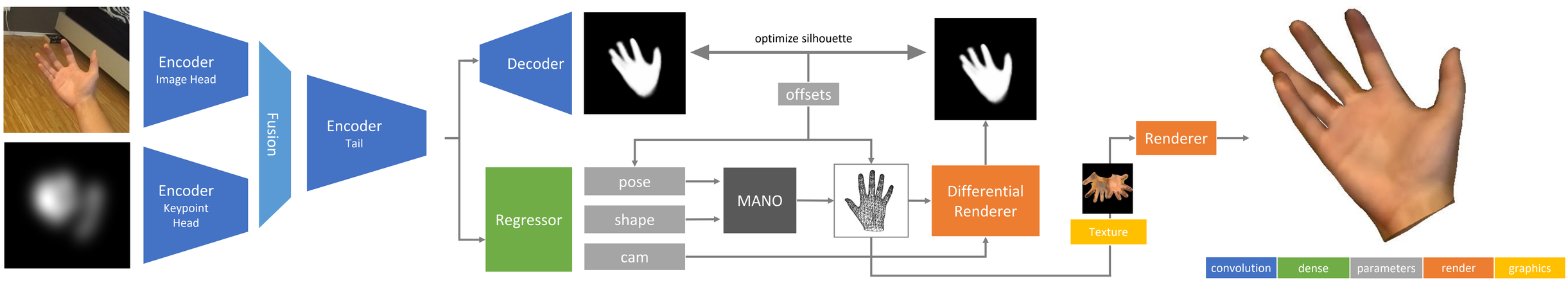
RealisticHands: A Hybrid Model for 3D Hand Reconstruction

Michael Seeber[1], Roi Poranne[2], Marc Polleyfeys[1,4] Martin R. Oswald[1,3]
[1]ETH Zurich, [2]University of Haifa, [3]University of Amsterdam, [4]Microsoft

REPLICATING THE USER'S PERSONAL HANDS IN 3D FOR BETTER EMBODIMENT

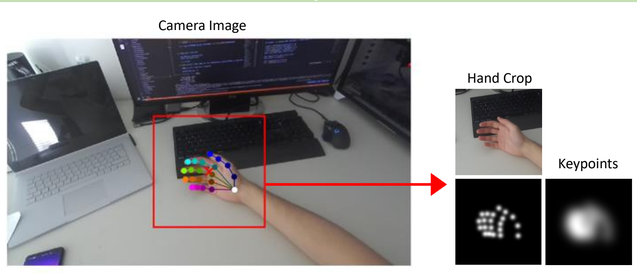
Summary: A novel hybrid approach for 3D hand reconstruction that combines the performance of learning based methods with the accurate image-model alignment of optimization based methods.



A deep encoder-decoder network regresses MANO & camera parameters as well as a segmentation mask. The hand mesh is refined via differential rendering to optimize the silhouette consistency.

Pipeline

Input



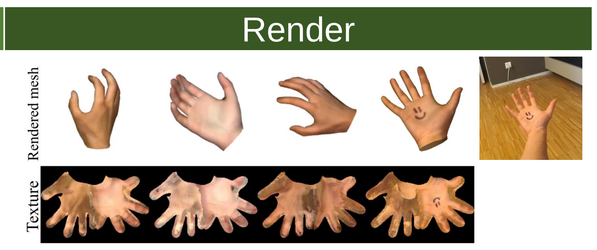
CNN

Problem: Fast, efficient and robust prediction of MANO model parameters.
Idea: Use a deep Encoder-Decoder with multiple fused input modalities.

Test time optimization

Problem: Minor errors accumulated along the kinematic chain.
Idea: Iteratively refine the hand mesh by adding offsets $\Delta p, \Delta s$ and Δv for the pose and shape and vertices respectively.

Render



Motivation & Challenges

Related Work:
- 3D hand pose estimation:
- 3D hand reconstruction:

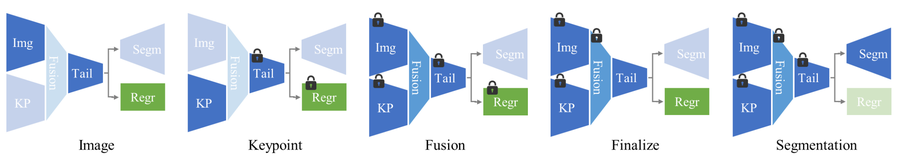
optimization ↔ learning
model based ↔ model free

