Seminar: Self-Supervised Learning, SS 2020

Abhinav Valada

Friday, June 15, 2020

Procedure

- Students should select three papers out of the list in preference order.
- Places will be assigned based on priority suggestions of HisInOne and motivation of the student by May 24, 2020.
- Students are requested to prepare a 20 minutes talk, write an abstract and a summary.
- The Seminar will be held as a virtual ”Blockseminar” in the last week of July.

Procedure

- The details of the presentation and the slides should be discussed with the supervisor two weeks before the presentation.
- The abstract should be two pages long and is due June 29, 2020.
- The summary is due two weeks after the presentation and should be even pages long at maximum (latex, a4wide, 11pt) not including the bibliography and figures. Significantly longer summaries will not be accepted.
- The final grade is based on the oral presentation, the written abstract, the summary, and participation in the blockseminar.

Classical Supervised Methods

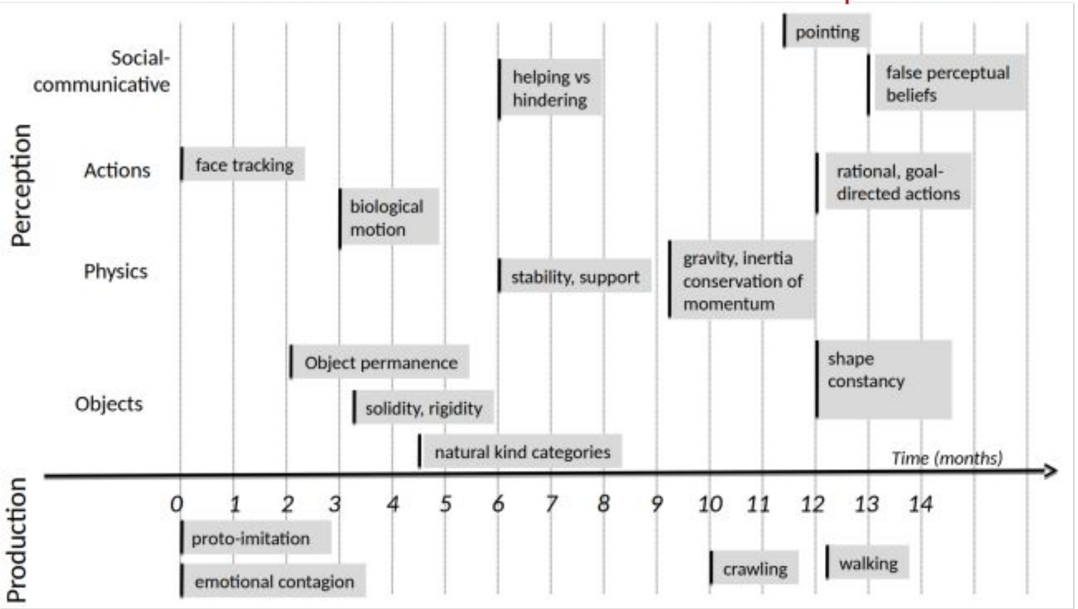
- Classical supervised learning methods strongly rely on large labeled datasets for training.
- Typical pipeline consists of collecting a dataset -> annotating the data -> selecting an architecture and an objective function -> training the model
- Pretraining such models on large datasets such as ImageNet has alleviated this limitation.
- Using layers from a pre-trained ImageNet model can have an important impact on the speed of training, and accuracy

Classical Supervised Methods

- In some cases, e.g. different types of data (robotics, medical imaging), pretraining in ImageNet may not lead to a large improvement.
- We would need to collect and annotate a huge amount of data.
- Creating datasets with sufficient annotated examples is a challenging task, and the process of labeling data is often arduous, expensive, and sometimes even infeasible.

Learning in Humans

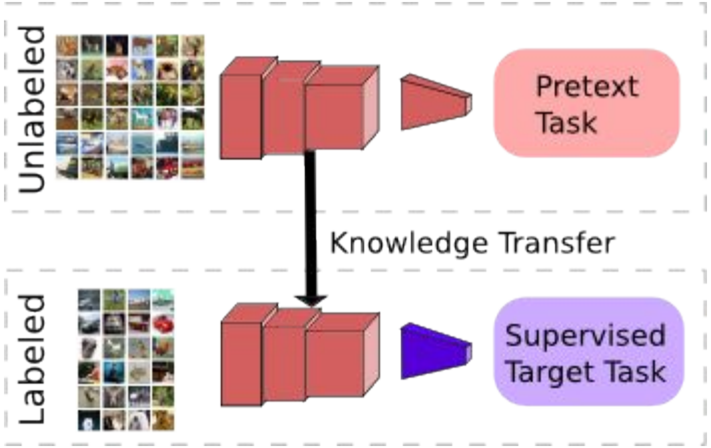
Observation Self-supervision → Language



Emmanuel Dupoux, "Cognitive Science in the Era of Artificial Intelligence", *Cognition*, vol. 173, pp. 43-59, 2018.

Self-supervised learning

- Self-supervised learning has emerged as an alternative to mitigate this problem
- In SS, we first learn a pretext task exploiting some property of the data
- We then use the learned semantically rich representations for fine-tuning on the target task.
- Advantages: Reduces or even eliminates the cost of labelling and exploit abundant unlabelled data

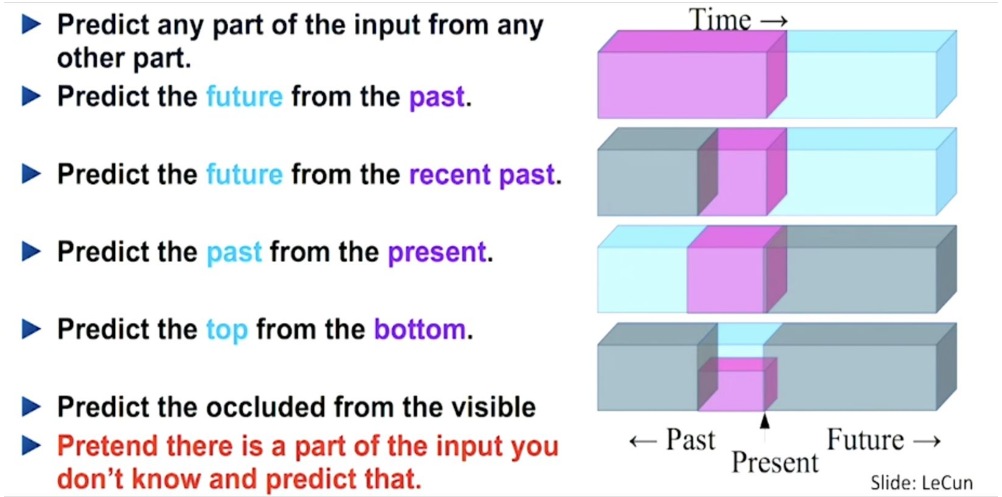


Self-supervised learning

- SS enables learning feature representation from unlabelled data by:
  - Exploiting dependencies within the data itself.
  - Generating labels based on those dependencies.
  - Solving a proxy objective on the generated labels.
- A portion of the input is used as a supervisory signal.
- Self-supervised learning allows using large amount of unlabelled data such as text, images, and audios on the Internet.
- The model learns meaningful patterns of the data when solving the pretext task.
- Then we can transfer the learned representation to solve the target task, also called the downstream task.

