

FACSIMILE: Fast and Accurate Scans From an Image in Less Than a Second

David Smith Matthew Loper Xiaochen Hu Paris Mavroidis Javier Romero
Amazon Body Labs
{dlsmith,mloper,sonnyh,parism,javier}@amazon.com



Figure 1: FAX converts a single RGB image (a) into a scan (b, d) with albedo texture (c, e)

Abstract

Current methods for body shape estimation either lack detail or require many images. They are usually architecturally complex and computationally expensive. We propose FACSIMILE (FAX), a method that estimates a detailed body from a single photo, lowering the bar for creating virtual representations of humans. Our approach is easy to implement and fast to execute, making it easily deployable. FAX uses an image-translation network which recovers geometry at the original resolution of the image. Counterintuitively, the main loss which drives FAX is on per-pixel surface normals instead of per-pixel depth, making it possible to estimate detailed body geometry without any depth supervision. We evaluate our approach both qualitatively and quantitatively, and compare with a state-of-the-art method.

1. Introduction

High resolution body capture has not seen widespread adoption, despite a myriad of applications in medicine, gaming, and shopping. Traditional methods for high-quality body estimation require expensive capture systems which are difficult to deploy [28, 8]. More affordable RGB-D sensors like kinect have tried to overcome this problem [47, 6], though those sensors are not as widespread as RGB cam-

eras. On the other hand, modern systems for single-photo body estimation lack detail [10, 31, 2, 22, 7, 33]. Our work is designed to help close the gap between an easily acquired image and a rich, detailed, reposeable avatar.

Systems targetted to recover shape from single images do a laudable job at recovering intermediate body representations. These include voxel-based reconstruction in [44], the synthetic-view generation system in [31], or the cross-modal neural nets in [10]. But inevitably, the fidelity of their capture is limited by the granularity of their representation.

To address this lack of representational power, we apply modern image-to-image translation techniques [19, 46] to geometry estimation. More concretely, we would like to estimate the depth corresponding to every foreground pixel in the image. But this presents a new problem: the naive estimation of depth via an image translation network creates noisy, unusable surfaces (Figure 2). This teaches us that when estimating depth with image-to-image translation, a direct loss on depth fails to give us a plausible surface.

The solution to this problem can be traced all the way back to Shape From Shading (SFS) literature by Horn [18], in which surface normals play a critical role in defining the relationship between a surface and its appearance. Work focused in the reconstruction of the face region [35] has shown that a loss on depth can benefit from an additional loss on normals. We go beyond this insight showing that a

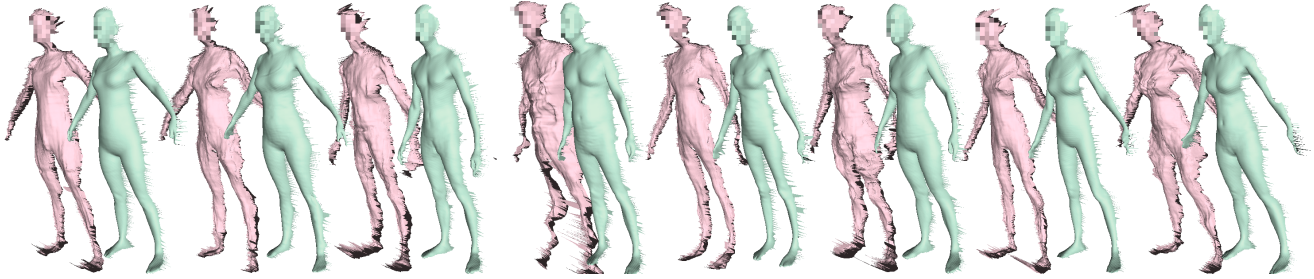


Figure 2: Frontal meshes estimated using (pink) an $L1$ loss on depth and (green) an $L1$ loss on normals.

loss just on normals can be *sufficient* to reconstruct a high-quality depth map up to scale, and that this applies for an articulated, far from spherical object.

Because a single depthmap is still far from an entire avatar, we extended the system to estimate front and back-facing geometry and albedo. Similar to the concurrent work in [31], we exploit the idea of obtaining two values per pixel by training the network to hypothesize the back side of the person (see Figure 3). Unlike [31], we do not restrict ourselves to texture and also estimate the back depth and normals. While current detailed methods like [31, 2] typically take several minutes to run, we compute an almost complete scan containing geometry and texture in less than one second. In this publication we assume a *cooperative subject* and focus on a specific type of image that maximizes information capture (frontal arms-down pose, minimal clothing), although we believe the method could be applied to other cases and will continue investigating them in future work.

We demonstrate three contributions. First, we compute full scans from a single image, orders of magnitude faster than current methods producing detailed scans. Although other methods also reproduce garments, our method extracts significantly more detail. We encourage the reader to review the scans in figures 1 and 7 and the supplementary material, paying special attention to subtle folds and compression artifacts in the chest, waist or hips, not present in any other methods. Second, we show how these scans can be converted into detailed deformable avatars with little additional time (less than 10 seconds), which can be valuable for applications like gaming, measurements from an image, and virtual telepresence. Finally, we illustrate the efficacy of our method by comparing it quantitatively against the state-of-the-art multi-image method [3] and performing a qualitative and quantitative ablation study.

