Grasping Pattern Estimation Based on Co-occurrence of Object and Hand Shape

Takuya Kawakami, Tadashi Matsuo, Yoko Ogawa, Nobutaka Shimada,

Ritsumeikan University

Introduction

- An object as a tool has its own function. The function is closely related to how a human grasp it [1].

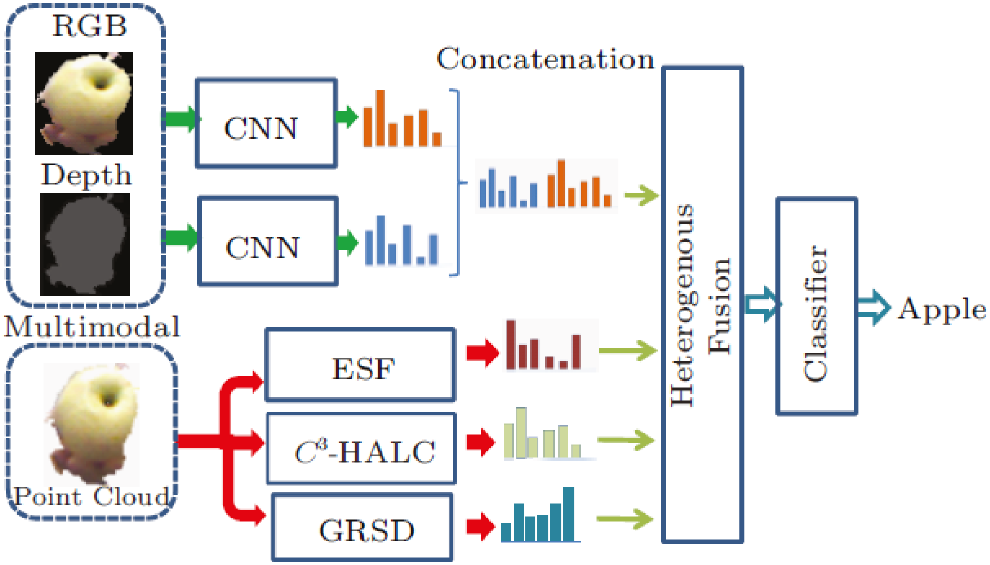
Can we estimate how to grasp an object from the object itself?

It will be useful for object recognition and robot manipulation.

[1] N. Kamakura, “Shape of hand and Hand motion”. Ishiyaku Publishers, 1989.

Related work

- Xiong Lv et al., "RGB-D Hand-Held Object Recognition Based on Heterogeneous Feature Fusion",
Journal of Computer Science and Technology(2015)

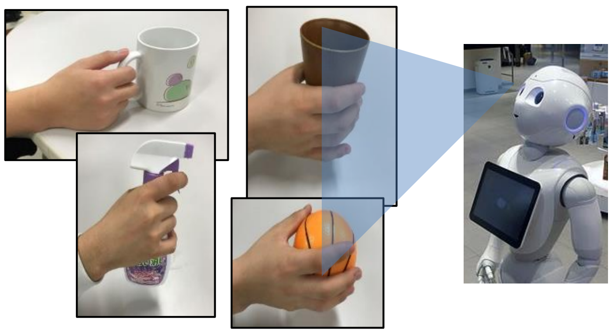


They achieved highly accurate classification by utilizing how to grasp an object, but…

- It estimates only an object label (not how to grasp it).
- All teacher labels must be given manually.

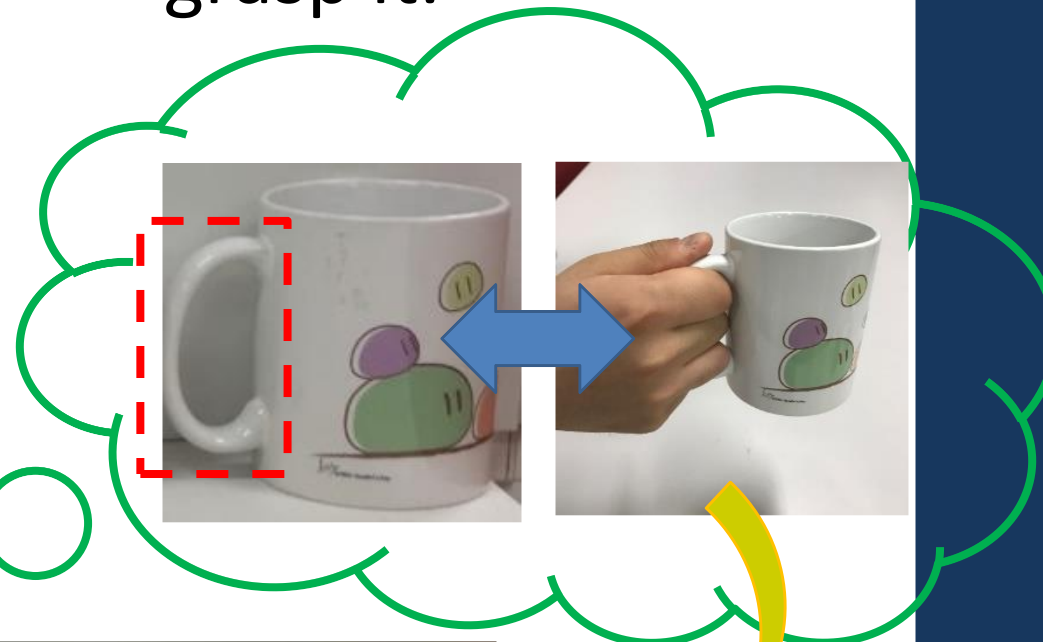
Our goal

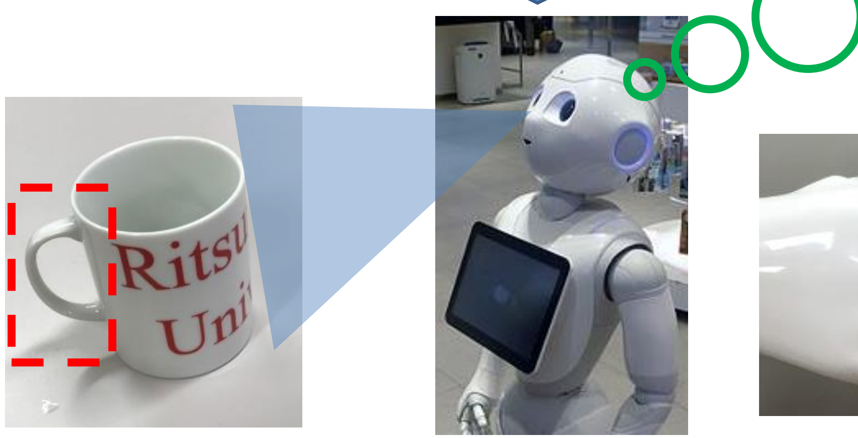
Training



Learn human interactions without teacher labels.

Recall how a human grasp it.

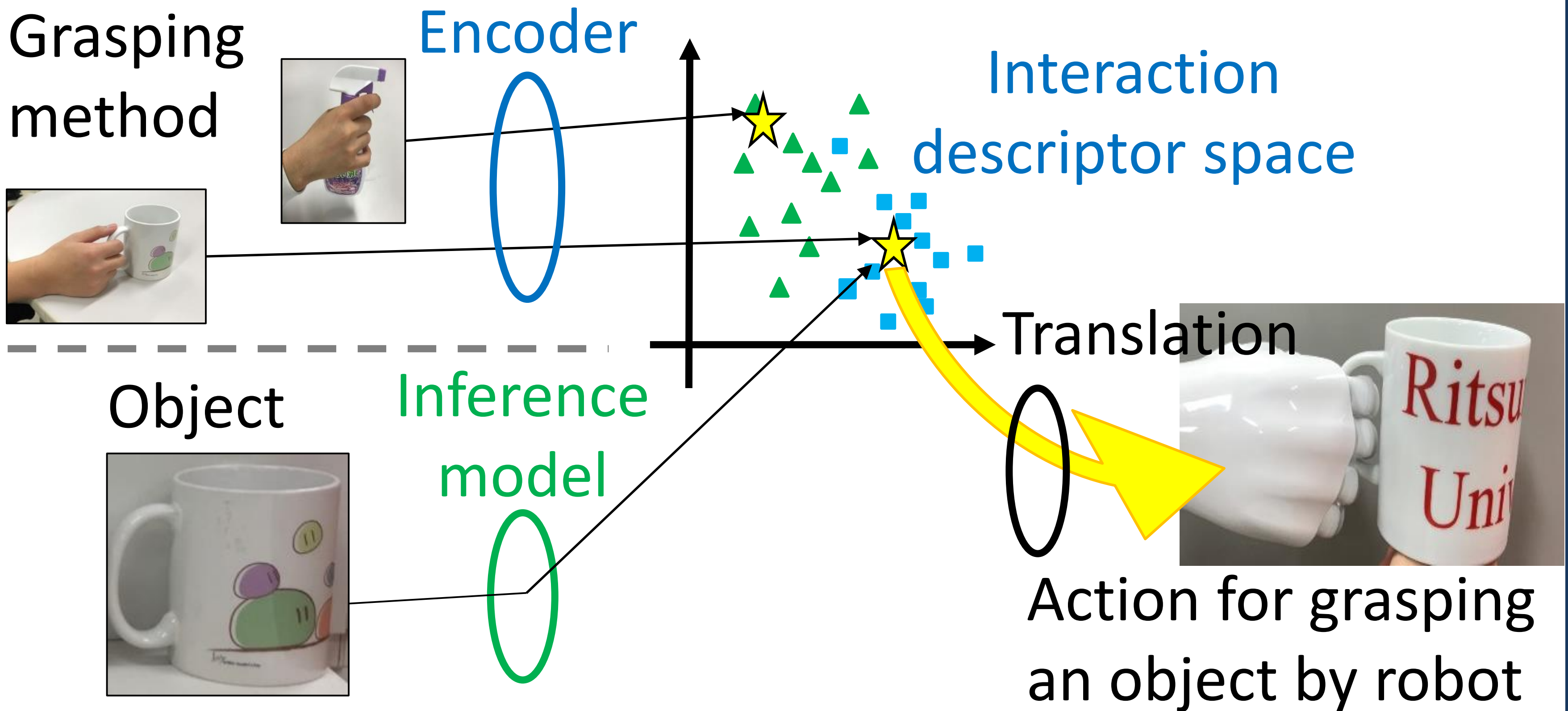




Make action to grasp it.

Proposed method

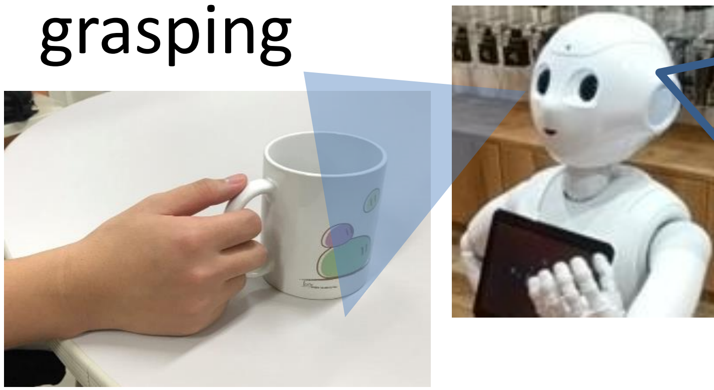
- We generate an interaction descriptor, a numeral representation of a human grasping method.
- And then we make an inference model to learn the relation between object and grasping method.



