Human Shape from Silhouettes using Generative HKS Descriptors and Cross-Modal Neural Networks

Endri Dibra[1], Himanshu Jain[1], Cengiz Öztireli[1], Remo Ziegler[2], Markus Gross[1]
[1]Department of Computer Science, ETH Zürich, [2]Vizrt
{edibra,cengizo,grossm}@inf.ethz.ch, jainh@student.ethz.ch, rziegler@vizrt.com

Abstract

In this work, we present a novel method for capturing human body shape from a single scaled silhouette. We combine deep correlated features capturing different 2D views, and embedding spaces based on 3D cues in a novel convolutional neural network (CNN) based architecture. We first train a CNN to find a richer body shape representation space from pose invariant 3D human shape descriptors. Then, we learn a mapping from silhouettes to this representation space, with the help of a novel architecture that exploits correlation of multi-view data during training time, to improve prediction at test time. We extensively validate our results on synthetic and real data, demonstrating significant improvements in accuracy as compared to the state-of-the-art, and providing a practical system for detailed human body measurements from a single image.

1. Introduction

Human body shape estimation has recently received a lot of interest. This partially relates to the growth in demand of applications such as tele-presence, virtual and augmented reality, virtual try-on, and body health monitoring. For such applications, having an accurate and practical system that estimates the 3D human body shape is of crucial importance. It needs to be accurate such that automated body measurements agree with the real ones, and needs to be practical such that it is fast and utilizes as few sensors as possible. With respect to the sensors utilized, in increasing order of simplicity, we can distinguish multiple cameras [10, 46], RGB and Depth [25] or a single image [20, 67, 29, 23, 5, 17].

In this work we tackle the problem of shape estimation from a single or multiple silhouettes of a human body with poses compliant with two main applications: virtual garment fitting assuming a neutral pose [13, 6, 16], and shape from individually taken pictures or "Selfies" (e.g. through a mirror or a long selfie stick), assuming poses that exhibit mild self occlusion [17]. Compared to state-of-the-art in this domain, we achieve significantly higher accuracy on the reconstructed body shapes and simultaneously improve in speed if a GPU implementation is considered (or obtain similar run-times as previous works [17] on the CPU). This is achieved thanks to a novel Neural Network architecture (Fig.1) consisting of various components that (a) are able to learn a body shape representation from 3D shape descriptors and map this representation to 3D shapes, (b) can successfully reconstruct a 3D body mesh from one or two given body silhouettes, and (c) can leverage multi-view data at training time, to boost predictions for a single view at test time through cross-modality learning.

Previous methods attempt to find a mapping from silhouettes to the parameters of a statistical body shape model [2], utilizing handcrafted features [17], silhouette PCA representations [13] possibly with local fine tuning [6], or CNNs [16]. Based on the obtained parameters, a least squares system is solved to obtain the final mesh. We also use CNNs to learn silhouette features, but unlike [16], we first map them to a shape representation space that is generated from 3D shape descriptors (Heat Kernel Signature (HKS) [57]) invariant to isometric deformations and maximizing intra-human-class variation, and then decode them to full body vertex positions. Regressing to this space improves the predictions and speeds up the computation.

Recently, Dibra et al. [17] demonstrated how to boost features coming from one view (scaled frontal) during test time, utilizing information from two views (front and side) at training time, by projecting features with Canonical Correlation Analysis (CCA) [26] for a regression task. CCA comes with shortcomings though as (1) it computes a linear projection, (2) it is hard in practice to extend it to more than two views, and (3) suffers from lack of scalability to large datasets as it has to "memorize" the whole training data set. As part of this work, we propose an architecture (which we call Cross-Modal Neural Network (CMNN)) that is able to overcome the mentioned challenges, by first generating features from various views separately, and then combining them through shared layers. This leads to improvements in

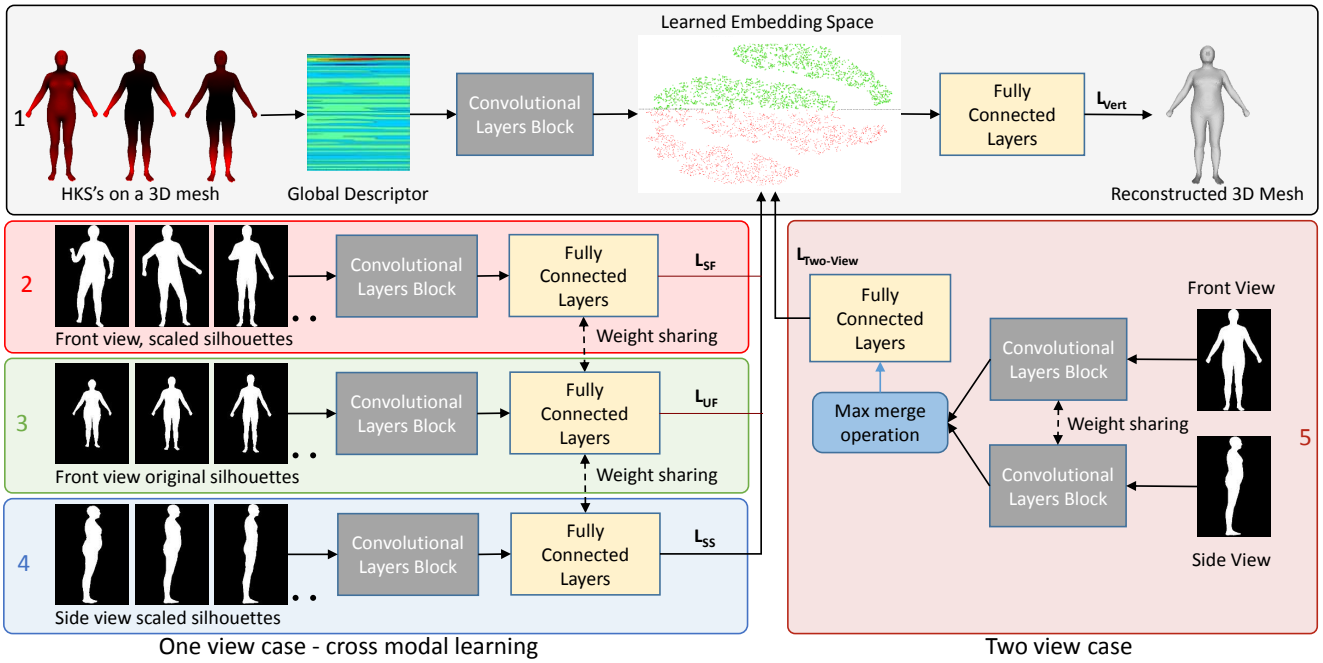

Figure 1. Our body shape estimation method. (1) HKS-Net: HKS projected features as input, generates an embedding space which is mapped to 3D meshes. (2),(3) and (4) Three modes of the Cross-Modal Neural Network (CMNN) (only (2) is used at test time). (5) An architecture that requires both views at test time. The method uses either CMNN or (5), depending on the number of available input views.

predictive capabilities with respect to the uni-modal case. Abstracting away from silhouettes, this network can be used as-is for other tasks where multiple views on the data are present, such as image and text retrieval, or audio and image matching.

