Robot Learning Seminar
WS 2022/23

Robot Learning Lab

Albert-Ludwigs-Universität Freiburg

Friday, 21 October 2021

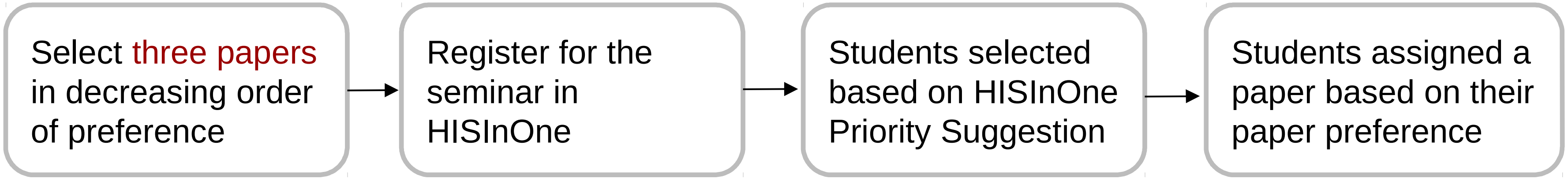
Evaluation

Evaluation	Due Date
Paper Abstract	13/01/2022
Seminar Presentation	10/02/2022
Paper Summary	24/02/2022

- Abstract → At most 2 pages
- Presentation → At most 20 minutes
- Summary → At most 7 pages excluding bibliography and figures
- Final grade → Abstract, Presentation, Summary, Seminar participation

Seminar: https://rl.uni-freiburg.de/teaching/ws22/seminar-robot-learning

Enrollment Procedure



Select three papers in decreasing order of preference

Register for the seminar in HISInOne

Students selected based on HISInOne Priority Suggestion

Students assigned a paper based on their paper preference

Until 24/10/2021
Form

Until 24/10/2021

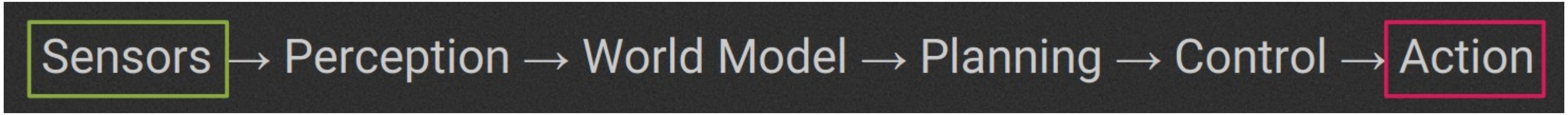
Until 25/10/2021

Until 25/10/2021

Robot Learning

- Tremendous progress on complex, high dimensional data
  - Speech Recognition
  - Natural Language Processing
  - Computer Vision

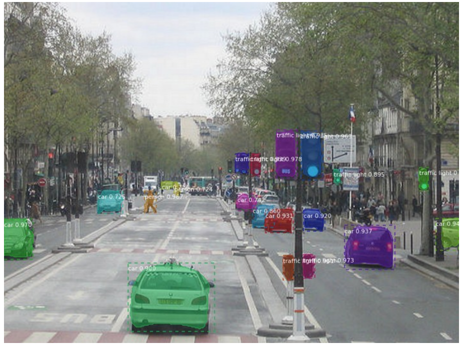
- Autonomous systems smart enough to operate in the real world

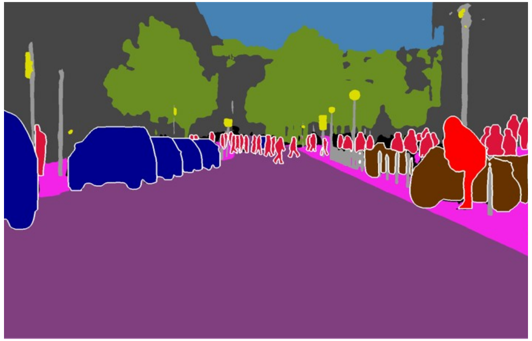


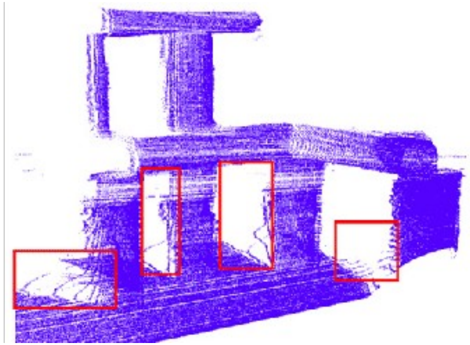
Sensors → Perception → World Model → Planning → Control → Action