Grasping image

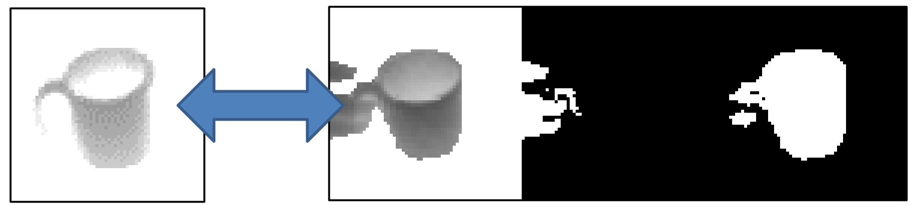
Automatically collect
Images for learning

Observing human
grasping



I_{obj}
Object image

I_{grasp}
Grasping image

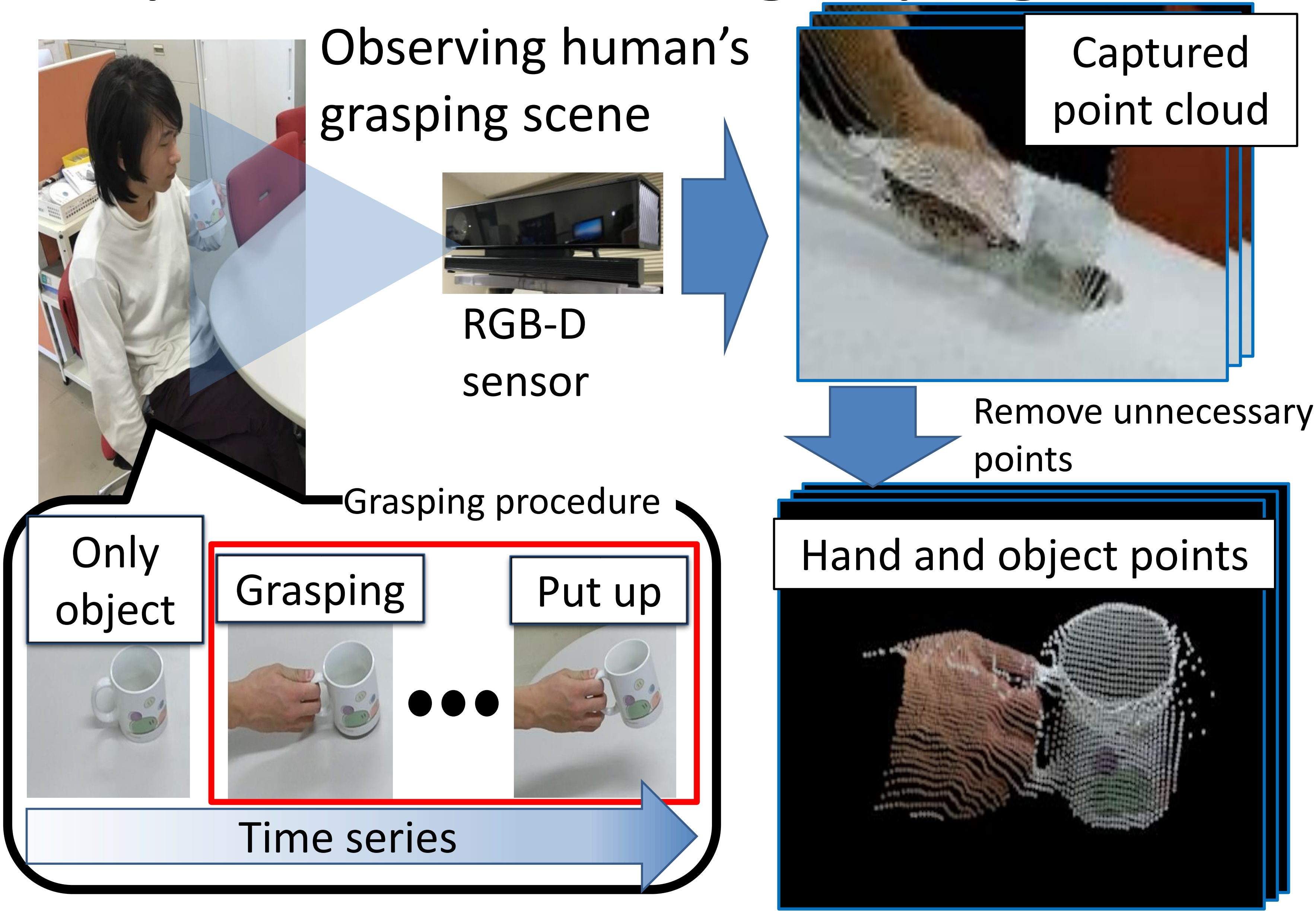


Depth

Depth Hand mask Object mask

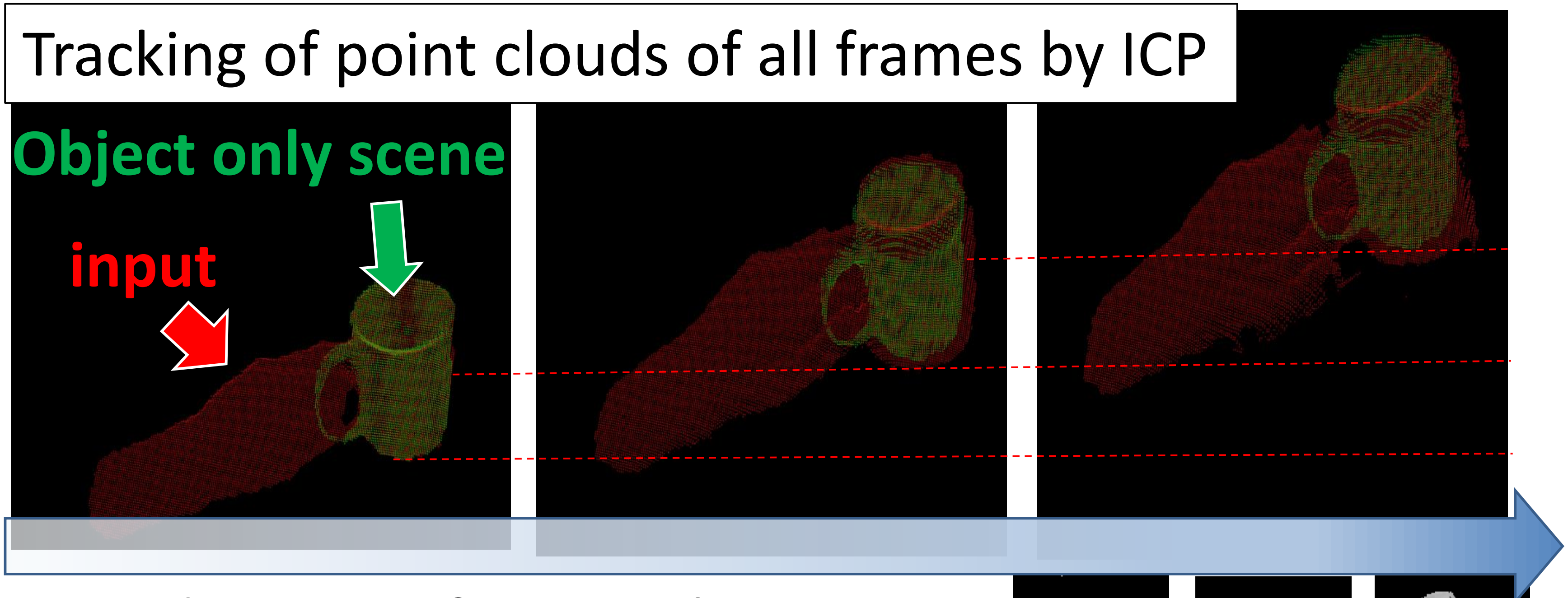
Grasping method is represented as a grasping image.
It consists of a depth image, hand mask and object mask.
It is paired with the corresponding object image.