In summary, the contributions of this paper are: (1) A novel neural network architecture for 3D body shape estimation from silhouettes consisting of three main components, (a) a generative component that can invert a pose-invariant 3D shape descriptor to reconstruct its neutral shape, (b) a predictive component that combines 2D and 3D cues to map silhouettes to human body shapes, (c) a cross-modal component that leverages multi-view information to boost single view predictions; and (2) a state-of-the-art system for human body shape estimation that significantly improves accuracy as compared to existing methods.

2. Related Work

Human Body Shape from an Image. Early works in estimating 3D body shapes make assumptions on the number of views [33] or simple geometric models [30, 41] often achieving coarse approximations of the underlying geometry. As scanning of a multitude of people in various poses and shapes was made possible [47], more complete, parametric human body shape models were learned [2, 24, 42, 36] that capture deformations due to shape and pose. The effectiveness of such models with human priors, gave rise to methods that try to estimate the human body shape from single [20, 67, 29, 23, 12, 46] or multiple input images [4, 10, 23, 46], by estimating the parameters of the model, through matching projected silhouettes of the 3D shapes to extracted image silhouettes by correspondence. Assumptions on the view, calibration, error metrics [10, 20, 29] and especially speed and manual interaction, needed to estimate pose and shape by silhouette matching, in the presence of occlusions and challenging poses [67, 29, 12], are common limitations of these methods, despite promising work to automatize the matching process [52, 53, 32]. A very recent work by Bogo et al. [5] attempts at estimating both the 3D pose and shape from a single 2D image with given 2D joints, making use of a 3D shape model based on skinning weights [36]. It utilizes a human body prior as a regularizer, for uncommon limb lengths or body interpenetrations, achieving excellent results on 3D pose estimations, however, lacking accuracy analysis on the generated body shapes.

While the abovementioned works tackle the shape estimation problem by iteratively minimizing an energy function, another body of works estimate the 3D body shape by first constructing statistical models of 2D silhouette features and 3D bodies, and then defining a mapping between the parameters of each model [62, 55, 13, 15, 14, 6, 17]. In terms of silhouette representation they vary from PCA learned silhouette descriptors [13, 6] to handcrafted features such as the Radial Distance Functions and Shape Contexts [55] or

the Weighted Normal Depth and Curvature [17]. The statistical 3D body model is learned by applying PCA on triangle deformations from an average human body shape [2]. With respect to the body parameter estimations, Xi et al. [62] utilize a linear mapping, Sigal et al. [55] a mixture of kernel regressors, Chen et al. [13] a shared Gaussian process latent variable model, Dibra et al. [17] a combination of projections at Correlated Spaces and Random Forest Regressors and Boisvert et al. [6] an initial mapping with the method from [13] which is further refined by an optimization procedure with local fitting. The mentioned methods target applications similar to ours, however except for [17], they are lacking practicality for interactive applications due to their running times, and have been evaluated under more restrictive assumptions with respect to the camera calibration, poses, and amount of views required. Under similar settings, a more recent work [16] attempts at finding a mapping from the image directly, by training an end-to-end Convolutional Neural Network to regress to body shape parameters.

In contrast to these methods, we first learn an embedding space from 3D shape descriptors, that are invariant to isometric deformations, by training a CNN to regress directly to 3D body shape vertices. Then we learn a mapping from 2D silhouette images to this new embedding space. We demonstrate improved performance over the previous methods [16, 6] working under restrictive assumptions (two views and known camera calibration) with this set-up. Finally, by incorporating cross-modality learning from multiple views, we also outperform Dibra et al. [17] under a more general setting (one view and unknown camera calibration).

CNN-s on 3D shapes. The improvement in accuracy and performance by utilizing Convolutional Neural Networks for 2D image related tasks is widely aknowledged in the community by now. Once one goes to 3D, one of the main paradigms utilized is to represent the data as a low resolution voxelized grid [61, 56, 48]. This representation has been mainly utilized for shape classification and retrieval tasks [61, 56, 51] or to find a mapping from 2D view representations of those shapes [48], and has been geared towards rigid objects (like chairs, tables, cars etc.). Another possibility to represent the 3D shape, stemming more from the Computer Graphics community is that of 3D Shape Descriptors, which have been extensively studied for shape matching and retrieval [28, 58, 59].

Various shape descriptors have been proposed, with most recent approaches being diffusion based methods [57, 9, 49]. Based on the Laplace-Beltrami operator that can robustly characterize the points on a meshed surface, some of the proposed descriptors are the global point signature (GPS) [49], the heat kernel signature (HKS) [57] and the Wave Kernel Signature (WKS) [3]. Further works build on these and related descriptors and learn better descriptors, mainly through CNN-s that are utilized in shape retrieval, classification and especially shape matching [44, 7, 8, 38, 39, 60, 63, 35, 18]. Their main objective is either to maximize the inter class variance or to design features that find intra-class similarities. We, on the other hand, want to find suitable descriptors that maximize intra-class variance (here human body shapes), and learn a mapping by regression to 3D body shapes, which to the best of our knowledge has not been explored.

Due to the properties of the HKS, such as invariance to isometric deformations and insensitivity to small perturbations on the surface, which are very desirable in order to consistently explain the same human body shape under varying non-rigid deformations, we start from the HKS and encode it into a new shape embedding space, from which we can decode the full body mesh or to which we can regress possible views of the bodies. In this way, our method can be thought of as a generative technique that learns an inverse mapping, from the descriptor space to the shape space.

Cross-Modality Learning. In the presence of multiple views or modalities representing the same data, unsupervised learning techniques have been proposed that leverage such modalities during training, to learn better representations that can be useful when one of them is missing at test time. There exist a couple of applications that rely on learning common representations, including 1) transfer learning, 2) reconstruction of a missing view, 3) matching accross views, and directly related to our work 4) boosting single view performance utilizing data from other views or otherwise called cross-modality learning.

Early works, like Canonical Correlation Analysis (CCA) [26] and it's kernelized version [22] find maximally correlated linear and non-linear projections of two random vectors with the intention of maximizing mutual information and minimizing individual noise. Fusing learned features for better prediction [50], hallucinating multiple modalities from a single view [40] as well as a generalized version of CCA [54] for a classification and retrieval task, have been proposed. Except for a few works [17, 40], utilizing cross-modality learning to improve regression has had little attention. To tackle the inability of CCA to scale well to large datasets, there have been recent attempts that utilize neural networks like Deep CCA [1] and its GPU counterpart [64], Multimodal Autoencoders [43] and Correlational Neural Networks [11] but these methods do not focus on boosting single view predictions.

Unlike these techniques, we present a way to perform cross-modality learning by first learning representative features through CNN-s, and then passing them through shared encoding layers, with the objective of regressing to the embedding space. We demonstrate significant increase in performance over uni-modal predictions, and scalability to higher dimensional large scale data.

3. The Generative and Cross-Modal Estimator

The main goal of our method is to accurately estimate a 3D body shape from a silhouette (or two) of a person adopting poses in compliance with two applications - virtual cloth fitting and self shape monitoring. On par with the related work, we consider either a single frontal silhouette scaled to a fixed size (no camera calibration information) with poses exhibiting mild self occlusions, or two views simultaneously (front and side, scaled or unscaled) of a person in a neutral pose. We propose to tackle this problem with a deep network architecture (Fig.1). Our network is composed of three core components: a generative component that can invert pose-invariant 3D shape descriptors, obtained from a multitude of 3D meshes (Sec.3.1) to their corresponding 3D shape, by learning an embedding space (Sec.3.2); a cross-modal component that leverages multi-view information at training time to boost single view predictions at test time (Sec.3.3); and a combination of losses to perform joint training over the whole network (Sec.3.4).

