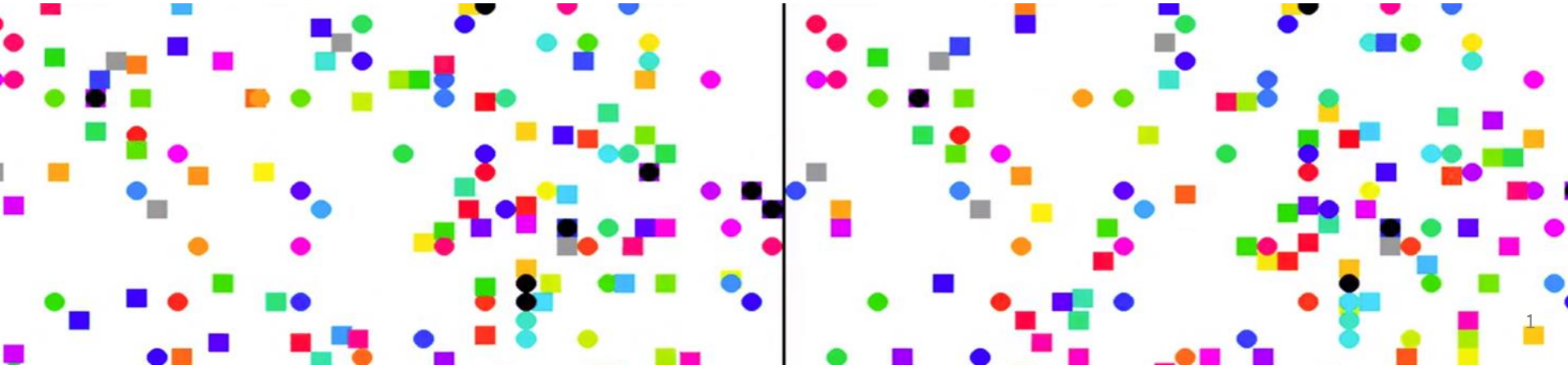
PRIMAL

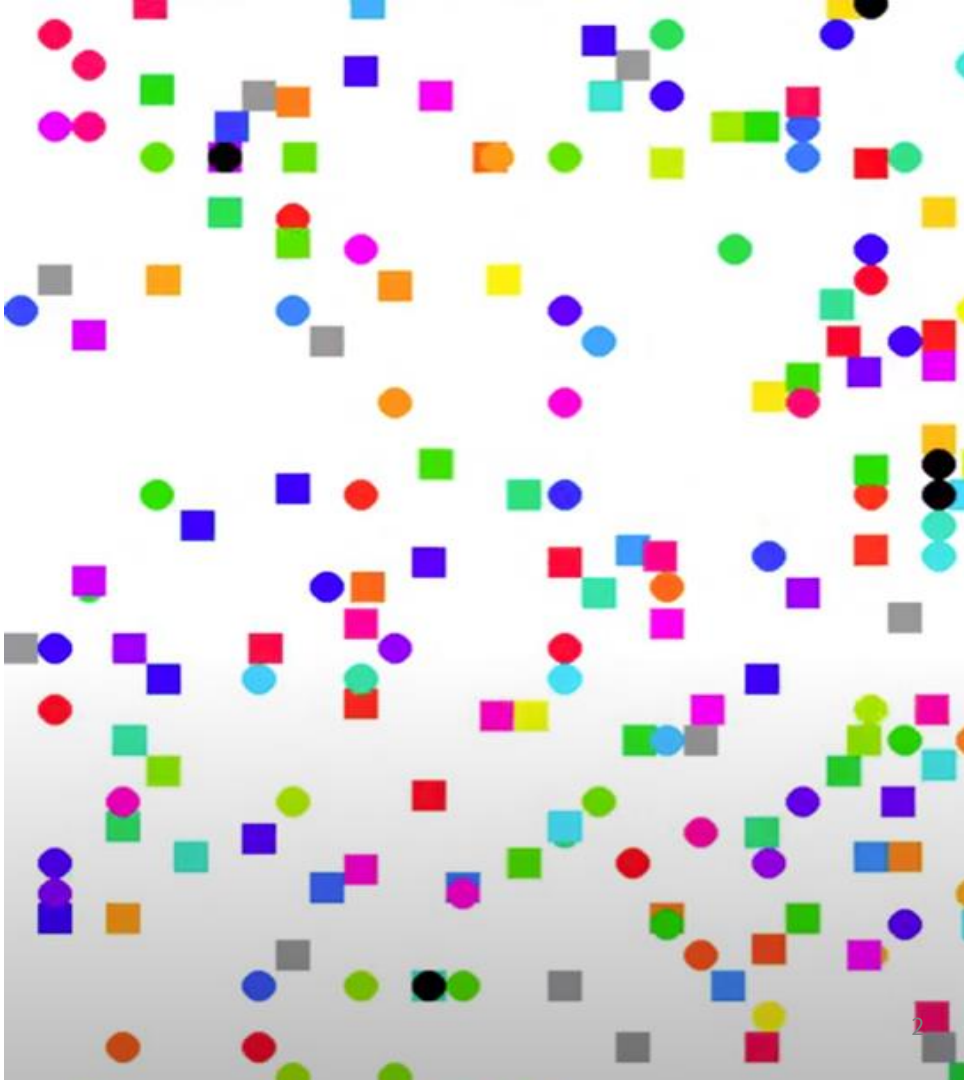
Pathfinding via Reinforcement and Imitation Multi-Agent Learning

