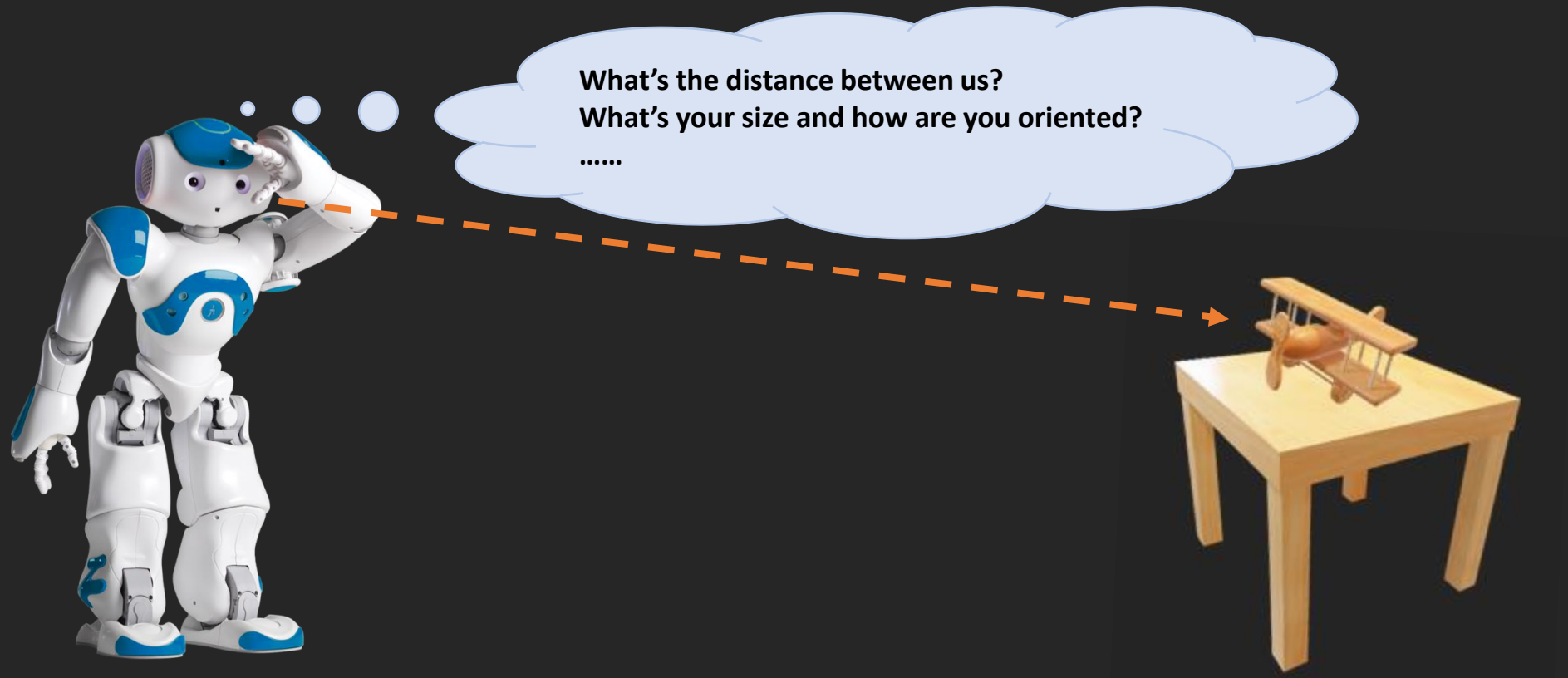
2D and 3D Geometric Attributes Estimation in Images via Deep Learning

Xuchong Qiu

Advisors: Renaud Marlet, Chaohui Wang

Motivations

- Understand the geometric attributes of objects



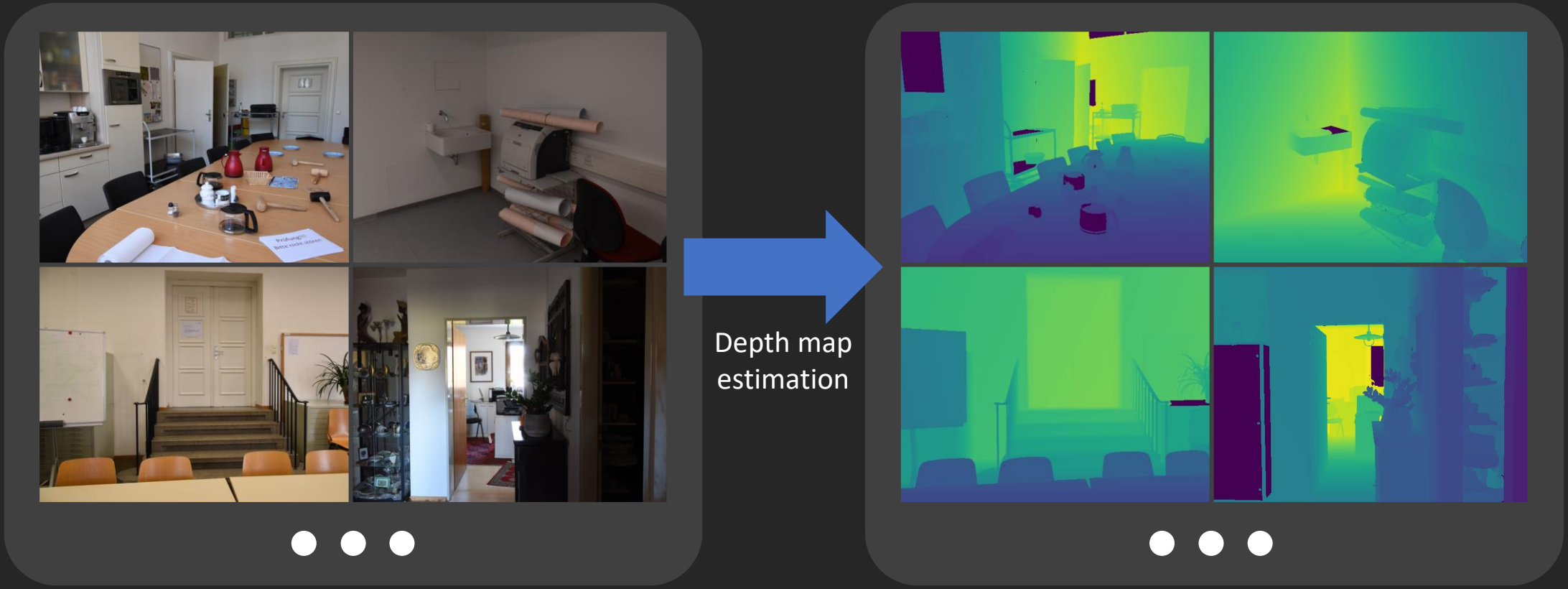
Motivations

- Understand the geometric attributes of the environment

Deep Learning

Iteratively adjusting the model parameters in datasets by learning.

For example:



Our research

1. Single Object Tracking with Structural Semantics
2. Object Pose Estimation with Object Shapes
3. Scene Occlusion Relationship and Depth Estimation

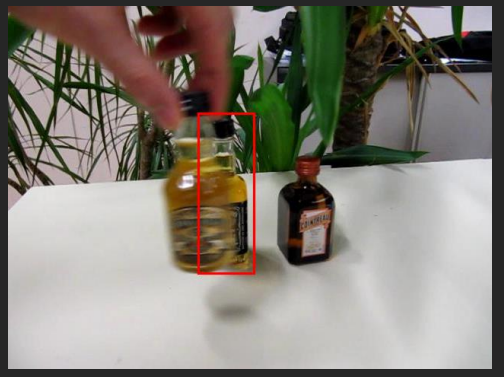
Single Object Tracking with Structural Semantics

The goal of single object tracking


The target in the first frame








Localize the target in the resting frames of the video

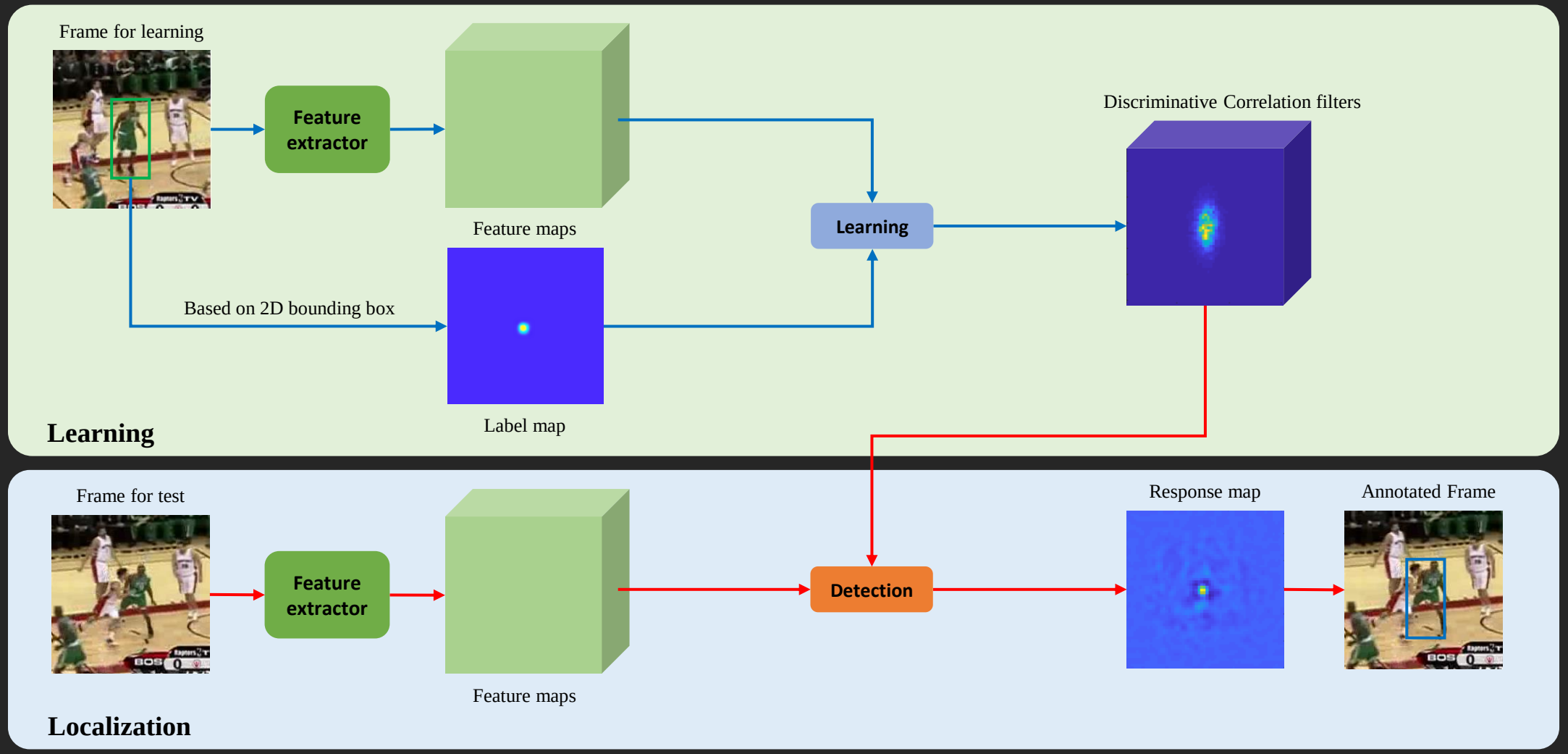
Challenges in object tracking

- Object appearance and size variations

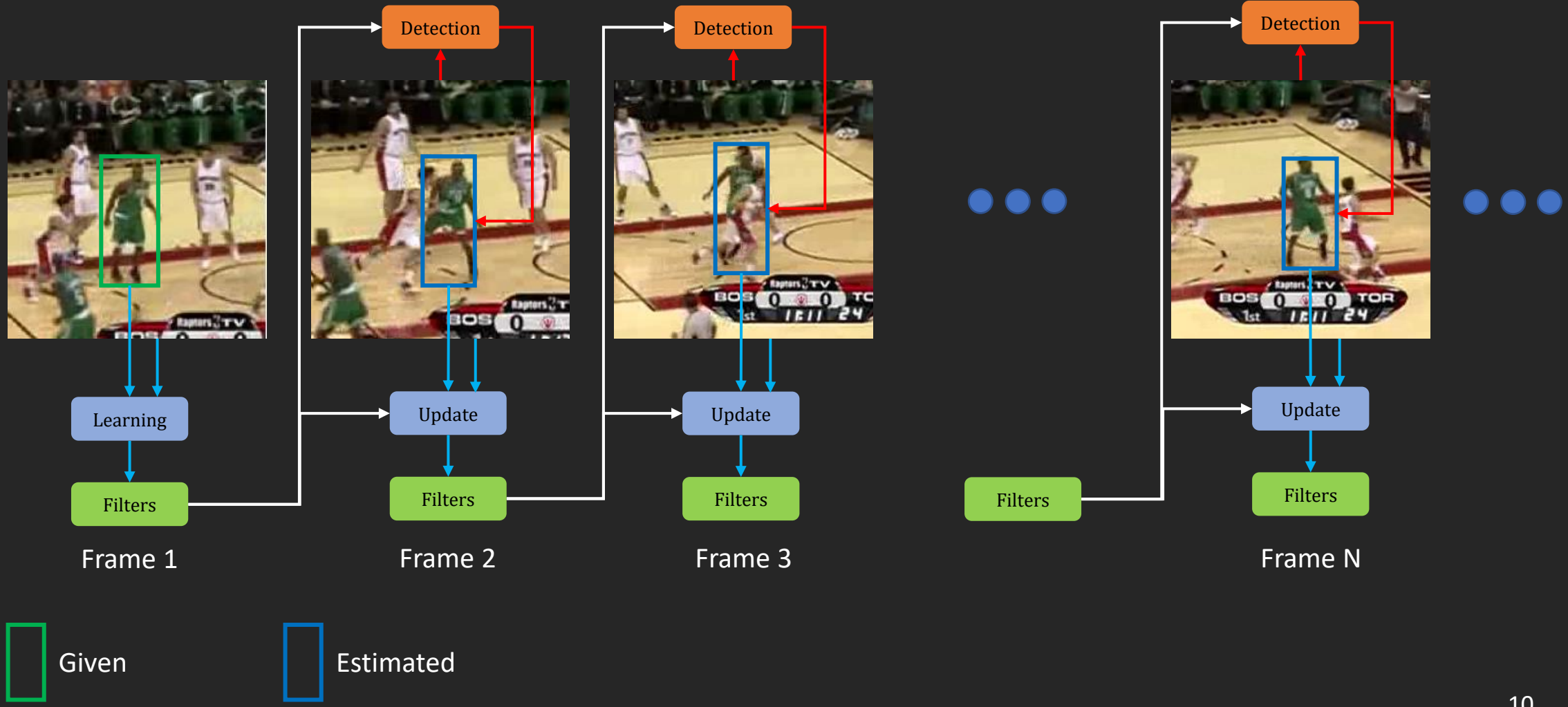
- Motion blur

- Occlusions, illumination changes, background clutter, etc.

DCF-based tracking framework [Henriques et al. PAMI 2015]



Online learning of a DCF-based tracker [Henriques et al. PAMI 2015]

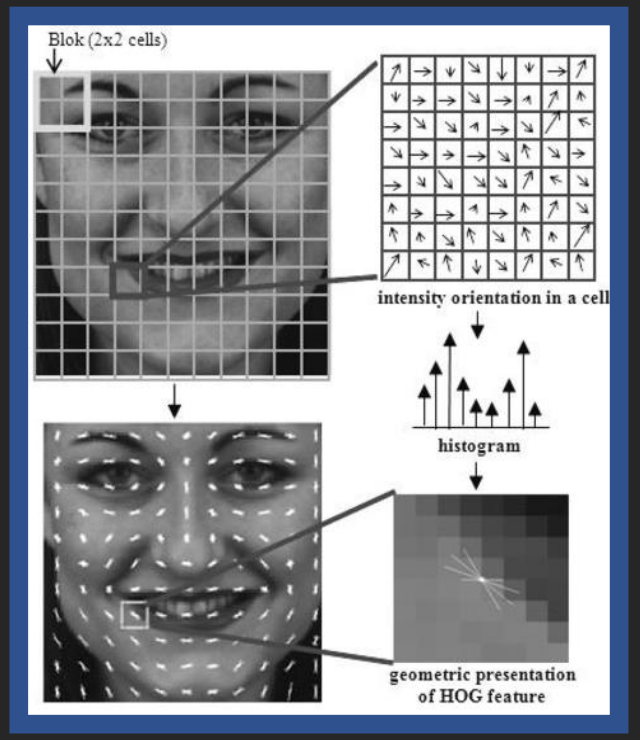


Features used by DCF-based trackers

2010 - 2013 → **2013 - 2015** → **2015 - now**

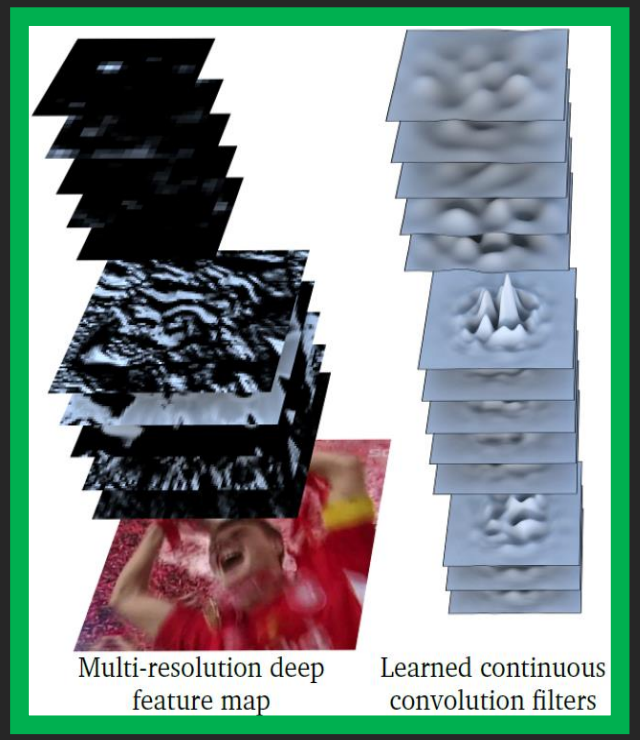
Single-channel features

Grayscale image [Bolme et al. CVPR 2010]



Multi-channel features

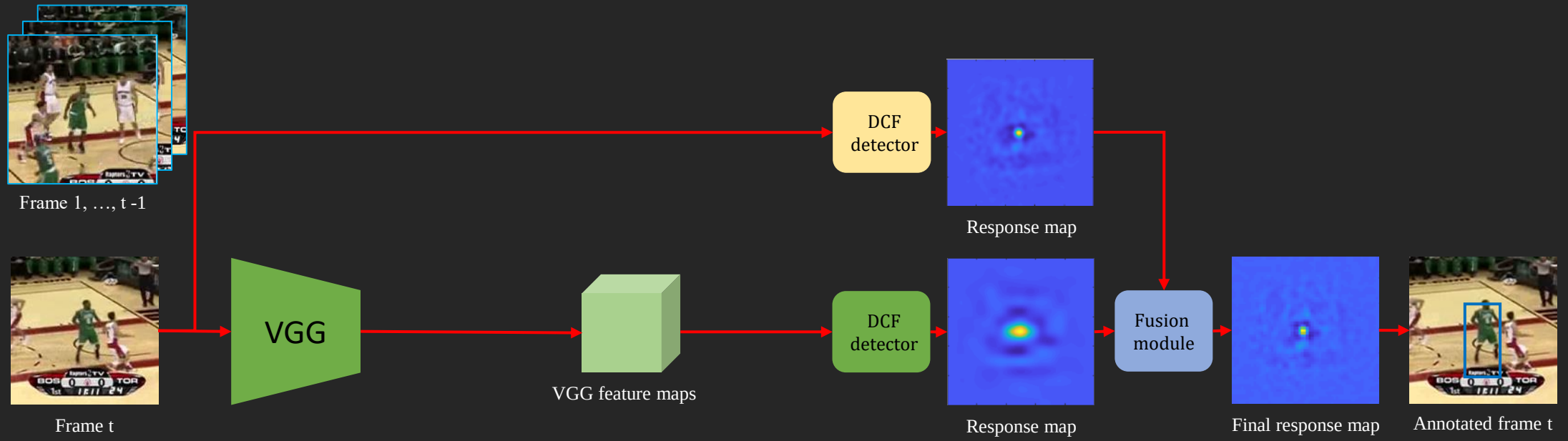
Color attributes [Danelljan et al. CVPR 2014]
HOG [Henriques et al. PAMI 2015]



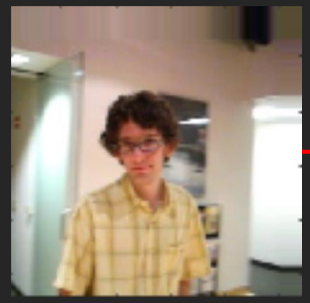
Deep features

Hierarchical deep features [Ma et al. ICCV 2015]
Continuous operator [Danelljan et al. ECCV 2016]
Patch-wise feature [Wang et al. CVPR 2018]

Baseline tracker [Danelljan et al. ECCV 2016]

