The (Un)Surprising Effectiveness of Pre-Trained Vision Models for Control

Meta AI

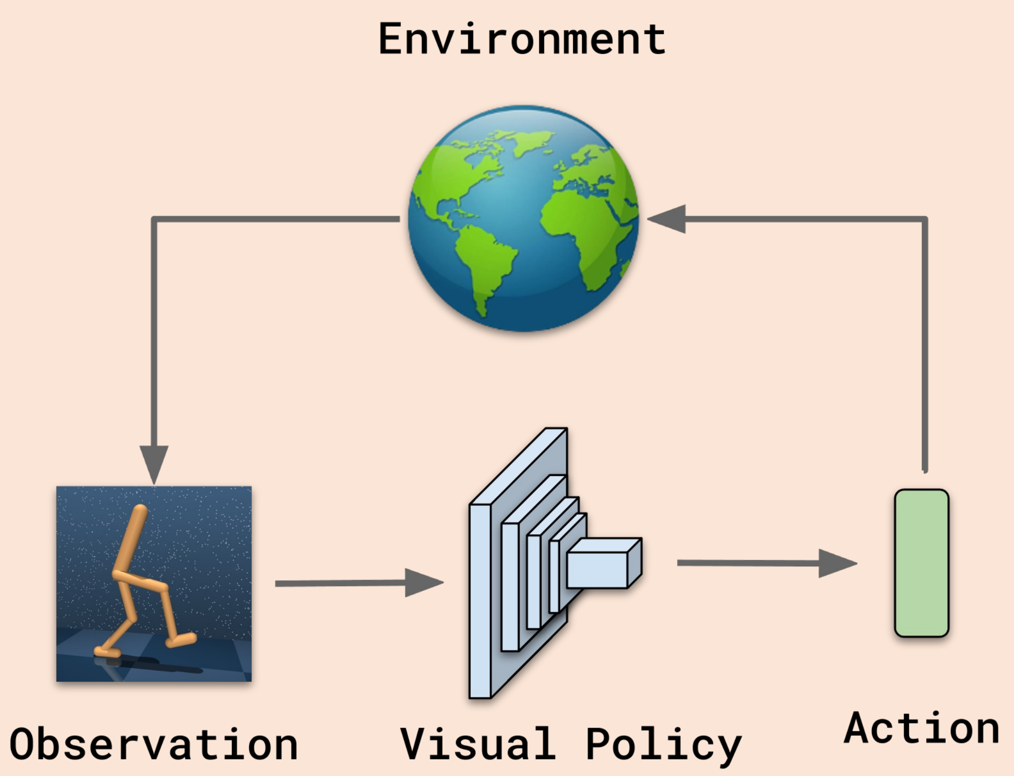
Simone Parisi*, Aravind Rajeswaran*,

Senthil Purushwalkam, Abhinav Gupta

CMU

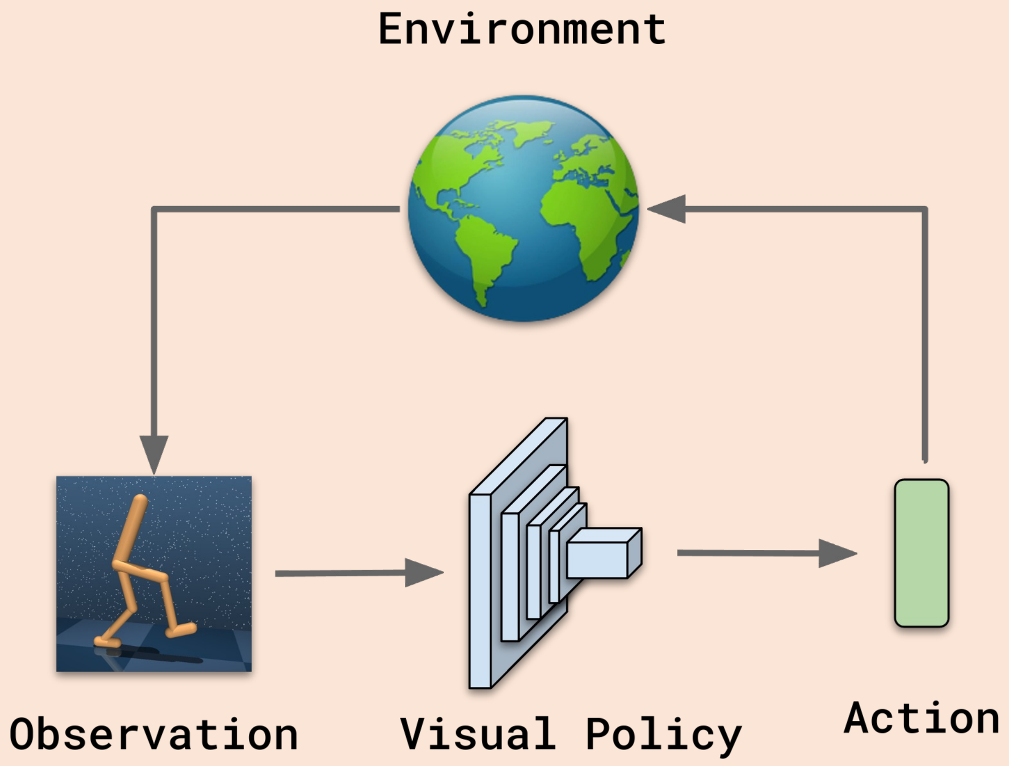
Policy Learning from Visual Inputs

Perception – Action Loop



Policy Learning from Visual Inputs

Perception – Action Loop



Applications

Robotics

(physical hardware)

Embodied AI agents

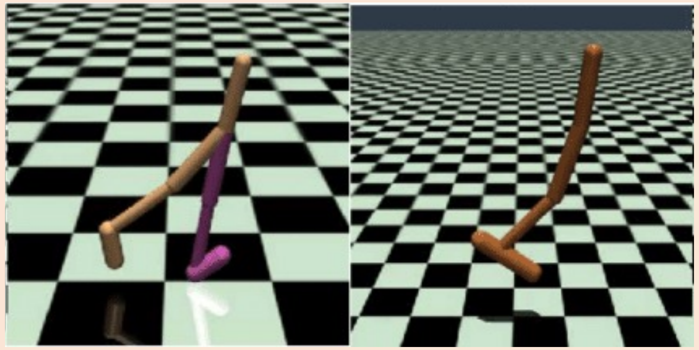
in virtual worlds

Others: content recommendation based on visual characteristics, egocentric personal assistants etc.