2. Related Work

Geometry estimation from a single-photo has been a topic of research for at least 50 years. Classic methods like *shape from shading* [17] take shading images and produce the underlying geometry. Modern solutions to this problem can be computationally efficient and intuitive [48, 4], but

the limitations of the light and distribution models applied to the data make them brittle in the presence of input noise, which is unavoidable in real data. Deep learning based methods have achieved impressive results in reducing this brittleness in outdoor depth reconstruction for autonomous driving [13] and indoor geometry reconstruction [12, 43].

Single-photo body estimation methods typically bottleneck through fixed intermediate representations, which while enabling piecewise modeling, ultimately limit the amount of achievable detail. Some methods bottleneck through segmented images [21, 15, 38, 10, 33], others through estimated keypoints positions [7, 26], and some through both [44, 34, 14, 1]. All such methods permit too much ambiguity to allow for dense surface reconstruction. Recent methods [22] avoid this limitation by using encoder-decoder representations directly on the image. They achieve remarkable robustness to images in the wild, but struggle to recover detailed shape and pose. Work on SURREAL [45] estimates depth directly, but with coarse detail. The SiC-LoPe [31] system tolerates greater clothing variation than our system, but its geometric detail is limited by the use of intermediate silhouettes. To the credit of these works, all but [10] were designed for capturing bodies “in the wild” with tolerance for pose variation, whereas our goal is to capture a detailed avatar from a restricted pose.

Single-photo face estimation methods have produced useful insights for body estimation. Early work by Blanz and Vetter [5] was ground-breaking but suffered from lack of detail and problems with robustness *in the wild*. Robustness was addressed by data-driven models [9, 11, 20, 37, 40, 39, 41]; detail was addressed first by shape from shading [24, 27], and then by deep learning [36, 42, 35]. A recent survey by Zollhoffer et al [49] has more specifics. FAX specifically shares themes with [36, 42], in which the image-to-image translation architecture from Isola et al [19] is successfully applied to detailed face geometry estimation.

Our focus is on avatar geometry estimation from a single color image. For a more general review of body estimation from multiple images, readers are advised to review the excellent summaries of previous work provided in Alldieck et al [2] and Bogo et al [8].

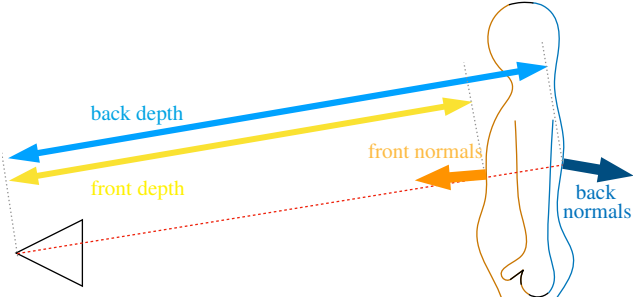


Figure 3: Depth, surface normals and albedo are computed for the body points closest and furthest from the camera along the optic ray entailed by each pixel. Note the presence of pixels which remain unobserved and become holes in the inferred scan (black contours in the image).

3. Method

Our goal is to estimate a detailed 3D scan from a single RGB image. We treat this as an image-to-image translation task, where we *translate* an image to depth and albedo values in image space. More specifically, we estimate those outputs for both the front- and back-facing portions of the body. The depth images form regular grids of vertices, which can be trivially triangulated to create a 3D surface.

We describe our depth estimation architecture in more detail in Section 3.2, but focus first on albedo estimation in Section 3.1, since the training protocol closely resembles the prior work of [46]. Finally, we explain how to obtain a complete, reposable and reshapable avatar in Section 3.3.

3.1. Albedo estimation

Our architecture of choice is based on the image-to-image translation work of [46]. We omit features specific to semantic segmentation and image editing, as well as their “enhancer” networks. Thus we define our generator using their “global generator”, which is composed of a downsampling section, followed by a number of residual [16] blocks, and completed with an upsampling section that restores the feature maps to the input resolution. We make one minor modification by replacing transposed convolutions with upsample-convolutions to avoid checkerboard artifacts [32].

The loss in [46] is composed of three terms: an adversarial loss, \mathcal{L}_{GAN} using a multi-scale PatchGAN [19] discriminator with an LSGAN [30] objective; a feature matching loss, \mathcal{L}_{FM} , which penalizes discrepancies between the internal discriminator activations from the generated G vs. real images y ; and a perceptual loss, \mathcal{L}_{VGG} , which uses a pre-trained VGG19 network, and similarly measures the different VGG activations from real and generated images:

$$\mathcal{L}_{alb}(x, y^i) = \mathcal{L}_{GAN}(x, y^i, G^i) + \lambda_{FM} \mathcal{L}_{FM}(x, y^i, G^i) + \lambda_{VGG} \mathcal{L}_{VGG}(y^i, G^i) \quad (1)$$

where i indexes front and back. Every generated image $G^i(x)$ depends on the input image x , so we drop this dependency from now on to simplify notation. Front and back albedo use the same loss components, though employ separate discriminators for the front and back estimates, enabling them to specialize. The application of this network to our problem of albedo estimation is straightforward. Given synthetic training data (see Section 4.2) of images and the corresponding front and back albedo, we estimate G with six channels corresponding to the two albedo sets (center of Figure 4). The total loss is the sum of losses applied to front and back, $\mathcal{L}_{alb}(x, y^f) + \mathcal{L}_{alb}(x, y^b)$.

