Inference of Grasping Pattern from Object Image Based on Interaction Descriptor

Tadashi Matsuo, Takuya Kawakami, 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].

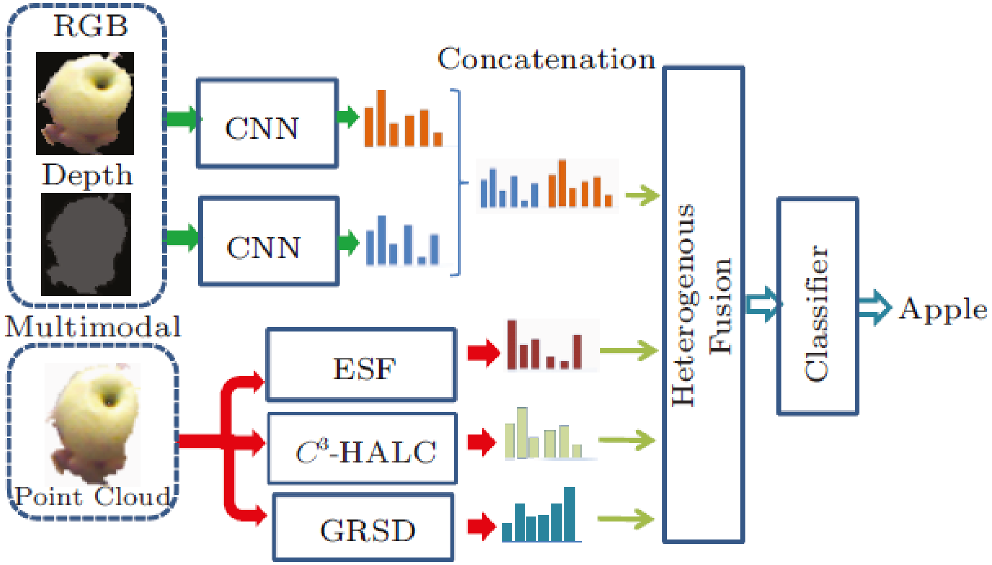
Can we estimate how to grasp an object from the object itself?

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)

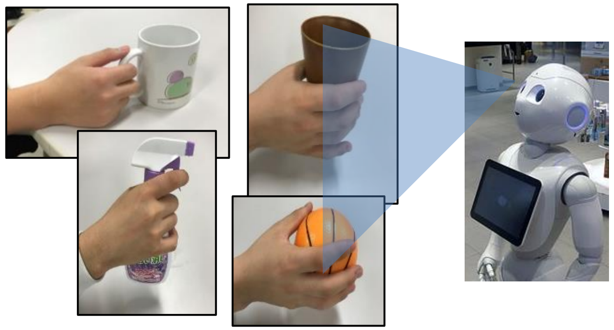


They achieved highly accurate classification by utilizing how to grasp an object, but…

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

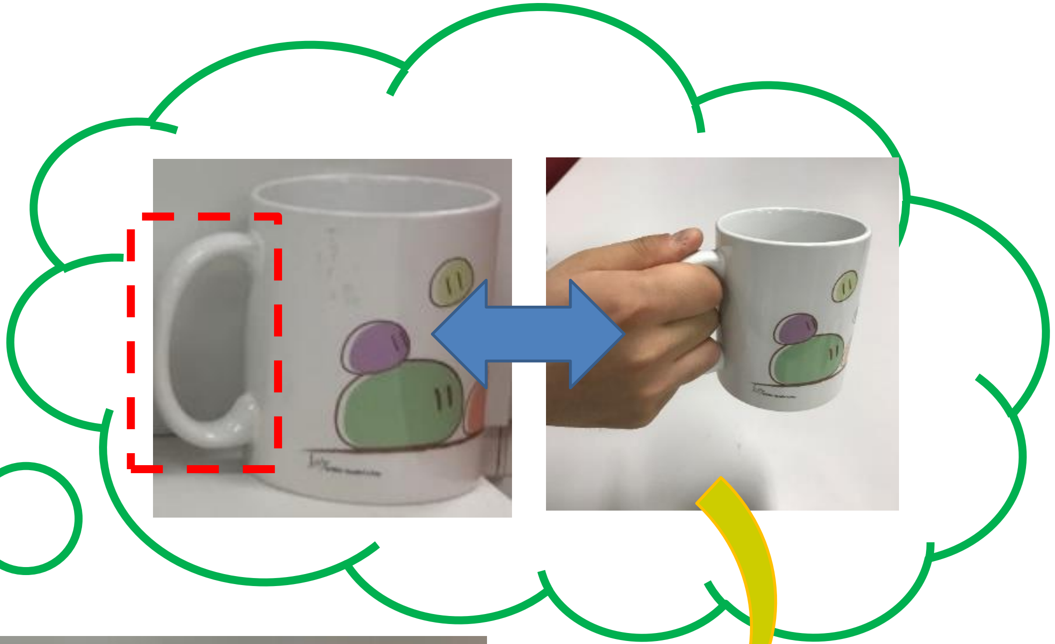
Our goal

Training



1.Learn human interactions without teacher labels.

2.Recall how a human grasp it from an object appearance.

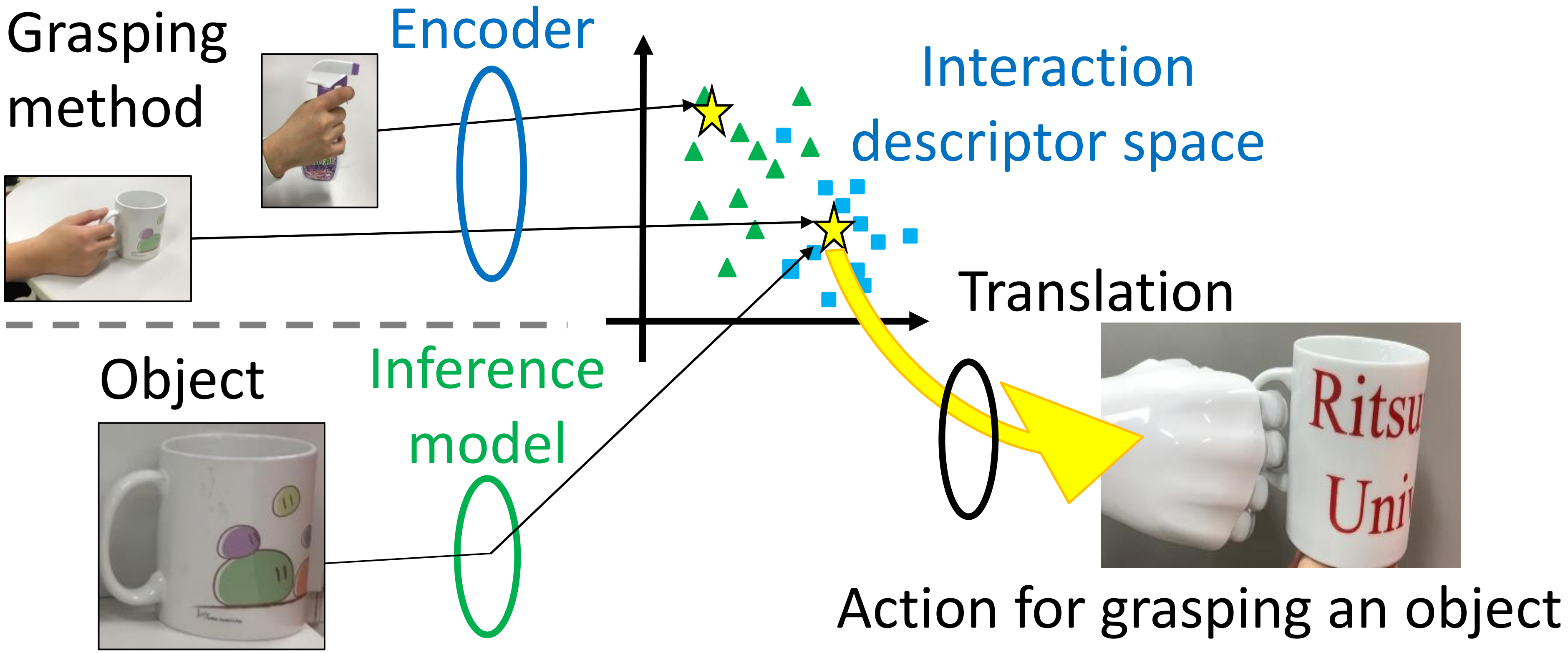


Object appearance

3.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.



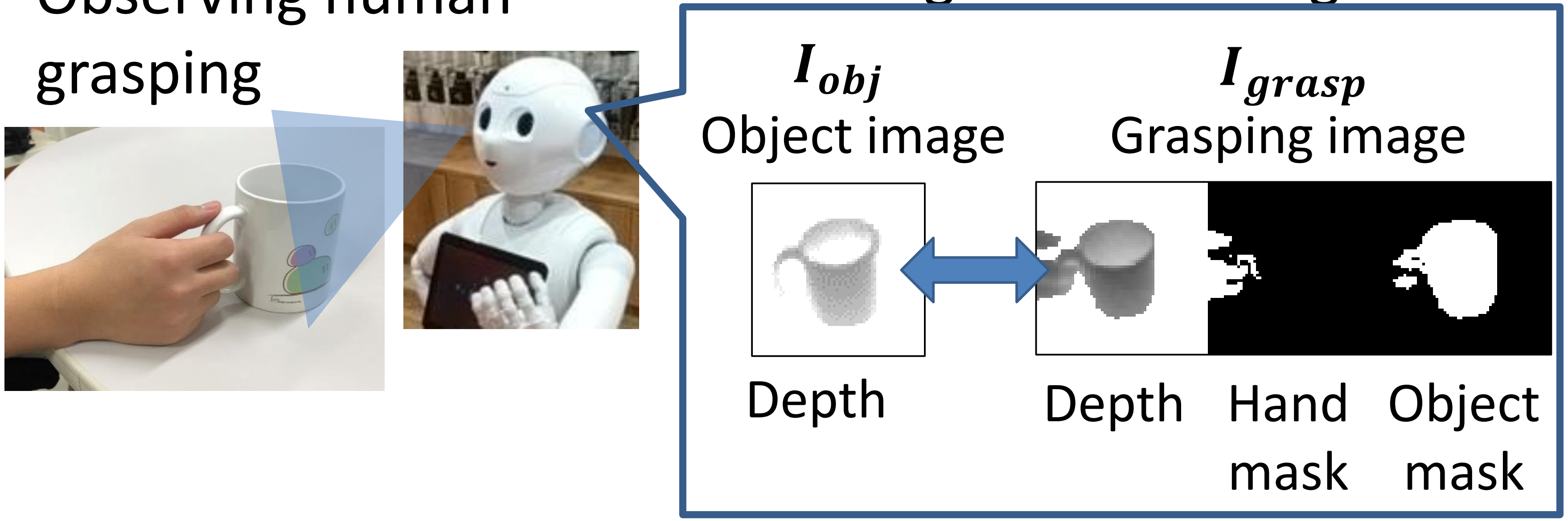
Flow of the presentation

1. Grasping image
2. Interaction descriptor
    1. Shift invariant auto-encoder
    2. Examples of shift invariant auto-encoder
    3. Results of interaction descriptor
3. Inference model
    1. Concept
    2. Recalled interactions from an object
    3. Interaction Map
4. Conclusion

Grasping image

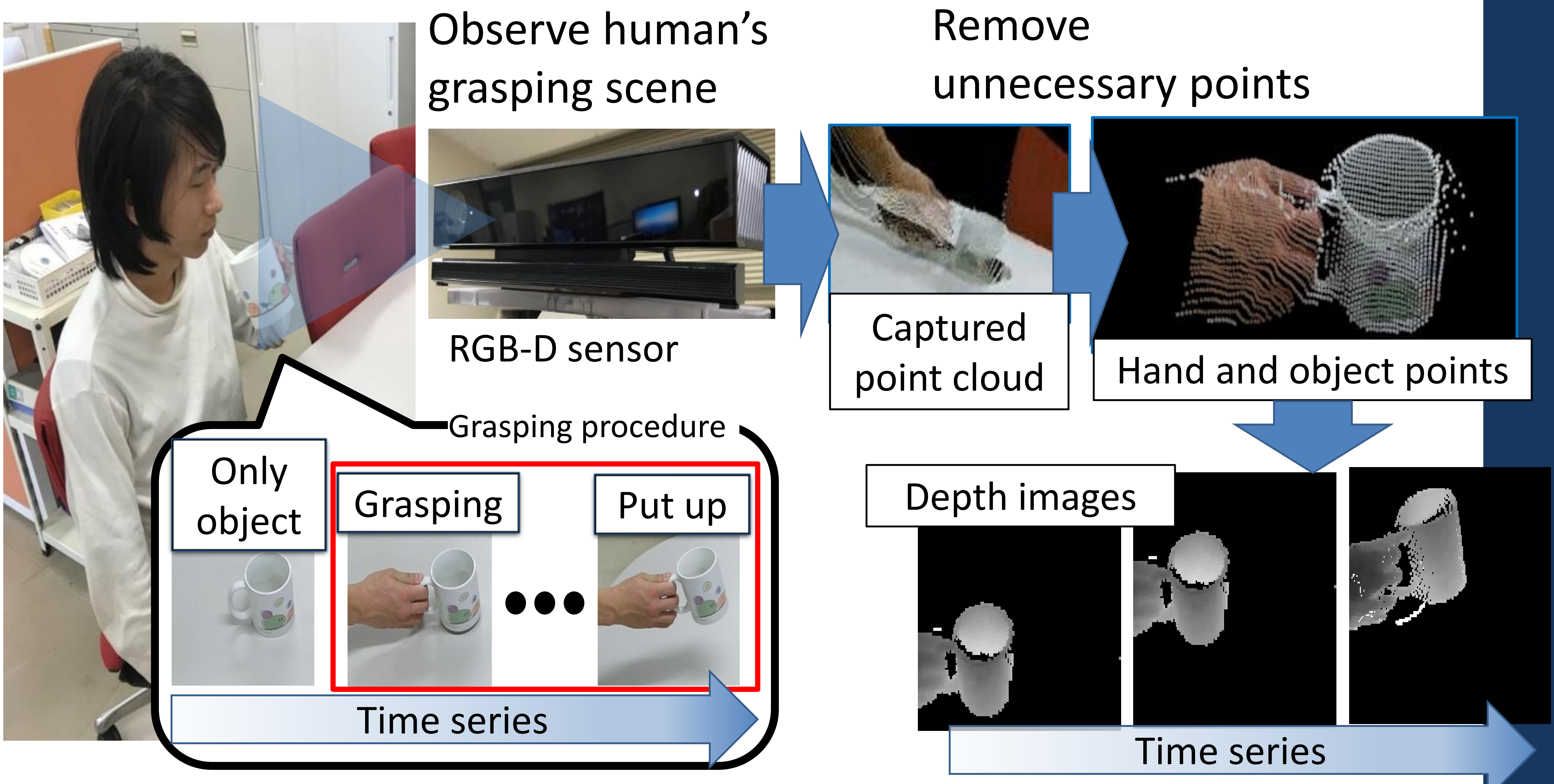
Automatically collect
Images for learning

