Reinforcement learning in robotics

Karel Zimmermann

CTU
CZECH TECHNICAL UNIVERSITY IN PRAGUE

Czech Technical University in Prague
Faculty of Electrical Engineering, Department of Cybernetics

Both tasks formalised as reinforcement learning problems

States: $\mathbf{x} \in \mathcal{R}^n$

Both tasks formalised as reinforcement learning problems

States: $\mathbf{x} \in \mathcal{R}^n$

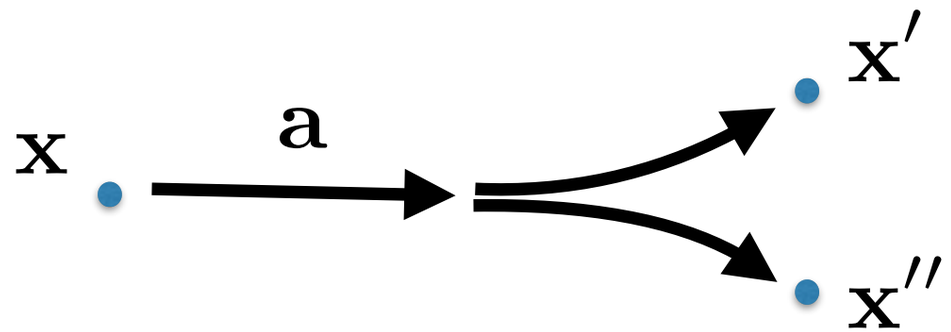
Actions: $\mathbf{a} \in \mathcal{R}^m$

Both tasks formalised as reinforcement learning problems

States: $\mathbf{x} \in \mathcal{R}^n$

Actions: $\mathbf{a} \in \mathcal{R}^m$

Model: $p(\mathbf{x}'|\mathbf{x}, \mathbf{a})$

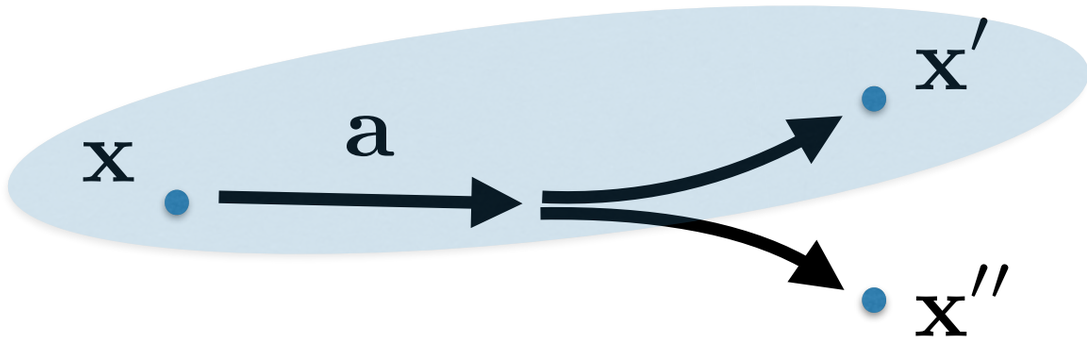


Both tasks formalised as reinforcement learning problems

States: $\mathbf{x} \in \mathcal{R}^n$

Actions: $\mathbf{a} \in \mathcal{R}^m$

Model: $p(\mathbf{x}'|\mathbf{x}, \mathbf{a})$



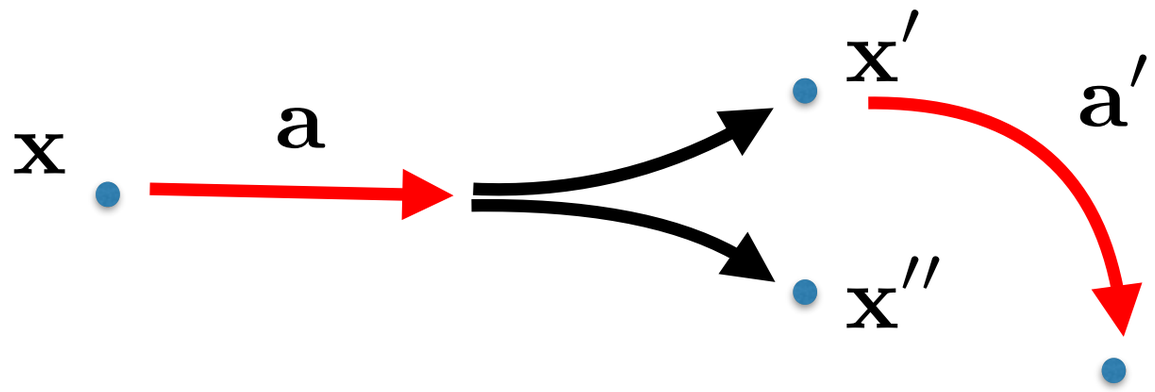
Rewards: $r(\mathbf{x}, \mathbf{a}, \mathbf{x}') \in \mathcal{R}$

Both tasks formalised as reinforcement learning problems

States: $\mathbf{x} \in \mathcal{R}^n$

Actions: $\mathbf{a} \in \mathcal{R}^m$

Model: $p(\mathbf{x}'|\mathbf{x}, \mathbf{a})$



Rewards: $r(\mathbf{x}, \mathbf{a}, \mathbf{x}') \in \mathcal{R}$

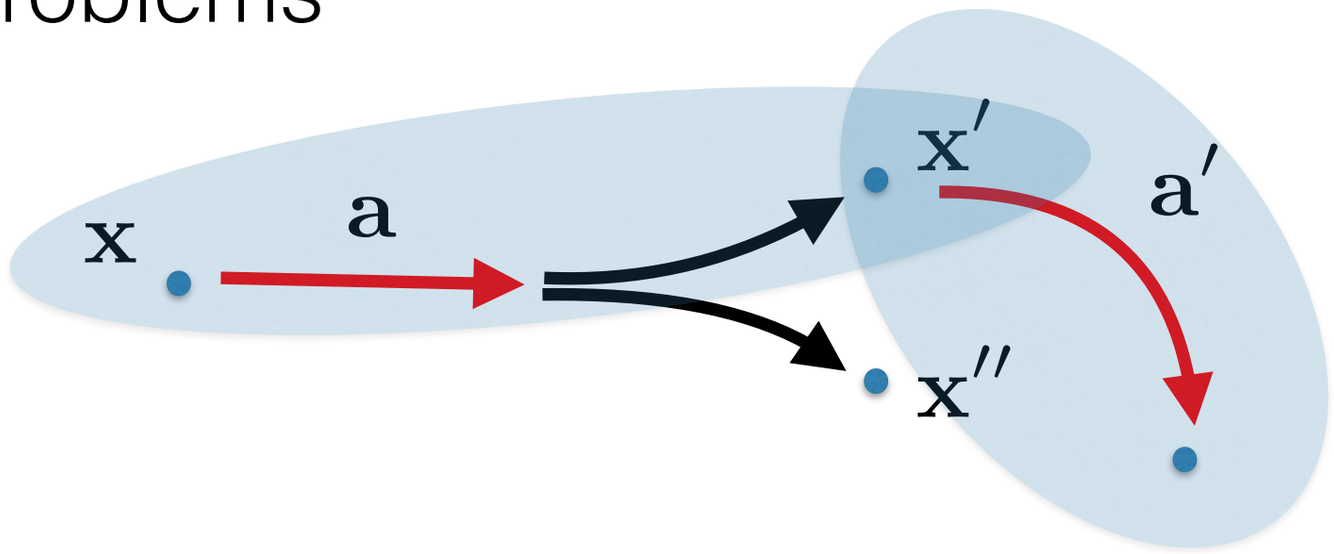
Policy: $\pi(\mathbf{a}|\mathbf{x})$

Both tasks formalised as reinforcement learning problems

States: $\mathbf{x} \in \mathcal{R}^n$

Actions: $\mathbf{a} \in \mathcal{R}^m$

Model: $p(\mathbf{x}'|\mathbf{x}, \mathbf{a})$



Rewards: $r(\mathbf{x}, \mathbf{a}, \mathbf{x}') \in \mathcal{R}$

Policy: $\pi(\mathbf{a}|\mathbf{x})$

Goal: $\pi^* = \arg\max_{\pi} J_\pi$ (e.g. $J_\pi = \mathtt{E}\left[\sum_{t=0}^{T} r_t\right]$)

Challenges of reinforcement learning for robotics

States: $\mathbf{x} \in \mathcal{R}^n$ incomplete, noisy

Actions: $\mathbf{a} \in \mathcal{R}^m$

Model: $p(\mathbf{x}'|\mathbf{x}, \mathbf{a})$

Rewards: $r(\mathbf{x}, \mathbf{a}, \mathbf{x}') \in \mathcal{R}$

Policy: $\pi(\mathbf{a}|\mathbf{x})$

Goal: $\pi^* = \arg\max_\pi J_\pi$ (e.g. $J_\pi = \mathtt{E}\left[\sum_{t=0}^{T} r_t\right]$)

Challenges of reinforcement learning for robotics

States: $\mathbf{x} \in \mathcal{R}^n$ incomplete, noisy

Actions: $\mathbf{a} \in \mathcal{R}^m$ continuous high-dimensional

Model: $p(\mathbf{x}'|\mathbf{x}, \mathbf{a})$

Rewards: $r(\mathbf{x}, \mathbf{a}, \mathbf{x}') \in \mathcal{R}$

Policy: $\pi(\mathbf{a}|\mathbf{x})$

Goal: $\pi^* = \arg\max_\pi J_\pi$ (e.g. $J_\pi = \mathtt{E}\left[\sum_{t=0}^{T} r_t\right]$)

Challenges of reinforcement learning for robotics

States: $\mathbf{x} \in \mathcal{R}^n$ incomplete, noisy

Actions: $\mathbf{a} \in \mathcal{R}^m$ continuous high-dimensional

Model: $p(\mathbf{x}'|\mathbf{x}, \mathbf{a})$ inaccurate model

Rewards: $r(\mathbf{x}, \mathbf{a}, \mathbf{x}') \in \mathcal{R}$

Policy: $\pi(\mathbf{a}|\mathbf{x})$

Goal: $\pi^* = \arg\max_\pi J_\pi$ (e.g. $J_\pi = \mathtt{E}\left[\sum_{t=0}^{T} r_t\right]$)

Challenges of reinforcement learning for robotics

States: $\mathbf{x} \in \mathcal{R}^n$ incomplete, noisy

Actions: $\mathbf{a} \in \mathcal{R}^m$ continuous high-dimensional

Model: $p(\mathbf{x}'|\mathbf{x}, \mathbf{a})$ inaccurate model

Rewards: $r(\mathbf{x}, \mathbf{a}, \mathbf{x}') \in \mathcal{R}$ hard to engineer

Policy: $\pi(\mathbf{a}|\mathbf{x})$

Goal: $\pi^* = \arg\max_\pi J_\pi$ (e.g. $J_\pi = \mathtt{E}\left[\sum_{t=0}^{T} r_t\right]$)

Challenges of reinforcement learning for robotics

States: $\mathbf{x} \in \mathcal{R}^n$ incomplete, noisy

Actions: $\mathbf{a} \in \mathcal{R}^m$ continuous high-dimensional

Model: $p(\mathbf{x}'|\mathbf{x}, \mathbf{a})$ inaccurate model

Rewards: $r(\mathbf{x}, \mathbf{a}, \mathbf{x}') \in \mathcal{R}$ hard to engineer

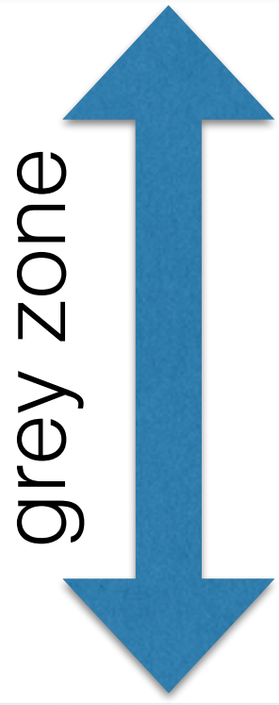
Policy: $\pi(\mathbf{a}|\mathbf{x})$ execution endanger the robot

Goal: $\pi^* = \arg\max_\pi J_\pi$ (e.g. $J_\pi = \mathtt{E}\left[\sum_{t=0}^{T} r_t\right]$)

Taxonomy of policy search methods

- Direct policy search

  e.g. gradient ascent for $\pi^* = \arg\max_{\pi} J_\pi$



Episodic REPS [Peters, 2010]

PILCO [Deisenroth, ICML 2011]

Actor-critic (e.g. DPG [Silver,JMLR 2014])

Deep Q-learning (e.g. [Mnih,Nature 2015])

