Hybride Lernstrategien für dateneffiziente Robotik

Jesse Marekwica

2026-01-21

Proseminar: Informatik trifft Maschinenbau

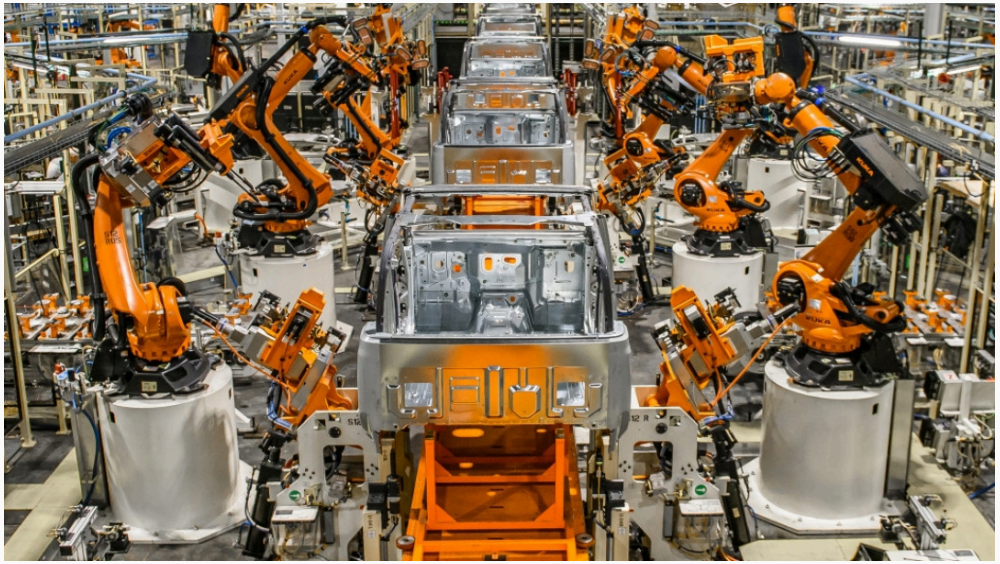
Inhalt

- Vorstellung des Papers
- Markov-Decision-Process (MDP)
- Reinforcement Learning with Prior Data (RLPD)
- Fazit zum Paper

Motivation

Problem der modernen Montage

- Autonome Montageprozesse sind verbeitet in der Industrie
- Prozesse sind jedoch meist statisch und unflexibel
- Roboter mit menschlichen Montagefähigkeiten sind eine Herausforderung

