Single Image 3D Hand Reconstruction with Mesh Convolutions

Dominik Kulon[1, 3]
d.kulon17@imperial.ac.uk
Haoyang Wang[1, 3]
haoyang.wang15@imperial.ac.uk
Riza Alp Güler[1, 3]
r.guler@imperial.ac.uk
Michael Bronstein[1, 2, 4]
m.bronstein@imperial.ac.uk
Stefanos Zafeiriou[1, 3]
s.zafeiriou@imperial.ac.uk

[1] Imperial College London
Department of Computing
London, UK

[2] Università della Svizzera italiana
Institute of Computational Science
Lugano, Switzerland

[3] Ariel AI
arielai.com

[4] Twitter, Inc.
twitter.com

Abstract

Monocular 3D reconstruction of deformable objects, such as human body parts, has been typically approached by predicting parameters of heavyweight linear models. In this paper, we demonstrate an alternative solution that is based on the idea of encoding images into a latent non-linear representation of meshes. The prior on 3D hand shapes is learned by training an autoencoder with intrinsic graph convolutions performed in the spectral domain. The pre-trained decoder acts as a non-linear statistical deformable model. The latent parameters that reconstruct the shape and articulated pose of hands in the image are predicted using an image encoder. We show that our system reconstructs plausible meshes and operates in real-time. We evaluate the quality of the mesh reconstructions produced by the decoder on a new dataset and show latent space interpolation results. Our code, data, and models will be made publicly available.

1 Introduction

Convolutional Neural Networks (CNNs) have been effectively used in computer vision tasks on human geometry understanding, such as 3D model fitting and surface correspondence estimation. Recent model based 3D reconstruction systems [22, 29, 30] predict parameters of a statistical deformable model of the human body and a weak perspective camera for the alignment. These systems rely on the SMPL model [26], where the shape of the person and deformations due to human pose are modelled with linear bases. In this paper, we take a radically different approach and propose a new paradigm for 3D deformable alignment and reconstruction of articulated shapes by exploiting the intrinsic structure of the 3D hand meshes using graph convolutions. Our system is presented in Figure 1.

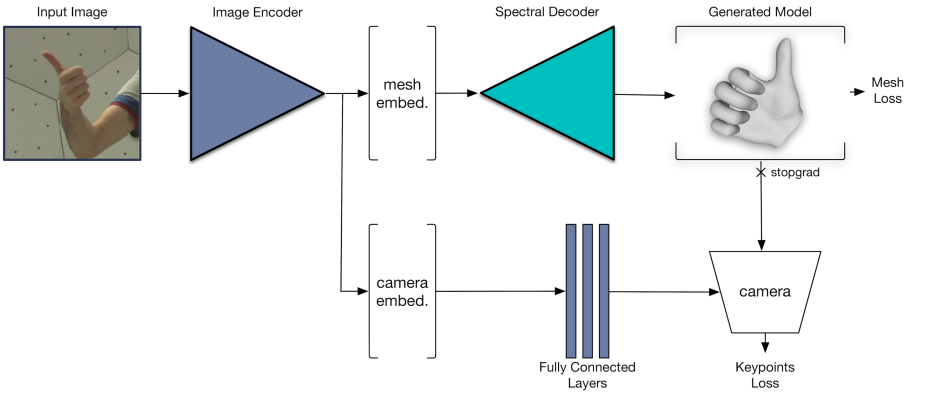

Figure 1: The presented system generates a 3D hand model corresponding to the hand in the image. The system consists of two branches trained simultaneously. The network in the top branch is capable of generating 3D hand models from a single image in a wide variety of shapes and poses. The ground truth meshes are aligned to the mean hand model in a canonical frame. The network in the lower branch regresses parameters of a weak perspective projection to align the generated mesh with the image. The gradient from the keypoints loss does not flow through the top branch to prevent the camera regressor from affecting the quality of mesh reconstructions.

To summarise, our contributions are as follows:

- In order to create the high-quality training data, we build a new high-resolution model of the hand. We fit the model to around thousand scans to compute a distribution of valid pose and shape parameters and to 3D annotations from the Panoptic Studio dataset [34] to obtain the ground truth data for the mesh recovery system.

- We train an autoencoder on sampled meshes with removed scale variations and global orientation. By reusing the decoder, we obtain the first graph morphable model of the human hand. Pose deformations are learned directly by a neural network and therefore the model does not suffer from computational drawbacks and training inefficiency of skinning methods with corrective offsets. It also has a significantly smaller number of parameters than the initial model and constraints the space of valid poses.

- We train an image encoder and take advantage of the pre-trained spectral decoder to recover 3D hand meshes (Figure 1). The resulting system is able to generate hand models in real time. Additionally, we incorporate a camera regression network to compute a weak projection. We find that the network outcompetes the baseline method on the mesh reconstruction task.

Finally, we make all the contributed items (i.e. model, ground truth data, source code) publicly available at https://github.com/dkulon/hand-reconstruction.

