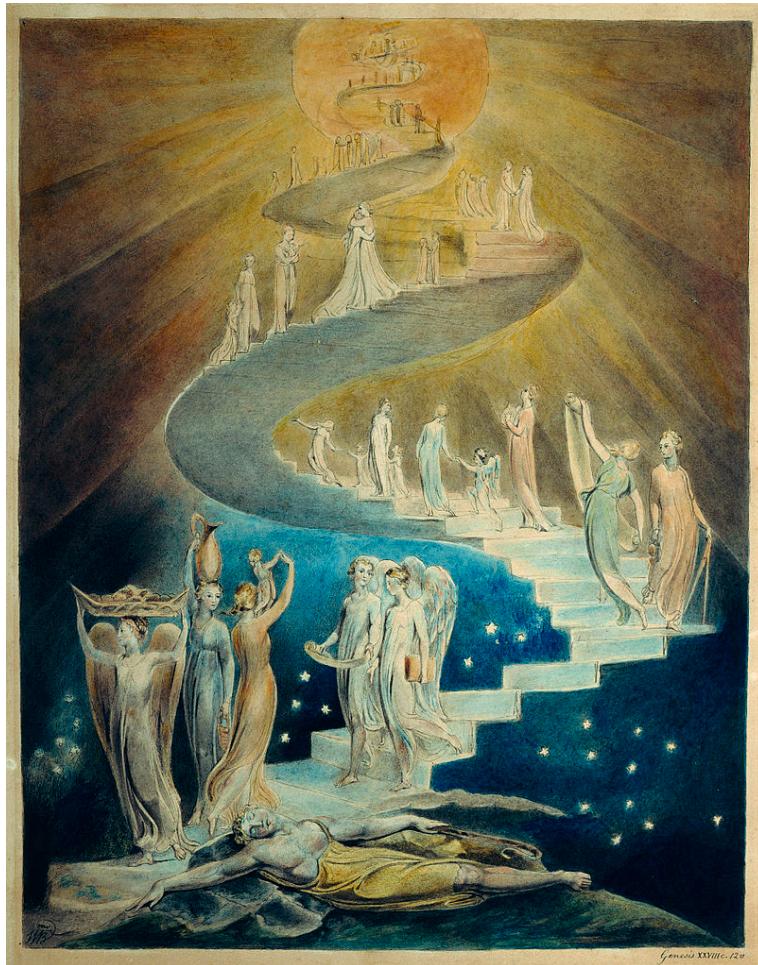


# Deep learning trends

(a non-comprehensive, non-objective view)



William Blake, Jacob's Ladder (1805)

# Outline

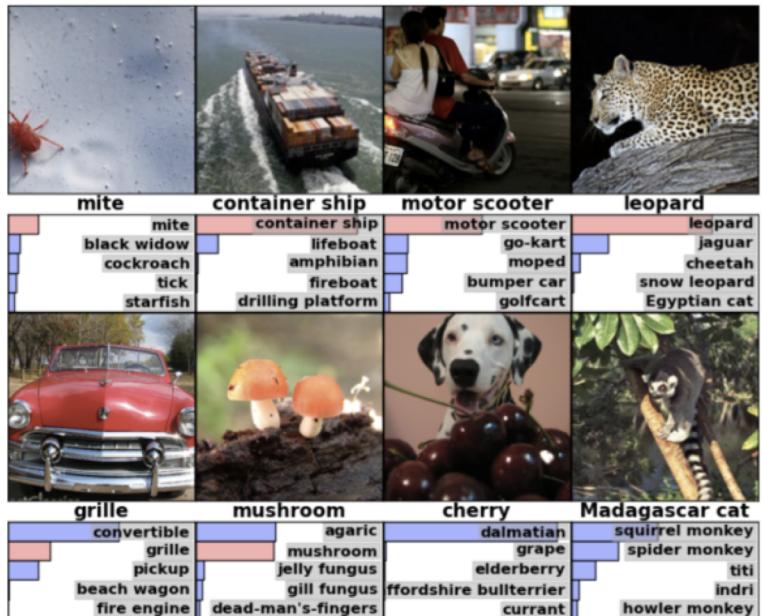
- Vision
- Embodied learning (RL, robotics)
- Language

# Deep learning in vision: Where are we now?

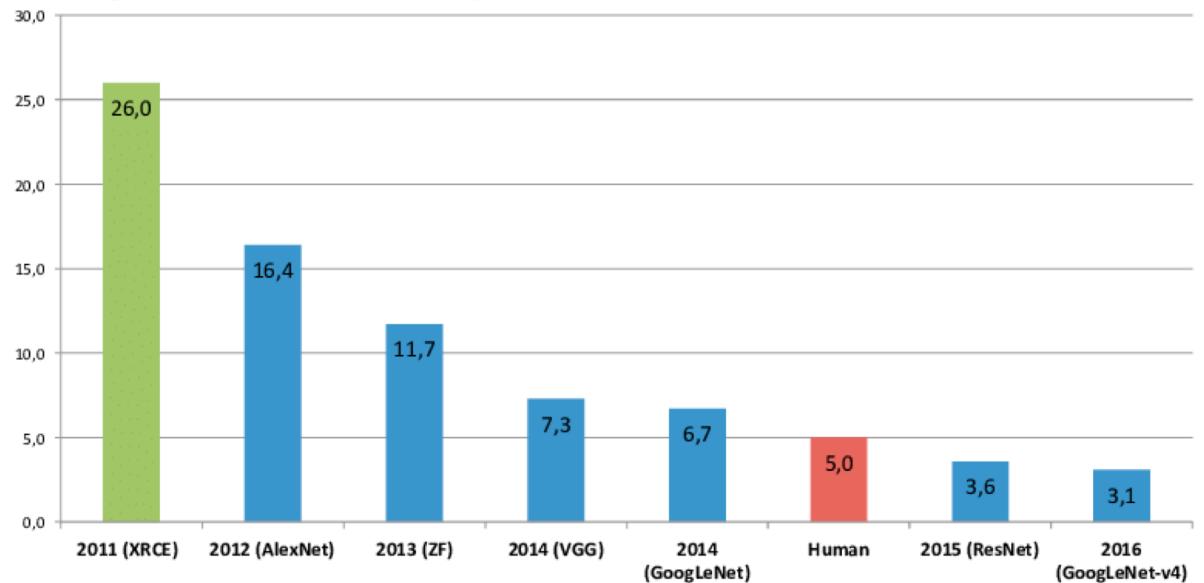
- Glass is half full?
  - DL methods *actually work* and are better than older methods in most ways
  - Good methodologies: standardized benchmarks and metrics for many problems, quantitative comparisons, ablation studies
  - Good infrastructure: deep learning packages, cloud services, etc.
  - Culture of code sharing and reproducibility
- Glass is half empty?
  - Too much focus on benchmarks and numbers
  - Number of papers is exploding but diversity of topics is not necessarily increasing
  - Cutting-edge research is becoming prohibitively resource-intensive
  - Core benchmarks (ImageNet, COCO) are likely saturating but bigger datasets (or larger amounts of computing power) are not yet generally accessible

