Practicum on machine learning and artificial intelligence on the visual data

Representation of 3D data and method of their acquisition
Reconstruction of 3D models
6 DoF position estimation

RNDr. Martin Madaras, PhD.
Mgr. Lukáš Gajdošech

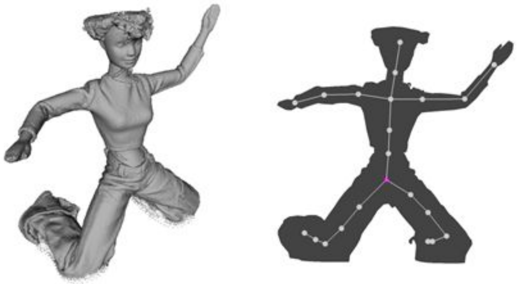
Overview

- Introduction & motivation
- 1. 3D scanners and point clouds
- 2. Reconstruction of 3D models
- 3. Processing of ordered and unordered point clouds
    - Segmentation
    - 6 DoF position estimation
- Project & homework

About me

- Comenius University, DAI
  - Research and theses supervision at university
  - Lectures @ FMFI - FCGIP, VAR, PMLAIVD
- Skeletex Research
  - Freelancing research and development company
  - Data processing for 3D scanners and cameras
- Timeline
  - 2014 - finished PhD @ FMFI
  - 2015 - research assistant @ FMFI / freelancer / SAIA post-doc project @ TU Wien
  - 2017 - co-founded Skeletex Research (3D scan processing)
  - 2018 - assistant professor @ FIIT / FMFI
  - 2020 - ML-based research on 3D scan processing
  - 2022 - habilitation @ FMFI

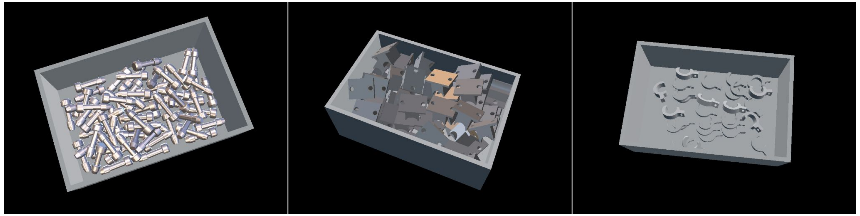
Research timeline



- Mesh processing / Skeleton extraction
- Skeletal animation / Skeletal-based parametrization and representation
- Motion capture / Hybrid optical-inertial systems
- 3D cameras / Skeleton-based human fusion
- 3D object reconstruction
- **3D scan processing (parallel CUDA-based pipeline, ML-based pipelines)**
- **Synthetic dataset creation / Physical simulation and rendering**

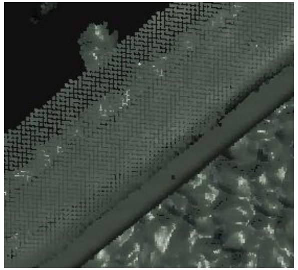
Motivation

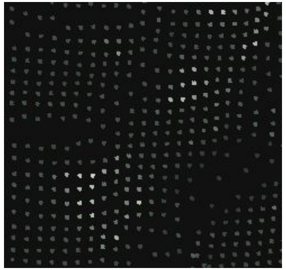
- Problems with robustness of analytical methods
- Prototyping based on ML models and training
- Development, deployment and maintenance costs efficiency
- Data-driven research and publications

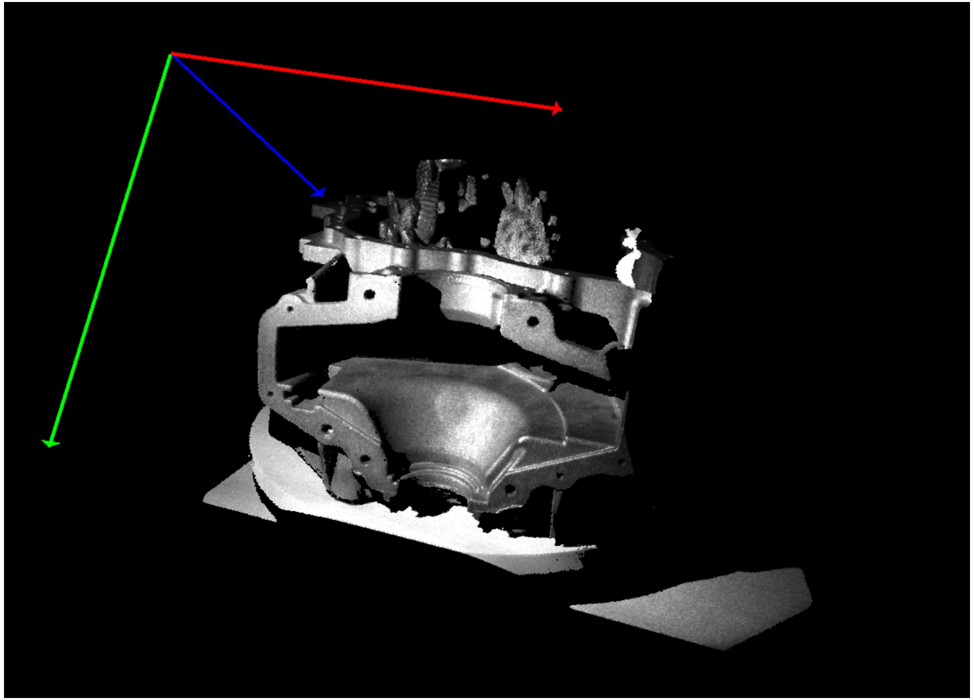


3D camera data

- Structured point cloud
- Camera-space
    - Cartesian coordinate basis
    - Right-handed basis