2 Related Work

Statistical Modelling of 3D Shapes. The work that popularized statistical modelling of 3D shape and texture came from Blanz and Vetter [4]. They constructed the first statistical 3D Morphable Model (3DMM) by performing dimensionality reduction on a set of meshes representing human faces with a fixed topological structure. It was a PCA-based implementation that produces new shape instances from a combination of linear basis.

However, pose changes of the body joints and soft tissue deformations are non-linear and thus cannot be captured by PCA. The first deformable model of the human body was created by Allen et al. [1]. They constructed a PCA model from 250 registered body scans in an A-pose coming from the CAESAR dataset. The pose changes are obtained by applying linear blend skinning which interpolates rotations matrices assigned to the joints to transform a vertex. This technique, widely used in computer graphics, causes the loss of volume of the surface close to the joints.

The follow-up work [2] tries to solve this issue by adding a corrective offset. The Skinned Multi-Person Linear model (SMPL) [26] uses a similar approach to produce more realistic results by learning pose-dependent corrective blend shapes from a large number of scans. MANO [32] is an analogously defined hand model with 778 vertices introduced because SMPL does not include hand blend shapes.

While SMPL-based models produce realistic results, they are heavyweight in terms of parameterization causing slow inference and the optimization process is time-consuming. Moreover, these models have a low number of vertices and it is arguable whether corrective offsets have enough representation power to capture fine-details of soft-tissue deformations. Nevertheless, they have become a standard tool for recovering human body parts where the problem is being solved by regressing axis-angles of joint rotations and blend shape weights given an image with keypoint annotations.

Body Mesh Recovery. Numerous methods have been proposed to recover a subject-specific body mesh from an RGB image by finding parameters of the SMPL model. In SMPLify, Bogo et al. [5] estimate the 2D body joint locations and use these estimates to iteratively fit the SMPL model. The space of plausible shapes is constrained by adding prior and interpenetration error terms to the objective function. The shape parameters are only derived from the joint locations and thus the shape of the predicted model is close to the average.

The issue of shape estimation inaccuracy is addressed in other works by introducing a silhouette error term. Lassner et al. [23] extend the SMPLify method by penalizing the bi-directional distance from points of the projected model to the ground-truth silhouette. This approach provides a better estimate of the body shape than its predecessor but it again requires numerically solving a complex optimization problem.

The Human Mesh Recovery framework [22] reconstructs the body mesh in an end-to-end manner by regressing the model parameters and minimizing the reprojection loss of the joint locations given by the SMPL model and the ground-truth 2D image annotations. The authors introduce an adversary to discriminate body parameters that constrains the pose space. Tan et al. [36] train a decoder from the SMPL parameters to a body silhouette and afterwards they train an encoder-decoder network that outputs a silhouette given an RGB image. Pavlakos et al. [30] train a neural network that consists of modules responsible for predicting the mask, landmarks, and model parameters. Varol et al. [37] combine segmentation, 2D pose, and 3D pose predictions to infer the volumetric representation of the body into which they fit the

SMPL model. Neural Body Fitting [29] is an end-to-end system that integrates the SMPL model within a CNN. The network is a sequence of layers responsible for computing the segmentation, model parameters, and joint positions. Finally, the joints are projected onto the image to compute the loss. Weng at al. [39] show that it is possible to create an animated character from a single photo.

All models described above have none or very limited variance of hand poses and face expressions. The authors of the Total Capture system [21] build a unified body model that is iteratively fit into the point cloud obtained from a multi-camera setup. The system produces impressive results but is limited to the laboratory setting. The follow-up work [41] enables predicting the model parameters in the wild by fitting the model into predicted joint confidence and orientation maps.

Hand Mesh Recovery. Concurrently to our work, five methods have been proposed to recover a hand mesh from a single image. Four of them rely on regressing parameters of the MANO model whereas our hand model is represented as a lightweight generator trained end-to-end in an unsupervised manner.

The first work by Boukhayma et al. [7] proposes a system analogous to the Human Mesh Recovery and is evaluated on images in the wild. The second work by Zhang et al. [43] introduces a method that iteratively regresses model parameters from the heatmaps. Baek et al. [3] combine an iterative refinement with a differentiable renderer and Hasson et al. [19] reconstruct hand meshes with objects they interact with.

The last work by Ge et al. [15] is similar to ours but uses depth maps for supervision and focuses on hand pose estimation while our goal is to recover a realistic hand model. All authors claim state-of-the-art performance on evaluated hand pose estimation benchmarks what shows the benefits of using a model-based representation.

Geometric Deep Learning Recent works in the area of geometric deep learning [9] generalized convolutions to graphs and Riemannian manifolds. Particularly, these methods allow us to extract features from a local neighbourhood of a vertex placed in a triangular mesh.

Bruna et al. [10] expressed convolutions in the spectral domain where the filter is parameterized using a smooth spectral transfer function. This method has multiple disadvantages. The filters are basis-dependent and thus do not generalize across domains. The method requires $O(n)$ parameters per layers and the computation of forward and inverse Fourier transforms has $O(n^2)$ complexity. Moreover, there is no guarantee of spatial localization of filters.

