DeepRob

[Group 5] Lecture 5
Multisensory Learning + Manipulation

by Mason Hawver, Ryan Diaz, and Hanchen Cui
University of Minnesota

Image Source: D. F. Gomes, P. Paoletti, and S. Luo, “Generation of gelsight tactile images for sim2real learning,” IEEE Robotics and Automation Letters, vol. 6, no. 2, pp. 4177–4184, 2021.

Outline

Each person has one theme and take away - think about that when adding slides to the deck

mason - datagen

hanchen- multimae

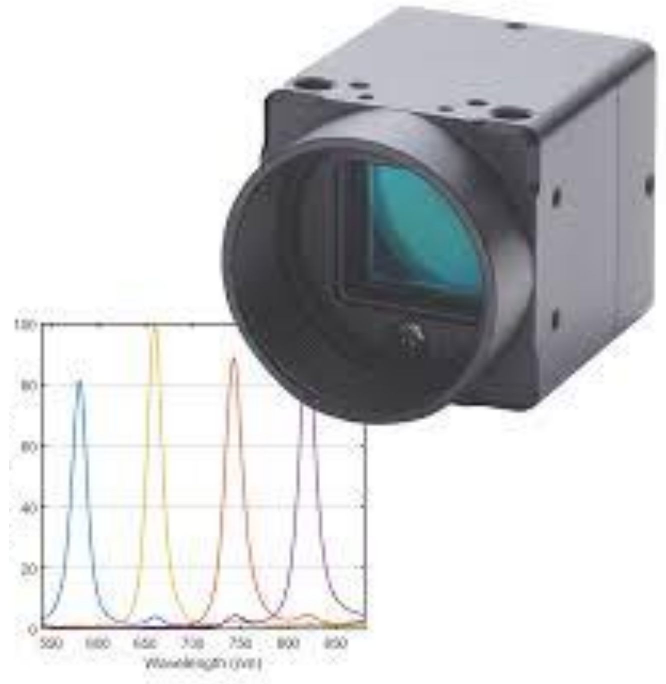
ryan - learning with visuotactile data

cite the image/vid src

slide for questions

practice slides

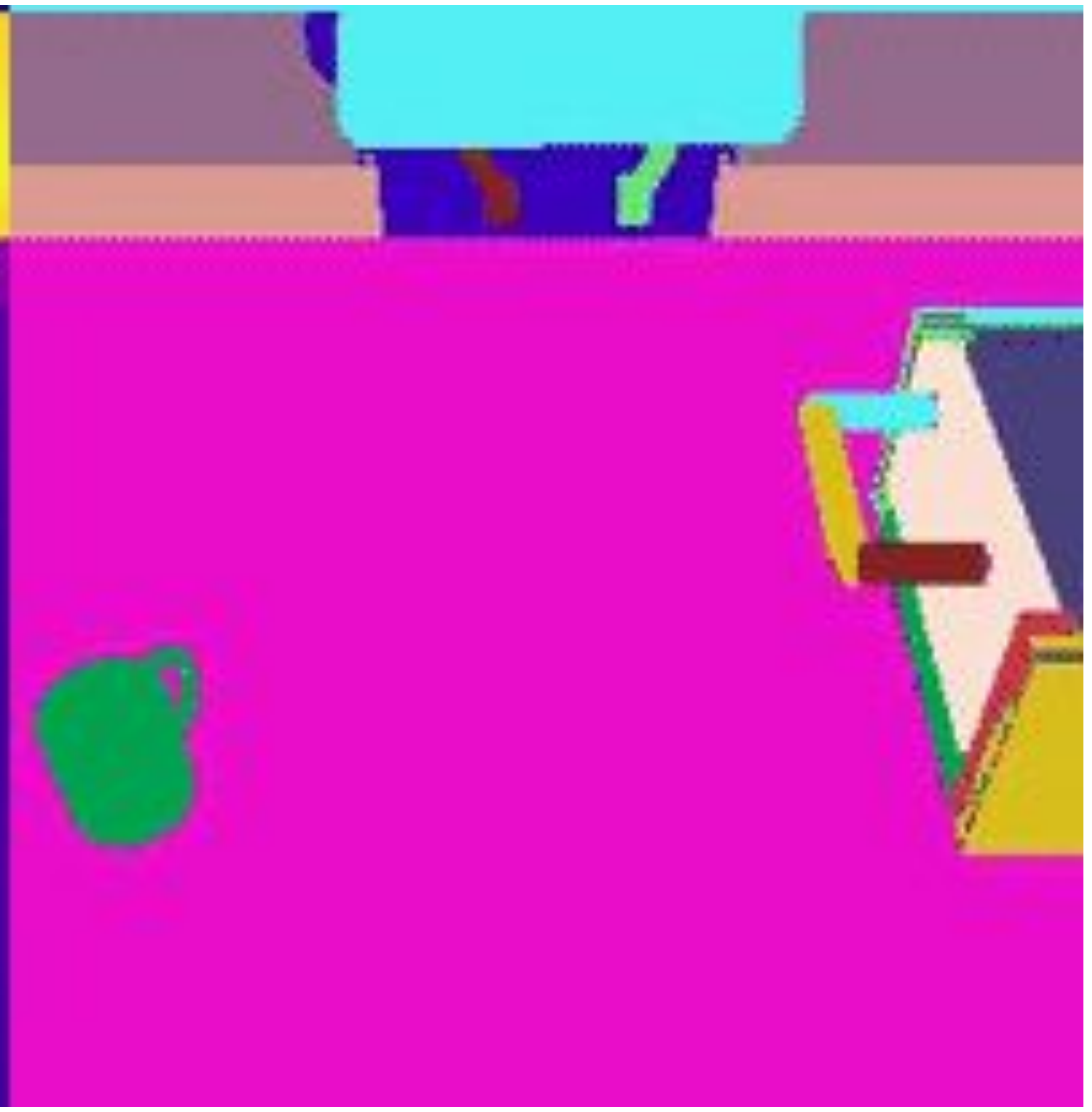
- intro to multimodal sensors
  - Different sensors (camera, depth, audio, contact, etc, force and torque, temporal)
    - Structures (Dimensions and examples) [one each]
    - Maps
      - Image, Depth, Segmentation, visual tactile (GelSight / bubble grippers!!!!) [Mason]
      - Multi Spectral images (all frequencies of light <- great for farming!) [Mason]
    - Clouds
      - Point cloud (Lidars, multi camera), gaussian splate [Hanchen]
    - Time series:
      - Force Torque, audio [Ryan]
    - This is not the only set of sensors…
- Intro to Multi Modal Learning wrt to Foundation models
  - Multi**modal** (vision+language) vs multi**sensory** (raw sensor data)
  - Inrto pretraining (foundation models) (quick) [mason]
    - individual [one each]
      - CNNs, ViTs for maps
      - Point++ for point clods
      - Time series models????? -> ask ryan (FFT, MLP
      - Individual encoders (images (ResNet, ViT), force-torque (MLP, CausalConv, etc.), audio)
  - pretraining with multimodal data (add timeline stuff here)
    - ~~LLM vs VLM stuff [Hanchen]~~
    - Multimae [hanchen]
    - MSVT, VTT, SVFL, AugInsert [Ryan] <- Using visuotactile data, [with audio] Hearing Touch, See Hear Feel



How do we sense and perceive the world?

How can **robots** sense and perceive the world?

2D Maps

RGB Color	Depth	Segmentation
		
3xWxH, uint8	1xWxH, float	1xWxH, uint8

Credit: Robomimic and Mason Hawver

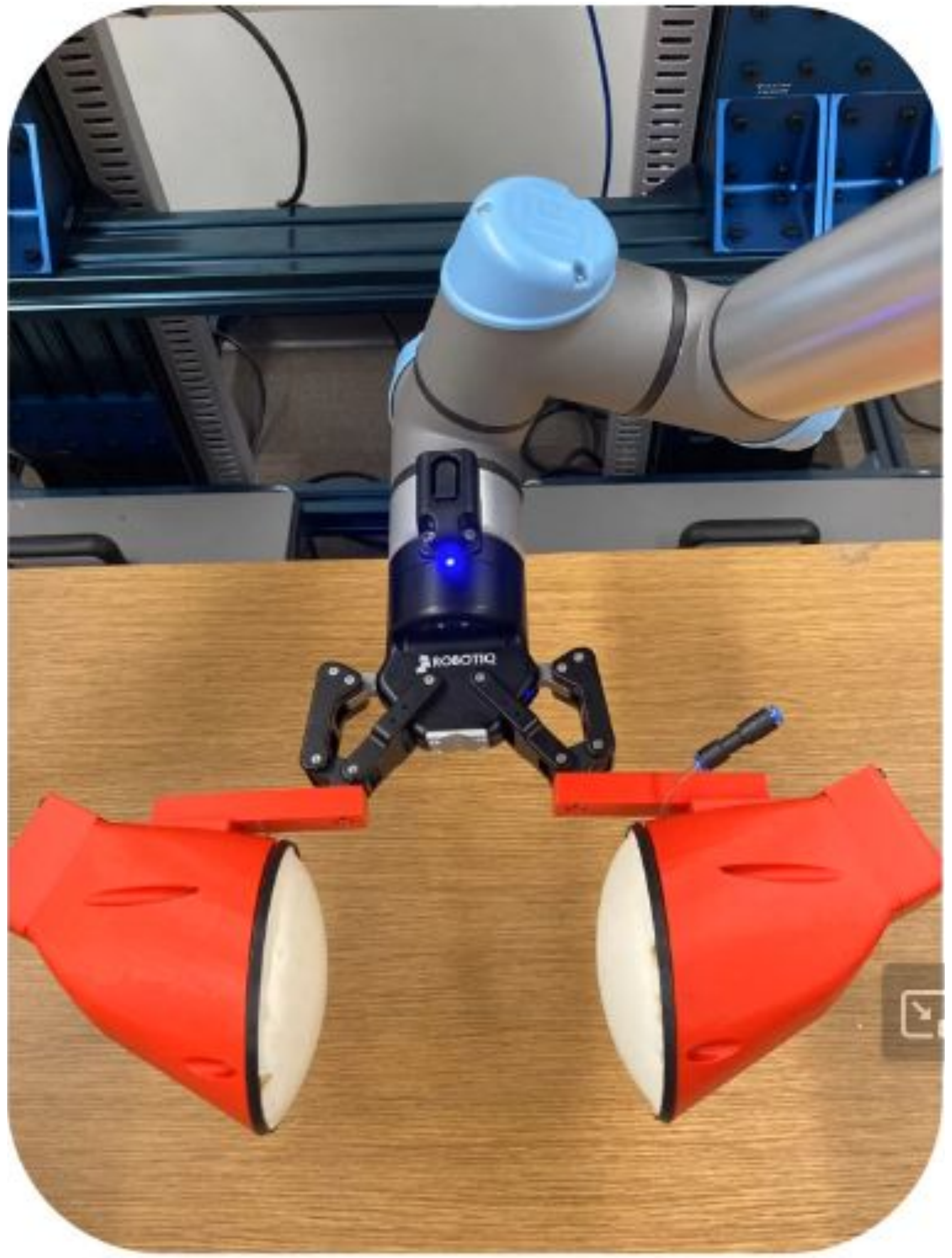
2D Maps - Visuotactile

Visuotactile: Represent Tactile Information with Vision

Credit: Robomimic and Mason Hawver

2D Maps - Visuotactile

Visuotactile: Represent Tactile Information with Vision

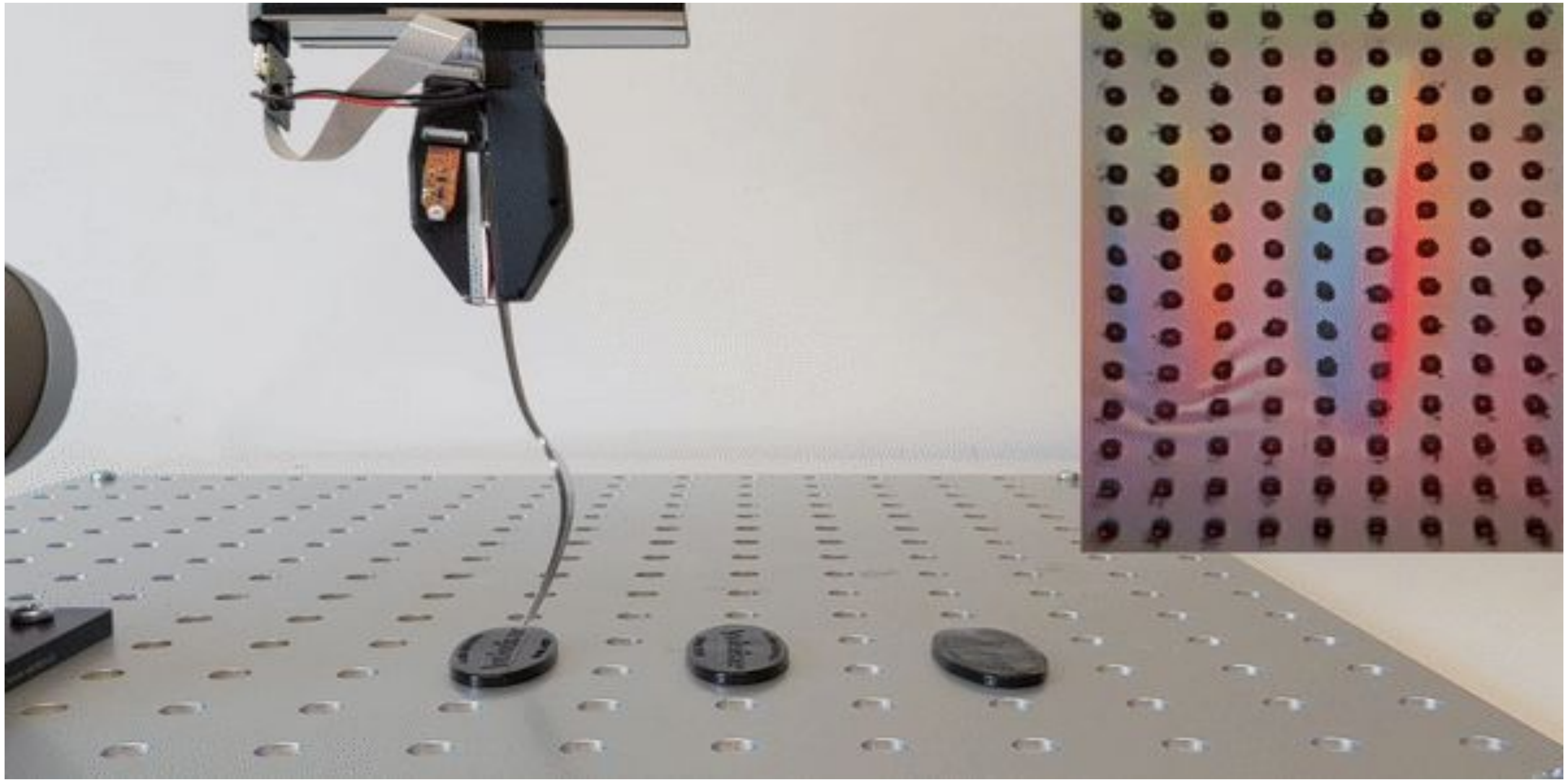


Two Grayscale Images
2x1xWxH, unit8

Credit: RPM Miles and Aaron, and TRI

2D Maps - Visuotactile

Visuotactile: Represent Tactile Information with Vision

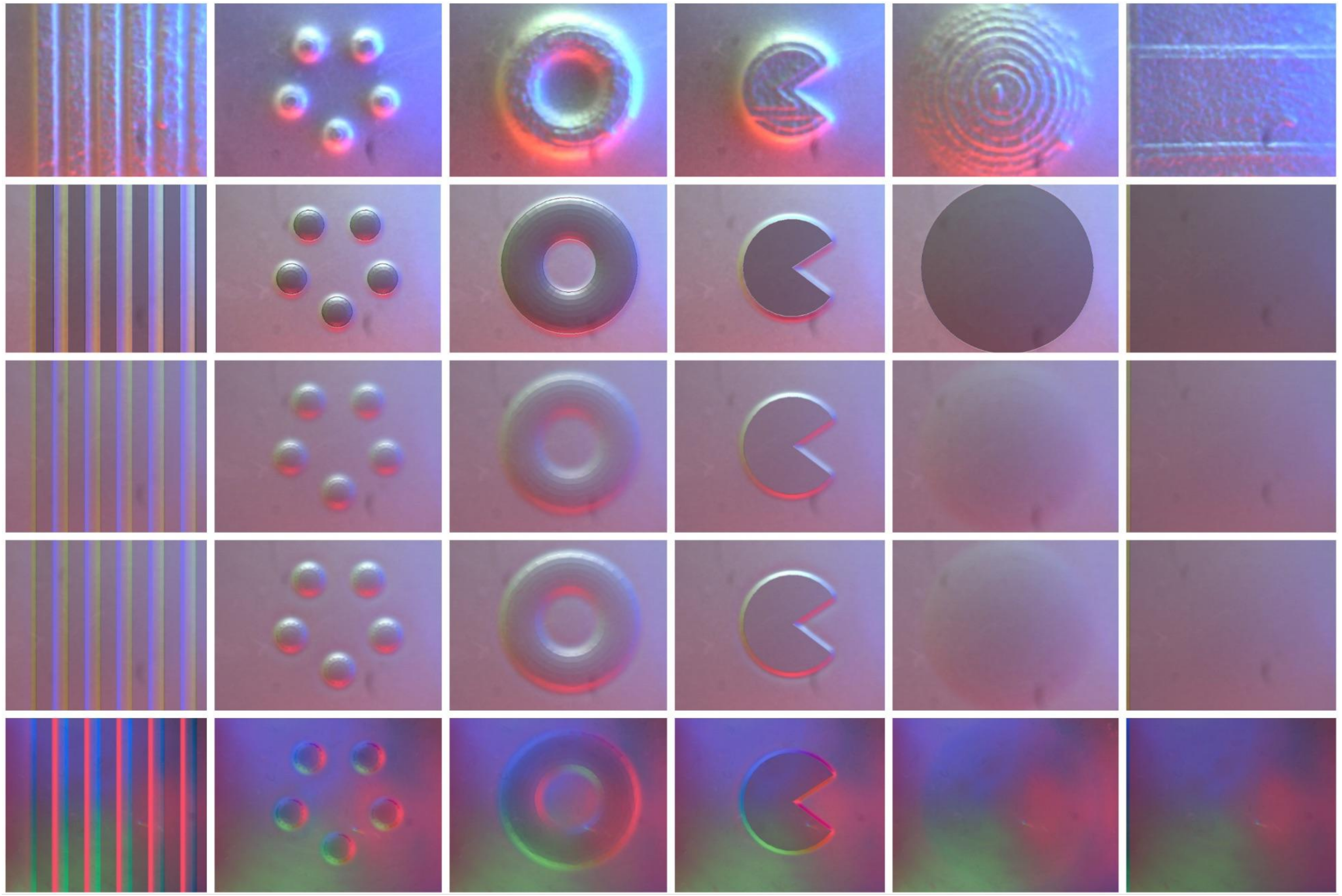


Vector Field, Map of Vectors:
2xWxH, floats

Credit: Rui Li and GelSight

2D Maps - Visuotactile

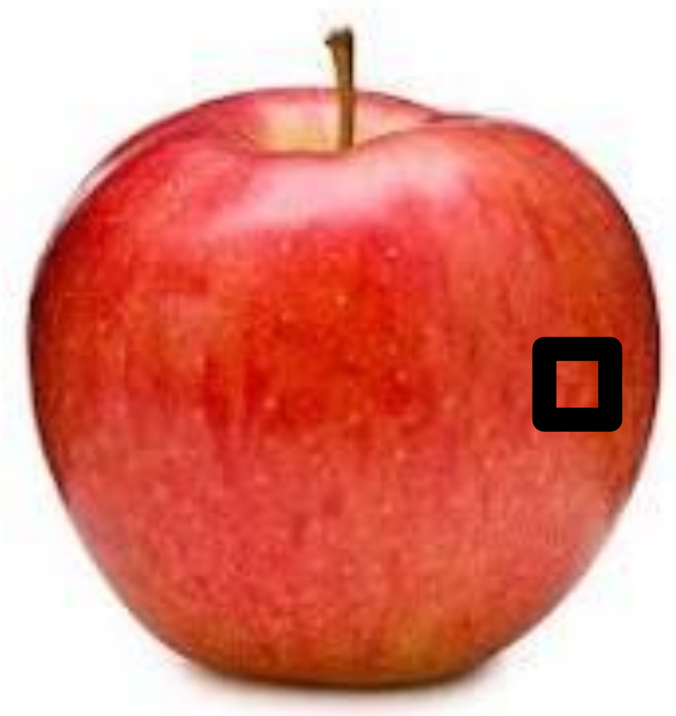
Visuotactile: Represent Tactile Information with Vision

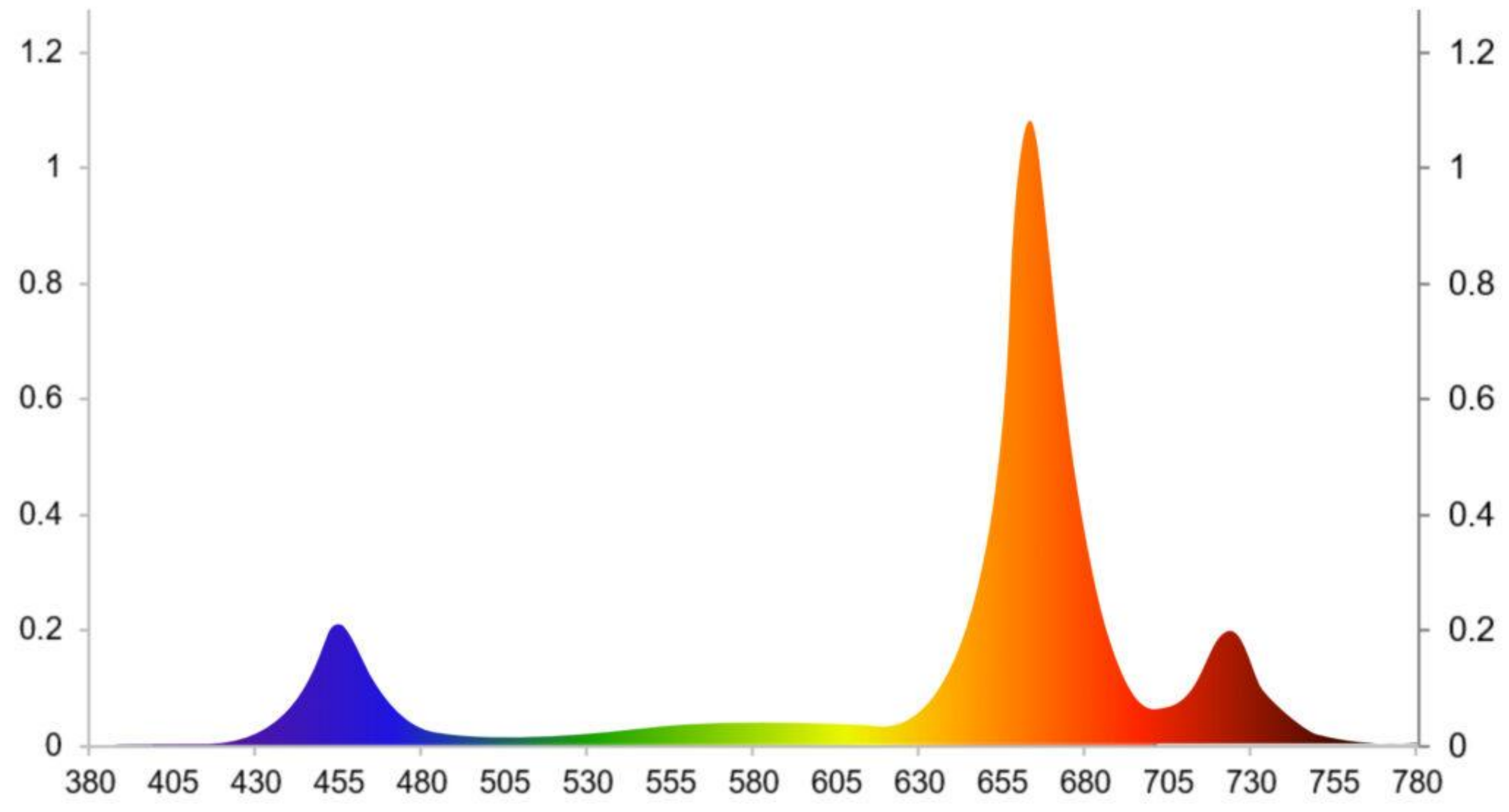


Gel Sight, RGB Image:
3xWxH, unit8

D. F. Gomes, P. Paoletti, and S. Luo, “Generation of gelsight tactile images for sim2real learning,” IEEE Robotics and Automation Letters, vol. 6, no. 2, pp. 4177–4184, 2021.

2D Maps - MultiSpectral





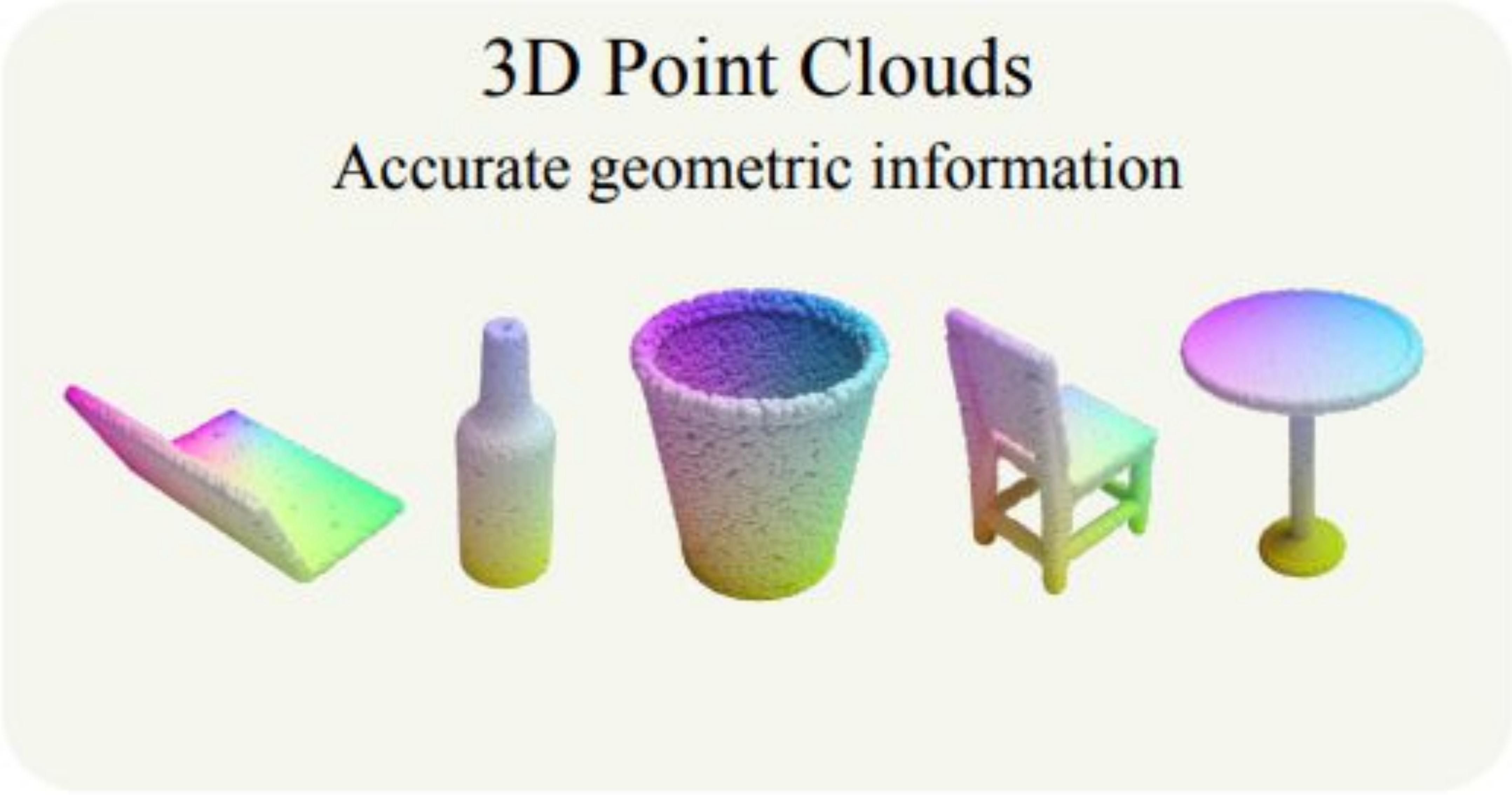
2D Map of Histograms
BinsxWxH, unit8

Credit: IdeaForge and Mathworks

2D Maps - MultiSpectral

Credit: IdeaForge and Mathworks

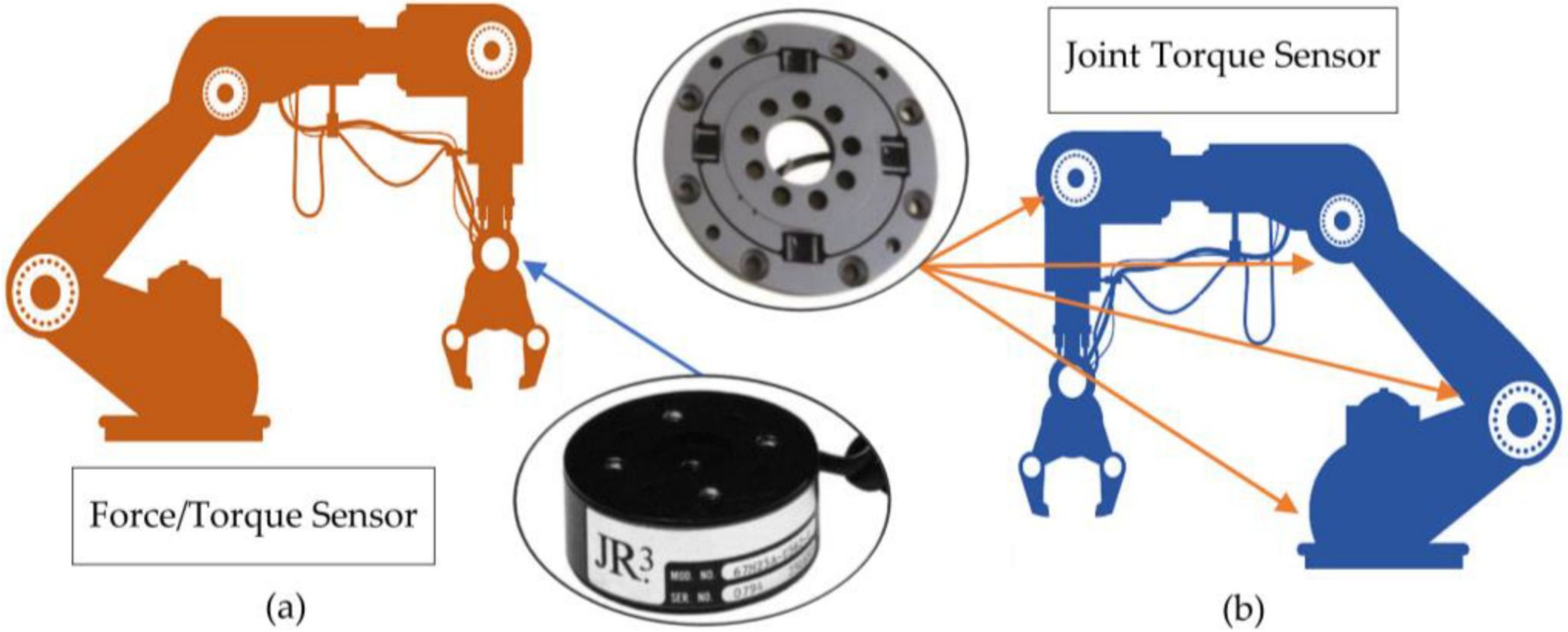
Perceive the 3D world



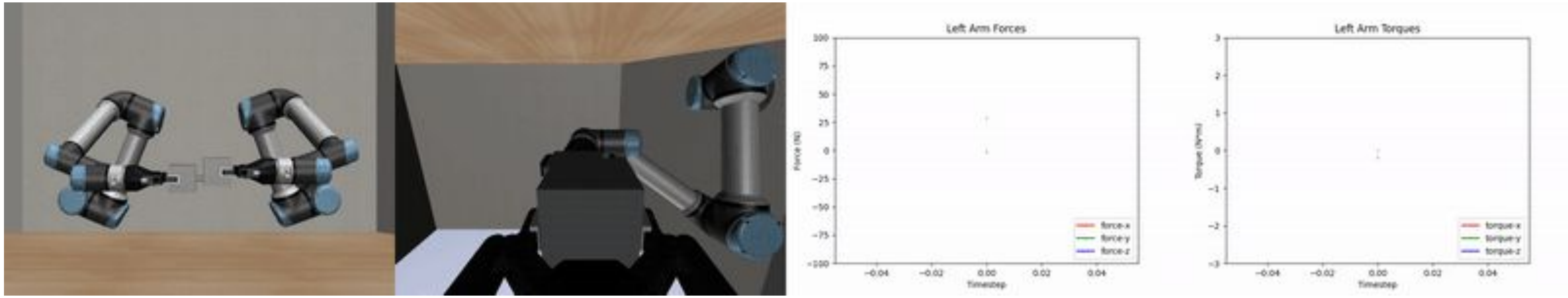
data structure: (N,X,Y,Z)
unordered: permutation invariant

Zhang, Tong, et al. "A universal semantic-geometric representation for robotic manipulation." *arXiv preprint arXiv:2306.10474* (2023).