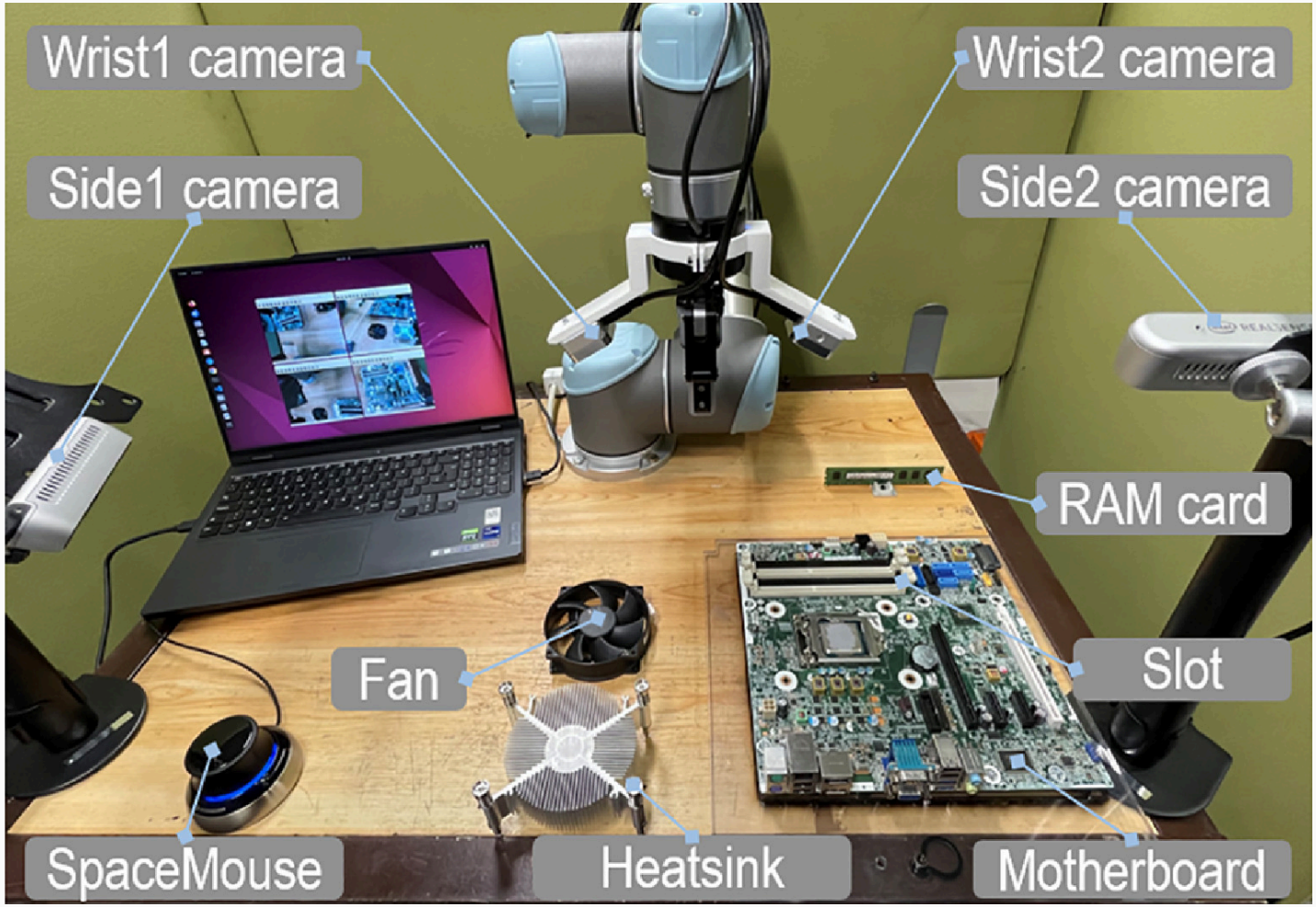


Fertigungsstraße in der Automobilindustrie - [1]

Sind autonome, flexible und präzise Montagen möglich?

