Instance Pose Estimation from RGB: A New State-of-the-Art

"Uncertainty-Driven 6D Pose Estimation of Objects and Scenes from a Single RGB Image"

Eric Brachmann, Frank Michel, Alexander Krull, Michael Ying Yang, Stefan Gumhold, Carsten Rother

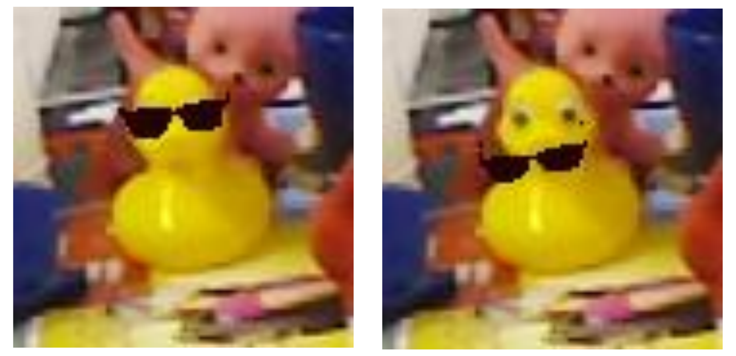
COMPUTER VISION LAB DRESDEN

Computergraphik und Visualisierung

Motivation:

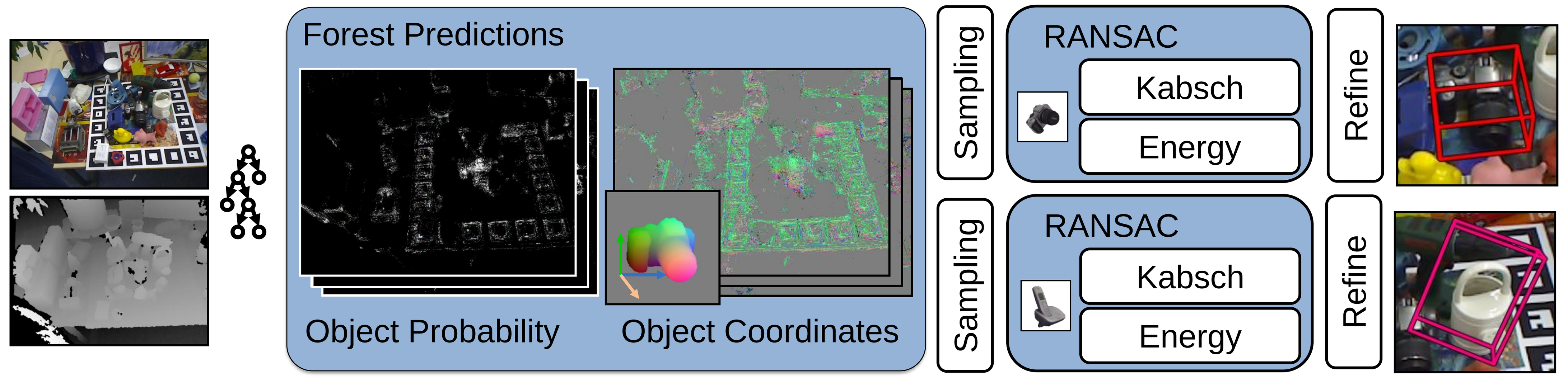
Object instance pose estimation and **camera localization** from RGBD images has been very successful.

Goal: To achieve visually convincing results using **RGB only**.



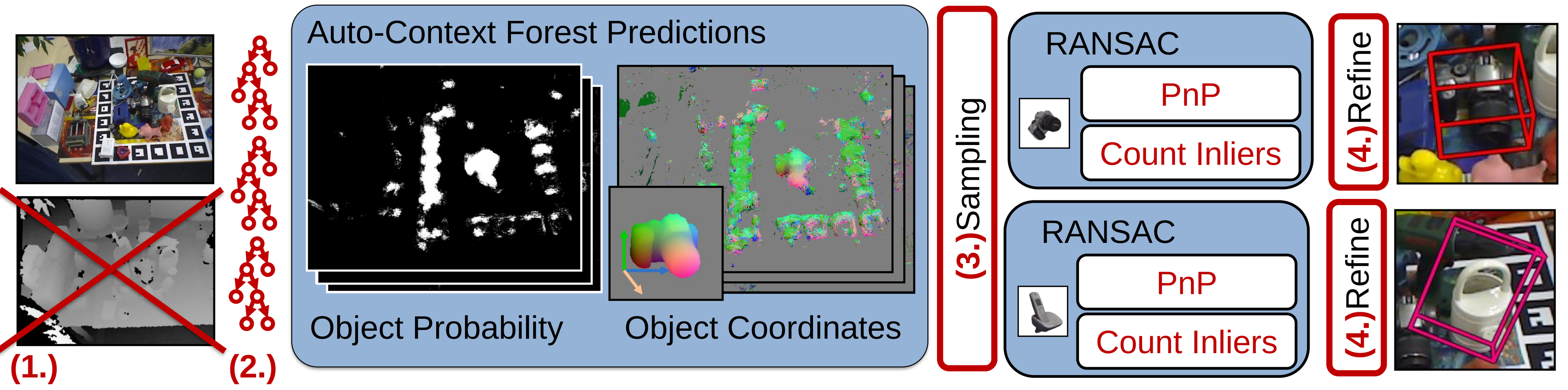
Ours Line2D [4]