Renáta Pivodová, březen 2022



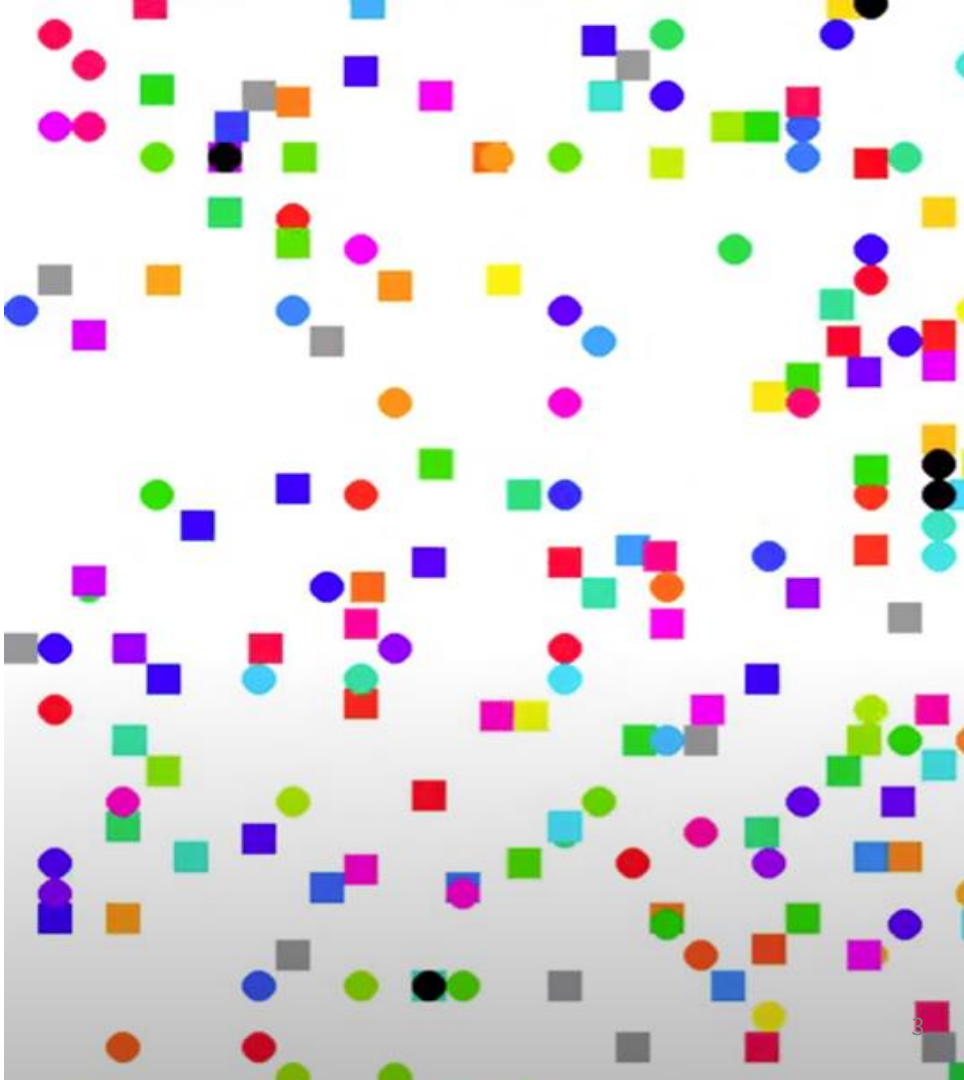
Osnova

1. Úvod
2. Důležité pojmy
3. MAPF a reinforcement learning
4. Učení
5. Experimenty
6. Závěr



Osnova

1. Úvod
2. Důležité pojmy
3. MAPF a reinforcement learning
4. Učení
5. Experimenty
6. Závěr

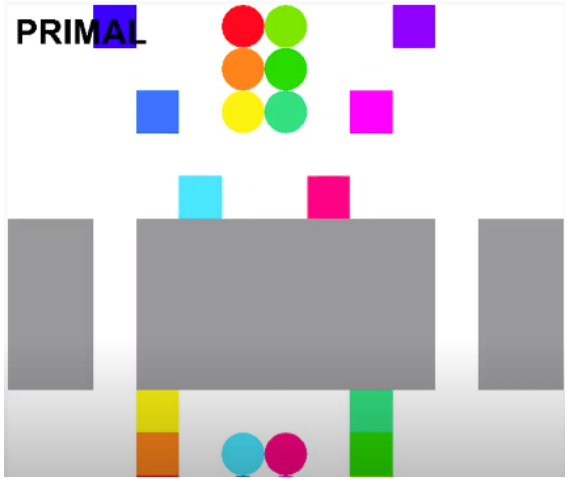


Nevýhody SOTA plánovačů

- Centralizované plánování.
- Maximálně stovky agentů.
- Online plánování v řádu sekund až minut.

Charakteristiky PRIMAL

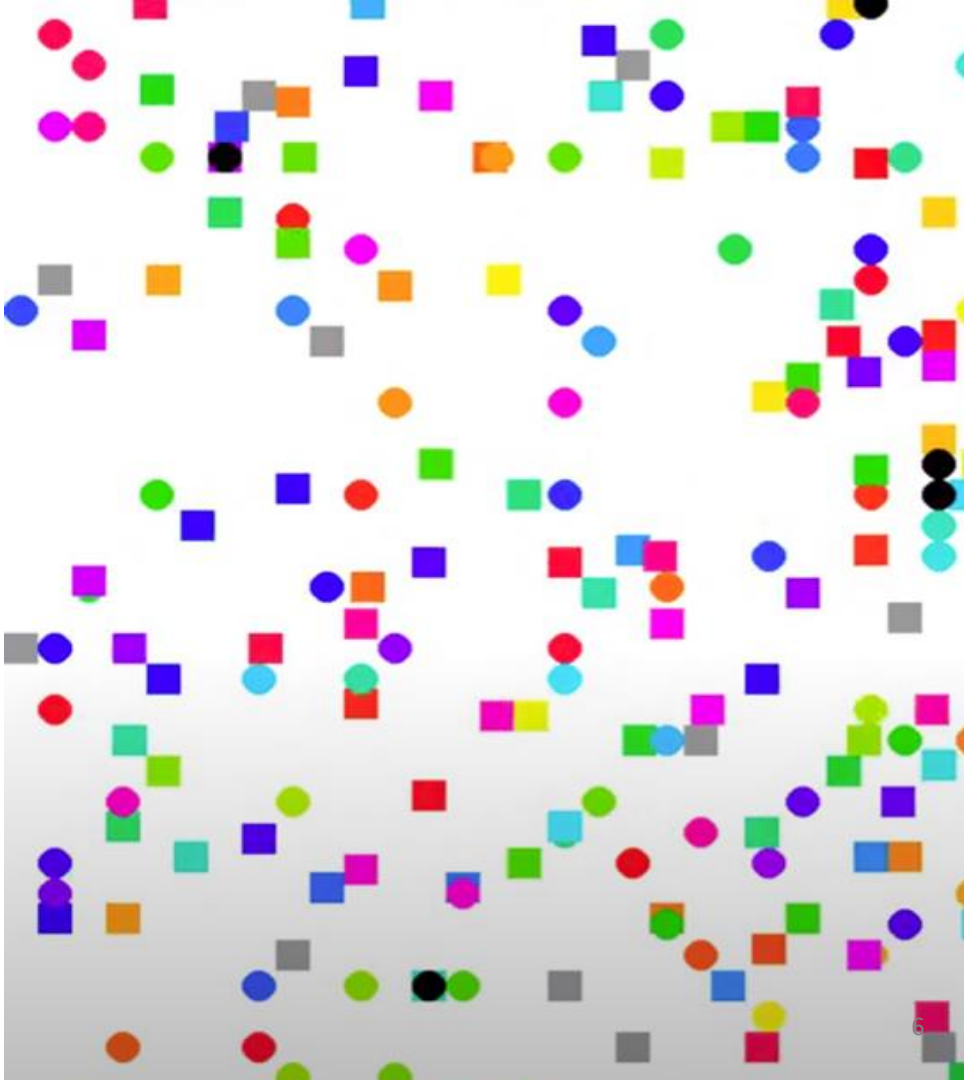
- PRIMAL = nový rámec pro MAPF.
  - Pathfinding via Reinforcement and Imitation Multi-Agent Learning
- Kombinuje reinforcement a imitation learning.
- Základní principy:
  - **Částečně** pozorovatelný svět.
  - Reaktivní **online** plánování.
  - **Implicitní** koordinace.