Perception

- Complex environments
- Noisy observations and sensors







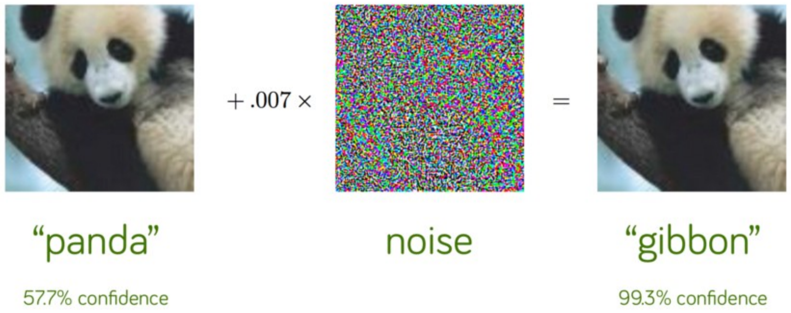
Mask R-CNN for object detection and instance segmentation on Keras and TensorFlow, Waleed et. al., 2017.

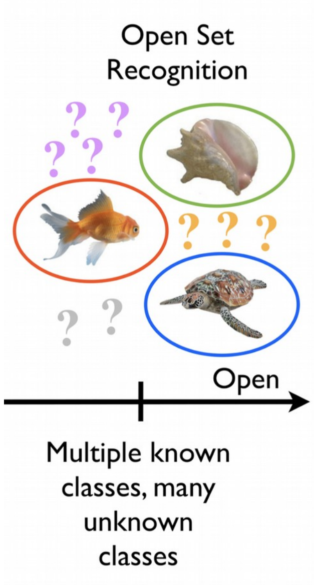
EfficientPS: Efficient Panoptic Segmentation, Mohan and Valada, 2020.

Hybrid approach for alignment of a pre-processed three-dimensional point cloud, video, and CAD model using partial point cloud in retrofitting applications, Patil et. Al., 2018.

Unknown, Open World

- Unknown world → Many unlabelled samples
- Uncertainty estimation
- Adversarial attacks