3.1. Shape Model and Data Generation

In order to properly train our network, we recur to synthetic data as in the previous works since they best approximate our real input requirements. We need to obtain a large number of meshes from which we can extract 3D descriptors and 2D silhouettes in various poses. We make use of existing datasets [66, 45] consisting of meshes fitted to the commercially available CAESAR [47] dataset that contains 3D human body scans. Starting from these datasets, we can generate hundreds of thousands of human body meshes by learning a statistical model. The methods we compare to [62, 13, 6, 16, 17] utilize a low-dimensional parametric human model (SCAPE [2]) that is based on triangle deformations learned from 3D range scans of people in various shapes and poses. Despite more recent body models [42, 36], for fair comparisons and evaluation, we also utilize SCAPE, which is defined as a set of 12894 triangle deformations applied to a reference template 3D mesh consisting of 6449 vertices, with parameters α and β representing pose and intrinsic body shape deformations, respectively. From these parameters, each edge $\mathbf{e}_{i1}$ and $\mathbf{e}_{i2}$ of the i^{th} triangle of the template mesh, defined as the difference vectors between the vertices of the triangle, can be transformed as

$$\mathbf{e}_{ij}^{'} = \mathbf{R}_i(\alpha)\mathbf{S}_i(\beta)\mathbf{Q}_i(\mathbf{R}_i(\alpha))\mathbf{e}_{ij}, \tag{1}$$

with $j \in \{1, 2\}$. Matrices $\mathbf{R}_i(\alpha)$, $\mathbf{Q}_i(\mathbf{R}_i(\alpha))$ and $\mathbf{S}_i(\beta)$ correspond to joint rotations, pose induced non-rigid deformations, and intrinsic shape variation, respectively. Similar to [16, 17], we learn a deformation space by applying PCA to the set of deformations for all meshes in the datasets, with respect to a template mesh, all in the same pose. To synthesize new meshes, we sample from a 20 dimensional multivariate normal distribution, given by the first 20 components obtained via PCA that capture 95% of the energy.

Under the common assumption that the intrinsic body shape does not change significantly due to pose changes [2], we decouple pose from shape deformations. Hence, for a neutral pose we have $\mathbf{e}_{ij}^{'} = \mathbf{S}_i(\beta)\mathbf{e}_{ij}$. To add pose variation to the mesh synthesis process, instead of the transformation $\mathbf{R}_i(\alpha)$ parametrized by $alpha$, we utilize Linear Blend Skinning (LBS) [34], as in previous works [29, 19, 65], which computes the new position of each restpose vertex $\mathbf{v}_1, ..., \mathbf{v}_n \in \mathbf{R}^4$ in homogenous coordinates, with a weighted combination of the bone transformation matrices $\mathbf{T}_1, ..., \mathbf{T}_m \in \mathbf{R}^{4\times4}$ of an embedded skeleton controlling the mesh, and skinning weights $w_{i,1}, ..., w_{i,m} \in \mathbf{R}$ for a vertex $\mathbf{v}_i$ and the m^{th} bone transformation, as follows:

$$\mathbf{v}_i^{'} = \sum_{j=1}^{m} w_{i,j}\mathbf{T}_j\mathbf{v}_i = \left(\sum_{j=1}^{m} w_{i,j}\mathbf{T}_j\right)\mathbf{v}_i. \tag{2}$$

Combining various intrinsic shapes and poses as generated above, we create a synthetic dataset consisting of half a million meshes, from which we extract HKS descriptors and silhouettes for training.

3.2. Generating 3D Shapes from HKS (HKS-Net)

The first part of our architecture aims at learning a mapping from 3D shape descriptors to 3D meshes via a shape embedding space. We start by extracting Heat Kernel Signatures (HKS) and then projecting them to the eigenvectors of the Laplace-Beltrami operator to obtain a global descriptor. This is used to learn the embedding space, as well as an inverse mapping that can generate 3D shapes in a neutral pose given the corresponding descriptor.

Heat Kernel Signatures (HKS). Let a 3D shape be represented as a graph $G = (V, E, W)$, where V, E and W represent the set of vertices, edges, and some weights on the edges, respectively. The weights encode the underlying geometry of the shape, and can be computed via standard techniques from the mesh processing literature [57]. Given such a graph constructed by connecting pairs of vertices on a surface with weighted edges, the heat kernel $H_t(x, y)$ is defined as the amount of heat that is transferred from the vertex x to vertex y at time t, given a unit heat source at x [57]:

$$H_t(x, y) = \sum_i e^{-\lambda_i t}\phi_i(x)\phi_i(y), \tag{3}$$

where H_t denotes the heat kernel, t is the diffusion time, λ_i and ϕ_i represent the i^{th} eigenvalue and the corresponding eigenvector of the Laplace-Beltrami operator, respectively, and x and y denote two vertices. Heat kernel has various nice properties that are desirable to represent human body

shapes under different poses. In particular, it is invariant under isometric deformations of the shape, captures different levels of detail and global properties of the shape, and it is stable under perturbations [57].

The heat kernel at vertex x and time t can be used to define the heat kernel signature $HKS_x(t)$ for this vertex:

$$HKS_x(t) = H_t(x,x) = \sum_i e^{-\lambda_i t}\phi_i^2(x). \qquad (4)$$

Hence, for each vertex x, we have a corresponding function $HKS_x(t)$ that provides a multi-scale descriptor for x. As the scale (i.e. t) increases, we capture more and more global properties of the intrinsic shape. In practice, the times t are sampled to obtain a vector $HKS_x(t_j), j \leq J$ for each vertex x. In our technique, we use $J = 100$ time samples. Then for each t_j, we can form the vectors $\mathbf{h}_j := [HKS_{x_1}(t_j), HKS_{x_2}(t_j)\cdots]^T$.

Projected HKS Matrix. To learn the embedding space, the HKS for all vertices at a given time t_j are projected onto the eigenvectors of the Laplace-Beltrami operator in order to obtain a 2D image capturing the global intrinsic shape. Specifically, we compute a matrix $\mathbf{M}$ with $\mathbf{M}_{ij} = \phi_i^T \mathbf{h}_j$, i.e. the dot product of the i^{th} eigenvector of the Laplace-Beltrami operator and the heat kernel vector defined over the vertices for time t_j. Since we use 300 eigenvectors ϕ_i, we thus get a 300×100 matrix $\mathbf{M}$.

This is then used as input to the top part of our network (that we call HKS-Net, Fig.1 (1)) to learn an embedding space of about 4000 dimensions, by minimizing the per-vertex squared norm loss L_{Vert}. A simplistic representation of this embedding, computed utilizing T-SNE [37], is also presented in Fig.1, where female meshes are depicted in green dots and male meshes in red. An important property of HKS-Net is that we can reconstruct a 3D mesh in a neutral pose when HKS-Net is presented with a computed $\mathbf{M}$. Hence, HKS-Net can invert the HKS descriptors. Although we do not utilize this property in the scope of this work, we believe that this could be a valuable tool for geometry processing applications. But instead, we use the embedding space with 4000 dimensions as the target space for the cross-modal silhouette-based training of our network, which we explain next.