1. Úvod

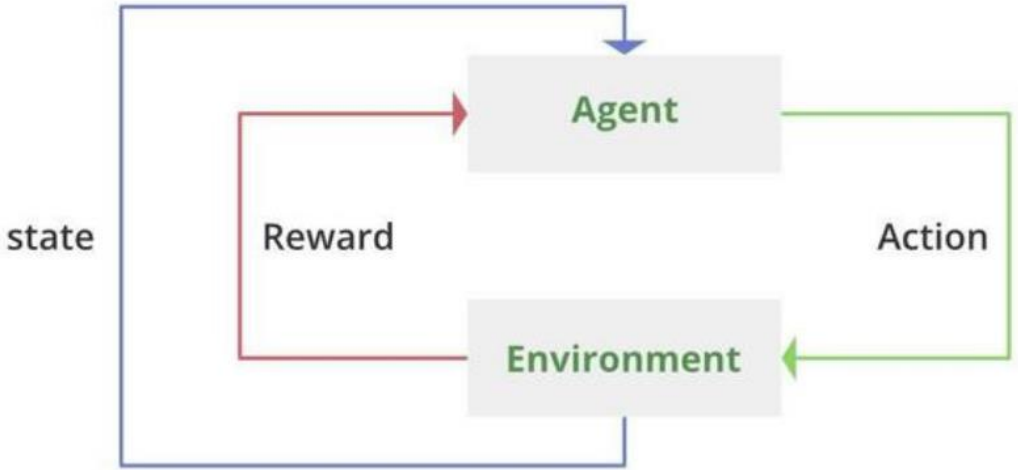
Osnova

1. Úvod
2. Důležité pojmy
3. MAPF a reinforcement learning
4. Učení
5. Experimenty
6. Závěr



Reinforcement Learning - I

- Agent vykonává **akce** podle **strategie.**
- Agent se učí pomocí **zpětné vazby** od prostředí.



2. Důležité pojmy

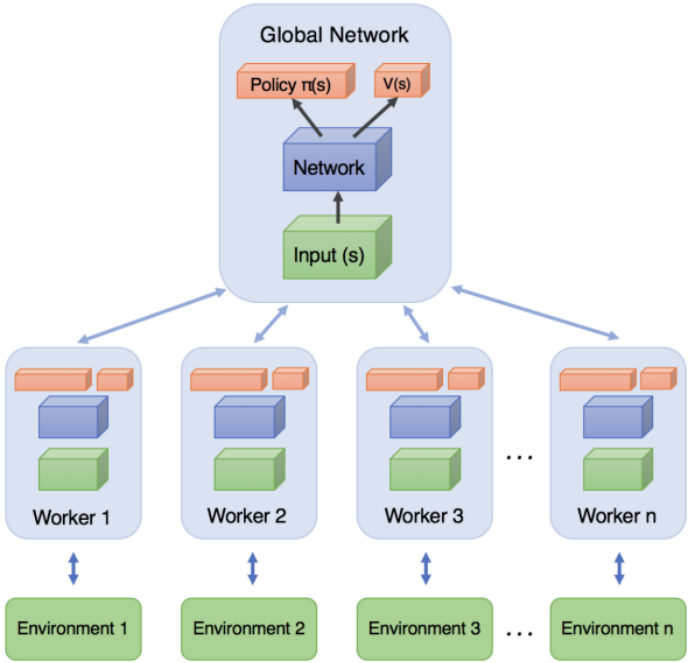
Reinforcement Learning - II (formalizace)

- Prostředí jako markovský rozhodovací proces = (S, A, P, R)
  - S … konečná množina stavů prostředí
  - A … konečná množina akcí
  - $P_a(s, s')$ … pravděpodobnost, že po akci *a* ve stavu *s* přejde prostředí do stavu *s'*
  - $R_a(s, s')$ … odměna za to, že po akci *a* ve stavu *s* bude prostředí ve stavu *s'*
- Chování agenta:
  - Strategie $\pi(s,a)$ … pravděpodobnost vykonání akce *a* ve stavu *s*
  - Agent maximalizuje celkovou odměnu.
- Hodnota stavu $V(s)$ je střední hodnota celkové diskontované odměny.
- Hodnota akce $Q(s,a)$

2. Důležité pojmy

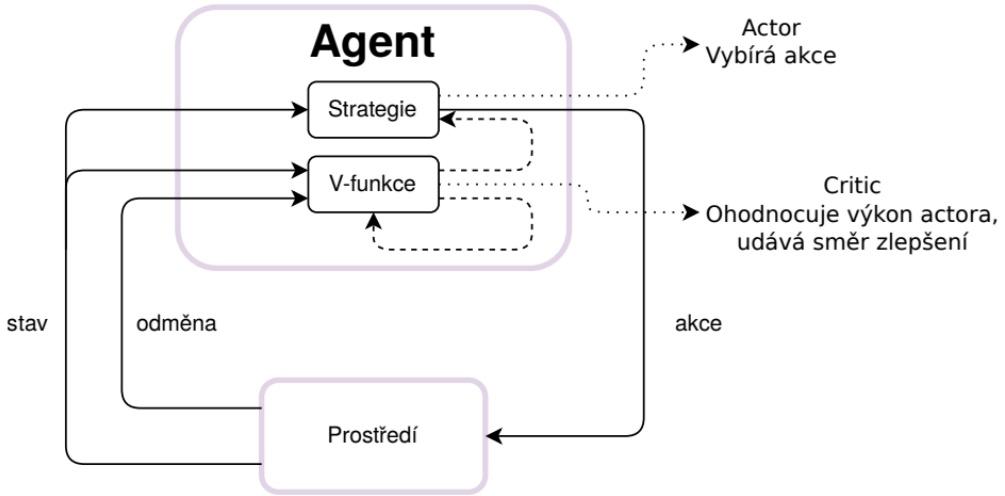
A3C algoritmus - I

- = Asynchronous Advantage Actor Critic algorithm
- Každý agent má vlastní informace, pracuje nezávisle na ostatních:
  - **parametry** neuronové sítě,
  - kopii **světa**.
- Všichni komunikují a vylepšují jednu **globální síť**.



A3C algoritmus - II

- Algoritmus predikuje *V(s)* i *π(s)*.
  - Hodnotou (kritikem) aktualizuje strategii (aktéra).
- Advantage: *A = Q(s, a) – V(s)*
  - O kolik se reálná odměna liší od očekávané odměny.