3D camera data

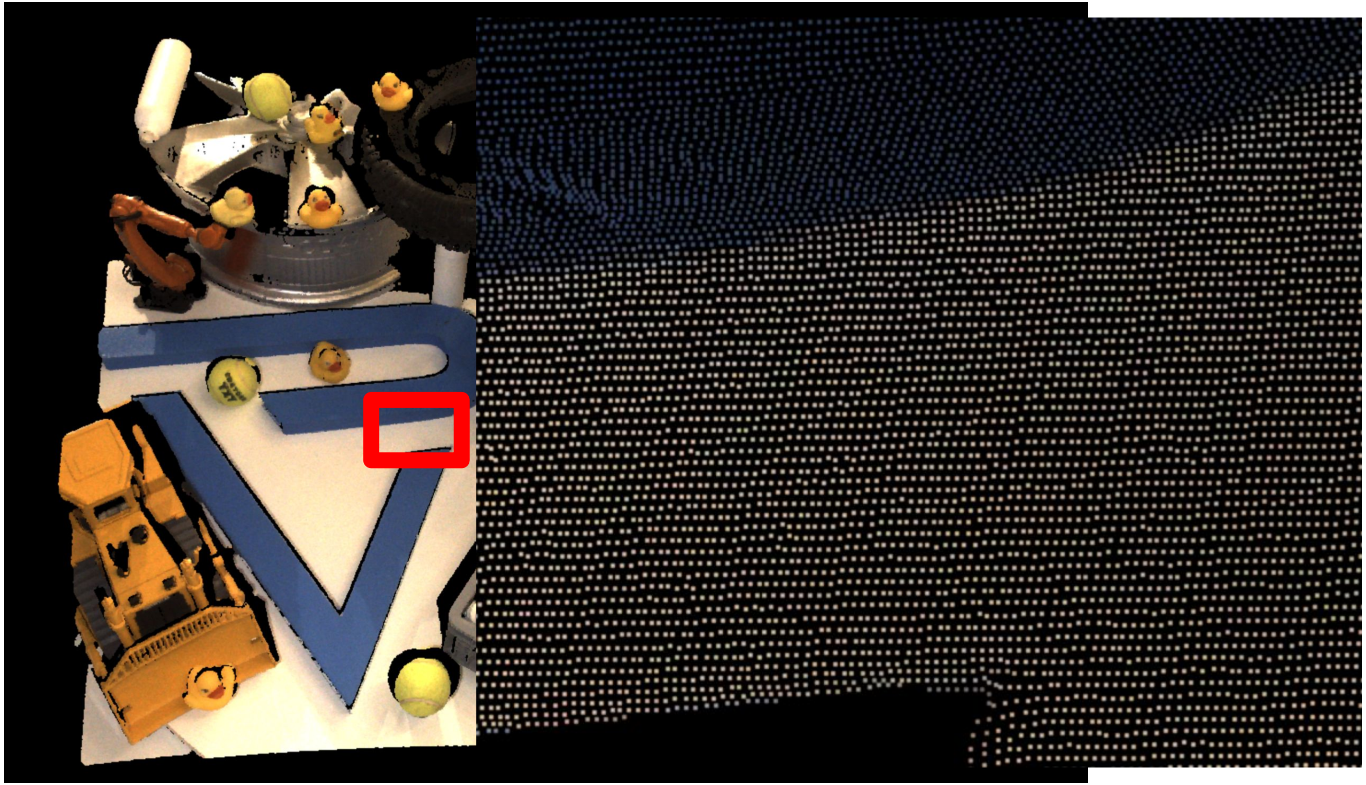
- Point cloud
- RGB texture

3D camera data

- Ordered
- Point cloud

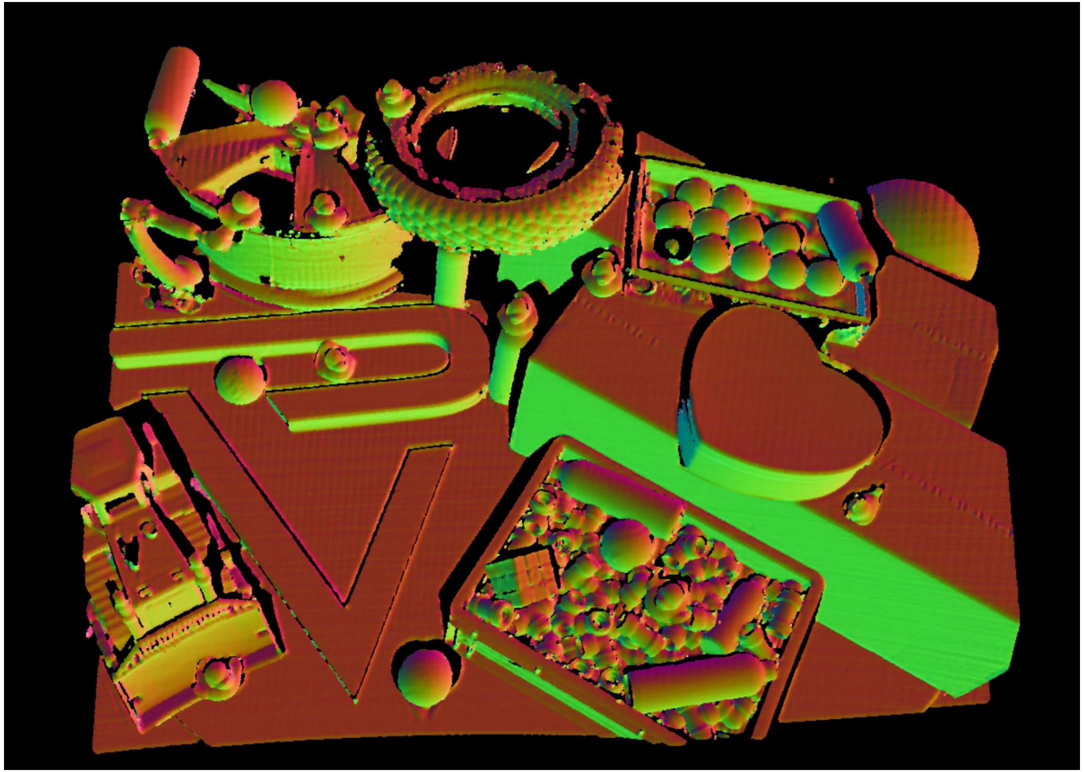
3D camera data

- Ordered
- Point cloud



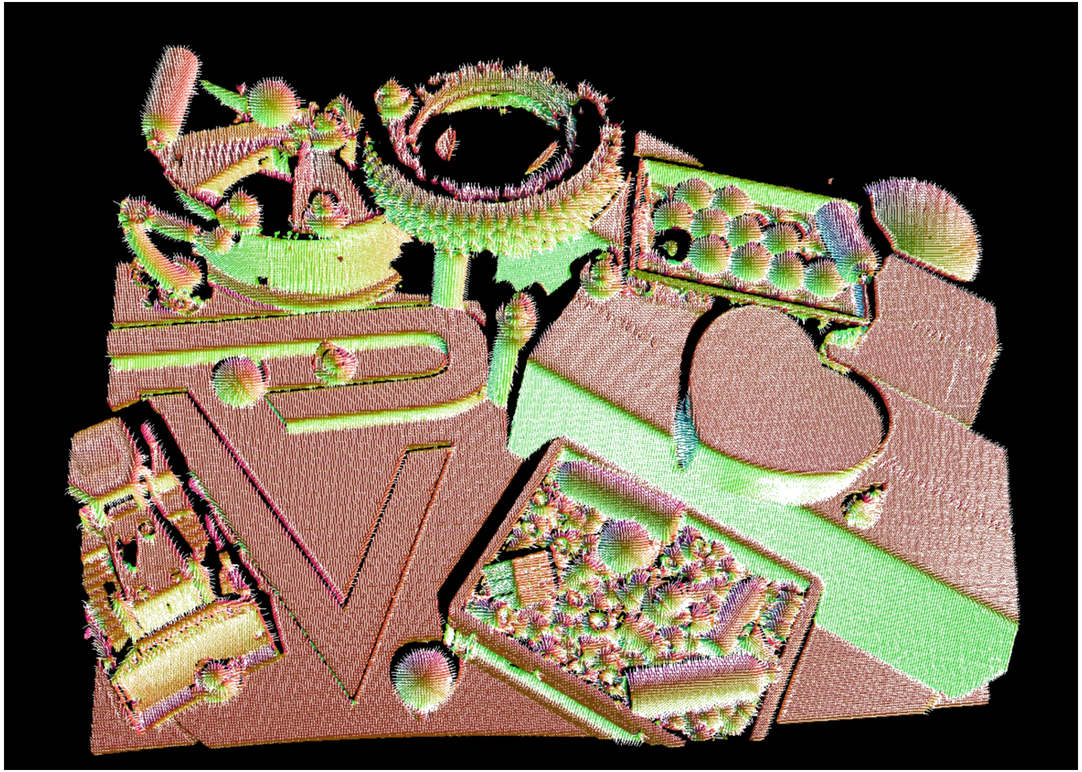
3D camera data

- Normals

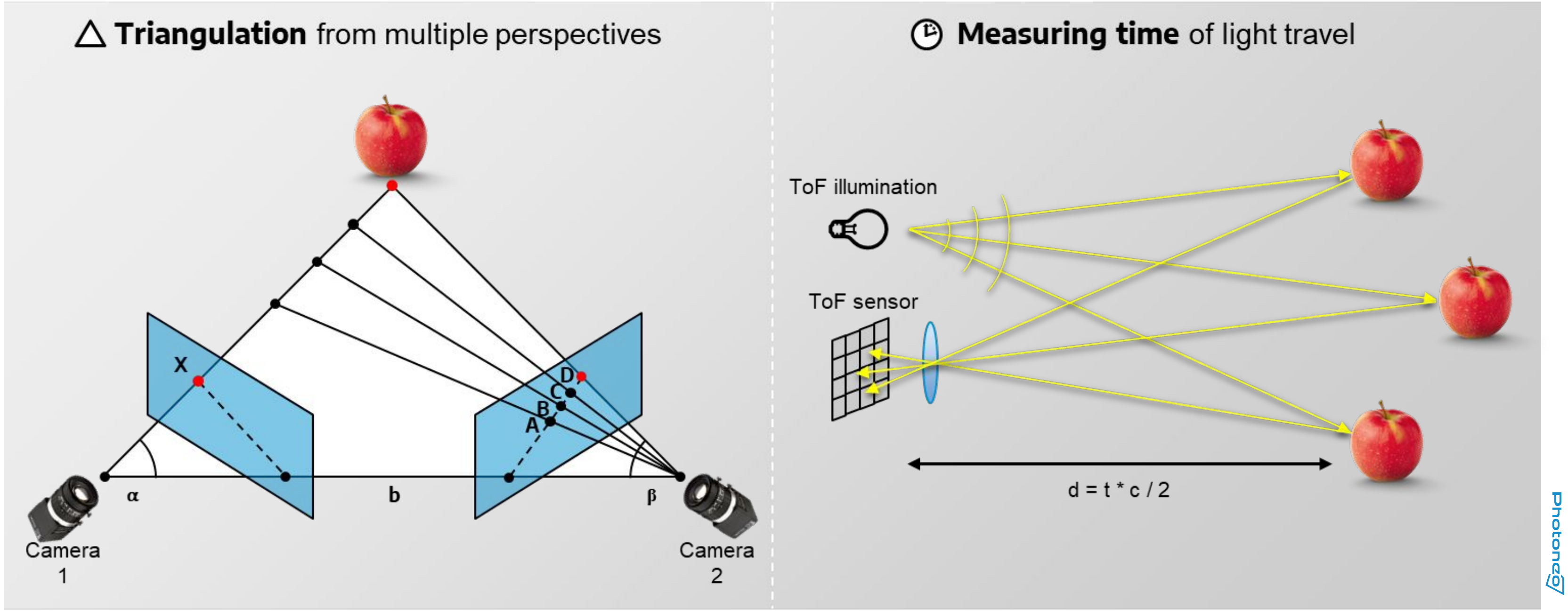


3D camera data

- Normals



3D cameras


△ **Triangulation** from multiple perspectives
X
D
C
B
A
α
b
β
Camera 1
Camera 2
Measuring time of light travel
ToF illumination
ToF sensor
d = t * c / 2
Photoneo

3D cameras

- Time measuring-based
  - Time of flight 3D cameras (Microsoft Kinect One)
  - LiDARs (pulsed laser, iPhone 6)
- Triangulation-based
  - Photogrammetry
  - Profile laser scanners
  - Structured light 3D scanners (HP Pro S3)
  - Structured light 3D camera (PrimeSense)
  - Parallel structured light 3D camera (Photoneo)

Time of flight

- High frame rate
- Noisy data, lower precision

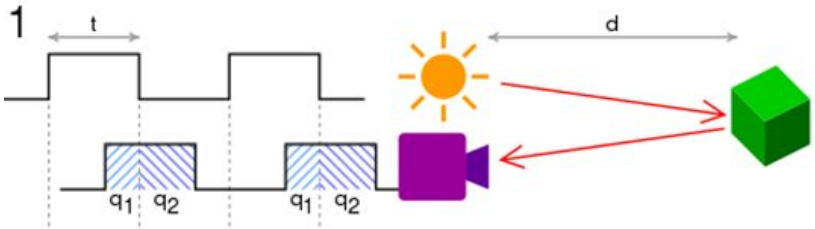
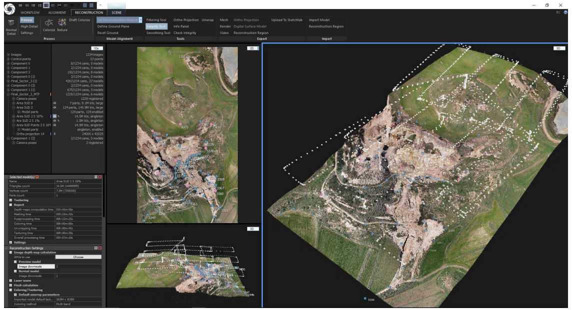


Image adopted from Buttgen et al. 2018

Image adopted from Li et al. 2014, Time-of-Flight Camera - An Introduction

Stereo vision

- 2 or more registered cameras
- Photogrammetry (bundle adjustment)
  - Global optimization



Images adopted from Capturing Reality

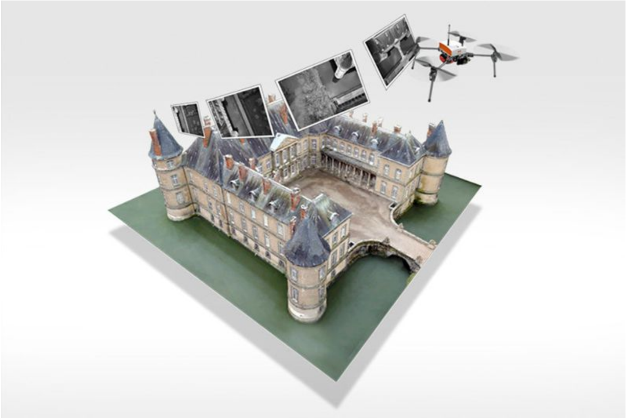
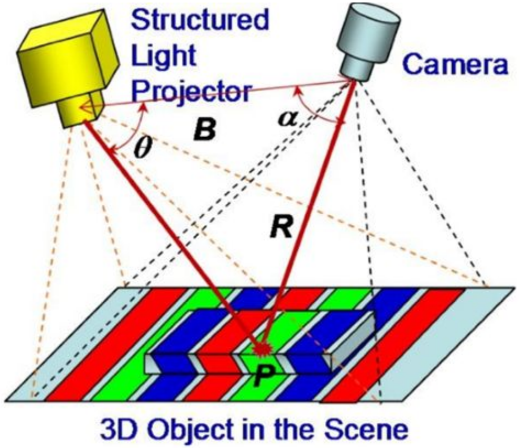


Image adopted from Ramırez-Hernandez et al. 2020

Structured light

- Pattern projection
  - Binary code encoding projectors angle
- 3D coordinate mapping calculation