3.2. Depth estimation

Motivation As previously explained, direct estimation of depth is challenging due to various reasons. First, there is an ambiguity between scale and distance to the camera difficult to resolve even by humans. And second, this distance to the camera entails a much larger data variance than shape details. Therefore, a loss on depth encourages the network to solve the overall distance to the camera, which is a very challenging and mostly irrelevant problem for our purpose. Instead, we focus on inferring *local* surface geometry, which is invariant to scale ambiguities.

In initial experiments we managed to estimate detailed surface normals through the direct application of the image-translation network described in Section 3.1. However, integrating normals into robust depth *efficiently* is a challenging problem at the core of shape from shading literature. While integration of inferred normal images is challenging and expensive, its inverse operator is simple: the spatial derivative. Spatial derivatives can be implemented simply as a fixed layer with a local difference filter. By placing such layer directly behind the estimated normals (see δ layer in Figure 4), we are implicitly forcing the previous result to correspond to depth. Similar to the classic integration approach, this allows us to infer depth even in the absence of depth ground truth data, but without the extra computational cost incurred by explicit integration.

Losses In our depth architecture (see Figure 4), the output is three channels and they represent the front and back depth G_d^i where i denotes front or back, as well as a mask G_m denoting where depth is valid. The front and back depth are processed with a spatial differentiation network δ that converts the depth into normals $G_n^i = \delta(G_d^i, G_m, f)$. This spatial differentiation depends on the focal length f (considered fixed in train and test data) to correct perspective distortion. Furthermore, the differentiation operator incorporates the mask G_m produced by the network, to ensure we do not differentiate through boundaries. In the areas where depth is not valid, a constant normal value is produced.

While albedo (or color in general) seems to clearly bene-

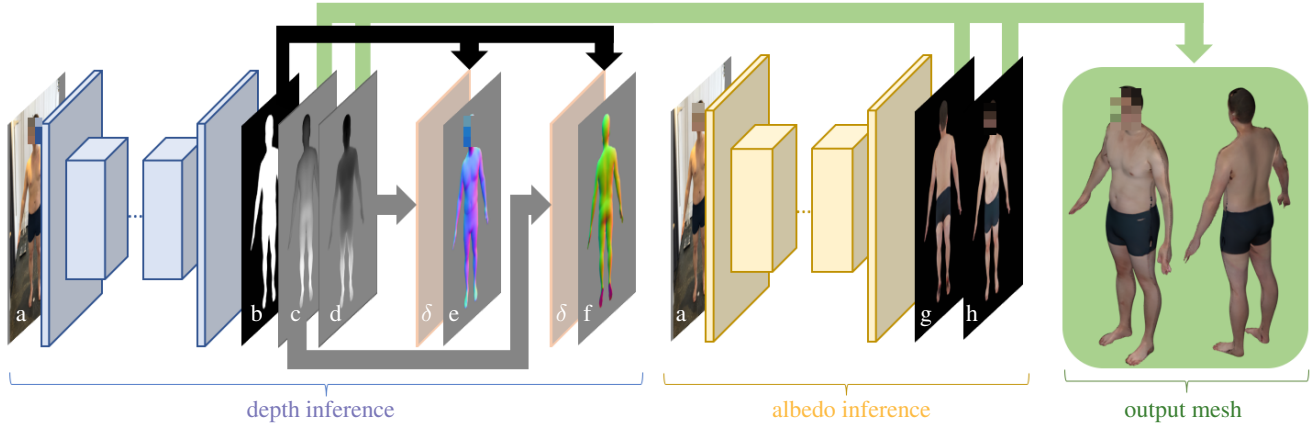


Figure 4: Network architecture for geometry inference (left) and albedo inference (right) from an image a . They share the architecture inspired by [46] in the first stage (blue and yellow, trained separately). In geometry, the network outputs three channels (mask b , front and back depth e and d), while in albedo six channels are produced (RGB back g and RGB front h). Depth channels are processed by fixed spatial differentiation layers δ which use the mask to limit its effect to the foreground area, resulting in front and back normals (e and f). For compactness, we do not show the front and back albedo discriminators.

fit from adversarial losses, the same does not seem to be true for recovering geometry. In our experience (similar to what is described in [36]), the adversarial loss in \mathcal{L}_{alb} introduces noise when applied to the problem of depth and normal estimation, and reduces its robustness to unseen conditions. For this reason, the depth \mathcal{L}_d and normal \mathcal{L}_n terms of our geometry estimation objective

$$\mathcal{L}_n^i(x, y) = \mathcal{L}_{L1}(y, G_n^i) + \lambda_{VGG} \mathcal{L}_{VGG}(y, G_n^i) \quad (2)$$

$$\mathcal{L}_d^i(x, y) = \mathcal{L}_{L1}(y, G_d^i) \quad (3)$$

replace the adversarial loss with an L1 loss. \mathcal{L}_{VGG} is not applied to the depth representation as this would require a normalization of the (unbounded) depth values that could cause training instability. The total loss can potentially include this geometric loss applied to normals and/or depth, as well as a binary cross entropy loss on the mask output

$$\mathcal{L}_{full}(x, y) = \lambda_d (\mathcal{L}_d^f + \mathcal{L}_d^b) + \lambda_n (\mathcal{L}_n^f + \mathcal{L}_n^b) + \lambda_{msk} \mathcal{L}_{msk}(y_m, G_m) \quad (4)$$

In Section 4.5 and Table 2, we study the contributions of these loss terms both qualitatively and quantitatively.