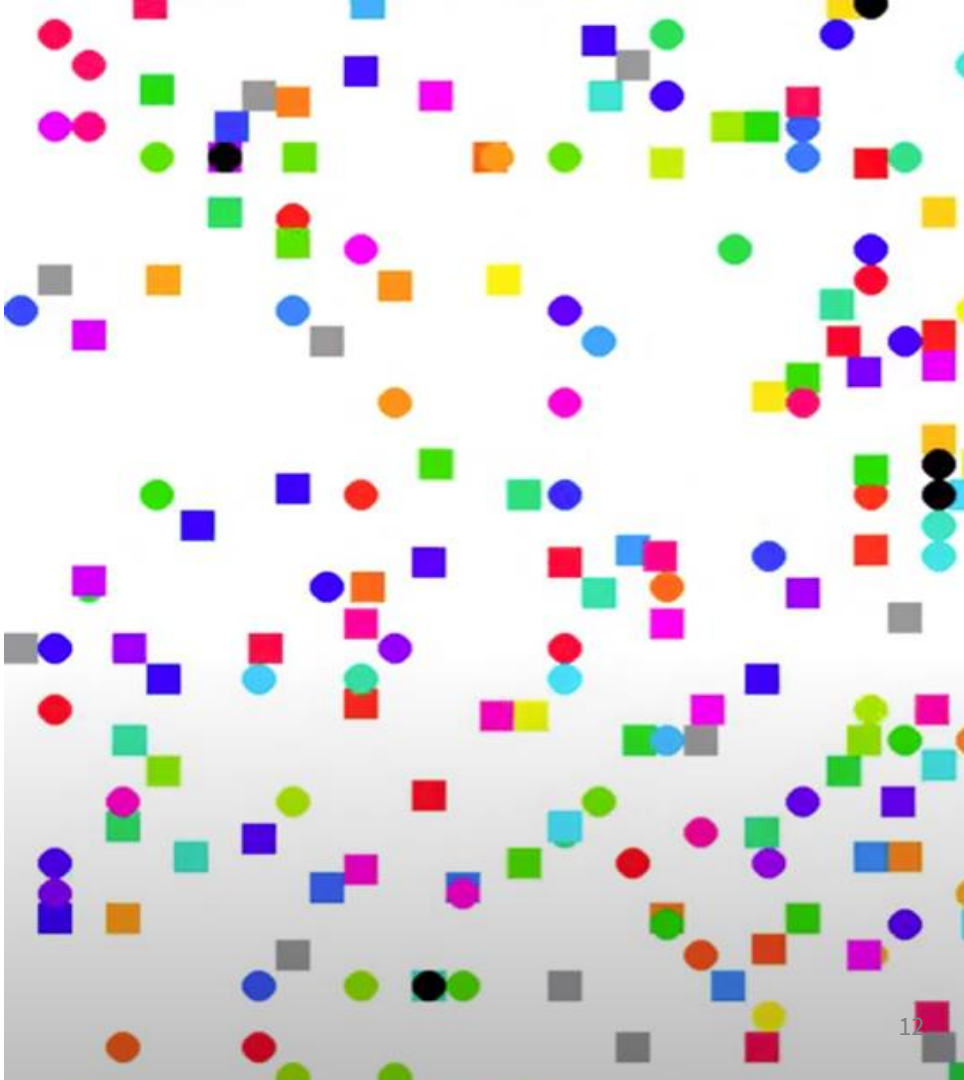
2. Důležité pojmy

Imitation learning

- Expert dodá několik **vzorových scénářů (demonstrations)**.
  - Slouží agentovy k naučení se optimální strategie.
- Behavioural cloning
  - Naučení expertovy strategie pomocí **supervised learning**.
- Algoritmus:
  - 1. Získejte od experta vzorové scénáře.
  - 2. Převeďte scénáře do dvojic (stav, akce).
  - 3. Pomocí supervised learning a minimalizace ztrátové funkce se naučte strategii.

Osnova

1. Úvod
2. Důležité pojmy
3. MAPF a reinforcement learning
4. Učení
5. Experimenty
6. Závěr

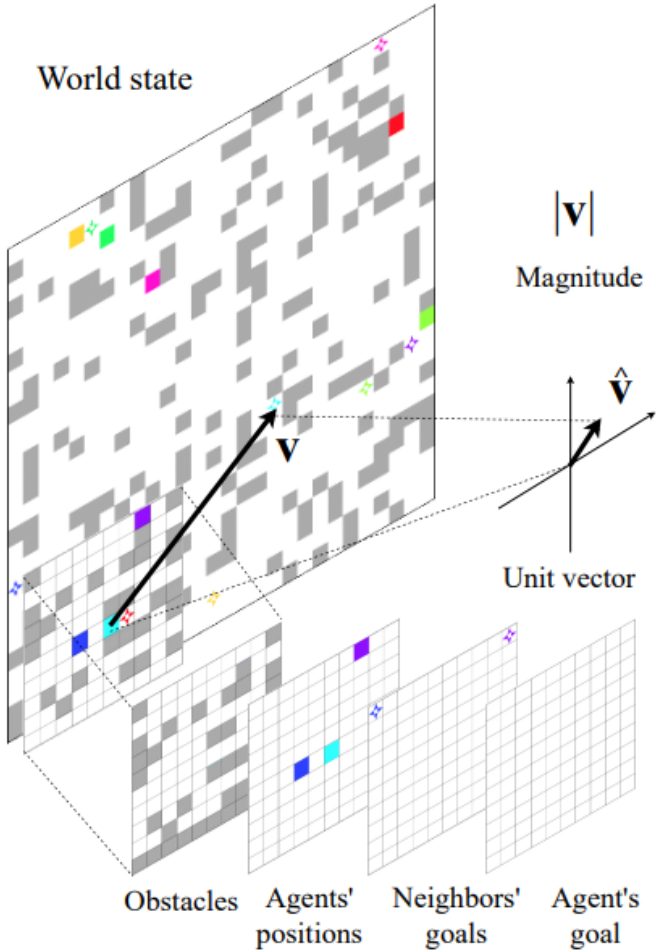


Pozorovací prostor agenta - I

- Svět je reprezentován pomocí **mřížky** (gridworld).
- Agent vidí pouze své okolí, tedy **malou část** mřížky.
  - V praxi o velikosti 10 x 10.
  - Agent je centrem této mřížky.
- Agent má navíc dvě přídavné informace o svém cíli:
  - jednotkový vektor **ukazující** na cíl,
  - Euklidovskou **vzdálenost** k cíli.