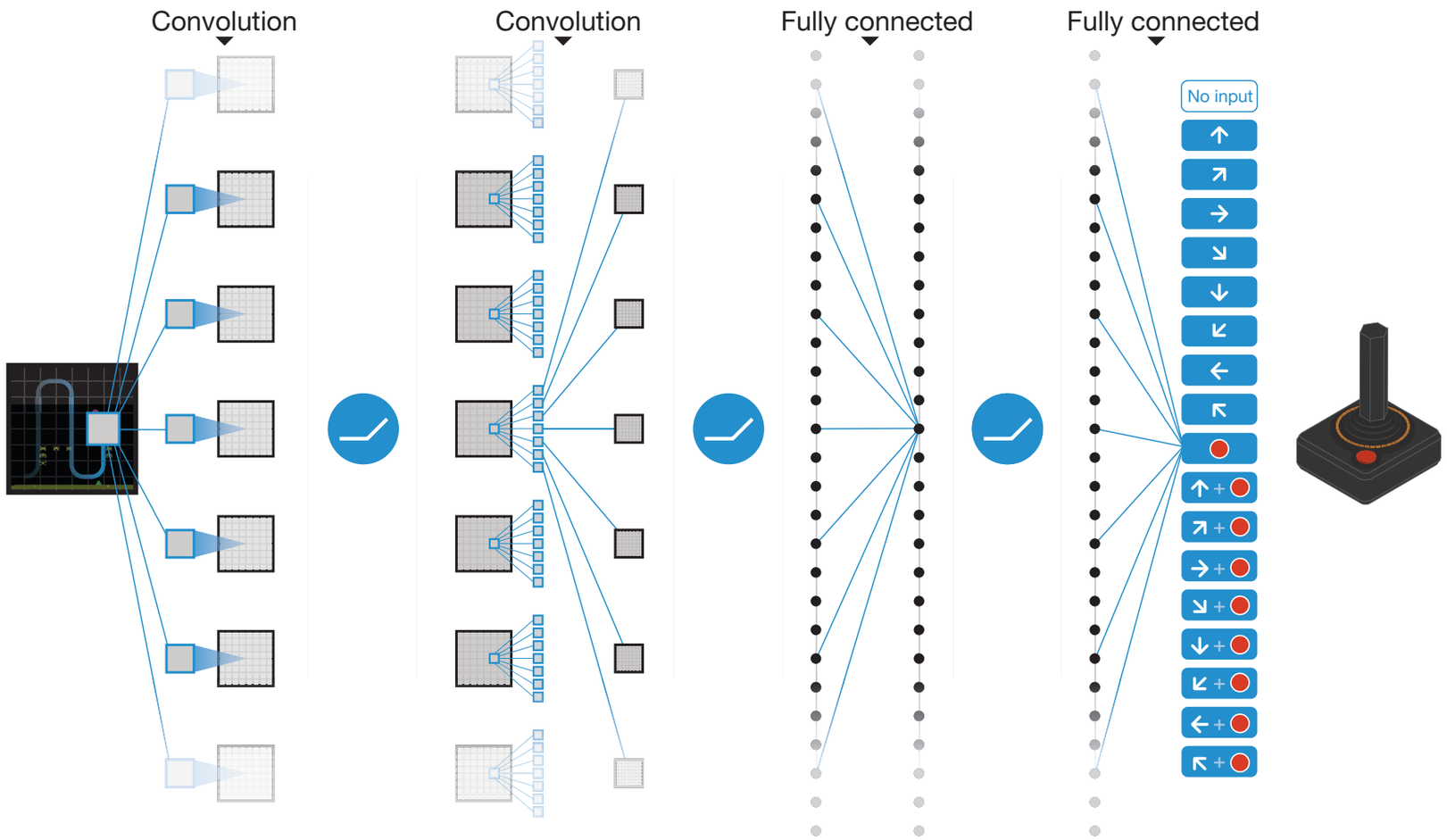
- Value-based methods (dual function [Kober, 2013])

  e.g. search for $Q(\mathbf{x}, \mathbf{a}) = r(\mathbf{x}, \mathbf{a}, \mathbf{x}') + \gamma \max_{\mathbf{a}'} Q(\mathbf{x}', \mathbf{a}')$

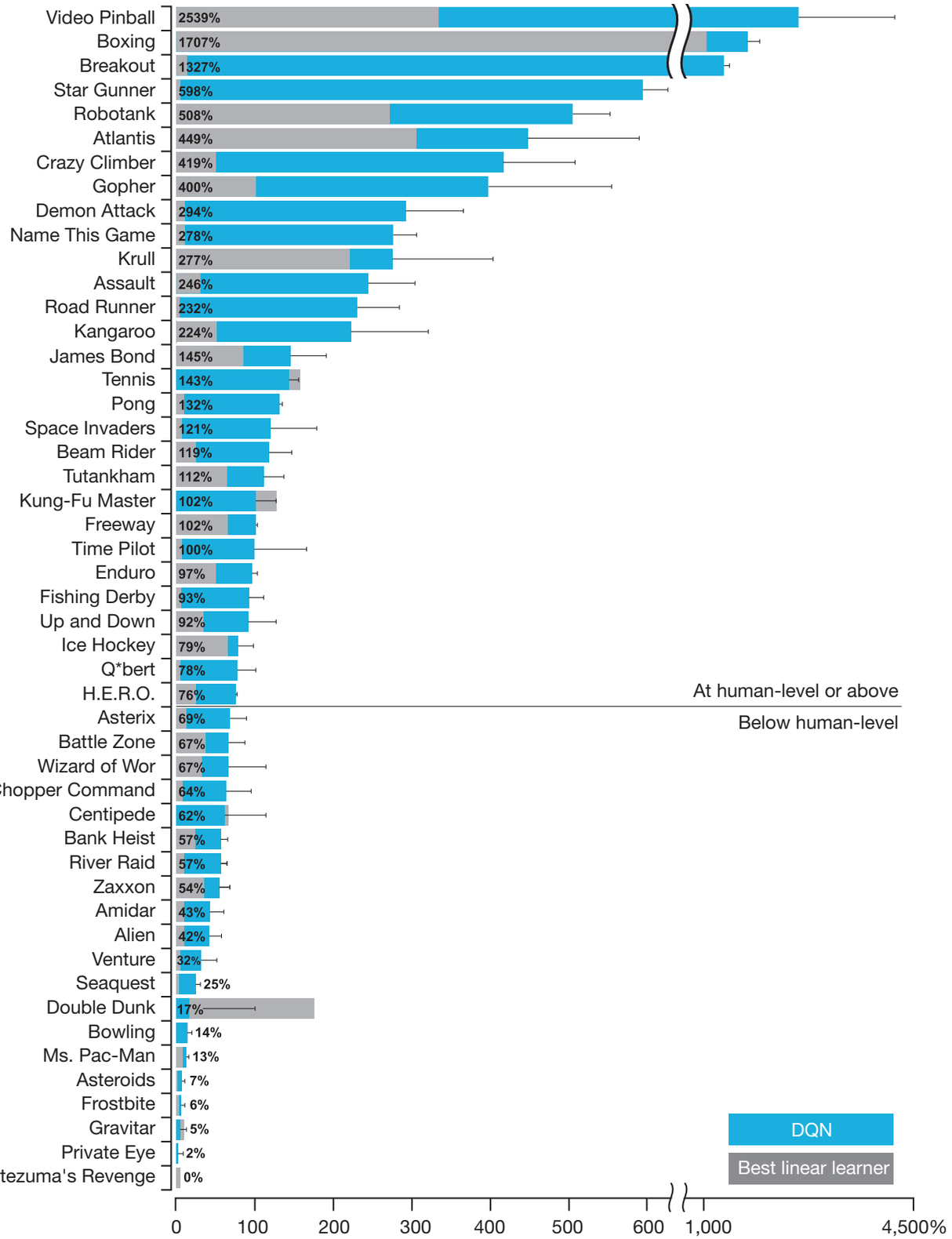
  $$\pi^* = \arg\max_{a} Q(\mathbf{x}, \mathbf{a})$$

Mnih et al. Nature 2015

- 2600 atari games
- **state space:** pixels (e.g. VGA resolution)
- **action space:** discrete joystic actions (8 direction + 8 direction with button)
- collection of control tasks: https://gym.openai.com



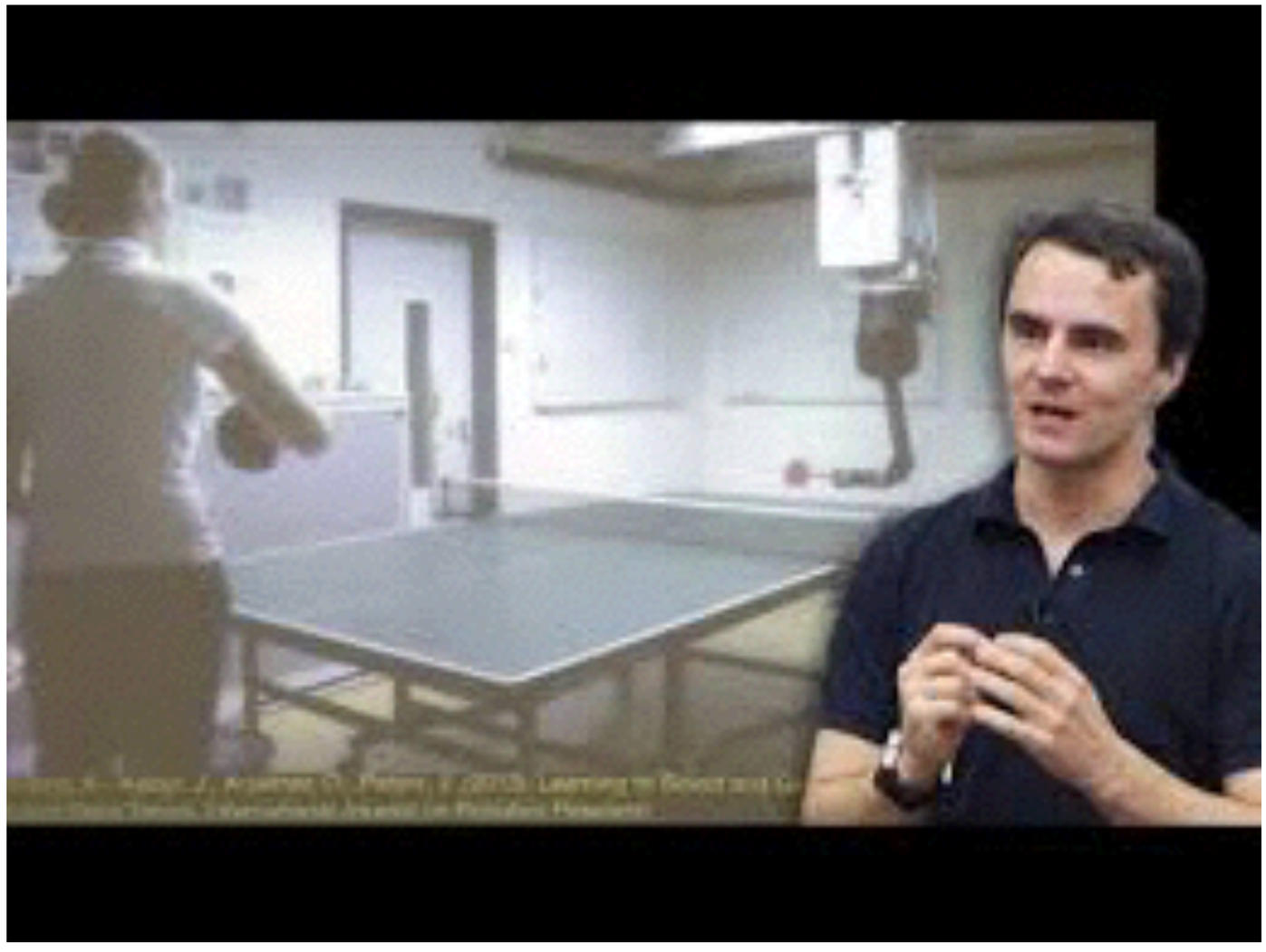
Mnih et al. Nature 2015



Czech Technical University in Prague
Faculty of Electrical Engineering, Department of Cybernetics

Peters et al. NOW 2013

- imitation learning from human demonstration
- **state space:** joint positions, velocities, acceler.
- **action space:** motor torques
- gradient minimization in policy parameter space

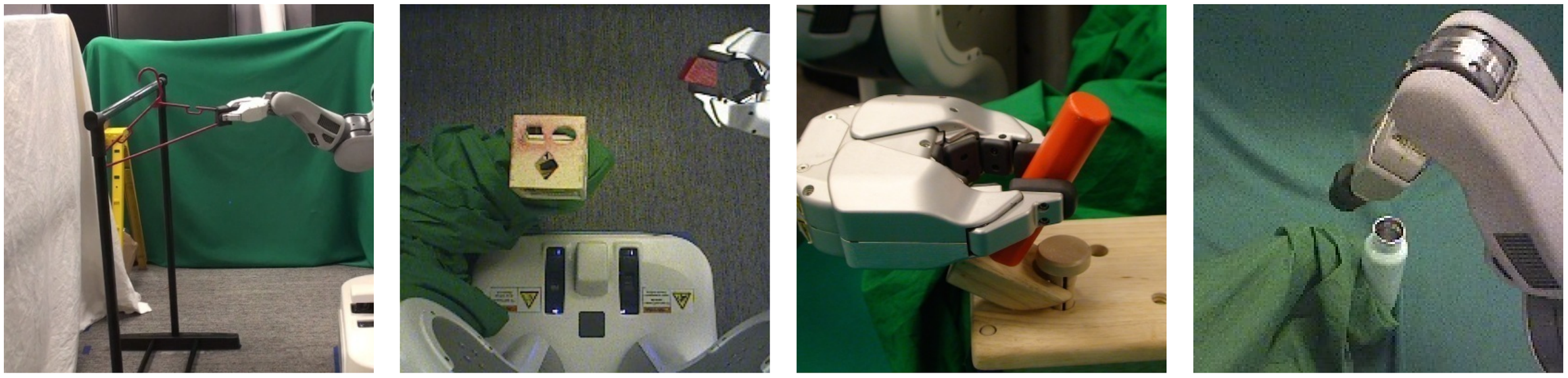


Abbeel et al. IJRR 2010

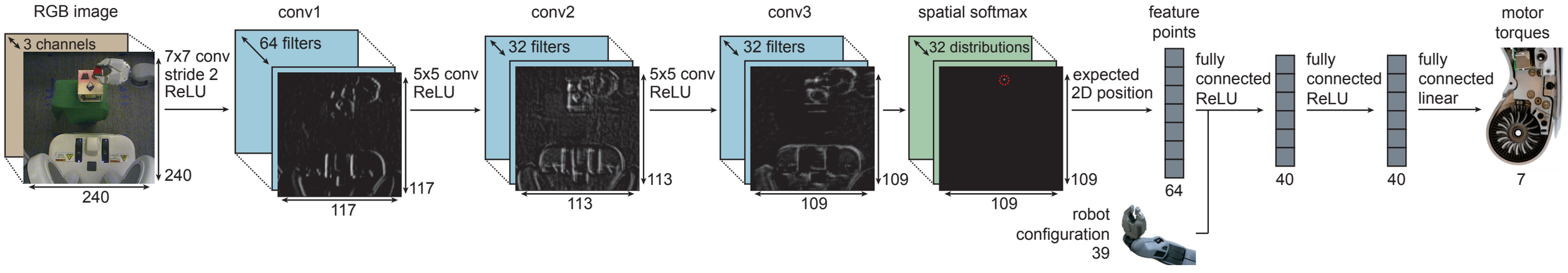
- inverse reinforcment learning
- **state space:** angular and euclidean position, velocity, acceleration
- **action space:** motor torques
- learning reward function from expert pilot