Self-supervised learning: Natural Language Processing

- Self-supervised learning is widely used in natural language processing.
- A pretext task example is to predict the next word of a sentence.
- To complete the pretext task, the model learns the nature of language.
- Then the pretext learned model is used to solve more complex supervised target tasks, such as sentiment analysis.



- Predict any part of the input from any other part.
- Predict the future from the past.
- Predict the future from the recent past.
- Predict the past from the present.
- Predict the top from the bottom.
- Predict the occluded from the visible
- Pretend there is a part of the input you don't know and predict that.

Slide: LeCun

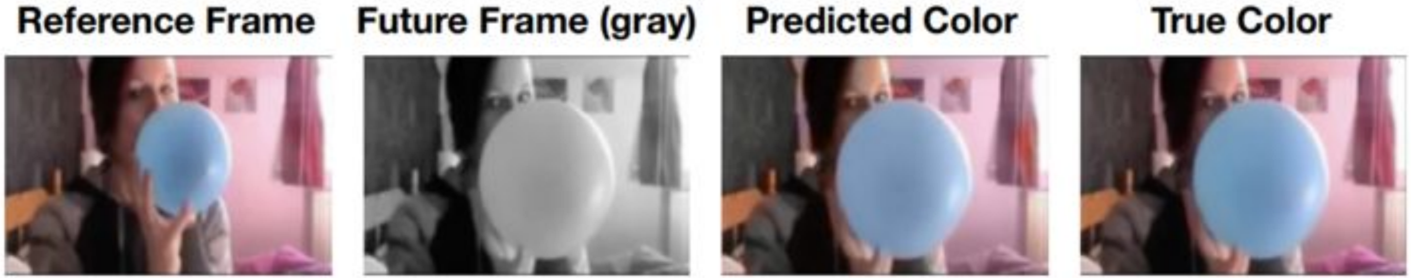
Source: Yann LeCun @EPFL - "Self-supervised learning: could machines learn like humans?"

Self-supervised learning: Colorization

Individual image Colorization: a model is trained to color a grayscale input image.

Video Colorization: the task is to copy colors from a normal reference frame in color to another target frame in grayscale by leveraging the natural temporal coherency of colors across video frames.

Solving the pretext task, the model learns to keep track of correlated pixels in different frames.

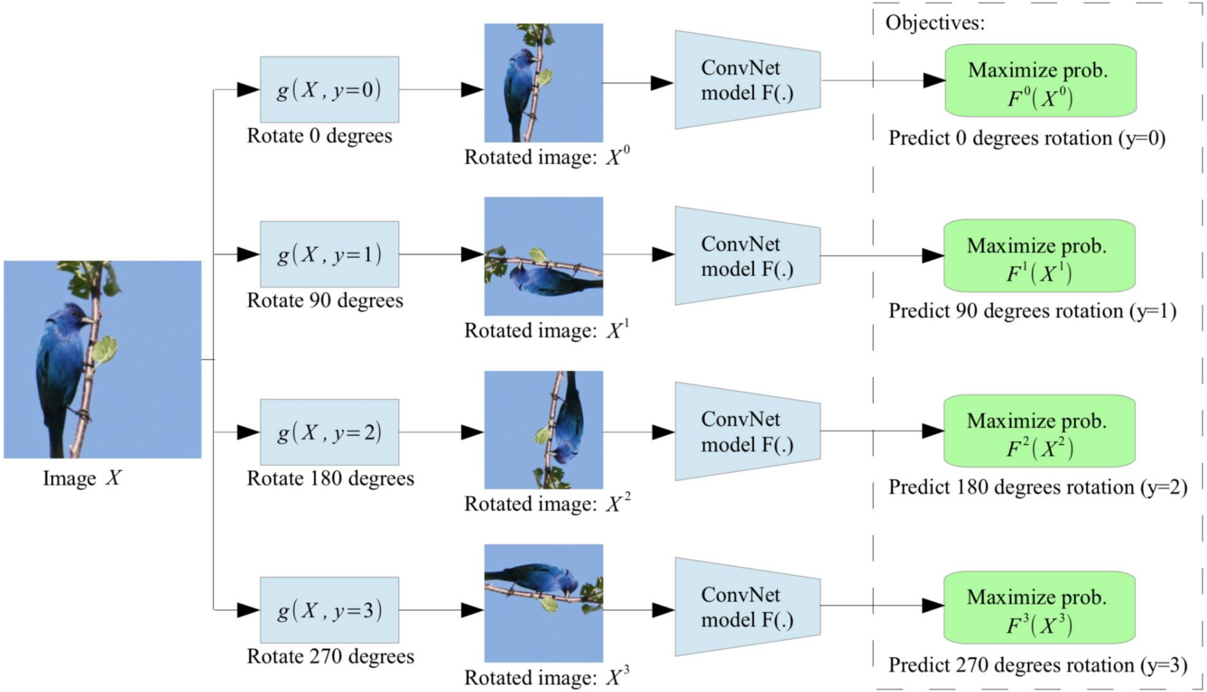


Source: Vondrick et al. 2018

Self-supervised learning: Distortion

Rotation: Each input image is first rotated.

The pretext task to predict which rotation has been applied.



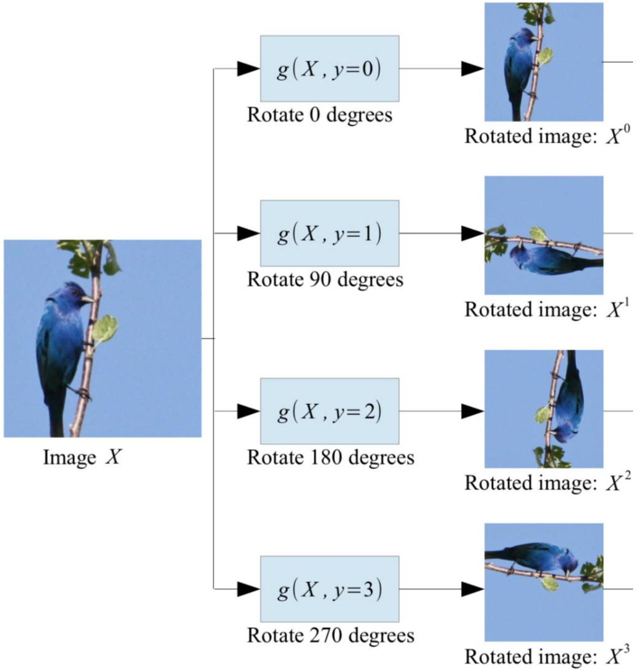
Source: Gidaris et al. 2018

Self-supervised learning: Distortion

Identify the same image with different rotations

The model learns to recognize high level object parts, such as heads, noses, and eyes, and the relative positions of these parts, rather than local patterns.

