GRAPH MINING

HOPE-Net: A Graph-based Model for Hand-Object Pose Estimation

GRAPH MINING

Group 6:

Doãn Anh Khoa - 21127076

Đoàn Việt Hưng - 21127289

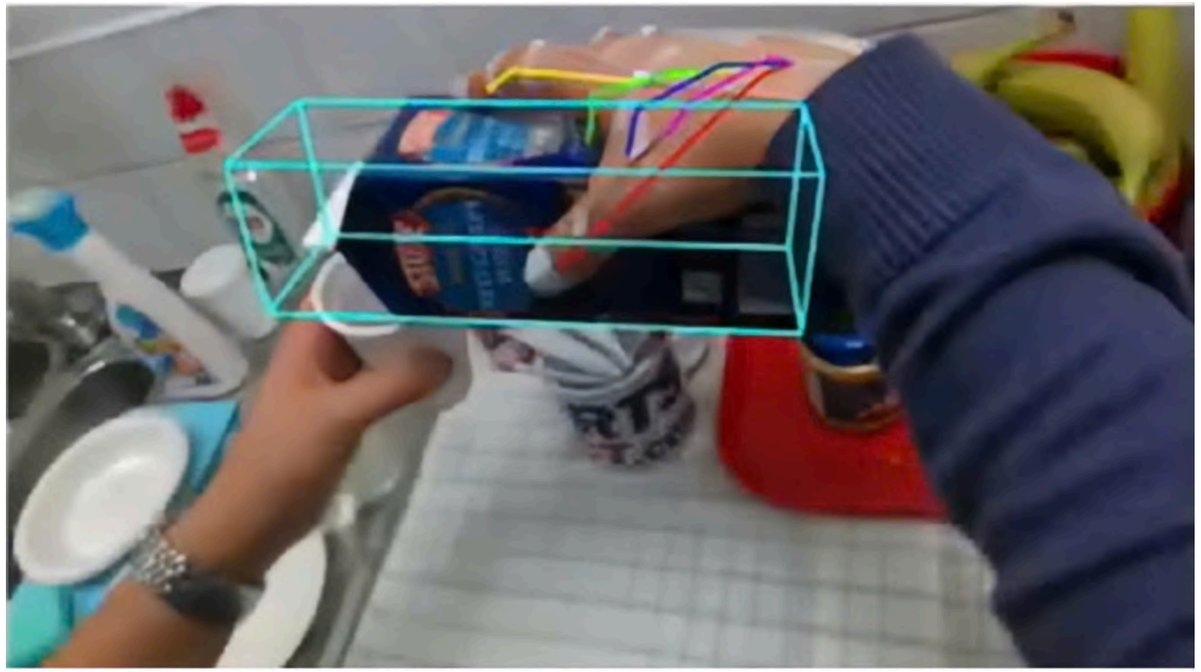
Đinh Bảo Trân - 21127454

Lê Nguyễn Phương Uyên - 21127476

Content

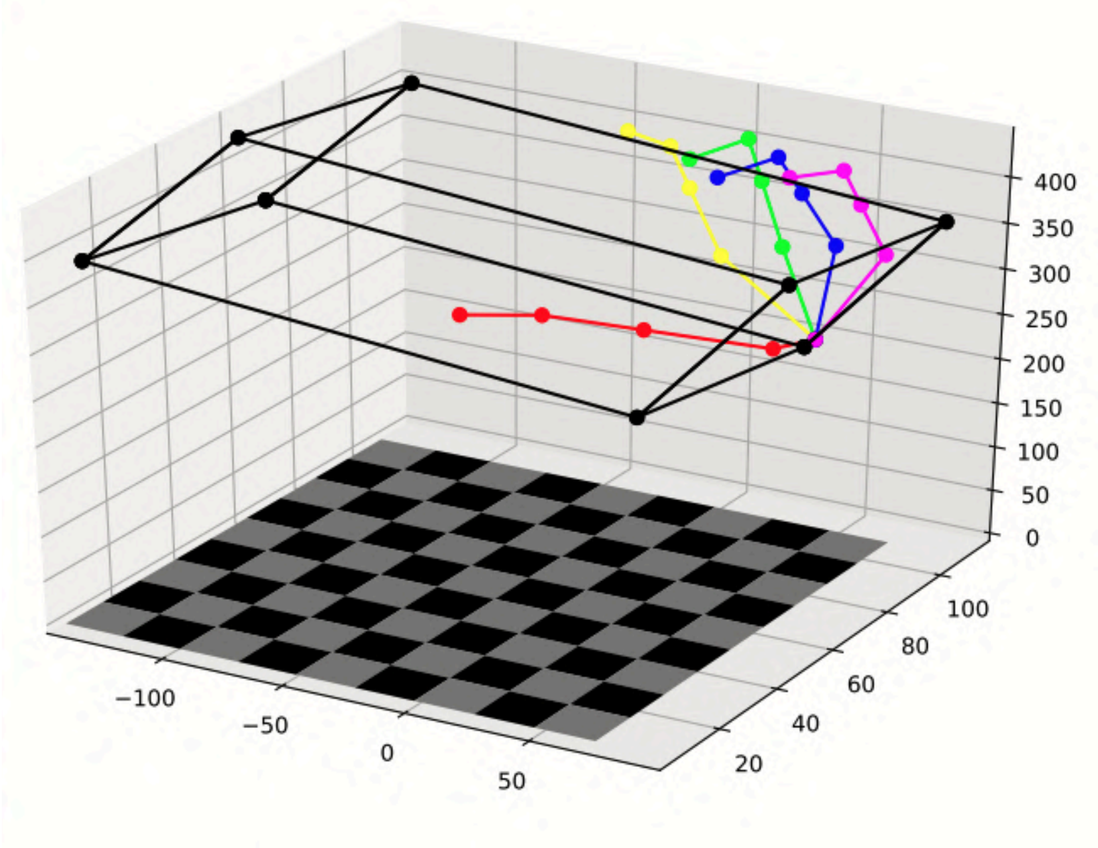
Overview

Introduction



- HOPE-NET MODEL
- GRAPH-BASED APPROACH
- IMPROVE ACCURACY

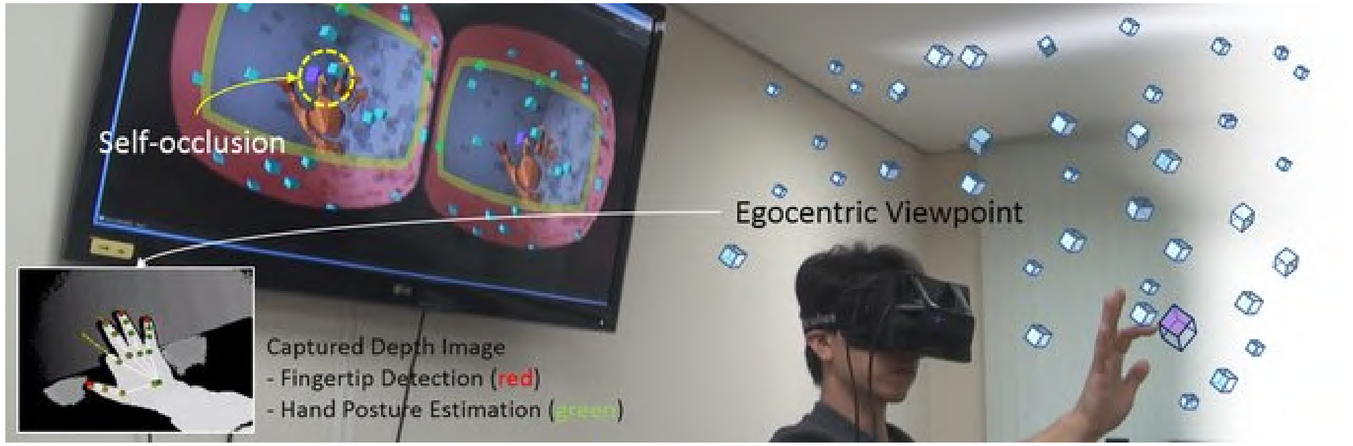
Theoretical meanings

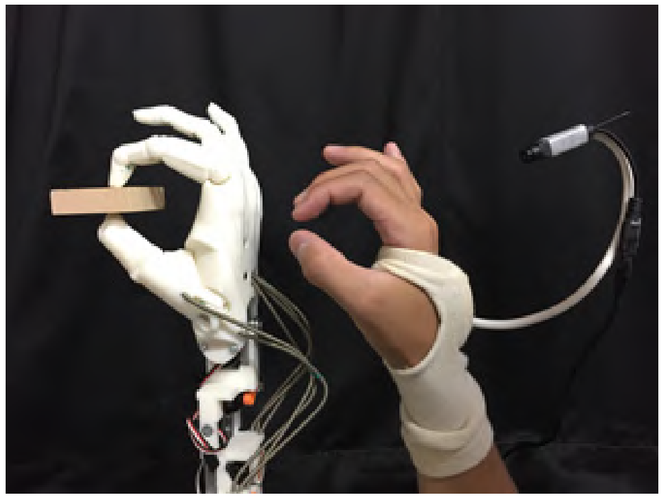


- GRAPH THEORY IN COMPUTER VISION
- FEATURE CONCATENATION AND CONDITIONING
- ADAPTIVE GRAPH U-NET

Practical meanings

- VR/AR APPLICATIONS
- ROBOTICS
- SECURITY AND SURVEILLANCE
- HEALTHCARE AND REHABILITATION





HAND-POSE AND OBJECT SHAPES