3.3. Estimating Dense Correspondence

The system described in the previous section produces per-pixel depth values, which are inherently incomplete. Moreover, since those values are created per pixel, they lack any semantic meaning (where is the nose, elbow, etc). In this section we adopt the mesh alignment process described in [6] to infer the non-visible (black parts in Figure 3) parts

of the body geometry based on SMPL [29], a statistical model of human shape and pose.

The alignment process deforms a set of free body vertices (referred to as *the mesh*) so that they are close to the pointcloud inferred in the previous section (referred to as *the scan*), while also being likely according to the SMPL body model. Similar to [6], we minimize a loss composed of a weighted average of a scan-to-mesh distance term E_s , a face landmark term E_{face} , two pose and shape priors E_{pose} and E_{shape} , and a term that couples the inferred free vertices with the model E_{cpt} . We provide some intuition about the terms in the following paragraphs, although the details can be obtained in the original publication.

E_s penalizes the squared 3D distance between the scan and closest points on the surface of the mesh. E_{face} penalizes the squared 3D distance between detected face landmarks [23] on the image (in implicit correspondence with the scan) and pre-defined landmark locations in SMPL. E_{cpt} encourages the mesh, which can deform freely, to stay close to the model implied by the optimized pose and shape parameters. E_{pose} and E_{shape} regularize pose and shape of the coupled model by penalizing the Mahalanobis distance between those SMPL parameters and their Gaussian distributions inferred from the CMU and SMPL datasets [7].

As it is common in single view and non-calibrated multi-view shape estimation, our results cannot recover the subjects scale accurately. Since SMPL cannot fit scan at arbitrary scales, we first scale the scan to a fixed height before optimizing the mesh, then apply the inverse scale to the optimized mesh, returning it to the original reference frame.

When training our depth estimator, the loss on depth acts as a global constraint, enforcing that the front and back



Figure 5: Each row shows two instances of synthetic data (one male, one female). For each example, left to right: RGB, front and back albedo, front and back depth, front and back normals, and segmentation. Note that these examples do not really belong to our training set, since the textures come only from test subjects who signed a special consent form.

scans be estimated at consistent scales. When this loss is omitted during training (see Section 4.5), the front and back scale are not necessarily coherent, and thus their relative scale must be optimized during mesh alignment. This can be accomplished by introducing a single additional free scale variable that is applied to the back vertices and optimized along with the mesh. When describing our experiments, we refer to this option as *opt back*.

4. Experiments

4.1. Training and evaluation details

For albedo estimation, we train on random crops of size 512×512 to comply with memory limitations. The multi-scale discriminators process images at $1 \times$, $\frac{1}{2} \times$, and $\frac{1}{4} \times$ resolutions. Losses are weighted as in [46]. For depth estimation, we train on 720×960 images, and work with a focal length of 720 pixels. We do not assume a fixed distance to the camera. Both albedo and depth estimation networks are trained for 180k steps with a batch size of one, and input images are augmented with gaussian blur, gaussian noise, hue, saturation, brightness, and contrast. The training process takes approximately 48 hours with a V100 Tesla GPU.

Evaluation is performed on 720×960 images. A single forward pass of either network takes about 100 milliseconds, while aligning SMPL to the scan takes 7 seconds.

4.2. Datasets

We train exclusively on synthetic datasets (Figure 5), and test on real images collected “in-lab” — i.e., in a well-lit, indoor environment, where images are captured by lab technicians, and subjects wear tight-fitting clothing and stand in an “A”-pose (see Figure 7).

We render 40,000 synthetic image tuples (1% held out each for validation and testing). The bodies have a base low-frequency geometry synthesized with SMPL, and high-frequency displacements captured in-lab. The SMPL shape

parameters are sampled from the CAESAR dataset and poses are sampled from a mix of (a) CAESAR poses and (b) a set of in-lab scan poses with arms varying from A-pose to relaxed. Textures and displacement maps, derived from 3D photogrammetry scans of people captured in-lab, are randomly sampled and applied to the base bodies, which increases the diversity of the input and output spaces.

The camera is fixed with zero rotation at the origin, and the body randomly translated and rotated to simulate a distance of roughly 2 meters with a slight downward tilt of the camera. Specifically, translation is sampled from $x \sim [-0.5, 0.5]$, $y \sim [0.0, 0.4]$, $z \sim [-2.2, -1.5]$ in meters and rotation as Euler angles in degrees from $x \sim [-9.0, 35]$, $y \sim [-7, 7]$, $z \sim [-2, 2]$, applied in yxz order. Background images are drawn from OpenImages [25], excluding images containing people.

We use three light sources: an image-based environment light (which uses the background image as a light source), a point light, and a rectangular area light. For each render, we randomly sample the intensity of all lights, the position and color temperature of the point and area lights, the orientation and size of the area light, and the specularity and roughness of the shader on the body. All light sources cast raytrace shadows, with the most visible generally coming from the area and point lights.