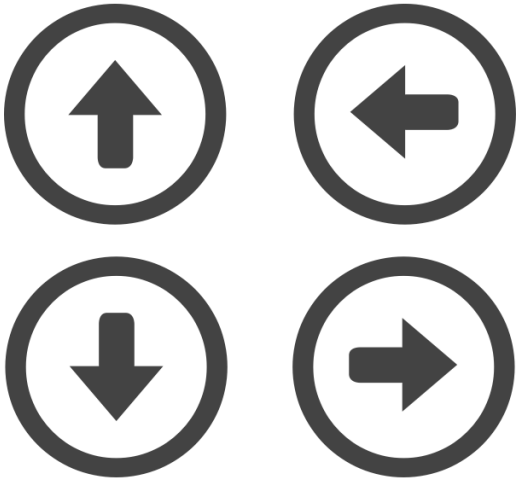
Pozorovací prostor agenta - II

- Agent vnímá své okolí na čtyřech úrovních:
  - překážky,
  - pozice ostatních agentů,
  - pozice vlastního cíle,
  - pozice cílů ostatních agentů.
- Jedná se o **binární matice**.



Agentovy akce - I

- Možné akce:
  - krok vpřed,
  - krok vzad,
  - krok vpravo,
  - krok vlevo,
  - stůj.
- Za jednotku času vykoná jednu akci (diskrétnost).



Agentovy akce - II (trénování)

- Agent vybírá **pouze** z validních akcí.
  - Učení podpořeno ztrátovou funkcí.
  - Stabilnější trénování než při penalizaci za výběr nevalidní akce.
- Pokud při testování agent zvolí nevalidní akci, stojí na místě.
- Předchází oscilujícím strategiím.
  - Agent se nesmí vrátit na **předchozí** místo.
  - Nutí k exploraci.

Odměňování agenta

- Trestání za každou chvíli mimo cíl.
  - Strategie “co nejrychleji do cíle”.
- Větší trest za stání než pohyb.
  - Podpora explorace.
- I přes filtraci nevalidních akcí mohou vzniknout **kolize** s ostatními agenty.
  - Agenti se pohybují **postupně** a v **náhodném pořadí**.

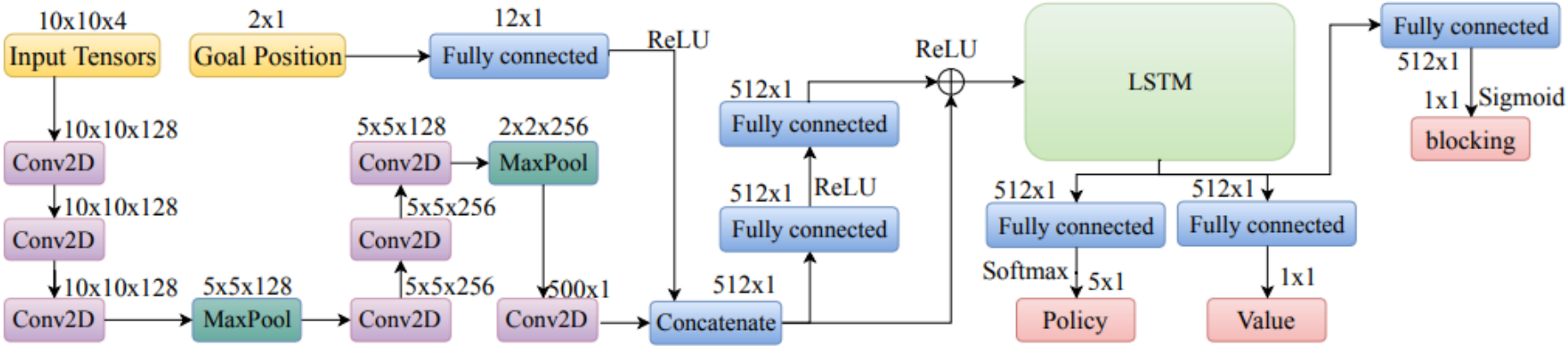
SIMPLE REWARD STRUCTURE.

Action	Reward
Move [N/E/S/W]	-0.3
Agent Collision	-2.0
No Movement (on/off goal)	0.0 / -0.5
Finish Episode	+20.0

Actor-critic síť - I

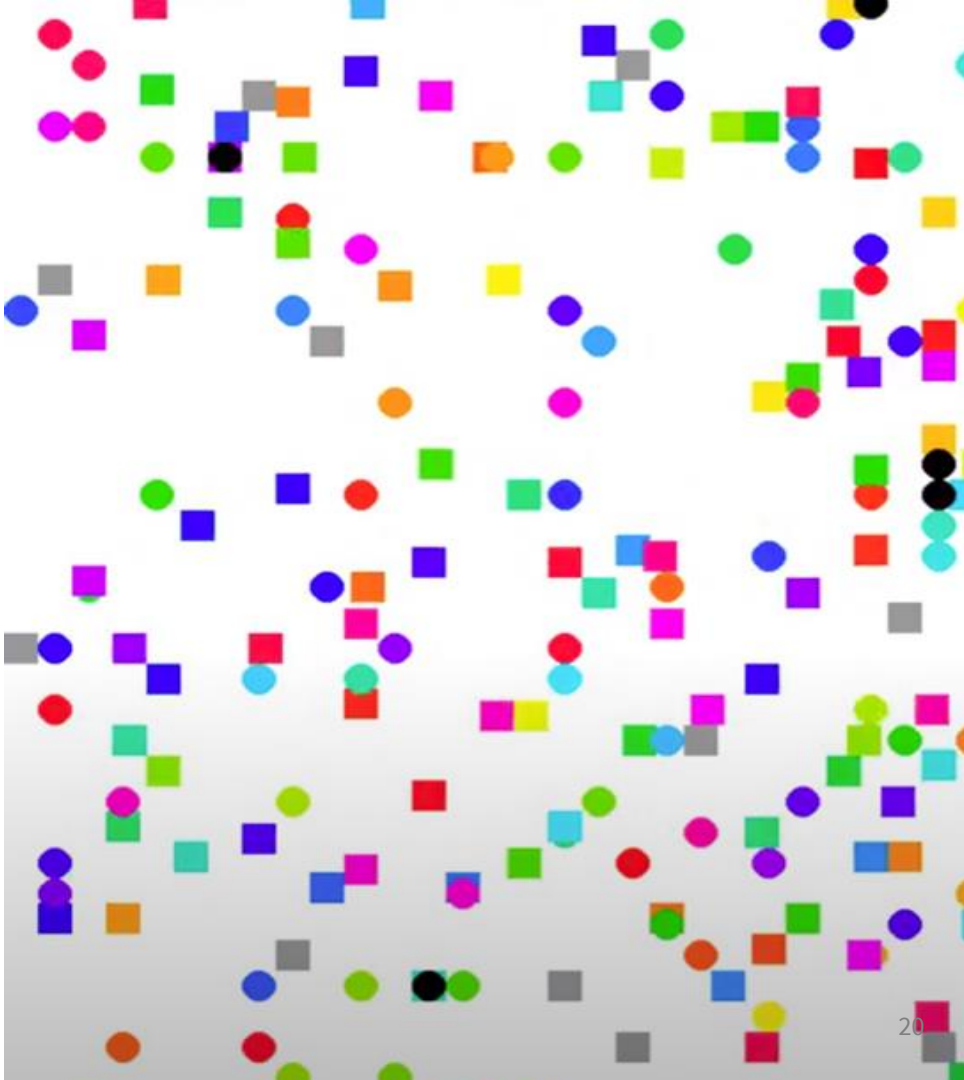
- PRIMAL je založen na **A3C** algoritmu.
- **Neuronová síť** doporučuje následující akci.
- Architektura:
  - šestivrstvá konvoluční síť,
  - dva vstupy: agentovo **pozorování** a **vzdálenost**/**směr** cíle,
  - tři výstupy: **strategie**, výsledná **hodnota** a “**blokuji** jiného agenta?”