3.3. Cross-Modal Neural Network (CMNN)

The second component thus consists of finding a mapping from silhouettes to the newly learned embedding space. We generate five types of silhouettes that can be referred to as *modes* : frontal view scaled in various poses with minor self occlusion, frontal view scaled in a neutral pose, side view scaled in a neutral pose and front and side view unscaled in a neutral pose (Fig.1).[1] Here, unscaled implies known camera calibration, and scaled means we resize the silhouettes such that they have the same height. Frontal means that the plane of the bones that form the torso is parallel to the camera plane, and side is a 90 degrees rotated version of the frontal view. At test time, our results are not affected by slight deviations from these views. We thus center the silhouettes, and resize them to an image of resolution 264×192 before inputting them to the CMNN. We, of course, do not expect to use all the modes/views at once during testing, but our intention is to leverage the vast amount of data from various modes at training time for robust predictions at test time.

We start by training a network similar in size to the previous works [16] (5 convolutional and 3 dense layers), with AdaMax optimizer [31], and learning rate of e^{-4}, to map each mode individually to the embedding space by minimizing squared losses on the 4000 embedding space parameters (Fig.1 (2),(3) and (4) with the respective losses L_{SF}, L_{UF} and L_{SS}). As shown in Tab.2, we already achieve better results for the one-view case as compared to related works. This pre-training serves as an initialization for the convolutional weights of the Cross-Modal Neural Network (CMNN). The final cross-modal training is performed by starting from the weights given by the pre-training, and optimizing for the shared weights for the fully connected layers with a combined loss, e.g. for scaled-front and scaled-side we minimize $L_{SF} + L_{SS}$, or for three modes, the loss is $L_{SF} + L_{UF} + L_{SS}$.

The idea is to let each single convolutional network compute silhouette features separately first, and then correlate these high-level features at later stages. We observed that we obtain significant improvements when cross-correlating various combinations of 2 modes and 3 modes during training (Tab.2) as compared to the uni-modal results. CMNN offers several advantages as compared to CCA. First, we obtain a non-linear correlation between high-level features. Second, we can add as many modes as we want, while it is not trivial to correlate more than two spaces with CCA. Finally, we do not need to store all training data in memory as in the case of CCA.

One of the main focuses of this paper is estimating a 3D shape for the scaled-frontal case, with similar application scenarios as in the previous works [17]. Hence, our desired test time mode, i.e. the desired input at test time, is a silhouette from a frontal view with unknown camera parameters. Without loss of generality, we consider the unscaled-frontal and scaled-side as the other additional modes. Note that this can be extended with more views and further variations.

3.4. Joint Training

Finally, we would like to jointly train HKS-Net and CMNN for obtaining the final generative network. This is done by using all losses at the same time and back-

[1] Please note that throughout the text *mode* and *view* are used interchangeably to emphasize different ways of representing the same 3D mesh.

Table 1. Nomenclature for the various experiments. For the architecture components highlighted in colors and with numbers, please refer to Fig.1.

Name	Training Input	Test Input	Architecture
SF-1	Scaled Frontal View (SFV), Neutral Pose	SFV	2
SF-1-P	SFV, Various Poses	SFV	2
SFU-1	SFV, Unscaled Frontal View (UFV)	SFV	2 3
SFS-1	SFV, Scaled Side View (SSV)	SFV	2 4
SFUS-1	SFV, UFV, SSV	SFV	2 3 4
SFUS-HKS-1	SFV, UFV, SSV, projected HKS (PHKS)	SFV	1 2 3 4
SF-SS-2	SFV, SSV	SFV, SSV	5
UF-US-2	UFV, Unscaled Side View (USV)	UFV, USV	5
UF-US-HKS-2	UFV, USV, PHKS	UFV, USV	1 5

propagating them to all parts of the architecture. We thus perform a joint training with the HKS-Net by minimizing $L_{SF} + L_{UF} + L_{SS} + L_{Vert}$. This training not only improves the mappings from 2D silhouettes to the 3D meshes, but also improves the generative capabilities of the HKS-Net by learning a better embedding space (Tab.2 and Tab.3).

Two-View Case. We also consider the case when two complementary input silhouette images (front and side) are given simultaneosuly, which further allows comparisons to some of the related works [62, 13, 6, 16]. For this case, we mainly consider neutral poses. As the architecture, we use the HKS-Net along with a network similar to the one used in a recent work [16] (Fig.1 (5)) where, unlike in CMNN, the weight sharing is performed at early stages during convolutions, and the last convolutional layers are merged through a max-pooling operation. This is then trained with the sum of squared losses $L_{Two-View} + L_{Vert}$, on the embedding space and the mesh vertex locations, as before. Similarly, the mapping to the embedding space is decoded to a 3D mesh space through a forward pass in the dense layers of the HKS-Net. This achieves better results than in the previous works [16], due to the newly learned embedding (Tab.3).

4. Experiments and Results

We have run an extensive set of experiments to ensure the reliability of our technique. In this section, we report results of our qualitative and quantitative tests, with thorough comparisons to the state-of-the-art. In order to quantitatively assess our method, we perform experiments on synthetic data similar to previous works [6, 17, 16] by computing errors on the same 16 body measurements widely utilized in tailor fitting, as shown in Fig. 2. Since all the methods we compare to, as well as ours, make use of the same shape model [2], the comparisons become more reliable through these measurements on estimated meshes in full correspondence.

From the combined datasets [66, 45] of meshes fitted to real body scans, where duplicate removal is ensured as in [16, 17], we set 1000 meshes apart for testing, and utilize the rest for generating the human body model and training data (Sec.3.1). For these left-out meshes we then extract HKS descriptors and silhouettes in various views and poses.

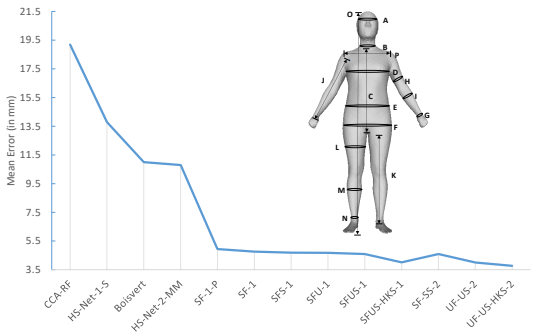

Figure 2. Plot of the mean error over all body measurements illustrated on a mesh, for the methods from Tab.2 abd Tab.3.

We apply LBS [34] to deform the meshes into desired poses compliant with our applications (see supplementary).

We run the methods from two previous works [17, 16] on the silhouettes extracted from these meshes, while for others [62, 13, 6], we report the numbers from their experiments performed on similar but fewer meshes (around 300). In addition to comparisons with the state-of-the-art, we thoroughly evaluate the added value of each component in our network. In the end we conclude with qualitative results and run-time evaluations.

