GANerated Hands for Real-Time 3D Hand Tracking from Monocular RGB

Franziska Mueller, et, al. CVPR 2018

2018.11.20

20185209 Sangyoon Lee

Table of contents

Motivation

- Hand pose estimation is available in many applications.

Natural interaction

Activity recognition

Information interpretation

Challenges

- **(Self-)occlusion** and **self-similarities**
- **Hard to annotate** data in **3D**

Background (1)

- Multi view method is used to overcome occlusions.
- Many studies have used 2-8 RGB cameras to overcome this problem.
  - R. Wang, S. Paris, and J. Popovic. **6d hands: markerless hand-tracking for computer aided design.** In Proc. of UIST, pages 549-558. ACM, 2011.
  - I. Oikonomidis, N. Kyriazis, and A. A. Argyros. **Full dof tracking of a hand interacting with an object by modeling occlusions and physical** constraints. In Computer Vision (ICCV), 2011 IEEE International Conference on, pages 2088-2095. IEEE, 2011**.**

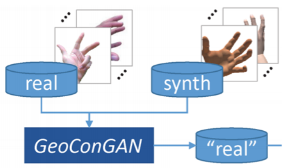
Background (2)

- Generate data set to support Learning based model.
  - J. Tompson, M. Stein, Y. Lecun, and K. Perlin. Real-time continuous pose recovery of human hands using convolutional networks. ACM Transactions on Graphics, 33, August 2014.
- Generation of synthetic hand in virtual environment.
  - F. Mueller, D. Mehta, O. Sotnychenko, S. Sridhar, D. Casas, and C. Theobalt. Real-time hand tracking under occlusion from an egocentric rgb-d sensor. In International Conference on Computer Vision (ICCV), 2017.

Contribution

- Real-time full 3D hand tracking from monocular RGB video.
- Technical Novelties

1)

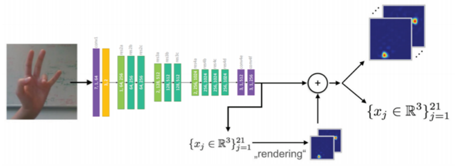


New GAN for **geometrically consistent unpaired image-to-image translation**

2)

Novel **enhanced RGB dataset** with **3D** hand joint **annotations** (>260k frames)

3)



CNN with **projection layer** for tightly coupled regression of **2D and 3D joint locations**

Solution : Hand tracking system

- Overview of the solution

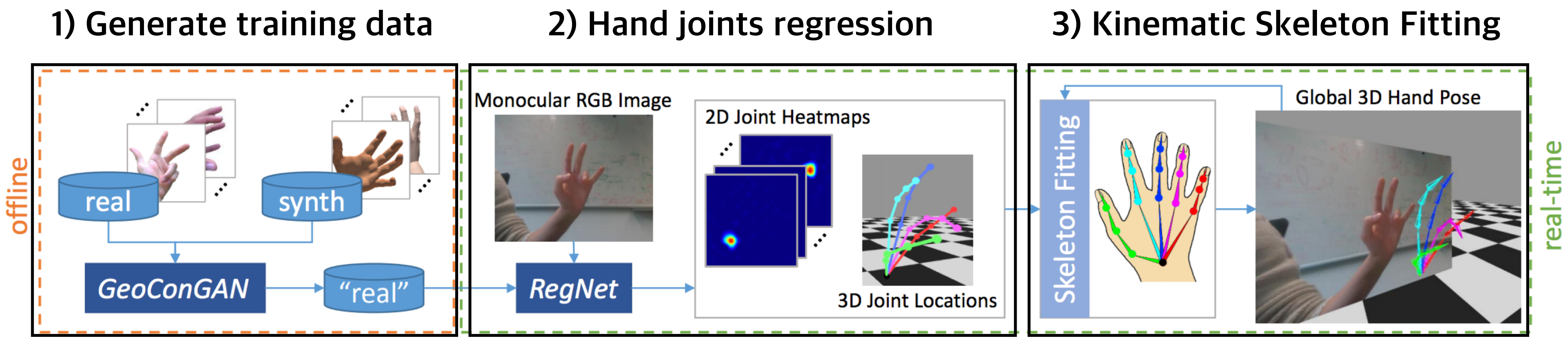


Figure 2: Pipeline of our real-time system for monocular RGB hand tracking in 3D.