Defferrard et al. [14] represent a spectral transfer function as a polynomial of degree r. This approach has a constant number of parameters per layer and filters are guaranteed to have r-hops support. There is also no explicit computation of the eigenvectors resulting in $O(nr)$ computational complexity. CayleyNet [24] replaced polynomial filters with rational complex functions to avoid the Laplacian eigendecomposition and achieve better spectral resolution.

The first mesh convolutional neural networks such as Geodesic CNN [27], Anisotropic CNN [6], and MoNet [28] define convolutions in the spatial domain where it is possible to orient the kernels. FeastNet [38] uses an attention mechanism to weight the neighbour selection whereas spiral filters traverse adjacent vertices in a fixed order [25].

Ranjan et al. [31] use fast spectral convolutions [14] to find a low-dimensional non-linear representation of the human face in an unsupervised manner. Their network has 75% less

parameters than a PCA-based morphable model while obtaining improved reconstruction results. Cheng et al. [12] define this architecture in an adversarial setting obtaining more detailed reconstructions and Zhou et al. [44] include additional features in the graph structure to predict the face shape with texture. In this work, we show for the first time that mesh autoencoders, if appropriately trained, can be used to represent highly-articulated objects such as hands.

3 Graph Morphable Model

We introduce the first hand model learned form a collection of meshes in an unsupervised way. The model is capable of generating realistic posed hand shapes with 7,907 vertices. We achieve this by training a mesh autoencoder with convolutions performed in the frequency domain. The decoder has only four layers resulting in low computational requirements. Interpolation in the latent space produces meaningful transitions enabling fast optimization of the objective function for the problem of model fitting by regressing latent parameters (Figure 1).

3.1 Graph Fourier Transform

Non-normalized graph Laplacian is defined as $\Delta = D - W$ where $D = diag(\sum_j W_{ij})$ is the degree matrix and W a weighted adjacency matrix. For any signal $\vec{f} : V \rightarrow R^N$ defined on an undirected graph $G = (V, \mathcal{E}, W)$, the graph Laplacian satisfies

$$(\Delta f)(i) = \sum_{j \in N_i} W_{ij}(f(i) - f(j)) \tag{1}$$

where $N_i = \{j \,|\, (i,j) \in \mathcal{E}\}$ it the neighbourhood of a vertex i, V a set of vertices, and $\mathcal{E}$ a set of edges [13, 33].

Graph Laplacians are symmetric and positive semi-definite. Therefore they have a complete set of orthogonal eigenvectors $\Phi = (\phi_1, ..., \phi_n)$ where $n = |V|$ with real, non-negative eigenvalues $\lambda_1, ..., \lambda_n$. Due to these properties, Δ admits an eigendecomposition $\Delta = \phi \Lambda \phi^T$ where $\Lambda = (\lambda_1, ..., \lambda_n)$. Graph Fourier transform is the expansion of $\vec{f}$ in terms of the eigenvectors of the graph Laplacian which can be written in the matrix form $\hat{\vec{f}} = \phi^T \vec{f}$. It follows that the inverse graph Fourier transform is given by $\vec{f} = \phi \hat{\vec{f}}$.

3.2 Spectral Convolutions

Bruna et al. [10] expressed convolution in Fourier space $(f \star h) = \hat{h}(\Delta) f$ where the filter $\hat{h}$ is parameterized using a smooth spectral transfer function

$$\hat{h}(\Delta) = \Phi \hat{h}(\Lambda) \Phi^T \tag{2}$$

$$\hat{h}(\Lambda) = diag(\hat{h}(\lambda_1), ..., \hat{h}(\lambda_n)). \tag{3}$$

Defferrard et al. [14] parameterized the filter with a Chebyshev expansion of order $r-1$ such that

$$h(\tilde{\Delta}) = \sum_{j=0}^{r-1} \alpha_j T_j(\tilde{\Delta}) \tag{4}$$

where $T_k(\lambda) = 2\lambda T_{k-1}(\lambda) - T_{k-2}(\lambda)$ with $T_0 = 1$ and $T_1 = \lambda$. $\tilde{\Delta} = 2\lambda_n^{-1}\Delta - \mathrm{I}$ denotes rescaling of the Laplacian eigenvalues from the interval $[0, \lambda_n]$ to $[-1, 1]$. As we discussed in the Related Work section, this approach is computationally faster and filters are localized with r-hops support.

3.3 Network Architecture

We follow design choices of CoMA [31] which is a mesh autoencoder with four layers of convolutions followed by downsampling. We start with input mesh with 7,907 vertices followed by sequence of convolutions (16, 32, 32, 48 filters) and downsampling after each convolutional layer (4, 4, 2, 2 graph reduction factors). Afterwards, we apply a fully connected layer to obtain a latent vector with 64 parameters. The decoder is symmetric. We choose leaky ReLU for the activation function based on experimental evaluation and we use a smaller filter with the Chebyshev polynomial of order $r = 3$. Downsampling approach is also adopted from CoMA which minimises the quadric error to decimate the template. However, we compute downsampled topology from a mesh in a half-closed hand position in Blender. We find that the network trained with the original implementation produces extremely noisy meshes because decimated templates do not have vertices around the joints. The choice of downsampled graph topology has the most significant effect on the quality of mesh reconstructions.