OCCLUSION

DATA SCARCITY AND ANNOTATION

LIGHTING AND ENVIRONMENTAL CONDITIONS

DYNAMIC AND REAL-TIME PROCESSING

Methodology

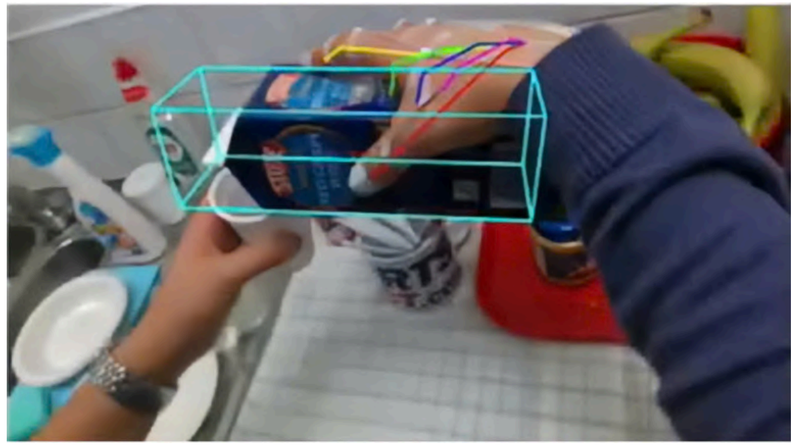
Problem statement

INPUT

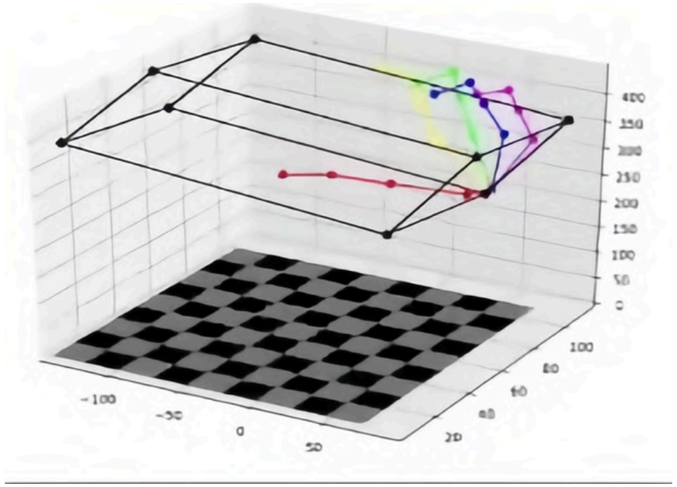
3D object models, hand joints coordinates, 6D pose information, and image frames from different viewpoints.

OUTPUT

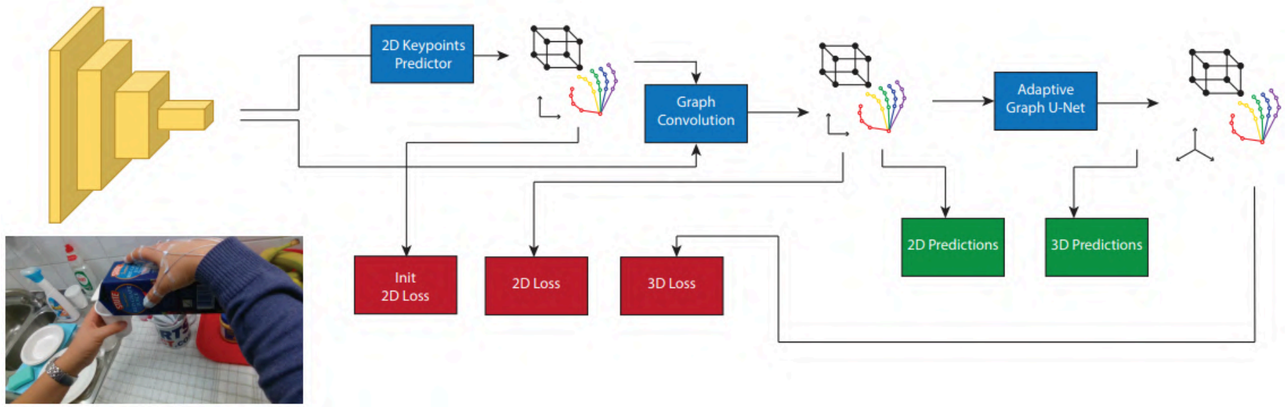
Pose Parameters



3D Keypoints

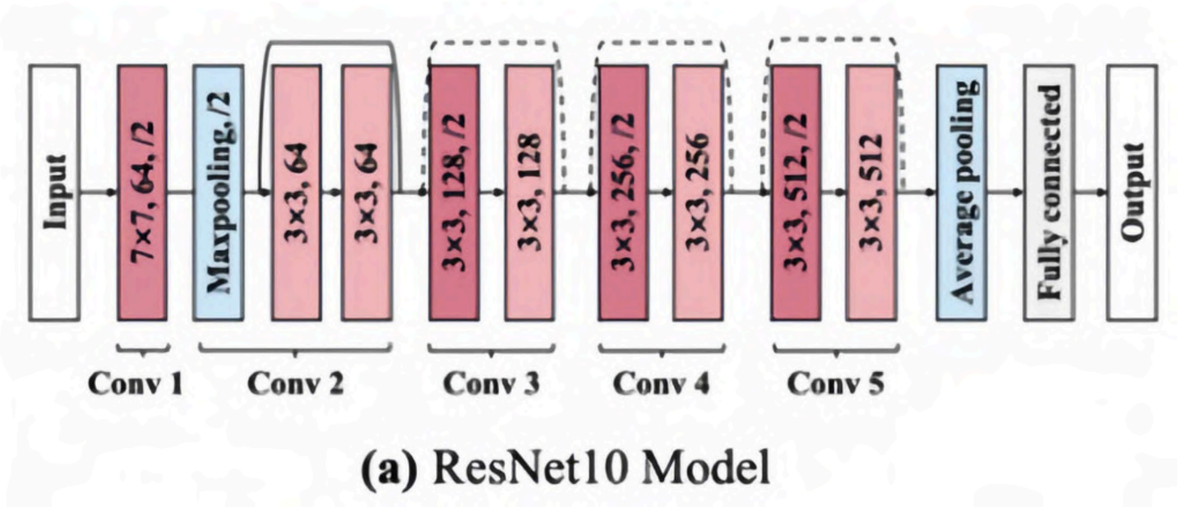


Model Architecture



- RESNET10
- GRAPH NETWORK
- ADAPTIVE GRAPH U-NET

RESNET10



(a) ResNet10 Model

THE IMAGE ENCODER

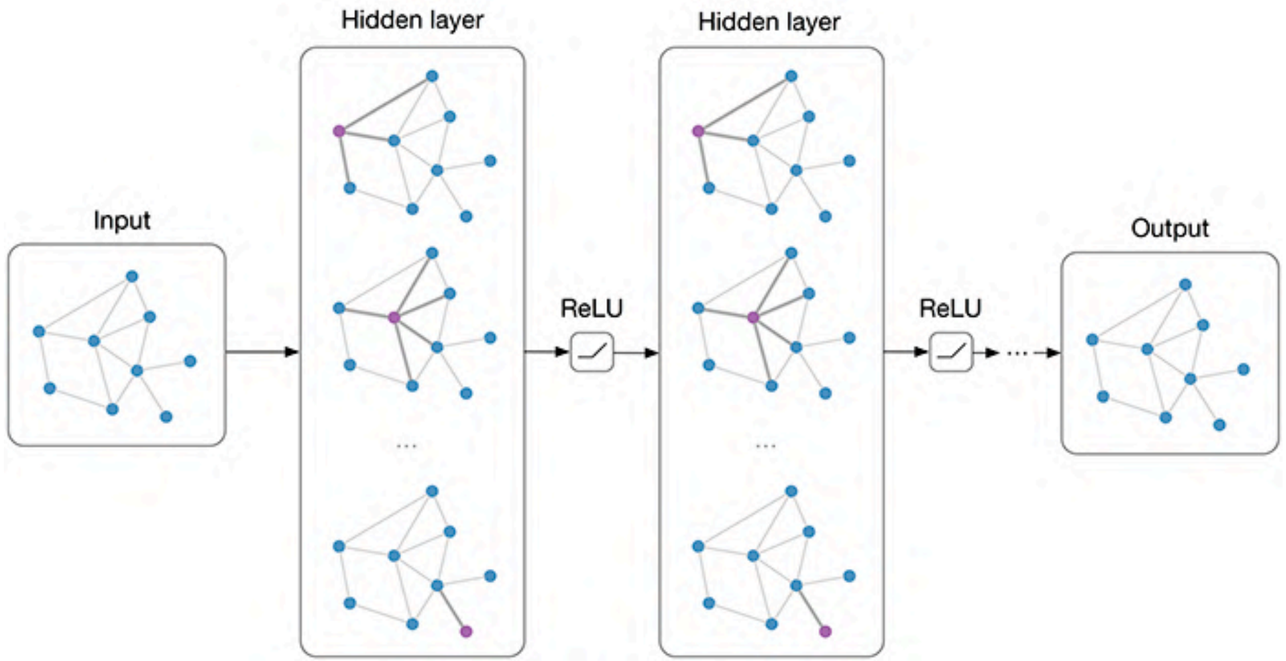
- 2048D feature vector for each image

PREDICT THE INITIAL 2D COORDINATES

- Using a fullyconnected layer.