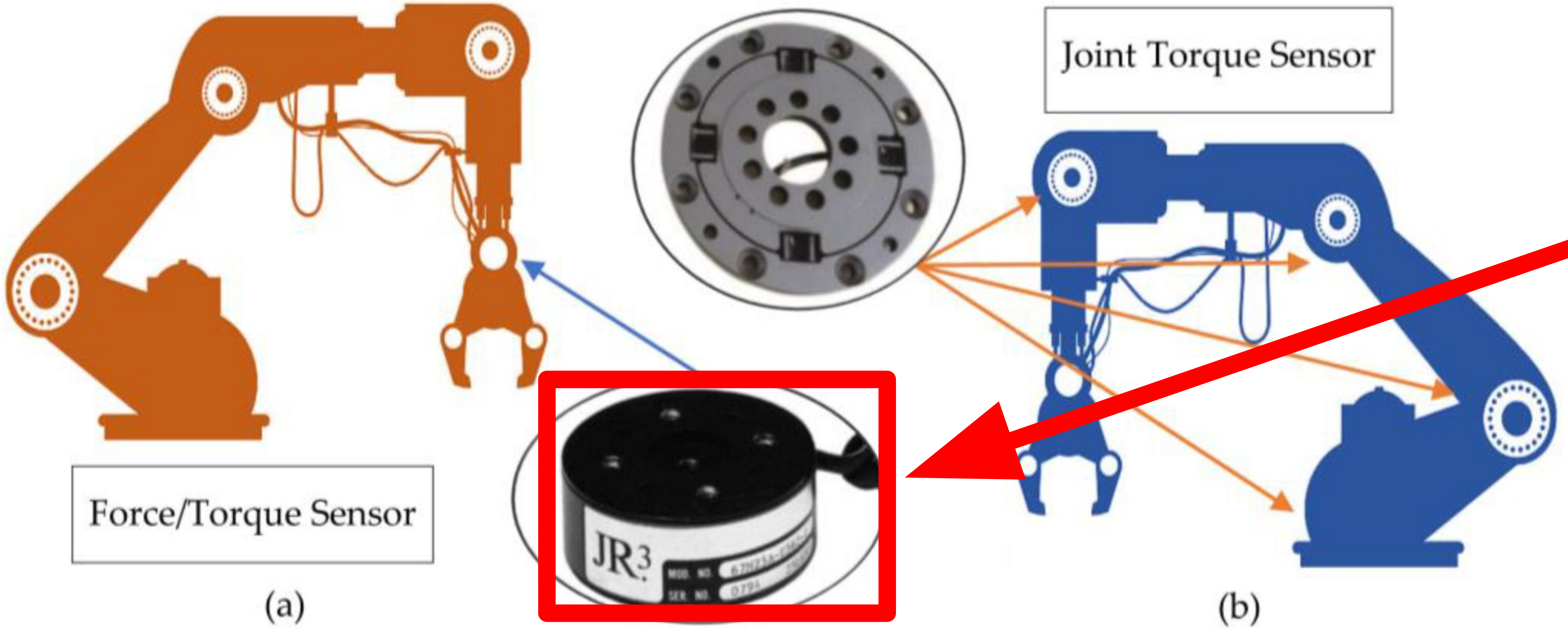
Force-Torque Sensing



R. P. Ubeda, S. C. Guti´errez Rubert, R. Zotovic Stanisic, and ´A. Perles Ivars, “Design and manufacturing of an ultra-low-cost custom torque sensor for robotics,” Sensors, vol. 18, no. 6, 2018.



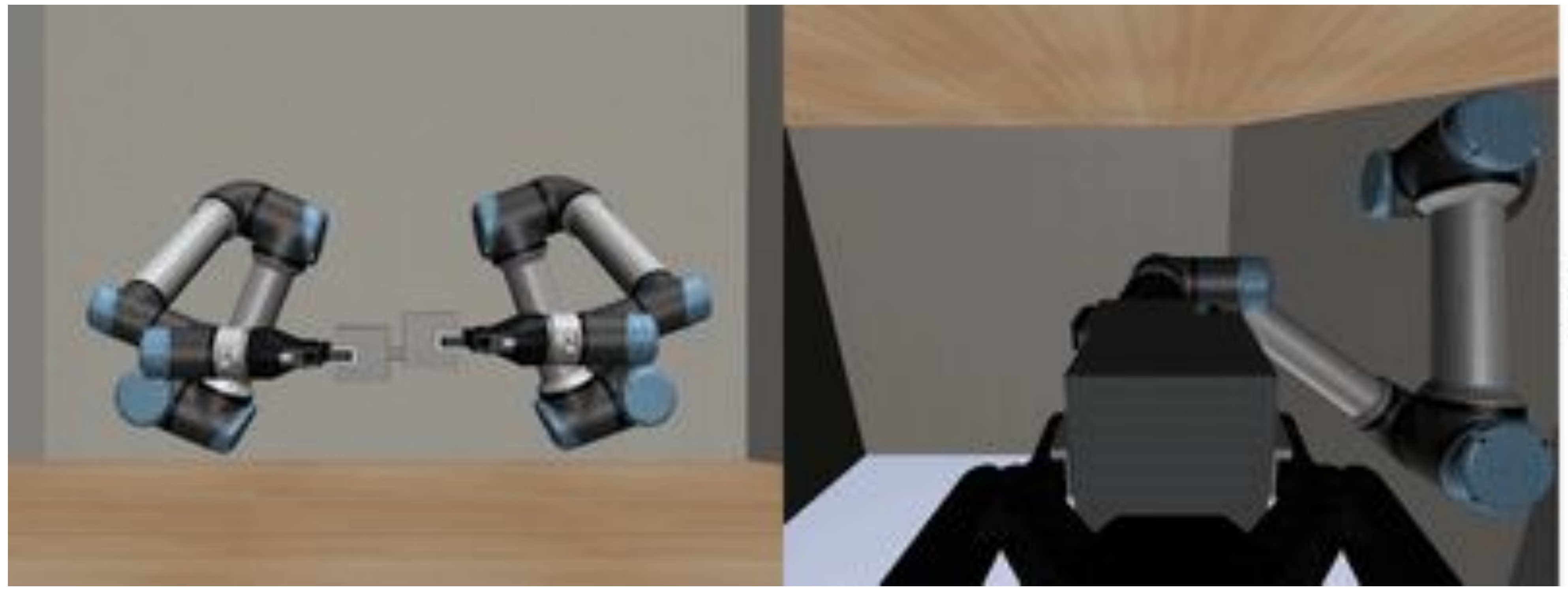
Force-Torque Sensing

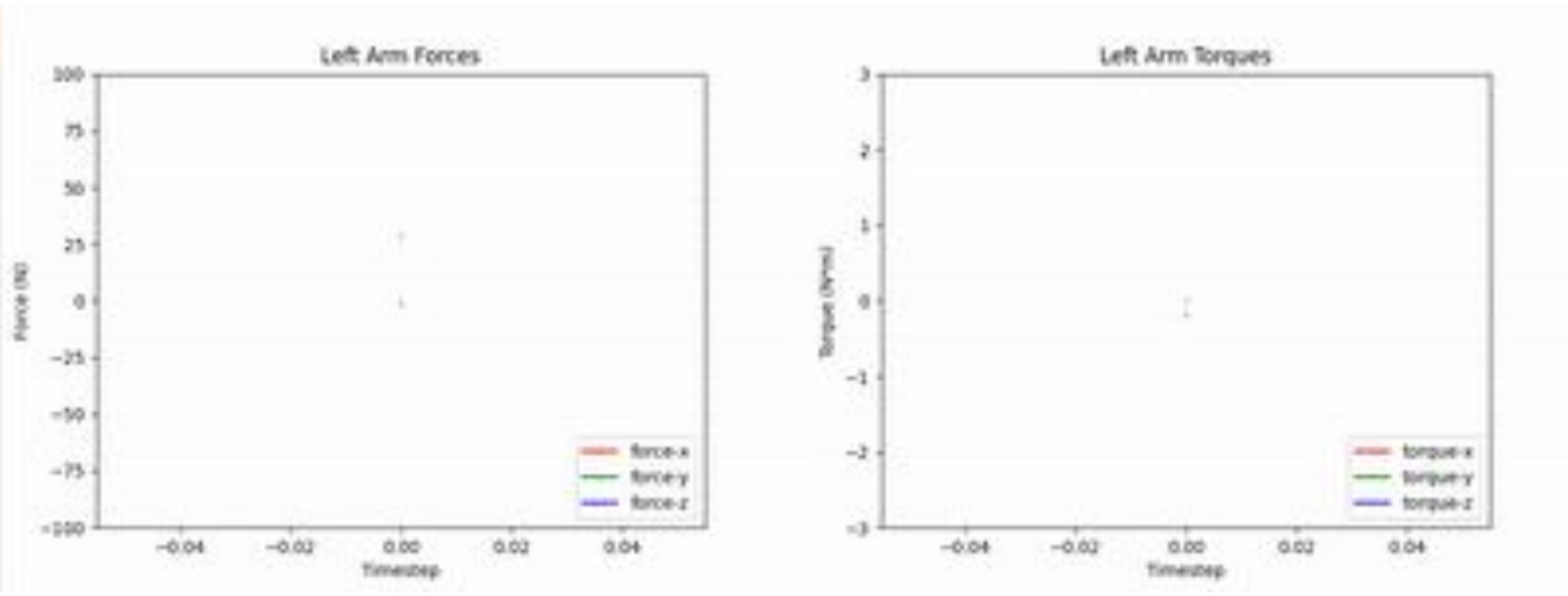


6-axis force-torque sensor along X,Y,Z axes relative to EEF frame

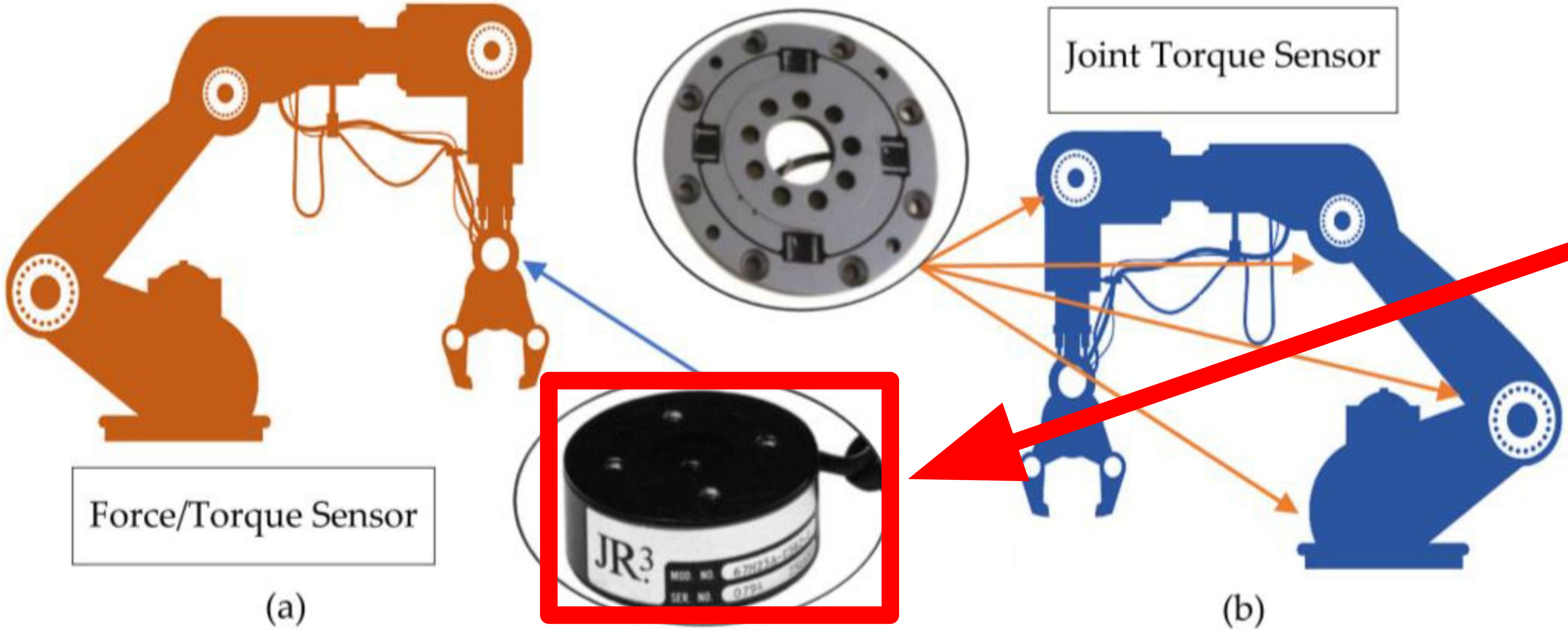
F/T measured using electrical signals

R. P. Ubeda, S. C. Guti´errez Rubert, R. Zotovic Stanisic, and ´A. Perles Ivars, “Design and manufacturing of an ultra-low-cost custom torque sensor for robotics,” Sensors, vol. 18, no. 6, 2018.





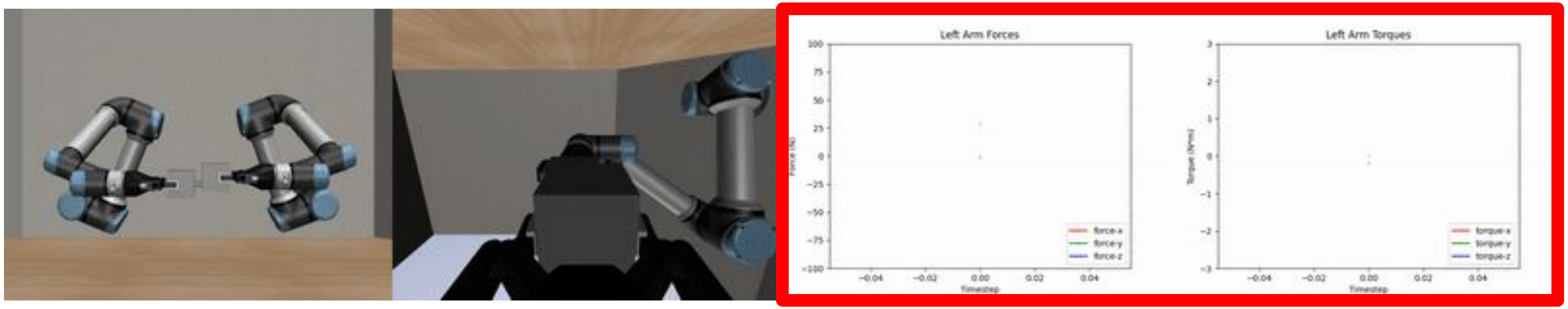
Force-Torque Sensing



R. P. Ubeda, S. C. Guti´errez Rubert, R. Zotovic Stanisic, and ´A. Perles Ivars, “Design and manufacturing of an ultra-low-cost custom torque sensor for robotics,” Sensors, vol. 18, no. 6, 2018.

6-axis force-torque sensor along X,Y,Z axes relative to EEF frame

F/T measured using electrical signals



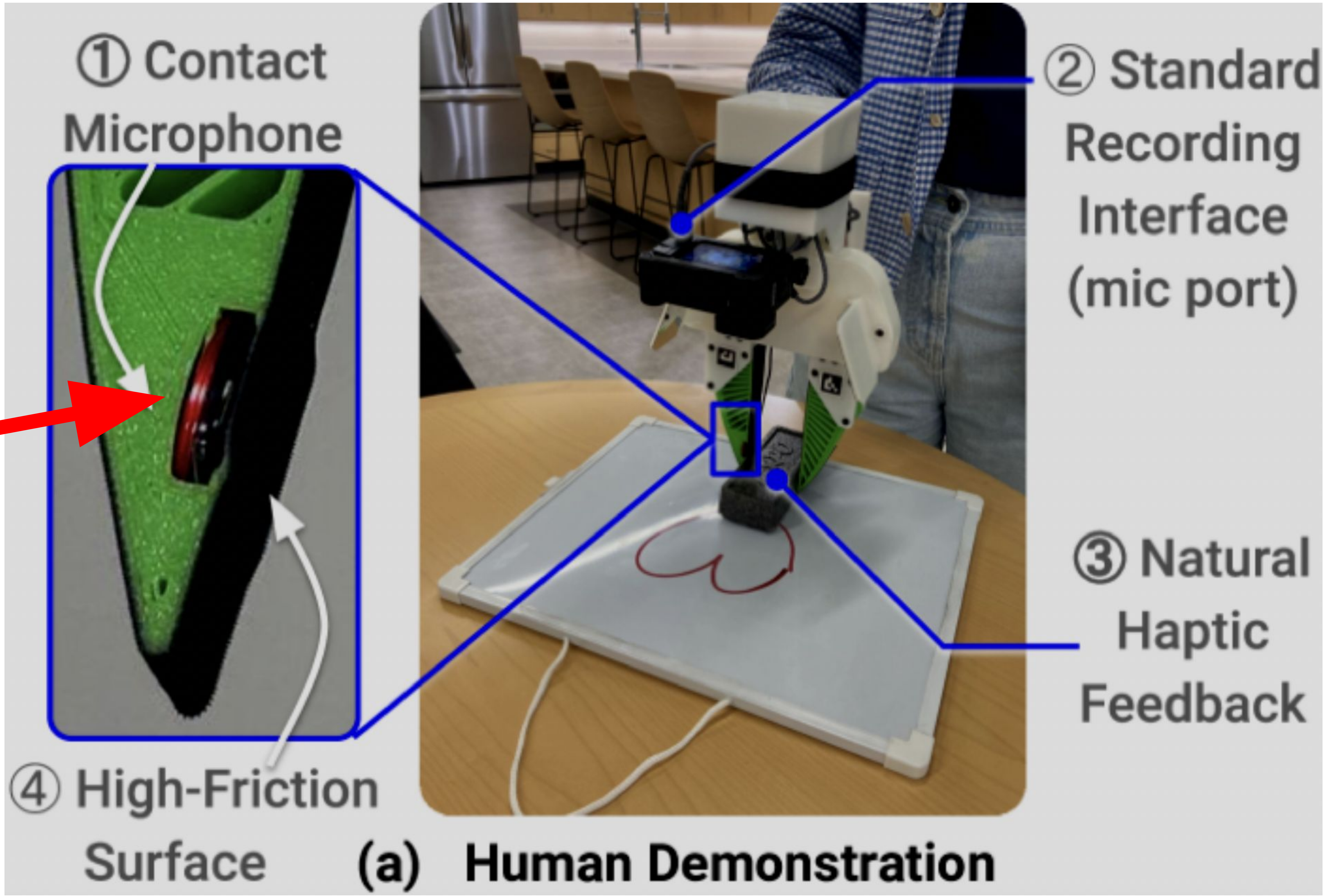
Hx6, float

Contact Audio Sensing

https://modularsynthlab.com/product/high-quality-piezo-contact-microphone-piezo-transducer-27mm-120cm-cable-mono-jack-3-5mm/?v=0b3b97fa6688

Contact Audio Sensing

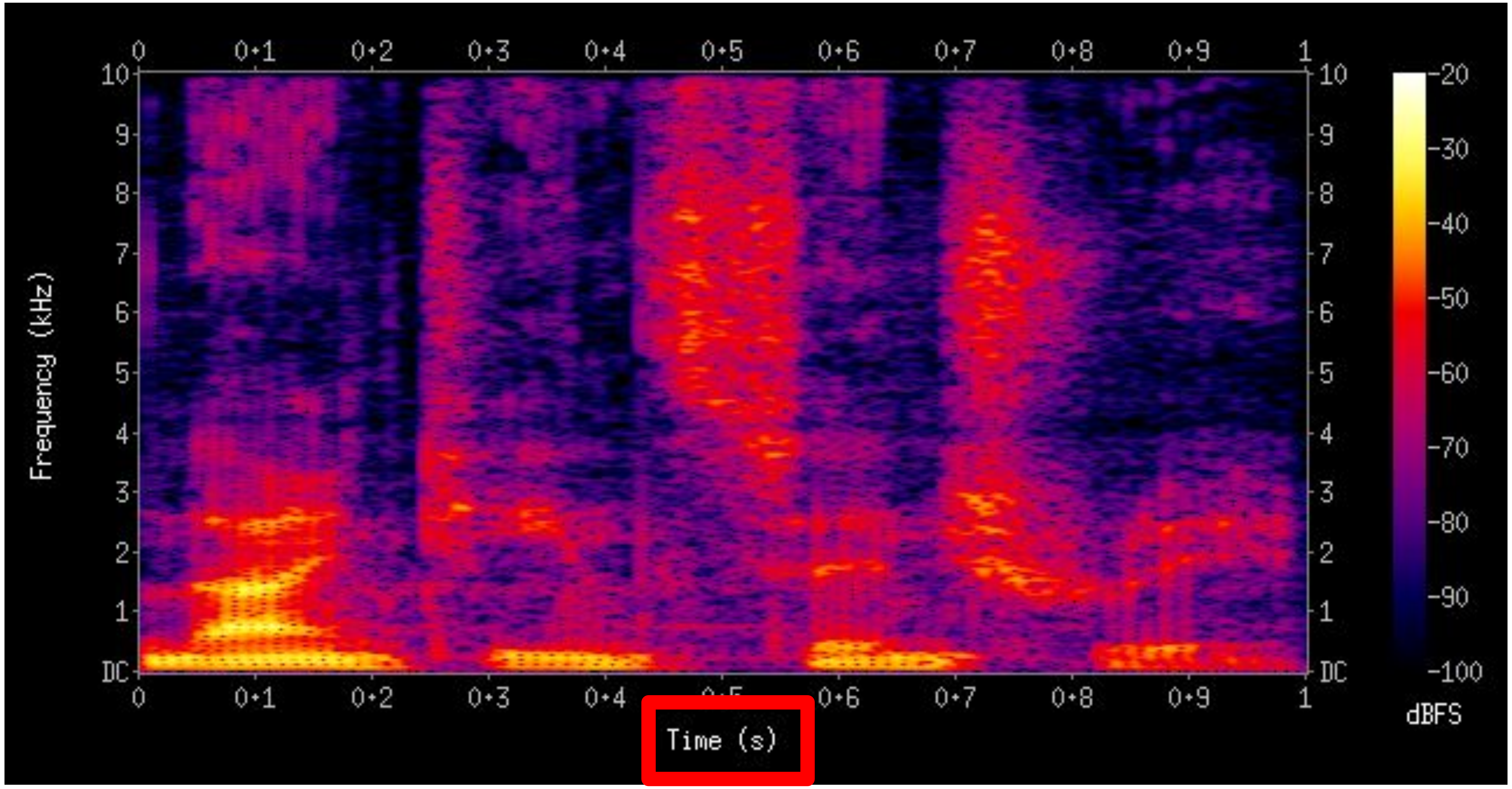
https://modularsynthlab.com/product/high-quality-piezo-contact-microphone-piezo-transducer-27mm-120cm-cable-mono-jack-3-5mm/?v=0b3b97fa6688



Z. Liu, C. Chi, E. Cousineau, N. Kuppuswamy, B. Burchfiel, and S. Song, “Maniwav: Learning robot manipulation from in-the-wild audio-visual data,” arXiv preprint arXiv:2406.19464, 2024.

Contact Audio Sensing

Spectrograms: Transferring audio into the vision domain

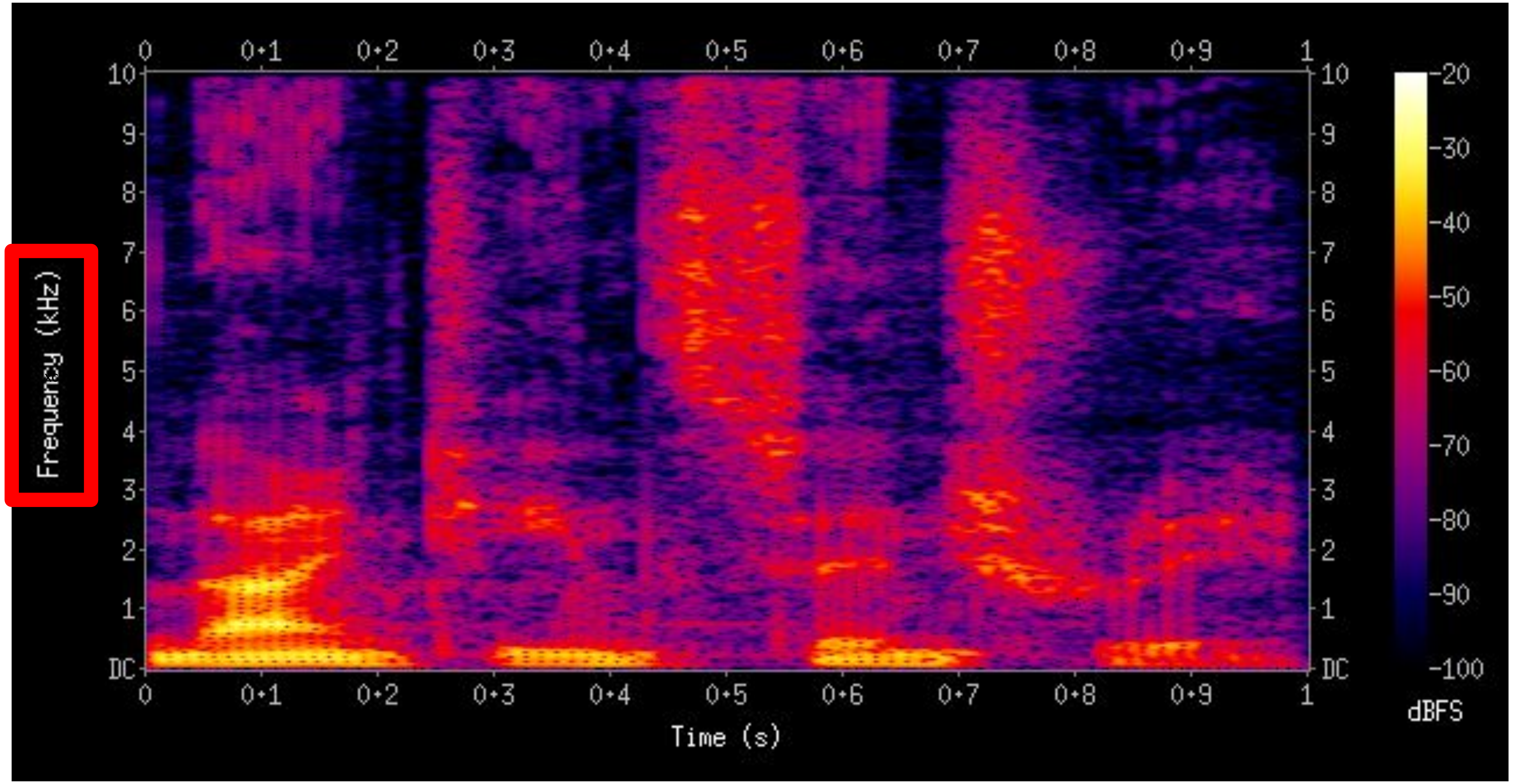


1. Record an audio sample over time (i.e. last few seconds from contact microphone)

Credit: By Aquegg - Own work, Public Domain, https://commons.wikimedia.org/w/index.php?curid=5544473

Contact Audio Sensing

Spectrograms: Transferring audio into the vision domain

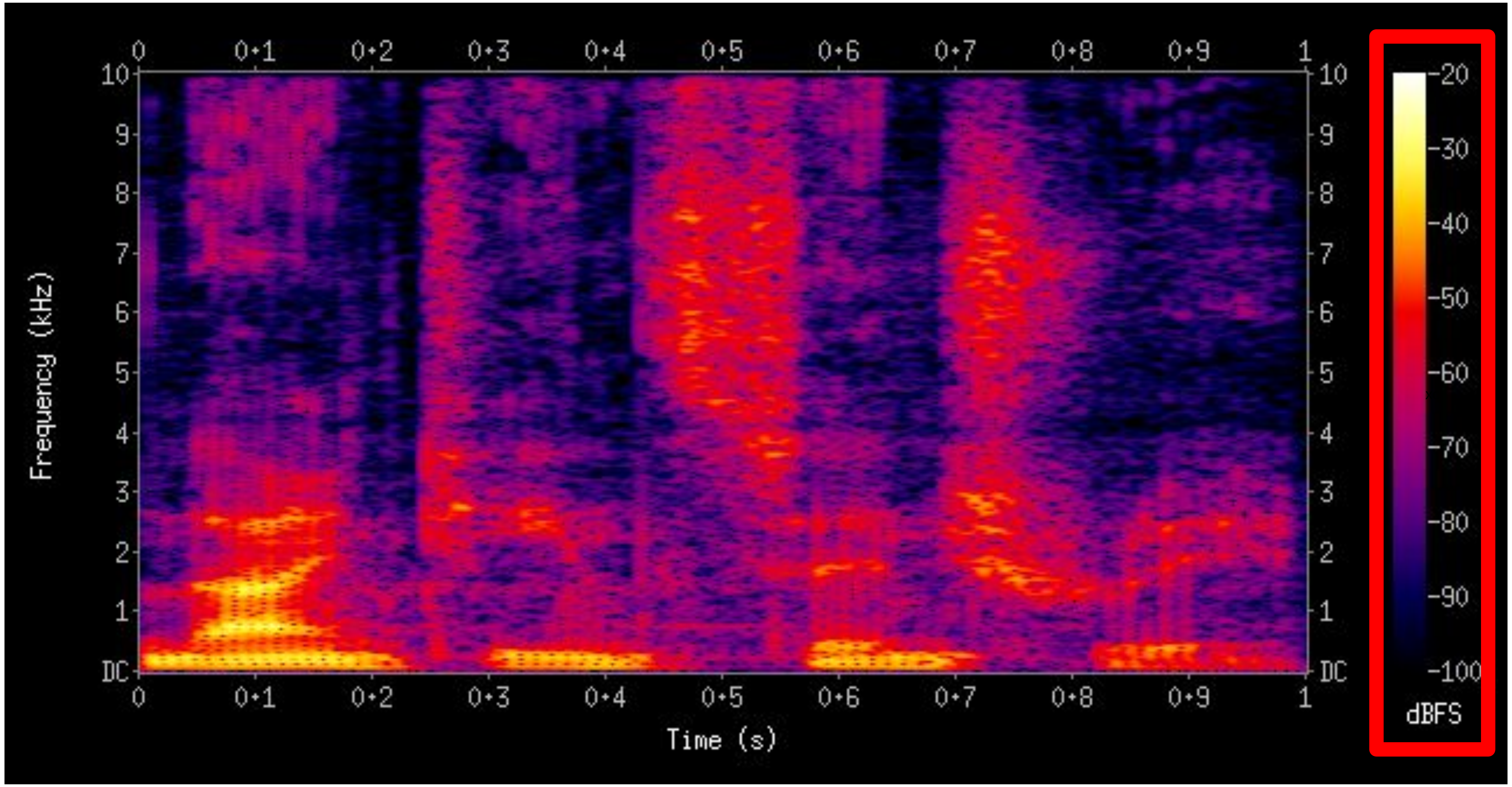


2. Extract frequencies for each time step

Credit: By Aquegg - Own work, Public Domain, https://commons.wikimedia.org/w/index.php?curid=5544473

Contact Audio Sensing

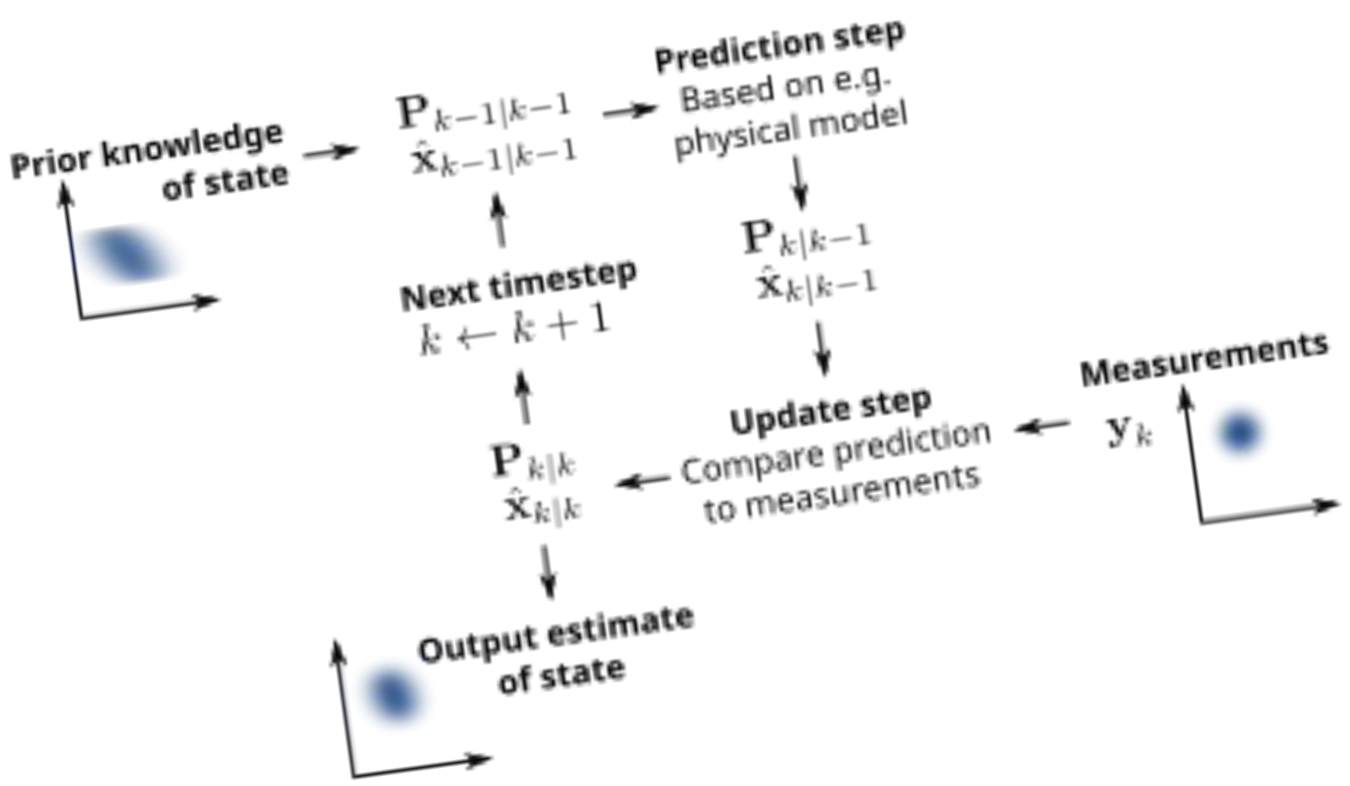
Spectrograms: Transferring audio into the vision domain

