no. 7, pp. 1325–1339, 2013.

[25] G. Varol, J. Romero, X. Martin, N. Mahmood, M. J. Black, I. Laptev, and C. Schmid, "Learning from synthetic humans," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 109–117.

[26] C. Lassner, J. Romero, M. Kiefel, F. Bogo, M. J. Black, and P. V. Gehler, "Unite the people: Closing the loop between 3d and 2d human representations," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 6050–6059.

[27] G. Varol, D. Ceylan, B. Russell, J. Yang, E. Yumer, I. Laptev, and C. Schmid, "Bodynet: Volumetric inference of 3d human body shapes," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 20–36.

[28] G. Pavlakos, L. Zhu, X. Zhou, and K. Daniilidis, "Learning to estimate 3d human pose and shape from a single color image," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 459–468.

[29] A. Kanazawa, M. J. Black, D. W. Jacobs, and J. Malik, "End-to-end recovery of human shape and pose," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 7122–7131.

[30] N. Kolotouros, G. Pavlakos, M. J. Black, and K. Daniilidis, "Learning to reconstruct 3d human pose and shape via model-fitting in the loop," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 2252–2261.

[31] M. Omran, C. Lassner, G. Pons-Moll, P. Gehler, and B. Schiele, "Neural body fitting: Unifying deep learning and model based human pose and shape estimation," in *2018 international conference on 3D vision (3DV)*. IEEE, 2018, pp. 484–494.

[32] R. A. Guler and I. Kokkinos, "Holopose: Holistic 3d human reconstruction in-the-wild," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 10 884–10 894.

[33] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh, "Realtime multi-person 2d pose estimation using part affinity fields," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 7291–7299.

[34] M. Fieraru, A. Khoreva, L. Pishchulin, and B. Schiele, "Learning to refine human pose estimation," in *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 2018, pp. 205–214.

[35] G. Pavlakos, X. Zhou, K. G. Derpanis, and K. Daniilidis, "Coarse-to-fine volumetric prediction for single-image 3d human pose," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 7025–7034.

[36] X. Sun, B. Xiao, F. Wei, S. Liang, and Y. Wei, "Integral human pose regression," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 529–545.

[37] D. Mehta, O. Sotnychenko, F. Mueller, W. Xu, M. Elgharib, P. Fua, H.-P. Seidel, H. Rhodin, G. Pons-Moll, and C. Theobalt, "Xnect: Real-time multi-person 3d human pose estimation with a single rgb camera," *arXiv preprint arXiv:1907.00837*, 2019.

[38] F. Bogo, A. Kanazawa, C. Lassner, P. Gehler, J. Romero, and M. J. Black, "Keep it smpl: Automatic estimation of 3d human pose and shape from a single image," in *European Conference on Computer Vision*. Springer, 2016, pp. 561–578.

[39] T. Alldieck, M. Magnor, B. L. Bhatnagar, C. Theobalt, and G. Pons-Moll, "Learning to reconstruct people in clothing from a single rgb camera," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 1175–1186.

[40] T. Alldieck, G. Pons-Moll, C. Theobalt, and M. Magnor, "Tex2shape: Detailed full human body geometry from a single image," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 2293–2303.

[41] W. Jiang, N. Kolotouros, G. Pavlakos, X. Zhou, and K. Daniilidis, "Coherent reconstruction of multiple humans from a single image," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 5579–5588.

[42] R. Natsume, S. Saito, Z. Huang, W. Chen, C. Ma, H. Li, and S. Morishima, "Siclope: Silhouette-based clothed people," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 4480–4490.

[43] R. Alp Güler, N. Neverova, and I. Kokkinos, "Densepose: Dense human pose estimation in the wild," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 7297–7306.

[44] D. Smith, M. Loper, X. Hu, P. Mavroidis, and J. Romero, "Facsimile: Fast and accurate scans from an image in less than a second," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 5330–5339.

[45] V. Gabeur, J.-S. Franco, X. Martin, C. Schmid, and G. Rogez, "Moulding humans: Non-parametric 3d human shape estimation from single images," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 2232–2241.

[46] Z. Li, T. Dekel, F. Cole, R. Tucker, N. Snavely, C. Liu, and W. T. Freeman, "Learning the depths of moving people by watching frozen people," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 4521–4530.

[47] F. Tan, H. Zhu, Z. Cui, S. Zhu, M. Pollefeys, and P. Tan, "Self-supervised human depth estimation from monocular videos," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 650–659.