Observing human
grasping



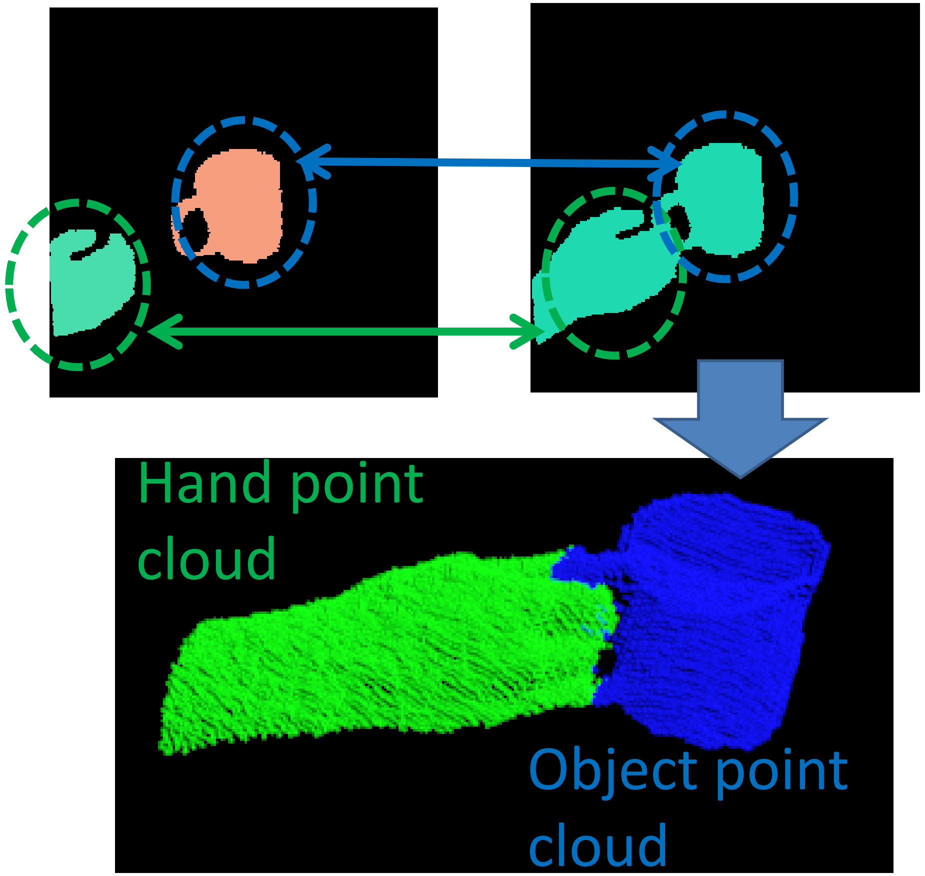
Grasping method is represented as a grasping image.
It consists of a depth image, hand mask and object mask.
It is paired with the corresponding object image.

Automatic collection of grasping images (1/2) - Capture

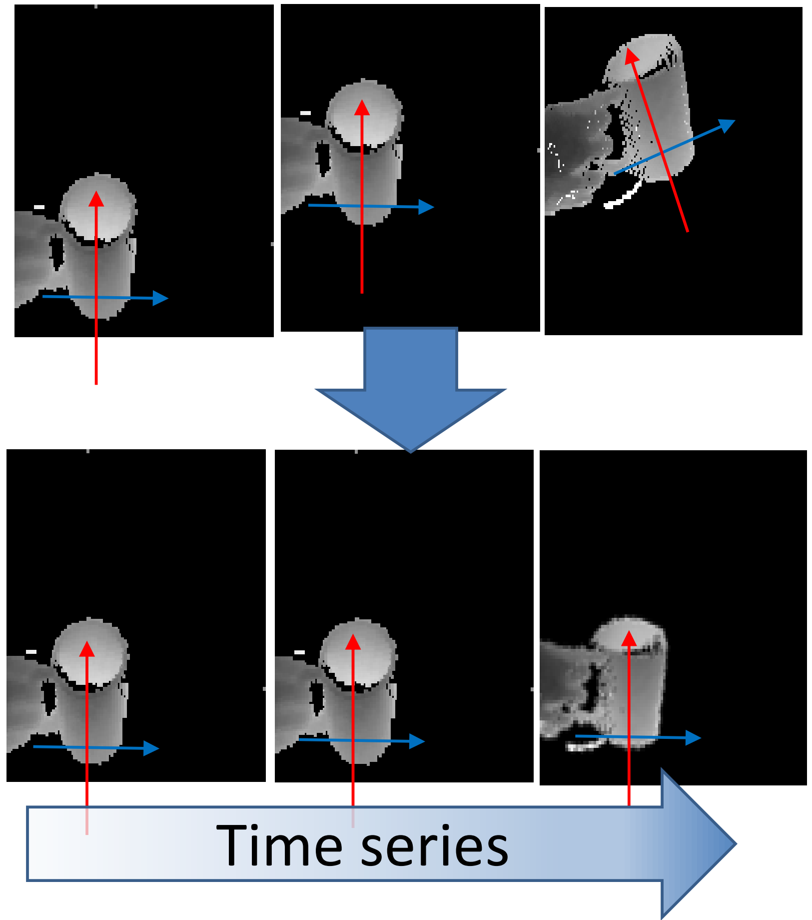


Automatic collection of grasping images (2/2) - Segmentation

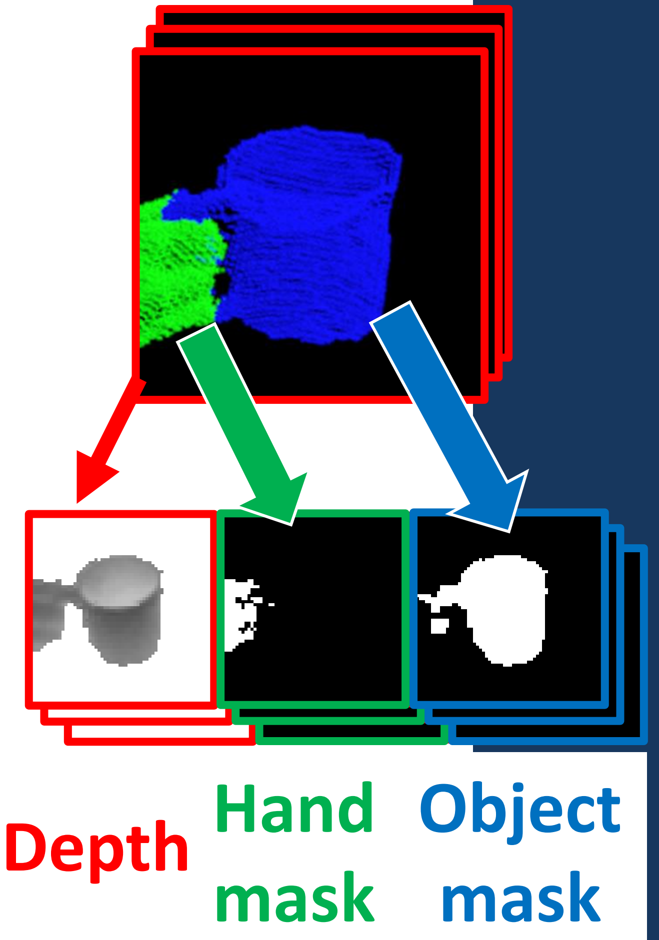
1. Segment **hand** and **object** points by using the image with isolated regions.



2. Align points based on the initial **object** points.



3. Generate grasping images

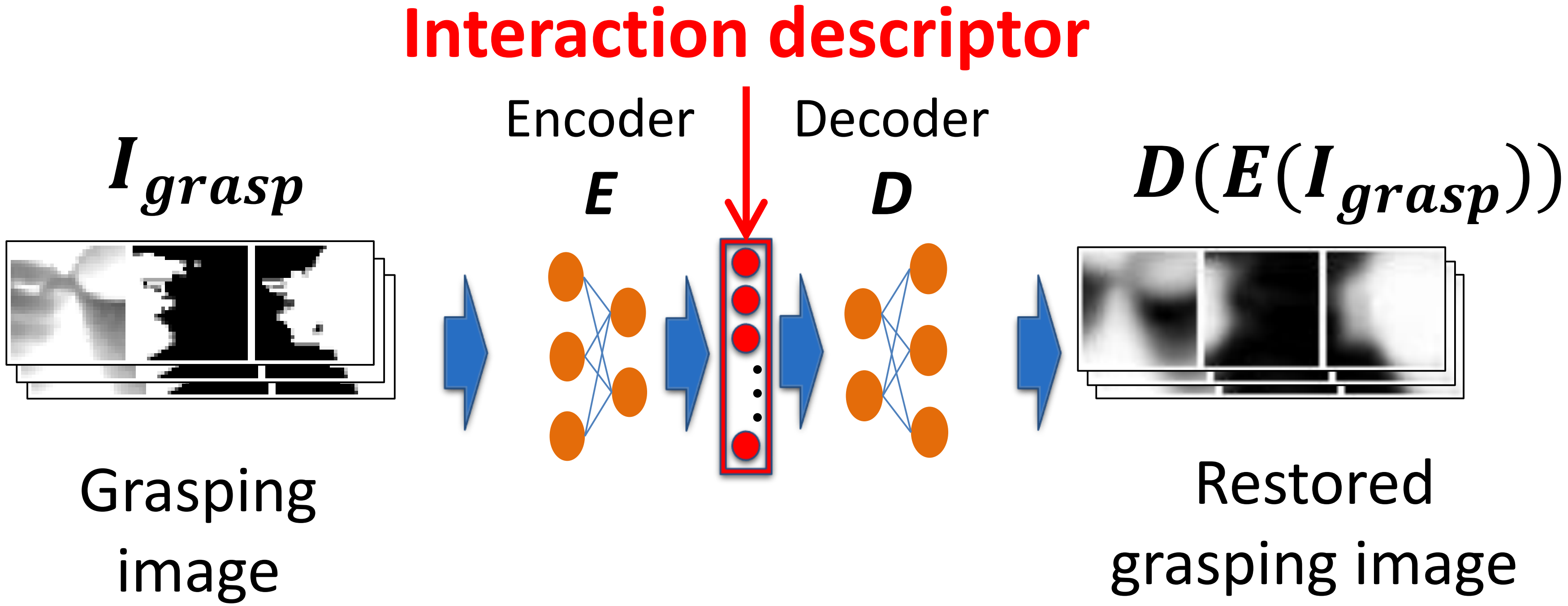


Flow of the presentation

1. Grasping image
2. Interaction descriptor
    1. Shift invariant auto-encoder
    2. Examples of shift invariant auto-encoder
    3. Results of interaction descriptor
3. Inference model
    1. Concept
    2. Recalled interactions from an object
    3. Interaction Map
4. Conclusion

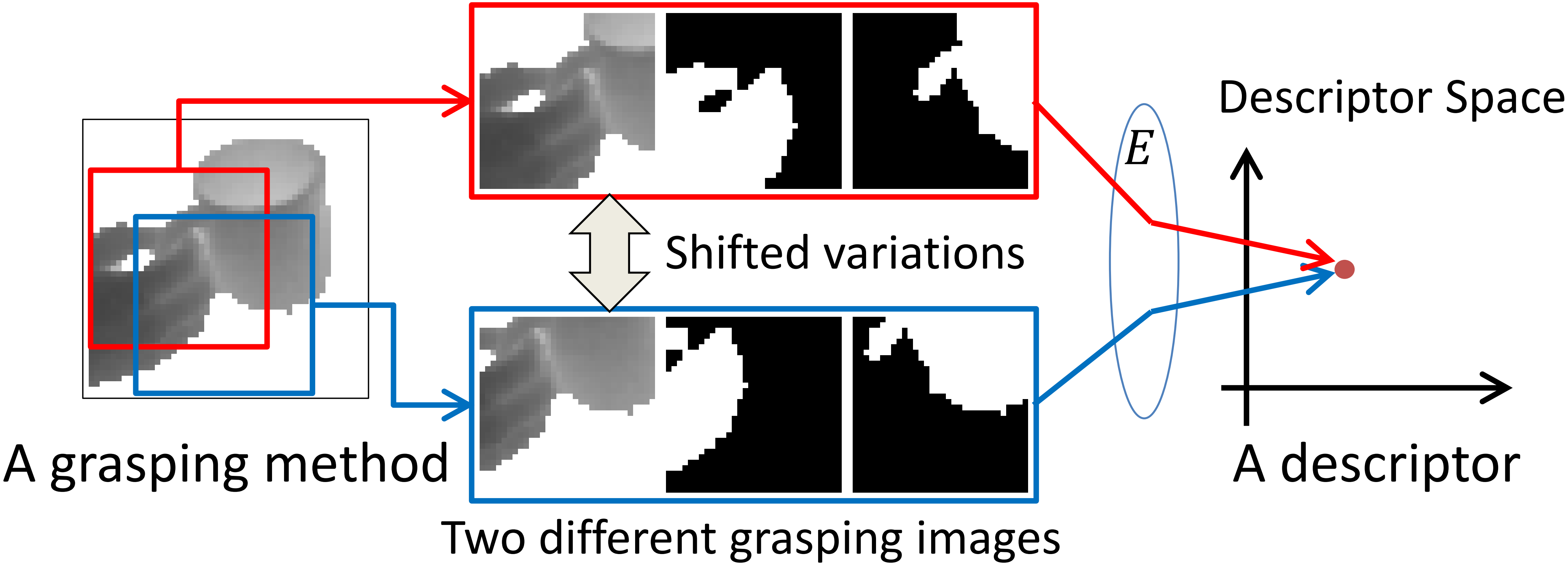
Interaction descriptor

We generate interaction descriptors by auto-encoder method;



- A low dimensional descriptor represents essence of an input.
- The auto-encoder (E and D) can be trained without teacher labels.