State-of-the-Art using RGBD (previous work [1],[2]):



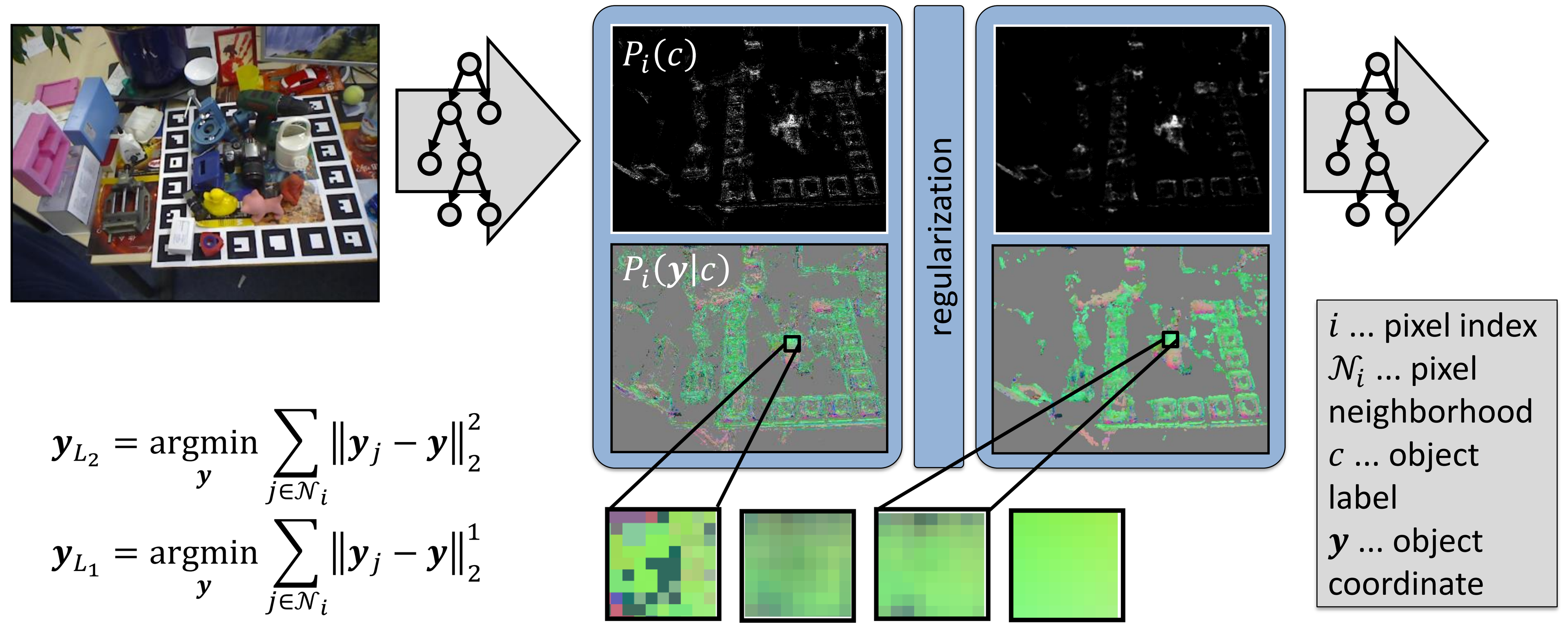
Contributions:

1. New **state-of-the-art** for 6D pose estimation from **RGB only**
2. Robust **auto-context** framework **for multi-dimensional continuous data**
3. Scalable RANSAC for **multiple objects**
4. Refinement using **uncertainty**
5. High **flexibility** (objects / scenes, RGB / RGB-D)



(2.) Object Coordinate Auto-Context:

The (auto-context) random forest predicts **object label distributions** $P_i(c)$ and **object coordinate distributions** $P_i(\boldsymbol{y}|c)$. We regularize predictions before each auto-context layer.



$$\boldsymbol{y}_{L_2} = \underset{\boldsymbol{y}}{\operatorname{argmin}} \sum_{j\in\mathcal{N}_i} \|\boldsymbol{y}_j - \boldsymbol{y}\|_2^2$$

$$\boldsymbol{y}_{L_1} = \underset{\boldsymbol{y}}{\operatorname{argmin}} \sum_{j\in\mathcal{N}_i} \|\boldsymbol{y}_j - \boldsymbol{y}\|_2^1$$