Quantitative Experiments. The 16 measurements are calculated as follows: straight line measurements are computed by Euclidean distances between two extreme vertices, while for the ellipsoidal ones, we calculate the perimeter on the body surface. For each measurement, we report the mean and standard deviations of the errors over all estimated meshes with respect to the ground truth ones. We report errors when only the frontal view silhouette is utilized at test time in Tab. 2, and if both frontal and side view silhouettes are available at test time in Tab. 3. For both tables, we distinguish between two cases: known camera distance (unscaled) and unknown camera distance (called scaled in the subsequent analysis, since we scale the silhouettes to have the same height in this case, as elaborated in Sec. 3.3). The nomenclature for our experiments is summarized in Tab. 1. Note that for all methods in the tables, the errors are for a neutral pose, except for $SF-1-P$, where we show the error measures when we train and test using different poses. The mean error over all body measurements for the methods we consider is depicted in Fig. 2. Our best mean error for the one view cross-modal case is 4.01 mm and for the two-view case is 3.77 mm, showing a very high accuracy for the tasks we consider. These are significantly better than the mean error of the previous works with 19.19 mm [17], 10.8 mm [16], 11 mm [6], and 10.1 mm [25], even though some of these methods operate under more restrictive assumptions. Our best results, that achieve state-of-the-art, are highlighted in bold.

For the one view case (Tab. 2), one can see that as we go

Table 2. Body measurement errors comparison with shapes reconstructed from one scaled frontal silhouette. The nomenclature is presented in Tab. 1. Last two columns show the results of the state-of-the-art methods. The measurements are illustrated in Fig. 2 (top-right). Errors are expressed as Mean±Std. Dev in millimeters. Our best achieving method *SFUS-HKS-1* is highlighted.

Measurements	SF-1-P	SF-1	SFS-1	SFU-1	SFUS-1	SFUS-HKS-1	HS-Net-1-S [16]	CCA-RF [17]
A. Head circumference	4.3±3.5	3.9±3.1	3.7±2.9	3.7±2.9	3.9±2.9	**3.1±2.6**	4±4	8±8
B. Neck circumference	2.2±1.8	2.3±1.8	2.3±1.8	2.3±1.8	2.2±1.7	**2.1±1.7**	8±5	7±7
C. Shoulder-blade/crotch length	6.2±4.9	6.1±4.8	5.3±4.2	5.3±4.1	5.4±4.1	**4.9±3.8**	20±15	18±17
D. Chest circumference	6.7±5.4	6.7±5.3	5.9±4.9	5.9±4.7	5.8±4.8	**5.8±4.8**	13±7	25±24
E. Waist circumference	8.1±6.1	7.8±6.2	7.5±5.9	7.5±5.9	7.5±5.7	**6.4±5.2**	19±13	24±24
F. Pelvis circumference	9.3±7.5	8.8±7.2	8.4±6.7	8.2±6.6	8.1±6.5	**7.1±5.9**	19±14	26±25
G. Wrist circumference	2.1±1.7	2.1±1.7	1.9±1.6	1.9±1.6	1.9±1.6	**1.7±1.5**	5±3	5±5
H. Bicep circumference	3.9±3.1	3.3±2.6	2.9±2.4	2.9±2.4	2.9±2.5	**2.9±2.5**	8±4	11±11
I. Forearm circumference	3.1±2.4	2.9±2.3	3.1±2.3	2.7±2.3	2.9±2.3	**2.6±2.2**	7±4	9±8
J. Arm length	4.1±3.1	3.8±2.9	3.3±2.5	3.3±2.5	3.2±2.5	**2.9±2.4**	12±8	13±12
K. Inside leg length	7.3±5.1	6.8±5.2	6.2±4.8	6.5±4.9	5.7±4.5	**5.4±4.3**	20±14	20±19
L. Thigh circumference	6.3±4.9	6.3±5.5	5.8±4.9	5.7±4.7	5.8±4.8	**5.8±4.9**	13±8	18±17
M. Calf circumference	3.6±2.9	3.5±3.1	3.3±2.7	3.3±2.6	3.5±2.8	**2.9±2.5**	12±7	12±12
N. Ankle circumference	2.1±1.5	2.1±1.7	1.9±1.5	1.8±1.4	2.1±1.5	**1.6±1.3**	6±3	6±6
O. Overall height	12.6±9.9	12.4±9.9	11.2±8.6	10.9±8.4	10.4±8.1	**9.8±7.7**	50±39	43±41
P. Shoulder breadth	2.3±1.9	2.3±1.8	2.2±1.2	2.2±1.9	2.1±1.7	**1.9±1.7**	4±4	6±6

Table 3. Same as in Tab. 2, however with shapes reconstructed from two views at the same time. Last four columns show the results of the other state-of-the-art methods for the same task. Our best achieving method *UF-US-HKS-2* is highlighted.

Measurements	SF-SS-2	UF-US-2	UF-US-HKS-2	HS-2-Net-MM [16]	Boisvert et al. [6]	Chen et al. [15]	Xi et al. [62]
A. Head circumference	3.9±3.2	3.3±2.6	**3.2±2.6**	7.4±5.8	10±12	23±27	50±60
B. Neck circumference	1.9±1.7	2.0±1.6	**1.9±1.5**	5.3±3.1	11±13	27±34	59±72
C. Shoulder-blade/crotch length	5.1±4.1	4.3±3.5	**4.2±3.4**	9.9±7.0	4±5	52±65	119±150
D. Chest circumference	5.4±4.8	5.8±4.3	**5.6±4.7**	19.1±12.5	10±12	18±22	36±45
E. Waist circumference	7.5±5.7	7.6±5.9	**7.1±5.8**	18.4±13.2	22±23	37±39	55±62
F. Pelvis circumference	8.0±6.4	8.0±6.4	**6.9±5.6**	14.9±11.3	11±12	15±19	23±28
G. Wrist circumference	1.9±1.6	1.6±1.4	**1.6±1.3**	3.8±2.7	9±12	24±30	56±70
H. Bicep circumference	3.0±2.6	2.6±2.1	**2.6±2.1**	6.5±4.9	17±22	59±76	146±177
I. Forearm circumference	3.0±2.4	2.9±2.1	**2.2±1.9**	5.5±4.2	16±20	76±100	182±230
J. Arm length	3.3±2.6	2.4±1.9	**2.3±1.9**	8.1±6.4	15±21	53±73	109±141
K. Inside leg length	5.6±5.1	4.3±3.8	**4.3±3.8**	15.6±12.4	6±7	9±12	19±24
L. Thigh circumference	5.8±5.1	5.1±4.3	**5.1±4.3**	13.7±10.8	9±12	19±25	35±44
M. Calf circumference	3.9±3.2	3.1±2.1	**2.7±1.9**	8.5±6.5	6±7	16±21	33±42
N. Ankle circumference	2.1±1.5	1.6±1.1	**1.4±1.1**	4.6±3.2	14±16	28±35	61±78
O. Overall height	10.6±8.6	7.2±6.1	**7.1±5.5**	25.9±20.4	9±12	21±27	49±62
P. Shoulder breadth	2.2±1.8	2.1±1.8	**2.1±1.8**	5.6±3.9	6±7	12±15	24±31

from uni-modal to cross-modal training, by using multiple views at training time and sharing weights in the fully connected layers, the errors constantly decrease. We show the effect of adding a side scaled view only ($SFS-1$), an unscaled frontal view only ($SFU-1$), and combining all three ($SFUS-1$). The lowest errors are achieved through joint training ($SFUS-HKS-1$) of the CMNN and HKS-Net (Sec. 3.4). In this case, not only the accuracy of predictions from silhouettes, but also the accuracy of the HKS-Net itself is improved as compared to when it is separately trained, reducing the mean error over all the meshes from 4.74 to 3.77 mm. We further report results when different poses are applied on the test meshes ($SF-1-P$), in contrast to all other methods considered. Even in this case, the errors do not differ much from the neutral pose case ($SF-1$), implying robustness to variations for the pose space we consider.