Levine et al JMLR 2016

- guides policy gradient method by optimal trajectories
- **state space:** RGB camera images
- **action space:** motor torques

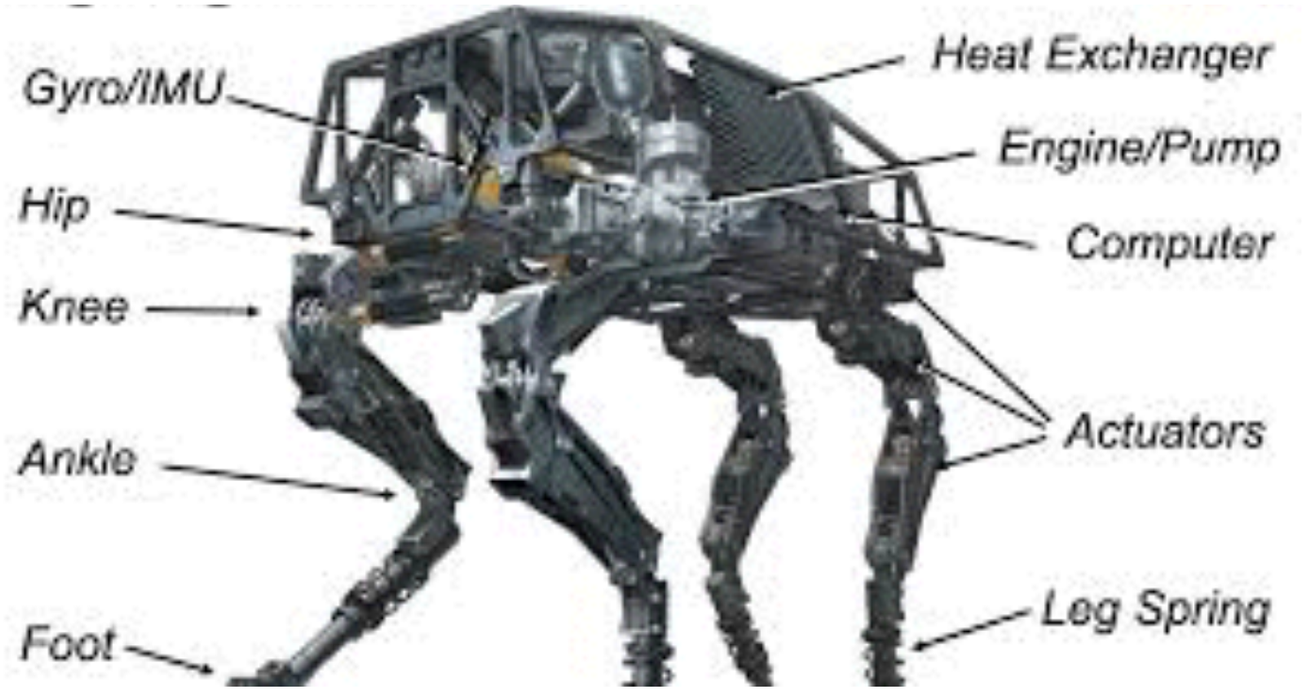


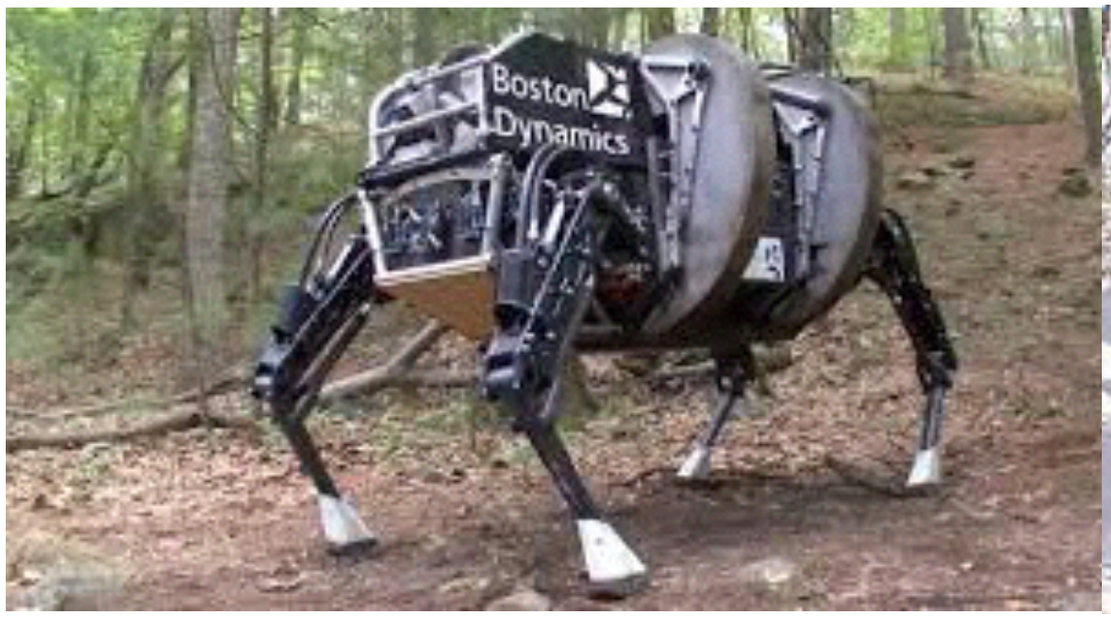
(a) hanger (b) cube (c) hammer (d) bottle

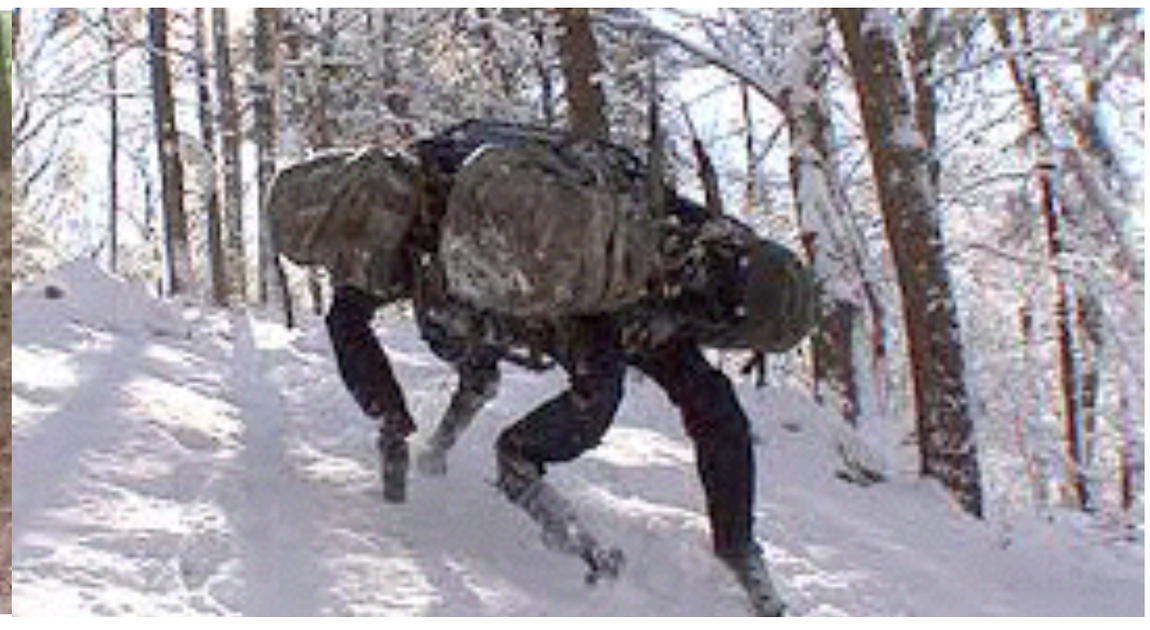


Czech Technical University in Prague
Faculty of Electrical Engineering, Department of Cybernetics

Boston dynamics - big dog





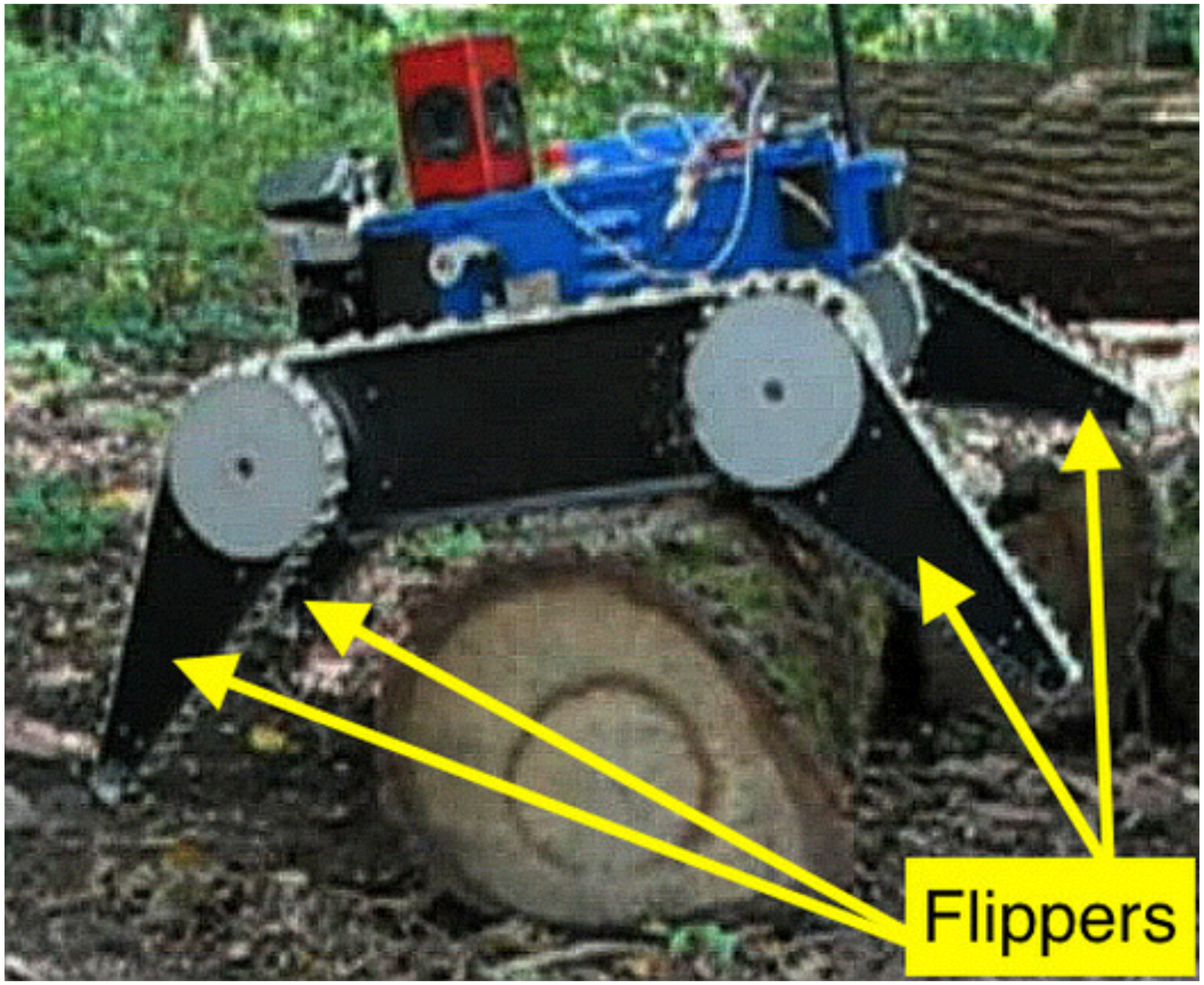


Taxonomy of policy search methods

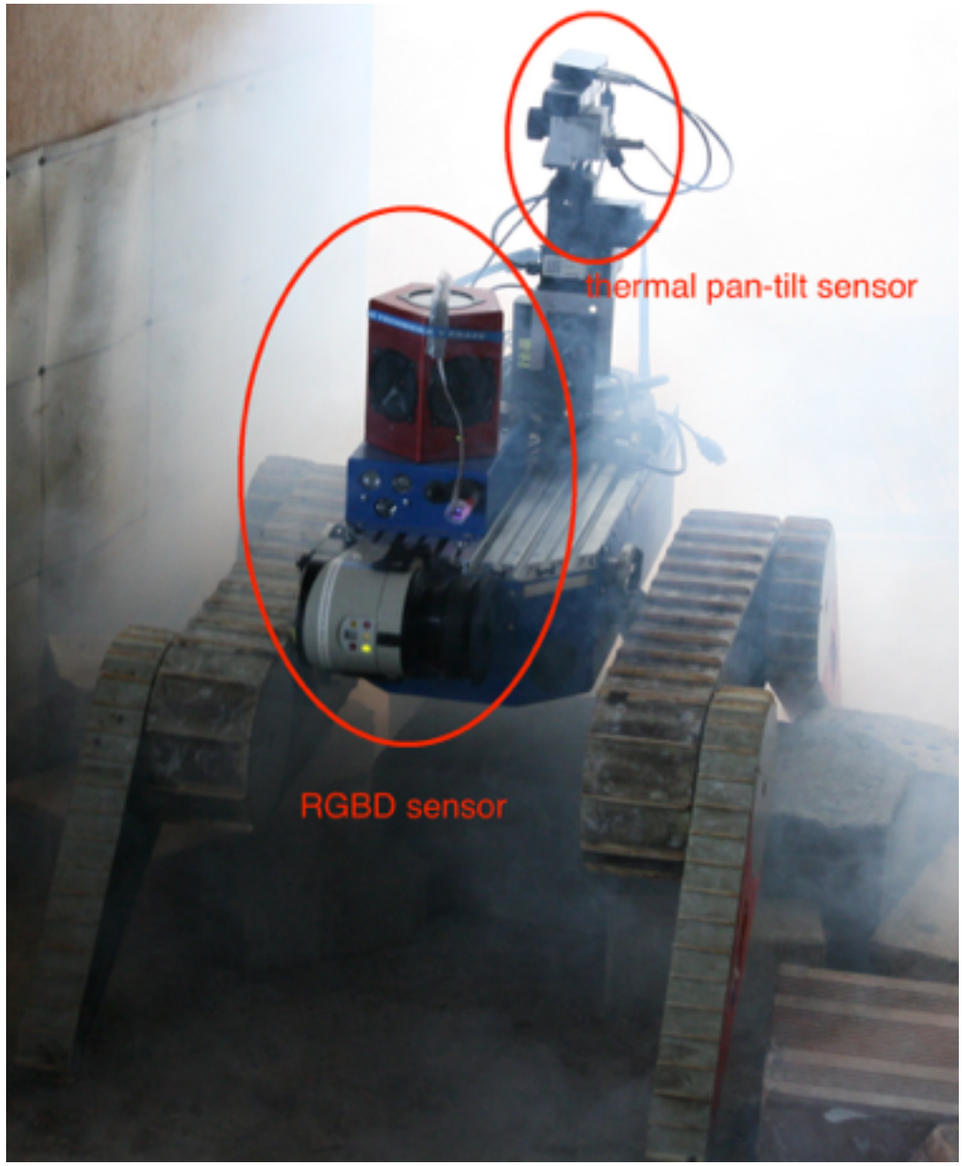
- Model-free methods
  - e.g. REINFORCE [Williams, 1992],
    natural gradients [Peters, 2013]
  - require many samples
  - do not introduce model bias