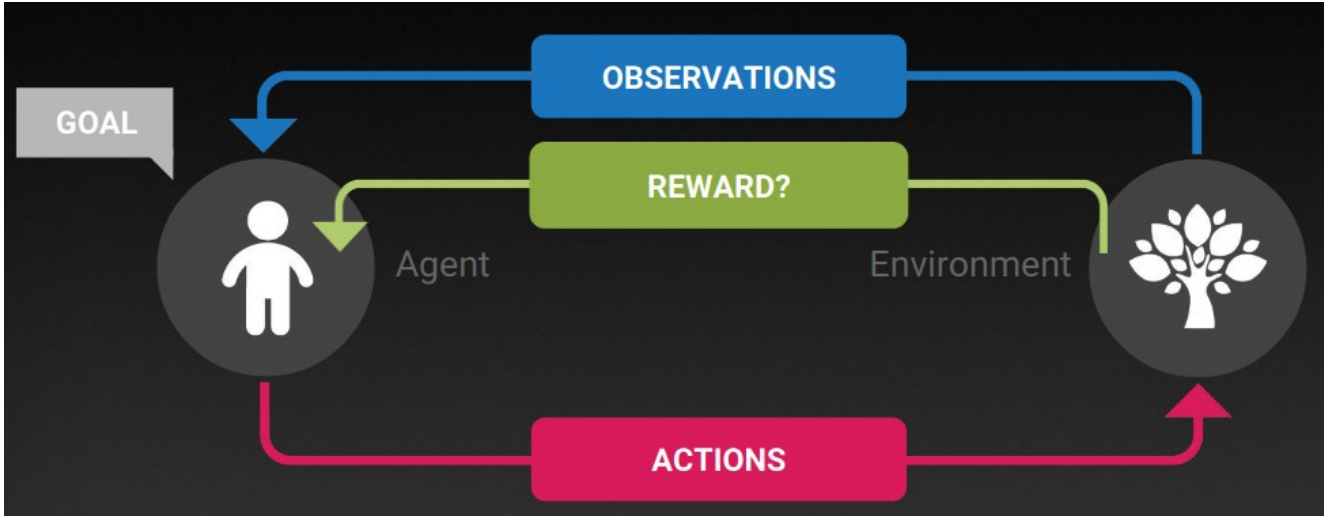


Towards Open Set Recognition, Scheirer et. al., 2012

Explaining and Harnessing Adversarial Examples, Goodfellow et. al., 2014

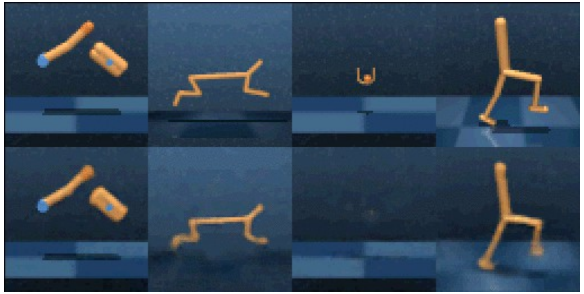
Autonomous Decision Making

- Reinforcement learning for short- and long-term decision making

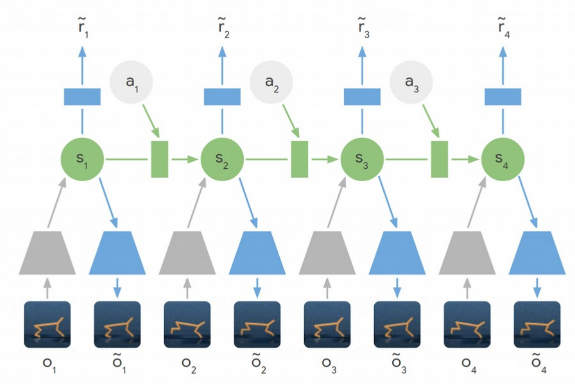


Reinforcement Learning

- Model free RL
  - Adapts to complex scenarios
  - Directly optimize policy
  - Data intensive



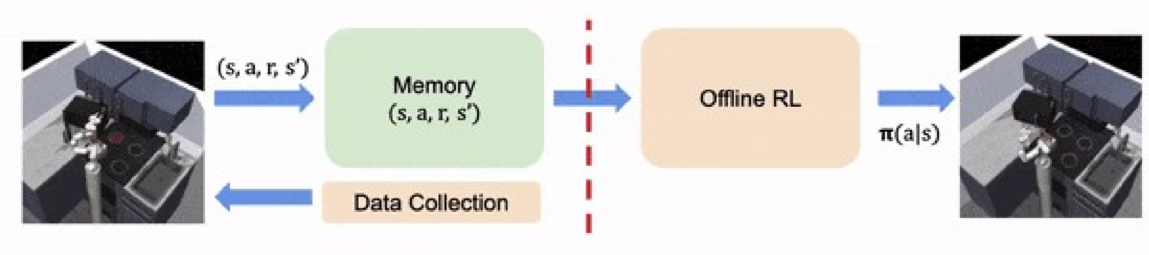
- Model-based RL
  - Learns a world model
  - Promise of better generalization



Learning Latent Dynamics for Planning from Pixels, Hafner et. al., 2019

Expensive Real World Data

- Sim2Real
  - Domain adaptation
  - Action and dynamics noise
- Offline RL
  - Large amounts of unstructured data
  - Little annotated / expert data

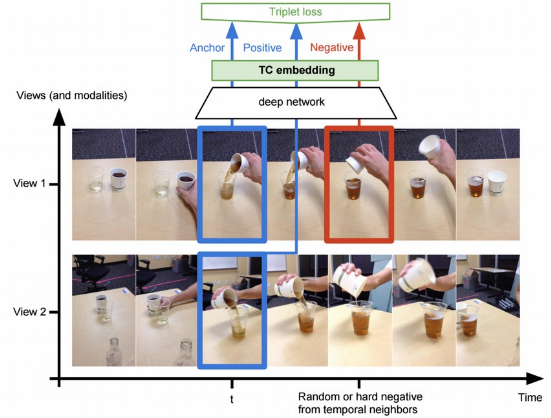


Learning Hand-Eye Coordination for Robotic Grasping with Deep Learning and Large-Scale Data Collection. Sergey Levine, Peter Pastor, Alex Krizhevsky, Deirdre Quillen

