CS688 Student Presentation

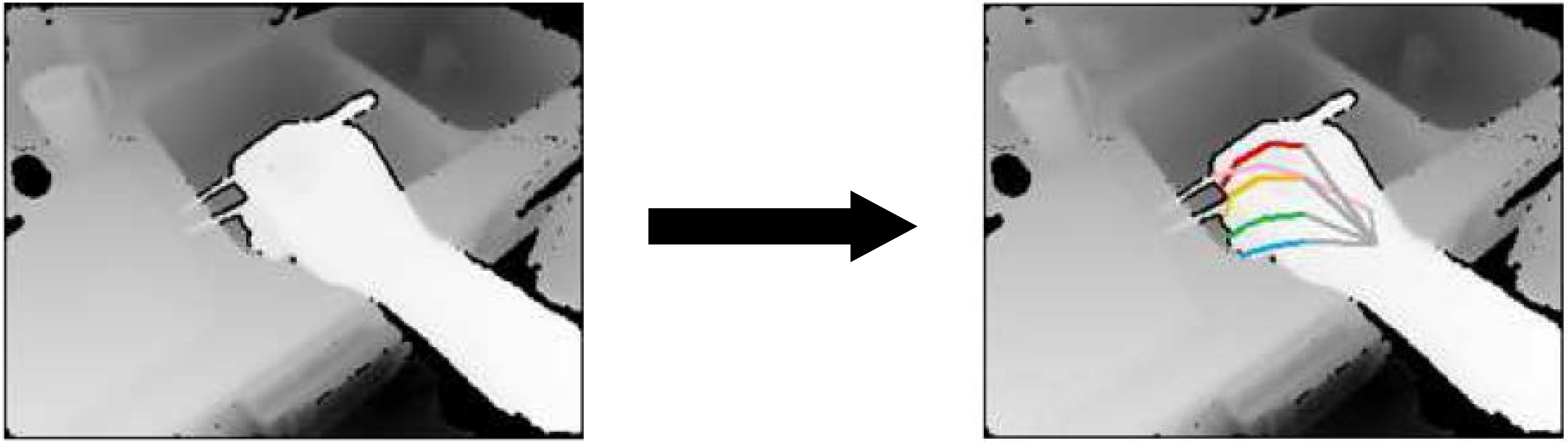
# Robust Hand Pose Estimation during the Interaction with an Unknown Object (ICCV17)

18.11.20

Youngbo Shim

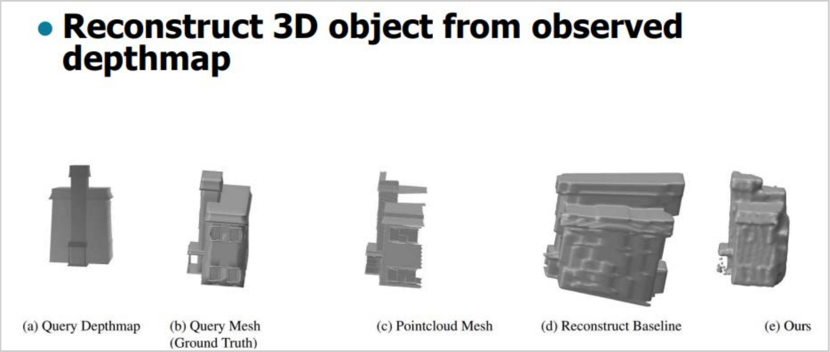
# Problem statement

- Detecting hand pose during interaction with an object
  - from a egocentric depth image

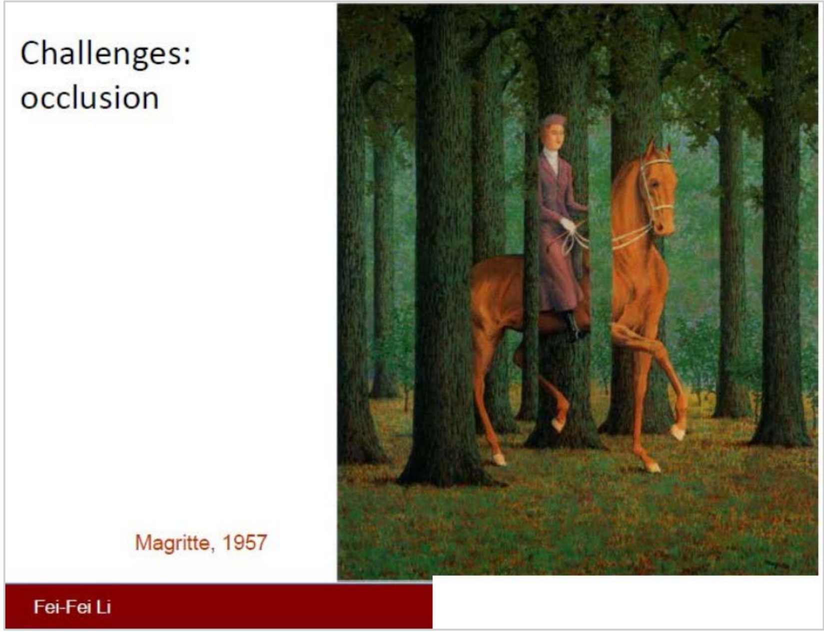


# Problem statement

- Challenges
  - Occlusion
    - Self-occlusion
    - Object occlusion
  - Lack of appropriate dataset




From Taehee Kim's slide



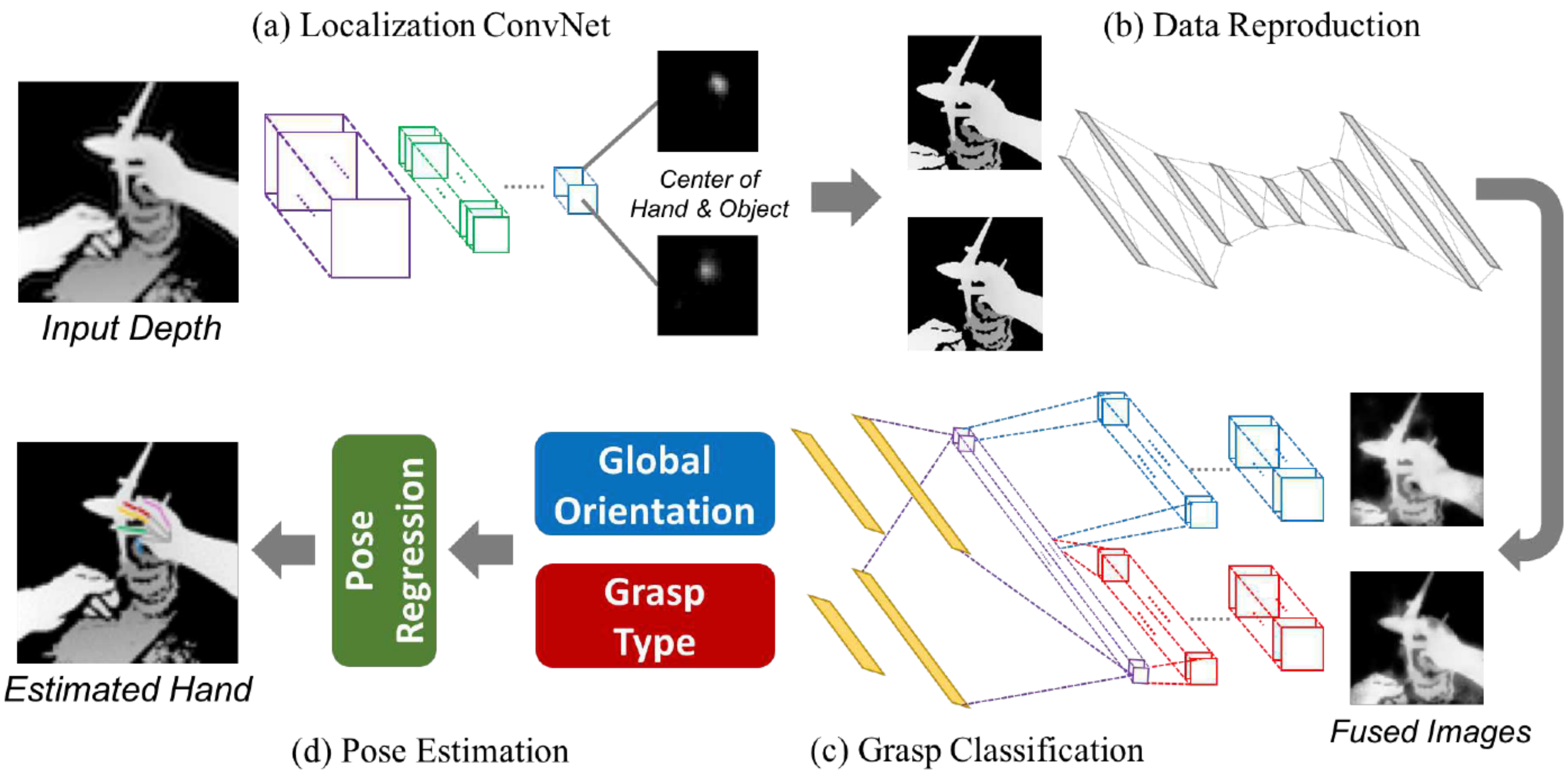

From the lecture note

# Overview

- Object occlusion: multi-channel pipeline (hand / object)
- Dataset synthesis & Data reproduction