# ImageNet: Asset or liability?

- Performance on the basic classification task has saturated



ImageNet Classification Error (Top 5)



[Figure source](#)

[ILSVRC Challenge](#)

# ImageNet: Asset or liability?

- Performance on the basic classification task has saturated

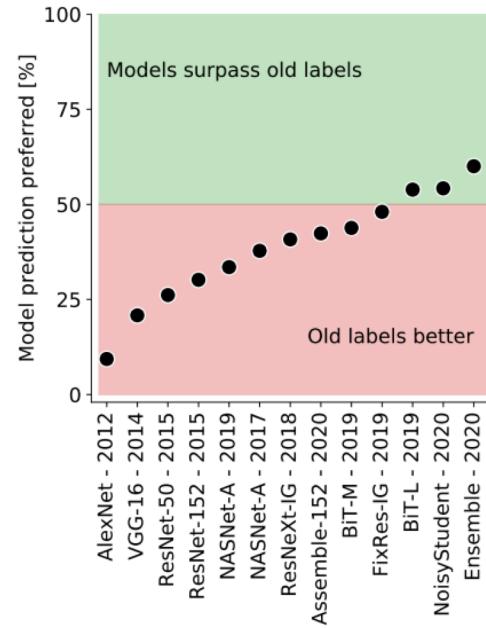


Figure 1: When presented with a model's prediction and the original ImageNet label, human annotators now prefer model predictions on average (Section 4). Nevertheless, there remains considerable progress to be made before fully capturing human preferences.

# ImageNet: Asset or liability?

- Attaching labels to images is not very meaningful in the first place

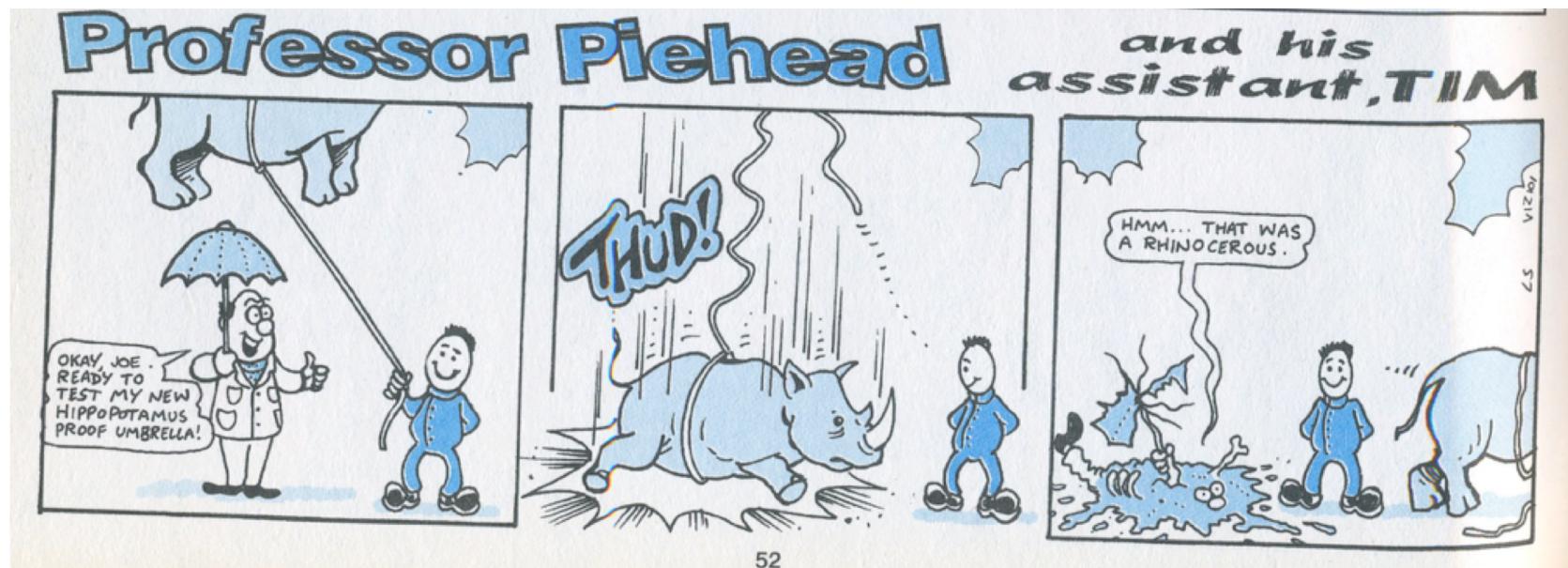


"Now! ... That should clear up  
a few things around here!"

