Séminaire de Mathématiques Appliquées du CERMICS

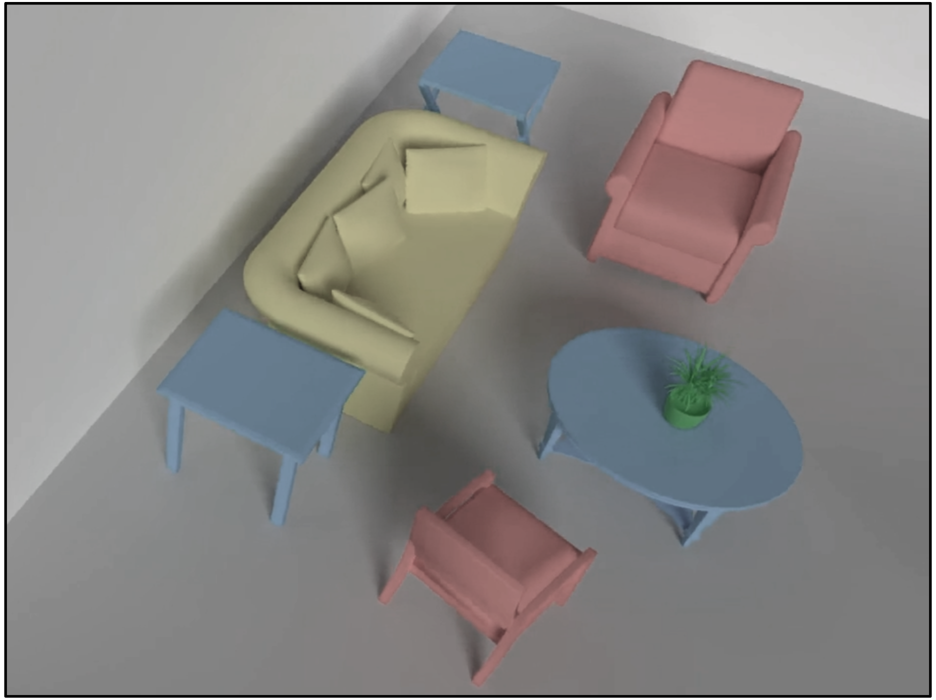
3D Scene Understanding from Images

Vincent Lepetit (École des Ponts ParisTech)

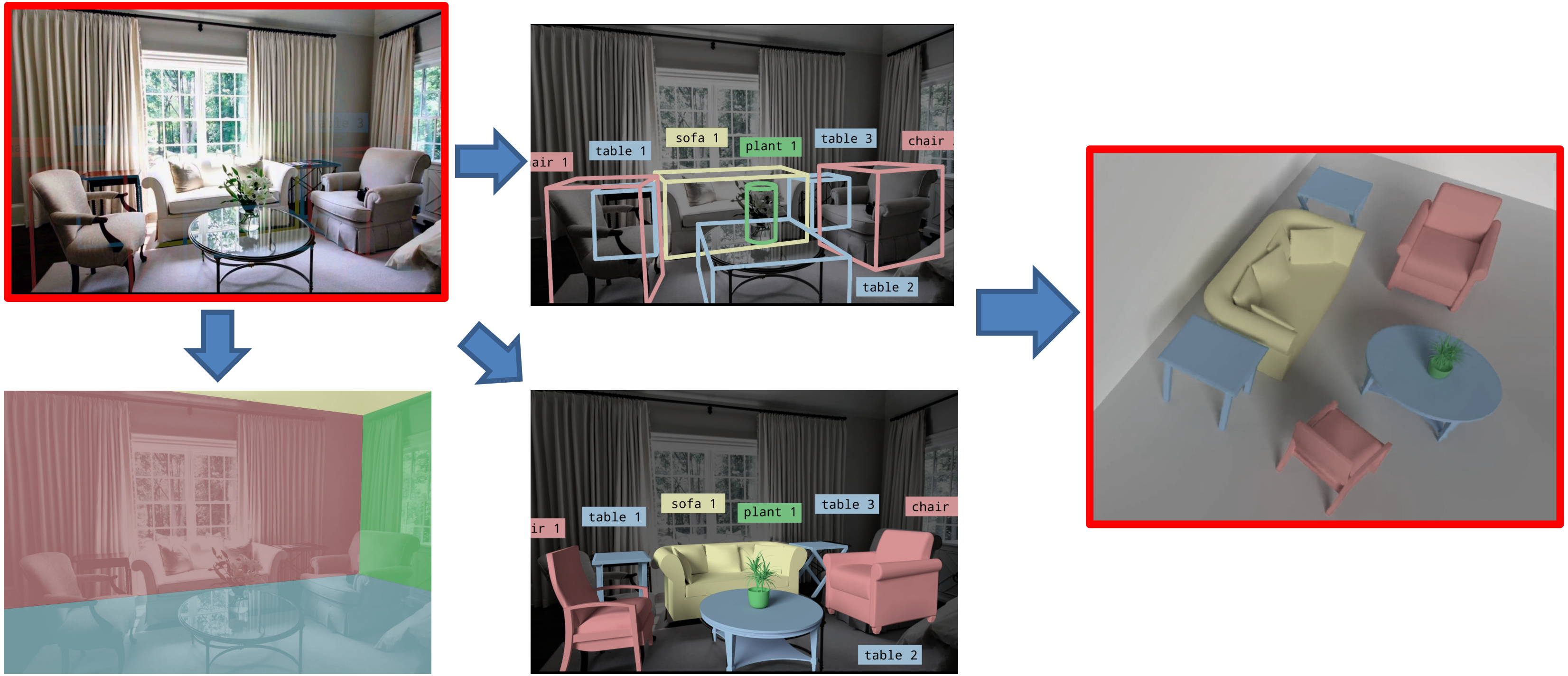
3 octobre 2019

3D Scene Understanding from Images

Vincent Lepetit
Imagine – LIGM – Ecole des Ponts ENPC



3D Scene Understanding from Images

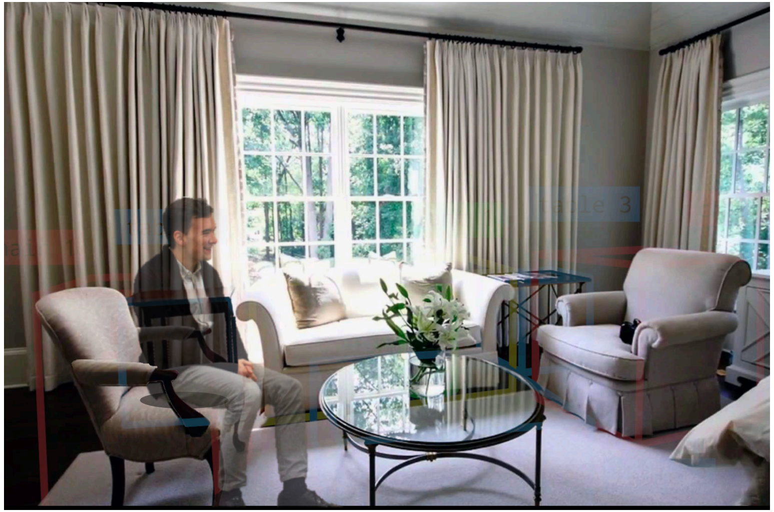


Old, fundamental problem [Roberts 65, Yakimovsky 73, Ohta 78]

Why It Is Useful

Possible applications:

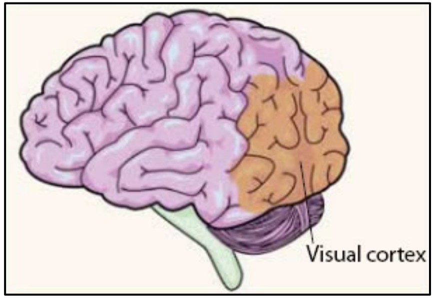
Robotics – Interaction with the environment



Augmented Reality – Realistic Augmentation

Why It Is Difficult

3D scene understanding, a **long-standing problem in computer vision, for many reasons**



Human vision is a (mostly) unconscious process that involves 20% of the brain

different chairs have different shapes, materials, ..

illumination effects

3D to 2D projection

occlusions

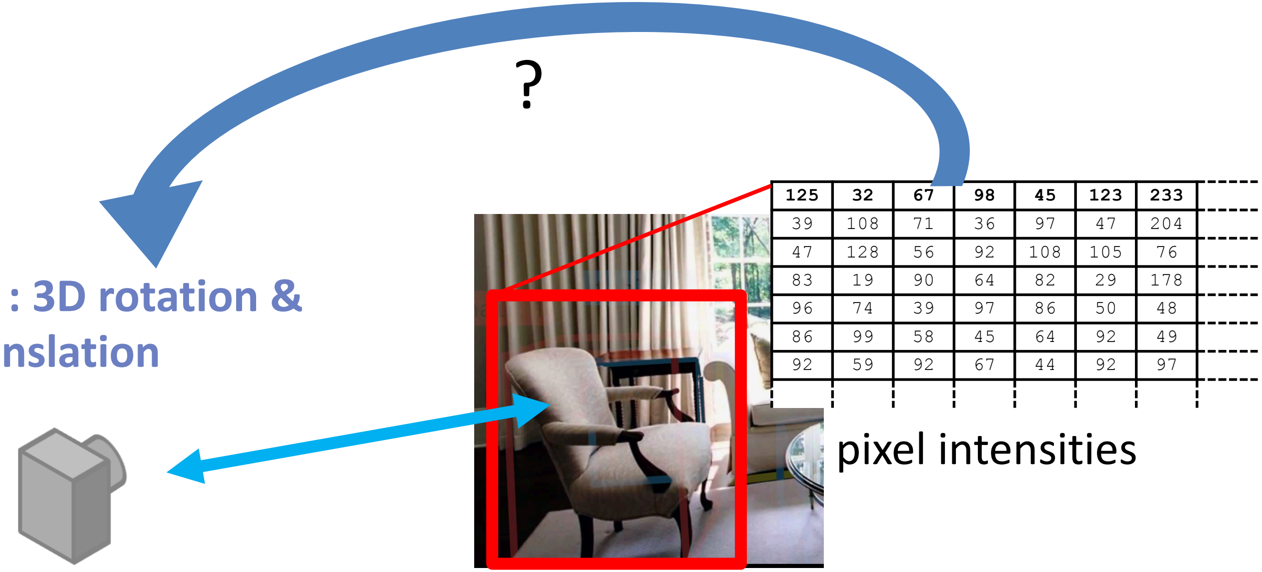
reflections

transparency

Why It Is Difficult, Example of Pose

Object pose estimation = estimating the 3D motion (= **a 3D rotation + a 3D translation**) between the object and the camera.

The function from the pixel intensities to the rotation and translation is extremely complex.