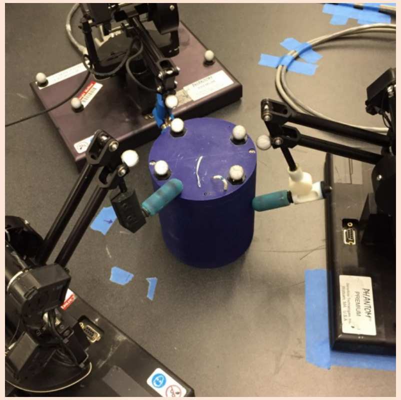
Policy Learning for Control/Robotics

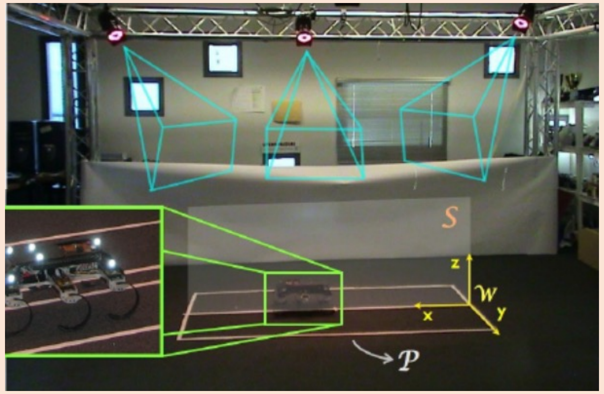
Type 1 : Compact State Spaces

Directly from simulators



From motion capture systems

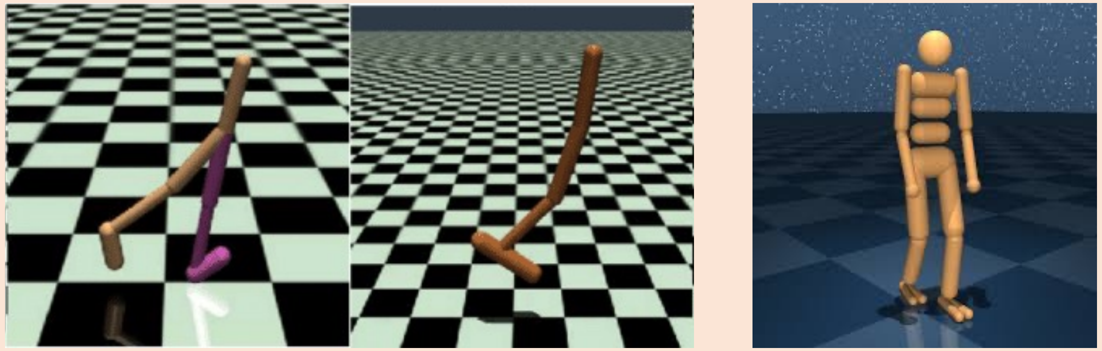




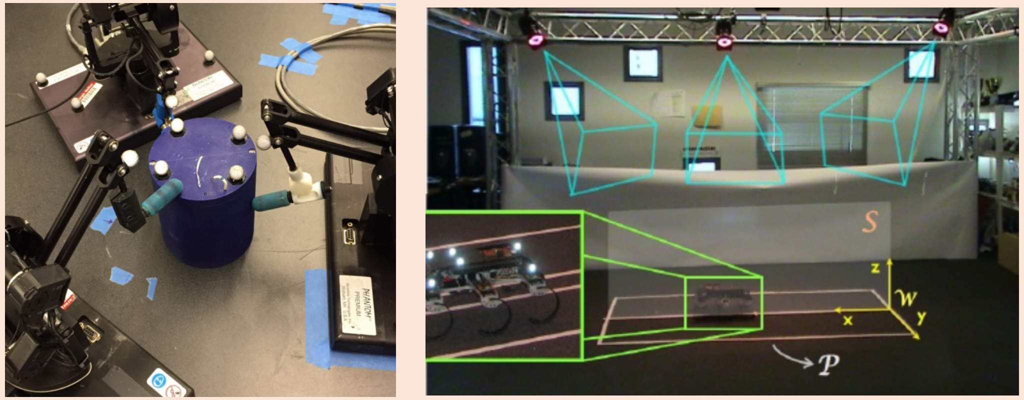
Policy Learning for Control/Robotics

Type 1 : Compact State Spaces

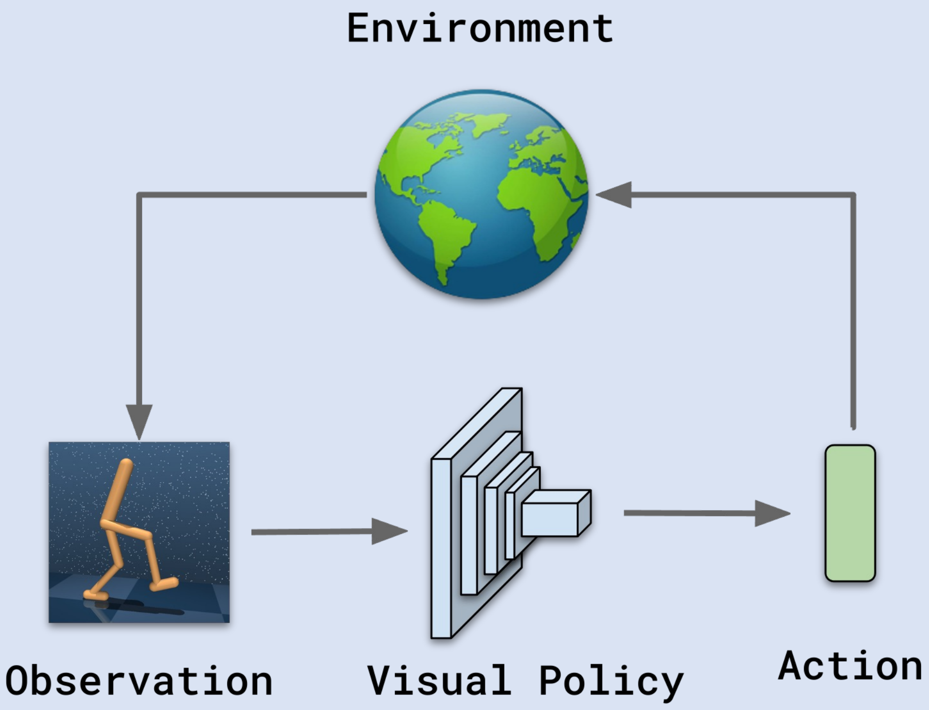
Directly from simulators



From motion capture systems



Type 2 : Tabula-Rasa End-to-End Policies

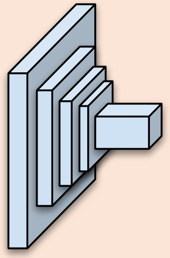


(Mostly) learn entire visuo-motor policy from scratch, or

(Sometimes) highly domain specific pretraining

