Single Image 3D Hand Reconstruction with Mesh Convolutions

Dominik Kulon[1, 3] Haoyang Wang[1, 3] Riza Alp Güler[1, 3] Michael Bronstein[1, 2] Stefanos Zafeiriou[1, 3]

[1]Imperial College London [2]USI, Lugano [3]Ariel AI

IDEA

Monocular 3D reconstruction of deformable objects, such as human body parts, has been typically approached by predicting parameters of heavyweight linear models.

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 images are predicted using an image encoder.

CONTRIBUTIONS

- In order to create the high-quality training data, we build a new high-resolution model of the hand.
- We train an autoencoder on sampled meshes. By reusing the decoder, we obtain **the first graph morphable model of a highly-articulated object**.
- We train an image encoder and take advantage of the pre-trained mesh decoder to **recover 3D hand meshes aligned with the image**.
- The resulting system is able to generate hand models in **real time** and outcompetes the baseline method on the mesh reconstruction task.
- We make the code, data, and network weights **publicly available**.

GRAPH MORPHABLE MODEL

Defferrard et al. [1] expressed convolution in Fourier space where **the filter is parameterized using the Chebyshev expansion** of order r-1. This approach is computationally fast and **filters are localized with r-hops support**.

We train an autoencoder [2] on sampled meshes with removed scale variations and global orientation. We start with an input mesh with 7,907 vertices followed by sequence of four Chebyshev convolutions of order r=3 and **downsampling after each convolutional layer**. Afterwards, we apply a fully connected layer to obtain a latent vector with 64 parameters. The decoder is symmetric.

Our **Graph Morphable Model is the decoder** with 64 latent parameters**.** 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 baseline model and **constraints the space of valid poses**.

TRAINING DATA

MANO [3] is the hand model where the subject-specific shape and deformations due to pose are expressed with linear bases. The number of vertices in MANO (778) is not sufficient to realistically model shape deformations of the human hand. **We introduce a similarly designed hand model with 7,907 vertices to generate the following training data**:

- **Graph Morphable Model. We fit the model to around thousand scans to compute a distribution of valid pose and shape parameters.** 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, for each 15 joint angles in the human hand kinematic tree, we compute euler-angle clusters via K-means.
- **Mesh Recovery System.** We optimize model parameters to match the 3D annotations from the Panoptic Studio dataset.

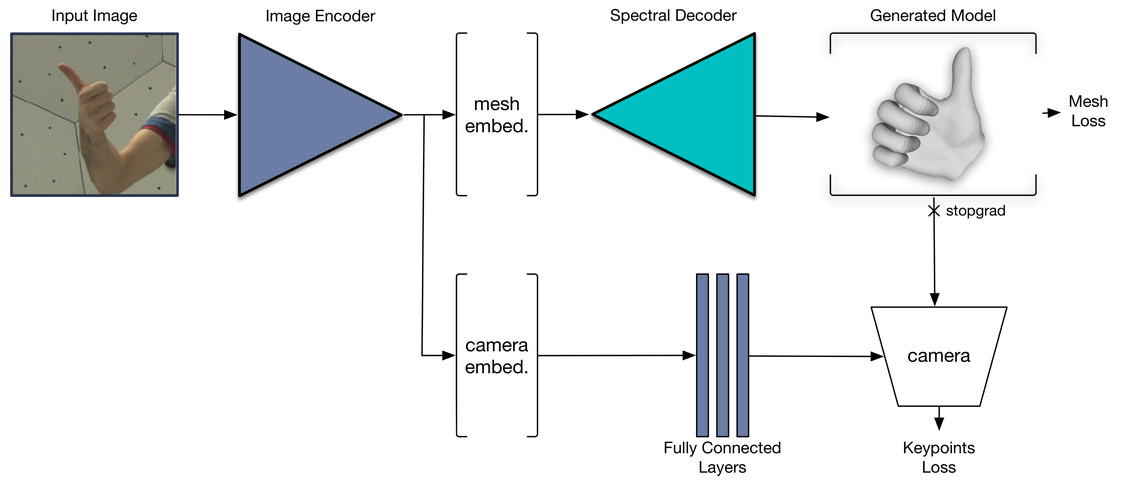
REFERENCES

1. Michaël Defferrard, Xavier Bresson, and Pierre Vandergheynst. Convolutional neural networks on graphs with fast localized spectral filtering., 2016.
2. Anurag Ranjan, Timo Bolkart, Soubhik Sanyal, and Michael J. Black. Generating 3D faces using convolutional mesh autoencoders., 2018.
3. Javier Romero, Dimitrios Tzionas, and Michael J. Black. Embodied hands: Modeling and capturing hands and bodies together., 2017.

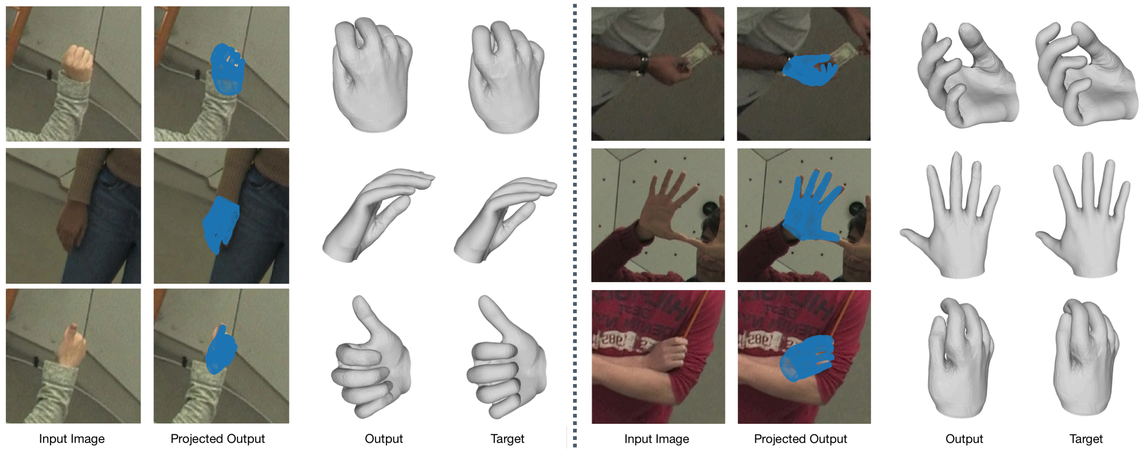
Imperial College London

EPSRC Engineering and Physical Sciences Research Council

SINGLE IMAGE MESH GENERATION



EVALUATION



	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

RESOURCES

https://arxiv.org/abs/1905.01326

https://github.com/dkulon/hand-reconstruction

CONTACT

dominik.kulon@outlook.com

http://dkulon.com