Favorite cartoon of J. Koenderink

# ImageNet: Asset or liability?

- Attaching labels to images is not very meaningful in the first place



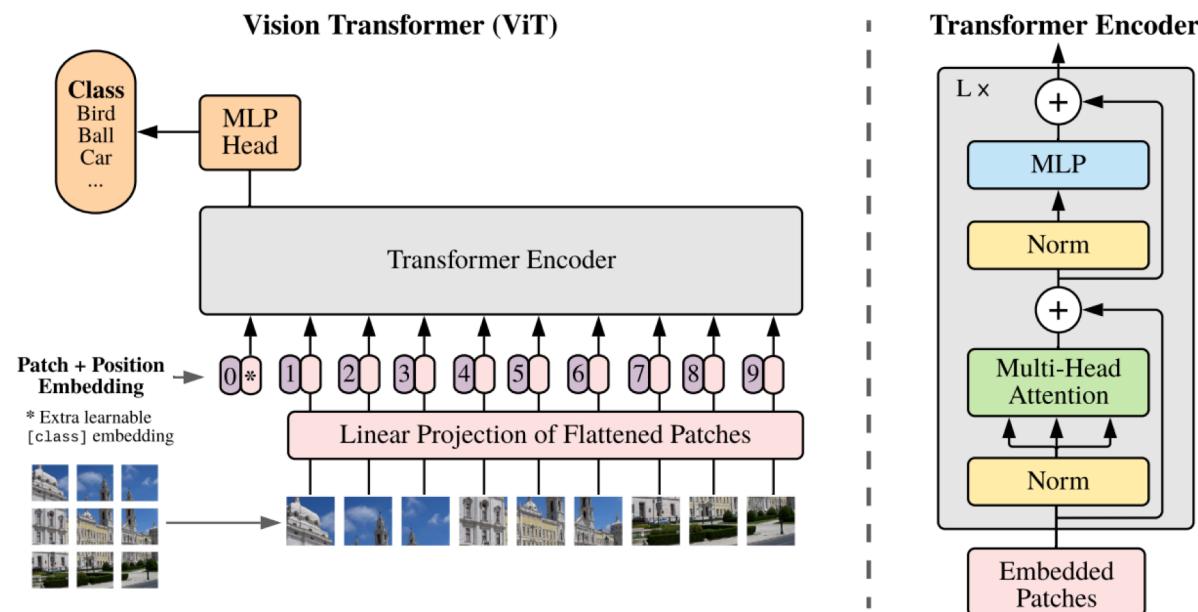
Favorite cartoon of D. Forsyth

# How can we move forward?

- Develop the next generation of architectures

# Beyond convolutional networks?

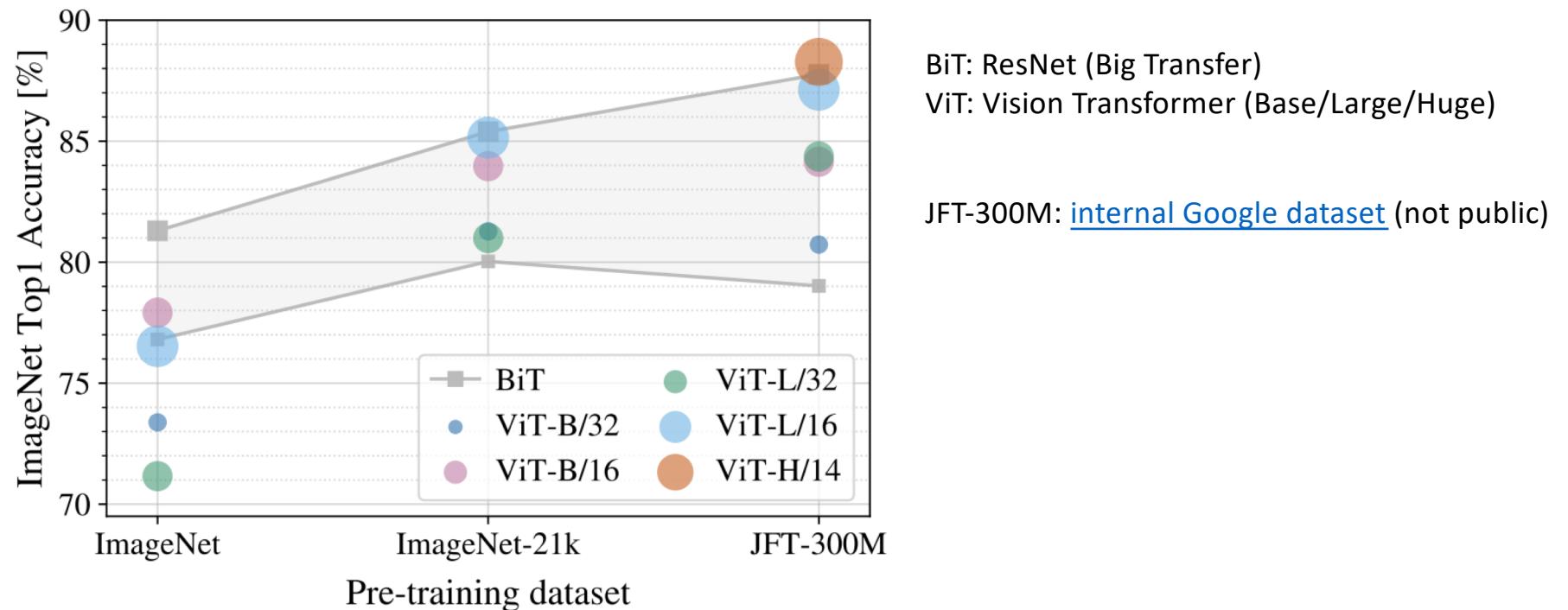
- Transformers for images



A. Dosovitskiy et al. [An image is worth 16x16 words: Transformers for image recognition at scale](#). arXiv 2020

# Beyond convolutional networks?

- Transformers for images



A. Dosovitskiy et al. [An image is worth 16x16 words: Transformers for image recognition at scale](#). arXiv 2020

# How can we move forward?

- Develop the next generation of architectures
- Go beyond classification
  - “Rich” prediction tasks

# “Rich” prediction tasks

**Class label to image**



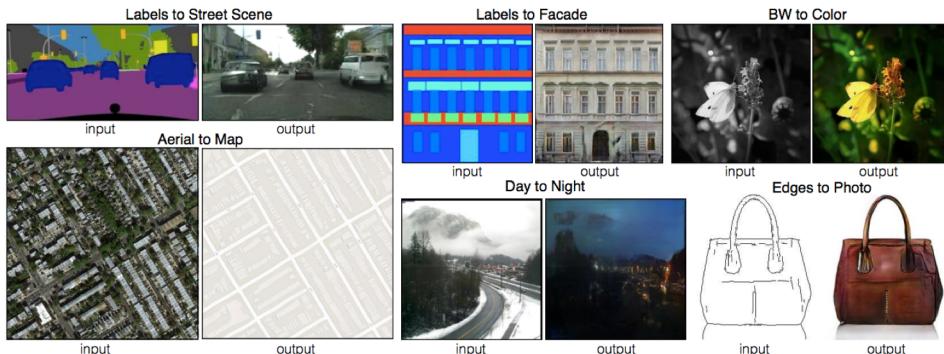
[Zhang et al. \(2019\)](#)

**Text to image**



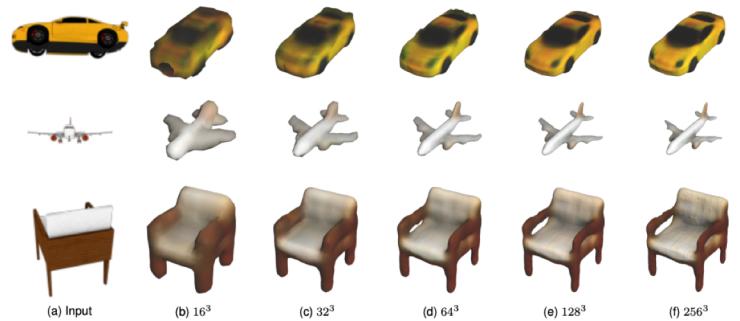
[Qiao et al. \(2019\)](#)

**Image to image**



[Isola et al. \(2017\)](#)

**Image to 3D**



[Hane et al. \(2019\)](#)