For the two view case, we compare to the results of the works that require two views at test time [6, 62, 13, 16]. We utilize the same camera calibration assumptions, and again achieve significant improvements in accuracy ($UF-US-HKS-2$), due to the new shape embedding space jointly trained with the prediction network. For the two view-case, we do not test on multiple poses, since the previous works we compare to are also tested on neutral poses for this particular application. One interesting observation here is that the results for the single view cross-modal case ($SFUS-1$ in Tab. 2) are comparable to, and in some measurements even better than those of the two-view network ($SF-SS-2$ in Tab. 3). Since no joint training was performed in either case, and the loss for both cases is in the shape embedding space, this demonstrates the importance of the shared fully connected layers and cross-modal training for boosting prediction performance at test time.

Qualitative Experiments. We evaluate our method on three test subjects from a previos work [17] in a neutral and selfie pose, and four new subjects with other poses. As can be observed in Fig. 3, our reconstructions resemble the real individuals more closely, as compared to those from Dibra

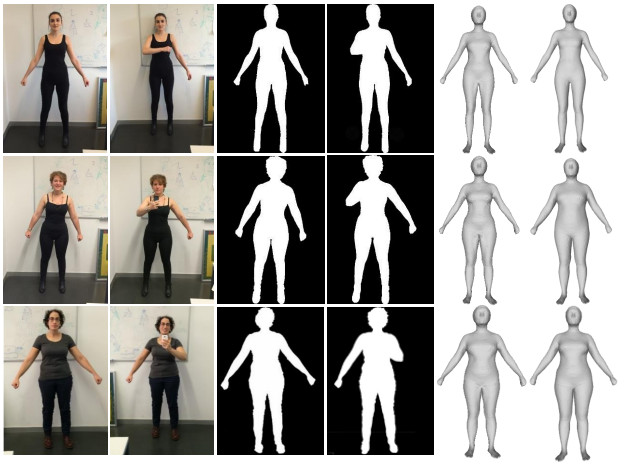

Figure 3. Results for predictions on the test images from Dibra et al. [17]. From left to right: the two input images in a rest and selfie pose, the corresponding silhouettes, the estimated mesh by our method $SF-1-P$, and by the method of Dibra et al. [17].

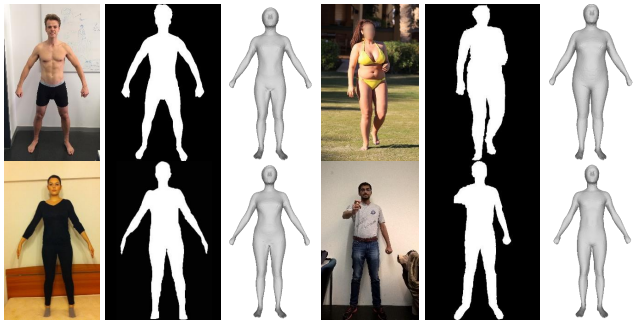

Figure 4. Predictions results on four test subjects in different poses and with clothes. From left to right: input image, the corresponding silhouette, the estimated mesh by our method $SF-1-P$.

et al. [17] (last column), especially for the second subject. We additionally show mesh overlays over the input images, applied also to the method from Bogo et al. [5] in the supplementary. The results in Fig. 4 illustrate harder cases, where the silhouettes differ more from those of the training data due to clothing, poses, and occlusions. Our results still explain the silhouettes well for all cases.

Speed. The training of our network was performed on an Intel(R) Core(TM) i7 CPU 4770 3.4 GHz with NVIDIA GTX 1080 (8G) GPU. It took around 50 min per epoch, with one epoch consisting of roughly $50,000$ samples. The total training time for the various architectures considered in the experiments varies from 15-30 epochs. We conducted our test time experiments on an Intel(R) Core(TM) i7 CPU 950 3.0 GHz with NVIDIA GTX 940 (2GB) GPU. Since our method directly outputs the vertices of a mesh, and does not need to solve a least squares system (Eq. 1), it is much faster (0.15 seconds) than other methods when using the GPU for prediction. Even when using a CPU, our method takes about 0.3 seconds, similar to the fastest method [17], and less than 6 seconds [6] and 0.45 seconds [16], as reported in other previous works. As a result, our method scales to higher mesh resolutions, and can be directly used as an end-to-end pipeline, outputting a full 3D mesh. With the advances in compressed deep networks (e.g. [21, 27]), this can potentially be ported to mobile devices, which is in line with our targeted application of shape from selfies.

Finally, we perform a further experiment with noise added to the silhouettes, as in the previous works [17, 16]. The method is robust to silhouette noise, with a mean error increase of 4.1 mm for high levels of noise. We present further results on poses, silhouette noise, failure cases and a comparison to CCA applied instead of our Cross-Modal Network in the supplementary material.

5. Conclusion and Discussion

We presented a novel method for capturing a 3D human body shape from a single silhouette with unknown camera parameters. This is achieved by combining deep correlated features capturing different 2D views, and embedding spaces based on 3D shape descriptors in a novel CNN-based architecture. We extensively validated our results on synthetic and real data, demonstrating significant improvement in accuracy as compared to the state-of-the-art methods. We illustrated that each component of the architecture is important to achieve these improved results. Combined with the lowest running times over all the state-of-the-art, we thus provide a practical system for detailed human body measurements with millimetric accuracy.

The proposed cross-modal neural network enhances features by incorporating information coming from different modalities at training time. The idea of such correlating networks can be extended for many other problems where privileged data is available, or correlations among different data types (e.g image, text, audio) are to be exploited. HKS-Net like architectures can be used for inverting shape descriptors, which can have various applications for understanding and generating shapes.

Inferring 3D shapes from 2D projections is an ill-posed problem. As in the previous works, we operate under mild occlusions and a certain level of silhouette noise, which are realistic assumptions for many scenarios including ours. However, especially for severe occlusions, we would need stronger priors to infer correct 3D shapes. We believe that extending our techniques for images with shading cues can provide accurate estimations even for such cases. A training covering different environments and textures would be necessary for this case.

Acknowledgment. This work was funded by the KTI-grant 15599.1. We would like to thank Wan-Chun Alex Ma for the help with the datasets and Brian McWilliams for the valuable discussions about CCA.