Solution : Generation of Training Data

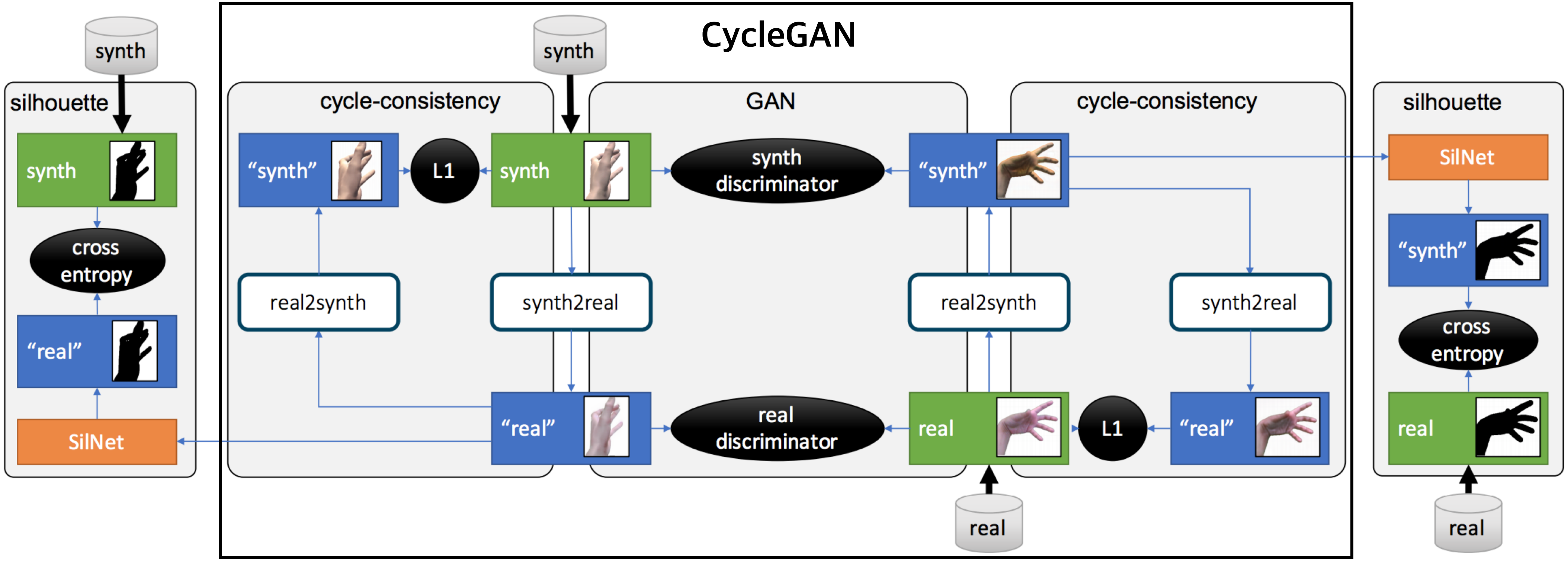
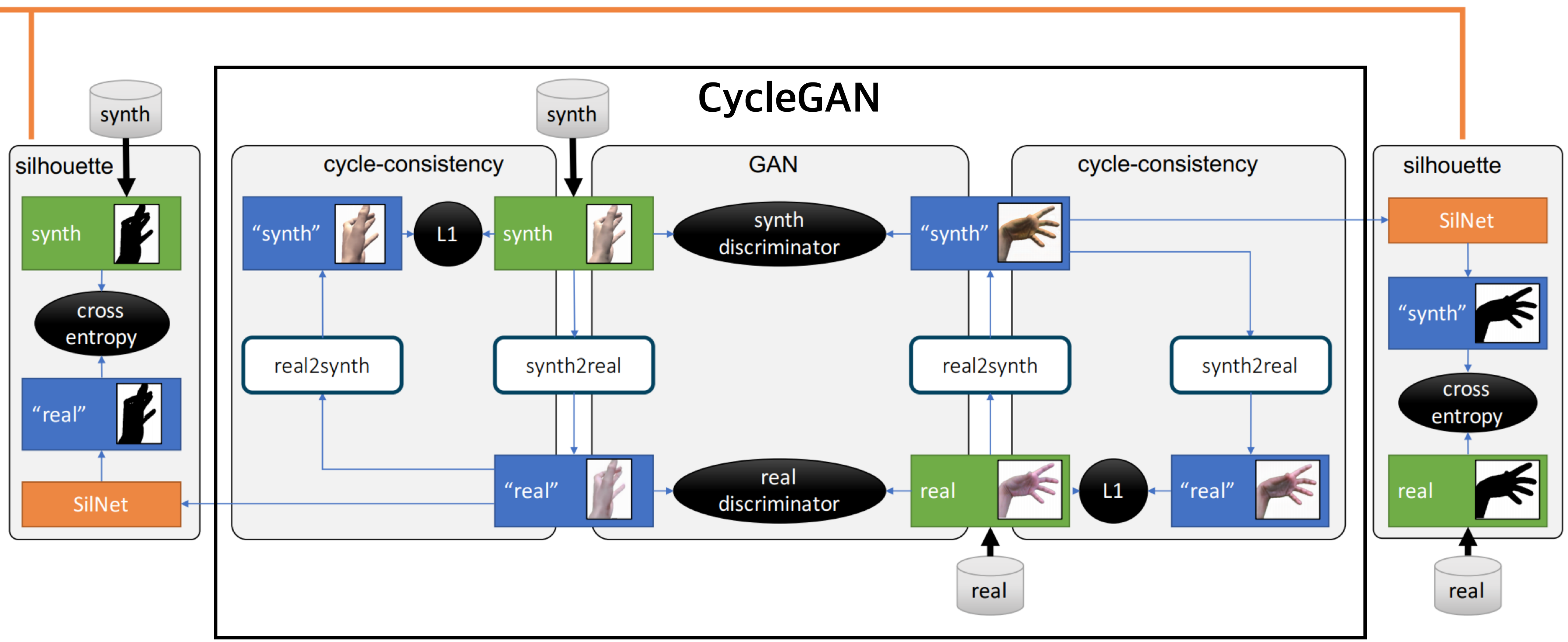