Pre-Training & Self-Supervision in Vision/NLP

Task-Agnostic Pre-Training



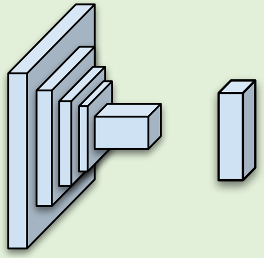
Pre-Training Objective

Large & Diverse Data Bank

Large & Flexible Neural Model

Generic Objective (Contrastive, masking..)

Task-Specific Adaptation



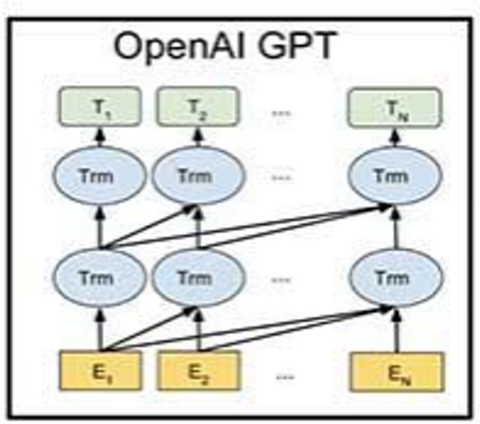
Classification

Translation

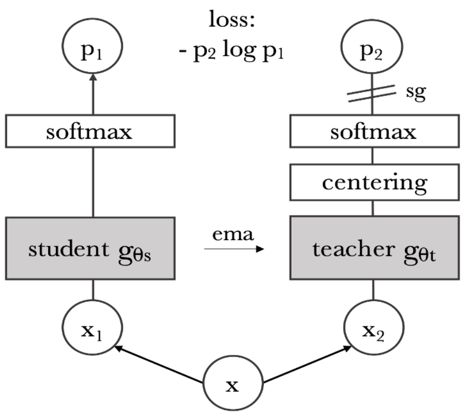
and more …

Pre-Trained “Foundation” Model

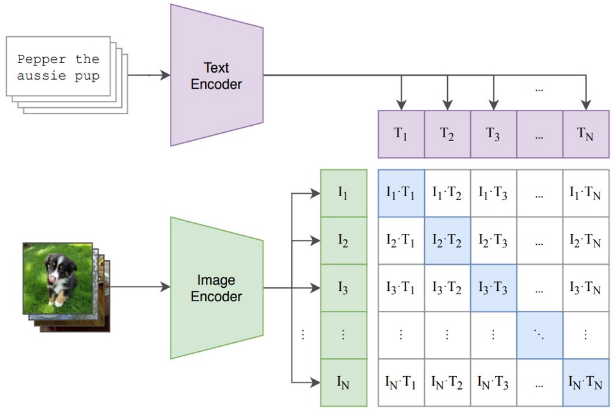
Downstream Tasks



GPT-X / BERT / RoBERTa
1+ trillion words

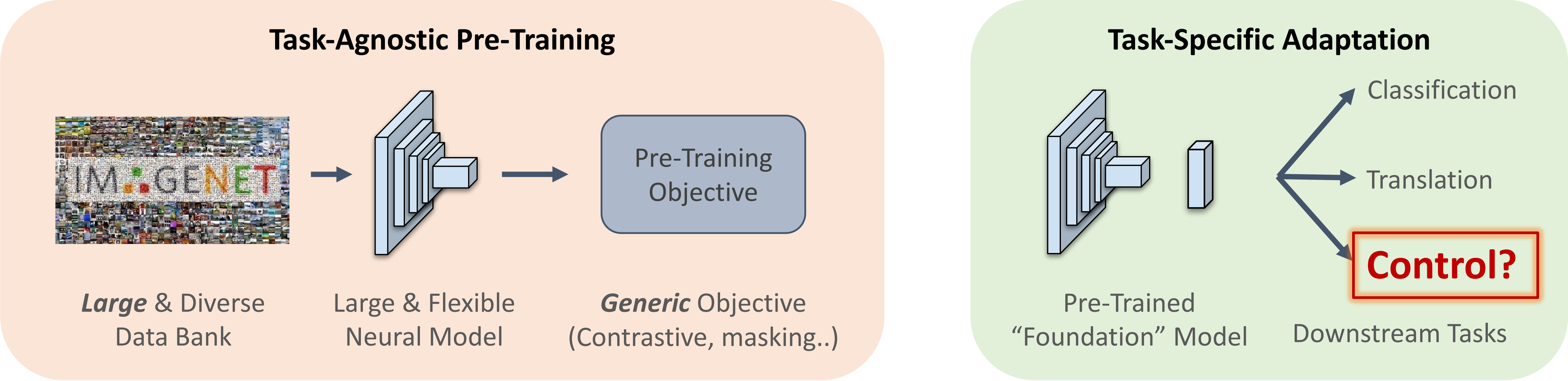


MoCo / SimCLR / DINO / MAE
(ImageNet without labels)



