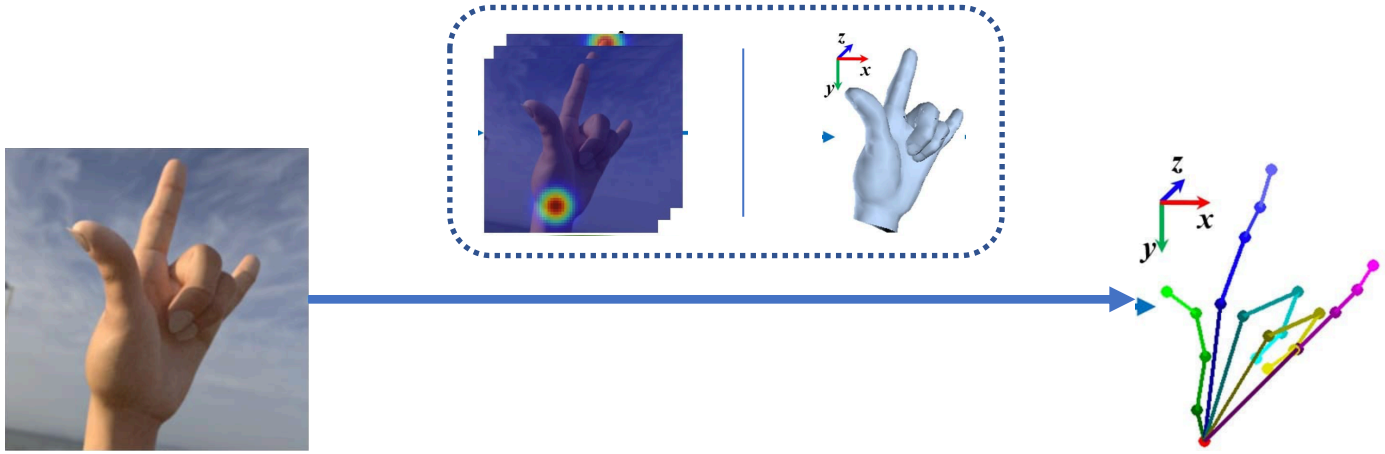
3D hand pose estimation from a single RGB



张蒙豪

2020/04/10

Challenges for hand pose estimation

- Occlusion
- Lack of characteristic local features
- Strong articulation
- Relatively low resolution

Challenges for hand pose estimation from a single RGB

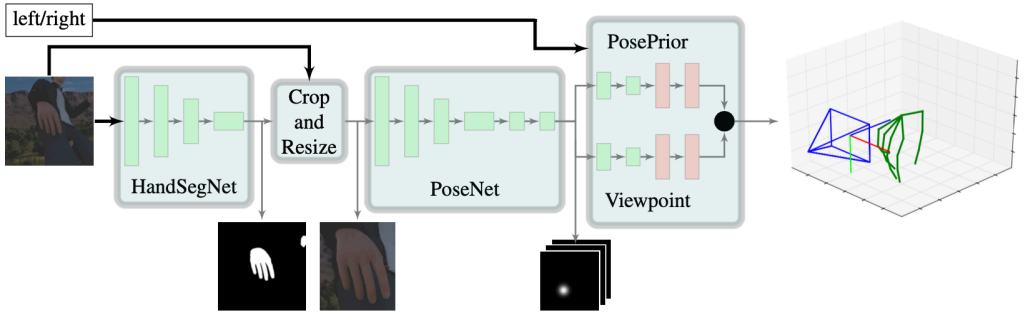
- Depth ambiguity (ill-posed problem)
- Dataset(2d/3d annotation, domain gap)

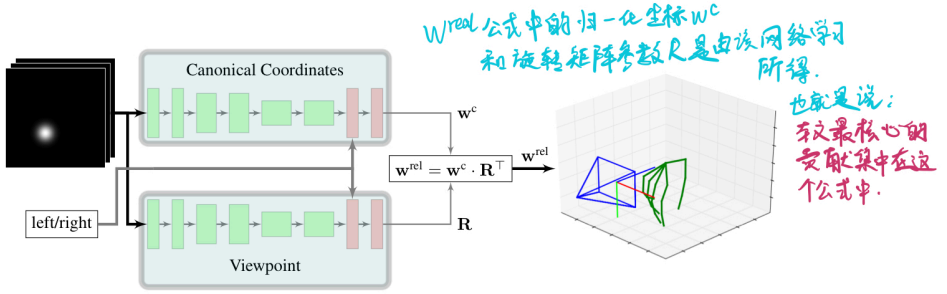
Paper list

- Learning to Estimate 3D Hand Pose from Single RGB Images(ICCV 2017)
- Weakly-supervised 3D Hand Pose Estimation from Monocular RGB Images(ECCV 2018)
- 3D Hand Shape and Pose from Images in the Wild(CVPR 2019)
- 3D Hand Shape and Pose Estimation from a Single RGB Image(CVPR 2019)
- Monocular Real-time Hand Shape and Motion Capture using Multi-modal Data(CVPR 2020)