Figure 3: Network architecture of our *GeoConGAN*. The trainable part comprises the *real2synth* and the *synth2real* components, where we show both components twice for visualization purposes. The loss functions are shown in black, images from our database in green boxes, images generated by the networks in blue boxes, and the existing *SilNet* in orange boxes.

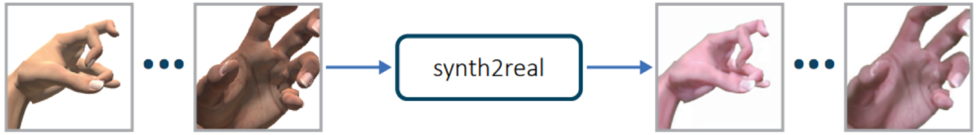
Solution : Generation of Training Data

Geometric consistency loss:

- forces hand pose to not change during translation, hence
- perfect synthetic annotation can be transferred



Solution : Generation of Training Data



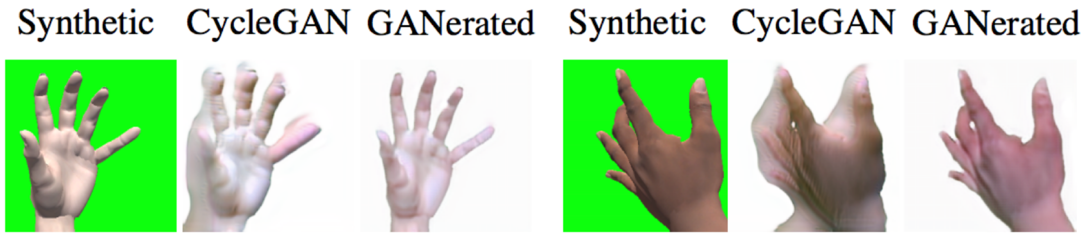


Figure 4: Our *GeoConGAN* translates from synthetic to real images by using an additional geometric consistency loss.

