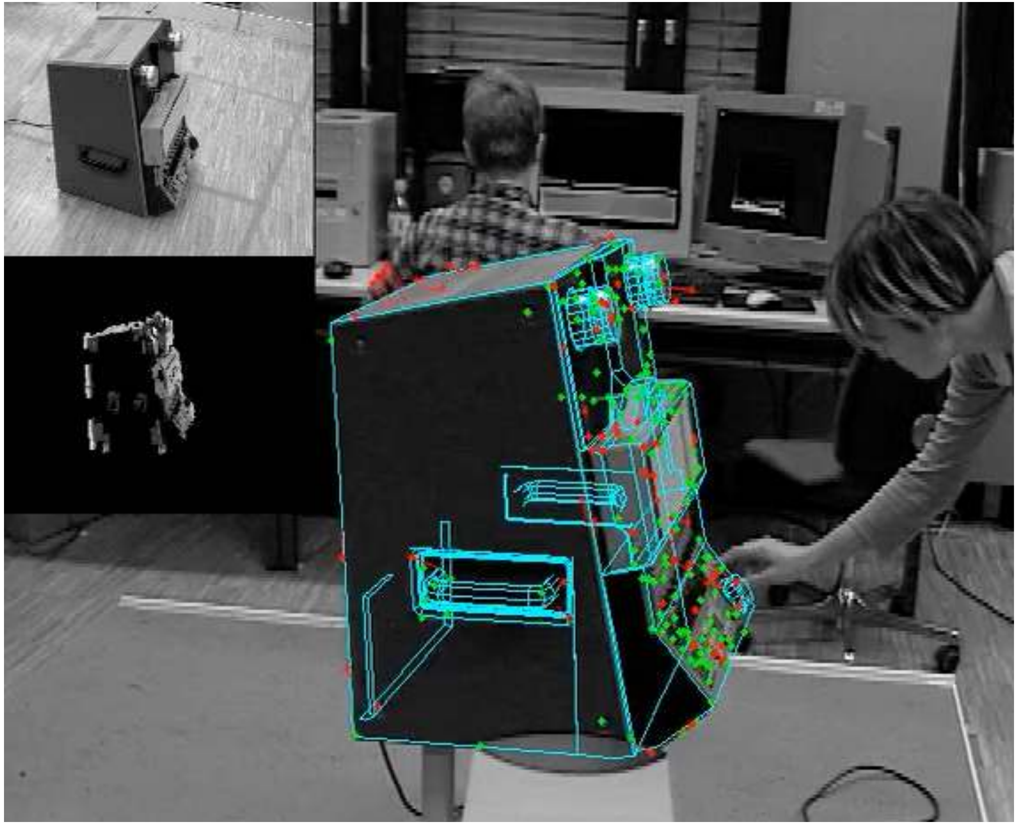
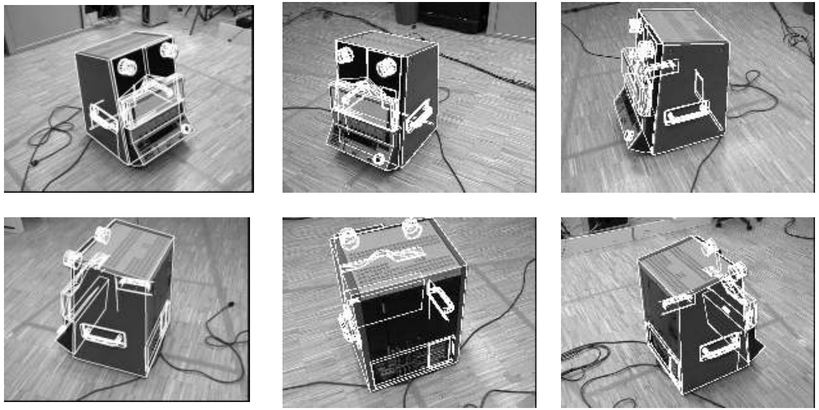
New Problems in 3D Object Pose Estimation

Vincent Lepetit
ENPC ParisTech, France

École des Ponts ParisTech

TU Graz
Graz University of Technology





reference frames

2003: Real-time tracking with feature points and offline and online reference frames.

2003: Real-time face tracking (with feature points)

but ..
how do we initialize object tracking?

→ let's work on 3D object *detection*

2005: Real-time feature point matching with randomized trees, and later binary descriptors

but ..
what if the object does not have enough feature points?

→ let's work on texture-less objects

École des Ponts ParisTech

2010: 3D object detection with templates

interactive template creation

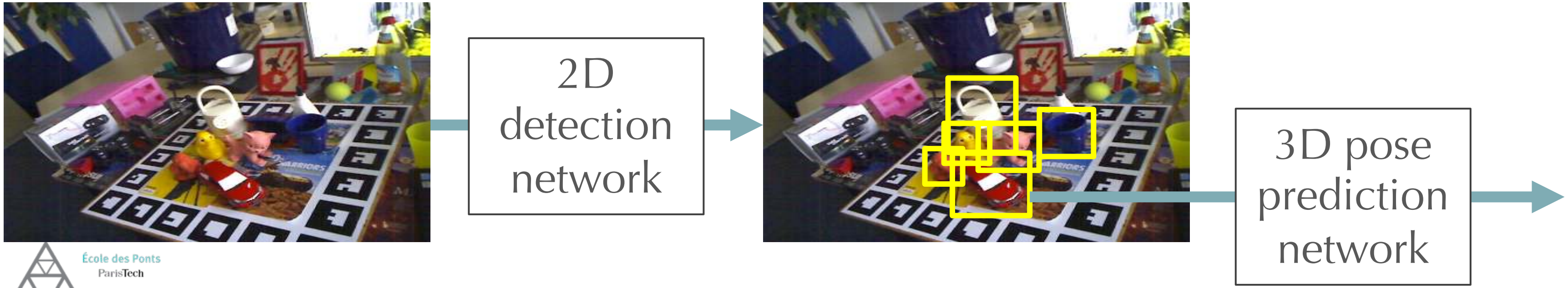
→ research driven by practical problems

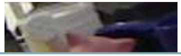
[also, real-time demos are cool]

current standard approach to 3D object pose estimation

1) network(s) training

2) inference

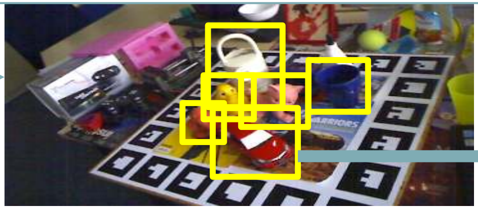




- needs a lot of annotated data (ok, you knew that already);

- needs *training time*;

- needs annotated real data for evaluation (and to help learning).

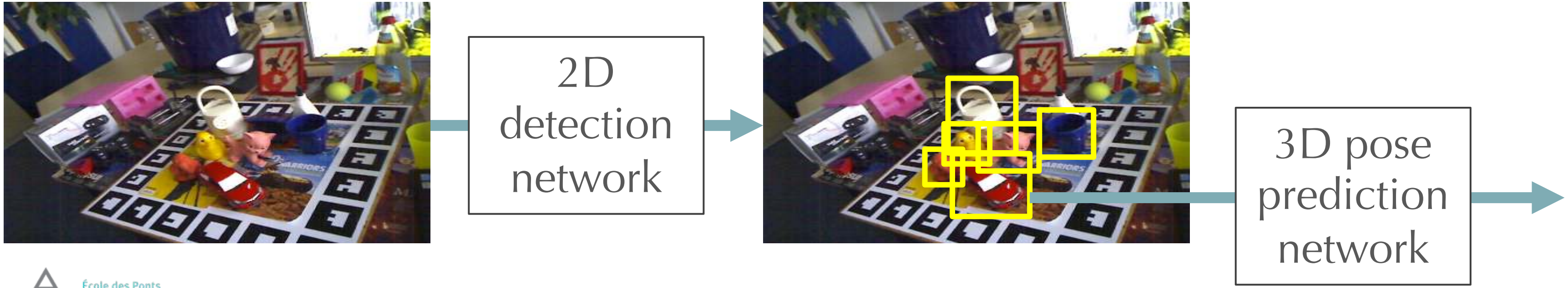


About **learning time** requirement: Can we

- **detect** and

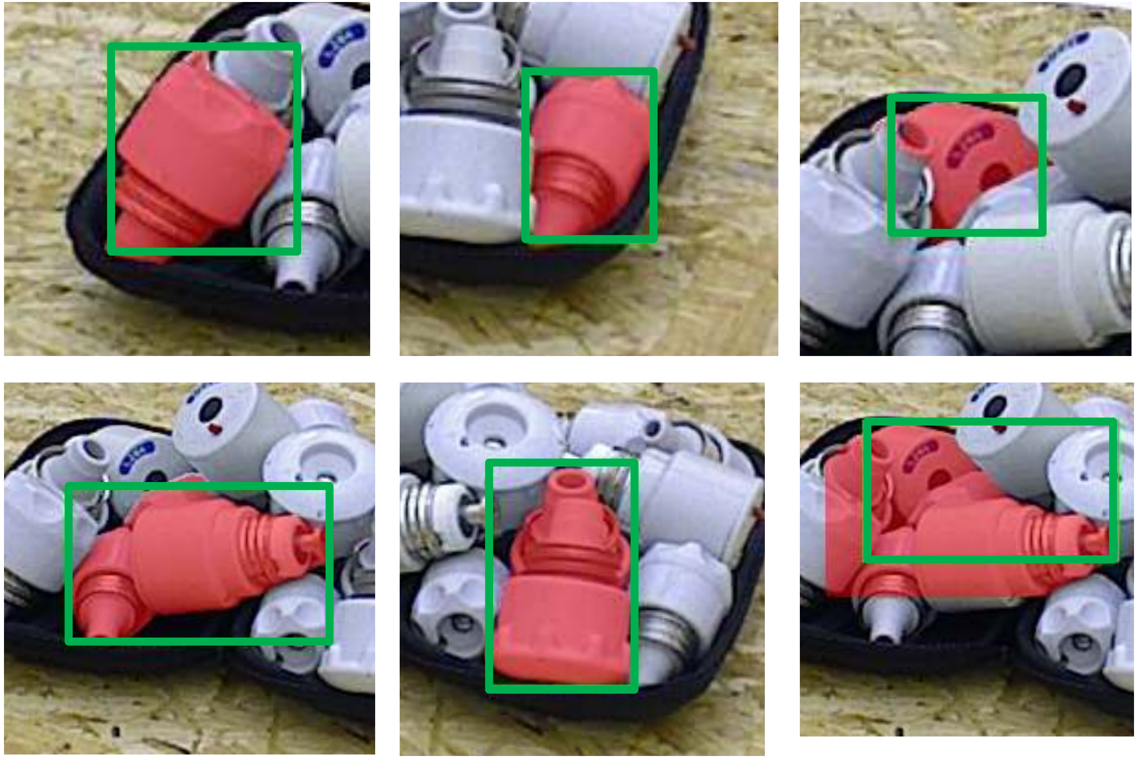
- predict the **3D pose** of objects

***without* learning** for these specific objects?
(and still use deep learning)



École des Ponts
ParisTech

Detecting unknown objects in 2D



Architecture trained on the UVO dataset (similar to COCO)

***not* trained on T-LESS**

1st Place Solution for the UVO Challenge on Image-based Open-World Segmentation 2021. Yuming Du, Wen Guo, Yang Xiao, and Vincent Lepetit. ICCV Workshop, 2021. (Code available)

Detecting unknown objects in 2D

2D detection with Swin-L Transformer → bounding box / bounding box / .. / bounding box → segmentation with Swin-L Transformer →

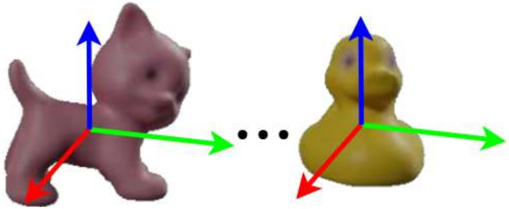
Trained in a class-agnostic way;

Training on many objects and the use of Transformers make the architecture generalize well to new objects.

