Self-supervised Visual Reinforcement Learning with Object-centric Representations

Andrii Zadaianchuk, Maximilian Seitzer, Georg Martius

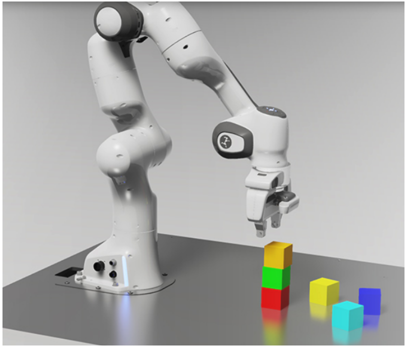
ETHZ and MPI IS

andrii.zadaianchuk@tuebingen.mpg.de

May 5, 2021

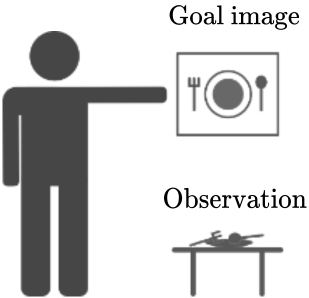
Real-life challenges for Autonomous Learning



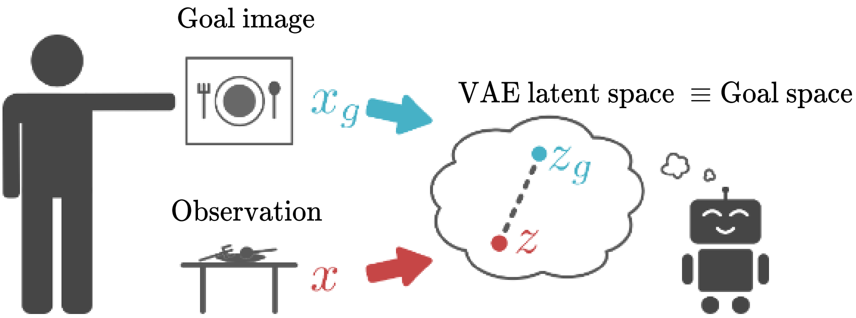
- Learning is **self-supervised** (no reward signal during training)
- Observations are **high-dimensional**
- Tasks and observations are **compositional**

image sources: parentmap.com, nvidia.com

Prior work: Self-supervised Visual RL



Prior work: Self-supervised Visual RL



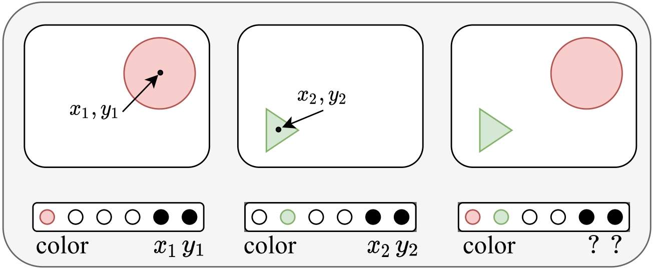
Usage of the VAE latent space as goal space

- VAE latent space is used as goal space
- Reward signal based on distance in VAE latent space

Image from Nair et al., 2018

Problems with VAE goal space

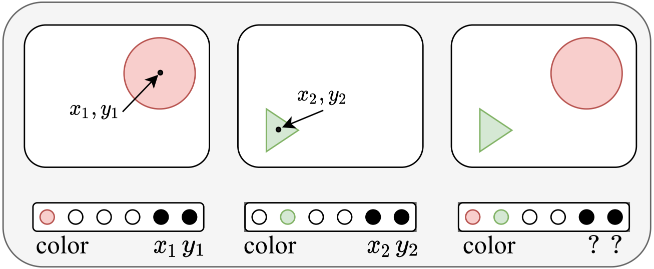
- Distributed VAE representation suffers from **binding problem** [1]



[1] [Greff et al., 2016, Greff et al., 2020]

Problems with VAE goal space

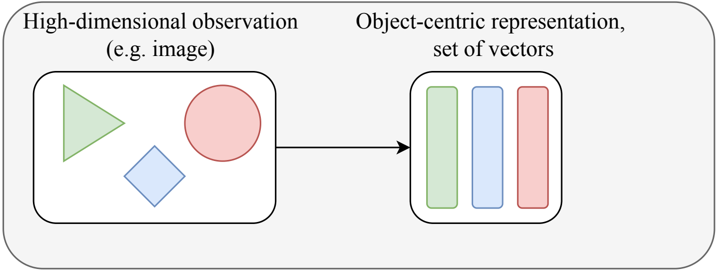
- Distributed VAE representation suffers from **binding problem** [1]