The training data comes from a MANO-like model with 7,907 vertices. The process of sampling realistic templates is described in Section 5. We train the network for 6 hours on a single GeForce RTX 2080 Ti with a batch size 64 and learning rate 0.001. In addition to the L1 reconstruction loss, we impose an L2 penalty on the latent vector weighted by 5e-7 and we use L2 regularization scaled by 5e-5. The loss function is minimized with AdamW opimizer with a decay factor 10e-6. Our Graph Morphable Model (GMM) is the decoder with $Z = 64$ latent parameters.

4 Single Image Mesh Generation

We take the dataset of hand images with corresponding 3D hand meshes aligned in canonical coordinates (Section 5). The single image mesh generator (Figure 1) consists of the image encoder E_{image}, for which we use the DenseNet-121 [20] network, pretrained on the ImageNet classification task. The outputs of the E_{image} are a mesh embedding that is passed to the pre-trained Graph Morphable Model $\mathcal{D}_{GMM}$, and a camera embedding fed into fully connected layers to estimate parameters of a weak perspective camera $\mathcal{D}_{camera}$. The hand joints, obtained by taking the average of the surrounding ring vertices in the generated mesh, are projected onto the image plane to compute the loss based on 2D keypoint annotations. More specifically, we minimize the loss:

$$\begin{aligned}\mathcal{L} = &\sum_i |\hat{\mathcal{Y}}_i - \mathcal{D}_{GMM}(E_{image}(X_i))|_1 \\ &+ \lambda_{kpts} \sum_i |\hat{\mathcal{J}}_i - \mathcal{D}_{camera}(E_{image}(X_i))|_1 \\ &+ \lambda_{embed} \sum_i ||E_{image}(X_i)_{1:Z}||_2\end{aligned} \tag{5}$$

Figure 2: a) Random shapes sampled from our high resolution linear model. Posed examples can be found in Figure 5. b) Samples from the MANO model (taken from [32]).

for ground truth meshes $\hat{\mathcal{Y}}_i$, keypoint annotations $\hat{\mathcal{J}}_i$, and input images X_i. We set hyper-parameters λ_{kpts} to 0.01 and λ_{embed} to 5e-5. We also add L2 regularization weighted by 1e-5.

During training, we freeze the weights of the GMM and train the image encoder and camera regressor simultaneously for 130 epochs with the same hardware setting and optimizer as the autoencoder but we set learning rate to 1e-4. The network is able to reproduce hand pose early in the training while longer optimization reduces noise around fingertips for extreme poses. We zero the gradient that flows from the camera regression module through $\mathcal{D}_{GMM}$ as we find it to provide better reconstruction results.

5 Evaluation

There are no existing benchmarks that contain images of hands with corresponding meshes or large collection of meshes that can be used to train the Graph Morphable Model and a body part recovery system. Therefore, to address both issues we build a new high resolution hand model following MANO with ten times more vertices and removed scale variations from PCA linear bases. Afterwards, we fit this model to 3D keypoints annotations from the Panoptic dome dataset [34]. We also compute a distribution of valid poses from registrations of around a thousand scans from the MANO dataset to sample realistic meshes for training the autoencoder.

Panoptic DomeDb. This dataset provides 3D annotations of the whole human body including 21 hand joints. It contains multiple sequences of videos with a large number of subjects captured from multiple views in a laboratory setting. We use a subset of around 30 synchronized HD cameras and sample best annotations from a single camera in terms of visibility after each 100th frame. Our preprocessed dataset contains 40360 training, 3000 validation, and 3000 test samples with 3D keypoint annotations projected to the image coordinates.

Hand Model. The number of vertices in MANO (778) is not sufficient to realistically model shape deformations of the human hand. A potential increase in the number of vertices introduces difficulties with scan registrations and significantly increases optimization time for the tasks of parameter training and pose fitting. Moreover, the shape model in MANO was created by applying a dimensionality reduction to rigidly aligned scans. This results in a first principal component that models scale deformations. This is undesirable for our

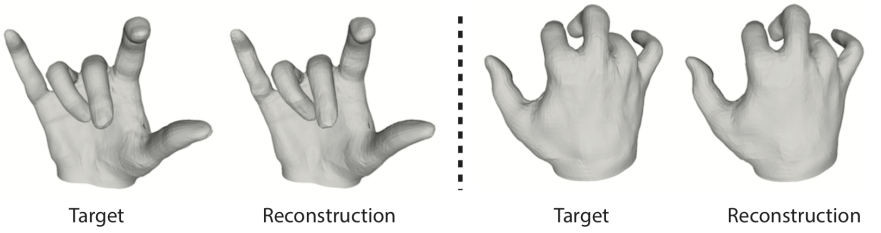

Figure 3: Autoencoder reconstruction results for challenging samples in terms of the pose (left pair) and shape (right pair).

model fitting procedure that applies a similarity alignment at the beginning of optimization. To address these issues, we introduce a hand model with 7,907 vertices which is more than SMPL and MANO combined. We register the reference template to a set of scans with selected keypoints. Then the shape components are computed by applying PCA to the set of Procrustes aligned registrations. The size of the model is controlled by a scale multiplier. Figure 2 shows that we are able to model fine-details of the hand surface. Finally, we optimize pose, shape, and perspective parameters with the Dogleg method to match the ground truth keypoint annotations [5].