[48] Y. Jafarian and H. S. Park, "Learning high fidelity depths of dressed humans by watching social media dance videos," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 12 753–12 762.

[49] Z. Zheng, T. Yu, Y. Wei, Q. Dai, and Y. Liu, "Deephuman: 3d human reconstruction from a single image," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 7739–7749.

[50] S. Saito, Z. Huang, R. Natsume, S. Morishima, A. Kanazawa, and H. Li, "Pifu: Pixel-aligned implicit function for high-resolution

clothed human digitization," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 2304–2314.

[51] S. Saito, T. Simon, J. Saragih, and H. Joo, "Pifuhd: Multi-level pixel-aligned implicit function for high-resolution 3d human digitization," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 84–93.

[52] J. Martinez, R. Hossain, J. Romero, and J. J. Little, "A simple yet effective baseline for 3d human pose estimation," in *ICCV*, 2017.

[53] K. Zhou, J. Huang, J. Snyder, X. Liu, H. Bao, B. Guo, and H.-Y. Shum, "Large mesh deformation using the volumetric graph laplacian," *ACM Transactions on Graphics (TOG)*, vol. 24, no. 3, pp. 496–503, 2005.

[54] E. S. Ho, T. Komura, and C.-L. Tai, "Spatial relationship preserving character motion adaptation," *ACM Transactions on Graphics (TOG)*, vol. 29, no. 4, pp. 1–8, 2010.

[55] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask r-cnn," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2961–2969.

[56] S. Xie and Z. Tu, "Holistically-nested edge detection," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1395–1403.

[57] R. Durstenfeld, "Algorithm 235: random permutation," *Communications of the ACM*, vol. 7, no. 7, p. 420, 1964.

[58] H. Chui and A. Rangarajan, "A new point matching algorithm for non-rigid registration," *Computer Vision and Image Understanding*, vol. 89, no. 2-3, pp. 114–141, 2003.

[59] F. L. Bookstein, "Principal warps: Thin-plate splines and the decomposition of deformations," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 11, no. 6, pp. 567–585, 1989.

[60] S. Gold, A. Rangarajan, C.-P. Lu, S. Pappu, and E. Mjolsness, "New algorithms for 2d and 3d point matching: pose estimation and correspondence," *Pattern recognition*, vol. 31, no. 8, pp. 1019–1031, 1998.

[61] V. Kazemi and J. Sullivan, "One millisecond face alignment with an ensemble of regression trees," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 1867–1874.

[62] C. Cao, Y. Weng, S. Zhou, Y. Tong, and K. Zhou, "Facewarehouse: A 3d facial expression database for visual computing," *IEEE Transactions on Visualization and Computer Graphics*, vol. 20, no. 3, pp. 413–425, 2013.

[63] J. Liu, J. Sun, and H.-Y. Shum, "Paint selection," in *ACM SIGGRAPH 2009 Papers*, ser. SIGGRAPH '09. New York, NY, USA: Association for Computing Machinery, 2009. [Online]. Available: https://doi.org/10.1145/1576246.1531375

[64] P. Pérez, M. Gangnet, and A. Blake, "Poisson image editing," *ACM Transactions on graphics (TOG)*, vol. 22, no. 3, pp. 313–318, 2003.

[65] G. Guennebaud, B. Jacob *et al.*, "Eigen v3," http://eigen.tuxfamily.org, 2010.

[66] S. Agarwal, K. Mierle, and Others, "Ceres solver," http://ceres-solver.org.

[67] G. Bradski, "The OpenCV Library," *Dr. Dobb's Journal of Software Tools*, 2000.

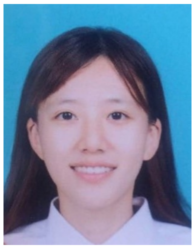
[68] M. K. Johnson and E. H. Adelson, "Shape estimation in natural illumination," in *CVPR 2011*. IEEE, 2011, pp. 2553–2560.

[69] J. T. Barron and J. Malik, "Shape, illumination, and reflectance from shading," *IEEE transactions on pattern analysis and machine intelligence*, vol. 37, no. 8, pp. 1670–1687, 2014.

[70] P. Wang and A. Yuille, "Doc: Deep occlusion estimation from a single image," in *European Conference on Computer Vision*. Springer, 2016, pp. 545–561.

[71] H. Joo, T. Simon, and Y. Sheikh, "Total capture: A 3d deformation model for tracking faces, hands, and bodies," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 8320–8329.