Images adopted from Gang et al. 2011, Structured-light 3D surface imaging

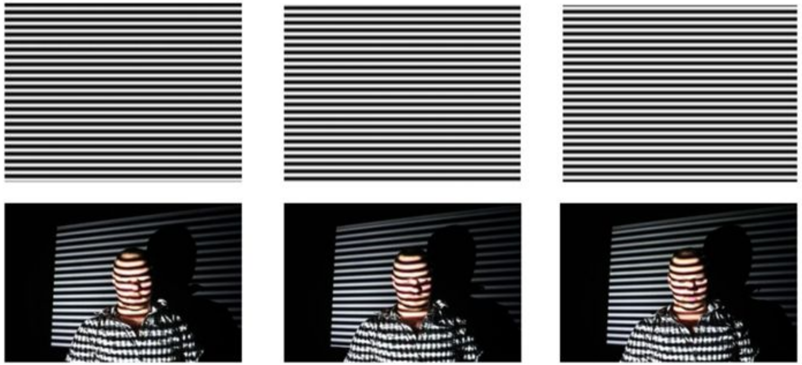
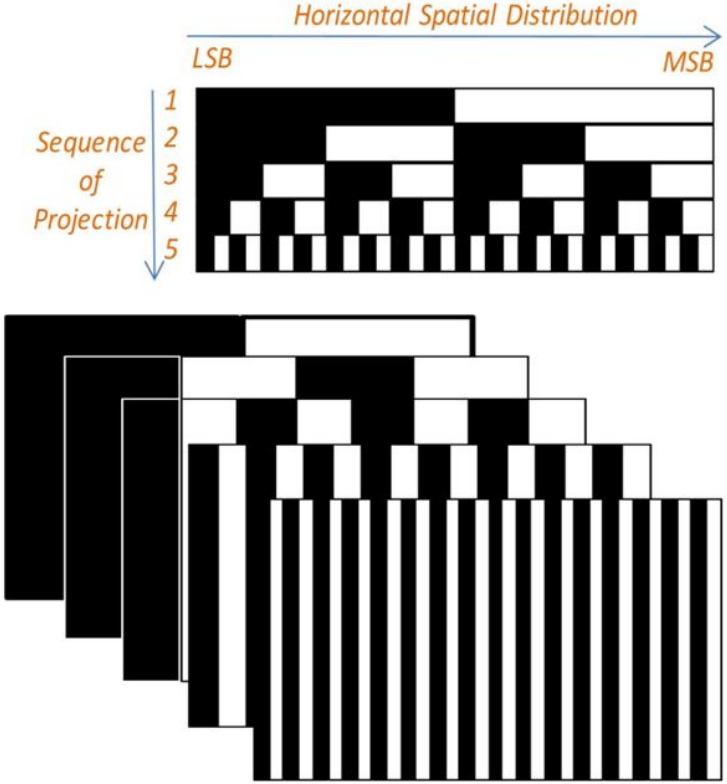


Image adopted from HP 3D Structured Light Scanner Pro S3

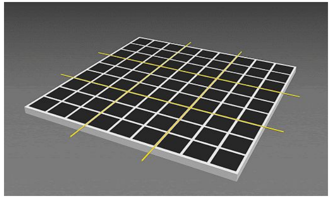
Structured light

- Less noisy, higher precision
- Multiple pattern projections for high precision
    - Full sensor-space
    - Gray code

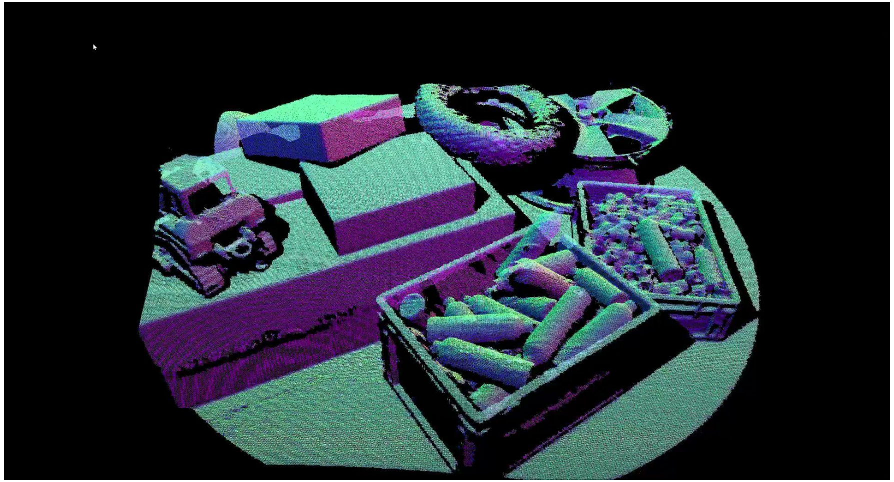


Images adopted from Gang et al. 2011, Structured-light 3D surface imaging

Parallel structured light 3D Camera



- New hardware enables new industry possibilities
- e.g., MotionCam-3D
  - Parallel structured light
  - Laser swipe & mosaic sensor



Object reconstruction

- Quality control
- Metrology (measurements)
- Virtual objects

Rotary table reconstruction

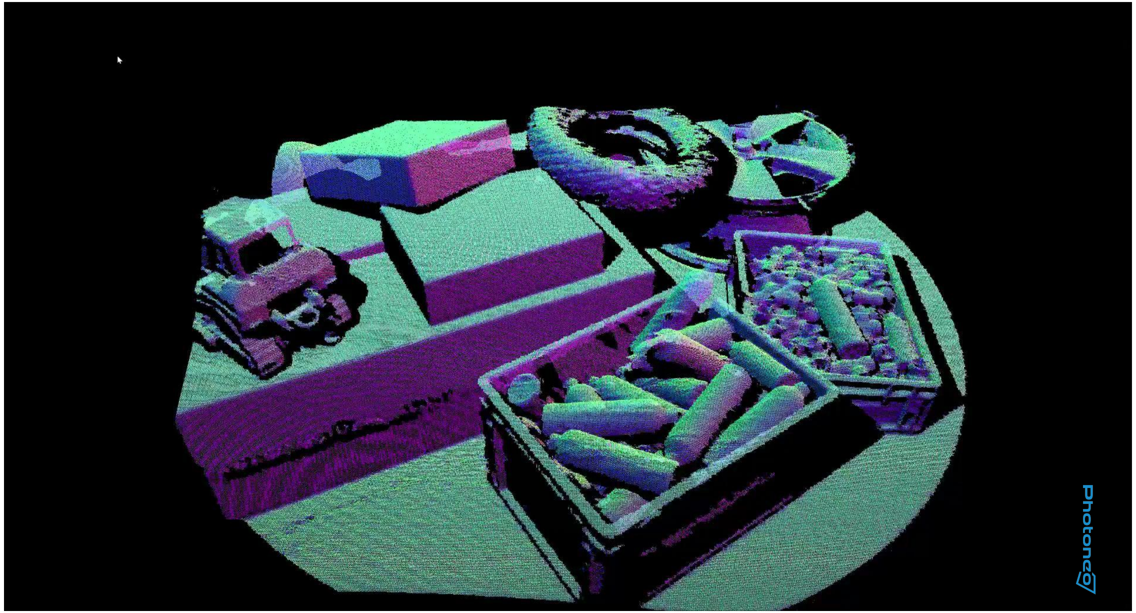
Reconstruction pipeline

- Point-cloud Rigid Alignment and Fusion of 3D Scans (PRAFOS)
- Offline reconstruction pipeline
    - Alignment, filtering and reconstruction

3D meshing pipeline

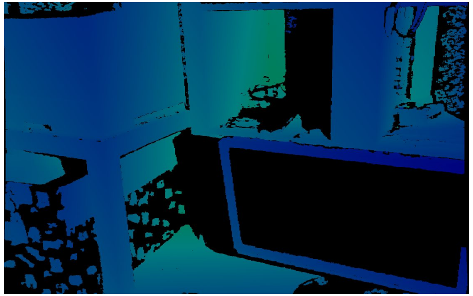
Parallel data processing for 3D camera

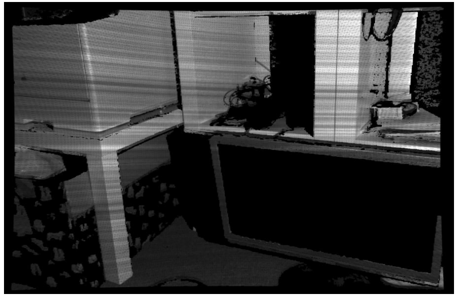
- Structured-light 3D camera
- 20+ FPS
- High-quality data



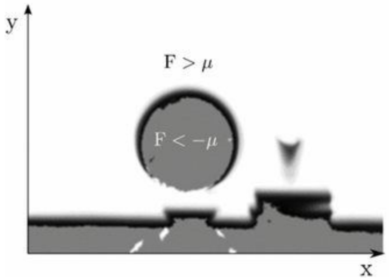
Fusion using 3D Camera

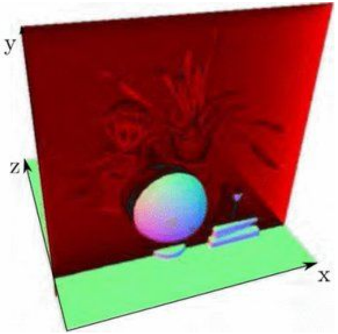
- Real-time reconstruction
- Point cloud in sensor space as input
- Scene is stored in TSDF structure
    - Truncated
    - Signed
    - Distance
    - Function





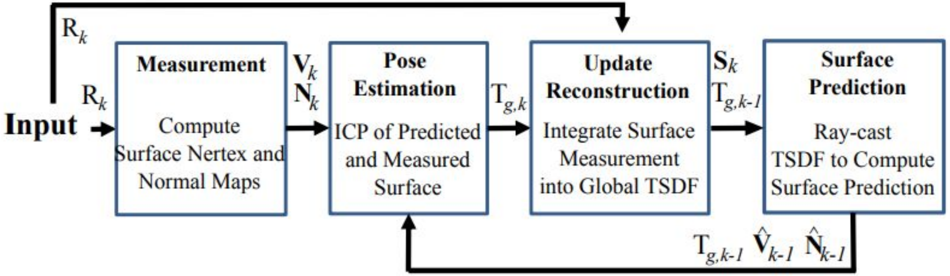
Images adopted from Newcombe 2011 et al., KinectFusion



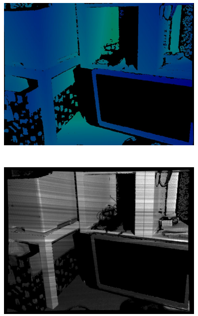


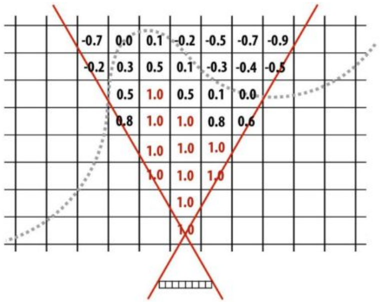
Fusion registration and update

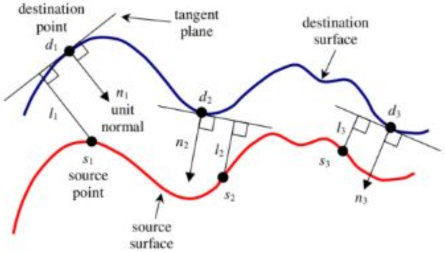
- For each incoming frame in the loop:
  - Sensor pose estimation (R + T)
  - Surface update
  - Surface prediction