Pose Sampler. We use samples obtained from our high-resolution linear model to train the proposed mesh autoencoder. The shape coefficients are sampled from a standard normal distribution. In order to sample plausible and diverse hand poses, we resort to the distribution of joint angles in the MANO database. For each 15 joint angles in the human hand kinematic tree, we compute euler-angle clusters via K-means [16]. During synthesis of training samples, we randomly select euler-angle cluster centers for each joint, effectively sampling shapes from the manifold of plausible shapes. In our experiments, we have used 64 rotation centers.

Results. We implement a baseline method that replaces a spectral decoder from Figure 1 with a TensorFlow implementation of our high resolution MANO model (MANO-like). We also fine-tune the network (Spectral, fine-tuned) by training our network (Spectral, fixed) for 170 more iterations without freezing the decoder's weights. Table 1 shows that we are able to obtain a lower reconstruction error despite the fact that the MANO model was used to generate the ground truth data in the first place. The table also presents that the spectral decoder is six times smaller in terms of number of parameters, obtains lower inference time, and operates in real-time. The DenseNet-121 image encoder runs at 53 FPS, therefore, the whole system maintains real-time performance. Qualitatively, we show that the autoencoder produces visually indistinguishable reconstructions from the input (Figure 3) and has a meaningful latent representation providing smooth transitions in the latent space (Figure 4). The visualization of the results of the mesh recovery system is presented in Figure 5.

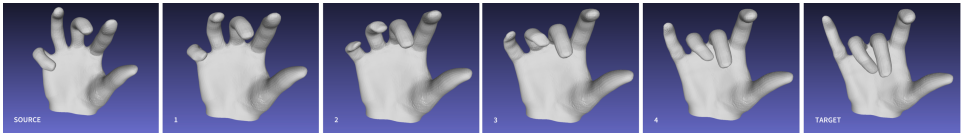

Figure 4: Interpolation in the latent space between two random samples. Please, note that source and target shapes (e.g. finger length) are also different.

	Spectral, fixed	Spectral, fine-tuned	MANO-like
Reconstruction error [mm]	2.33	**2.30**	2.56
Inference time (generator) [ms]	-	**3.04**	4.64
Inference time (generator) [fps]	-	**329**	216
Number of params. (generator)	**393,080**	**393,080**	2,498,612

Table 1: Mesh L1 reconstruction error, inference time, and number of parameters for systems with different types of generators evaluated on Panoptic DomeDb. The scale of the target meshes is smaller than in real world.

6 Conclusion

We proposed a system for generating a subject-specific hand model from a single image. To achieve this, we trained a graph morphable model in an unsupervised manner obtaining a lightweight non-linear representation of hand shapes. The resulting generator was connected to the image encoder and camera regression networks to produce meshes aligned with images. Our system is able to produce realistic hand shapes in real-time that match the target models. To train the networks, we generated a dataset of images with corresponding 3D meshes using a high-vertex count hand model with blend shapes learned from scans.

The morphable model could benefit from defining kernels in a different domain. Spectral methods do not impose canonical ordering of neighbours because the kernels are isotropic due to rotation invariance of the graph Laplacian. In the spatial domain, we can address the issue of orientation ambiguity by imposing a canonical direction or angular max pooling. We expect that spatial localization of filters would allow the network to model fine-details of body part deformations in contrast to spectral methods that average the neighbours [8]. We observed that the choice of a downsampled graph structure has the utmost importance on the performance of the network. Therefore, the need for a learnable graph coarsening mechanism [11, 42] naturally arises in our scenario. We will explore spatial mesh convolutions and differentiable pooling in the future works. In terms of applications beyond computer graphics, our system can be modified to solve the problems of hand pose estimation [34, 35, 45], fingertip detection [40], or dense correspondence computation [17, 18].

7 Acknowledgements

The work of S. Zafeiriou and R. A. Güler has been partially funded by the EPSRC Fellowship Deform (EP/S010203/1). The work of M. Bronstein has been partially funded by ERC Consolidator Grant No. 724228 (LEMAN). Finally, the work of D. Kulon was funded by a PhD scholarship.