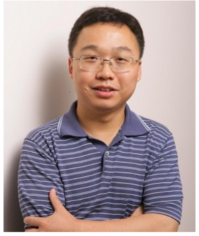
Zhenjie Yang received the bachelor's degree in Computer Science from Sun Yat-sen University in 2017. Currently, he is working toward the MSc degree at the State Key Lab of CAD&CG, Zhejiang University. His research interests include geometry processing and image analysis.



Beijia Chen is a Ph.D. candidate at the State Key Lab of CAD&CG, Zhejiang University. She obtained her B.S. and M.S. degree in Nanjing University of Science and Technology. Her research interests include 3d human recovery, image manipulation, and deep learning.

Youyi Zheng is a Researcher (PI) at the State Key Lab of CAD&CG, College of Computer Science, Zhejiang University. He obtained his Ph.D. from the Department of Computer Science and Engineering at Hong Kong University of Science and Technology, and his M.Sc. and B.Sc. degrees in Mathematics, both from Zhejiang University. His research interests include geometric modeling, imaging, and human-computer interaction.

Xiang Chen is an Associate Professor in the State Key Lab of CAD&CG, Zhejiang University. He received his Ph.D. in Computer Science from Zhejiang University in 2012. His current research interests mainly include fabrication-aware design, physics-based simulation, image analysis/editing, shape modeling/retrieval and computer-aided design.



Kun Zhou is a Cheung Kong Professor in the Computer Science Department of Zhejiang University and the Director of the State Key Lab of CAD&CG. He received his BS degree and PhD degree in computer science, both from Zhejiang University. After graduation he spent six years with Microsoft Research Asia, and was a lead researcher of the graphics group before moving back to Zhejiang University. He was elected as an IEEE Fellow in 2015, and an ACM Fellow in 2020.

Human Bas-Relief Generation from A Single Photograph

Supplemental Material

Zhenjie Yang* Beijia Chen* Youyi Zheng* Xiang Chen* Kun Zhou*

* Zhejiang University

1 Robust Contour Points Matching

1.1 Solving the *tps* transform with a known correspondence

Given two 2d point sets $\{\mathbf{x}_i\}$ and $\{\mathbf{y}_i\}$ of size N with one-to-one correspondence, we find a transform function f by minimizing:

$$E_{\text{general}}(f) = \sum_{i=1}^{K} (\mathbf{y}_i - f(\mathbf{x}_i))^2 + \lambda \mathcal{J}(f), \tag{1}$$

where the second term controls the smoothness of the transform.

A *tps* function can be represented as two matrices $\mathbf{d}$ and $\mathbf{c}$:

$$f_{\text{tps}}(\mathbf{x}) = \mathbf{x} \cdot \mathbf{d} + \phi(\mathbf{x}) \cdot \mathbf{c}, \tag{2}$$

where $\mathbf{x}$ is the homogeneous coordinates of an arbitrary point in the space of $\mathbf{x}_i$, $\mathbf{d}$ is a 3×3 matrix for the global affine transform and $\mathbf{c}$ is a $N \times 3$ matrix for the local non-rigid transform. The kernel function $\phi(\mathbf{x})$ a $1 \times N$ vector for each point $\mathbf{x}$, where $\phi_i(\mathbf{x}) = ||\mathbf{x} - \mathbf{x}_i||^2 log||\mathbf{x} - \mathbf{x}_i||$.

By substituting f_{tps} into Equation 1, E_{general} can be written as:

$$E_{\text{tps}}(\mathbf{d}, \mathbf{c}) = ||\mathbf{Y} - (\mathbf{X}\mathbf{d} + \mathbf{\Phi}\mathbf{c})||^2 + \lambda \text{trace}(\mathbf{c}^T \mathbf{\Phi} \mathbf{c}), \tag{3}$$

where $\mathbf{X}$, $\mathbf{Y}$ and $\mathbf{\Phi}$ are the stack version of $\mathbf{x}$, $\mathbf{y}$ and $\phi(\mathbf{x})$, and the entry in $\mathbf{\Phi}$ is $\mathbf{\Phi}_{ij} = ||\mathbf{x}_i - \mathbf{x}_j||^2 \log||\mathbf{x}_i - \mathbf{x}_j||$.