- Some dimensions encode **task-irrelevant information**

[1] [Greff et al., 2016, Greff et al., 2020]

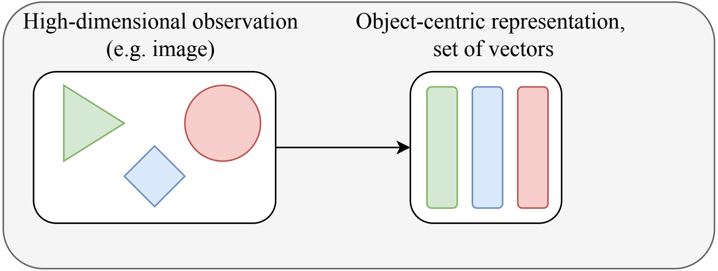
Object-centric representations provide better goals



Object-centric representation for multi-object observations

- Observation is represented as set of (low-dimensional) vectors

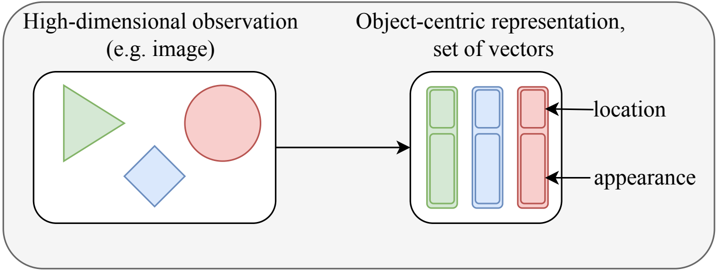
Object-centric representations provide better goals



Object-centric representation for multi-object observations

- Observation is represented as set of (low-dimensional) vectors
- Learning of representations is fully unsupervised

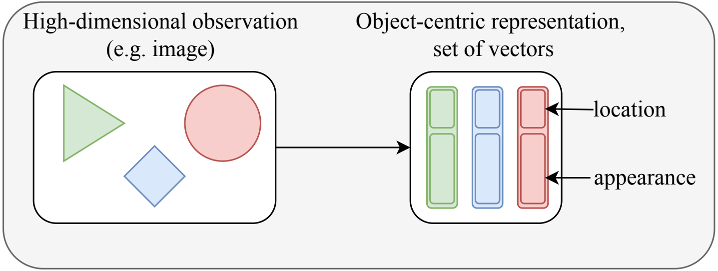
Object-centric representations provide better goals



Object-centric representation for multi-object observations

- Observation is represented as set of (low-dimensional) vectors
- Learning of representations is fully unsupervised
- Each object representations can be additionally structured

Object-centric representations provide better goals

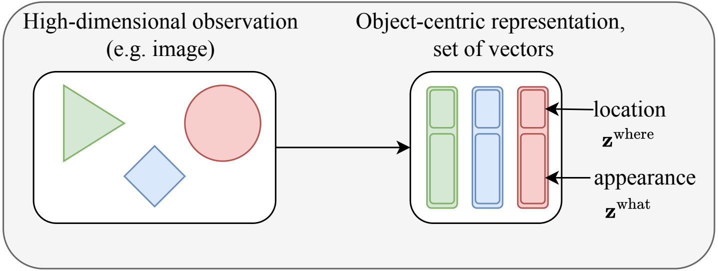


Object-centric representation for multi-object observations

SCALable sequential **O**bject-oriented **R**epresentations (**SCALOR**) [2]

[2] [Jiang et al., 2019]

Object-centric representations provide better goals



Object-centric representation for multi-object observations

SCALable sequential **O**bject-oriented **R**epresentations (**SCALOR**) [2]

[2][Jiang et al., 2019]

Our contributions

- Developed **S**elf-supervised **M**ulti-**O**bject **RL** (**SMORL**) agent that autonomously learns skills in compositional environments
- Designed **goal-conditioned attention policy** compatible with object-centric representations
- Proposed **efficient self-supervised training** that exploits structured latent space