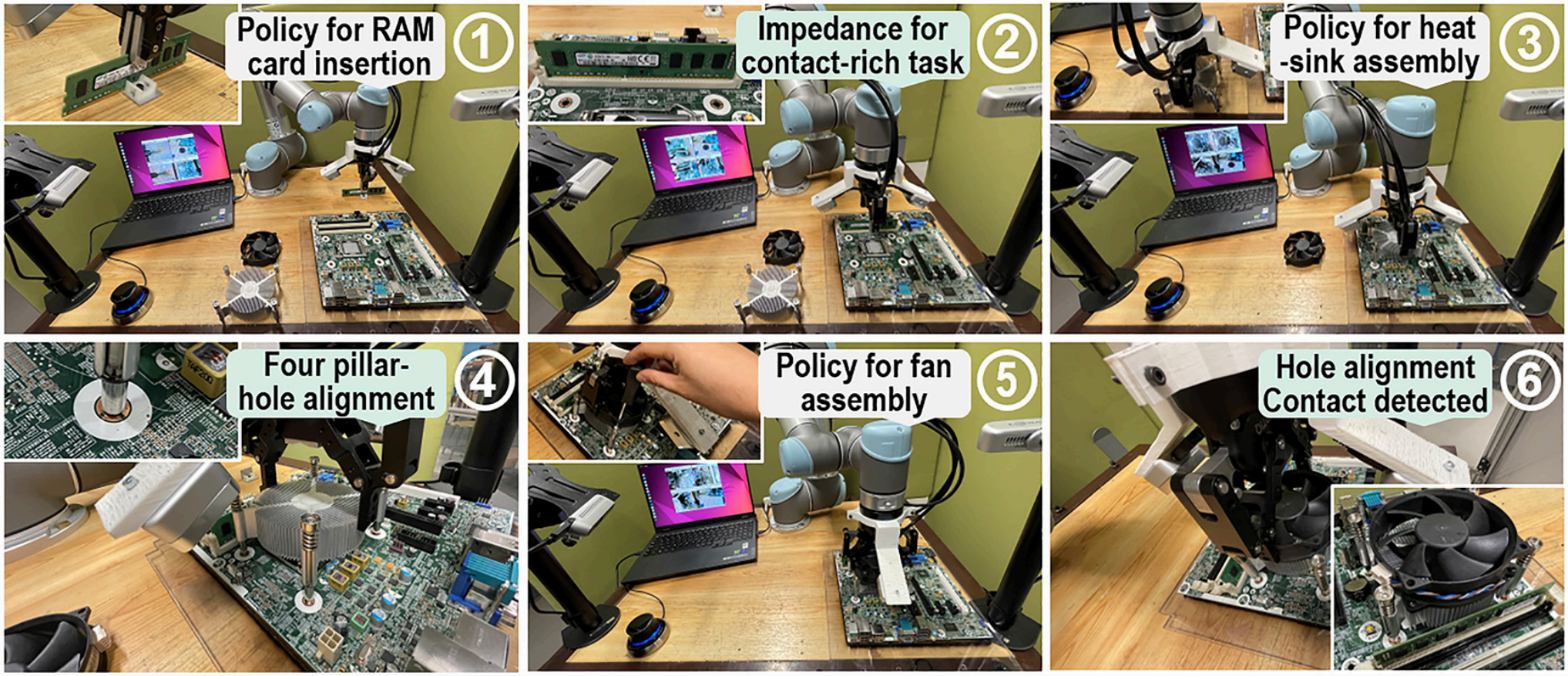
Lösung des Papers

Präzisionsmontage mit RL & Bilder



Montageaufbau (Montiert wird RAM, Kühlkörper und Lüfter) - [2]

Präzisionsmontage mit RL & Bilder

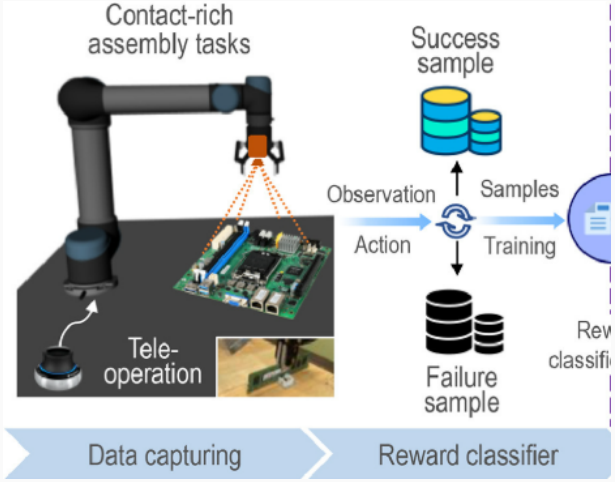


Montageablauf (Einsetzen und Ausrichten) - [2]

Präzisionsmontage mit RL & Bilder

Datenerhebung

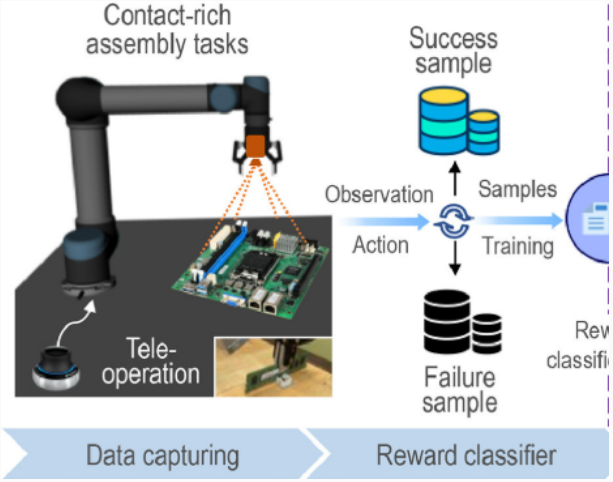
- Menschliche Demos
- Extraktion von Daten
- Einbettung in System