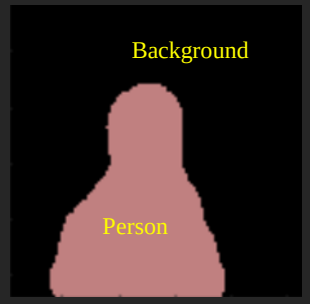


Key idea: using semantic segmentation scores

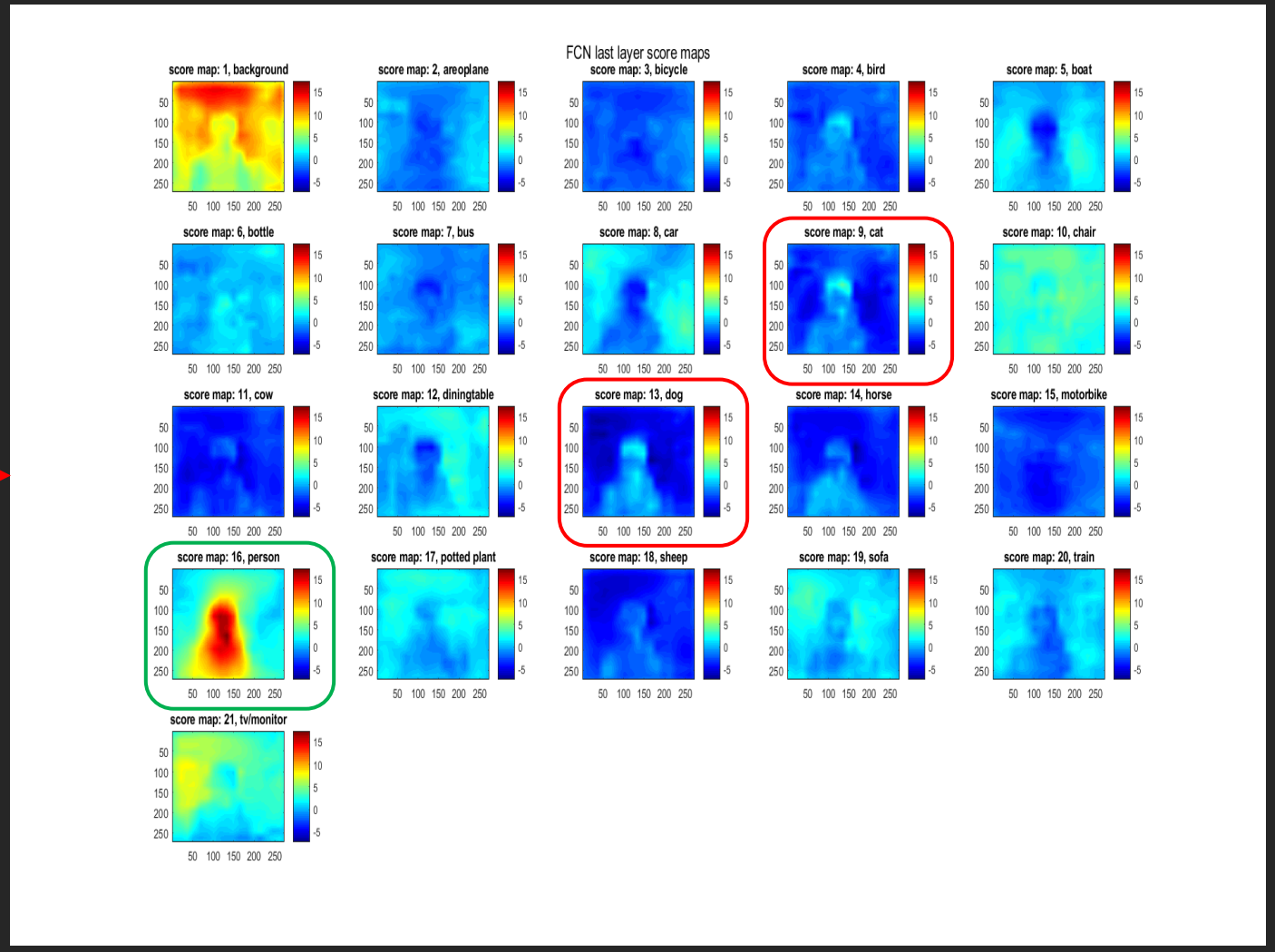


FCN

Deep semantic segmentation model

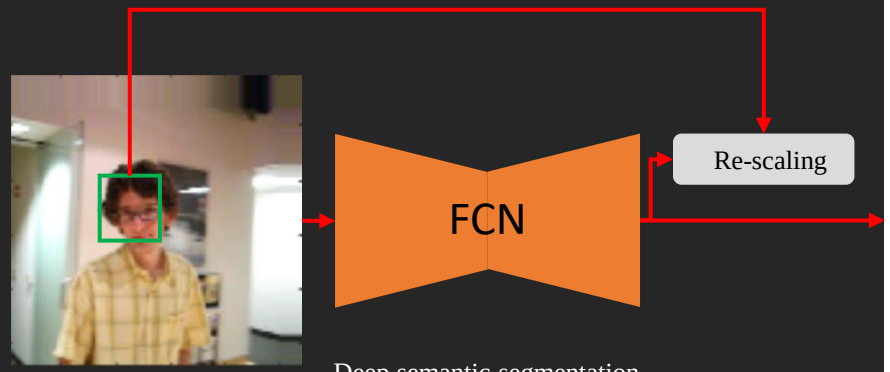


Semantic segmentation

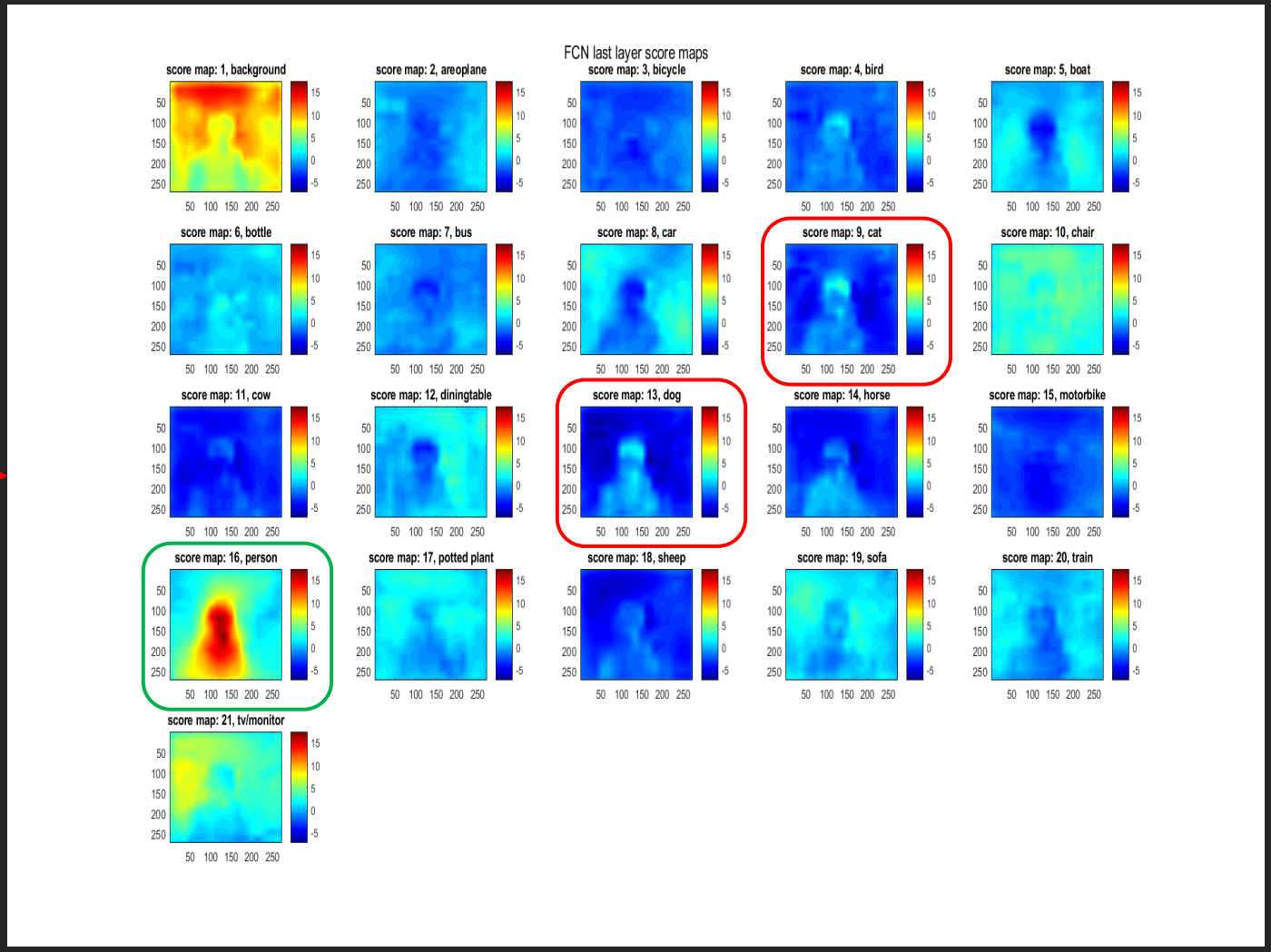


Segmentation score maps

Key idea: using semantic segmentation scores

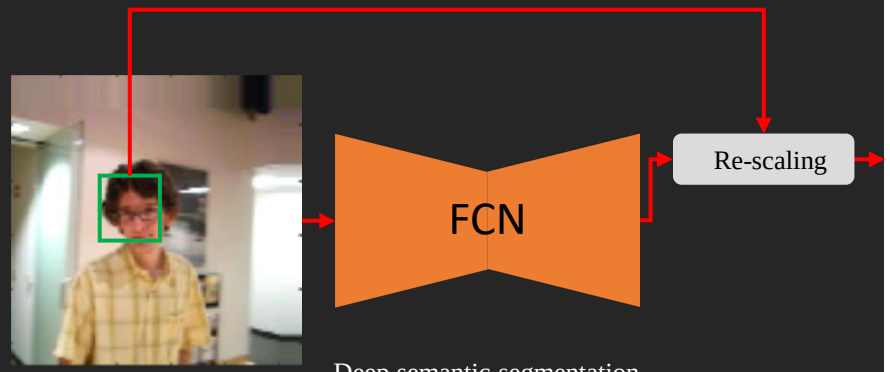


Deep semantic segmentation model

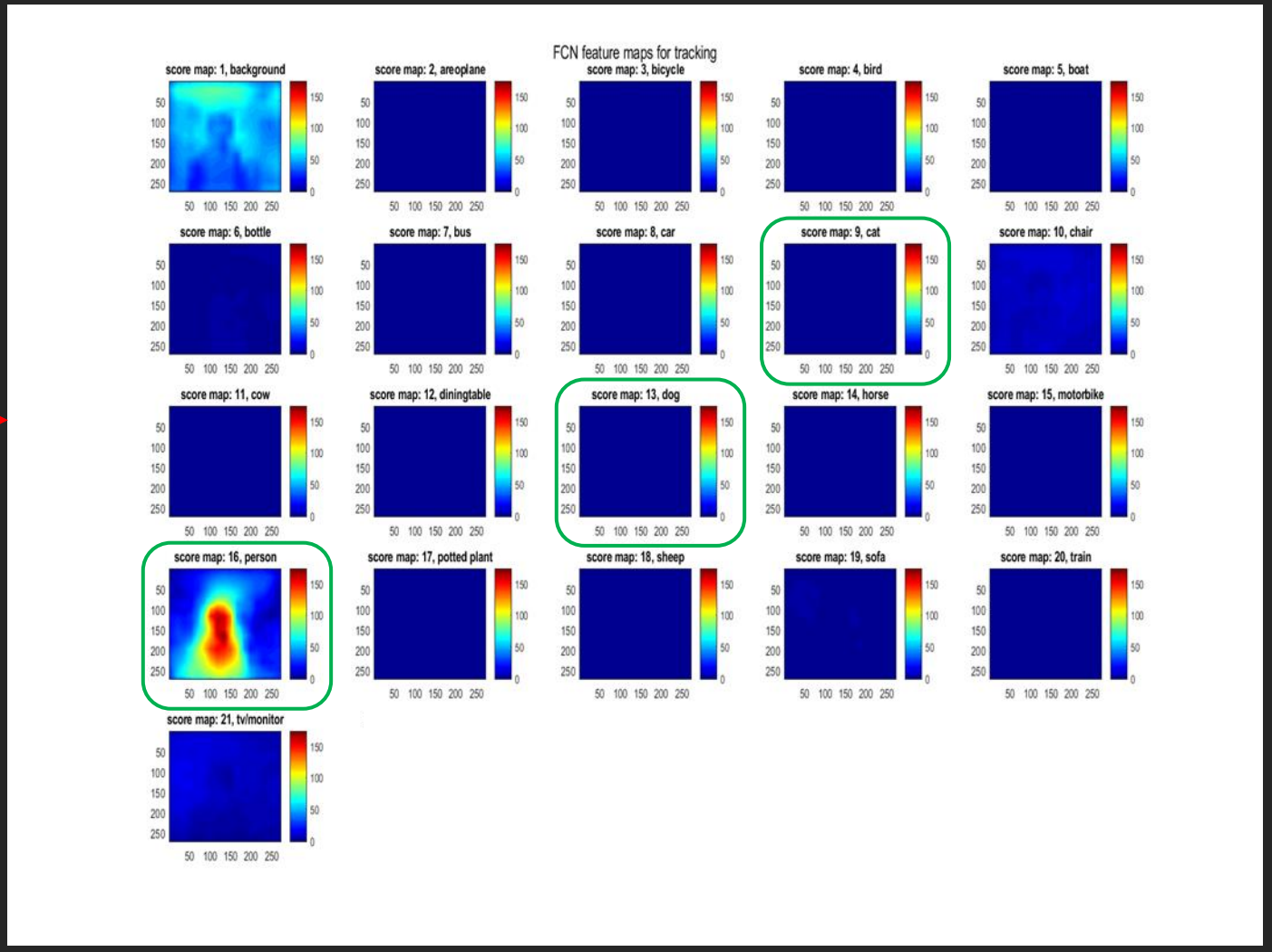


Segmentation score maps

Key idea: using semantic segmentation scores

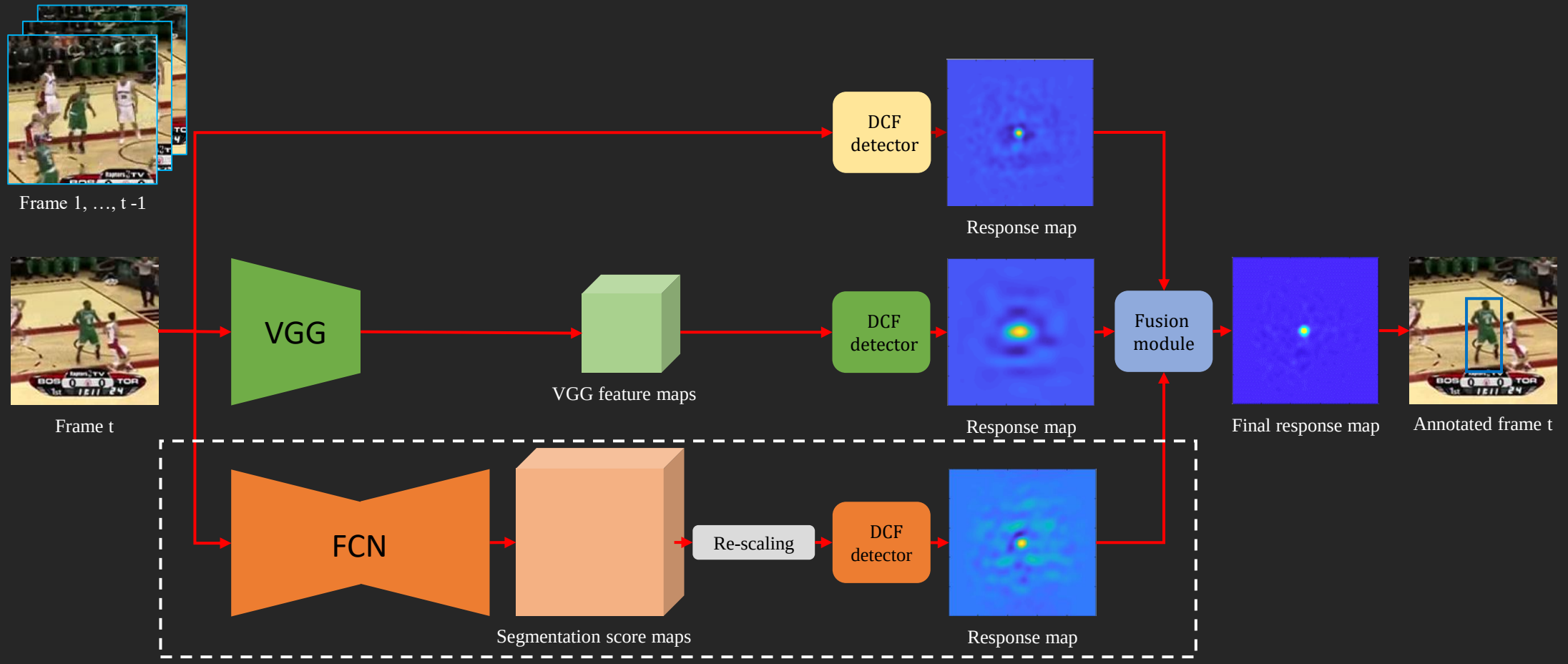


Deep semantic segmentation model



Structural semantic feature maps for tracking

Tracking with structural semantics



Quantitative results on benchmark TB-50

Method	Localization accuracy
Baseline [1]	0.834
Ours (w/o re-scaling)	0.837
Ours	**0.857**

Method	Overlap accuracy
Baseline [1]	0.605
Ours (w/o re-scaling)	0.612
Ours	**0.620**

[1] Danelljan et al. ECCV 2016

Qualitative results on benchmark TB-50

Ours

Baseline [1]

[1] Danelljan, M., *et al.*: Beyond correlation filters: learning continuous convolution operators for visual tracking. In: ECCV, 2016

Qualitative results on benchmark TB-50

Ours

Baseline [1]

[1] Danelljan, M., *et al.*: Beyond correlation filters: learning continuous convolution operators for visual tracking. In: ECCV, 2016

Conclusion

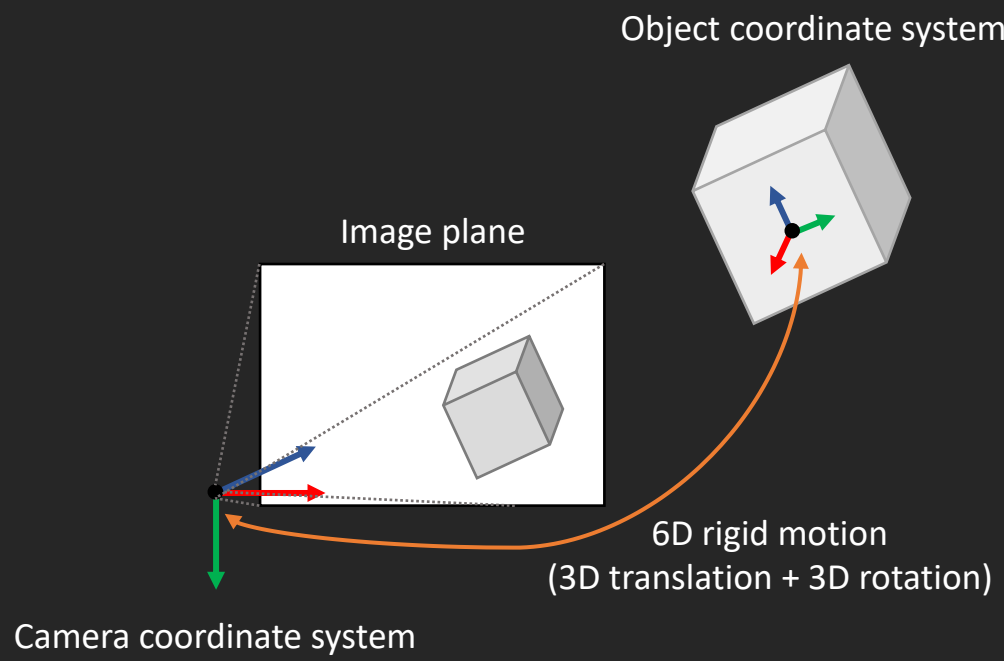
1. Semantic segmentation scores can help tracking
2. Post-processing segmentation scores leads to better results

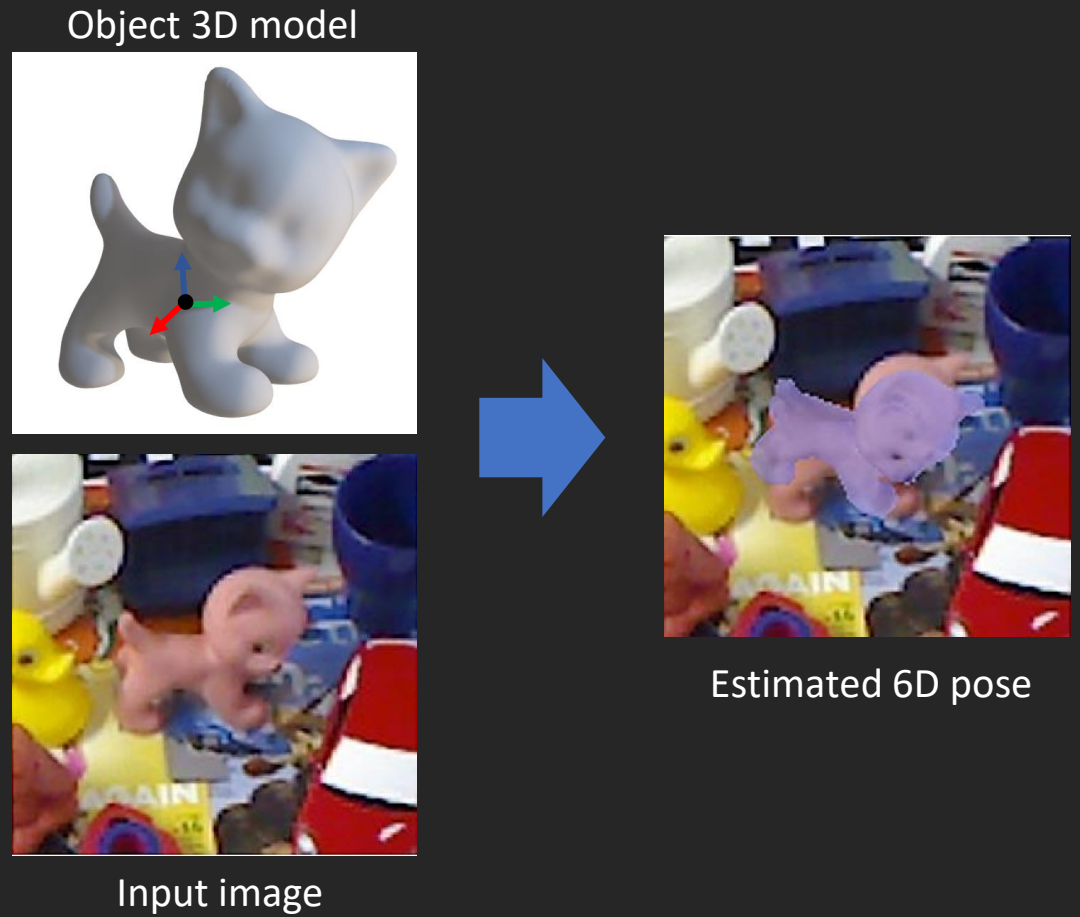
Object Pose Estimation with Object Shapes

Pose From Shape: Deep Pose Estimation for Arbitrary 3D Objects. Yang Xiao, Xuchong Qiu, Pierre-Alain Langlois, Mathieu Aubry, and Renaud Marlet. In: BMVC 2019.