First try

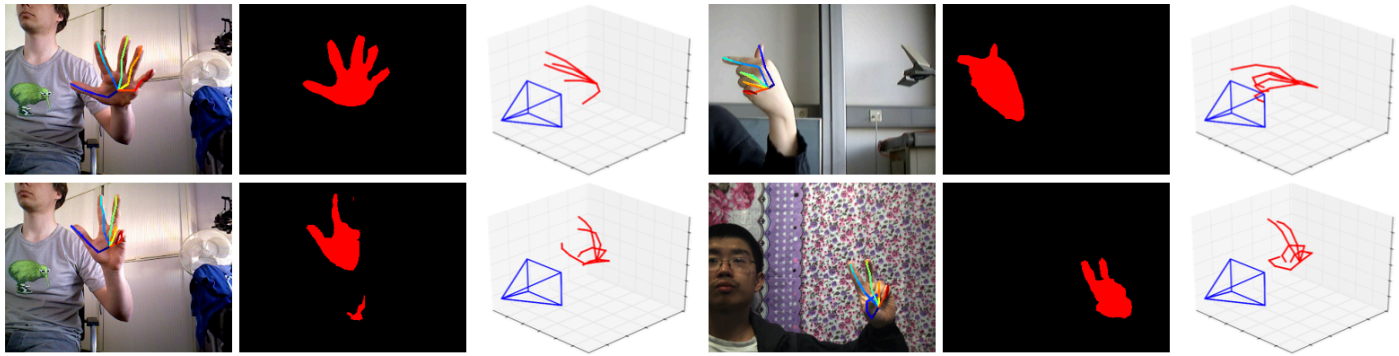
- A new dataset(RHD)
- A baseline method. (first deep learning shot in this field)

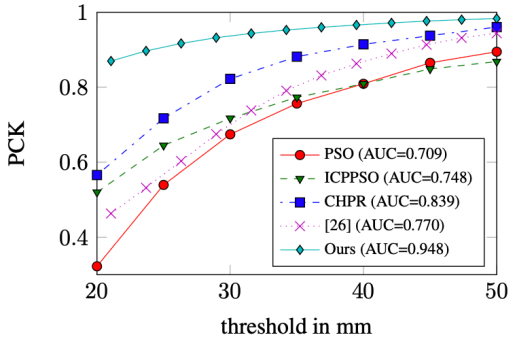




First try

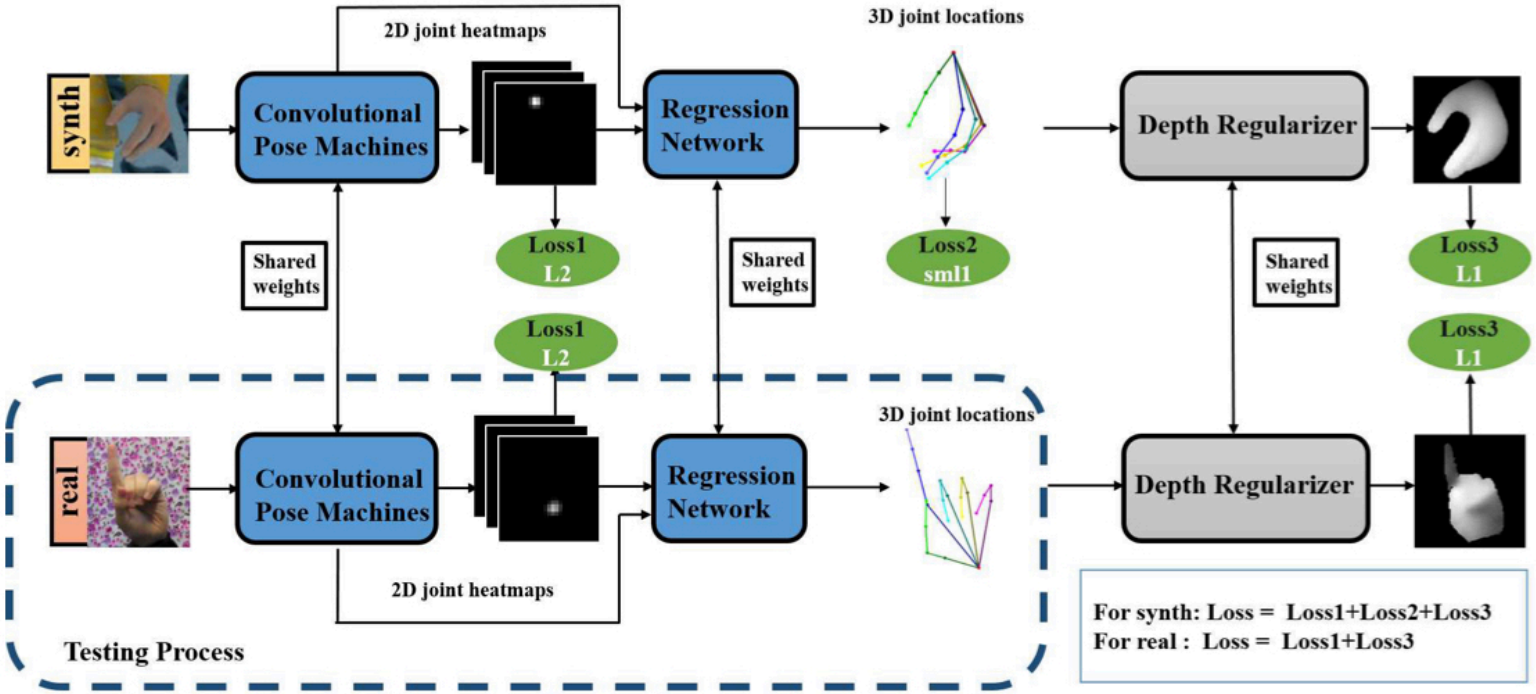
- Results:



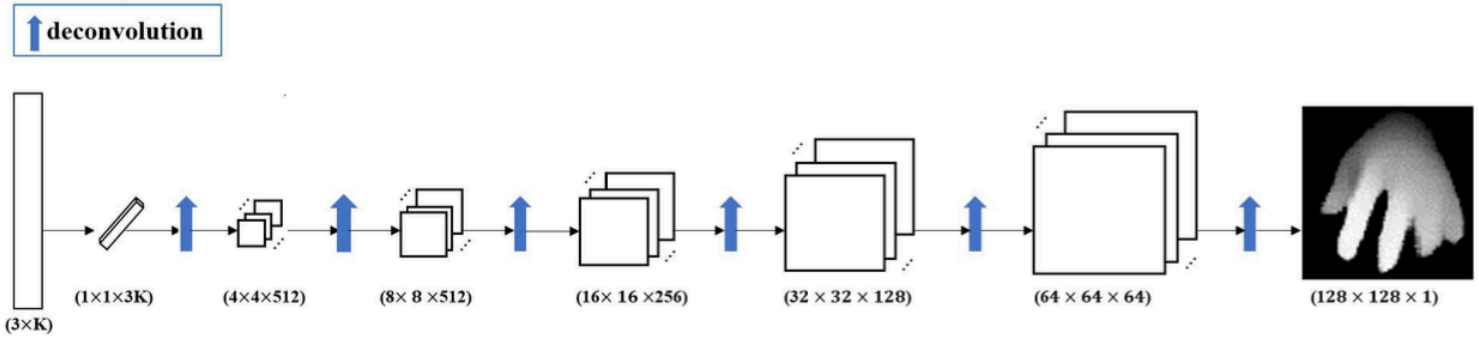


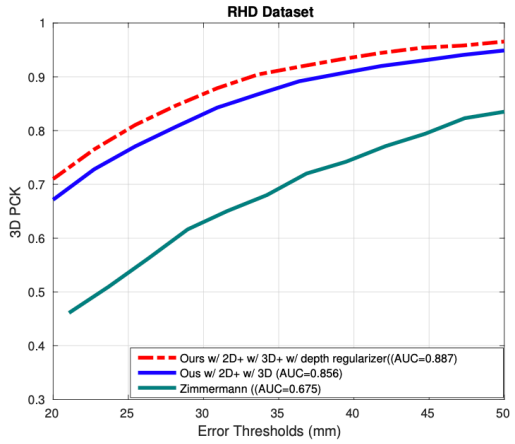
Adopt depth as weakly supervise

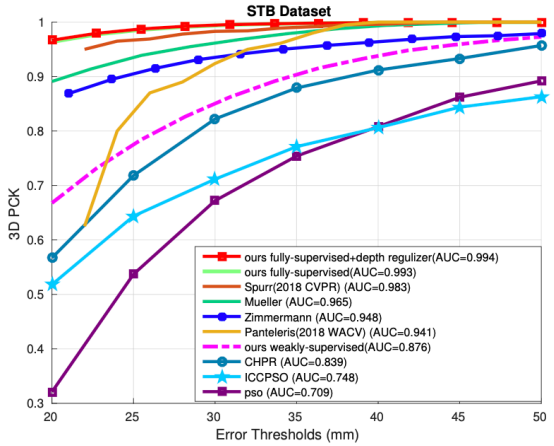
- If 3d annotated, full supervise learning
- If not, weakly supervise learning