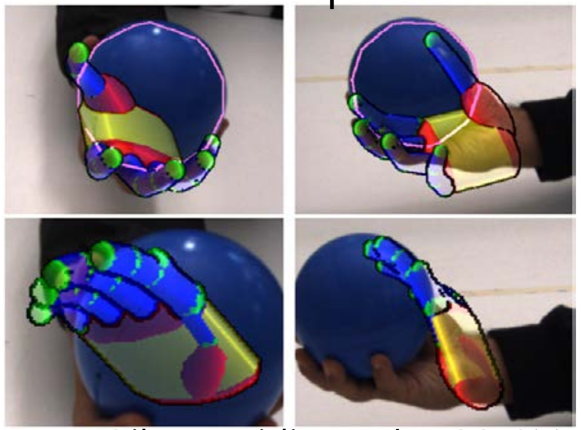
# Overview

- Object occlusion: multi-channel pipeline (hand / object)
- Dataset synthesis & Data reproduction



# Related work

## Hand model optimization



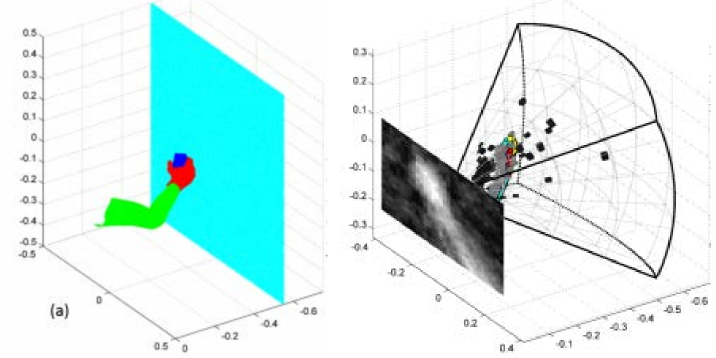
Oikonomidis et al. (ICCV'11)

## NN search from templates

(a) Original image (b) Segmented hand, HOG (c) NN in database, HOG

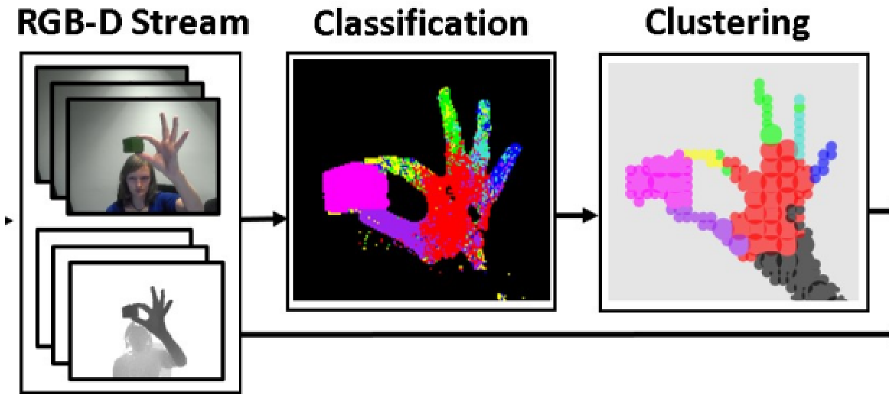
Romero et al. (ICRA'10)

## Segmentation & SVM



Rogez et al. (CVPR'15)

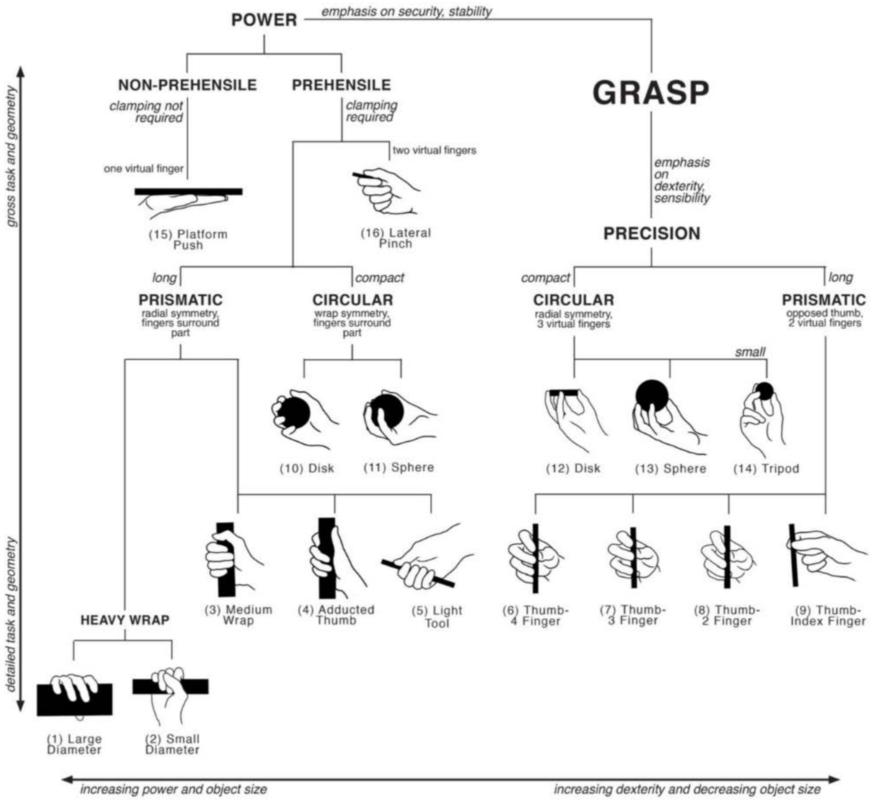
## Segmentation & pixel-wise classification




Sridhar et al. (ECCV'16)

# Related work

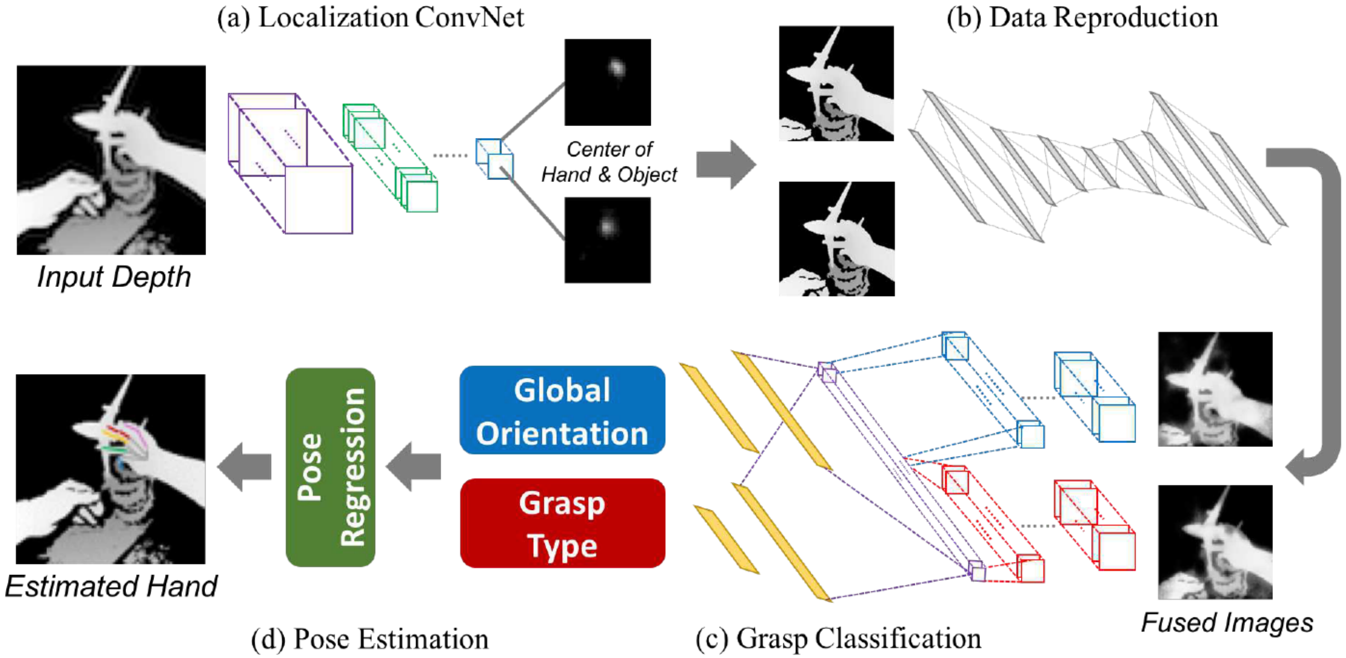
- Understanding Everyday Hands in Action from RGB-D Images
  - CNN framework for hand grasp classification
  - GUN-71 dataset based on hand grasp taxonomy