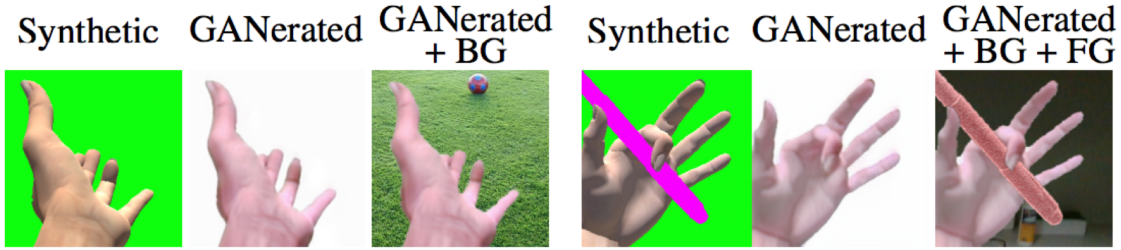


Figure 5: Two examples of synthetic images with background/object masks in green/pink.

Solution : Hand Joints Regression

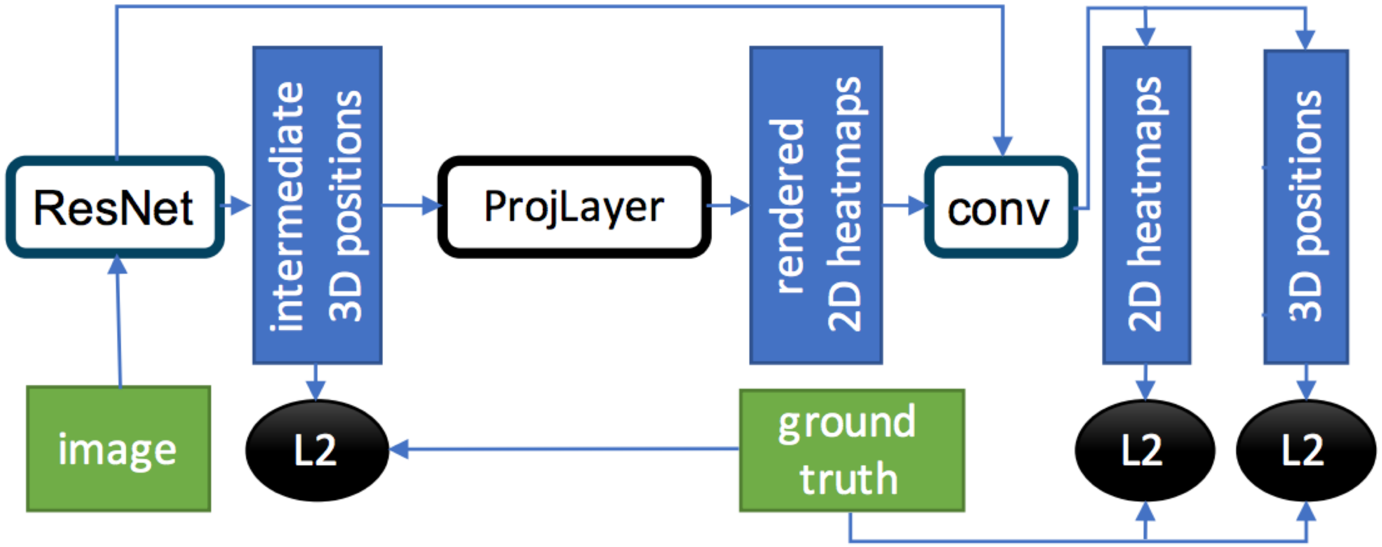
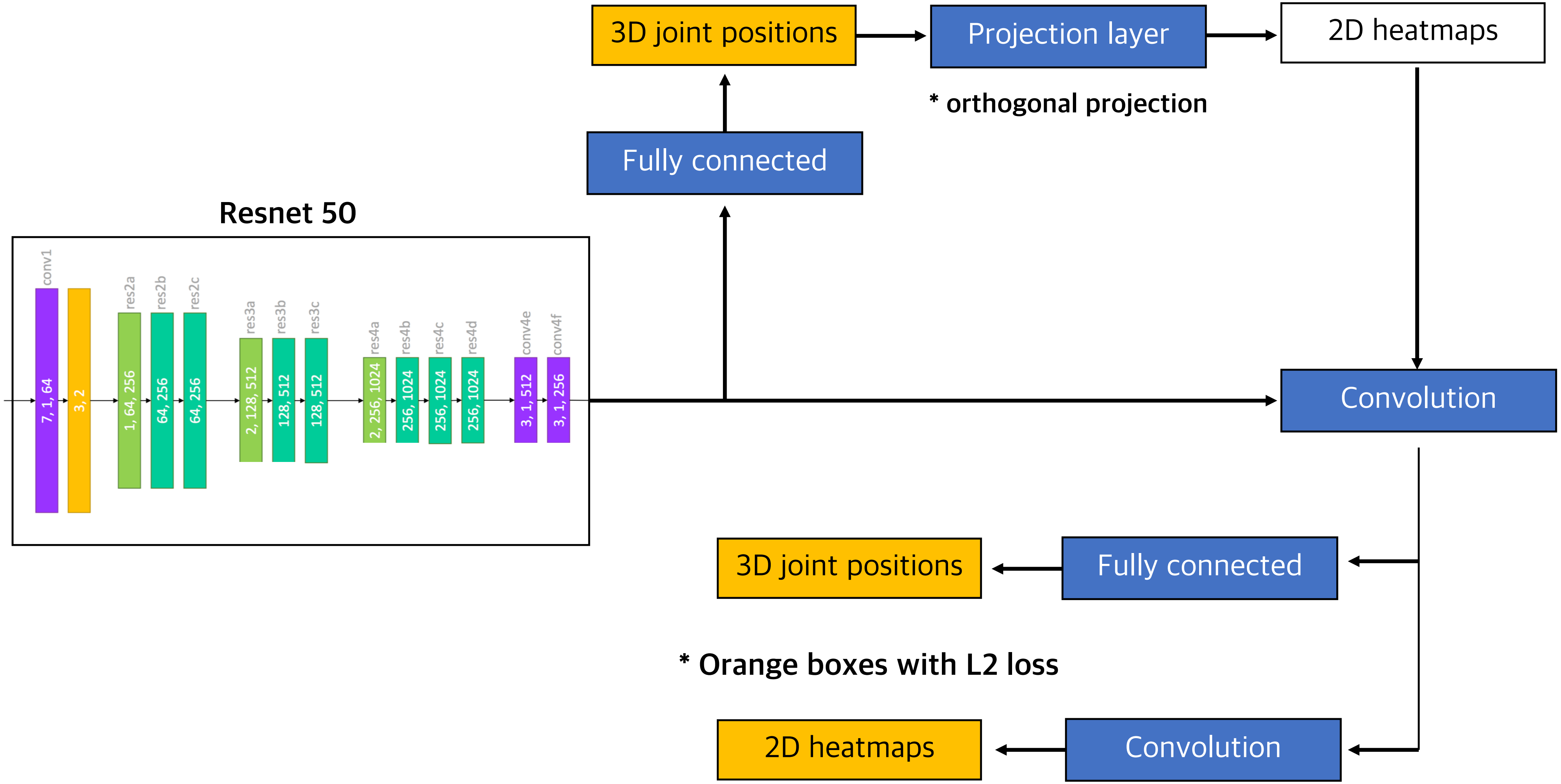