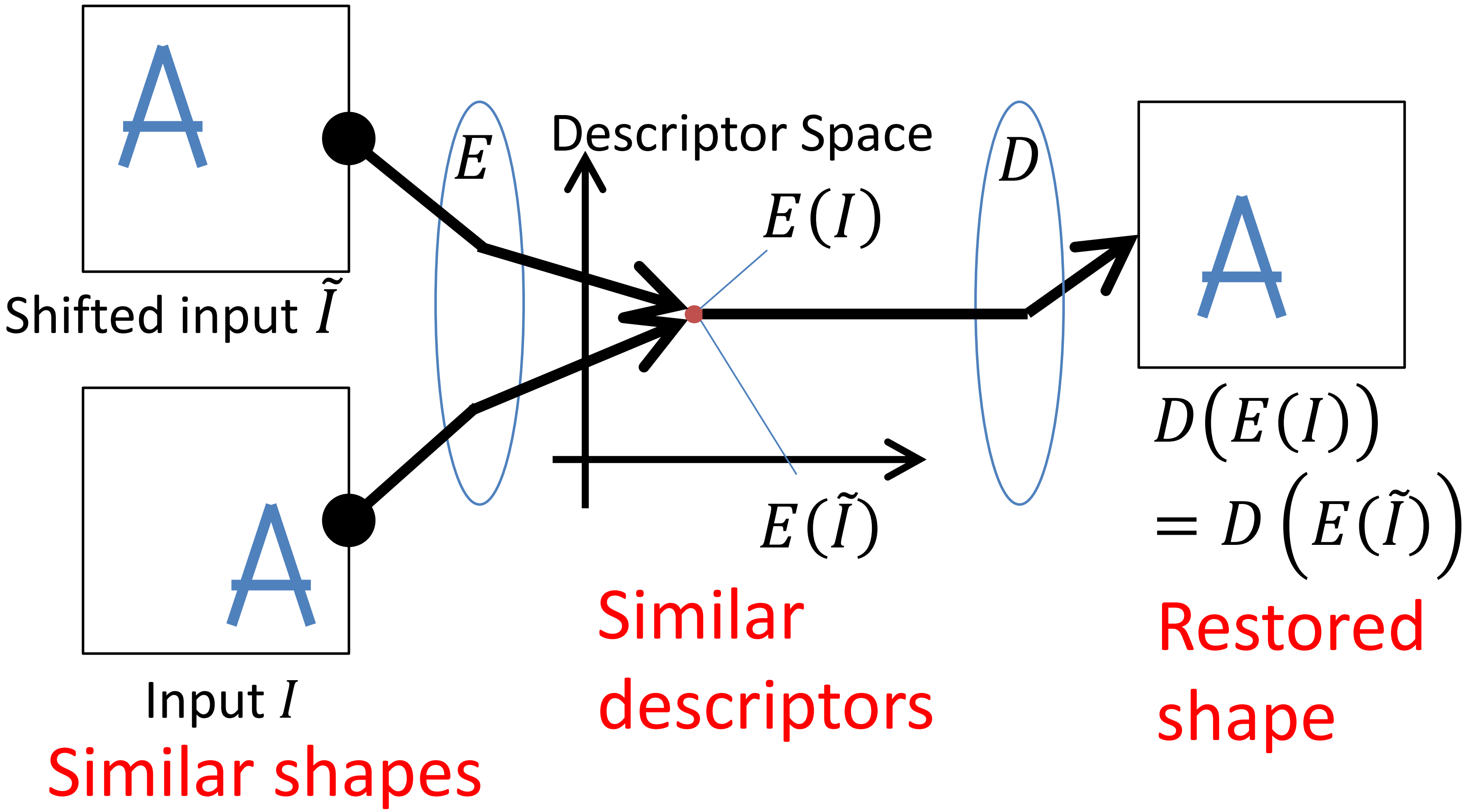
Desirable property of the encoder



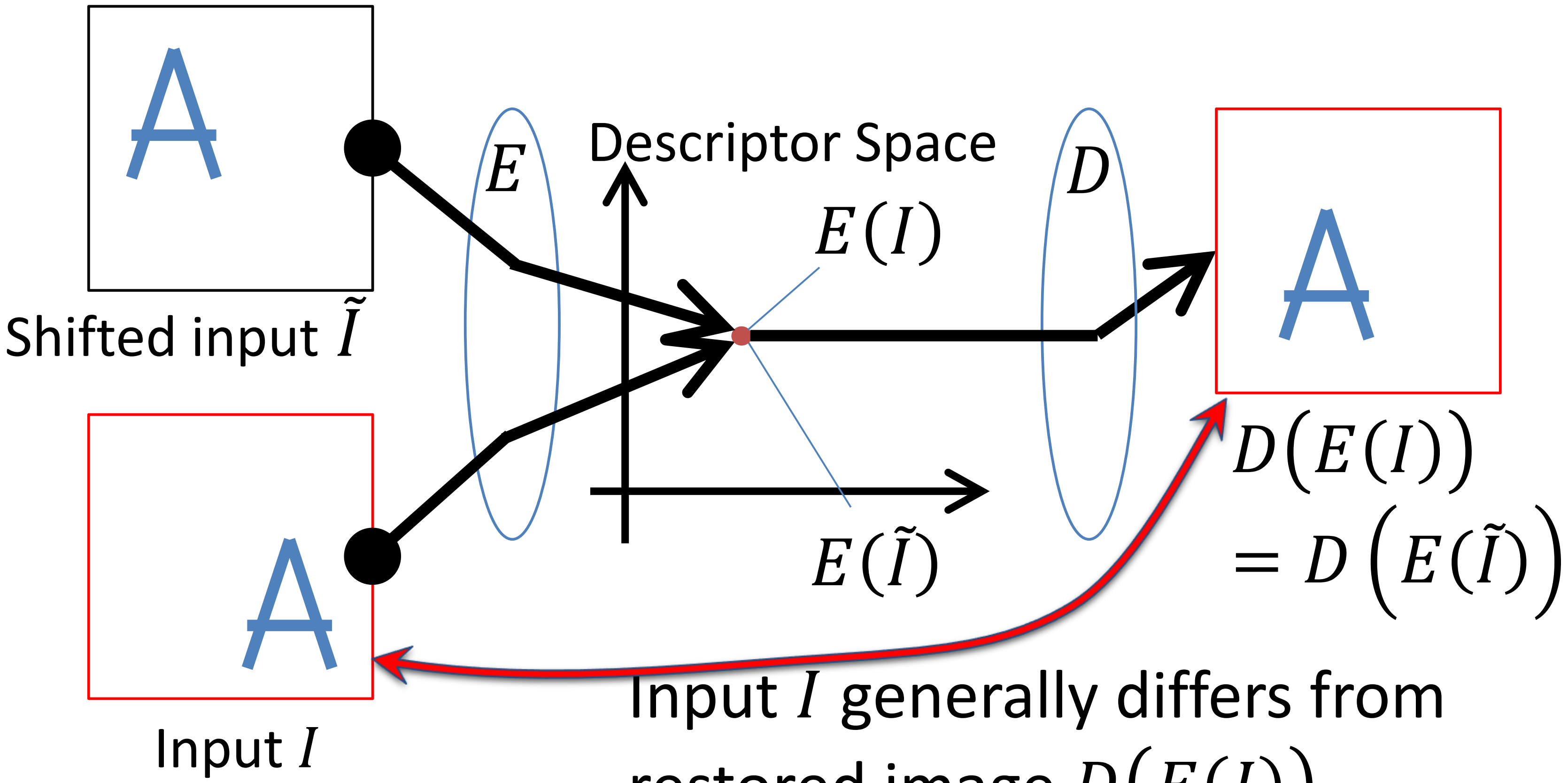
A grasping method should correspond to an interaction descriptor.

The encoder should ignore spatial shifts. **Shift invariant auto-encoder**

Shift invariant auto-encoder



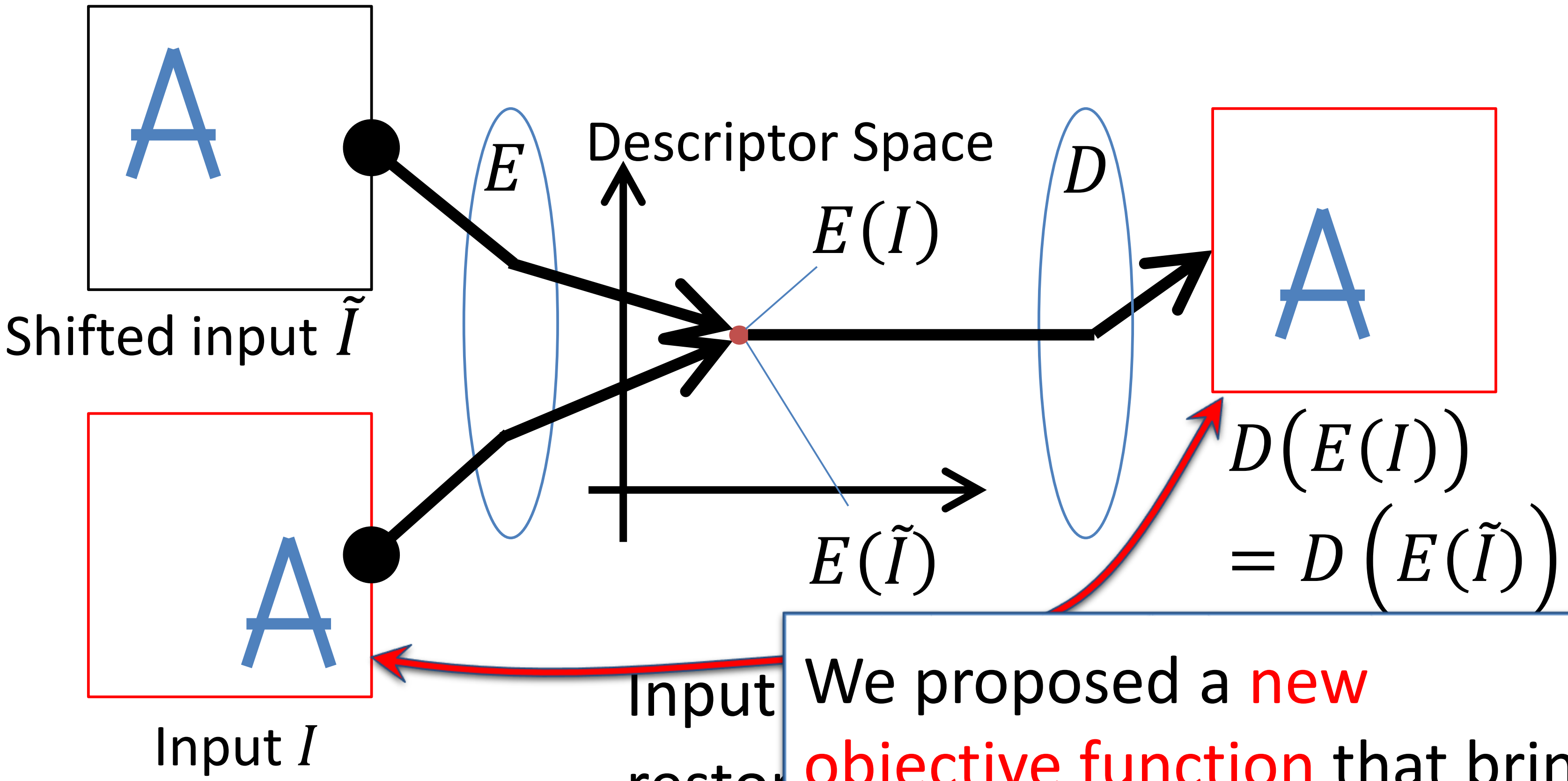
Shift invariant auto-encoder



Input I generally differs from restored image $D(E(I))$.

It cannot be trained with ordinary objective function $\sum_{I} \left\| D(E(I)) - I \right\|_{L2}^{2}$

Shift invariant auto-encoder



Input
restor

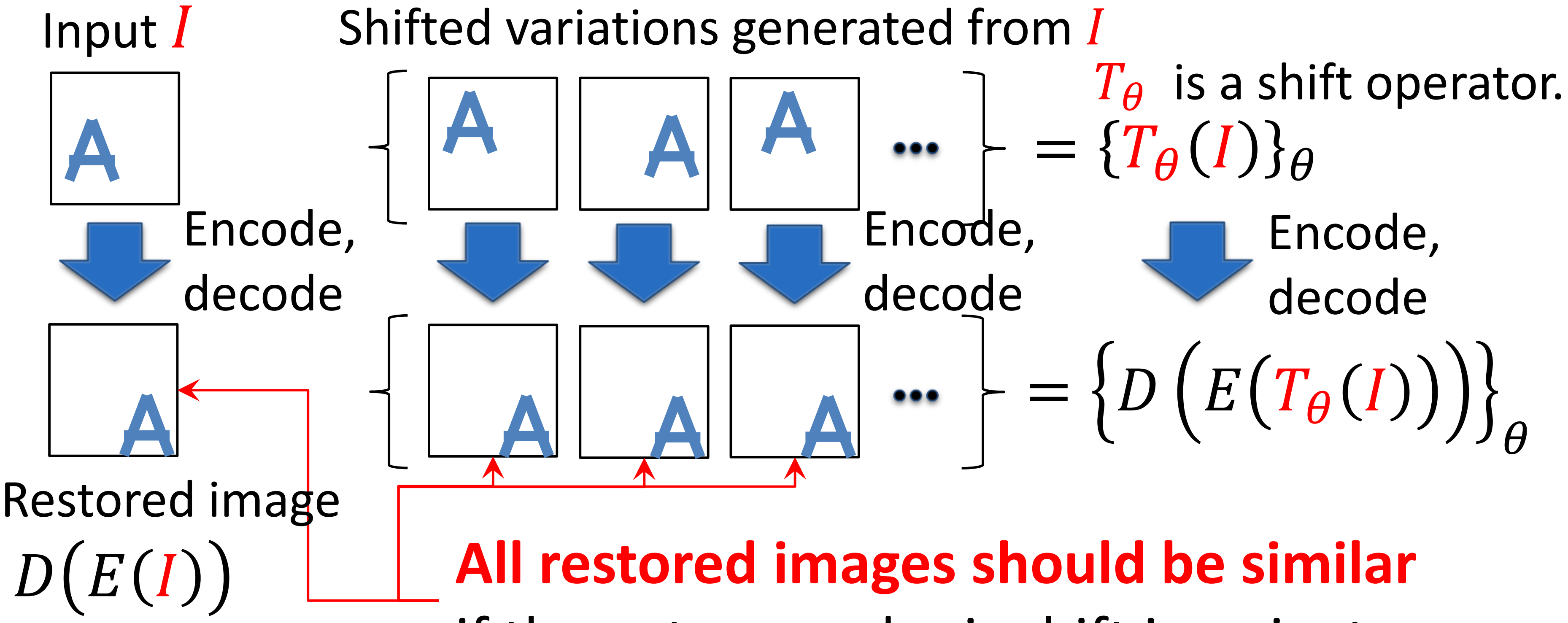
We proposed a new objective function that brings
- **shift invariance**,
- **shape restoration**.

It cannot be trained wi
ordinary objective functi

I

Cost function for shift invariant auto-encoder (1/3)

Evaluation of shift invariance

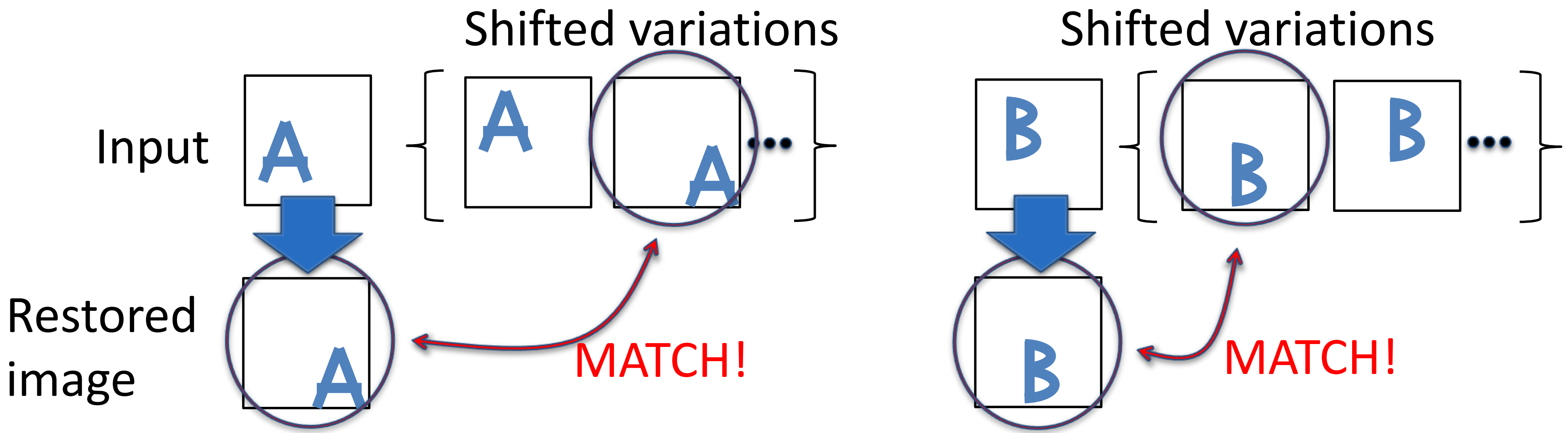


All restored images should be similar if the auto-encoder is shift invariant.

$$\sum_I \sum_\theta \left\| D\left(E(I)\right) - D\left(E\left(T_\theta(I)\right)\right) \right\|_{L2}^2$$

Train E, D so that they minimize this term.