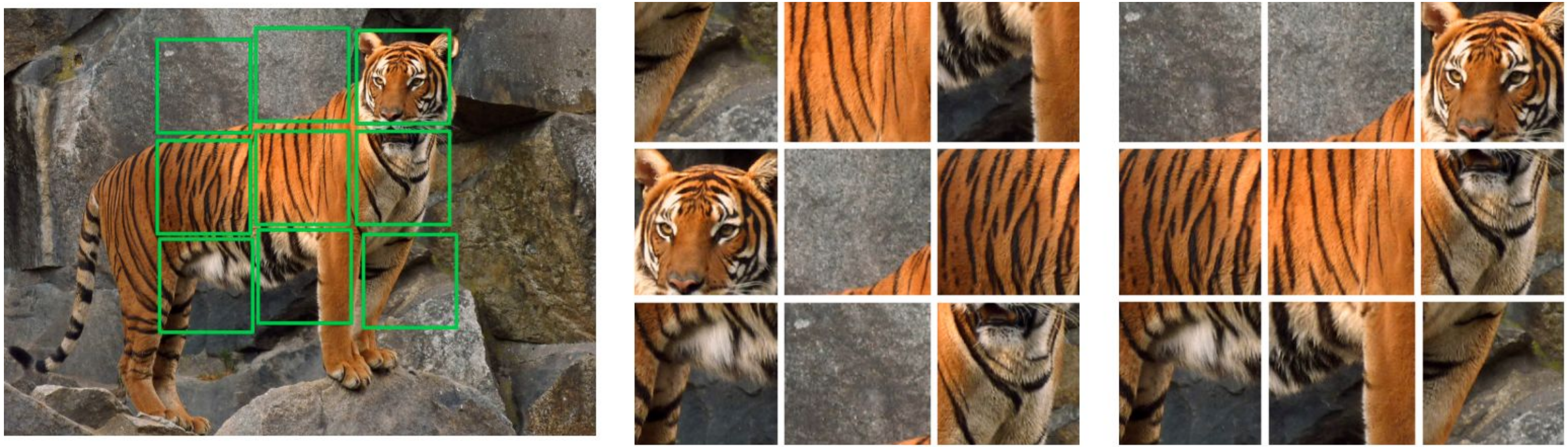
This pretext task drives the model to **learn semantic concepts of objects**.



Source: Gidaris et al. 2018

Self-supervised learning: Jigsaw

The model is trained to place shuffled patches back to the original locations. Solving Jigsaw puzzles can be used to teach a system that an object is made of parts and what these parts are.



Source: Noroozi et al. 2017

Self-supervised learning: Jigsaw

Two separate instances within the same categories have similar features (shape). However, some low-level features are different (color and texture).

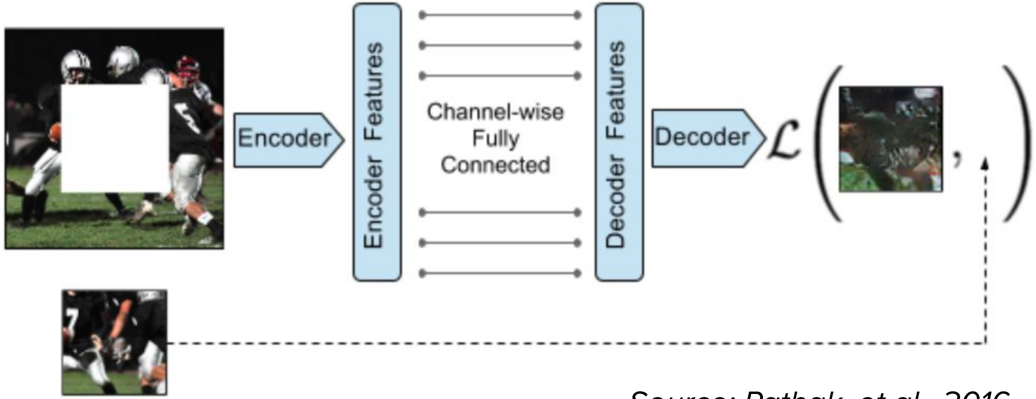
The Jigsaw puzzle solver learns to ignore such features when they do not help the localization of parts.

Source: Noroozi et al. 2017

Self-supervised learning: Inpainting

Humans are able to understand this structure and make visual predictions even when seeing only parts of the scene.

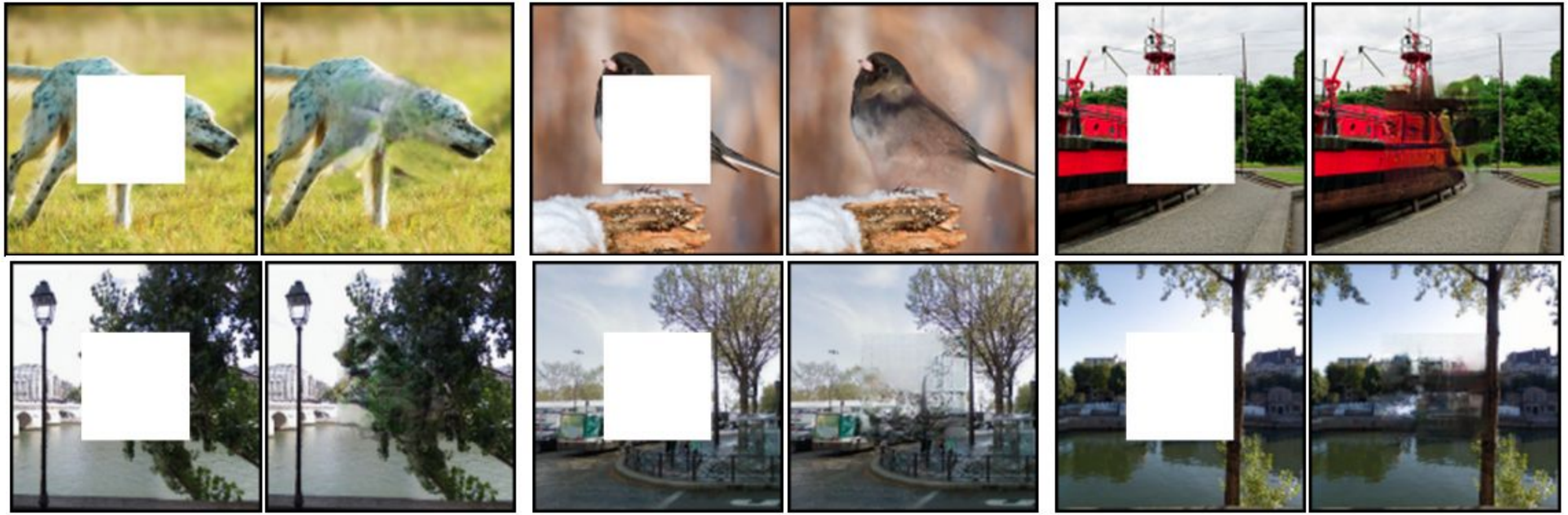
The pretext task (generative modeling) is to reconstruct the original input while learning meaningful latent representation.