Capture of human's grasping scene

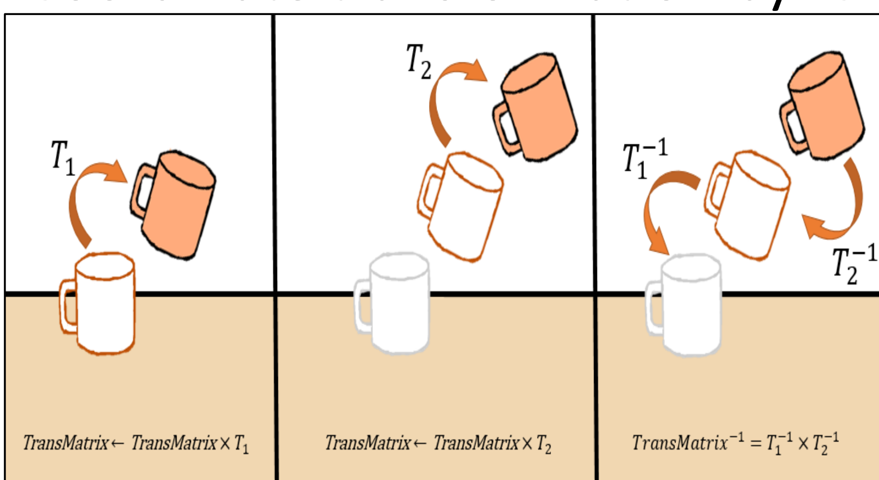


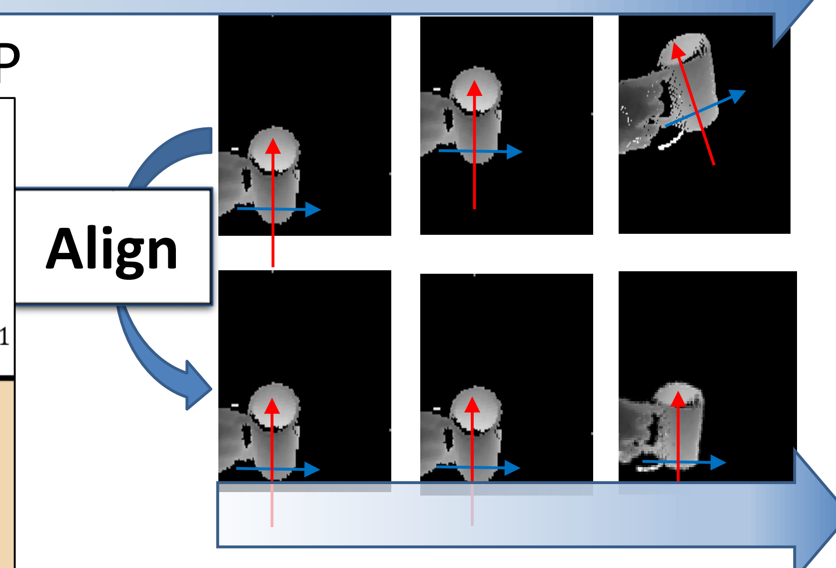
Alignment based on object

Tracking of point clouds of all frames by ICP



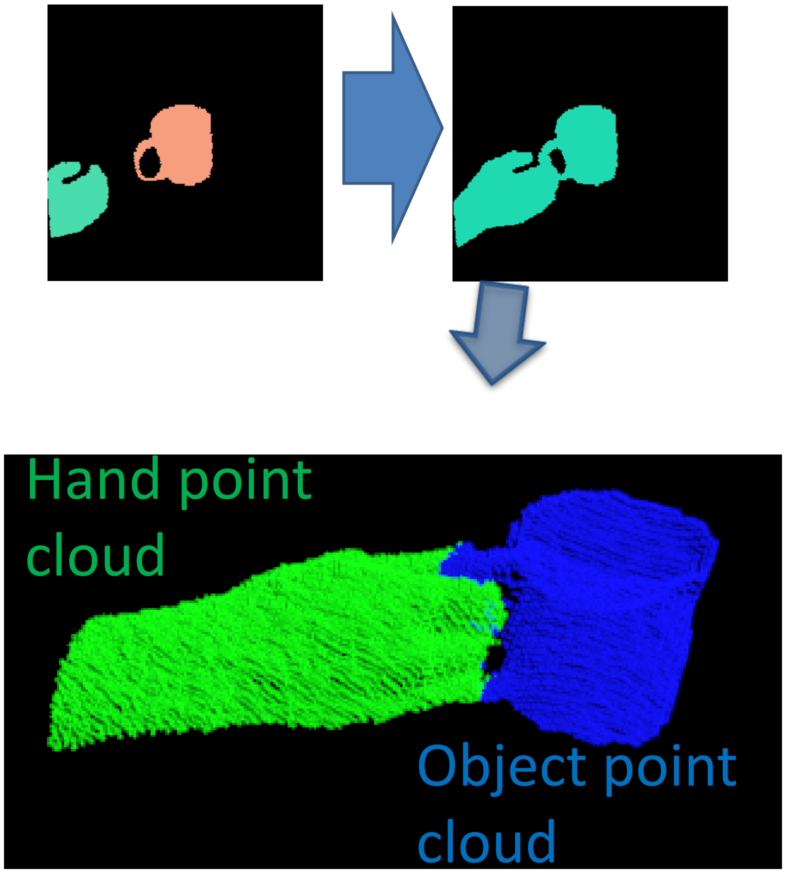
Coordinate transformation by ICP

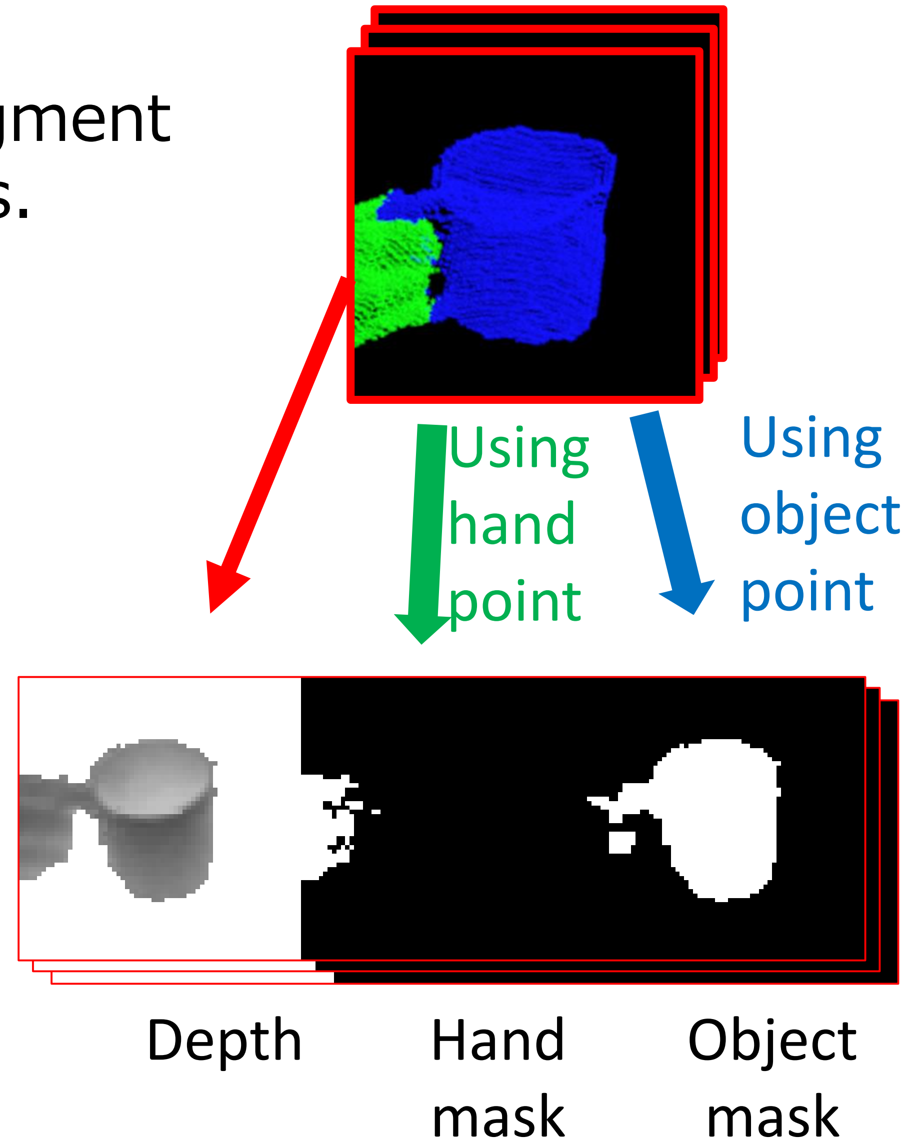




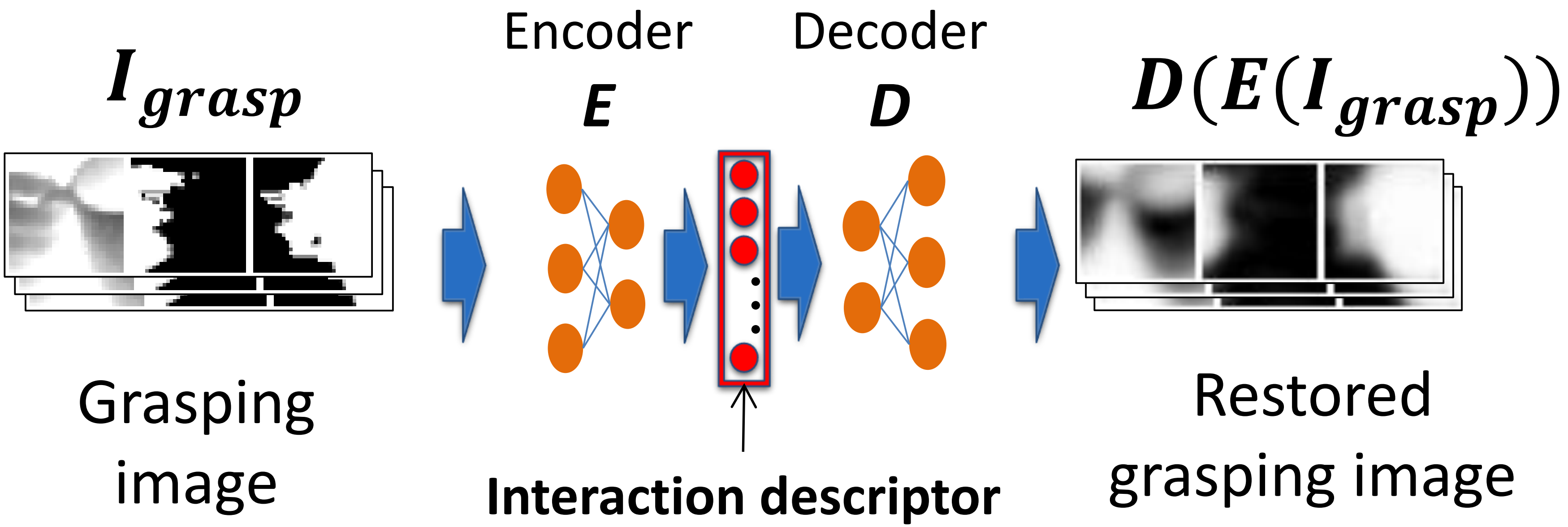
Segmentation of a hand and an object

Just before changing the number of regions, we segment hand points / object points.





Interaction descriptor

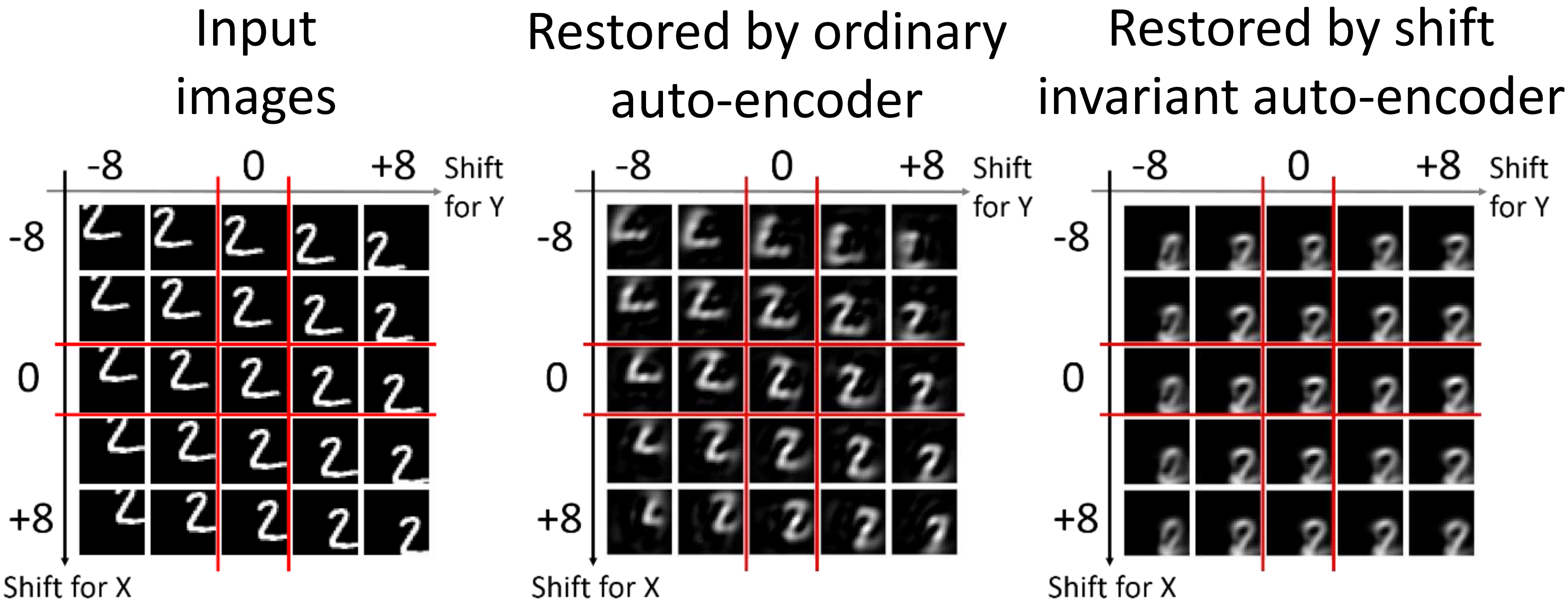


E and D are trained by minimizing restoration error without teacher labels.

A descriptor represents essence of an input.

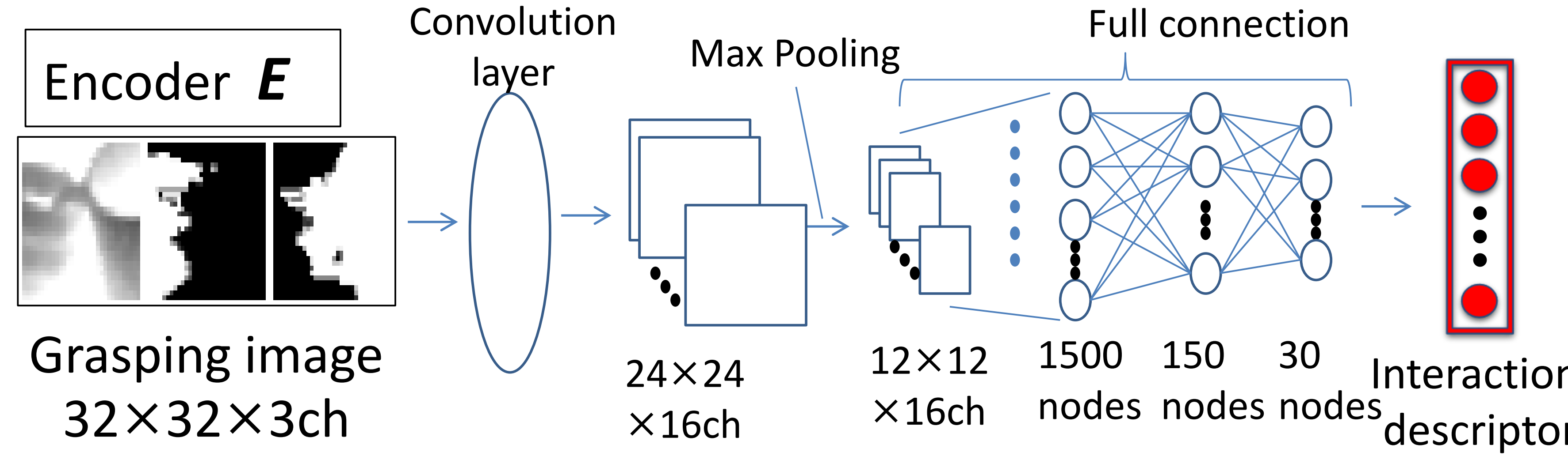
Shift invariant auto-encoder

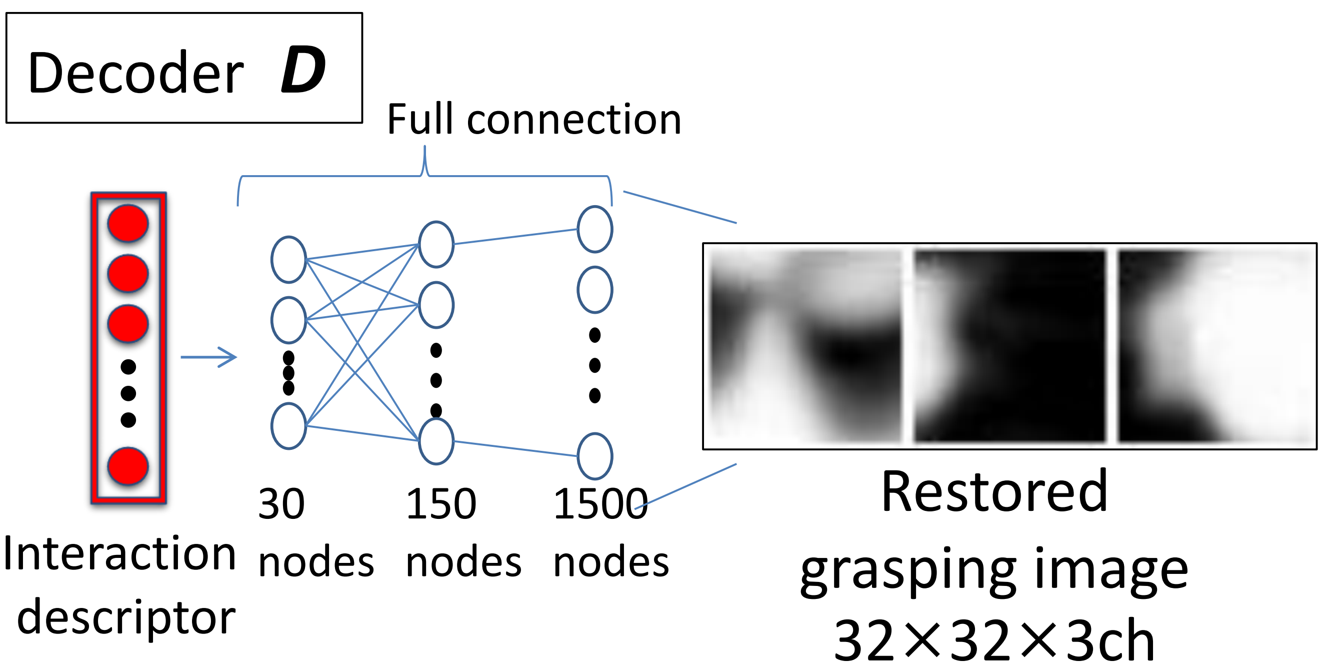
An ordinary auto-encoder encodes shape and position.
But spatial shift in grasping images is not important.
We use shift invariant auto-encoder to encode a shape itself. (descriptor includes shape information only)