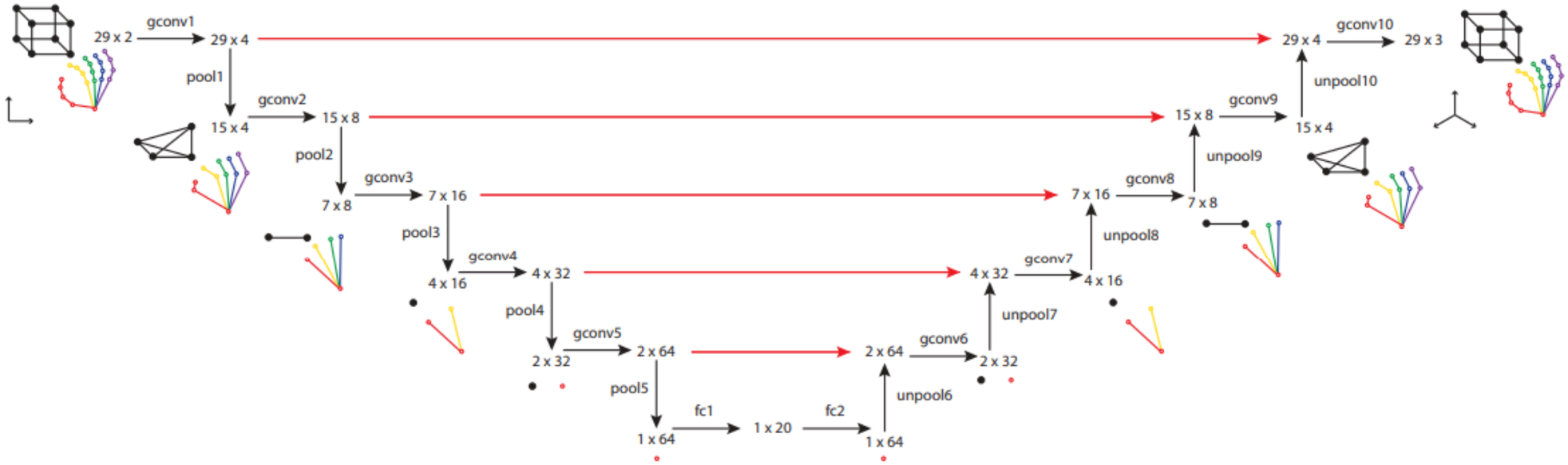
YIELDING A GRAPH WITH 2050 FEATURES

Graph Network



- A 3-layer adaptive graph convolution network
- Use adjacency information
- Modify the 2D coordinates of the keypoints.

ADaptive Graph U-Net



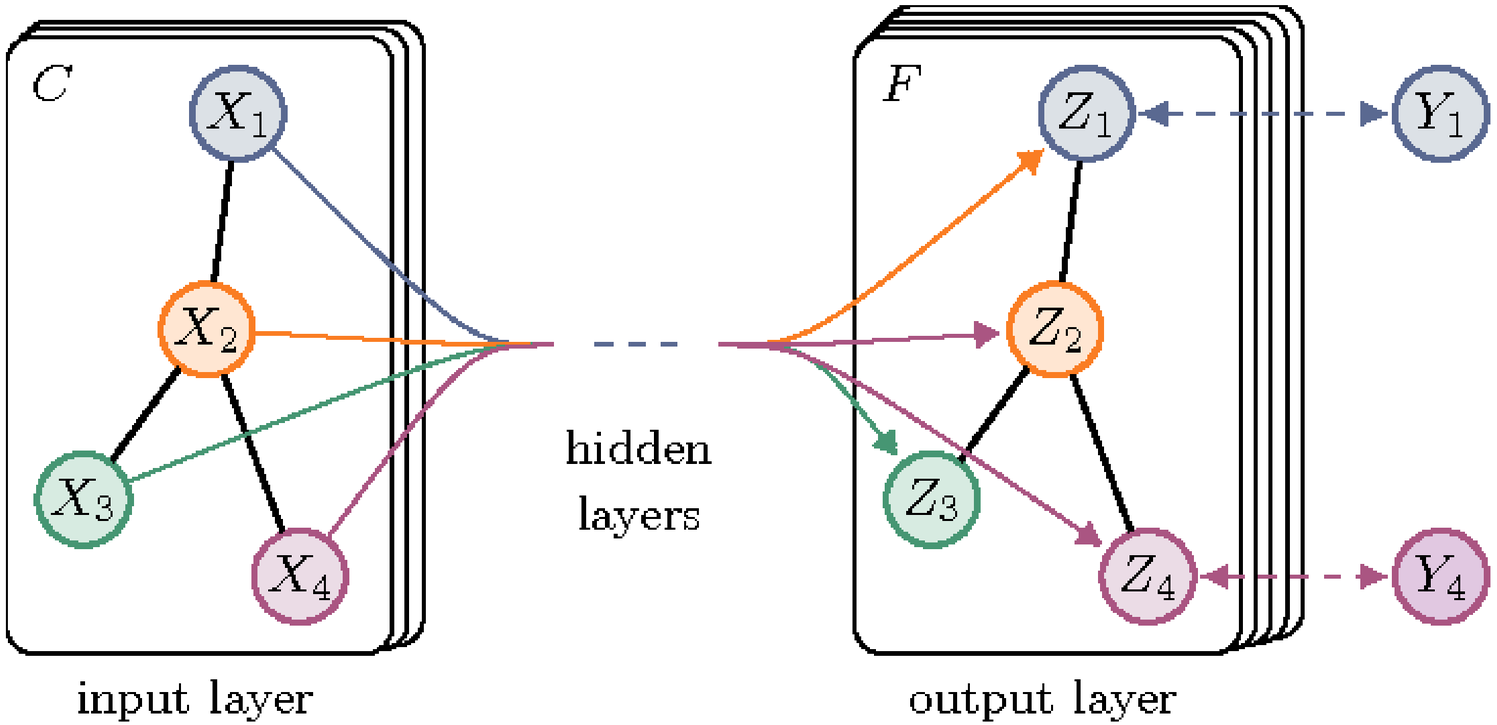
Graph convolution

Graph pooling

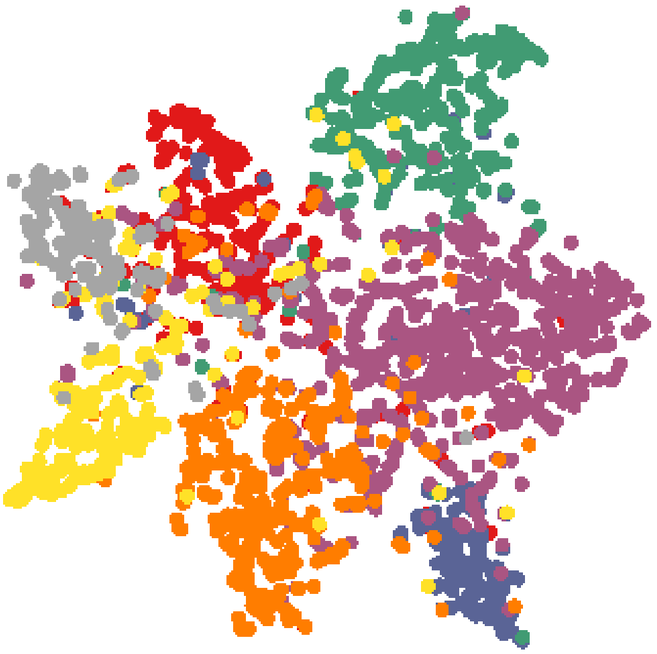
ReLu

Graph unpooling

Graph convolution



(a) Graph Convolutional Network



(b) Hidden layer activations

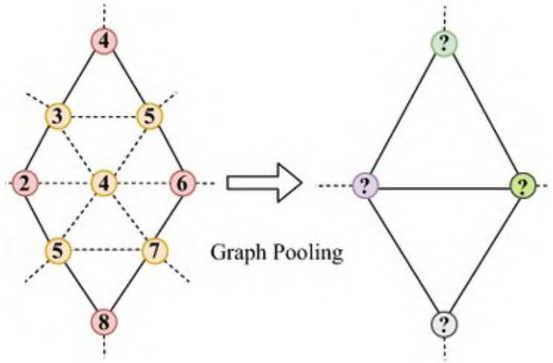
Graph convolution

This paper’s convolution based on the Renormalization Trick:

$$Y = \sigma(\tilde{A}XW)$$

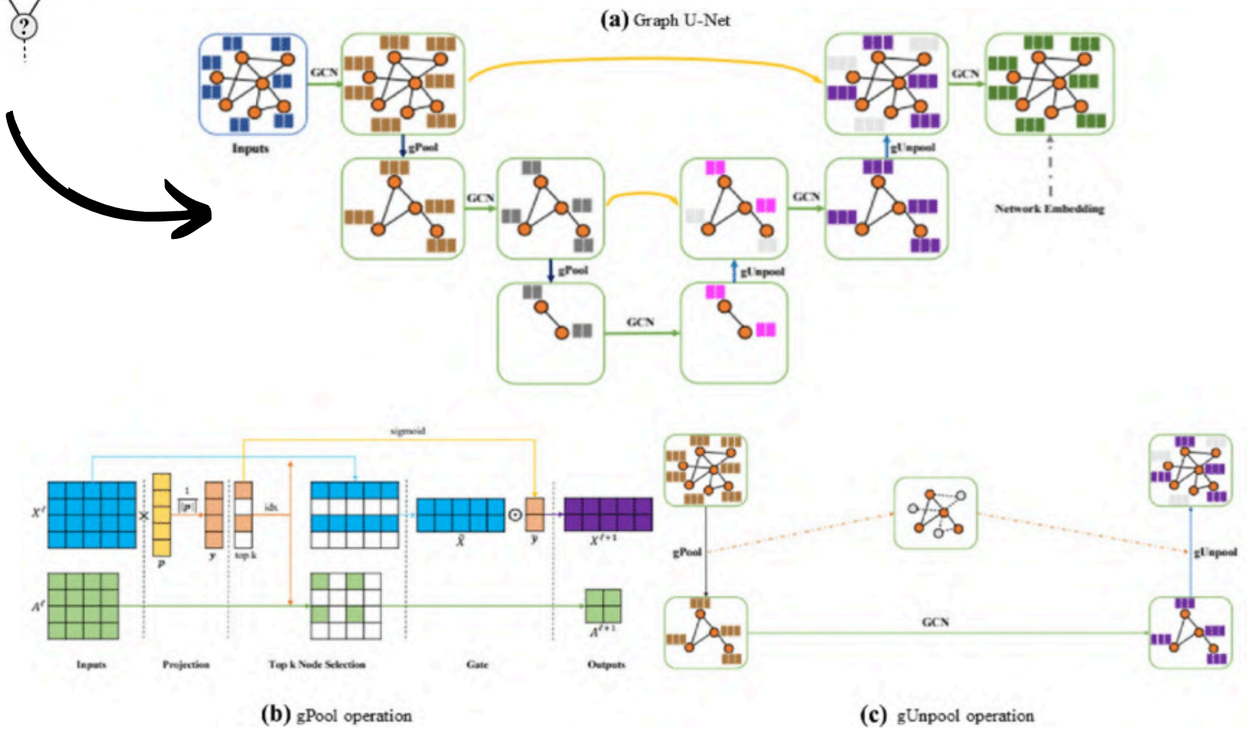
Where:

- σ: ReLU activation function
- W: trainable weights matrix
- X: matrix of input features
- Ã: row-normalized adjacency matrix (from Laplacian formula)



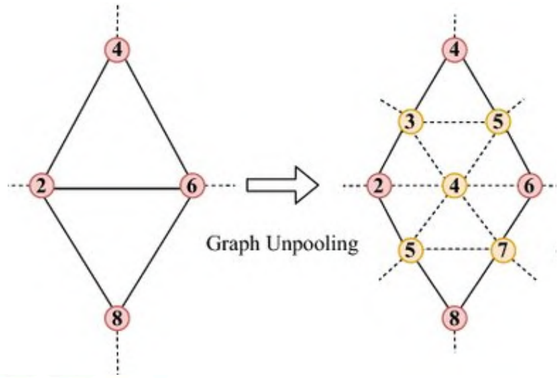
Graph pooling

Due to the weaknesses of the sigmoid function, they used a fully connected layer and applied it to the transposed feature matrix.

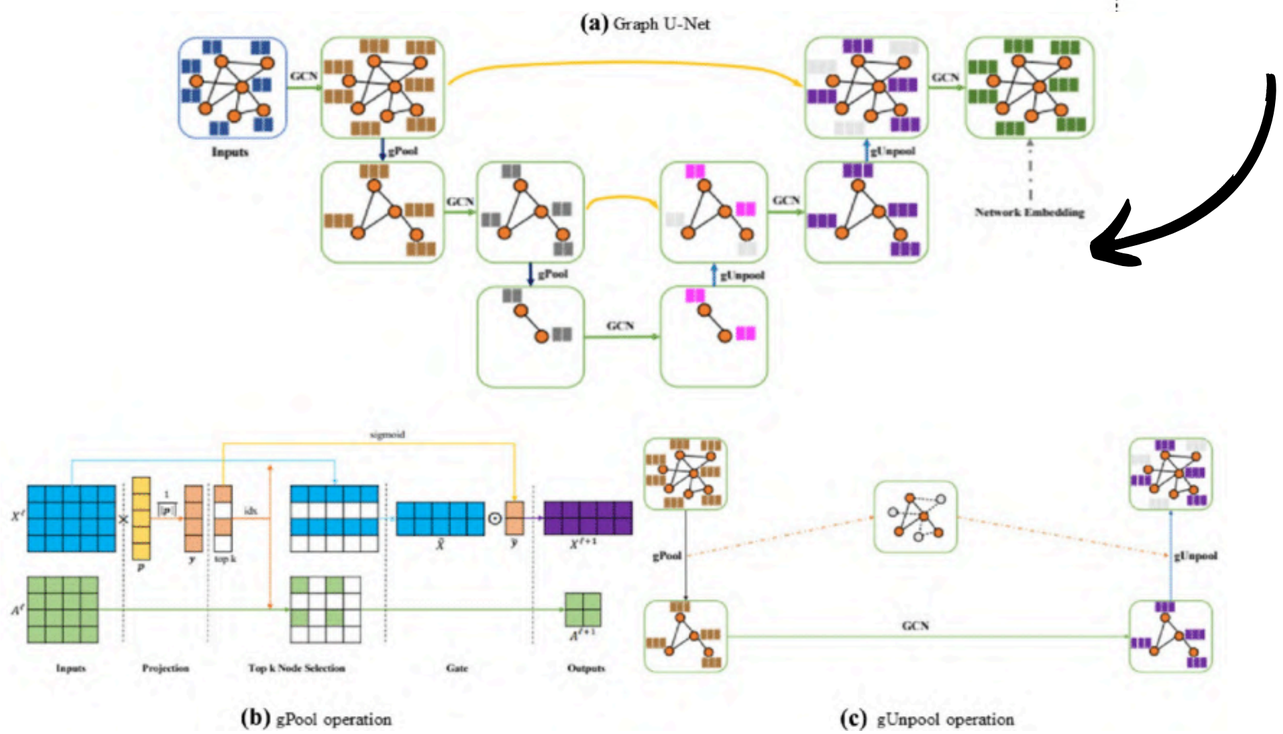


Graph unpooling

- That approach adds the pooled nodes to the graph with empty features and uses the subsequent graph convolution to fill those features



- Using a layer for gUnpool is similar to the pooling layer.



Loss Function and Training the Model

$$\mathcal{L} = \alpha\mathcal{L}_{init2D} + \beta\mathcal{L}_{2D} + \mathcal{L}_{3D}$$

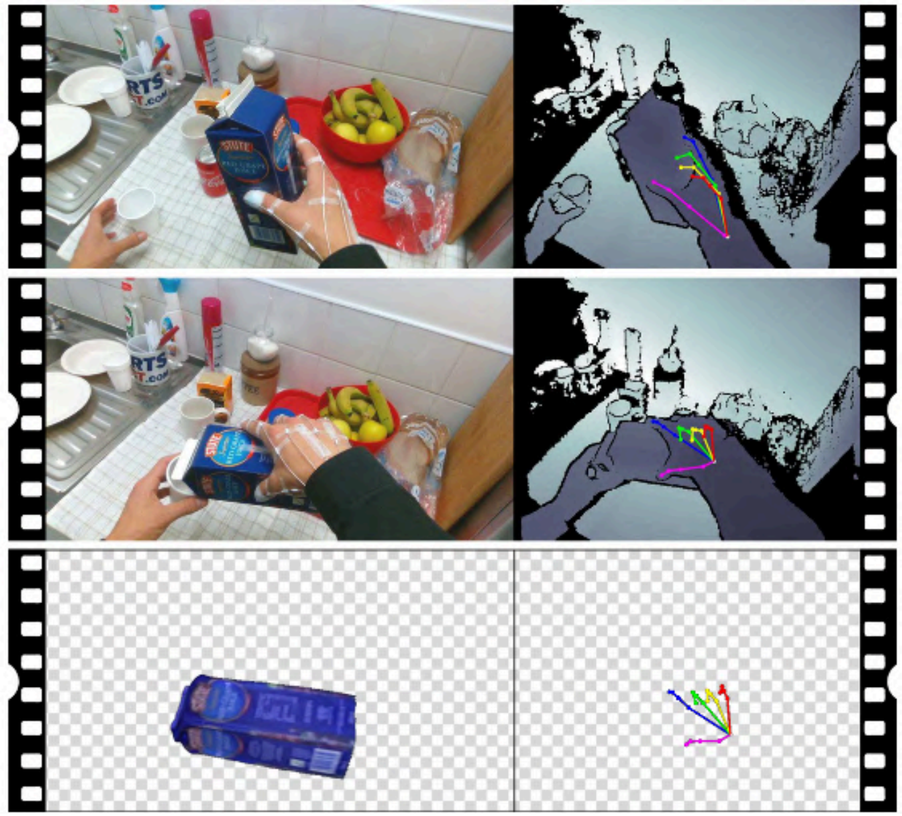
- Set α and β to 0.1
  - Purpose: bring the 2D error (in pixels) and 3D error (in millimeters) into a similar range.
- For each of the loss functions, we used Mean Squared Error.