Präzisionsmontage mit RL & Bilder

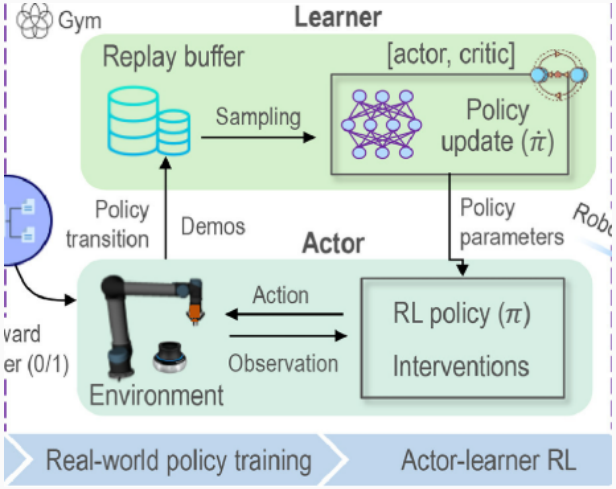
Datenerhebung

- Menschliche Demos
- Extraktion von Daten
- Einbettung in System



Reinforcement Learning

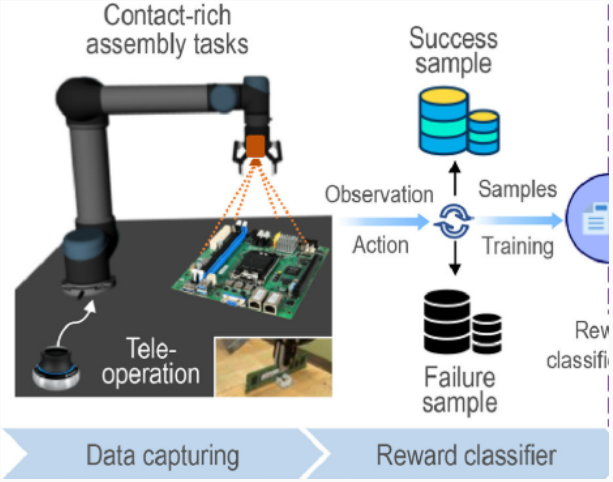
- RLPD + HIL
- Actor & Critic
- Policy Optimierung