T. Matsuo, et al., “Transform invariant auto-encoder,” IROS 2017, https://doi.org/10.1109/IROS.2017.8206047

Structure of auto-encoder





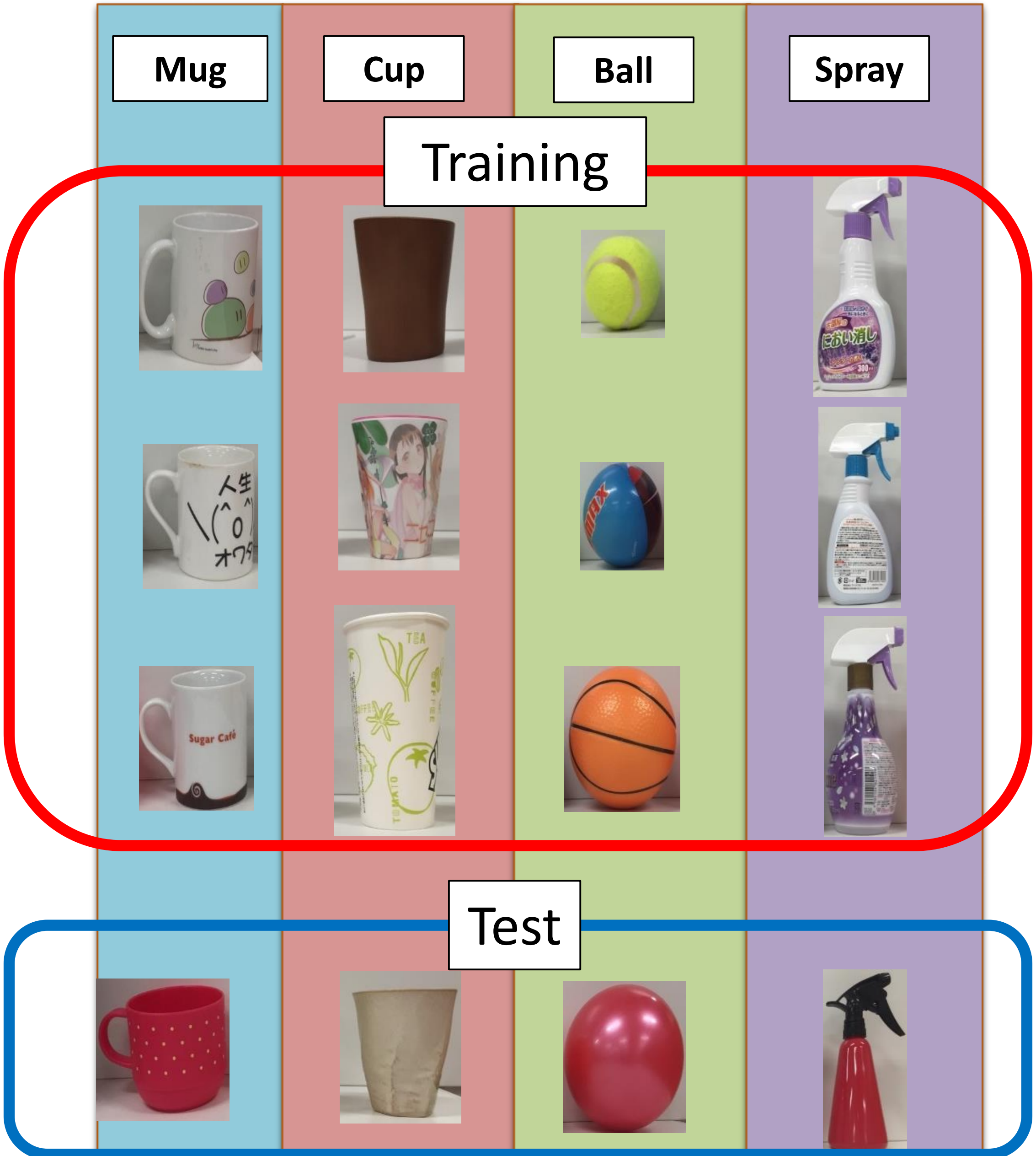
Parameters

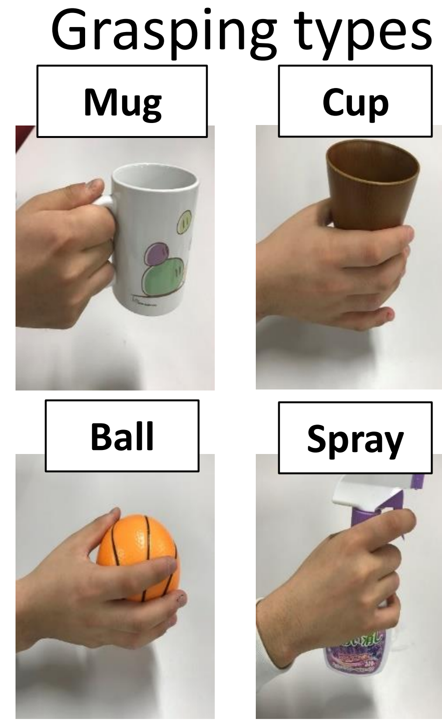
Input size: 32×32×3

Descriptor dimension: 30

Kernel size: 9×9

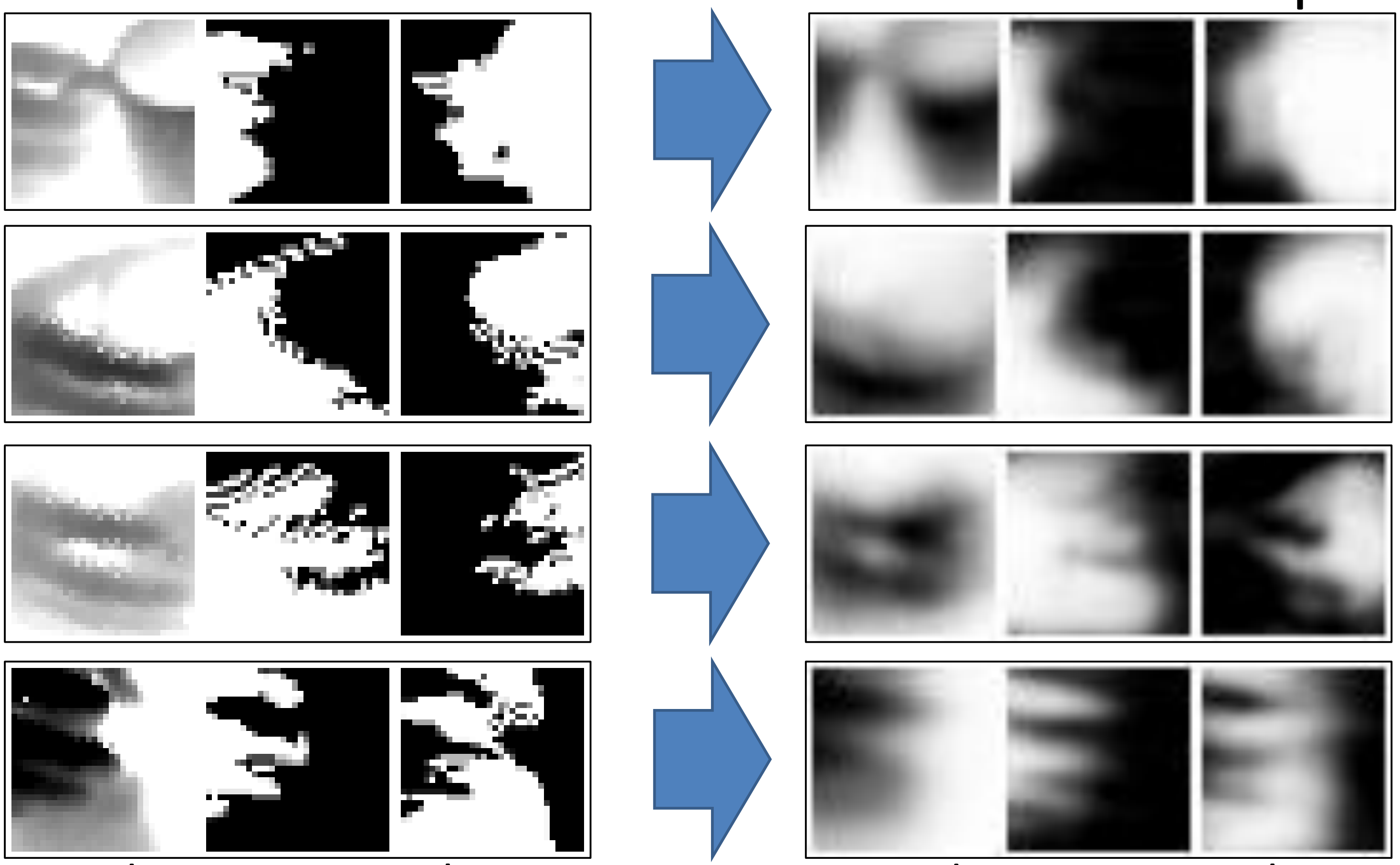
Objects and grasping types





Training images:
80scenes×12kinds
= 960
Test images:
80scenes×4kinds
= 320

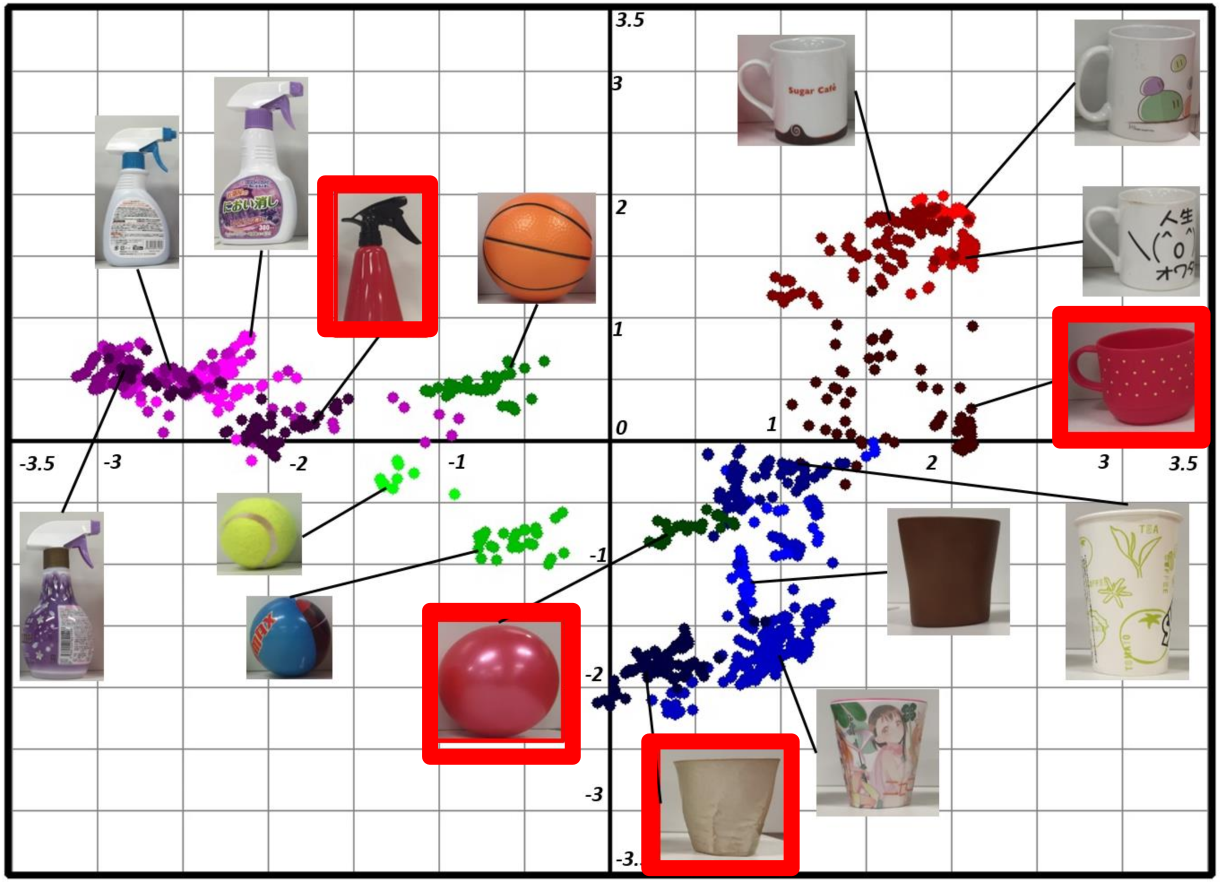
Restored grasping images



Interaction descriptor has approximate shape information.