- Rogez, Grégory, James S. Supancic, and Deva Ramanan. "Understanding everyday hands in action from rgb-d images." *Proceedings of the IEEE international conference on computer vision*. 2015.

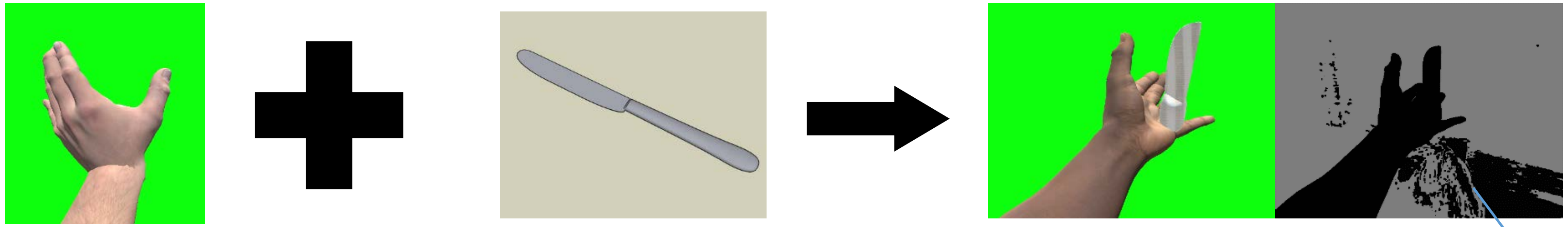
# The flow of this work

- Main idea
  - *The shape of an object causes a configuration of the hand grasp*
- Simultaneously train DNN using paired depth images for each hand and object