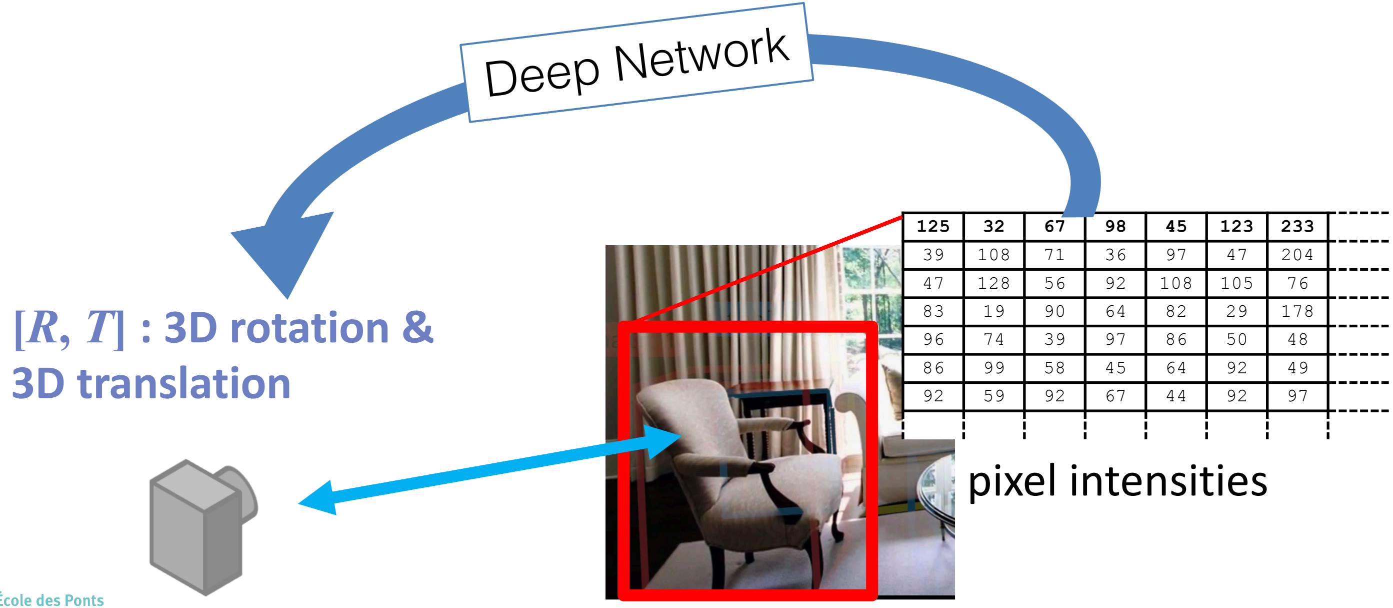
125	**32**	**67**	**98**	**45**	**123**	**233**
39	108	71	36	97	47	204
47	128	56	92	108	105	76
83	19	90	64	82	29	178
96	74	39	97	86	50	48
86	99	58	45	64	92	49
92	59	92	67	44	92	97

École des Ponts ParisTech

Why It Is Difficult, Example of Pose

Object pose estimation = estimating the 3D motion (= **a 3D rotation + a 3D translation**) between the object and the camera.

The function from the pixel intensities to the rotation and translation is extremely complex.



125	**32**	**67**	**98**	**45**	**123**	**233**
39	108	71	36	97	47	204
47	128	56	92	108	105	76
83	19	90	64	82	29	178
96	74	39	97	86	50	48
86	99	58	45	64	92	49
92	59	92	67	44	92	97

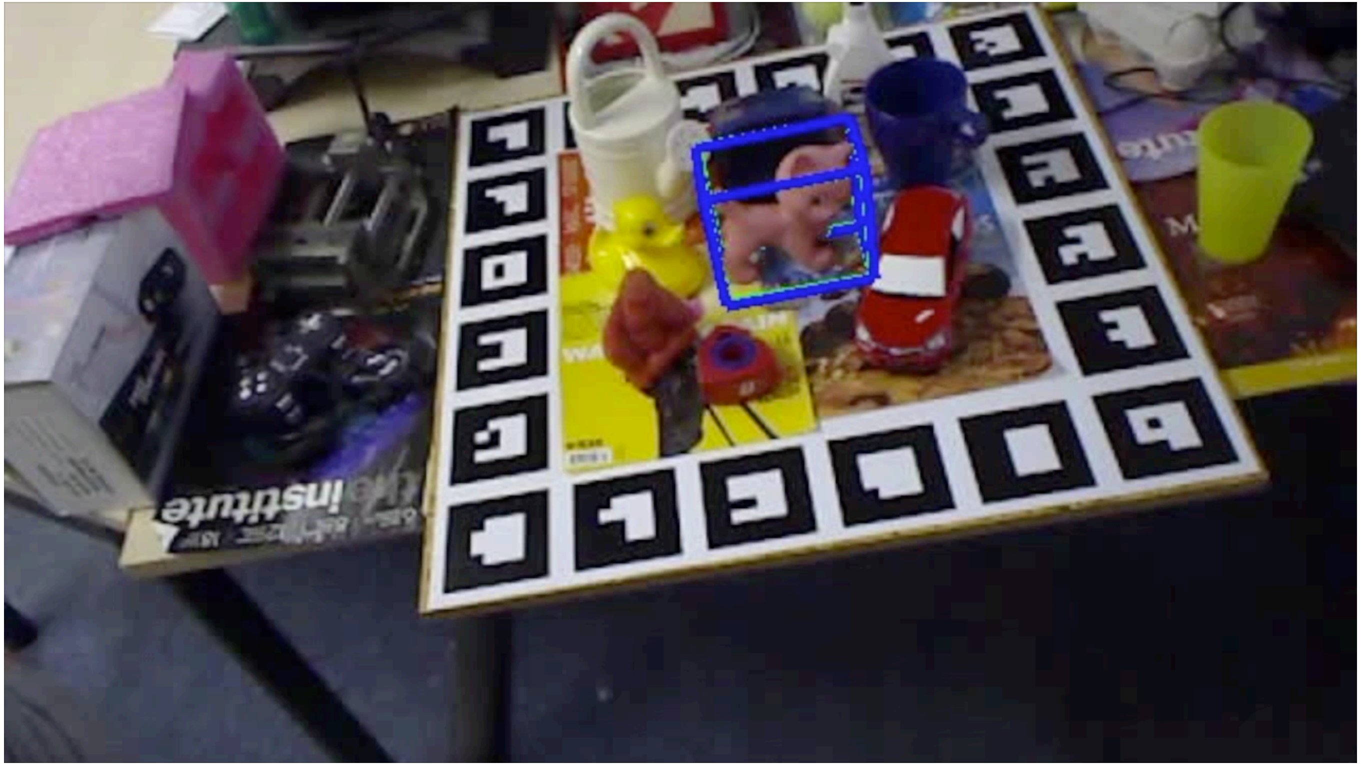
Flexibility of Deep Learning

- We can use a Deep Network to approximate any continuous function;

$$\mathbf{x} \longrightarrow \boxed{\text{Deep Network } f(\mathbf{x};\, \Theta)} \longrightarrow \mathbf{o}$$

- We can use any loss function as long as it is differentiable to find parameters Θ;

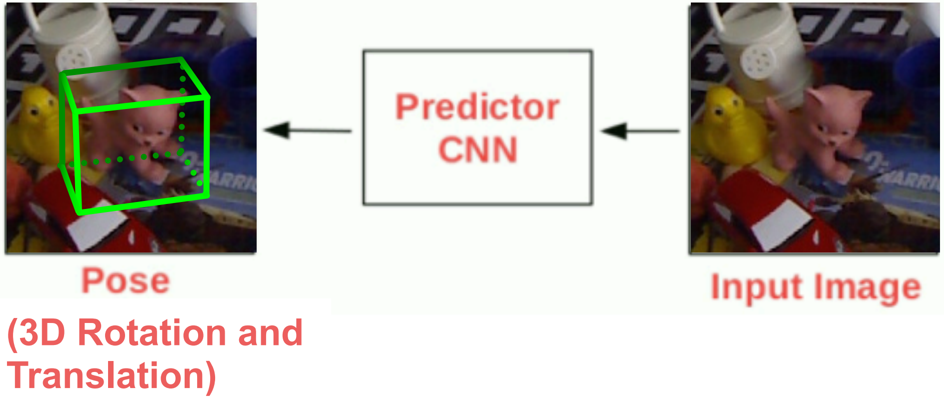
3D Pose Estimation of Rigid Objects



BB8: A Scalable, Accurate, Robust to Partial Occlusion Method for Predicting the 3D Poses of Challenging Objects without Using Depth. Mahdi Rad and Vincent Lepetit. ICCV 2017.

3D Pose Estimation of Rigid Objects

Training set: images with known rotations and translations
(about 200 images in practice)



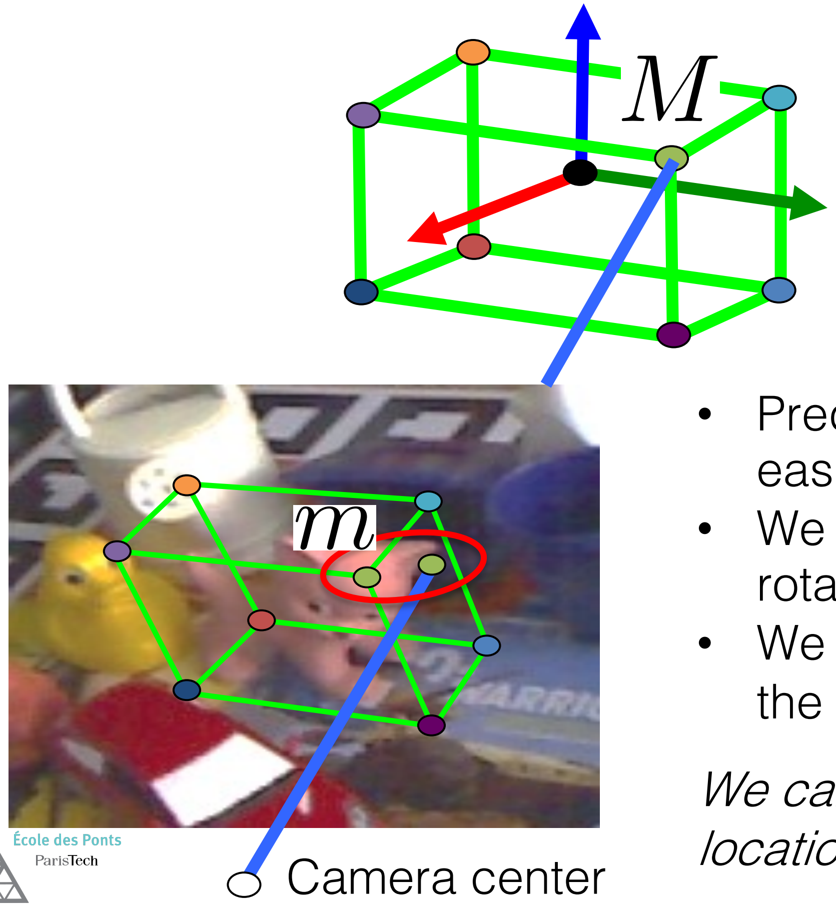
Possible Loss Function

Training set

$(I_i, (R_i, T_i))$

$$\min_{\Theta} \sum_i \mathrm{dist}_R(R_i, f_R(I_i; \Theta)) + \lambda \mathrm{dist}_T(T_i, f_T(I_i; \Theta))$$

3D Pose Estimation from Correspondences



- Predicting 2D locations from an image is an easier regression task;
- We do not need a representation of the 3D rotation;
- We do not need to balance the rotation and the translation.