Actor-critic síť - II



Osnova

1. Úvod
2. Důležité pojmy
3. MAPF a reinforcement learning
4. Učení
5. Experimenty
6. Závěr



Učení koordinace

- Blokování cesty ostatním, když je agent ve svém cíli.
- Výzva: **nesobecké** chování agentů.
  - I přes snížení okamžité odměny.
- Učení je postaveno na třech metodách:
  - blocking penalty,
  - kombinace RL a IL,
  - náhodnost prostředí.

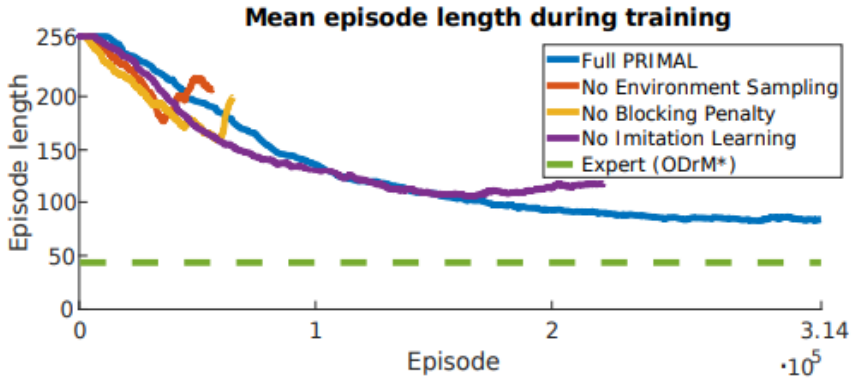


Fig. 5. Mean episode length during training, lower is better. The dotted line shows the baseline, obtained from the expert ODrM* planner. When we remove either environment sampling, the blocking penalties, or imitation learning from our approach, the policy converges to a worse solution.

Blocking penalty

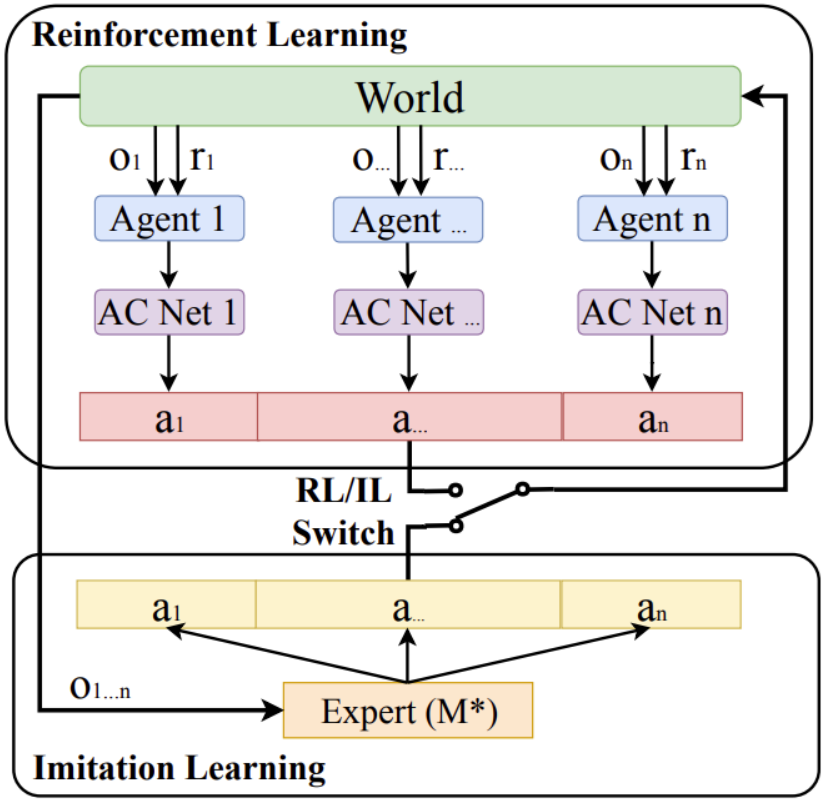
- = penalizace za prodloužení cesty ostatních agentů.
  - Tzn. úplné **zablokování**, nebo výrazné **zpomalení**.
  - - 2 body
- Motivace k opuštění cíle.
  - A opuštění lokálního maxima jejich odměny.

SIMPLE REWARD STRUCTURE.

Action	Reward
Move [N/E/S/W]	-0.3
Agent Collision	-2.0
No Movement (on/off goal)	0.0 / -0.5
Finish Episode	+20.0

Kombinace RL a IL

- IL rychle **identifikuje** dobré oblasti state-action prostoru.
- RL vylepšuje strategii volnou **explorací** těchto oblastí.
- Na začátku každé epizody se RL nebo IL volí **náhodně**.
- RL používá **A3C** algoritmus.
- Expertem pro IL je **ODrM***.
  - Levné online generování vzorů.