Source: Pathak, et al., 2016

Self-supervised learning: Inpainting

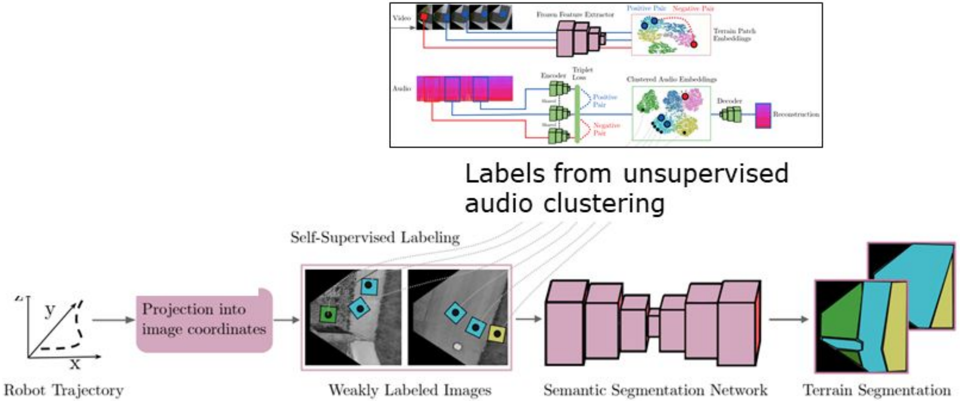
The model learns to fill in a missing piece in the image.



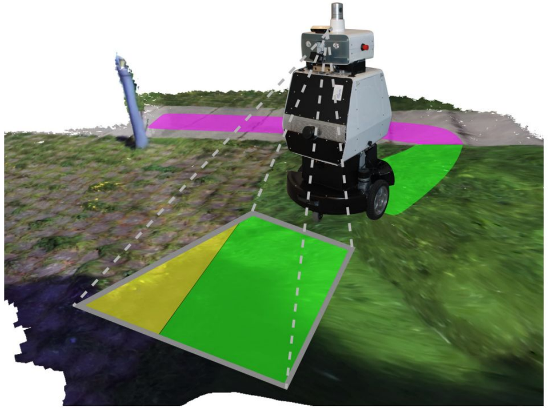
Source: Pathak, et al., 2016

Self-Supervised Learning: Cross-Modality and Robotics

- Self-supervision from cross-modality learning
- Goal: Reliably classify terrains for safe & efficient navigation
  - Equip the robot with a microphone and a camera
  - Leverage labels obtained from a audio classifier for self-supervision of the visual classifier



Labels from unsupervised audio clustering



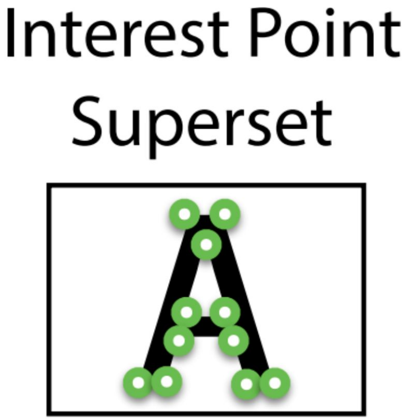
Source: Zurn, et al., 2019

Topic 1: SuperPoint: Self-Supervised Interest Point Detection and Description

Interest points are 2D locations in an image which are stable and repeatable from different lighting conditions and viewpoints.

Main task: Interest point detection.

This work uses a self-supervised approach to create a large dataset of pseudo-ground truth interest point locations in real images.



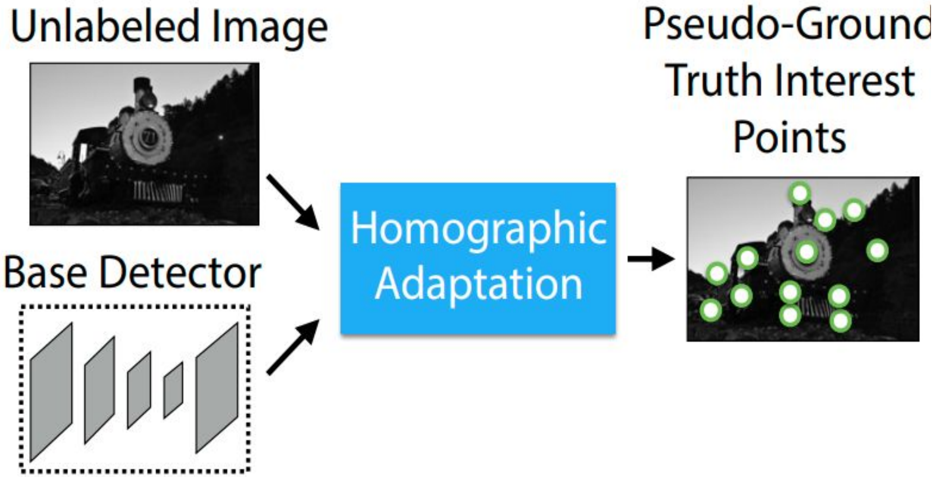
Source: DeTone et al. 2018

Supervisor: Dr. Daniele Cattaneo - Paper link: https://arxiv.org/abs/1712.07629

Topic 1: SuperPoint: Self-Supervised Interest Point Detection and Description

First, train a ConvNet based **detector** on a synthetic dataset with simple geometric shapes with no ambiguity in the interest points.

Then, they combine the trained detector and Homographic Adaptation to boost the performance on image textures and patterns.