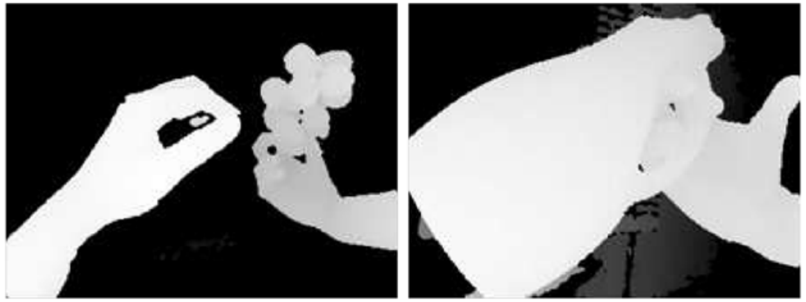
# Synthetic dataset

- virtual 3D CAD model of hand and objects
- model fitting method



Images are from SynthHands [ICCV'17], not from this paper.
Only for explanation

background: real depth image

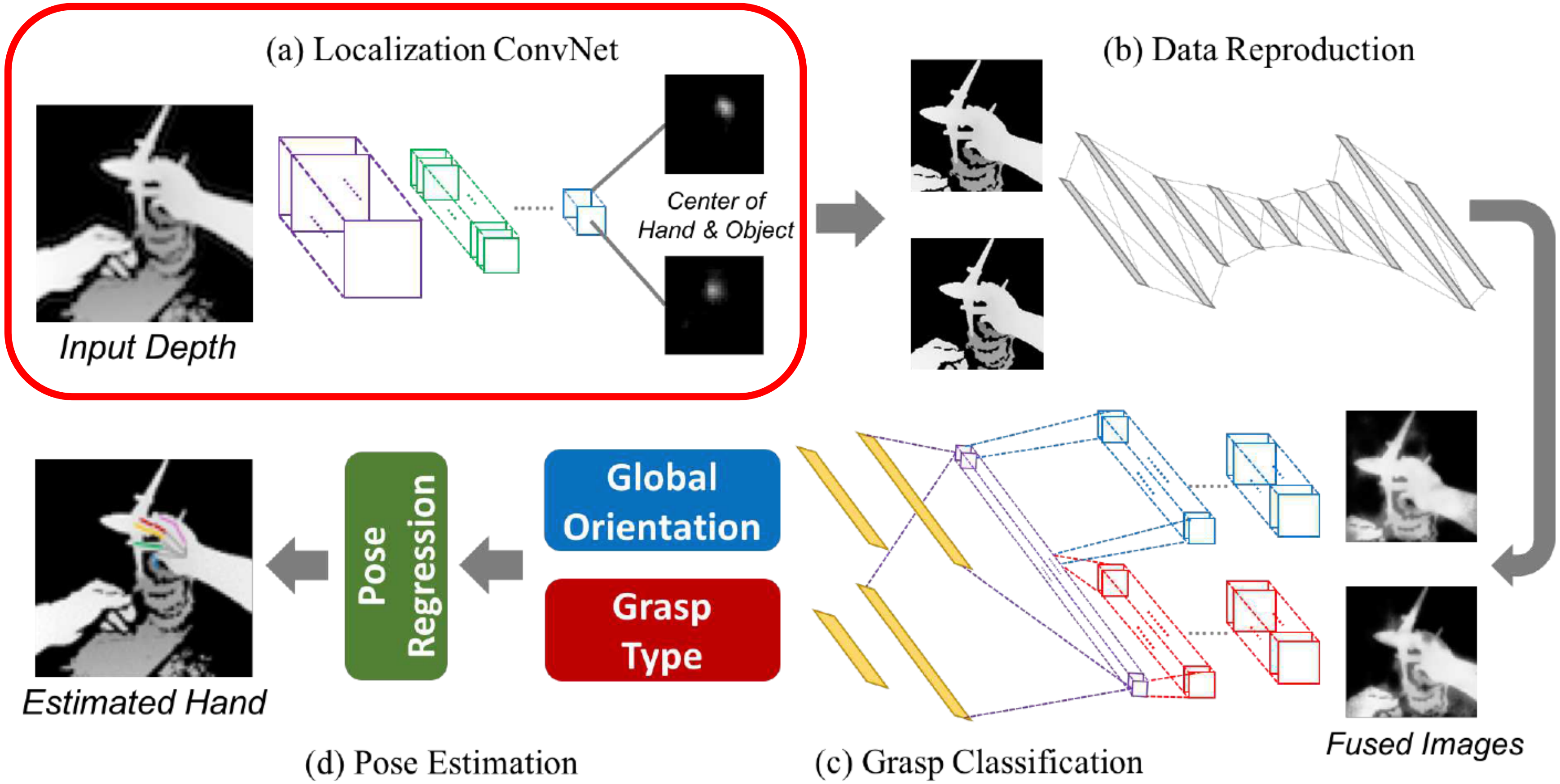


Input images from the paper

- 33 grasps × 40 objects × 48 rotations × 5 populations = 330K depth images

# Localization network

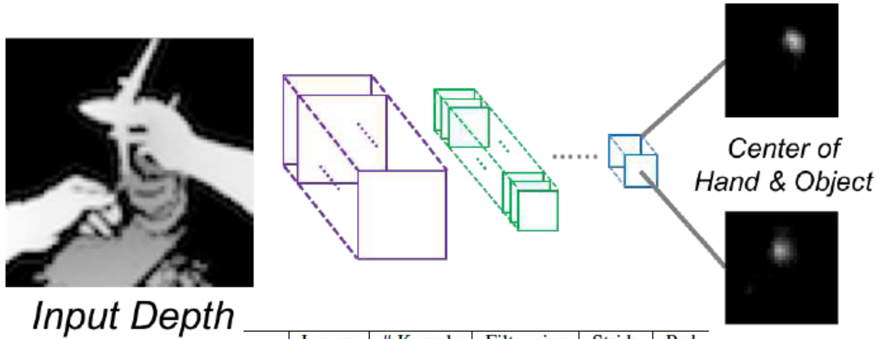
- Heatmap generation from ConvNet



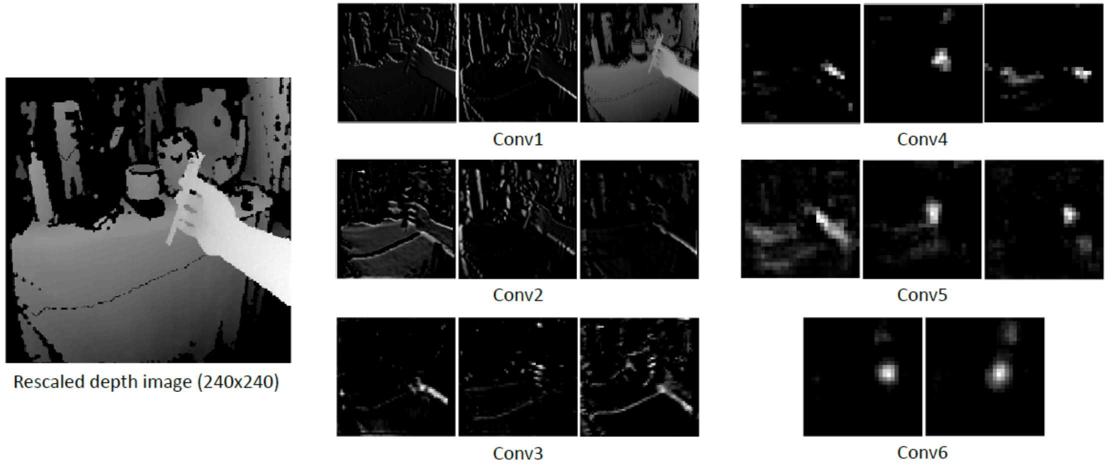
# Localization network

- Heatmap generation from ConvNet
  - Detect center position of each hand and object
  - Crop hand & object images based on detected position

(a) Localization ConvNet

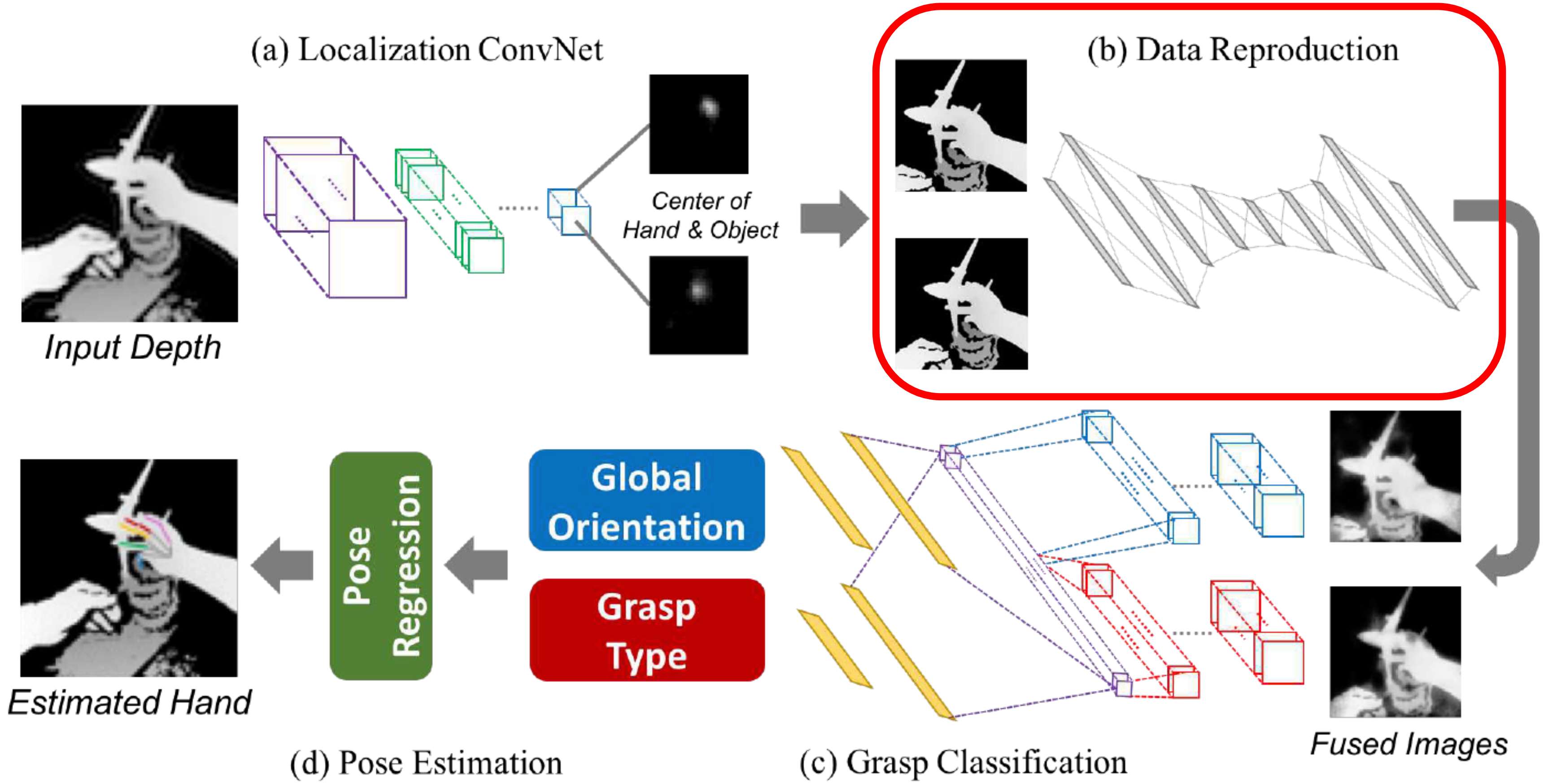



| | Layers | # Kernels | Filter size | Stride | Pad |
|---|---|---|---|---|---|
| 1 | Conv | 16 | 5×5×1 | 1 | 2 |
| 2 | ReLU | | | | |
| 3 | Pmax | | | 2 | 0 |
| 4 | Conv | 32 | 5×5×16 | 1 | 2 |
| 5 | ReLU | | | | |
| 6 | Pmax | | | 2 | 0 |
| 7 | Conv | 64 | 5×5×32 | 1 | 2 |
| 8 | ReLU | | | | |
| 9 | Pmax | | | 2 | 0 |
| 10 | Conv | 128 | 5×5×64 | 1 | 2 |
| 11 | ReLU | | | | |
| 12 | Conv | 256 | 5×5×128 | 1 | 2 |
| 13 | ReLU | | | | |
| 14 | Conv | 2 | 5×5×256 | 1 | 2 |
| 15 | ReLU | | | | |
| 16 | L2 | | | | |