We can compute the 3D pose from these 2D locations.

École des Ponts ParisTech

New Loss Function

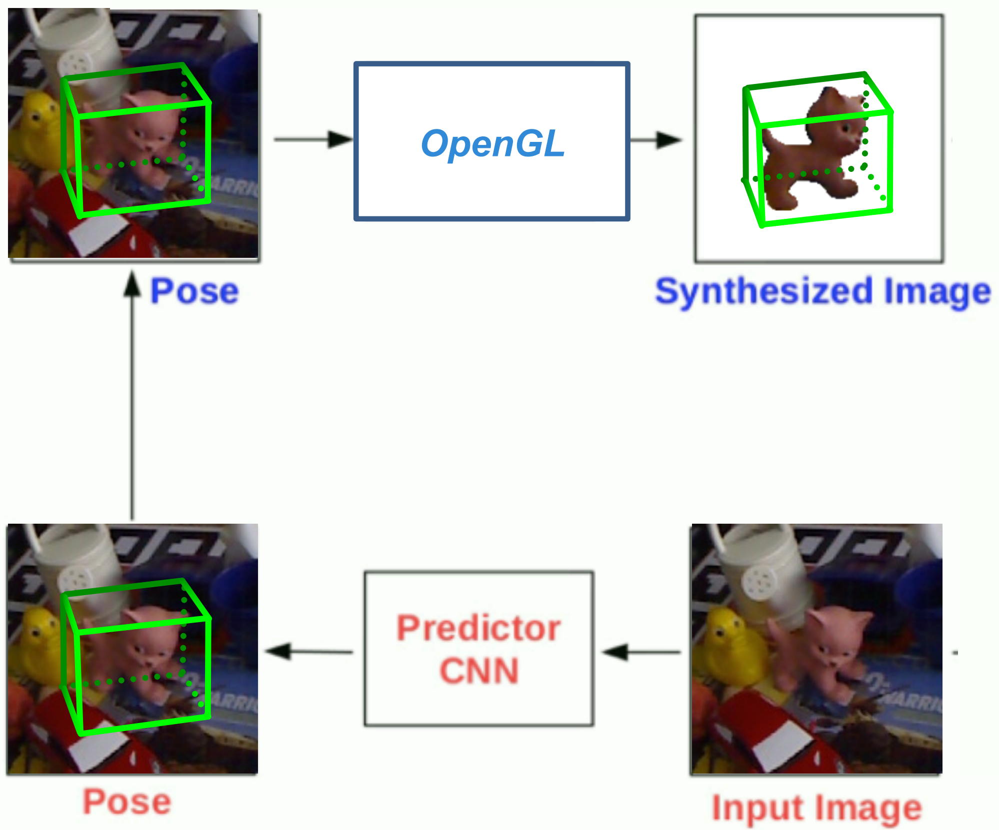
Training set

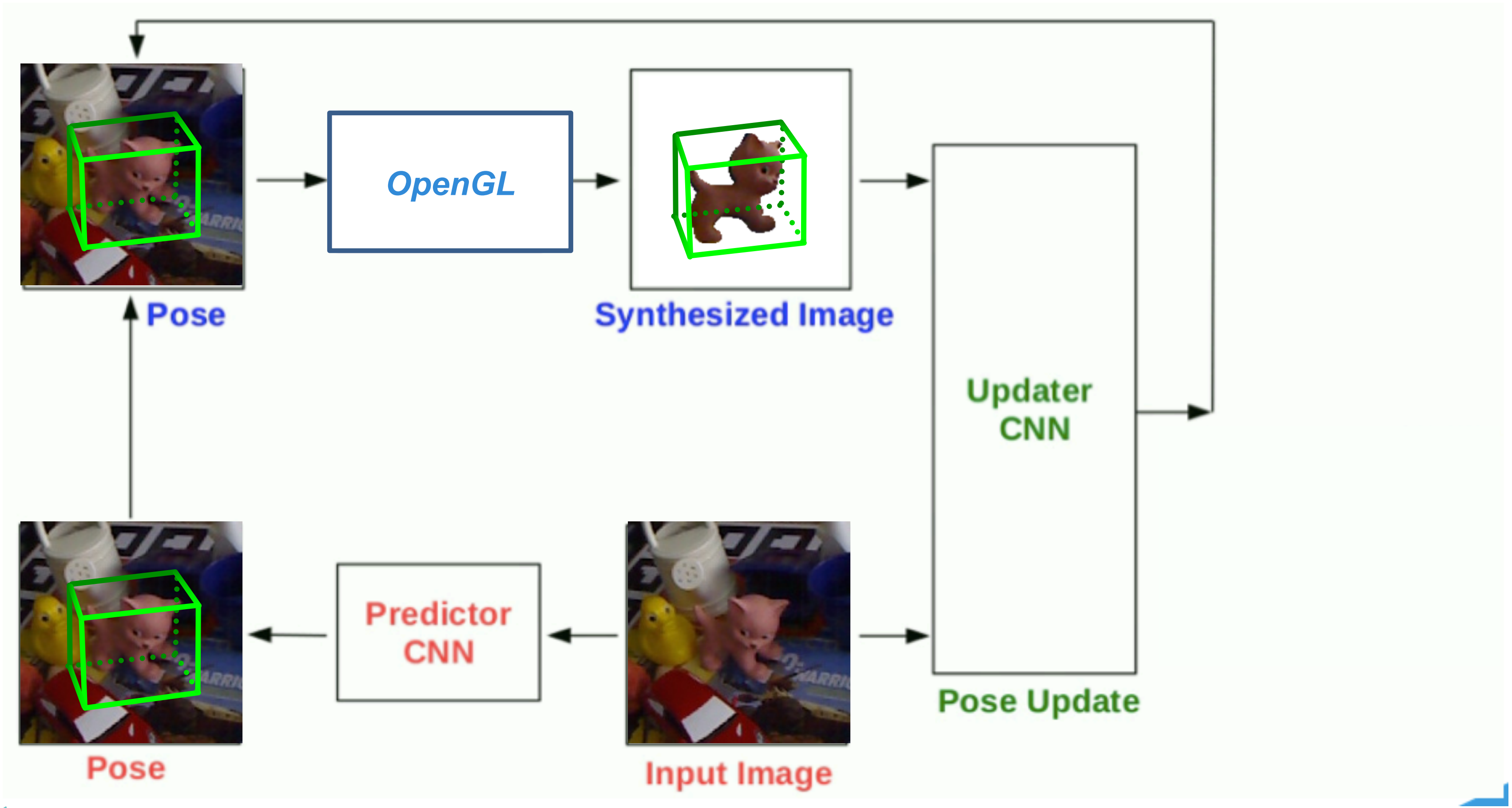
$$\begin{aligned} &(I_i, (R_i, T_i)) \\ \rightarrow \quad &(I_i, m_i = (m_{i1}, .., m_{i8})) \end{aligned}$$

$$\min_{\Theta} \sum_i \mathrm{dist}(m_i, f(I_i; \Theta))$$

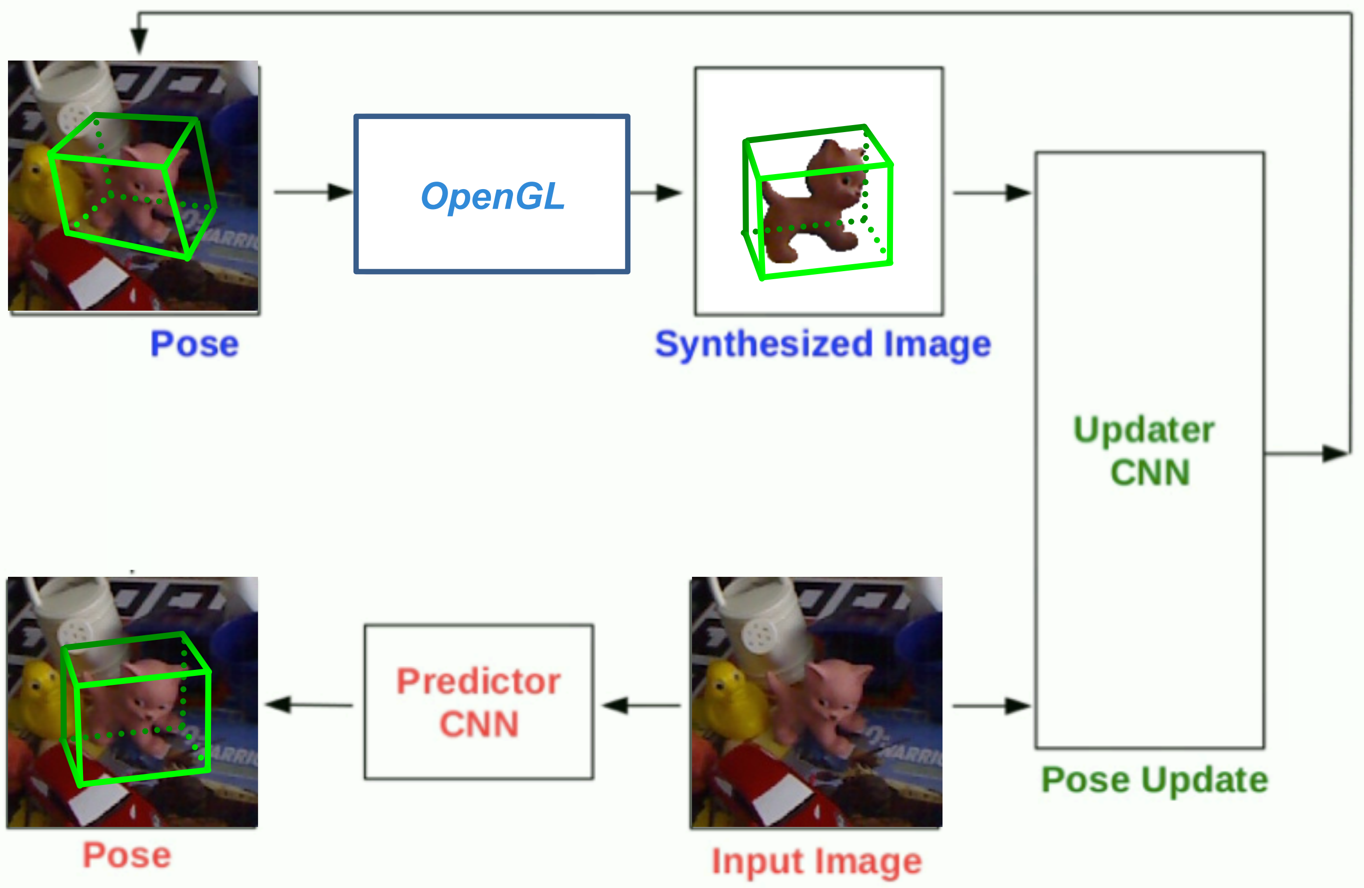
Remark

Can we predict which functions are easy to approximate with Deep Networks?