Predicting the 6D pose of *new* objects without learning

Scenario:

- we just got the 3D models for new objects:

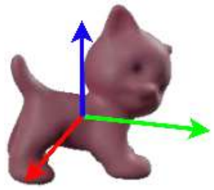


- we want to predict the pose of these objects NOW (*i.e.*, without retraining a deep network):

Templates for 3D Object Pose Estimation Revisited: Generalization to New Objects and Robustness to Occlusions. Van Nguyen Nguyen, Yinlin Hu, Yang Xiao, Mathieu Salzmann, and Vincent Lepetit. CVPR 2022.

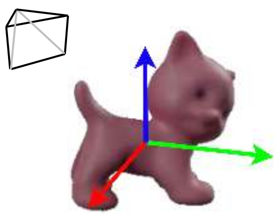
Using templates for new objects

Incoming new objects:



Using templates for new objects

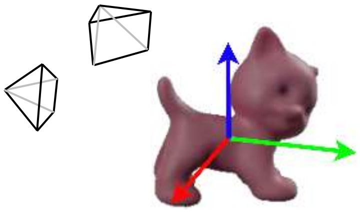
Incoming new objects:

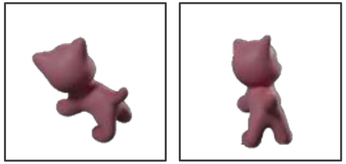


École des Ponts
ParisTech

Using templates for new objects

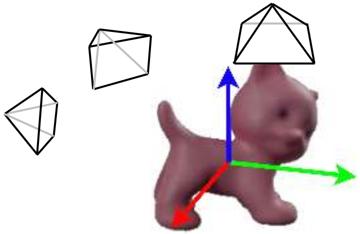
Incoming new objects:





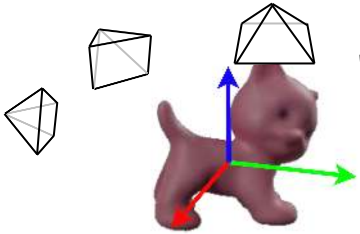
Using templates for new objects

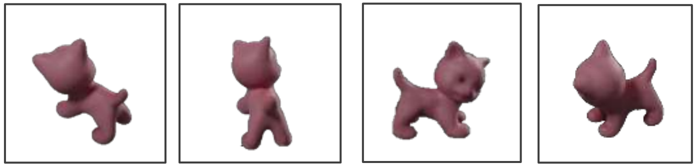
Incoming new objects:



Using templates for new objects

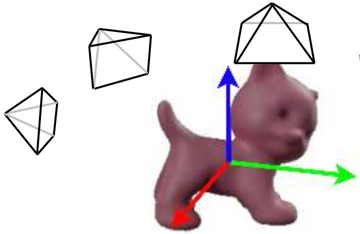
Incoming new objects:



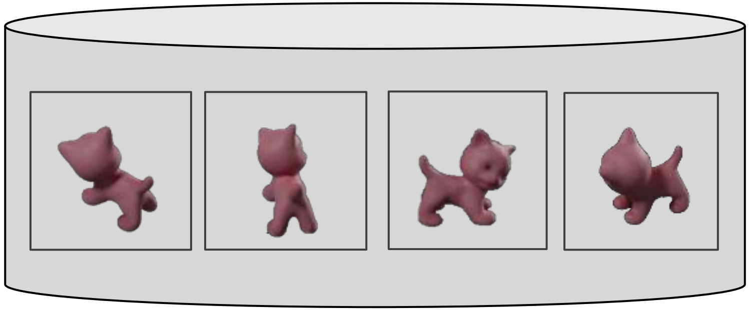


Using templates for new objects

Incoming new objects:

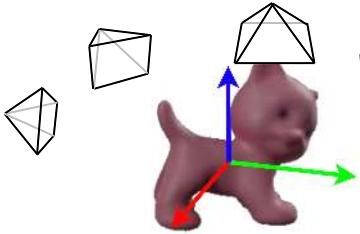


"short-term memory"

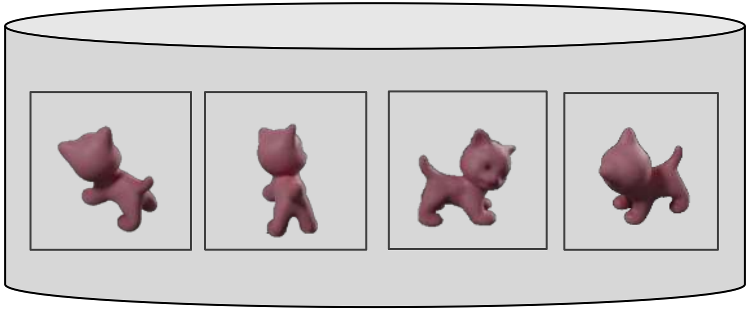


Using templates for new objects

Incoming new objects:



"short-term memory"

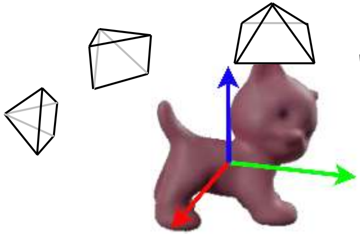


Inference:

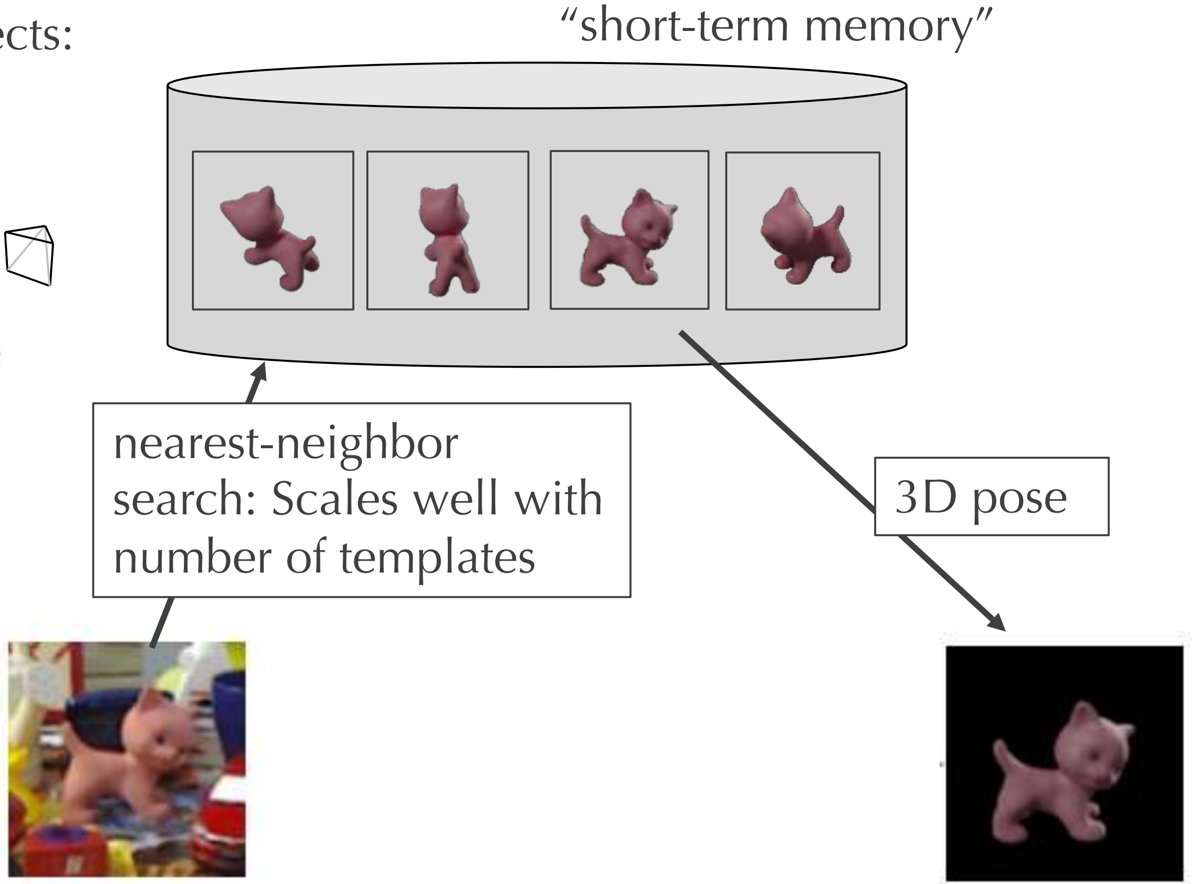
École des Ponts ParisTech

Using templates for new objects

Incoming new objects:



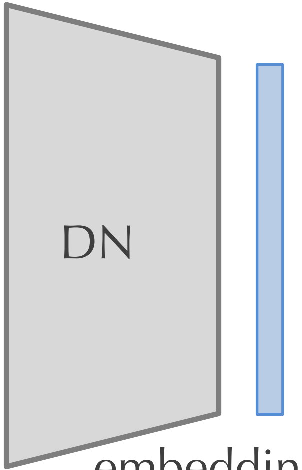
Inference:



École des Ponts ParisTech

template embeddings

template

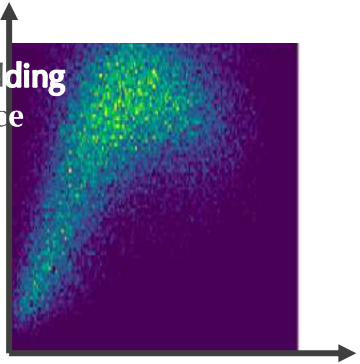


embedding

learned with the InfoNCE loss on real and synthetic images of **known objects**

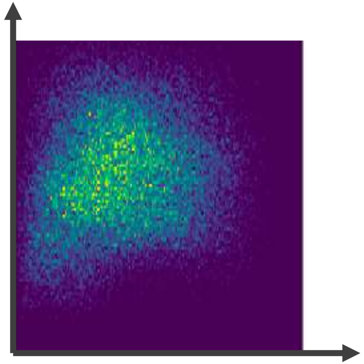
correlation between '**pose distances**' and '**embedding distances**':

embedding distance



pose distance

for a **known object**, on which the embedding was trained

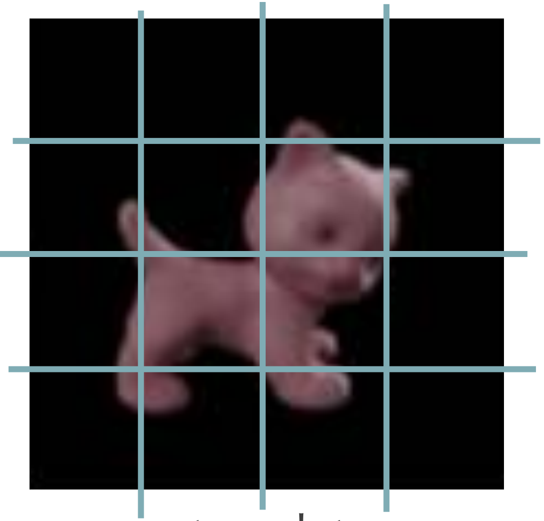


for a **new object**

DN does not generalize well ☹

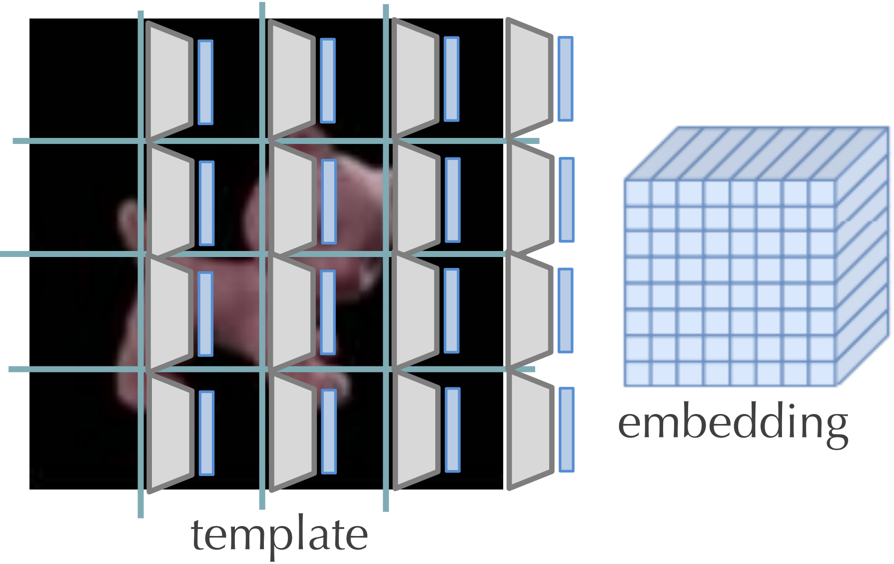
École des Ponts ParisTech

template embeddings

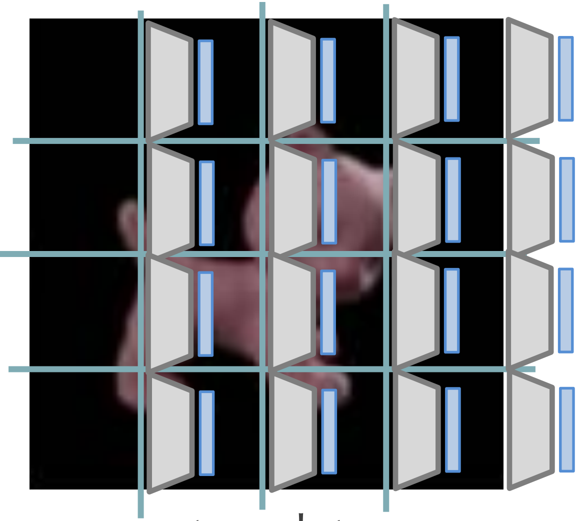


template

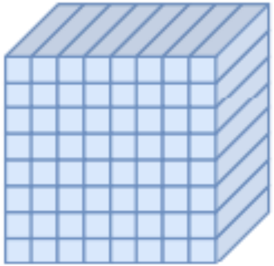
template embeddings



template embeddings

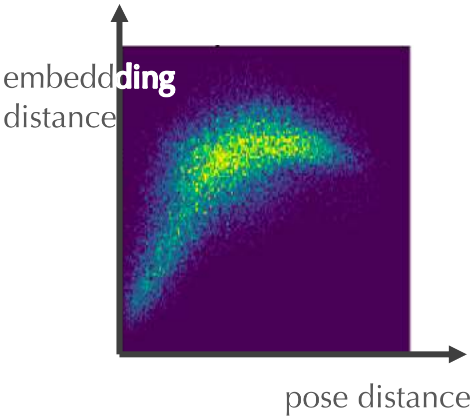


template

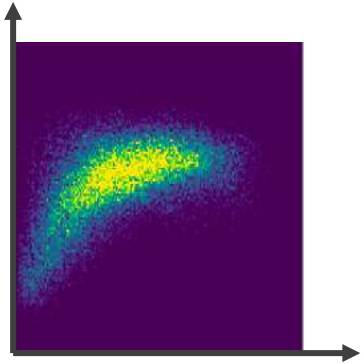


embedding

correlation between 'pose distances' and 'embedding distances':



for an object on which the embedding was trained



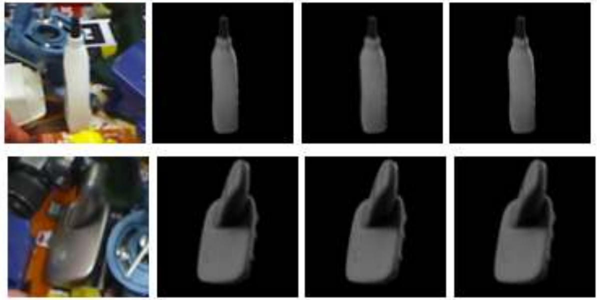
for a new object

Embedding computed from local parts generalizes better to new objects

École des Ponts
ParisTech

Some results

objects used to learn the embedding:



new objects:

the new objects can be very different from the objects used to learn the embedding

objects used to learn the embedding under occlusions:

query | ground truth | 'global' embedding | 'local parts' embedding

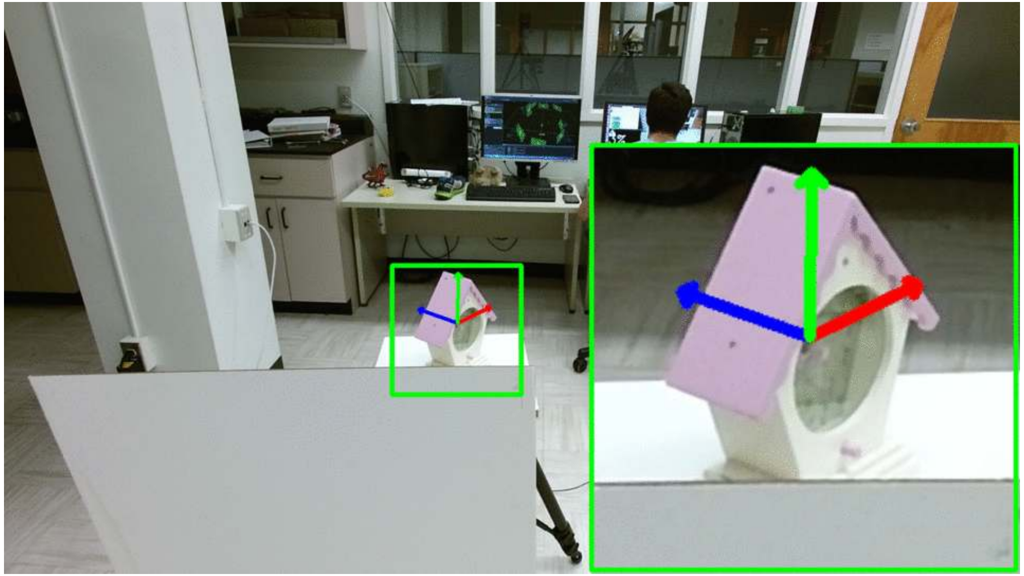
new objects under occlusions:

query | ground truth | 'global' embedding | 'local parts' embedding

bonus: comparing the 'local parts' embeddings is robust to occlusions

Predicting the 6D *motion* of *unknown* objects without learning

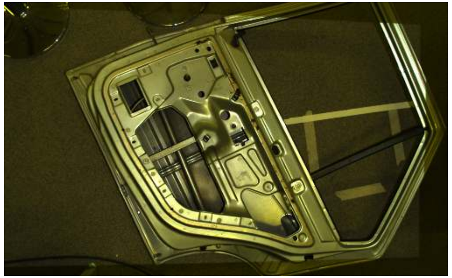
no depth maps, no CAD models, no prior images, no retraining:



oral at 3DV
presented on Thursday morning

PIZZA: A Powerful Image-only Zero-Shot Zero-CAD Approach to 6DoF Tracking. Van Nguyen Nguyen, Yuming Du, Yang Xiao, Michaël Ramamonjisoa, Vincent Lepetit. Oral at 3DV 2022.