A good way to minimize E_{tps} (Equation 3) is to solve the following equations [1]:

$$\begin{cases} \mathbf{P}\mathbf{c} + \mathbf{X}\mathbf{d} = \mathbf{Y} & \text{(4a)} \\ \mathbf{X}^T \mathbf{c} = \mathbf{0} & \text{(4b)} \end{cases}$$

in which $\mathbf{P} = \mathbf{\Phi} + \lambda \mathbf{I}$. By performing Qr decomposition of $\mathbf{X}$, we can compute $\mathbf{c}$ and $\mathbf{d}$ as follows [1]:

$$\mathbf{X} = (\mathbf{Q}_1 : \mathbf{Q}_2) \begin{pmatrix} \mathbf{R} \\ \mathbf{0} \end{pmatrix} \tag{5}$$

$$\begin{cases} \mathbf{c} = \mathbf{Q}_2(\mathbf{Q}_2^T \mathbf{P} \mathbf{Q}_2)^{-1} \mathbf{Q}_2^T \mathbf{Y} & \text{(6a)} \\ \mathbf{d} = \mathbf{R}^{-1} \mathbf{Q}_1^T (\mathbf{Y} - \mathbf{T}\mathbf{c}) & \text{(6b)} \end{cases}$$

1.2 Solving the *tps* transform and the unknown correspondence

Now, we have two sets $\{\mathbf{x}_i\}$ and $\{\mathbf{y}_i\}$, but their correspondence is unknown. If we want to find a transform f, we have to incorporate an additional robustness-term of the correspondence $\mathbf{Z}$ into the following minimization

$$\min_{\mathbf{Z}, f} \sum_{i=1}^{N} \sum_{j=1}^{K} Z_{ij} \left\|\mathbf{y}_j - f(\mathbf{x}_i)\right\|^2 + \lambda \left\|\mathsf{L}(f)\right\|^2 - \xi \sum_{i=1}^{N} \sum_{j=1}^{K} Z_{ij}, \tag{7}$$

subject to $Z_{ij} \in \{0, 1\}$, $\sum_{j=1}^{K+1} Z_{ij} = 1$, and $\sum_{i=1}^{N+1} Z_{ij} = 1$. Here we have $N = \left|\mathcal{P}_{\mathcal{G}}\right|, K = |\mathcal{P}_{\mathcal{I}}|$ and $\mathbf{x}_i \in \mathcal{P}_{\mathcal{G}}, \mathbf{y}_j \in \mathcal{P}_{\mathcal{I}}$. The first term represents the approximation fidelity, the second term represents a constraint on the smoothness of f, and the third term penalized the number of outliers.

It is natural to minimize Equation 7 alternatively on f and $\mathbf{Z}$.

When only considering $\mathbf{Z}$, we minimize

$$E(\mathbf{Z}) = \sum_{i=1}^{N}\sum_{j=1}^{K} Z_{ij} D_{ij}, \tag{8}$$

where $D_{ij} = \xi - \|\mathbf{y}_j - f(\mathbf{x}_i)\|^2$.

When only considering f, similar to subsection 1.1, we have following equations:

$$\begin{cases} \hat{\mathbf{P}}\mathbf{c} + \mathbf{X}\mathbf{d} = \hat{\mathbf{Y}} & \text{(9a)} \\ \mathbf{X}^T\mathbf{c} = 0 & \text{(9b)} \end{cases}$$

in which

$$\begin{cases} \hat{\mathbf{P}} = \mathbf{\Phi} + \lambda \mathbf{W}^{-1} & \text{(10a)} \\ \mathbf{W} = \mathrm{diag}(w_1^2, ..., w_n^2) & \text{(10b)} \\ w_i = 1/\sum_{j=1}^{K} Z_{ij}, i = 1, ..., N & \text{(10c)} \\ \hat{\mathbf{Y}} = (\hat{\mathbf{y}}_1^T, ..., \hat{\mathbf{y}}_n^T) & \text{(10d)} \\ \hat{\mathbf{y}}_i = \sum_{j=1}^{K} Z_{ij}\mathbf{y}_j / \sum_{j=1}^{K} Z_{ij}, i = 1, ..., N & \text{(10e)} \end{cases}$$

To avoid the space-jumping in $\mathbf{Z}$, we adopt the TPS-RPM [2] to process Equation 7. In Algorithm 1, the temperature T is for relaxing the binary correspondence matrix $\mathbf{Z}$ to the doubly stochastic matrix $\mathbf{M}$. The variables f and $\mathbf{M}$ are alternatively optimized via an annealing schedule.