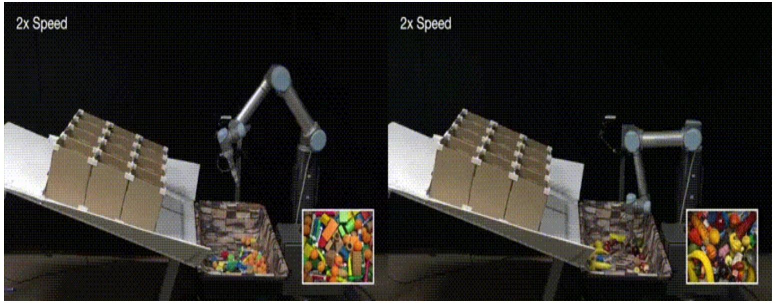
D4RL: Datasets for Deep Data-Driven Reinforcement Learning, Justin Fu, Aviral Kumar, Ofir Nachum, George Tucker, Sergey Levine

Weak- and Self-Supervision

- Provide labels for simpler tasks
  - Object presence and absence
  - Consistency over time
  - Viewpoint invariance



- Reduce oversight
  - Automatic resets
  - Reward labelling



Time-Contrastive Networks: Self-Supervised Learning from Video, Sermanet et. al., 2018

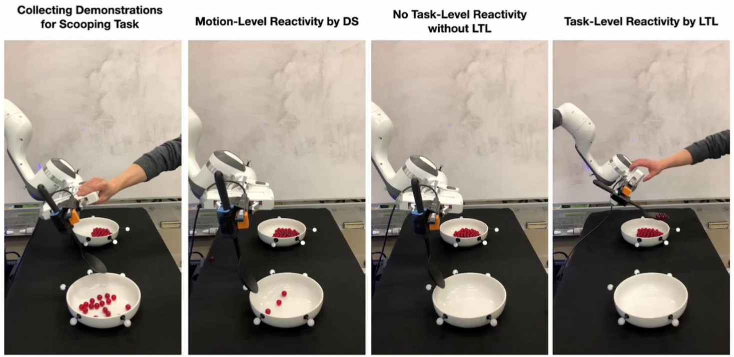
TossingBot: Learning to Throw Arbitrary Objects, Zeng et. al., 2019.

Seminar Topics

Temporal Logic Imitation: Learning Plan-Satisficing Motion Policies from Demonstrations

Supervisor: Adrian Röfer

- Learning from demonstration is not reliable in reproducing multi-step tasks
  - Cannot recover from task-level perturbations (accidentally dropping an object)
  - Point-sized sub-goals are too narrow
- In this work:
  - Learning to identify different *modes* of a task describable by temporal linear logic
  - Learn motions with mode invariance and reachability properties
  - Achieve task level and motion level robustness to perturbations

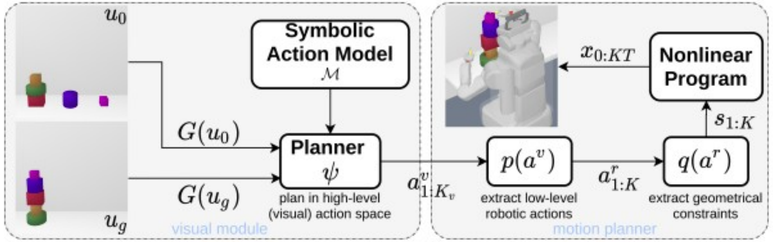

Collecting Demonstrations for Scooping Task
Motion-Level Reactivity by DS
No Task-Level Reactivity without LTL
Task-Level Reactivity by LTL

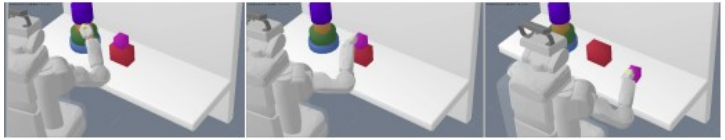
Self-Supervised Learning of Scene-Graph Representations for Robotic Sequential Manipulation Planning

Supervisor: Adrian Röfer

- Long sequential manipulation tasks are difficult
  - Which actions should be taken?
  - Are these feasible?

- In this work:
  - Scene Graph representation to ease planning
  - Learned self-supervised
  - Integration with motion synthesis to ground feasibility of plans

