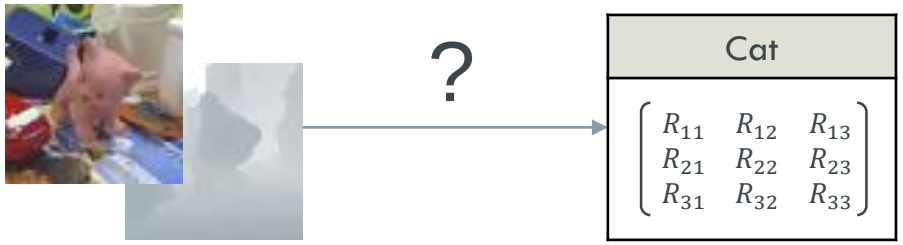
Technische Universität München

EXERCISE 3: 3D OBJECT INSTANCE RECOGNITION AND POSE ESTIMATION

MOTIVATION

Problem

- Classify and estimate a 3D pose of the object given its RGB, Depth, or RGB-D image
- Approach should be extensible and work on a large number of objects

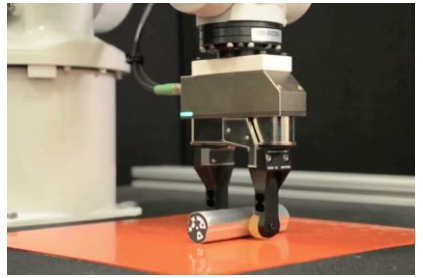


Input: Depth, RGB or RGB-D image

Output: instance and object pose

Applications

- Robotics
- Augmented Reality
- Tracking

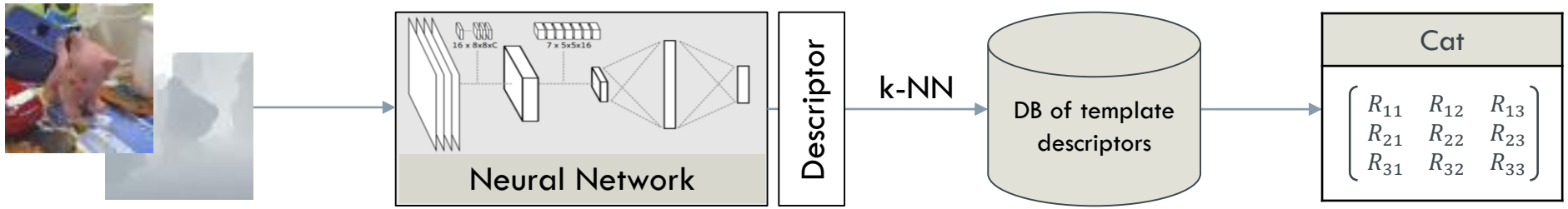


Robotic hand (MIT News)

Challenges

- Scalability with respect to the number of classes
- Scarcity of reliable training data
- Lack of powerful features
- Illumination, noise, background changes and occlusions

METHOD: WORKING PRINCIPLE



Input:
depth, RGB or RGB-D image

Descriptor CNN:
maps an image to a lower-dimensional space

k-NN:
finds a similar descriptor in the DB

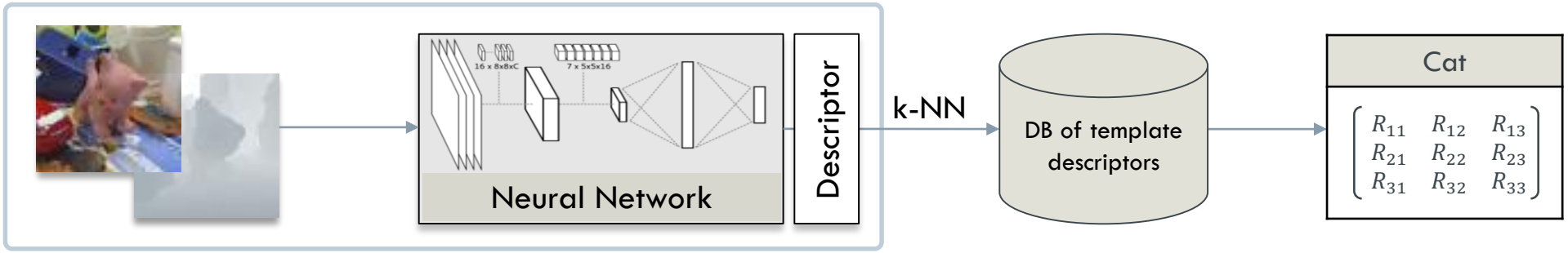
Output:
instance and object pose

Components:

- Descriptor CNN mapping an image to a descriptor space
- Database of template descriptors
- k-NN search on the database