The goal of object pose estimation

- Estimate the 6D rigid motion between the object and the camera





Challenges in object pose estimation

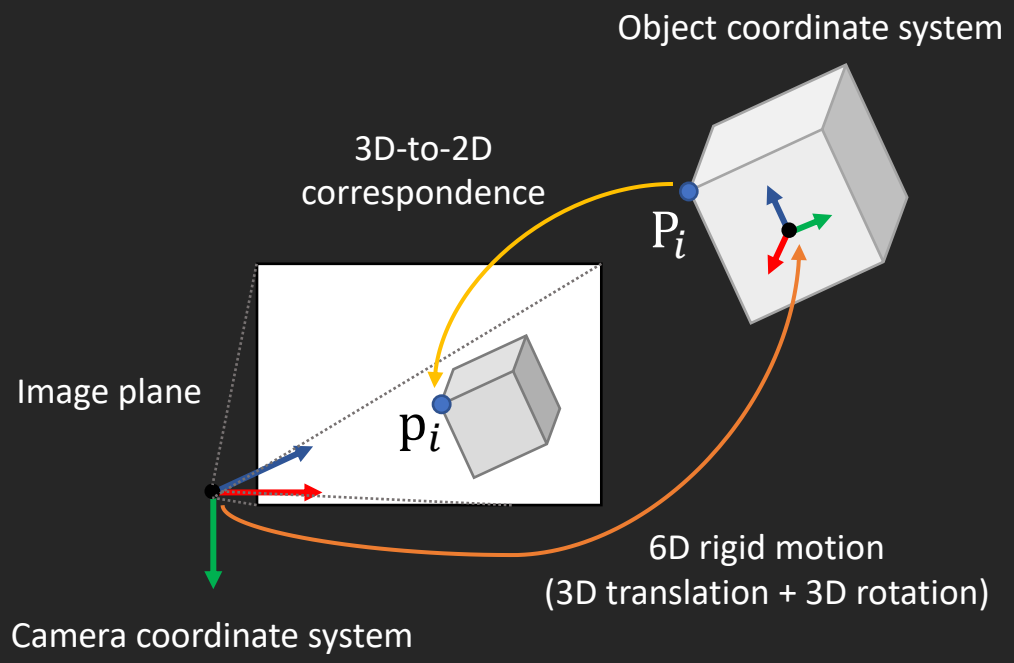
- Textured/weakly-textured object appearance variations [1]

- Unconstrained backgrounds for objects in the wild [2]

- High degrees of freedom, occlusions, changing illumination conditions, etc.

[1] Hinterstoisser, S., *et al.*: Model based training, detection and pose estimation of texture-less 3d objects in heavily cluttered scenes. In: ACCV, 2012
[2] Xiang, Y., *et al.*: ObjectNet3D: A large scale database for 3D object recognition. In: ECCV, 2016

Pose estimation with 3D-to-2D point correspondences



- **Perspective projection model of pinhole camera:**

$$s_i \bar{\mathbf{p}}_i = \mathbf{K}[\mathbf{R}|\mathbf{t}]\bar{\mathbf{P}}_i$$

where

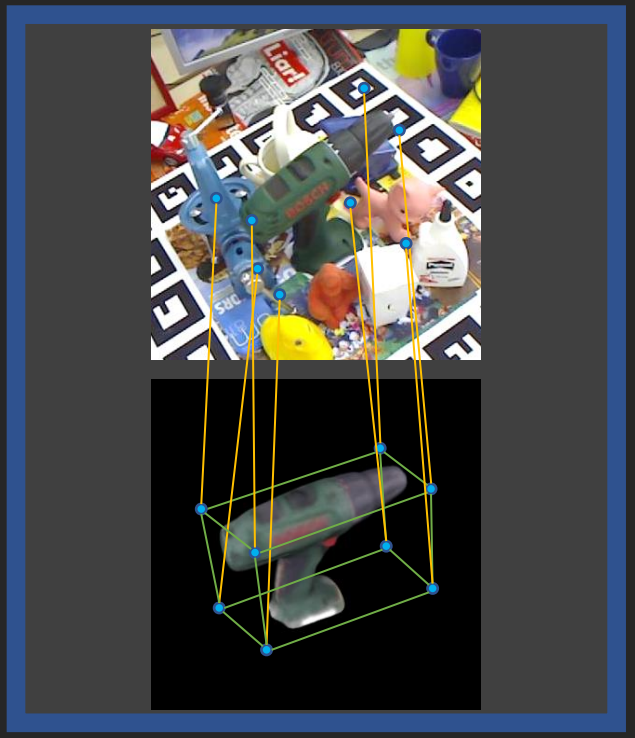
$\bar{\mathbf{P}}_i = [x\, y\, z\, 1]^T$	3D point in the object coordinate system
$\bar{\mathbf{p}}_i = [u\, v\, 1]^T$	2D point in the image coordinate system
$\mathbf{R}$	3D rotation of the object w.r.t. the camera
$\mathbf{t}$	3D translation of the object w.r.t. the camera
$\mathbf{K}$	Camera intrinsic matrix
s_i	Scale factor of the point

- **PnP algorithm:**

$$\{\mathbf{p}_i\}_{i=1}^n, \{\mathbf{P}_i\}_{i=1}^n, \mathbf{K} \rightarrow \boxed{\text{PnP}} \rightarrow \mathbf{R}, \mathbf{t}$$

Deep pose estimators using 3D-to-2D correspondences

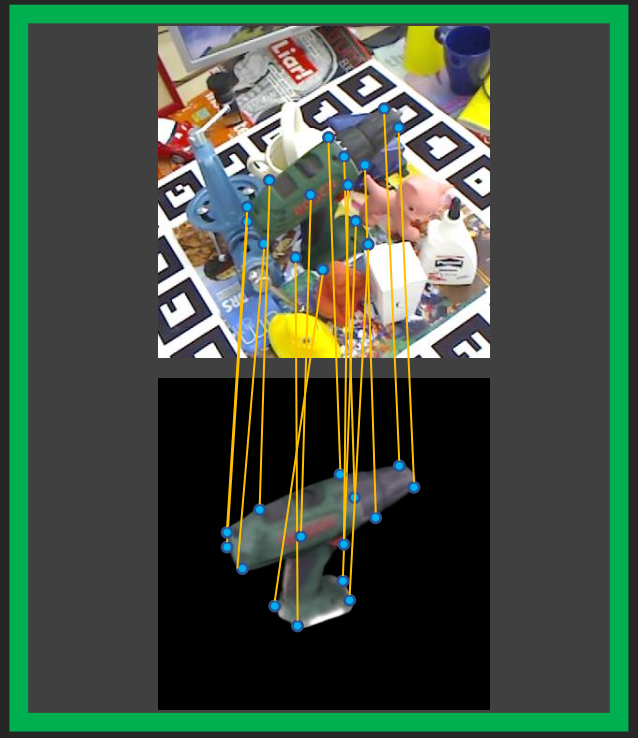
2017-2018



Object bounding box corners

BB8 [Rad et al. ICCV 2017]
YOLO-6D [Tekin et al. CVPR 2018]

2018 - 2019



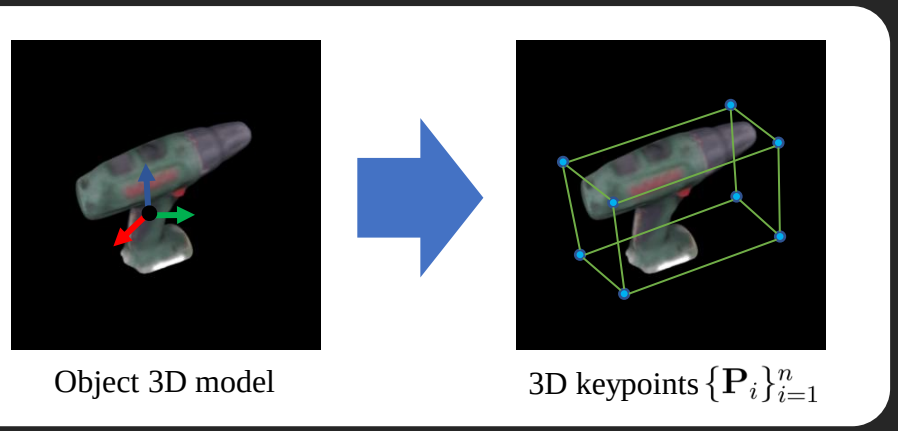
Object surface points

PVNet [Peng et al. CVPR 2019]
Our work

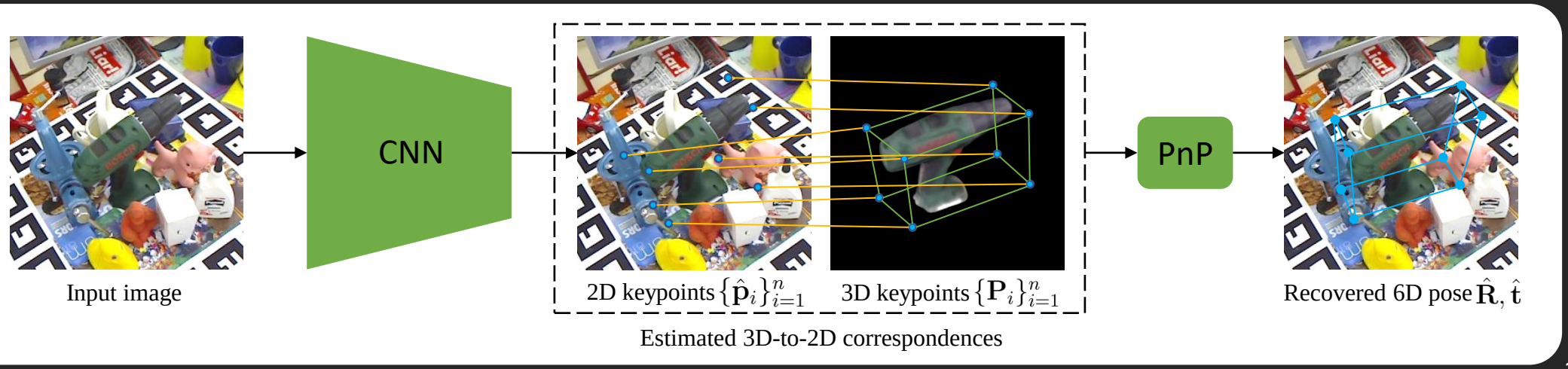
2019 - now

Pose estimation with 3D-to-2D point correspondences

- Baseline: using the object 3D bounding box corners as 3D keypoints [YOLO-6D Tekin et al. CVPR 2018]

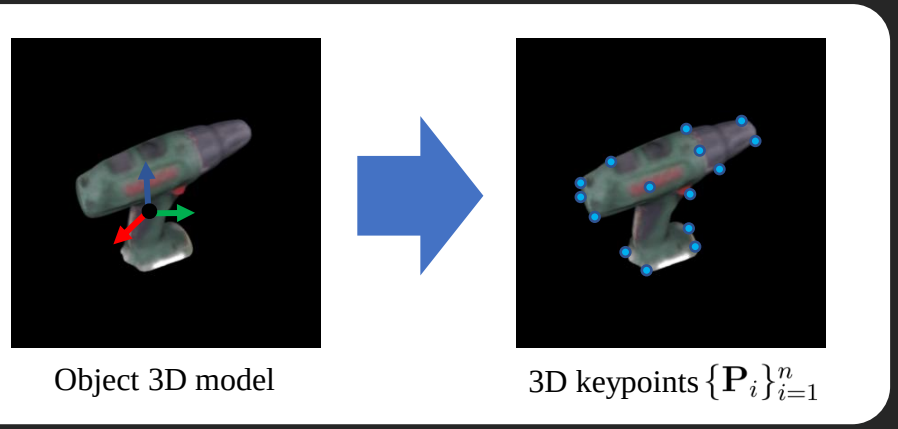


- Object pose estimation by predicting 2D keypoints in the input image

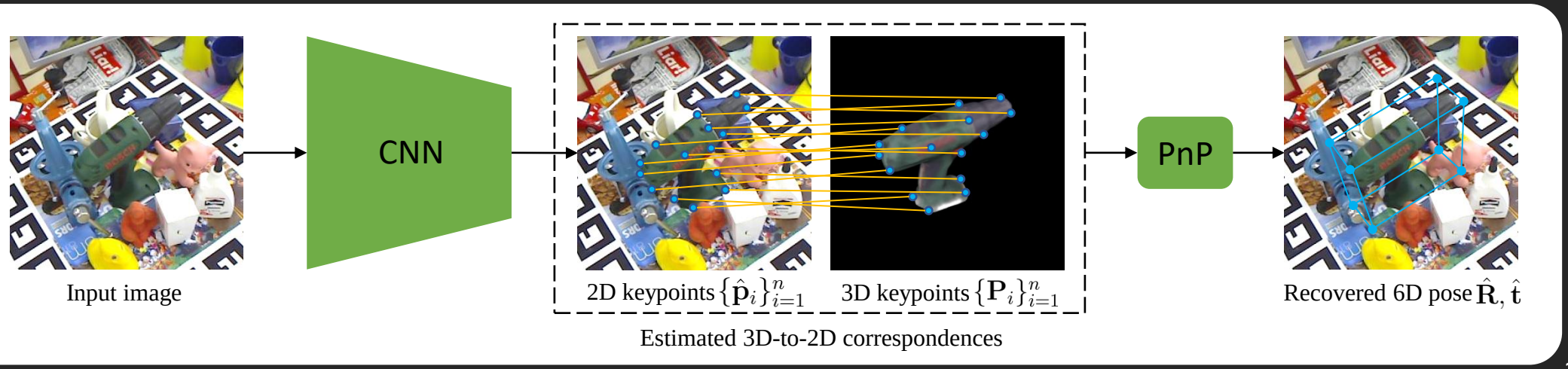


Key idea: using the object surface points

- Ours: randomly sampling the object surface points as 3D keypoints



- Object pose estimation by predicting 2D keypoints in the input image



Quantitative results on dataset LINEMOD

- 2D point estimation error

LINEMOD	ape	bvise	cam	can	cat	drill	duck	ebox	glue	holep	iron	lamp	phone	**mean**
2D Projection Error in pixels (lower is better)														
Baseline [1]	4.05	5.00	4.59	4.08	3.84	5.58	4.89	4.38	3.57	5.62	7.08	7.05	7.86	5.20
Ours (100 points)	**2.28**	**3.06**	**3.26**	**2.59**	**2.38**	**4.32**	**3.71**	**3.76**	**2.51**	**4.01**	**5.97**	**4.26**	**5.41**	**3.66**

- 6D pose estimation accuracy with the ADD-0.1d metric

LINEMOD		ape	bvise	cam	can	cat	drill	duck	ebox*	glue*	holep	iron	lamp	phone	**mean**
ADD-0.1d in percentages (higher is better)															
	Branchmann [2]	-	-	-	-	-	-	-	-	-	-	-	-	-	32.2
	SSD-6D [3]	0	0.2	0.4	1.4	0.5	2.6	0	8.9	0	0.3	8.9	8.2	0.2	2.4
w/o Refine.	BB8 [4]	27.9	62.0	40.1	48.1	**45.2**	58.6	**32.8**	40.0	27.0	42.4	67.0	39.9	35.2	43.6
	Baseline [1]	21.6	78.7	33.4	67.5	41.8	63.5	25.6	65.5	80.1	42.3	**69.3**	67.9	**47.7**	54.2
	Ours	**33.9**	**80.6**	**43.9**	**75.2**	44.2	**66.1**	29.2	**66.7**	**81.1**	**48.6**	63.9	**71.4**	42.0	**57.5**

* ADD-S-0.1d used for symmetric objects eggbox and glue.

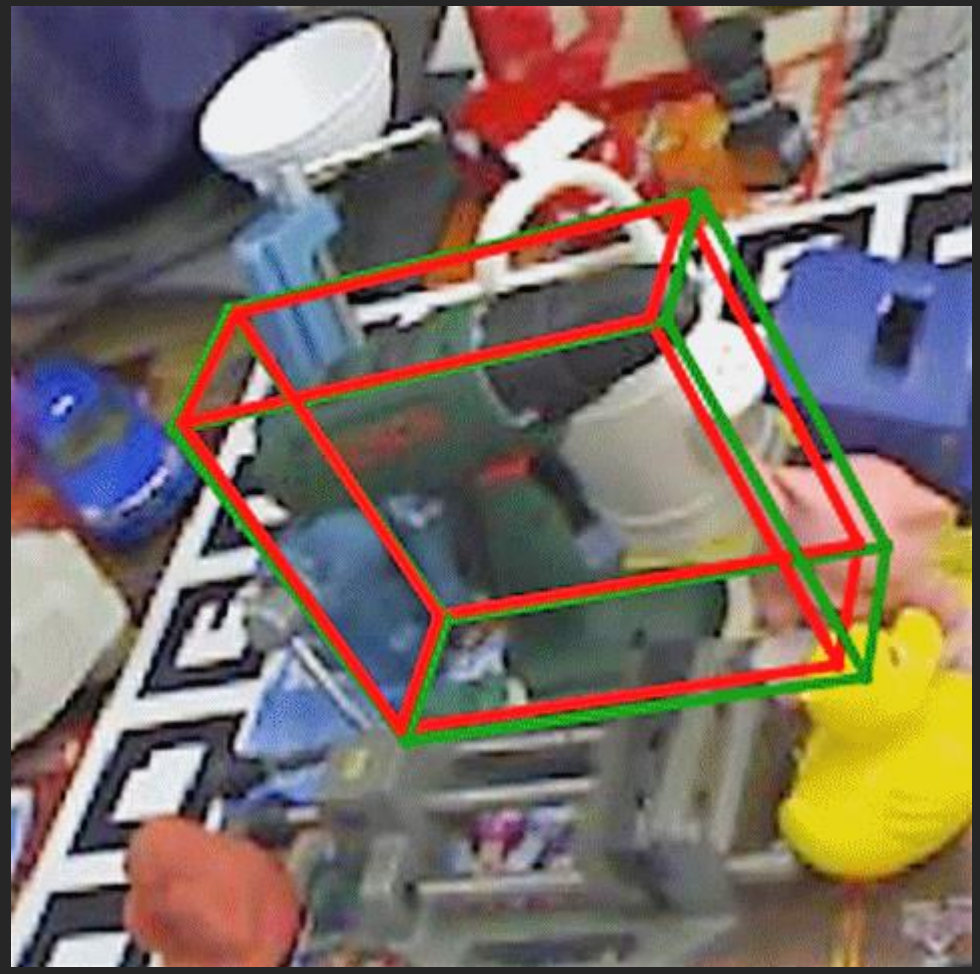
[1] Tekin et al. CVPR 2018, [2] Branchmann et al. CVPR 2016, [3] Kehl et al. ICCV 2017, [4] Rad et al. ICCV 2017

Qualitative results on dataset LINEMOD

Ground truth

Baseline [1]

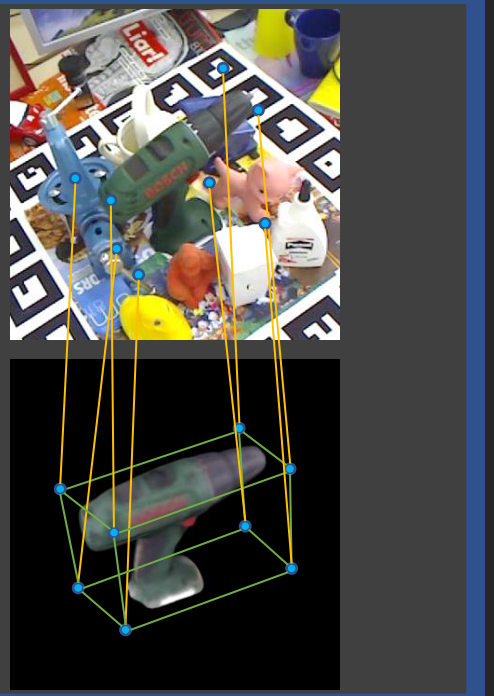
Ours



[1] Tekin, B., *et al.*: Real-time seamless single shot 6D object pose prediction. In: CVPR, 2018