From the supplement

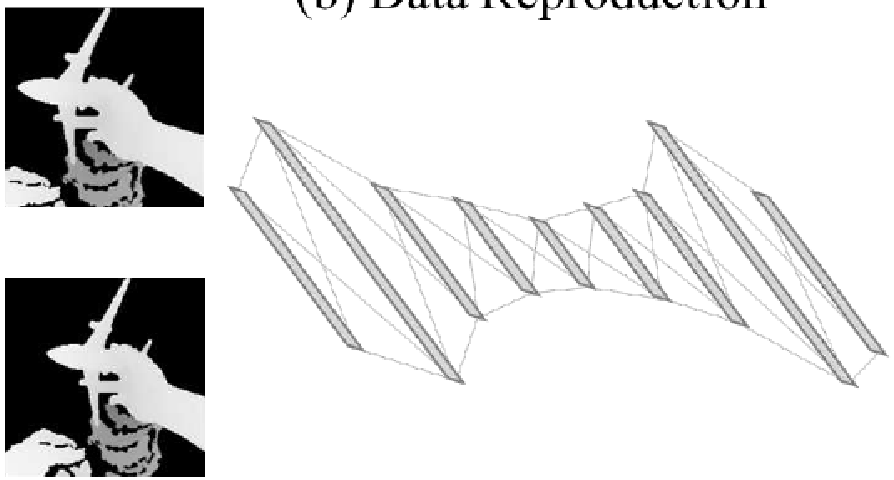
# Reproduction of realistic dataset

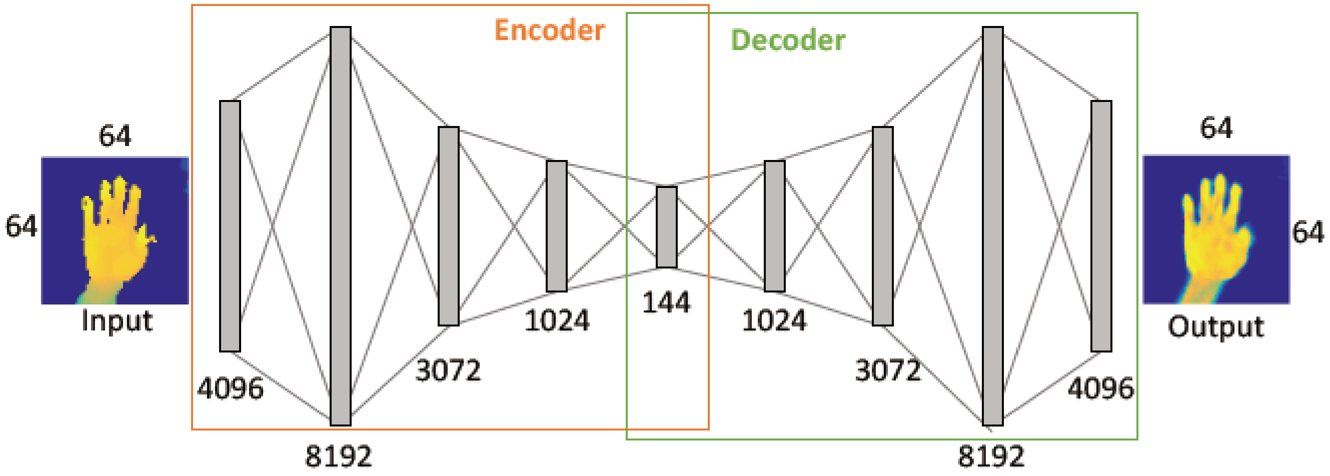


# Reproduction of realistic dataset

- To add “realism” from the network trained with synthetic images
- Signal reconstruction through an autoencoder
  - Trained with (160K real + 80K synthetic) depth images
  - Mimics the actual sensor image

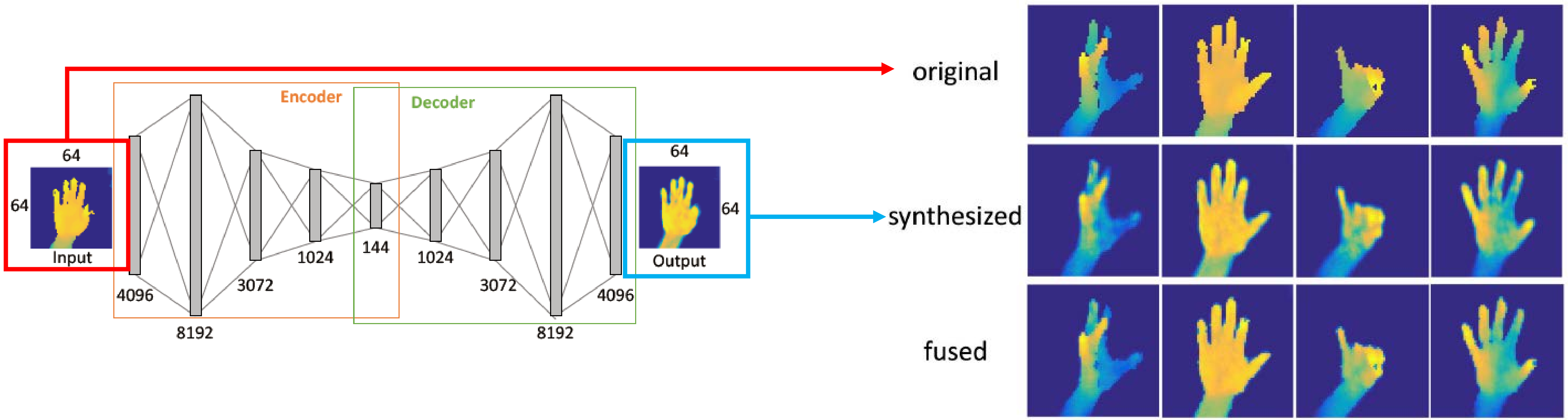
(b) Data Reproduction



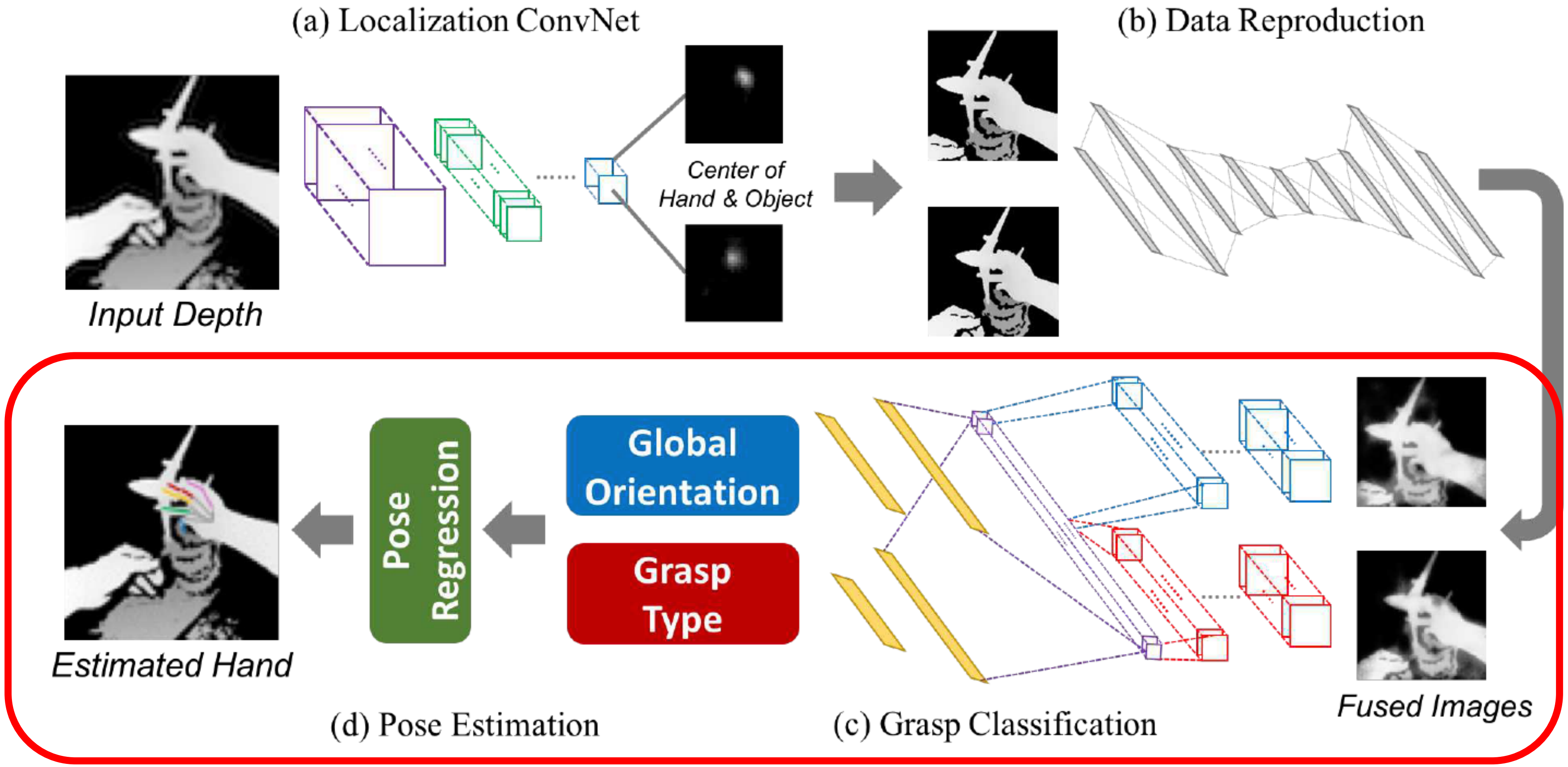


# Reproduction of realistic dataset

- Design choice: Original / Synthesized / **fused**
  - Original: pixel-wise defects due to real sensor artifacts (e.g. holes or missing pixels)
  - Synthesized: Compression distortion occurs