OpenGL
Pose
Synthesized Image
Predictor
CNN
Pose
Input Image

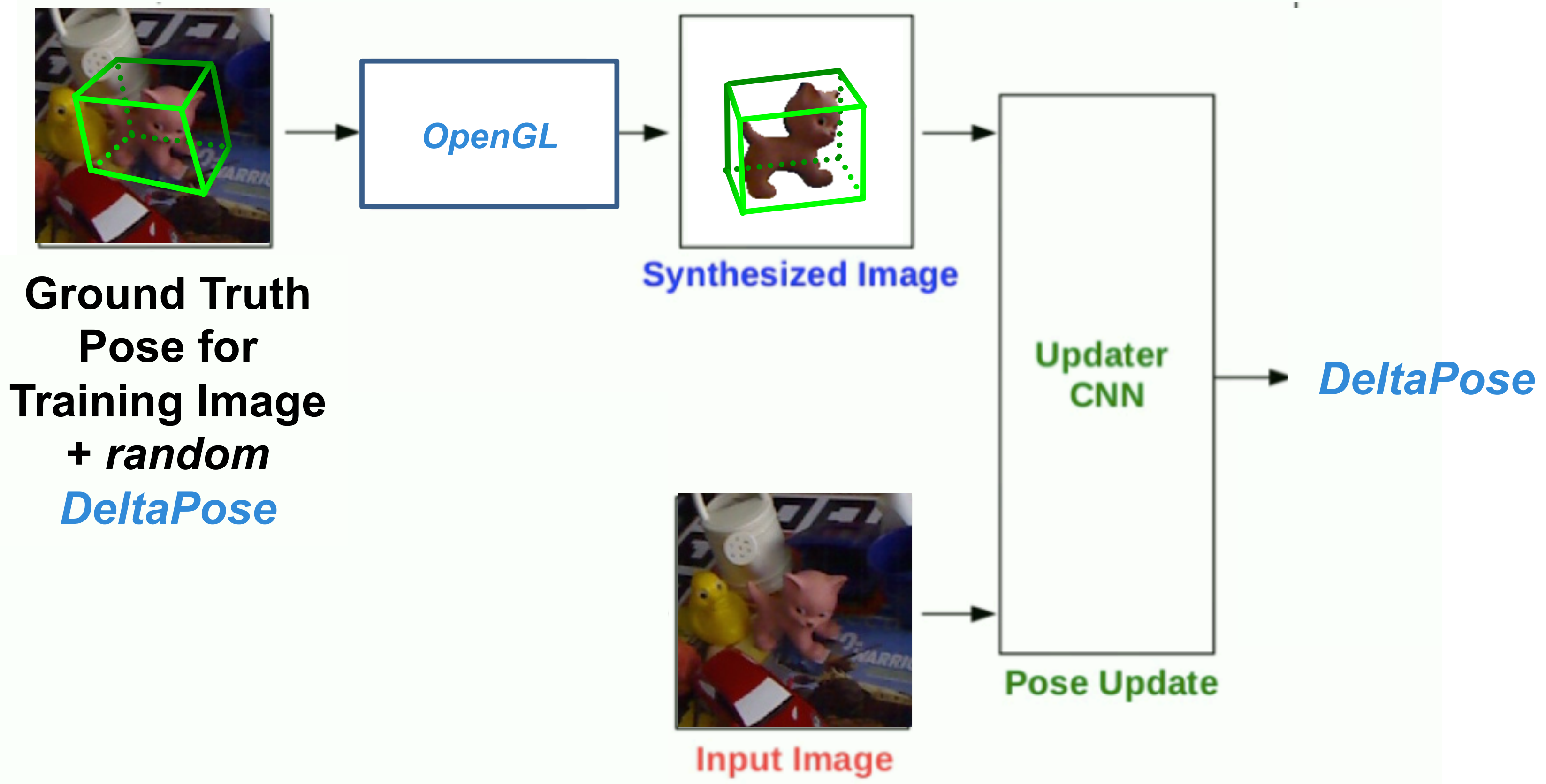


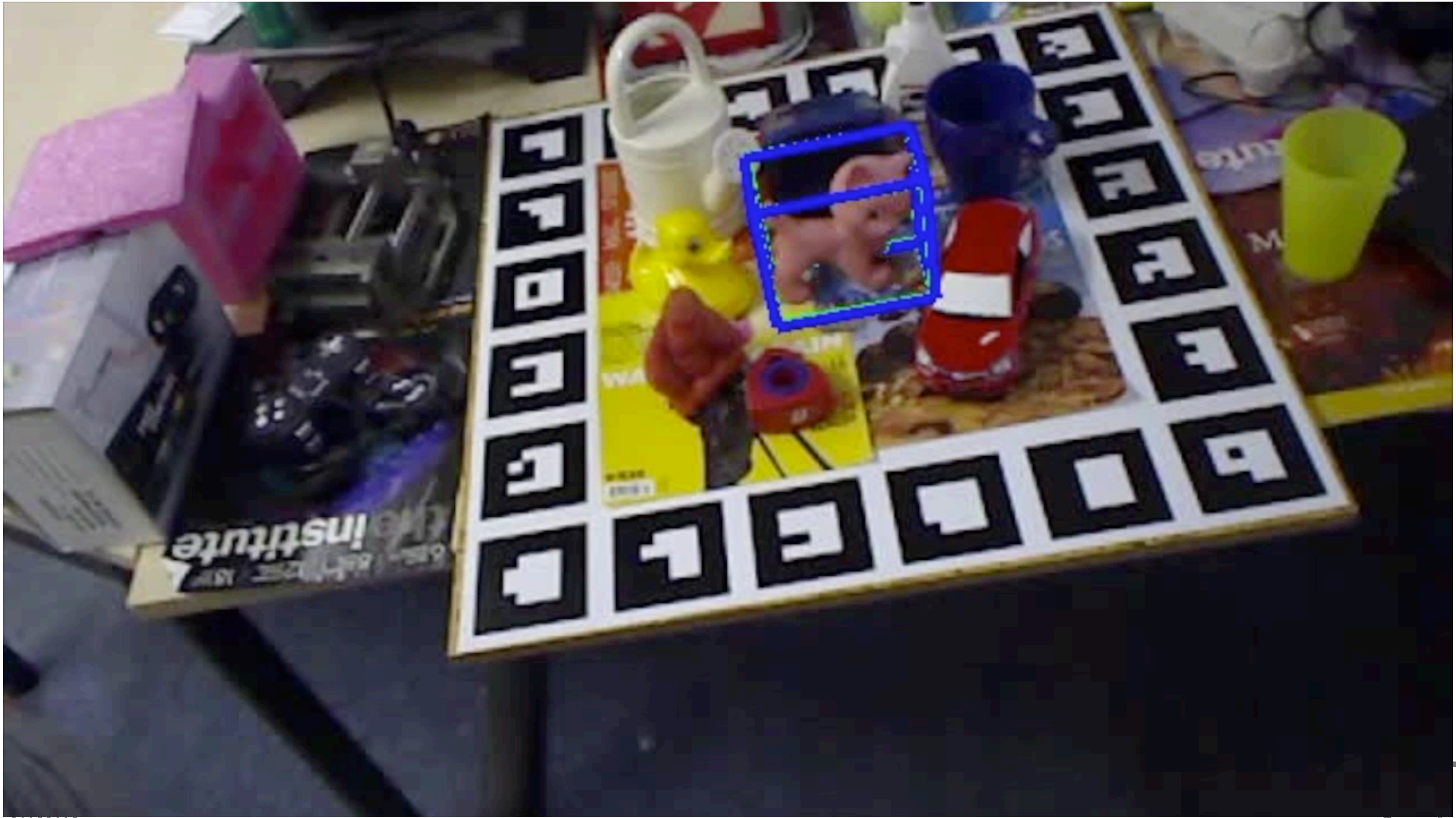
École des Ponts
ParisTech

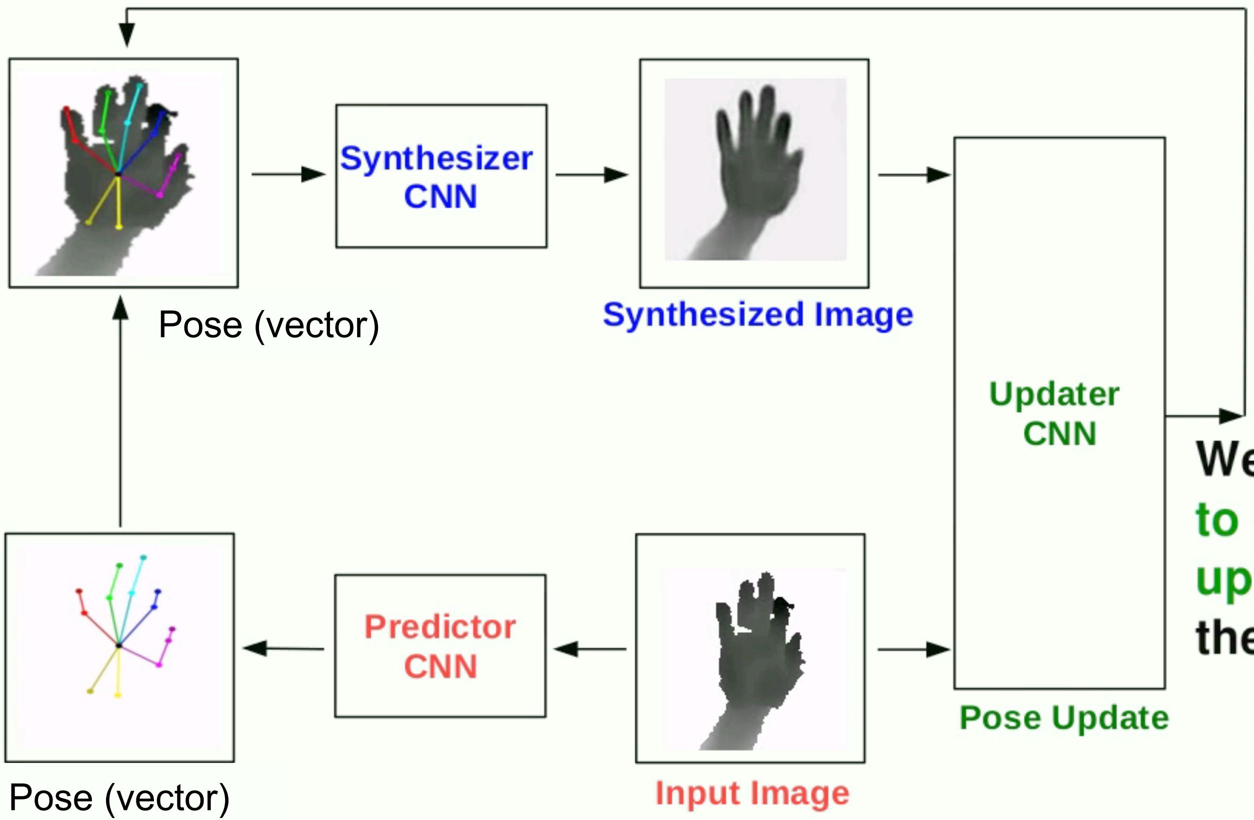


OpenGL
Pose
Synthesized Image
Updater
CNN
Pose Update
Predictor
CNN
Pose
Input Image