new dataset, by CEA



Large range of scales:
→ more ambiguities
→ bounding boxes outside the images (this makes the current approach fail)

automatically annotating 3D data

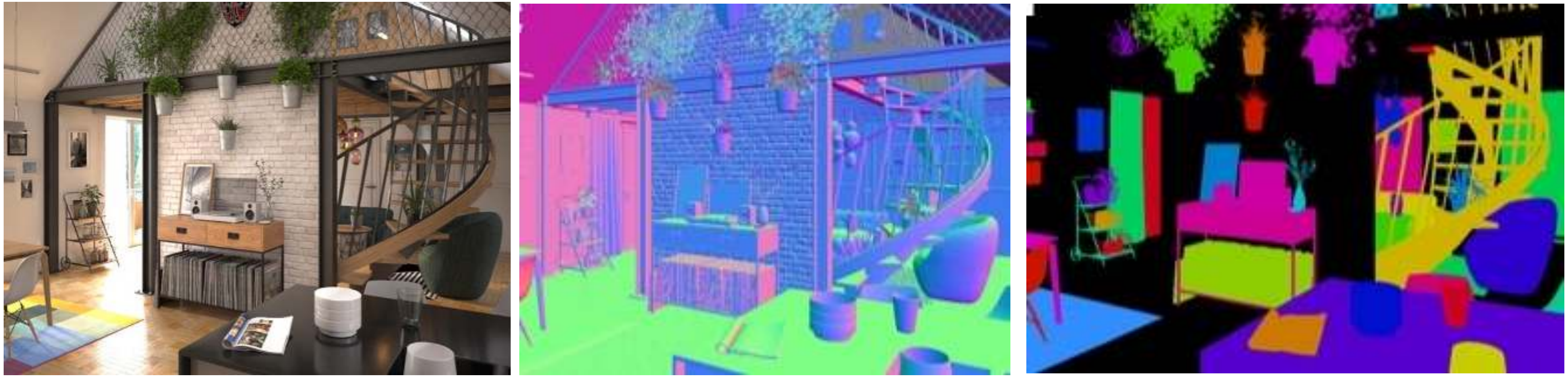
École des Ponts
ParisTech

annotated training data

SUN-RGBD dataset, ~10’000 images annotated manually
2,051 hours for annotations by oDesk workers +
corrections by the paper’ authors

École des Ponts
ParisTech

synthetic images?



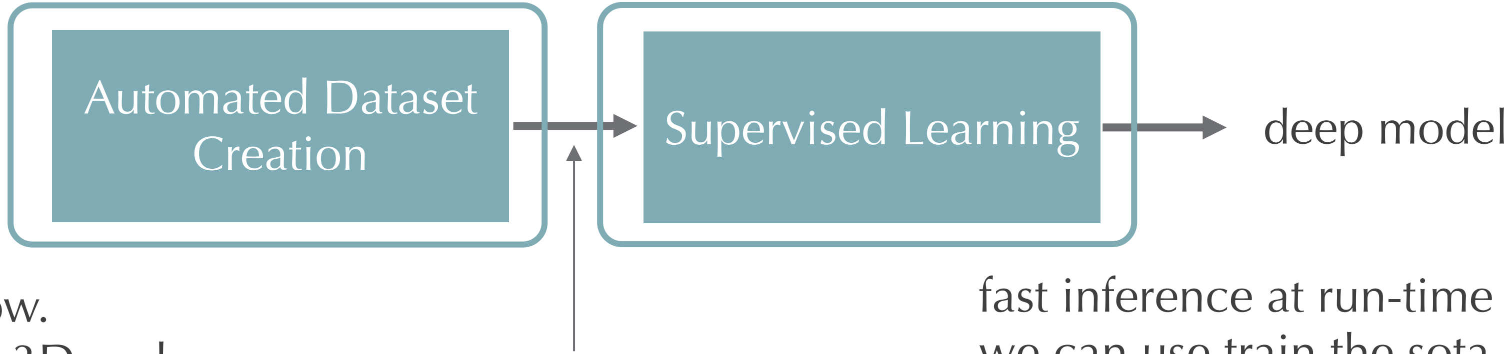
About 71′000 synthetic images. Costs $57K to create (scene creation + image rendering), and took 231 vCPU years (2.4 years of wall-clock time on a large compute node).

[Mike Roberts and Nathan Paczan. Hypersim: A Photorealistic Synthetic Dataset for Holistic Indoor Scene Understanding. In *arXiv*, 2020]

École des Ponts
ParisTech

Generating 3D labels automatically

(almost) solved once the dataset is large enough



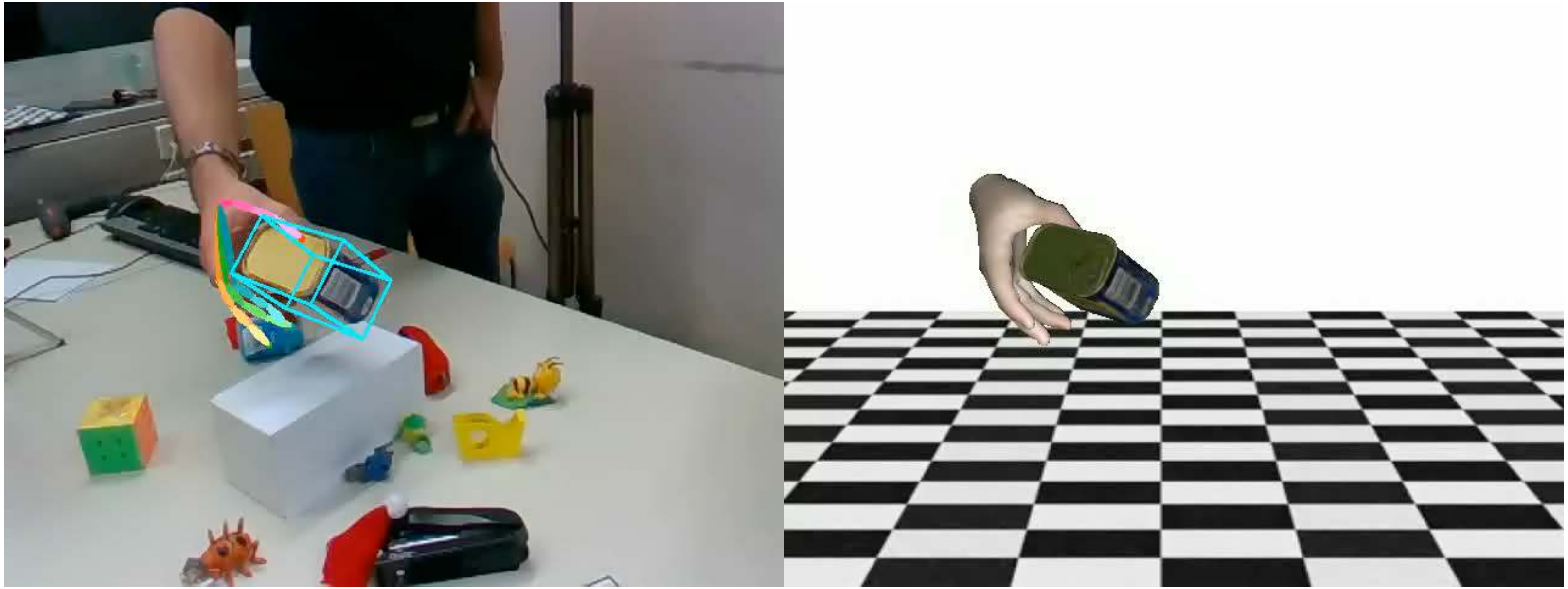
ok if slow.
can use 3D and temporal constraints, additional sensors, etc.

we can check the quality of the self-annotations

fast inference at run-time
we can use train the sota methods on our annotations

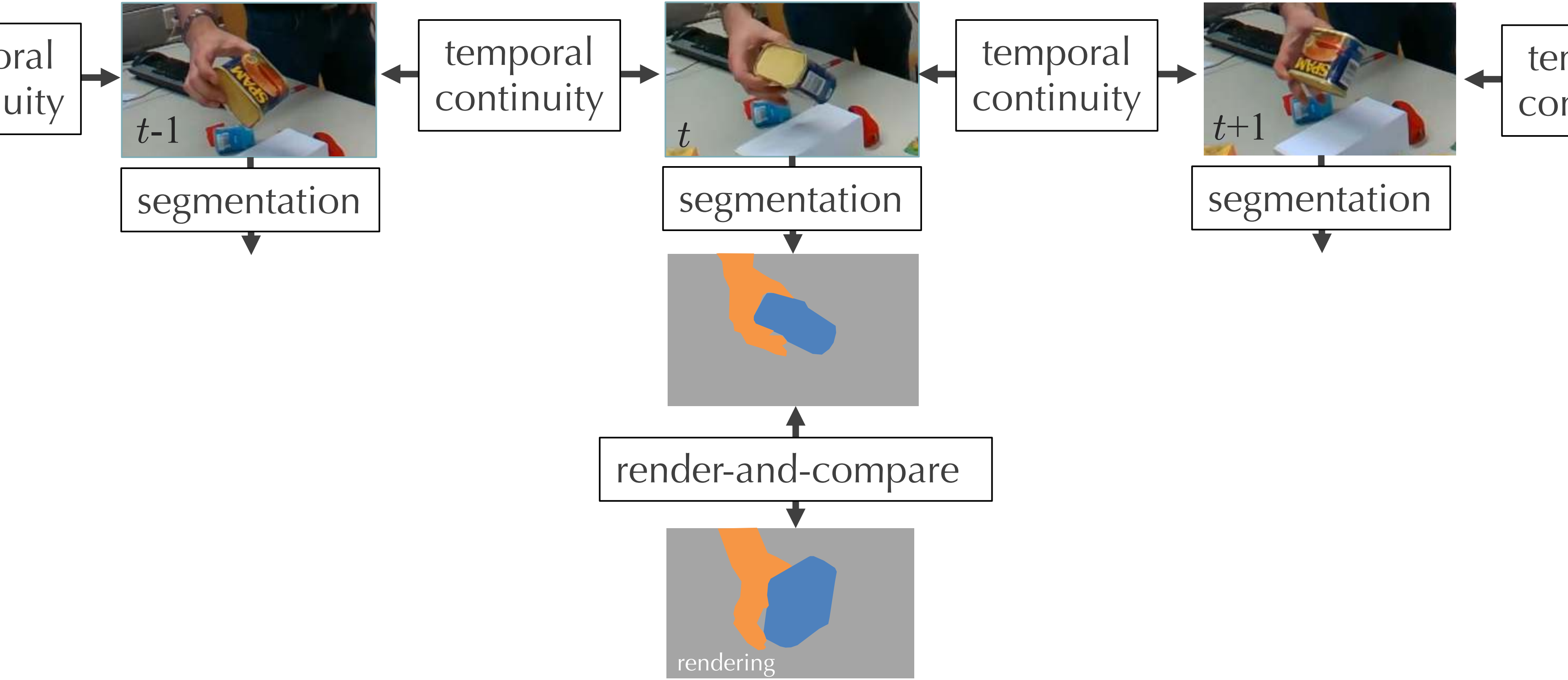
École des Ponts
ParisTech

HOnnotate



HOnnotate: A Method for 3D Annotation of Hand and Object Poses. Shreyas Hampali, Mahdi Rad, Markus Oberweger, and Vincent Lepetit. CVPR 2020.