Präzisionsmontage mit RL & Bilder

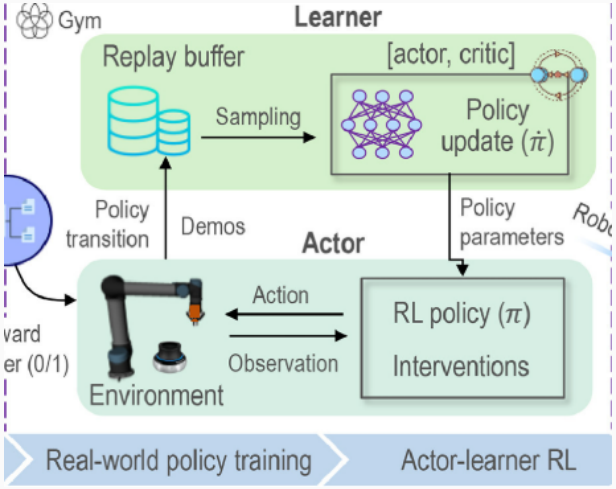
Datenerhebung

- Menschliche Demos
- Extraktion von Daten
- Einbettung in System



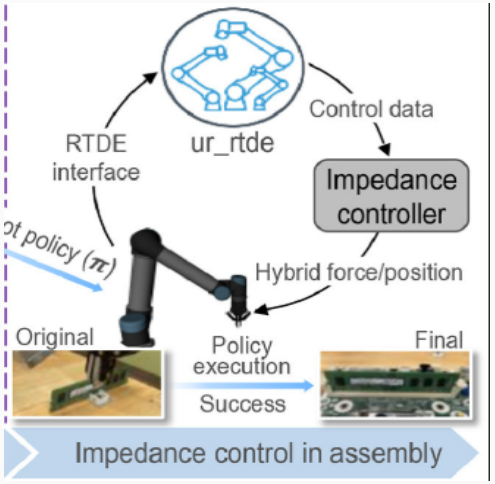
Reinforcement Learning

- RLPD + HIL
- Actor & Critic
- Policy Optimierung



Impedanzcontroller

- Nachgiebigkeit
- $F = k \cdot e$
- RL + Physik

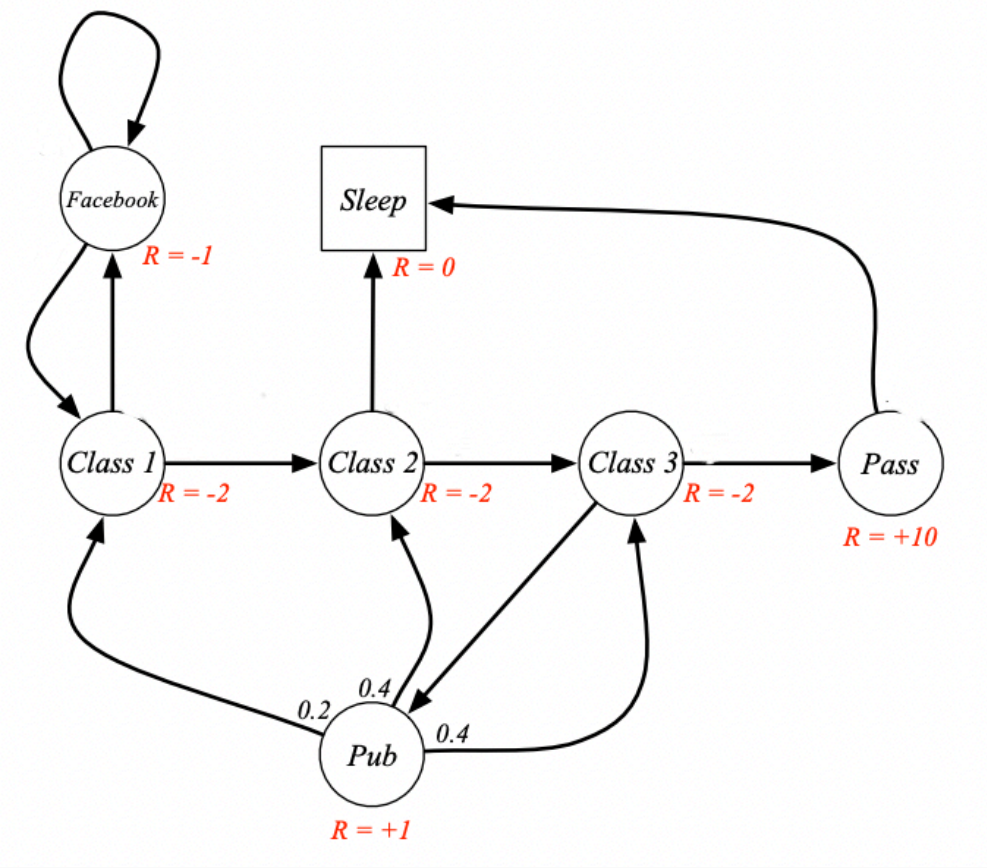


Wie formulieren wir eine vage Problemstellung in etwas algorithmisch berechenbares?

Markov-Decision-Process (MDP)

Markov-Decision-Process

- Vage Probleme werden zu berechenbare Mathematik
- $M = \{S, A, P, R, \gamma\}$
  - S - Zustand
  - A - Entscheidung/Handlung
  - P - Wahrscheinlichkeit
  - R - Belohnung
  - γ - Diskontfaktor
- Ziel: Optimale Strategie (π) finden



MDP als Graph eines Studentenlebens - [3]

Markov-Decision-Process

$M = \{S, A, P, R, \gamma\}$

- S (State): 2x RGB-Bilder + Roboter-Gelenkdaten + Griffstärke
- A (Action): 6D-Pose (Position & Rotation) + Greifer-Status (Auf/Zu)
- P (Probability): Sensorrauschen + Reibung + Widerstand + Toleranz
- R (Reward): Neuronales Netz (Binary Classifier) + Menschliche Demos
- γ (Discount): Verhindert Divergenz + Weitsichtigkeit

Markov-Decision-Process

$M = \{S, A, P, R, \gamma\}$

- S (State): 2x RGB-Bilder + Roboter-Gelenkdaten + Griffstärke
- A (Action): 6D-Pose (Position & Rotation) + Greifer-Status (Auf/Zu)
- P (Probability): Sensorrauschen + Reibung + Widerstand + Toleranz
- R (Reward): Neuronales Netz (Binary Classifier) + Menschliche Demos
- γ (Discount): Verhindert Divergenz + Weitsichtigkeit

Warum Reinforcement Learning für Policy (π)?

Da die Kontaktphysik (P) und Bilddaten (S) zu komplex für Formeln sind, muss der Roboter die Lösung erlernen.

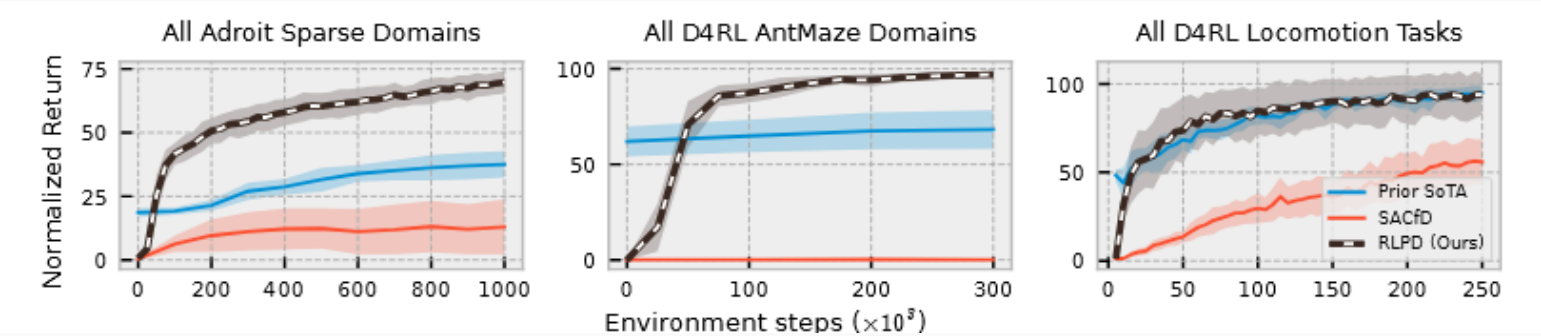
Wie ist es möglich bei so hoher Datenkomplexität trotzdem Effizient bleiben?