- Model-based methods
  - e.g. PILCO [Deisenroth, 2011],
    GPREPS [Kupcsik, 2015]
  - suffer from model bias

Search & Rescue mobile robotic platform



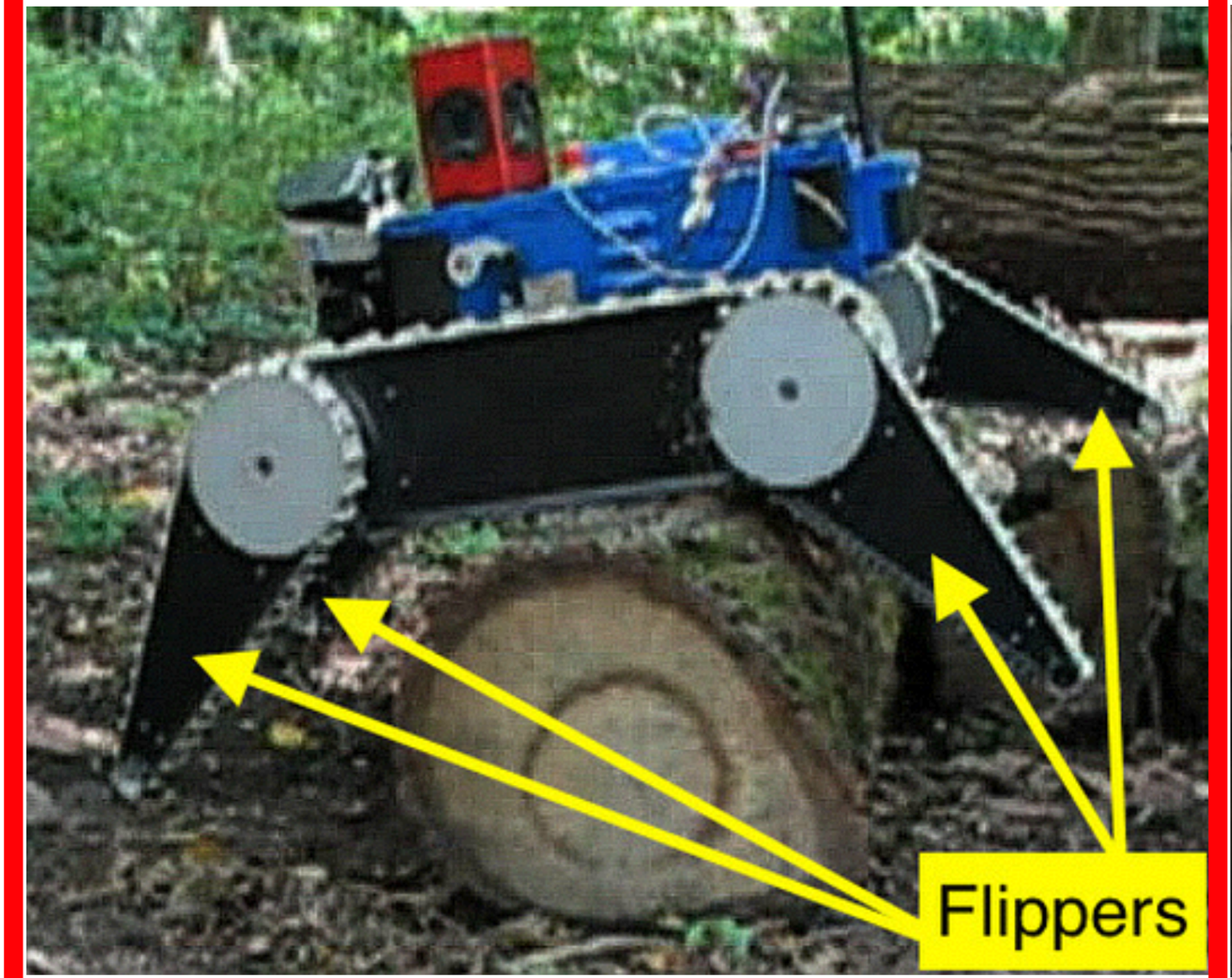
Motion and compliance control of flippers for terrain smooth traversal



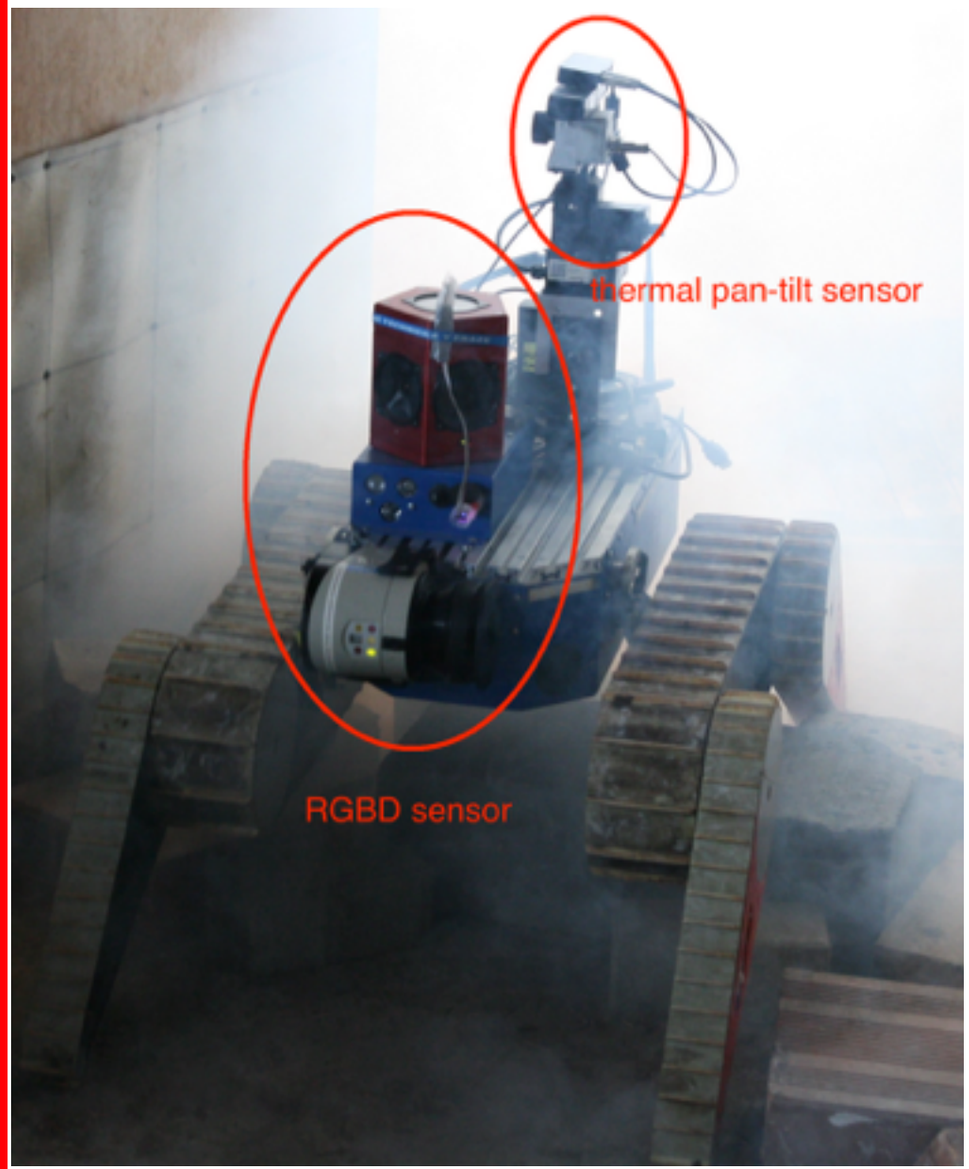
Motion control of thermal camera for semantic segmentation

Czech Technical University in Prague
Faculty of Electrical Engineering, Department of Cybernetics

Search & Rescue mobile robotic platform



Motion and compliance control of flippers for terrain smooth traversal



Motion control of thermal camera for semantic segmentation

Czech Technical University in Prague
Faculty of Electrical Engineering, Department of Cybernetics

Motion and compliance control of flippers

Actions:

- torques in flipper engines
- complience of flippers

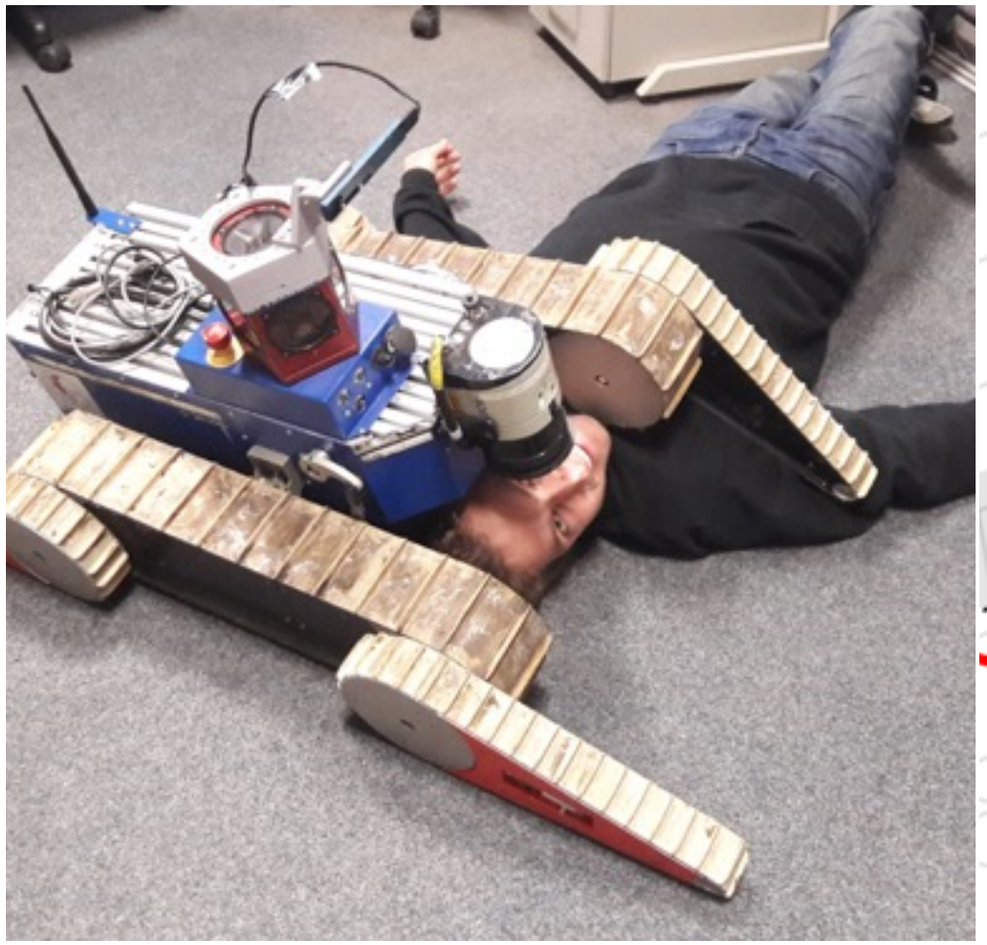
Proprioceptive states: $\mathbf{x}^p$

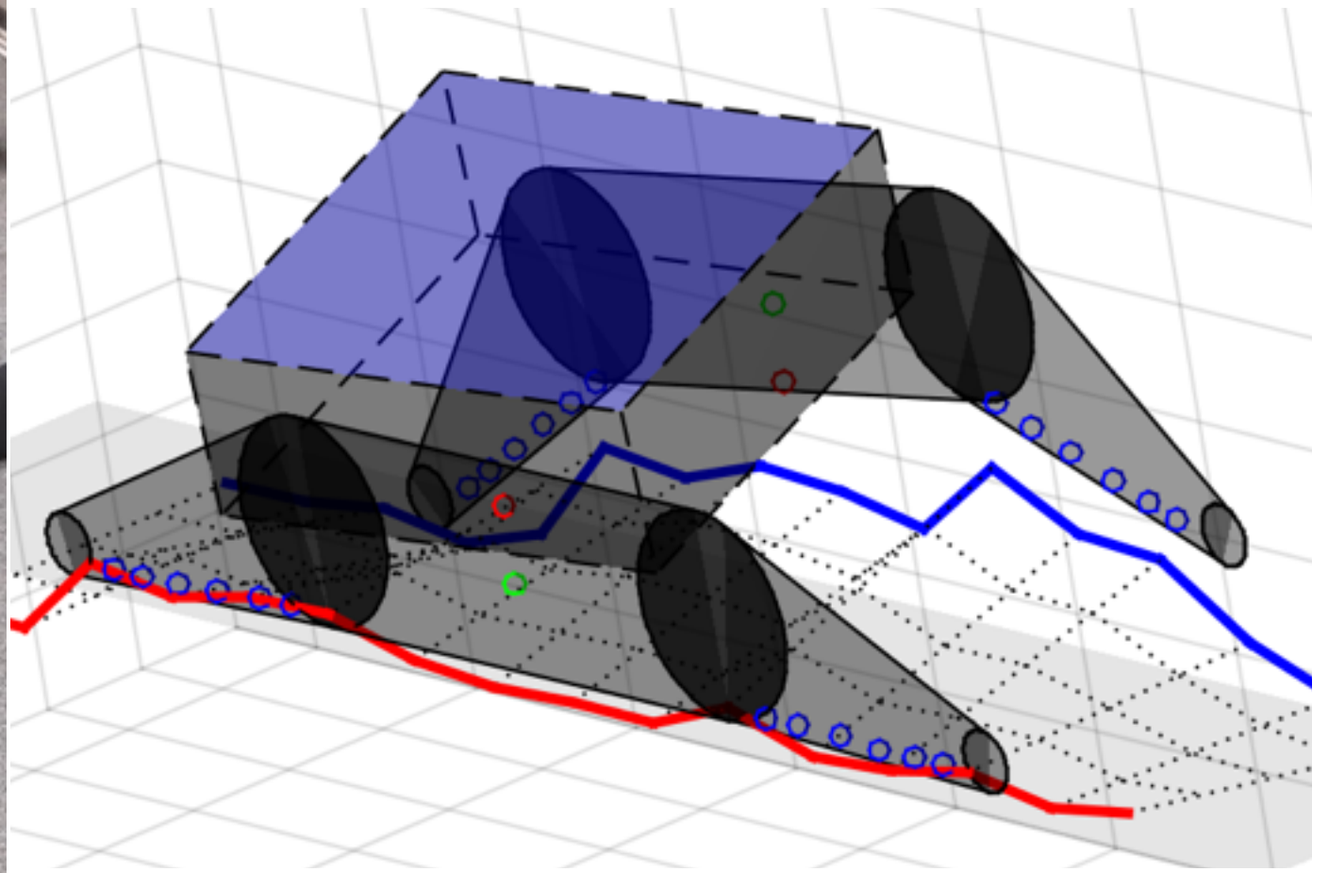
- robot's roll,pitch and current flipper configuration
- torques in engines (4 flippers+2 main tracks)

Motion and compliance control of flippers

Exteroceptive states: $\mathbf{x}^e$

- incomplete local height map obtained by successive mapping from depth data.





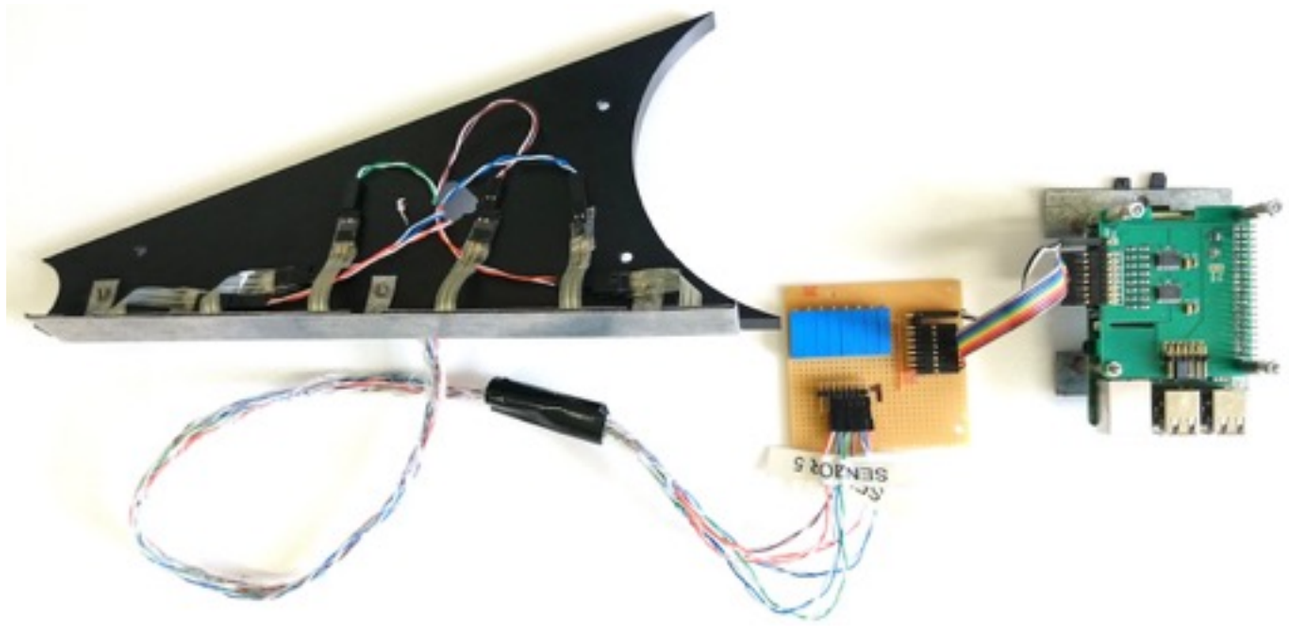
Motion and compliance control of flippers

Czech Technical University in Prague
Faculty of Electrical Engineering, Department of Cybernetics

Motion and compliance control of flippers

Proprioceptive states: $\mathbf{x}^p$

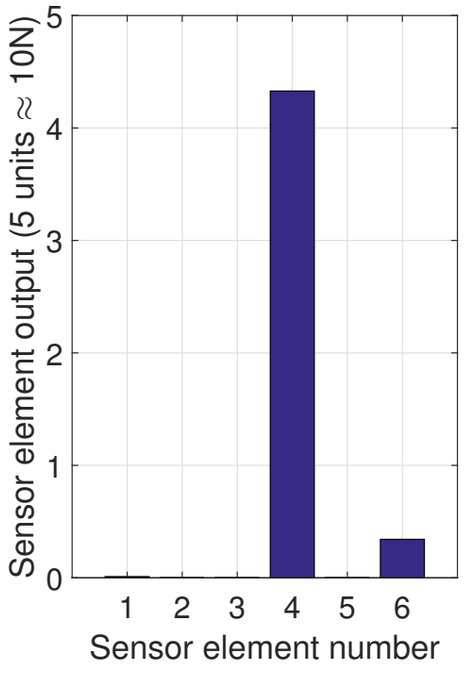
- robot's roll,pitch and current flipper configuration
- torques in engines (4 flippers+2 main tracks)
- **24 tactile sensors (6 per flipper)**

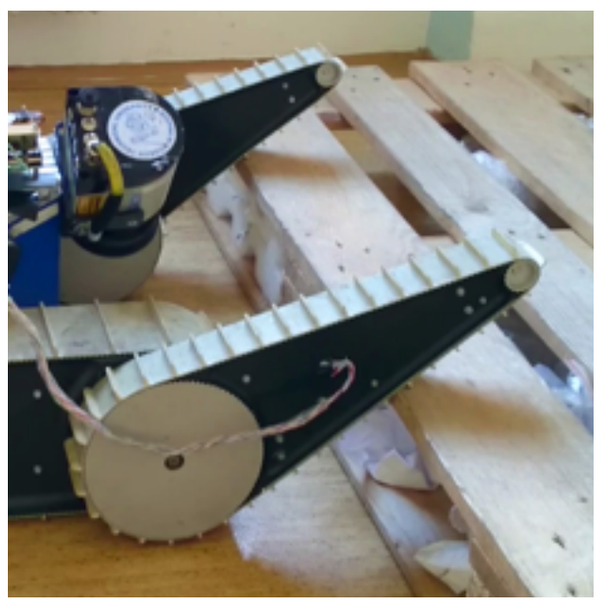


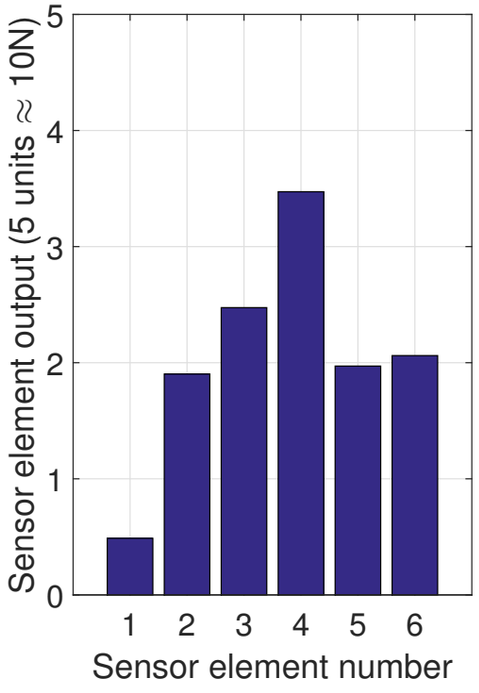
Czech Technical University in Prague
Faculty of Electrical Engineering, Department of Cybernetics

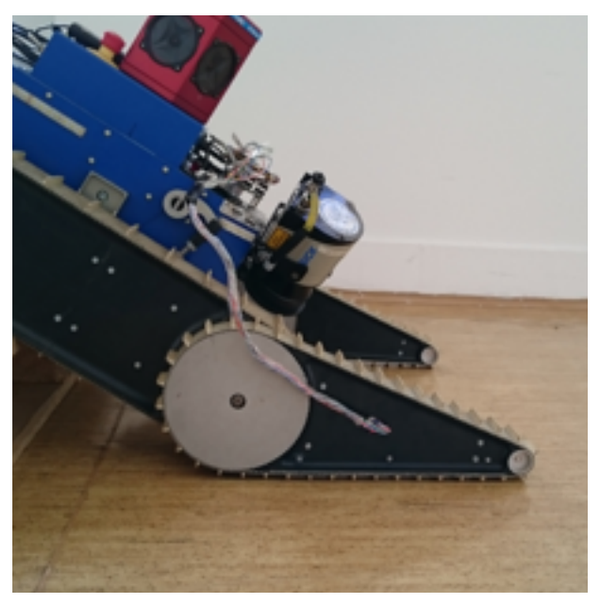
Motion and compliance control of flippers

- rich proprioceptive data often allows tactile reconstruction





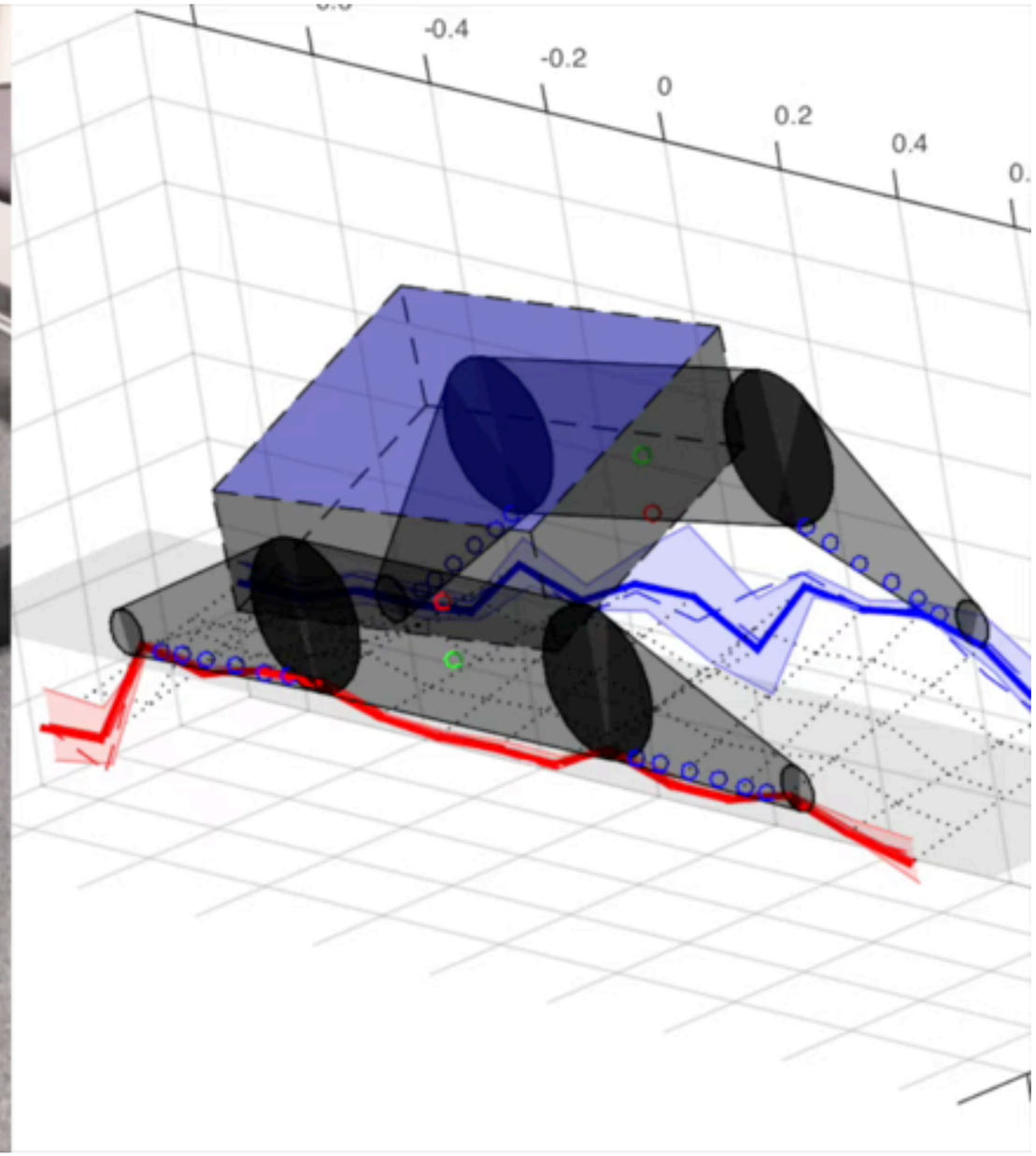




- **offline:** learn conditional probabilities $p(x_i^e | \mathbf{x}^e_{\setminus i}, \mathbf{x}^p)$ from collected trajectories