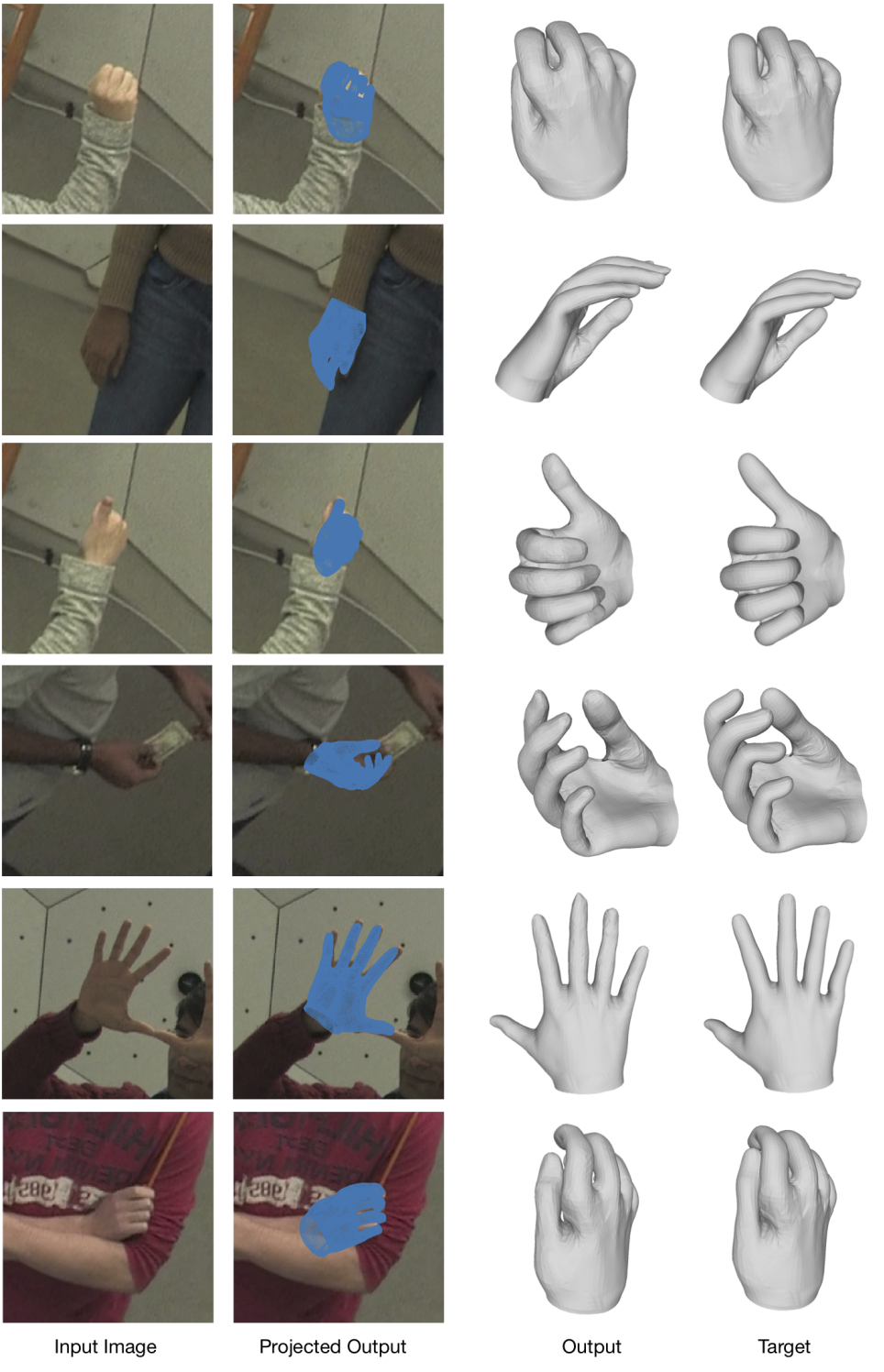

Figure 5: Qualitative results of our system.

References

[1] Brett Allen, Brian Curless, Brian Curless, and Zoran Popović. The space of human body shapes: Reconstruction and parameterization from range scans. *ACM Trans. Graph.*, 22(3):587–594, July 2003. ISSN 0730-0301. doi: 10.1145/882262.882311. URL http://doi.acm.org/10.1145/882262.882311.

[2] Brett Allen, Brian Curless, Zoran Popović, and Aaron Hertzmann. Learning a correlated model of identity and pose-dependent body shape variation for real-time synthesis. In *Proceedings of the 2006 ACM SIGGRAPH/Eurographics symposium on Computer animation*, pages 147–156, Aire-la-Ville, Switzerland, 2006. Eurographics Association.

[3] Seungryul Baek, Kwang In Kim, and Tae-Kyun Kim. Pushing the envelope for rgb-based dense 3d hand pose estimation via neural rendering. *CoRR*, abs/1904.04196, 2019. URL http://arxiv.org/abs/1904.04196.

[4] 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*, pages 187–194. ACM Press/Addison-Wesley Publishing Co., 1999.

[5] 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.

[6] Davide Boscaini, Jonathan Masci, Emanuele Rodolà, and Michael M. Bronstein. Learning shape correspondence with anisotropic convolutional neural networks. *CoRR*, abs/1605.06437, 2016. URL http://arxiv.org/abs/1605.06437.