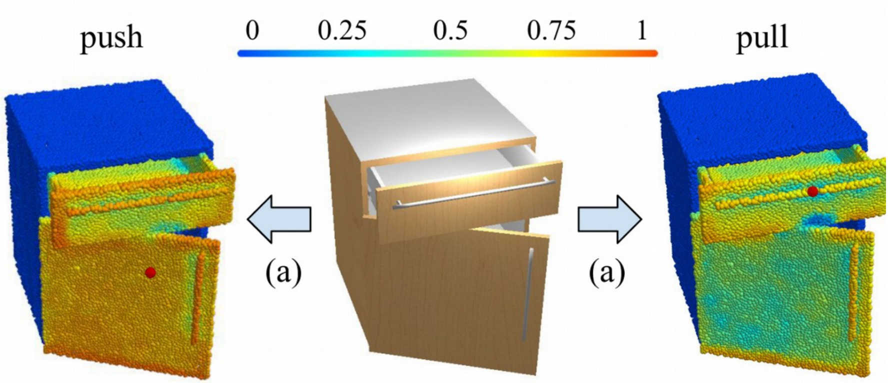




Where2Act: From Pixels to Actions for Articulated 3D Objects

Supervisor: Nick Heppert

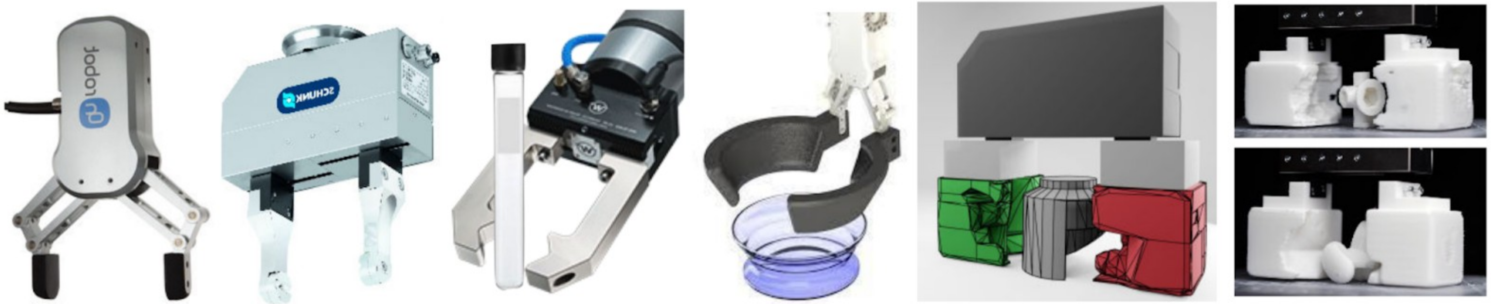
- Interacting with articulated objects (cabinets, ovens, microwaves, etc.) is a core robotic task
- In this work:
  - generative network architecture to generate interaction points
  - online data sampling method used in a learning-from-interaction framework



Fit2Form: 3D Generative Model for Robot Gripper Form Design

Supervisor: Nick Heppert

- Robotic gripper design has a huge influence on grasp performance
- In this work:
  - data-driven gripper design generator
  - learned fitness function

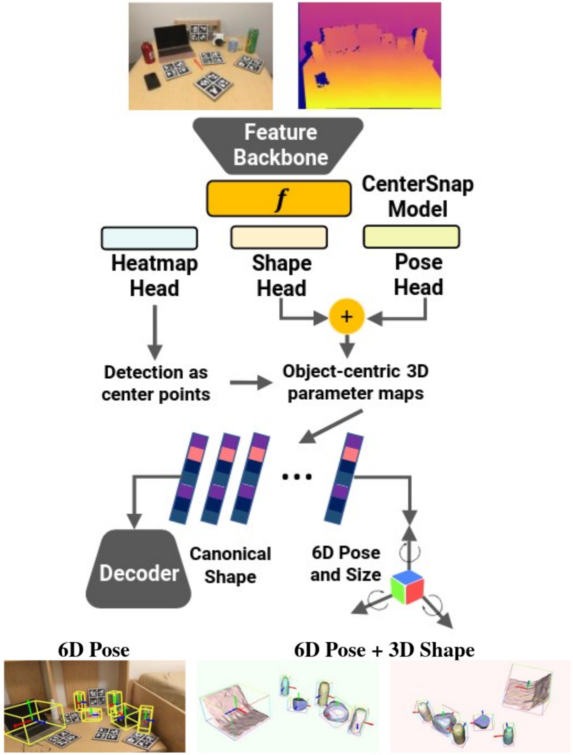


(a) General purpose grippers (b) Task specific grippers (c) Generated gripper designs

CenterSnap: Single-Shot Multi-Object 3D Shape Reconstruction and Categorical 6D Pose and Size Estimation