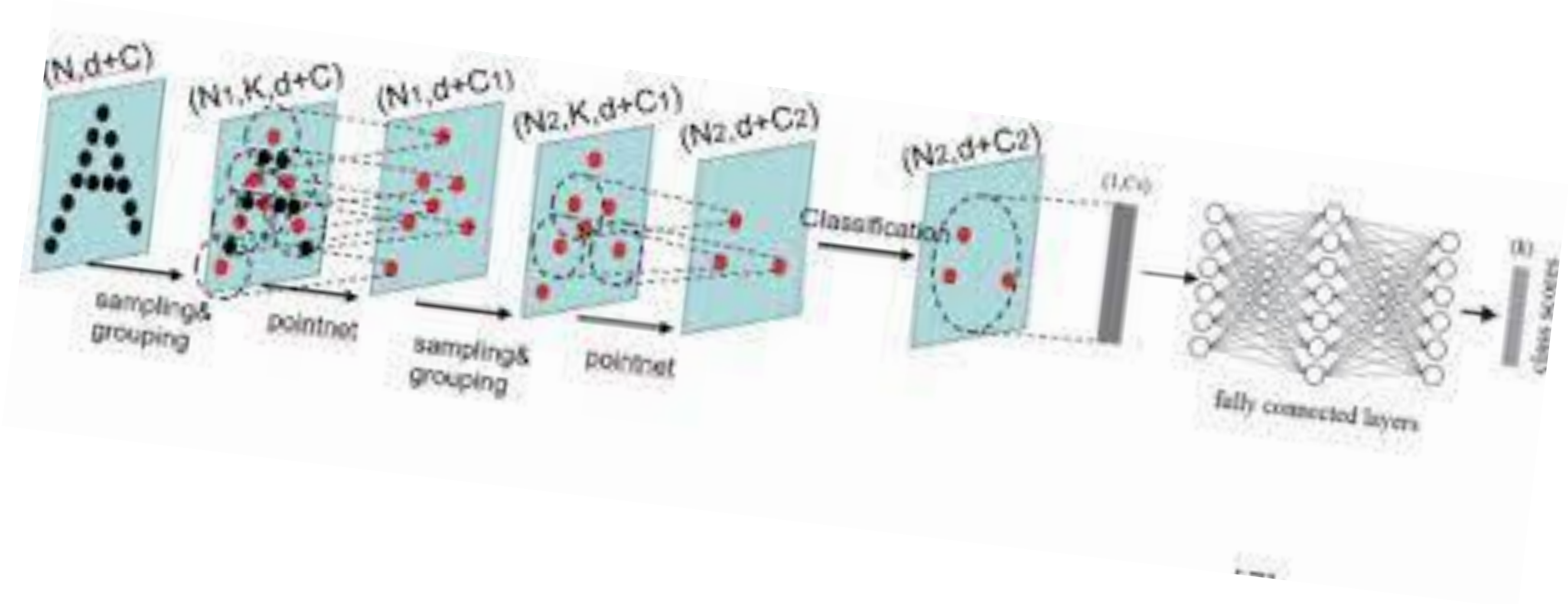


3. Plot amplitudes of each frequency for each timestep

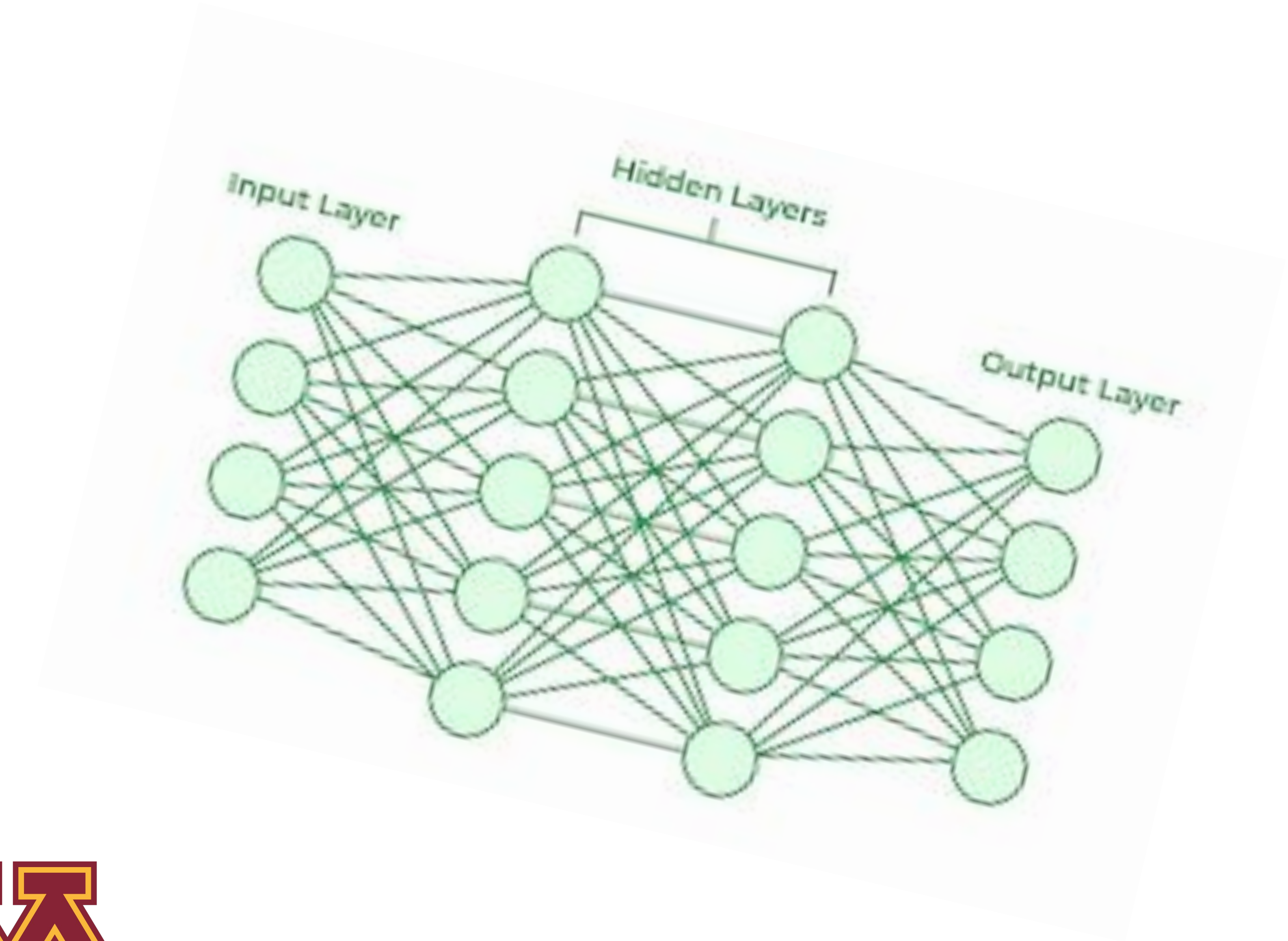
Credit: By Aquegg - Own work, Public Domain, https://commons.wikimedia.org/w/index.php?curid=5544473

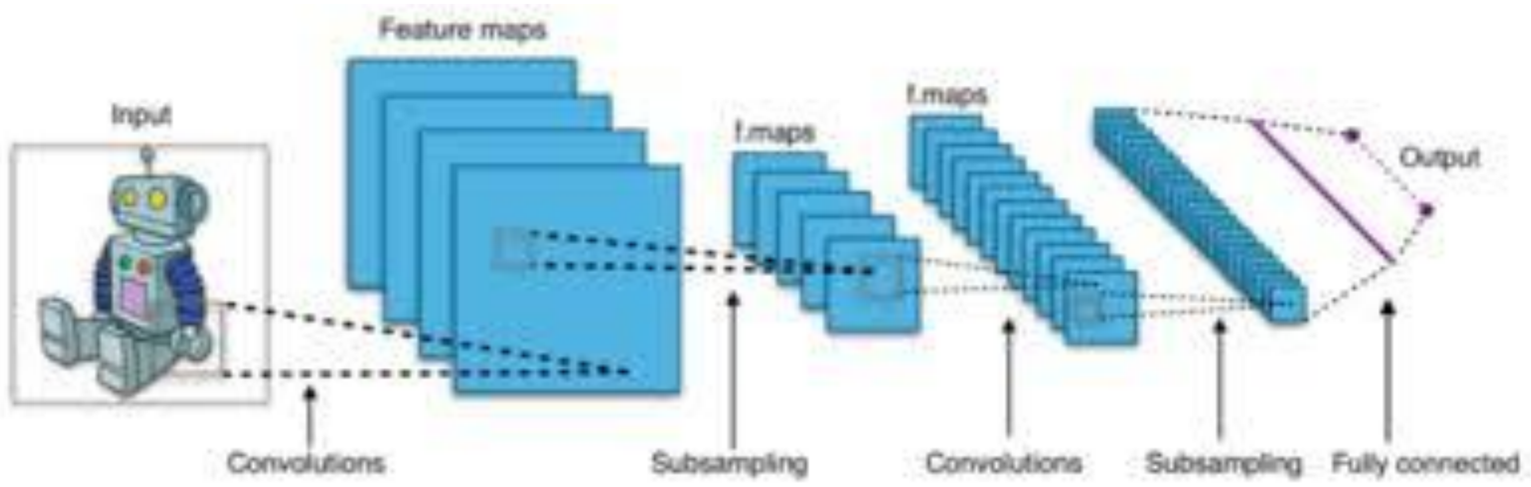
These are not all the types of sensors…

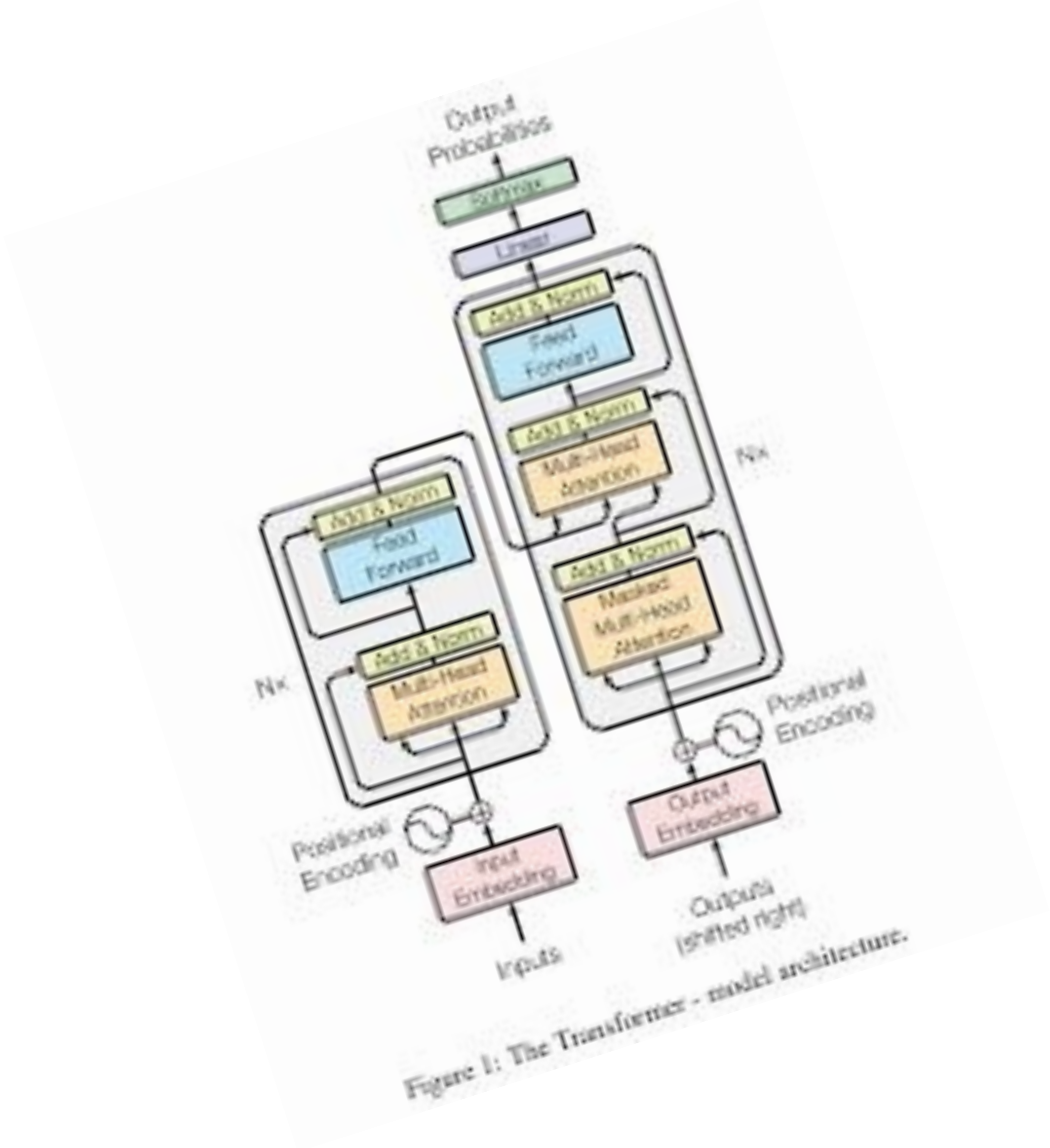




How can we **encode** each sensor data type?

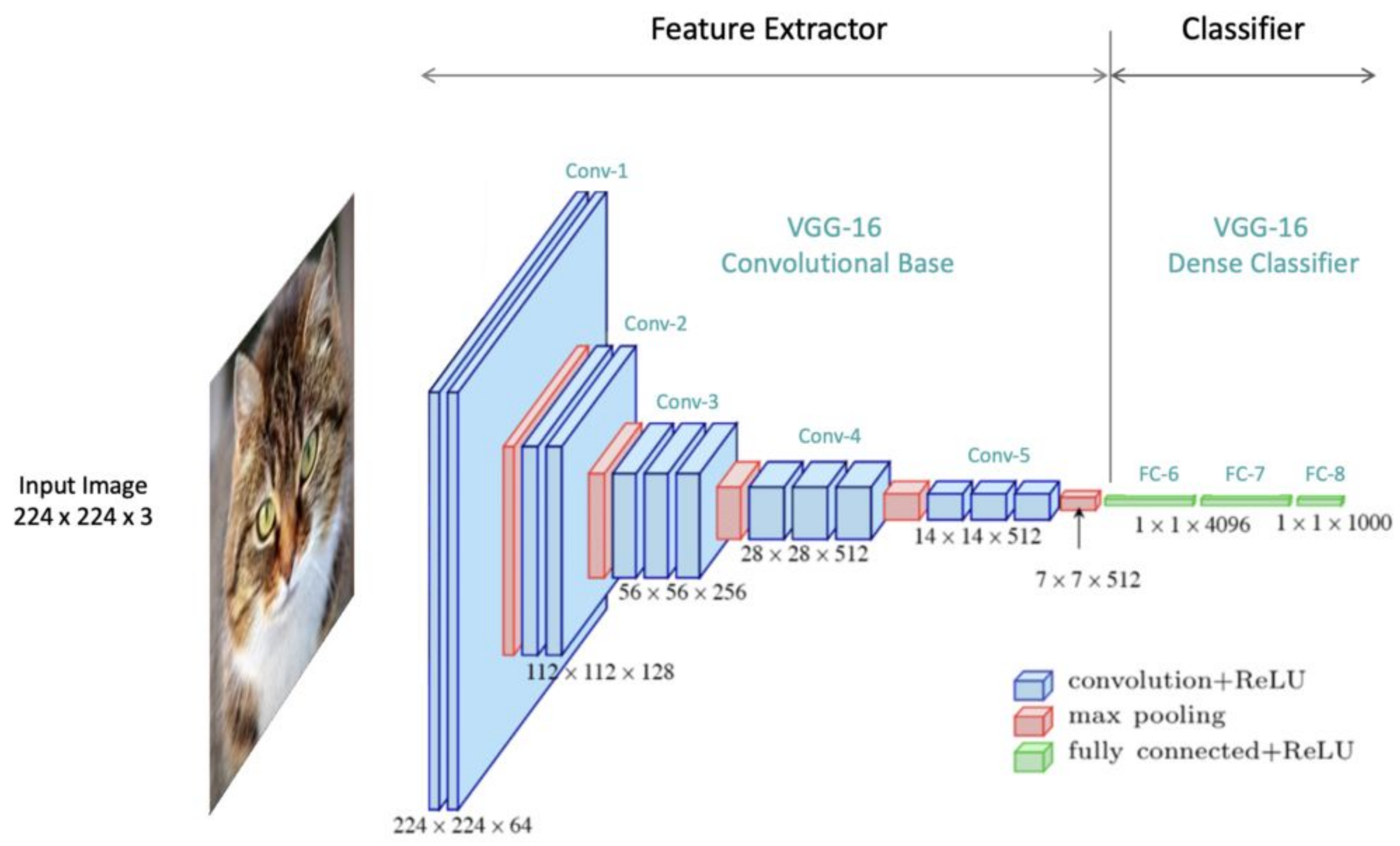




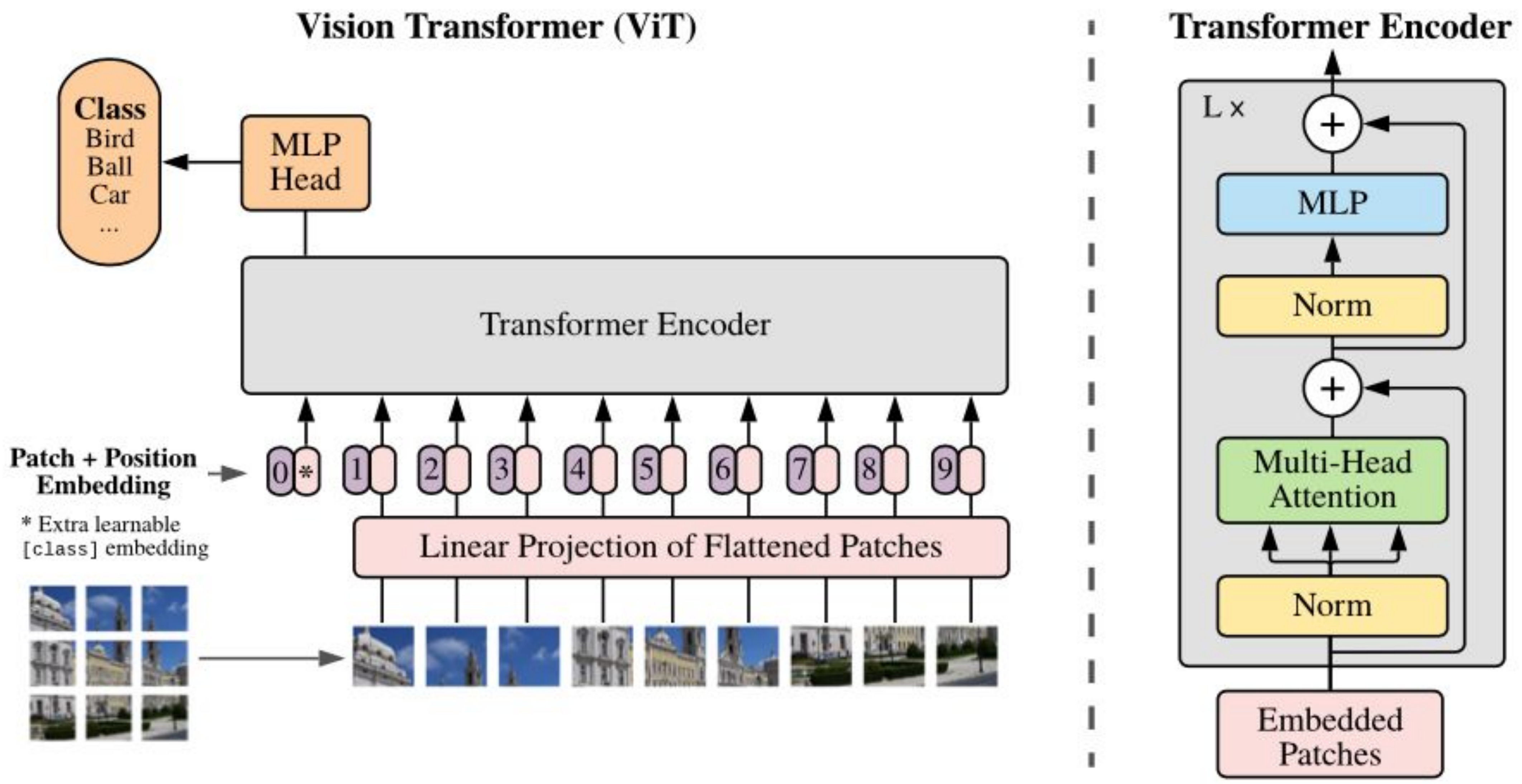


Credit: Attention Is All You Need and Wikipedia and PointNet: Deep Learning on Point Sets for 3D Classification and Segmentation

Classifying 2D Maps



CNNs

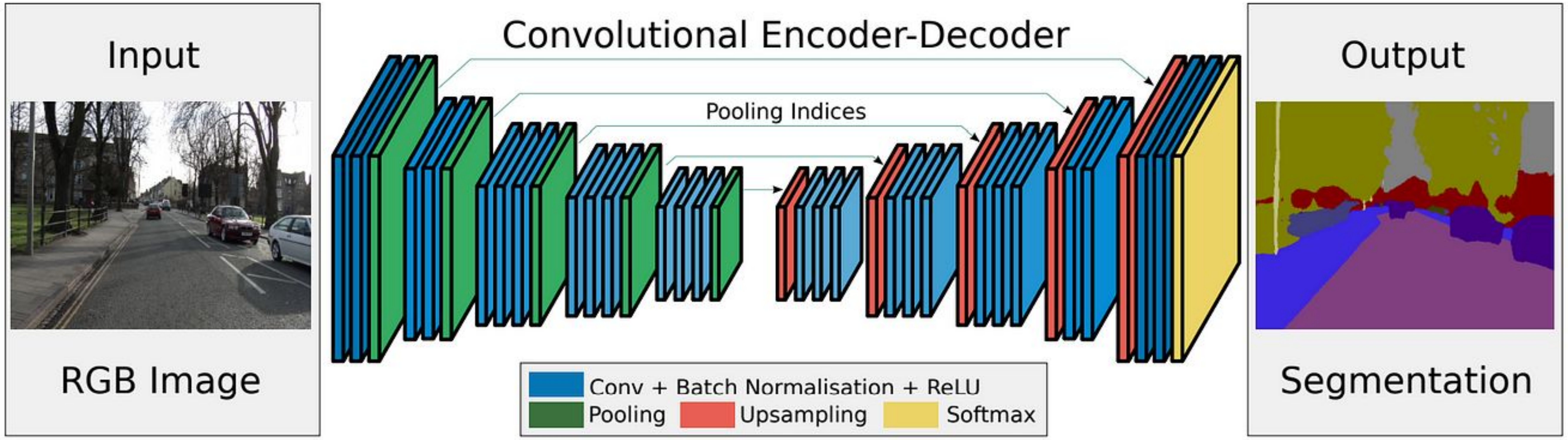


ViTs

Credit: Learn OpenCV and Unite.AI

Encoding 2D Maps

Task Encoding

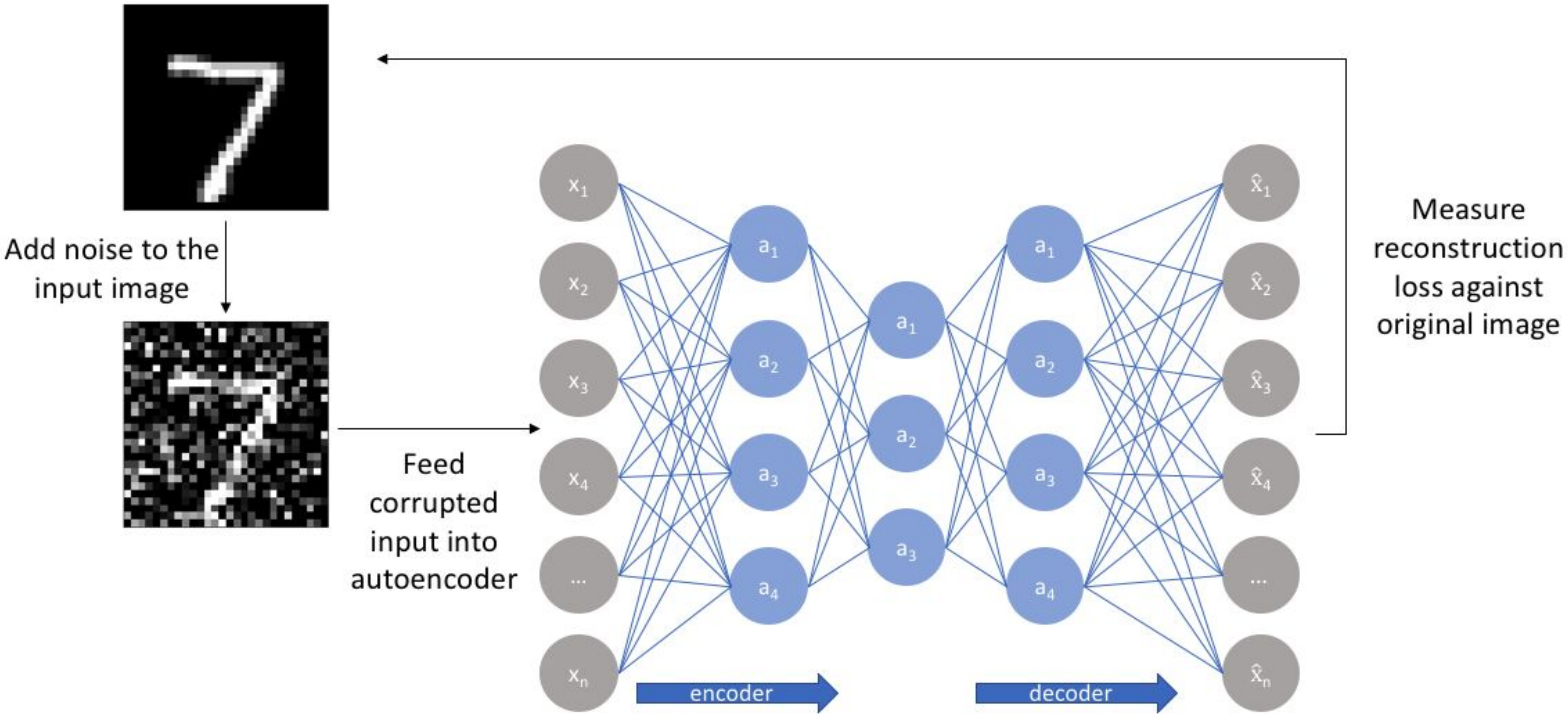


Latent Vector

Credit: Jeremy Jordan

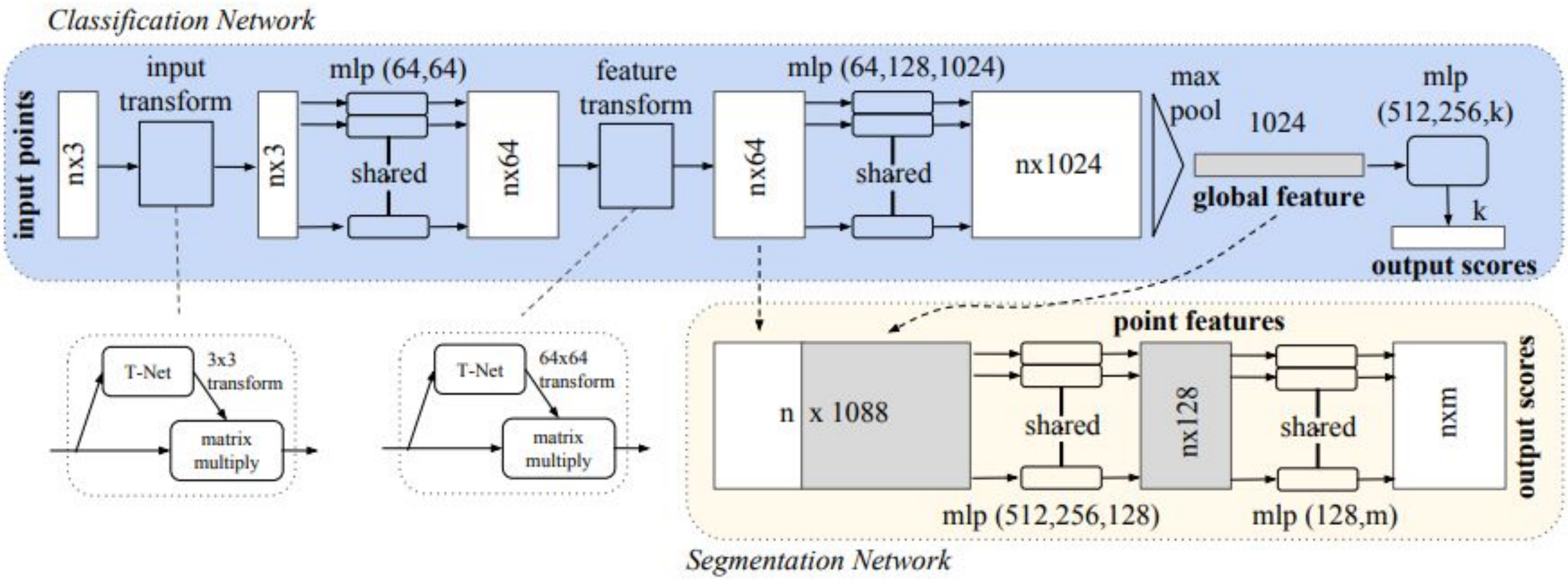
Encoding 2D Maps

Autoregressive Encoding



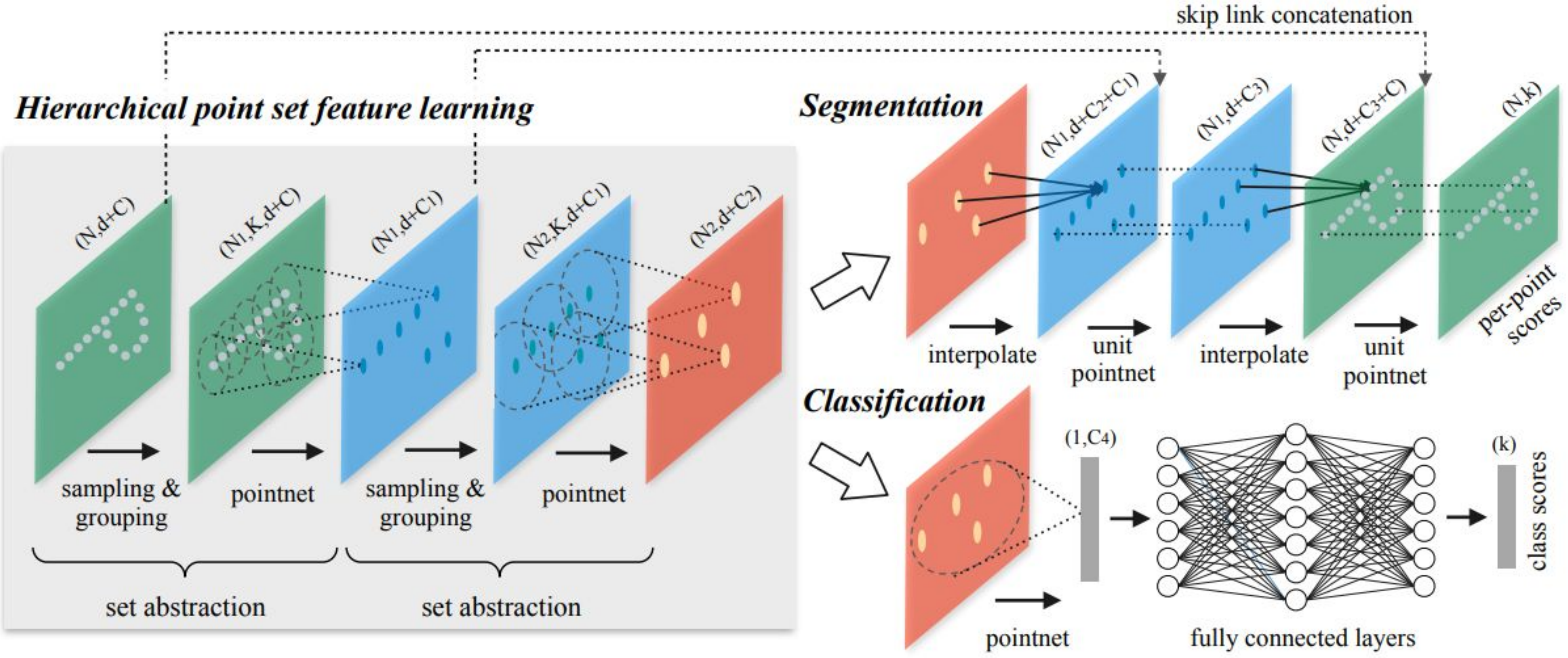
Credit: Jeremy Jordan

How to process point cloud data—pointnet



Qi, Charles R., et al. "Pointnet: Deep learning on point sets for 3d classification and segmentation." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017.