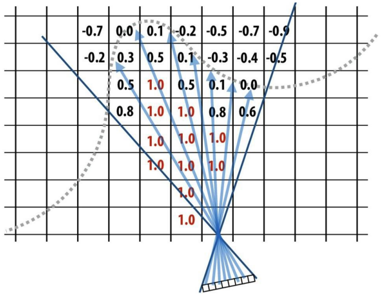


Images adopted from Newcombe 2011 et al., KinectFusion



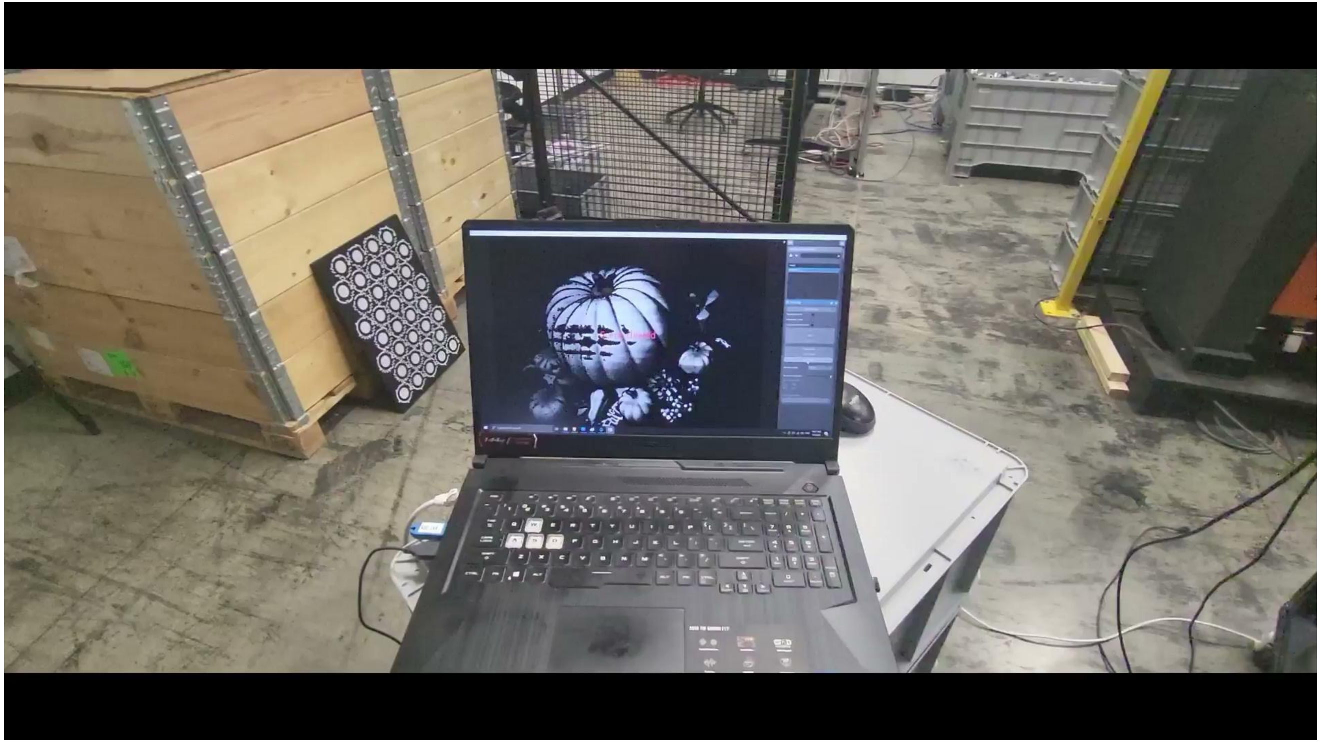






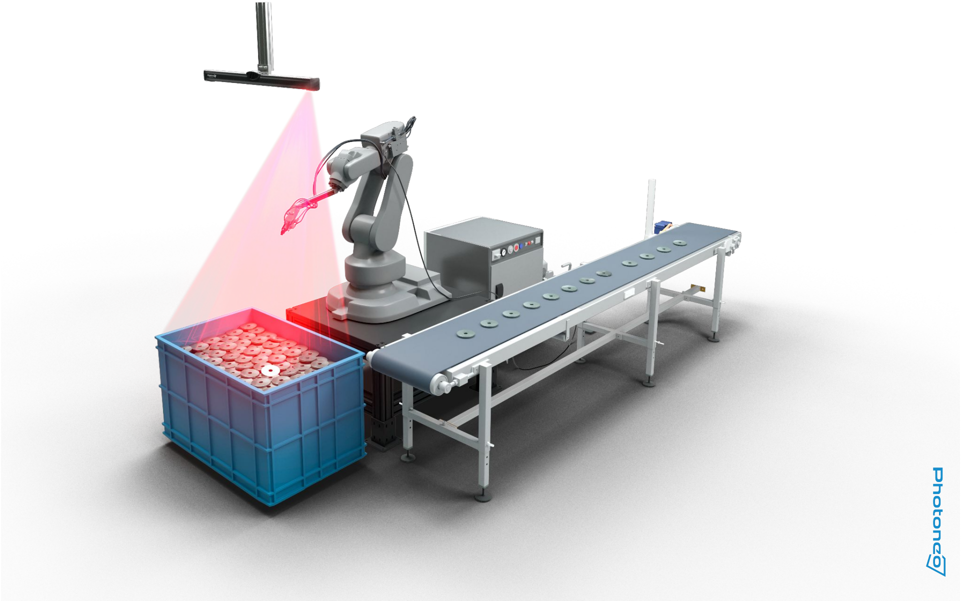
Online fusion by robotic arm

-



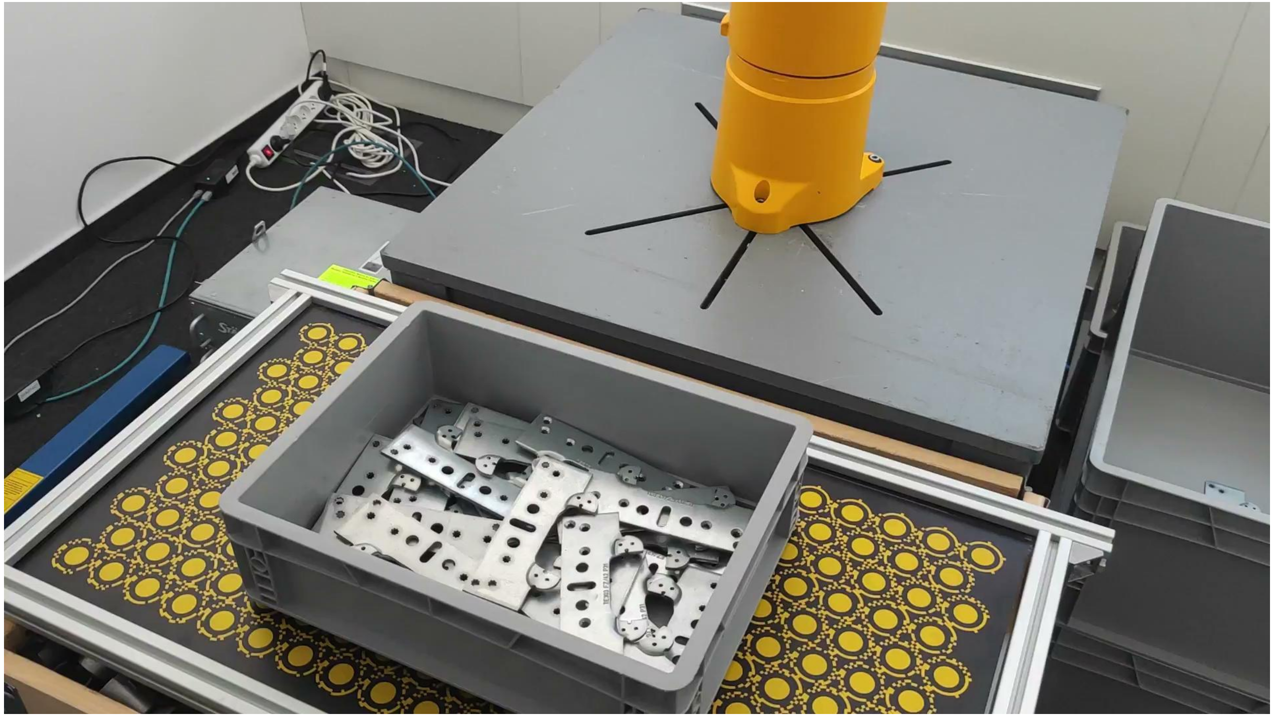
Vision-guided robotics

- Visual information processing
- Localization
  - Segmentation
  - 6D pose estimation
- Picking and handling
- Kinematics optimization



Bin picking

-



Point cloud processing

Image adopted from: Katsoulas et al. 2003, Localization of piled boxes by means of the hough transform, DAGM

(a) Intensity image (b) Edge Map (c) Lines

- Analytical methods
  - Analysis based on local geometrical informations
  - Detection of feature points / edges
  - Precise transformation calculations
- Data-driven ML methods
  - Analysis based on statistics and annotated training data
  - Global semantic information can be used
  - Robust to noise and outliers in the data

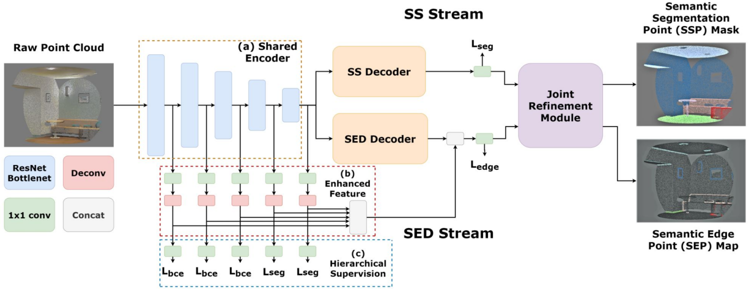


Image adopted from: Hu et al. 2020. JSENet: Joint Semantic Segmentation and Edge Detection Network for 3D Point Cloud, ECCV”

Parallel data processing for 3D scanners

- Madaras, M. Stuchlík, M. and Talčík, M. (2021)
- **Fast Bridgeless Pyramid Segmentation for Organized Point Clouds**
- In Proceedings of the 16th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications - Volume 4: VISAPP, ISBN 978-989-758-488-6; ISSN 2184-4321, pages 205-210.

Fast Bridgeless Pyramid Segmentation for Organized Point Clouds

Martin Madaras[1,2,a], Martin Stuchlík[1,b] and Matúš Talčík[1,3,c]
[1] Skeletex Research, Slovakia
[2] Faculty of Mathematics, Physics and Informatics, Comenius University Bratislava, Slovakia
[3] Masaryk University Brno, Czech Republic

Keywords: Point Cloud, Segmentation, Parallel, Pyramid, GPU, CUDA.