global optimization



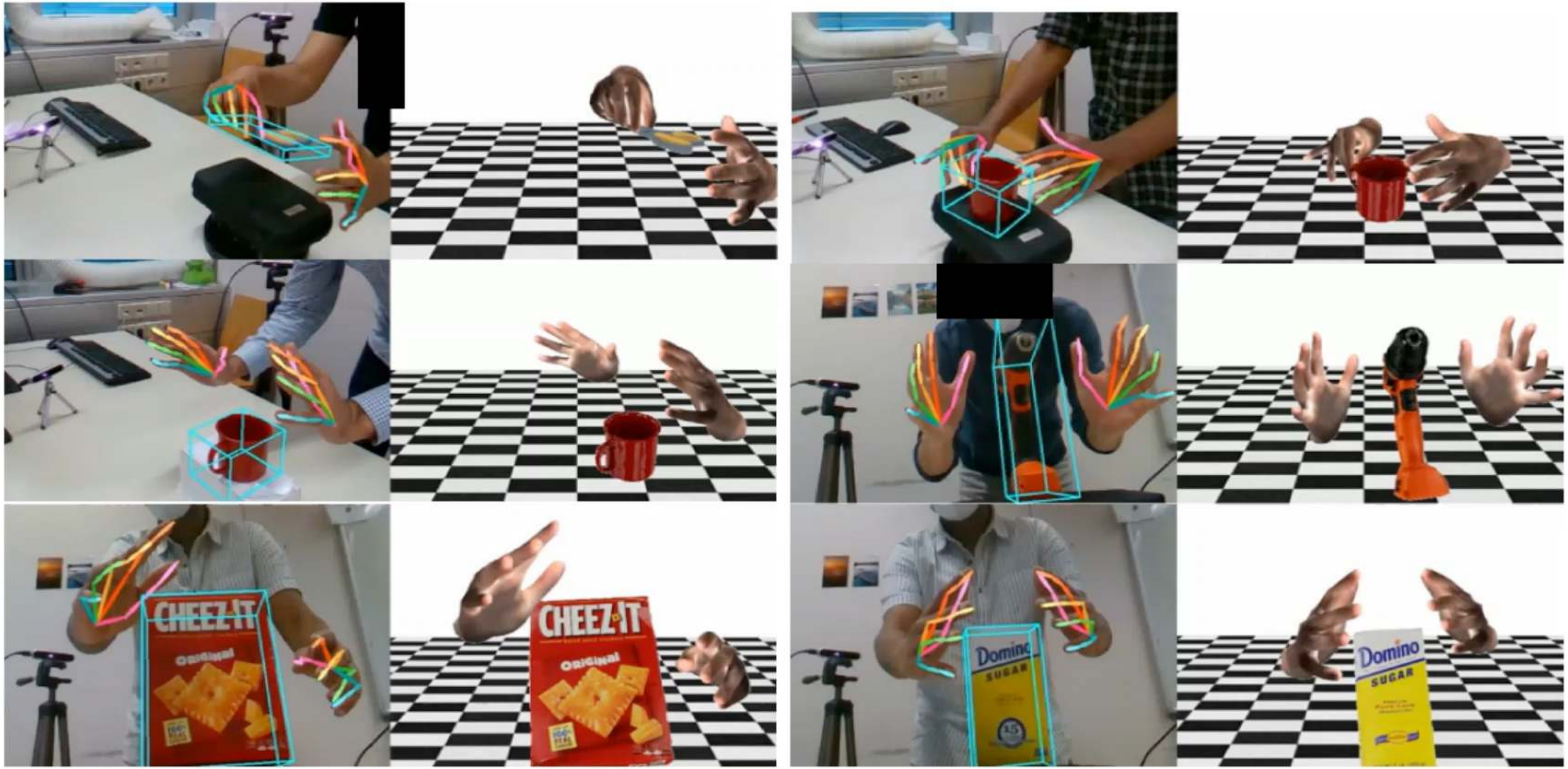
Vision as Bayesian inference: analysis by synthesis? Alan Yuille and Daniel Kersten. Trends in Cognitive Science, 2006.

validation

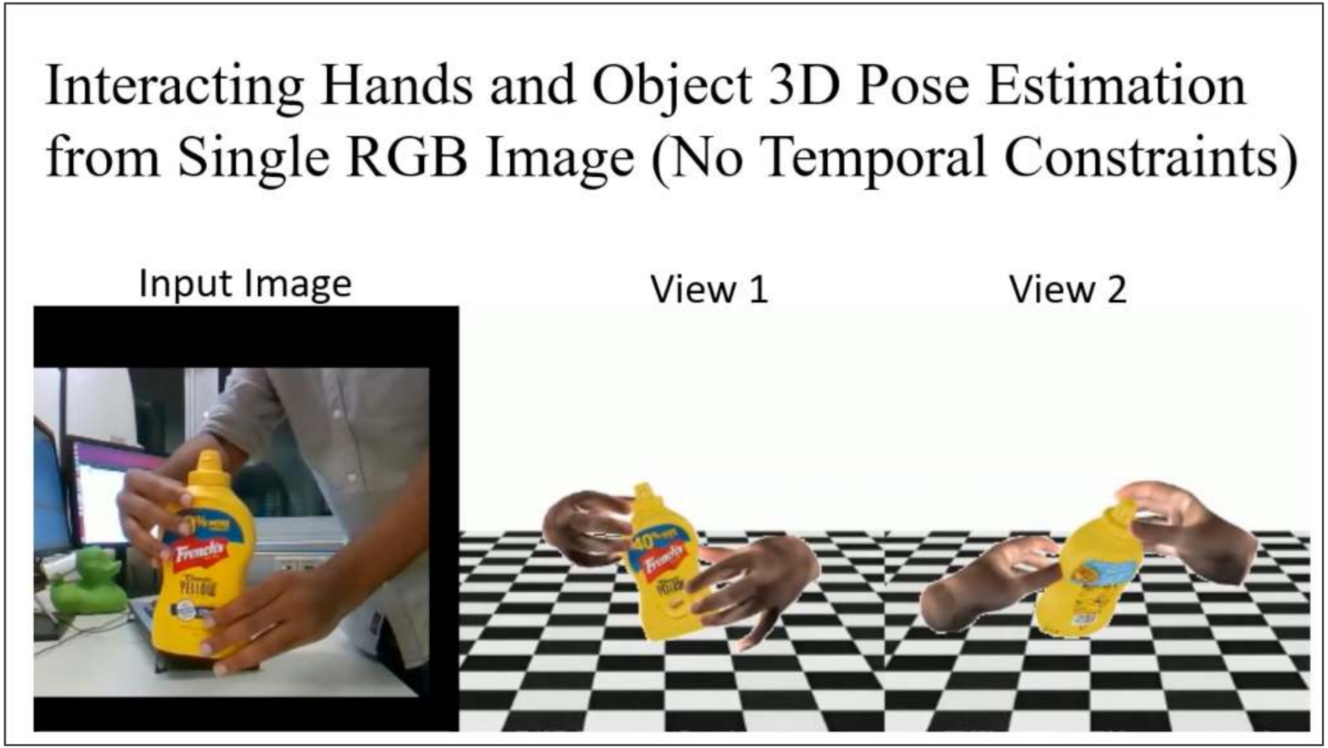
joints localized manually using the point cloud, and compared to the retrieved joint locations


SPA

H2O-3D


CHEEZ-IT
ORIGINAL
CHEEZ-IT
ORIGINAL
Domino
SUGAR
Domino
SUGAR

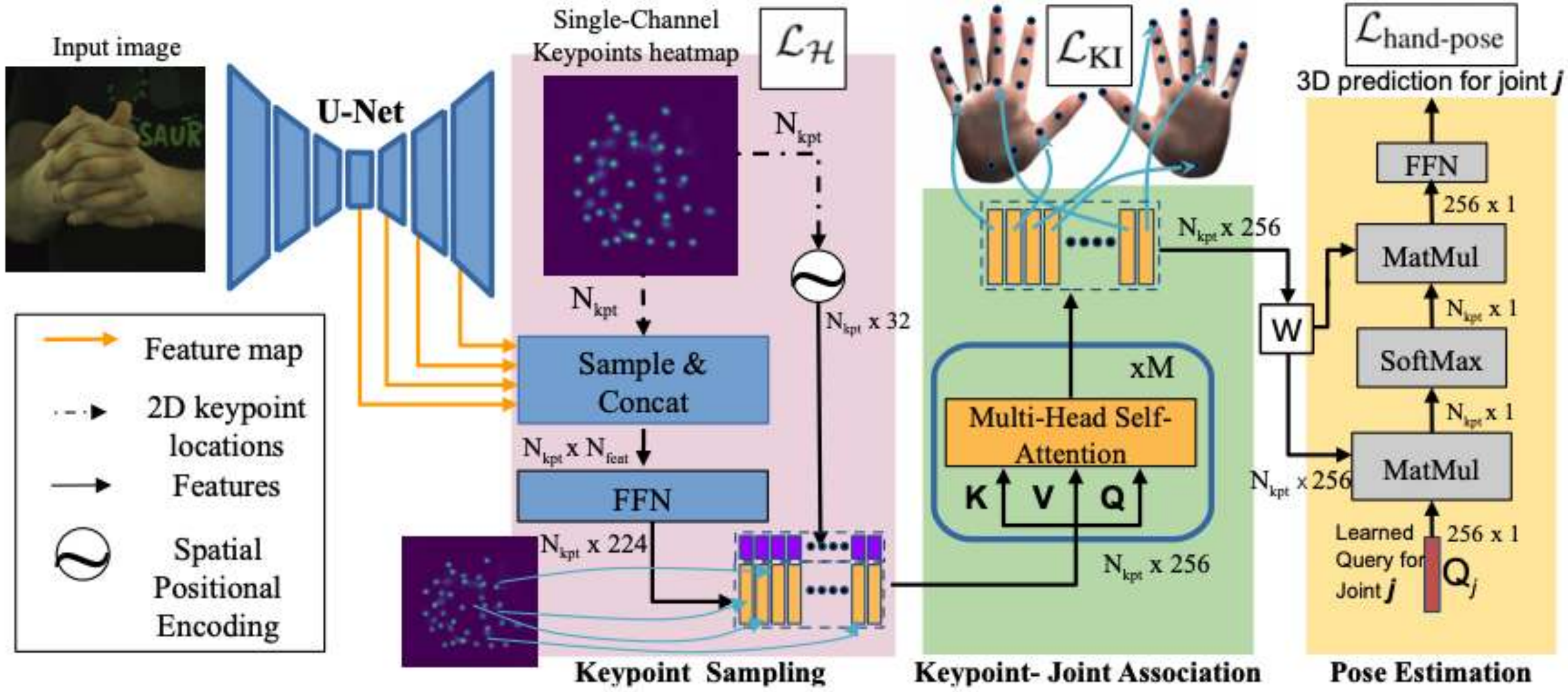
Keypoint Transformer



Keypoint Transformer: Solving Joint Identification in Challenging Hands and Object Interactions for Accurate 3D Pose Estimation. Shreyas Hampali, Sayan Deb Sarkar, Mahdi Rad, Vincent Lepetit. Oral at CVPR 2022