# Learning 3D structure

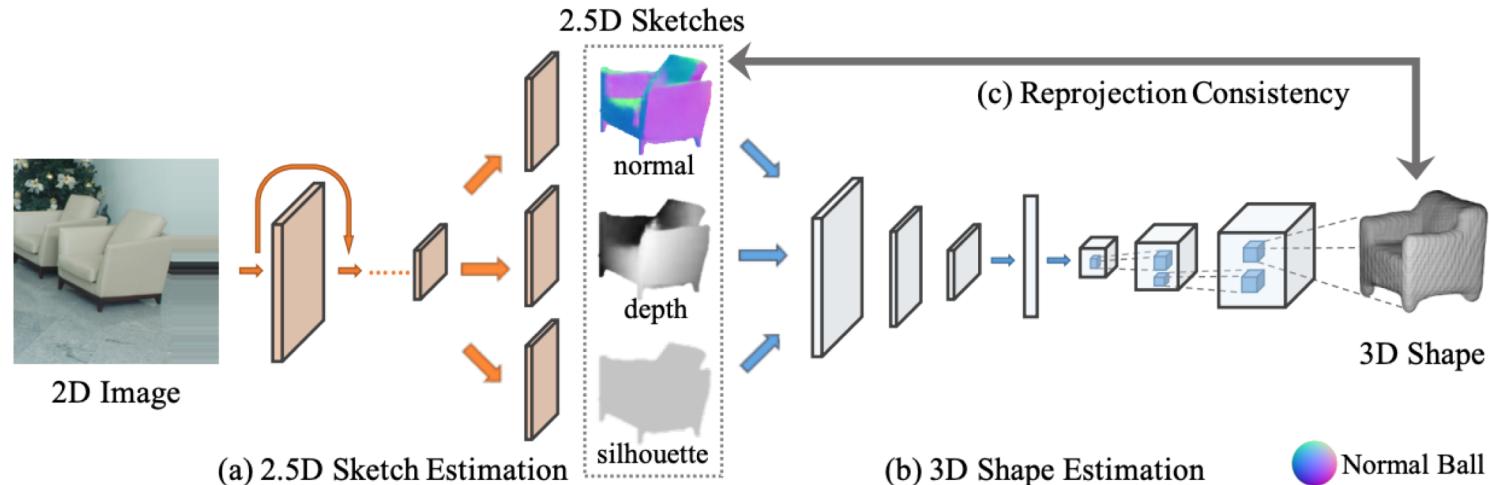
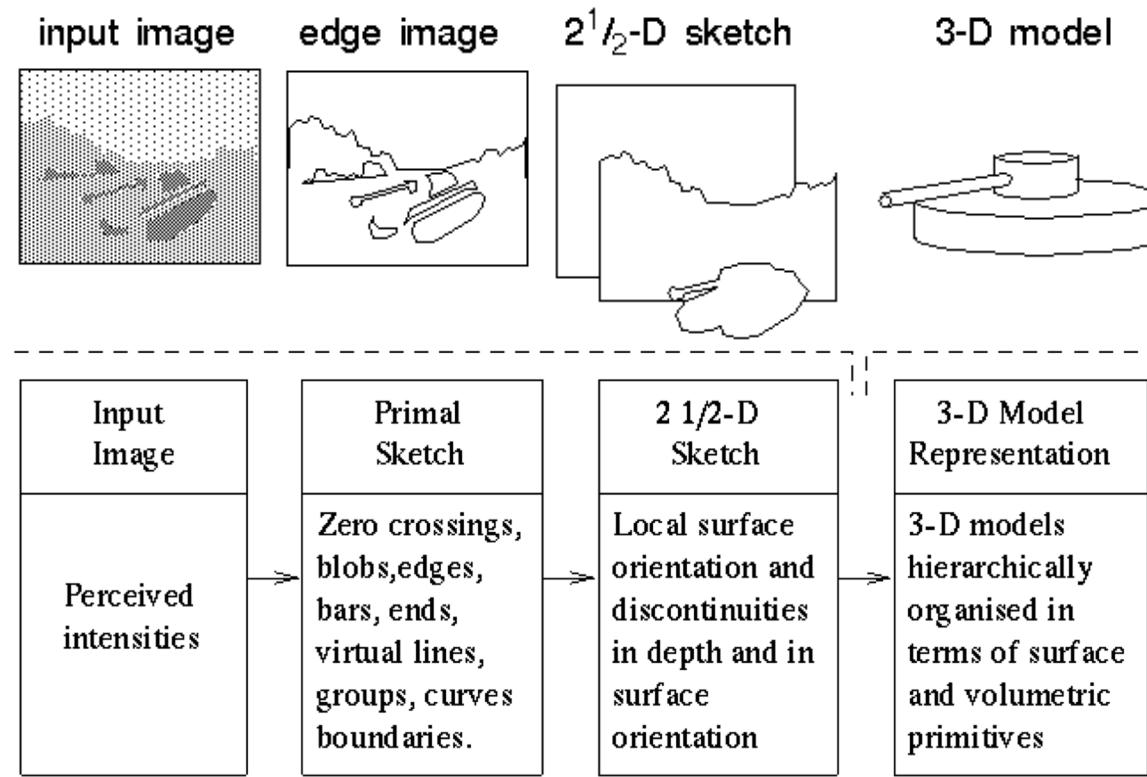


Figure 2: Our model (MarrNet) has three major components: (a) 2.5D sketch estimation, (b) 3D shape estimation, and (c) a loss function for reprojection consistency. MarrNet first recovers object normal, depth, and silhouette images from an RGB image. It then regresses the 3D shape from the 2.5D sketches. In both steps, it uses an encoding-decoding network. It finally employs a reprojection consistency loss to ensure the estimated 3D shape aligns with the 2.5D sketches. The entire framework can be trained end-to-end.

# Historical context: Marr's visual processing pipeline

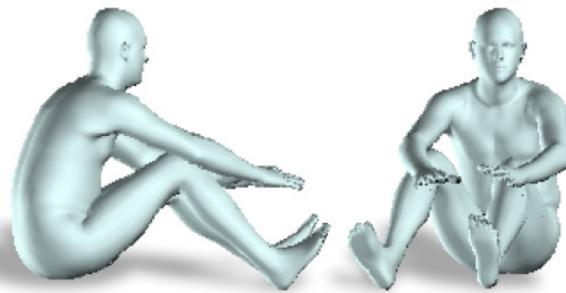
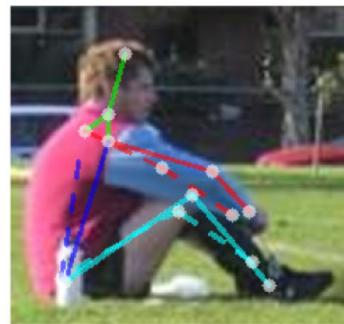


[David Marr \(1945-1980\)](#)