Abstract: An intelligent automatic robotic system needs to understand the world as fast as possible. A common way to capture the world is to use a depth camera. The depth camera produces an organized point cloud that later needs to be processed to understand the scene. Usually, segmentation is one of the first preprocessing steps for the data processing pipeline. Our proposed pyramid segmentation is a simple, fast and lightweight split-and-merge method designed for depth cameras. The algorithm consists of two steps, edge detection and a hierarchical method for bridgeless labeling of connected components. The pyramid segmentation generates the seeds hierarchically, in a top-down manner, from the largest regions to the smallest ones. The neighboring areas around the seeds are filled in a parallel manner, by rendering axis-aligned line primitives, which makes the performance of the method fast. The hierarchical approach of labeling enables to connect neighboring segments without unnecessary bridges in a parallel way that can be efficiently implemented using CUDA.

1 INTRODUCTION

The world is moving towards automation. Robots are picking parts from trays to assemble larger components, cars can park themselves and cruise on highways and 3D printers are self-correcting printing mistakes. All these applications are based on machines understanding the surrounding world. One of the principles is to determine where and what kind of objects are positioned in the space around the machine. This is traditionally done by a depth camera and subsequent processing of the input point clouds. When capturing a stream of organized point clouds from a 3D camera, we need to process the stream of data as fast as possible. The captured point cloud is immediately processed by an image processing pipeline, where segmentation is one of the most important processing steps. The segmented point cloud is used as an input to following processing algorithms and thus the segmentation is used for the input data reduction. Because the scanning artifacts result in bridges in the thresholded pseudo-curvature metrics, it is not possible to use methods based on thresholding the metrics and flood-filling directly. Furthermore, segmentation methods based on hierarchical clustering cannot be efficiently implemented in a parallel way. The following algorithm in the processing pipeline is performed only using the subset of the input point cloud that is needed for the optimal performance of the algorithm. Therefore, efficiency and low execution time of the segmentation process is crucial in all these robotics applications.

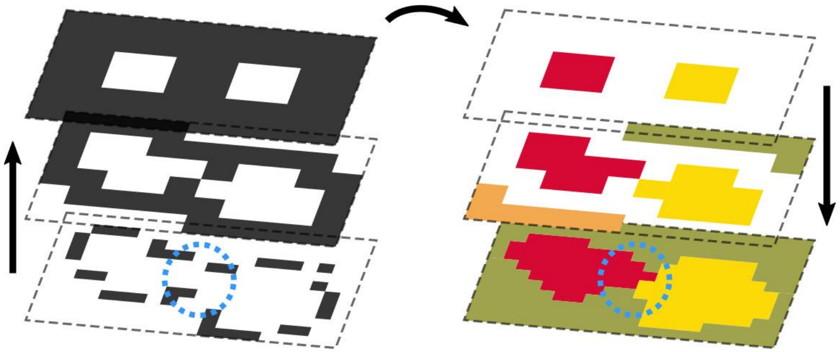
The main contribution of the paper is a novelty hierarchical parallel filling method that was designed for fast pyramid segmentation of organized point clouds (or directly depth maps with computed normal approximations). The proposed two-step method modifies the region filling algorithms such as a watershed or connected component labeling, into a fast and robust method that fills the regions bridgelessly. The filling is bridgeless even if the segmentation metrics contains bridges between the regions after the thresholding (bridges can be seen in Figure 1 and Section 3.3). The proposed method is directly applied on a pseudo-curvature metrics computed from a set of input organized point clouds captured by a 3D scanner. The segmentation results are evaluated qualitatively and the processing times are quantitatively compared with other state of the art filling methods, benchmarked on different CUDA compatible hardware.

[a] https://orcid.org/0000-0003-3917-4510
[b] https://orcid.org/0000-0001-8556-8364
[c] https://orcid.org/0000-0002-3865-4507

Fast hierarchical segmentation

- Fast and bridgeless segmentation
- Metrics based on geometry and thresholding
- Structured point cloud segmentation
- Parallel filling of regions using pyramid hierarchy

TX2	Scan01	12.10	69.25	81.35
	Scan02	10.40	73.21	83.62
	Scan03	13.50	56.59	70.09
	Scan04	10.40	54.09	64.50
	Scan05	3.96	50.94	54.90
1050Ti	Scan01	1.30	15.83	17.13
	Scan02	1.25	19.72	20.97
	Scan03	1.29	15.22	16.51
	Scan04	1.29	16.54	17.83
	Scan05	0.90	11.94	12.84



6D pose estimation of bins

- Lukáš Gajdošech, Viktor Kocúr, Martin Stuchlík, Lukáš Hudec and Martin Madaras
- **Towards Deep Learning-based 6D Bin Pose Estimation in 3D Scans**
- Proceedings of the 17th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications, VISIGRAPP 2022 (Volume 4: VISAPP) 2022, pp. 545-552, February 2022

Towards Deep Learning-based 6D Bin Pose Estimation in 3D Scans

Lukáš Gajdošech[1,2][a], Viktor Kocur[2,4][b], Martin Stuchlík[1][c], Lukáš Hudec[3][d] and Martin Madaras[1,2][e]
[1]*Skeletex Research, Slovakia*
[2]*Faculty of Mathematics, Physics and Informatics, Comenius University Bratislava, Slovakia*
[3]*Faculty of Informatics and Information Technologies, Slovak Technical University Bratislava, Slovakia*
[4]*Faculty of Information Technology, Brno University of Technology, Czech Republic*
{gajdosech, kocur}@fmph.uniba.sk, {madaras, stuchlik}@skeletex.xyz

Keywords: Computer Vision, Bin Pose Estimation, 6D Pose Estimation, Deep Learning, Point Clouds.

Abstract: An automated robotic system needs to be as robust as possible and fail-safe in general while having relatively high precision and repeatability. Although deep learning-based methods are becoming research standard on how to approach 3D scan and image processing tasks, the industry standard for processing this data is still analytically-based. Our paper claims that analytical methods are less robust and harder for testing, updating, and maintaining. This paper focuses on a specific task of 6D pose estimation of a bin in 3D scans. Therefore, we present a high-quality dataset composed of synthetic data and real scans captured by a structured-light scanner with precise annotations. Additionally, we propose two different methods for 6D bin pose estimation, an analytical method as the industrial standard and a baseline data-driven method. Both approaches are cross-evaluated, and our experiments show that augmenting the training on real scans with synthetic data improves our proposed data-driven neural model. This position paper is preliminary, as proposed methods are trained and evaluated on a relatively small initial dataset which we plan to extend in the future.

1 INTRODUCTION

Capturing a scene with 3D scanners is a standard for automatized systems analyzing a scene. To pick mechanical parts from a bin by a robotic arm equipped with a gripper, the parts need to be localized. First, the localization of bin is essential to restrain the robot from collisions. Then, the kinematics of the robot is optimized for path planning. The problem of bin localization can be defined as a 6 DoF pose estimation of a template 3D model of the bin in the 3D scan.

Nowadays, analytical methods are still the industrial standard for the processing of 3D scans. On the contrary, the academic and research standards have evolved to data-driven or hybrid approaches. Analytical computation of bin transformation in captured point clouds might be vulnerable to missing critical information in the captured scans, like corners and edges, yielding lower robustness than expected. The computation precision of a hard-defined analytical algorithm might be higher but at the cost of lower robustness if a key content is missing. In applications of automated intelligent systems, it may be interesting to lower its precision to increase the robustness in some scenarios. The other possible approach is to split the pipeline into two steps - the first part of the pipeline orients on the robustness and raw data-driven localization. The second part focuses on the precision-based analytical solution starting from the predicted pose estimations, thus having the robustness properties inherited from the data-driven approach.

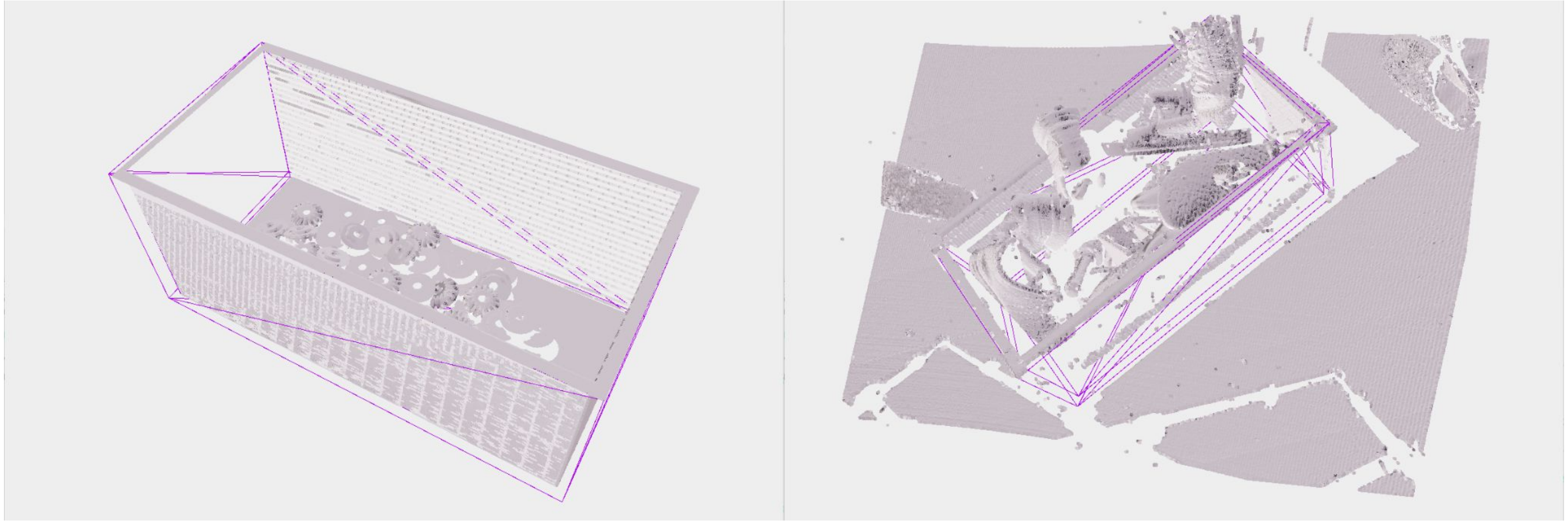
In this paper, we present a novel dataset containing high-quality real and synthetic 3D scans of different bins in various poses containing a variety of items captured by structured light scanners. We publish the dataset[1] for further research. We propose an analytical method and a conceptually simple deep convolutional neural network for 6D bin pose estimation. We experimentally evaluate it and show that our network is more robust than the analytical method.