Training the Updater as Data Augmentation







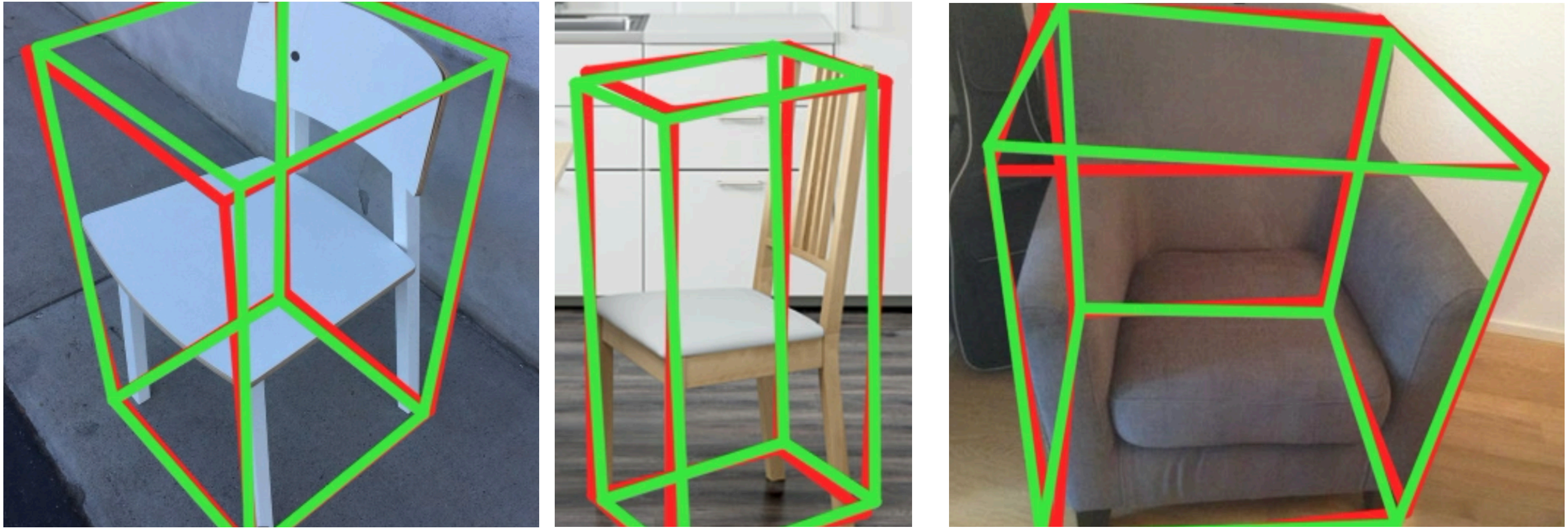
We train a CNN to predict updates for the pose.

3D hand pose estimation
and tracking

Training a Feedback Loop for Hand Pose Estimation. *Markus Oberweger, Paul Wohlhart, and Vincent Lepetit.* ICCV'15. Oral

3D Pose Retrieval for Object Categories

3D Pose Retrieval for Object Categories

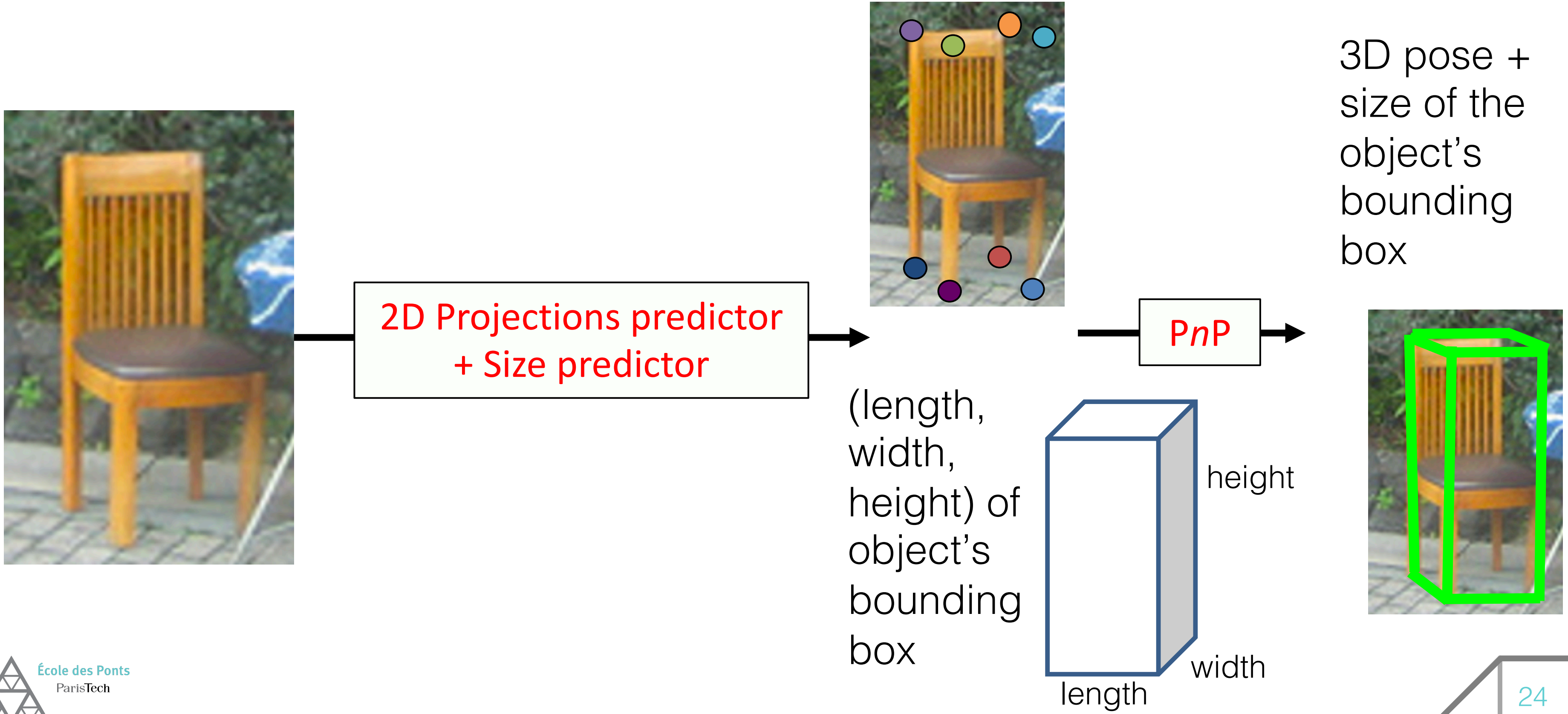


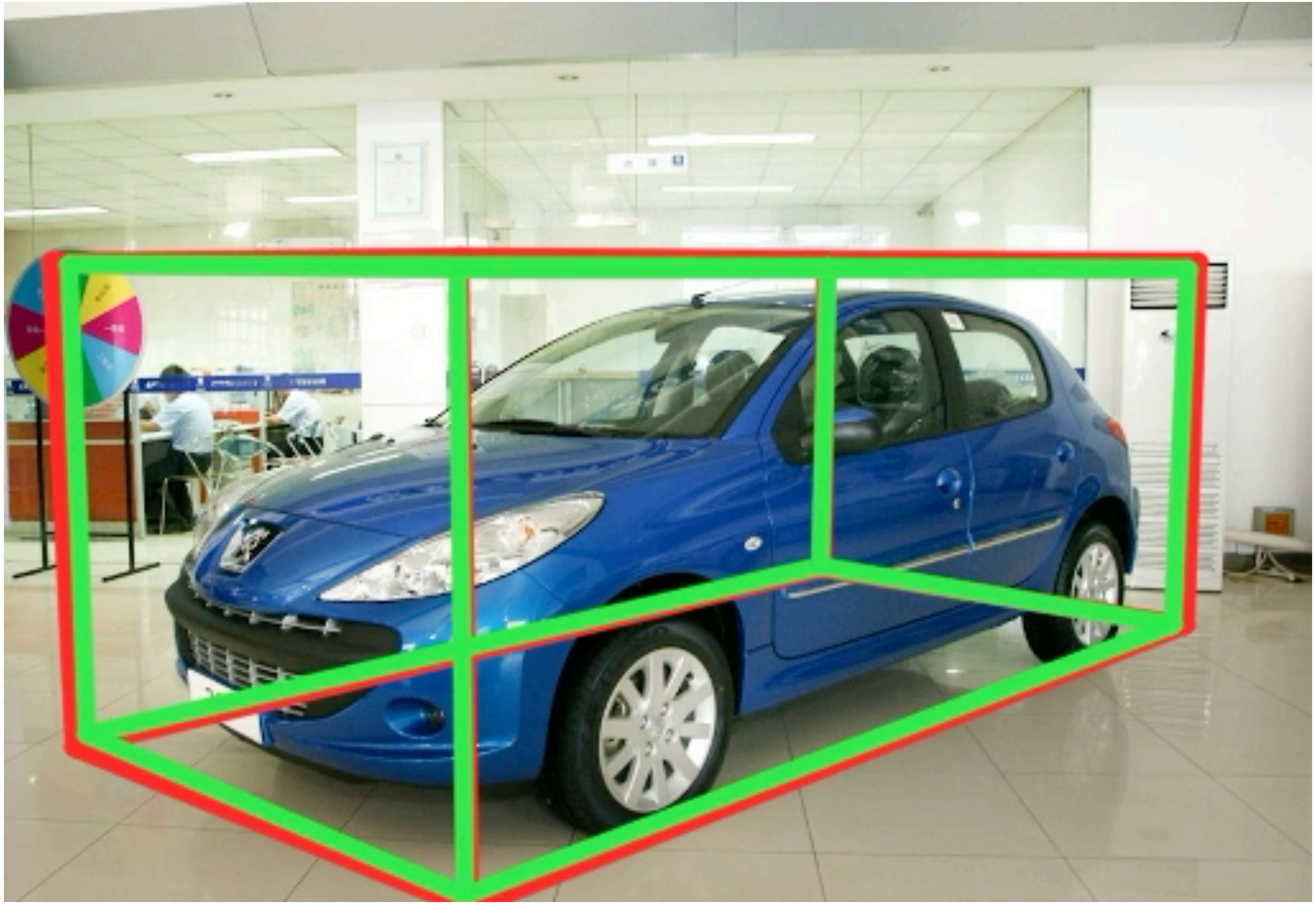
3D Pose Estimation and 3D Model Retrieval for Objects in the Wild. Alexander Grabner, Peter M. Roth, and Vincent Lepetit. CVPR 2018.