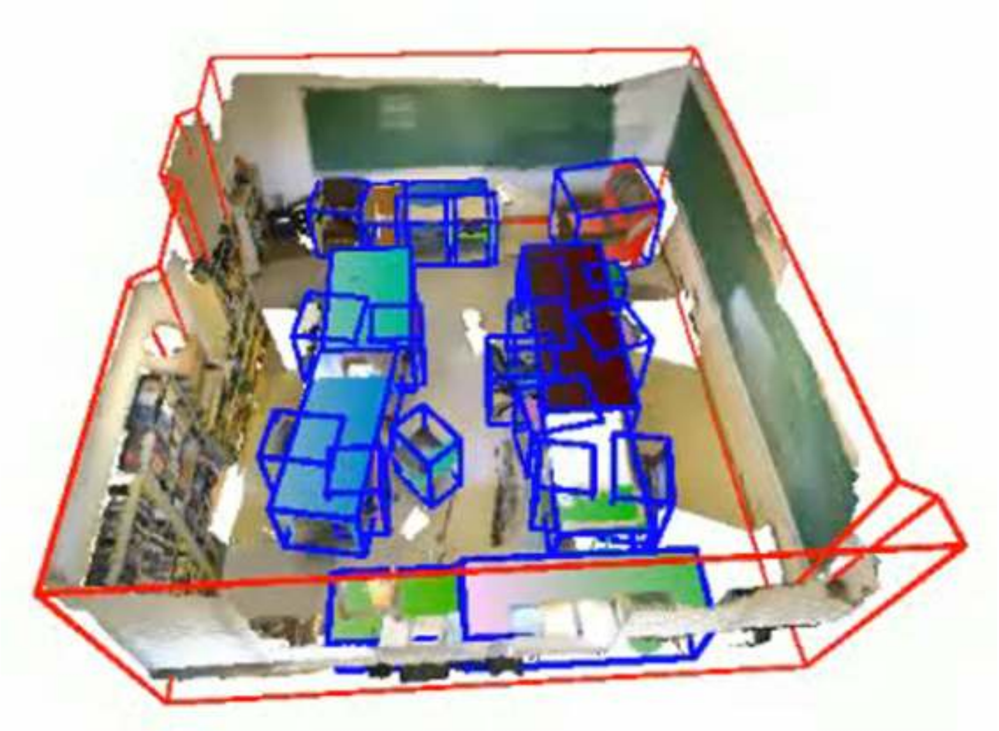
École des Ponts ParisTech

Keypoint Transformer



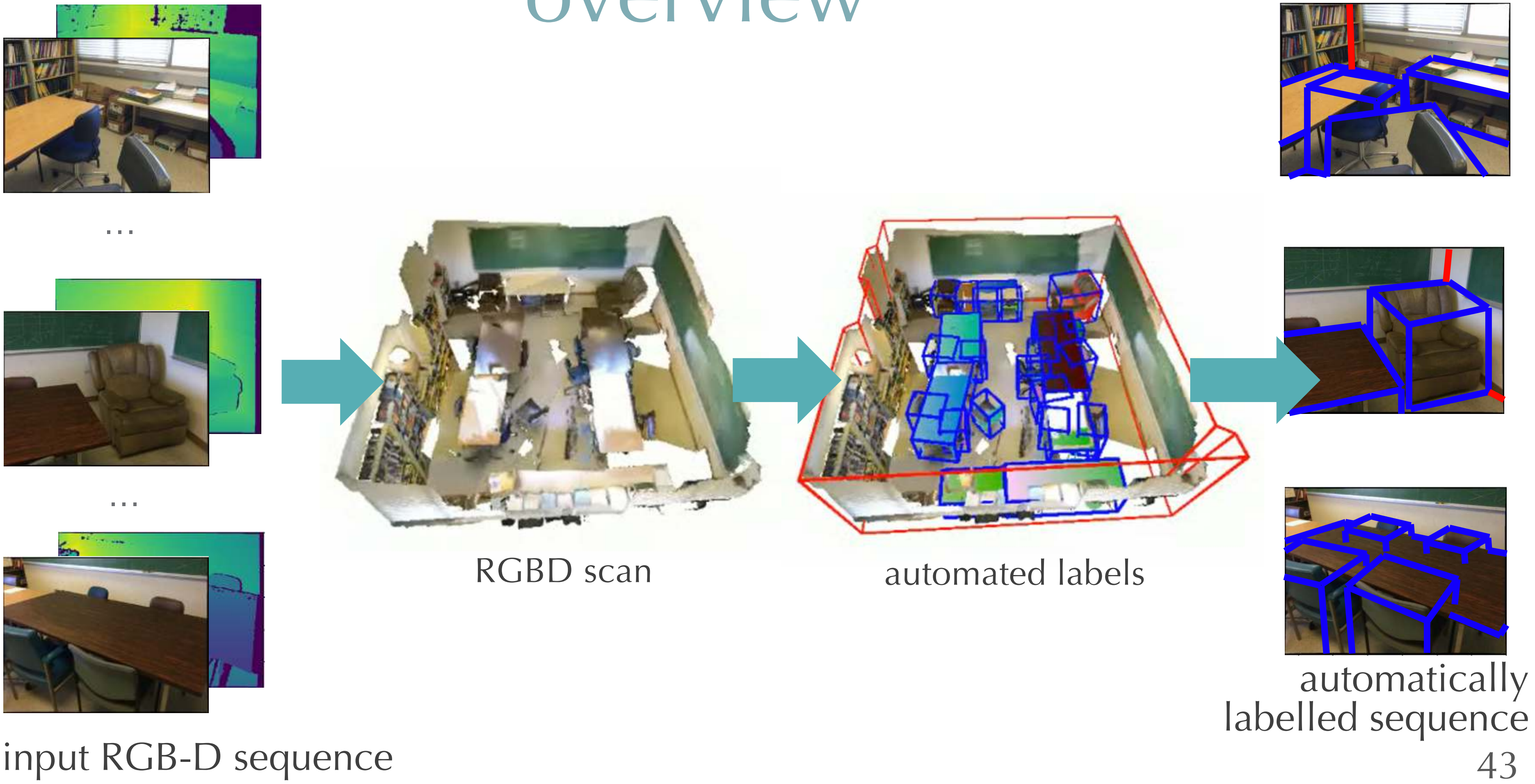
Keypoint Transformer: Solving Joint Identification in Challenging Hands and Object Interactions for Accurate 3D Pose Estimation. Shreyas Hampali, Sayan Deb Sarkar, Mahdi Rad, Vincent Lepetit. Oral at CVPR 2022

Indoor Scenes



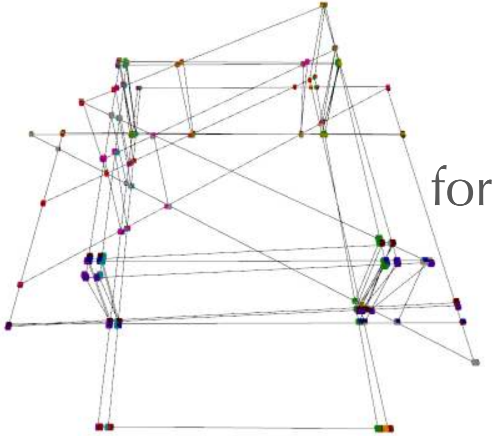
Monte Carlo Scene Search for 3D Scene Understanding. Shreyas Hampali, Sinisa Stekovic, Sayan Deb Sarkar, Chetan Srinivasa Kumar, Friedrich Fraundorfer, and Vincent Lepetit. CVPR 2021. (The two first authors have equal contributions)

overview

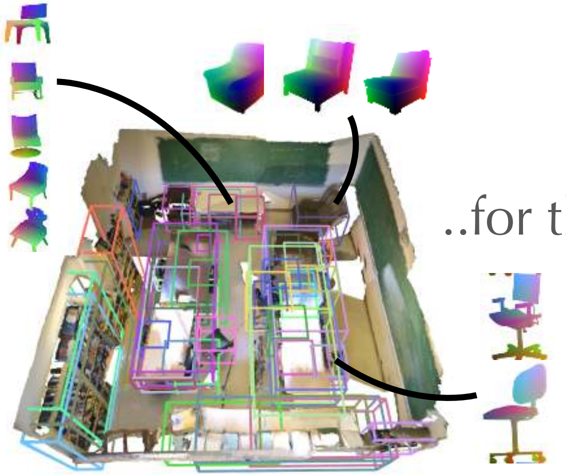

...
...
input RGB-D sequence
RGBD scan
automated labels
automatically labelled sequence

General Idea

Step #1: Make proposals



for the **walls** and..