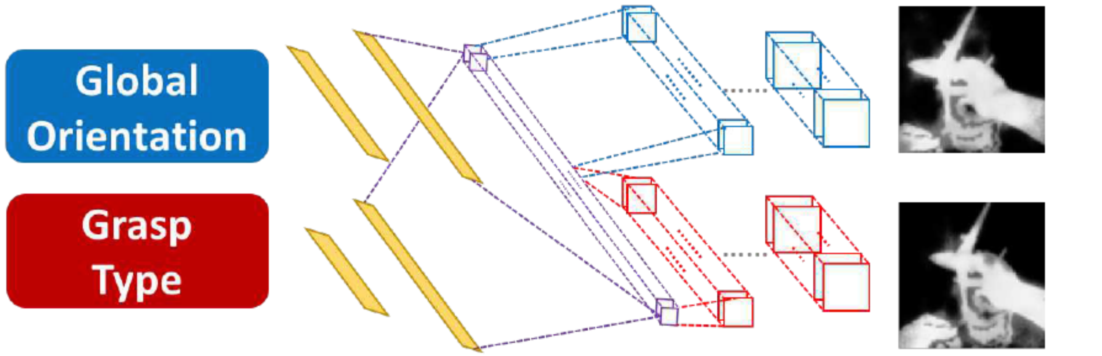


# Grasp classification & Pose estimation



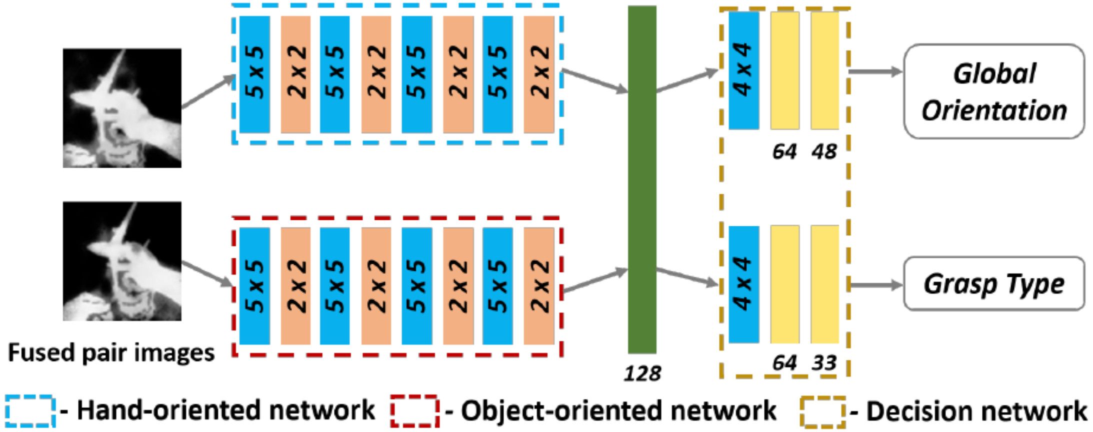
# Grasp classification

- Collaboratively learn convolutional features
- Share features about grasps from each hand and object perspective




Fused Images

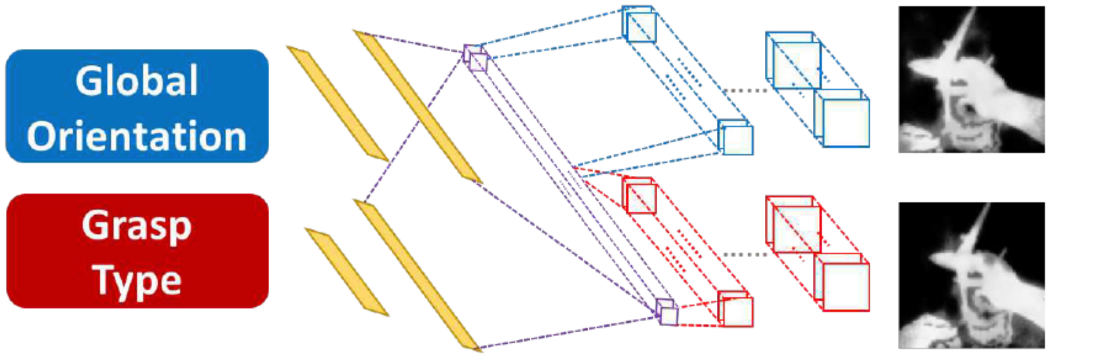
(c) Grasp Classification



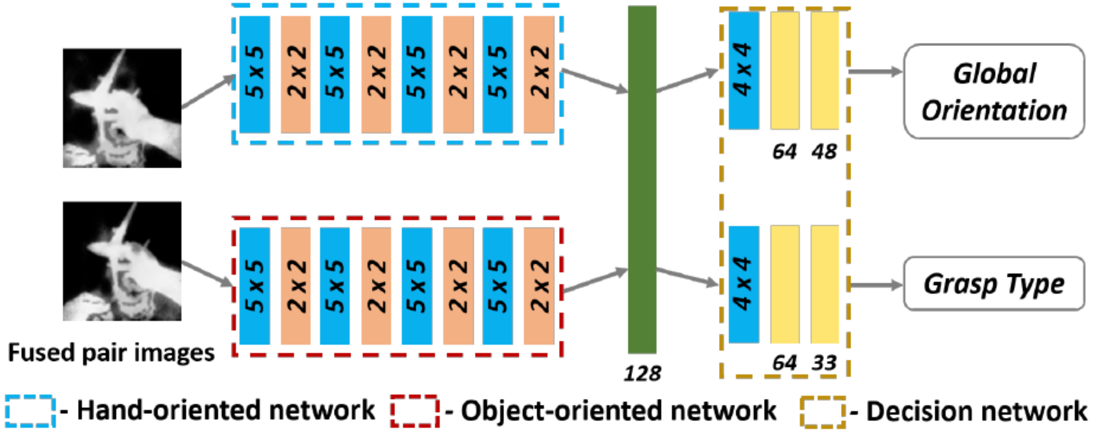

Blue: Conv+ReLu, Orange: Pmax
Green: Concat, Yellow: FC (ReLu between FC)

# Grasp classification

- Global orientation
  - Top 5 wrist orientations using Softmax function
- Grasp type
  - Top 1 grasp type



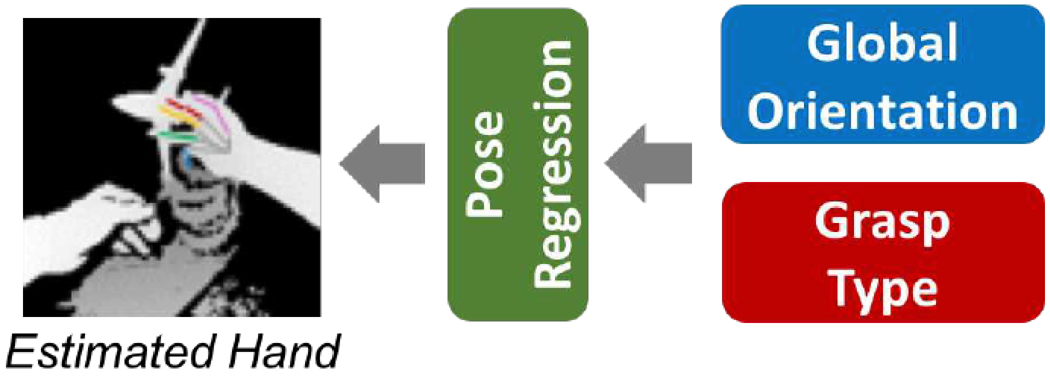
(c) Grasp Classification




Blue: Conv+ReLu, Orange: Pmax
Green: Concat, Yellow: FC (ReLu between FC)