Deep pose estimators using 3D-to-2D correspondences

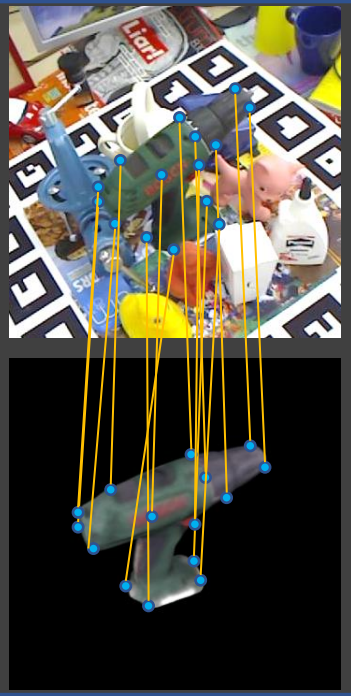
2017-2018



Object bounding box corners

BB8 [Rad et al. ICCV 2017]
YOLO-6D [Tekin et al. CVPR 2018]

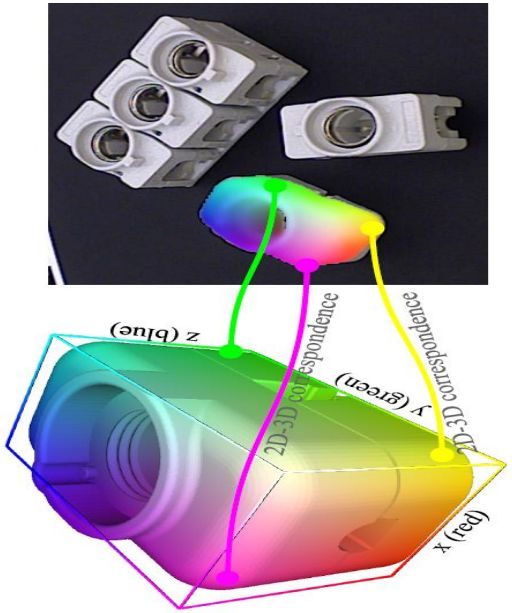
2018 - 2019



Object surface points

PVNet [Peng et al. CVPR 2019]
Our work

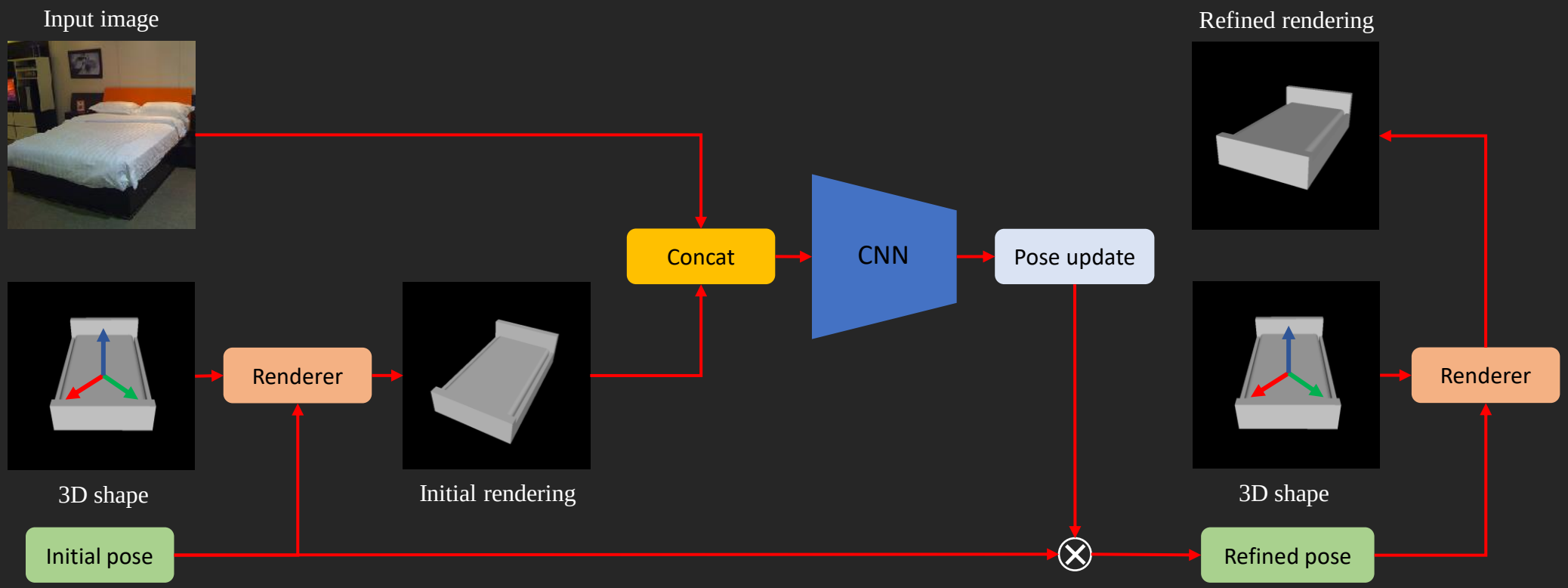
2019 - now



Dense correspondences

DPOD [Zakharov et al. ICCV 2019]
Pix2Pose [Park et al. ICCV 2019]

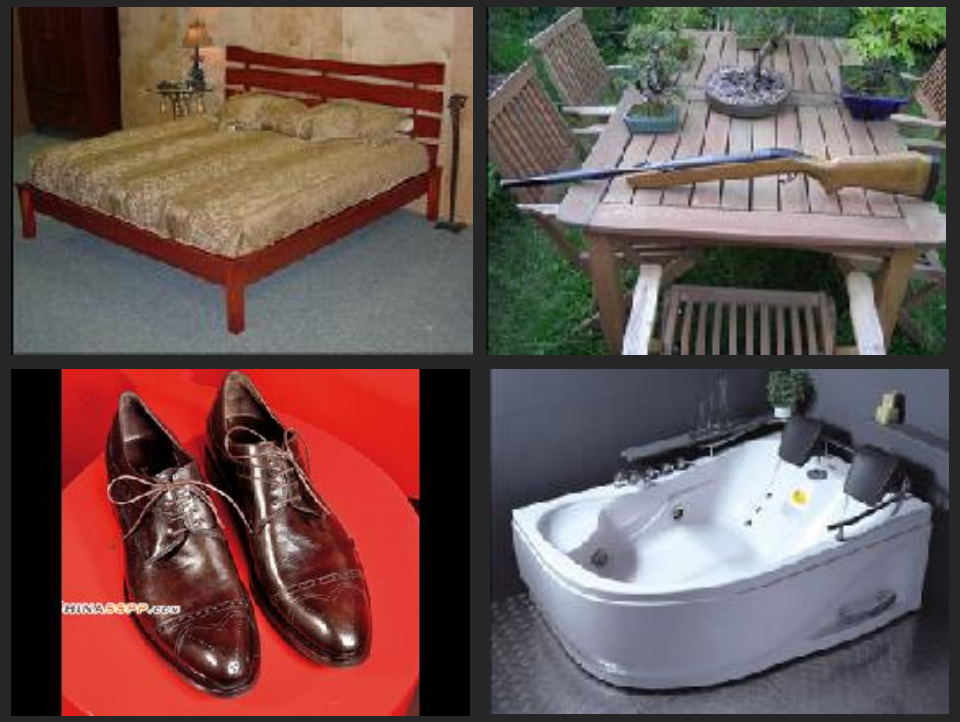
Object pose refinement pipeline [1]



[1] Li, Y., *et al.*: DeepIM: Deep iterative matching for 6D pose estimation. In: ECCV, 2018

LINEMOD dataset vs. Objects in the wild

Objects in LINEMOD [1]

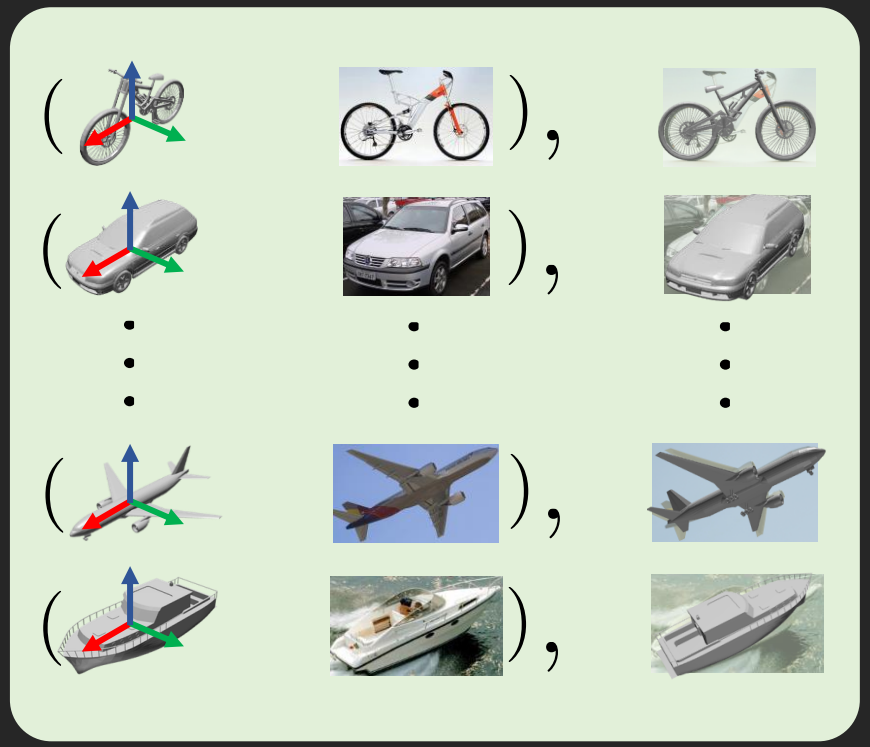


Objects in the wild

[1] Hinterstoisser et al. ACCV 2012

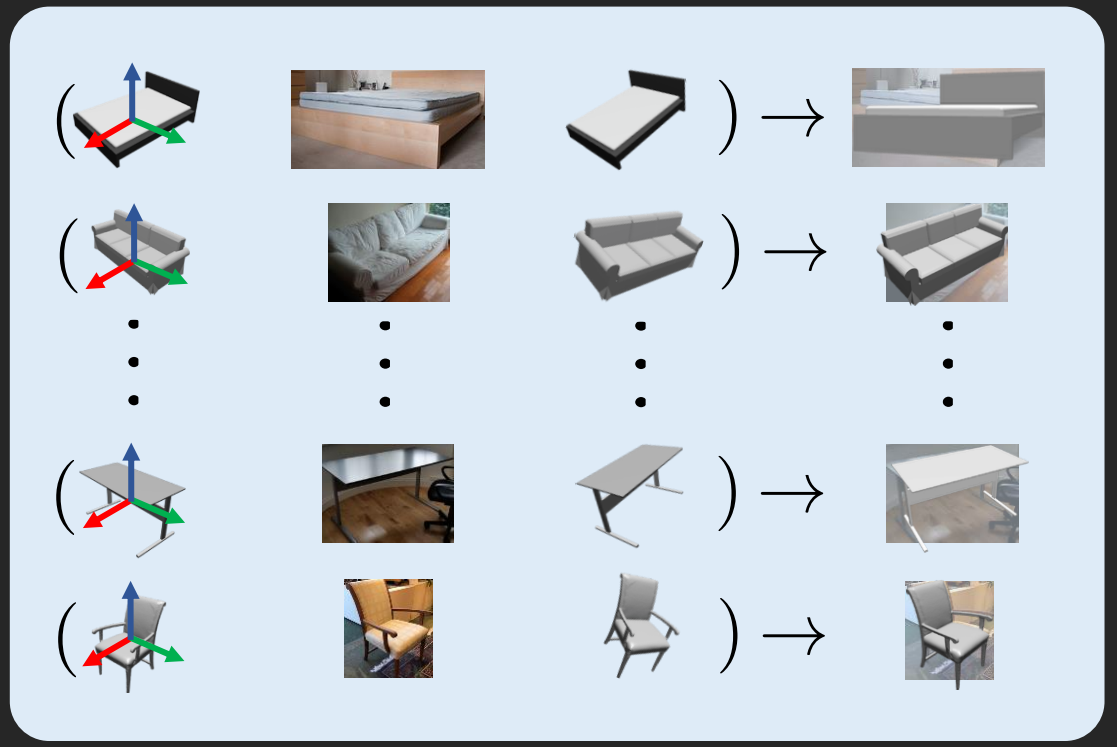
Key idea: pose refinement for unseen objects in the wild

Training set



3D shape | Input image | Ground-truth pose

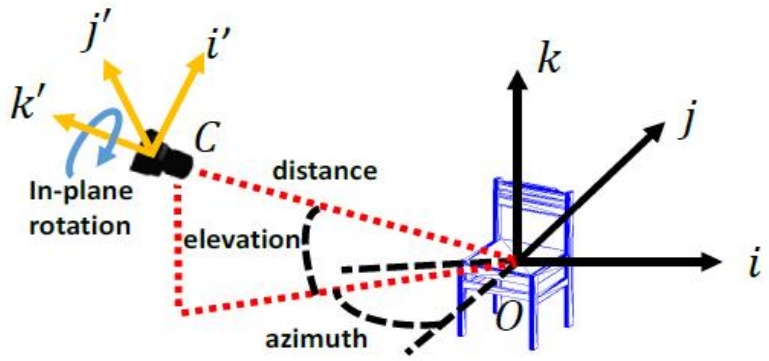
Test set (unseen object categories)



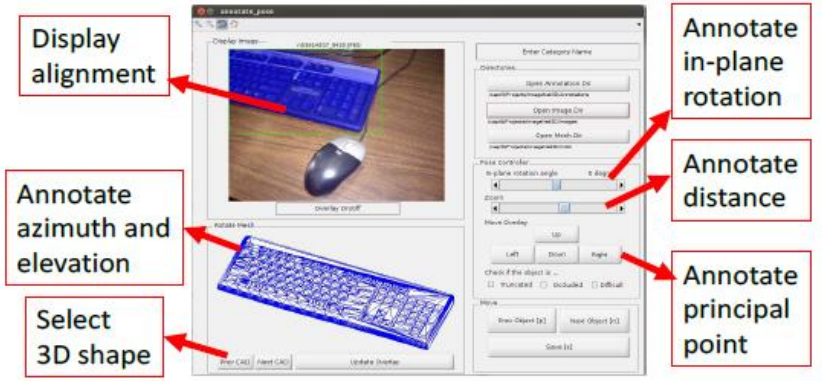
3D shape | Input image | Initial pose | Refined pose

Data preparation for pose refinement

- Labeling object poses in real images [1]



Camera frame and object frame in the virtual 3D scene



Annotation interface

- Synthetic data generation by domain randomization (texture + background)

[1] Xiang, Y., *et al.*: ObjectNet3D: A large scale database for 3D object recognition. In: ECCV, 2016

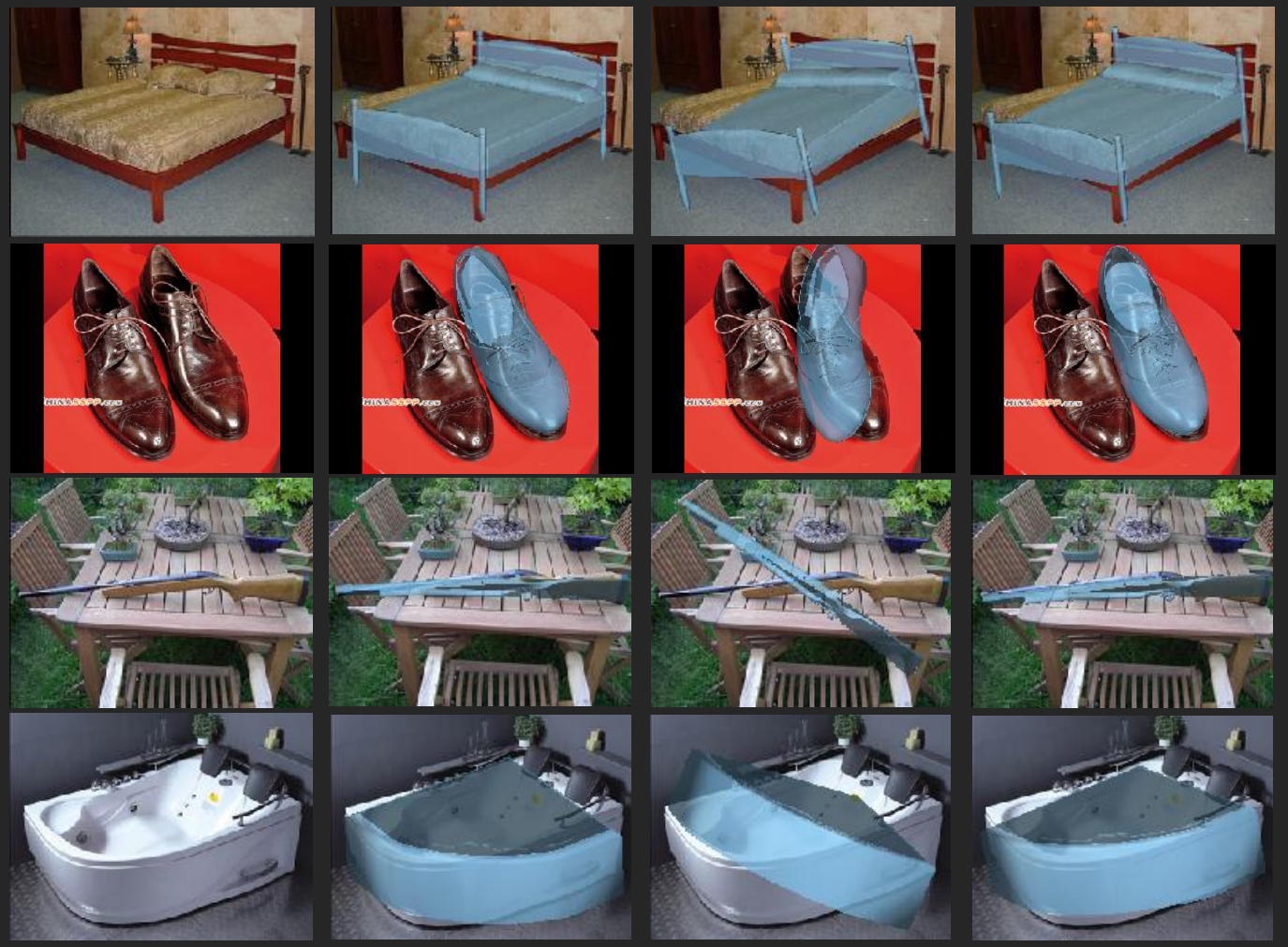
Quantitative results on dataset ObjectNet3D

ObjectNet3D	bed	bookcase	calc	cellphone	comp	door	cabinet	guitar	iron	knife	micro
					***Acc*$_{\frac{\pi}{6}}$ in percentages (higher is better)**						
Init	76	73	79	80	80	78	77	82	76	**72**	81
Ours (synthetic)	82	74	77	74	82	80	83	84	82	71	**94**
Ours (real)	**85**	**85**	**91**	**85**	**90**	**92**	**89**	**86**	**93**	**72**	91
					***MedErr* in degrees (lower is better)**						
Init	21	24	22	21	22	21	21	19	19	24	23
Ours (synthetic)	20	23	20	21	16	19	17	18	20	24	**14**
Ours (real)	**16**	**19**	**17**	**18**	**15**	**15**	**15**	**16**	**17**	**21**	**14**

	pen	pot	rifle	shoe	slipper	stove	toilet	tub	wheelchair	**mean**
					***Acc*$_{\frac{\pi}{6}}$ in percentages (higher is better)**					
Init	79	76	79	76	**80**	72	76	72	80	77
Ours (synthetic)	77	74	86	78	74	84	73	75	76	79
Ours (real)	**81**	**80**	**89**	**85**	78	**88**	**82**	**84**	**84**	**86**
					***MedErr* in degrees (lower is better)**					
Init	22	21	18	23	**21**	22	22	24	22	22
Ours (synthetic)	21	19	17	19	22	17	23	21	20	20
Ours (real)	**18**	**18**	**16**	**17**	**21**	**15**	**18**	**19**	**18**	**17**

[images: 90,127 in the wild | objects: 201,888 | categories: 100 | 3D models: 791]

Qualitative results on dataset ObjectNet3D



Input image | Ground-truth pose | Initial pose | Refined pose

Evaluation on unseen dataset LINEMOD

LINEMOD	ape	bvise	cam	can	cat	drill	duck	ebox*	glue*	holep	iron	lamp	phone	**mean**
ADD-0.1d in percentages (higher is better)														
Init	4.3	19.0	14.9	8.8	12.2	18.8	6.4	100	81.85	20.9	9.8	25.2	10.8	25.6
w/ our refinement	**24.7**	**24.7**	**21.1**	**15.9**	**17.5**	**36.1**	**13.0**	100	54.63	**32.9**	**27.6**	**25.0**	**20.1**	**31.8**

* ADD-S-0.1d used for symmetric objects eggbox and glue.