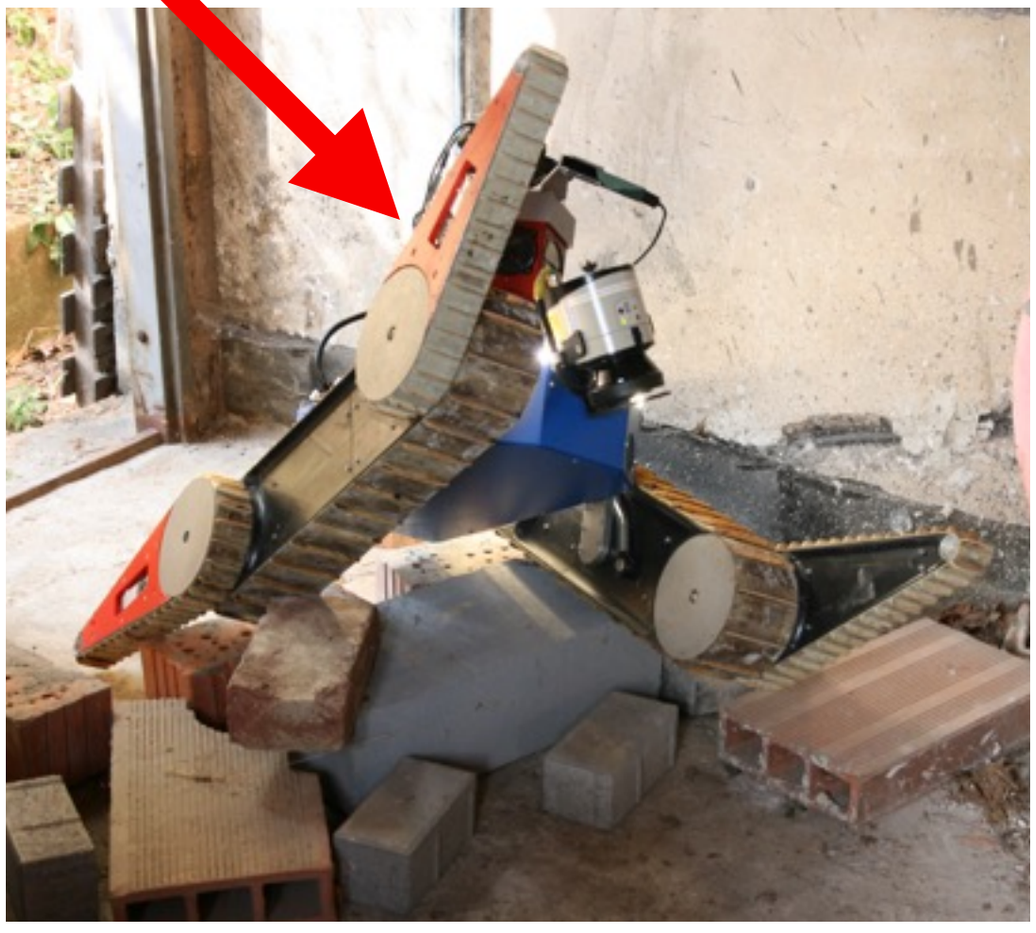
Motion and compliance control of flippers

- **online:** Gibbs sampling from conditional probabilities.



Motion and compliance control of flippers

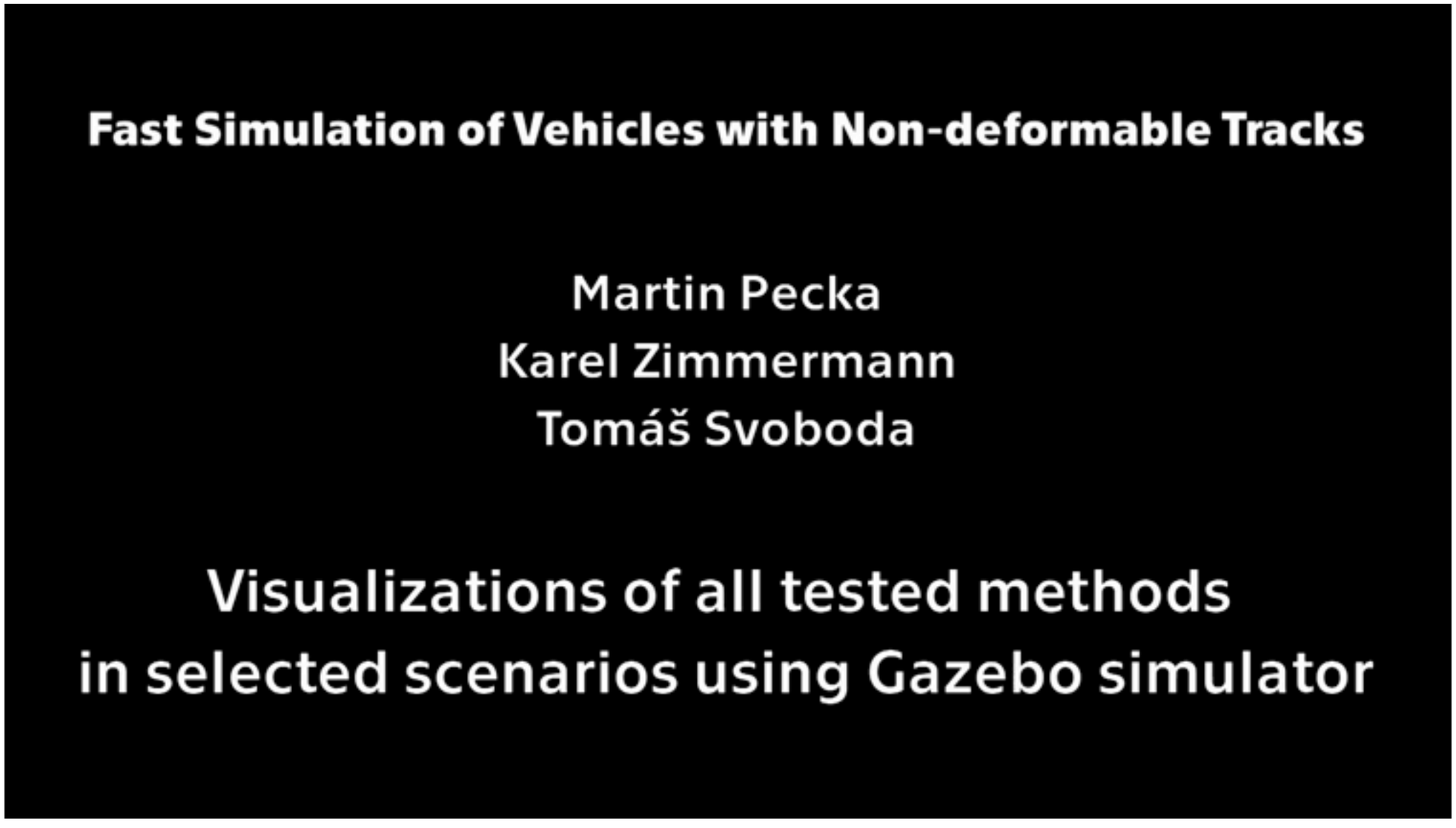
- Flippers provides both the motion and the perception.
- Learn perception-friendly policy for traversing obstacles.
- Real robot, real danger, limited number of real-world trials



Czech Technical University in Prague
Faculty of Electrical Engineering, Department of Cybernetics

Motion and compliance control of flippers

- We speed up and control the learning process by:
  - Initialize policy on physical simulator
  - Incorporating expert heuristics (feasible trajectories for tough obstacles, motion roughness, safety in simulator)



Czech Technical University in Prague
Faculty of Electrical Engineering, Department of Cybernetics

Motion and compliance control of flippers

We propose constrained policy gradient search.

[1] M.Pecka, V.Šalanský, K.Zimmermann, T.Svoboda.
Autonomous flipper control with safety constraints, **IROS**, 2016.

Gradient maximisation of rewards:

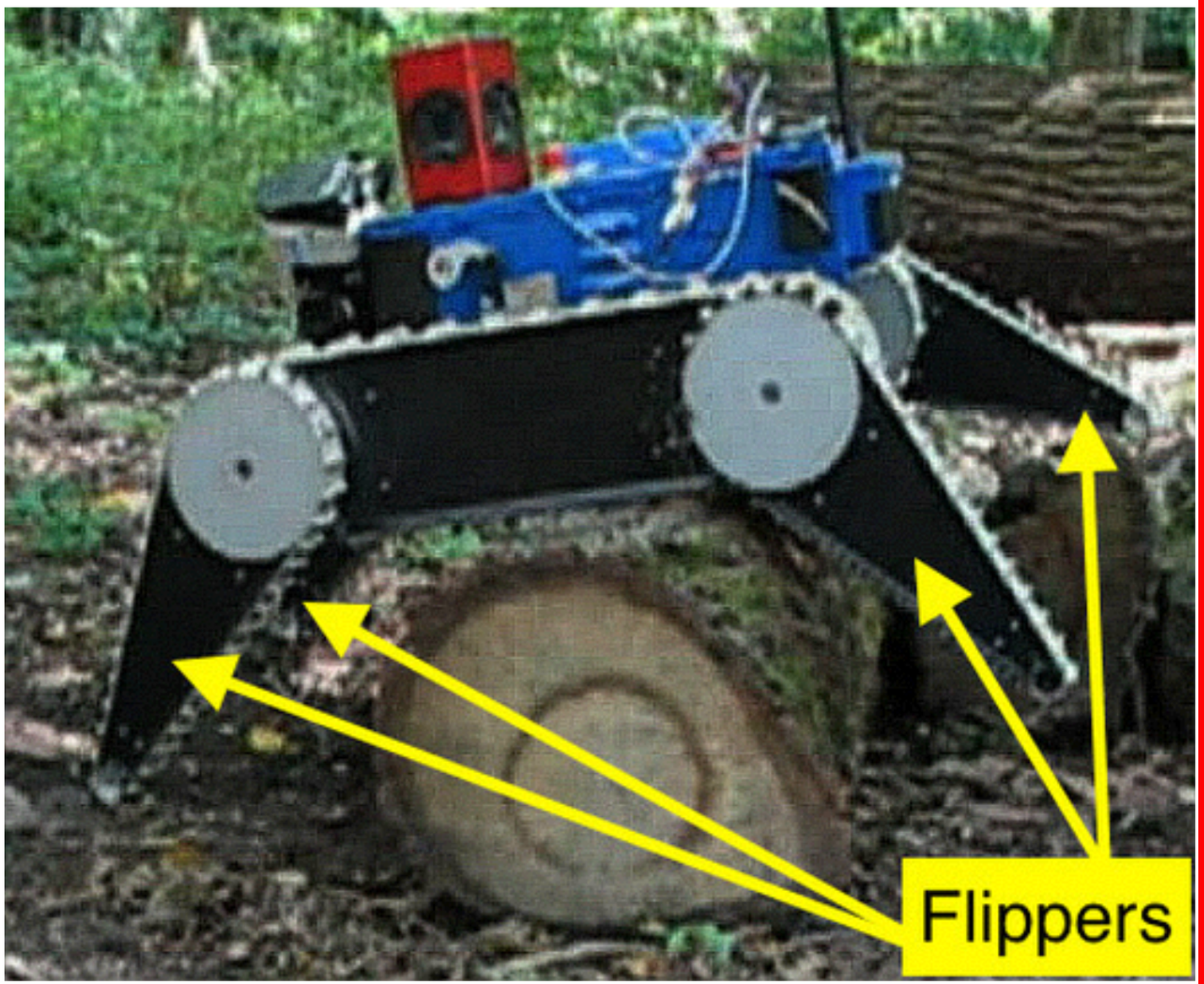
- forward speed

Subject to heuristic constraints:

- tactile reconstruction accuracy
- pitch/roll angle limit (preventing robot's flip-over)
- motion roughness limit measured by accelerometers
- optimal action in a particular state given by an *expert*

Constraints allow for better control of learning process than ad-hoc sum of penalties in the reward function.