[a] https://orcid.org/0000-0002-8646-2147
[b] https://orcid.org/0000-0001-8752-2685
[c] https://orcid.org/0000-0001-8556-8364
[d] https://orcid.org/0000-0002-1659-0362
[e] https://orcid.org/0000-0003-3917-4510

[1] http://skeletex.xyz/portfolio/datasets

Gajdošech, L., Kocur, V., Stuchlík, M., Hudec, L. and Madaras, M.
Towards Deep Learning-based 6D Bin Pose Estimation in 3D Scan.
In *Proceedings of the 17th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP 2022) - Volume 4: VISAPP*, pages 545-552
ISBN: 978-989-758-555-5; ISSN: 2184-4321

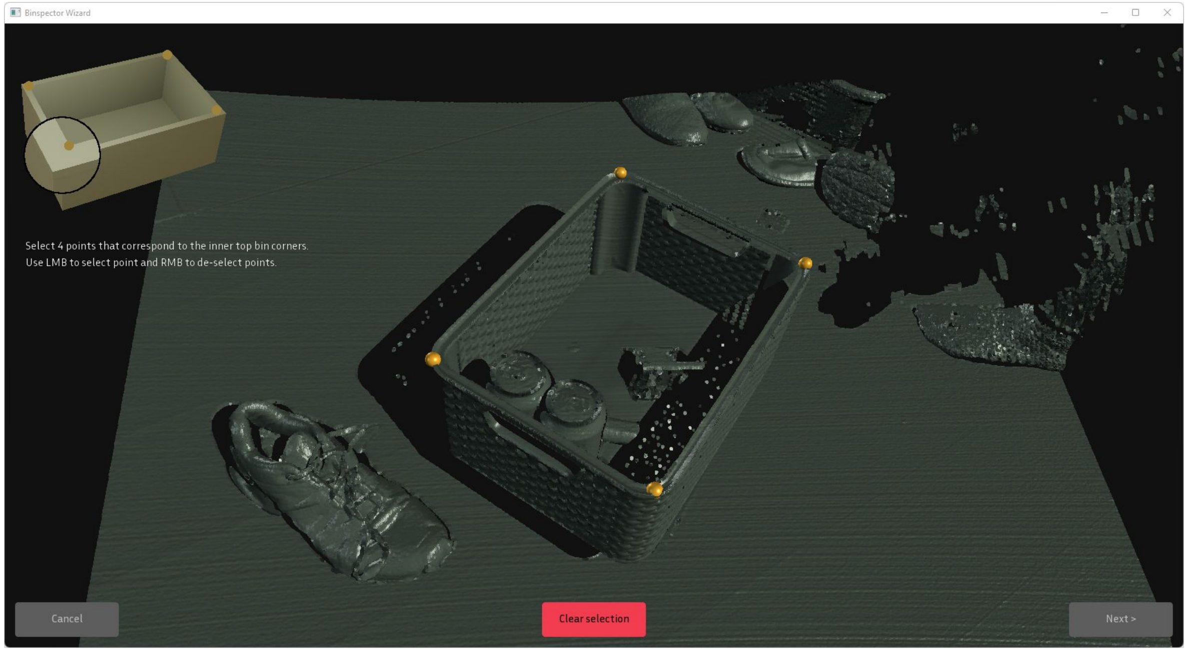
6D pose estimation of bin



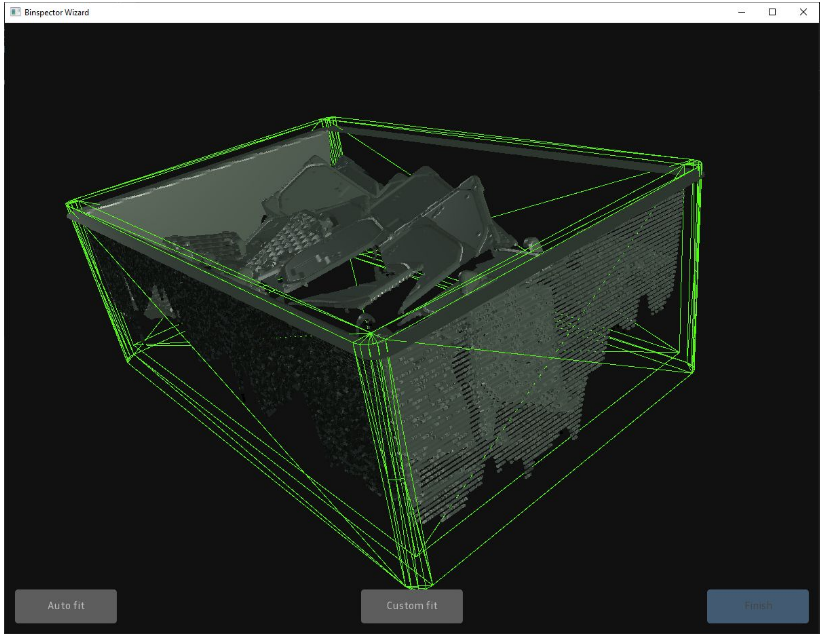
Lukáš Gajdošech & Martin Stuchlík & Viktor Kocur & Lukáš Hudec & Martin Madaras, Towards Deep Learning-based 6D Bin Pose Estimation in 3D Scans, Visigrapp 2022

Data annotation tool

-

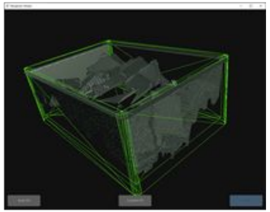


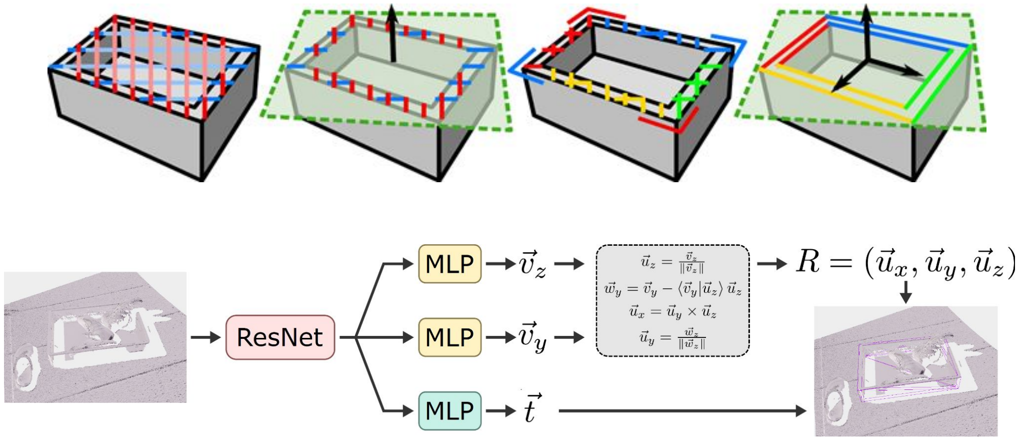
Synthetic data generation



6D bin pose estimation

- ResNet network architecture
- Two heads for rotation and one head (branch) for translation

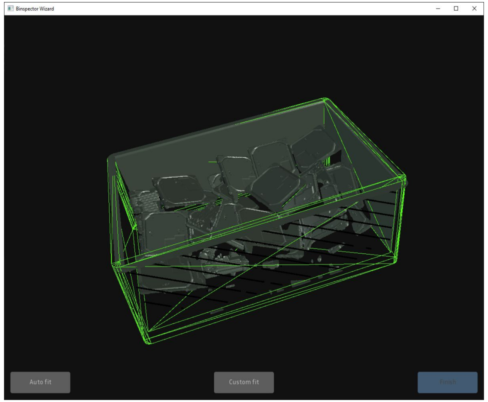


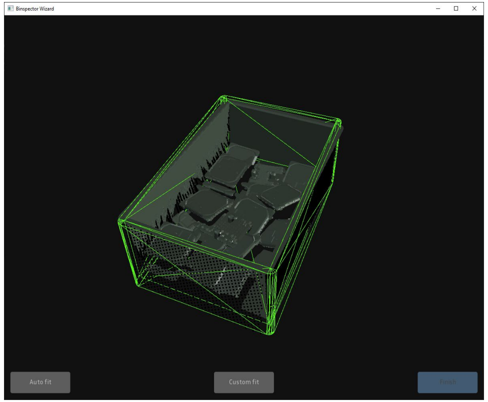


ResNet34 architecture, 34 layers, 8.3M parameters

Data-driven results on synthetic data

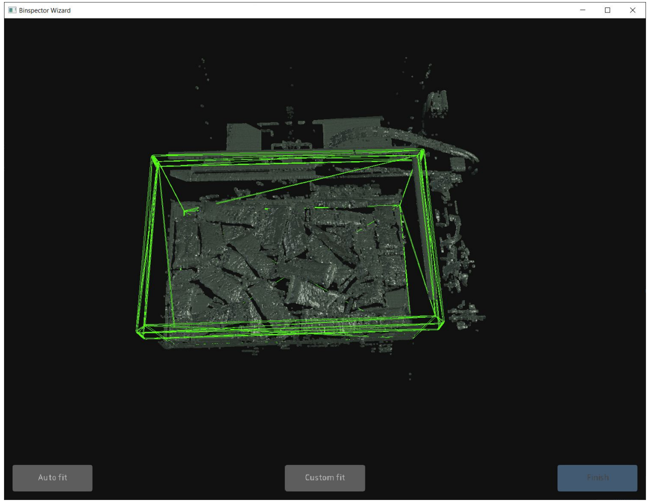
- Approximation of the transformation
- Works well for same data type as training data





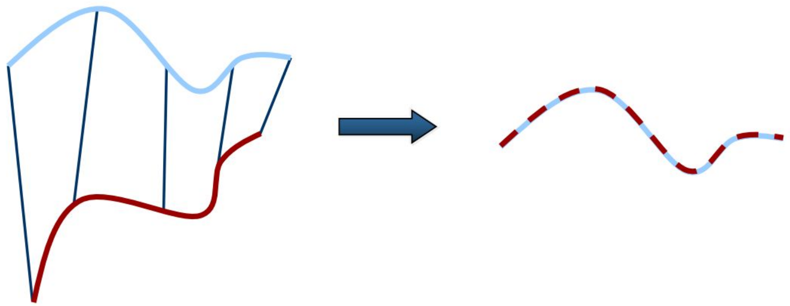
Data-driven results on real data

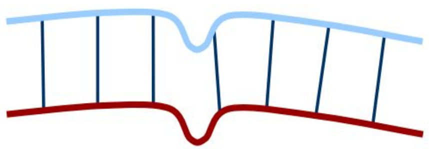
- Robust to noise and outliers
- Precision is lower on real data
- Analytical post-processing can be used (ICP)



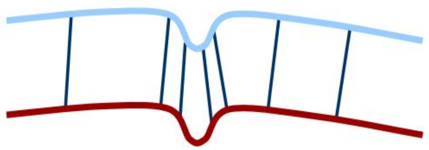
Iterative closest point

- Projection-based ICP pairing





uniform sampling



normal-space sampling

Image adopted from: Sebastian Thrun, Wolfram Burgard and Dieter Fox, Probabilistic Robotics, 2005