Algorithm 1 TPS-RPM

```
Require: $T_0$ initial temperature,$T_1$ final temperature,$r$ attenuation rate
$T \leftarrow T_0$
while $T > T_1$ do                                                        ▷ Deterministic annealing
    $\beta \leftarrow \frac{1}{T}$
    $\hat{M}^0_{ij} \leftarrow \exp(\beta D_{ij})$
    while $\hat{\mathbf{M}}$ not converges do                                  ▷ Sinkhorn method
        $\hat{M}^1_{ij} \leftarrow \hat{M}^0_{ij} / \sum_{k=1}^{K+1} \hat{M}^0_{ik}, i = 1, ..., N$          ▷ Nomalize rows
        $\hat{M}^0_{ij} \leftarrow \hat{M}^1_{ij} / \sum_{k=1}^{N+1} \hat{M}^1_{kj}, j = 1, ..., K$          ▷ Nomalize cols
    end while
    $\mathbf{M} \leftarrow \hat{\mathbf{M}}^0$
    $\mathbf{c} \leftarrow \mathbf{Q}_2(\mathbf{Q}_2^T\hat{\mathbf{P}}\mathbf{Q}_2)^{-1}\mathbf{Q}_2^T\hat{\mathbf{Y}}$
    $\mathbf{d} \leftarrow \mathbf{R}^{-1}\mathbf{Q}_1^T(\hat{\mathbf{Y}} - \mathbf{Tc})$
    $T \leftarrow rT$
end while
```

2 The Algorithms Used in Section 6.1

See Algorithm 3 for the overall pipeline of base shape generation, Algorithm 3 for the warping of base normal map, and Algorithm 4 for the normal map construction from the segmentation mask, *i.e.*, the inflation process.

Algorithm 2 Base Shape Generation

```
Warp normal map rendered from the guide model
Generate height field of body region from the warped normal map
Fit a 3D face model to the detected facial landmarks
Generate height field of the face region
Compute normal maps of user-masked regions
Generate height field from the normal maps
Compose these height fields with Poisson Editing
```

Algorithm 3 Normal Map Warping

```
function WarpNormalMap($normalMap[\ ][\ ]$, $f_{\mathsf{tps}}$)
    Initialize $warpedNormalMap[\ ][\ ]$ with $(0, 0, 0)$
    for each point $\mathbf{p}_d$ in $warpedNormalMap$ do
        $\mathbf{p} \leftarrow f_{\mathsf{tps}}(\mathbf{p}_d)$
        if $\mathbf{p}$ is valid in $normalMap$ then
            $warpedNormalMap(\mathbf{p}_d) \leftarrow$ BilinearInterpolateNormal($normalMap[\ ][\ ], \mathbf{p}$)
        end if
    end for
    return $warpedNormalMap$
end function

function BilinearInterpolateNormal($normalMap[\ ][\ ]$, $\mathbf{p}$)
    $x \leftarrow p_x$, $y \leftarrow p_y$
    $x_1 \leftarrow \mathrm{floor}(x)$, $x_2 \leftarrow \mathrm{ceil}(x)$, $y_1 \leftarrow \mathrm{floor}(y)$, $y_2 \leftarrow \mathrm{ceil}(y)$
    $\mathbf{q}_{11} \leftarrow (x_1, y_1)$, $\mathbf{q}_{12} \leftarrow (x_1, y_2)$, $\mathbf{q}_{21} \leftarrow (x_2, y_1)$, $\mathbf{q}_{22} \leftarrow (x_2, y_2)$
    $\mathbf{n}_{11} \leftarrow normalMap(\mathbf{q}_{11})$, $\mathbf{n}_{12} \leftarrow normalMap(\mathbf{q}_{12})$
    $\mathbf{n}_{21} \leftarrow normalMap(\mathbf{q}_{21})$, $\mathbf{n}_{22} \leftarrow normalMap(\mathbf{q}_{22})$
    $\mathbf{n} \leftarrow \begin{bmatrix} x_2 - x & x - x_1 \end{bmatrix} \begin{bmatrix} \mathbf{n}_{11} & \mathbf{n}_{12} \\ \mathbf{n}_{21} & \mathbf{n}_{22} \end{bmatrix} \begin{bmatrix} y_2 - y \\ y - y_1 \end{bmatrix}$
    if $||\mathbf{n}|| \neq 0$ then
        $\mathbf{n} \leftarrow \mathbf{n}/||\mathbf{n}||$
    end if
    return $\mathbf{n}$
end function
```