3D Pose Retrieval for Object Categories

pose predictor → 3D pose ?

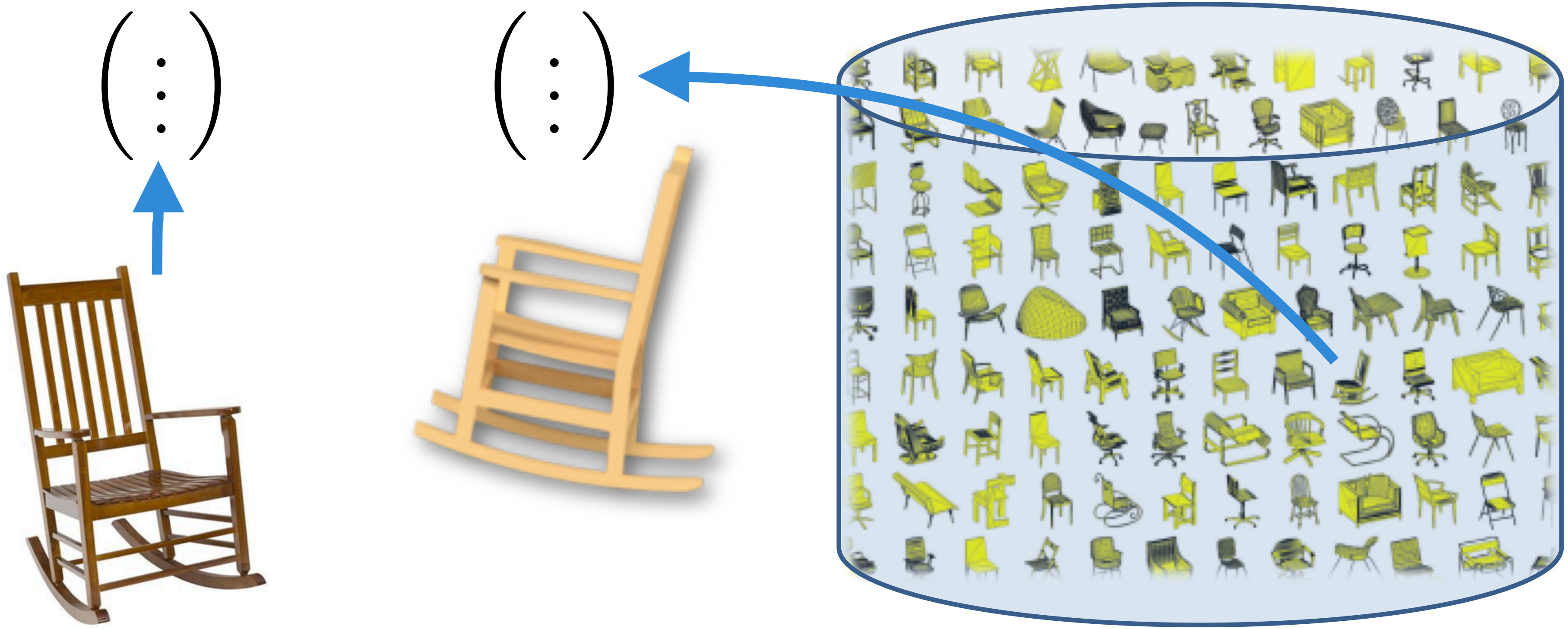
3D Pose Retrieval for Object Categories





3D Model Retrieval for Object Categories

3D Model Retrieval for Object Categories



ShapeNet

Location Fields

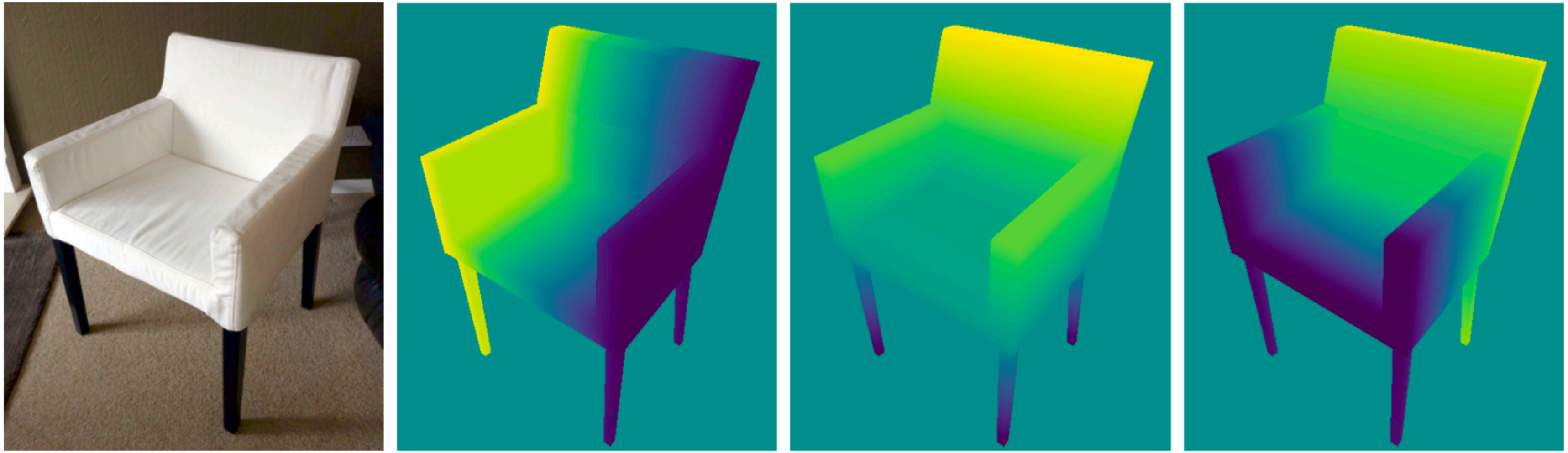


Image LF (X) LF (Y) LF (Z)

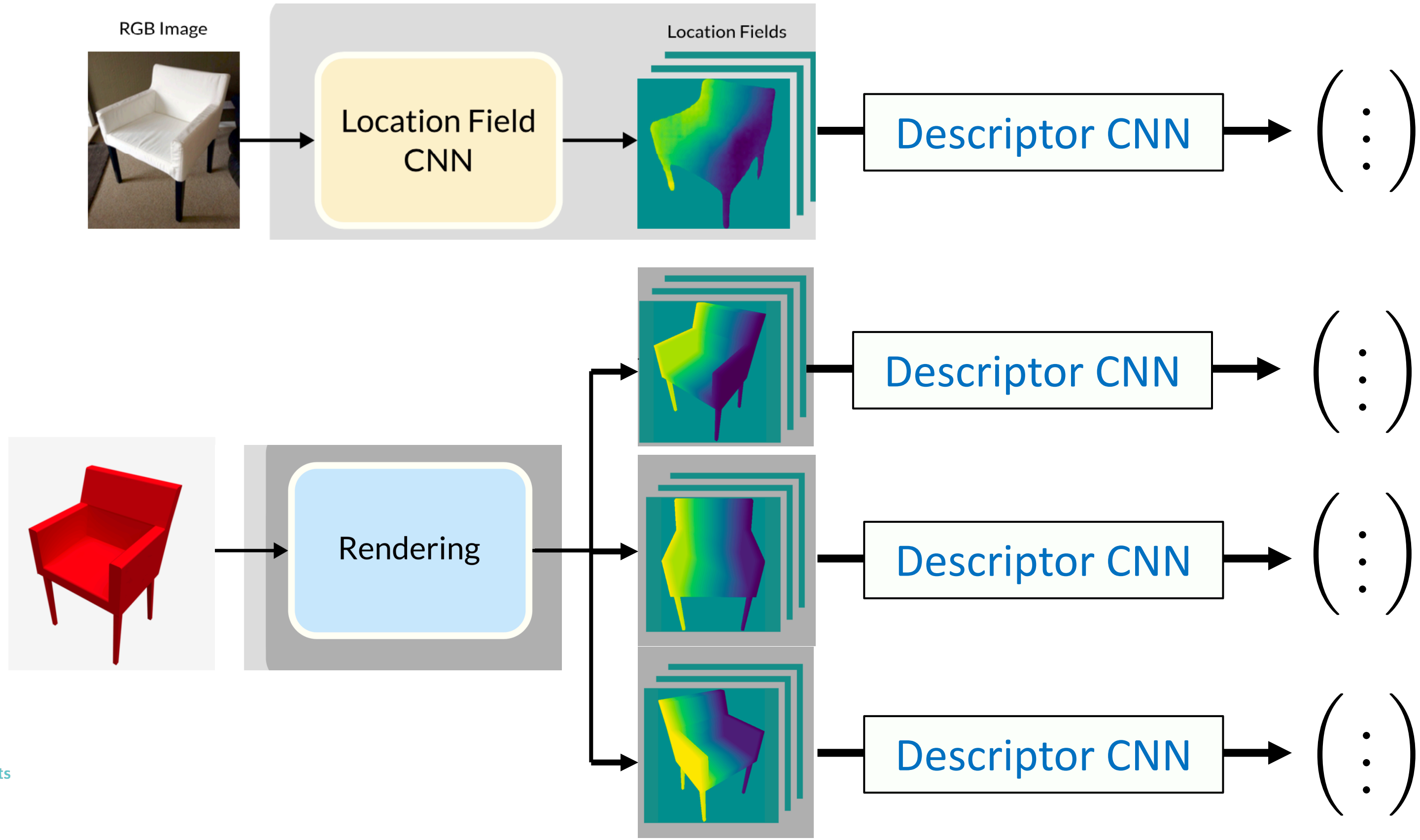
Flexibility of Deep Learning

- We can use a Deep Network to approximate any continuous function;

$$\mathbf{x} \longrightarrow \boxed{\text{Deep Network } f(\mathbf{x};\, \Theta)} \longrightarrow \mathbf{o}$$

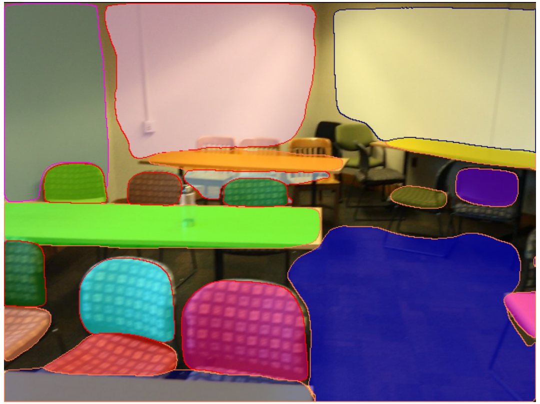
- We can use any loss function as long as it is differentiable to find parameters Θ;

Learning the Descriptors



École des Ponts
ParisTech

Room Layout



Need for Training Data…

Examples of training data from the Pix3D dataset

…

What can we do when there is no annotated training data?