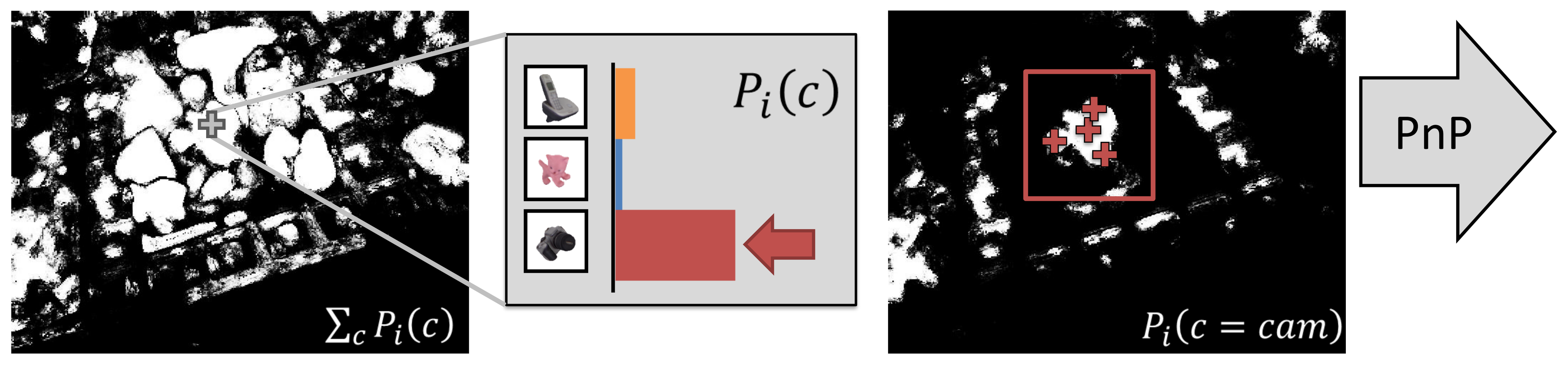
i ... pixel index
$\mathcal{N}_i$ ... pixel neighborhood
c ... object label
$\boldsymbol{y}$ ... object coordinate

We use **fast pixel difference features** on the RGB image and the regularized predictions of the previous forest in the auto-context stack.

(3.) Multi-Object RANSAC:

When sampling a pose hypothesis H_c, we decide on the fly which object c it belongs to. We avoid sampling hypotheses for objects not visible.

Sampling one hypothesis:



Draw first pixel i_1 Draw object c Draw pixel i_1, i_2, i_3 Calculate H_c

Distribution after sampling n_H hypotheses:

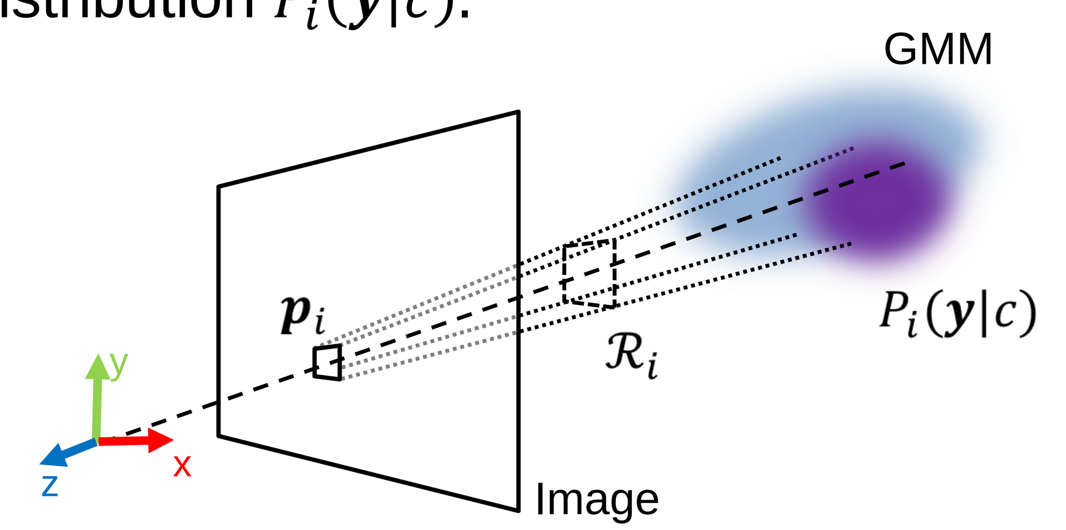
Pre-emptive optimization and refinement per object.