Supervisor: Eugenio Chisari

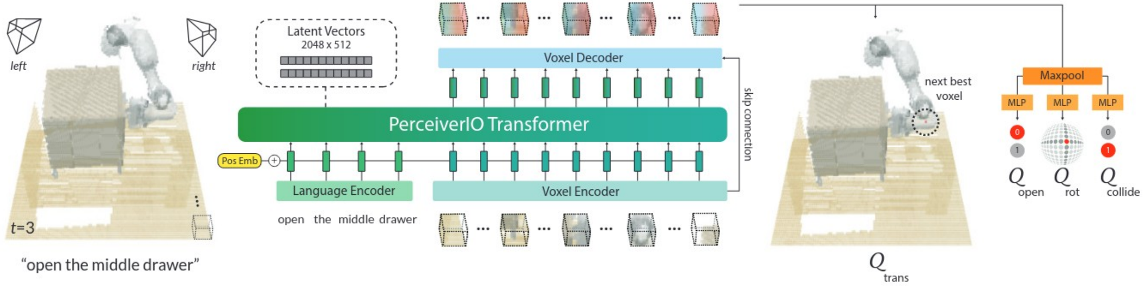
- Many object pose estimation methods assume the availability of a CAD model of the object
- Existing approaches first detect each object instance in the image and then regress their pose or shape
- In this work:
    - Simple one-stage approach to predict both pose and shape of each object in the scene
    - No model of the object is required at inference time. The network can generalize to unseen object instances
    - Bounding-box free detection, by treating each object as a spatial center



Perceiver-Actor: A Multi-Task Transformer for Robotic Manipulation

Supervisor: Eugenio Chisari

- Transformers have revolutionized vision and natural language processing with their ability to scale with large datasets
- In robotic manipulation, data is both limited and expensive
- In this work:
  - PerAct is proposed, a language-conditioned BC agent based on the Perceiver Transformer
  - Voxelized observations and action space provide a strong structural and spatial prior
  - With just few demos per task, a single policy is trained to solve 18 sim and 7 real-world tasks

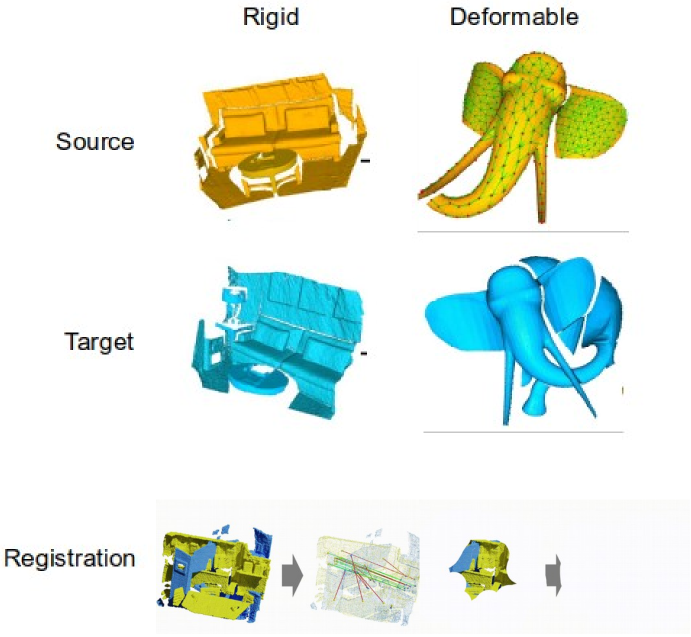


Lepard: Learning partial point cloud matching in rigid and deformable scenes

Supervisor: José Arce

- Matching partial point clouds from range sensors lies at the core of many 3D computer vision applications.
  - Point clouds are usually assumed to be rigid.

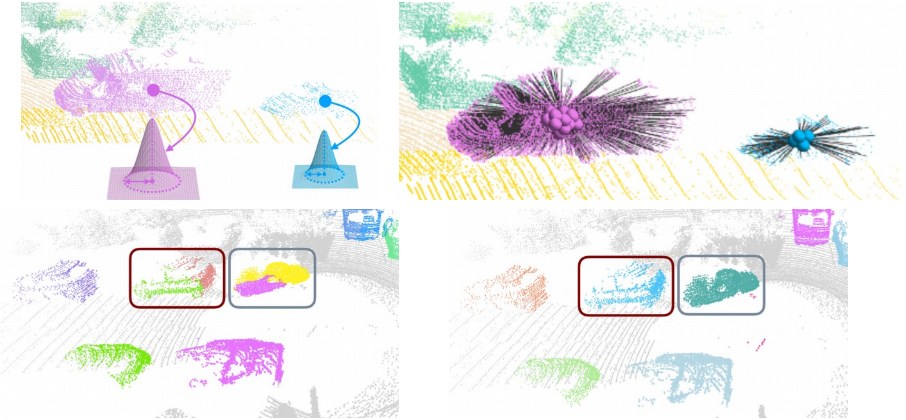
- In this work:
  - Disentanglement of feature and 3D spaces
  - Relative 3D Positional encoding
  - Self- and Cross- attention point matching
  - Repositioning technique for deformations
  - 4DMatch: Benchmark for non-rigid registration