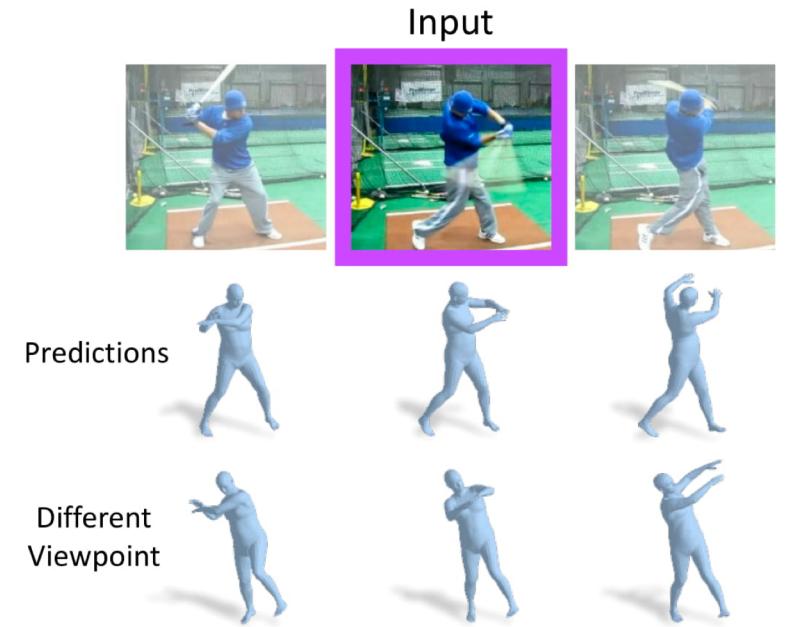


[Image source](#)

# Learning 3D structure



F. Bogo, A. Kanazawa, C. Lassner, P. Gehler, J. Romero and M. Black,  
[Keep it SMPL: Automatic Estimation of 3D Human Pose and Shape  
from a Single Image](#), ECCV 2016



A. Kanazawa, J. Zhang, P. Felsen, J. Malik,  
[Learning 3D Human Dynamics from Video](#),  
CVPR 2019

# Learning skills from video

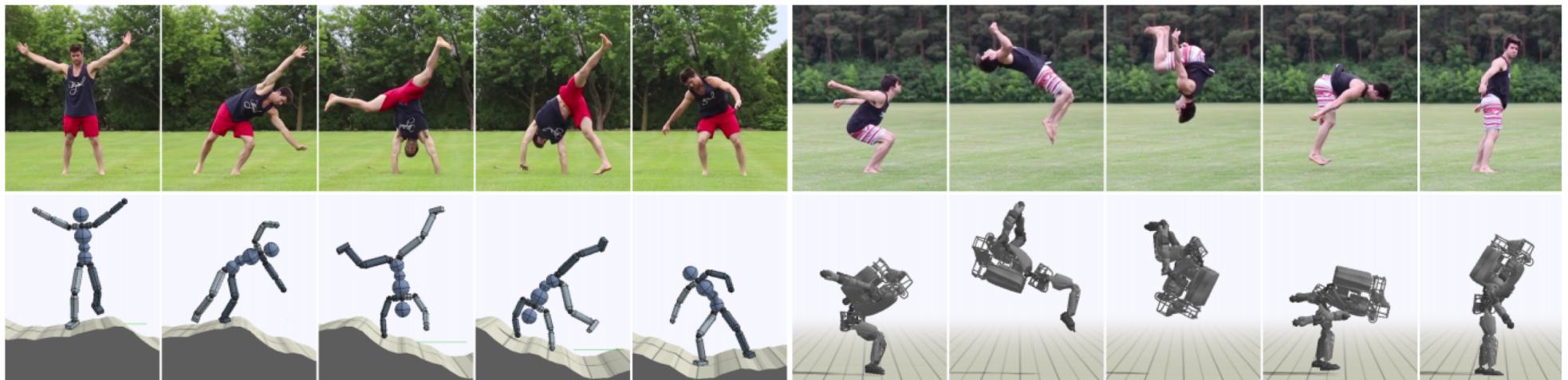
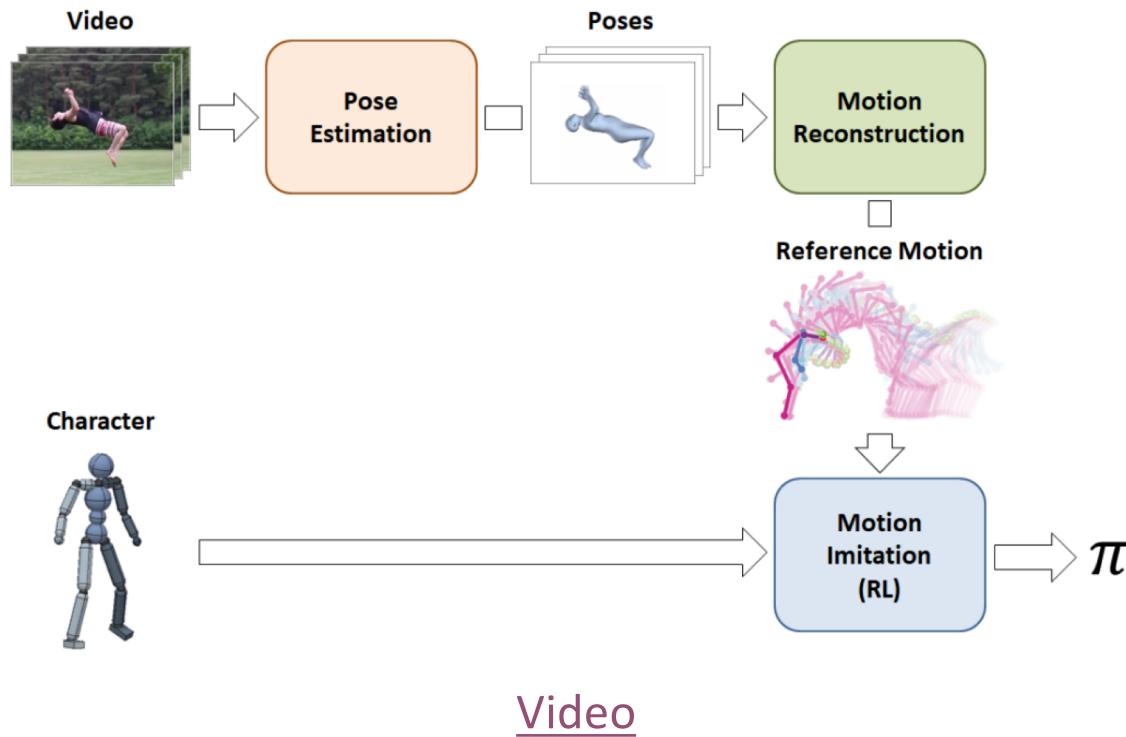


Fig. 1. Simulated characters performing highly dynamic skills learned by imitating video clips of human demonstrations. **Left:** Humanoid performing cartwheel B on irregular terrain. **Right:** Backflip A retargeted to a simulated Atlas robot.