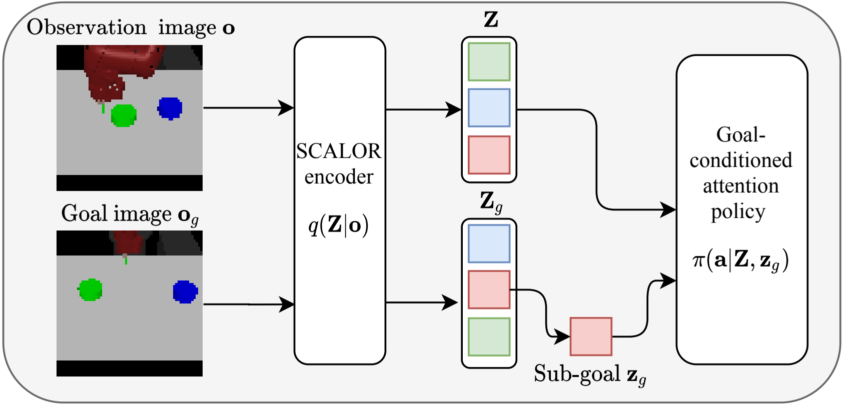
Our contributions

- Developed **S**elf-supervised **M**ulti-**O**bject **RL** (**SMORL**) agent that autonomously learns skills in compositional environments
- Designed **goal-conditioned attention policy** compatible with object-centric representations
- Proposed **efficient self-supervised training** that exploits structured latent space

Our contributions

- Developed **S**elf-supervised **M**ulti-**O**bject **RL (SMORL)** agent that autonomously learns skills in compositional environments
- Designed **goal-conditioned attention policy** compatible with object-centric representations
- Proposed **efficient self-supervised training** that exploits structured latent space

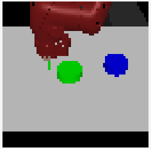
Self-supervised Multi-Object RL (**SMORL**)



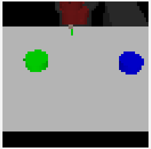
SMORL pipeline during evaluation

Self-supervised Multi-Object RL (**SMORL**)

Observation image $\mathbf{o}$

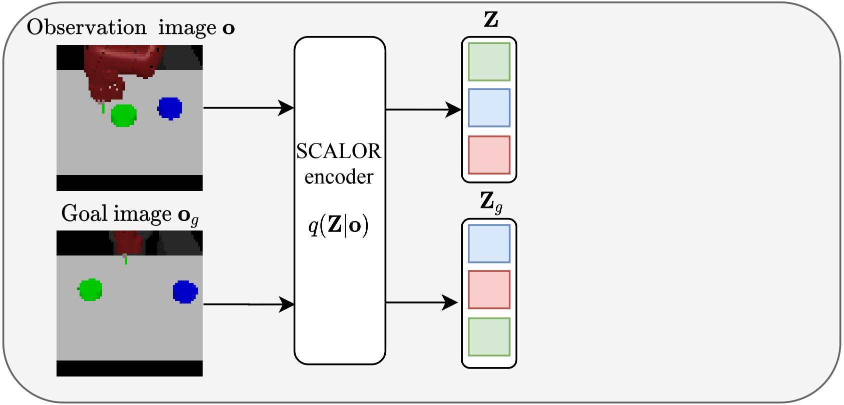


Goal image $\mathbf{o}_g$



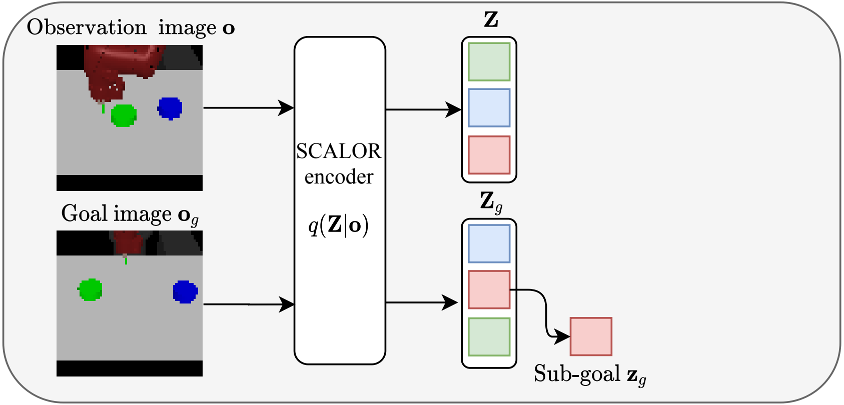
SMORL pipeline during evaluation

Self-supervised Multi-Object RL (**SMORL**)