4D-StOP: Panoptic Segmentation of 4D LiDAR using Spatio-temporal Object Proposal Generation and Aggregation

Supervisor: José Arce

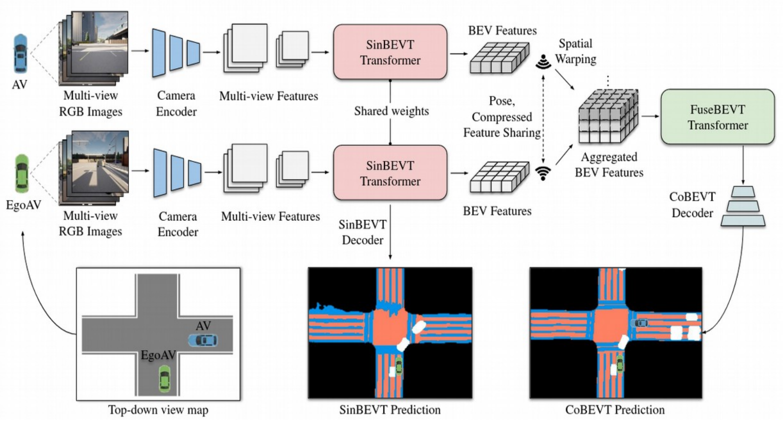
- Class and instance segmentation of point clouds across space-time
  - Temporally consistent instance ID and preserved semantic labels
- In this work:
  - End-to-end 4D Panoptic Segmentation
  - Voting-based center predictions
  - Tracklet aggregation with geometric features



CoBEVT: Cooperative Bird’s Eye View Semantic Segmentation with Sparse Transformers

Supervisor: Nikhil Gosala

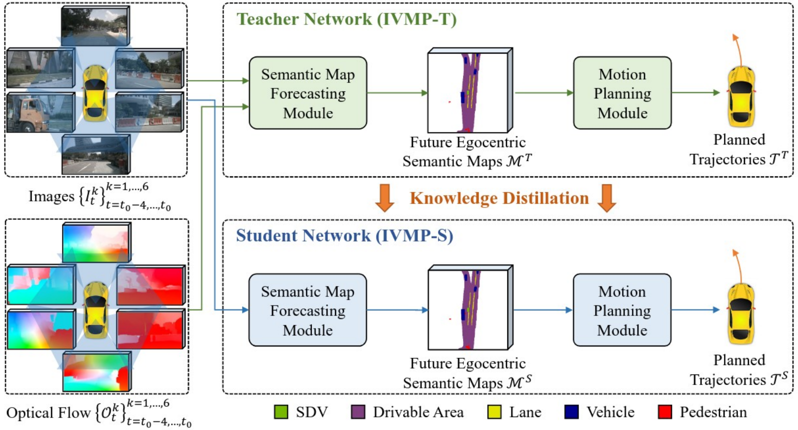
- Prior work on BEV semantic segmentation focuses on single agent systems. These systems struggle with occlusions and distance objects
- Vehicle-to-Vehicle (V2V) communication technologies have enabled autonomous vehicles to share sensing information
- In this work:
  - generic multi-agent multi-camera perception framework that can cooperatively generate BEV map predictions
  - fused axial attention module (FAX), which captures sparsely local and global spatial interactions across views and agents



Learning Interpretable End-to-End Vision-Based Motion Planning for Autonomous Driving with Optical Flow Distillation