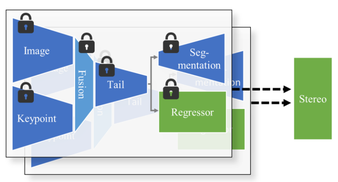
hybrid approach

Challenges:
- *High DoF*: hands are highly articulated with multiple degrees of freedom
- *Self Similarity*: hard to distinguish between fingers due to similar appearance
- *Occlusion*: especially the egocentric setting comes with lot of self-occlusion
- *Variation* of hands: human hands exhibit a lot of variation in size, shape etc.
- *Fast motion:* Fast moving fingers lead to motion blur & dilution of contours
- *Limited data:* Most datasets on hand shapes include only a few participants.

Training Framework

In our training scheme selective components are frozen / removed to maximize performance of individual network branches.
→ Significant performance increase *(Table 1)*



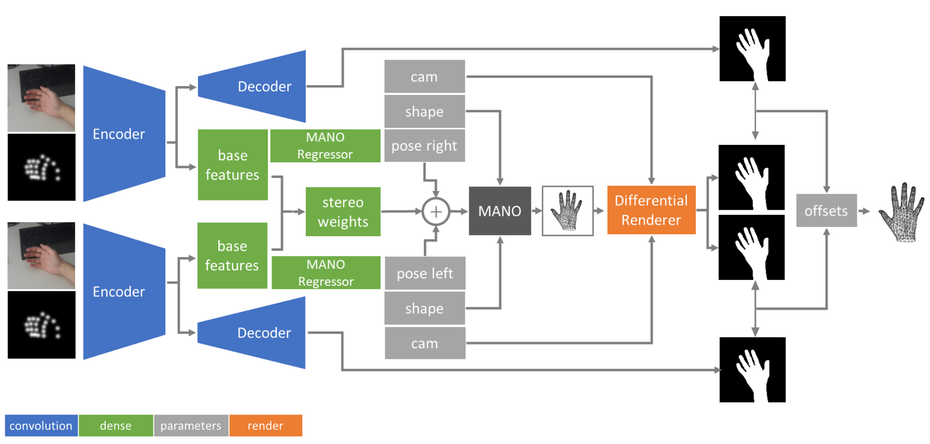


	MPVE↓		F@5mm↑		F@15mm↑	
	no KP	with KP	no KP	with KP	no KP	with KP
single-stage	1.79	1.16	0.31	0.49	0.82	0.92
multi-stage	**0.85**	**0.78**	**0.45**	**0.66**	**0.96**	**0.97**

Table 1: The multi-stage training outperforms a joint training scheme that tends to emphasize the dominant branch.

Stereo Extension

Estimation a 3D hand pose from a single RGB image is an ill-posed problem due to depth and scale ambiguities. We regress stereo weights $\boldsymbol{w}$ that are used to fuse the predicted poses from the left and right in a meaningful way.



Results

MANO Regression

We outperform baselines, such as the mean shape and inverse kinematic fits, and also rank higher than other model based approaches such as [1, 2, 3, 4]. Further, we perform quantitative similarly to I2L [5], a model-free approach that occassionaly produces collapsed results. The qualitative superiority of our methods over I2L is also well demonstrated in the supplementary video.

Methods	MPVE (PA)↓	F@5mm (PA)↑	F@5mm (PA)↑	Model based	GT scale
Mean shape	1.64	0.336	0.837	✓	
Inverse Kinematics [1]	1.37	0.439	0.892	✓	
Hasson *et al.* [2]	1.33	0.429	0.907	✓	✓
Boukhayma *et al.* [3]	1.32	0.427	0.894	✓	✓
Mano CNN [1]	1.09	0.516	0.934	✓	✓
Kulon *et al.* [4]	0.86	0.614	0.966	✗	✗
I2L [5]	0.76	0.681	0.973	✗	✗
Ours (Mono, MP Hands)	0.97	0.575	0.949	✓	✗
Ours (Mono, PoseNet(I2L))	0.78	0.662	0.971	✓	✗

Table 2: Quantitative comparison of our approach with other methods on the task of monocular hand pose and shape estimation.