SMORL pipeline during evaluation

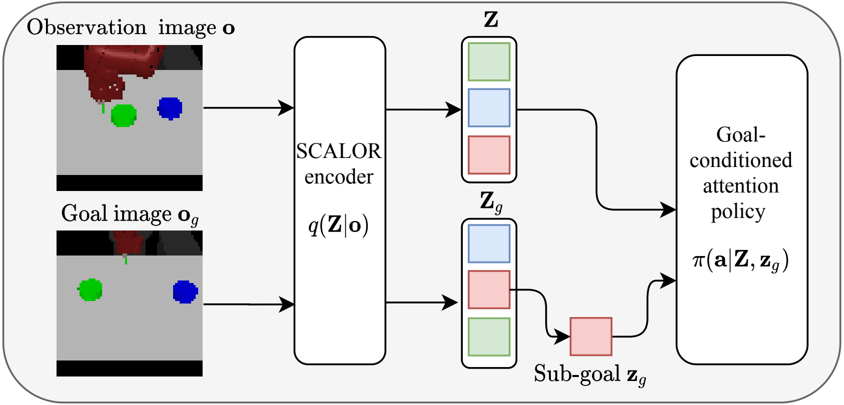
Self-supervised Multi-Object RL (**SMORL**)



SMORL pipeline during evaluation

- Learned policy is sequentially achieving all the recognised sub-goals.

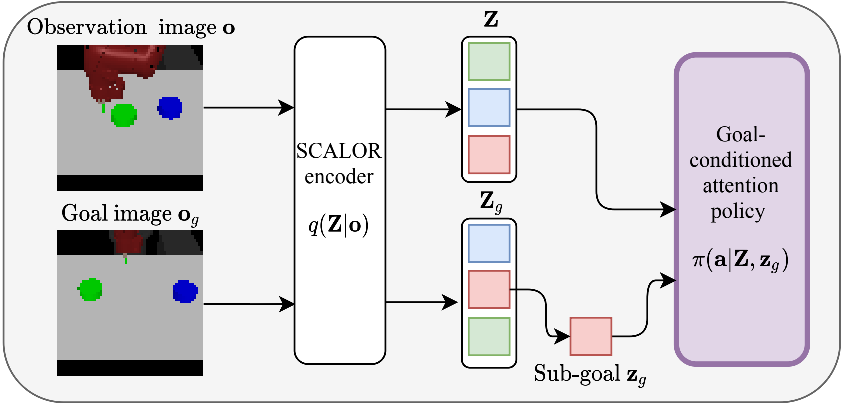
Self-supervised Multi-Object RL (**SMORL**)



SMORL pipeline during evaluation

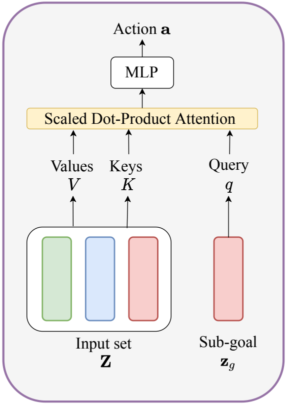
- Learned policy is sequentially achieving all the recognised sub-goals.

Self-supervised Multi-Object RL (**SMORL**)