Reinforcement Learning with Prior Data (RLPD)

RLPD

RLPD entspringt dem Ansatz von Soft-Actor-Critic (SAC) mit essenziellen Designerweiterungen.

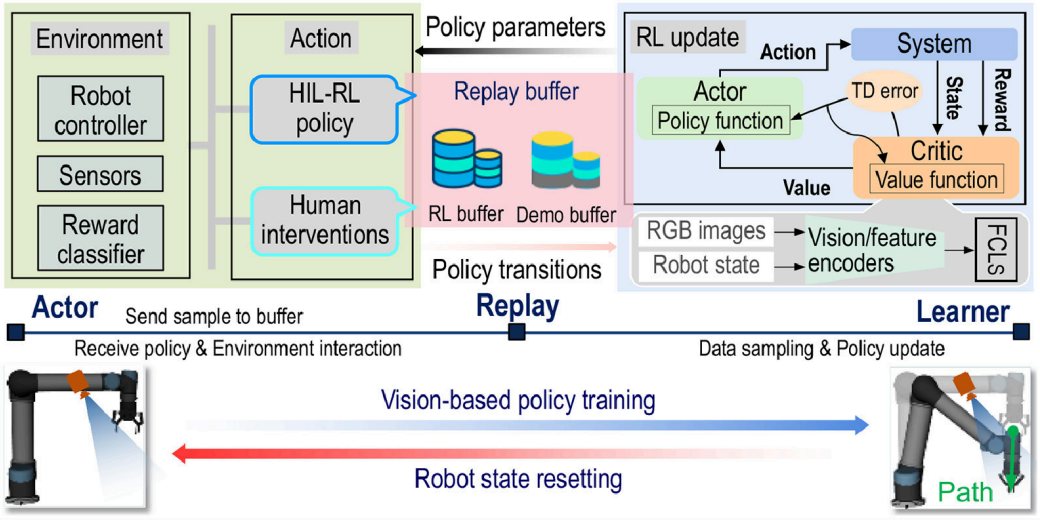
1. Eine einfache und effiziente Methode zur Einbindung von Offline-Daten
2. Normalisierung von Ebenen zur Milderung von Überschätzungen
3. Effiziente Entnahme von Datenpunkten (Sample-Efficiency)



Vergleich verschiedener Algorithmen gegen RLPD von Ball et al. - [4]

RLPD - Zwei Buffer System

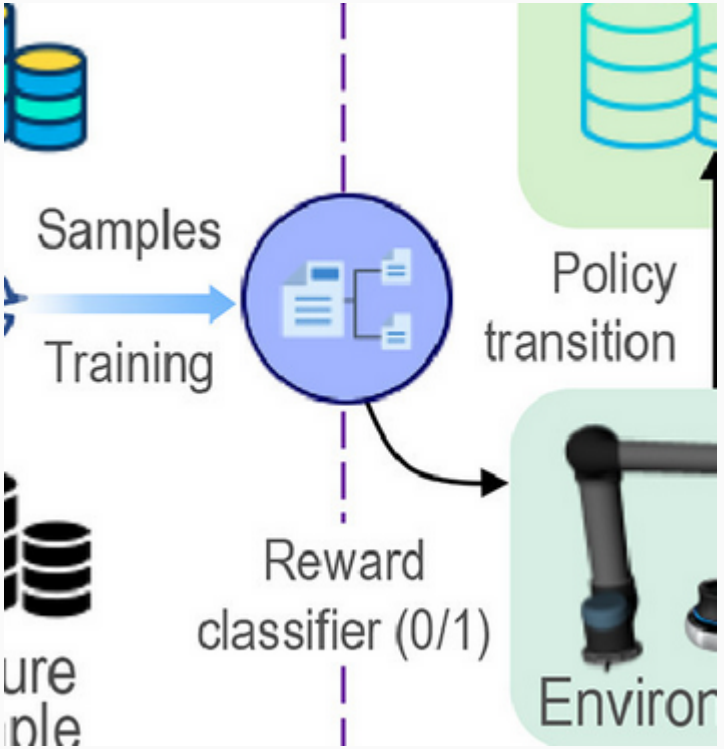
- Richtungsfindung zu Beginn bei RL sehr Zeit- und Rechenintensiv
- Menschliche Demonstrationen oder suboptimale Policies können Richtung vorgeben (PD)
- Offline und Online Daten trennen
  - Zwei Buffer + 50/50 Datenpunkte
- Umsetzung im Paper: Vorhanden



Lui & Wang verwenden ebenfalls zwei Buffer - [2]