Figure 6: Architecture of *RegNet*. While only *ResNet* and *conv* are trainable, errors are still back-propagated through our *ProjLayer*. Input data is shown in green, data generated by the network in blue, and the loss is shown in black.

Solution : Hand Joints Regression



Solution : Hand Joints Regression

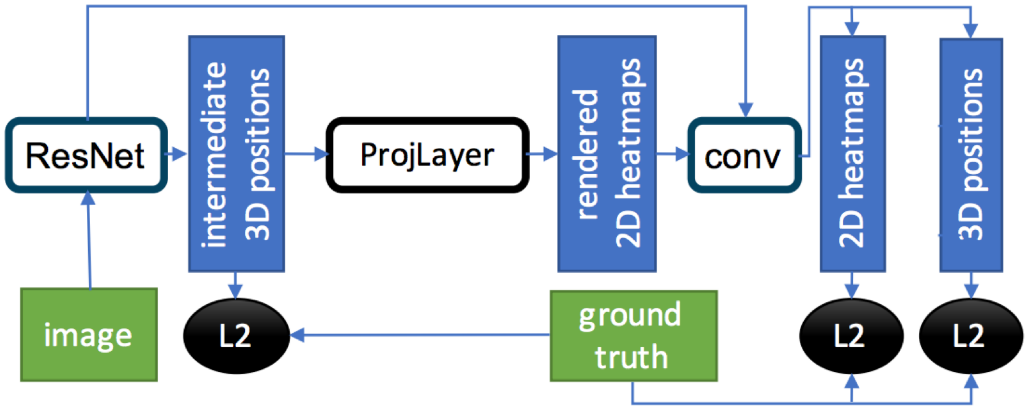
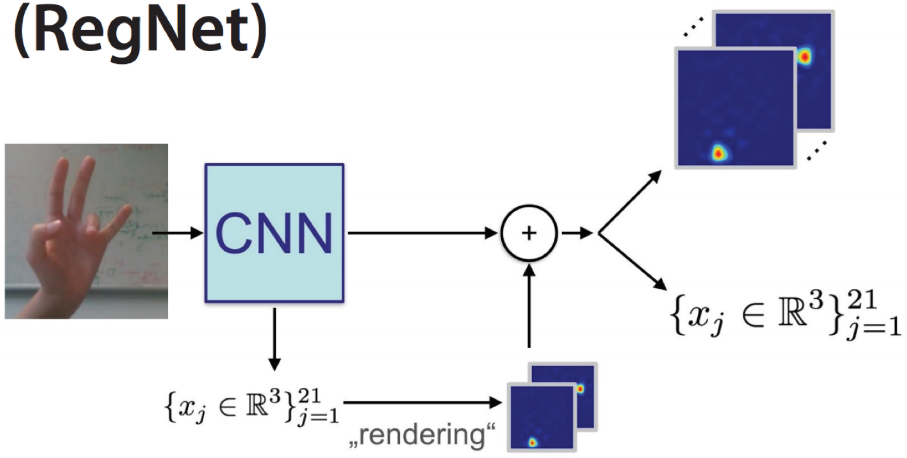


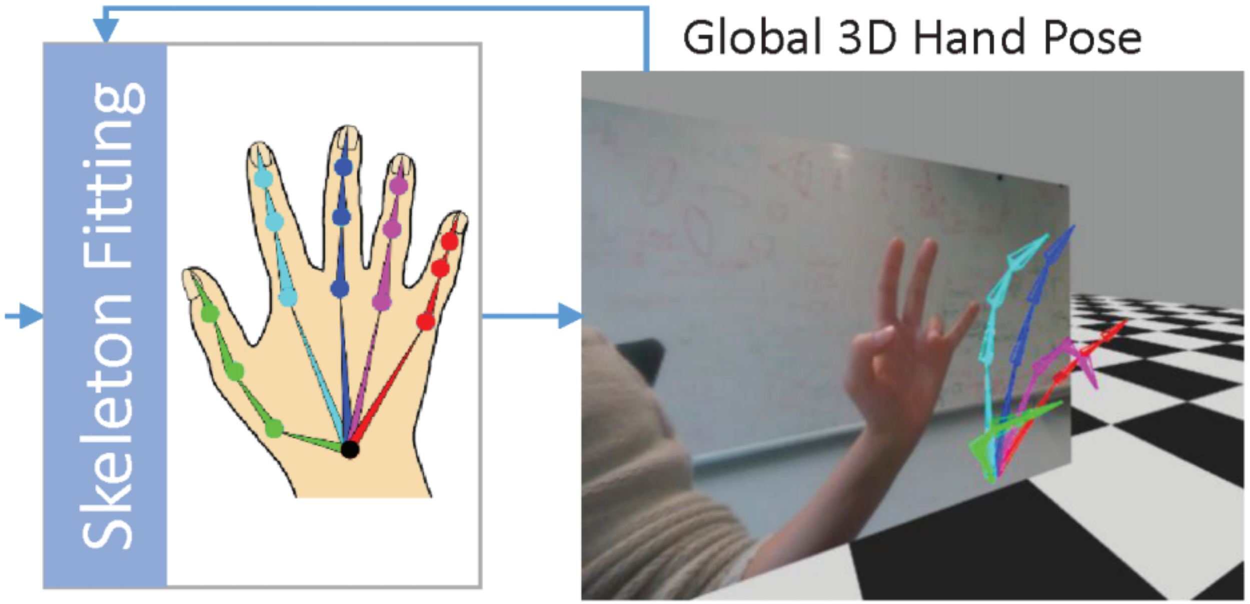
Figure 6: Architecture of *RegNet*. While only *ResNet* and *conv* are trainable, errors are still back-propagated through our *ProjLayer*. Input data is shown in green, data generated by the network in blue, and the loss is shown in black.