Search & Rescue mobile robotic platform

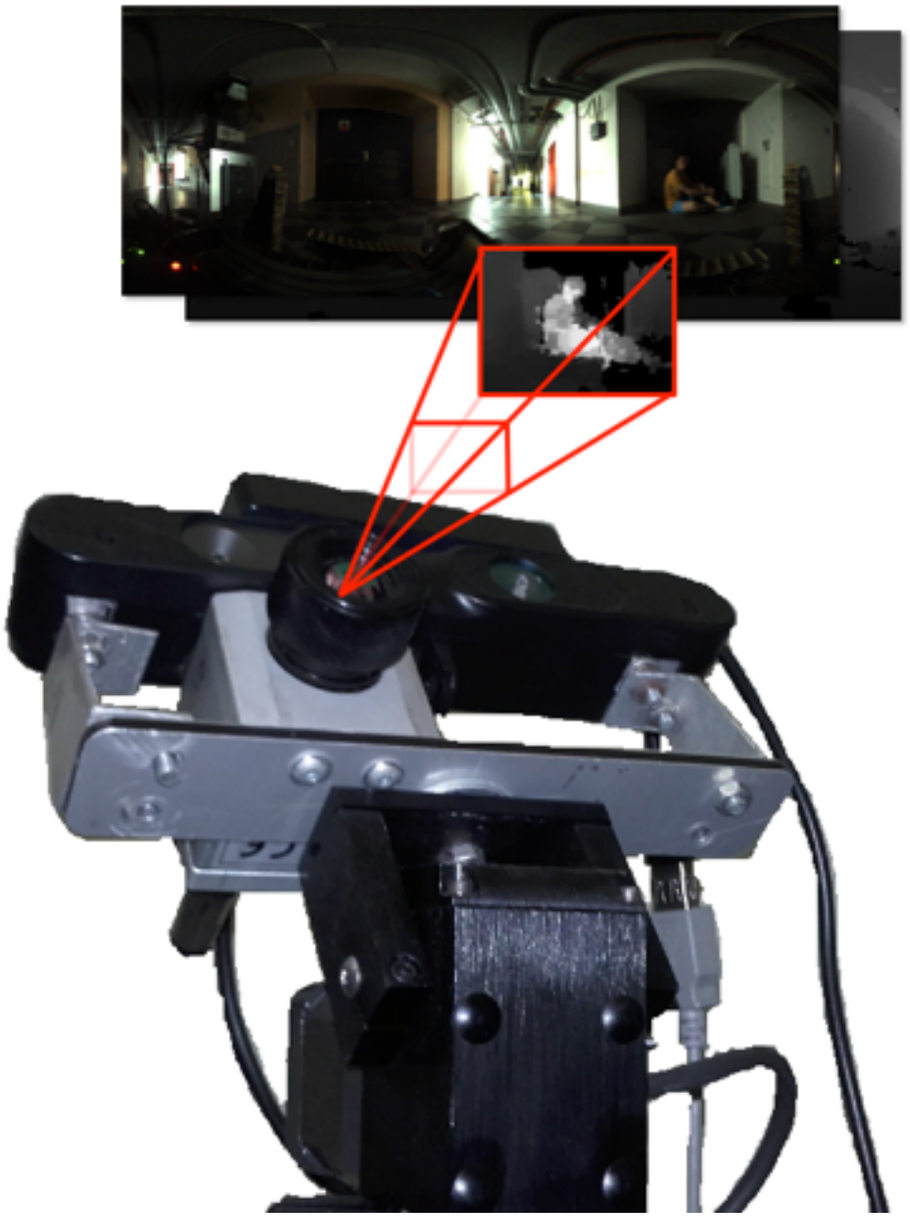


Motion and compliance control of flippers for terrain smooth traversal

Motion control of thermal camera for semantic segmentation

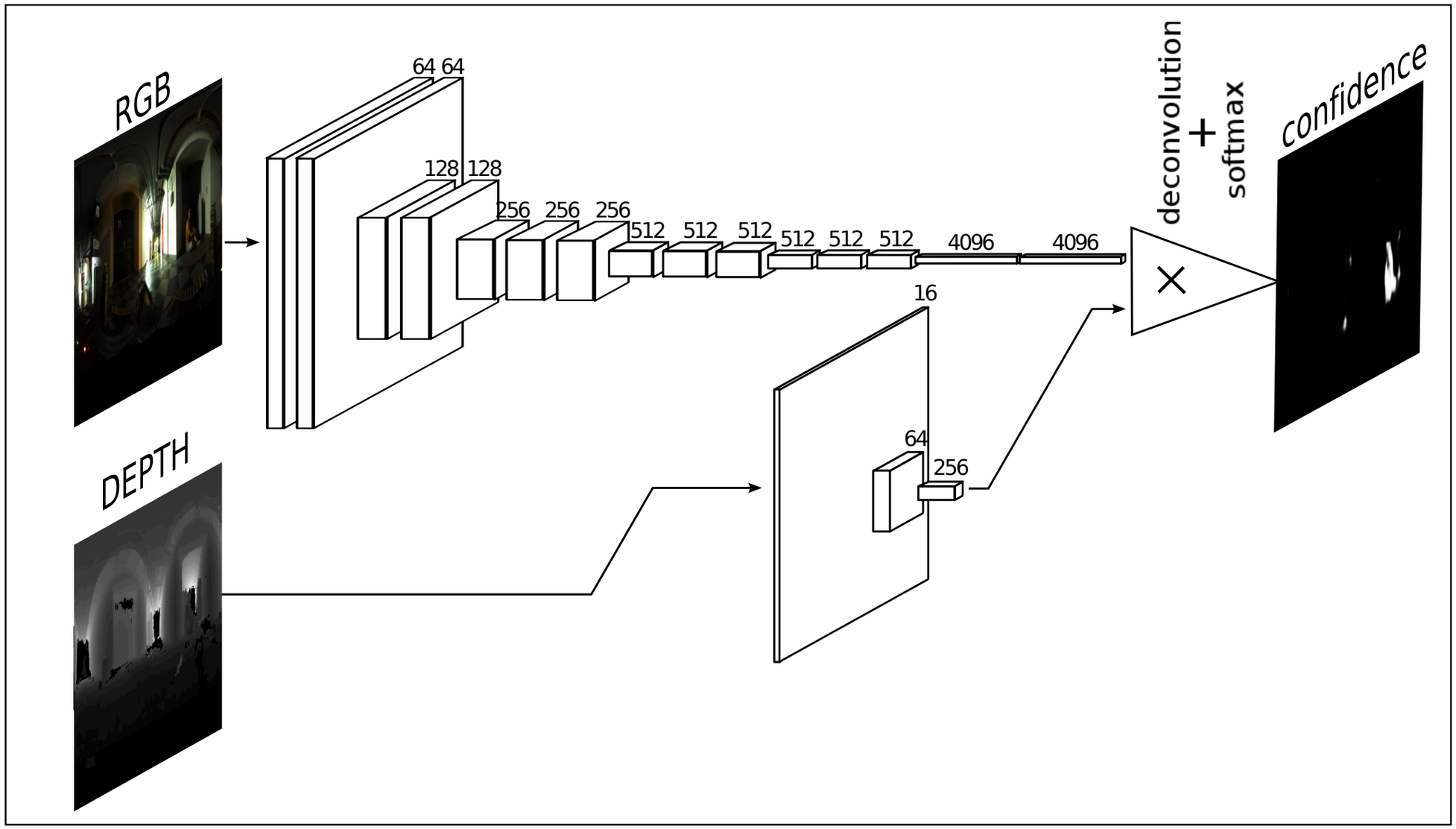
Czech Technical University in Prague
Faculty of Electrical Engineering, Department of Cybernetics

Motion control of thermal camera for segmentation



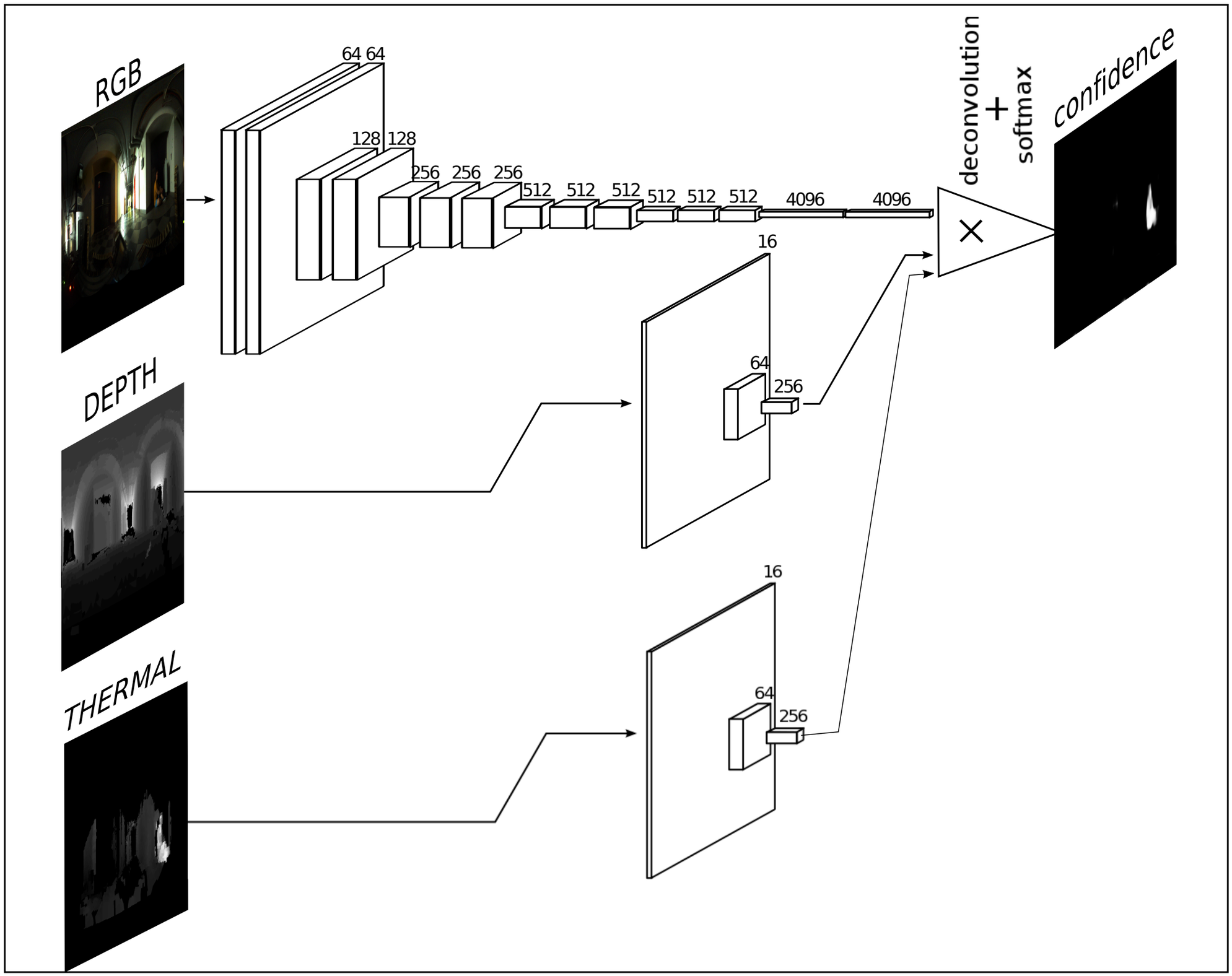
- Robot follows a known exploration path into unknown environment
- **Problem 1**: Segment humans from captured incomplete RGBDT data
- **Problem 2:** Where to look with the thermal sensor to minimize segmentation error?
- **Approach:** Learn simultaneously segmentation and policy deep CNN

Segmentation network RGBD->H

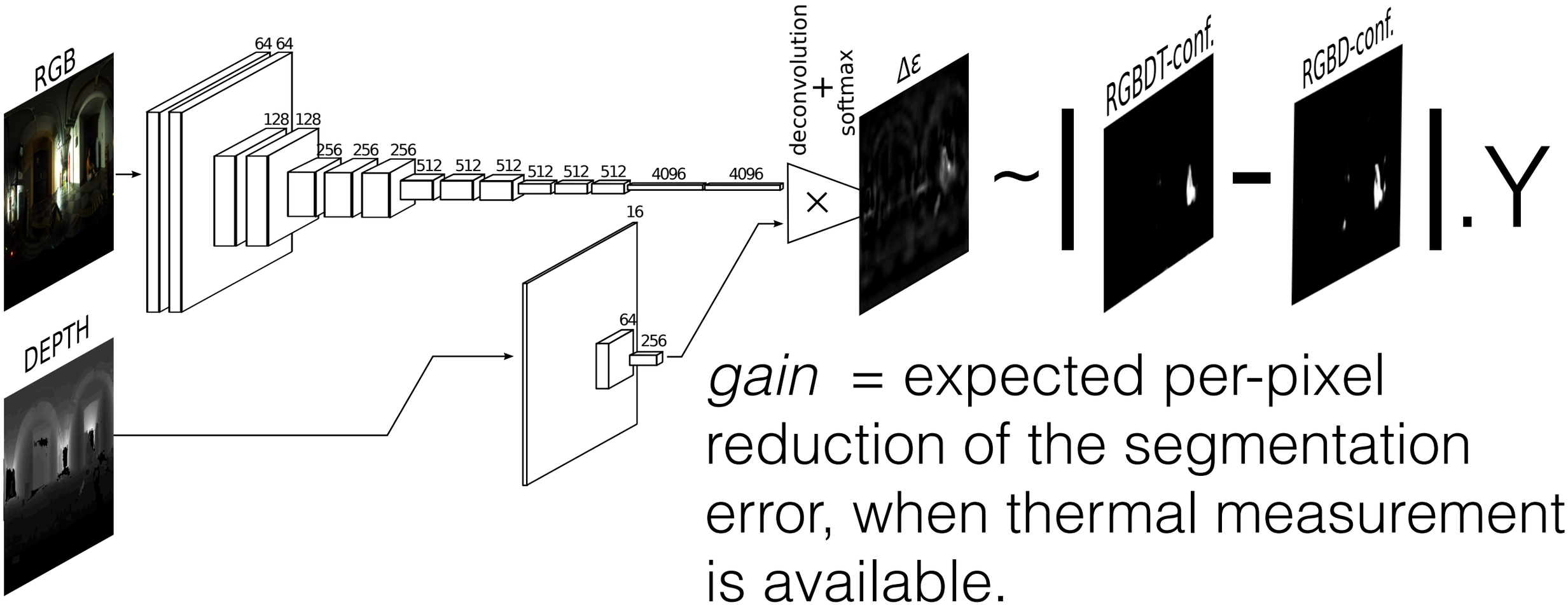


[Long CVPR 2015]'s segmentation network extended by depth and thermal modalities.