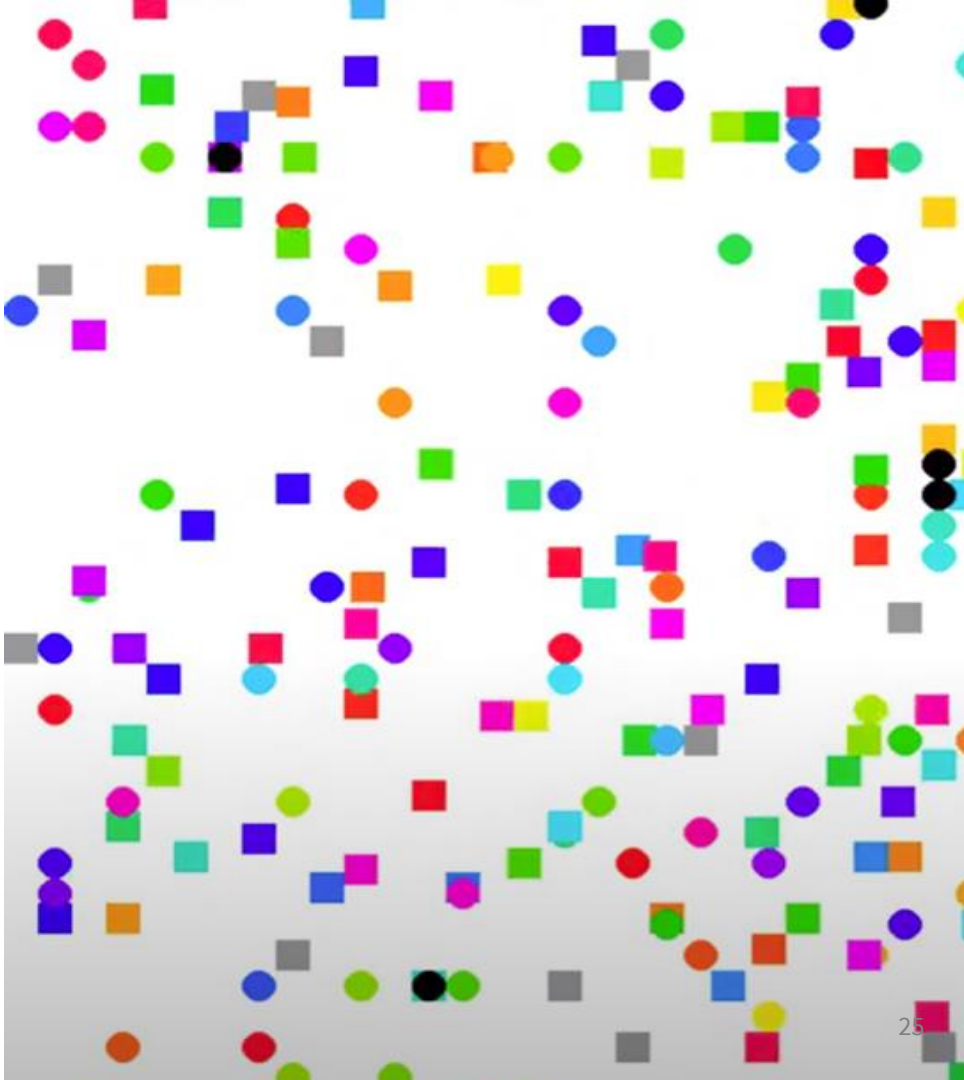
4. Učení

Náhodnost prostředí

- Pro každou epizodu trénování se generuje nové prostředí.
- Náhodná změna:
  - **velikosti** světa,
  - **hustoty** překážek.
- Parametry trénování:
  - velikost světa: 10, 40, 70
  - hustota překážek 0 - 50%

Osnova

1. Úvod
2. Důležité pojmy
3. MAPF a reinforcement learning
4. Učení
5. Experimenty
6. Závěr



Parametry

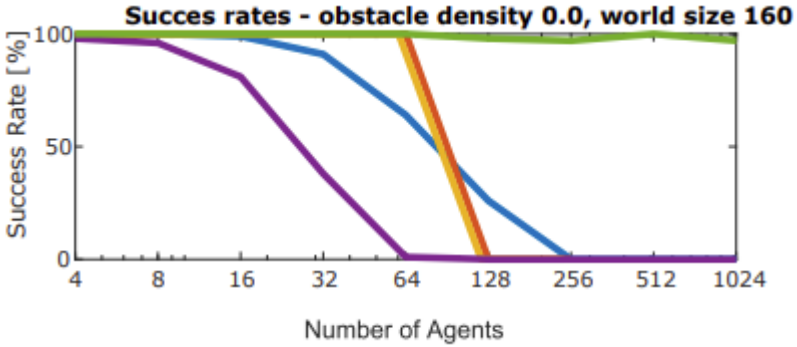
- Trénování:
  - discount faktor = 0,95
  - délka epizody = 256
  - velikost batche = 128
  - tři nezávislá prostředí
  - osm agentů

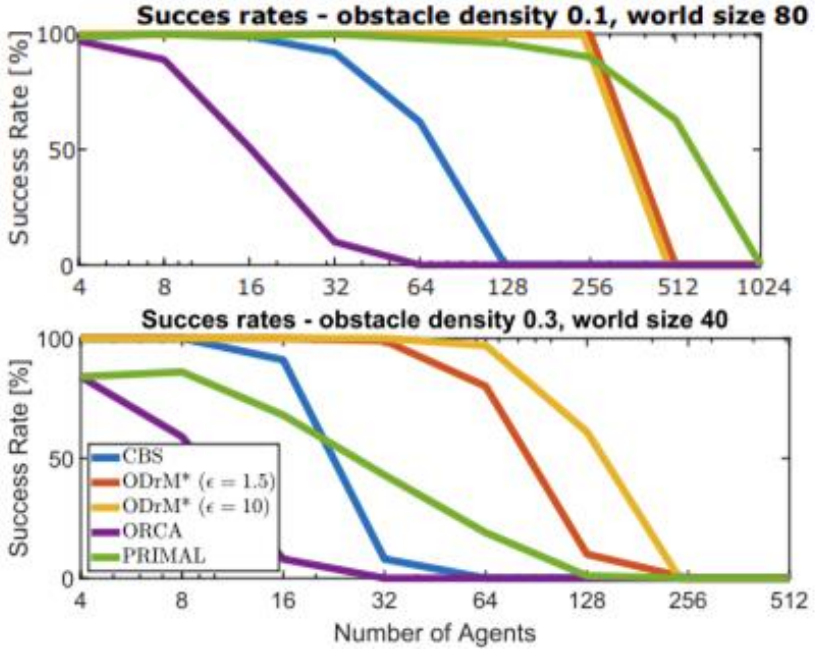
- Experimenty:
  - Velikost světa: 10; 20; 40; 80; 160
  - Hustota překážek: 0; 0,1; 0,2; 0,3
  - Velikost týmu: 4; 8;...; 1024

Porovnávané plánovače

- CBS
  - = Conflict Based Search
  - Hledání cesty pomocí stromu konfliktů.
- ODrM*
  - Kombinace M*, A* a CBS.
- ORCA
  - = Optimal Reciprocal Collision Avoidance
  - Kombinuje individuální hledání cesty a reciproční vyhýbání kolizí.