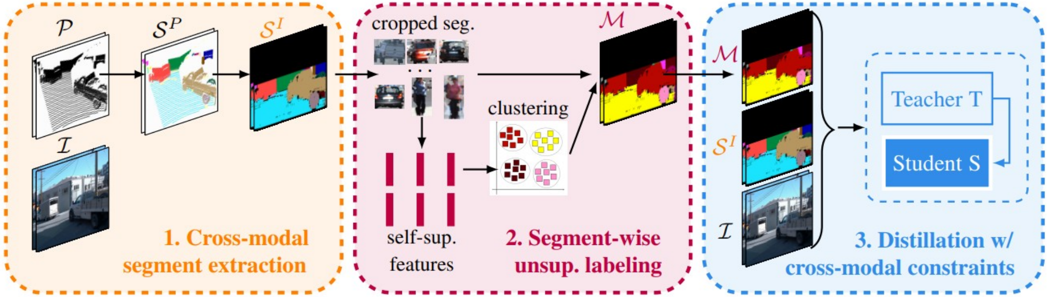
Supervisor: Nikhil Gosala

- End-to-end vision-based methods typically have limited interpretability, limiting their applicability in practice
- In this work:
    - An interpretable end-to-end vision-based motion planning approach is proposed, IVMP
    - IVMP predicts future egocentric maps in BEV space, which are then employed to plan trajectories
    - Additional optical flow distillation paradigm, to enhance the network, still running in real-time


Teacher Network (IVMP-T)
Semantic Map Forecasting Module
Future Egocentric Semantic Maps $\mathcal{M}^T$
Motion Planning Module
Planned Trajectories $\mathcal{T}^T$
Images $\{I_t^k\}_{t=t_0-4,\dots,t_0}^{k=1,\dots,6}$
Knowledge Distillation
Student Network (IVMP-S)
Semantic Map Forecasting Module
Future Egocentric Semantic Maps $\mathcal{M}^S$
Motion Planning Module
Planned Trajectories $\mathcal{T}^S$
Optical Flow $\{O_t^k\}_{t=t_0-4,\dots,t_0}^{k=1,\dots,6}$
🟩 SDV
🟪 Drivable Area
🟨 Lane
🟦 Vehicle
🟥 Pedestrian

Drive&Segment: Unsupervised Semantic Segmentation of Urban Scenes via Cross-modal Distillation

Supervisor:
Rohit Mohan

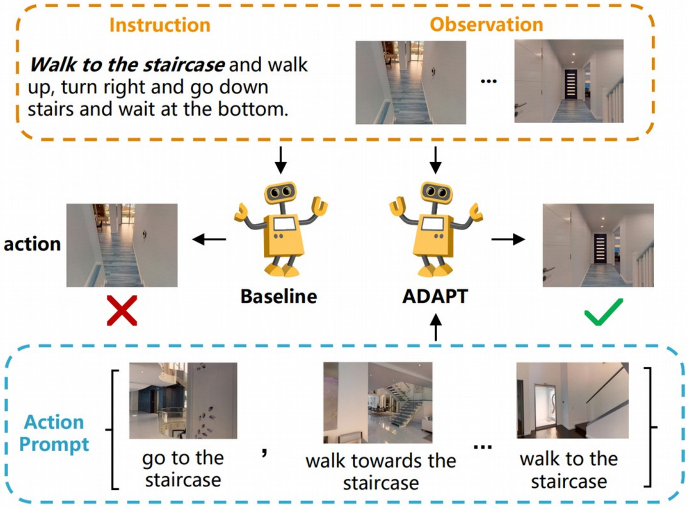


- High-quality labels are cost and time intensive
- Unlabeled data is available in abundance
- In this work:
    - A novel method that leverages synchronized LiDAR and image data
    - Show 3D object proposals can be aligned with the input images and reliably clustered into semantically meaningful pseudo-classes
    - Develop a crossmodal distillation approach that leverages image data partially annotated with the resulting pseudo-classes to train a transformer-based model for image semantic segmentation

ADAPT: Vision-Language Navigation and Modality-Aligned Action Prompts

Supervisor: Rohit Mohan

- Vision-Language Navigation requires an embodied agent to perform action-level modality alignment
- Existing VLN agents learn the instruction-path data directly and cannot sufficiently explore action-level alignment knowledge inside the multi-modal inputs
- In this work:
  - Action prompts enable the explicit learning of action-level modality alignment for successful navigation
  - To collect high-quality action prompts, a Contrastive Language-Image Pretraining (CLIP) model is used, which has powerful cross-modality alignment ability
  - A modality alignment loss and a sequential consistency loss are further introduced to enhance the alignment of the action prompt



Instruction
Observation
Walk to the staircase and walk up, turn right and go down stairs and wait at the bottom.
...
action
Baseline
ADAPT
Action Prompt
go to the staircase
walk towards the staircase
...
walk to the staircase

Questions

Announcement: Hiwi open position

- We have an open position for a hiwi, good opportunity to work on practical robotics, get to know the lab and possibly be considered for master project/thesis

- Good knowledge of C++ and ROS required