4.3. Visual Evaluation

As a baseline, we consider direct estimation of frontal depth with an $L1$ loss function. Figure 2 shows meshes estimated from natural test images, comparing models trained with an $L1$ loss on depth vs. an $L1$ loss on normals. Results with the depth-only loss appear unusable, while results with the normals-only loss are smooth, robust, and capture an impressive amount of detail. Thus, for detailed depth estimation of human bodies, a direct loss on depth is insufficient, whereas a loss on surface normals is sufficient to produce robust and detailed depth estimates. However, since the loss

Subject ID	FAX (mm)	FAX (mm) (opt pose)	[3]
50002	9.46	6.56	5.13
50004	7.90	4.19	4.36
50009	5.23	3.86	3.72
50020	6.60	3.85	3.32
50021	4.76	3.27	4.45
50022	5.08	3.50	5.71
50025	5.03	3.02	4.84
50026	7.83	4.87	4.56
50027	8.21	4.34	3.89

Table 1: Bi-directional mesh-to-mesh error on subjects from D-FAUST dataset using our baseline method. For each subject we report average error across multiple instances rendered with random environment configurations, using the methodology described in Section 4.2.

on normals only constrains the output locally, the geometry will not be true to scale. A loss on depth, while not crucial for the quality of the geometry, encourages the output toward a space of plausible human scales.

One advantage of FAX is its ability to extract subtle shape detail from a single image. Recovered shapes are intricate and personal, as observed in the waist, hips and chest of almost every example in Figure 7. This is hard to achieve by methods based on convex-hull [31], voxels [44] or SMPL shape parameters [22]. Even methods optimizing explicitly the shape to fit the image contour, like [3], fail to recover this level of detail because the underlying optimization has to find a compromise between the data and the underlying (overly smooth) model. Detail obtained from FAX is mostly visible in the contours, but the side renders show that this detail is reconstructed in a coherent manner across the body shape, recreating bust and stomach shape that is coherent with the silhouette and image shading.

Visual discontinuities such as shadows and tattoos are a challenge. Classic shape-from-shading methods are notorious for introducing ridge artifacts at misleading visual boundaries. As shown in Figure 7 (row 3, on right), our methods produce clean geometry in the presence of tattoos. And in Figure 6, our method exhibits invariance to sharp shadows. We credit this invariance almost entirely to the diversity in our training dataset; before introducing sharp shadows in our training (Figure 5: row 3 on left), ridge artifacts around shadows were common in our test output.

Spatial scan holes are an additional challenge. Like many high-quality scanner setups, our raw estimated scan does not capture all geometry, noticeably visible as the seam between front and rear-facing depth maps. This problem is one motivation for fitting an avatar: beyond providing repositability, it provides hole closure and scan completion. Figures 1 and 7 illustrates our scans, their seams, and the avatars that provide hole closure.

Our front albedo estimation network is resilient to soft shadows. To see this, consider the RGB input and frontal textured scan in Figure 7, which is illuminated with the same light as the grey scans. In particular, observe the removal of skin highlights in row 4 right, and much more even skin tone in legs and torso in most of them, e.g. row 7 right. Removing sharp casted shadows is extremely challenging, but reasonable results are achieved in row 1, 2 and 5 right.

Our back albedo estimator exhibits pleasing front/back consistency, including skin tone and garment continuity. Some bra straps (e.g. row 7 left in 7) show a continuous but physically implausible configuration, while garments in skin-tone colors (row 3 left in 7) blend into the skin texture. Improvements to training data should address this.

4.4. Quantitative evaluation on Dynamic FAUST

We compare our system quantitatively with [3], which is one of the state of the art systems in estimating shape from multiple images. Following [3], we generate synthetic renders from the subjects in Dynamic FAUST, estimate their shape, and evaluate it against the synthetic data. Unlike [3], we only require one image for each subject. We should also note that since our system works with RGB images, the authors of [8] kindly provided us with one natural texture for each subject in their dataset.

We follow the procedure described in [3] to compute the errors in Table 1. First, we estimate the scan and alignment as described in Sections 3.2 and 3.3. Using SMPL, we unpose the alignment and scale it to make it as tall as the groundtruth shape. Using this fixed shape, we optimize translation and scale to minimize the average bidirectional distance between vertices in each mesh and the surface of the other mesh, initializing the translation and pose from groundtruth. We repeat this procedure over N synthetic images per subject to obtain more reliable estimations of the error. This average bidirectional distance is reported in the left column from Table 1. This procedure is comparable to the *full method* reported in [3]. Our errors are larger than in [3], which can be attributed to two factors. First, we have access to a single image while [3] used hundreds of them. Second, applying the groundtruth pose from the scan can be suboptimal, since SMPL conflates pose and shape to some extent. To decouple this problem, we also optimized the pose together with scale and translation (keeping shape fixed at all times), which is shown in the middle column of Table 1. Note however that we believe this result is not directly comparable to [3].

4.5. Ablation Study

Here we study factors that contribute to our method performance. We first consider the individual contribution of our loss terms. We next vary the number of residual blocks in the network, which affects network depth. Similarly,

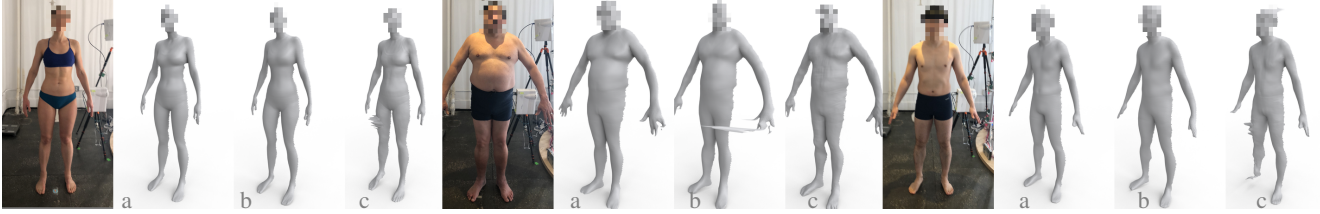


Figure 6: Visual comparison of the ablation experiments (a) Baseline, (b) without $\mathcal{L}_{L1}(y, G_n^i)$ term, and (c) 2 scales.