Pointnet++

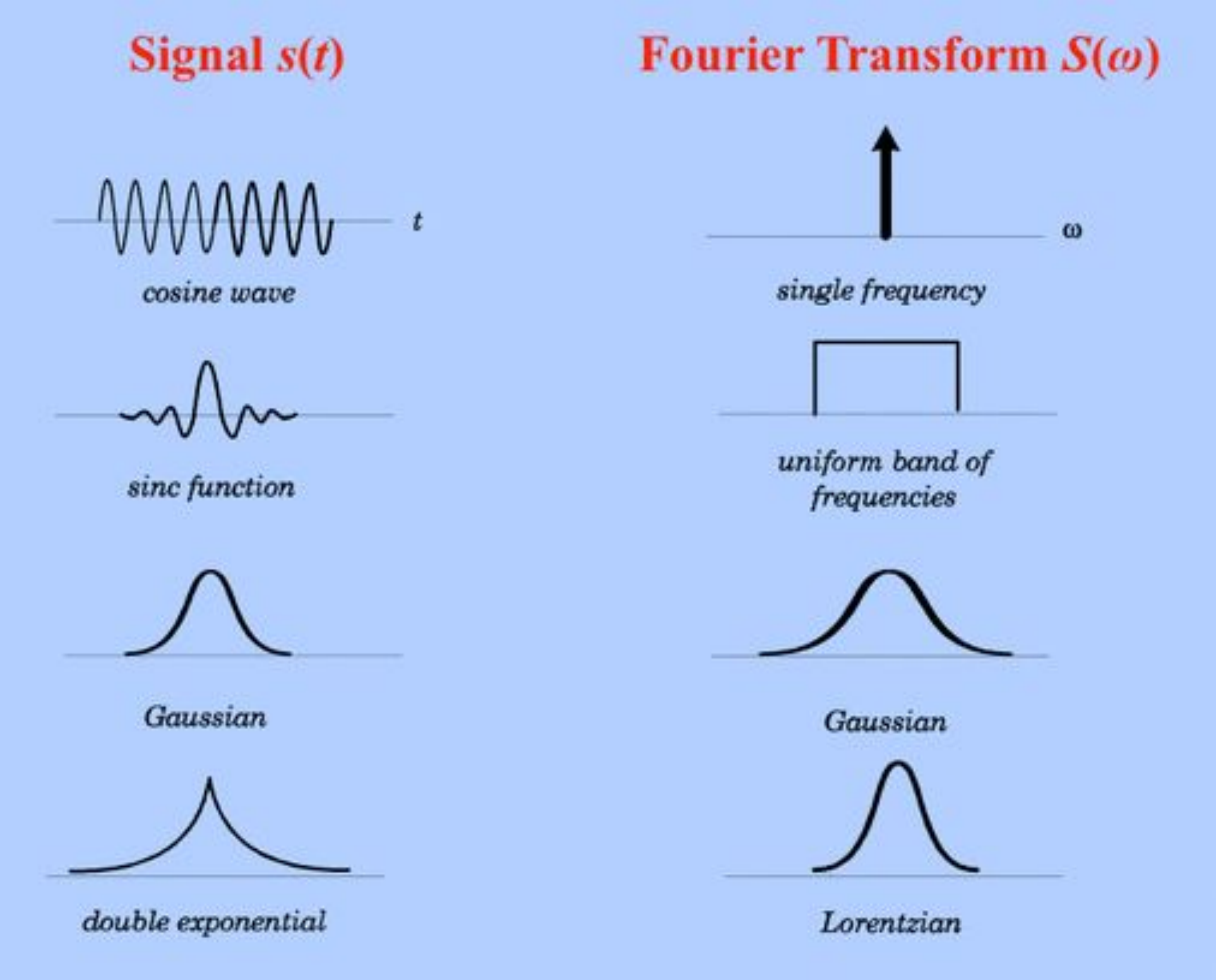


capture local geometric features

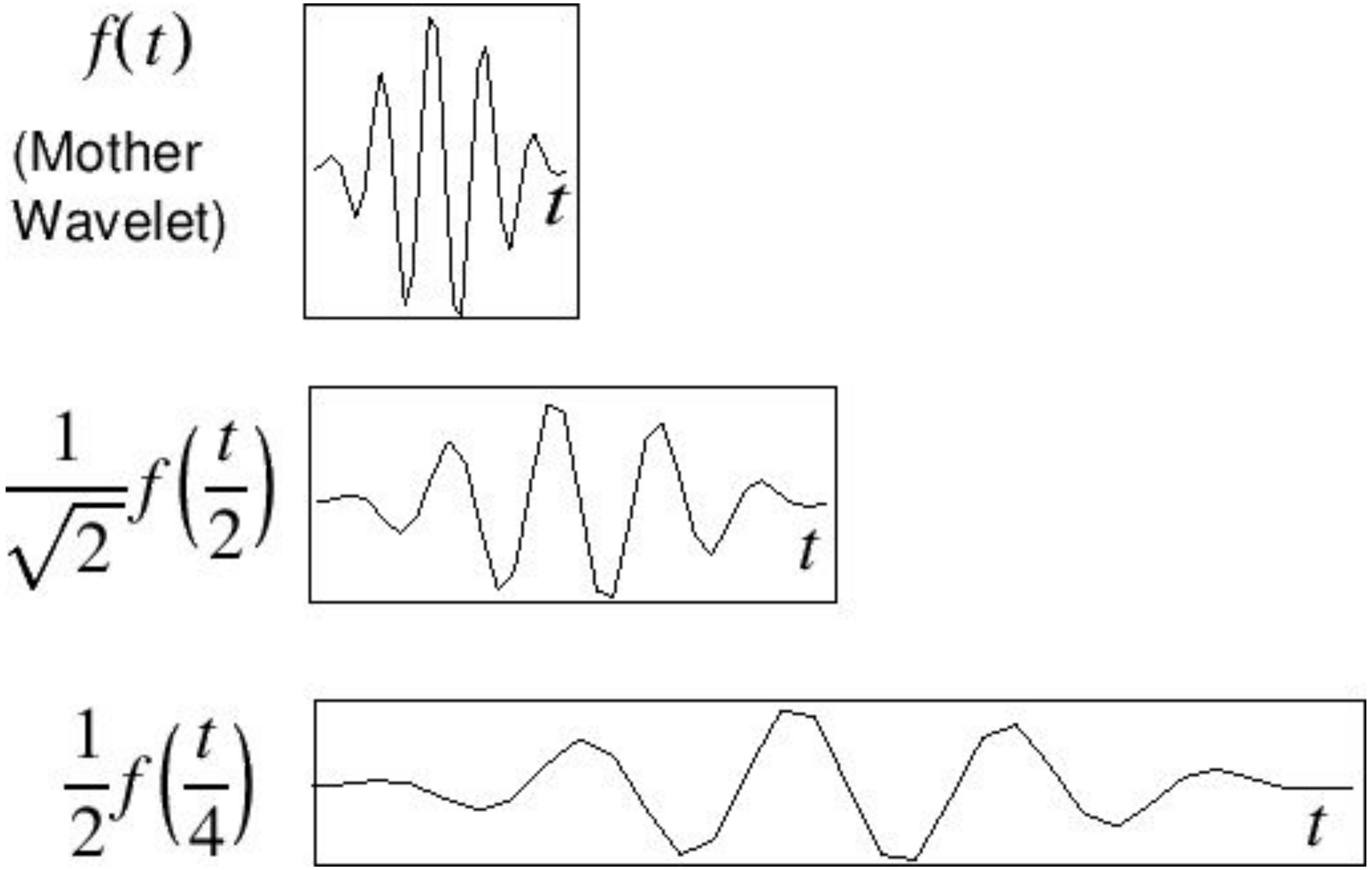
1 .hierarchical structure   2. multi-scale feature aggregation

Qi, Charles Ruizhongtai, et al. "Pointnet++: Deep hierarchical feature learning on point sets in a metric space." *Advances in neural information processing systems* 30 (2017).

Processing Time Series Data



Fourier Transform

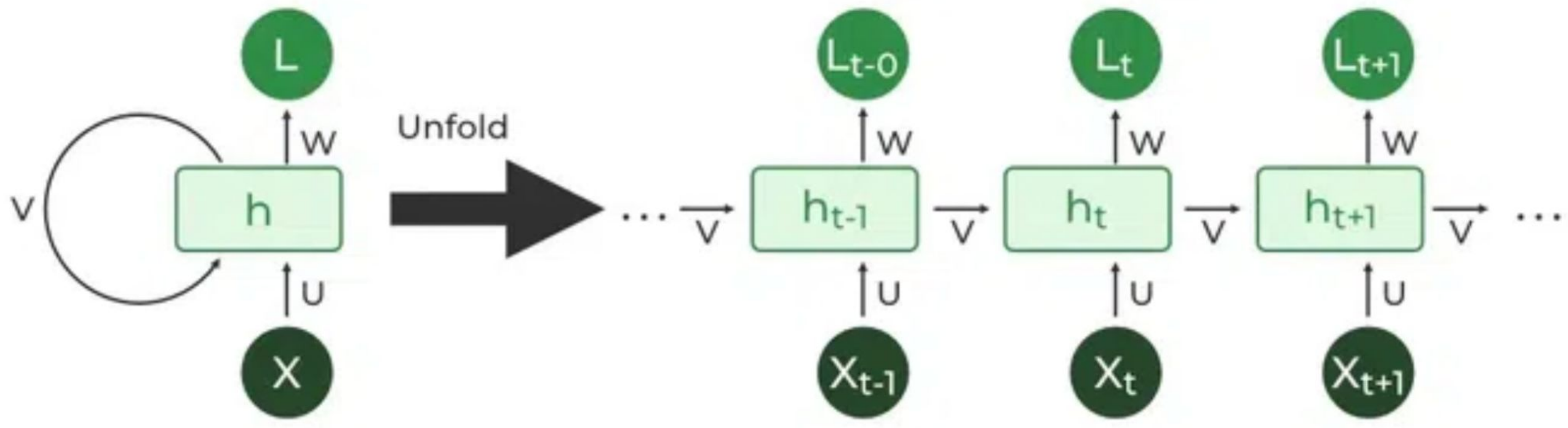


Wavelet Transform

https://mriquestions.com/fourier-transform-ft.html

https://ccrma.stanford.edu/~jos/sasp/Continuous_Wavelet_Transform.html

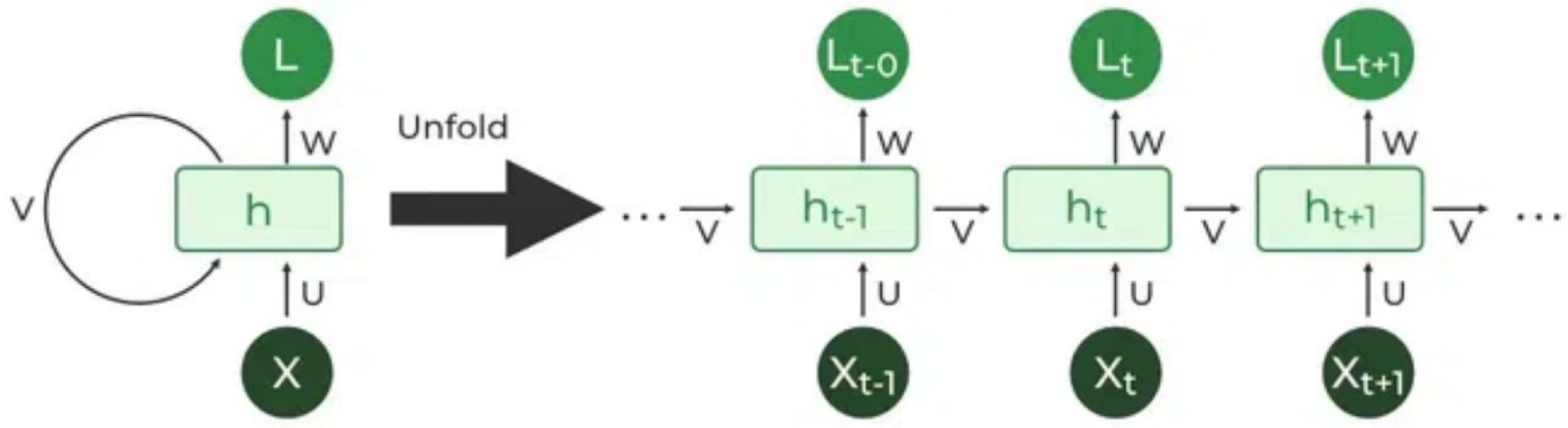
Processing Time Series Data



RNNs / Attention

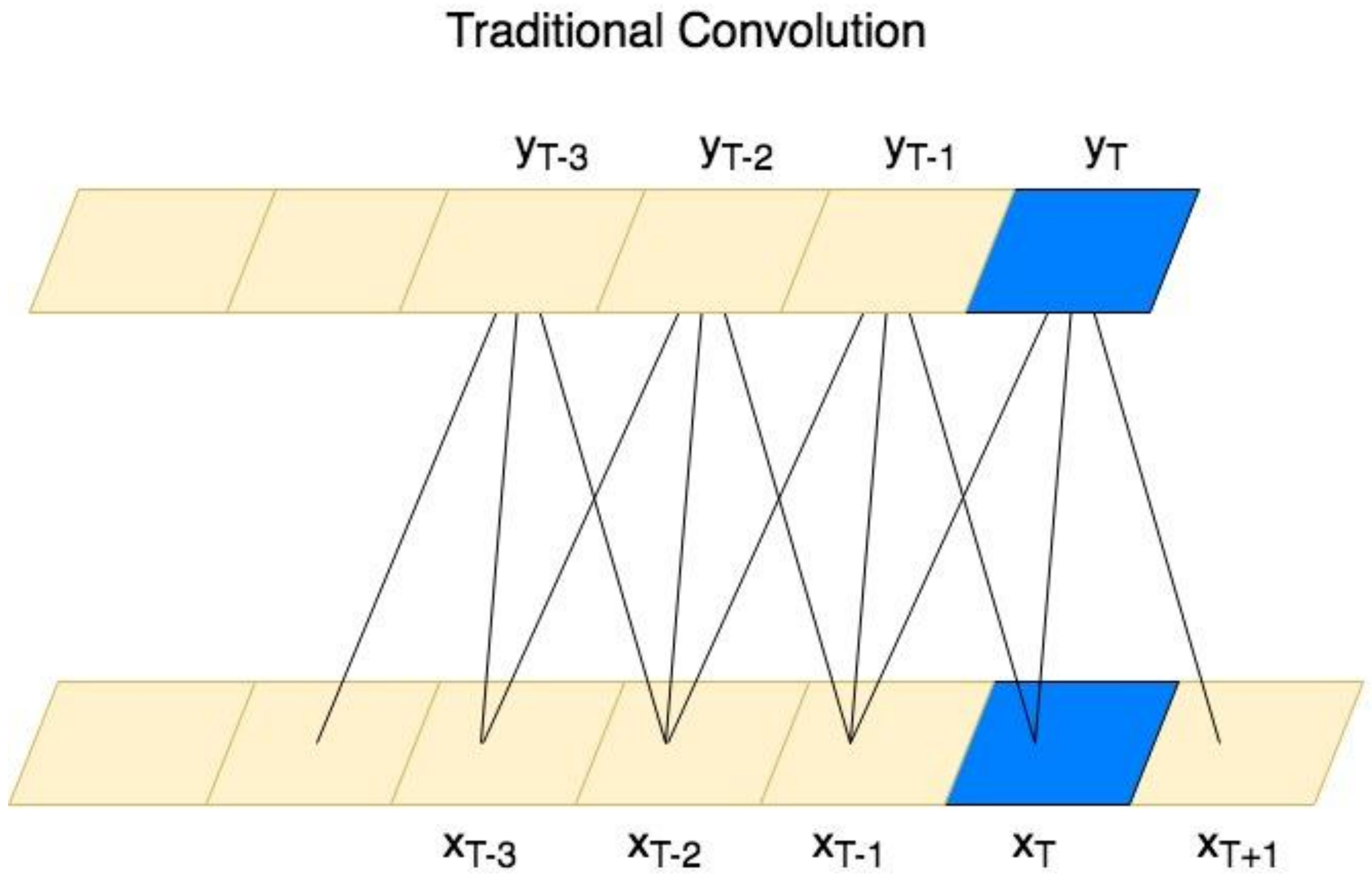
https://www.geeksforgeeks.org/introduction-to-recurrent-neural-network/#

Processing Time Series Data



RNNs / Attention

https://www.geeksforgeeks.org/introduction-to-recurrent-neural-network/#

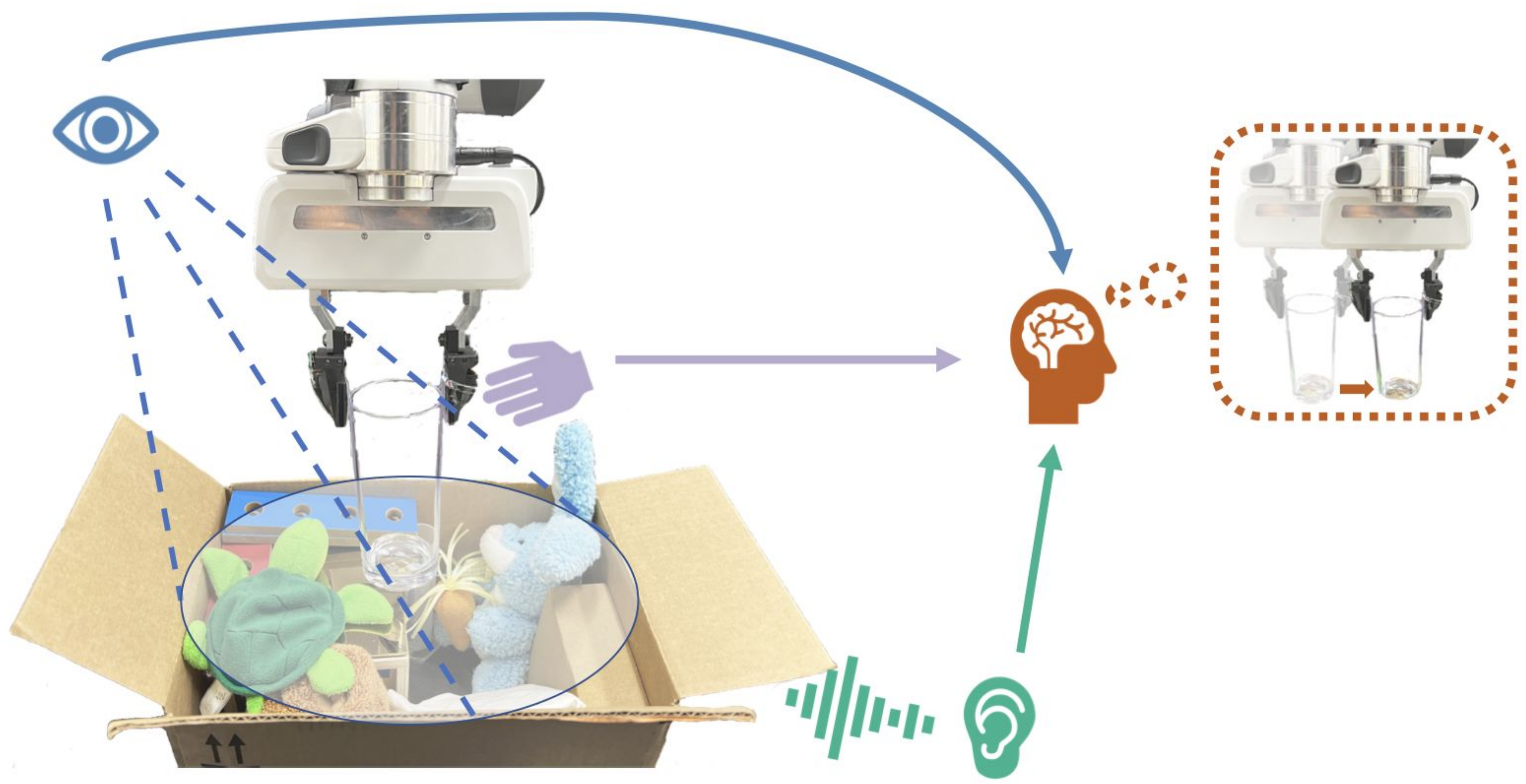


Causal Convolution

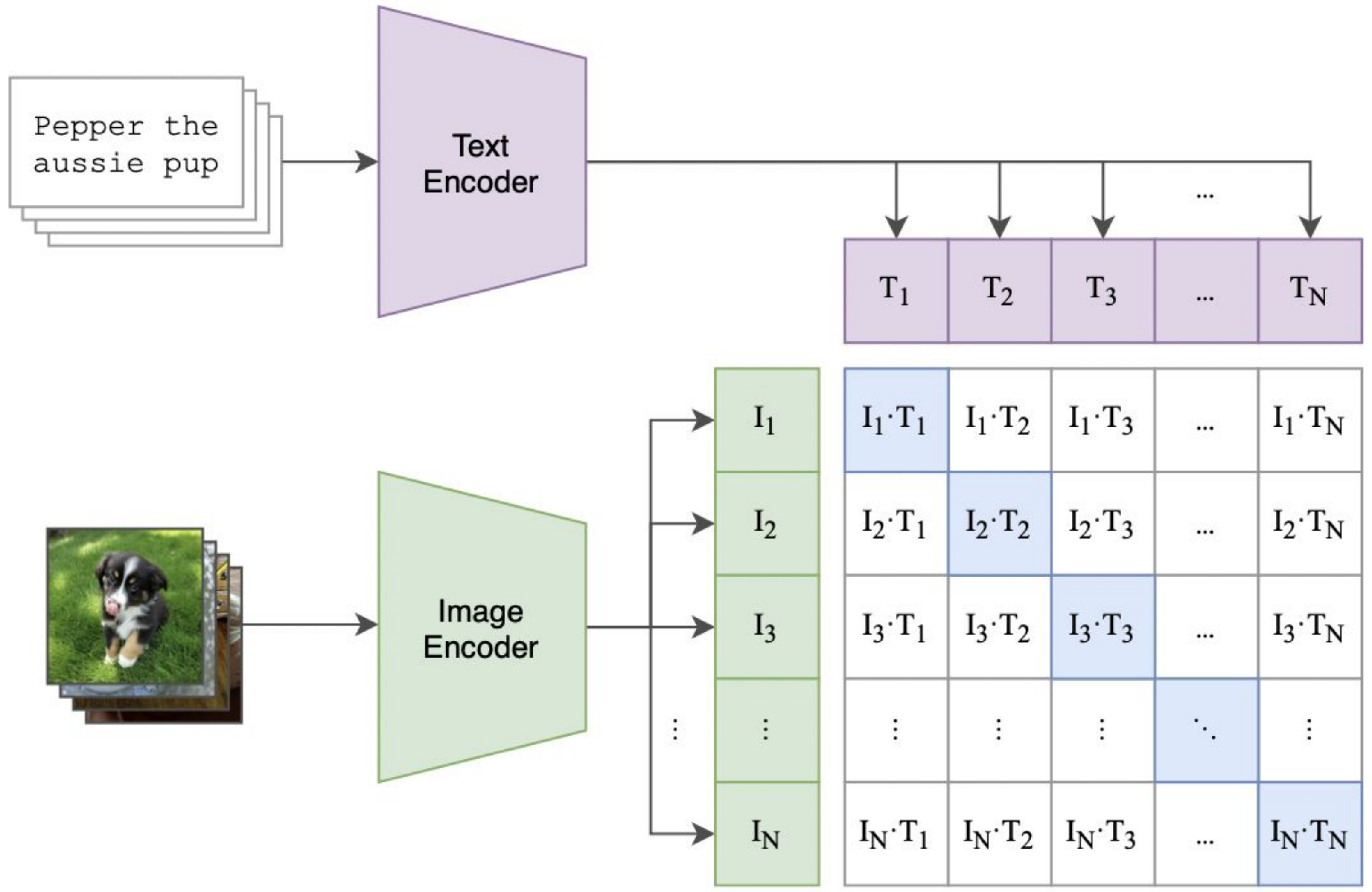
https://blog.fastforwardlabs.com/2018/05/31/convolve-all-the-things.html

Questions?

H. Li, Y. Zhang, J. Zhu, S. Wang, M. A. Lee, H. Xu, E. Adelson, L. Fei-Fei, R. Gao, and J. Wu, “See, hear, and feel: Smart sensory fusion for robotic manipulation,” in Conference on Robot Learning, pp. 1368–1378, PMLR, 2023.



What is the difference between Multi***sensory*** vs. Multi***modal*** ?



A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al., “Learning transferable visual models from natural language supervision,” in International conference on machine learning, pp. 8748–8763, PMLR, 2021.

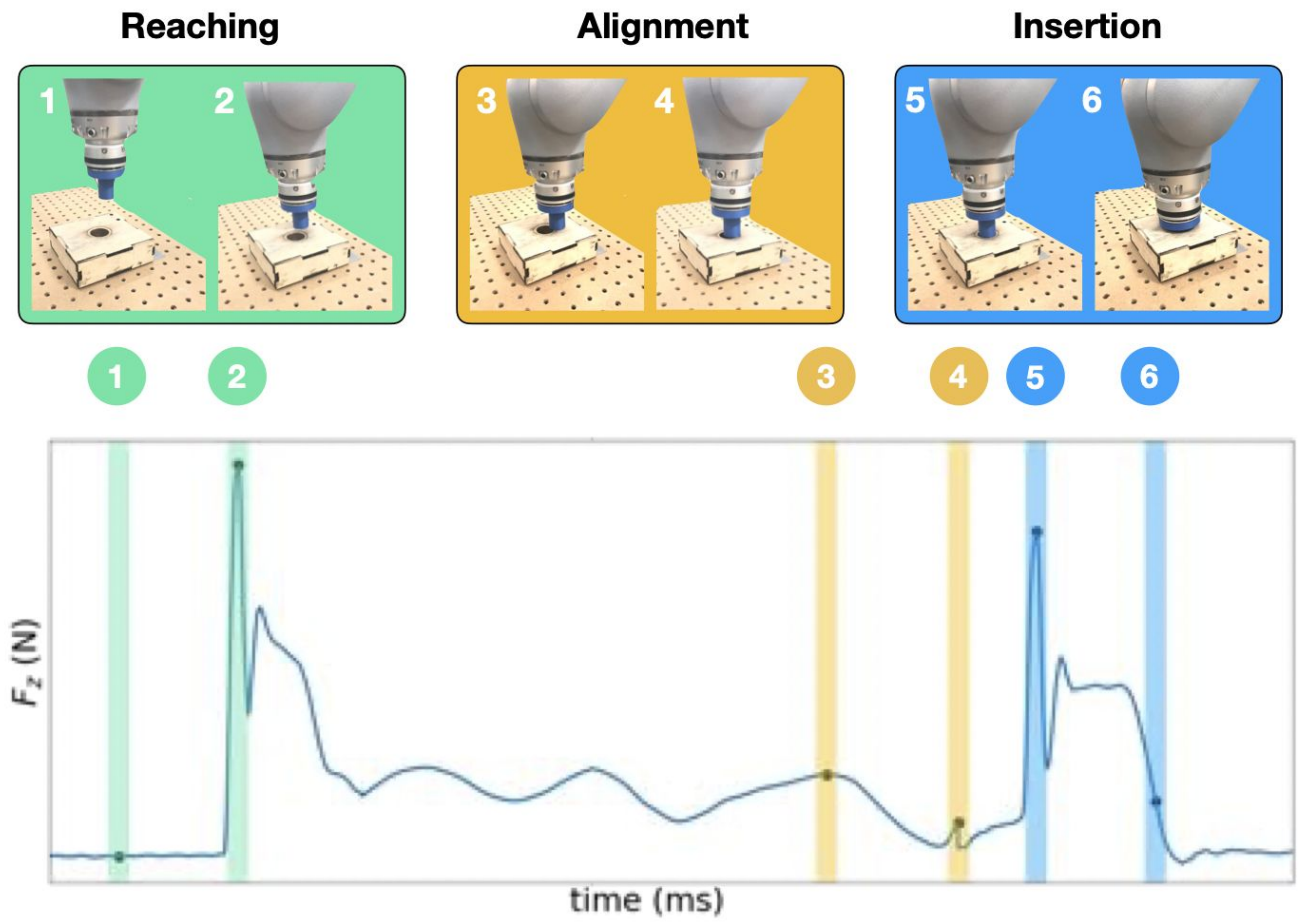
Putting it all together

Making Sense of Vision and Touch: Self-Supervised Learning of Multimodal Representations for Contact-Rich Tasks

Michelle A. Lee, Yuke Zhu, Krishnan Srinivasan, Parth Shah, Silvio Savarese, Li Fei-Fei, Animesh Garg, Jeannette Bohg

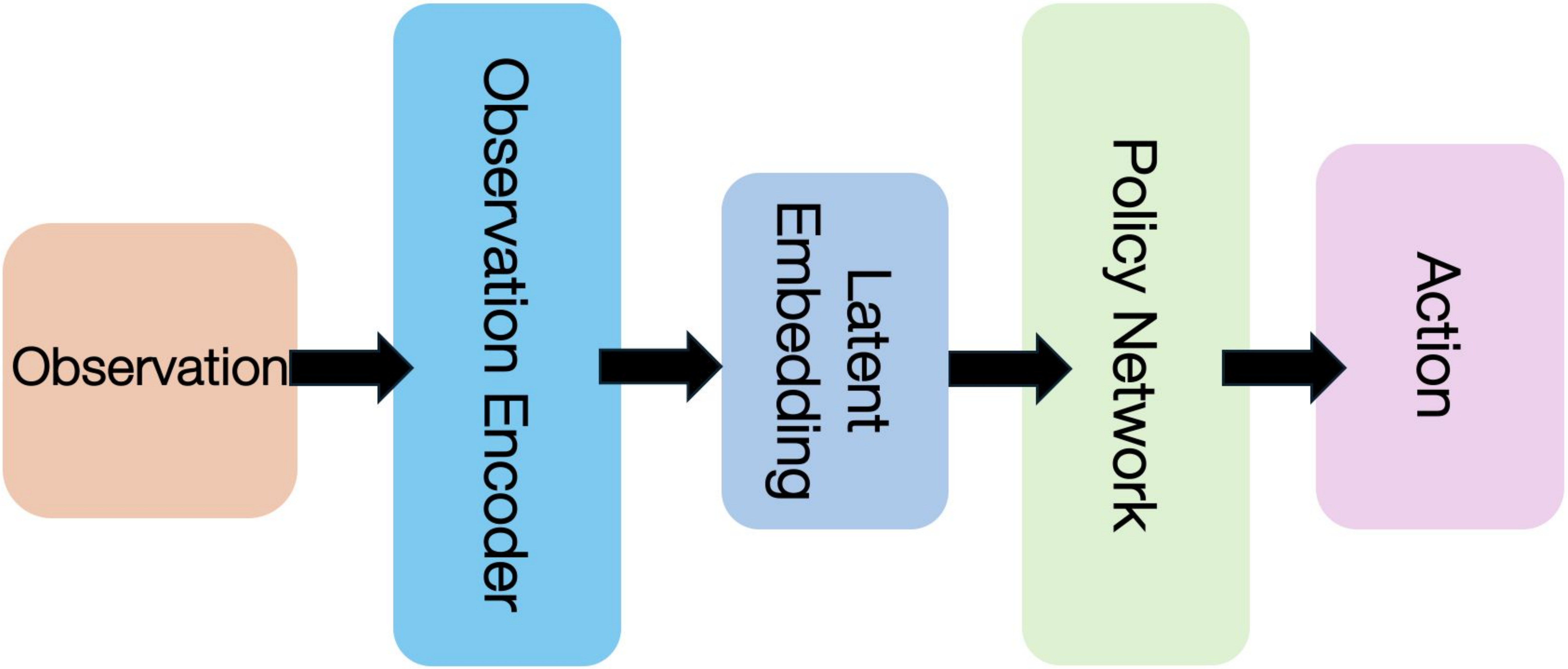
ICRA 2019 [Best Paper Award]
T-RO 2020 [Extended Version]

How can we learn good latent representations for contact-rich tasks?



M. A. Lee, Y. Zhu, K. Srinivasan, P. Shah, S. Savarese, L. Fei-Fei, A. Garg, and J. Bohg, “Making sense of vision and touch: Self-supervised learning of multimodal representations for contact-rich tasks,” in 2019 International conference on robotics and automation (ICRA), pp. 8943–8950, IEEE, 2019.

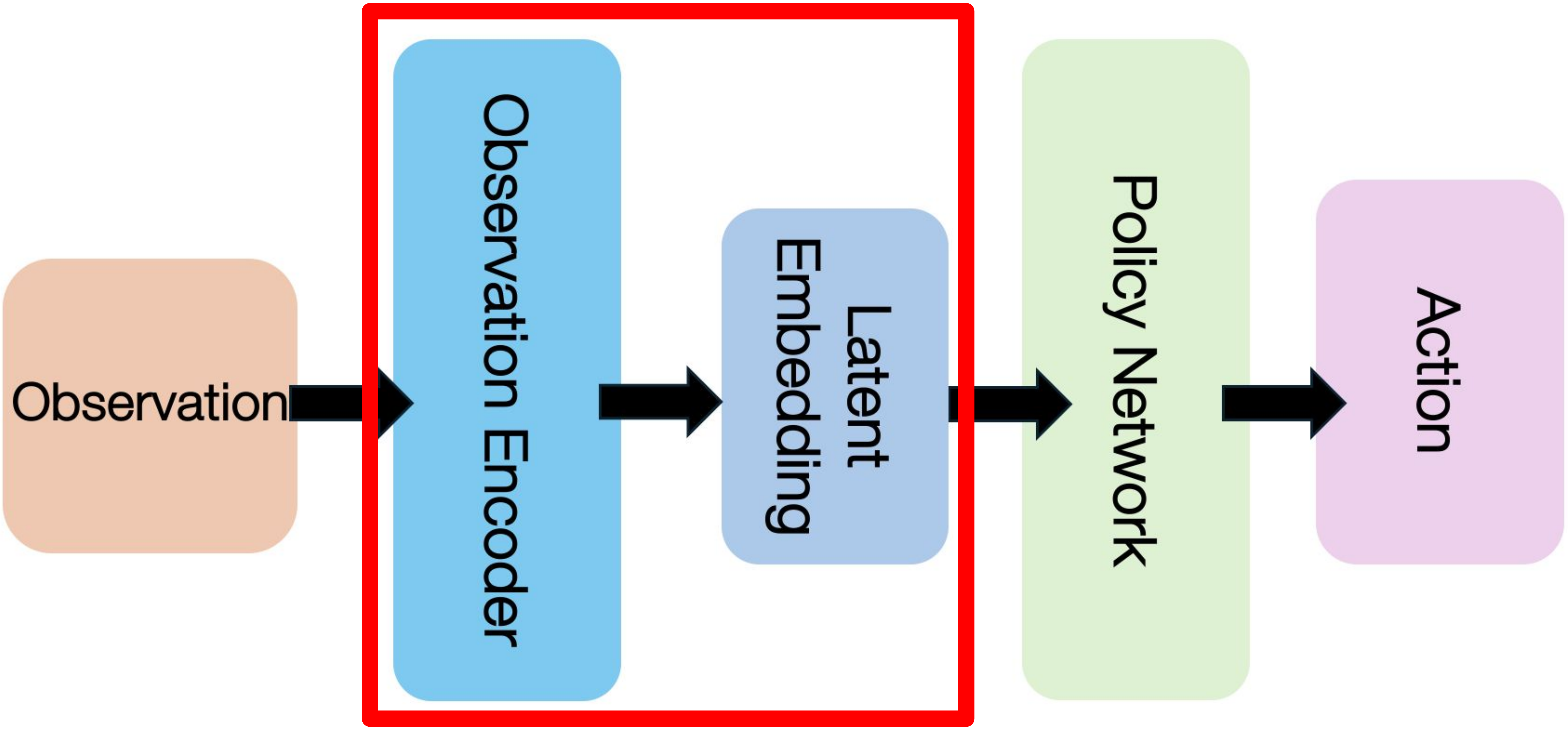
Idea: Decouple representation and policy learning



M. A. Lee, Y. Zhu, K. Srinivasan, P. Shah, S. Savarese, L. Fei-Fei, A. Garg, and J. Bohg, "Making sense of vision and touch: Self-supervised learning of multimodal representations for contact-rich tasks," in 2019 International conference on robotics and automation (ICRA), pp. 8943–8950, IEEE, 2019.

Idea: Decouple representation and policy learning

1. Learn latent embedding space through self-supervised learning

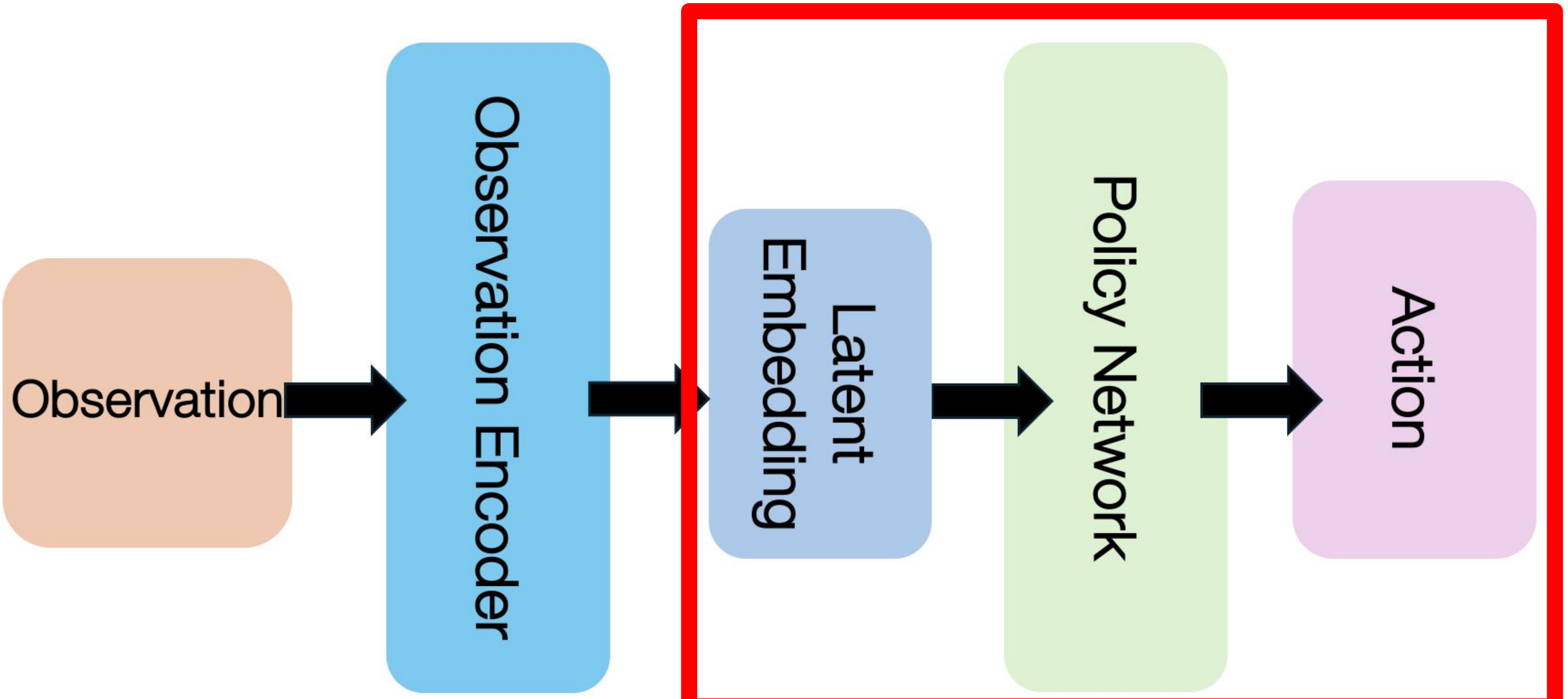


M. A. Lee, Y. Zhu, K. Srinivasan, P. Shah, S. Savarese, L. Fei-Fei, A. Garg, and J. Bohg, "Making sense of vision and touch: Self-supervised learning of multimodal representations for contact-rich tasks," in 2019 International conference on robotics and automation (ICRA), pp. 8943–8950, IEEE, 2019.