Projection-based ICP

- Given two set of points: $X = \{x_1, ..., x_n\}$

  $P = \{p_1, ..., p_n\}$

- Compute translation and rotation that minimizes the sum of the squared error (MSE):

$$E(R,t) = \frac{1}{N_p} \sum_{i=1}^{N_p} ||x_i - Rp_i - t||^2$$

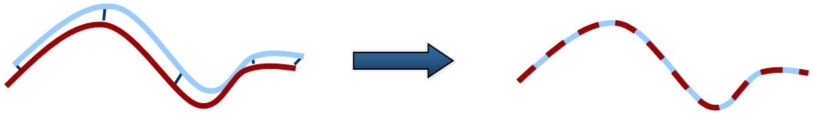
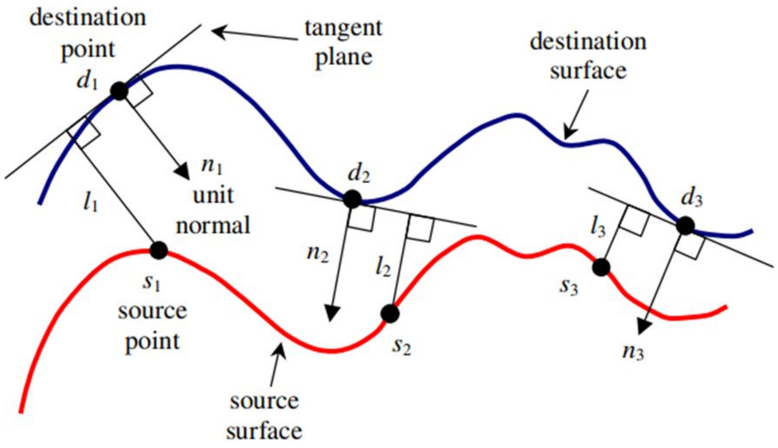


Image adopted from: Sebastian Thrun, Wolfram Burgard and Dieter Fox, Probabilistic Robotics, 2005

Point to plane projections

- Point-to-plane error metrics
  - Minimize the sum of the squared distance
  - Projection of a point to the tangent plane at its correspondence point [2]

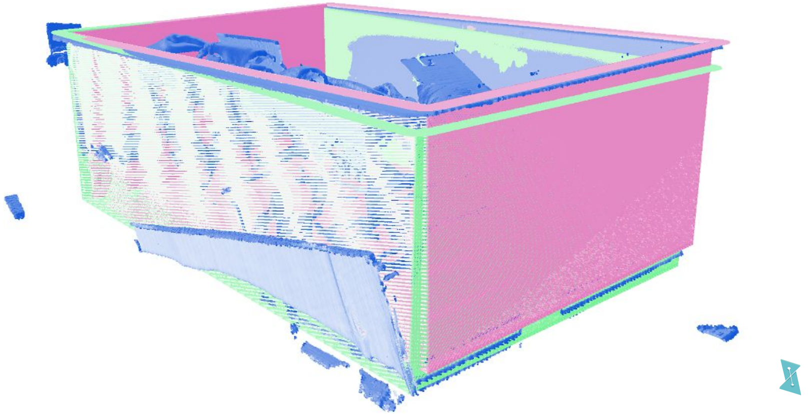
Image adopted from: Low 2004, Linear Least-Squares Optimization for Point-to-Plane ICP Surface Registration



[2] Chen & Medioni,1991, Object Modeling by Registration of Multiple Range Images, IEEE International Conference on Robotics and Automation

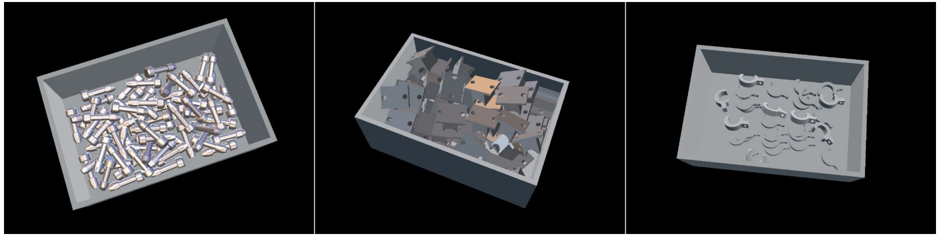
Hybrid two-step approach

- Neural network inference for pre-alignment
- Analytical post-processing using ICP (iterative closest point) for post-alignment
  - good results if there is a good starting approximation



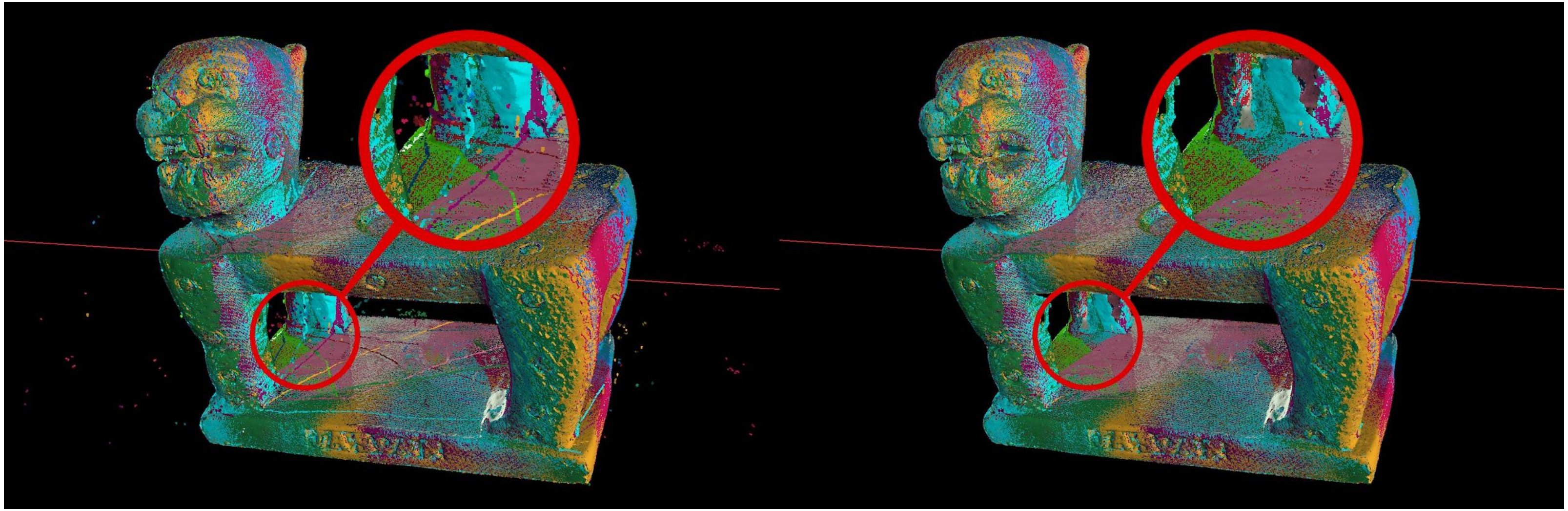
Hybrid methodology

- Combination of data-driven approximations and analytical alignments
- Pre-alignment and final alignment
    - Pose estimation
    - Registration / alignment
- Higher robustness and precision of the approach



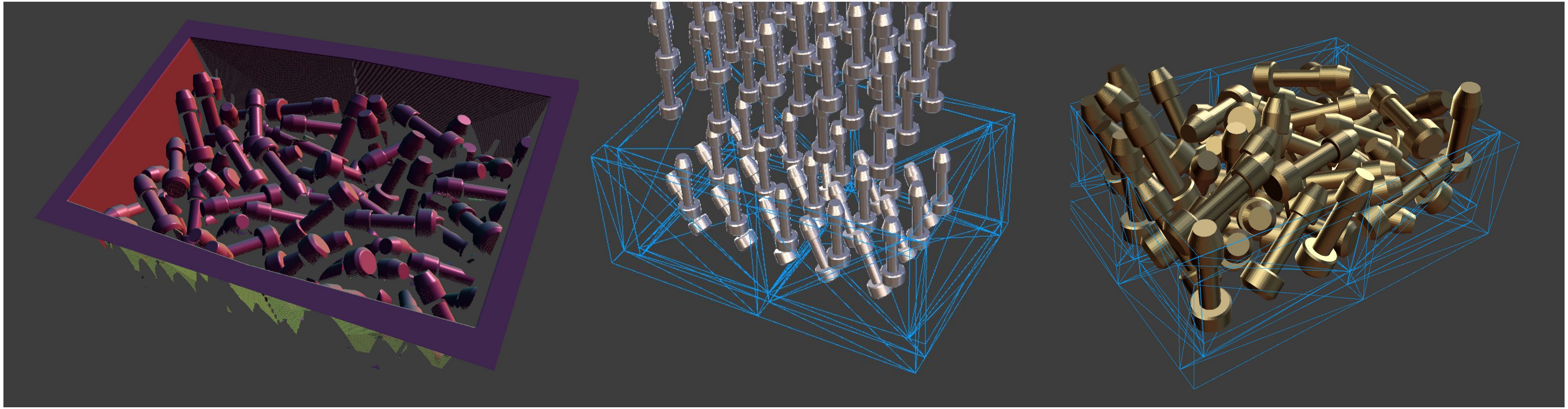
Point cloud processing using ML

- Full-stack point cloud processing pipeline can be moved to data-driven methods
  - Filtering, detection, segmentation, 6D pose estimation, etc.



Synthetic data generation for machine learning

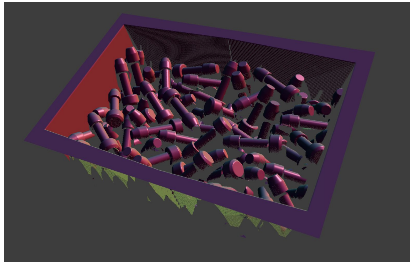
- Huge training datasets available for standard 2D computer vision problems
- Annotated 3D datasets are missing, we need to render synthetic ones

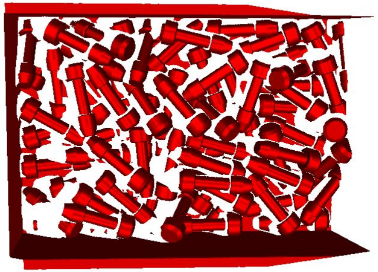


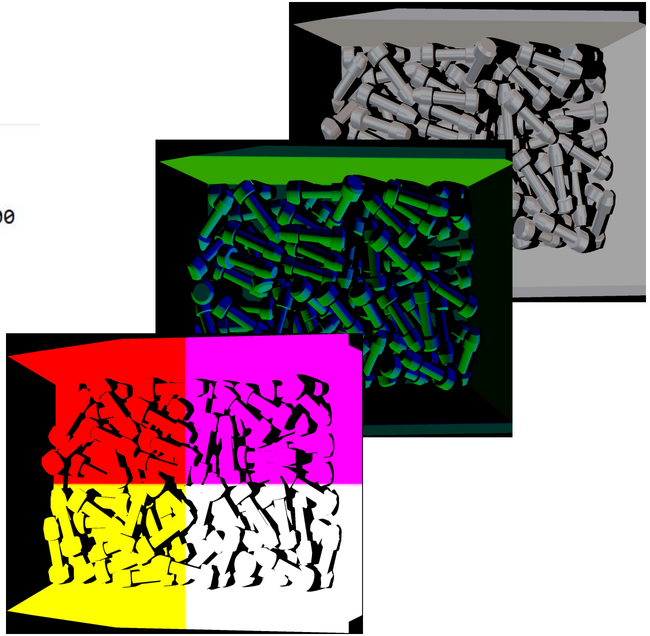
Synthetic data generator

- Bin generator
  - BinSim
  - Virtual scanner

```
0.999277 -0.002169 0.037963 0.000000
0.000000 -0.998371 -0.057050 0.000000
0.038025 0.057009 -0.997649 0.000000
-4.253108 -6.800327 855.200745 1.000000
```