How can we automatically learn new objects?

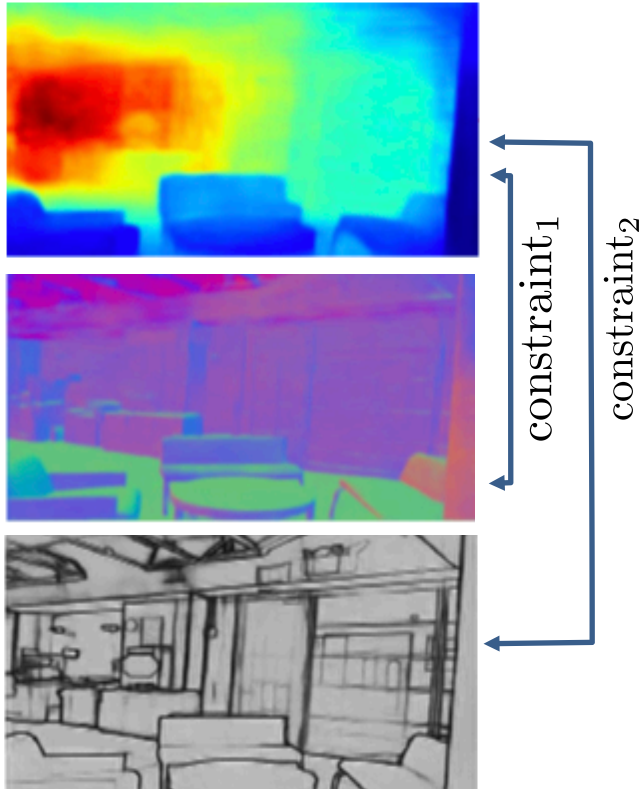
École des Ponts
ParisTech

Learning to Predict Depth

$$\min_{\Theta} \sum_i \mathrm{dist}(D_i, f(I_i; \Theta))$$

Learning to Predict Depth, Normals, Object Contours

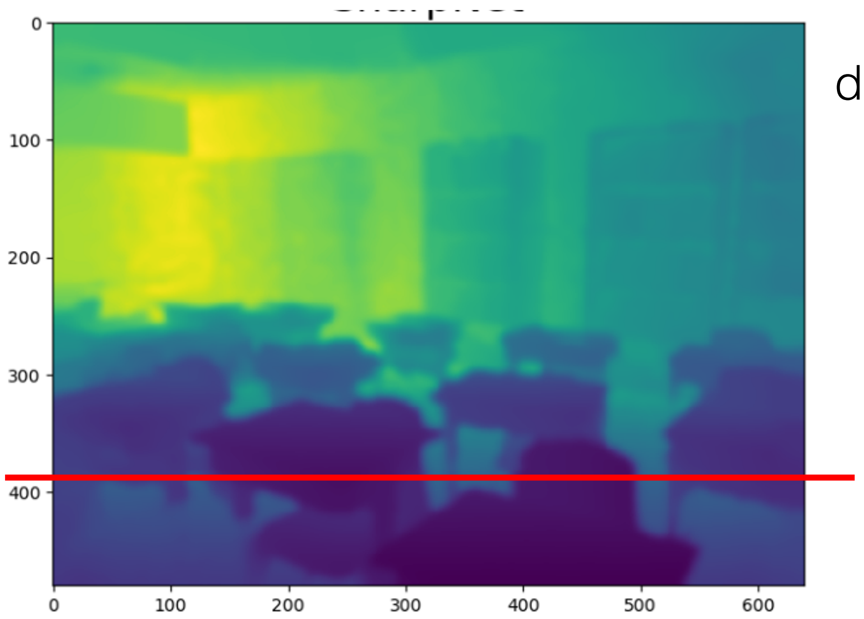
Deep Network

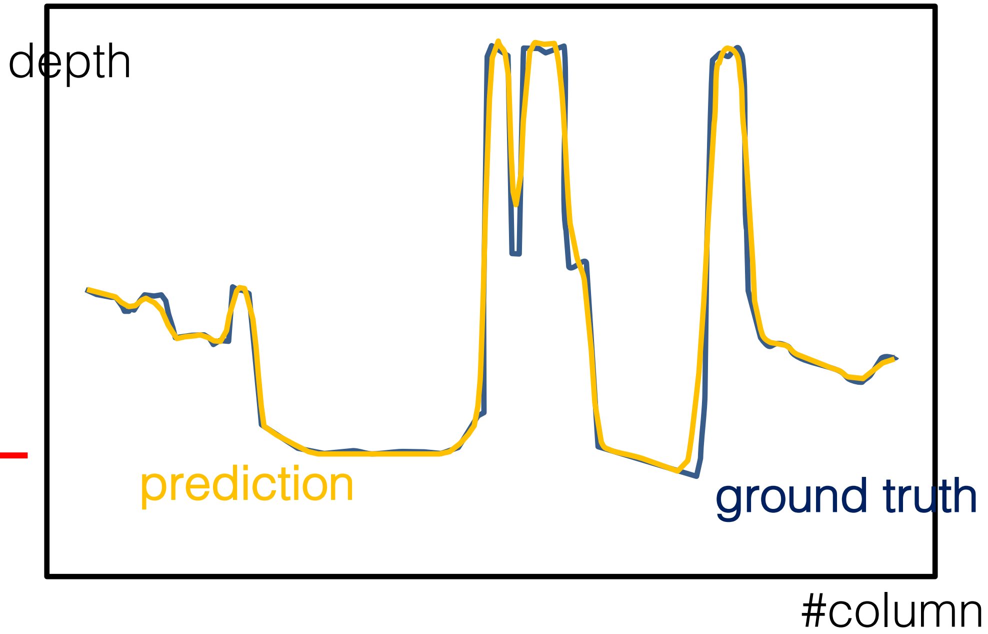


$$
\begin{aligned}
\min_\Theta \sum_i \quad & \mathrm{dist}\left((D_i, N_i, C_i), f(I_i;\Theta)\right) + \\
& \lambda_1 \mathrm{constraint}_1(D(f(I_i;\Theta)), N(f(I_i;\Theta))) + \\
& \lambda_2 \mathrm{constraint}_2(D(f(I_i;\Theta)), C(f(I_i;\Theta))) +
\end{aligned}
$$

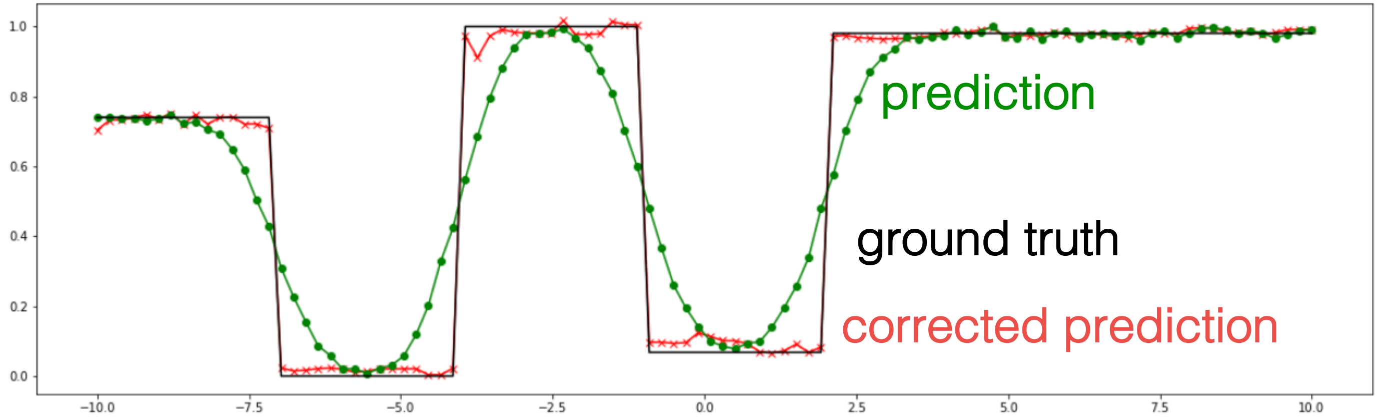
Depth, Normals, Contours Prediction from RGB