Label	\mathcal{L}_n^i	\mathcal{L}_d^i	Blur aug.	# Res. blocks	# scales	Error	Error (opt back)	Error (opt pose)	Error (opt back, pose)
Baseline	✓	✓	✓	9	4	6.89	6.66	3.77	3.65
5 res blocks	✓	✓	✓	5	4	6.76	6.63	3.62	3.60
No blur aug.	✓	✓		9	4	6.99	6.97	3.83	3.85
2 scales	✓	✓	✓	9	2	8.21	7.88	4.50	4.34
No depth	✓		✓	9	4	-	8.57	-	3.87
No normals		✓	✓	9	4	9.02	9.04	5.28	5.36
No VGG	✓	✓	✓	9	4	7.80	6.69	4.18	3.60

Table 2: Ablation study on our depth estimator, using mesh distance for evaluation. See Section 4.5 for more details.

we change how many downsampling operations (*scales*) are performed. These operations involve learned convolutions, and thus add capacity and depth to the network. Finally, we test the role of blur data augmentation performed on our synthetic training data. We run this experiment on images from 87 subjects (see Figure 6 for four subject examples).

Results of the ablation study are summarized in Table 2. For compatibility with [3], we perform all comparisons with estimated alignments instead of scans, using the procedure described in Section 4.4, reporting average bi-directional point-to-mesh distances. However, fitting a model to our scan regularizes problems in less robust variants of our pipeline (e.g., “No blur aug.”) and the imperfections in the unposing process may introduce subtle and potentially misleading inaccuracies, thus the tradeoffs in model variants will not necessarily be well represented by this metric.

Columns labeled with *opt pose* relate to pose optimized to minimize distance, similar to the previous section. We also consider the independent optimization of front and back scale (as described in Section 3.3, labeled as *opt back*), since experiments with no depth show differences in scale in the front and back that render quantitative evaluation useless without such independent optimization.

Most noticeable is the importance of normals in this loss. Removing normal terms (both L1 and VGG) is more detrimental than removing the depth term, which is consistent with the intuition provided in Figure 2. Removing depth or normal terms incurs a negative effect compared with the baseline. Reducing downsampling makes the network shallower, allowing it to keep more detail (see Table 2) but also noise, incurring a big accuracy penalty. Although blur aug-

mentation has a small numerical impact, we observe that it creates spikes and holes, making it unusable for the rapid creation of a textured scan. Lastly, omitting the VGG loss on normals causes a minor loss in accuracy.

We add an extra configuration in Figure 6: removing the L1 loss on normals but keeping VGG results in an over-smoothed scan with more shading artifacts. Finally, while it’s surprising that reducing the number of residual blocks improves accuracy, we consider the difference negligible.

5. Conclusions

FAX estimates full body geometry and albedo from a single RGB image at a level of detail previously unseen. This quality depends critically on two main factors. First, we do not indirect our output through representations like voxels, convex hulls or body models, which allow us to recover detail at the original pixel definition with an image-translation network, orders of magnitude faster than competing methods. Second, our geometry estimation depends critically on the role of surface normals, and we show how even surface normals alone can produce plausible bodies in the absence of depth information. We evaluate our system using two datasets, perform an ablation study, and extensively illustrate the visual performance of our system.

For future work, we believe improving our training data can overcome many restrictions of the current method, like the frontal pose or minimal clothing. We would like to eliminate the seams in scan geometry and texture in a rapid, data-driven manner. Finally, we believe incorporating an additional view can help reduce the inherent ambiguity present in the shapes estimated from a single view.



Figure 7: Two columns with RGB image, scan with and without texture and alignment. Pay close attention to variation in shape, pose and ethnicity, as well as the fidelity of detail in hips, waist and chest, specially in the silhouette region. Note that most test subjects in this figure are wearing similar clothes to the garments present in the synthetic training data.