Distribution of interaction descriptors

2nd principal component

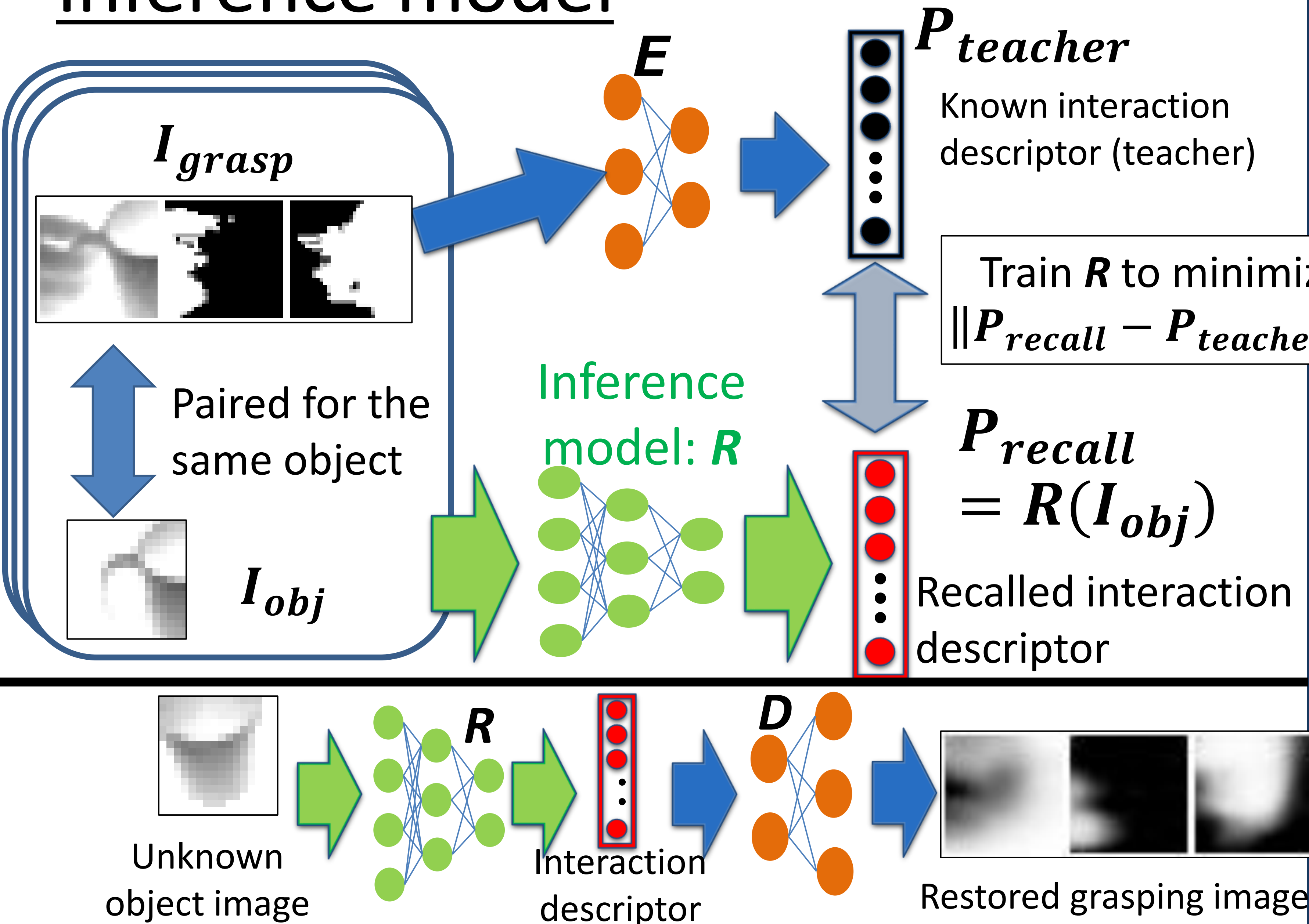


1st principal component

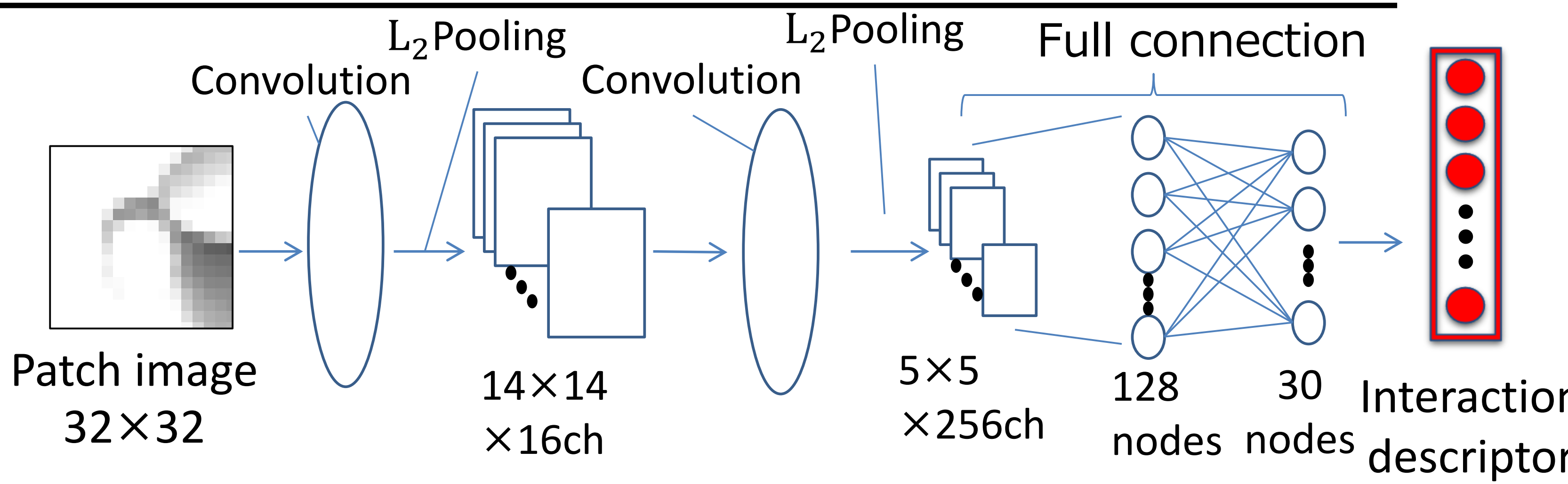
Non-frame:
Object for training

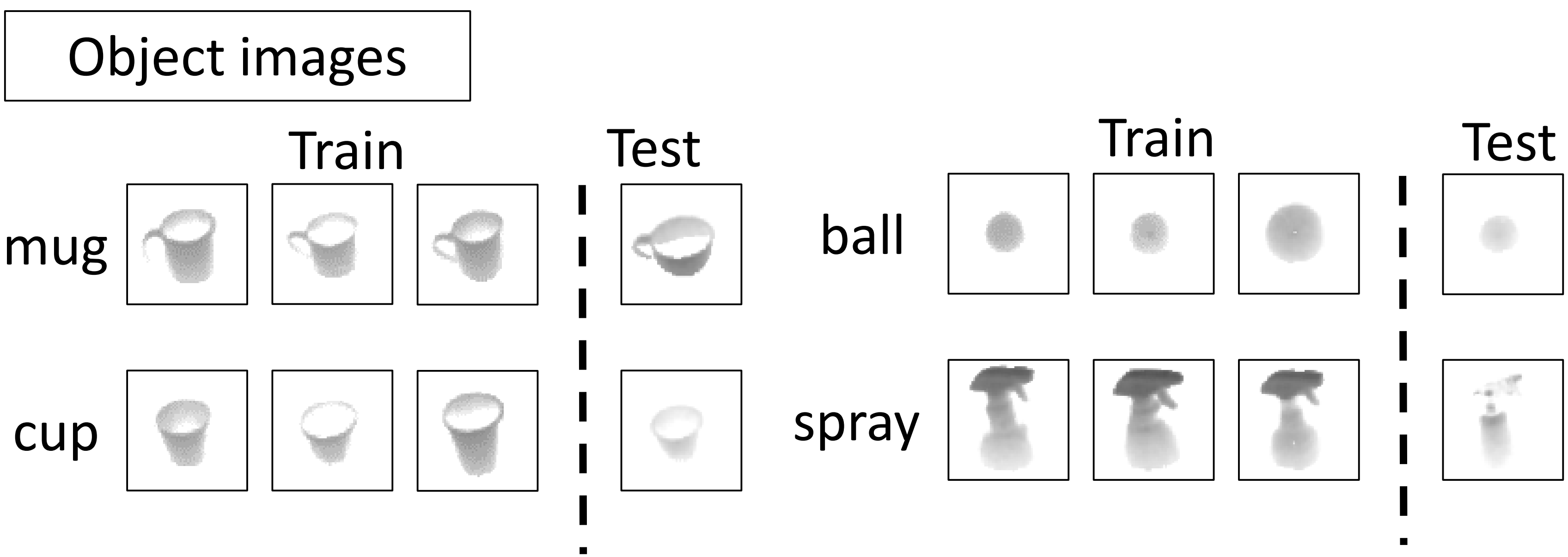
Red frame:
Object for test

Inference model

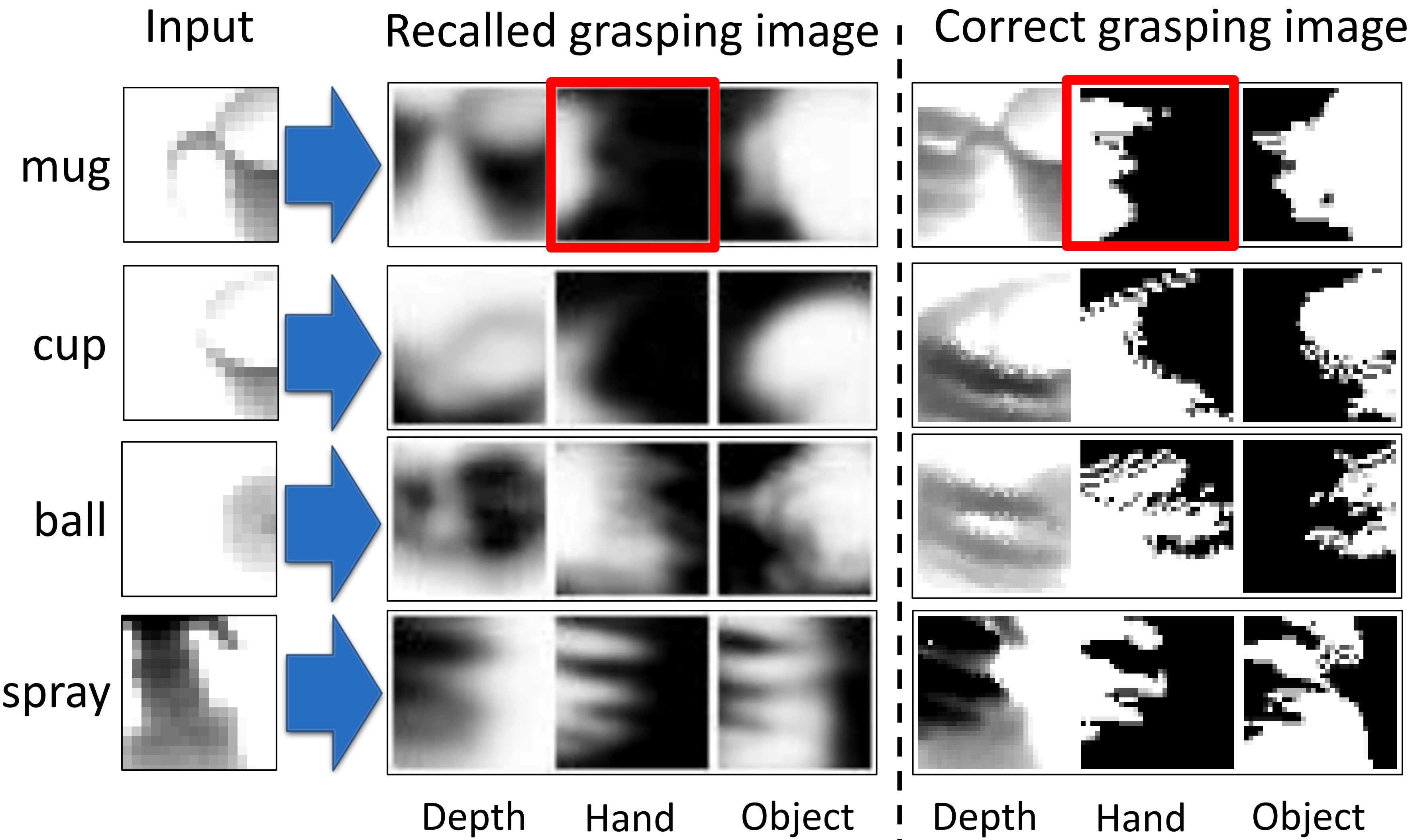


Structure of the inference model



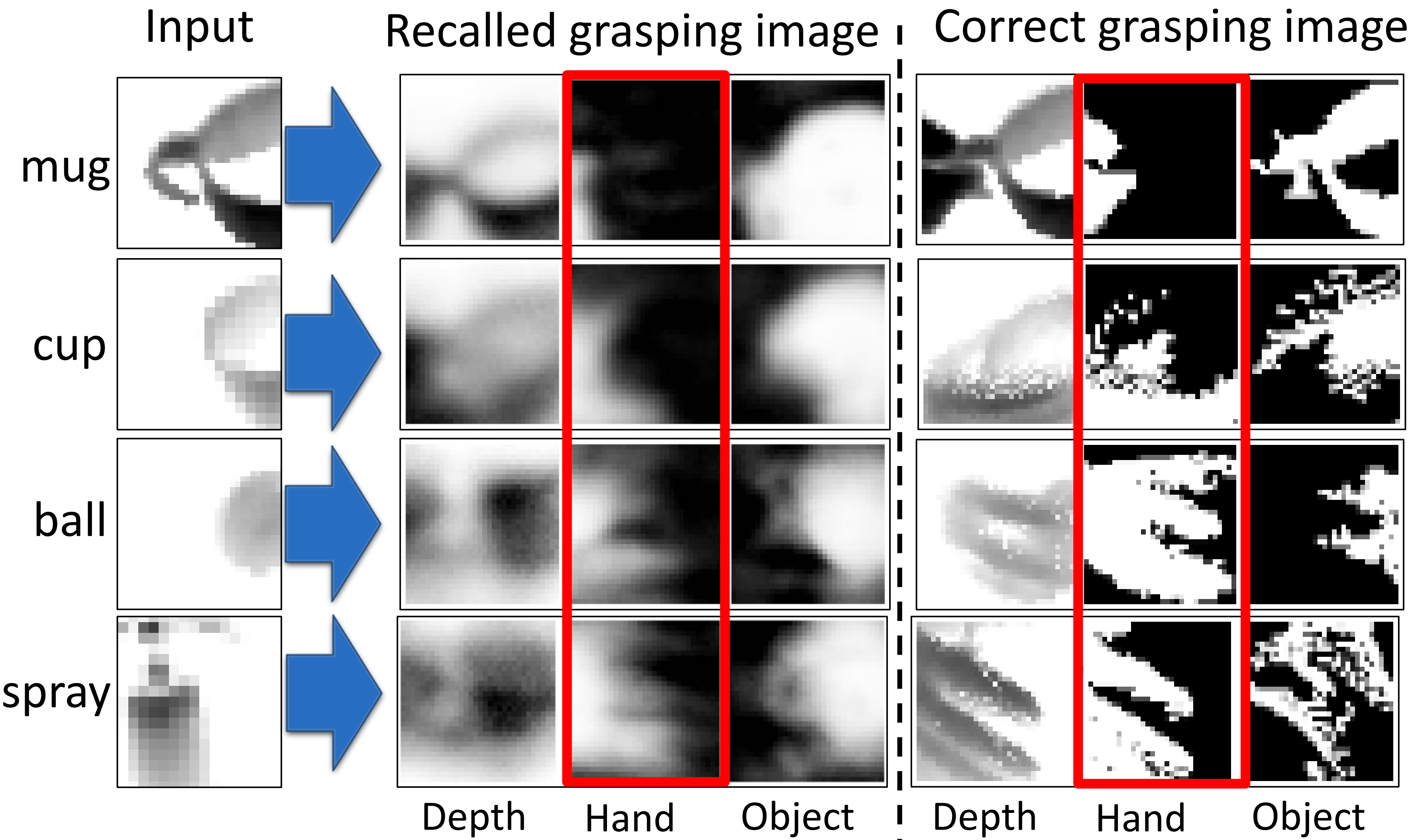