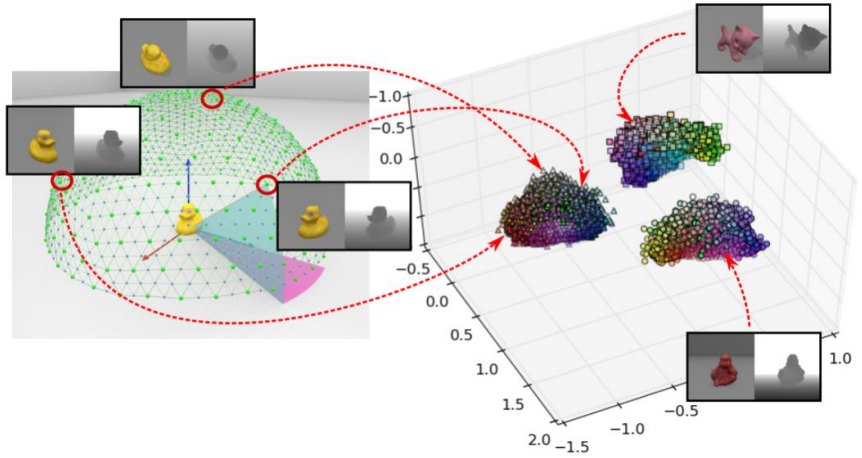
[4] Wohlhart, Paul, and Vincent Lepetit. "Learning descriptors for object recognition and 3d pose estimation." *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2015.

METHOD: MANIFOLD LEARNING



Descriptor properties:

- *Same object:*
  - Small Euclidian distance between the descriptors
  - Representative of the difference in pose
- *Different object:*
  - Large Euclidian distance between the descriptors



Mapping images to 3D descriptors [4]

[4] Wohlhart, Paul, and Vincent Lepetit. "Learning descriptors for object recognition and 3d pose estimation." *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2015.

METHOD: TRIPLET- AND PAIR-WISE TERMS

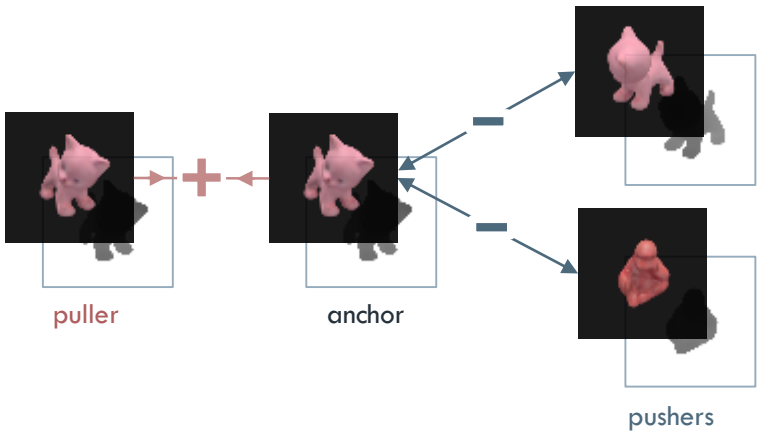
Triplet-wise terms

Define a **triplet** (s_a, s_+, s_-), where

- s_a and s_+ are the images of the same object and a similar pose
- s_a and s_- are the images of different objects or of the same object but with less similar poses

Cost function:

- $L_{triplets} = \sum_{(s_a,s_+,s_-)\in T} \max\left(0, 1 - \frac{\|f(x_a)-f(x_-)\|_2}{\|f(x_a)-f(x_+)\|_2+m}\right)$,
  - where f(x) is the output of the CNN for image x and m is a margin



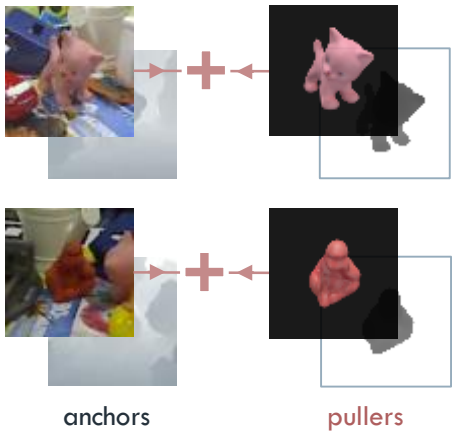
Pair-wise terms

Define a **pair** (s_a, s_+), where

- s_a and s_+ are the images of the same object and a similar pose
- Different background conditions, illumination, noise

Cost function:

- $L_{pairs} = \sum_{(s_a,s_+)\in P} \|f(x_a) - f(x_+)\|_2^2$,
  - where f(x) is the output of the CNN for image x



[4] Wohlhart, Paul, and Vincent Lepetit. "Learning descriptors for object recognition and 3d pose estimation." *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2015.