Experiment & Evaluations

Dataset

First-Person Hand Action Dataset

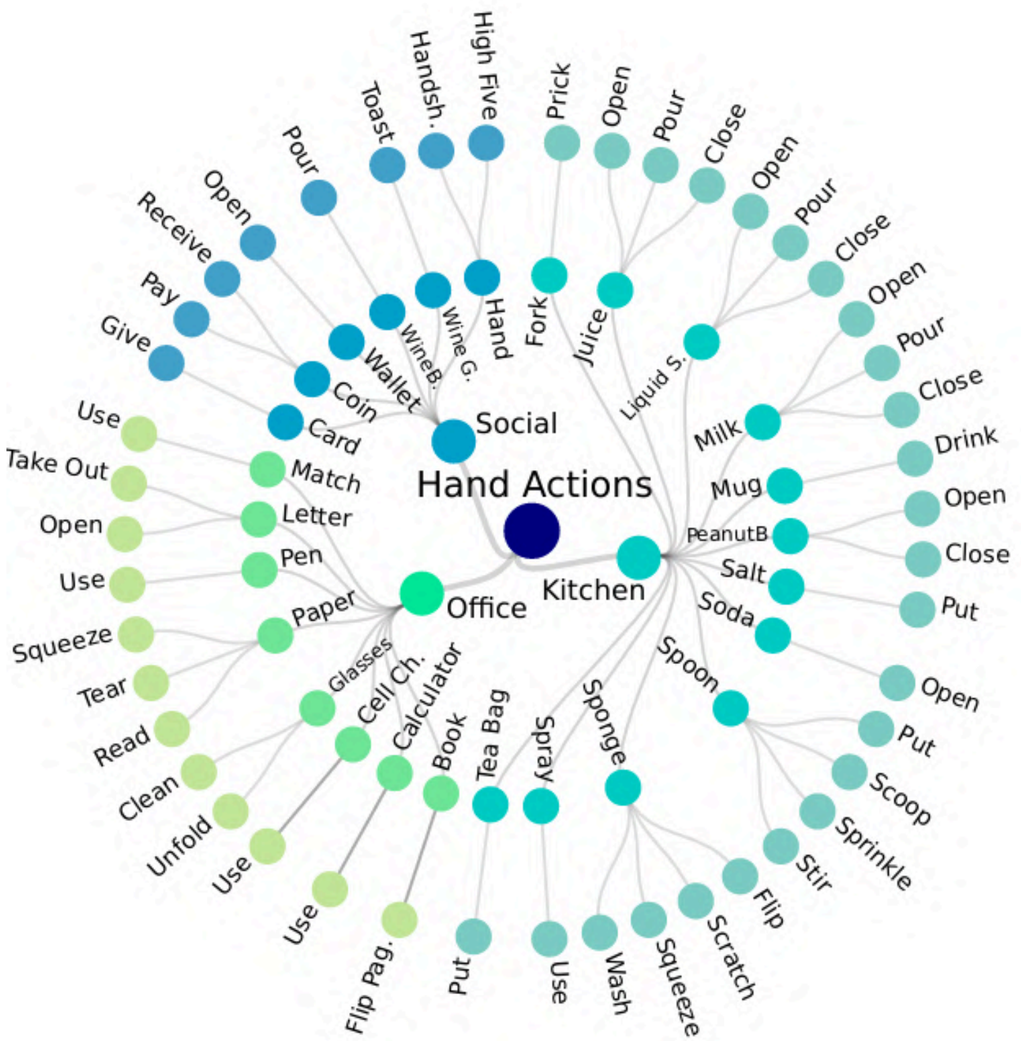
This dataset features videos of hand actions performed from a first-person perspective



Dataset

First-Person Hand Action Dataset

- The actions involve manipulating everyday objects.
- The dataset captures a variety of actions on these objects.

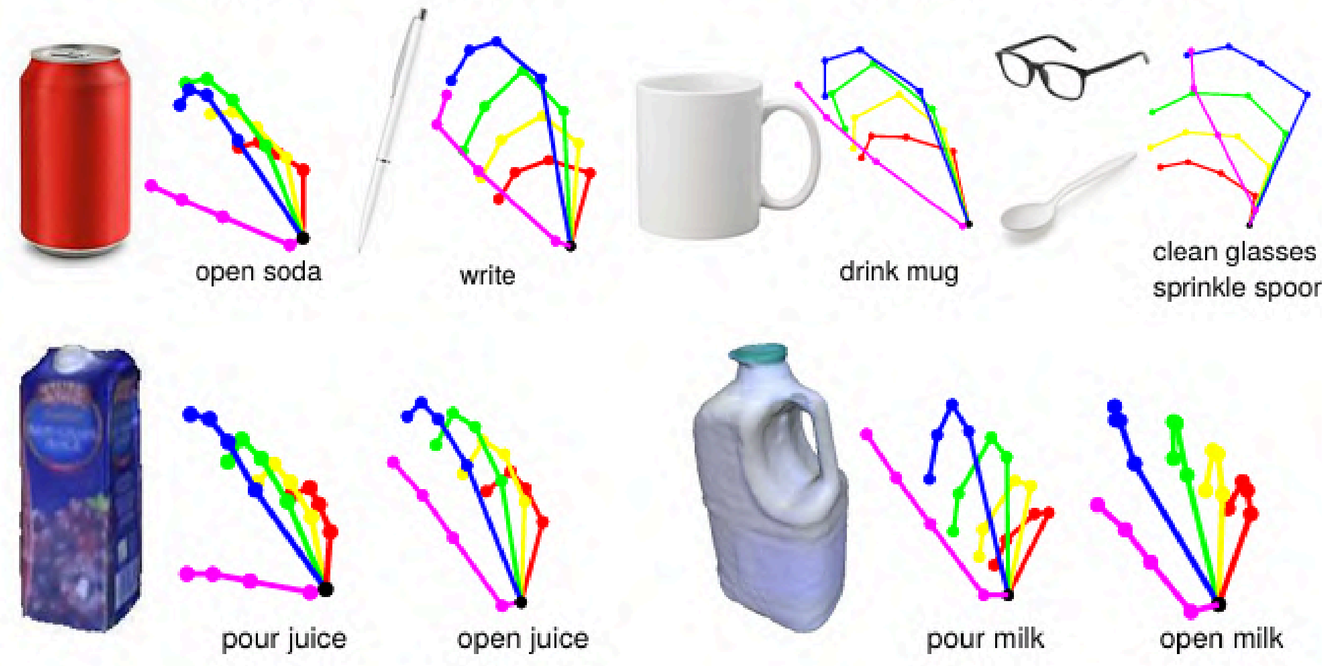


Dataset

First-Person Hand Action Dataset

Average Poses: Represents the average pose for each action class.

- Objects and Grasps: Objects can have multiple grasps depending on the action.
- Grasps and Actions: A single grasp can be used for multiple actions.



Dataset

First-Person Hand Action Dataset

Only a subset of 21,501 frames includes:

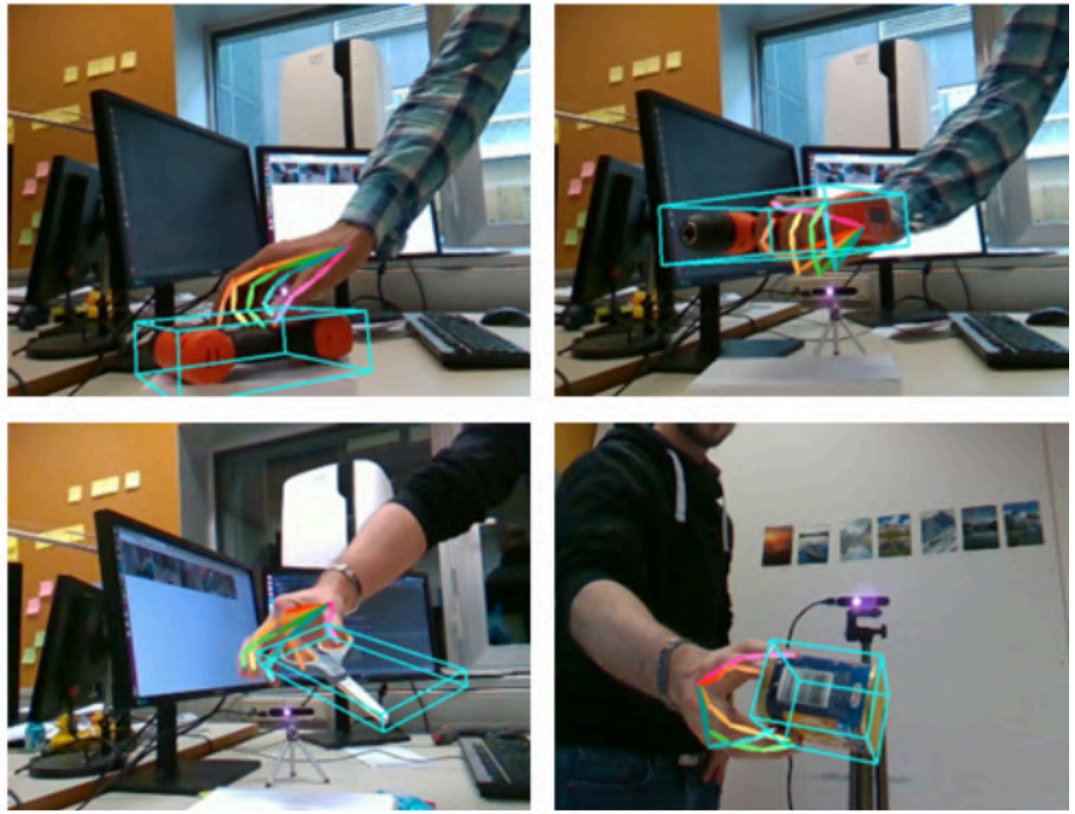
- 11,019 frames for training.
- 10,482 frames for evaluation.

Dataset

HO-3D Dataset

Provides a third-person view:

- The camera is placed separately from the subject
- Capturing the interactions from a distance.