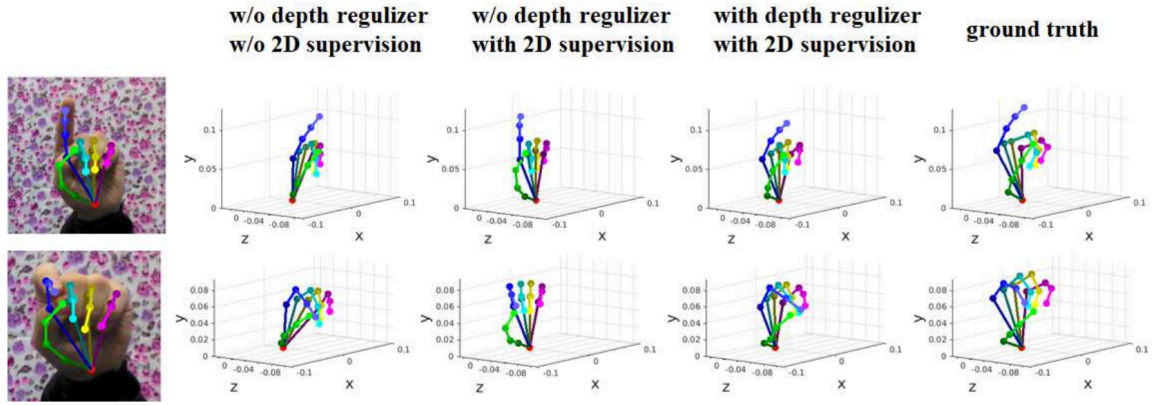


Adopt depth as weakly supervise





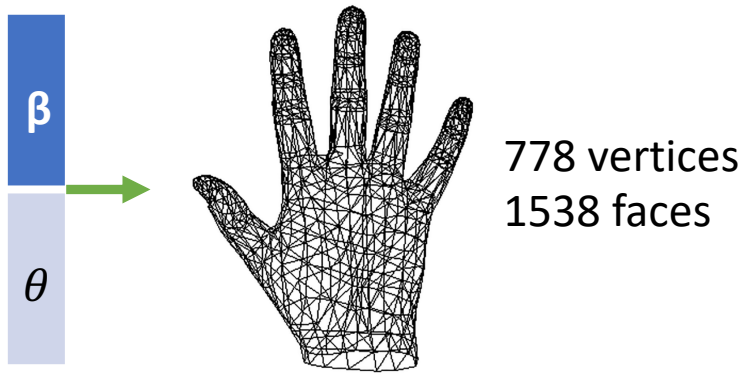




Adopt MANO model

- What is MANO model?
  - A hand model
  - Shape para β (10)
  - Pose para θ (3+K*3)
  - Shape blend func S
  - Pose blend func P
  - Joints regressor J

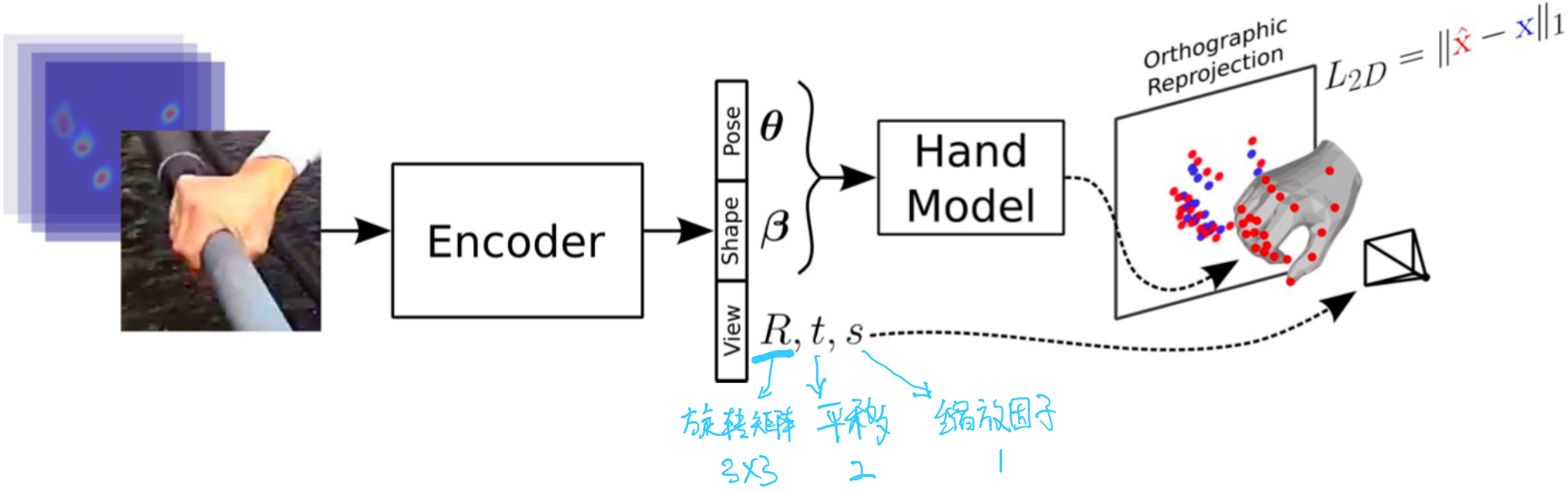
  - Given β,θ, you can get a mesh.



Embodied hands: modeling and capturing hands and bodies together; Javier Romero , Dimitrios Tzionas , Michael J. Black; ACM Transactions on Graphics (ToG), 2017

Adopt MANO model(CVPR 2019)

- A good prior
- Reconstruction first, then 2d/3d pose
- End2end



Adopt MANO model (CVPR 2019)

- First to pre train the encoder(a self made pre train dataset)
- Then train on public dataset(PANOPTIC with only 2d anno, STB with 3d anno)

We combine multiple losses to train our pipeline: A 2D joint re-projection loss L_{2D}, a 3D joint loss L_{3D}, a hand mask loss L_{mask} and a model parameter regularization loss L_{reg}.

$$L = L_{2D} + \alpha_{3D} L_{3D} + \alpha_{mask} L_{mask} + \alpha_{reg} L_{reg}, \quad (5)$$

where $\alpha_{3D} = 10^2$, $\alpha_{mask} = 10^2$ and $\alpha_{reg} = 10^1$ are weighting factors.

3D Hand Shape and Pose from Images in the Wild; Adnane Boukhayma, Rodrigo de Bem, Philip H.S. Torr

Adopt MANO model (CVPR 2019)

- Results

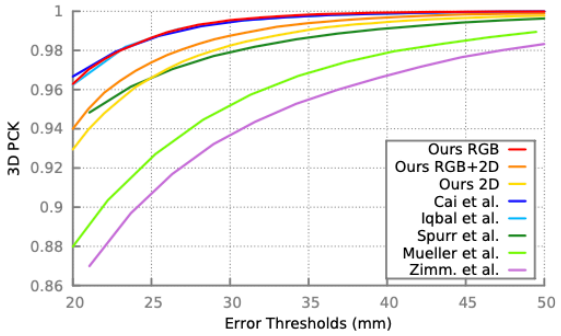


Figure 4: 3D PCK for STEREO.