[Video](#)

X. B. Peng, A. Kanazawa, J. Malik, P. Abbeel, S. Levine, [SFV: Reinforcement Learning of Physical Skills from Videos](#),  
SIGGRAPH Asia 2018

# Learning skills from video



X. B. Peng, A. Kanazawa, J. Malik, P. Abbeel, S. Levine, [SFV: Reinforcement Learning of Physical Skills from Videos](#),  
SIGGRAPH Asia 2018

# How can we move forward?

- Develop the next generation of architectures
- Go beyond classification
  - “Rich” prediction tasks
  - Generation

# Generation

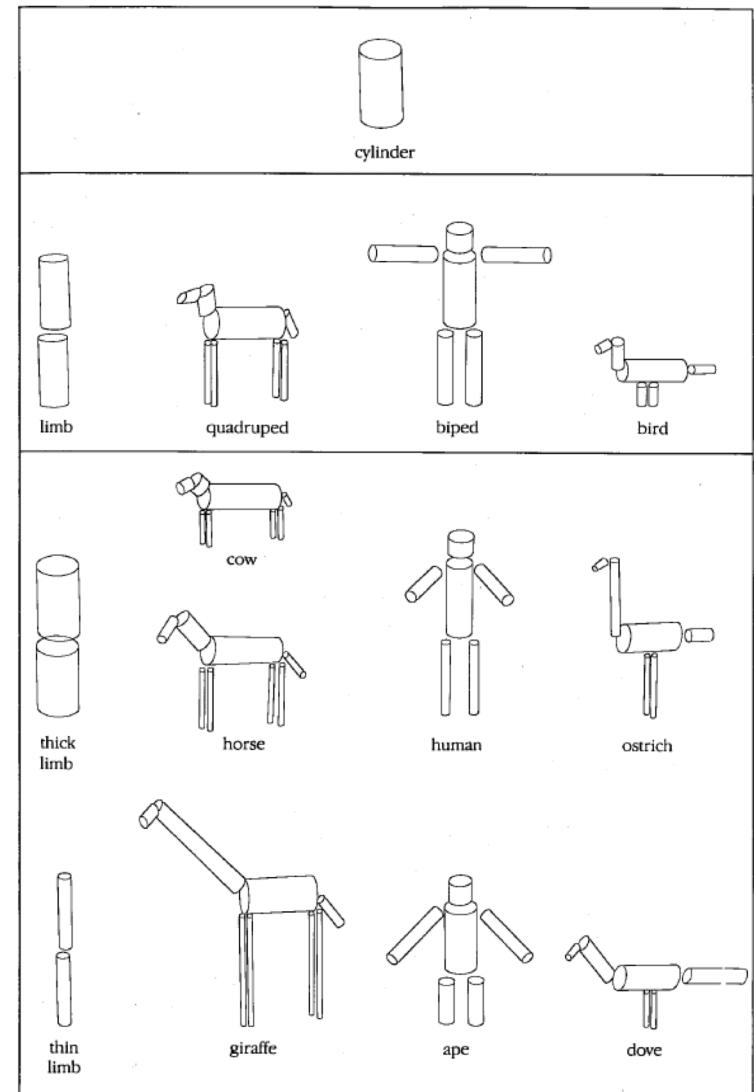
- State of the art: StyleGAN



T. Karras, S. Laine, T. Aila. [A Style-Based Generator Architecture for Generative Adversarial Networks](#). CVPR 2019

# Generation

- May shed light on important open questions, such as:
  - Do we need explicitly compositional (part-based) object representations?



Source: D. Marr, Vision, 1982

# Generation

- May shed light on important open questions, such as:
  - Do we need explicitly compositional (part-based) object representations?
- BigGAN proves that we don't?



A. Brock, J. Donahue, K. Simonyan, [Large scale GAN training for high fidelity natural image synthesis](#), ICLR 2019

# Generation

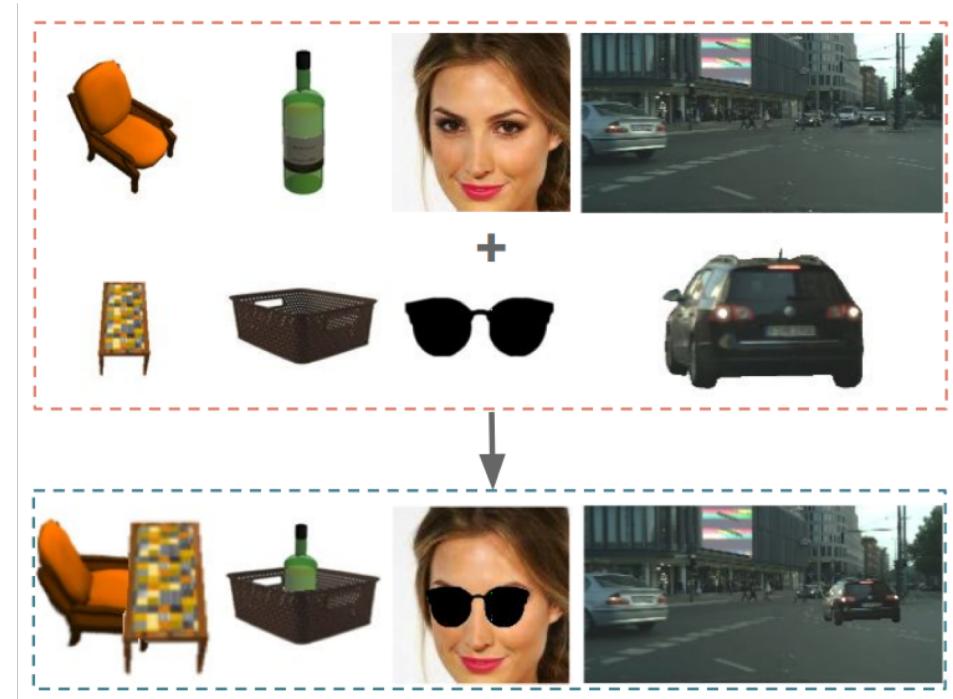
- May shed light on important open questions, such as:
  - Do we need explicitly compositional (part-based) object representations?
- Or maybe we do?



A. Brock, J. Donahue, K. Simonyan, [Large scale GAN training for high fidelity natural image synthesis](#), ICLR 2019

# Generation

- May shed light on important open questions, such as:
  - Do we need explicitly compositional (part-based) object representations?