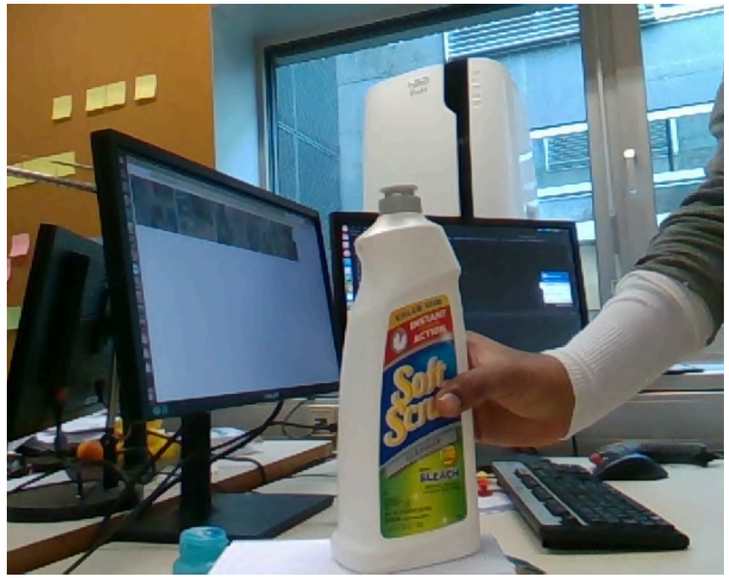


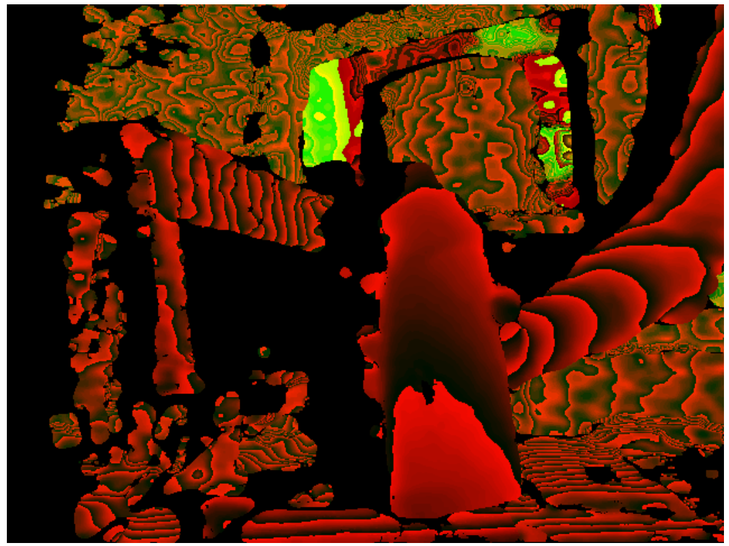
Dataset

HO-3D Dataset:

The dataset contains 77,558 frames in total, includes:

- 66,034 frames for training.
- 11,524 frames for evaluation.
- Subjects include 10 different people interacting with 10 distinct objects.



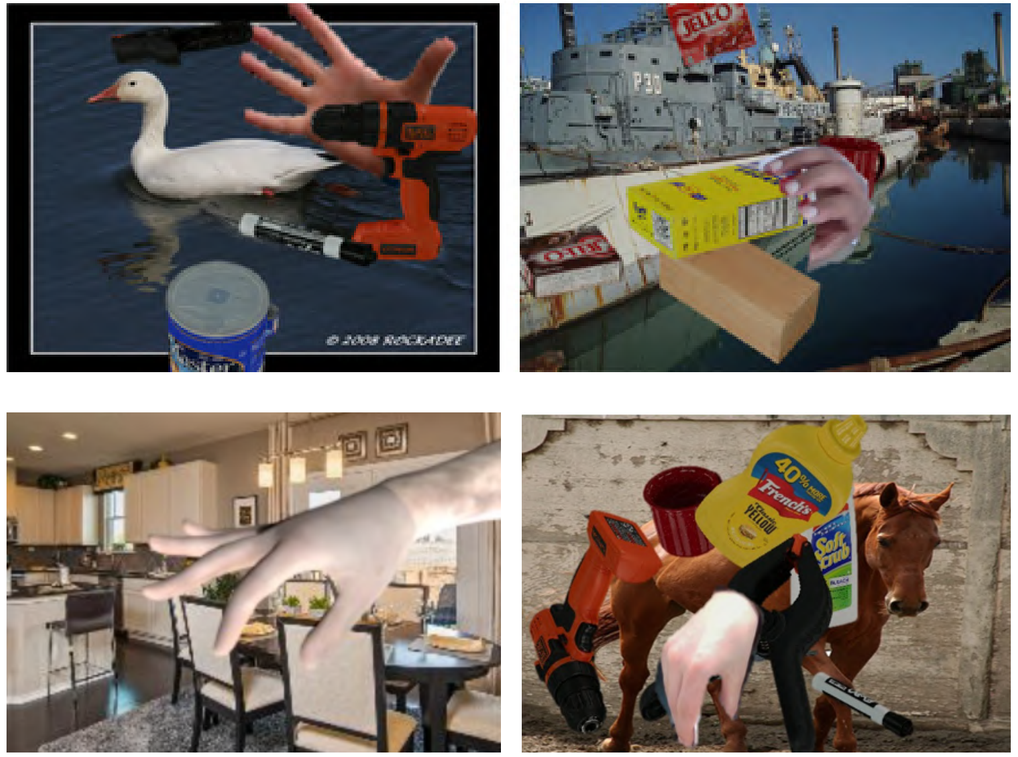


Dataset

ObMan Dataset:

A Synthetic dataset:

- The images are computer-generated
- Focuses on hand-object interactions



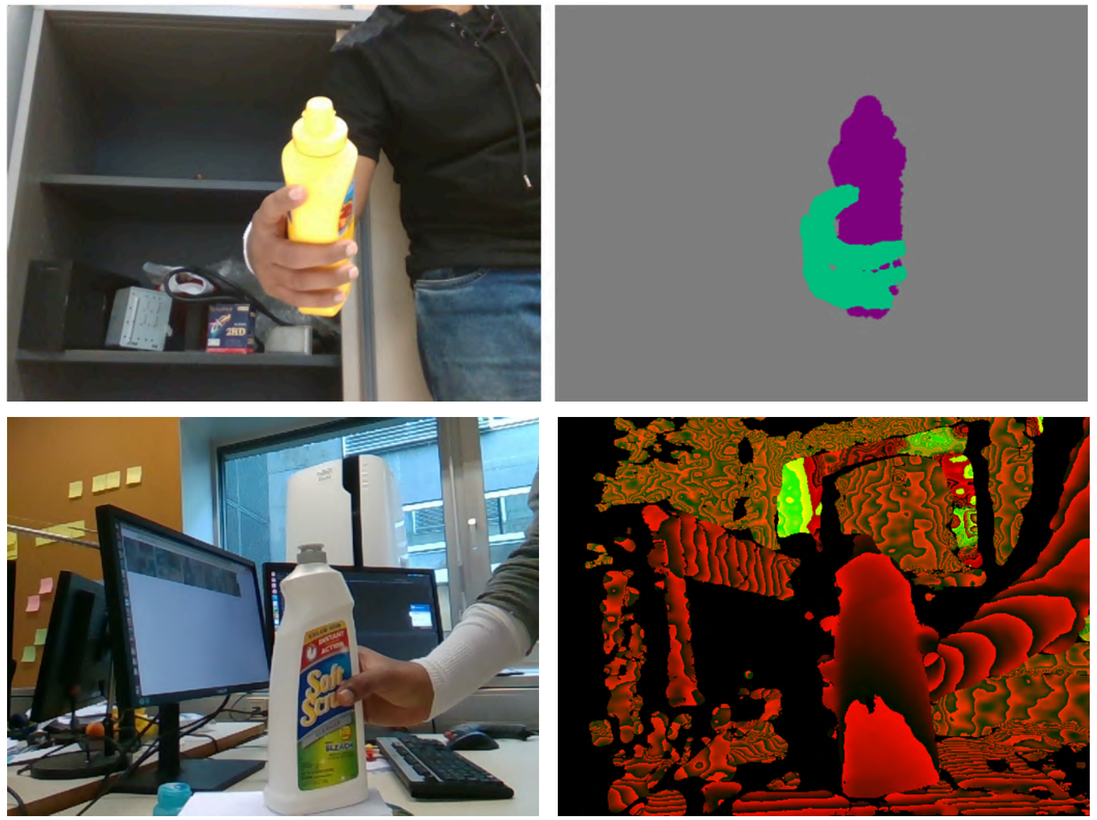
Dataset

ObMan Dataset:

ObMan includes:

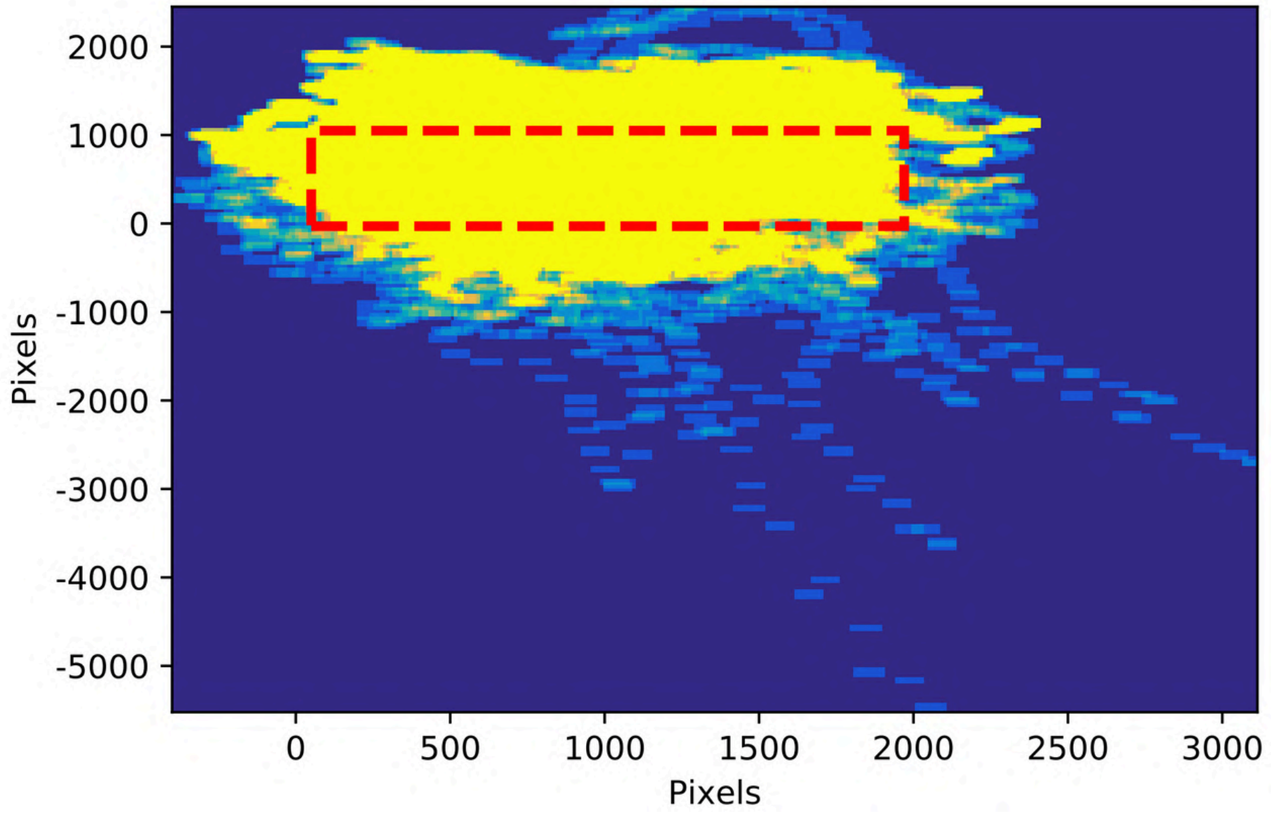
- 141,550 frames for training.
- 6,463 frames for validation.
- 6,285 frames for evaluation.