References

- [1] Thiemo Alldieck, Marcus Magnor, Bharat Lal Bhatnagar, Christian Theobalt, and Gerard Pons-Moll. Learning to reconstruct people in clothing from a single RGB camera. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, jun 2019. 2
- [2] Thiemo Alldieck, Marcus Magnor, Weipeng Xu, Christian Theobalt, and Gerard Pons-Moll. Detailed human avatars from monocular video. In *2018 International Conference on 3D Vision (3DV)*, pages 98–109. IEEE, 2018. 1, 2
- [3] Thiemo Alldieck, Marcus Magnor, Weipeng Xu, Christian Theobalt, and Gerard Pons-Moll. Video based reconstruction of 3d people models. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018. 2, 6, 7
- [4] Jonathan T. Barron and Jitendra Malik. Shape, illumination, and reflectance from shading. *IEEE Trans. Pattern Anal. Mach. Intell.*, 37(8):1670–1687, 2015. 2
- [5] Volker Blanz and Thomas Vetter. A morphable model for the synthesis of 3d faces. In *Proceedings of the 26th Annual Conference on Computer Graphics and Interactive Techniques*, SIGGRAPH '99, pages 187–194, New York, NY, USA, 1999. ACM Press/Addison-Wesley Publishing Co. 2
- [6] Federica Bogo, Michael J. Black, Matthew Loper, and Javier Romero. Detailed full-body reconstructions of moving people from monocular rgb-d sequences. In *Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV)*, ICCV '15, pages 2300–2308, Washington, DC, USA, 2015. IEEE Computer Society. 1, 4
- [7] Federica Bogo, Angjoo Kanazawa, Christoph Lassner, Peter Gehler, Javier Romero, and Michael J Black. Keep it smpl: Automatic estimation of 3d human pose and shape from a single image. In *European Conference on Computer Vision*, pages 561–578. Springer, 2016. 1, 2, 4
- [8] Federica Bogo, Javier Romero, Gerard Pons-Moll, and Michael J. Black. Dynamic FAUST: registering human bodies in motion. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 5573–5582, 2017. 1, 2, 6
- [9] James Booth, Epameinondas Antonakos, Stylianos Ploumpis, George Trigeorgis, Yannis Panagakis, and Stefanos Zafeiriou. 3d face morphable models "in-the-wild". In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 00, pages 5464–5473, July 2017. 2
- [10] Endri Dibra, Himanshu Jain, A. Cengiz Öztireli, Remo Ziegler, and Markus H. Gross. Human shape from silhouettes using generative hks descriptors and cross-modal neural networks. In *CVPR*, pages 5504–5514. IEEE Computer Society, 2017. 1, 2
- [11] Pengfei Dou, Shishir K. Shah, and Ioannis A. Kakadiaris. End-to-end 3d face reconstruction with deep neural networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1503–1512, July 2017. 2
- [12] David Eigen and Rob Fergus. Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture. *CoRR*, abs/1411.4734, 2014. 2
- [13] Huan Fu, Mingming Gong, Chaohui Wang, Kayhan Batmanghelich, and Dacheng Tao. Deep ordinal regression network for monocular depth estimation. *CoRR*, abs/1806.02446, 2018. 2
- [14] Yu Guo, Xiaowu Chen, Bin Zhou, and Qinqing Zhao. Clothed and naked human shapes estimation from a single image. In *CVM*, 2012. 2
- [15] Nils Hasler, Hanno Ackermann, Bodo Rosenhahn, Thorsten Thormählen, and Hans-Peter Seidel. Multilinear pose and body shape estimation of dressed subjects from image sets. In *The Twenty-Third IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2010, San Francisco, CA, USA, 13-18 June 2010*, pages 1823–1830, 2010. 2
- [16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 3
- [17] Berthold K. P. Horn. Shape-from-shading: A method for obtaining the shape of a smooth opaque object from one view. Technical Report MAC-TR-79 and AI-TR-232, AI Laboratory, MIT, Nov. 1970. 2
- [18] Berthold K. P. Horn and Michael J. Brooks. The variational approach to shape from shading. *Computer Vision, Graphics, and Image Processing*, 33(2):174–208, 1986. 1
- [19] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. *arXiv preprint*, 2017. 1, 2, 3
- [20] Aaron S. Jackson, Adrian Bulat, Vasileios Argyriou, and Georgios Tzimiropoulos. Large pose 3d face reconstruction from a single image via direct volumetric cnn regression. In *The IEEE International Conference on Computer Vision (ICCV)*, Oct 2017. 2
- [21] Arjun Jain, Thorsten Thormählen, Hans-Peter Seidel, and Christian Theobalt. Moviereshape: Tracking and reshaping of humans in videos. *ACM Trans. Graph.*, 29(6):148:1–148:10, Dec. 2010. 2
- [22] Angjoo Kanazawa, Michael J Black, David W Jacobs, and Jitendra Malik. End-to-end recovery of human shape and pose. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 1, 2, 6
- [23] Vahid Kazemi and Josephine Sullivan. One millisecond face alignment with an ensemble of regression trees. In *CVPR*, pages 1867–1874. IEEE Computer Society, 2014. 4
- [24] Ira Kemelmacher-Shlizerman and Ronen Basri. 3d face reconstruction from a single image using a single reference face shape. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(2):394–405, Feb 2011. 2
- [25] Ivan Krasin, Tom Duerig, Neil Alldrin, Andreas Veit, Sami Abu-El-Haija, Serge Belongie, David Cai, Zheyun Feng, Vittorio Ferrari, Victor Gomes, et al. Openimages: A public dataset for large-scale multi-label and multi-class image classification. *Dataset available from <https://github.com/openimages>*, 2(6):7, 2016. 5
- [26] Christoph Lassner, Javier Romero, Martin Kiefel, Federica Bogo, Michael J. Black, and Peter V. Gehler. Unite the people: Closing the loop between 3d and 2d human representations. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, July 2017. 2