(4.) Refinement Using Uncertainty:

We exploit the full object coordinate distribution $P_i(\boldsymbol{y}|c)$.

$$H^* = \underset{H}{\operatorname{argmax}} \sum_i \log P_i(H^{-1}\boldsymbol{e}_i|c)$$

Camera coordinate $\boldsymbol{e}_i$ is unknown, hence we integrate over the pixel volume $\mathcal{R}_i$.



$$H^* = \underset{H}{\operatorname{argmax}} \sum_i \log \iiint_{\mathcal{R}_i} P_i\left(H^{-1}\begin{pmatrix}x\\y\\z\end{pmatrix}\middle|c\right) dxdydz$$

$$H^* \approx \underset{H}{\operatorname{argmax}} \sum_i \log \sum_{(m,\boldsymbol{\mu},\Sigma)\in\mathcal{M}_i} m \int_{-\infty}^{\infty} z^2 \mathcal{N}\left(\begin{pmatrix}0\\0\\z\end{pmatrix}; \boldsymbol{\mu}_e, \Sigma_e\right) dz$$

$\boldsymbol{p}_i$ ... pixel position
$\mathcal{M}_i$ ... GMM component
m ... component weight
$\boldsymbol{\mu}$ ... mean
Σ ... covariance
$\square_e$... in camera space

Experimental Results

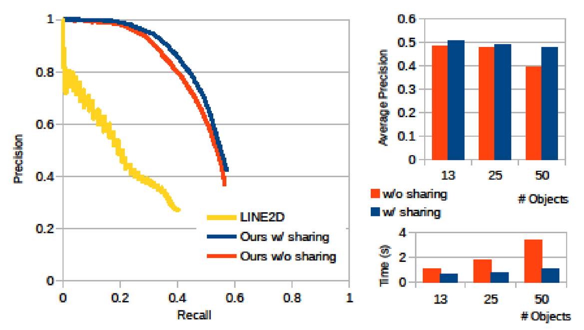
Single Object Pose Estimation from RGB images (dataset of [3]):

	Ours, no AC	Ours, no reg.	Ours, L_2 reg.	Ours, L_1 reg., no ref.	Ours, L_1 reg.	Line2D [4]
3D	27.3%	19.6%	46.0%	32.3%	**50.2%**	24.2%
2D	59.3%	38.0%	68.6%	69.5%	**73.7%**	20.9%

and from RGBD images:

	Brachmann et al. [1]	Krull et al. [2]	Ours, L_1 reg.
3D	97.4%	93.9%	**99.0%**
2D	81.7%	82.6%	**95.7%**

Multi-Object Detection (RGB) (dataset of [1]):



Camera localization (RGB) (dataset of [5]):

	5cm 5°	Avg. Med. Err.
Sparse RGB [5]	40.7%	-
PoseNet [6]	-	46.9cm, 5.4°
Ours	**55.2%**	**6.1cm, 2.7°**

Measure of [3] (3D):

$$\frac{1}{|\mathcal{V}|}\sum_{\boldsymbol{v}\in\mathcal{V}} \|H\boldsymbol{v} - \tilde{H}\boldsymbol{v}\| < 0.1d$$

Our measure (2D):

$$\frac{1}{|\mathcal{V}|}\sum_{\boldsymbol{v}\in\mathcal{V}} \|KH\boldsymbol{v} - K\tilde{H}\boldsymbol{v}\| < 5\text{px}$$

[1] E. Brachmann, A. Krull, F. Michel, S. Gumhold, J. Shotton, and C. Rother, "Learning 6D Object Pose Estimation using 3D Object Coordinates", ECCV 2014.
[2] A. Krull, E. Brachmann, F. Michel, M. Y. Yang, S. Gumhold, C. Rother, "Learning Analysis-by-Synthesis for 6D Pose Estimation in RGB-D Images", ICCV 2015.
[3] S. Hinterstoisser, V. Lepetit, S. Ilic, S. Holzer, G. Bradski, K. Konolige, N. Navab, "Model Based Training, Detection and Pose Estimation of Texture-Less 3D Objects in Heavily Cluttered Scenes", ACCV 2012
[4] S. Hinterstoisser, S. Holzer, C. Cagniart, S. Ilic, K. Konolige, N. Navab, V. Lepetit, "Multimodal Templates for Real-Time Detection of Texture-less Objects in Heavily Cluttered Scenes", ICCV 2011
[5] J. Shotton, B. Glocker, C. Zach, S. Izadi, A. Criminisi, A. Fitzgibbon, "Scene Coordinate Regression Forests for Camera Relocalization in RGB-D Images", CVPR 2013
[6] A. Kendall, M. Grimes, R. Cipolla "PoseNet: A Convolutional Network for Real-Time 6-DOF Camera Relocalization", ICCV 2015

This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program (grant agreement No 647769)

European Commission | Horizon 2020 European Union funding for Research & Innovation

erc TECHNISCHE UNIVERSITÄT DRESDEN