=> Over 150.000 frames in total!



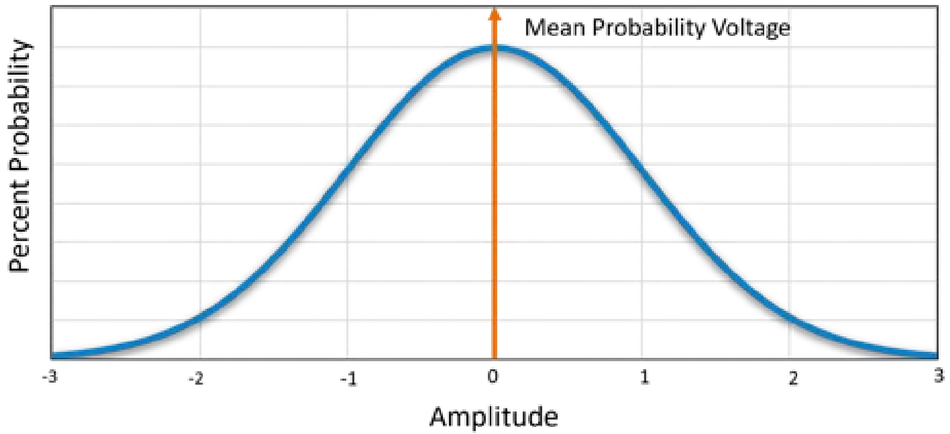
Implementation

- Dataset Issue of the First-Person Hand Action dataset
- Model Choice: Regression-based model + ResNet



Implementation

- 2D points are augmented with Gaussian noise (mean = 0, std = 10) for better robustness.



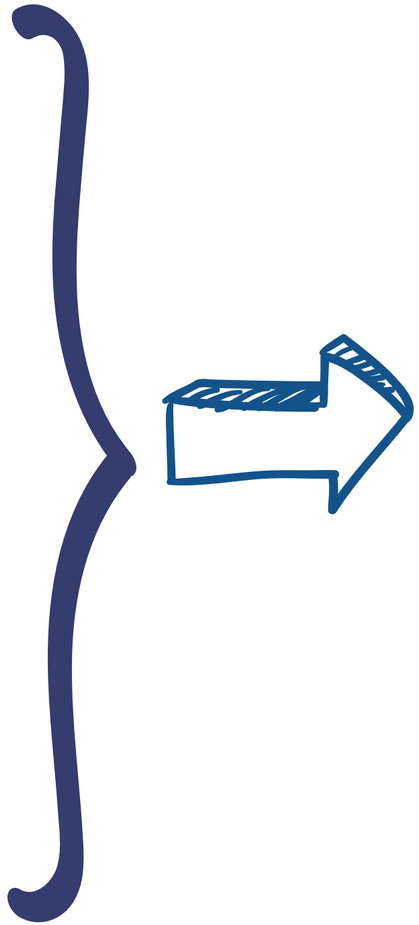
- Image Preprocessing: Images resized to 224×224 pixels before feeding into ResNet.

Implementation

Training

- **ResNet**: Initial learning rate = 0.001, decreased 0.1/100 steps for 5,000 epochs.

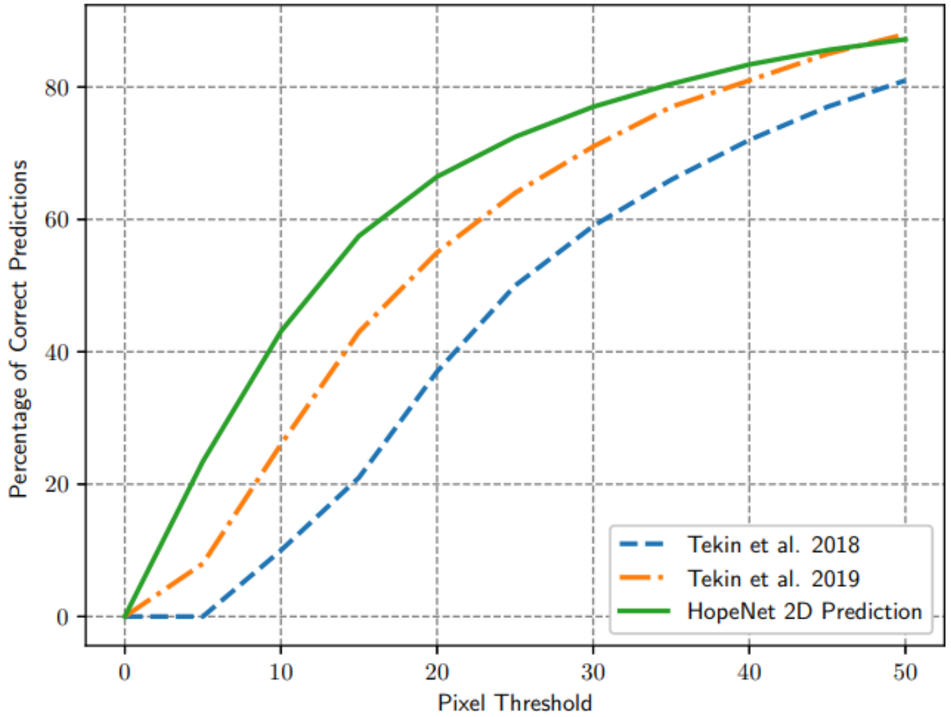
- **Graph Convolutional Network**: Initial learning rate = 0.001, decreased 0.9/4000 steps for 10,000 epochs.