Recalled grasping images (train)



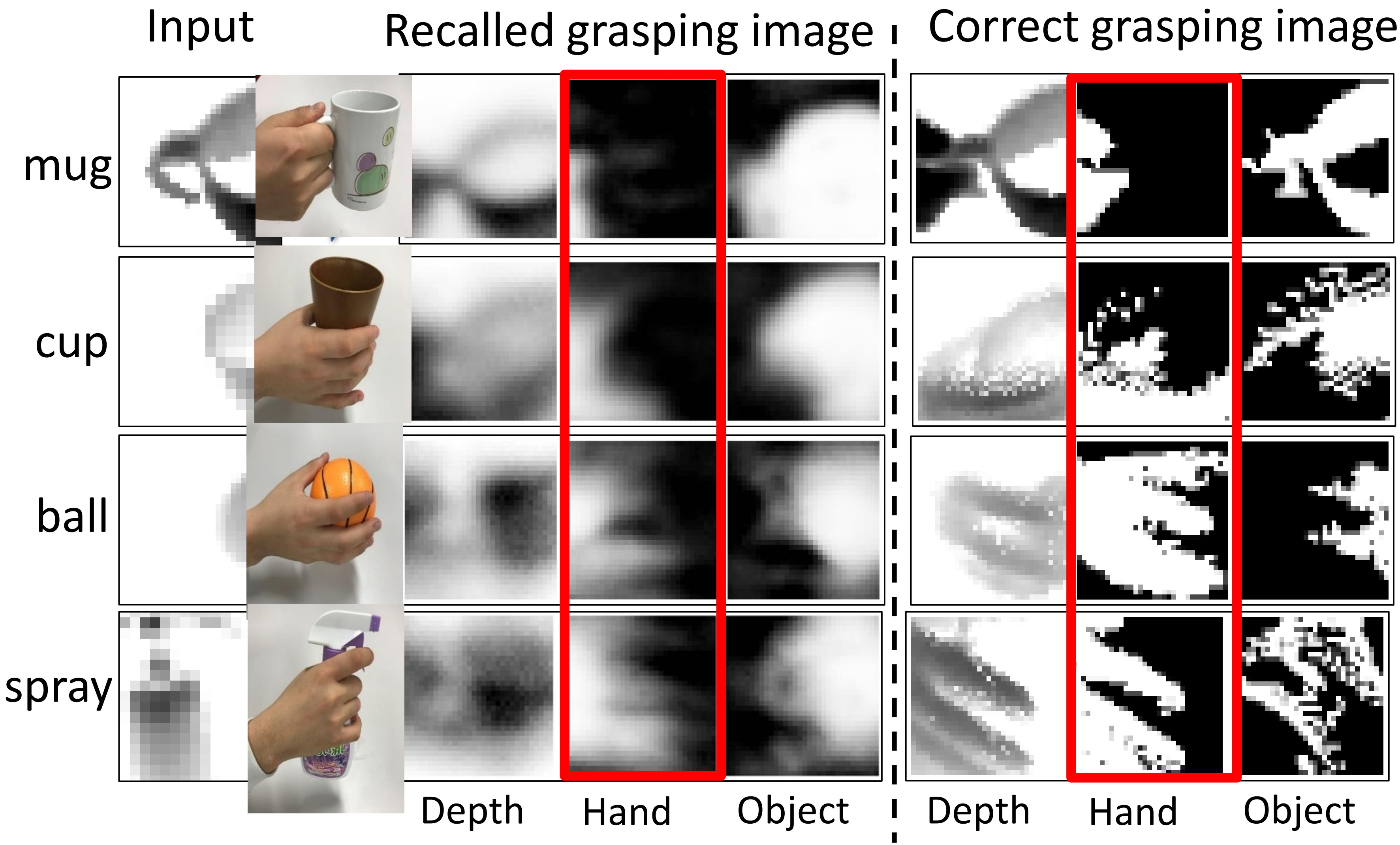
The inference model successfully recalls grasping images.

Recalled grasping images (test)



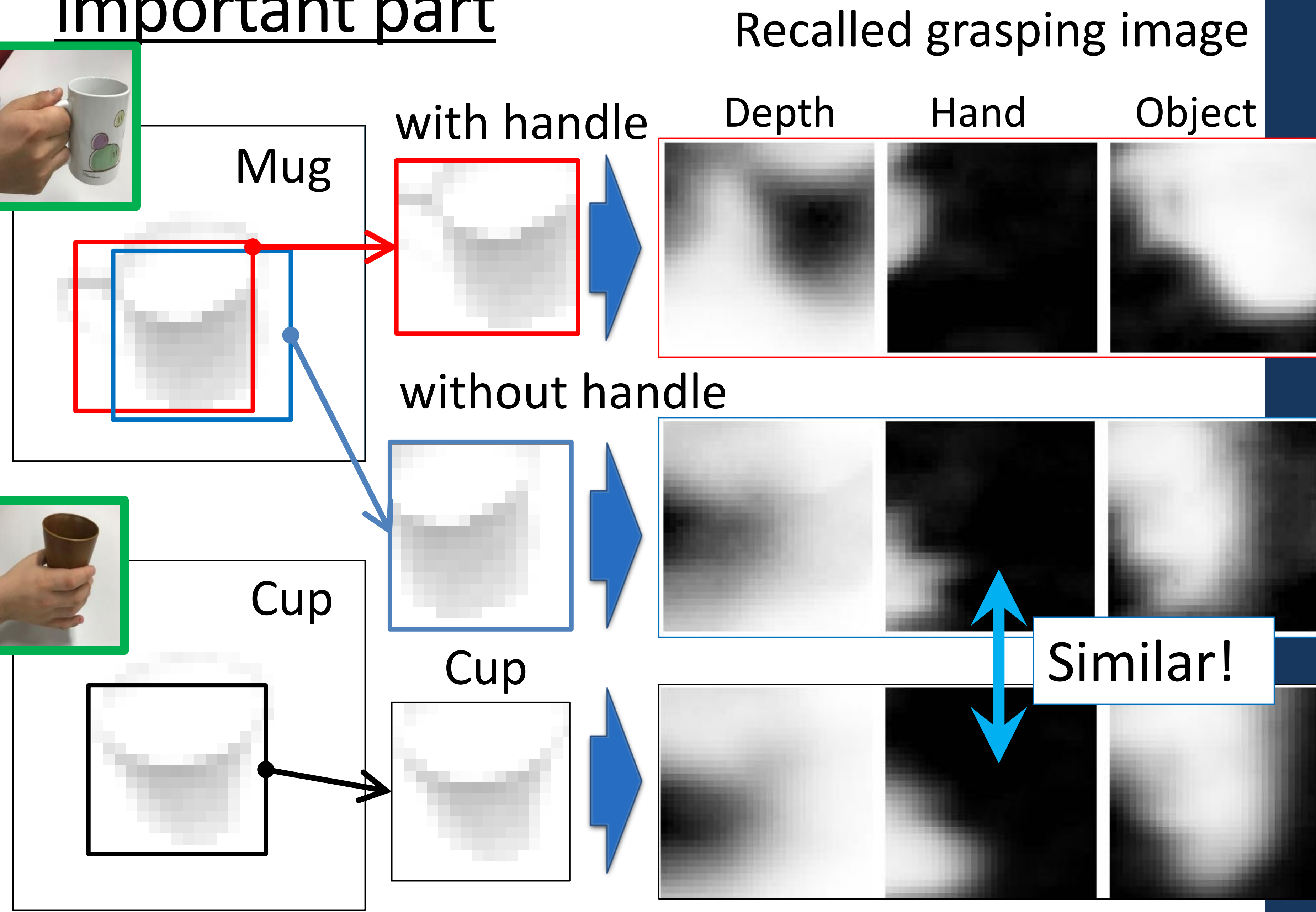
The model approximately recalls hand region masks.

Recalled grasping images (test)

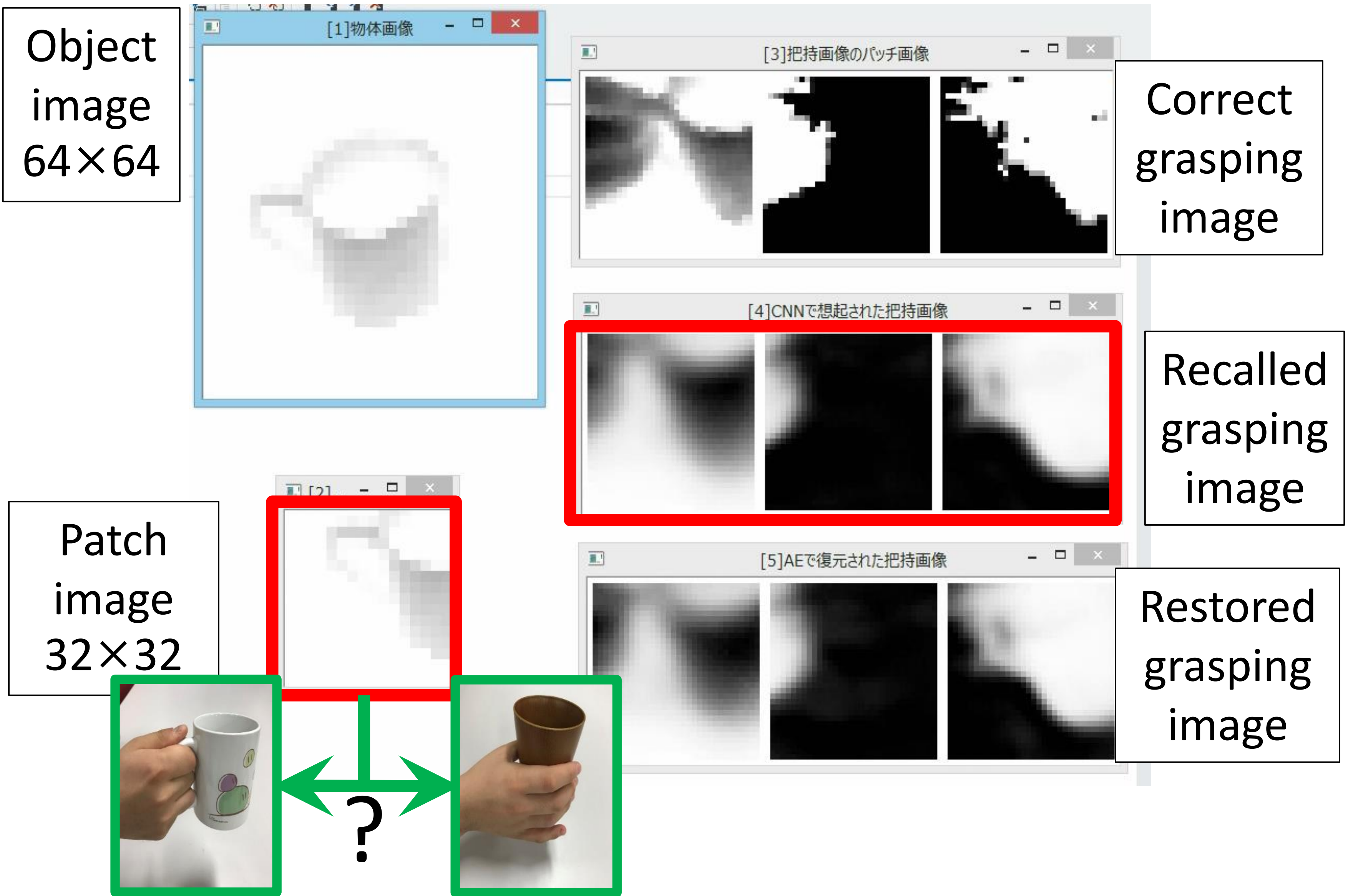


The model approximately recalls hand region masks.