RLPD - Normalisierung

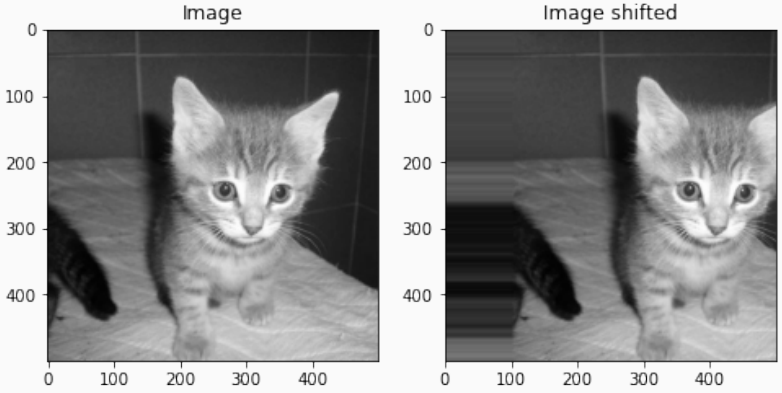
- Out-of-Distribution (OOD) Daten bereiten RL-Algorithmen Probleme
- OOD-Daten können vom Critic stark “überschätzt” werden $\rightarrow$ Divergenz
- Normalisierung in Ebenen von Neuronalen Netzen verhindert dies
- $\|Q(s,a)\| \leq \|w\|$ - Netzwerkgewichte limitieren Q-Wert
- Umsetung im Paper: Implizit



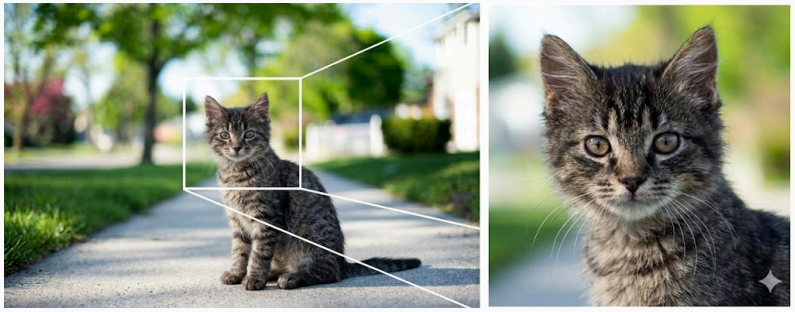
Reward Classifier gibt 1 und 0 aus - [2]

RLPD - Effizientes Samplen

- Nutzung von zwei Buffern erhöht deutlich Aufbereitung von Daten
- Mögliche Gegenwirkungen
  - Höhere Lerngeschwindigkeit
  - Qualitätssteigerung der Daten
- Gefahr: Überanpassung (Overfitting)
- Präventionsmethoden nutzen
  - Random Shift Augmentations
  - Random Ensemble Distillation
- Umsetzung im Paper: Unklar



Beispiel für Random Shift Augmentations - [5]



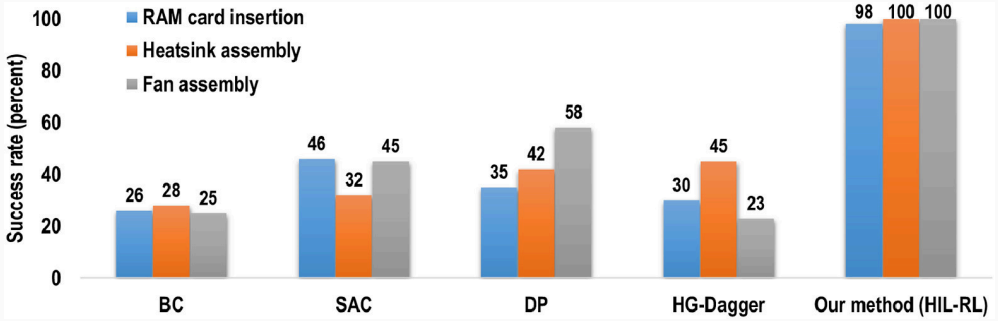
Beispiel für Image Cropping - Nano Banana Pro

Wenn Code zur Bewegung wird

Fazit

Ergebnisse

- Ergebnisse sind eindeutig
- Über 98% erfolgreiche Montage mit RLPD + HIL
- Verglichen wurde:
  - Behaviour cloning (BC)
  - Soft actor critic (SAC)
  - Diffusion policy (DP)
  - HG-Dagger
- Alle wurden mit 100 “human demonstrations” trainiert, bis auf HG und RLPD → “same number of interventions”