Input image | Ground-truth pose | Initial pose | Refined pose

Conclusion

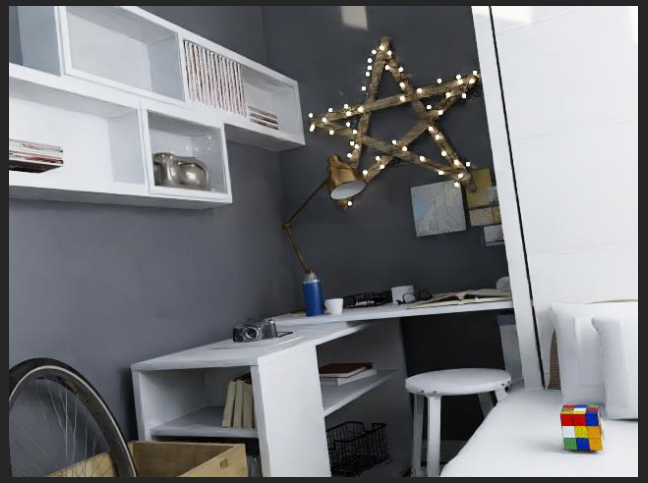
1. Predicting object 3D surface points in images leads to better results
2. Pose refiners can work well for unseen objects in the wild with a proper training

Scene Occlusion Relationship and Depth Estimation

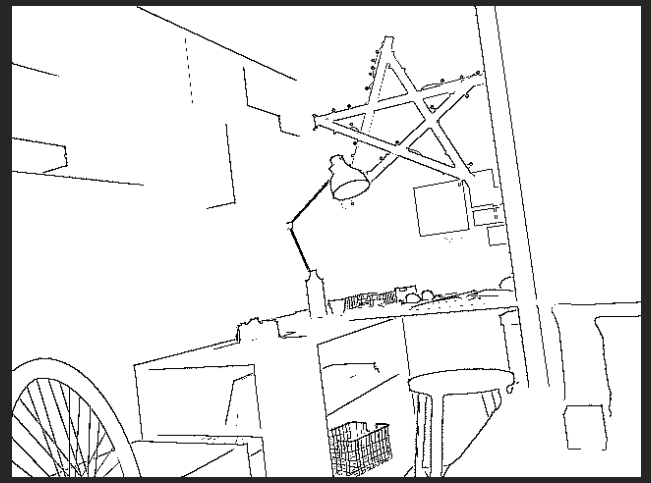
Pixel-Pair Occlusion Relationship Map (P2ORM): Formulation, Inference & Application.
Xuchong Qiu, Yang Xiao, Chaohui Wang, and Renaud Marlet. In: ECCV 2020 (Spotlight).

Goals

- Estimate geometric occlusion and depth map from an image



Input image

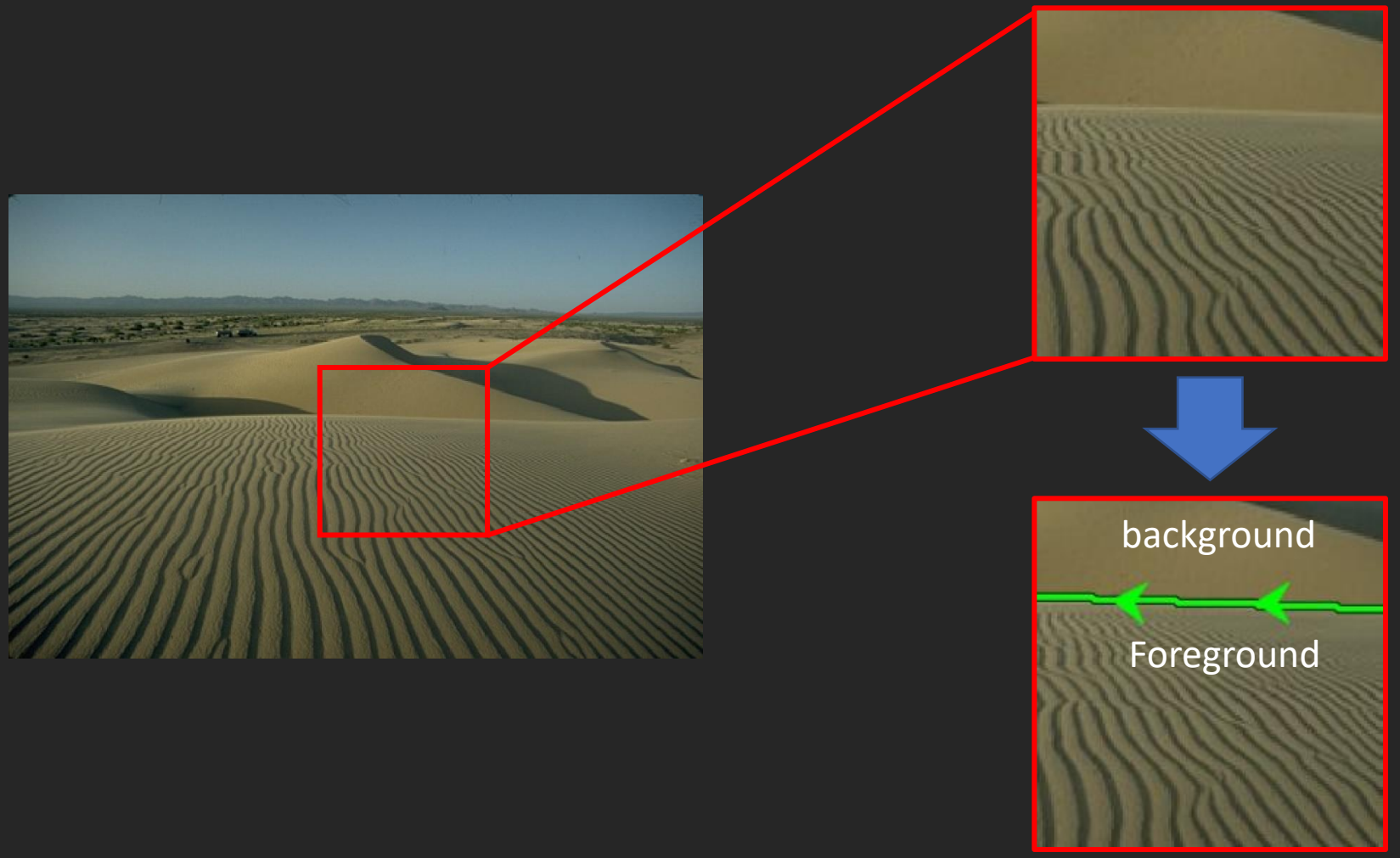


Geometric occlusion boundaries

Depth map

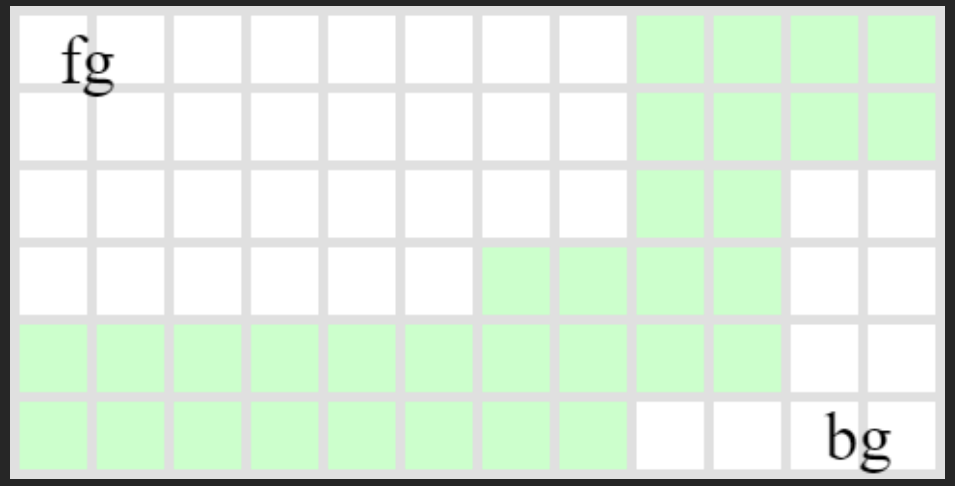
Occlusion boundary with directions

For example, oriented occlusion boundary estimation

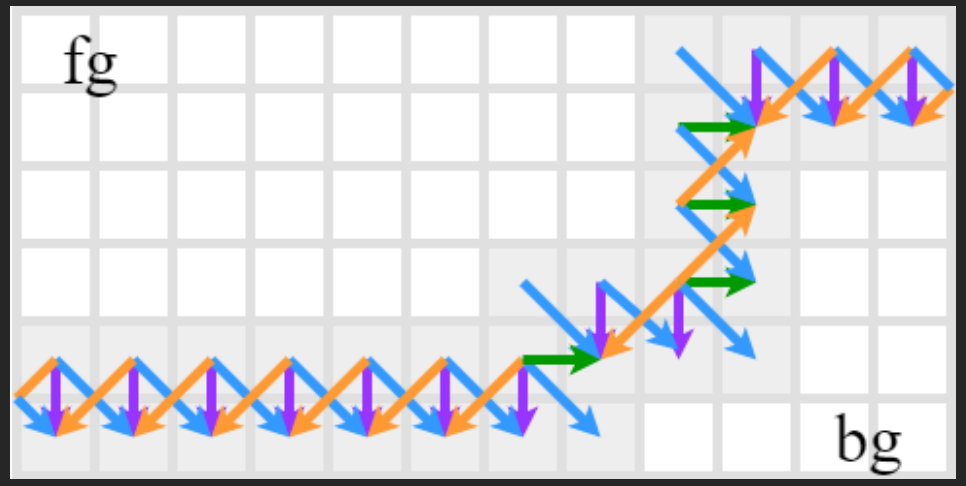


Key idea: classifying pairs of pixels

Pixel-wise occlusion boundary detection

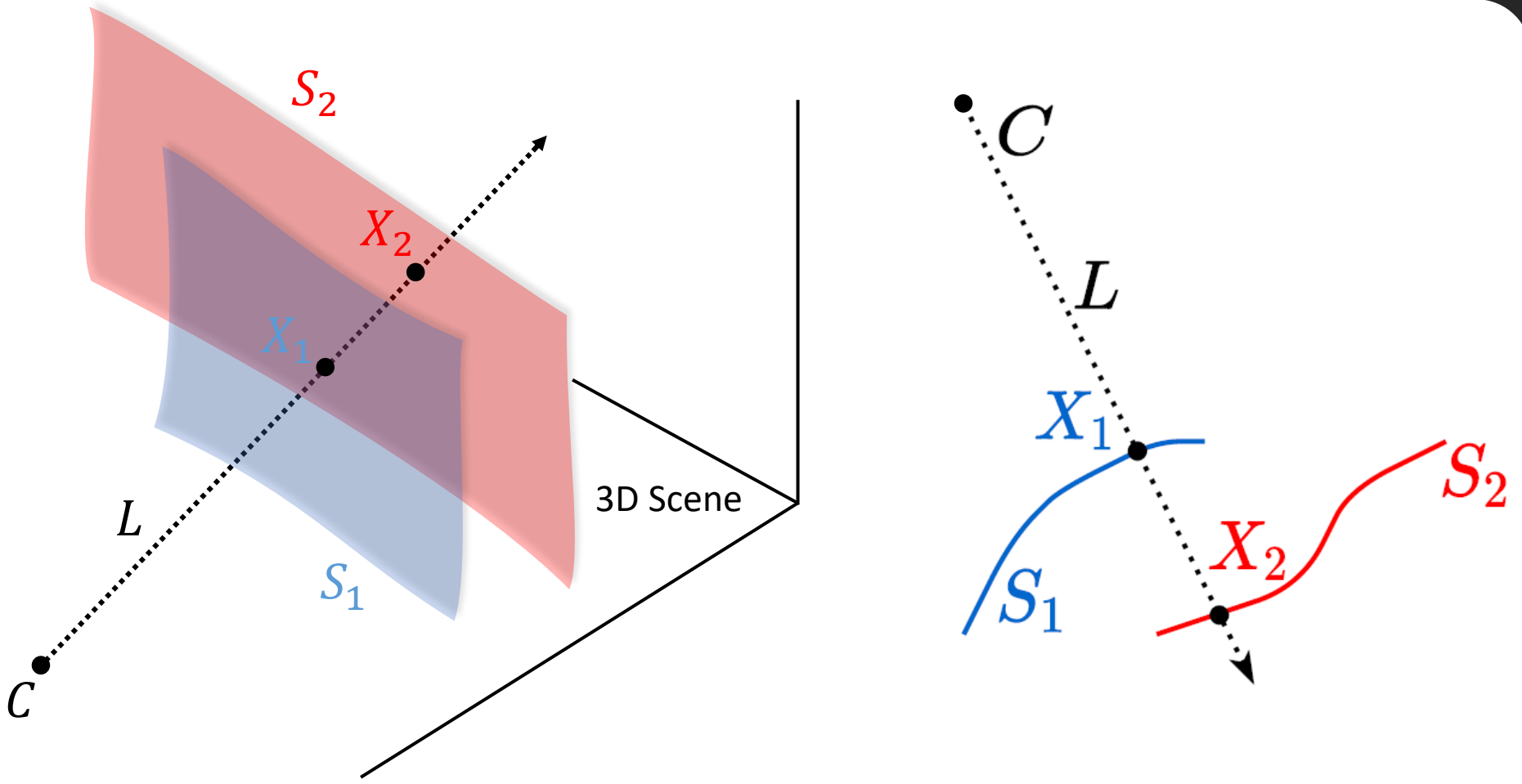


Classifying the occlusion status of neighbor pixel pairs



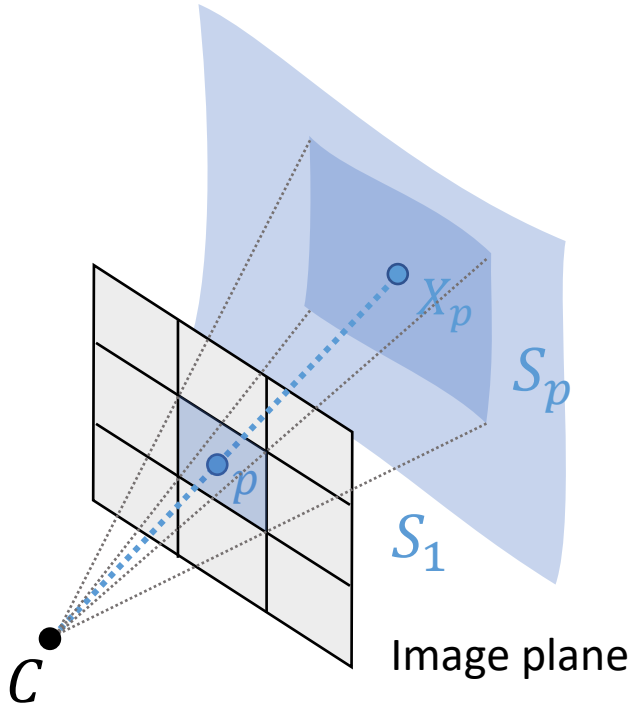
Unfortunately, there is no such type of pairwise occlusion annotations …