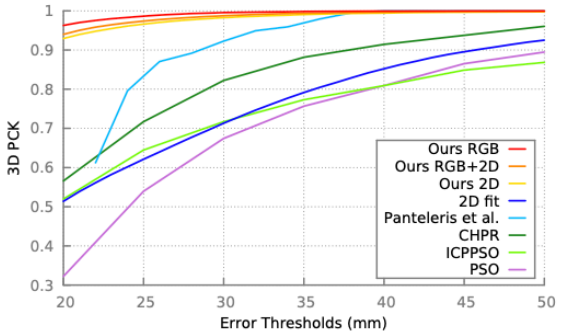


Figure 5: 3D PCK for STEREO.

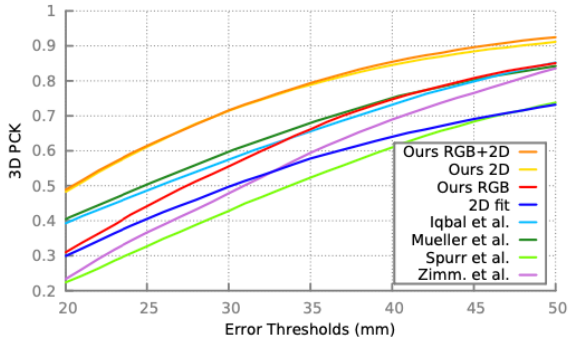


Figure 6: 3D PCK for DEXTER+OBJECT.

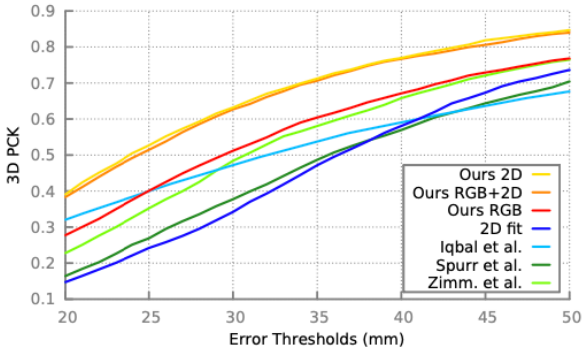


Figure 7: 3D PCK for EGODEXTER.

	Ours RGB	Ours RGB+2D	Ours 2D	2D fit
3D distance	**9.76**	10.18	10.46	23.21

Table 1: Average 3D joint distance (mm) to ground-truth for STEREO.

	Ours RGB	Ours RGB+2D	Ours 2D	2D fit	Spurr et al.	Zimm. et al.
3D distance	33.16	**25.53**	25.93	41.18	40.20	34.75

Table 2: Average 3D joint distance (mm) to ground-truth for DEXTER+OBJECT.

	Ours RGB	Ours RGB+2D	Ours 2D	2D fit	Spurr et al.	Zimm. et al.
3D distance	51.87	45.58	**45.33**	56.59	56.92	52.77

Table 3: Average 3D joint distance (mm) to ground-truth for EGODEXTER.

	Ours RGB	Ours RGB+2D	Ours 2D	2D fit	Zimm. et al.
2D distance	23.04	**18.95**	20.65	22.36	59.40

Table 4: Average re-projected 2D joint distance (px) to ground-truth for MPII+NZSL

Adopt MANO model (CVPR 2020)