Výsledky





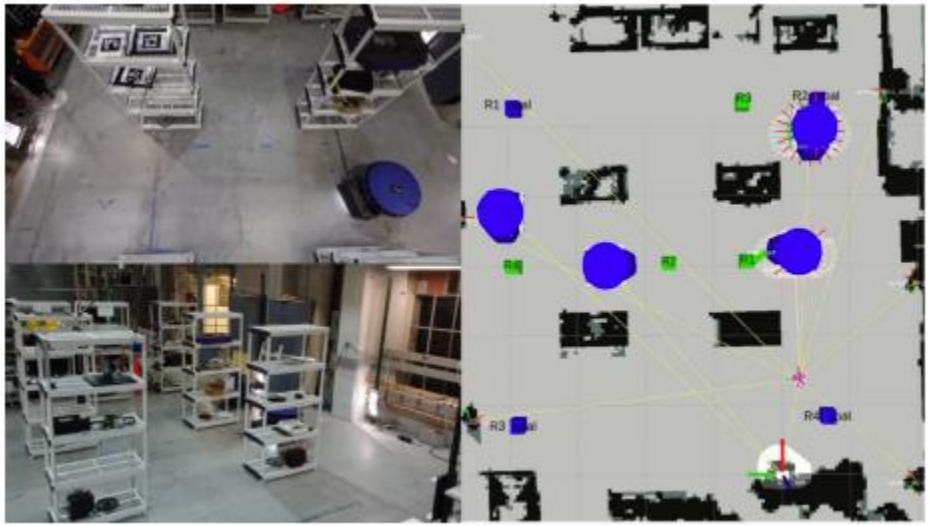
5. Experimenty

Porovnání s ostatními plánovači

- PRIMAL je **lepší** v prostředí s málo překážkami.
- PRIMAL je **horší** v prostředí s více překážkami.
  - https://youtu.be/uNVW1qe2dhQ
- Více agentů v okolí - špatné chování.
  - Malý svět a velký tým.
  - Chyba v trénování (konstantní počet agentů).
- Agenti plánují až dvakrát delší cesty.
  - Odlišná omezení MAPF.
  - Agenti se mohou bezprostředně následovat, mohou se prohazovat,...

Experiment s roboty

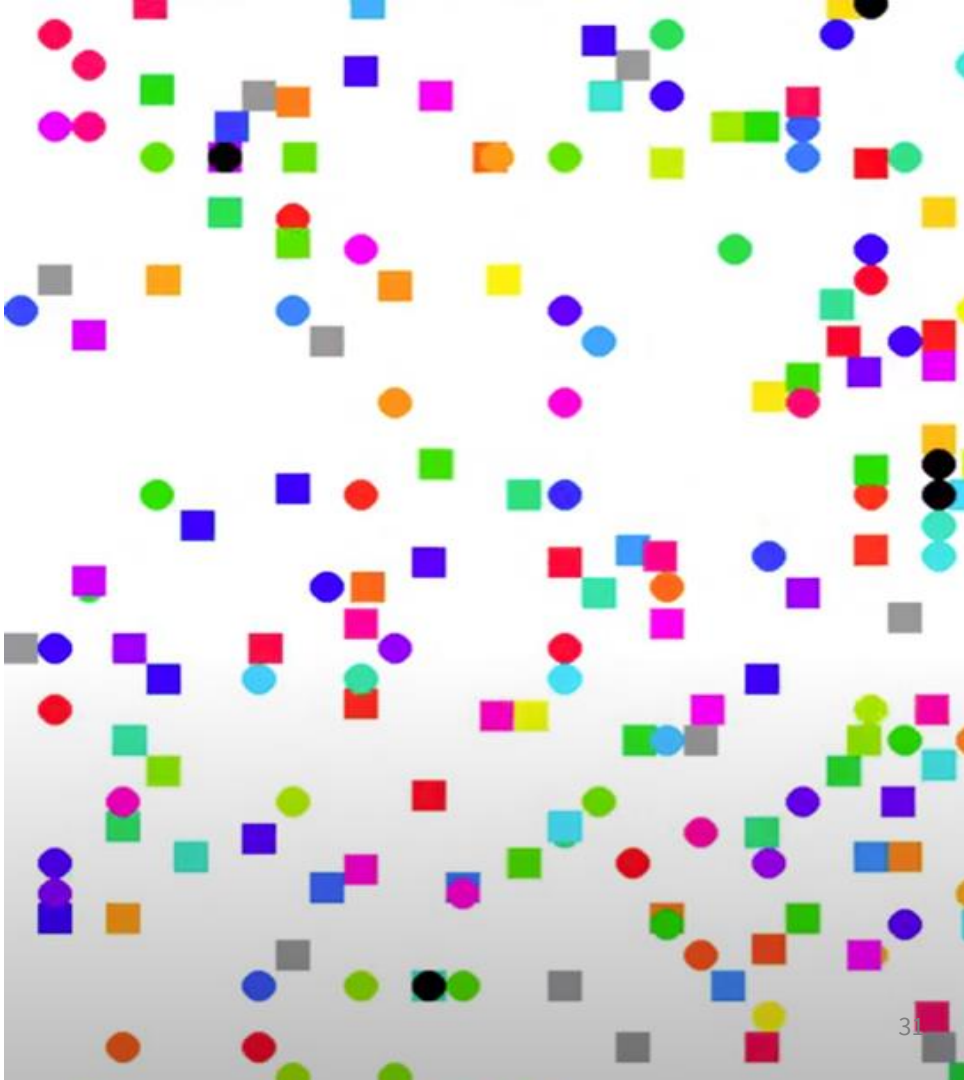
Hybrid Simulation - 2 Physical Robots and up to 3 Simulated Ones



5. Experimenty

Osnova

1. Úvod
2. Důležité pojmy
3. MAPF a reinforcement learning
4. Učení
5. Experimenty
6. Závěr



Shrnutí

- PRIMAL kombinuje distribuovaný reinforcement learning a imitation learning.
- Agent pracuje pouze s lokálními informacemi o světě.
- Pracuje velmi dobře v prostředí s nízkou hustotou překážek.
- Autoři se chtějí v budoucnu zaměřit na trénování agentů v továrnách apod.

- SARTORETTI, Guillaume, et al. Primal: Pathfinding via reinforcement and imitation multi-agent learning. IEEE Robotics and Automation Letters, 2019, 4.3: 2378-2385.
- Videa ke článku: https://youtube.com/playlist?list=PLt2UiOV2mr9lujyYrtrgXt8CF1ORd7CHa

BONUS: Učení actor-critic sítě

- Celková diskontovaná odměna:

$$R_t = \sum_{i=0}^{k} \gamma^i r_{t+i}$$

- Minimalizace ztrátové funkce hodnoty stavu:

$$L_V = \sum_{t=0}^{T} (V(o_t;\theta) - R_t)^2$$

- Minimalizace ztrátové funkce strategie:

$$L_\pi = \sigma_H \cdot H(\pi(o)) - \sum_{t=0}^{T} \log(P(a_t|\pi, o;\theta) A(o_t, a_t;\theta))$$