Geometric occlusion in a single image

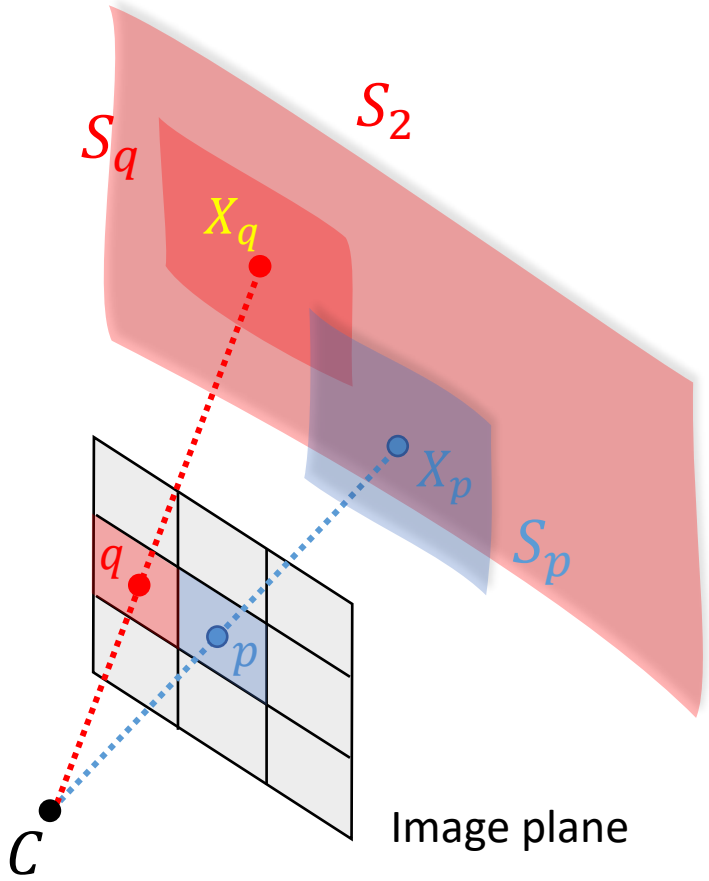


X_1, X_2 3D points S_1, S_2 3D surfaces C Camera center

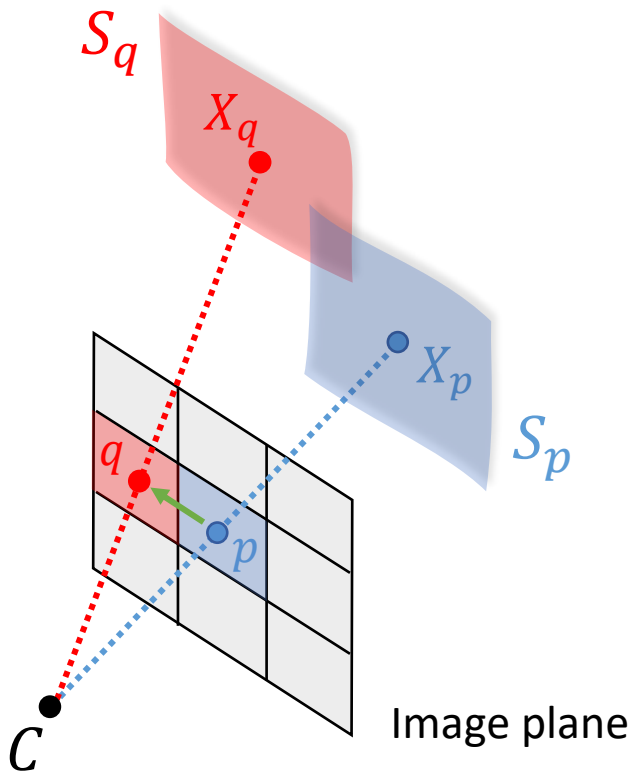
Geometric occlusion in a single image

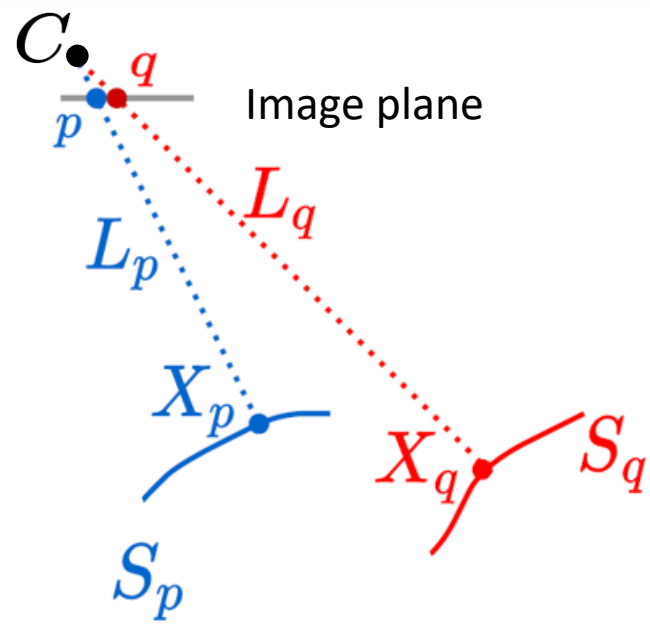


Geometric occlusion in a single image



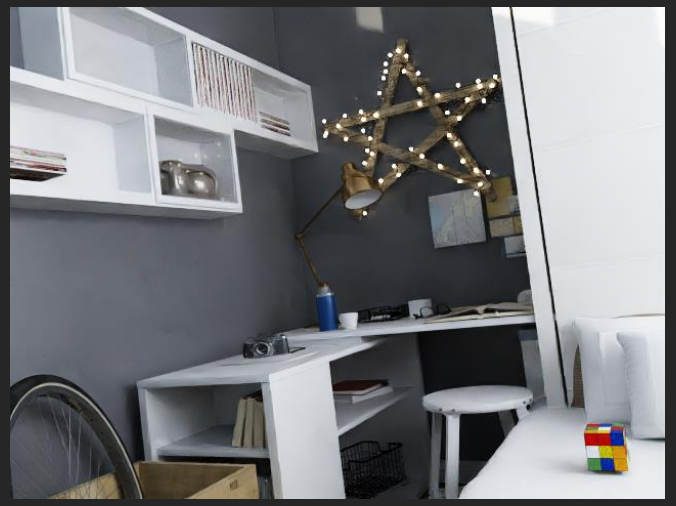
Geometric occlusion in a single image



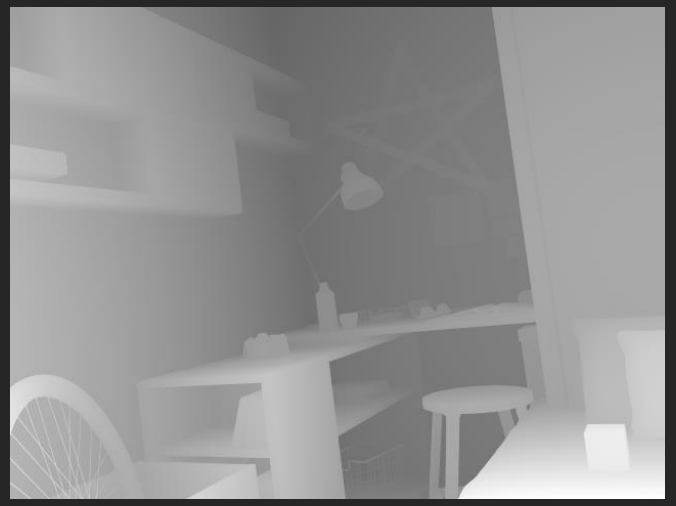


What we have

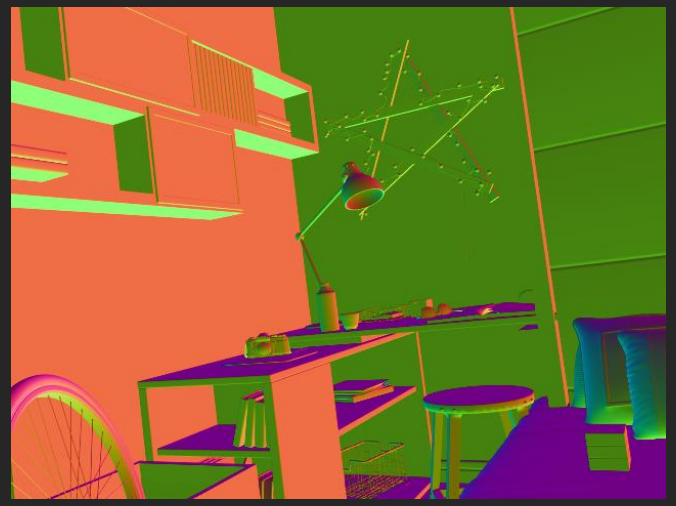
- A typical RGBD dataset



RGB image

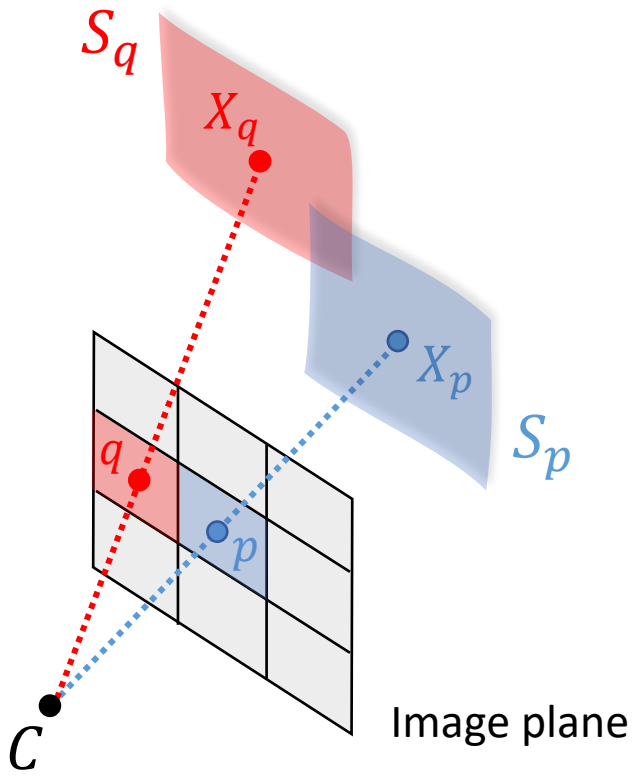


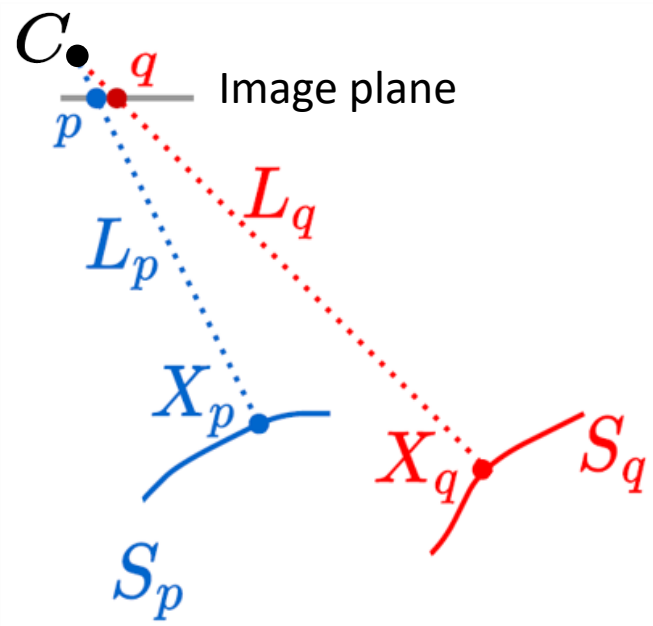
Depth map



Surface normal map

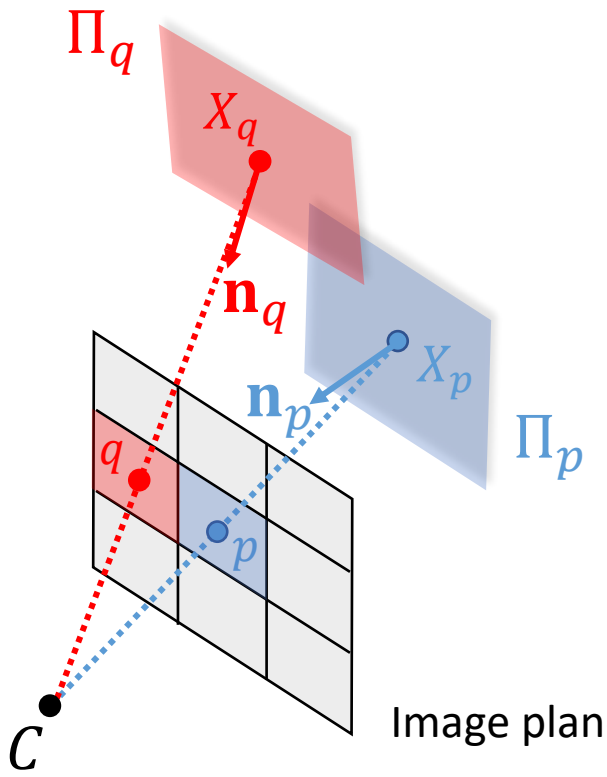
Ground truth generation

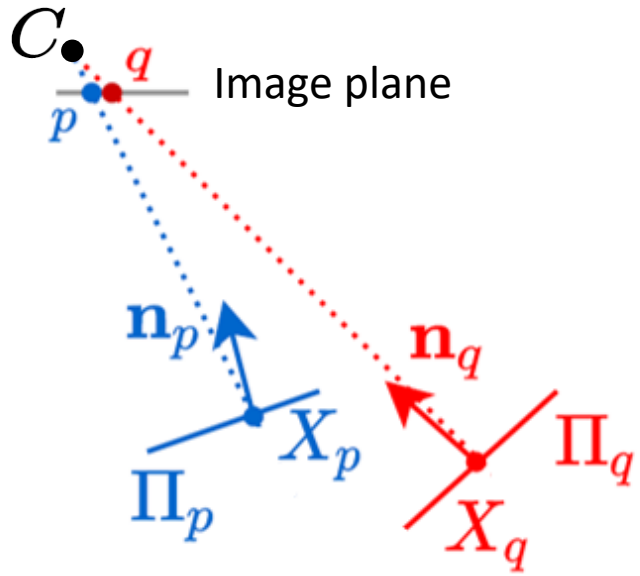




Ground truth generation

- Order-1 approximation of 3D surfaces

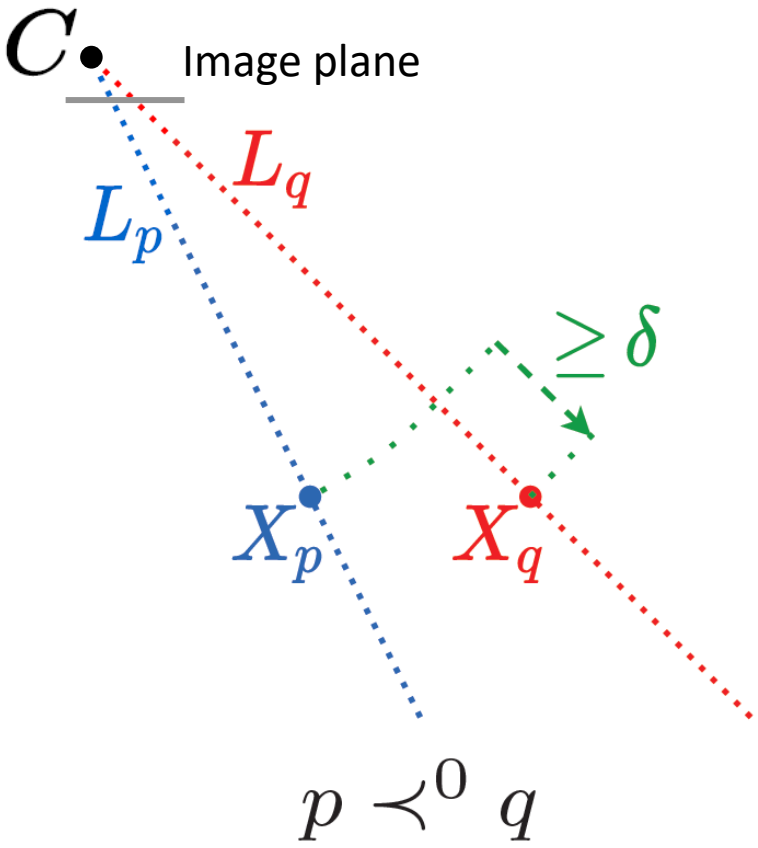




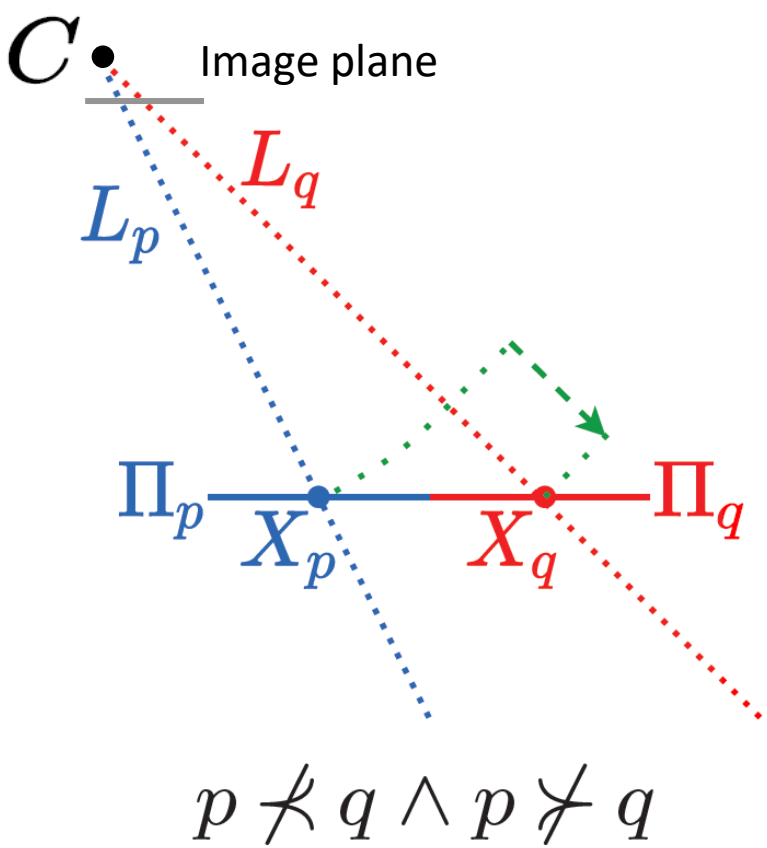
Π_p, Π_q Tangent planes $\mathbf{n}_p, \mathbf{n}_q$ Surface normals

Ground truth generation

- Order-0 occlusion

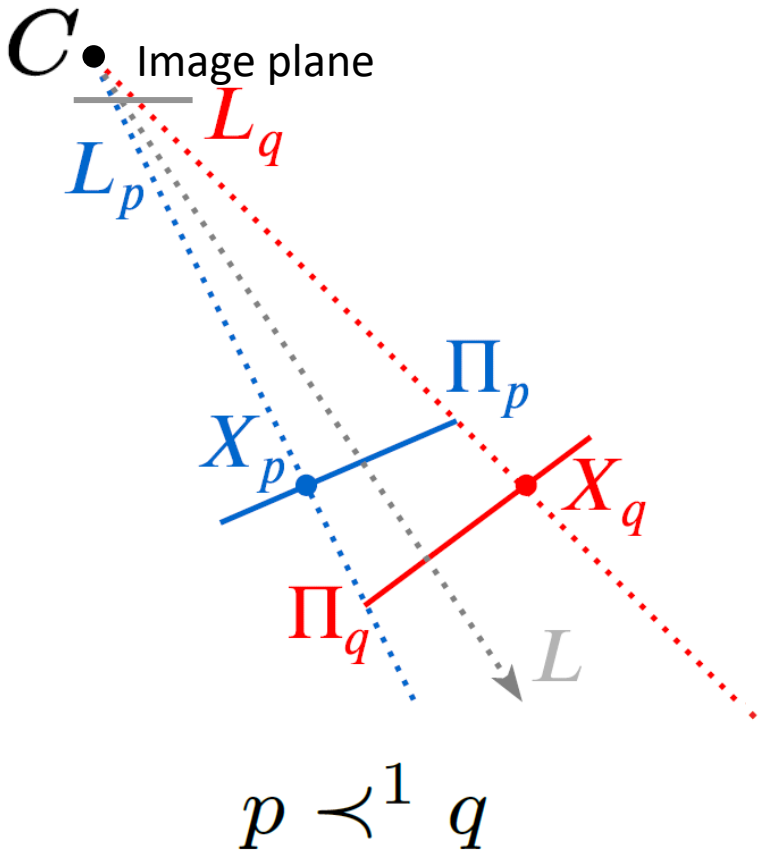


- Order-0 wrong occlusion condition

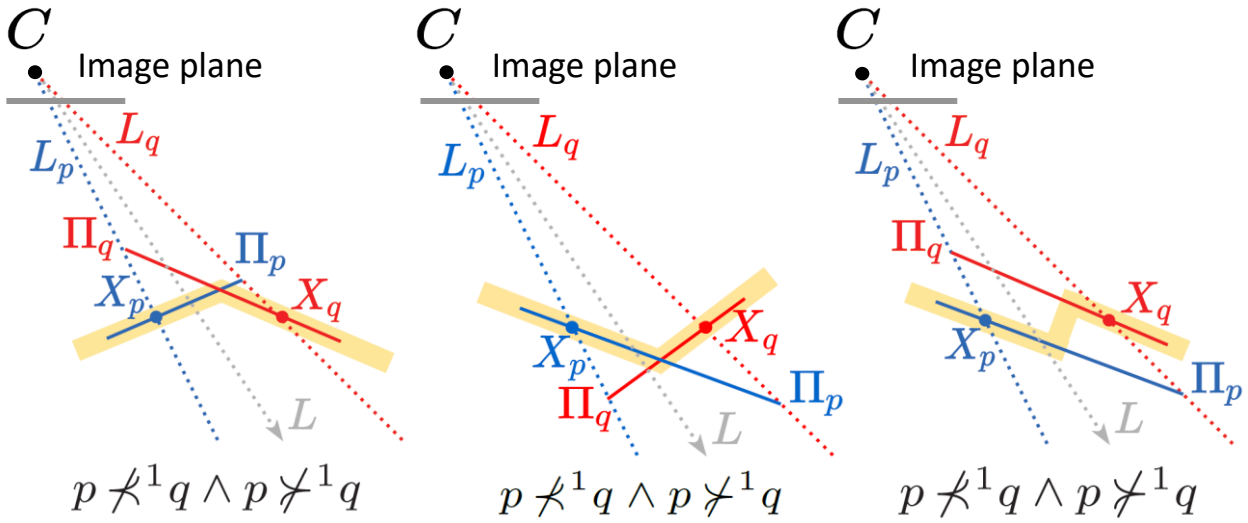


Ground truth generation

- Order-1 occlusion



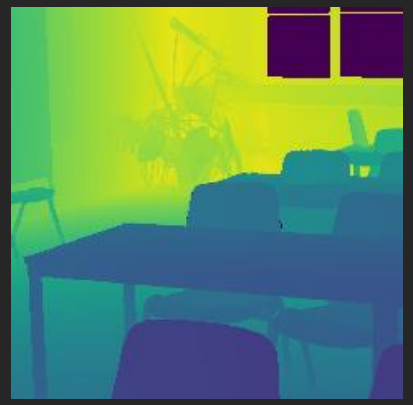
- Order-1 wrong occlusion conditions



Real 3D surface

Generated annotations

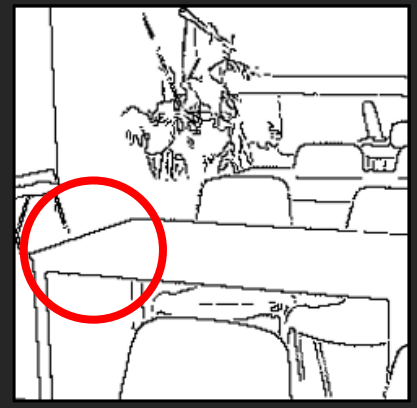
Image



Depth map

Surface normal map

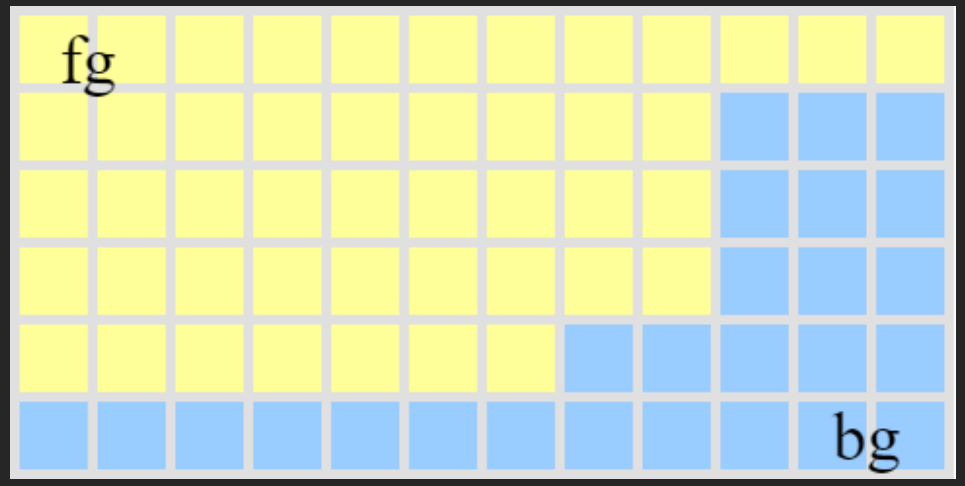
0-order occlusion boundary



Ours

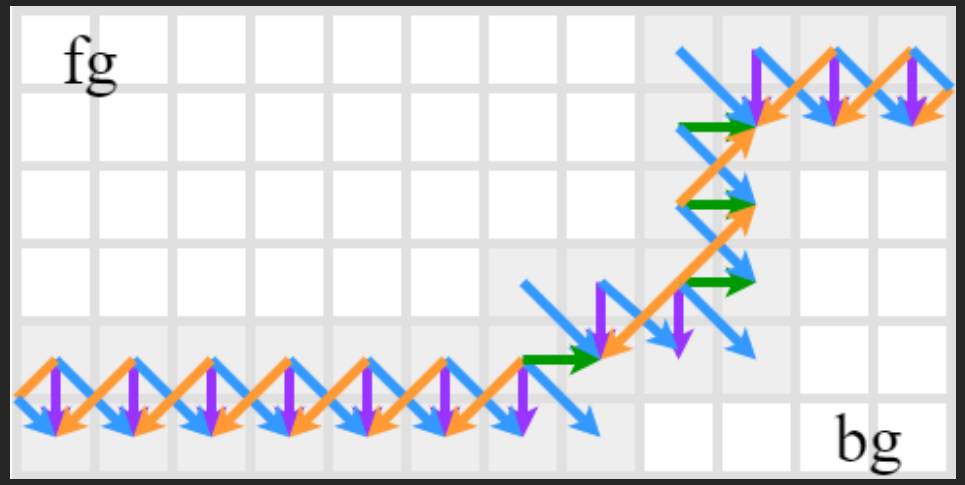
Pixel-Pair occlusion relationship map

Foreground/background mask



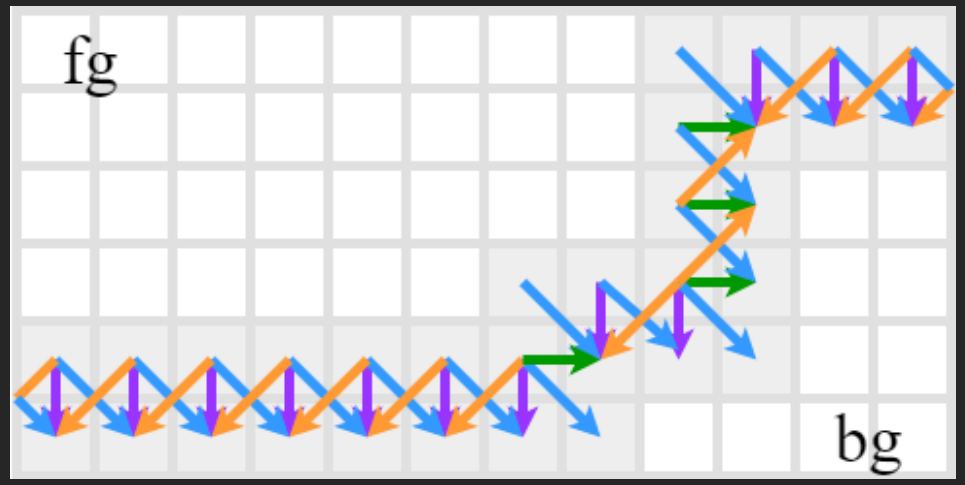
Pixel-Pair occlusion relationship map

Pixel-Pair Occlusion Relationship Map (P2ORM)