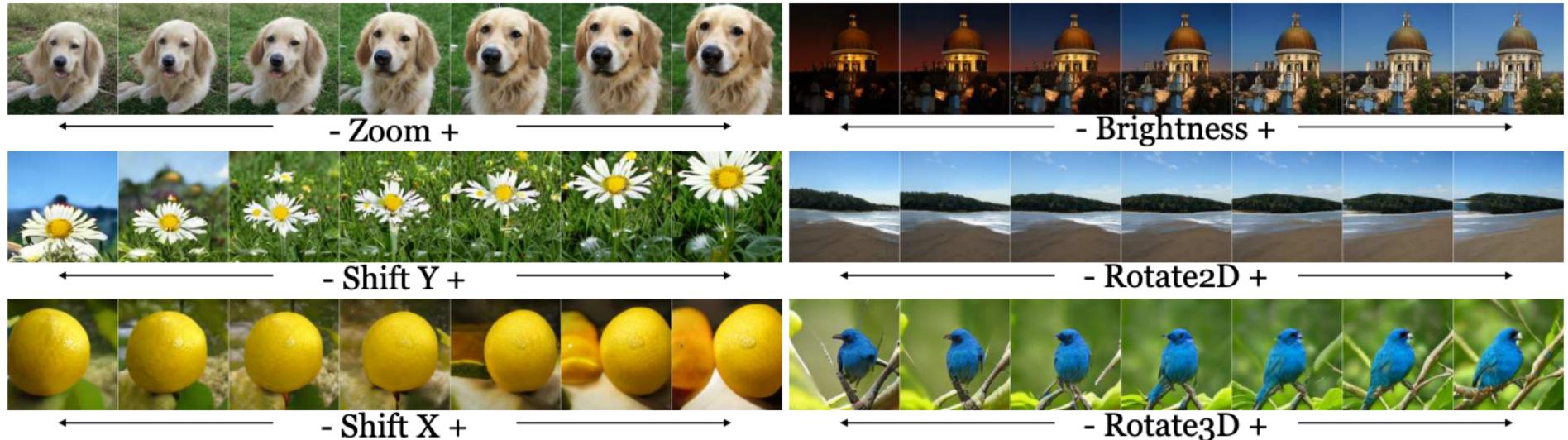


S. Azadi, D. Pathak, S. Ebrahimi, T. Darrell, [Compositional GAN: Learning Conditional Image Composition](#), arXiv 2019

# Generation

- May shed light on important open questions, such as:
  - Do we need explicit 3D object representations?

Maybe we don't?



A. Jahanian, L. Chai, and P. Isola. [On the "steerability" of generative adversarial networks](#). ICLR 2020

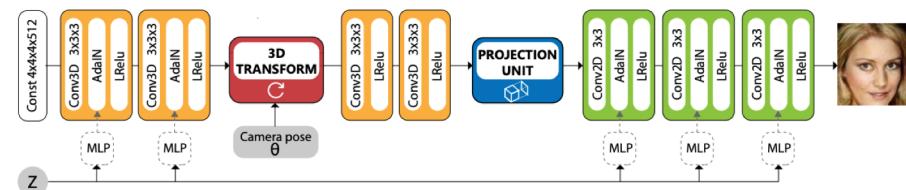
# Generation

- May shed light on important open questions, such as:
  - Do we need explicit 3D object representations?

Or maybe we do?



Figure 1. HoloGAN learns to separate pose from identity (shape and appearance) only from unlabelled 2D images without sacrificing the visual fidelity of the generated images. All results shown here are sampled from HoloGAN for the same identities in each row but in different poses.