Cost function for shift invariant auto-encoder (2/3)
Evaluation of shape restoration



A restored image should match with one of shifted images
if the auto-encoder can restore a shape.

$$\sum_I \left\| D\big(E(I)\big) - T_{\hat{\theta}(I)}(I) \right\|_{L2}^2$$

Train E, D so that they minimize this term.

where $\hat{\theta}(I)$ means the "best" transform parameter:

$$\hat{\theta}(I) = \arg\min_{\theta} \left\| D\big(E(I)\big) - T_{\theta}(I) \right\|_{L2}^2$$

Cost function for shift invariant auto-encoder (3/3) Total form

$$C(E,D) = \lambda_{inv} C_{inv}(E,D) + \lambda_{res} C_{res}(E,D) + \lambda_{spa} C_{spa}(E)$$

Invariance term $C_{inv}(E,D)$

$$\sum_{I}\sum_{\theta} \left\| D(E(I)) - D\left(E(T_{\theta}(I))\right) \right\|_{L2}^{2}$$

Restored image should be unchanged even if inputs are transformed with any parameter.

Restoration term $C_{res}(E,D)$

$$\sum_{I} \left\| D(E(I)) - T_{\hat{\theta}(I)}(I) \right\|_{L2}^{2}$$

Restored image should match with one of transformed images.

Sparseness term

$C_{spa}(E)$

$$\sum_{I \in S} \frac{\|E(I)\|_{L1}^{2}}{\|E(I)\|_{L2}^{2}}$$

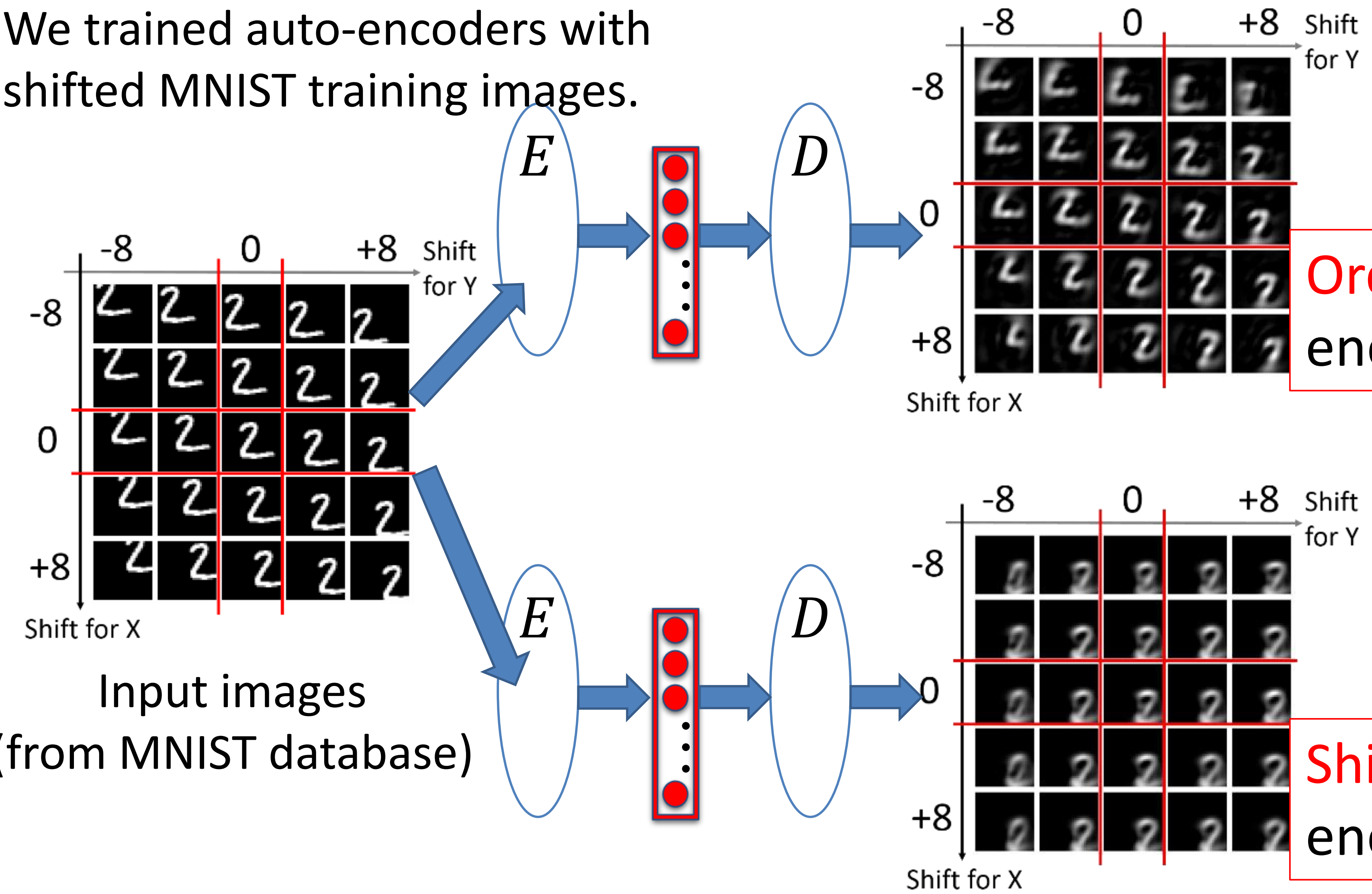
Descriptor $E(I)$ should be a sparse vector.

Flow of the presentation

1. Grasping image
2. Interaction descriptor
   1. Shift invariant auto-encoder
   2. Examples of shift invariant auto-encoder
   3. Results of interaction descriptor
3. Inference model
   1. Concept
   2. Recalled interactions from an object
   3. Interaction Map
4. Conclusion

Example of shift invariant auto-encoder (1/2)

We trained auto-encoders with shifted MNIST training images.



Input images (from MNIST database)

Restored by ordinary auto-encoder

Ordinary auto-encoder encodes shape and position.

Restored by shift invariant auto-encoder

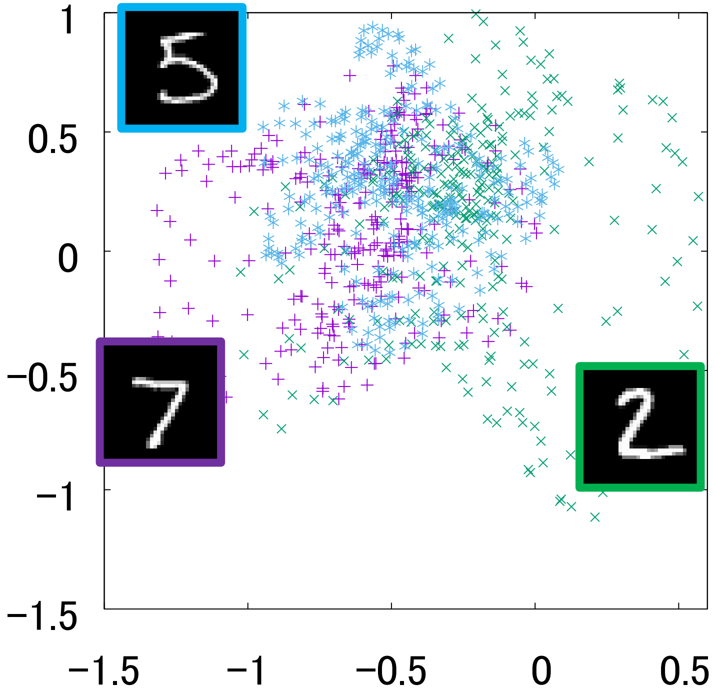
Shift invariant auto-encoder encodes shape itself.

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

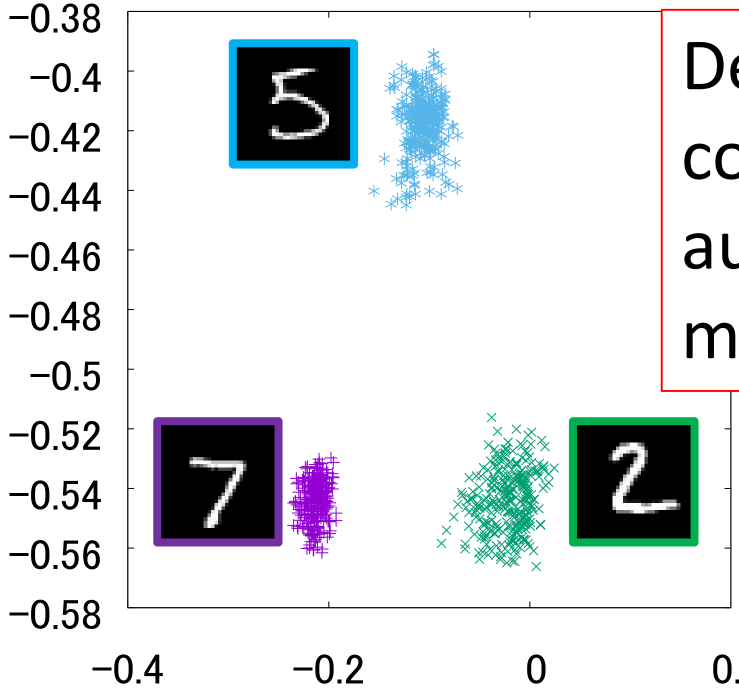
Example of shift invariant auto-encoder (2/2)
Distribution of descriptors

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

Distributions of descriptors from shifted test images such as  .



Ordinary auto-encoder



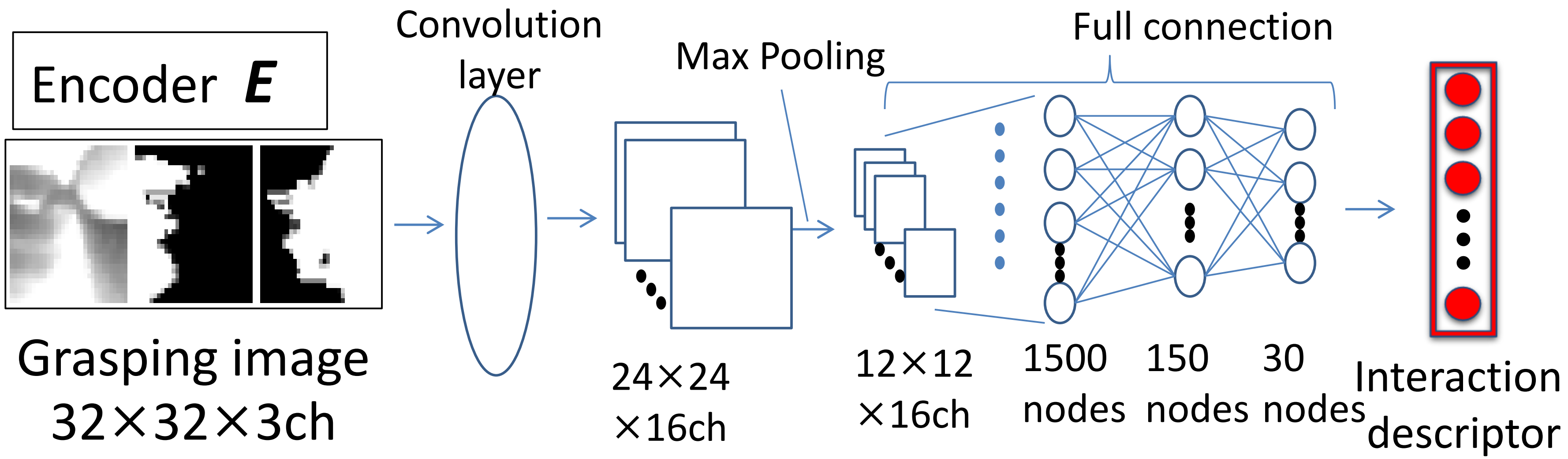
Shift invariant auto-encoder

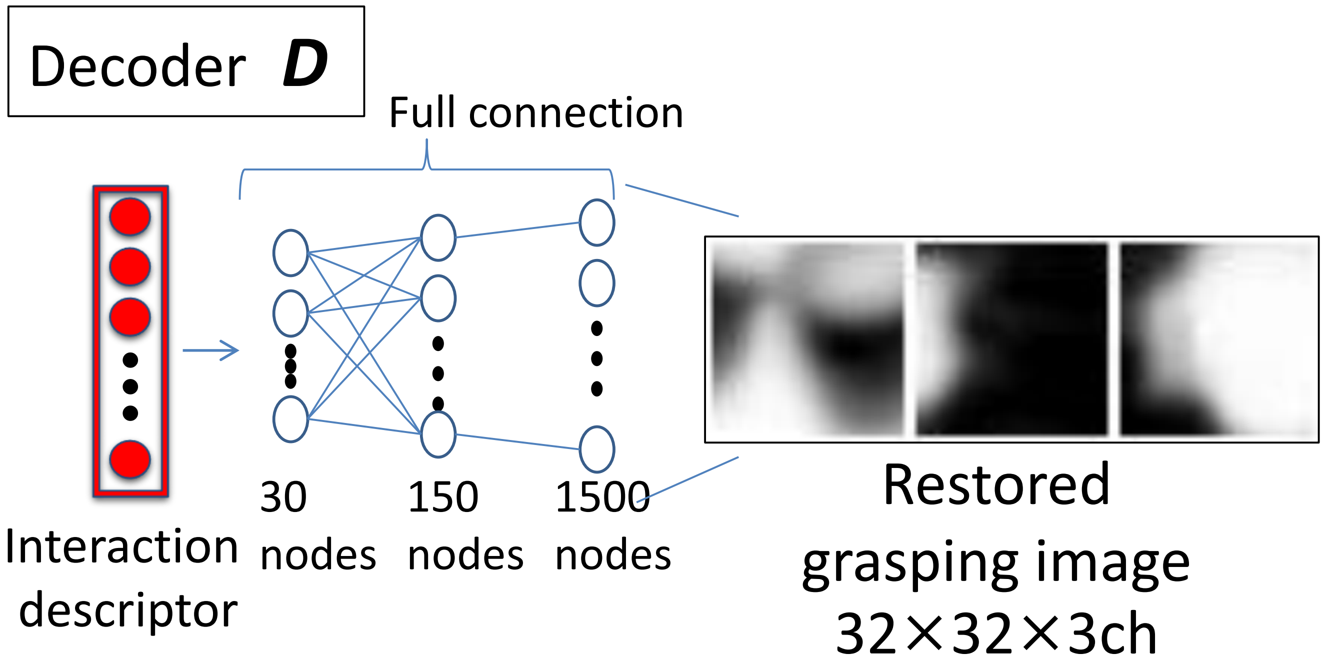
Descriptors from common shapes automatically make a cluster.