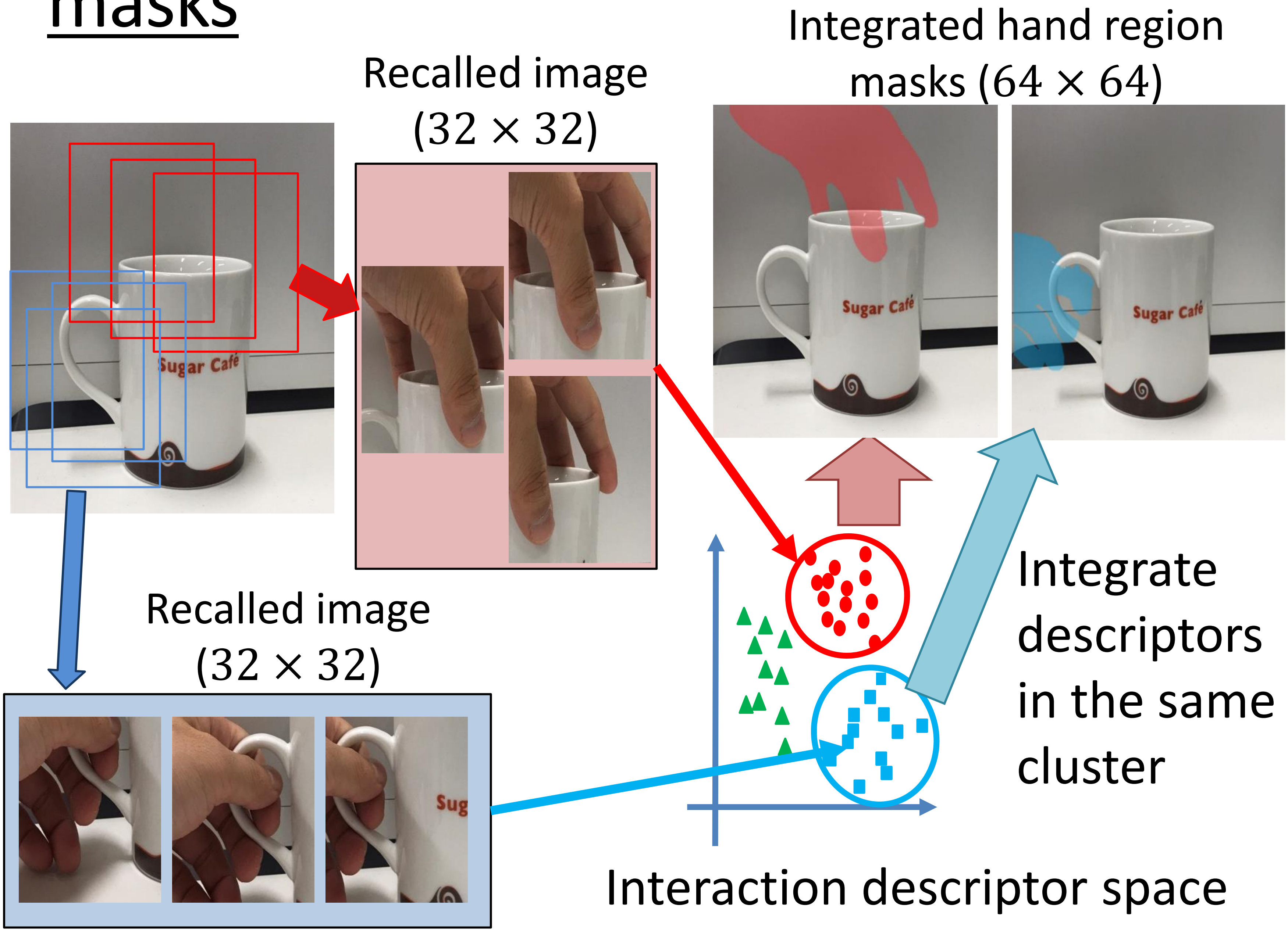
Recall from images with/without an important part



Recall from images with/without an important part



Integration of recalled hand region masks



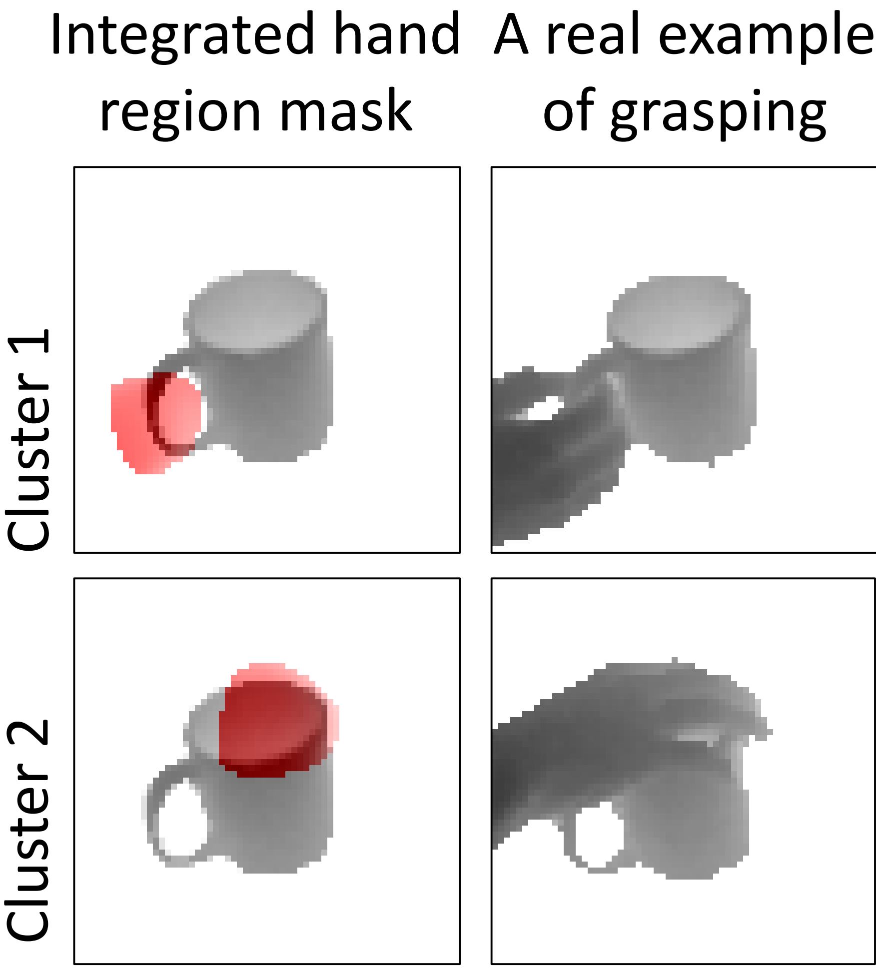
Multiple grasping types for object

To see part-specific inference, we train auto-encoder and inference model with below grasping types.

	Grasping type 1	Grasping type 2
mug		
cup		

	Grasping type 1	Grasping type 2
ball		
spray		

Integrated hand region mask



The integrated hand mask for cluster i is defined as:

$$P_i(x, y) = \frac{S_i(x, y)}{N_i(x, y)}$$

$S_i(x, y)$: Sum of recalled hand mask in the i-th cluster

$N_i(x, y)$: Number of non-zero at (x, y) of recalled hand mask in the i-th cluster

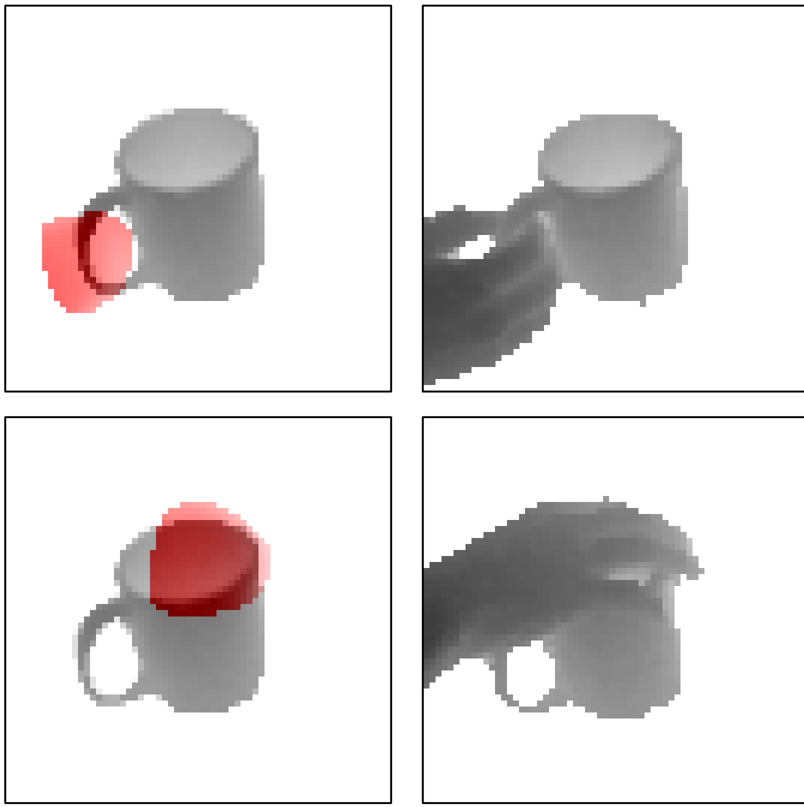
Integrated hand region mask indicates hand region when human grasps the object.

Integrated hand region mask

Integrated hand region mask | A real example of grasping

Cluster 1

Cluster 2



The integrated hand mask for cluster i is ... as:

$$\ldots) = \frac{S_i(x,y)}{N_i(x,y)}$$

$S_i(x,y)$: Sum of recalled

ha... in the i-th cluster

N... Number of non-

ze...) of recalled

hand mask in the i-th cluster

Integrated hand region mask indicate hand region when a human grasps the object.

Conclusion

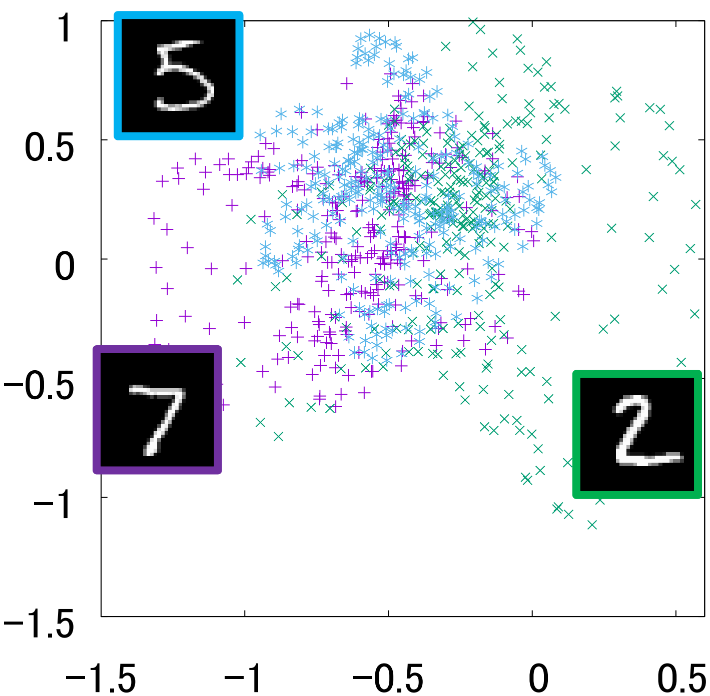
- We proposed a method to recall grasping method from an object. It is based on:
  - Interaction descriptor by shift invariant auto-encoder
    We can generate numeral representation of grasping method without teacher labels.
  - Inference model by CNN
    The relation between object shape and grasping method can be modeled by utilizing interaction descriptor.
- The proposed method can estimate hand region for grasping an object from the object itself.
- The proposed method will be useful for robot manipulator.

Distribution of descriptors from shift invariant auto-encoder

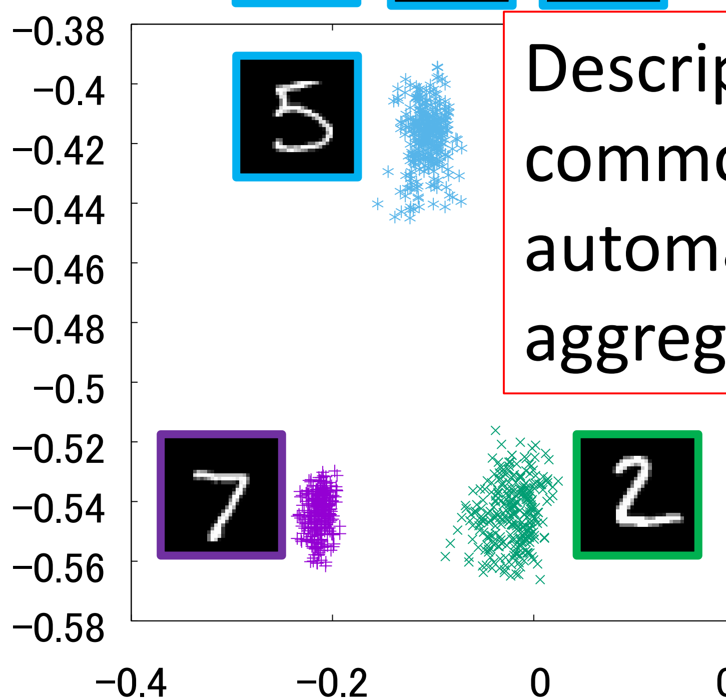
We trained auto-encoders with shifted MNIST training images.