SMORL pipeline during evaluation

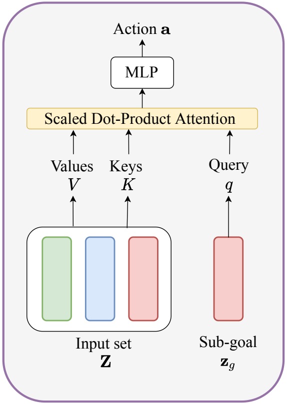
Goal-conditioned attention policy



Goal-conditioned attention policy

- Compatible with variable-size input sets Z

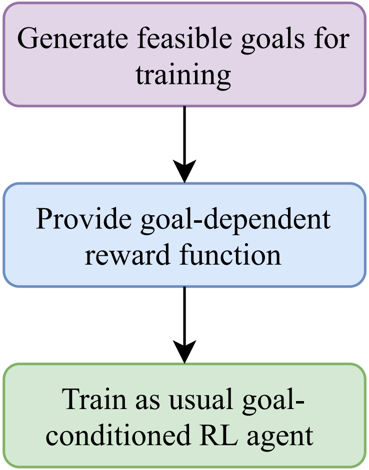
Goal-conditioned attention policy



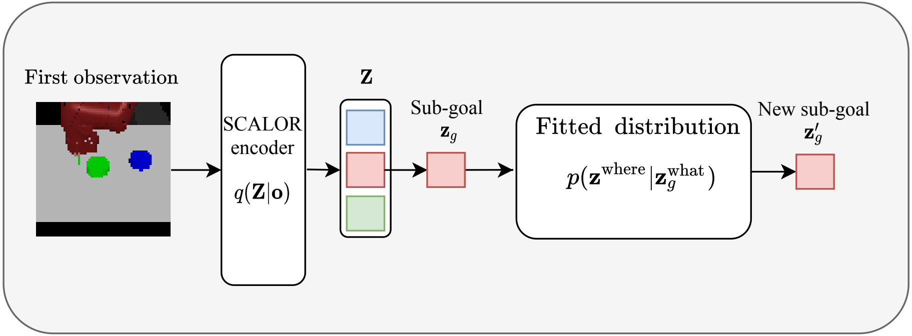
Goal-conditioned attention policy

- Compatible with variable-size input sets Z
- Attend to elements of the input set Z that are important for current goal z_g

SMORL training is self-supervised

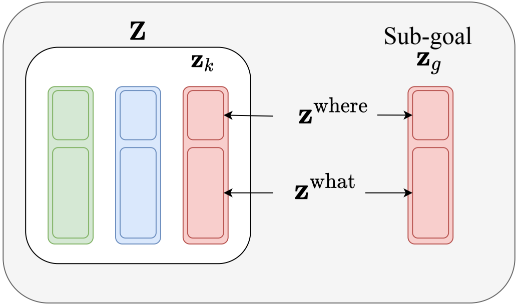


Goal generation during training



- Fit $p(z^{\text{where}}|z^{\text{what}})$ to observed data to estimate valid locations
- Pick random object representation $z_g = (z_g^{\text{where}}, z_g^{\text{what}})$
- Sample new z^{where} from $p(z^{\text{where}}|z_g^{\text{what}})$

Reward function in structured latent space



- Find most similar object: $k = \arg\min_i ||z_i^{\text{what}} - z_g^{\text{what}}||$
- Reward in subspace of locations: $r(Z, z_g) = -||z_k^{\text{where}} - z_g^{\text{where}}||$

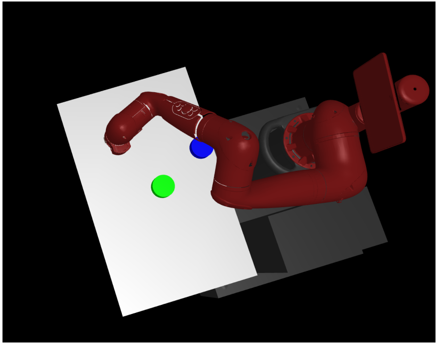
SMORL training combines SAC with object-centric representations

Algorithm 1 Self-Supervised Multi-Object RL (SMORL) training

Require: SCALOR encoder q_ϕ, goal-conditioned policy π_θ, goal-conditioned SAC trainer, number of training episodes K.

1: Train SCALOR on sequences data $\mathcal{D}$ from random policy.
2: **for** $n = 1, \ldots, K$ episodes **do**
3: Sample goal $\mathbf{z}_g = (\hat{\mathbf{z}}_g^{\text{where}}, \mathbf{z}_g^{\text{what}})$
4: Collect episode data with $\pi_\theta(\mathbf{a}_t | q_\phi(\mathbf{o}_t), \mathbf{z}_g)$ and $q_\phi(\mathbf{Z}_t | \mathbf{o}_t)$.
5: Store transitions $(\mathbf{Z}_t, \mathbf{a}_t, \mathbf{Z}_{t+1}, \mathbf{z}_g)$ into replay buffer $\mathcal{R}$.
6: Sample transitions from replay buffer $(\mathbf{Z}, \mathbf{a}, \mathbf{Z}', \mathbf{z}_g) \sim \mathcal{R}$
7: Compute matching reward signal $r = r(\mathbf{Z}', \mathbf{z}_g)$.
8: Update policy $\pi_\theta(\mathbf{Z}_t | q_\phi(\mathbf{o}_t), \mathbf{z}_g)$ with SAC trainer.
9: **end for**