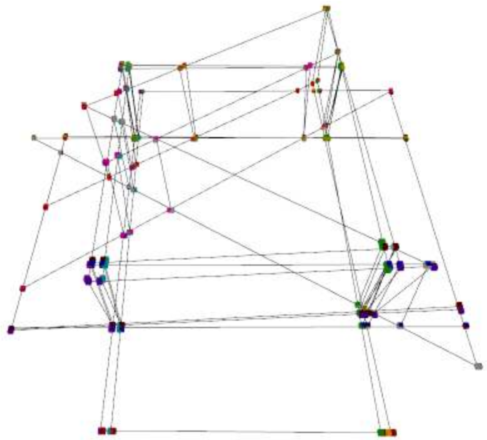


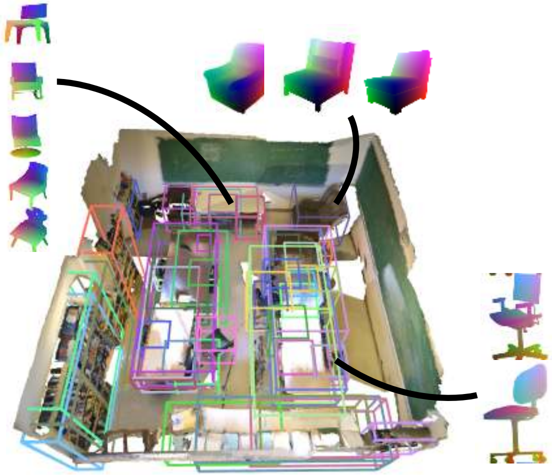
..for the **objects**

École des Ponts
ParisTech

General Idea

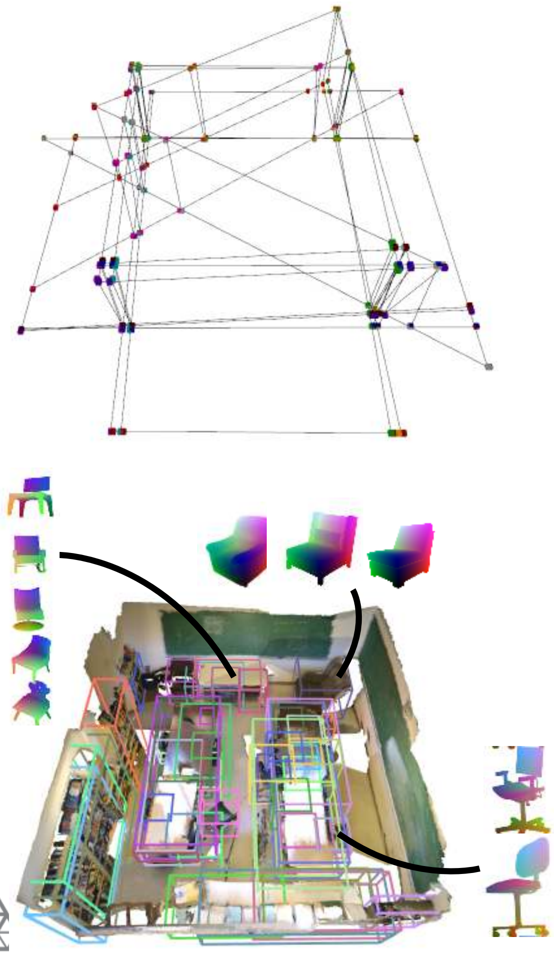
Step #1: Make proposals





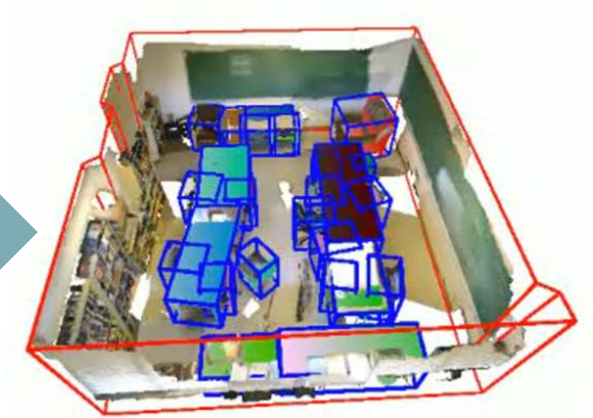
this step does not have to be perfect,
the next step will filter the false positives!

General Idea



Step #2: Select the correct proposals

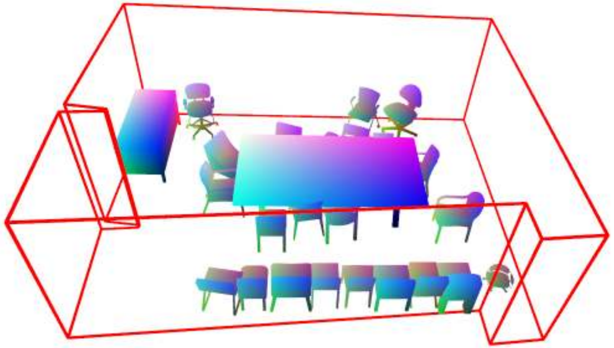
How can we select the correct proposals?

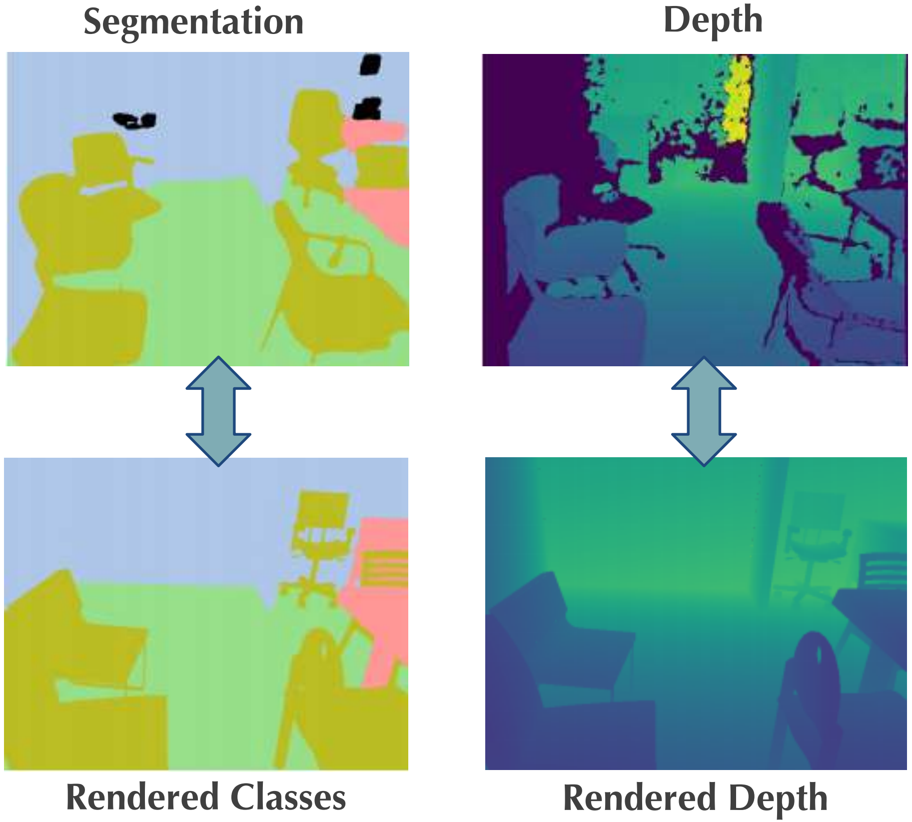


automated labels

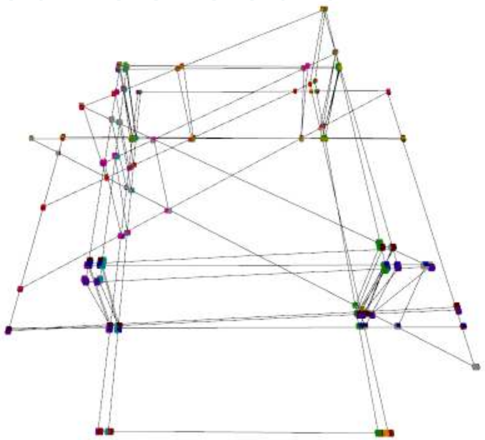
objective function: same as for the hand+object problem

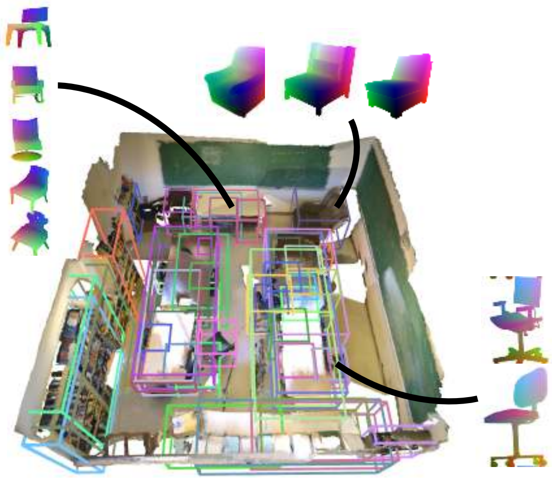
How good is this possible solution?





proposal selection





If we have 100 proposals, an exhaustive search would require 2^{100} (~10^{12}) evaluations!

The objective function is not differentiable.

It has no special form we can exploit for efficient optimization.

Let's try using a tree search algorithm...

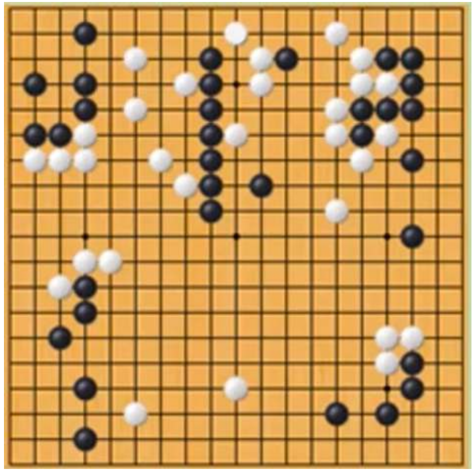
École des Ponts ParisTech

Monte Carlo Tree Search

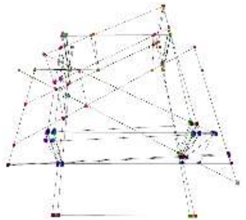
- Originates from the work of Bruce Abramson in the 1980′s;
- Name ‘MCTS′ coined by Rémi Coulom in 2006;

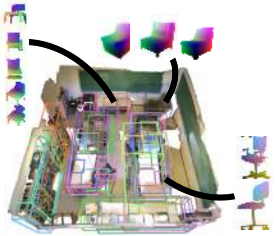
- Combined with Deep Learning by DeepMind in 2016 to play Go.

- Deals well with high-complexity games.
- No heuristics, exploration based on the objective.