Idea: Decouple representation and policy learning

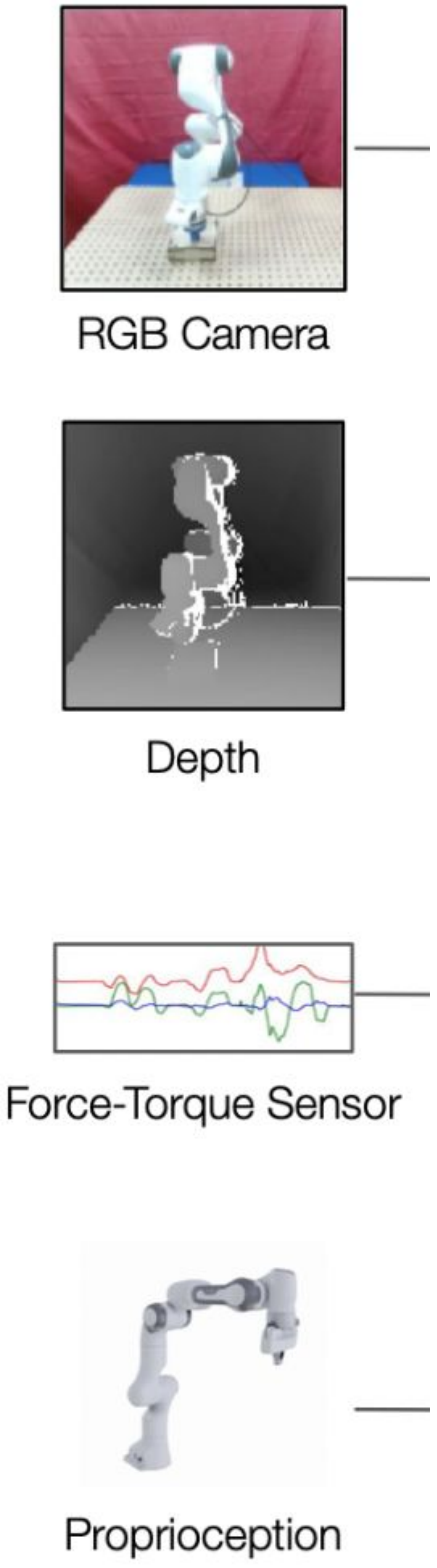
1. Learn latent embedding space through self-supervised learning

2. Use pretrained representation for policy learning



M. A. Lee, Y. Zhu, K. Srinivasan, P. Shah, S. Savarese, L. Fei-Fei, A. Garg, and J. Bohg, "Making sense of vision and touch: Self-supervised learning of multimodal representations for contact-rich tasks," in 2019 International conference on robotics and automation (ICRA), pp. 8943–8950, IEEE, 2019.

1. Learn latent embedding space through self-supervised learning

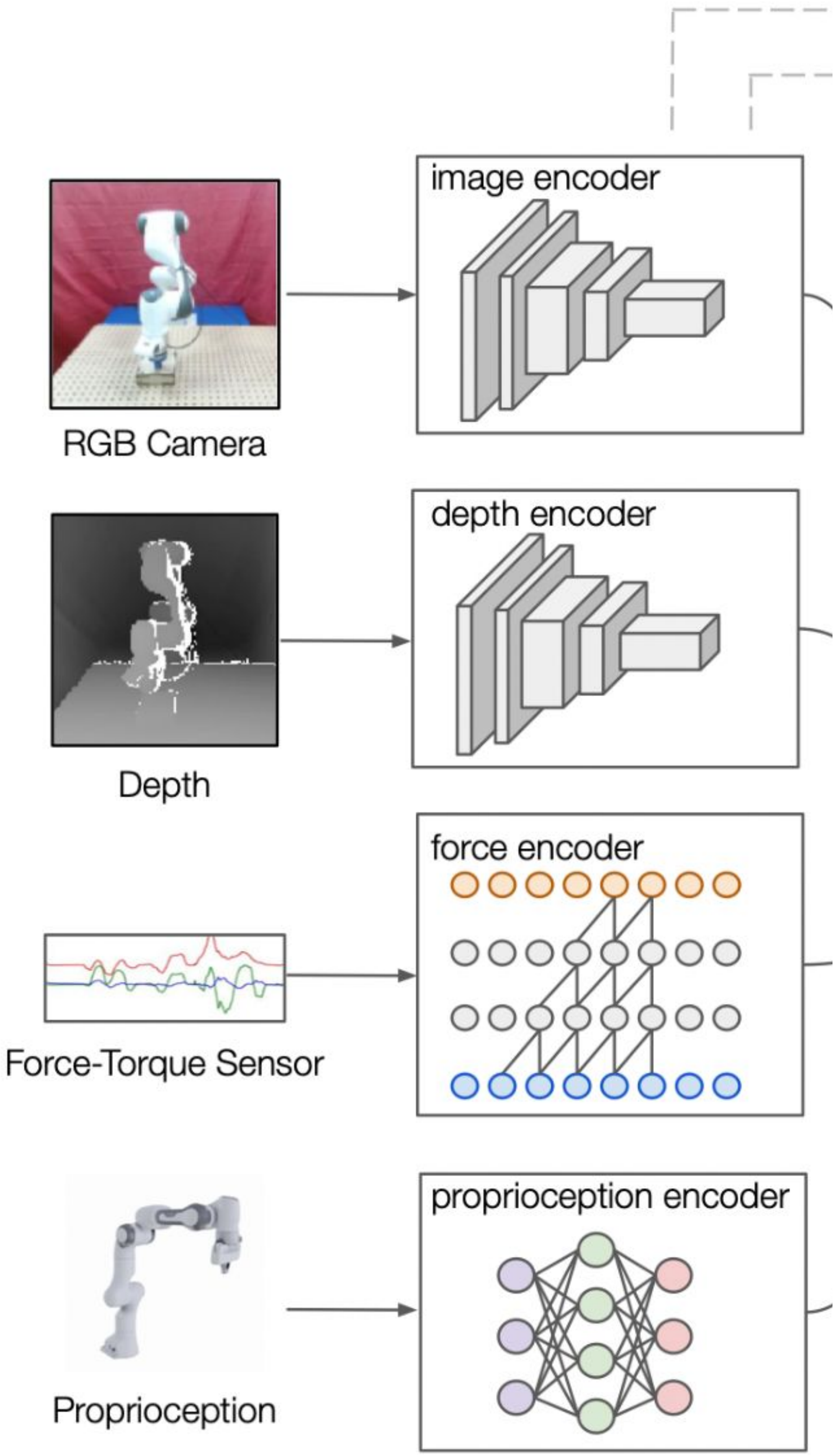


Sensing modes

M. A. Lee, Y. Zhu, K. Srinivasan, P. Shah, S. Savarese, L. Fei-Fei, A. Garg, and J. Bohg, “Making sense of vision and touch: Self-supervised learning of multimodal representations for contact-rich tasks,” in 2019 International conference on robotics and automation (ICRA), pp. 8943–8950, IEEE, 2019.

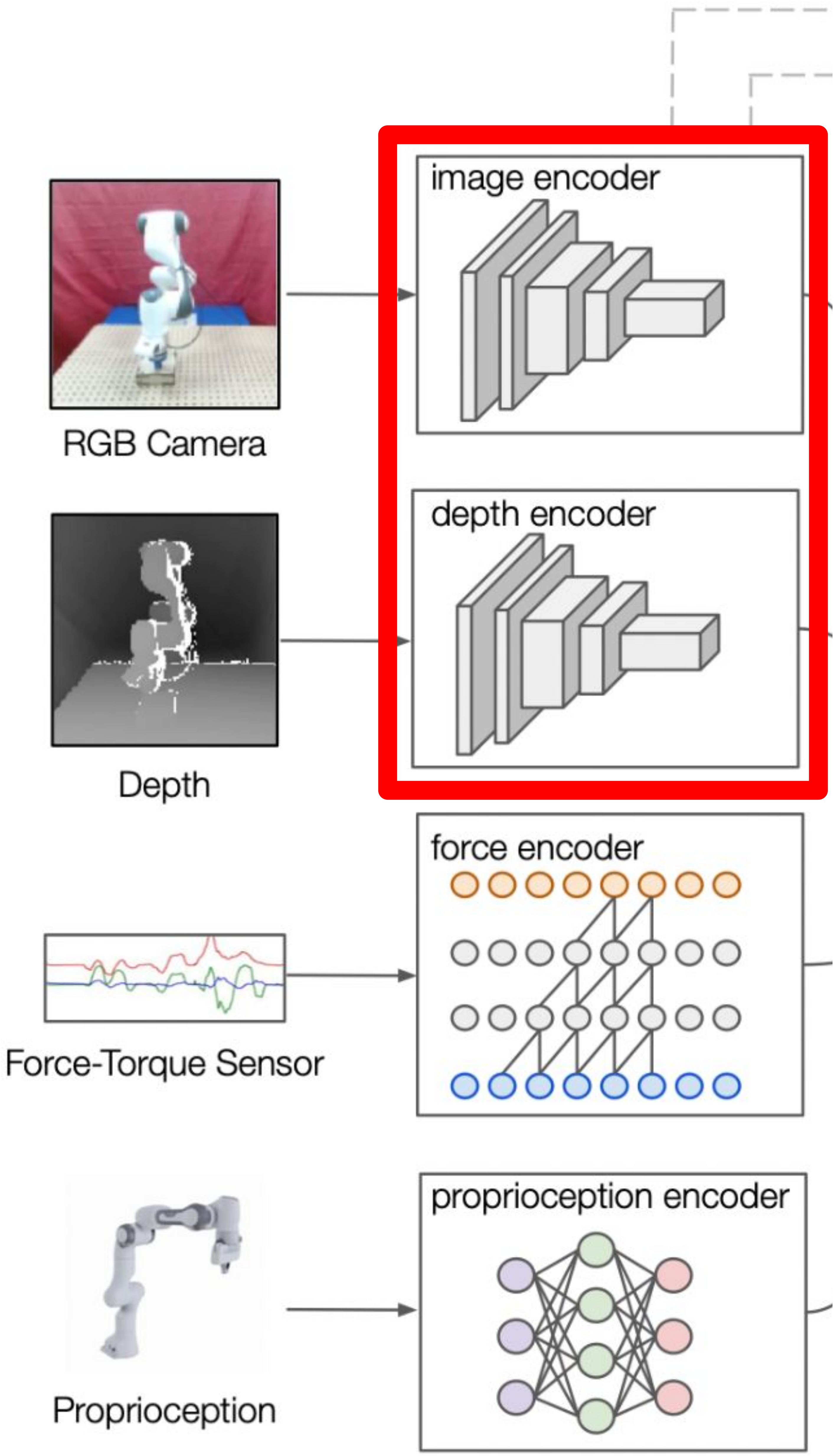
1. Learn latent embedding space through self-supervised learning

Domain-specific encoders



M. A. Lee, Y. Zhu, K. Srinivasan, P. Shah, S. Savarese, L. Fei-Fei, A. Garg, and J. Bohg, "Making sense of vision and touch: Self-supervised learning of multimodal representations for contact-rich tasks," in 2019 International conference on robotics and automation (ICRA), pp. 8943–8950, IEEE, 2019.

1. Learn latent embedding space through self-supervised learning



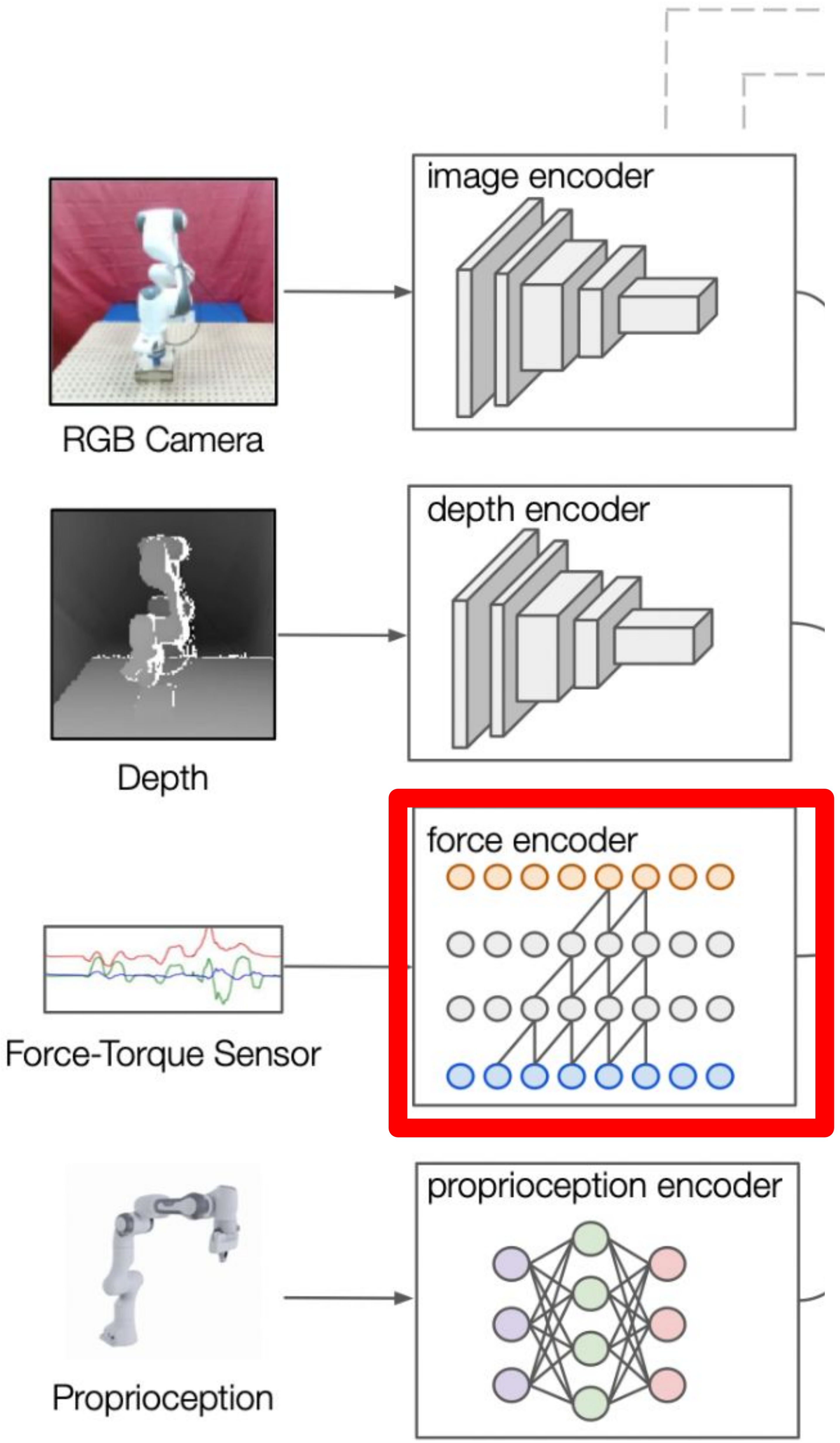
Domain-specific encoders

CNN (FlowNet)

CNN (VGG-16)

M. A. Lee, Y. Zhu, K. Srinivasan, P. Shah, S. Savarese, L. Fei-Fei, A. Garg, and J. Bohg, "Making sense of vision and touch: Self-supervised learning of multimodal representations for contact-rich tasks," in 2019 International conference on robotics and automation (ICRA), pp. 8943–8950, IEEE, 2019.

1. Learn latent embedding space through self-supervised learning



Domain-specific encoders

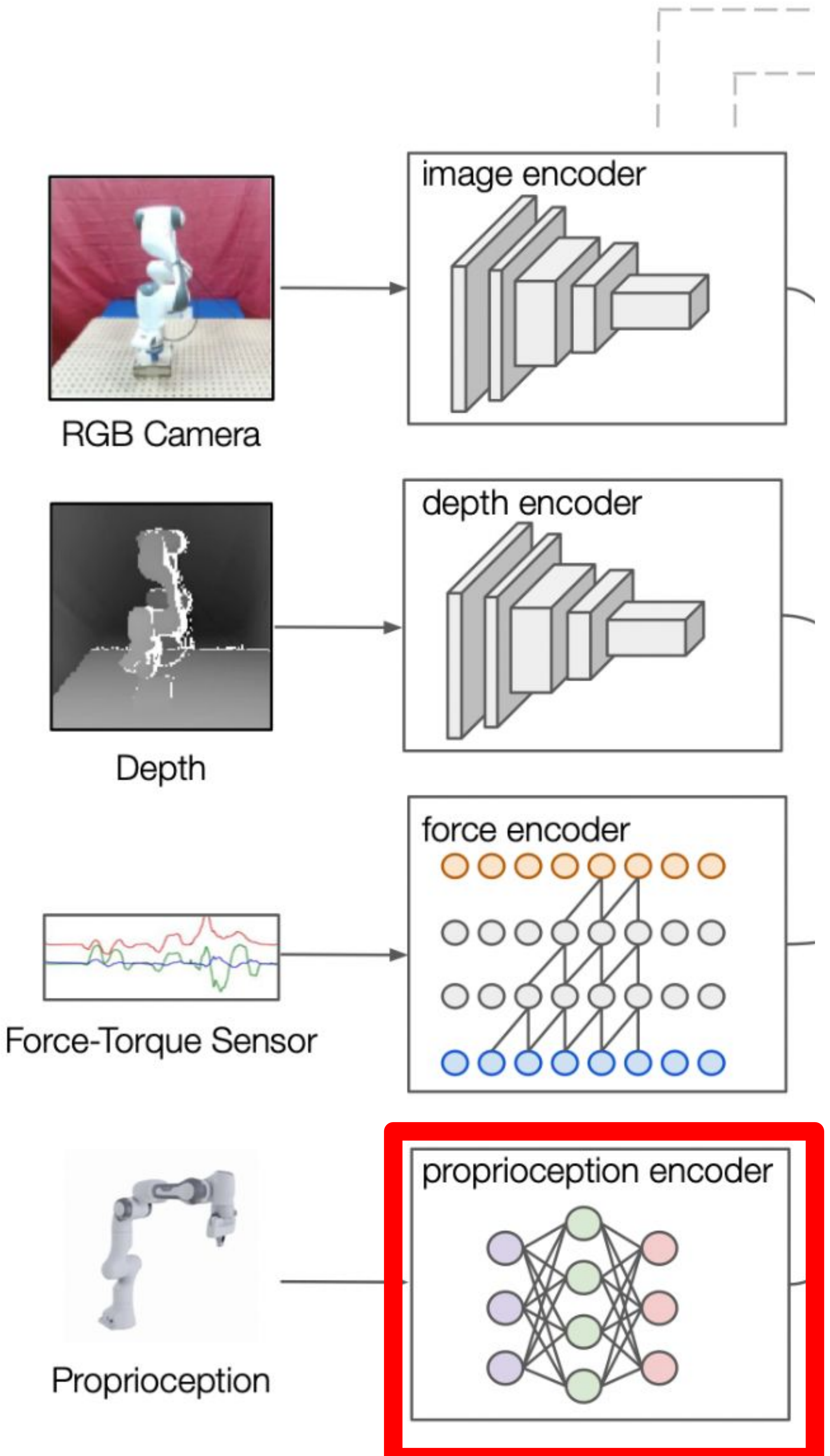
CNN (FlowNet)

CNN (VGG-16)

Causal Convolution Layer

M. A. Lee, Y. Zhu, K. Srinivasan, P. Shah, S. Savarese, L. Fei-Fei, A. Garg, and J. Bohg, "Making sense of vision and touch: Self-supervised learning of multimodal representations for contact-rich tasks," in 2019 International conference on robotics and automation (ICRA), pp. 8943–8950, IEEE, 2019.

1. Learn latent embedding space through self-supervised learning



Domain-specific encoders

CNN (FlowNet)

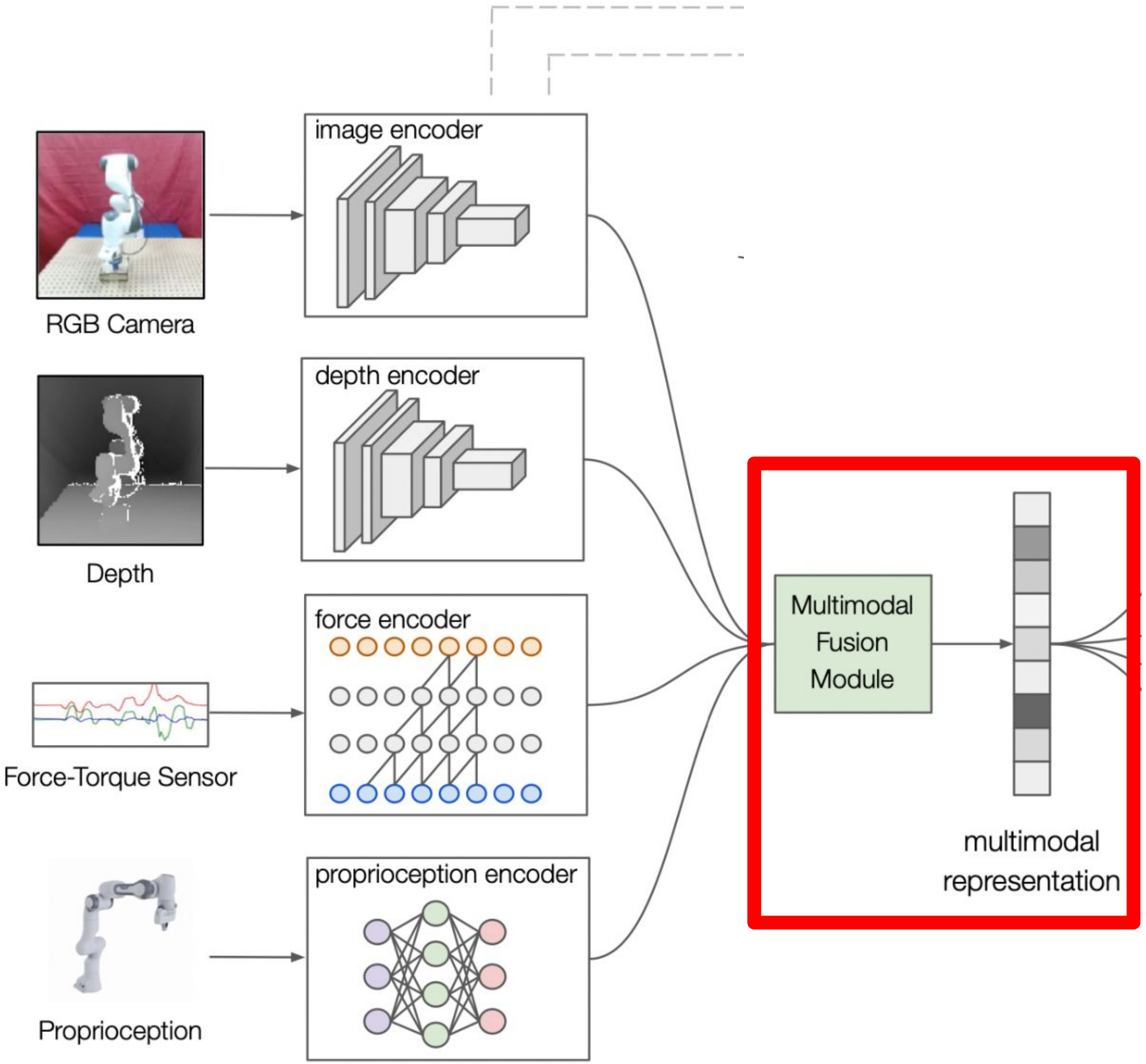
CNN (VGG-16)

Causal Convolution Layer

Fully-Connected Network (MLP)

M. A. Lee, Y. Zhu, K. Srinivasan, P. Shah, S. Savarese, L. Fei-Fei, A. Garg, and J. Bohg, "Making sense of vision and touch: Self-supervised learning of multimodal representations for contact-rich tasks," in 2019 International conference on robotics and automation (ICRA), pp. 8943–8950, IEEE, 2019.

1. Learn latent embedding space through self-supervised learning

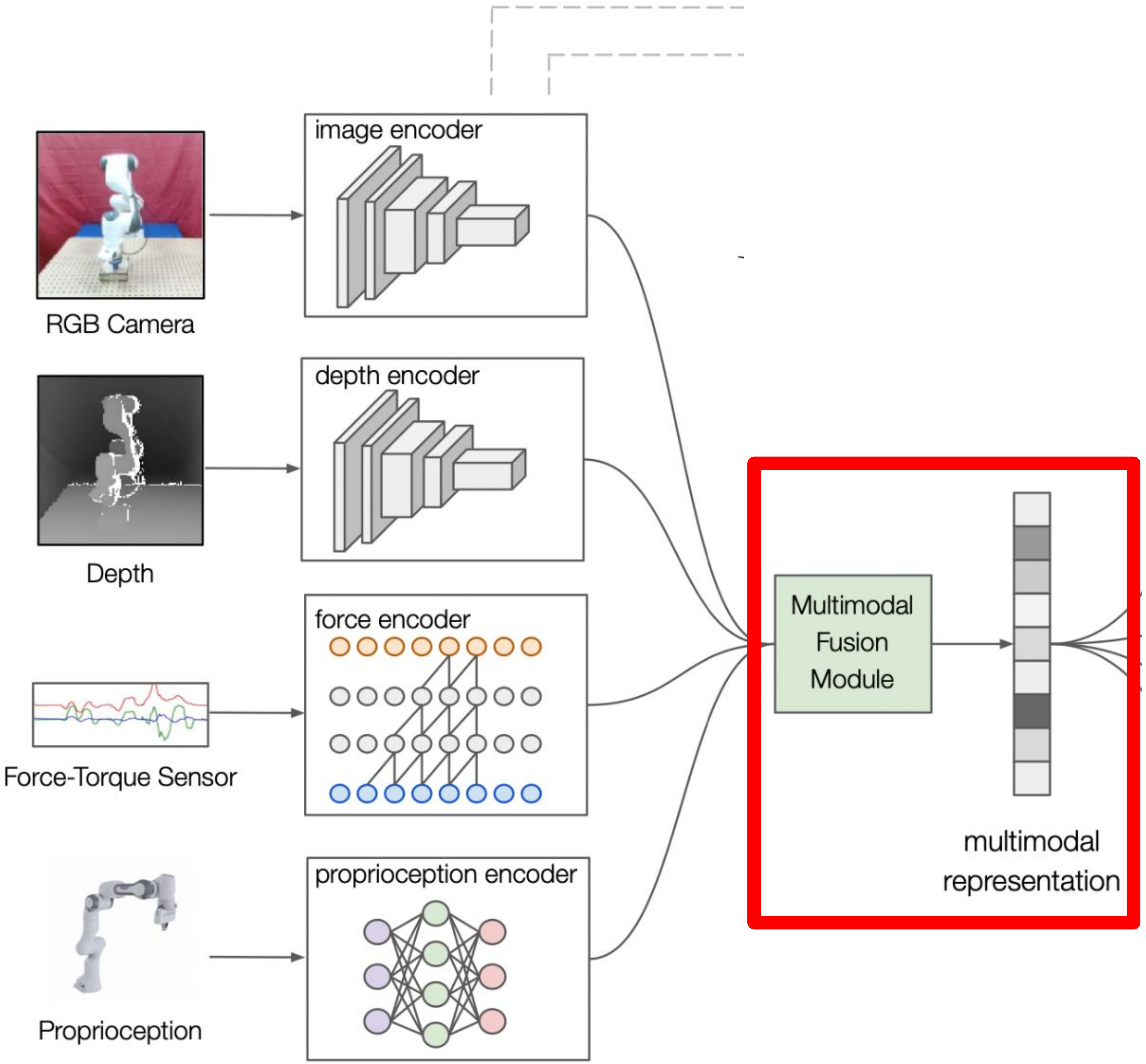


Multimodal Fusion

1. Simple concatenation

M. A. Lee, Y. Zhu, K. Srinivasan, P. Shah, S. Savarese, L. Fei-Fei, A. Garg, and J. Bohg, "Making sense of vision and touch: Self-supervised learning of multimodal representations for contact-rich tasks," in 2019 International conference on robotics and automation (ICRA), pp. 8943–8950, IEEE, 2019.

1. Learn latent embedding space through self-supervised learning



Multimodal Fusion

1. Simple concatenation

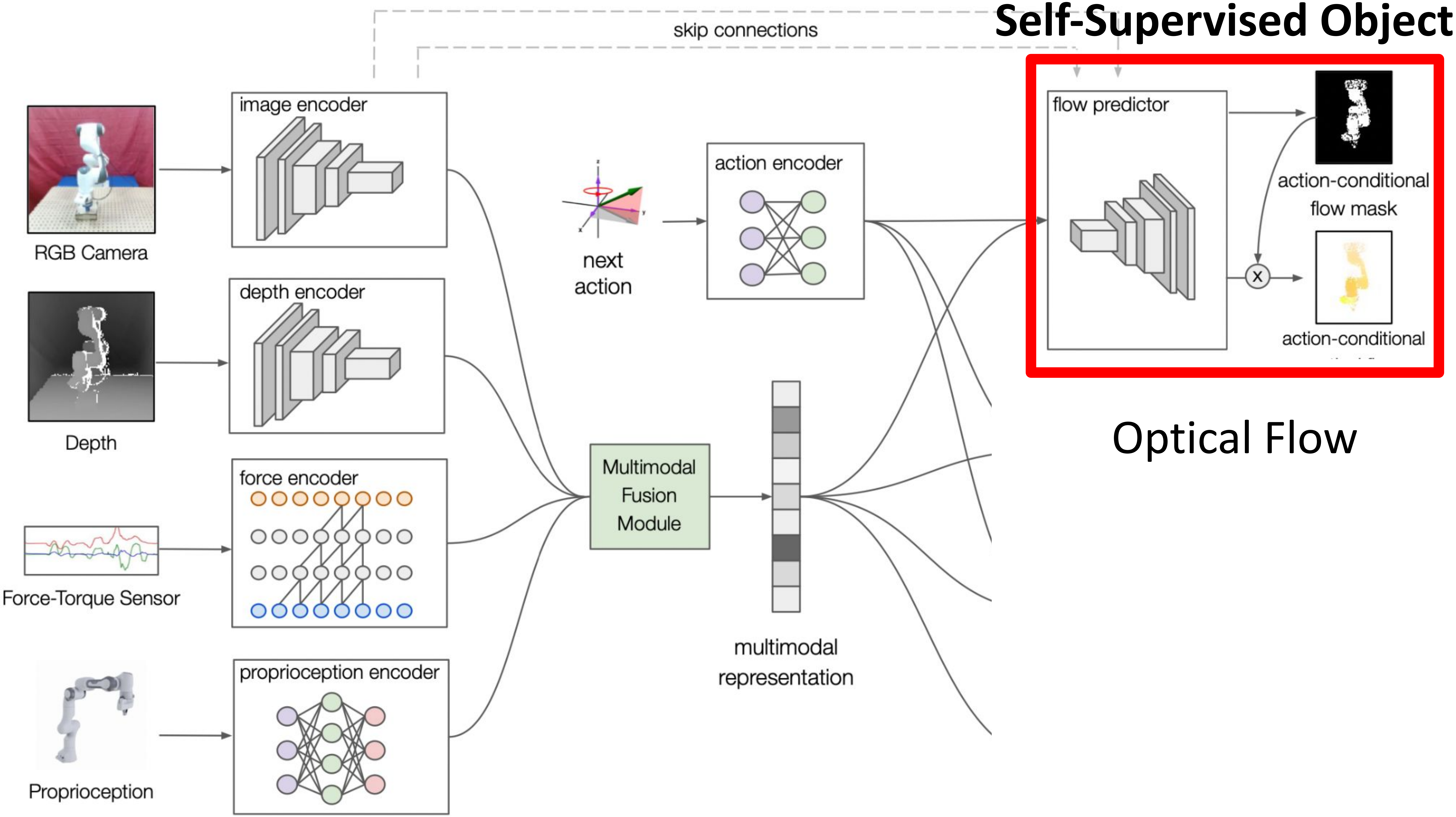
2. Variational inference + product of experts

(very similar to VAEs)

M. A. Lee, Y. Zhu, K. Srinivasan, P. Shah, S. Savarese, L. Fei-Fei, A. Garg, and J. Bohg, "Making sense of vision and touch: Self-supervised learning of multimodal representations for contact-rich tasks," in 2019 International conference on robotics and automation (ICRA), pp. 8943–8950, IEEE, 2019.