References

[1] G. Andrew, R. Arora, J. A. Bilmes, and K. Livescu. Deep canonical correlation analysis. In *Proceedings of the 30th International Conference on Machine Learning, ICML 2013, Atlanta, GA, USA, 16-21 June 2013*, pages 1247–1255, 2013. 3

[2] 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*, SIGGRAPH '05, pages 408–416, New York, NY, USA, 2005. ACM. 1, 2, 3, 4, 6

[3] M. Aubry, U. Schlickewei, and D. Cremers. The wave kernel signature: A quantum mechanical approach to shape analysis. In *Computer Vision Workshops (ICCV Workshops), 2011 IEEE International Conference on*, pages 1626–1633. IEEE, 2011. 3

[4] A. O. Balan, L. Sigal, M. J. Black, J. E. Davis, and H. W. Haussecker. Detailed human shape and pose from images. In *2007 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2007), 18-23 June 2007, Minneapolis, Minnesota, USA*, 2007. 2

[5] 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*, pages 561–578. Springer, 2016. 1, 2, 8

[6] J. Boisvert, C. Shu, S. Wuhrer, and P. Xi. Three-dimensional human shape inference from silhouettes: reconstruction and validation. *Mach. Vis. Appl.*, 24(1):145–157, 2013. 1, 2, 3, 4, 6, 7, 8

[7] D. Boscaini, J. Masci, E. Rodolà, and M. M. Bronstein. Learning shape correspondence with anisotropic convolutional neural networks. Technical Report arXiv:1605.06437, 2016. 3

[8] D. Boscaini, J. Masci, E. Rodolà, M. M. Bronstein, and D. Cremers. Anisotropic diffusion descriptors. volume 35, pages 431–441, 2016. 3

[9] A. M. Bronstein, M. M. Bronstein, R. Kimmel, M. Mahmoudi, and G. Sapiro. A gromov-hausdorff framework with diffusion geometry for topologically-robust non-rigid shape matching. *International Journal of Computer Vision*, 89:266–286, 2010. 3

[10] A. O. Bălan and M. J. Black. The naked truth: Estimating body shape under clothing. In *Proceedings of the 10th European Conference on Computer Vision: Part II*, ECCV '08, pages 15–29, Berlin, Heidelberg, 2008. Springer-Verlag. 1, 2

[11] S. Chandar, M. M. Khapra, H. Larochelle, and B. Ravindran. Correlational neural networks. *Neural Computation*, 28(2):257–285, 2016. 3

[12] X. Chen, Y. Guo, B. Zhou, and Q. Zhao. Deformable model for estimating clothed and naked human shapes from a single image. *The Visual Computer*, 29(11):1187–1196, 2013. 2

[13] Y. Chen and R. Cipolla. Learning shape priors for single view reconstruction. In *ICCV Workshops*. IEEE, 2009. 1, 2, 3, 4, 6, 7

[14] Y. Chen, T. Kim, and R. Cipolla. Silhouette-based object phenotype recognition using 3d shape priors. In *IEEE International Conference on Computer Vision, ICCV 2011, Barcelona, Spain, November 6-13, 2011*, pages 25–32, 2011. 2

[15] Y. Chen, T.-K. Kim, and R. Cipolla. Inferring 3d shapes and deformations from single views. In K. Daniilidis, P. Maragos, and N. Paragios, editors, *ECCV (3)*, volume 6313 of *Lecture Notes in Computer Science*, pages 300–313. Springer, 2010. 2, 7

[16] E. Dibra, H. Jain, C. Oztireli, R. Ziegler, and M. Gross. Hs-nets : Estimating human body shape from silhouettes with convolutional neural networks. In *Int. Conf. on 3D Vision*, October 2016. 1, 3, 4, 5, 6, 7, 8

[17] E. Dibra, C. Öztireli, R. Ziegler, and M. Gross. Shape from selfies: Human body shape estimation using cca regression forests. In *Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part V*, pages 88–104. Springer, 2016. 1, 2, 3, 4, 5, 6, 7, 8

[18] Y. Fang, J. Xie, G. Dai, M. Wang, F. Zhu, T. Xu, and E. Wong. 3d deep shape descriptor. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015. 3

[19] A. Feng, D. Casas, and A. Shapiro. Avatar reshaping and automatic rigging using a deformable model. In *Proceedings of the 8th ACM SIGGRAPH Conference on Motion in Games*, pages 57–64, Paris, France, Nov. 2015. ACM Press. 4

[20] P. Guan, A. Weiss, A. O. Balan, and M. J. Black. Estimating human shape and pose from a single image. In *IEEE 12th International Conference on Computer Vision, ICCV 2009, Kyoto, Japan, September 27 - October 4, 2009*, pages 1381–1388, 2009. 1, 2

[21] S. Han, H. Mao, and W. J. Dally. Deep compression: Compressing deep neural network with pruning, trained quantization and huffman coding. *CoRR, abs/1510.00149*, 2, 2015. 8

[22] D. R. Hardoon, S. R. Szedmak, and J. R. Shawe-taylor. Canonical correlation analysis: An overview with application to learning methods. *Neural Comput.*, 16(12):2639–2664, Dec. 2004. 3

[23] N. Hasler, H. Ackermann, B. Rosenhahn, T. Thormählen, and H. 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. 1, 2

[24] N. Hasler, C. Stoll, M. Sunkel, B. Rosenhahn, and H. Seidel. A statistical model of human pose and body shape. *Comput. Graph. Forum*, 28(2):337–346, 2009. 2

[25] T. Helten, A. Baak, G. Bharaj, M. Müller, H. Seidel, and C. Theobalt. Personalization and evaluation of a real-time depth-based full body tracker. In *2013 International Conference on 3D Vision, 3DV 2013, Seattle, Washington, USA, June 29 - July 1, 2013*, pages 279–286, 2013. 1, 6

[26] H. Hotelling. Relations between two sets of variates. *Biometrika*, 28(3/4):321–377, Dec. 1936. 1, 3

[27] F. N. Iandola, M. W. Moskewicz, K. Ashraf, S. Han, W. J. Dally, and K. Keutzer. Squeezenet: Alexnet-level accuracy with 50x fewer parameters and¡ 1mb model size. *arXiv preprint arXiv:1602.07360*, 2016. 8

[28] N. Iyer, S. Jayanti, K. Lou, Y. Kalyanaraman, and K. Ramani. Three-dimensional shape searching: state-of-the-art review and future trends. *Computer-aided Design*, 37(5):509–530, 2005. 3

[29] A. Jain, T. Thormählen, H.-P. Seidel, and C. Theobalt. Moviereshape: Tracking and reshaping of humans in videos. *ACM Trans. Graph. (Proc. SIGGRAPH Asia 2010)*, 29(5), 2010. 1, 2, 4

[30] I. A. Kakadiaris and D. Metaxas. Three-dimensional human body model acquisition from multiple views. *International Journal of Computer Vision*, 30(3):191–218, 1998. 2

[31] D. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 5

[32] Z. Lahner, E. Rodola, F. R. Schmidt, M. M. Bronstein, and D. Cremers. Efficient globally optimal 2d-to-3d deformable shape matching. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016. 2

[33] A. Laurentini. The visual hull concept for silhouette-based image understanding. *IEEE Trans. Pattern Anal. Mach. Intell.*, 16(2):150–162, Feb. 1994. 2

[34] J. P. Lewis, M. Cordner, and N. Fong. Pose space deformation: A unified approach to shape interpolation and skeleton-driven deformation. In *Proceedings of the 27th Annual Conference on Computer Graphics and Interactive Techniques*, SIGGRAPH '00, pages 165–172, New York, NY, USA, 2000. ACM Press/Addison-Wesley Publishing Co. 4, 6

[35] R. Litman and A. M. Bronstein. Learning spectral descriptors for deformable shape correspondence. *IEEE Trans. Pattern Anal. Mach. Intell.*, 36(1):171–180, 2014. 3