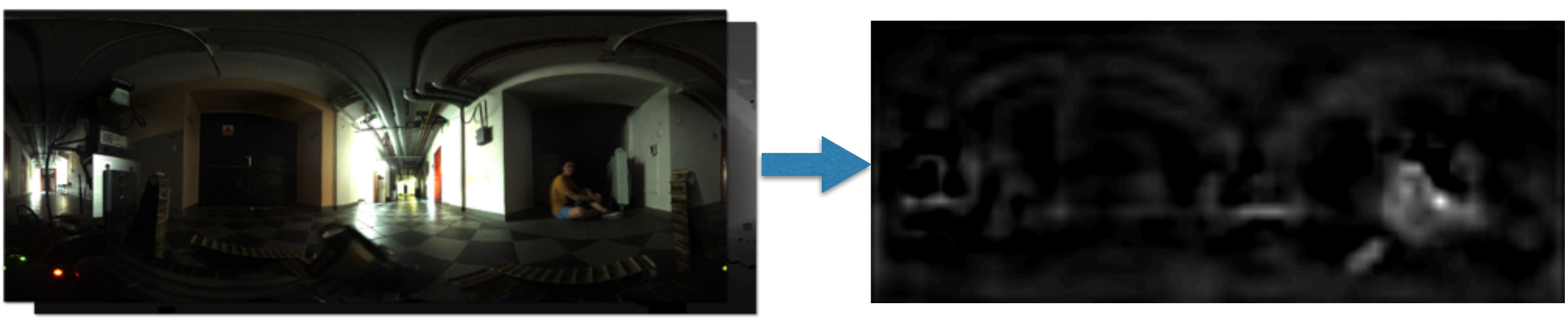
Segmentation network RGBDT->H



Self-Supervised training of *gain* predicting network

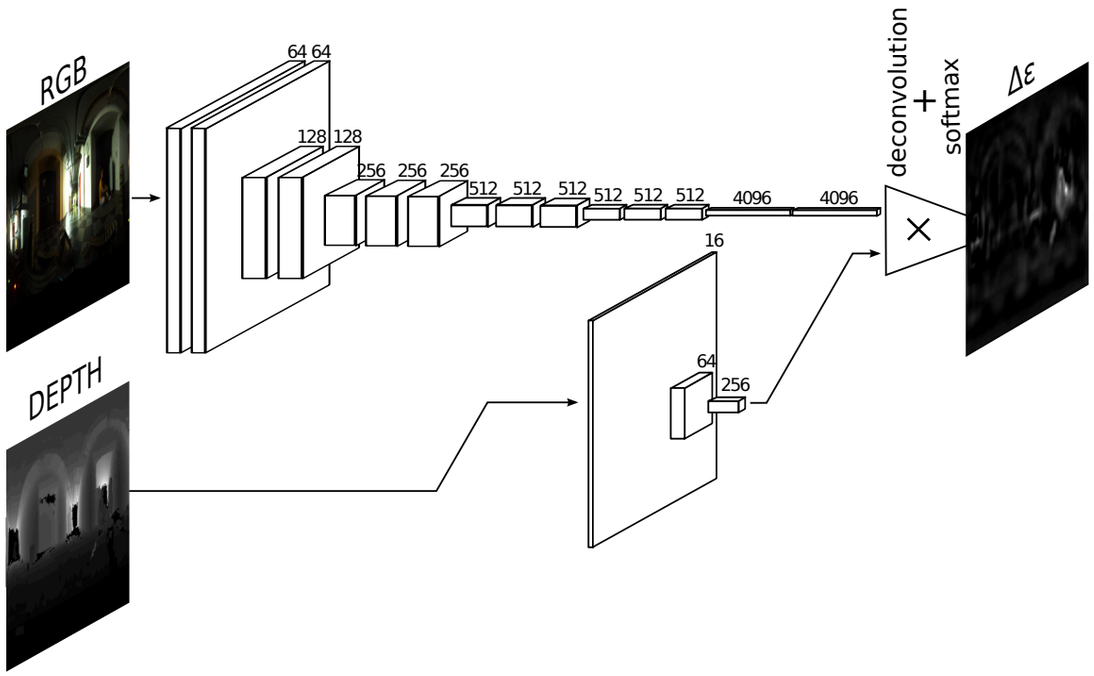


gain = expected per-pixel reduction of the segmentation error, when thermal measurement is available.

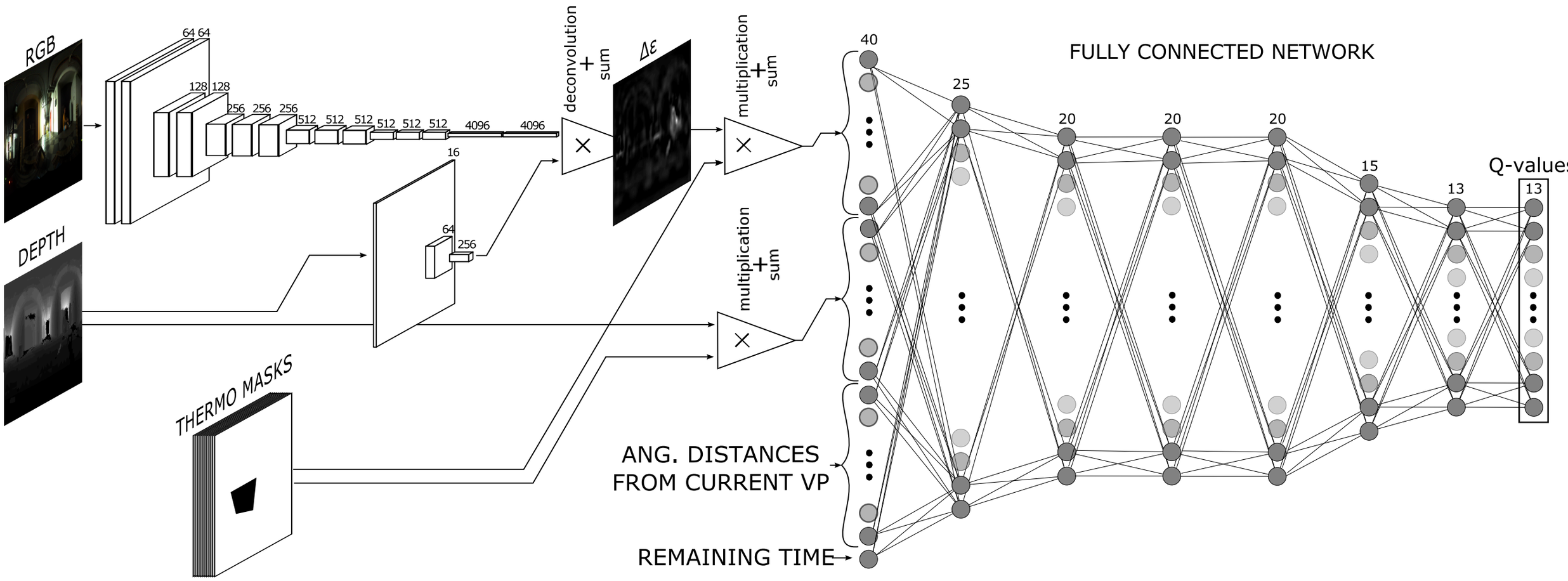


Czech Technical University in Prague
Faculty of Electrical Engineering, Department of Cybernetics

Gain predicting network RGBD->gain

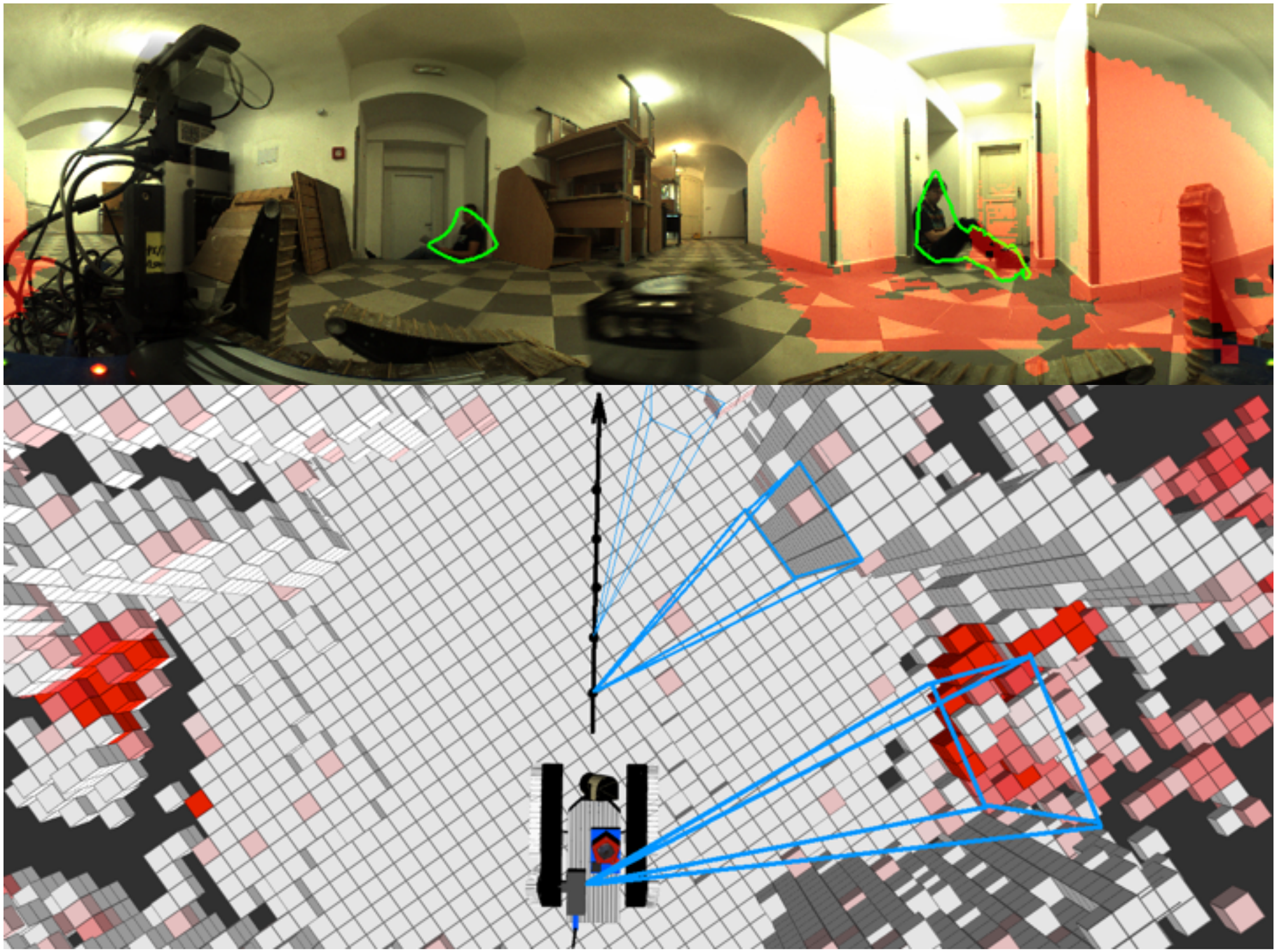


Policy network initialisation



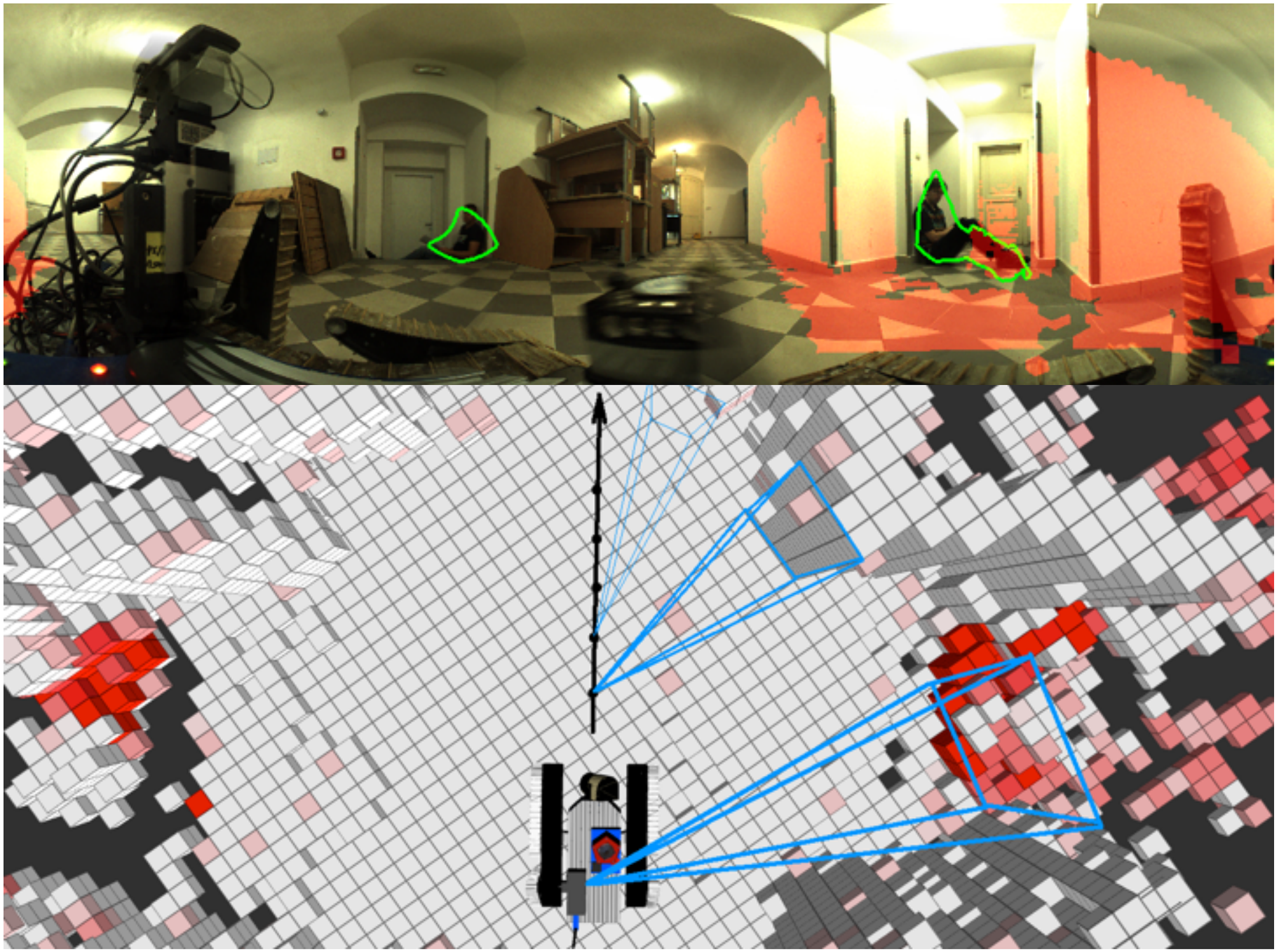
Initialize RGBD->Q by extending the RGBD->gain by fully connected layers (outputs corresponds to actions)

Policy network guided learning



Project gain on a complete voxel map and use MILP to get the optimal pan-tilt control wrt the long-term sum of accumulated gains (i.e. Q-values).

Policy network guided learning



Use optimal trajectories to guide learning of the policy network.

Czech Technical University in Prague
Faculty of Electrical Engineering, Department of Cybernetics

Motion control of thermal camera for segmentation

Czech Technical University in Prague
Faculty of Electrical Engineering, Department of Cybernetics