Flow of the presentation

1. Grasping image
2. Interaction descriptor
    1. Shift invariant auto-encoder
    2. Examples of shift invariant auto-encoder
    3. Examples of interaction descriptor
3. Inference model
    1. Concept
    2. Recalled interactions from an object
    3. Interaction Map
4. Conclusion

Structure of auto-encoder

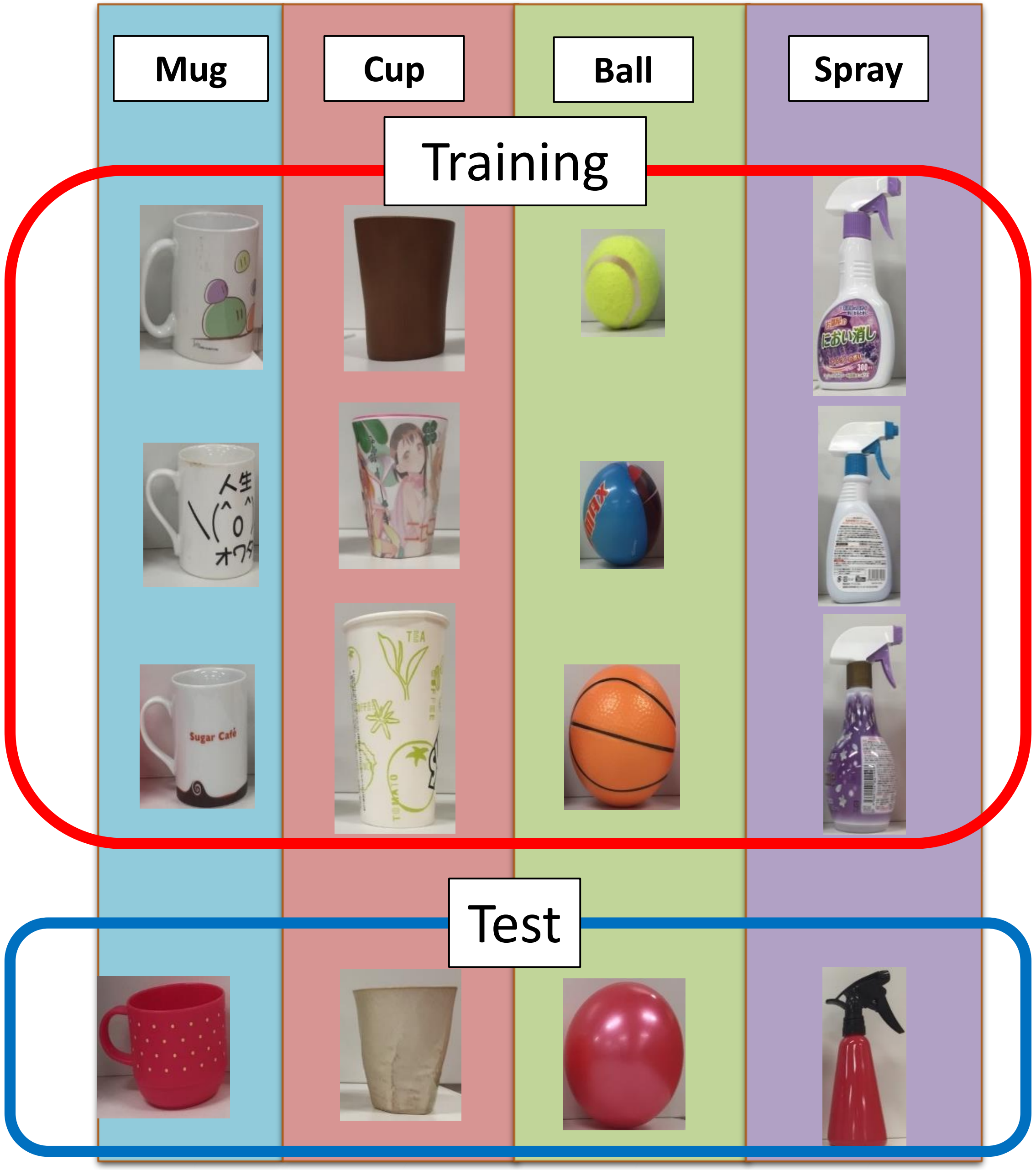




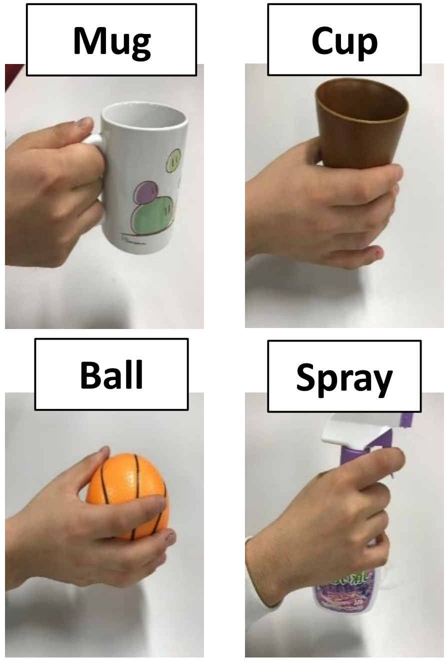
Parameters

Input size: 32×32×3
Descriptor dimension: 30
Kernel size: 9×9

Objects and grasping methods

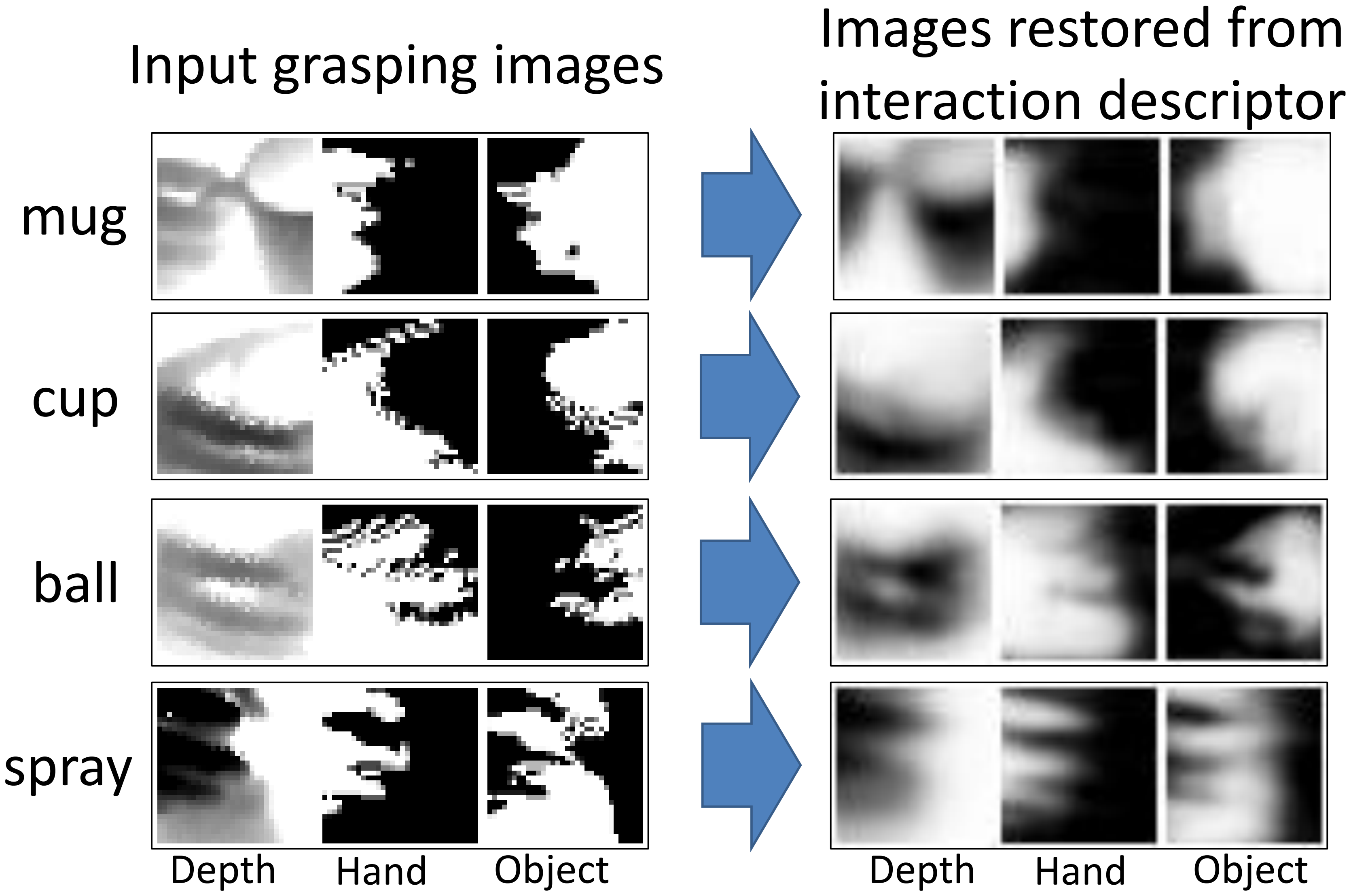


Grasping methods



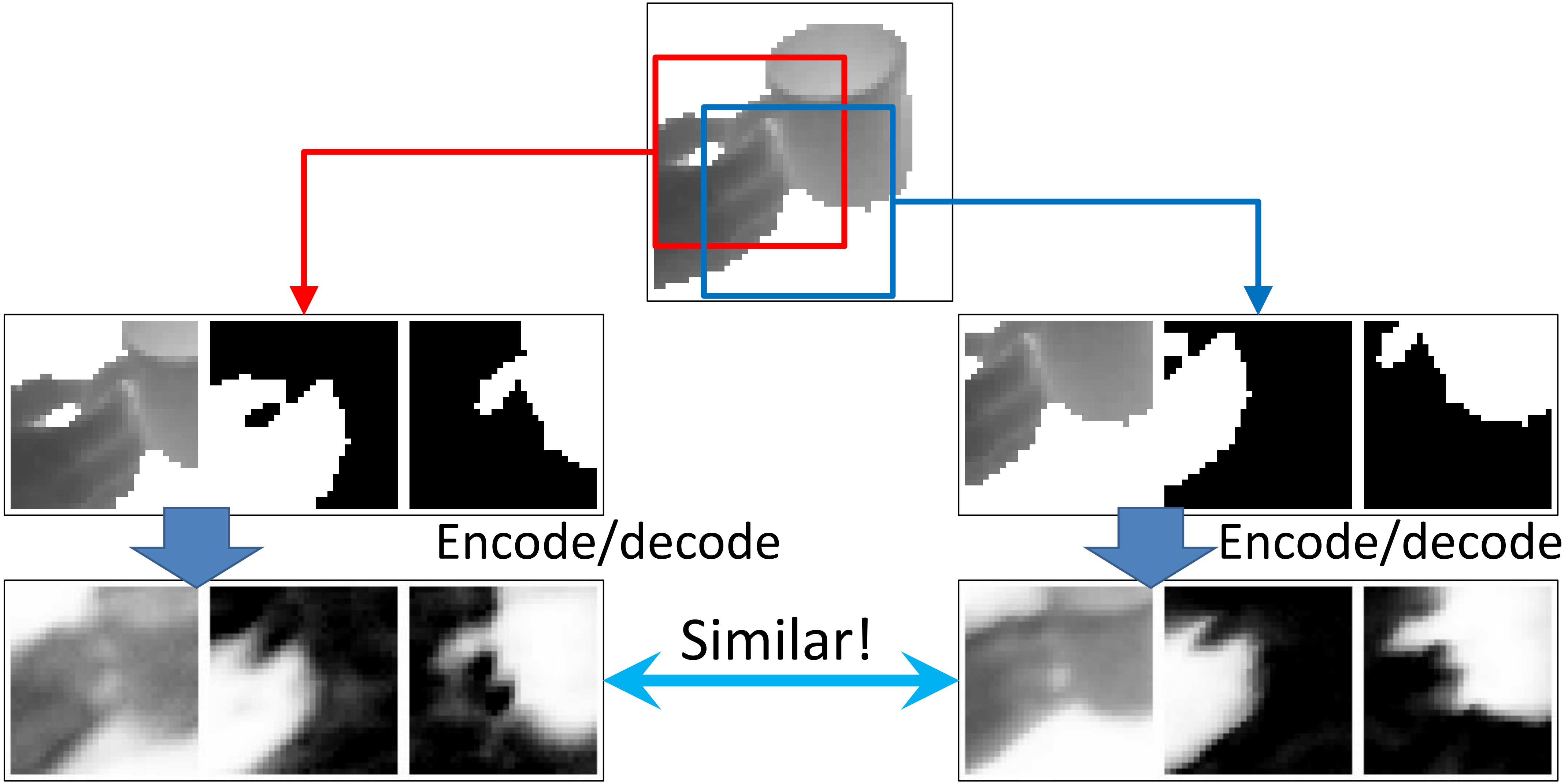
Training images:
80scenes×12kinds
= 960
Test images:
80scenes×4kinds
= 320

Restored grasping images



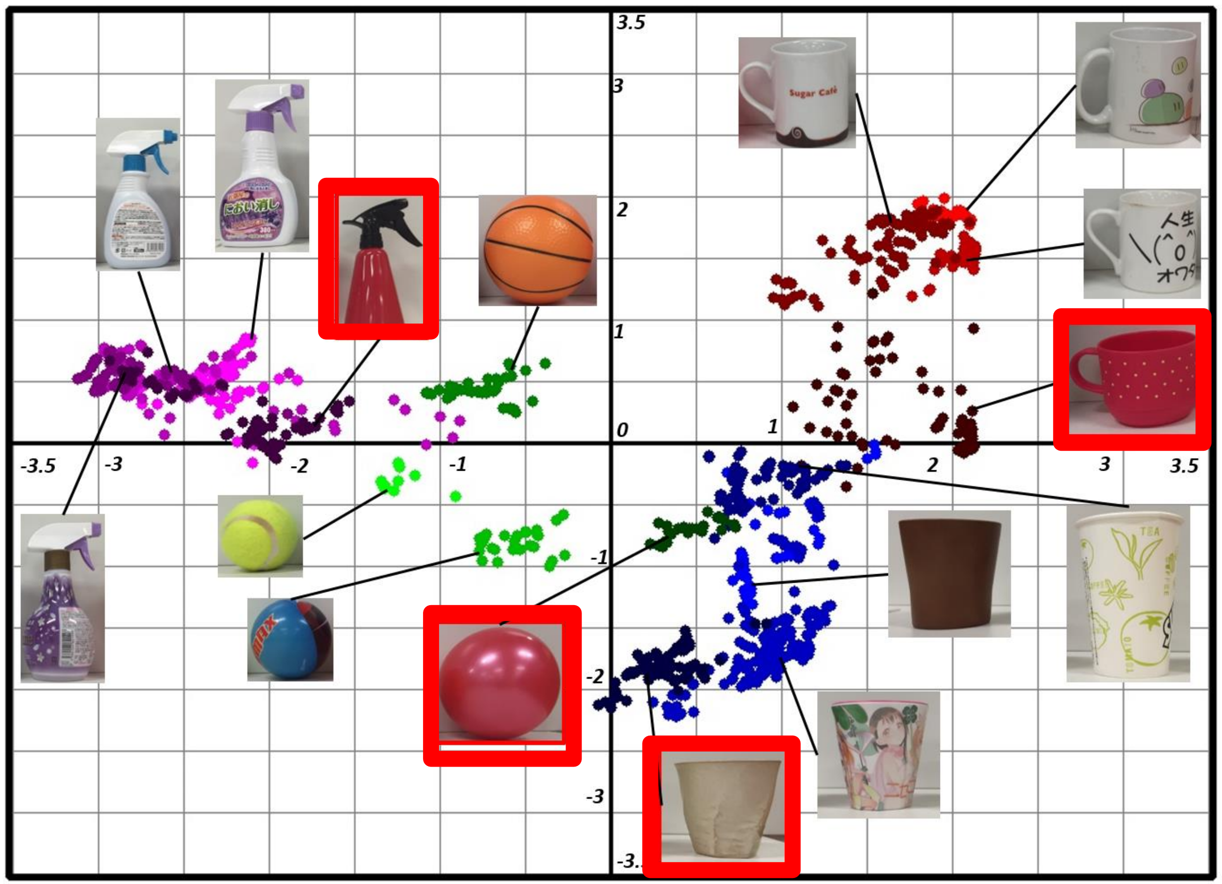
Interaction descriptor has approximate shape information.