1. Learn latent embedding space through self-supervised learning

Self-Supervised Objectives

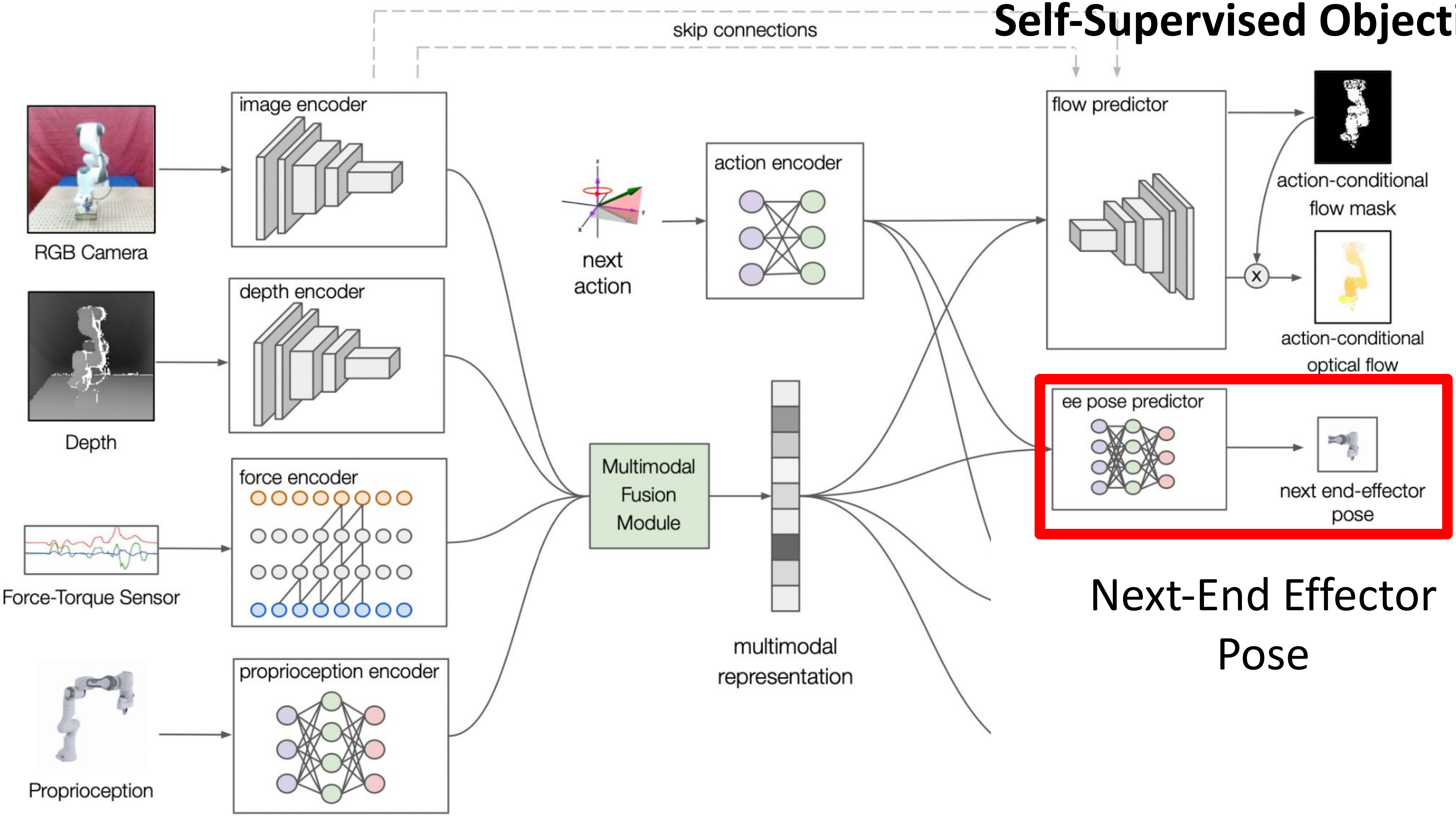


Optical Flow

M. A. Lee, Y. Zhu, K. Srinivasan, P. Shah, S. Savarese, L. Fei-Fei, A. Garg, and J. Bohg, "Making sense of vision and touch: Self-supervised learning of multimodal representations for contact-rich tasks," in 2019 International conference on robotics and automation (ICRA), pp. 8943–8950, IEEE, 2019.

1. Learn latent embedding space through self-supervised learning

Self-Supervised Objectives

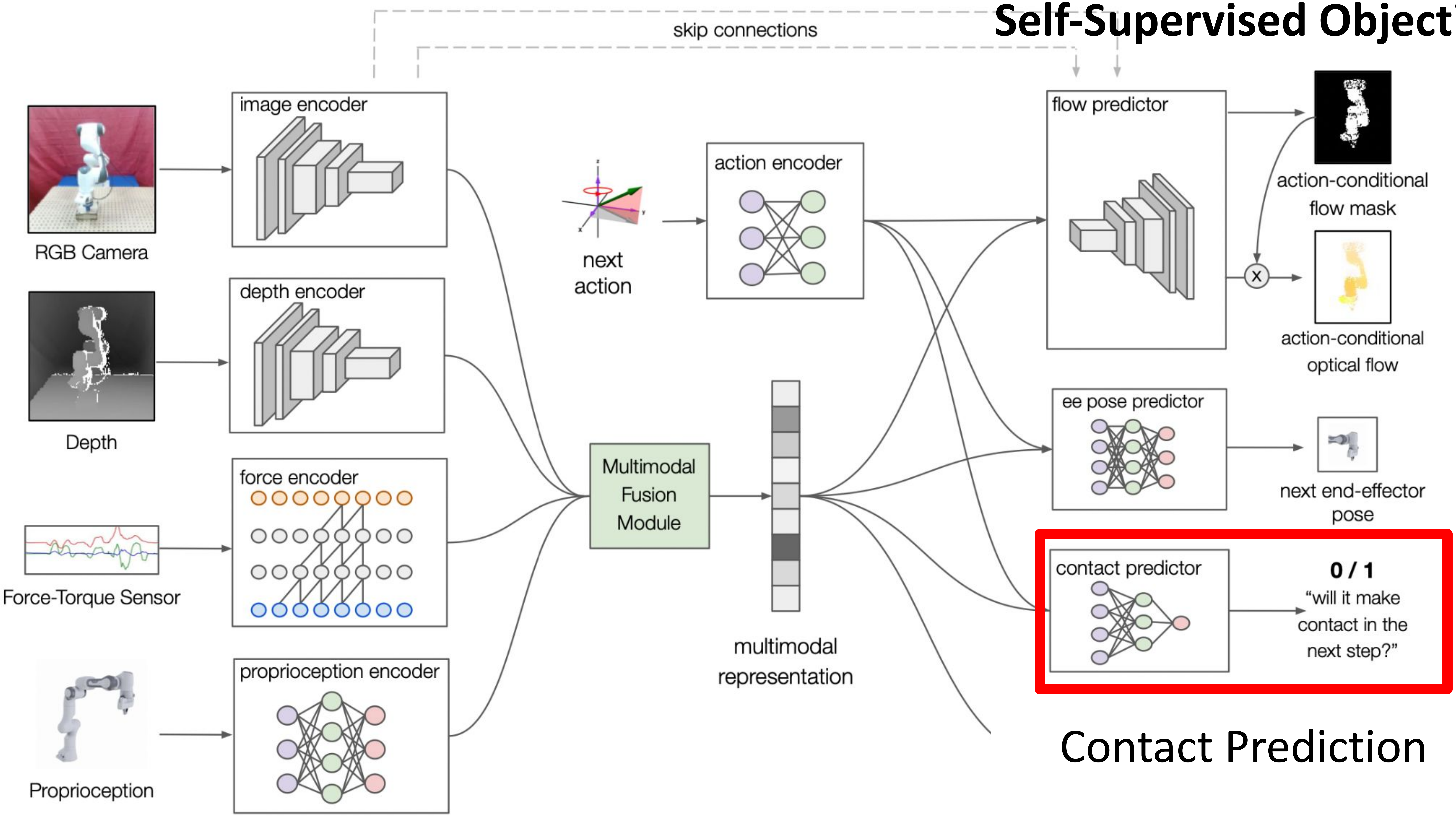


Next-End Effector Pose

M. A. Lee, Y. Zhu, K. Srinivasan, P. Shah, S. Savarese, L. Fei-Fei, A. Garg, and J. Bohg, "Making sense of vision and touch: Self-supervised learning of multimodal representations for contact-rich tasks," in 2019 International conference on robotics and automation (ICRA), pp. 8943–8950, IEEE, 2019.

1. Learn latent embedding space through self-supervised learning

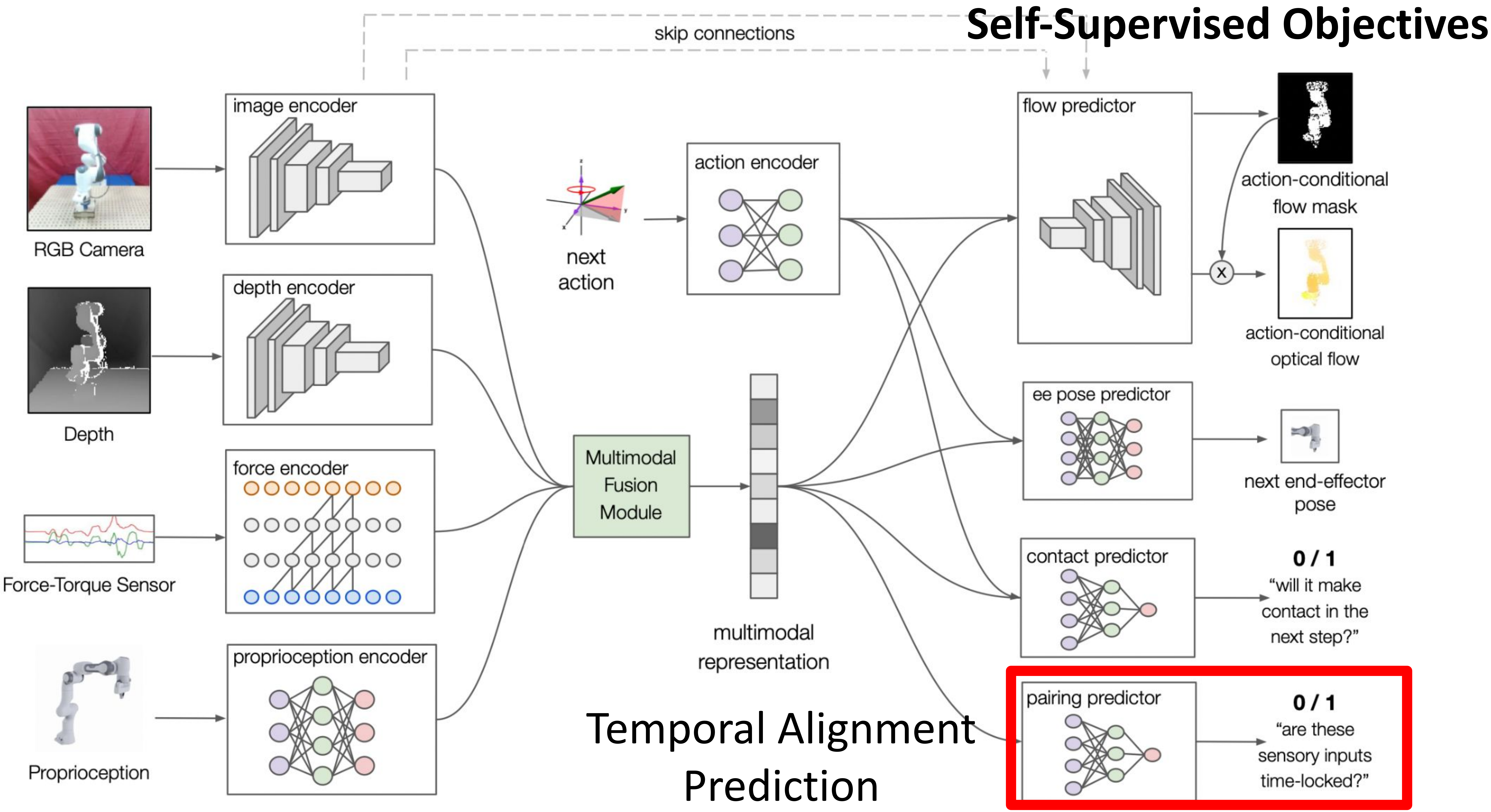
Self-Supervised Objectives



Contact Prediction

M. A. Lee, Y. Zhu, K. Srinivasan, P. Shah, S. Savarese, L. Fei-Fei, A. Garg, and J. Bohg, "Making sense of vision and touch: Self-supervised learning of multimodal representations for contact-rich tasks," in 2019 International conference on robotics and automation (ICRA), pp. 8943–8950, IEEE, 2019.

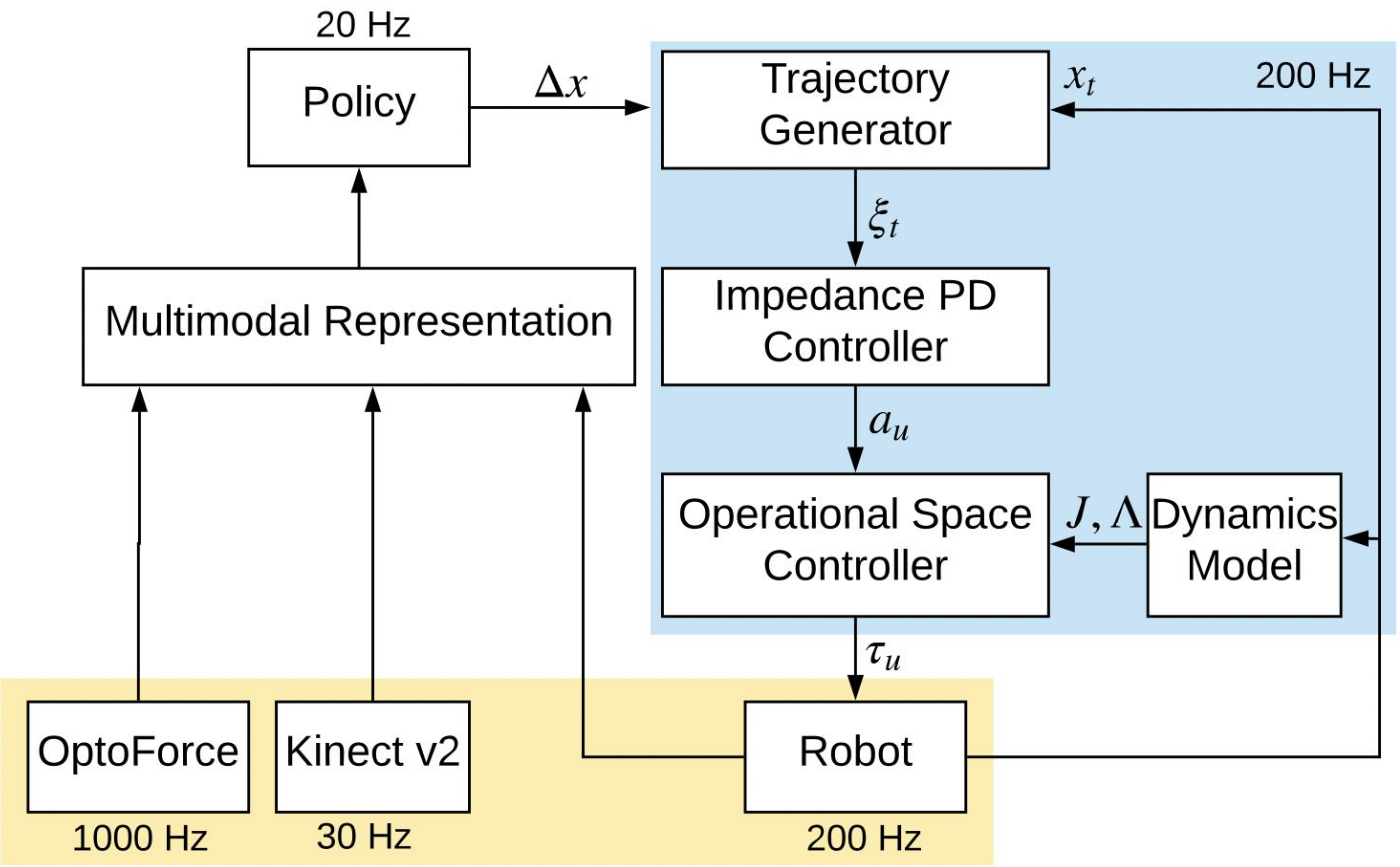
1. Learn latent embedding space through self-supervised learning



Temporal Alignment
Prediction

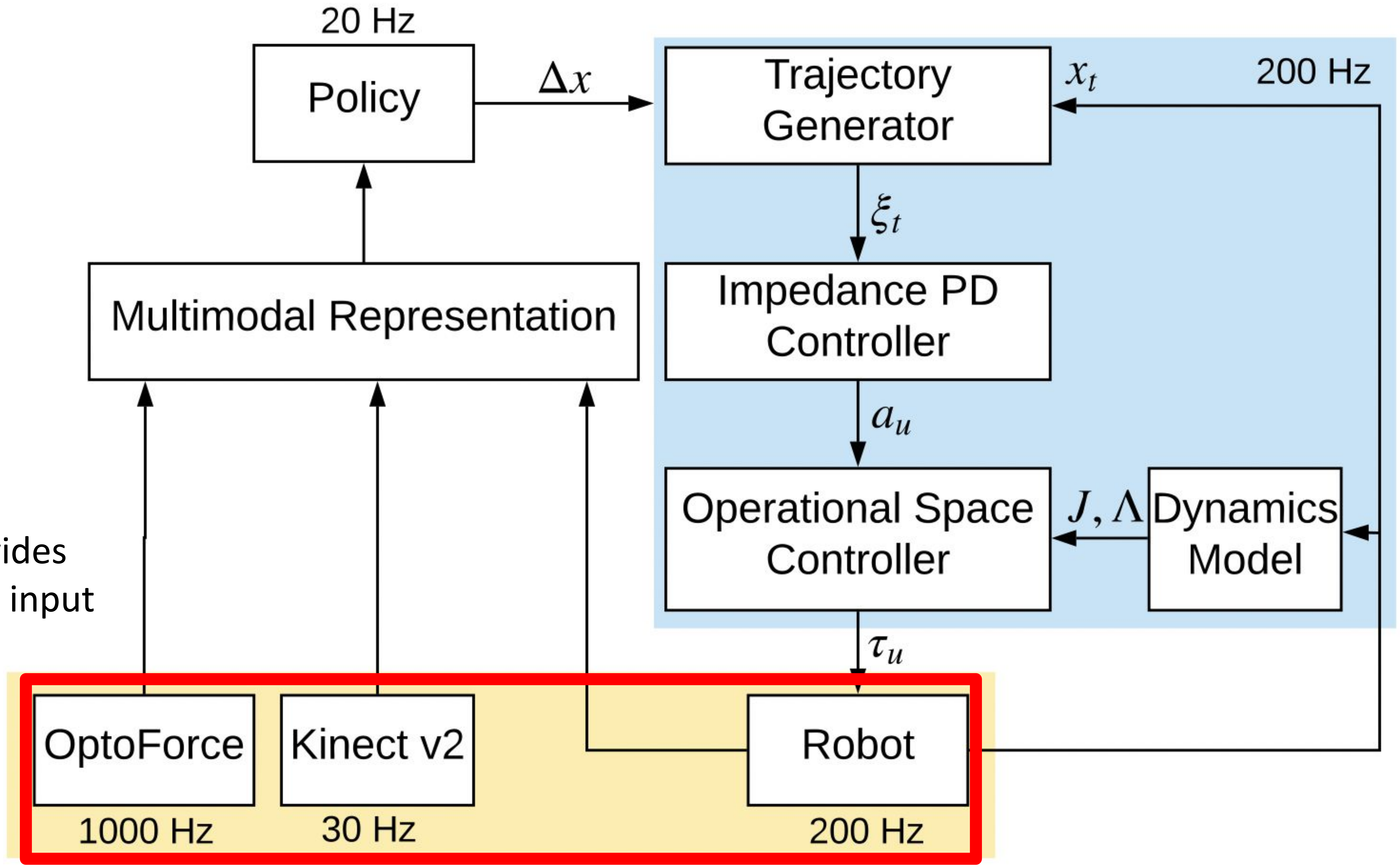
M. A. Lee, Y. Zhu, K. Srinivasan, P. Shah, S. Savarese, L. Fei-Fei, A. Garg, and J. Bohg, "Making sense of vision and touch: Self-supervised learning of multimodal representations for contact-rich tasks," in 2019 International conference on robotics and automation (ICRA), pp. 8943–8950, IEEE, 2019.

2. Use pretrained representation for policy learning



M. A. Lee, Y. Zhu, K. Srinivasan, P. Shah, S. Savarese, L. Fei-Fei, A. Garg, and J. Bohg, "Making sense of vision and touch: Self-supervised learning of multimodal representations for contact-rich tasks," in 2019 International conference on robotics and automation (ICRA), pp. 8943–8950, IEEE, 2019.

2. Use pretrained representation for policy learning

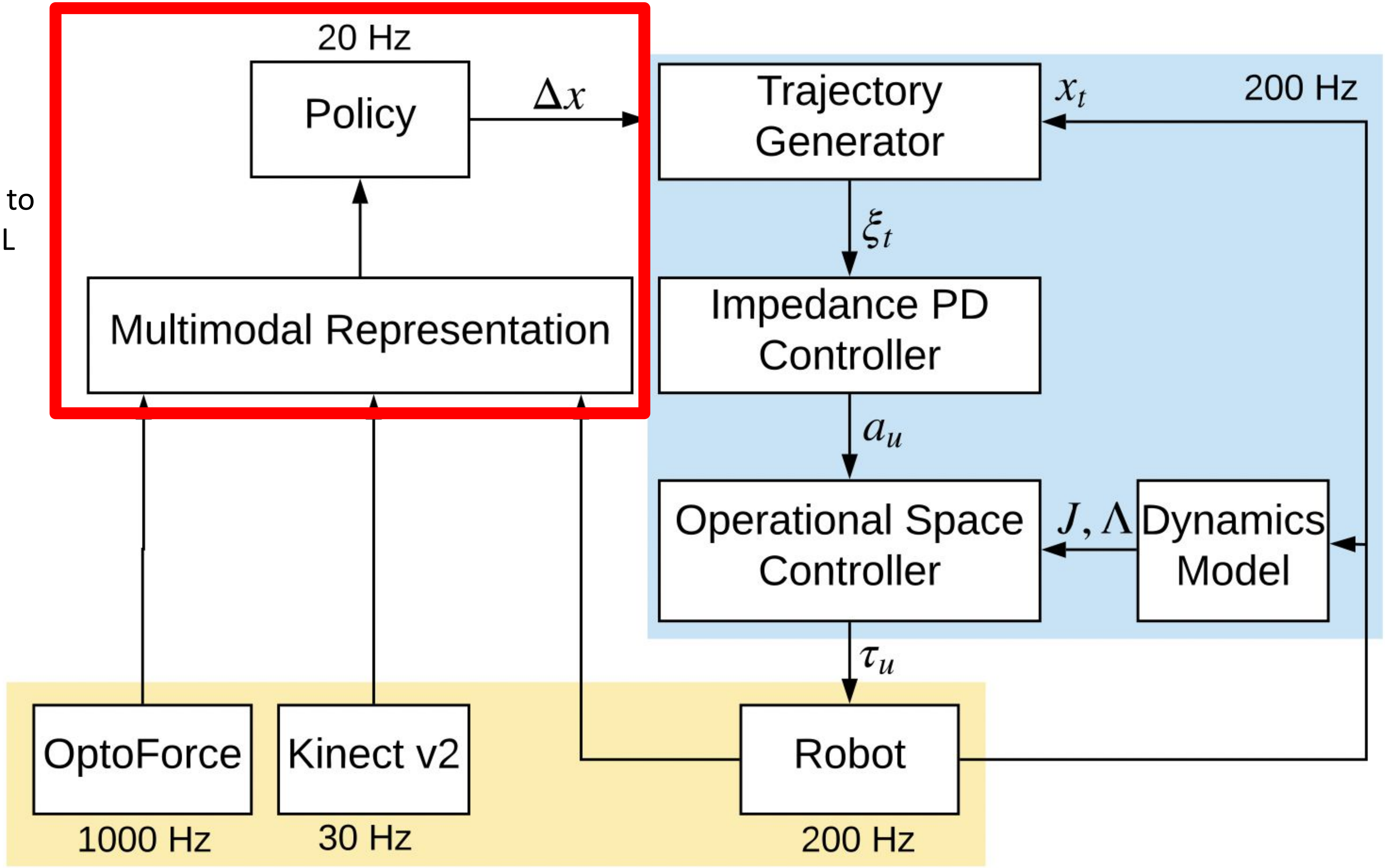


Robot provides sensory data input

M. A. Lee, Y. Zhu, K. Srinivasan, P. Shah, S. Savarese, L. Fei-Fei, A. Garg, and J. Bohg, "Making sense of vision and touch: Self-supervised learning of multimodal representations for contact-rich tasks," in 2019 International conference on robotics and automation (ICRA), pp. 8943–8950, IEEE, 2019.

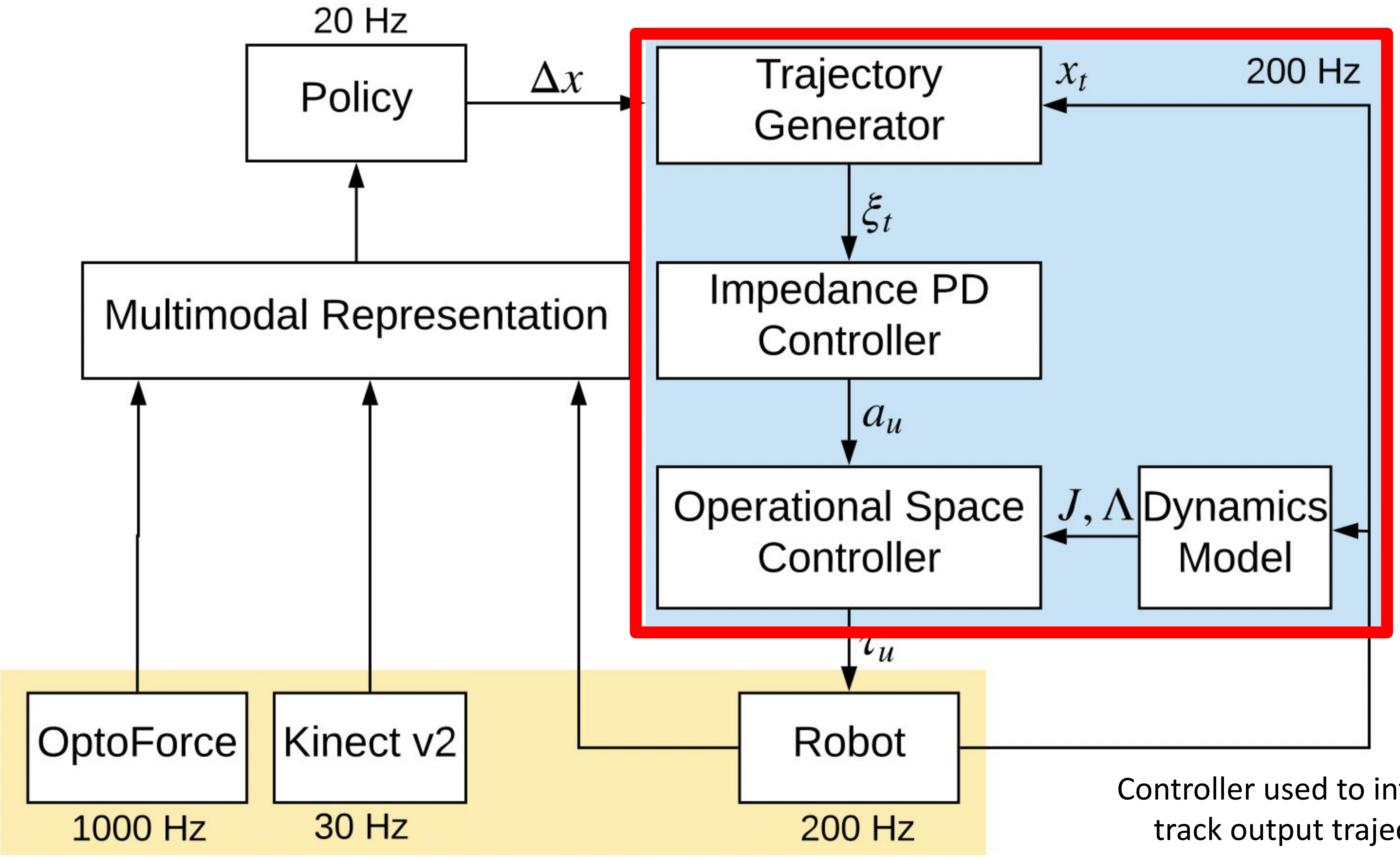
2. Use pretrained representation for policy learning

Pretrained representation used to train policy using RL



M. A. Lee, Y. Zhu, K. Srinivasan, P. Shah, S. Savarese, L. Fei-Fei, A. Garg, and J. Bohg, "Making sense of vision and touch: Self-supervised learning of multimodal representations for contact-rich tasks," in 2019 International conference on robotics and automation (ICRA), pp. 8943–8950, IEEE, 2019.

2. Use pretrained representation for policy learning



Controller used to interpolate and track output trajectory while remaining compliant

M. A. Lee, Y. Zhu, K. Srinivasan, P. Shah, S. Savarese, L. Fei-Fei, A. Garg, and J. Bohg, "Making sense of vision and touch: Self-supervised learning of multimodal representations for contact-rich tasks," in 2019 International conference on robotics and automation (ICRA), pp. 8943–8950, IEEE, 2019.

Visuo-Tactile Transformers for Manipulation

Yizhou Chen, Andrea Sipos, Mark Van der Merwe, Nima Fazeli

CoRL 2022

Focusing on contact-rich tasks

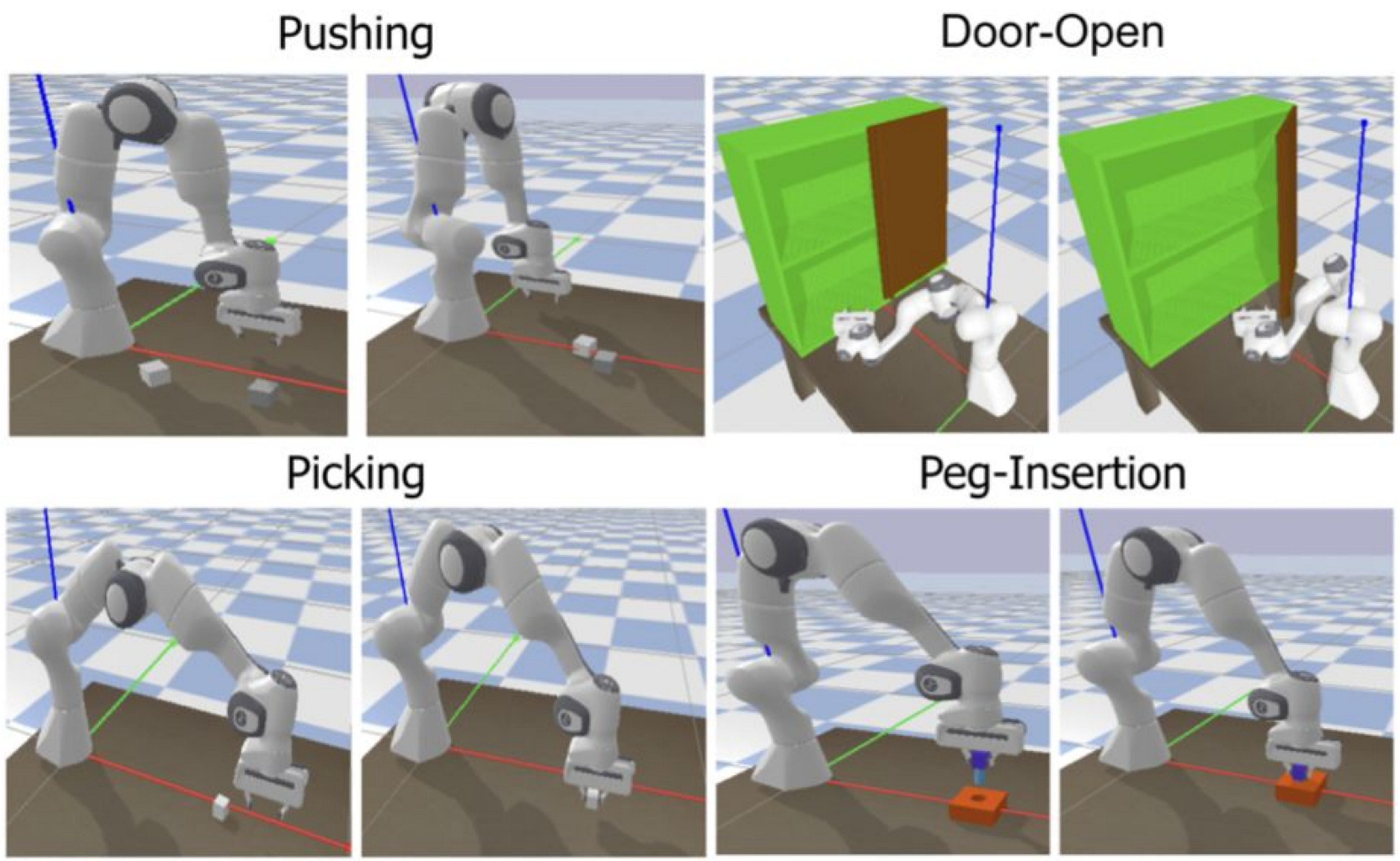
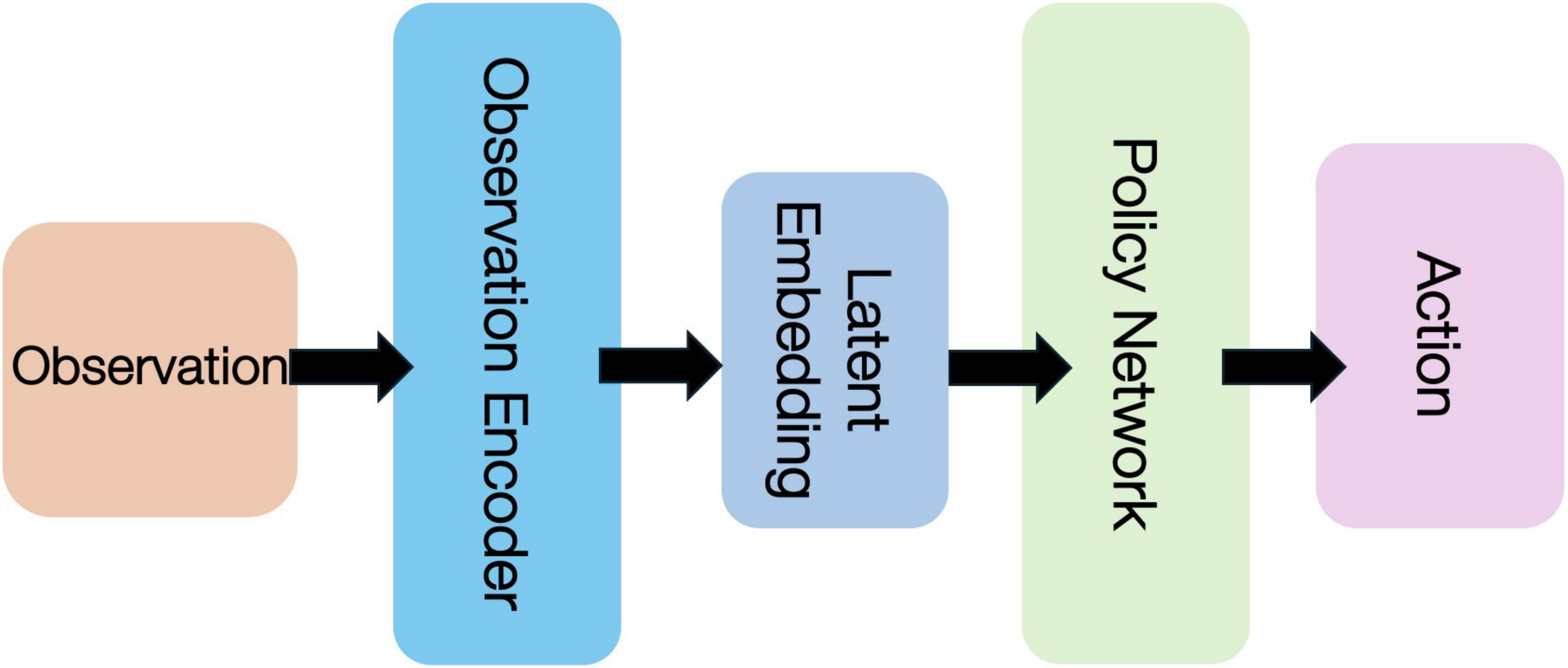


Figure 3: **Manipulation Tasks**: We evaluate VTT on four tasks in Pybullet. We vary visual and physical parameters in each task.