Algorithm 4 Normal Map from Mask

```
Apply Gaussian filter to mask $\mathsf{m}$ (with a nonzero region $\Omega$) to obtain $\mathsf{s}$
Compute the gradient of $\mathsf{s}$ as $\mathsf{g}$
Compute $\mathsf{n}_x$ and $\mathsf{n}_y$ by solving the Laplace's equation with the boundary condition $\mathsf{n}(p) = \mathsf{g}(p), \forall p \in \partial\Omega$
for each point $(u, v)$ in $\Omega$ do
    $x \leftarrow \mathsf{n}_x(u, v),\ y \leftarrow \mathsf{n}_y(u, v)$
    $\mathbf{n} \leftarrow [x, y, 1]^{\mathsf{T}}$
    $\mathbf{n} \leftarrow \mathbf{n}/||\mathbf{n}||$
    $\mathsf{n}(u, v) \leftarrow \mathbf{n}$
end for
Return the normal map $\mathsf{n}$
```

3 Integration

Slightly different from the original method [3], we conduct the computations with orthogonal projection since the perspective effect is weak in our scenario. Then for each 2D point $\mathbf{p} = [u, v]^T$, the relation between the 3D coordinate $\mathbf{V}(\mathbf{p})$ and the captured height map $h(\mathbf{p})$ is:

$$\mathbf{V}(\mathbf{p}) = [u, v, h(\mathbf{p})]^T$$

Under this assumption, the tangent vectors at each point can be computed as

$$\mathbf{T}_u(\mathbf{p}) = \frac{\partial \mathbf{V}(\mathbf{p})}{\partial u} = \left[1, 0, \frac{\partial h(\mathbf{p})}{\partial u}\right]^T$$

$$\mathbf{T}_v(\mathbf{p}) = \frac{\partial \mathbf{V}(\mathbf{p})}{\partial v} = \left[0, 1, \frac{\partial h(\mathbf{p})}{\partial v}\right]^T$$

which lead to a simpler linear system to solve.

4 More comparisons with PIFuHD [4]

The method presented in PIFuHD [4] does not directly support multi-person images, where body parts are often missed. The results also exhibit depth artifacts and fine-scale noise (see Figure 1).

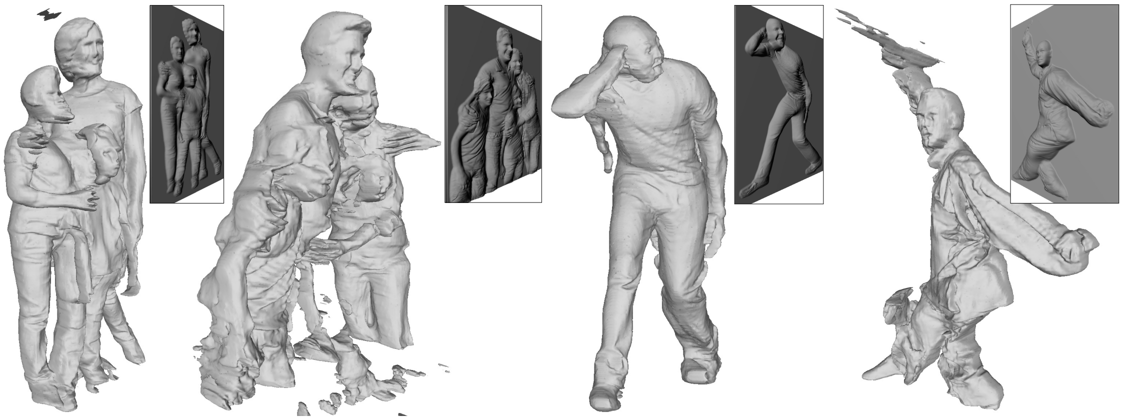


Figure 1: The side views of full-body reconstruction results of PIFuHD [4]. The corresponding frontal views are shown in the 3rd, 5th, 8th, and 9th columns of Figure 11 in the paper. The insets show the side views of our bas-reliefs for reference.

References

[1] G. Wahba, *Spline models for observational data*. Siam, 1990, vol. 59.

[2] H. Chui and A. Rangarajan, "A new point matching algorithm for non-rigid registration," *Computer Vision and Image Understanding*, vol. 89, no. 2-3, pp. 114–141, 2003.

[3] D. Nehab, S. Rusinkiewicz, J. Davis, and R. Ramamoorthi, "Efficiently combining positions and normals for precise 3d geometry," *ACM transactions on graphics (TOG)*, vol. 24, no. 3, pp. 536–543, 2005.

[4] S. Saito, T. Simon, J. Saragih, and H. Joo, “Pifuhd: Multi-level pixel-aligned implicit function for high-resolution 3d human digitization,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 84–93.