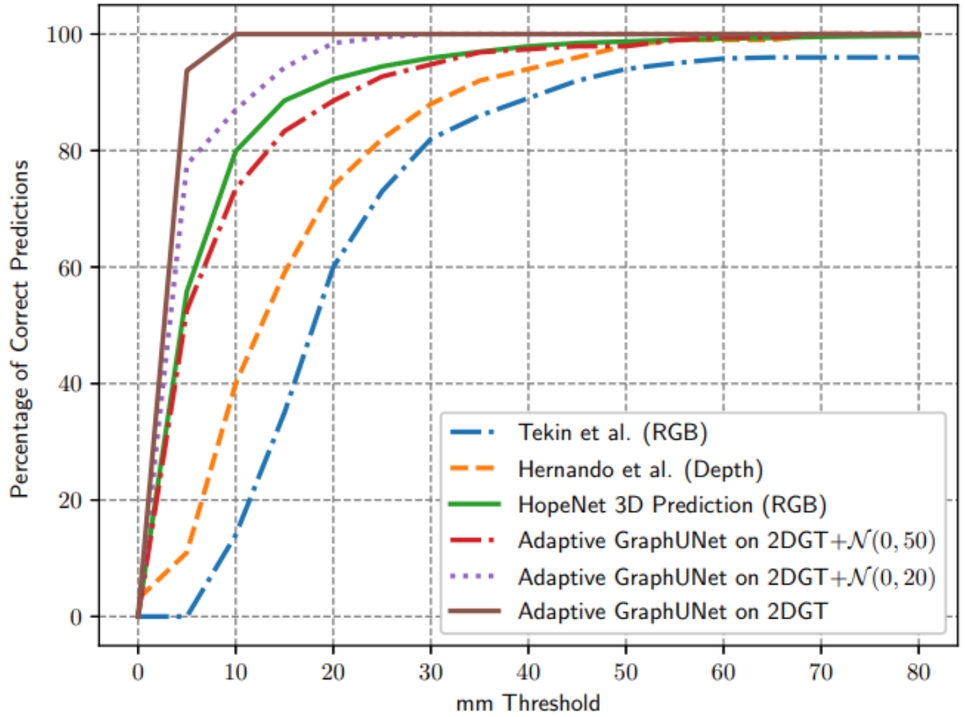


End-to-End Training: Additional 5,000 epochs with adjusted learning rates

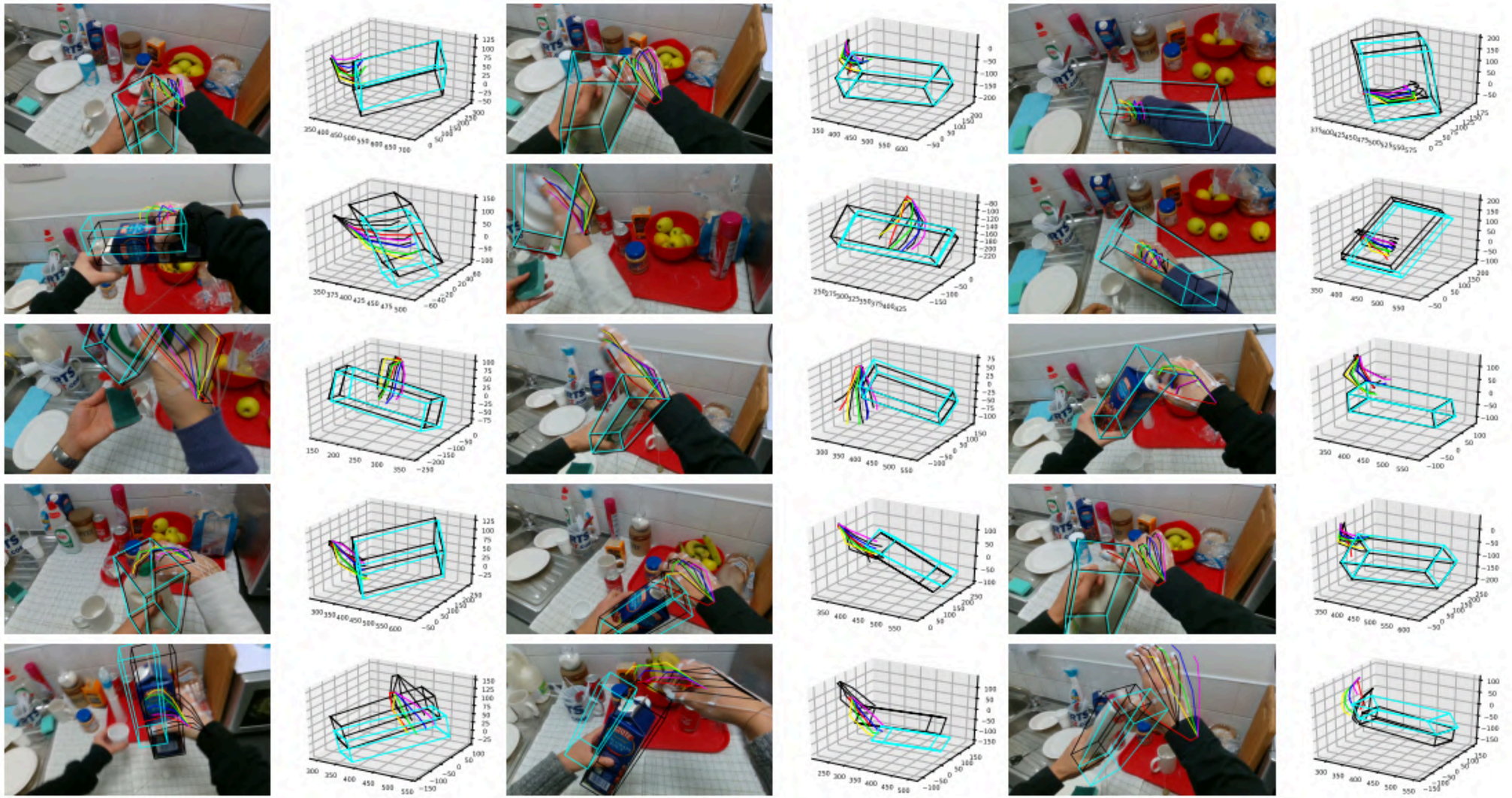
Evaluations

First-Person Hand Action Dataset

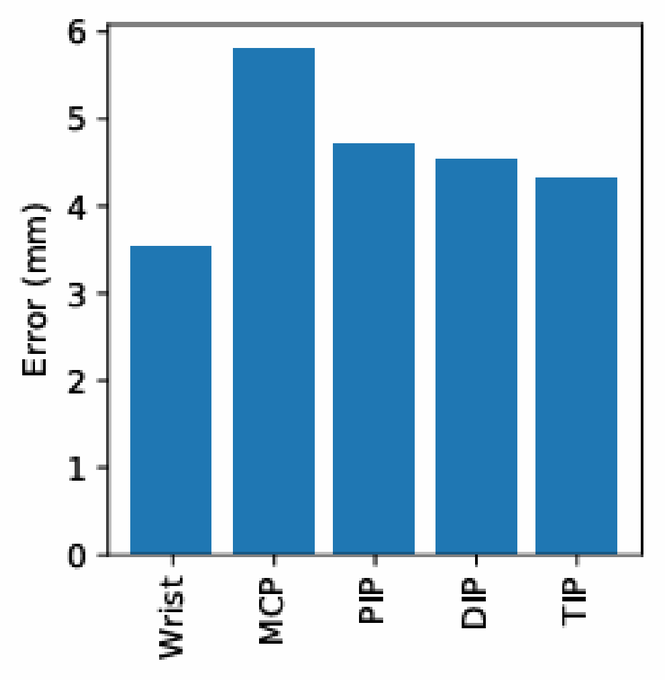


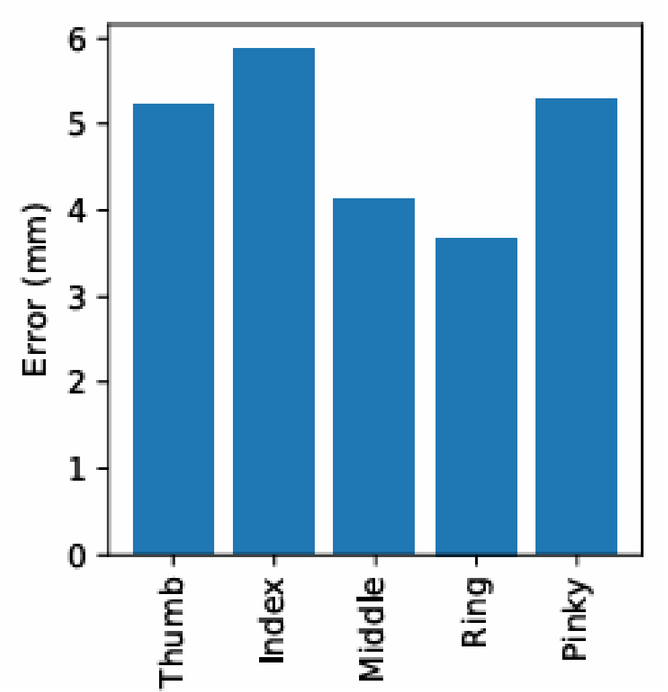


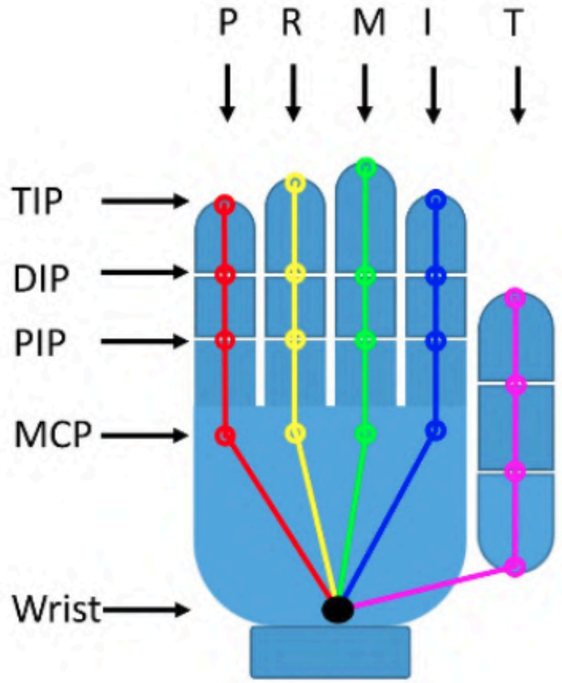
Evaluations

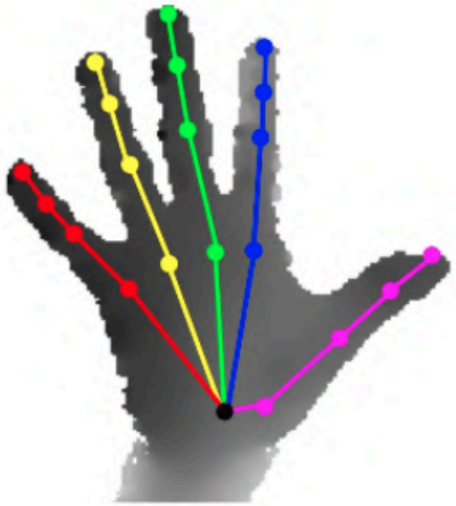


Evaluations









Evaluations

Table 1

Architecture	Average Error (mm)
Fully Connected	185.18
Adaptive Graph Convolution	68.93
Adaptive Graph U-Net	**6.81**

→ The Adaptive Graph U-net performs better than the other methods by a large margin

Evaluations

Table 2

Pooling method	Average Error (mm)
gPool [5]	153.28
Fixed Pooling Layers	7.41
Trainable Pooling	**6.81**

→ We see that by using a trainable pooling more efficiently and also by not breaking apart the graph after pooling, our pooling layer performs better than gPool

Evaluations

Table 3

Initial Adjacency Matrix	Average Error (mm)
Zeros ($\mathbf{0}_{n\times n}$)	92805.02
Random Initialization	94.42
Ones ($\mathbf{1}_{n\times n}$)	63.25
Skeleton	12.91
Identity ($\mathbf{I}_{n\times n}$)	**6.81**

→ The model seems to learn best when it finds the relationship between the nodes starting with an unbiased (uninformative)

Conclusion

Insights

Their approach beats the state-of-the-art while also **running in real-time.**

Limitations: their model is well-suited for objects that are of similar size or shape to those seen in the dataset during training

For real-world applications, using **a larger dataset** including a greater variety of shapes and environments would help to improve the **estimation accuracies.**

Limitations: the objects with a non-convex geometry lacking a tight 3D bounding box would be a challenge for their technique.

Future Work

Temporal Information	Integrate into the graph-based model to enhance pose estimation and advance action detection.
Graph Classification	Incorporate for inferring categorical semantic information, useful for sign language detection and gesture understanding.
Adaptive Graph U-Net	Apply to diverse problems including graph completion, protein classification, mesh classification, and body pose estimation.

THANK YOU