École des Ponts
ParisTech


0
100
200
300
400
0
100
200
300
400
500
600

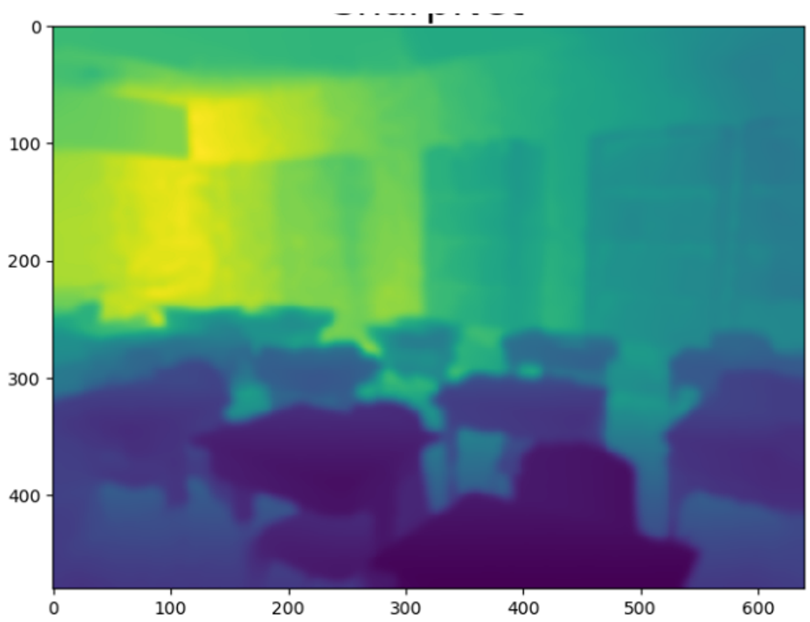

depth
prediction
ground truth
#column

Our Solution

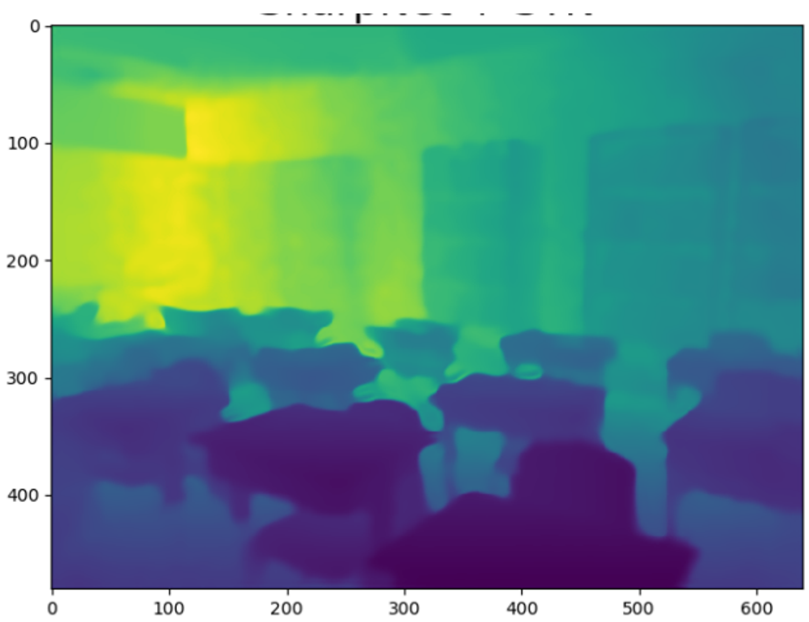


$$\forall \text{ pixel location } x \;\; \text{Corrected Prediction}(x) = \text{Prediction}(x + g(\text{Prediction}; \Theta)(x))$$

École des Ponts
ParisTech



Prediction



Corrected Prediction

PhD Students

Giorgia Pitteri

Hugo Germain

Michael Ramamonjisoa

Yuming Du

Qualcomm

Google

oculus

magic leap

FNSNF

Christian Doppler Forschungsgesellschaft

chist-era

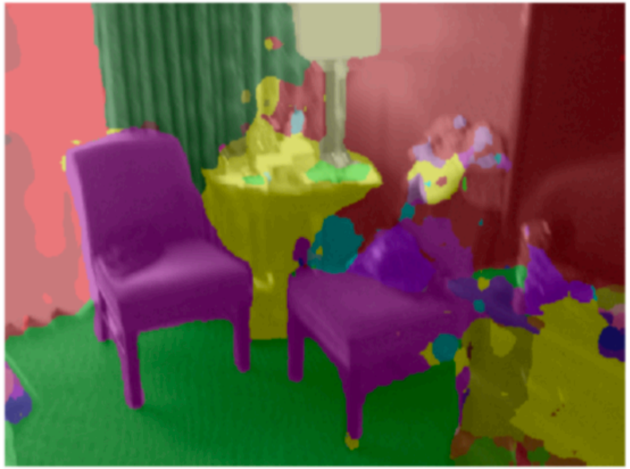
Robert Bosch Stiftung

SIEMENS

Thanks for listening!

Questions?

Learning Semantic Segmentation with Less Training Data

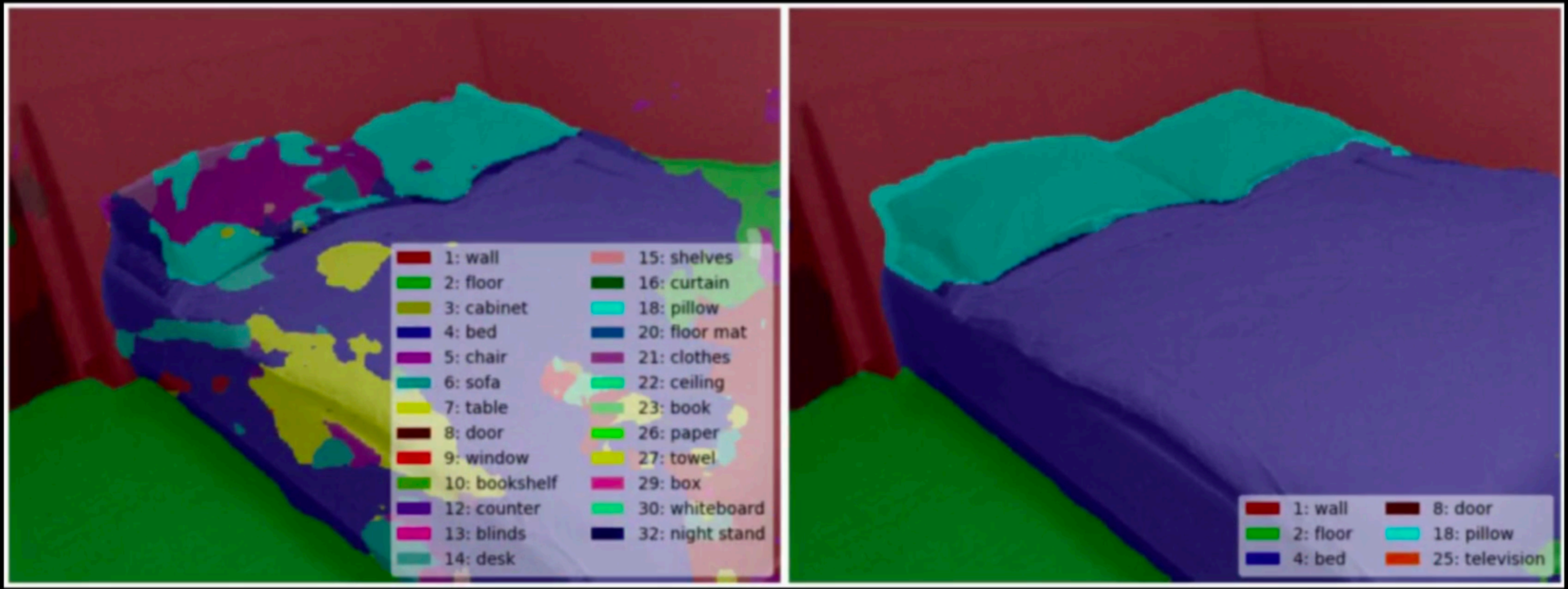


Supervised Learning

Learning with Geometric Constraints

Ground Truth

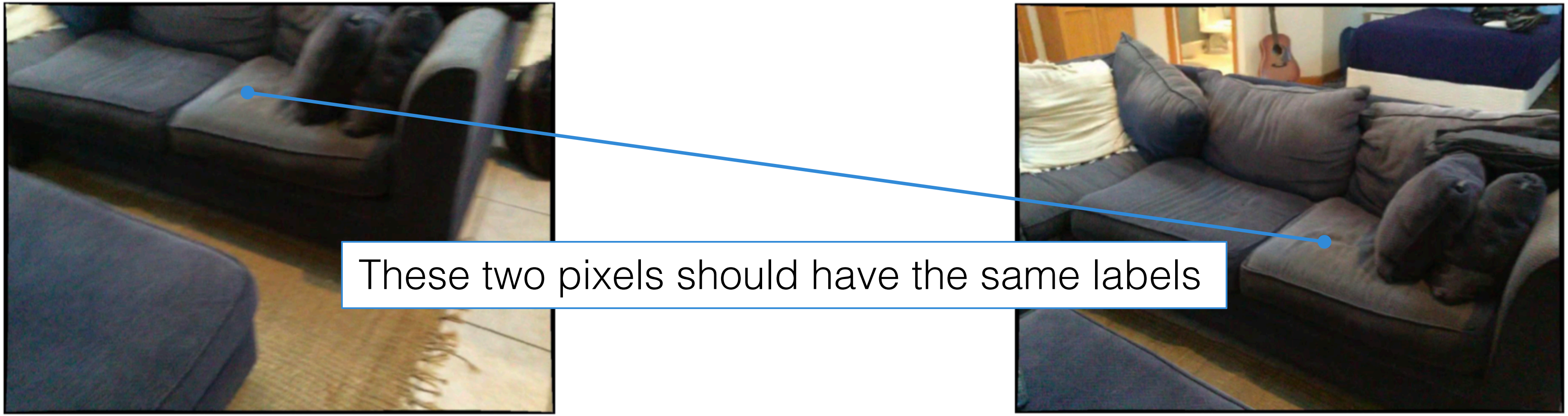
Learning Semantic Segmentation with Less Training Data



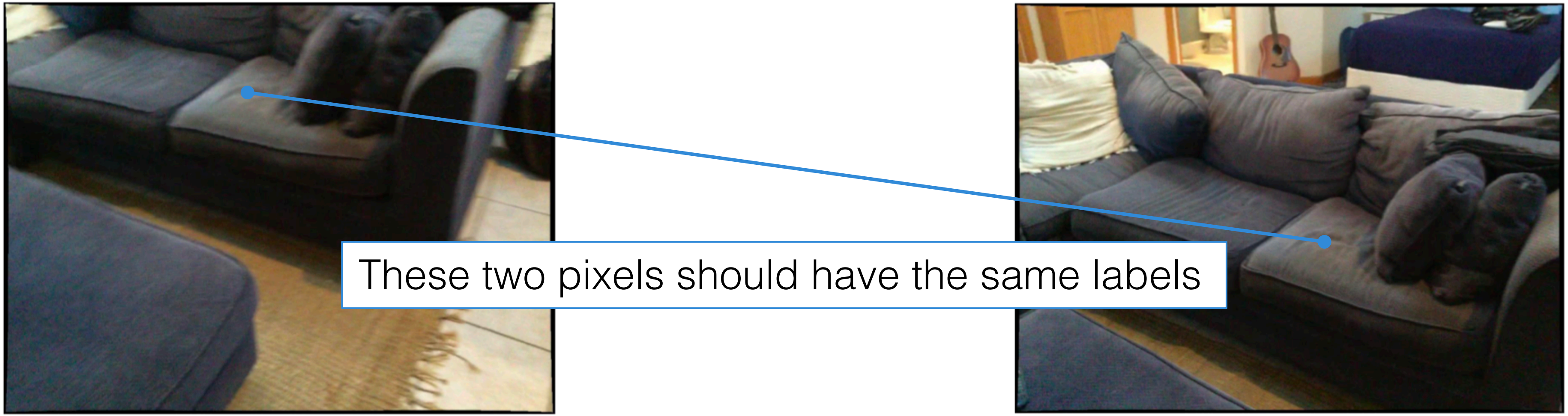
DeepLabV3+ trained supervised

DeepLabV3+ trained with S4-Net

Geometric Constraints



Geometric Constraints as Unsupervised Learning



loss term: $\text{cross-entropy}(\text{Segmenter}(I_1)[\mathbf{m}_1], \text{Segmenter}(I_2)[\mathbf{m}_2])$

Casting Geometric Constraints in Semantic Segmentation as Semi-Supervised Learning

Supplementary Material

Paper Index: 5701