Input:32 × 32
Descriptor dim: 30
Max shift width: 8

Distributions of descriptors from shifted test images such as .



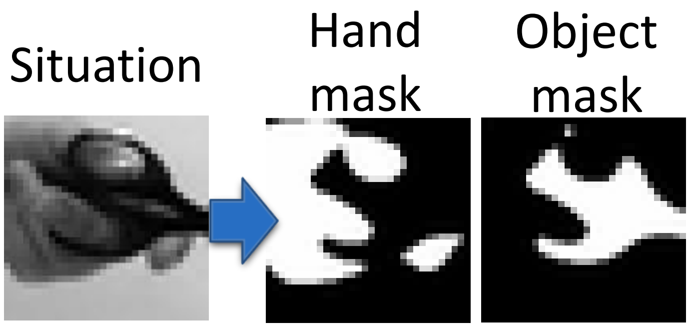
Ordinary auto-encoder



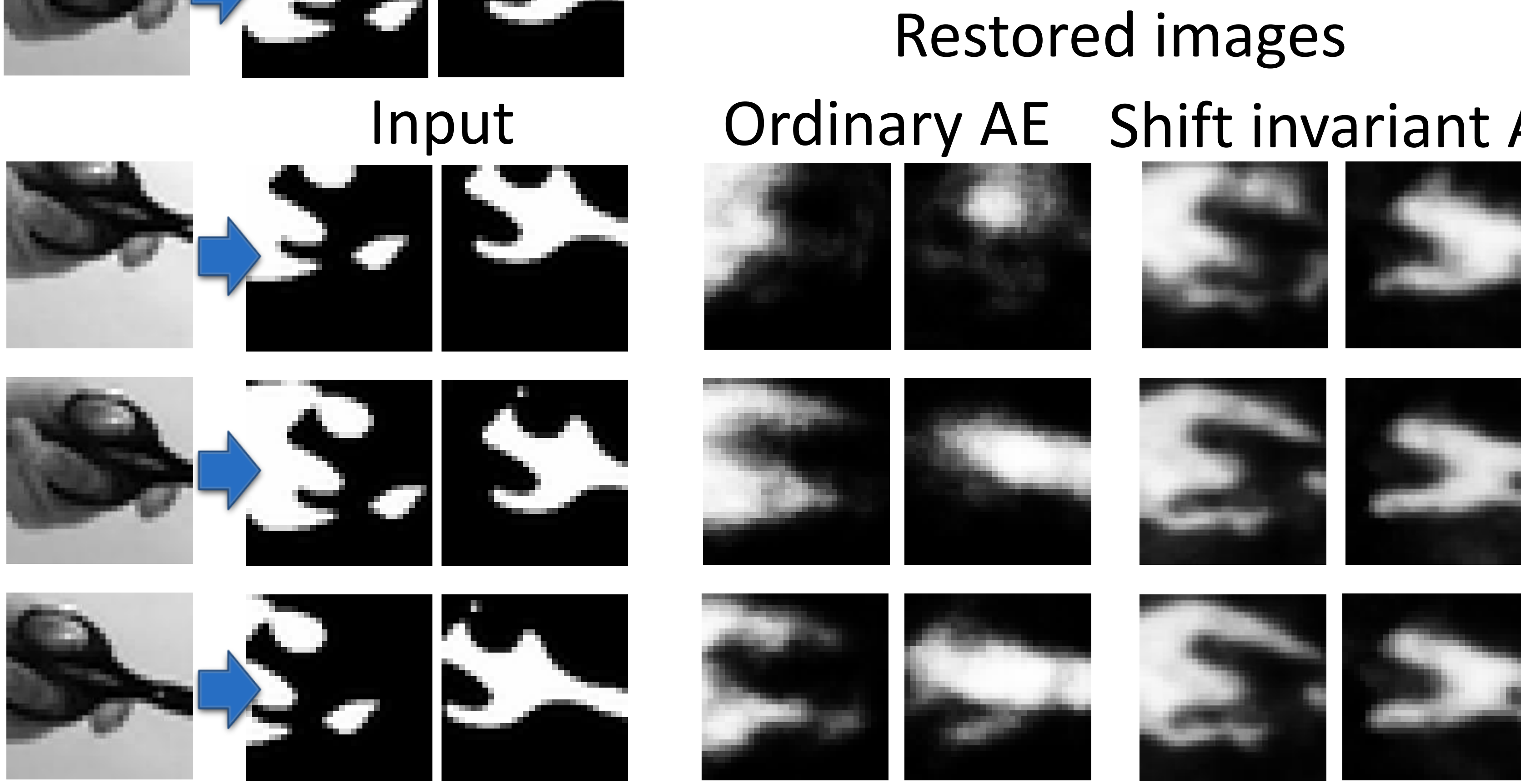
Descriptors from common shapes automatically aggregate.

Shift invariant auto-encoder

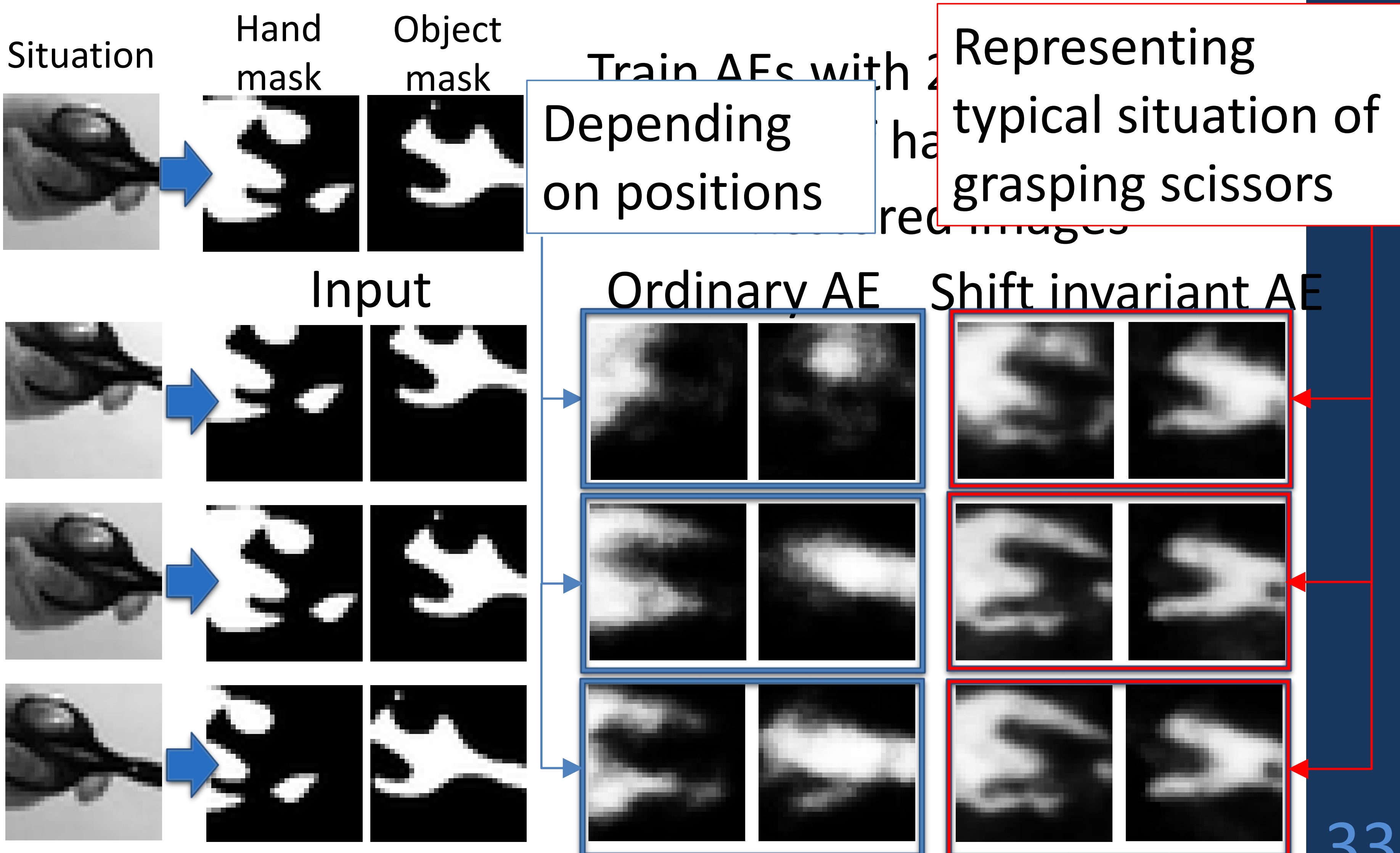
Example for hand-object interaction

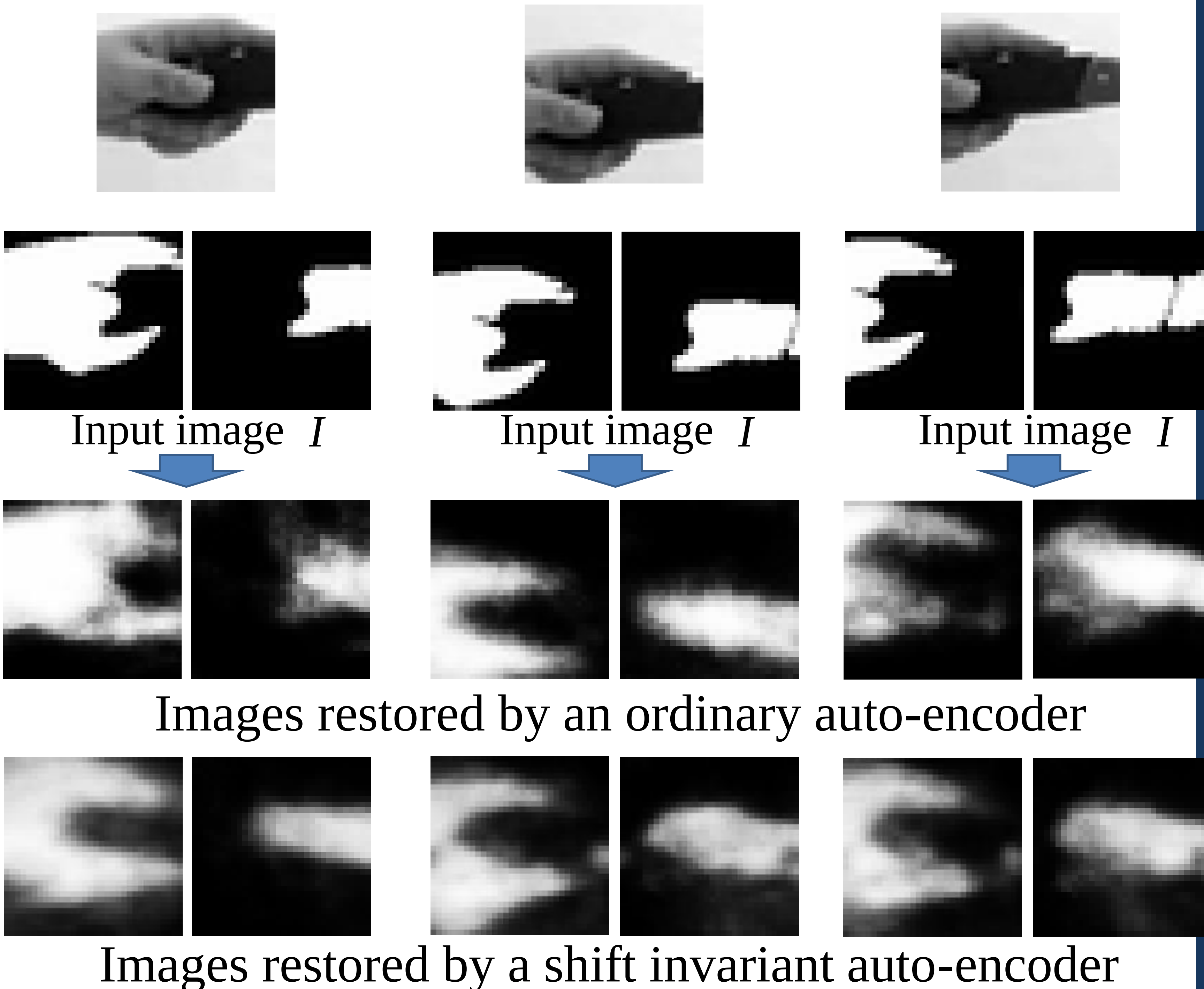


Train AEs with 2-channel images consisting of hand/object masks.



Example for hand-object interaction





Input image I
Input image I
Input image I
Images restored by an ordinary auto-encoder
Images restored by a shift invariant auto-encoder