Joint Position Regression (RegNet)



- **2D joint location heatmaps:** determine global position
- **Root-relative 3D joint positions:** distinguish hand poses with the same 2D projection

Solution : Kinematic Skeleton Fitting



Kinematic Skeleton Fitting

Minimize energy:

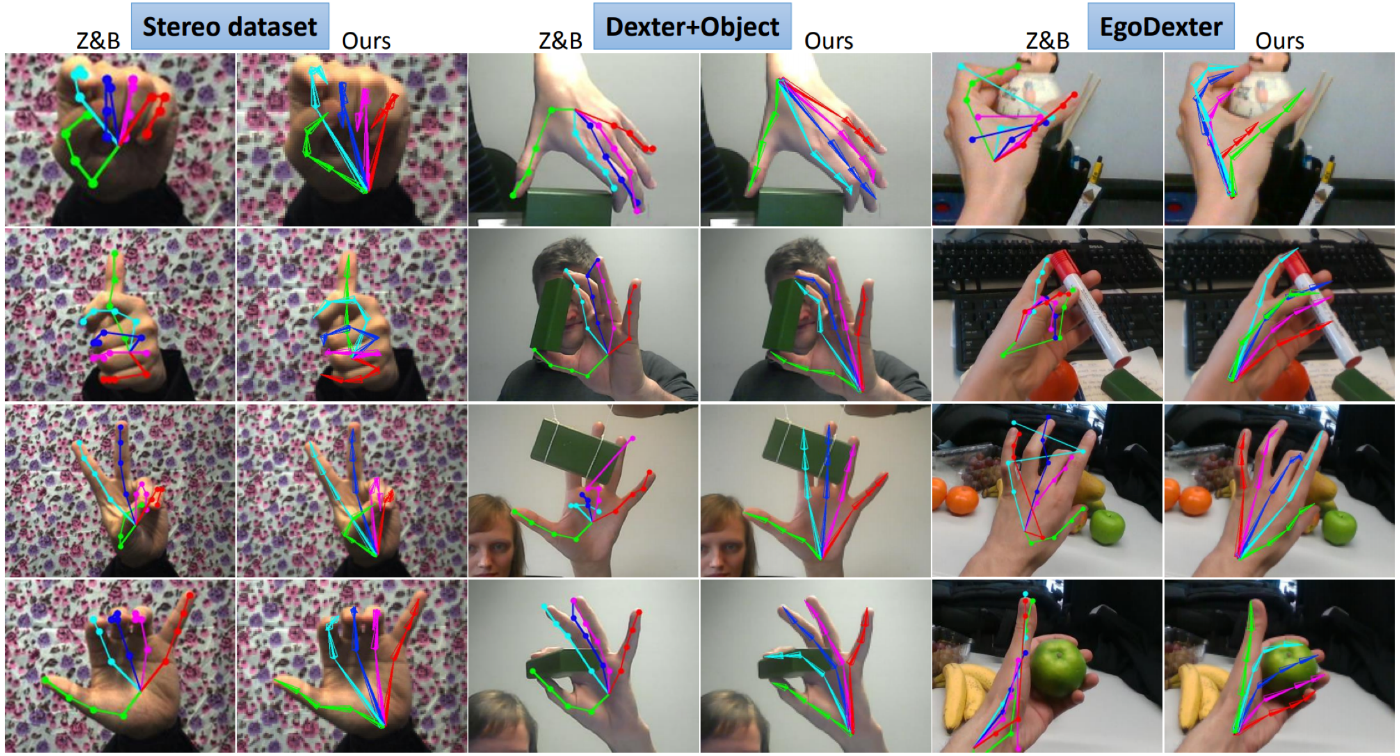
$$
\begin{aligned}
E(\Theta) = & \; E_{2D}(\Theta) && \text{2D Inverse Kinematics} \\
& + E_{3D}(\Theta) && \text{3D Inverse Kinematics} \\
& + E_{\text{limits}}(\Theta) && \text{Joint Angle Limits} \\
& + E_{\text{temp}}(\Theta) && \text{Temporal Smoothness}
\end{aligned}
$$