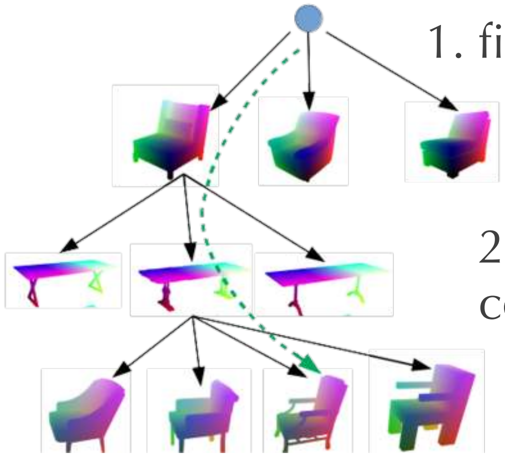


How to turn auto-labeling into a single-player game





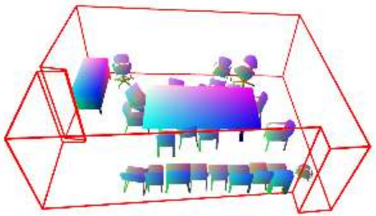
set of proposals



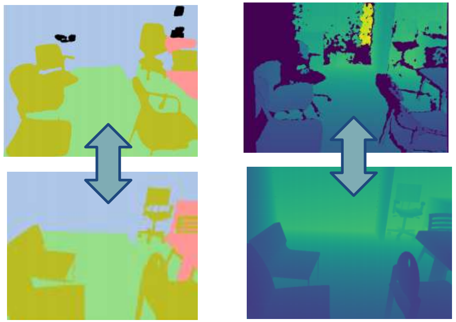
1. first move: pick one proposal

2. second move: pick another proposal compatible with the first one

3. and so on (we can skip some proposals)



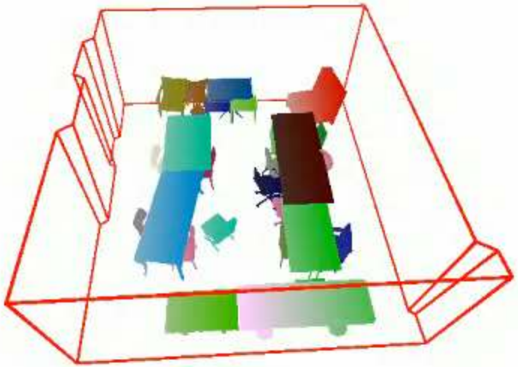
4. we measure how good the 'end-game' is with our objective function



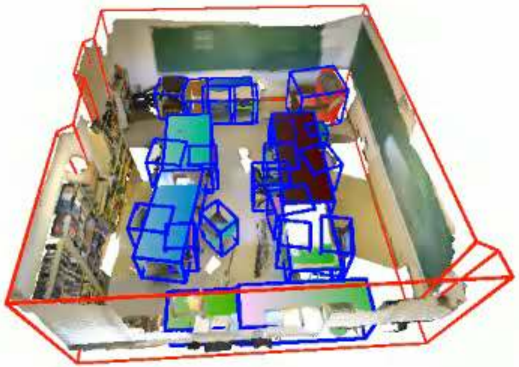
automated indoor annotations

RGBD scan

automated labels



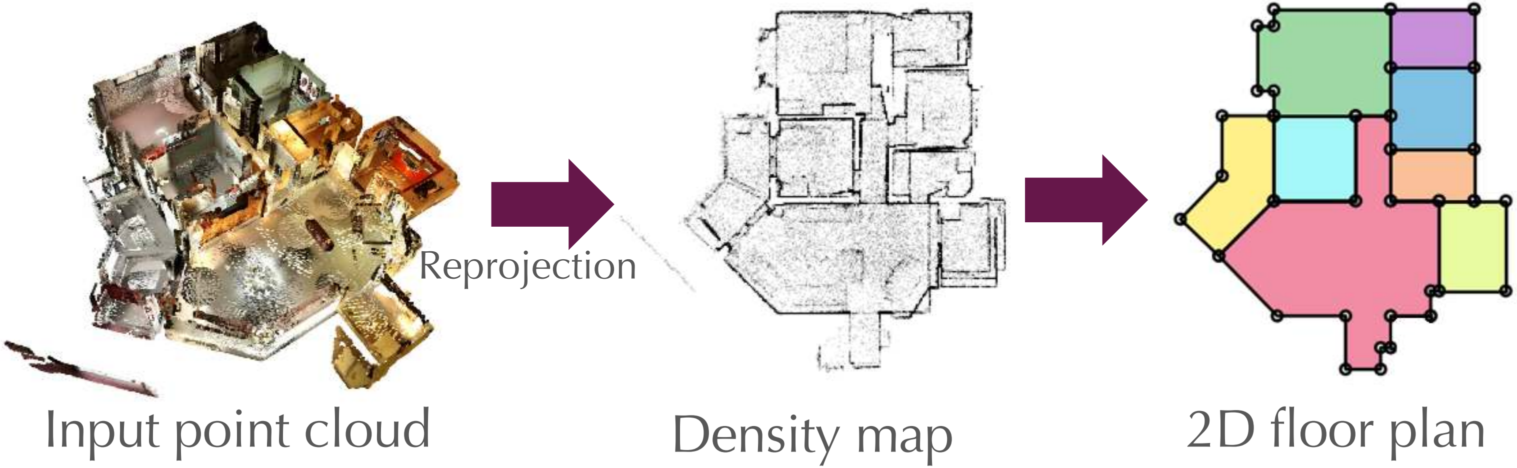
automated labels
with scan



Monte Carlo Scene Search for 3D Scene Understanding. Shreyas Hampali, Sinisa Stekovic, Sayan Deb Sarkar, Chetan Srinivasa Kumar, Friedrich Fraundorfer, and Vincent Lepetit. CVPR 2021. (The two first authors have equal contributions)

extension

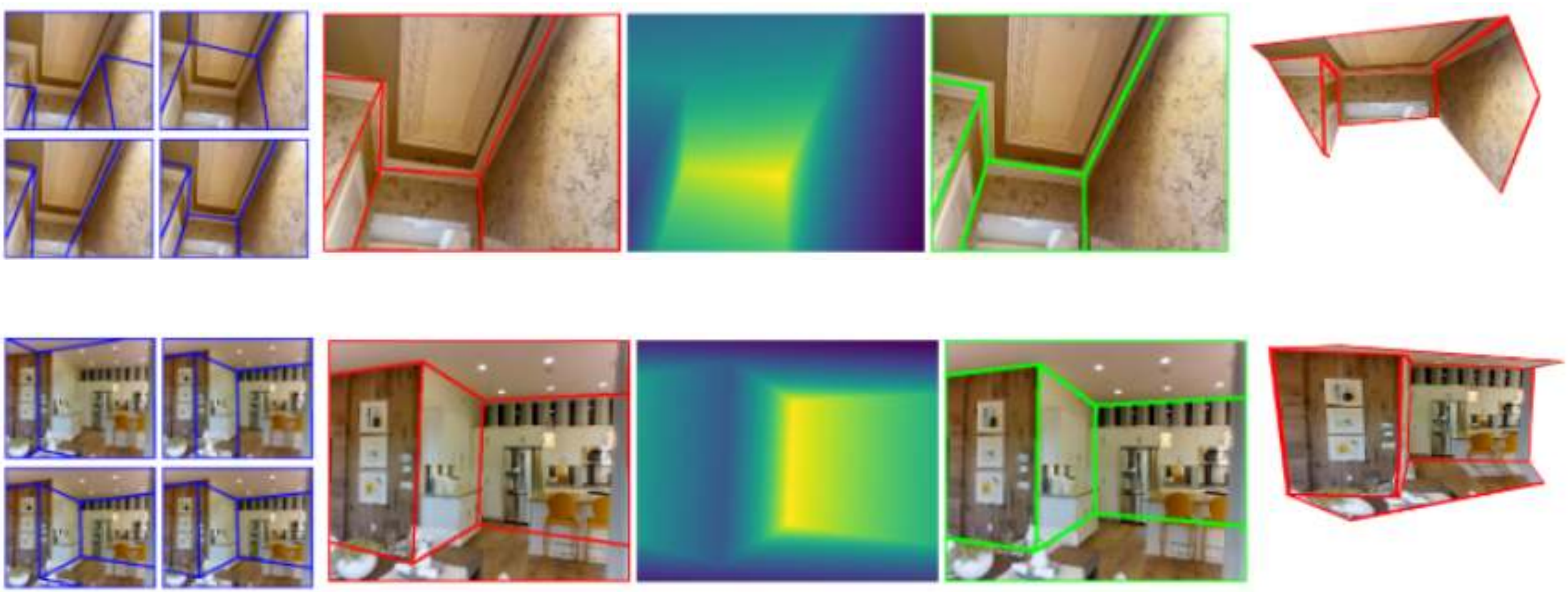
- learned objective function
- discrete search combined with continuous optimization



MonteFloor: Extending MCTS for Reconstructing Accurate Large-Scale Floor Plans. Sinisa Stekovic, Mahdi Rad, Friedrich Fraundorfer, and Vincent Lepetit. Oral at ICCV 2021.

École des Ponts
ParisTech

another problem, same solution



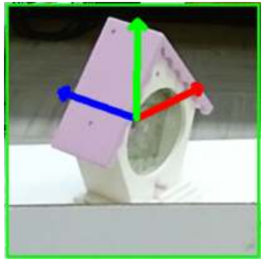
MCTS with Refinement for Proposals Selection Games in Scene Understanding. Sinisa Stekovic, Mahdi Rad, Alireza Moradi, Friedrich Fraundorfer, and Vincent Lepetit. IEEE TPAMI 2022.

École des Ponts ParisTech

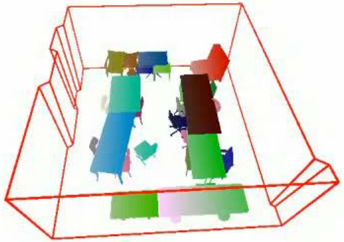
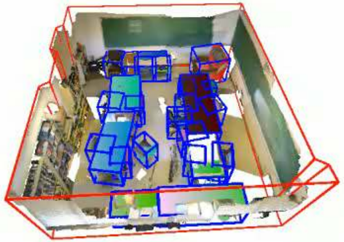
summary

- dealing with new objects without training time:

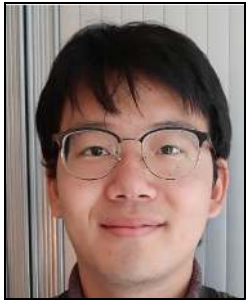
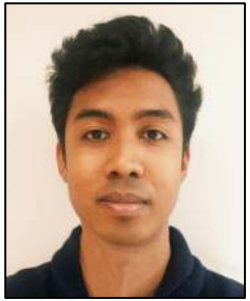




- MCTS for auto-labelling:



Sinisa Stekovic

Shreyas Hampali

Van Nguyen Nguyen

Yuming Du

Michaël Ramamonjisoa

Madhi Rad

Friedrich Fraundorfer

QUALCOMM®

chist-era

Thanks for listening!

Questions?