- To choose one or two? No, we use them all!
- Multi-task

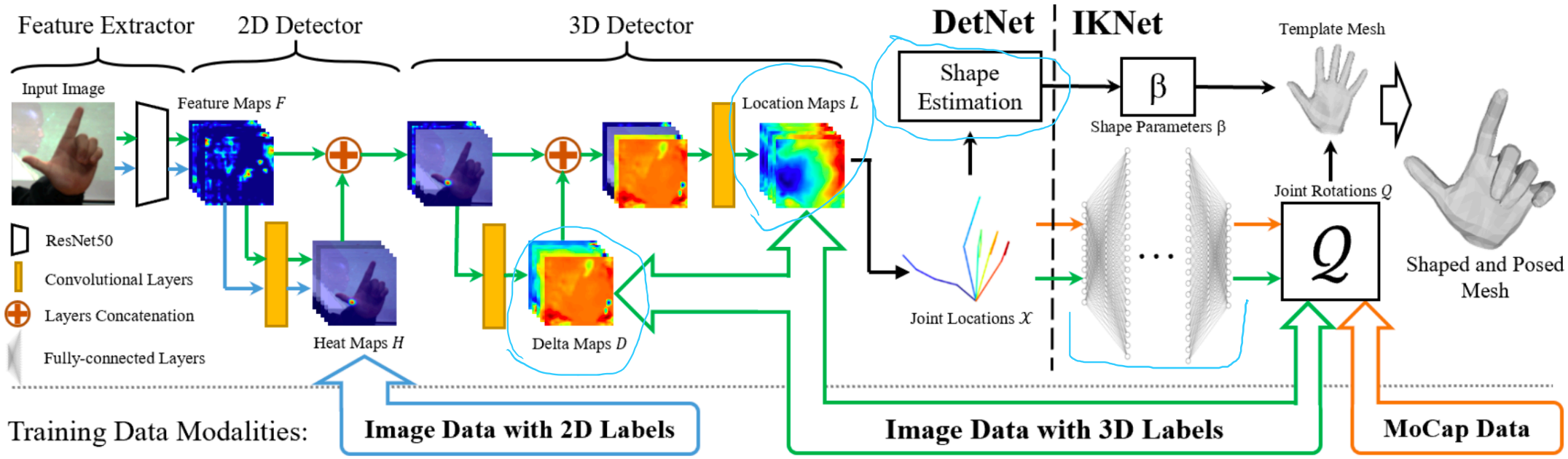


Figure 2. Overview of our architecture. It comprises two modules: first, our DetNet predicts the 2D and 3D joint positions from a single RGB image. Second, our IKNet takes the 3D joint predictions of DetNet and maps them to joint angles.

Adopt MANO model (CVPR 2020)

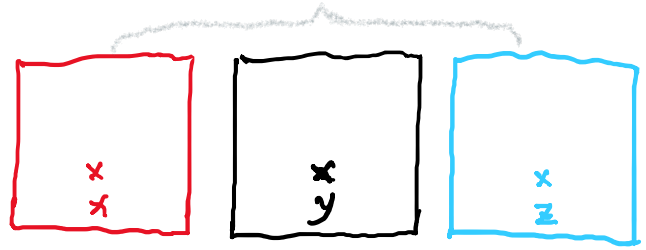
- 2D detector
  - ResNet50 extracts feature maps
  - 2 conv layers get 2d joints heatmaps hxwxJ
- 3D detector
  - 2 conv layers to get the D(Delta maps) → joints rotation R3 vector
  - 2 conv layers to get the L(Location maps) → x, y, z coord R^3 vector
- Shape estimation

para shape L2 regularizer

$$E(\beta) = \sum_b || \frac{l_b(\beta)}{l_{\text{ref}}(\beta)} - l_b^{\text{pred}} ||_2^2 + \lambda_\beta ||\beta||_2^2. \quad (8)$$

→ use traditional ML method to regress the β

for one joint (D, L)



Adopt MANO model (CVPR 2020)

- IKNet
- Data
  - MANO data & expansion
  - 3D annotated data
- Training
  - Input: $\mathcal{I} = [\mathcal{X}, \mathcal{D}, \mathcal{X}_{\text{ref}}, \mathcal{D}_{\text{ref}}] \in \mathbb{R}^{4 \times J \times 3}$,
  - Output: a quaternion $\hat{\mathcal{Q}} \in \mathbb{R}^{J \times 4}$,
  - Network: 7 layers FC with BN and sigmoid

for rest pose

$\mathcal{X}$: 3d joints position

$\mathcal{D}$: orientation of each bone

Adopt MANO model (CVPR 2020)

- Results

	Variants of our Method	AUC of PCK		
		DO	ED	STB
1)	**Ours**	**.948**	.811	**.898**
2)	w/o IKNet	.923	.804	.891
3)	w/o L_{12} and $L_{\cos}$	.933	.823	.869
4)	w/o 3DPosData	.926	.809	.873
5)	w/o L_{12}	.943	**.812**	.890
5)	w/o $L_{\cos}$	.840	.782	.808

Table 2. Ablation study. We evaluate the influence of: *2)* IKNet *3)* Direct rotational supervision on joint rotations. *4)* Weak supervision on joint rotations. *5)* Loss terms on the quaternions.

Method	AUC of PCK			
	DO	ED	STB	RHD
Ours	**.948**	**.811**	.898	.856*
Ge et al. [10]	-	-	**.998***	.920*
Zhang et al. [51]	.825	-	.995*	.901*
Yang et al. [48]	-	-	.996*	**.943***
Baek et al. [1]	.650	-	.995*	.926*
Xiang et al. [47]	.912	-	.994*	-
Boukhayma et al. [2]	.763	.674	.994*	-
Iqbal et al. [17]	.672	.543	.994*	-
Cai et al. [3]	-	-	.994*	.887*
Spurr et al. [34]	.511	-	.986*	.849*
Mueller et al. [23]	.482	-	.965*	-
Z&B [54]	.573	-	.948*	.670*

Table 1. Comparison with state-of-the-art methods on four public datasets. We use "*" to note that the model was trained on the dataset, and use "-" for those who did not report the results. Our system outperforms others by a large margin on the DO and ED dataset which we argue is the most fair comparison as none of the models are trained on these datasets. As [17] only reports results without alignment, we report the absolute values for this method.