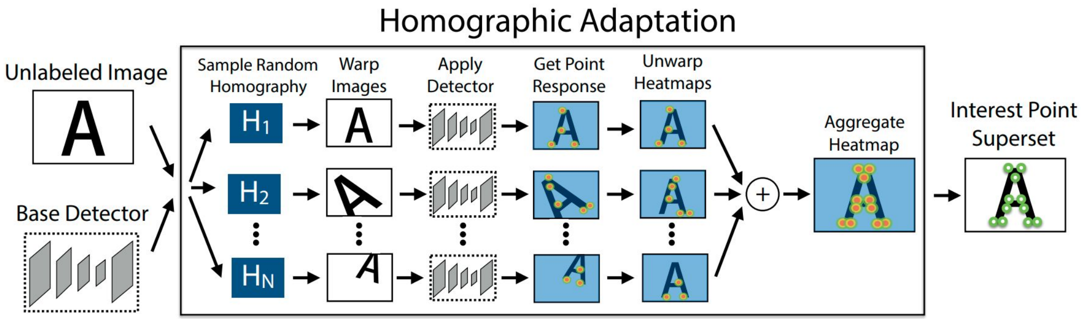
Source: DeTone et al. 2018

Supervisor: Dr. Daniele Cattaneo - Paper link: https://arxiv.org/abs/1712.07629

Topic 1: SuperPoint: Self-Supervised Interest Point Detection and Description

Homographic Adaptation is designed to enable self supervised training of interest point detectors.

It warps the input image multiple times to help an interest point detector see the scene from many different viewpoints and scales.



Source: DeTone et al. 2018

Supervisor: Dr. Daniele Cattaneo - Paper link: https://arxiv.org/abs/1712.07629

Topic 2: GANVO: Unsupervised Deep Monocular Visual Odometry and Depth Estimation with Generative Adversarial Networks

Tasks: Visual odometry (VO) and depth recovery.

Uses adversarial and recurrent unsupervised learning approaches for joint pose and depth map estimation.

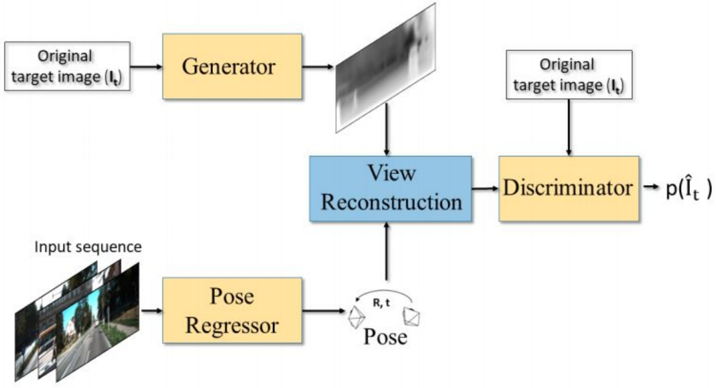
Generates depth images without any need for depth ground truth information.

Supervisor: Dr. Daniele Cattaneo - Paper link: https://arxiv.org/abs/1809.05786

Topic 2: GANVO: Unsupervised Deep Monocular Visual Odometry and Depth Estimation with Generative Adversarial Networks

Network consists of a pose regressor (CNN-RNN modules) and depth generator network (GAN).

Takes a sequence of monocular images to estimate 6-DoF camera motion and depth map that is sampled from the same input data distribution.



Source: Almalioglu et al. 2019

Supervisor: Dr. Daniele Cattaneo - Paper link: https://arxiv.org/abs/1809.05786

Topic 3: SelFlow: Self-Supervised Learning of Optical Flow

Task: optical flow estimation (pattern of apparent motion of objects).

- The basic idea behind unsupervised optical flow learning is to warp the target image towards the reference image according to the estimated optical flow
- Then minimize the difference between the reference image and the warped target image using a photometric loss.
- This idea could provide misleading information for occluded pixels.

Source: Liu et al. 2019

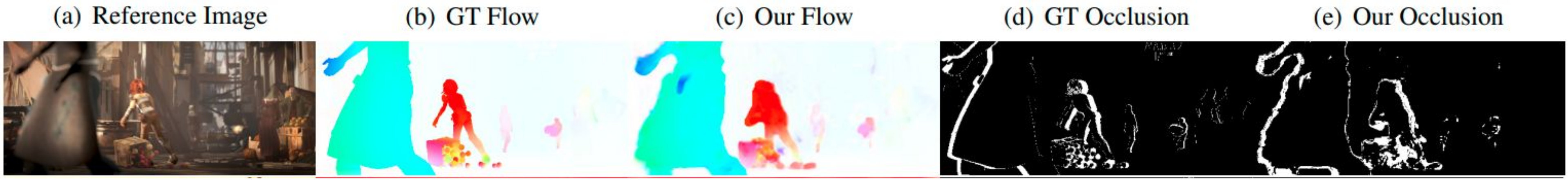
Supervisor: Dr. Daniele Cattaneo - Paper link: https://arxiv.org/abs/1712.07629

Topic 3: SelFlow: Self-Supervised Learning of Optical Flow

Pretext task: Distilling reliable flow estimations from non-occluded pixels.

These predictions are used to guide the optical flow learning for hallucinated occlusions.

This paper shows that a self-supervised approach can learn to estimate optical flow with any form of occlusions from unlabeled data.



Source: Liu et al. 2019

Supervisor: Dr. Daniele Cattaneo - Paper link: https://arxiv.org/abs/1904.09117

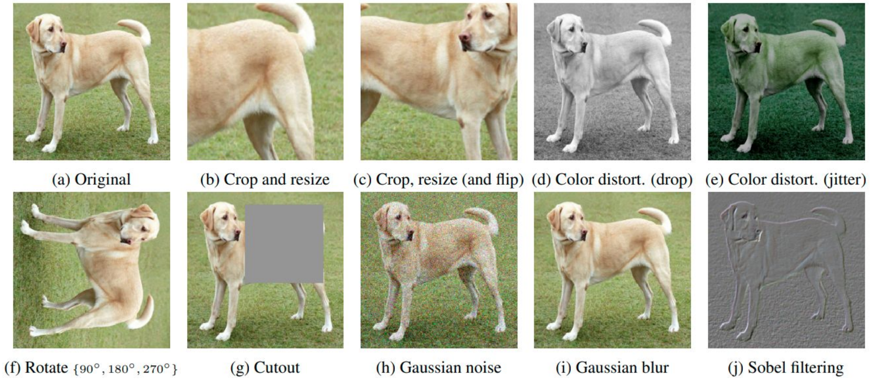
Topic 4: A Simple Framework for Contrastive Learning of Visual Representations

- Self-supervised discriminative approaches learn representations using objective functions similar to those used for supervised learning,
- They train networks to perform pretext tasks where both the inputs and labels are derived from an unlabeled dataset.
- Such approaches have relied on heuristics to design pretext tasks which could limit the generality of the learned representations.
- Discriminative approaches based on contrastive learning in the latent space have recently shown great promise, achieving state-of-the-art results.

Supervisor: Daniel Honerkamp - Paper link: https://arxiv.org/abs/2002.05709

Topic 4: A Simple Framework for Contrastive Learning of Visual Representations

The model learns representations by maximizing agreement between differently augmented views of the same data example via a contrastive loss in the latent space.