OpenAI CLIP, 400 million
Image-Caption pairs

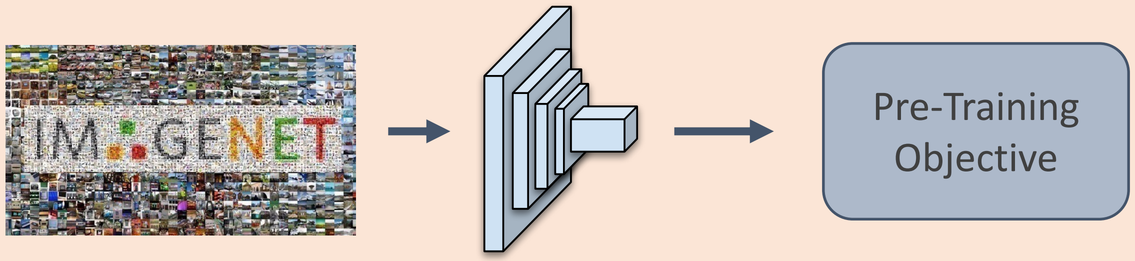
Pre-Training & Self-Supervision in Vision/NLP



*Can a **single vision model**, pre-trained entirely on **out-of-domain** passive datasets, work for diverse control tasks?*

Pre-Training & Self-Supervision in Vision/NLP

Task-Agnostic Pre-Training

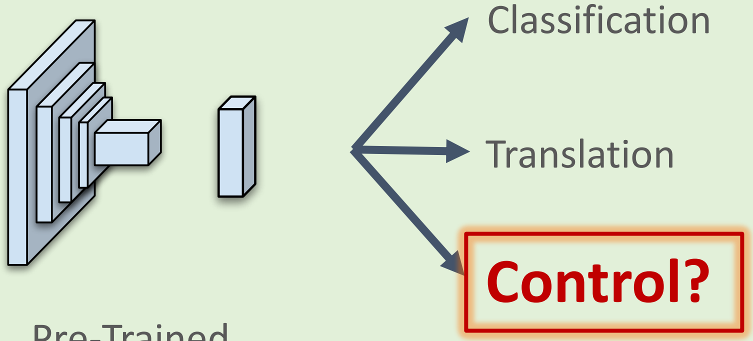


Large & Diverse Data Bank

Large & Flexible Neural Model

Generic Objective (Contrastive, masking..)

Task-Specific Adaptation

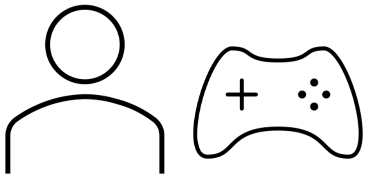


Pre-Trained "Foundation" Model

Downstream Tasks

We will evaluate pre-trained visual representations with *few-shot imitation learning*

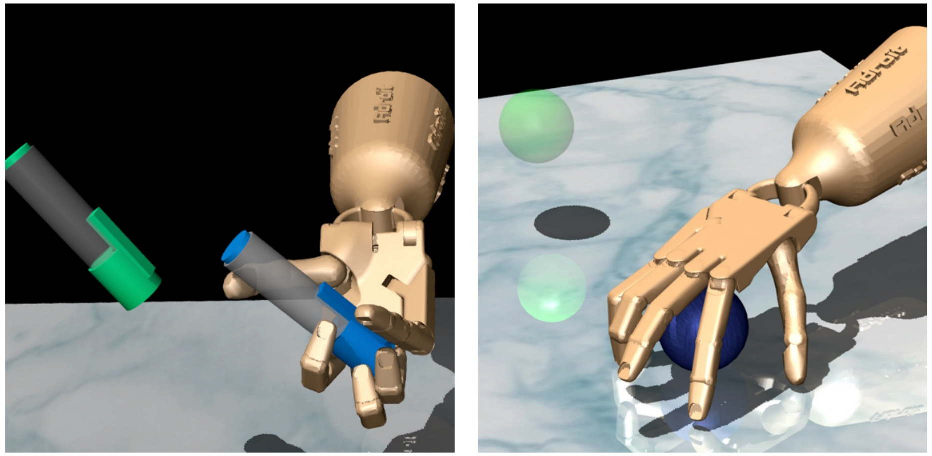
Use Frozen Pre-Trained Representation → Collect Few Demonstrations → Train small MLP Policy with frozen PVR embeddings → Deploy Policy with Representation



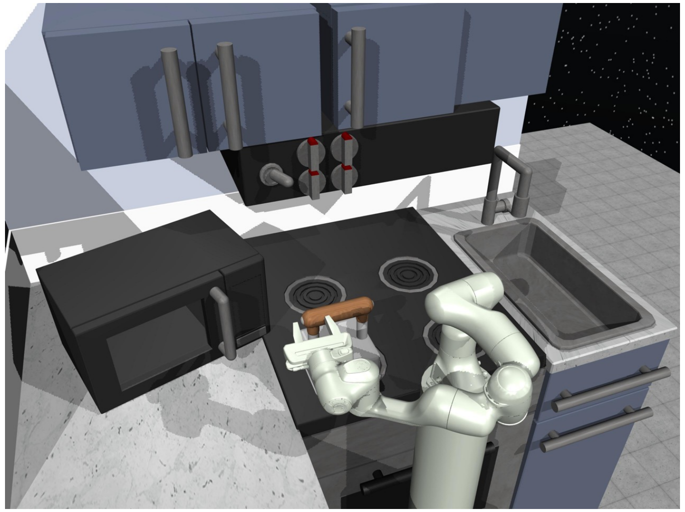
Evaluation Domains

Habitat ImageNav (Replica Dataset; 5 scenes)

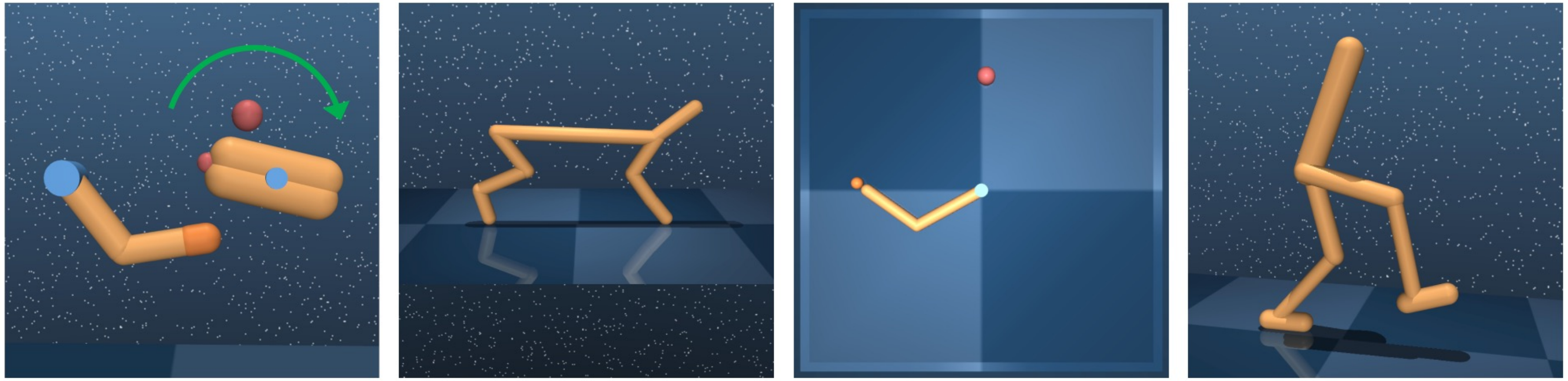
Adroit Dexterous Manipulation (2 hardest tasks)