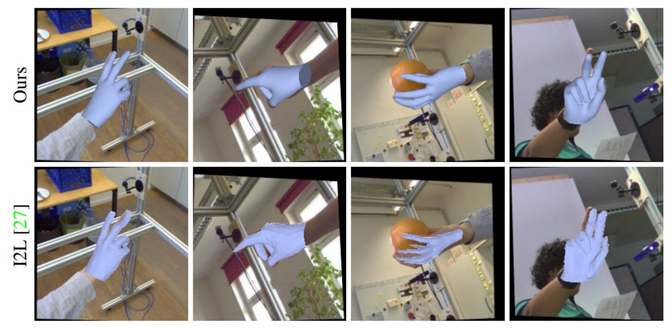


Figure 1: Qualitative results on the FreiHAND test set in comparision to I2L [5].

Segmentation

We compared the segmentation performance to various baseline such as HSV, CNN [6] and UMA [6].

	Combined		FreiHAND		Synth	
	IoU↑	Accuracy↑	IoU↑	Accuracy↑	IoU↑	Accuracy↑
HSV	0.26	0.81	0.33	0.90	0.18	0.72
CNN [6]	0.85	0.98	0.77	0.97	0.93	0.99
UMA [6]	0.86	0.98	0.78	0.98	0.93	0.99
Ours	**0.88**	**0.99**	**0.80**	**0.98**	**0.95**	**0.99**

Table 3: Quantitative segmentation results on the FreiHAND and Synthetic Stereo dataset.

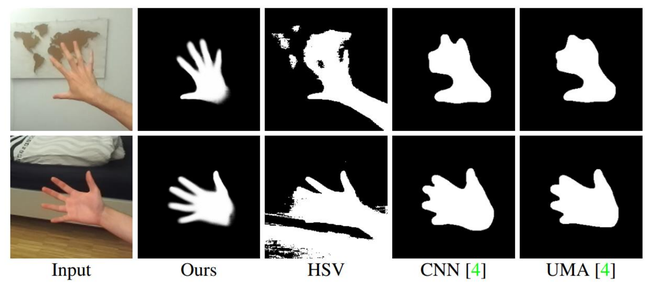


Figure 2: Qualitative segmentation results showing our method performs best.v

Test Time Optimization

To evaluate the benefit of the optimization stage, we compare the IoU of projected silhouettes with the segmentation mask.

	IoU↑	
	MP Hands	PoseNet
no refinement	0.677	0.722
3 iterations	0.706	0.758
10 iterations	0.735	0.793
15 iterations	**0.748**	**0.806**
I2L [27]		0.737

Figure 3: The test time optimization improves the image-model alignment.

Stereo Extension

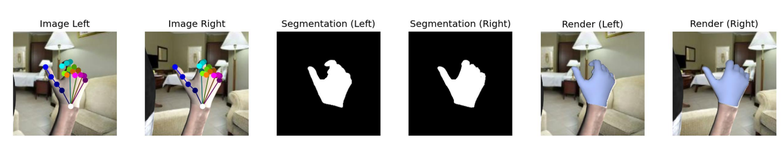
The stereo version greatly outperforms the monocular, showing that multiple views can resolve ambiguities.

Version	MPVE (PA)↓	F@5mm (PA)↑	F@15mm (PA)↑
Monocular	1.65	0.352	0.860
Stereo	**0.94**	**0.578**	**0.967**

Figure 4: Quantitative comparison results of monocular architecture and stereo architecture.

Synthetic Stereo Hands Dataset

Stereo hand datasets from an egocentric viewpoint are lacking.
→ We generated a large scale synthetic dataset



References

[1] Christian Zimmermann, Duygu Ceylan, Jimei Yang, Bryan Russell, Max Argus, and Thomas Brox. Freihand: A dataset for markerless capture of hand pose and shape from single rgb images.
[2] Yana Hasson, Gul Varol, Dimitris Tzionas, Igor Kalevatykh, Michael J. Black, Ivan Laptev, and Cordelia Schmid. Learning joint reconstruction of hands and manipulated objects.
[3] Adnane Boukhayma, Rodrigo de Bem, and Philip HS Torr. 3d hand shape and pose from images in the wild.
[4] Dominik Kulon, Riza Alp Guler, Iasonas Kokkinos, Michael M Bronstein, and Stefanos Zafeiriou. Weaklysupervised mesh-convolutional hand reconstruction in the wild.
[5] Gyeongsik Moon and Kyoung Mu Lee. I2l-meshnet: Imageto-lixel prediction network for accurate 3d human pose and mesh estimation from a single rgb image.
[6] Minjie Cai, Feng Lu, and Yoichi Sato. Generalizing hand segmentation in egocentric videos with uncertainty-guided model adaptation.