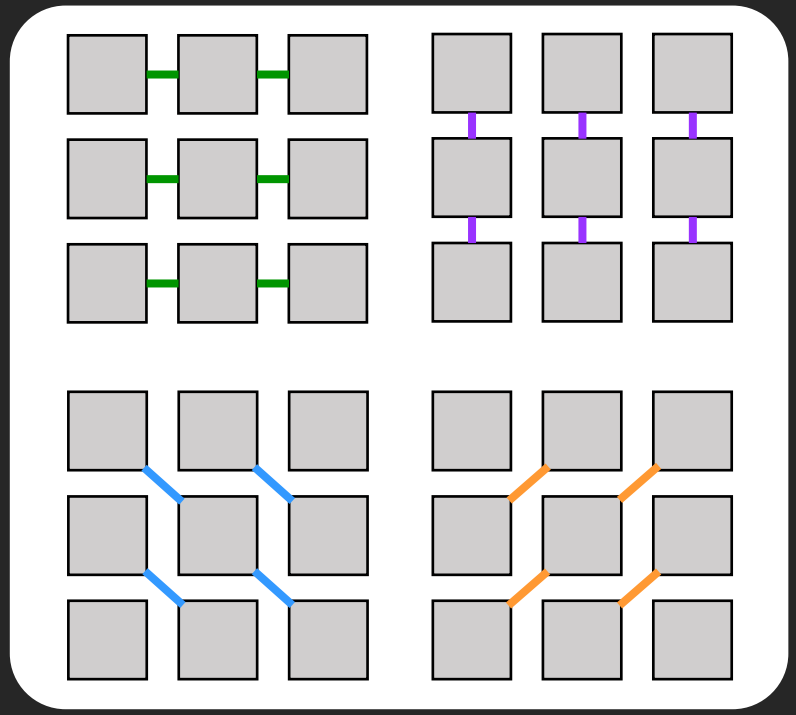


Pixel-Pair occlusion relationship map

Pixel-Pair Occlusion Relationship Map (P2ORM)

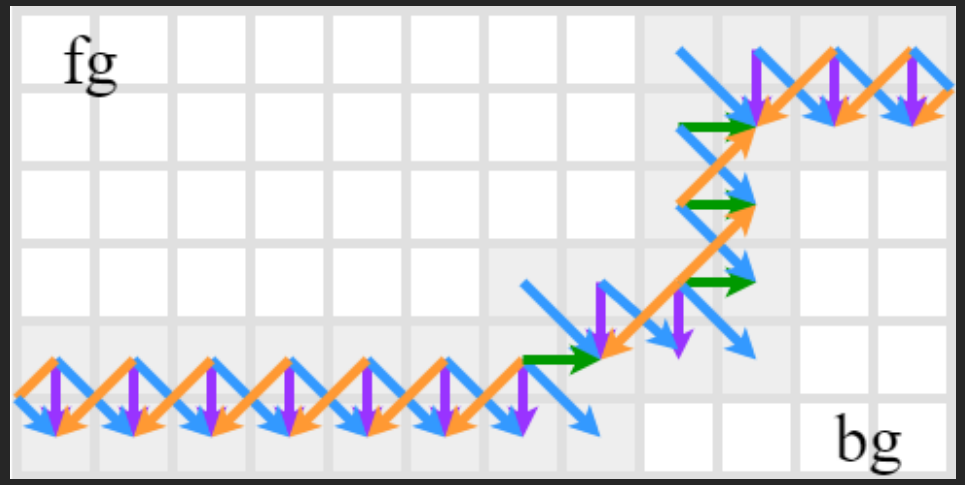


Four inclinations in 8-connectivity

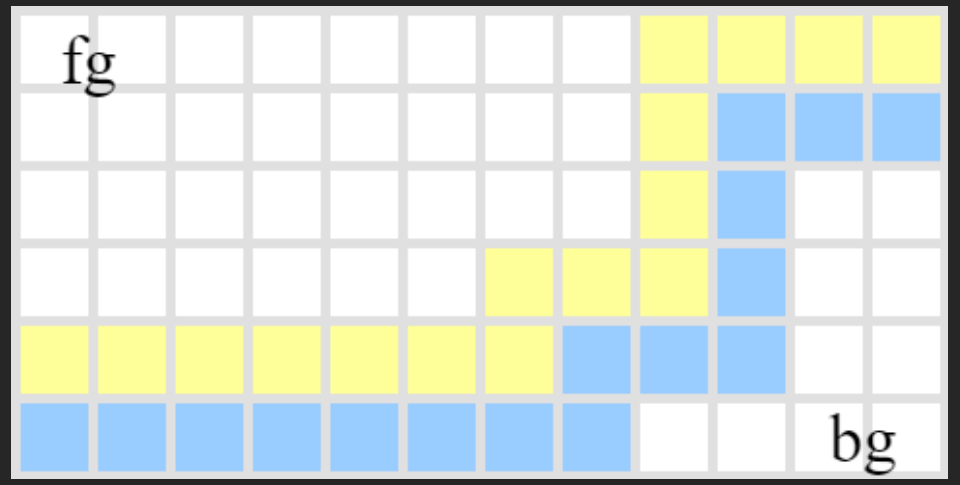


Pixel-Pair occlusion relationship map

Pixel-Pair Occlusion Relationship Map (P2ORM)



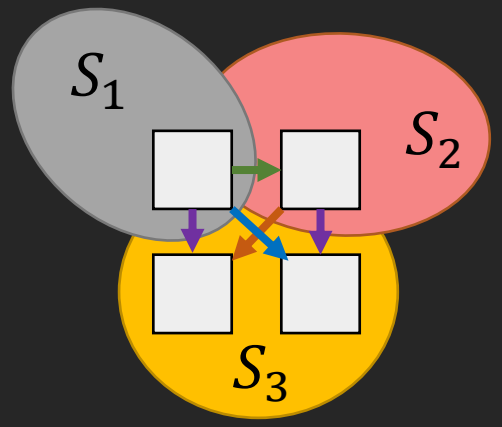
Figure/Ground notion [1]



[1] Ren et al. ECCV 2006

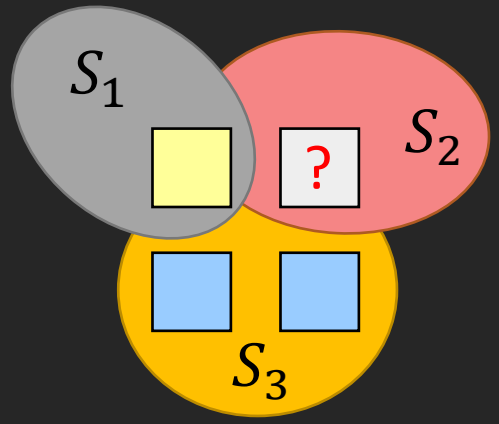
Pixel-Pair occlusion relationship map

Pixel-Pair Occlusion Relationship Map (P2ORM)



$S_1\ occludes\ S_2, S_3$
$S_2\ occludes\ S_3$

Figure/Ground notion

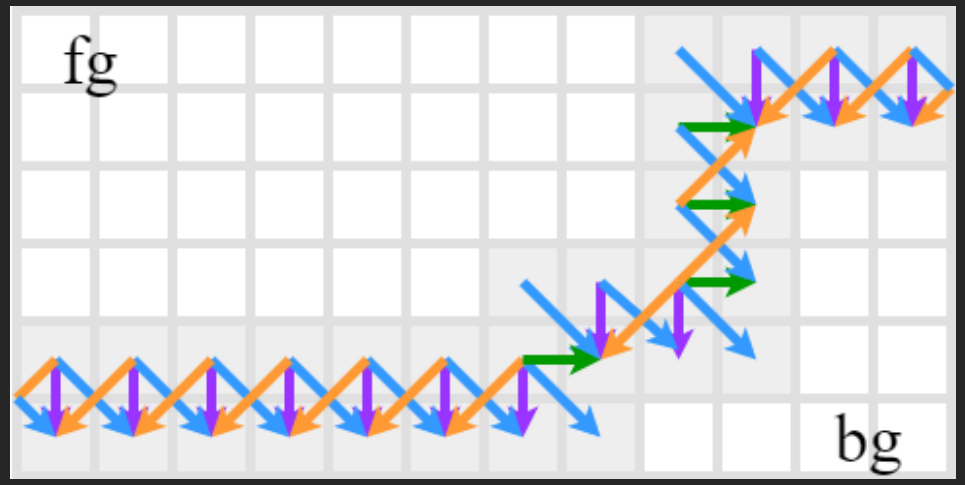


$S_1\ occludes\ S_2, S_3$
$S_2\ occludes\ S_3$

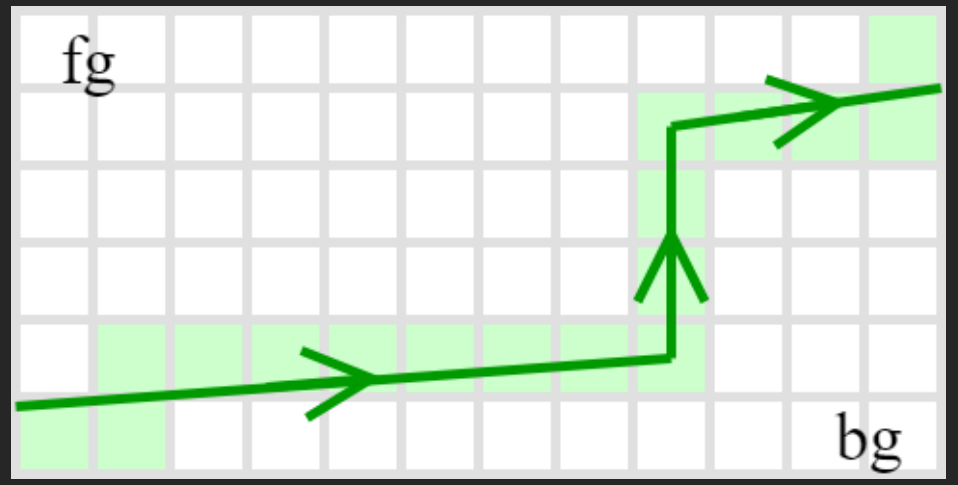
* S_1, S_2, S_3 are scene 3D surfaces.

Pixel-Pair occlusion relationship map

Pixel-Pair Occlusion Relationship Map (P2ORM)



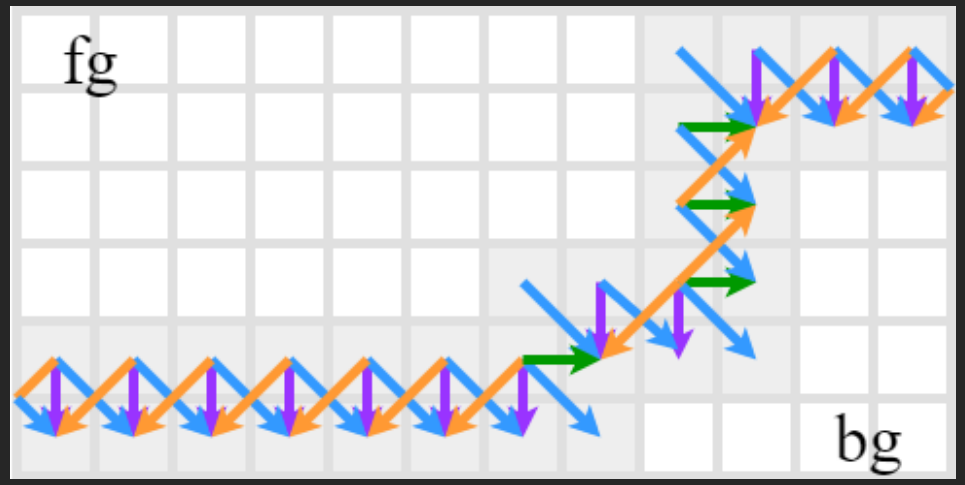
Oriented occlusion boundary notion [1]



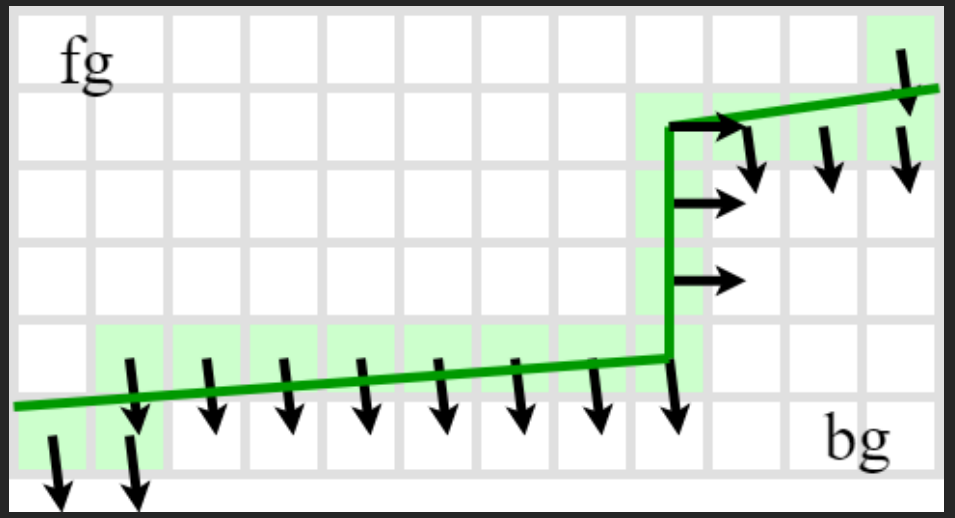
[1] Teo et al. CVPR 2015

Pixel-Pair occlusion relationship map

Pixel-Pair Occlusion Relationship Map (P2ORM)



Oriented occlusion boundary notion



Modeling P2ORM

- Pixel-pair occlusion status label

$$\omega_{p,q} = r = \begin{cases} 1 \\ 0 \\ -1 \end{cases}$$

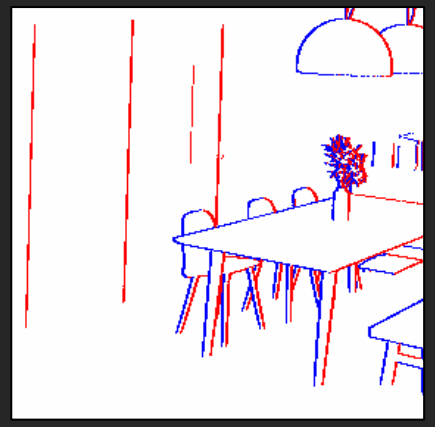
p occludes *q*
no occlusion between *p,q*
p is occluded by *q*

- Four inclinations for pixel pairs

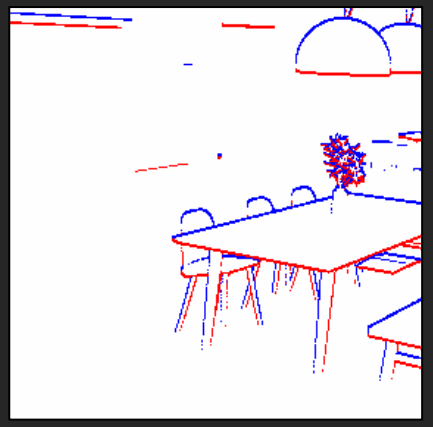
Horizontal Vertical Diagonal Antidiagonal

Modeling P2ORM

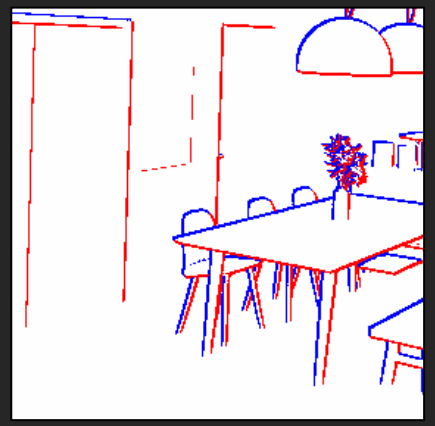
Horizontal



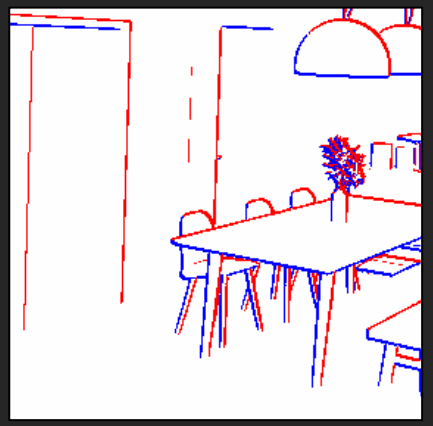
Vertical



Diagonal



Antidiagonal



p occludes q

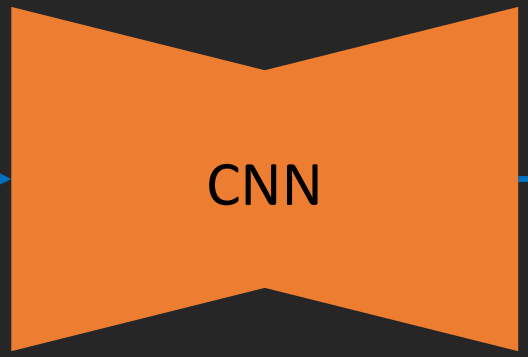
q occludes p

no occlusion

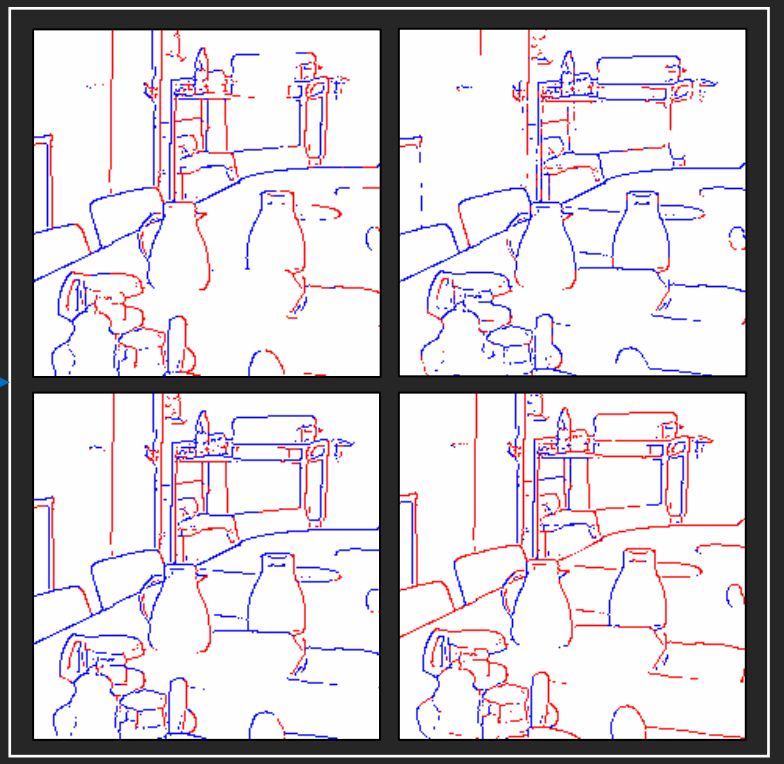
Estimating P2ORM: a classification task

Input image

Standard encoder/decoder



Pairwise occlusion relationship maps



Loss: cross-entropy

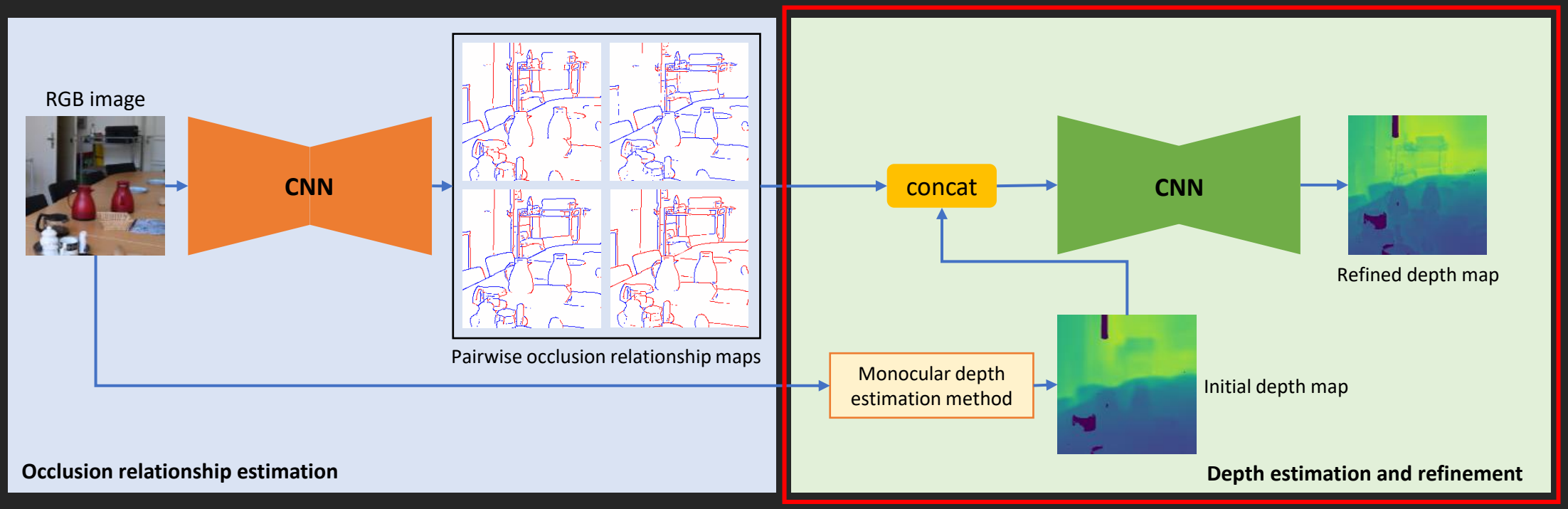
Quantitative results on oriented occlusion boundary estimation

Oriented occlusion boundary estimation. *Our re-implementation.