Franka Kitchen (5 tasks)



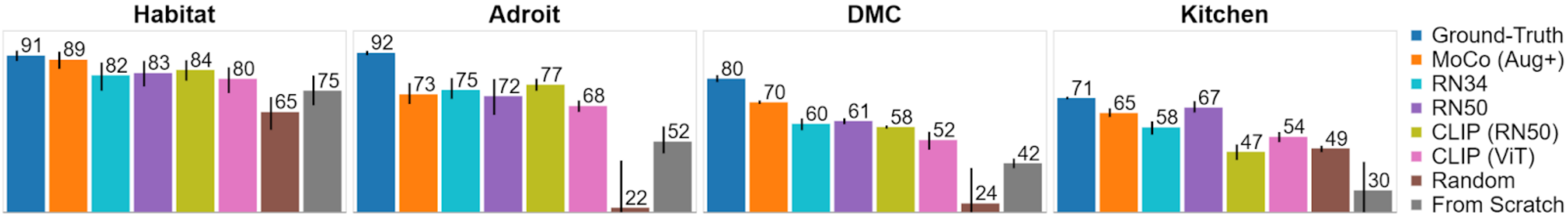
DeepMind Control Suite (5 tasks)



Results with Frozen PVRs

Q: How well do pre-trained vision models work off-the-shelf?

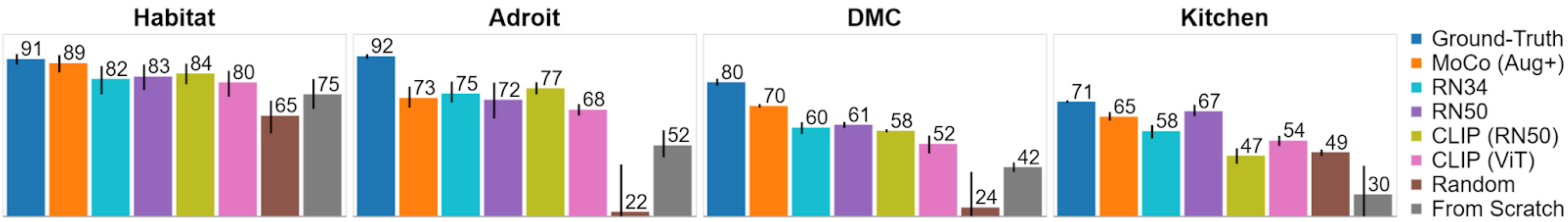
- Frozen PVRs (off-the-shelf) > frozen random features / end-to-end learning
- Self-supervised representations > supervised representations



Results with Frozen PVRs

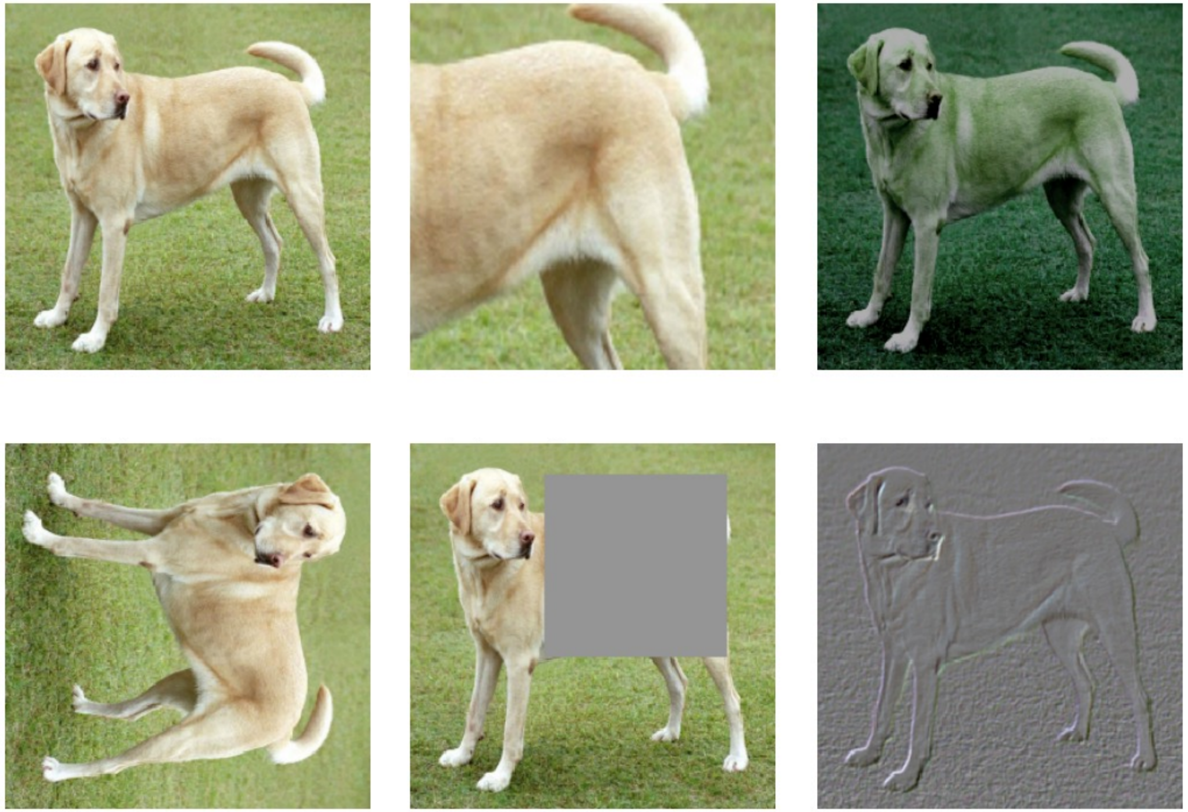
Q: How well do pre-trained vision models work off-the-shelf?