[7] Adnane Boukhayma, Rodrigo de Bem, and Philip H. S. Torr. 3d hand shape and pose from images in the wild. *CoRR*, abs/1902.03451, 2019. URL http://arxiv.org/abs/1902.03451.

[8] Giorgos Bouritsas, Sergiy Bokhnyak, Stylianos Ploumpis, Michael Bronstein, and Stefanos Zafeiriou. Neural 3D Morphable Models: Spiral Convolutional Networks for 3D Shape Representation Learning and Generation. *arXiv e-prints*, art. arXiv:1905.02876, May 2019.

[9] Michael M Bronstein, Joan Bruna, Yann LeCun, Arthur Szlam, and Pierre Vandergheynst. Geometric deep learning: going beyond euclidean data. *IEEE Signal Processing Magazine*, 34(4):18–42, 2017.

[10] Joan Bruna, Wojciech Zaremba, Arthur Szlam, and Yann LeCun. Spectral networks and locally connected networks on graphs. *CoRR*, abs/1312.6203, 2013. URL http://arxiv.org/abs/1312.6203.

[11] Cătălina Cangea, Petar Veličković, Nikola Jovanović, Thomas Kipf, and Pietro Liò. Towards Sparse Hierarchical Graph Classifiers. *arXiv e-prints*, art. arXiv:1811.01287, Nov 2018.

[12] Shiyang Cheng, Michael Bronstein, Yuxiang Zhou, Irene Kotsia, Maja Pantic, and Stefanos Zafeiriou. Meshgan: Non-linear 3d morphable models of faces. *CoRR*, abs/1903.10384, 2019. URL http://arxiv.org/abs/1903.10384.

[13] F. R. K. Chung. *Spectral Graph Theory*. American Mathematical Society, 1997.

[14] Michaël Defferrard, Xavier Bresson, and Pierre Vandergheynst. Convolutional neural networks on graphs with fast localized spectral filtering. *CoRR*, abs/1606.09375, 2016. URL http://arxiv.org/abs/1606.09375.

[15] Liuhao Ge, Zhou Ren, Yuncheng Li, Zehao Xue, Yingying Wang, Jianfei Cai, and Junsong Yuan. 3d hand shape and pose estimation from a single RGB image. *CoRR*, abs/1903.00812, 2019. URL http://arxiv.org/abs/1903.00812.

[16] Riza Alp Guler and Iasonas Kokkinos. Holopose: Holistic 3d human reconstruction in-the-wild. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[17] Riza Alp Güler, George Trigeorgis, Epameinondas Antonakos, Patrick Snape, Stefanos Zafeiriou, and Iasonas Kokkinos. Densereg: Fully convolutional dense shape regression in-the-wild. *CoRR*, abs/1612.01202, 2016. URL http://arxiv.org/abs/1612.01202.

[18] Riza Alp Güler, Natalia Neverova, and Iasonas Kokkinos. Densepose: Dense human pose estimation in the wild. *CoRR*, abs/1802.00434, 2018. URL http://arxiv.org/abs/1802.00434.

[19] Yana Hasson, Gül Varol, Dimitrios Tzionas, Igor Kalevatykh, Michael J. Black, Ivan Laptev, and Cordelia Schmid. Learning joint reconstruction of hands and manipulated objects. *CoRR*, abs/1904.05767, 2019. URL http://arxiv.org/abs/1904.05767.

[20] Gao Huang, Zhuang Liu, and Kilian Q. Weinberger. Densely connected convolutional networks. *CoRR*, abs/1608.06993, 2016. URL http://arxiv.org/abs/1608.06993.

[21] Hanbyul Joo, Tomas Simon, and Yaser Sheikh. Total capture: A 3d deformation model for tracking faces, hands, and bodies. *CoRR*, abs/1801.01615, 2018. URL http://arxiv.org/abs/1801.01615.

[22] Angjoo Kanazawa, Michael J. Black, David W. Jacobs, and Jitendra Malik. End-to-end recovery of human shape and pose. *CoRR*, abs/1712.06584, 2017.

[23] 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. *CoRR*, abs/1701.02468, 2017.

[24] Ron Levie, Federico Monti, Xavier Bresson, and Michael M. Bronstein. Cayleynets: Graph convolutional neural networks with complex rational spectral filters. *CoRR*, abs/1705.07664, 2017. URL http://arxiv.org/abs/1705.07664.

[25] Isaak Lim, Alexander Dielen, Marcel Campen, and Leif Kobbelt. A simple approach to intrinsic correspondence learning on unstructured 3d meshes. *CoRR*, abs/1809.06664, 2018. URL http://arxiv.org/abs/1809.06664.

[26] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. SMPL: A skinned multi-person linear model. *ACM Trans. Graphics (Proc. SIGGRAPH Asia)*, 34(6):248:1–248:16, October 2015.

[27] J. Masci, D. Boscaini, M. M. Bronstein, and P. Vandergheynst. Geodesic convolutional neural networks on riemannian manifolds. In *2015 IEEE International Conference on Computer Vision Workshop (ICCVW)*, pages 832–840, Dec 2015. doi: 10.1109/ICCVW.2015.112.