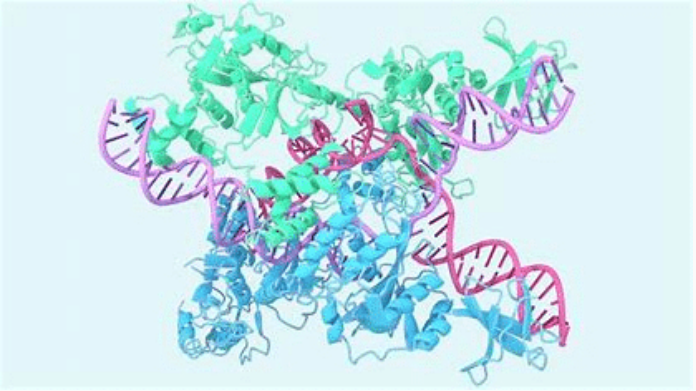
Ergebnisse erfolgreicher Montage während Lernprozess - [2]

Task	SAM card insertion	Heatsink assembly	Fan assembly
Training time (h)	1.1	0.9	0.6
Cycle time (s)	5.2	4.7	3.8

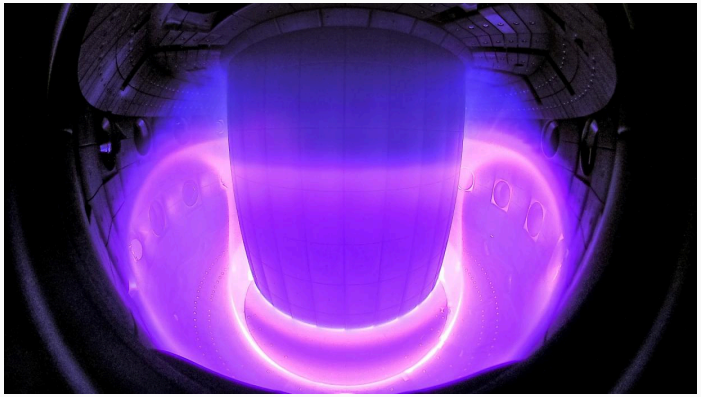
Benötigte Zeit und Ressourcen von RLPD zum Erlernen - [2]

Würdigung

- Informatik wird in interdisziplinären Bereichen immer relevanter
- “Über der Computerwelt heraus Dinge bewegen”
- Paper ist ein wunderbares Beispiel für Relevanz von Infromatik in Robotik
- Andere Bereiche benötigen ebenfalls qualifizierte Informatiker



AlphaFold entschlüsselte 200 Millionen Proteine - [6]



DeepMind steuert Tokamak zur Plasmaerzeugung - [7]

Kritik

- Erschwerte Validierung und Reproduzierbarkeit durch Informationslücken
  - *Wann und wie wurde “Human-in-the-Loop” umgesetzt?*
  - *Wie wurde “Layer Normalization” umgesetzt?*
  - *Wie wurde “Overfitting” präventiert?*
  - *Wie genau wurden die Algorithmen verglichen (Unterschiede)?*

> [...] used to compare the training performance over these tasks and perform ablation studies with the same number of human demonstrations but different interventions. Specifically, BC, SAC and DP are trained with 100 human demonstrations, while HG-Dagger has the same number of interventions as RL [2].

Bibliography

[1] "KR QUANTEC." Accessed: Jan. 07, 2026. [Online]. Available: https://www.kuka.com/de-de/produkte-leistungen/robotersysteme/industrieroboter/kr-quantec

[2] S. Liu and L. Wang, "Vision intelligence-conditioned reinforcement learning for precision assembly," *CIRP Annals*, vol. 74, no. 1, pp. 13–17, Jan. 2025, doi: 10.1016/j.cirp.2025.04.016.

[3] R. Khandelwal, "An Introduction to Markov Decision Process." Accessed: Jan. 19, 2026. [Online]. Available: https://arshren.medium.com/an-introduction-to-markov-decision-process-8cc36c454d46

[4] P. J. Ball, L. Smith, I. Kostrikov, and S. Levine, "Efficient Online Reinforcement Learning with Offline Data."

Bibliography

[5] S. Thrun and A. Schwartz, “Issues in Using Function Approximation for Reinforcement Learning,” *Proceedings of the 1993 Connectionist Models Summer School*. Psychology Press, 1994.

[6] J. Cheng *et al.*, “Accurate proteome-wide missense variant effect prediction with AlphaMissense,” *Science*, vol. 381, no. 6664, p. eadg7492, Sept. 2023, doi: 10.1126/science.adg7492.

[7] “DeepMind steuert Fusionsreaktor mit künstlicher Intelligenz.” Accessed: Jan. 21, 2026. [Online]. Available: https://futurezone.at/science/deepmind-kuenstliche-intelligenz-ki-fusionsreaktor-kernfusion-plasma/401908981