# Pose Estimation

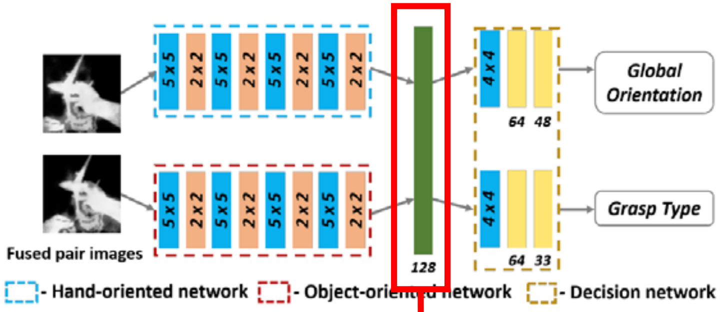
- Regression in a subset of classified orientation & grasp type
- Nearest neighbor search in the subset
- Regression through matrix completion




(d) Pose Estimation

# Pose Estimation

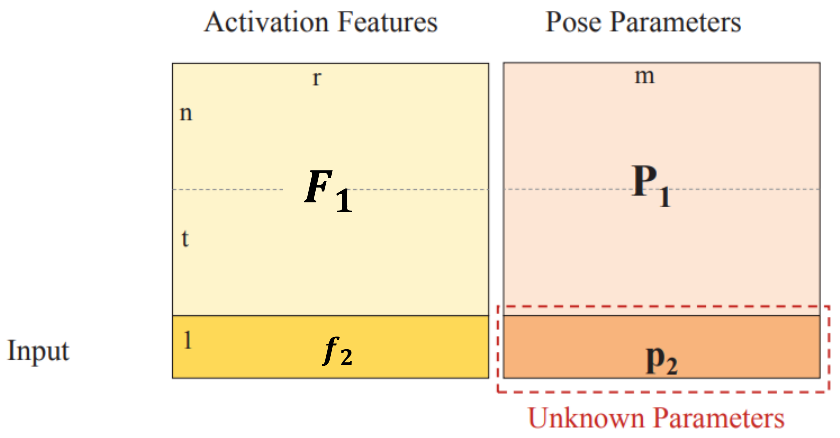
- Nearest neighbor search
  - Concatenated feature vector used
- Regression in a subset of classified orientation & grasp type
  - 1 grasp × 40 objects × 5 orientations × 5 populations ≈ 1K




Subset of synthetic dataset → NN search → 32 neighbor hand poses

# Pose Estimation

- Matrix completion
  - $F_1 \in \mathbb{R}^{l \times n}$, feature vectors of neighbor poses
  - $f_2 \in \mathbb{R}^{1 \times n}$, feature vector of input pose
  - $P_1 \in \mathbb{R}^{l \times m}$, joint angles of neighbor poses
  - **$\boldsymbol{p_2} \in \mathbb{R}^{\mathbf{1 \times m}}$, unknown angles of input pose**
  - $l = 32$, # of neighbors
  - $n = 64$, dimensionality of feature vector
  - $m = 18$, # of joint angles
- Interpolation of interest angles
  - weighted by feature vector



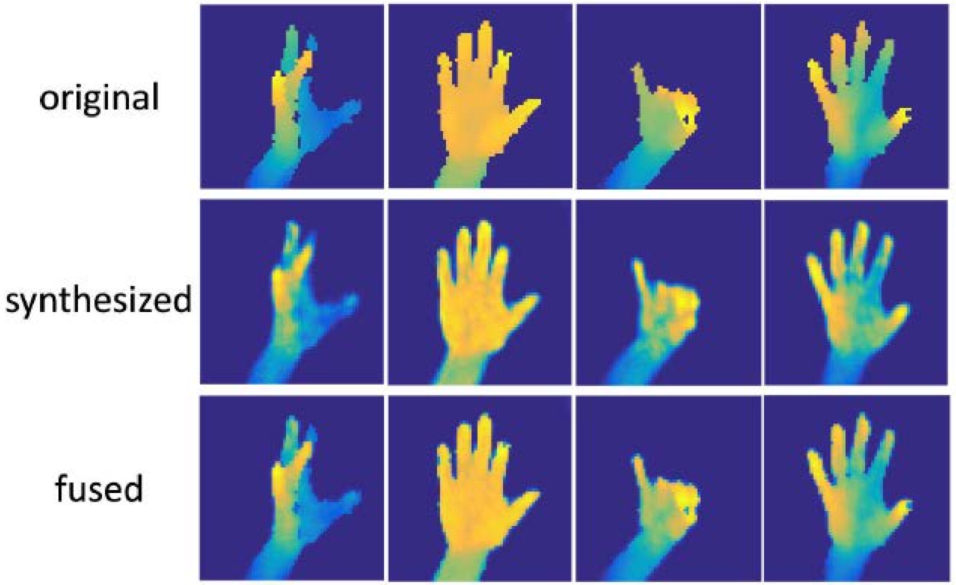

$$\mathbf{M} = \begin{bmatrix} \mathbf{F_1} & \mathbf{P_1} \\ \mathbf{f_2} & \mathbf{p_2} \end{bmatrix}, \qquad (1)$$

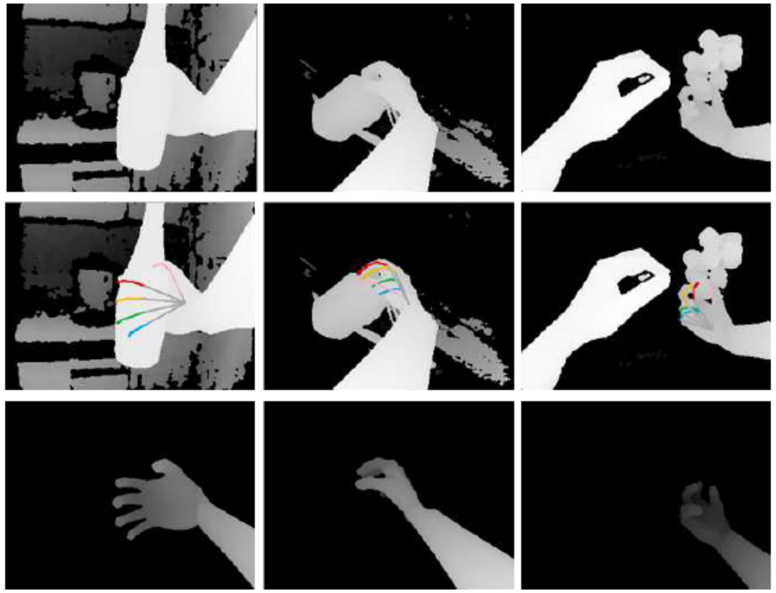
$$\mathbf{p_2} = \mathbf{f_2}(\mathbf{F_1})^{+}\mathbf{P_1}, \qquad (2)$$

- Sinha, Ayan, Chiho Choi, and Karthik Ramani. "DeepHand: Robust Hand Pose Estimation by Completing a Matrix Imputed with Deep Features-Supplementary Material."

# Evaluation

- Design choice evaluation
- Evaluation of multi-channel approach
- Pose estimation results



# Evaluation-Design choice analysis

- Original / Synthesized / **Fused**
  - Tested with GUN-71 dataset

| Train set \ Test set | Original GUN-71 | Synthesized GUN-71 | Fused GUN-71 |
|---|---|---|---|
| Original | 39.75% | 16.87% | 31.71% |
| Synthesized | 32.86% | 37.75% | 36.51% |
| Fused | 36.43% | 29.31% | **41.00%** |

Table 3: Grasp classification results for 33 grasps evaluated on GUN-71 dataset [19]. The use of reproduction network (spatially fused) improves overall classification results. Note that *Train* denotes the type of training dataset used to train our model and *Test* denotes the format of GUN-71 dataset used for testing our networks.

| Model | Classification accuracy |
|---|---|
| Rogez et al. [19] | 20.50 % |
| *Original* | 39.75 % |
| *Synthesized* | 37.75 % |
| **Ours (*Fused*)** | **41.00 %** |

Table 4: Accuracy comparison of grasp classification on GUN-71 dataset.