- [27] Chen Li, Kun Zhou, and Stephen Lin. Intrinsic face image decomposition with human face priors. In David J. Fleet, Tomás Pajdla, Bernt Schiele, and Tinne Tuytelaars, editors, *Computer Vision - ECCV 2014 - 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V*, volume 8693 of *Lecture Notes in Computer Science*, pages 218–233. Springer, 2014. 2
- [28] Guannan Li, Chenglei Wu, Carsten Stoll, Yebin Liu, Kiran Varanasi, Qionghai Dai, and Christian Theobalt. Capturing reliable human performances under general uncontrolled illumination. *Comput. Graph. Forum*, 32(2):275–284, 2013. 1
- [29] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. Smpl: A skinned multi-person linear model. *ACM Transactions on Graphics (TOG)*, 34(6):248, 2015. 4
- [30] Xudong Mao, Qing Li, Haoran Xie, Raymond YK Lau, Zhen Wang, and Stephen Paul Smolley. Least squares generative adversarial networks. In *Computer Vision (ICCV), 2017 IEEE International Conference on*, pages 2813–2821. IEEE, 2017. 3
- [31] Ryota Natsume, Shunsuke Saito, Zeng Huang, Weikai Chen, Chongyang Ma, Hao Li, and Shigeo Morishima. Siclope: Silhouette-based clothed people. *CoRR*, abs/1901.00049, 2019. 1, 2, 6
- [32] Augustus Odena, Vincent Dumoulin, and Chris Olah. Deconvolution and checkerboard artifacts. *Distill*, 2016. 3
- [33] Mohamed Omran, Christoph Lassner, Gerard Pons-Moll, Peter Gehler, and Bernt Schiele. Neural body fitting: Unifying deep learning and model based human pose and shape estimation. In *2018 International Conference on 3D Vision (3DV)*, pages 484–494. IEEE, 2018. 1, 2
- [34] Georgios Pavlakos, Luyang Zhu, Xiaowei Zhou, and Kostas Daniilidis. Learning to estimate 3d human pose and shape from a single color image. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 459–468, 2018. 2
- [35] Elad Richardson, Matan Sela, Roy Or-El, and Ron Kimmel. Learning detailed face reconstruction from a single image. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 5553–5562, 2017. 1, 2
- [36] Matan Sela, Elad Richardson, and Ron Kimmel. Unrestricted facial geometry reconstruction using image-to-image translation. In *Computer Vision (ICCV), 2017 IEEE International Conference on*, pages 1585–1594. IEEE, 2017. 2, 4
- [37] Soumyadip Sengupta, Angjoo Kanazawa, Carlos D. Castillo, and David W. Jacobs. Sfsnet: Learning shape, reflectance and illuminance of faces in the wild. In *Computer Vision and Pattern Recognition (CVPR)*, 2018. 2
- [38] Vince Tan, Ignas Budvytis, and Roberto Cipolla. Indirect deep structured learning for 3d human body shape and pose prediction. In *British Machine Vision Conference 2017, BMVC 2017, London, UK, September 4-7, 2017*, 2017. 2
- [39] Ayush Tewari, Michael Zollhöfer, Pablo Garrido, Florian Bernard, Hyeonwoo Kim, Patrick Pérez, and Christian Theobalt. Self-supervised multi-level face model learning for monocular reconstruction at over 250 Hz. In *Proceedings of Computer Vision and Pattern Recognition (CVPR 2018)*, 2018. 2
- [40] Ayush Tewari, Michael Zollhöfer, Hyeonwoo Kim, Pablo Garrido, Florian Bernard, Patrick Pérez, and Theobalt Christian. MoFA: Model-based Deep Convolutional Face Autoencoder for Unsupervised Monocular Reconstruction. In *The IEEE International Conference on Computer Vision (ICCV)*, 2017. 2
- [41] Anh Tuấn Tran, Tal Hassner, Iacopo Masi, and Gérard G. Medioni. Regressing robust and discriminative 3d morphable models with a very deep neural network. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 1493–1502, 2017. 2
- [42] Anh Tuấn Tran, Tal Hassner, Iacopo Masi, Eran Paz, Yuval Nirkin, and Gérard Medioni. Extreme 3d face reconstruction: Seeing through occlusions. In *Proc. CVPR*, 2018. 2
- [43] Shubham Tulsiani, Saurabh Gupta, David F. Fouhey, Alexei A. Efros, and Jitendra Malik. Factoring shape, pose, and layout from the 2d image of a 3d scene. *CoRR*, abs/1712.01812, 2017. 2
- [44] Gül Varol, Duygu Ceylan, Bryan Russell, Jimei Yang, Ersin Yumer, Ivan Laptev, and Cordelia Schmid. Bodynet: Volumetric inference of 3d human body shapes. In *Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part VII*, pages 20–38, 2018. 1, 2, 6
- [45] Gül Varol, Javier Romero, Xavier Martin, Naureen Mahmood, Michael J Black, Ivan Laptev, and Cordelia Schmid. Learning from synthetic humans. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2017)*, pages 4627–4635. IEEE, 2017. 2
- [46] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. High-resolution image synthesis and semantic manipulation with conditional gans. *arXiv preprint arXiv:1711.11585*, 2017. 1, 3, 4, 5
- [47] Qing Zhang, Bo Fu, Mao Ye, and Ruigang Yang. Quality dynamic human body modeling using a single low-cost depth camera. In *CVPR*, pages 676–683. IEEE Computer Society, 2014. 1
- [48] Ruo Zhang, Ping-Sing Tsai, James Edwin Cryer, and Mubarak Shah. Shape from shading: A survey. *IEEE Trans. Pattern Anal. Mach. Intell.*, 21(8):690–706, 1999. 2
- [49] Michael Zollhöfer, Justus Thies, Pablo Garrido, Thabo Bradley, Derek Beeler, Patrick Pérez, Marc Stamminger, Matthias Nießner, and Christian Theobalt. State of the art on monocular 3d face reconstruction, tracking, and applications. *Computer Graphics Forum*, 37(2):523–550, 2018. 2