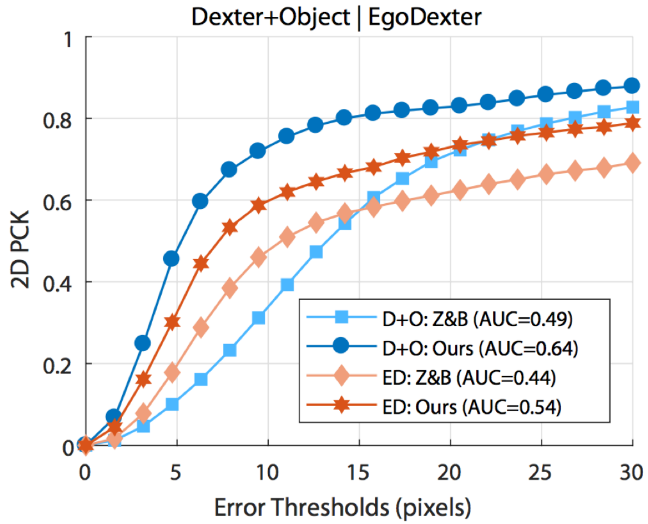
Final output: parameters of kinematic model (global transform, joint angles)

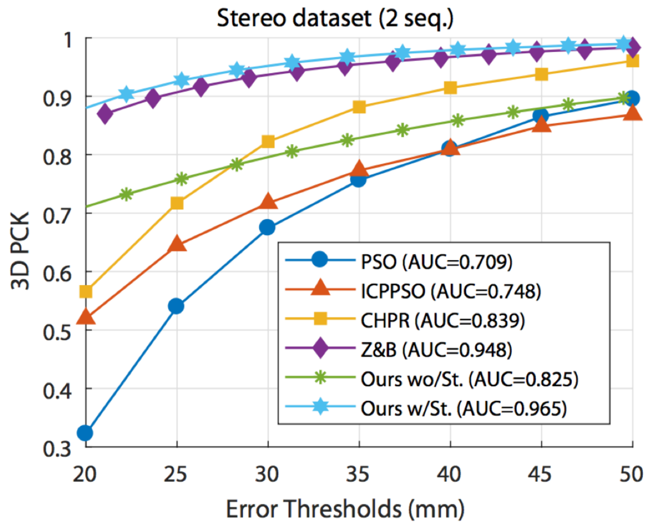
Evaluation

PCK : the Percentage of Correct Keypoints score

Comparison to Zimmermann and Brox, ICCV 2017



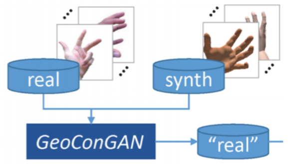




Conclusion & Summary

- Presents a more robust model for occlusions
- Presents
  - a data set similar to the real hand domain
  - a model that can create the data set
- Demonstrates these benefits in the evaluation
  - particularly in difficult occlusion scenarios.
- Summary
  - Real-time full 3D hand tracking from single monocular RGB video.
  - Technical Novelties

1)

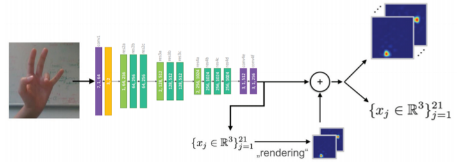


New GAN for **geometrically consistent unpaired image-to-image translation**

2)

Novel **enhanced RGB dataset** with **3D** hand joint **annotations** (>260k frames)

3)



CNN with **projection layer** for tightly coupled regression of **2D and 3D joint locations**

Q & A

- Thank you for listening

Quiz

- Q1
  - **What is the newly proposed loss function in this paper?**
    - A) Cycle Consistency
    - B) Rectangle Consistency
    - C) Triangle Consistency
    - D) Geometric consistency loss
- Q2
  - **Which of the following is not related to the contribution of this paper?**
    - A) Presents a more robust model for occlusions
    - B) Present a data set similar to the real hand domain
    - C) Presents a model that can create data similar to the real hand domain
    - D) Presents multi view method to overcome occlusions.