Effect of shift invariant auto-encoder



Interaction descriptors represent a typical shape without position.

Distribution of interaction descriptors



2nd principal component

1st principal component

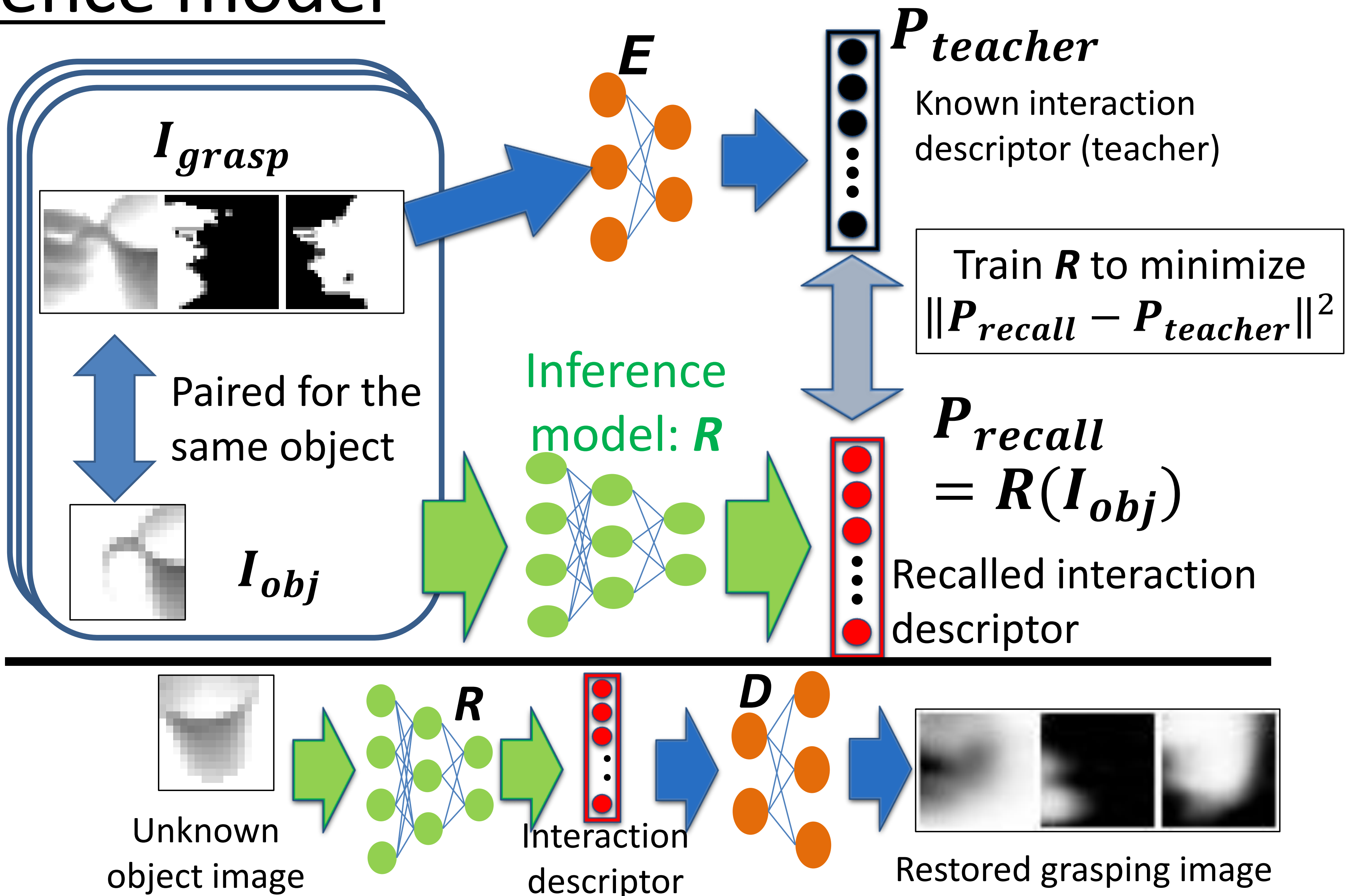
Non-frame:
Object for
training

Red frame:
Object for test

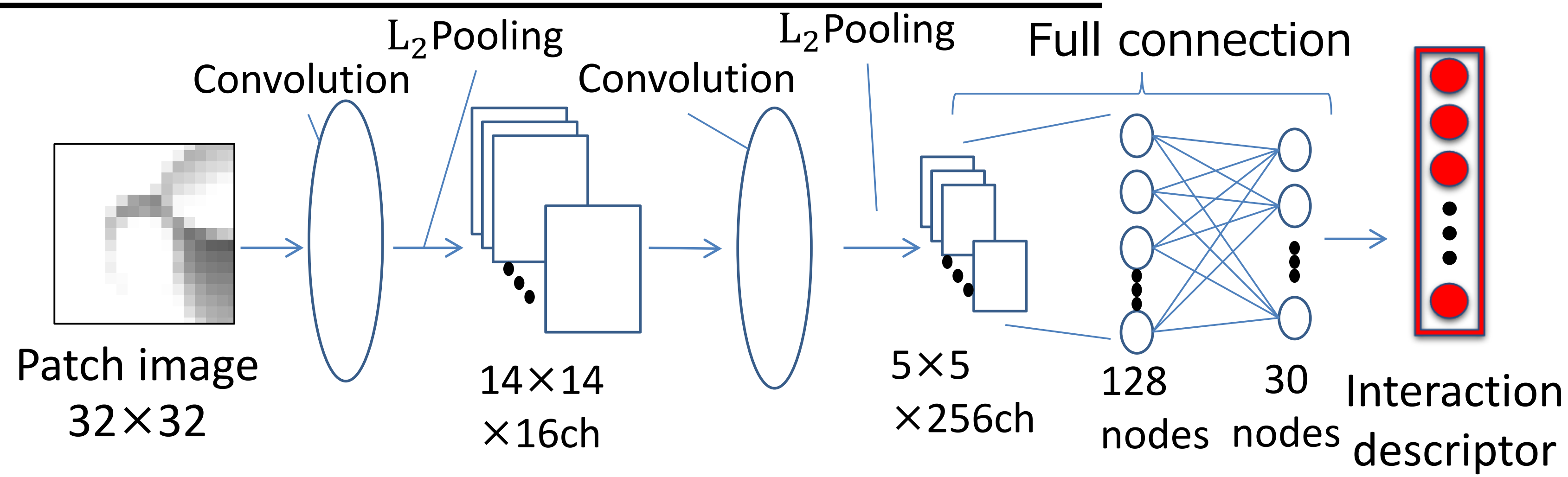
Flow of the presentation

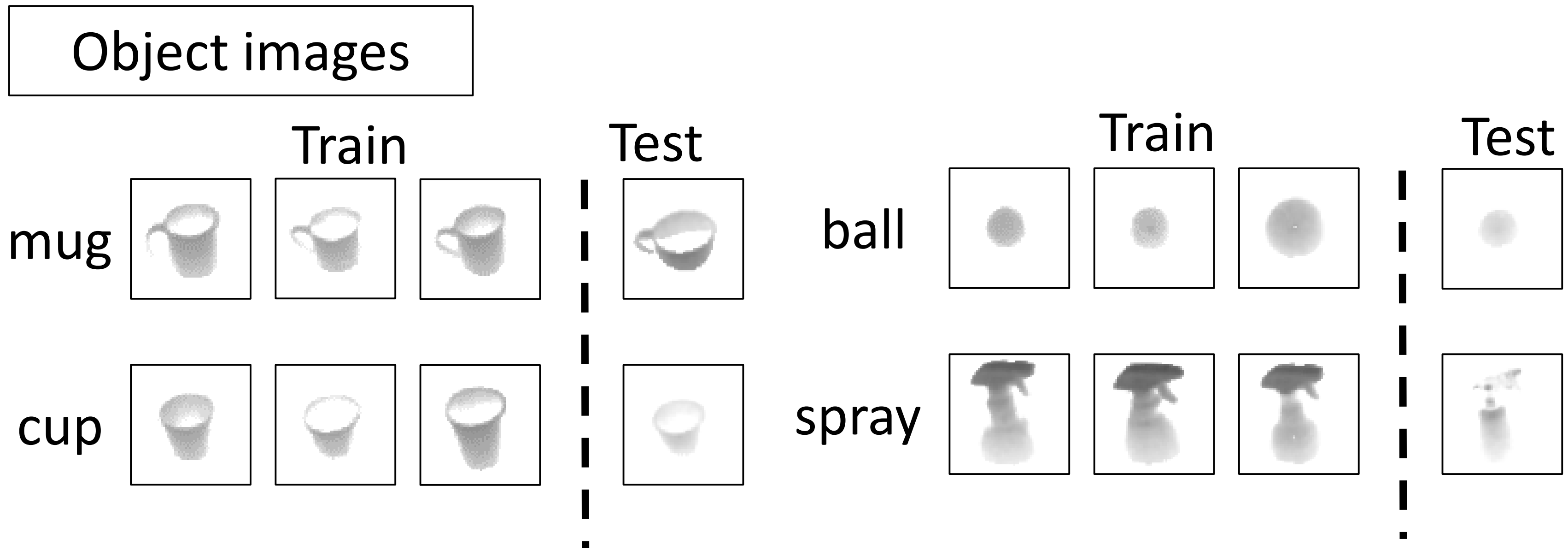
1. Grasping image
2. Interaction descriptor
   1. Shift invariant auto-encoder
   2. Examples of shift invariant auto-encoder
   3. Results of interaction descriptor
3. Inference model
   1. Concept
   2. Recalled interactions from an object
   3. Interaction Map
4. Conclusion