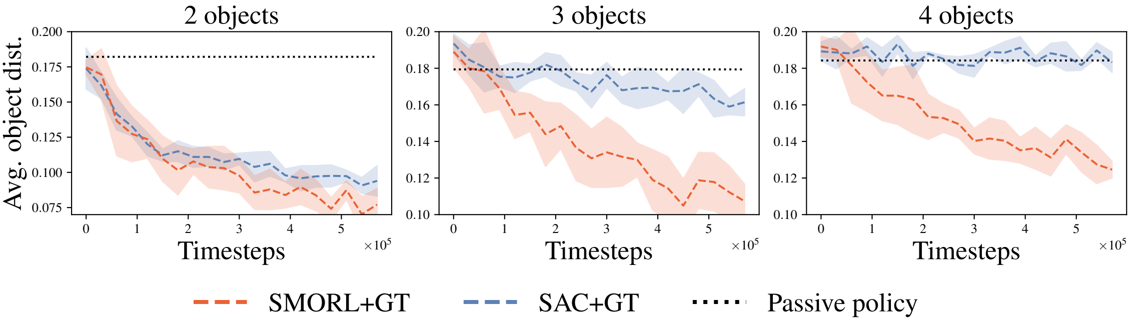
Visual Multi-object Rearrange Environment



Visual Multi-object Rearrange Environment

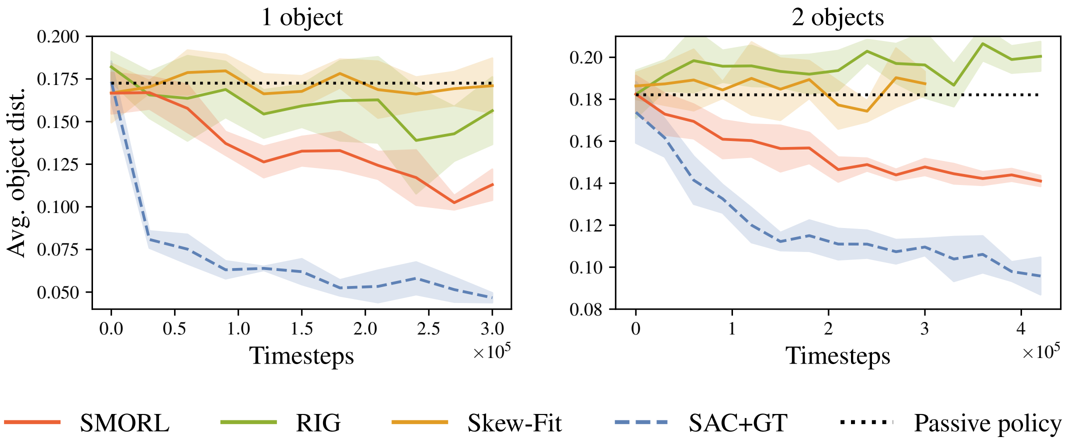
- Multi-object version of `multiworld` environment
- Objects placed randomly each episode, so that agent can not just memorize initial optimal actions.

SMORL with GT representation



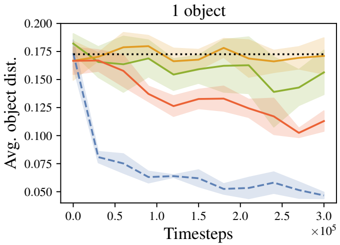
SMORL with high-dimensional observations

Visual Rearranging



Qualitative results on Visual Rearrange environment

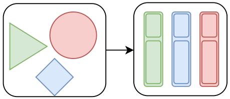
Visual Rearrange environment



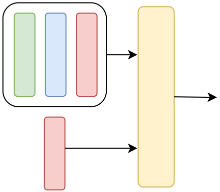
For more videos visit:

martius-lab.github.io/SMORL

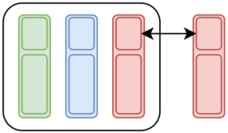
Conclusions



Object-centric representations improve performance of self-supervised visual RL agent



Goal-conditioned attention policy aggregates object-centric representations with focus on current goal



Additional structure in each object representation exploited for goal generation and reward function

Questions?

Poster session 8: May 5, 9-11 a.m. PDT

Project website: martius-lab.github.io/SMORL

Contact: andrii.zadaianchuk@tuebingen.mpg.de

References

- Greff, K., Srivastava, R. K., and Schmidhuber, J. (2016).
  Binding via reconstruction clustering.
- Greff, K., van Steenkiste, S., and Schmidhuber, J. (2020).
  On the binding problem in artificial neural networks.
- Jiang, J., Janghorbani, S., de Melo, G., and Ahn, S. (2019).
  Scalable object-oriented sequential generative models.
  arXiv preprint arXiv:1910.02384.