METHOD: DATASET GENERATION

Datatypes:

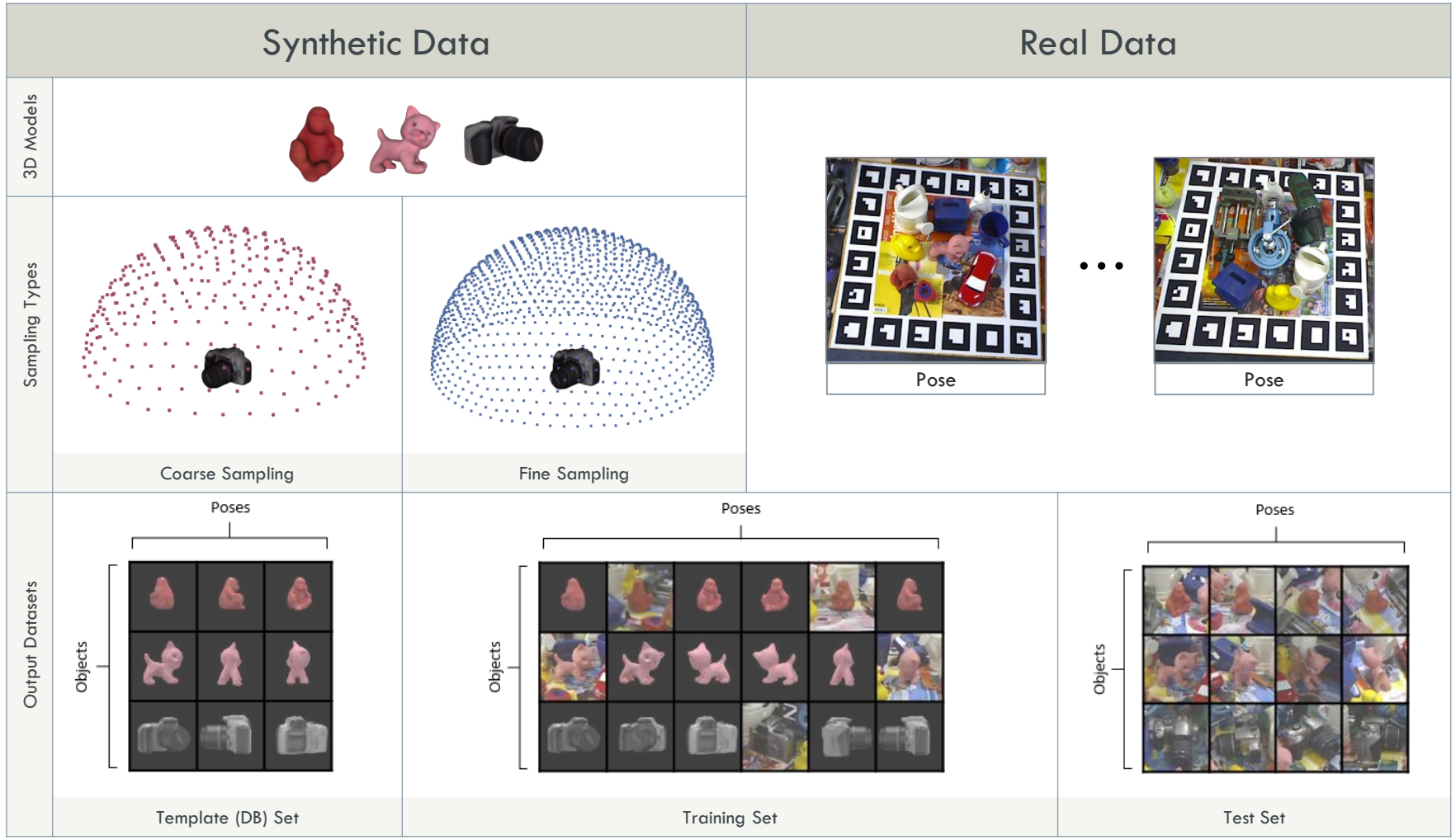
- Synthetic – render 3D models
- Real – use provided RGB-D images

1. Generate samples:

- Sample: patch + pose + object ID
- Patch: crop an image to get a patch of a certain size with an object located in the center

2. Generate datasets:

- Template: synthetic (coarse sampling) samples
- Training: synthetic (fine sampling) + real samples
- Test: real samples



METHOD: DATASET GENERATION – DEPTH MAP CROPPING

Datatypes:

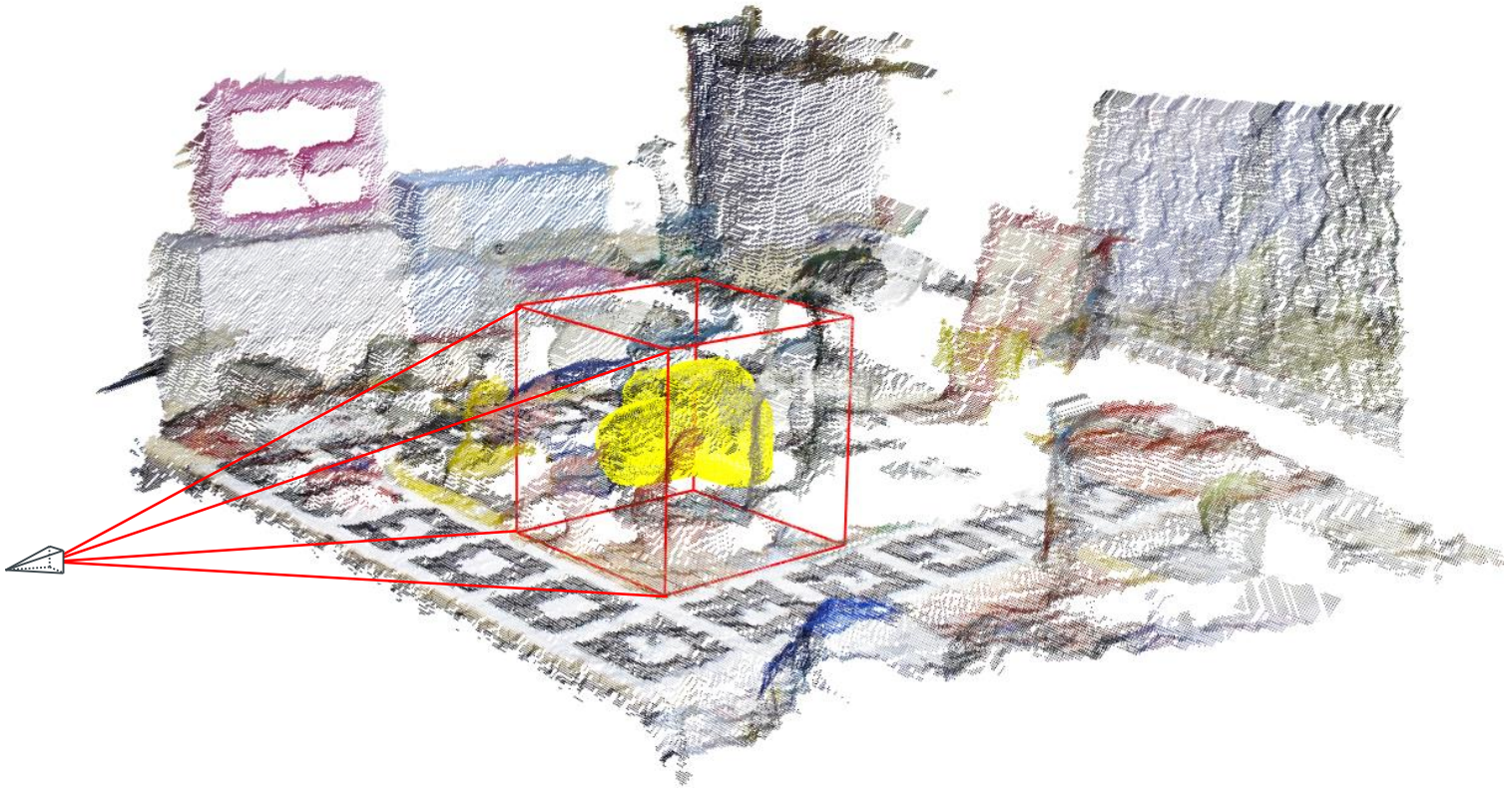
- Synthetic – render 3D models
- Real – use provided RGB-D images

1. Generate samples:

- Sample: patch + pose + object ID
- Patch: crop an image to get a patch of a certain size with an object located in the center

2. Generate datasets:

- Template: synthetic (coarse sampling) samples
- Training: synthetic (fine sampling) + real samples
- Test: real samples



METHOD: DATASET GENERATION

Datatypes:

- Synthetic – render 3D models
- Real – use provided RGB-D images

1. Generate samples:

- Sample: patch + pose + object ID
- Patch: crop an image to get a patch of a certain size with an object located in the center

2. Generate datasets:

- Template: synthetic (coarse sampling) samples
- Training: synthetic (fine sampling) + real samples
- Test: real samples