(a) Original (b) Crop and resize (c) Crop, resize (and flip) (d) Color distort. (drop) (e) Color distort. (jitter)

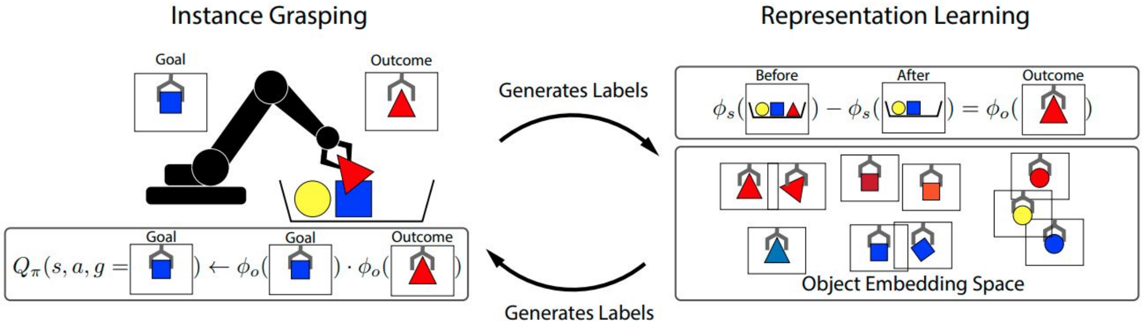
(f) Rotate $\{90°, 180°, 270°\}$ (g) Cutout (h) Gaussian noise (i) Gaussian blur (j) Sobel filtering

Source: Chen et al. 2019

Supervisor: Daniel Honerkamp - Paper link: https://arxiv.org/abs/2002.05709

Topic 5: Grasp2Vec: Learning Object Representations from Self-Supervised Grasping

Task: Acquiring object-centric representations through autonomous robotic interaction with the environment.



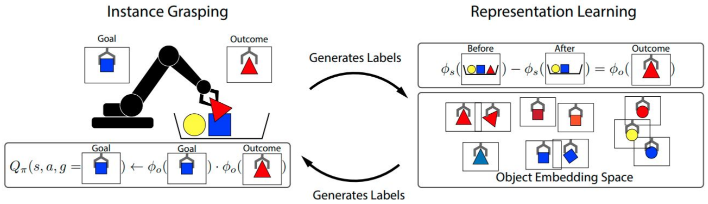
Source: Jang et al. 2019

Instance grasping and representation learning processes generate each other's labels in a fully self-supervised manner.

Supervisor: Daniel Honerkamp - Paper link: https://arxiv.org/abs/1811.06964

Topic 5: Grasp2Vec: Learning Object Representations from Self-Supervised Grasping

- Representation learning from grasping:
  - A robot arm removes an object from the scene.
  - Observes the resulting scene and the object in the gripper.
  - Then it is enforced that the difference of scene embeddings matches the object embedding.



Source: Jang et al. 2019

Supervising grasping with learned representations: This work proposes to use a similarity metric between object embeddings as a reward for instance grasping, removing the need to manually label grasp outcomes.

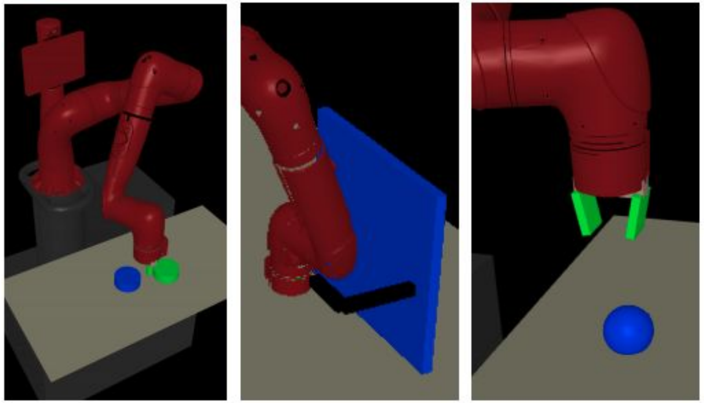
Supervisor: Daniel Honerkamp - Paper link: https://arxiv.org/abs/1811.06964

Topic 6: Visual Reinforcement Learning with Imagined Goals

For an autonomous agent to fulfill a wide range of user-specified goals at test time, it must be able to learn broadly applicable and general-purpose skill repertoires.

The particular goals that might be required at test-time are not known in advance.

Pusher, door opening, and pick-and-place.



Source: Nair et al. 2019

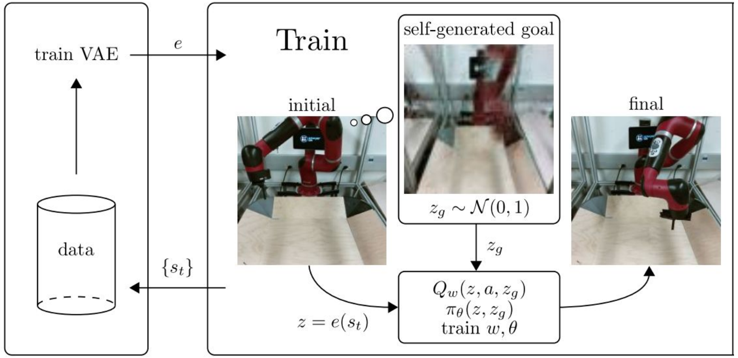
Supervisor: Daniel Honerkamp - Paper link: https://arxiv.org/abs/1807.04742

Topic 6: Visual Reinforcement Learning with Imagined Goals

In this work the agent performs a self-supervised “practice” phase where it imagines goals and attempts to achieve them.

Combines goal-conditioned reinforcement learning with unsupervised representation learning.

Representation learning is used to acquire a latent distribution that can be used to sample goals for unsupervised practice.