Y. Chen, M. Van der Merwe, A. Sipos, and N. Fazeli, "Visuo-tactile transformers for manipulation," in Conference on Robot Learning, pp. 2026–2040, PMLR, 2023.

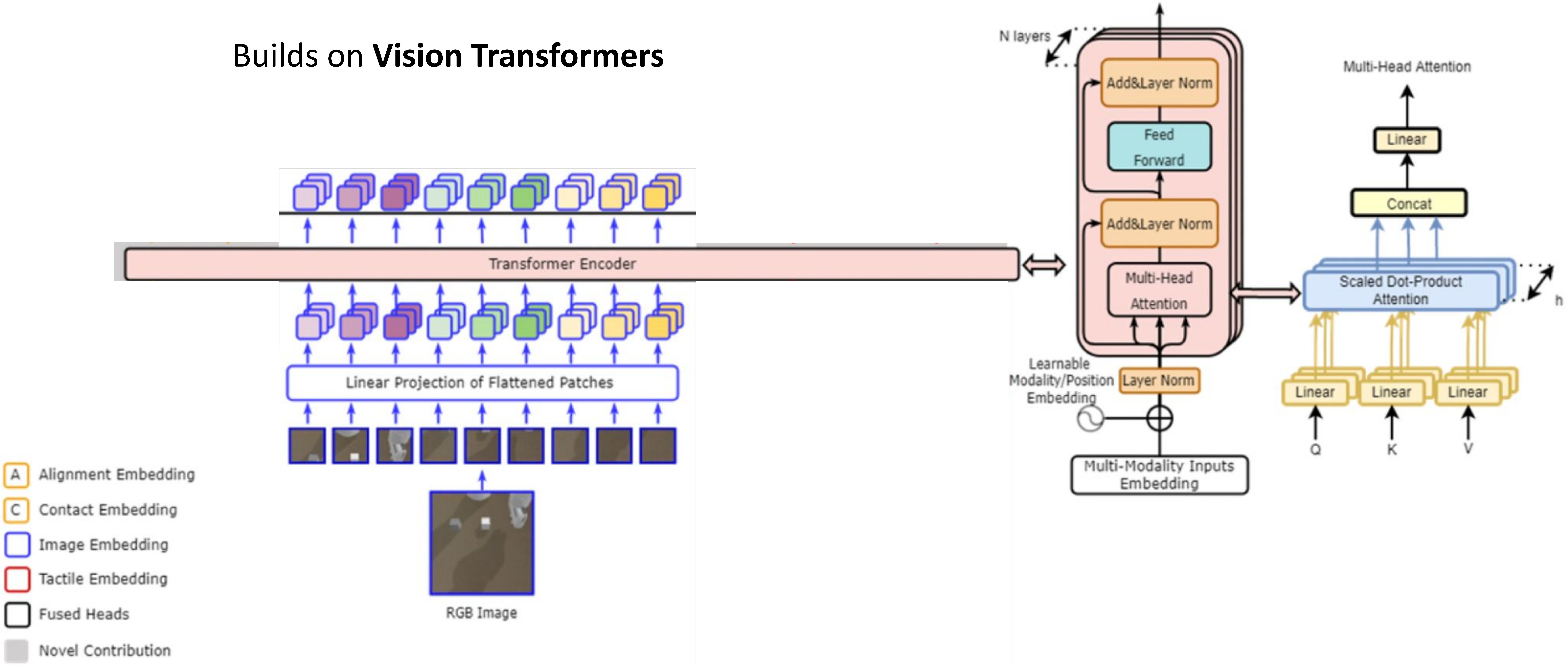
Same decoupling with a new encoder!



Y. Chen, M. Van der Merwe, A. Sipos, and N. Fazeli, “Visuo-tactile transformers for manipulation,” in Conference on Robot Learning, pp. 2026–2040, PMLR, 2023.

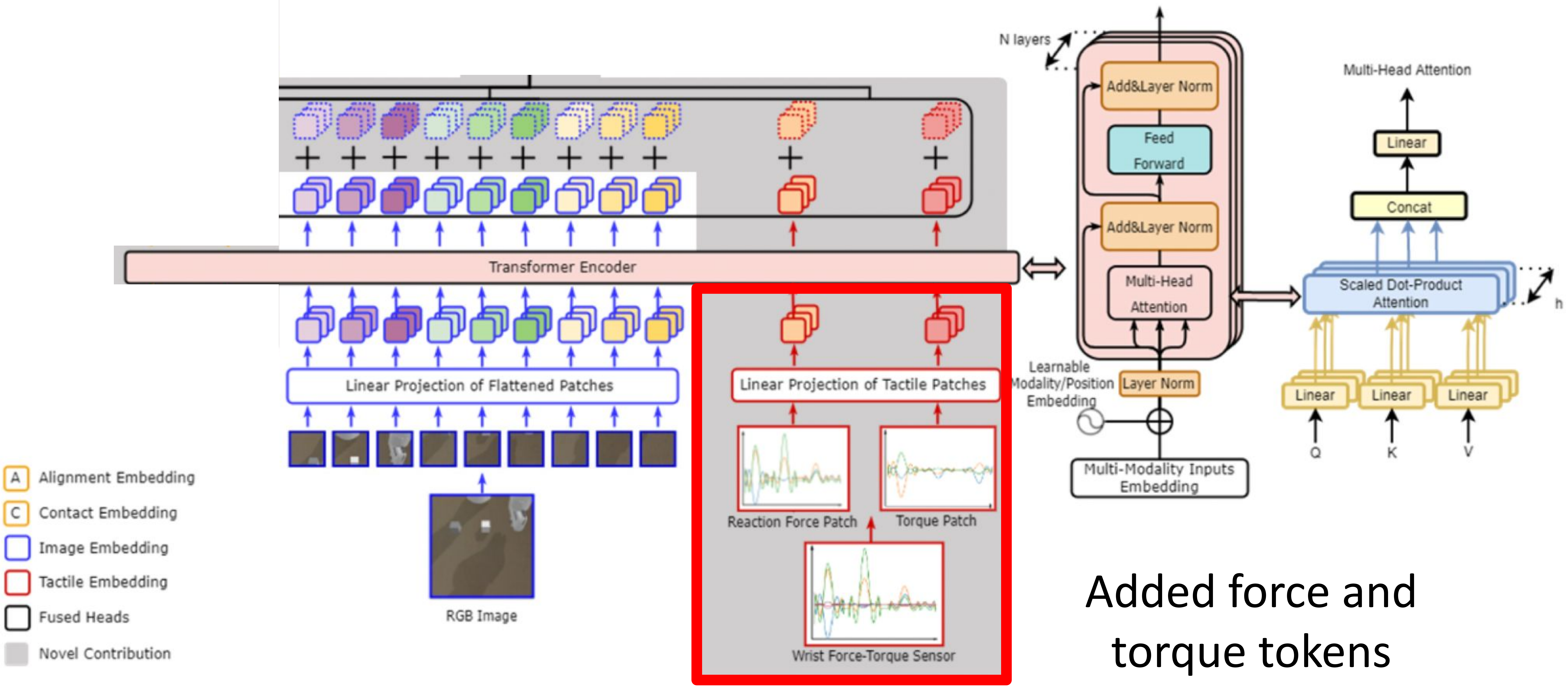
A shared (not separate) visuotactile encoder

Builds on **Vision Transformers**



Y. Chen, M. Van der Merwe, A. Sipos, and N. Fazeli, "Visuo-tactile transformers for manipulation," in Conference on Robot Learning, pp. 2026–2040, PMLR, 2023.

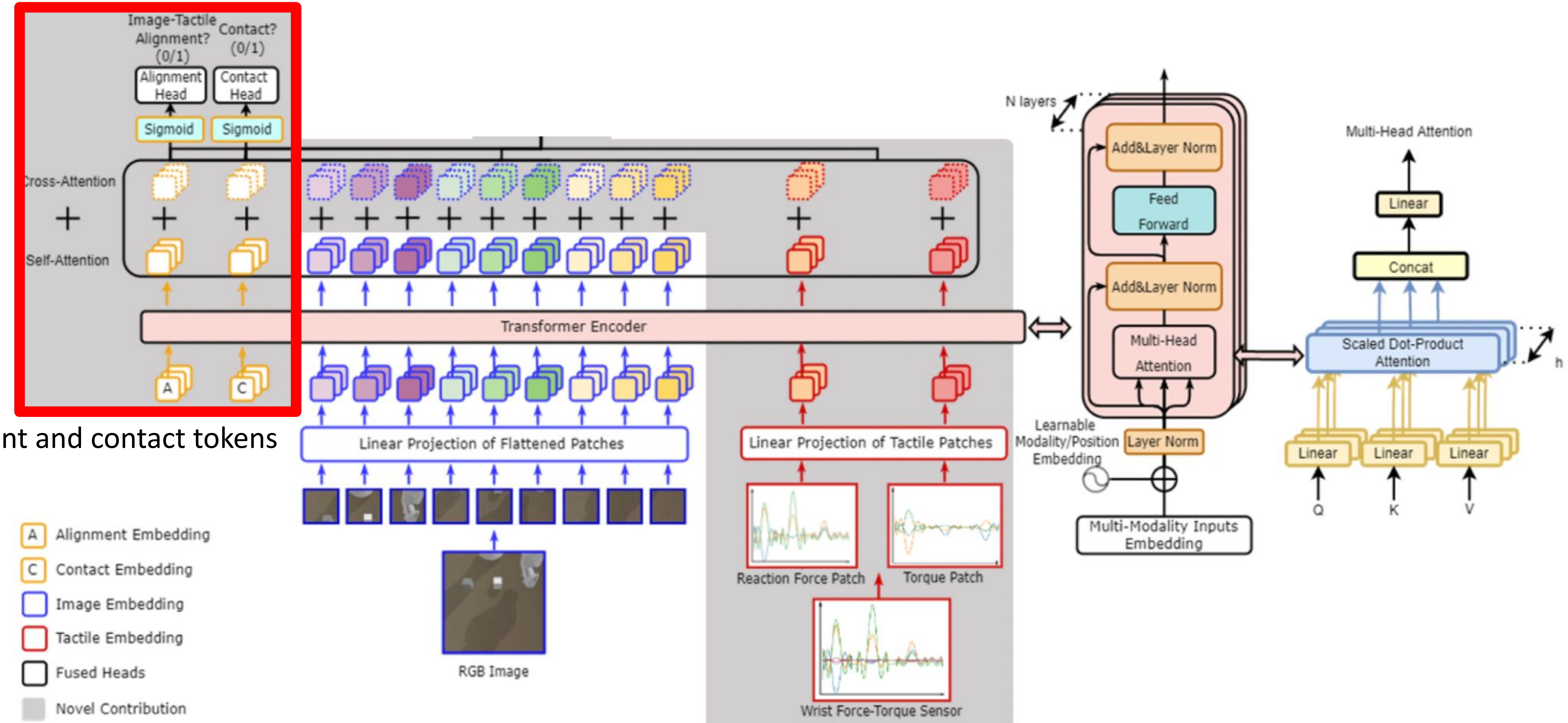
A shared (not separate) visuotactile encoder



Added force and torque tokens

Y. Chen, M. Van der Merwe, A. Sipos, and N. Fazeli, “Visuo-tactile transformers for manipulation,” in Conference on Robot Learning, pp. 2026–2040, PMLR, 2023.

A shared (not separate) visuotactile encoder

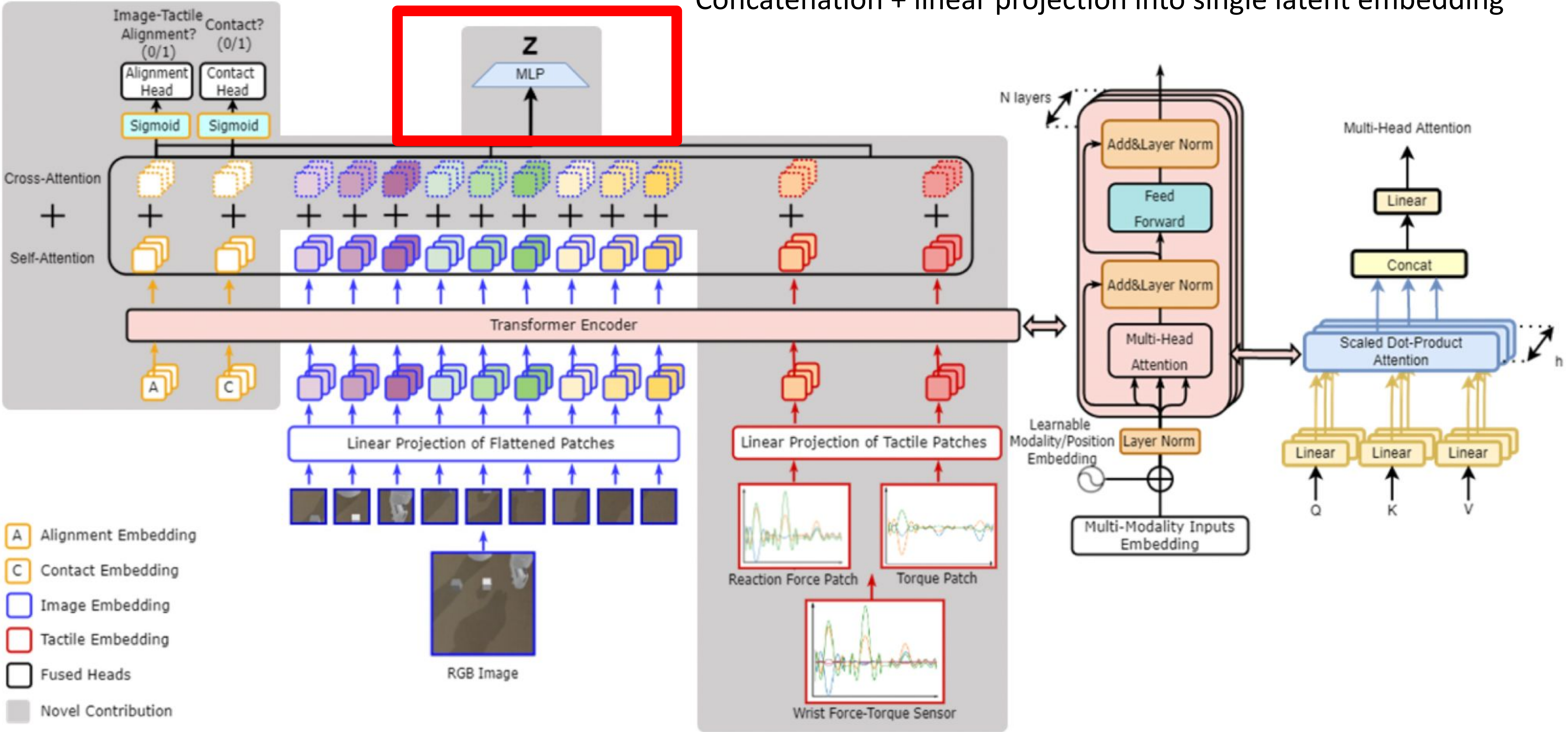


Alignment and contact tokens

Y. Chen, M. Van der Merwe, A. Sipos, and N. Fazeli, “Visuo-tactile transformers for manipulation,” in Conference on Robot Learning, pp. 2026–2040, PMLR, 2023.

A shared (not separate) visuotactile encoder

Concatenation + linear projection into single latent embedding



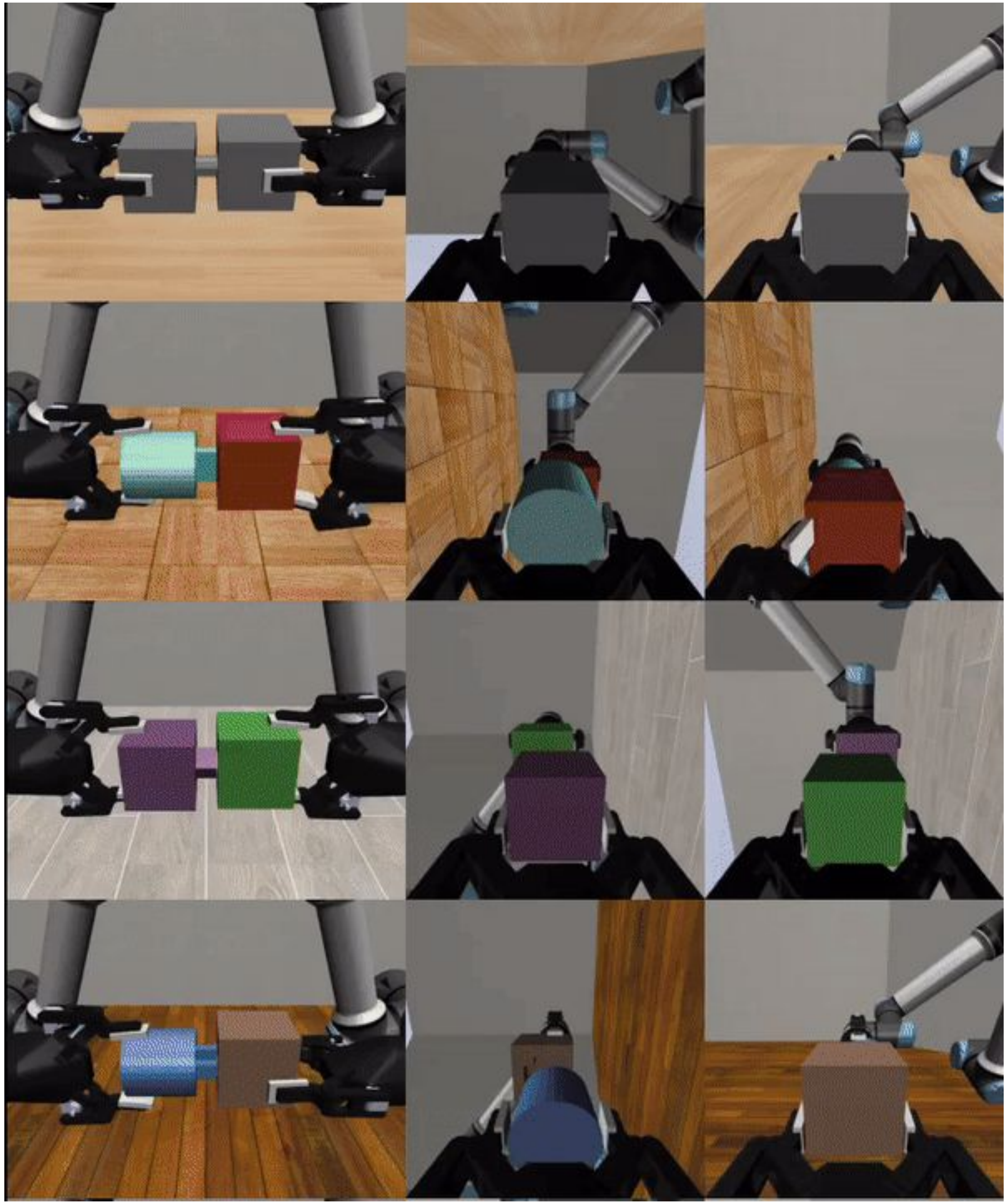
Y. Chen, M. Van der Merwe, A. Sipos, and N. Fazeli, "Visuo-tactile transformers for manipulation," in Conference on Robot Learning, pp. 2026–2040, PMLR, 2023.

AugInsert:
Learning Robust Visual-Force Policies via Data Augmentation for Object Assembly Tasks

Ryan Diaz, Adam Imdieke, Vivek Veeriah, Karthik Desingh

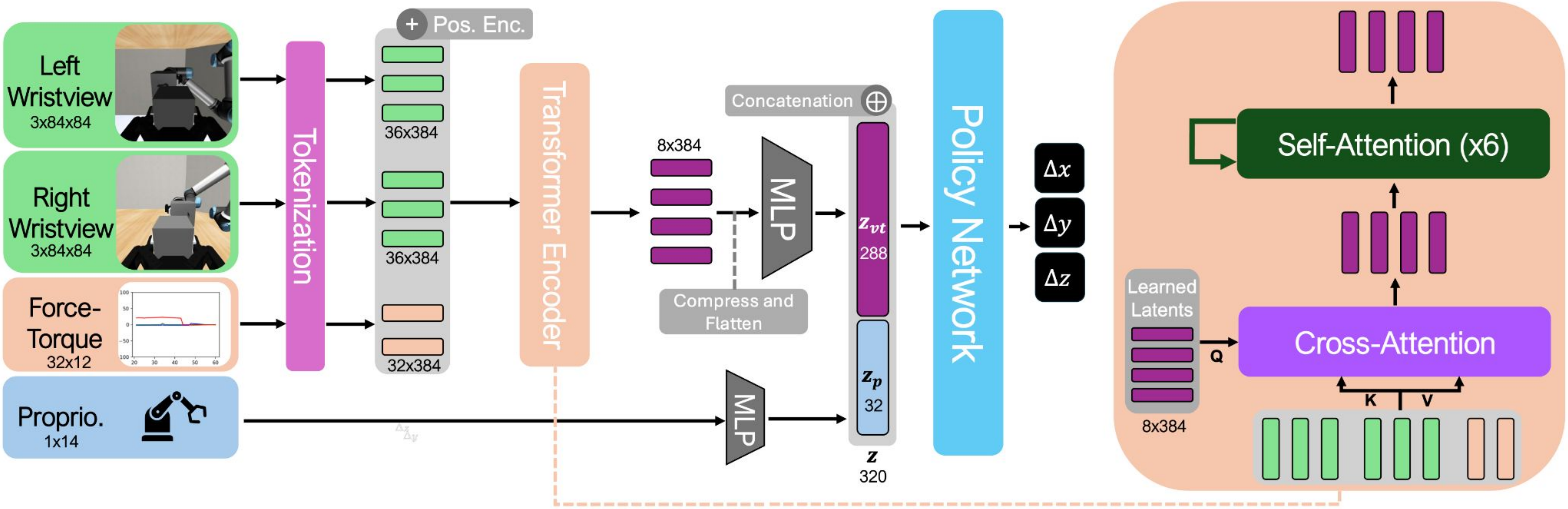
arXiv Preprint 2024 [Under Review]

Different flavors of the same task



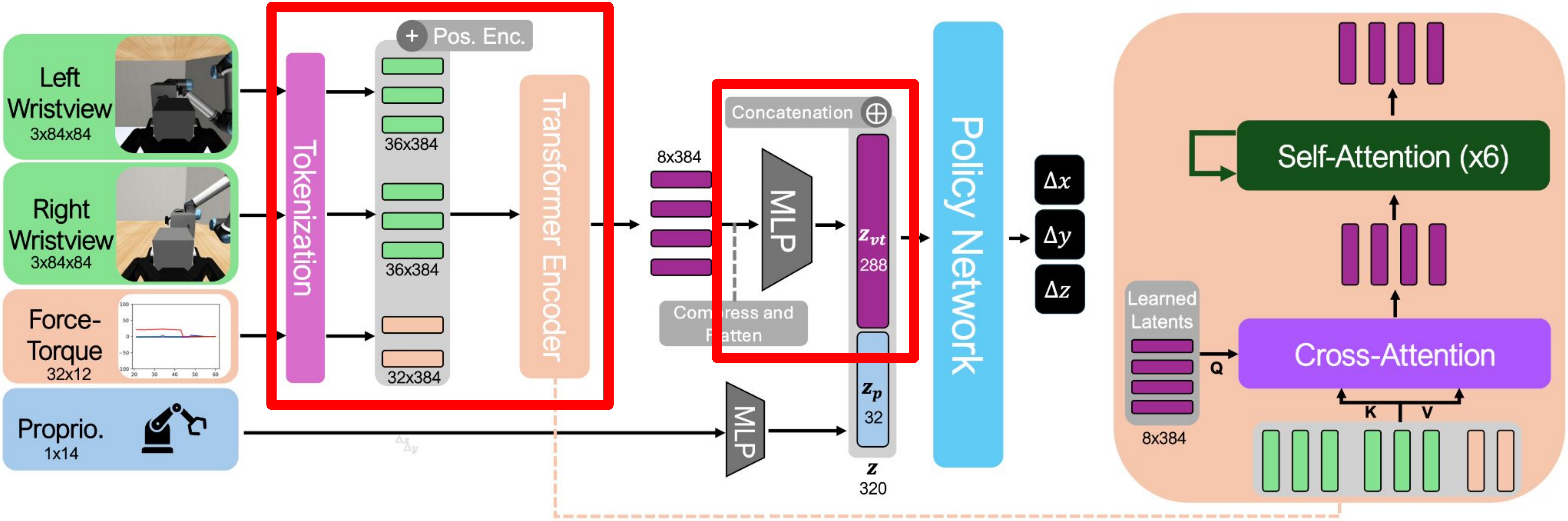
R. Diaz, A. Imdieke, V. Veeriah, and K. Desingh, “Auginsert: Learning robust visual-force policies via data augmentation for object assembly tasks,” arXiv preprint arXiv:2410.14968, 2024.

Multisensory Encoding Architecture



R. Diaz, A. Imdieke, V. Veeriah, and K. Desingh, "Auginsert: Learning robust visual-force policies via data augmentation for object assembly tasks," arXiv preprint arXiv:2410.14968, 2024.

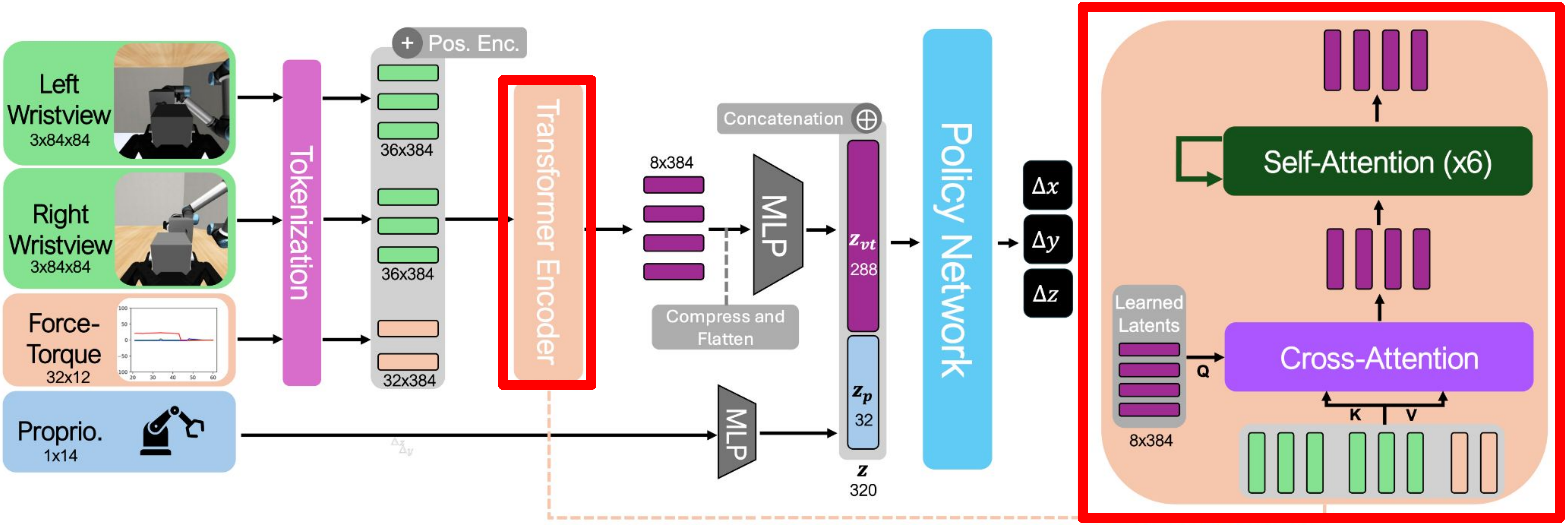
Multisensory Encoding Architecture



Tokenization and output representation inspired by Visuotactile Transformers

R. Diaz, A. Imdieke, V. Veeriah, and K. Desingh, "Auginsert: Learning robust visual-force policies via data augmentation for object assembly tasks," arXiv preprint arXiv:2410.14968, 2024.

Multisensory Encoding Architecture



Latent cross-attention inspired by Perceiver/PerceiverIO

R. Diaz, A. Imdieke, V. Veeriah, and K. Desingh, "Auginsert: Learning robust visual-force policies via data augmentation for object assembly tasks," arXiv preprint arXiv:2410.14968, 2024.

Hearing Touch:
Audio-Visual Pretraining for Contact-Rich Manipulation

Jared Mejia, Victoria Dean, Tess Hellebrekers, Abhinav Gupta

ICRA 2024

Motivation

Two key ingredients for improved manipulation: **pretraining on large datasets** and using **multisensory input** (with tactile data)

J. Mejia, V. Dean, T. Hellebrekers, and A. Gupta, "Hearing touch: Audio-visual pretraining for contact-rich manipulation," in 2024 IEEE International Conference on Robotics and Automation (ICRA), pp. 6912–6919, 2024.

Motivation

Two key ingredients for improved manipulation: **pretraining on large datasets** and using **multisensory input** (with tactile data)

How can we combine the two?

J. Mejia, V. Dean, T. Hellebrekers, and A. Gupta, “Hearing touch: Audio-visual pretraining for contact-rich manipulation,” in 2024 IEEE International Conference on Robotics and Automation (ICRA), pp. 6912–6919, 2024.

Motivation

Two key ingredients for improved manipulation: **pretraining on large datasets** and using **multisensory input** (with tactile data)

How can we combine the two?

Large internet-scale image datasets exist, but not much for tactile

J. Mejia, V. Dean, T. Hellebrekers, and A. Gupta, “Hearing touch: Audio-visual pretraining for contact-rich manipulation,” in 2024 IEEE International Conference on Robotics and Automation (ICRA), pp. 6912–6919, 2024.

Motivation

Two key ingredients for improved manipulation: **pretraining on large datasets** and using **multisensory input** (with tactile data)

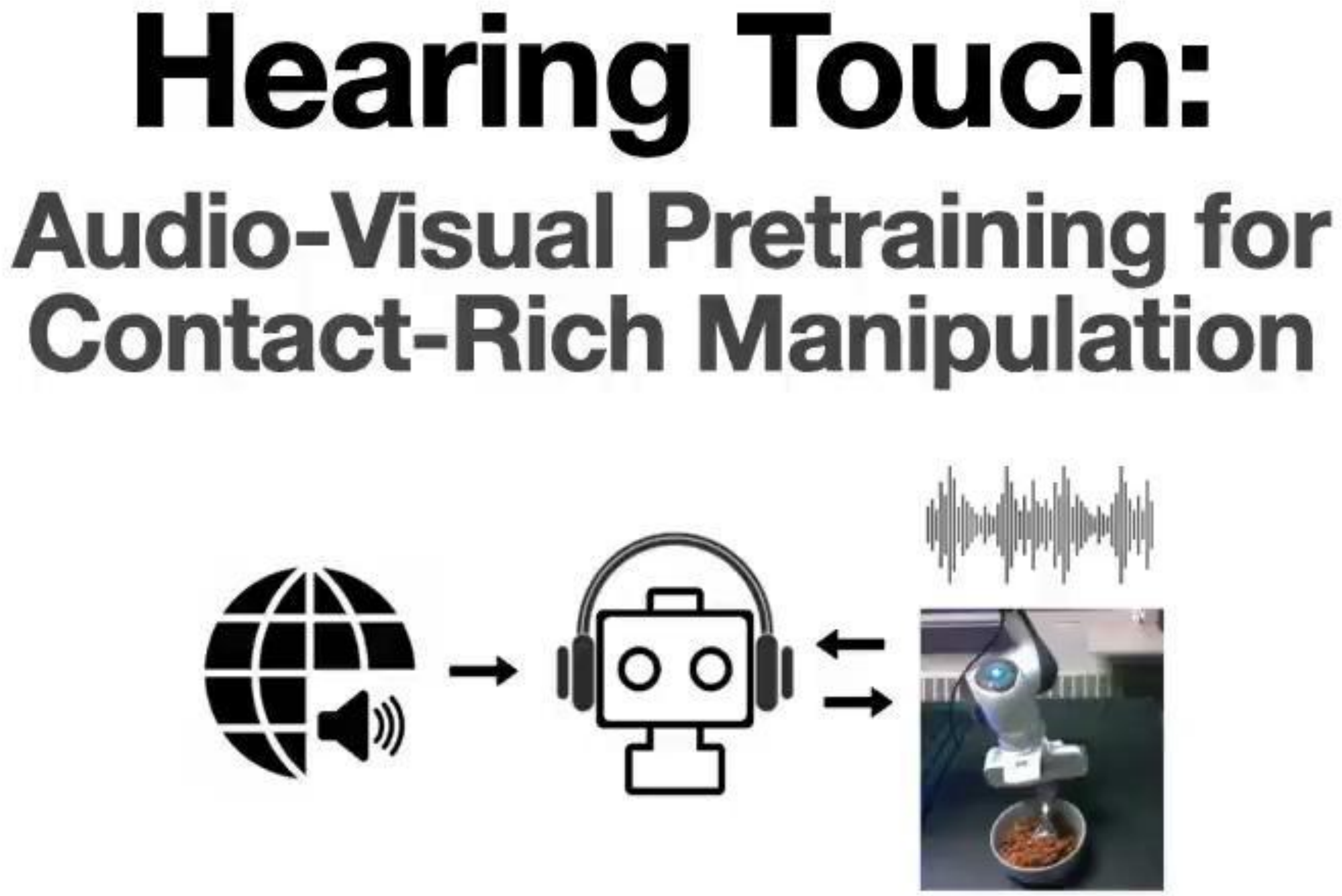
How can we combine the two?