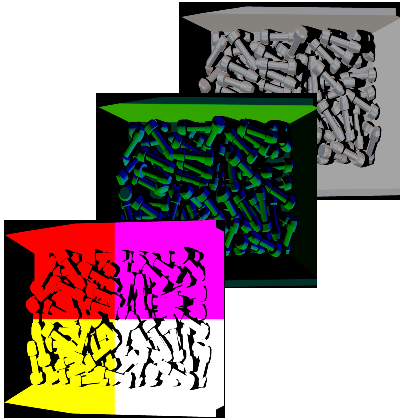






Synthetic data generator

- Pre-generated (.exr) datasets download:
- https://drive.google.com/drive/folders/1EX3GeXtZEJMD5y1Qh3wUF0R_3rY0nU1i?usp=sharing
-
- Bingenerator (.exr) download:
- https://drive.google.com/file/d/19n8_Sjb-t1JNWWQBCAQHd4LuDAH6fOBC/view?usp=share_link



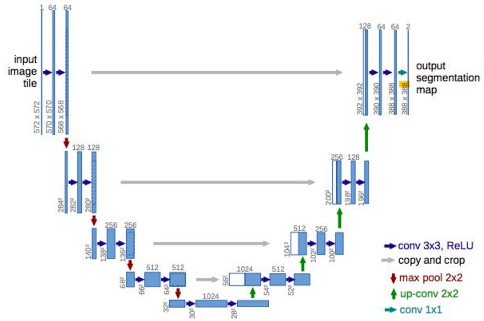
Google colab

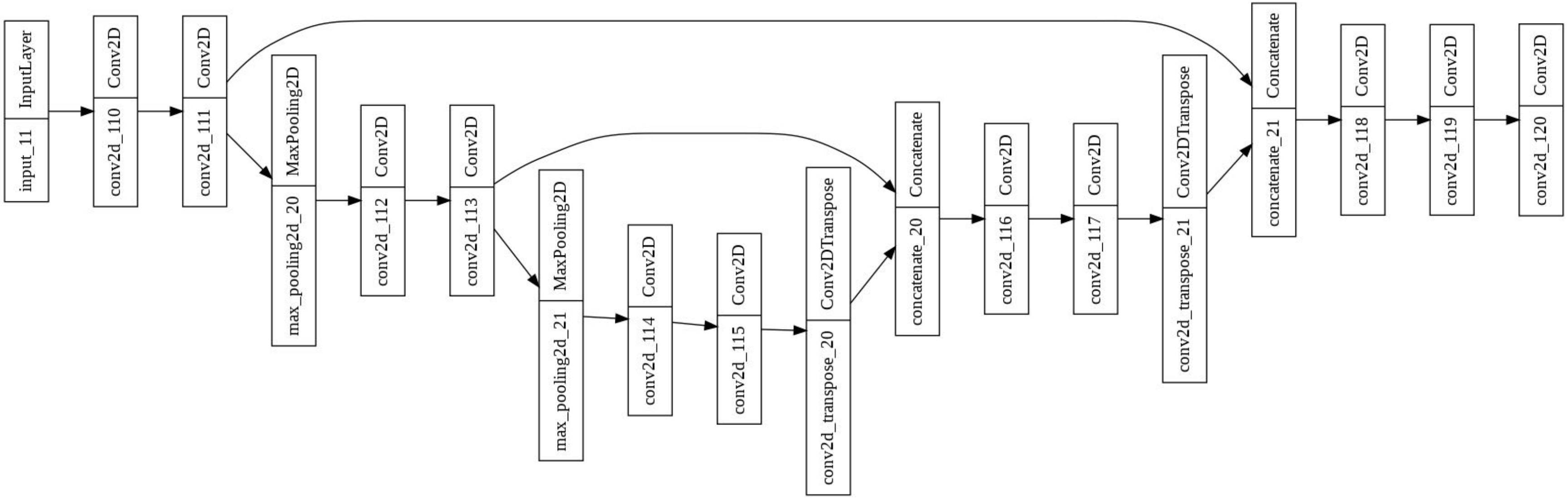
- U-Net segmentation (binary)
    - Ordered point clouds
    - Synthetic bin with objects dataset
    - https://colab.research.google.com/drive/1Oo8BqlbRwW7nu4e4veziyhWIMT2xAMTf?usp=sharing

- PointNet segmentation (semantic, low-res 1024 points)
    - Unordered point clouds
    - ShapeNet dataset + 3D scans of object
    - https://colab.research.google.com/drive/15Ug_SF1gNq2xAH1f6ZIOTrB4Dgk7vDhd?usp=sharing

Neural network model 1

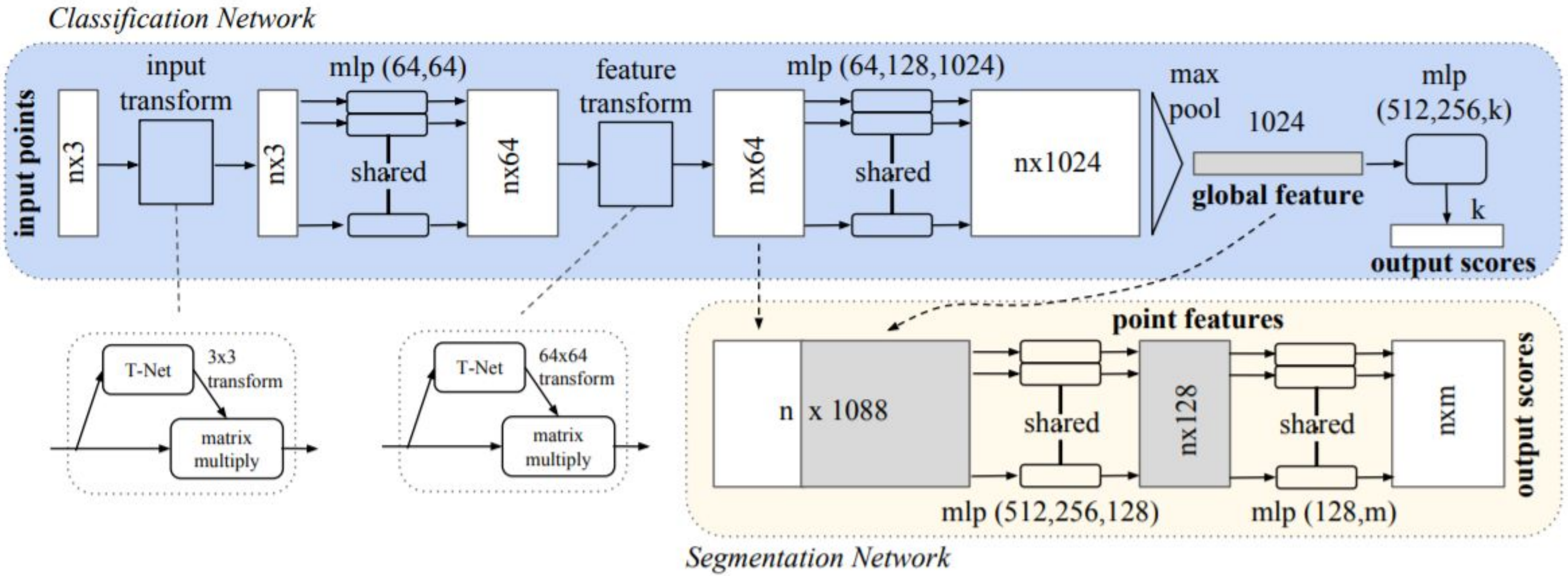
- U-Net
  - Total number of layers: 18
  - Trainable params: 126,841





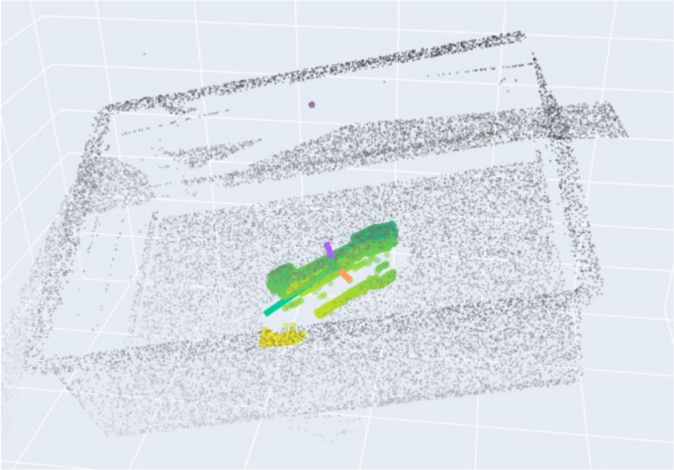
Neural network model 2

- PointNet



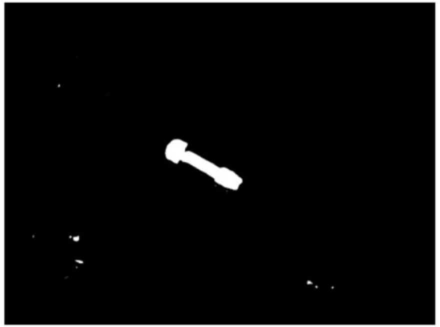
Homework project

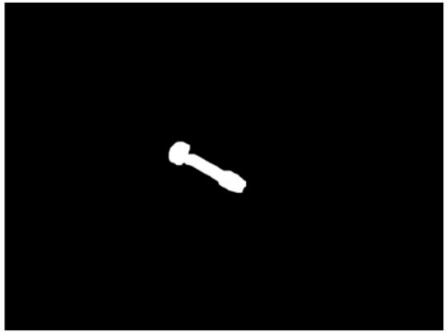
- 6 pose estimation of object in bins
  - Use U-Net segmentation (binary)
  - Find orientation of the segmented object using PCA
  - Visualize oriented bounding box or principal vectors



Homework project

- 6 pose estimation of object in bins
    - If the inferred segmentation mask is not composed of one component only
    - Use morphological operations - opening (erosion + dilation)
        - cv2.erode() , cv2.dilate()





Homework project (assignment in Slovak)

Úlohou je nájsť 6D orientáciu (pozícia + rotácia) súčiastky v 3D skene reprezentovanom ako mračno bodov. V prvej časti spravte segmentáciu pomocou konvolučnej neurónovej siete U-Net. Sieť natrénujte na syntetických dátach s objektom “thruster” a následne inferujte segmentáciu tiež na syntetickom 3D skene obsahujúcom iba jednu súčiastku. Pokiaľ je segmentačná maska z viacerých ako jedného komponentu, použite morfologické otvorenie na odstránenie malých komponentov. Morfologickú masku použite na vybratie podmnožiny bodov tvoriacich vzorky súčiastky, nad ktorými aplikujete Principal Component Analysis (PCA). Pri PCA získate centroid objektu a hlavné vektory, ktoré definuju orientáciu súčiastky. Vizualizujte tieto hlavné vektory, alebo orientovaný obdĺžnik ohraničujúci objem (boundig box) objektu.

Questions ?

madaras@skeletex.xyz

martin.madaras@fmph.uniba.sk