Method	BSDS ownership			NYUv2-OR			iBims-1-OR		
Metric	ODS	OIS	AP	ODS	OIS	AP	ODS	OIS	AP
SRF-OCC [1]	.419	.448	.337	-	-	-	-	-	-
DOC-DMLFOV [2]	.463	.491	.369	-	-	-	-	-	-
DOC-HED [2]	.522	.545	.428	-	-	-	-	-	-
DOOBNet [3]	.555	.570	.440	-	-	-	-	-	-
OFNet [4]	.583	.607	.501	-	-	-	-	-	-
DOOBNet*	.529	.543	.433	.343	.370	.263	.421	.440	312
OFNet*	.553	.577	.520	.402	.431	.342	.488	.513	.432
baseline	.571	.605	.524	.396	.428	.343	.482	.507	.431
ours (4-connectivity)	.590	.612	.512	.500	.522	.477	.575	.599	.508
ours (8-connectivity)	**.607**	**.632**	**.598**	**.520**	**.540**	**.497**	**.581**	**.603**	**.525**

[1] Teo et al. CVPR 2015, [2] Wang et al. ECCV 2016, [3] Wang et al. ACCV 2018, [4] Lu et al. ICCV 2019

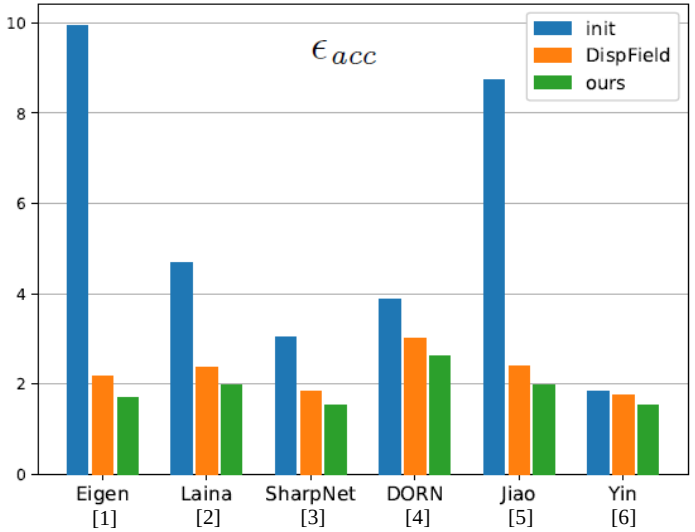
P2ORM: Application in depth map refinement

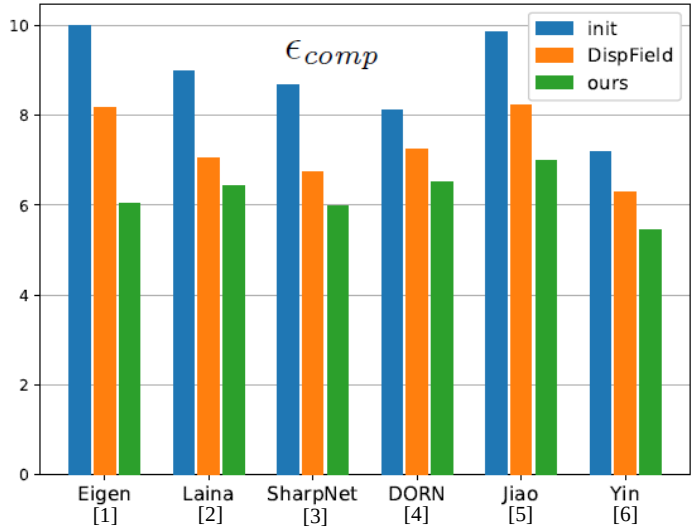


Loss: $L = L_{occonsist} + L_{regul}$

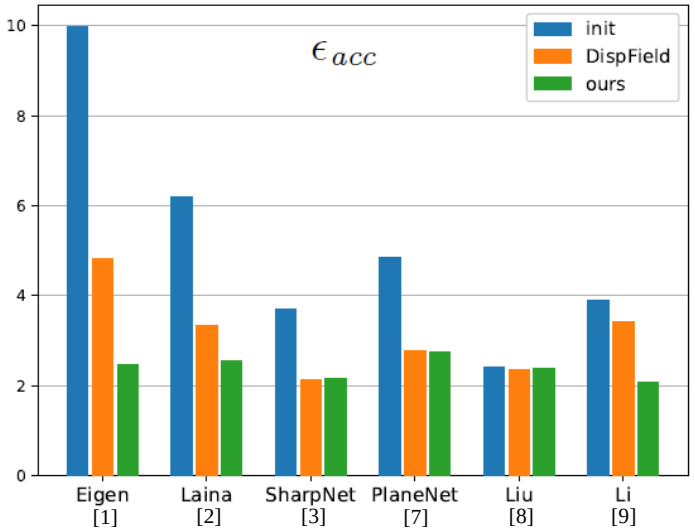
Quantitative results on depth map refinement

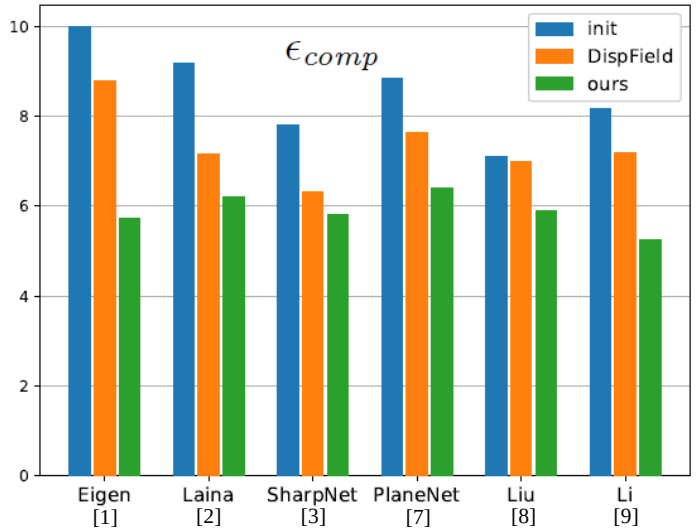
NYUv2





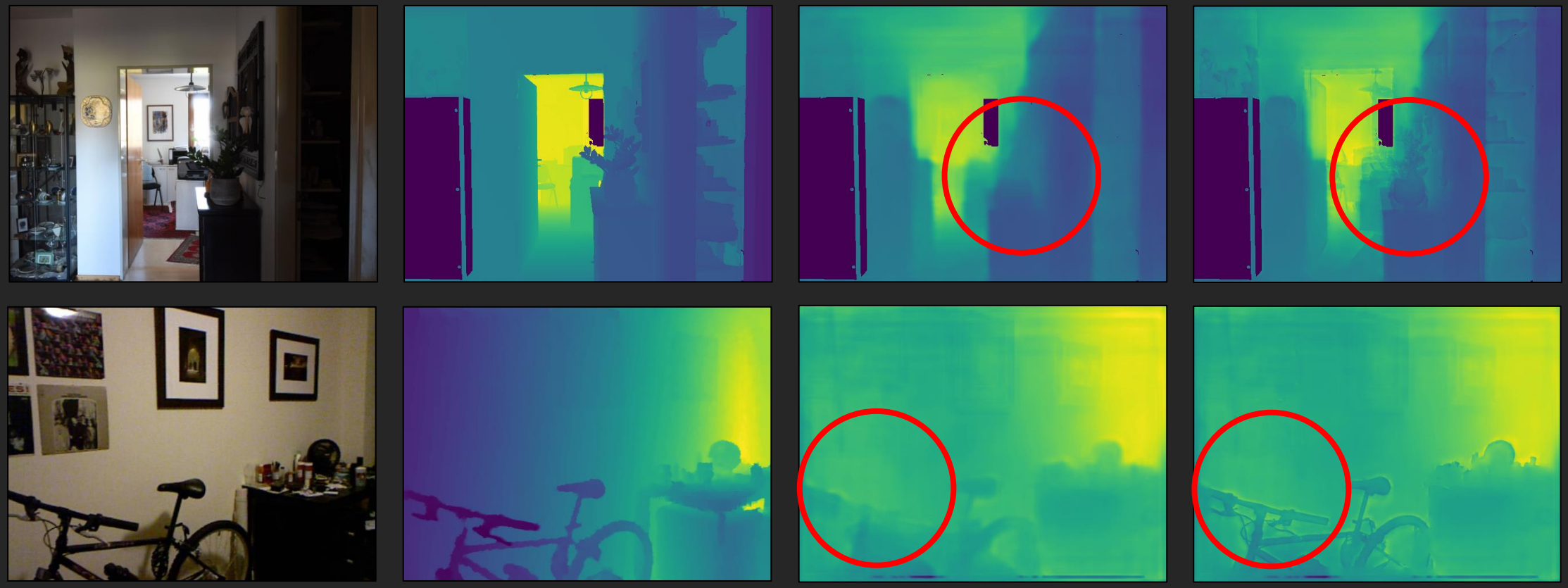
iBims-1





[1] Eigen et al. NeurIPS 2014
[2] Laina et al. 3DV 2016
[3] Ramamonjisoa and Lepetit ICCV Workshops 2019
[4] Fu et al. CVPR 2018
[5] Jiao et al. ECCV 2018
[6] Yin et al. ICCV 2019
[7] Liu et al. CVPR 2018
[8] Liu et al. CVPR 2015
[9] Li et al. ICCV 2016
DispField [Ramamonjisoa et al. CVPR 2020]

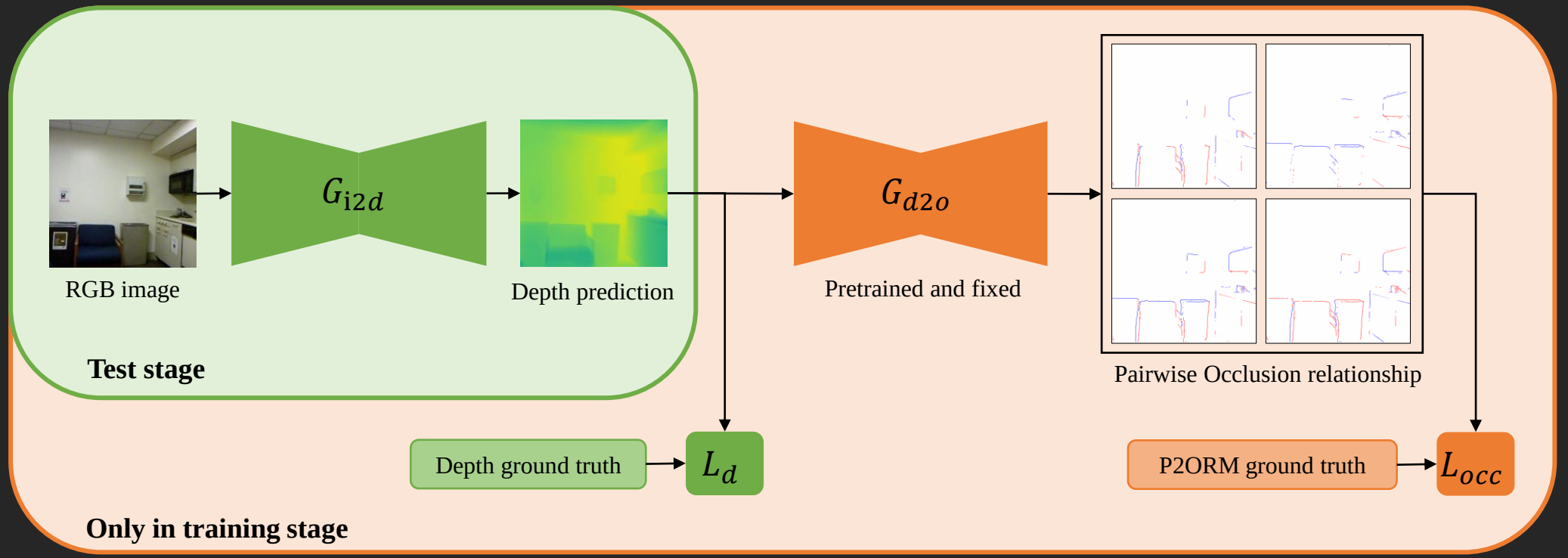
Qualitative results on depth map refinement



Input image | Ground truth | SharpNet prediction [1] | Our refined depth

[1] Ramamonjisoa, M., *et al.*: Sharpnet: Fast and accurate recovery of occluding contours in monocular depth estimation. In: ICCV Workshops, 2019

Single-stage monocular depth estimator using P2ORM



Loss for backpropagation: $L = L_d + L_{occ}$

Quantitative results on monocular depth estimation

Method	Boundaries(↓) ϵ_{acc}	Boundaries(↓) ϵ_{comp}	Depth Error(↓) rel	Depth Error(↓) $\log_{10}$	Depth Error(↓) RMS_{lin}	Depth Accuracy(↑) σ_1	Depth Accuracy(↑) σ_2	Depth Accuracy(↑) σ_3
Baseline	2.171	6.387	0.116	0.048	0.526	0.888	0.980	0.993
Ours	**1.830**	**5.965**	0.110	0.044	0.492	0.891	0.981	0.993

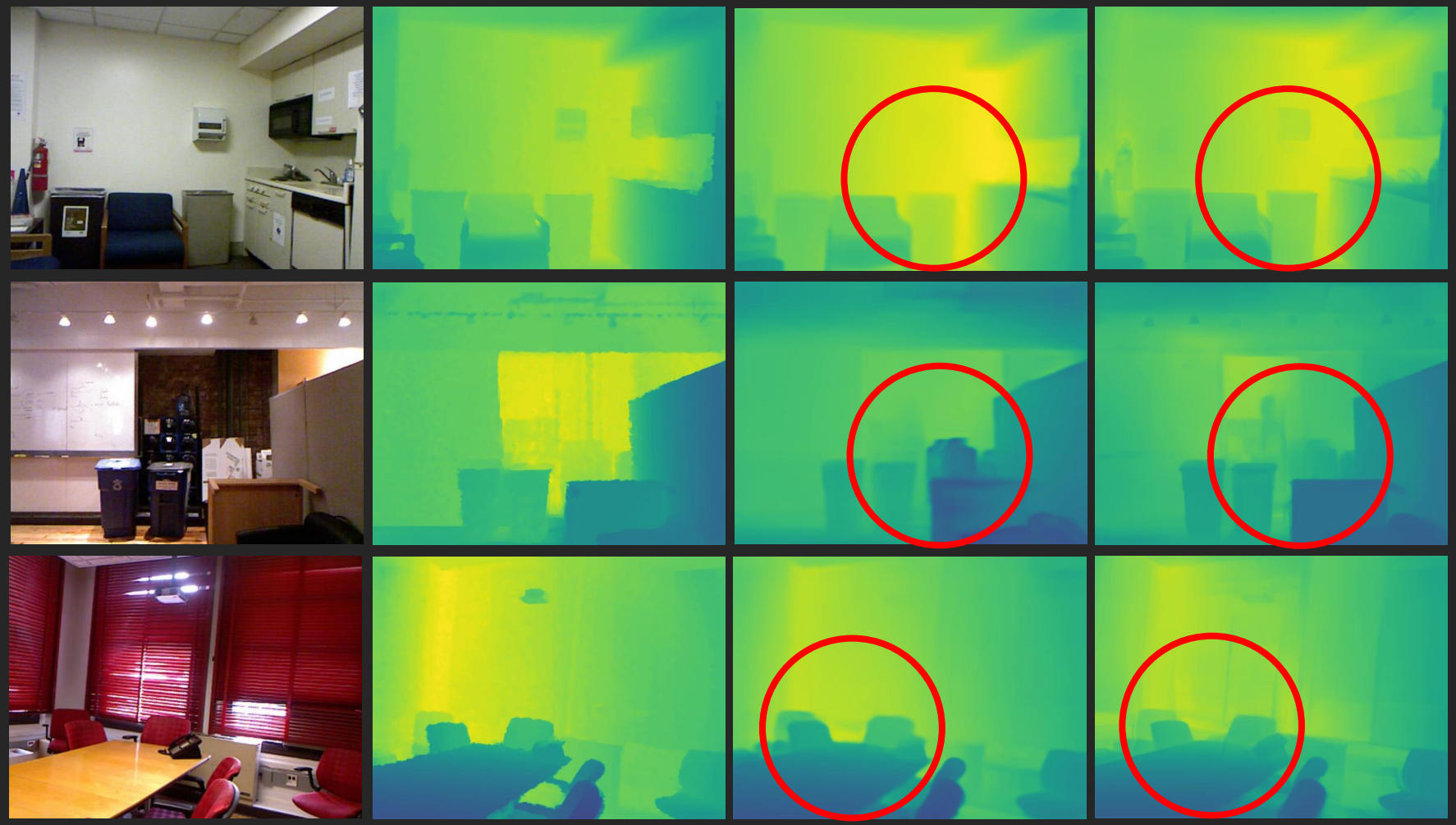
Comparison with our baseline method for monocular depth estimation on SceneNet [1]. Best depth boundaries results in **bold**.

Depth estimation method	Boundaries(↓) ϵ_{acc}	Boundaries(↓) ϵ_{comp}	Depth Error(↓) rel	Depth Error(↓) $\log_{10}$	Depth Error(↓) RMS_{lin}	Depth Accuracy(↑) σ_1	Depth Accuracy(↑) σ_2	Depth Accuracy(↑) σ_3
Eigen et al. [2]	9.926	9.993	0.236	0.095	0.765	0.611	0.887	0.971
Laina et al. [3]	4.702	8.982	0.142	0.059	0.510	0.818	0.955	0.988
Fu et al. [4]	3.872	8.117	0.131	0.053	0.493	0.848	0.956	0.984
Ramamonjisoa and Lepetit [5]	3.041	8.692	0.116	0.053	0.448	0.853	0.970	0.993
Yin et al. [6]	1.854	7.188	0.112	0.047	0.417	0.880	0.975	0.994
Ours	**1.398**	**6.414**	0.130	0.056	0.483	0.834	0.966	0.992

Evaluation of monocular depth estimation on NYUv2 [7], cropped within valid regions as in [2]. Best depth boundaries results in **bold**.

[1] McCormac et al. ICCV 2017, [2] Eigen et al. NeurIPS 2014, [3] Laina et al. 3DV 2016, [4] Fu et al. CVPR 2018,
[5] Ramamonjisoa and Lepetit ICCV Workshops 2019, [6] Yin et al. ICCV 2019, [7] Silberman et al. ECCV 2012,

Qualitative results on monocular depth estimation



Input image | Ground truth | Prediction by Yin et al.[1] | Ours

[1] Yin, W., *et al.*: Enforcing geometric constraints of virtual normal for depth estimation. In: ICCV, 2019

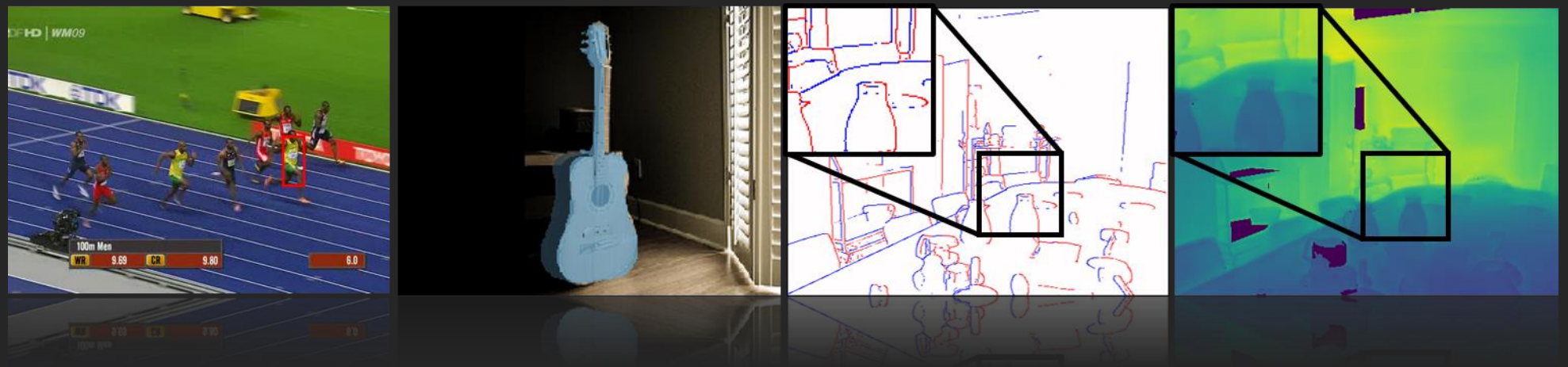
Conclusion

1. A new occlusion relationship representation based on pixel pairs
2. A depth map refinement method
3. A monocular depth map estimation method

Code and datasets: https://github.com/tim885/P2ORM

Summary

1. Single object tracking with structural semantics
2. Object pose estimation with object shapes
3. Scene occlusion relationship and depth estimation



École des Ponts
ParisTech

Imagine
Research Group


École des Ponts
ParisTech
Imagine
Research Group
Many thanks to all of you!

Thank You