- ➢ Frozen PVRs (off-the-shelf) > frozen random features / end-to-end learning
- ➢ Self-supervised representations > supervised representations
- ➢ ✅ Habitat: MoCo features competitive with states out-of-the-box!
- ➢ ❌ Remaining domains: Still sizable gap between states and PVRs

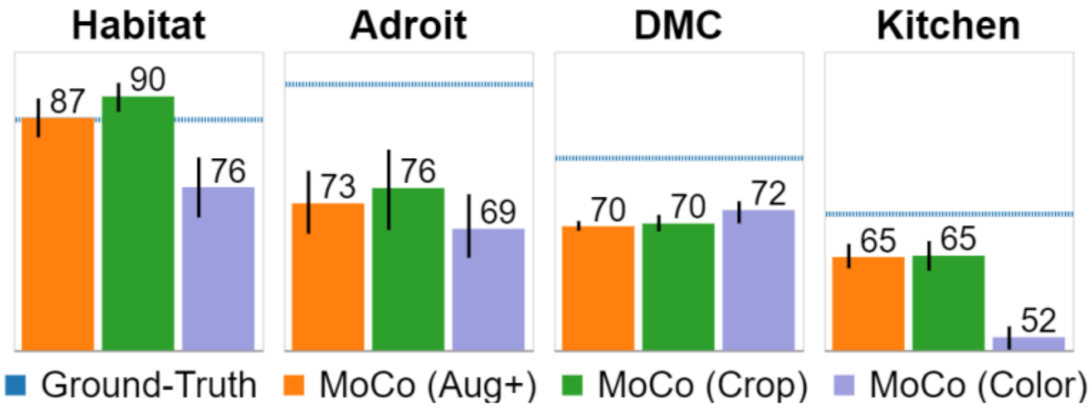


Recognition vs Control

Q: Does augmentations make a difference in SSL?



Increase similarity between embeddings of all these images

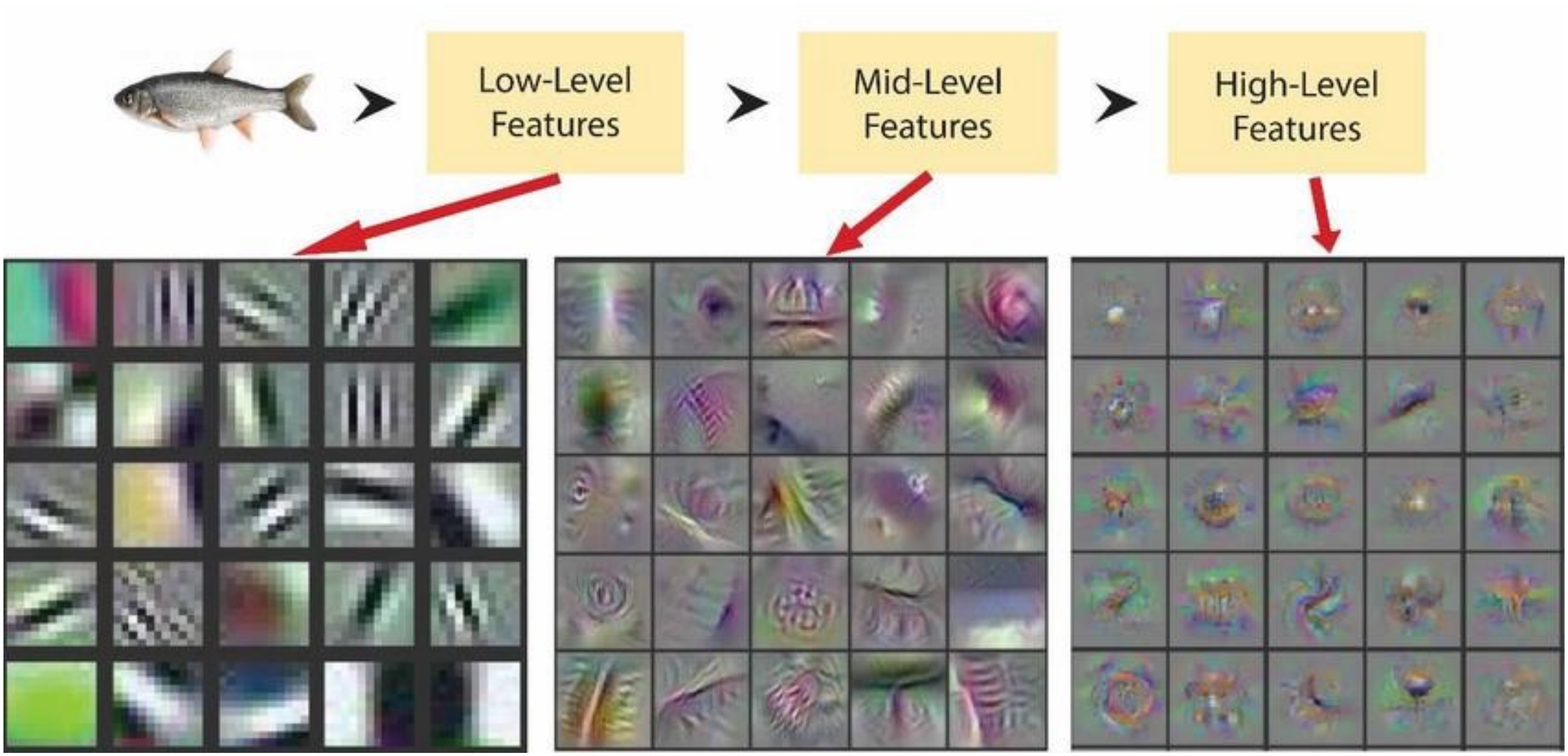


- Crop augmentations are most important (consistent with prior works, e.g. CURL, DrQ)
- Removing color aug helps in most cases