[36] M. Loper, N. Mahmood, J. Romero, G. Pons-Moll, and M. J. Black. Smpl: A skinned multi-person linear model. *ACM Trans. Graph.*, 34(6):248:1–248:16, Oct. 2015. 2, 4

[37] L. v. d. Maaten and G. Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(Nov):2579–2605, 2008. 5

[38] J. Masci, D. Boscaini, M. M. Bronstein, and P. Vandergheynst. Geodesic convolutional neural networks on riemannian manifolds. In *Proc. of the IEEE International Conference on Computer Vision (ICCV) Workshops*, pages 37–45, 2015. 3

[39] J. Masci, D. Boscaini, M. M. Bronstein, and P. Vandergheynst. Shapenet: Convolutional neural networks on non-euclidean manifolds. *CoRR*, abs/1501.06297, 2015. 3

[40] B. McWilliams, D. Balduzzi, and J. M. Buhmann. Correlated random features for fast semi-supervised learning. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 26*, pages 440–448. Curran Associates, Inc., 2013. 3

[41] I. Mikic, M. Trivedi, E. Hunter, and P. Cosman. Human body model acquisition and tracking using voxel data. *International Journal of Computer Vision*, 53(3):199–223, 2003. 2

[42] A. Neophytou and A. Hilton. Shape and pose space deformation for subject specific animation. In *Proceedings of the 2013 International Conference on 3D Vision*, 3DV '13, pages 334–341, Washington, DC, USA, 2013. IEEE Computer Society. 2, 4

[43] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, and A. Y. Ng. Multimodal deep learning. In *Proceedings of the 28th International Conference on Machine Learning, ICML 2011, Bellevue, Washington, USA, June 28 - July 2, 2011*, pages 689–696, 2011. 3

[44] D. Pickup, X. Sun, P. L. Rosin, R. R. Martin, Z. Cheng, Z. Lian, M. Aono, A. B. Hamza, A. Bronstein, M. Bronstein, S. Bu, U. Castellani, S. Cheng, V. Garro, A. Giachetti, A. Godil, J. Han, H. Johan, L. Lai, B. Li, C. Li, H. Li, R. Litman, X. Liu, Z. Liu, Y. Lu, A. Tatsuma, and J. Ye. Shape retrieval of non-rigid 3d human models. In *Proceedings of the 7th Eurographics Workshop on 3D Object Retrieval*, 3DOR '15, pages 101–110, Aire-la-Ville, Switzerland, Switzerland, 2014. Eurographics Association. 3

[45] L. Pishchulin, S. Wuhrer, T. Helten, C. Theobalt, and B. Schiele. Building statistical shape spaces for 3d human modeling. *CoRR*, abs/1503.05860, 2015. 4, 6

[46] H. Rhodin, N. Robertini, D. Casas, C. Richardt, H. Seidel, and C. Theobalt. General automatic human shape and motion capture using volumetric contour cues. In *Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part V*, pages 509–526, 2016. 1, 2

[47] K. M. Robinette and H. A. M. Daanen. The caesar project: A 3-d surface anthropometry survey. In *2nd International Conference on 3D Digital Imaging and Modeling (3DIM '99), 4-8 October 1999, Ottawa, Canada*, pages 380–387, 1999. 2, 4

[48] M. R. A. K. G. Rohit Girdhar, David F Fouhey. Learning a predictable and generative vector representation for objects. *European Conference on Computer Vision*, 2016. 3

[49] R. M. Rustamov. Laplace-beltrami eigenfunctions for deformation invariant shape representation. 2007. 3

[50] M. E. Sargin, Y. Yemez, E. Erzin, and A. M. Tekalp. Audiovisual synchronization and fusion using canonical correlation analysis. *IEEE Trans. Multimedia*, 9(7):1396–1403, 2007. 3

[51] M. Savva, F. Yu, H. Su, M. Aono, B. Chen, D. Cohen-Or, W. Deng, H. Su, S. Bai, X. Bai, et al. Shrec16 track large-scale 3d shape retrieval from shapenet core55. 3

[52] F. R. Schmidt, D. Farin, and D. Cremers. Fast matching of planar shapes in sub-cubic runtime. In *ICCV*, pages 1–6. IEEE Computer Society, 2007. 2

[53] F. R. Schmidt, E. Töppe, and D. Cremers. Efficient planar graph cuts with applications in computer vision. In *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, Miami, Florida, Jun 2009. 2

[54] A. Sharma, A. Kumar, H. D. III, and D. W. Jacobs. Generalized multiview analysis: A discriminative latent space. In *2012 IEEE Conference on Computer Vision and Pattern Recognition, Providence, RI, USA, June 16-21, 2012*, pages 2160–2167, 2012. 3

[55] L. Sigal, A. O. Balan, and M. J. Black. Combined discriminative and generative articulated pose and non-rigid shape estimation. In J. C. Platt, D. Koller, Y. Singer, and S. T. Roweis, editors, *NIPS*. Curran Associates, Inc., 2007. 2, 3

[56] H. Su, S. Maji, E. Kalogerakis, and E. Learned-Miller. Multi-view convolutional neural networks for 3d shape recognition. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 945–953, 2015. 3

[57] J. Sun, M. Ovsjanikov, and L. J. Guibas. A concise and provably informative multi-scale signature based on heat diffusion. 28(5):1383–1392, 2009. 1, 3, 4, 5

[58] J. W. H. Tangelder and R. C. Veltkamp. A survey of content based 3d shape retrieval methods. *Multimedia Tools and Applications*, 39(3):441, 2008. 3

[59] D. V. Vranic, D. Saupe, and J. Richter. Tools for 3d-object retrieval: Karhunen-loeve transform and spherical harmonics. 2001. 3

[60] L. Wei, Q. Huang, D. Ceylan, E. Vouga, and H. Li. Dense human body correspondences using convolutional networks. In *Computer Vision and Pattern Recognition (CVPR)*, 2016. 3

[61] Z. Wu, S. Song, A. Khosla, F. Yu, L. Zhang, X. Tang, and J. Xiao. 3d shapenets: A deep representation for volumetric shapes. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015. 3

[62] P. Xi, W. Lee, and C. Shu. A data-driven approach to human-body cloning using a segmented body database. In *Proceedings of the Pacific Conference on Computer Graphics and Applications, Pacific Graphics 2007, Maui, Hawaii, USA, October 29 - November 2, 2007*, pages 139–147, 2007. 2, 3, 4, 6, 7

[63] J. Xie, M. Wang, and Y. Fang. Learned binary spectral shape descriptor for 3d shape correspondence. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016. 3

[64] F. Yan and K. Mikolajczyk. Deep correlation for matching images and text. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015. 3

[65] J. Yang, J. Franco, F. Hétroy-Wheeler, and S. Wuhrer. Estimation of human body shape in motion with wide clothing. In *Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part IV*, pages 439–454, 2016. 4

[66] Y. Yang, Y. Yu, Y. Zhou, S. Du, J. Davis, and R. Yang. Semantic parametric reshaping of human body models. In *2nd International Conference on 3D Vision, 3DV 2014, Tokyo, Japan, December 8-11, 2014, Volume 2*, pages 41–48, 2014. 4, 6

[67] S. Zhou, H. Fu, L. Liu, D. Cohen-Or, and X. Han. Parametric reshaping of human bodies in images. *ACM Trans. Graph.*, 29(4), 2010. 1, 2