Source: Nair et al. 2019

Supervisor: Daniel Honerkamp - Paper link: https://arxiv.org/abs/1807.04742

Topic 7: Improving Semantic Segmentation through Spatio-Temporal Consistency Learned from Videos

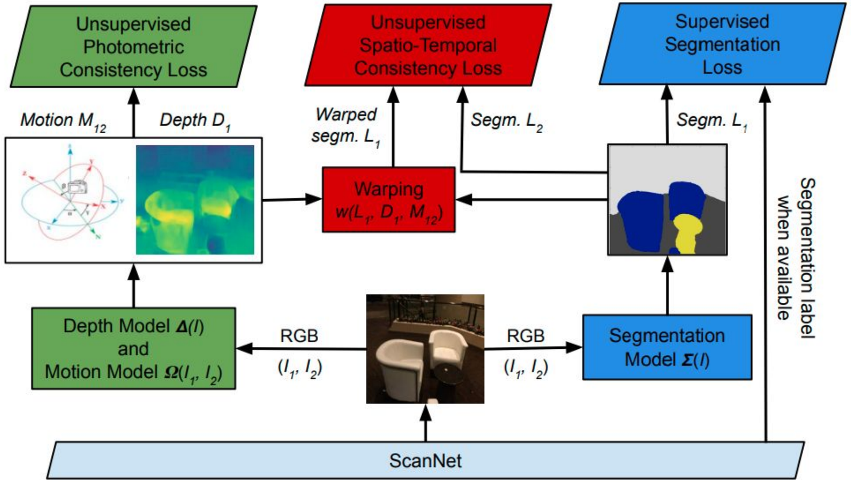
Task: Semantic Segmentation

1. Learning from video streams, as opposed to images, offers temporal coherency as a strong cue that can significantly enhance segmentation.
2. 3D Multiview consistency is as an additional supervision signal to train a single-frame segmenter, and as an additional signal at multi-frame inference time.
3. Unsupervised depth and egomotion estimation can bring together temporal continuity and multiview consistency as supervision signals for improving segmentation models.

Supervisor: Juana Valeria Hurtado - Paper link: https://arxiv.org/abs/2004.05324

Topic 7: Improving Semantic Segmentation through Spatio-Temporal Consistency Learned from Videos

Depth, egomotion, and camera intrinsics to improve the performance of single image semantic segmentation, by enforcing 3D-geometric and temporal consistency of segmentation masks across video frames.


Unsupervised Photometric Consistency Loss
Unsupervised Spatio-Temporal Consistency Loss
Supervised Segmentation Loss
Motion M_{12}
Depth D_1
Warped segm. L_1
Segm. L_2
Segm. L_1
Warping $w(L_1, D_1, M_{12})$
Segmentation label when available
Depth Model $\Delta(I)$ and Motion Model $\Omega(I_1, I_2)$
RGB (I_1, I_2)
RGB (I_1, I_2)
Segmentation Model $\Sigma(I)$
ScanNet

Source: Pasad et al. 2020

Supervisor: Juana Valeria Hurtado - Paper link: https://arxiv.org/abs/2004.05324

Topic 8: VideoBERT: A Joint Model for Video and Language Representation Learning

Tasks: action classification and video captioning.

Traina joint visual-linguistic model to learn high-level features without any explicit supervision.

GT: add some chopped basil leaves into it

VideoBERT: chop the basil and add to the bowl

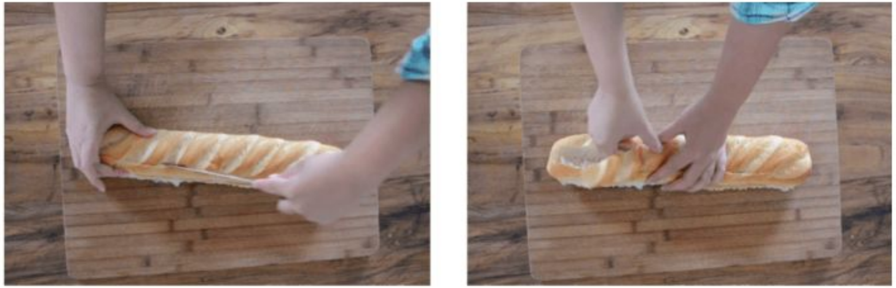
S3D: cut the tomatoes into thin slices

Source: Sun et al. 2019

Supervisor: Juana Valeria Hurtado - Paper link: https://arxiv.org/abs/1904.01766

Topic 8: VideoBERT: A Joint Model for Video and Language Representation Learning

Using videos where the spoken words are more likely to refer to visual content, this work presents a way to model the relationship between the visual domain and the linguistic domain.



GT: cut the top off of a french loaf

VideoBERT: cut the bread into thin slices

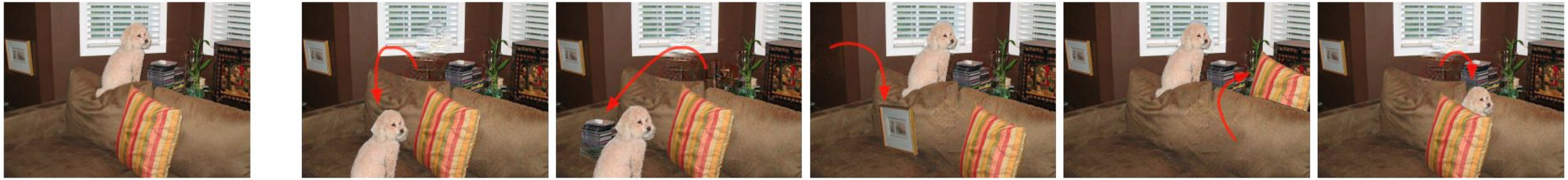
S3D: place the bread on the pan

Source: Sun et al. 2019

Supervisor: Juana Valeria Hurtado - Paper link: https://arxiv.org/abs/1904.01766

Topic 9: Self-Supervised Scene De-occlusion

Task: recover the underlying occlusion ordering and complete the invisible parts of occluded objects.



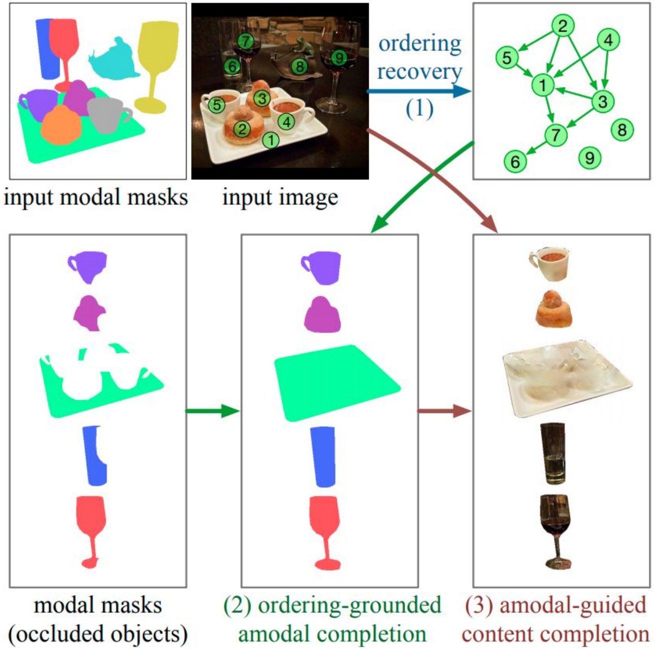
Source: Zhan et al. 2020

Supervisor: Juana Valeria Hurtado - Paper link: https://arxiv.org/abs/2004.02788

Topic 9: Self-Supervised Scene De-occlusion

Pretext tasks:

1. Progressive ordering recovery.
2. Modal completion.
3. Content completion.

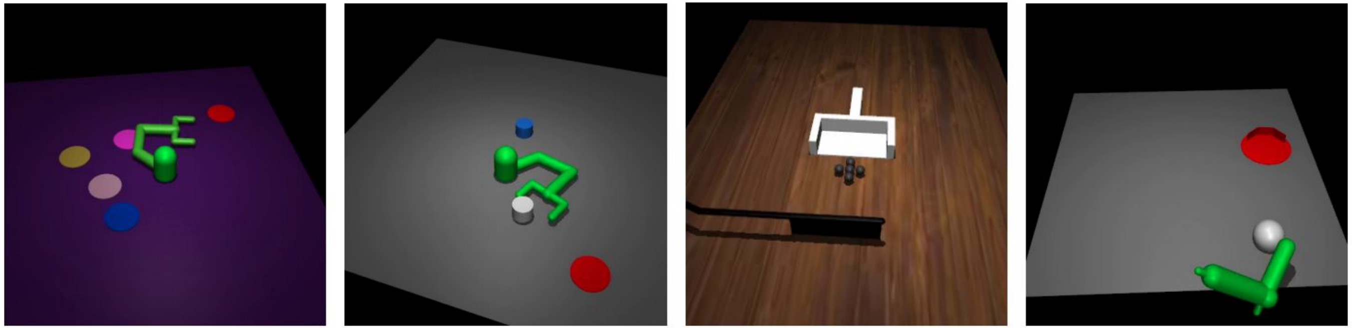

input modal masks
input image
ordering recovery (1)
1
2
3
4
5
6
7
8
9
modal masks (occluded objects)
(2) ordering-grounded amodal completion
(3) amodal-guided content completion

Source: Zhan et al. 2020

Supervisor: Juana Valeria Hurtado - Paper link: https://arxiv.org/abs/2004.02788

Topic 10: Imitation from Observation: Learning to Imitate Behaviors from Raw Video via Context Translation

The goal in imitation-from-observation is to learn policies only from a sequence of observations of the desired behavior, with each sequence obtained under differences in context.

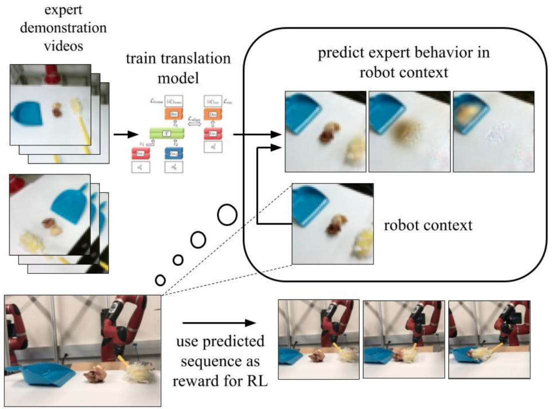


Source: Liu et al. 2018

Supervisor: Tim Welschehold - Paper link: https://arxiv.org/abs/1707.03374

Topic 10: Imitation from Observation: Learning to Imitate Behaviors from Raw Video via Context Translation

- Use videos of an expert demonstrator to train a context translation model.
- At learning time, robot sees the context of the task it needs to perform.
- Then, the model predicts what an expert would do in the robot context.
- This predicted sequence is used to define a cost function for RL thus enabling imitation from observation.


expert demonstration videos
train translation model
predict expert behavior in robot context
robot context
use predicted sequence as reward for RL

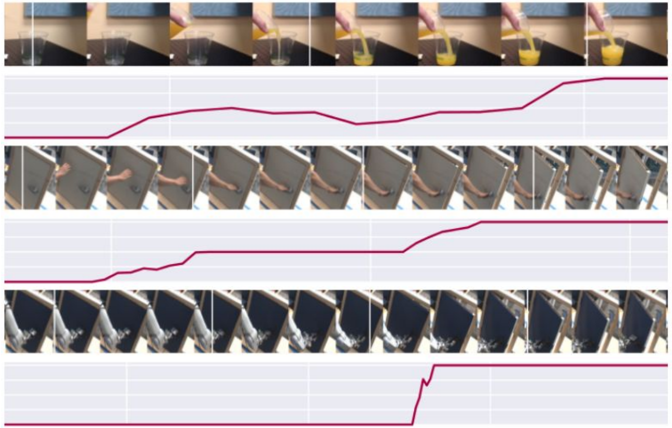
Source: Liu et al. 2018

Supervisor: Tim Welschehold - Paper link: https://arxiv.org/abs/1707.03374

Topic 11: Unsupervised Perceptual Rewards for Imitation Learning

Complex robotic manipulation skills learned directly and without supervised labels from a video of a human performing the task.

Reward function for reinforcement learning learned from human demonstrations.



Source: Sermanet et al. 2017

Supervisor: Tim Welschehold - Paper link: https://arxiv.org/abs/1612.06699

Topic 12: Time-Contrastive Networks: Self-Supervised Learning from Video

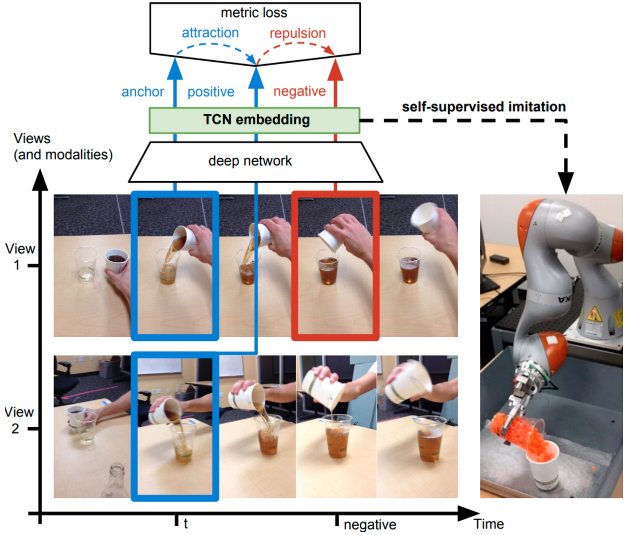
Learning representations and robotic behaviors entirely from unlabeled videos recorded from multiple viewpoints.

Source: Sermanet et al. 2018

Supervisor: Tim Welschehold - Paper link: https://arxiv.org/abs/1704.06888

Topic 12: Time-Contrastive Networks: Self-Supervised Learning from Video

- Learning signal from unlabeled multi-viewpoint videos of interaction scenarios.
- The learned representations effectively disentangle functional attributes such as pose while being viewpoint and agent invariant.
- The robot can learn to link this visual representation to a corresponding motor command using either reinforcement learning.



Source: Sermanet et al. 2018

Supervisor: Tim Welschehold - Paper link: https://arxiv.org/pdf/1704.06888.pdf

Assessing Interest

https://forms.gle/BvMtCXkKJt9XsaxB7