Semantic Recognition and Control require different visual invariances

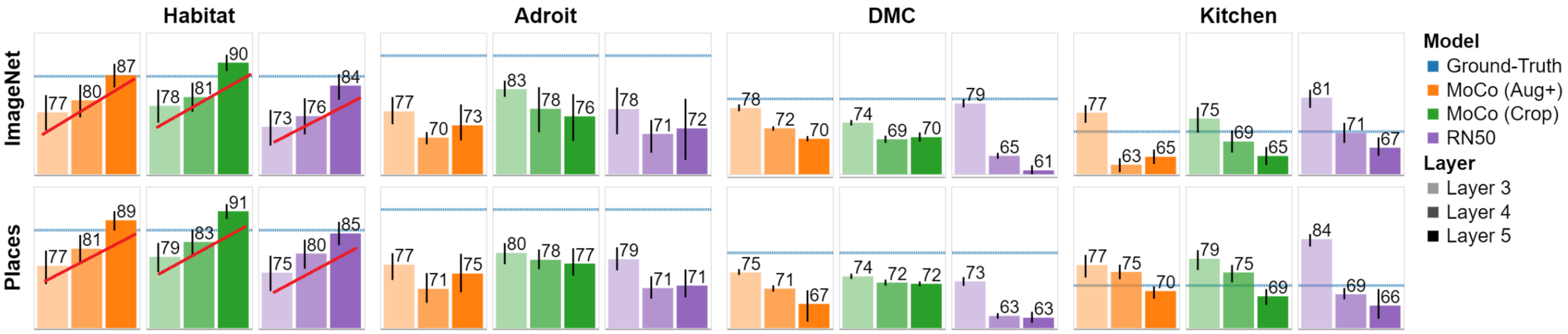
Image Credit: Chen et al. 2020 (SimCLR)

Different Layers Encode Different Invariances



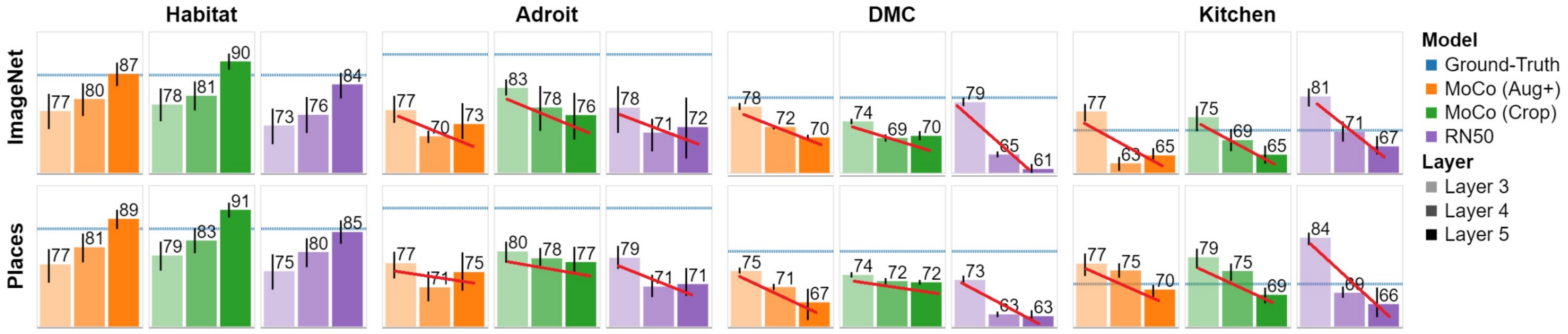
Results

➤ ✅ Later layer features are better for high-level semantic tasks (Habitat ImageNav)



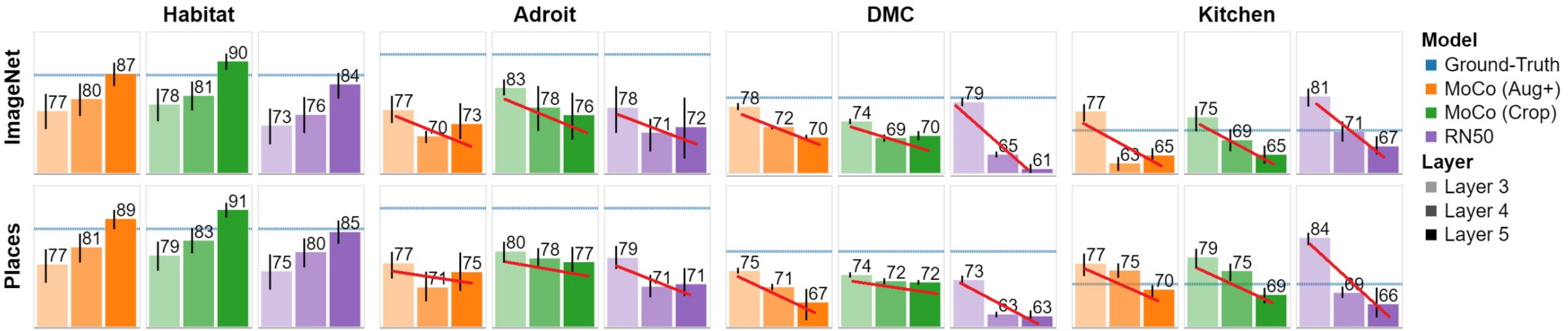
Results

- ✅ Later layer features are better for high-level semantic tasks (Habitat ImageNav)
- ✅ ***Early layer*** features are better for fine-grained control tasks (manipulation in MuJoCo)