Large internet-scale image datasets exist, but not much for tactile

Idea: Leverage *contact audio* as a tactile sensing mode to enable the use of internet-scale audio datasets for pretraining

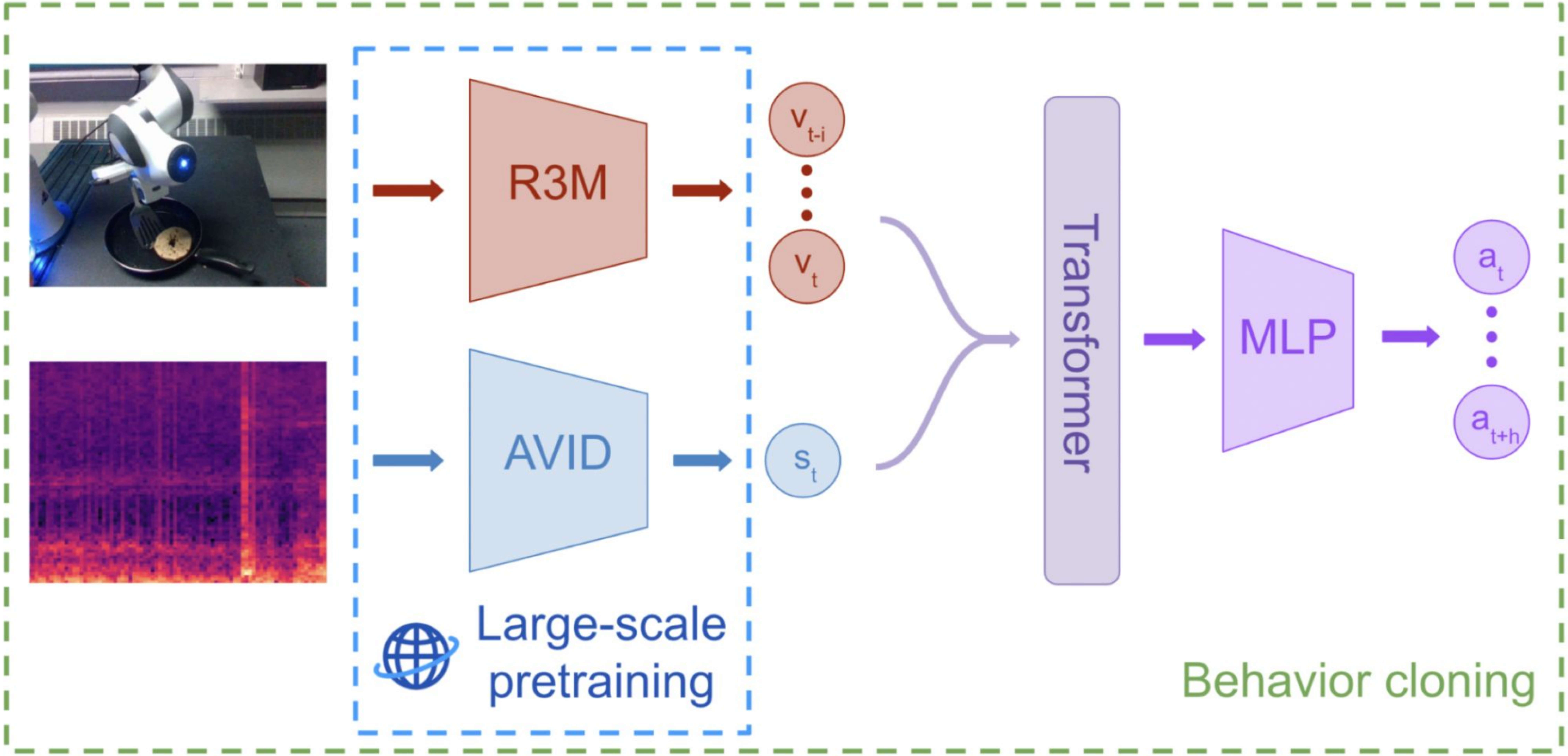
J. Mejia, V. Dean, T. Hellebrekers, and A. Gupta, "Hearing touch: Audio-visual pretraining for contact-rich manipulation," in 2024 IEEE International Conference on Robotics and Automation (ICRA), pp. 6912–6919, 2024.

Sound in Robotic Manipulation



J. Mejia, V. Dean, T. Hellebrekers, and A. Gupta, "Hearing touch: Audio-visual pretraining for contact-rich manipulation," in 2024 IEEE International Conference on Robotics and Automation (ICRA), pp. 6912–6919, 2024.

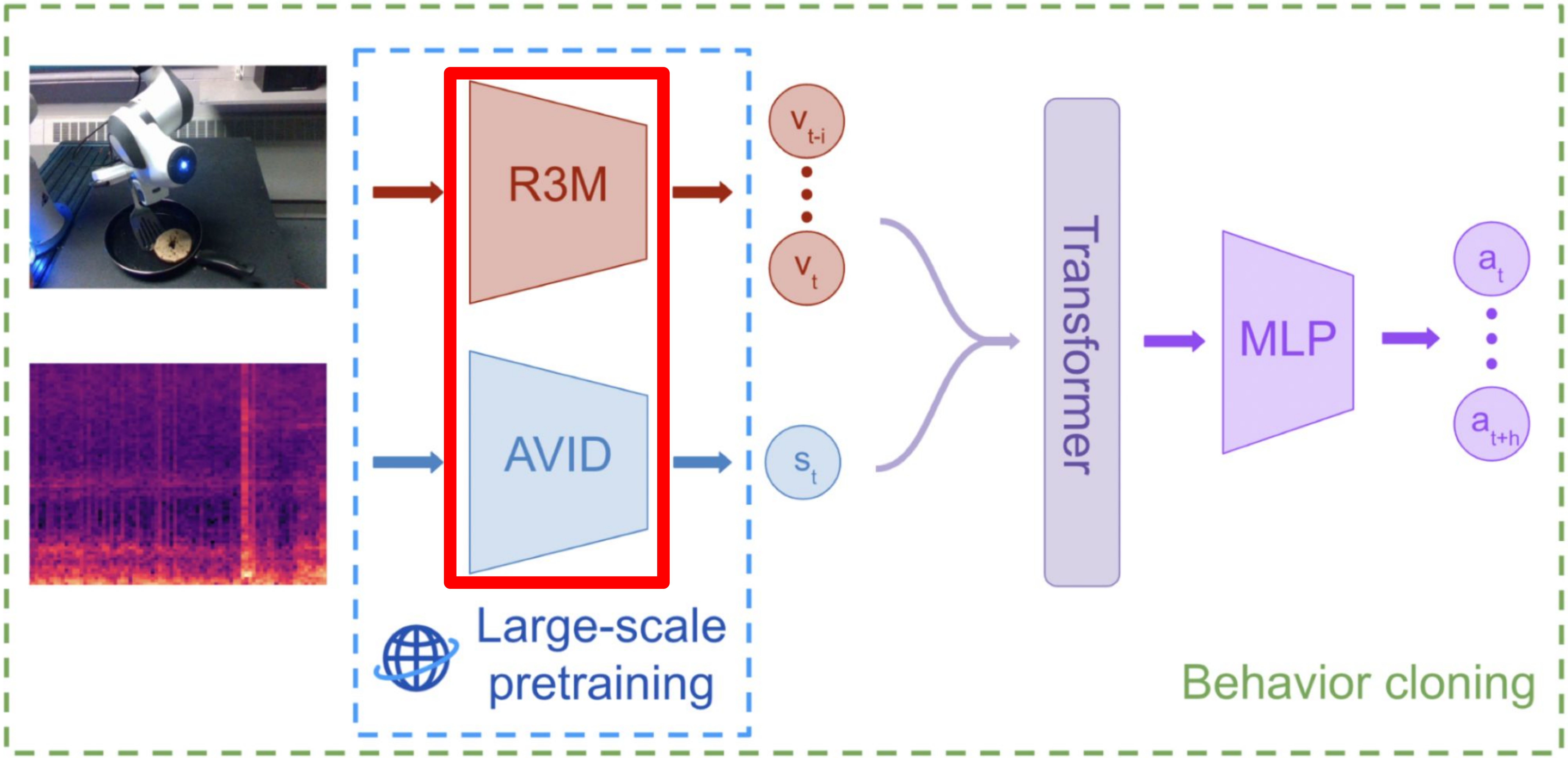
Framework



J. Mejia, V. Dean, T. Hellebrekers, and A. Gupta, "Hearing touch: Audio-visual pretraining for contact-rich manipulation," in 2024 IEEE International Conference on Robotics and Automation (ICRA), pp. 6912–6919, 2024.

Framework

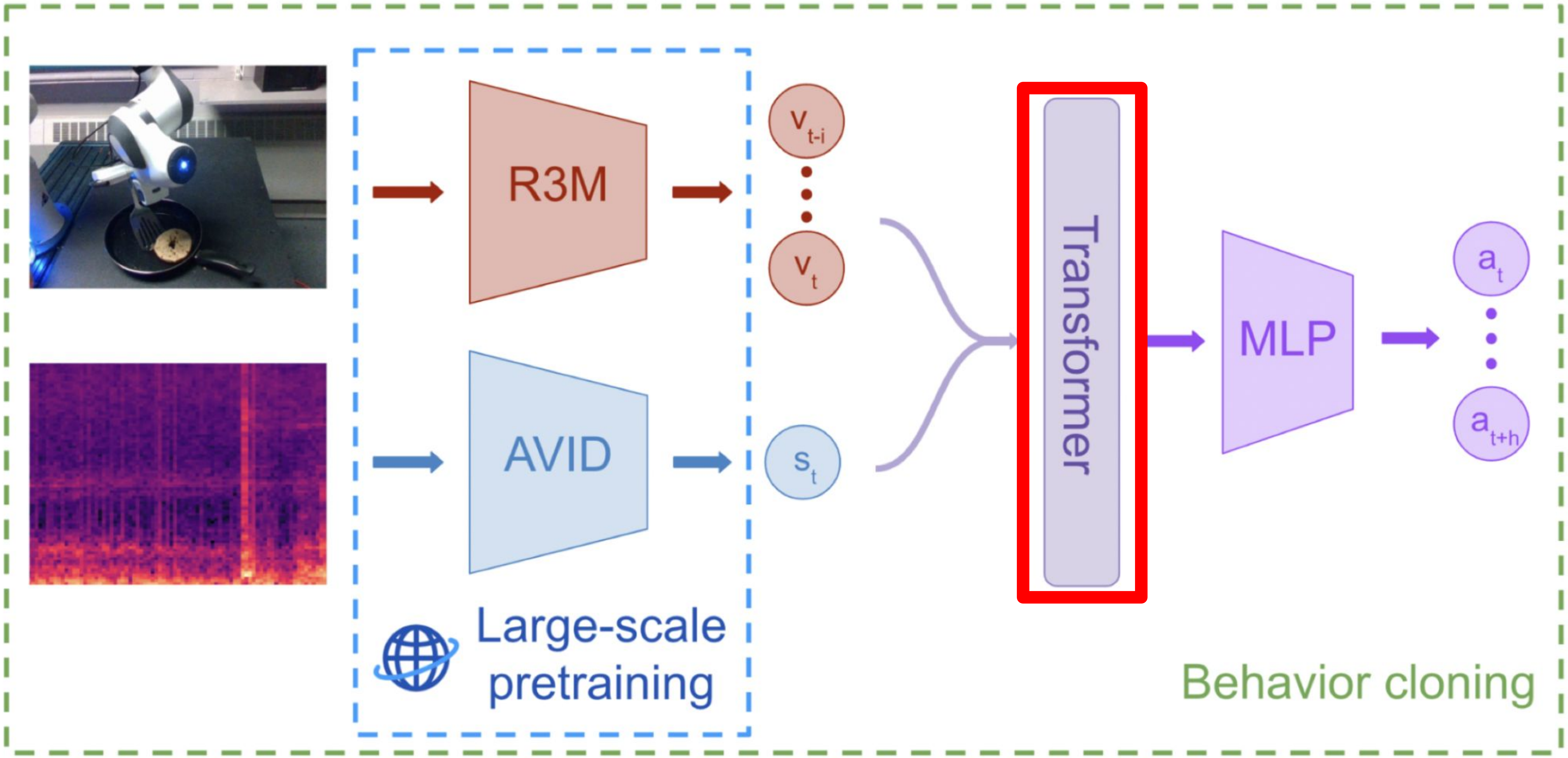
R3M: Trained on **Ego4D** (>3670 hours of egocentric human video)
AVID: Trained on **Audioset** (>2 million 10-sec audio clips from YouTube)



J. Mejia, V. Dean, T. Hellebrekers, and A. Gupta, "Hearing touch: Audio-visual pretraining for contact-rich manipulation," in 2024 IEEE International Conference on Robotics and Automation (ICRA), pp. 6912–6919, 2024.

Framework

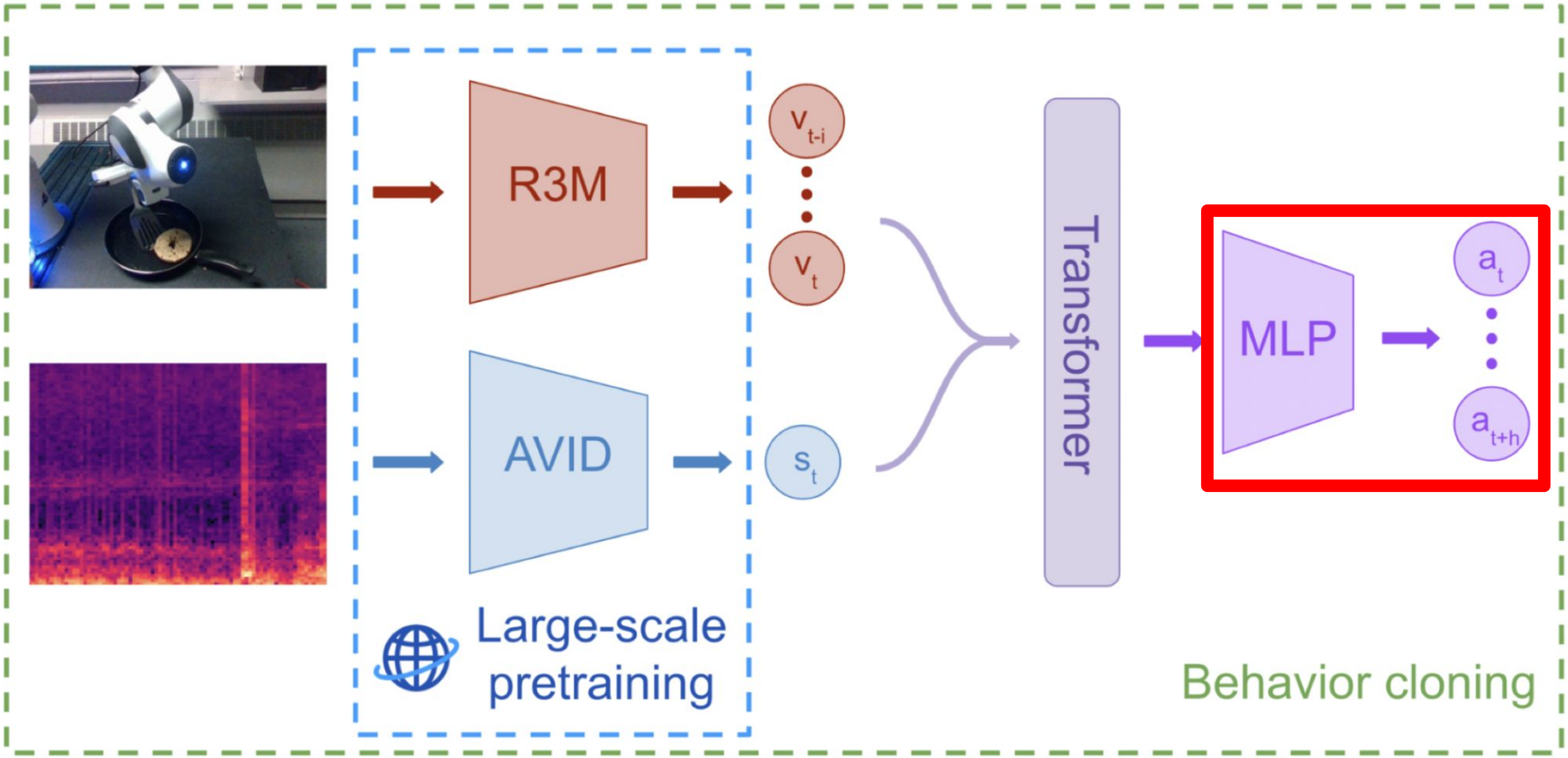
Multisensory transformer for modality fusion



J. Mejia, V. Dean, T. Hellebrekers, and A. Gupta, "Hearing touch: Audio-visual pretraining for contact-rich manipulation," in 2024 IEEE International Conference on Robotics and Automation (ICRA), pp. 6912–6919, 2024.

Framework

MLP policy network trained via behavior cloning



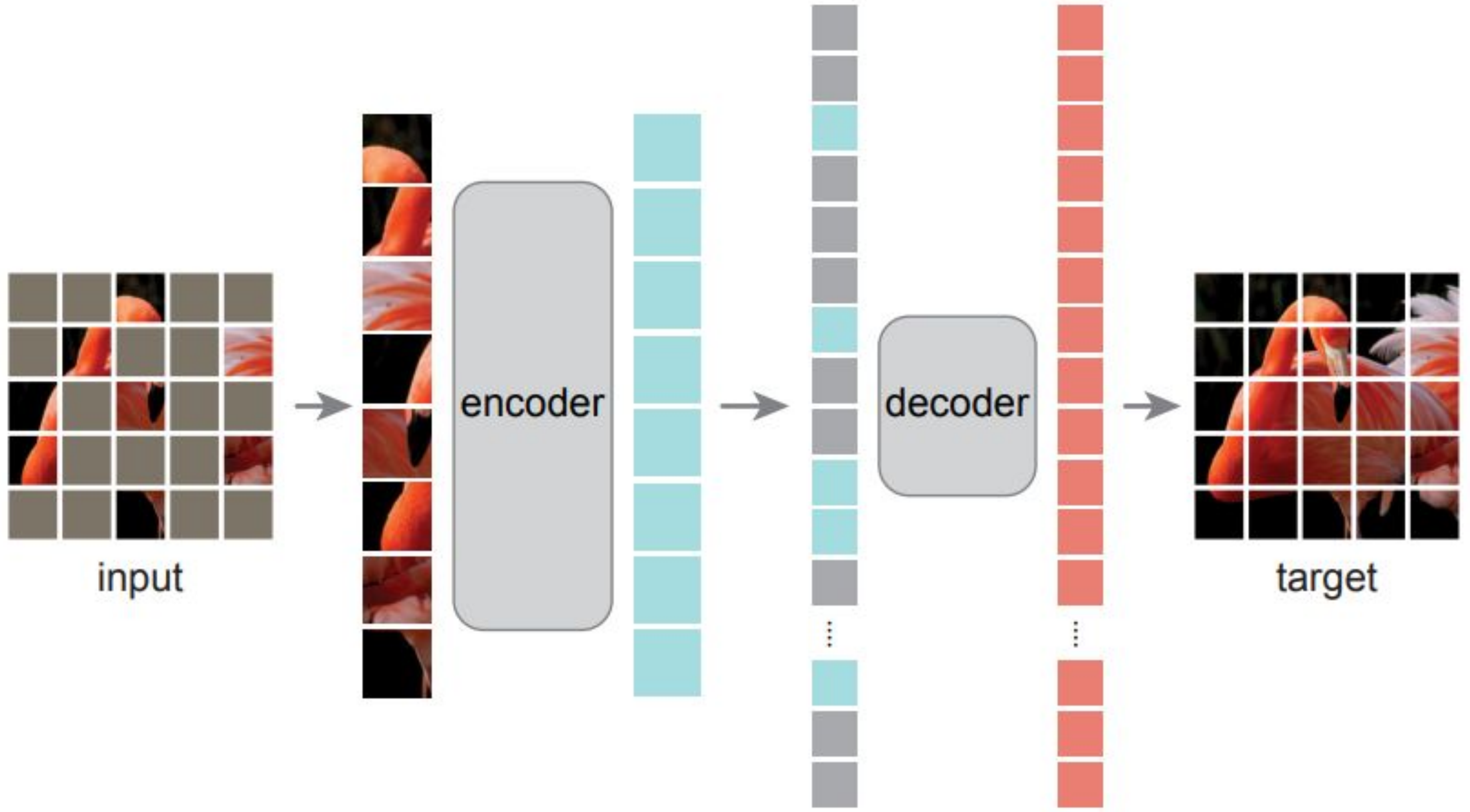
J. Mejia, V. Dean, T. Hellebrekers, and A. Gupta, “Hearing touch: Audio-visual pretraining for contact-rich manipulation,” in 2024 IEEE International Conference on Robotics and Automation (ICRA), pp. 6912–6919, 2024.

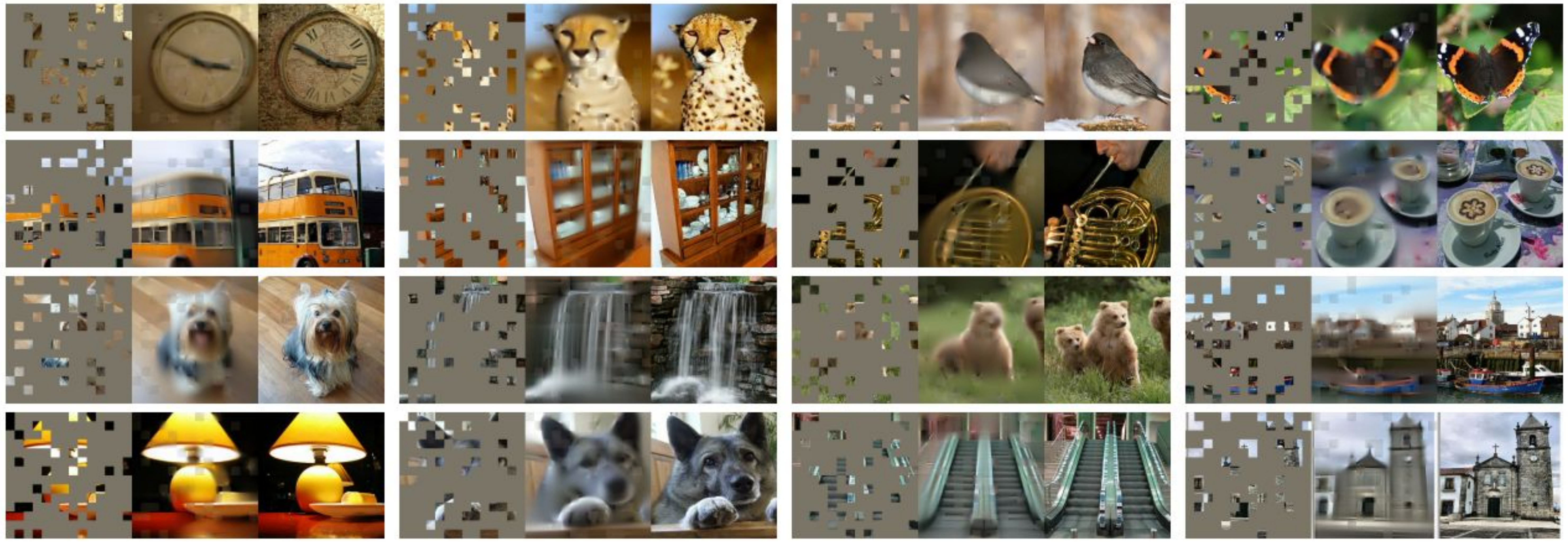
MultiMAE:
Multi-modal Multi-task Masked Autoencoders

Roman Bachmann, David Mizrahi, Andrei Atanov, Amir Zamir

ECCV 2022

Masked autoencoder



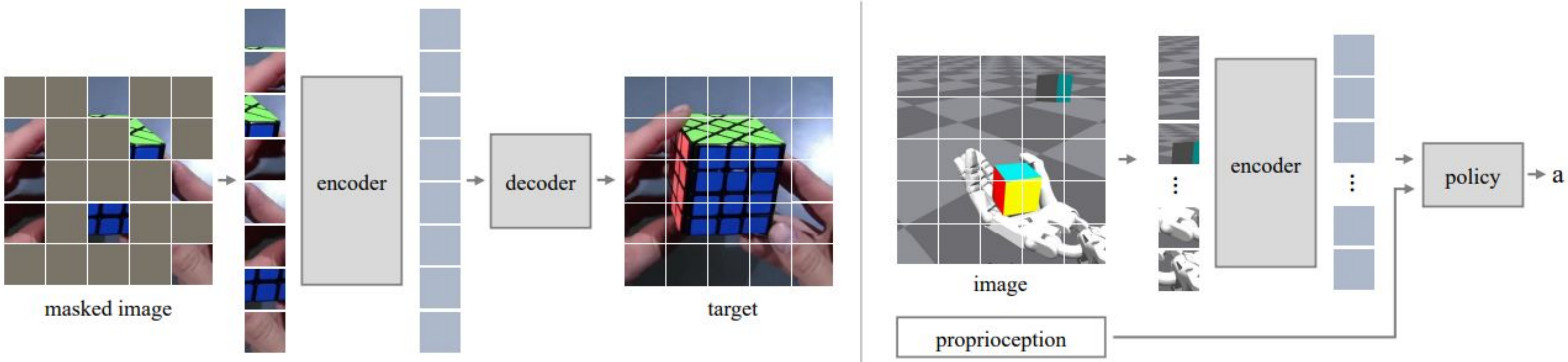


motivation:
1. image is heavily redundant
2. computationally efficient

implementation:
1. mask rate: 75%
2. a light weight decoder
3. loss only calculate on masked patches

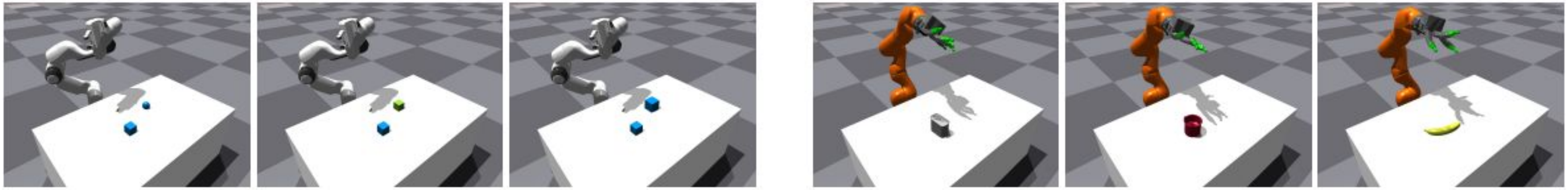
He, Kaiming, et al. "Masked autoencoders are scalable vision learners." *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2022.

Masked Visual Pre-training for Motor Control



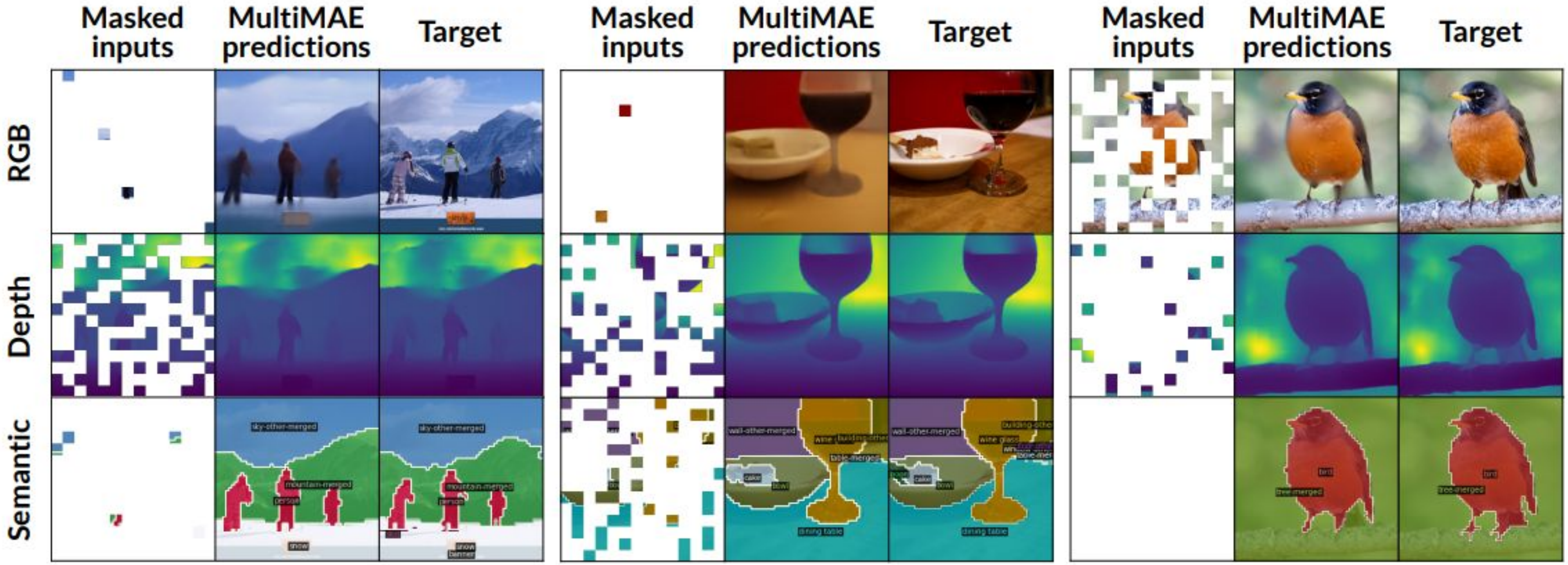
(a) masked visual pretraining

(b) learning motor control



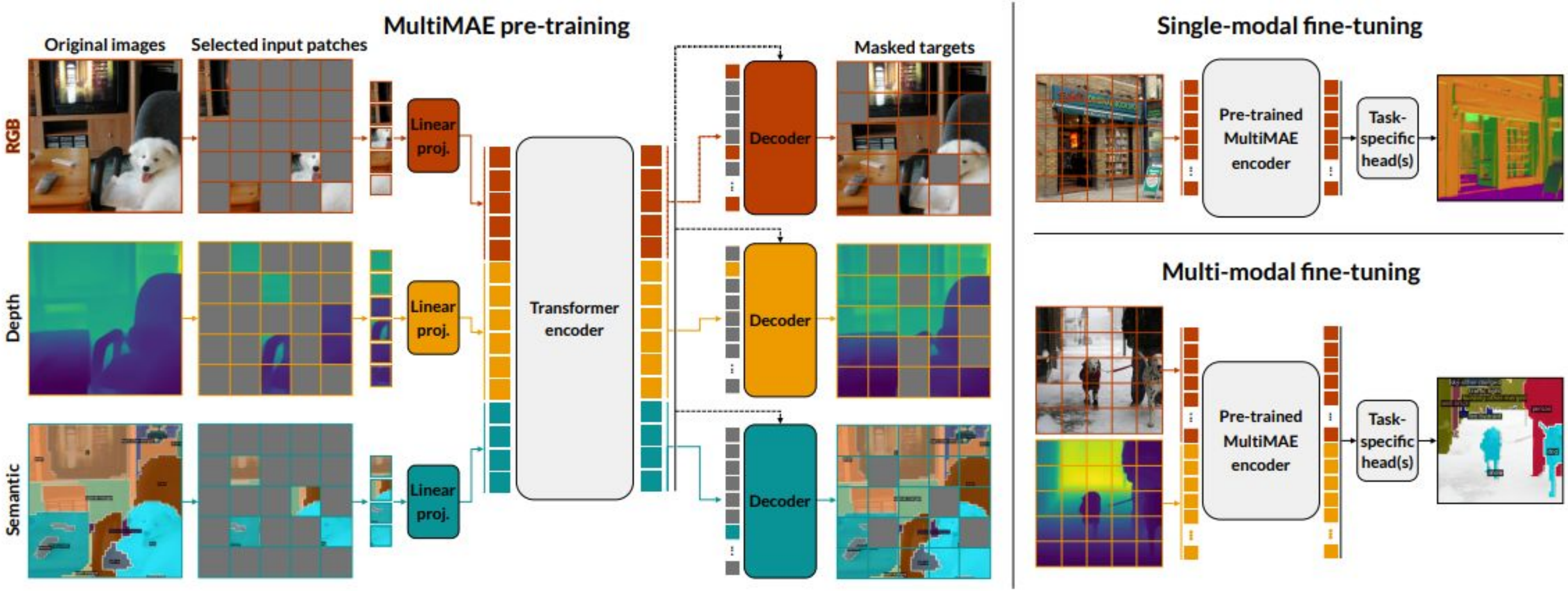
Xiao, Tete, et al. "Masked visual pre-training for motor control." *arXiv preprint arXiv:2203.06173* (2022).

MultiMAE pre-training objective



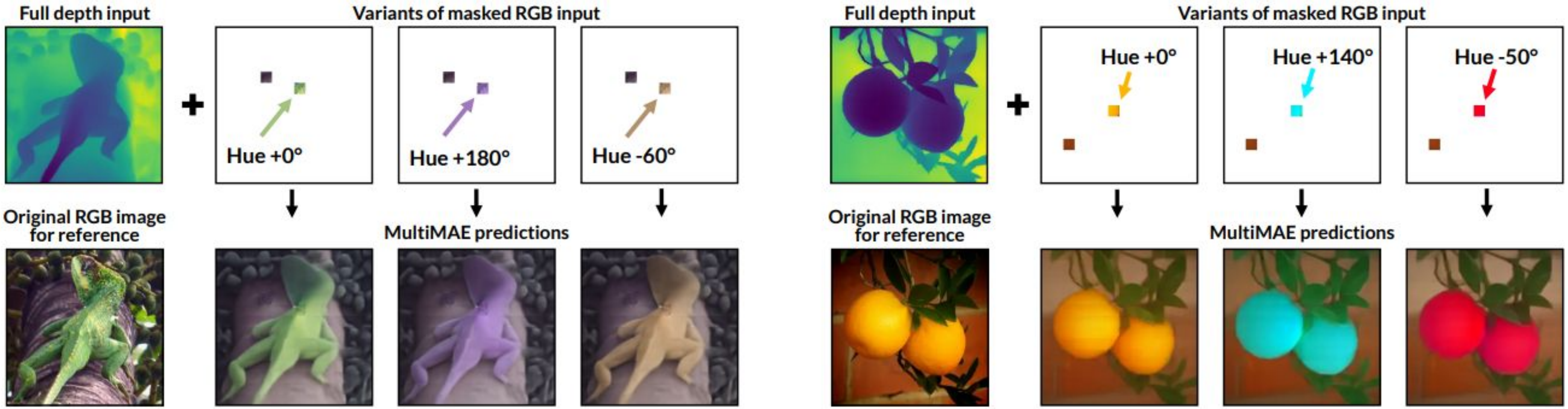
Bachmann, Roman, et al. "Multimae: Multi-modal multi-task masked autoencoders." *European Conference on Computer Vision*. Cham: Springer Nature Switzerland, 2022.

MultiMAE pre-training



Bachmann, Roman, et al. "Multimae: Multi-modal multi-task masked autoencoders." *European Conference on Computer Vision*. Cham: Springer Nature Switzerland, 2022.

Demonstration of cross-modal interaction



Bachmann, Roman, et al. "Multimae: Multi-modal multi-task masked autoencoders." *European Conference on Computer Vision*. Cham: Springer Nature Switzerland, 2022.

Thank you!

Next Lecture:
Student Lecture 6
Diffusion Models and Policy Learning

DR

DeepRob

[Group 5] Lecture 5
Multisensory Learning + Manipulation

by Mason Hawver, Ryan Diaz, and Hanchen Cui
University of Minnesota

Image Source: D. F. Gomes, P. Paoletti, and S. Luo, “Generation of gelsight tactile images for sim2real learning,” IEEE Robotics and Automation Letters, vol. 6, no. 2, pp. 4177–4184, 2021.