Adopt MANO model (CVPR 2020)

- Results

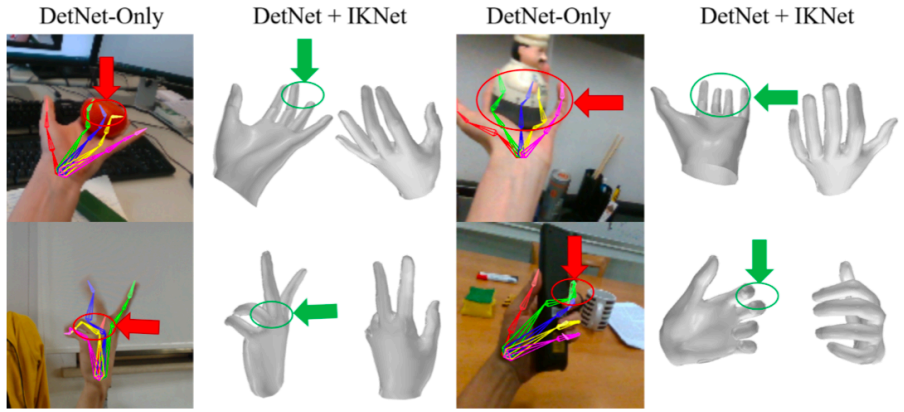


Figure 7. Our IKNet is able to compensate some errors from the DetNet based on the prior learned from the MoCap data.

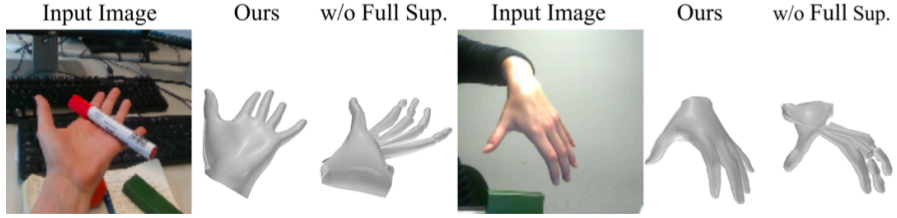


Figure 8. Comparison between IKNets with and without rotational supervision from MoCap data. Note that even though 3D joint positions match the ground truth, without this supervision unnatural poses are estimated.

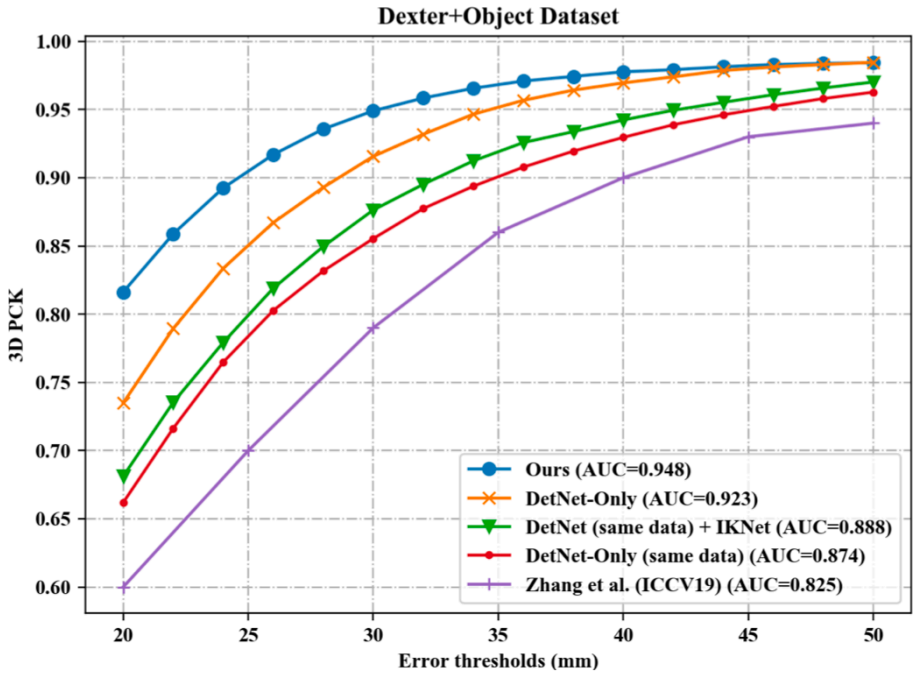
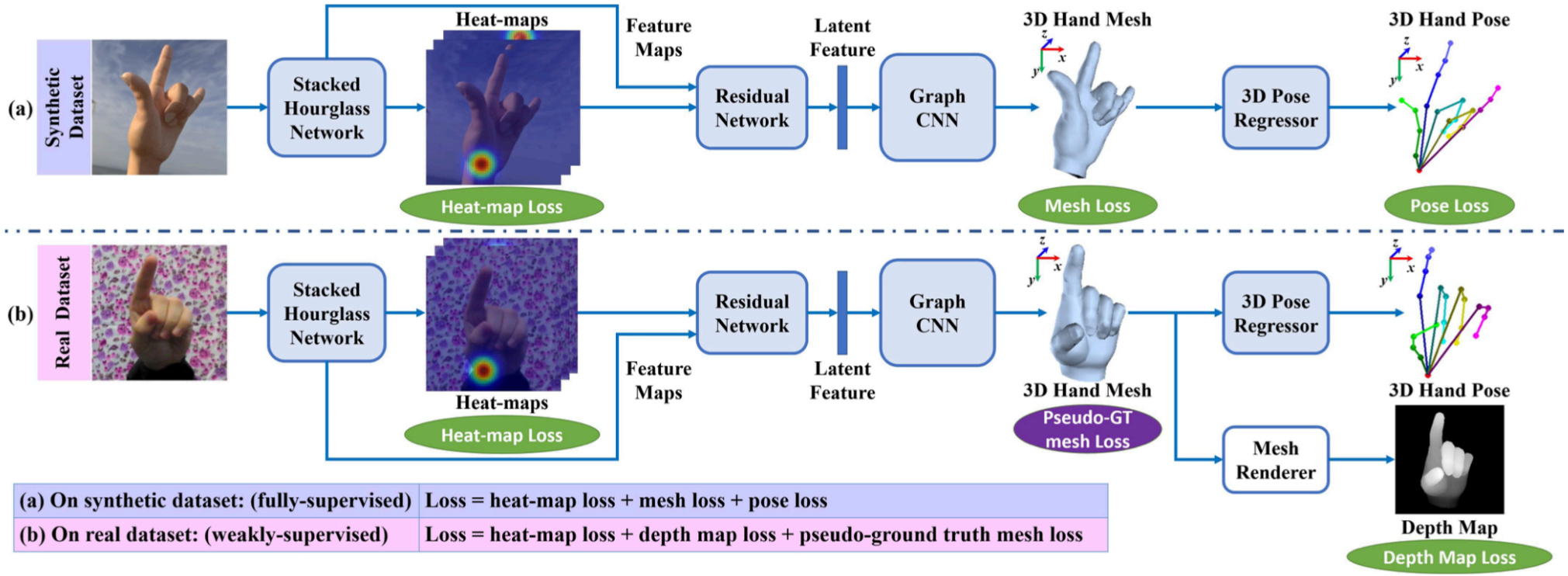