Inference model

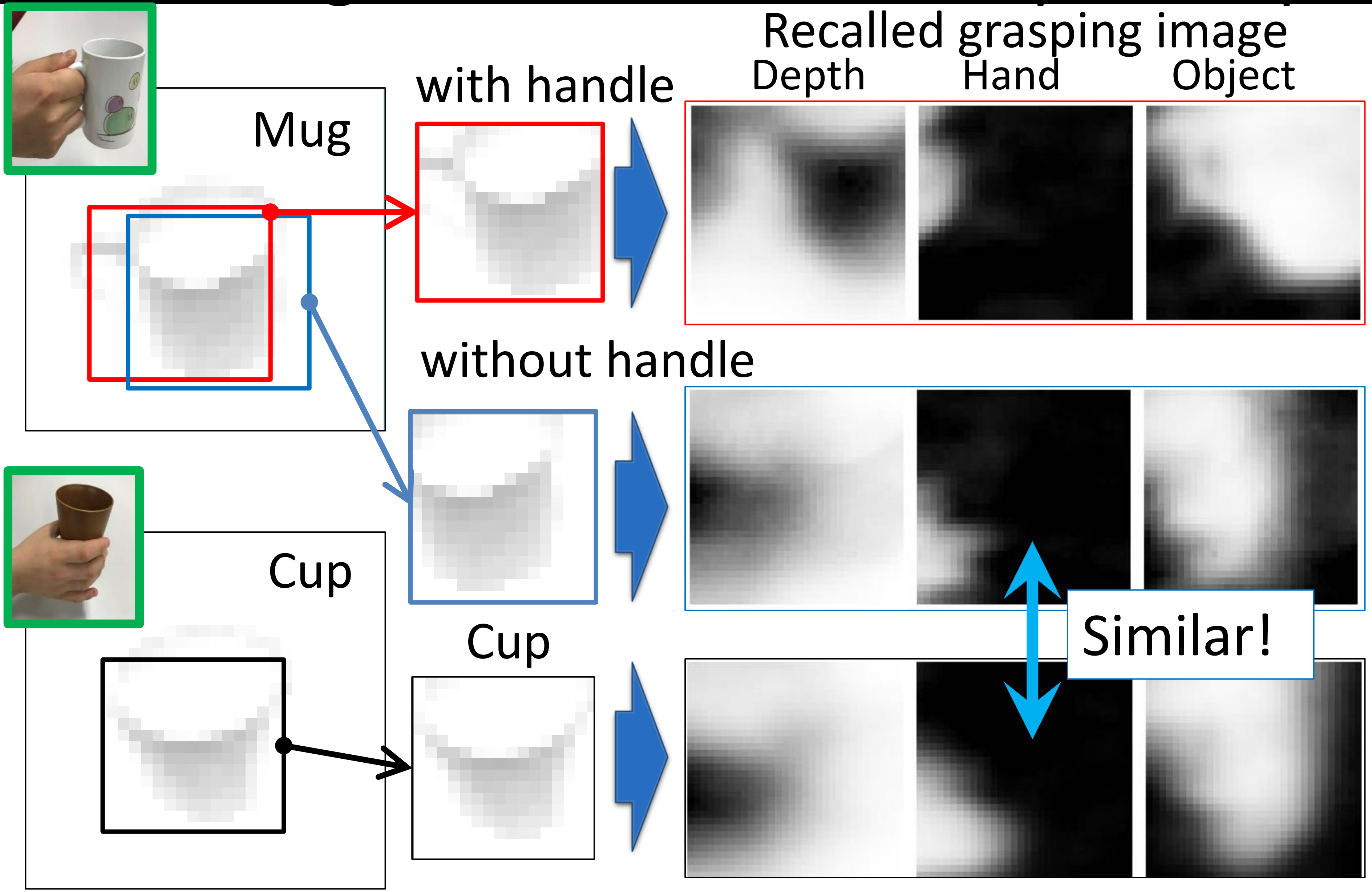


Structure of the inference model

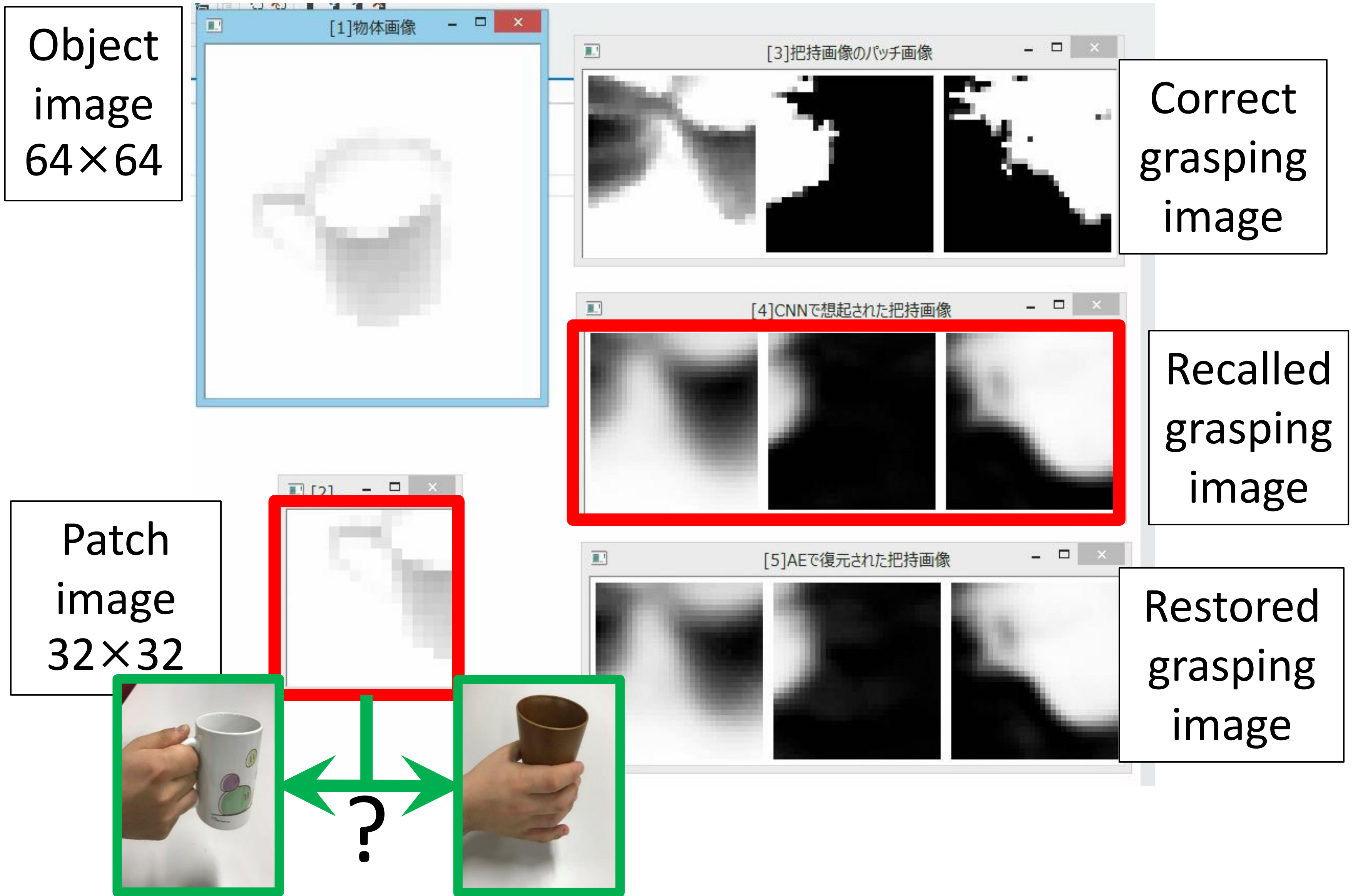




Recall from images with/without an important part



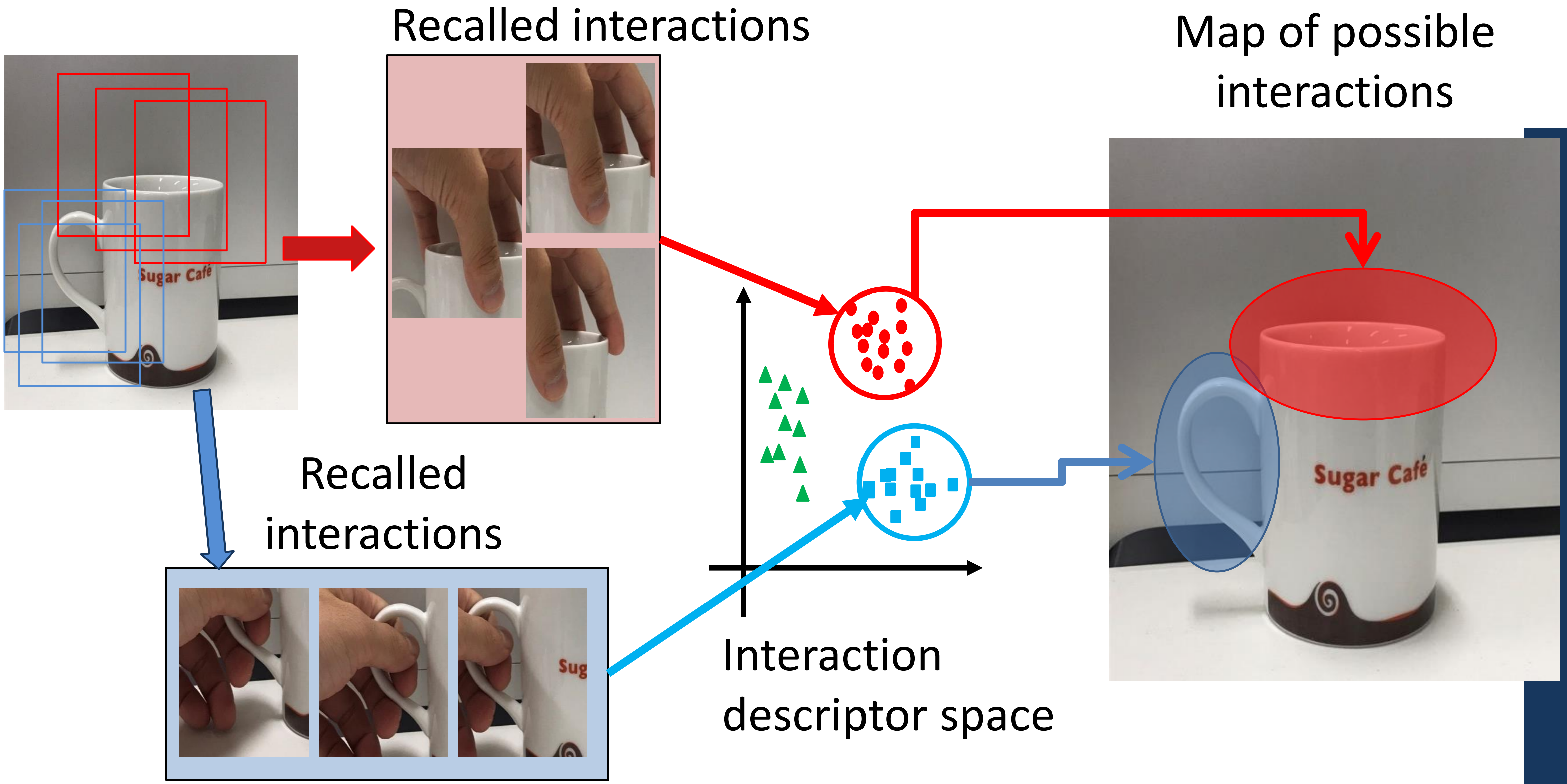
Recall from images with/without an important part



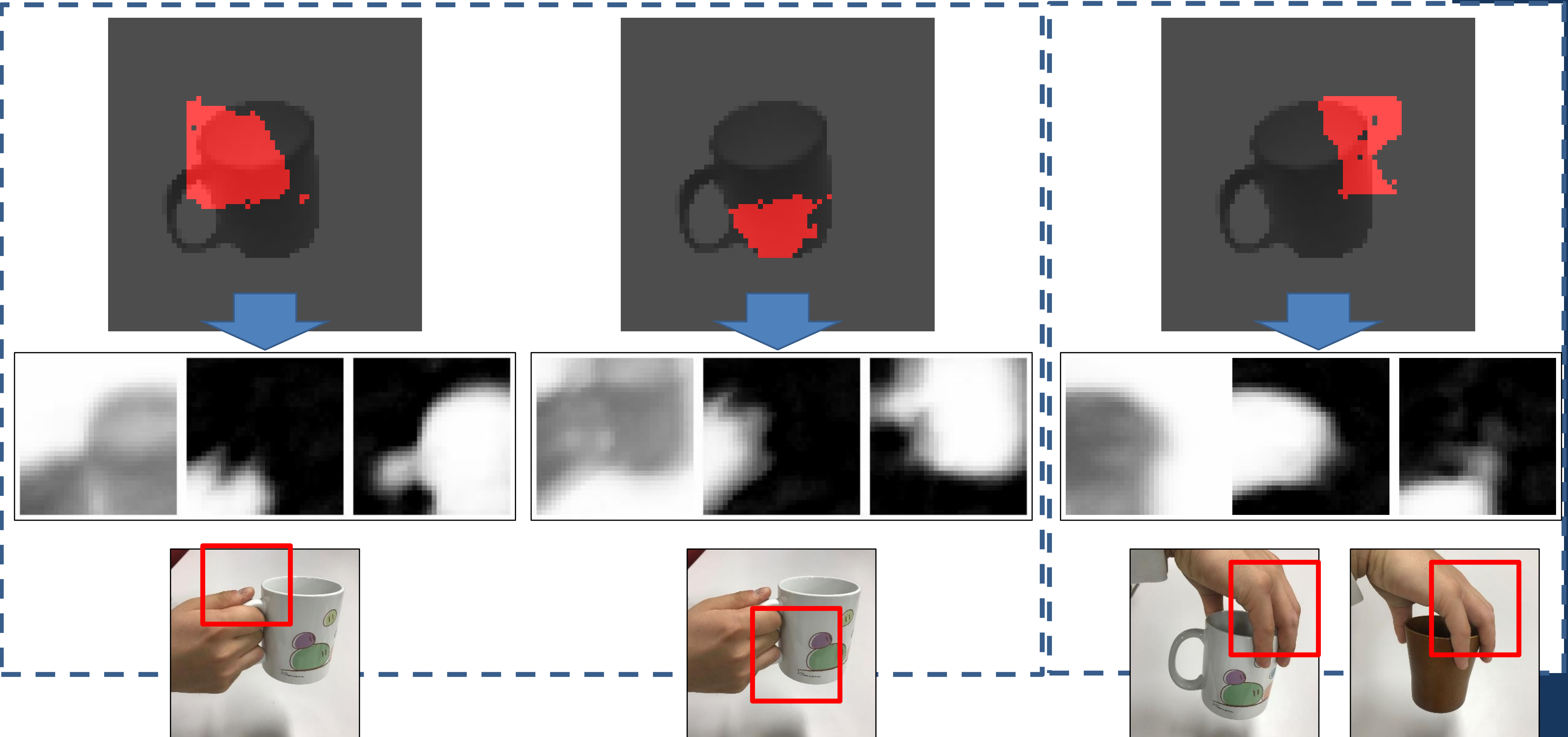
Flow of the presentation

1. Grasping image
2. Interaction descriptor
    1. Shift invariant auto-encoder
    2. Examples of shift invariant auto-encoder
    3. Results of interaction descriptor
3. Inference model
    1. Concept
    2. Recalled interactions from an object
    3. Interaction Map
4. Conclusion

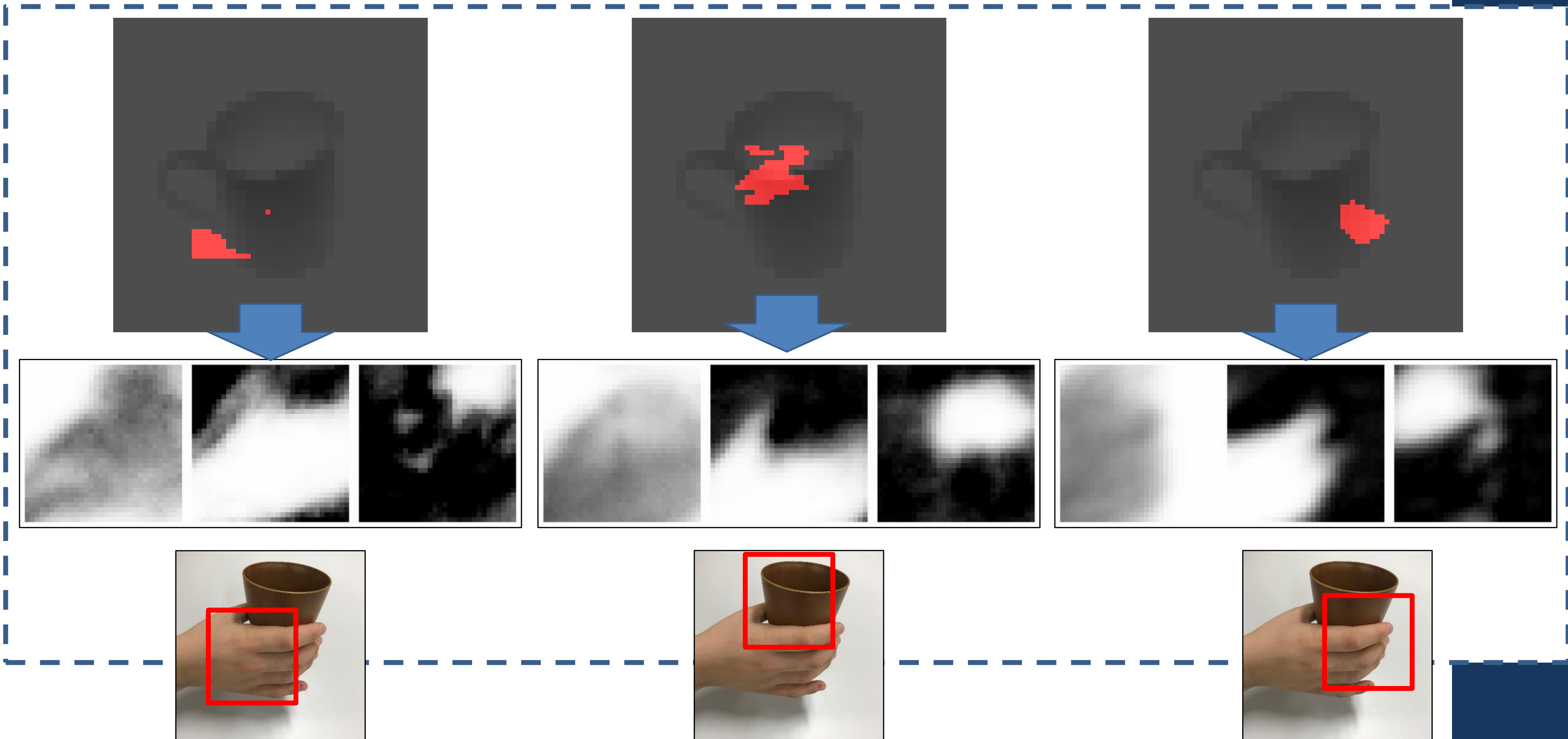
Interaction map on an object



Clusters of recalled descriptors (train)



Clusters of recalled descriptors(test)



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.