[28] Federico Monti, Davide Boscaini, Jonathan Masci, Emanuele Rodolà, Jan Svoboda, and Michael M. Bronstein. Geometric deep learning on graphs and manifolds using mixture model cnns. *CoRR*, abs/1611.08402, 2016. URL http://arxiv.org/abs/1611.08402.

[29] Mohamed Omran, Christoph Lassner, Gerard Pons-Moll, Peter V. Gehler, and Bernt Schiele. Neural body fitting: Unifying deep learning and model-based human pose and shape estimation. *CoRR*, abs/1808.05942, 2018. URL http://arxiv.org/abs/1808.05942.

[30] Georgios Pavlakos, Luyang Zhu, Xiaowei Zhou, and Kostas Daniilidis. Learning to estimate 3d human pose and shape from a single color image. *CoRR*, abs/1805.04092, 2018. URL http://arxiv.org/abs/1805.04092.

[31] Anurag Ranjan, Timo Bolkart, Soubhik Sanyal, and Michael J. Black. Generating 3d faces using convolutional mesh autoencoders. *CoRR*, abs/1807.10267, 2018. URL http://arxiv.org/abs/1807.10267.

[32] Javier Romero, Dimitrios Tzionas, and Michael J. Black. Embodied hands: Modeling and capturing hands and bodies together. *ACM Transactions on Graphics, (Proc. SIGGRAPH Asia)*, 36(6):245:1–245:17, November 2017. URL http://doi.acm.org/10.1145/3130800.3130883. (*) Two first authors contributed equally.

[33] David I. Shuman, Sunil K. Narang, Pascal Frossard, Antonio Ortega, and Pierre Vandergheynst. Signal processing on graphs: Extending high-dimensional data analysis to networks and other irregular data domains. *CoRR*, abs/1211.0053, 2012. URL http://arxiv.org/abs/1211.0053.

[34] Tomas Simon, Hanbyul Joo, Iain A. Matthews, and Yaser Sheikh. Hand keypoint detection in single images using multiview bootstrapping. *CoRR*, abs/1704.07809, 2017. URL http://arxiv.org/abs/1704.07809.

[35] Adrian Spurr, Jie Song, Seonwook Park, and Otmar Hilliges. Cross-modal deep variational hand pose estimation. *CoRR*, abs/1803.11404, 2018. URL http://arxiv.org/abs/1803.11404.

[36] Vince Tan, Ignas Budvytis, and Roberto Cipolla. Indirect deep structured learning for 3d human body shape and pose prediction. *BMVC*, 2017.

[37] Gül Varol, Duygu Ceylan, Bryan Russell, Jimei Yang, Ersin Yumer, Ivan Laptev, and Cordelia Schmid. Bodynet: Volumetric inference of 3d human body shapes. *CoRR*, abs/1804.04875, 2018. URL http://arxiv.org/abs/1804.04875.

[38] Nitika Verma, Edmond Boyer, and Jakob Verbeek. Dynamic filters in graph convolutional networks. *CoRR*, abs/1706.05206, 2017. URL http://arxiv.org/abs/1706.05206.

[39] Chung-Yi Weng, Brian Curless, and Ira Kemelmacher-Shlizerman. Photo wake-up: 3d character animation from a single photo. *CoRR*, abs/1812.02246, 2018. URL http://arxiv.org/abs/1812.02246.

[40] Aaron Wetzler, Ron Slossberg, and Ron Kimmel. Rule of thumb: Deep derotation for improved fingertip detection. *CoRR*, abs/1507.05726, 2015. URL http://arxiv.org/abs/1507.05726.

[41] Donglai Xiang, Hanbyul Joo, and Yaser Sheikh. Monocular total capture: Posing face, body, and hands in the wild. *CoRR*, abs/1812.01598, 2018. URL http://arxiv.org/abs/1812.01598.

[42] Rex Ying, Jiaxuan You, Christopher Morris, Xiang Ren, William L. Hamilton, and Jure Leskovec. Hierarchical graph representation learning with differentiable pooling. *CoRR*, abs/1806.08804, 2018. URL http://arxiv.org/abs/1806.08804.

[43] Xiong Zhang, Qiang Li, Wenbo Zhang, and Wen Zheng. End-to-end hand mesh recovery from a monocular RGB image. *CoRR*, abs/1902.09305, 2019. URL http://arxiv.org/abs/1902.09305.

[44] Yuxiang Zhou, Jiankang Deng, Irene Kotsia, and Stefanos Zafeiriou. Dense 3d face decoding over 2500fps: Joint texture & shape convolutional mesh decoders. *CoRR*, abs/1904.03525, 2019. URL http://arxiv.org/abs/1904.03525.

[45] Christian Zimmermann and Thomas Brox. Learning to estimate 3d hand pose from single RGB images. *CoRR*, abs/1705.01389, 2017. URL http://arxiv.org/abs/1705.01389.