Figure 6. Ablation study for the training data on DO. We use "same data" to indicate our model trained with RHD and STB, which is the same as Zhang et al. [51]. We demonstrate that our architecture is superior to theirs by design. Integrating more data further boosts the results.

Adopt GCN

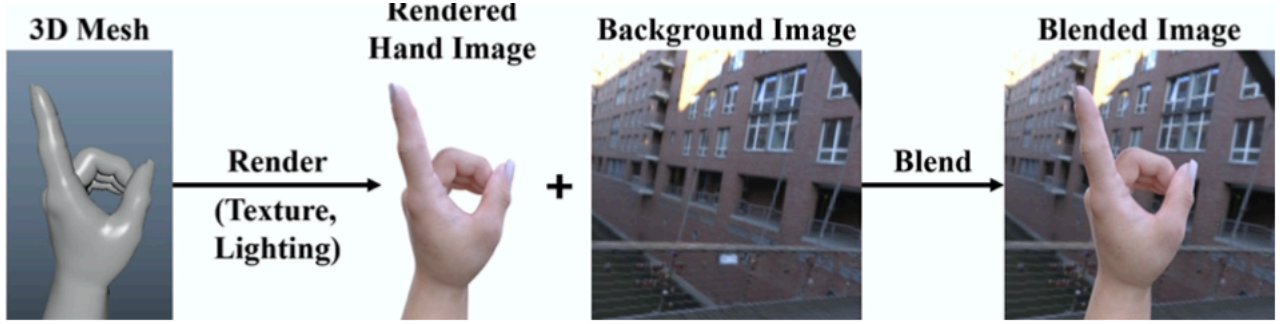
- MANO has limitation, use Graph.



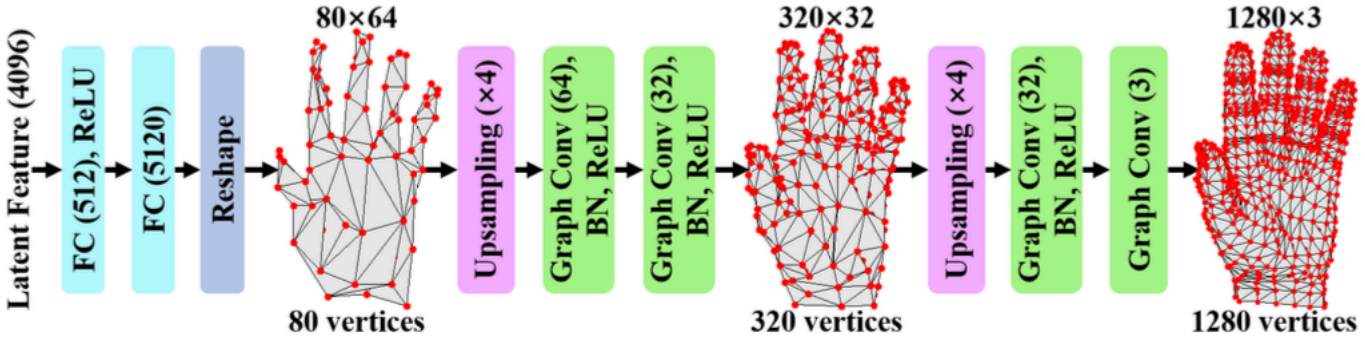
3D Hand Shape and Pose Estimation from a Single RGB Image; Liuhao Ge, Zhou Ren, Yuncheng Li, Zehao Xue, Yingying Wang, Jianfei Cai, Junsong Yuan;(CVPR 2019)

Adopt GCN

- New dataset.



- GCN



3D Hand Shape and Pose Estimation from a Single RGB Image; Liuhao Ge, Zhou Ren, Yuncheng Li, Zehao Xue, Yingying Wang, Jianfei Cai, Junsong Yuan;(CVPR 2019)

What’s next?

- Two hands in one model?
- Better way to handle occlusion(self/mutual occlusion) problem?
- New dataset?
- Novel GNN network?
- Any other prior could be used?

Thanks