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

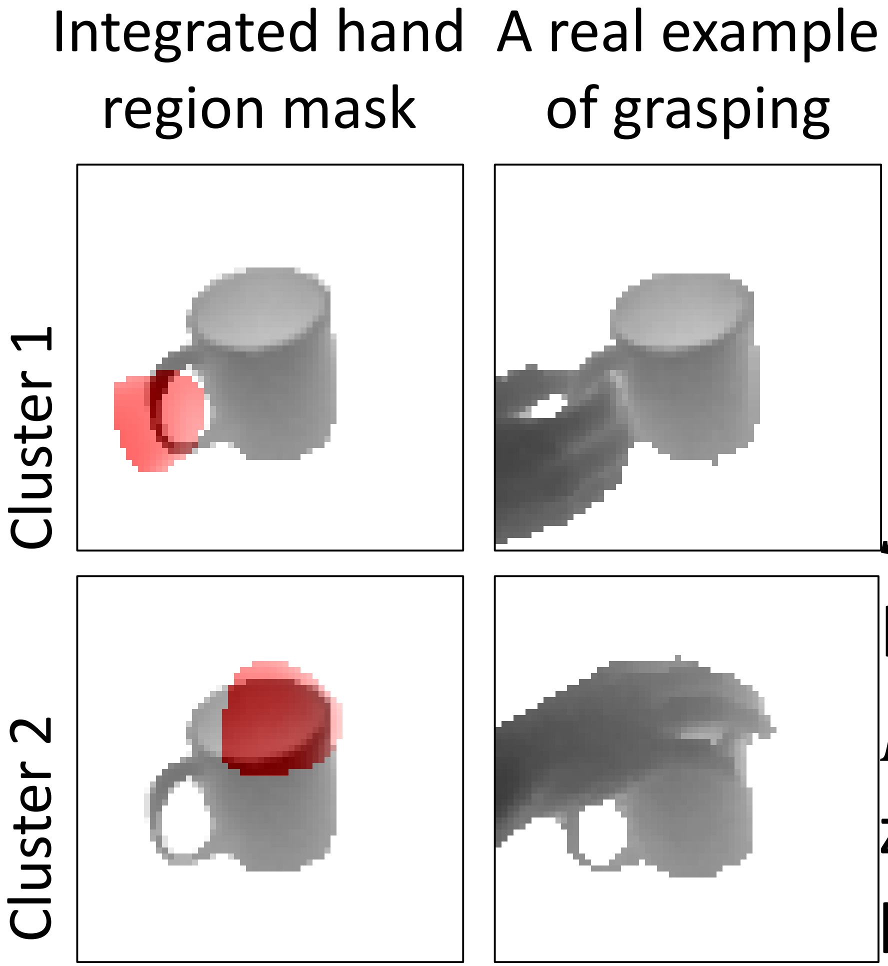
cup

Grasping type 1 | Grasping type 2

ball

spray

Integrated hand region mask



The integrated hand mask for cluster i is defined as:

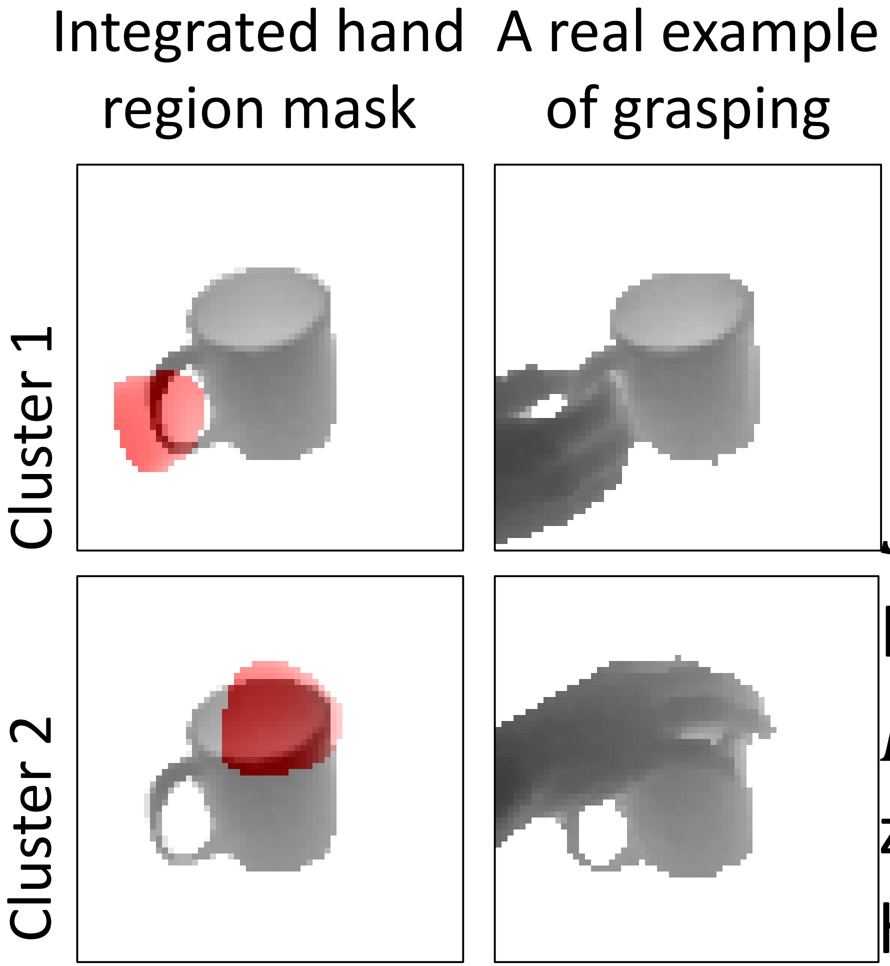
$$P_i(x,y) = \frac{S_i(x,y)}{N_i(x,y)}$$

$S_i(x,y)$: 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



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

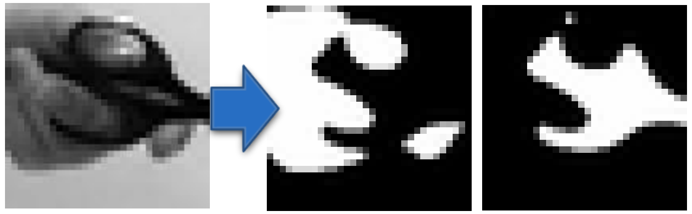
$$\ldots) = \frac{S_i(x,y)}{N_i(x,y)}$$

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

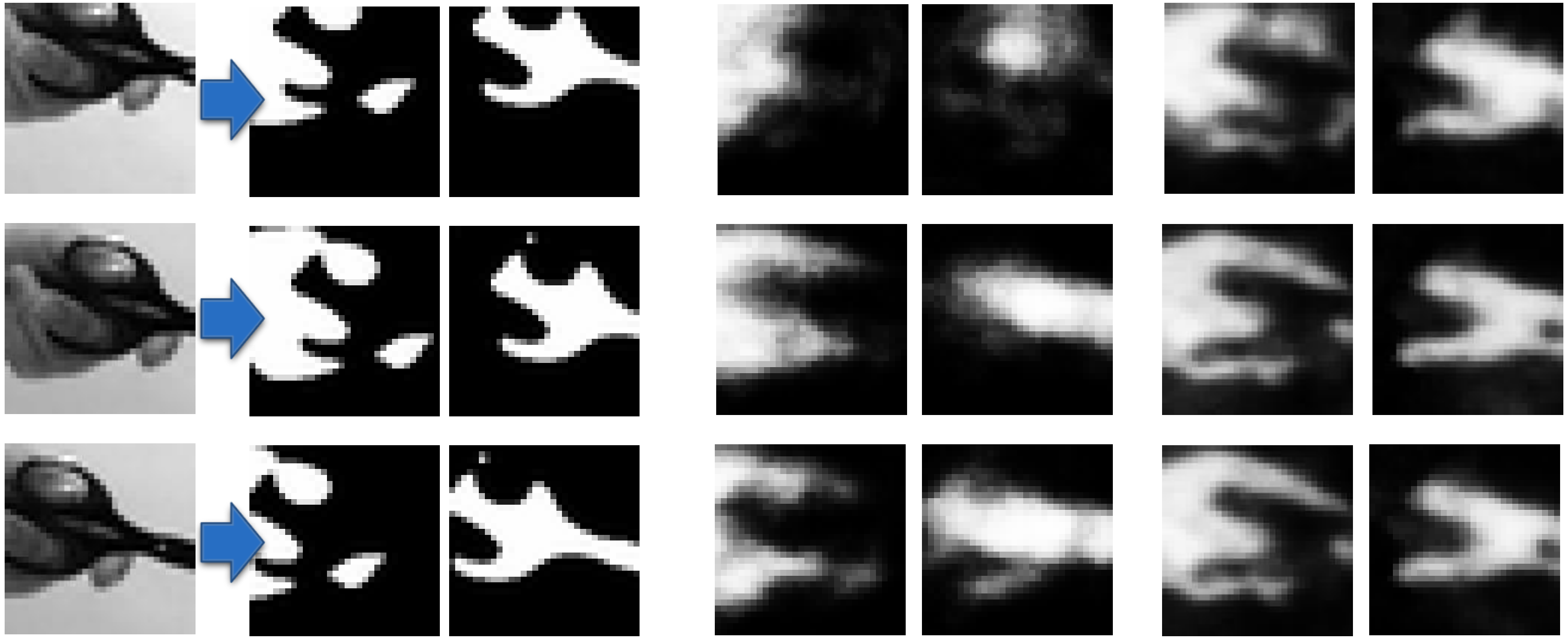
N... Number of non-ze...) of recalled hand mask in the i-th cluster

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

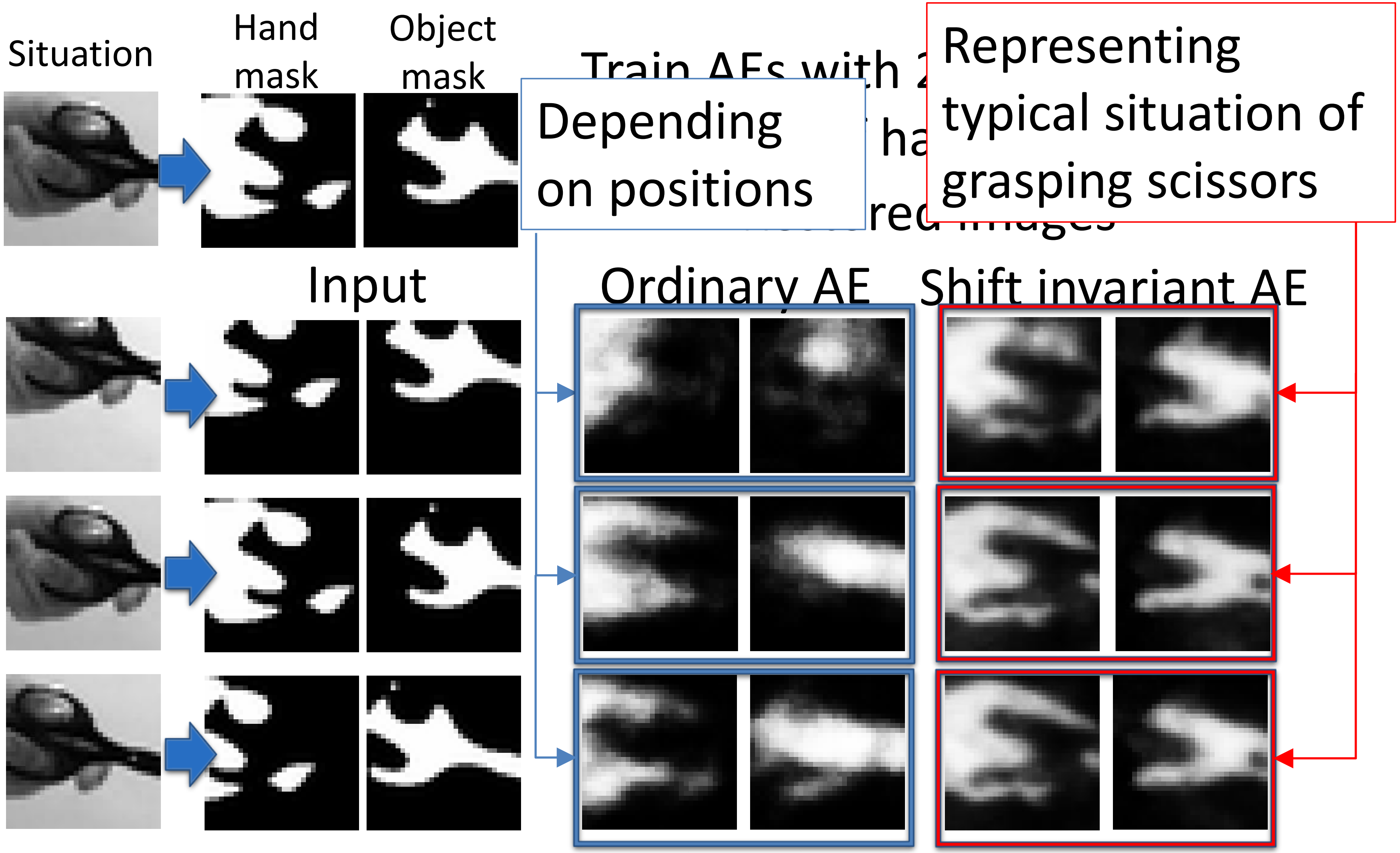
Example for hand-object interaction

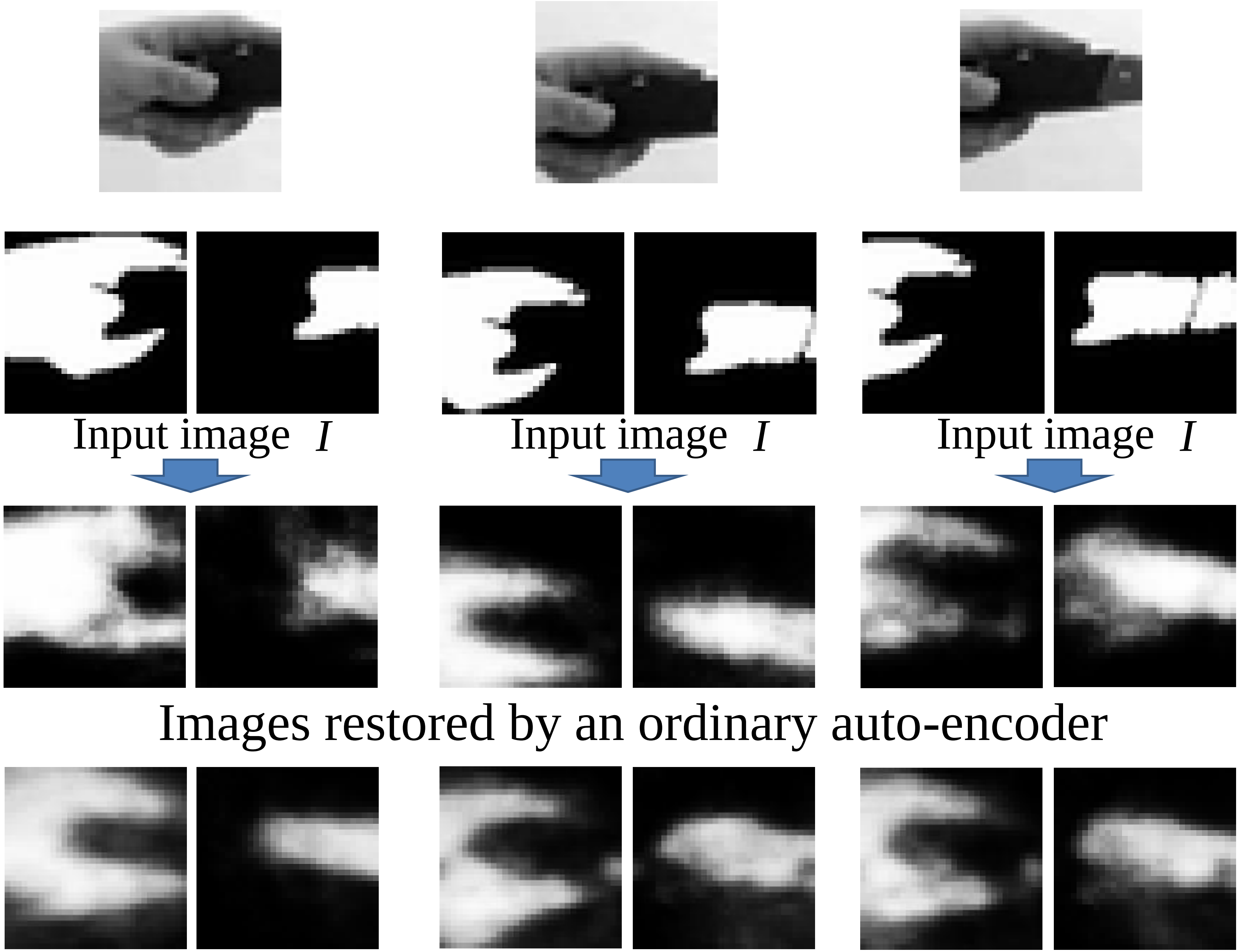


Train AEs with 2-channel images consisting of hand/object masks.



Example for hand-object interaction





Images restored by a shift invariant auto-encoder