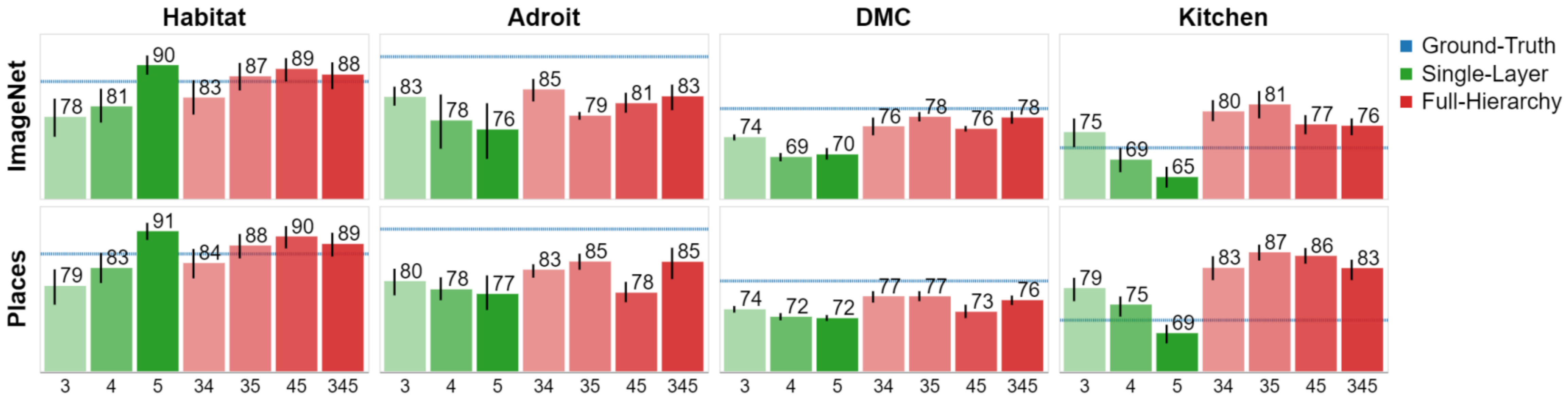
Results

- ➢ Later layer features are better for high-level semantic tasks (Habitat ImageNav)
- ➢ ***Early layer*** features are better for fine-grained control tasks (manipulation in MuJoCo)
- ➢ ✅ Early layer features are competitive with ground truth states in MuJoCo tasks
- ➢ Trends consistently true across multiple models, environments, and datasets



Results

- Combine features from multiple layers ➔ single vision model that works across the board?
- MoCo with Layer 5 : ❌ MuJoCo ✅ Habitat
- MoCo with Layer 3 : ✅ MuJoCo ❌ Habitat
- MoCo layers 3-4-5 : ✅ MuJoCo ✅ Habitat

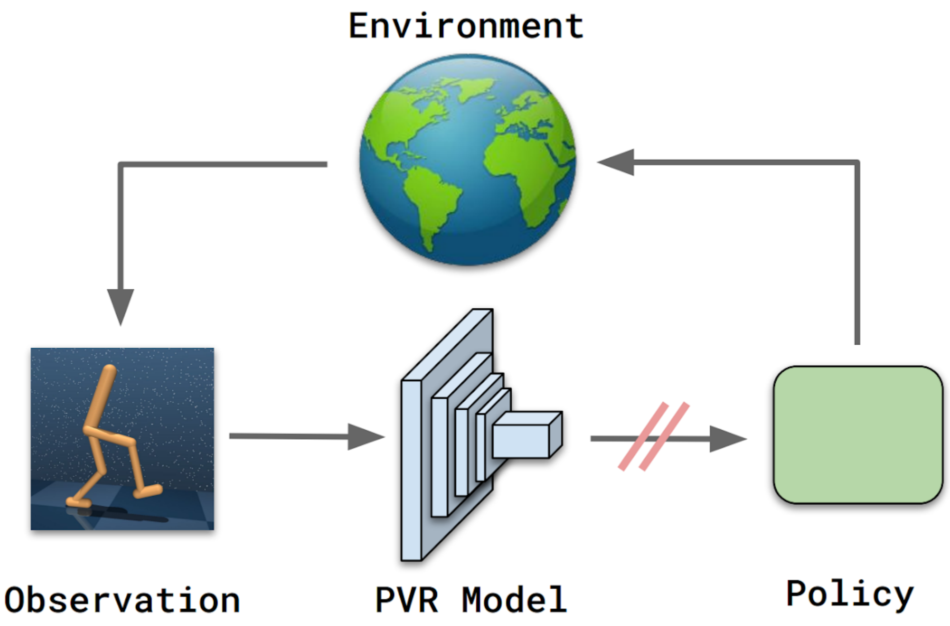


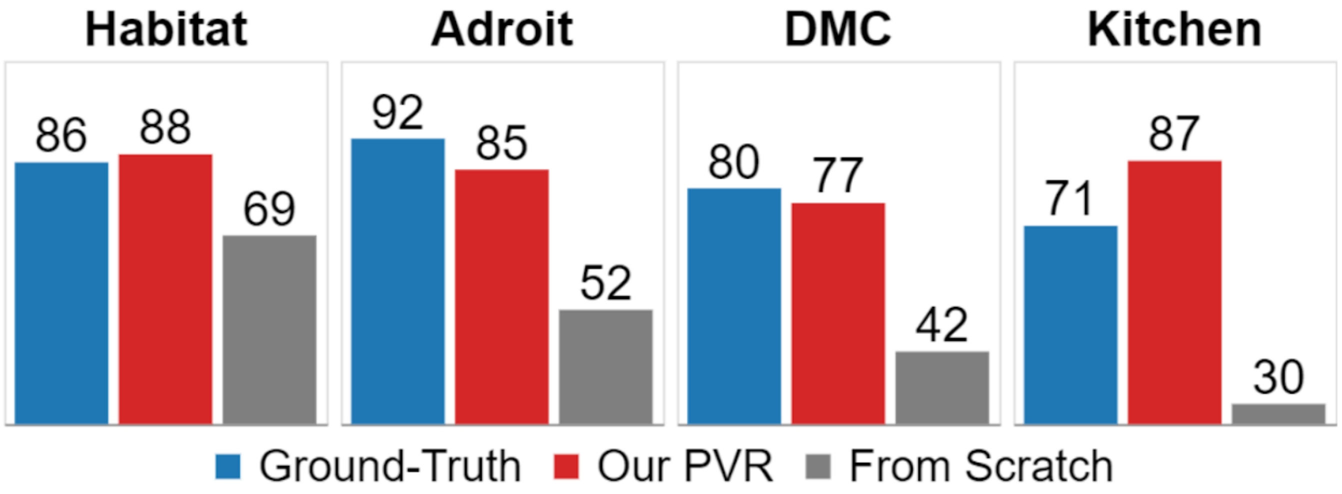
Summary

*Can a **single vision model**, pre-trained entirely on*

***out-of-domain** passive datasets, work for diverse control tasks?*

YES !!!





Take Home Message

Move away from tabula-rasa training

$\Downarrow$

Train control policies using pre-trained perception modules

$\Downarrow$

Save time, data, expertise

https://sites.google.com/view/pvr-control