# Evaluation-Design choice analysis

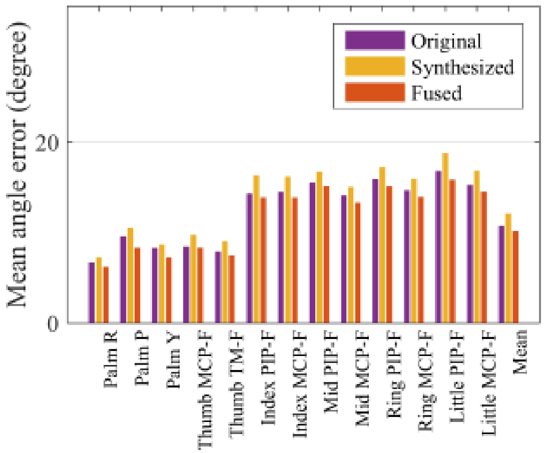
- Hand & Object is better
- Considering both perspectives could improve performance

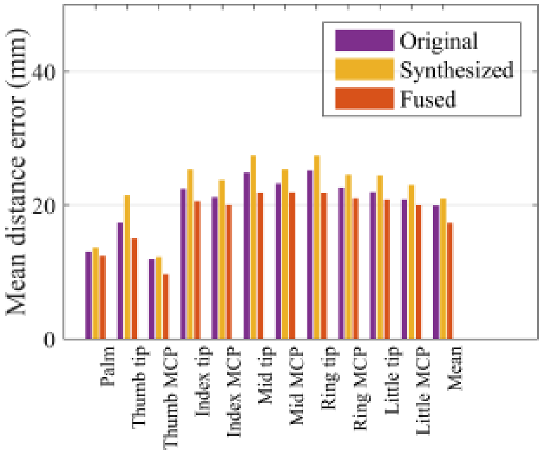
| Network | *Hand-only* | *Ojbect-only* | **Ours** |
|---|---|---|---|
| Orientation Acc. | 59.31% | 51.12% | **60.50%** |
| Grasp Acc. | 43.87% | 49.12% | **55.56%** |

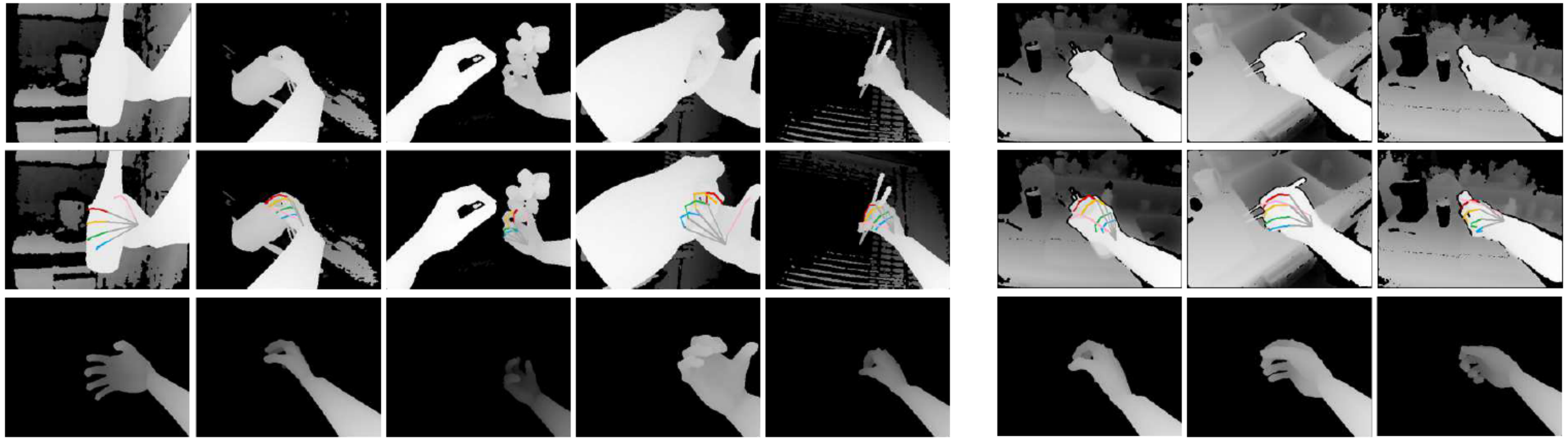
Table 5: Classification accuracy for the orientation of the hand and the grasp type. *Hand only* achieves higher performance to orientation classification than *Object only* but has less impact on grasp classification.

# Evaluation-Pose estimation

- Quantitative
  - Mean error with ground truth
  - Angle & distance
  - Smaller the better
- Qualitative
  - (a) Dataset from the paper
  - (b) GUN-71



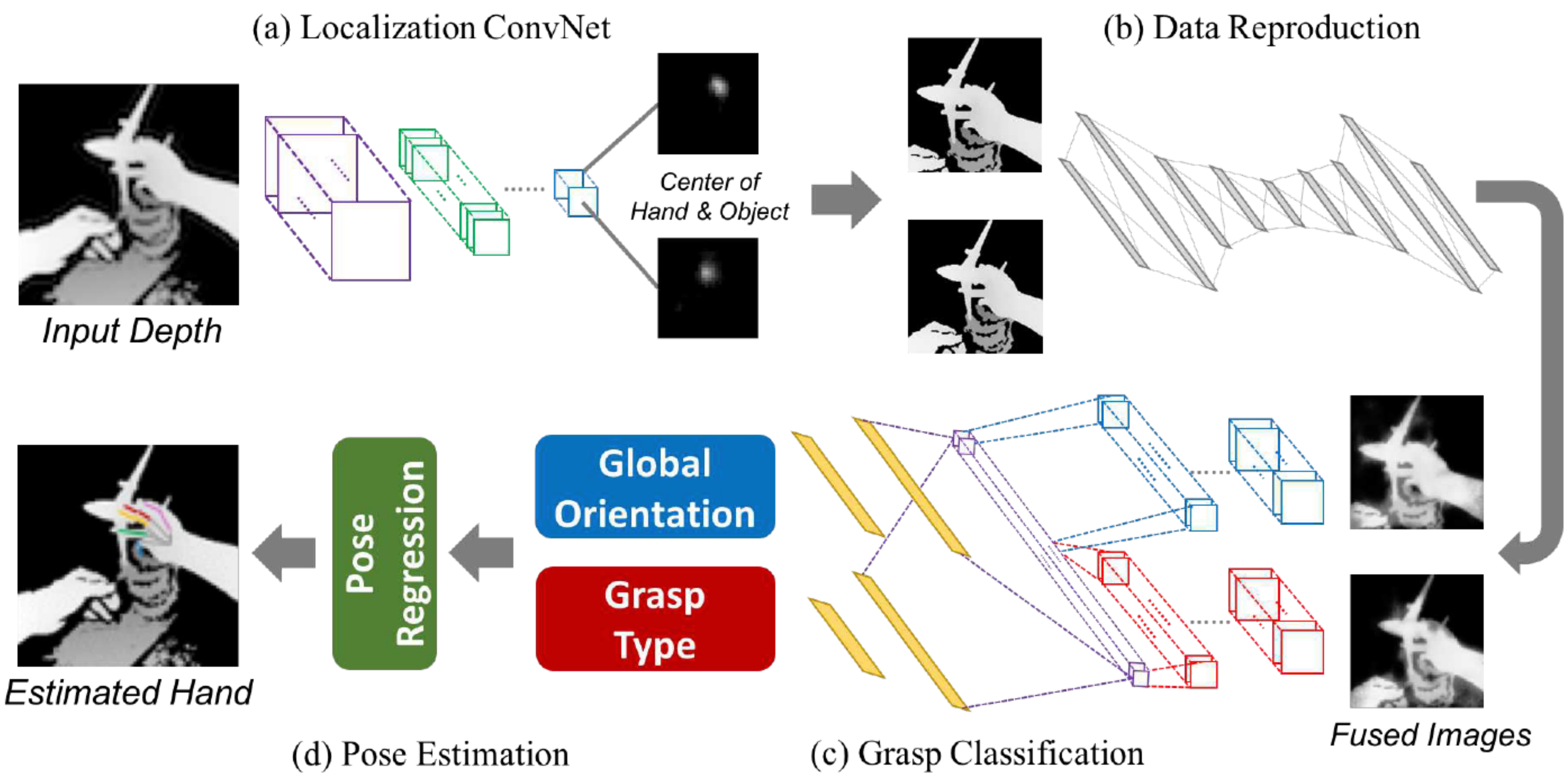



(a) (b)

# Wrap-up

Synthetic dataset



Thank you for listening

# Quiz

- Q1.
- Q2.