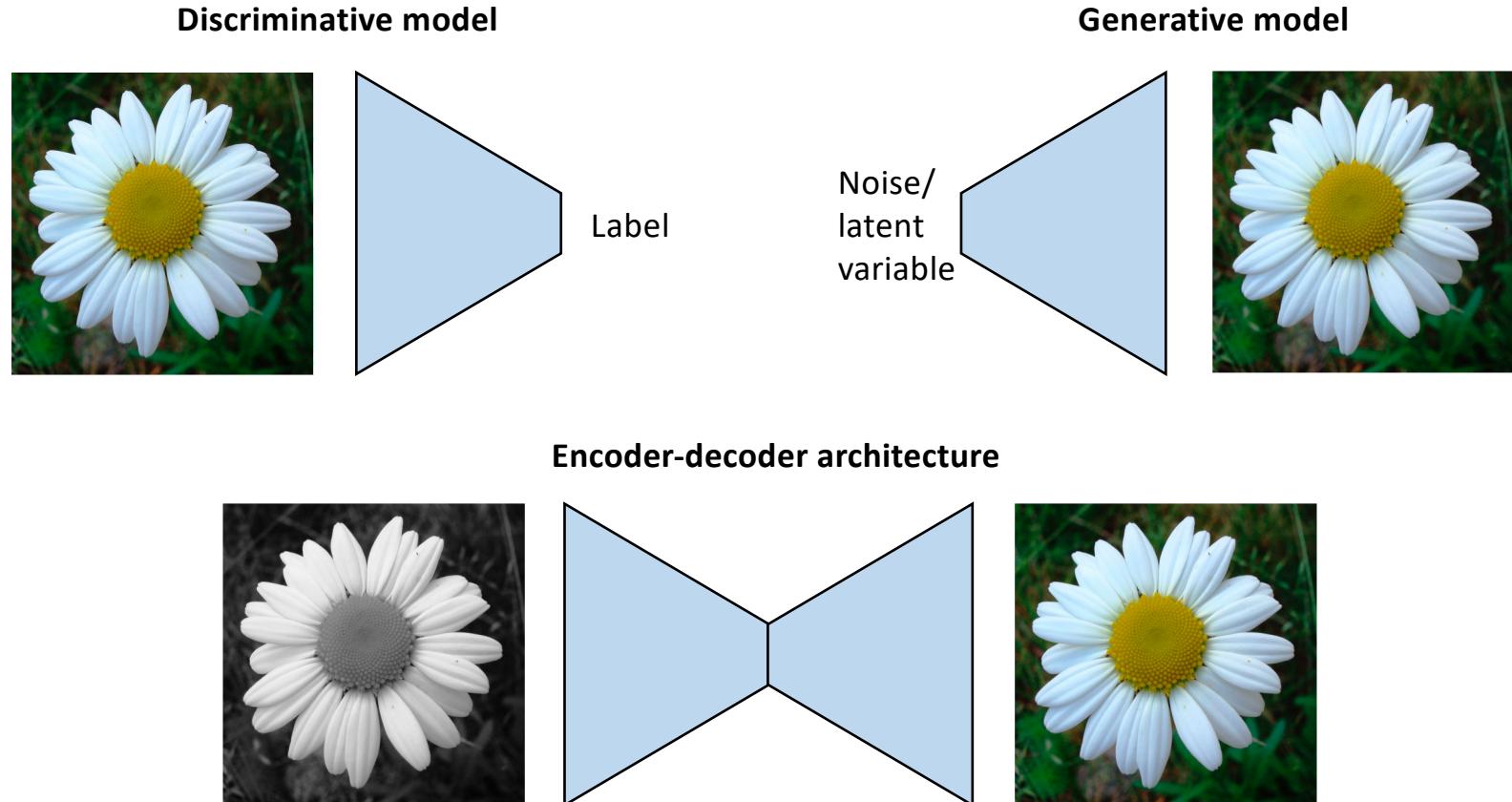


T. Nguyen-Phuoc et al., [HoloGAN: Unsupervised Learning of 3D Representations From Natural Images](#), 2019

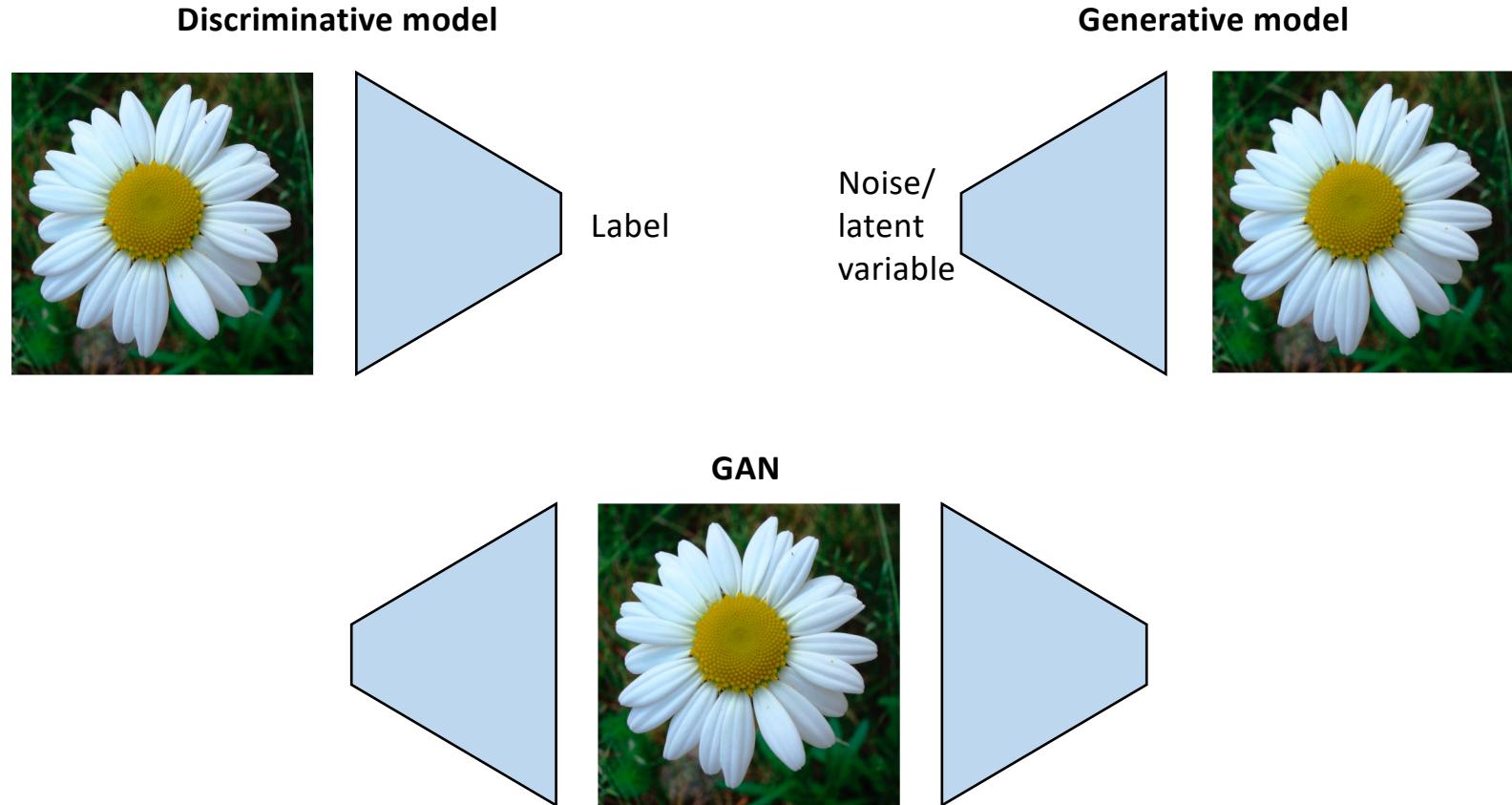
# Possible ways forward

- Develop the next generation of architectures
- Go beyond classification
  - “Rich” prediction tasks
  - Generation
- Work on integrating discriminative and generative models

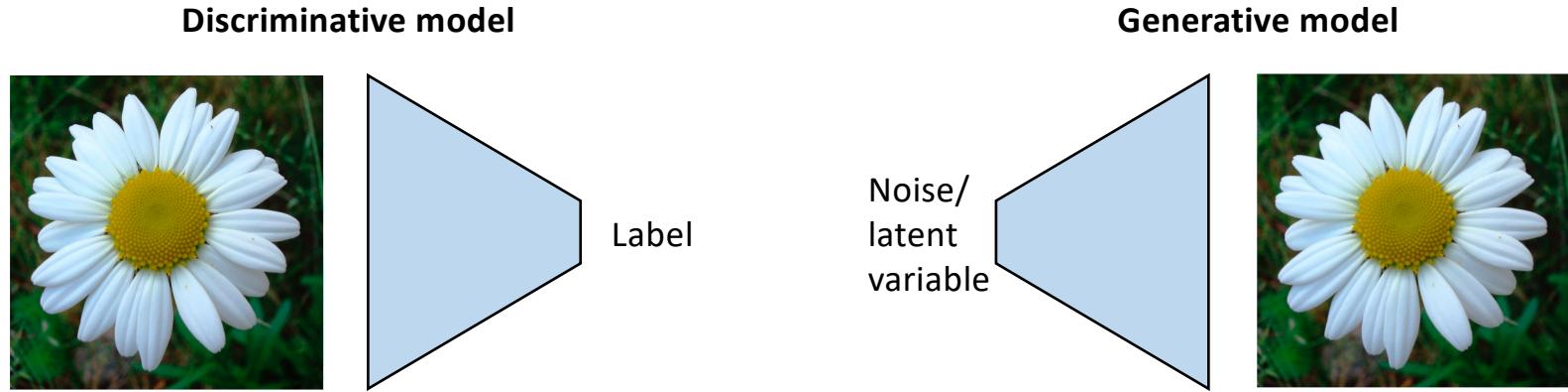
# Integrating discriminative and generative models?



# Integrating discriminative and generative models?



# Integrating discriminative and generative models?



**What else is there?**

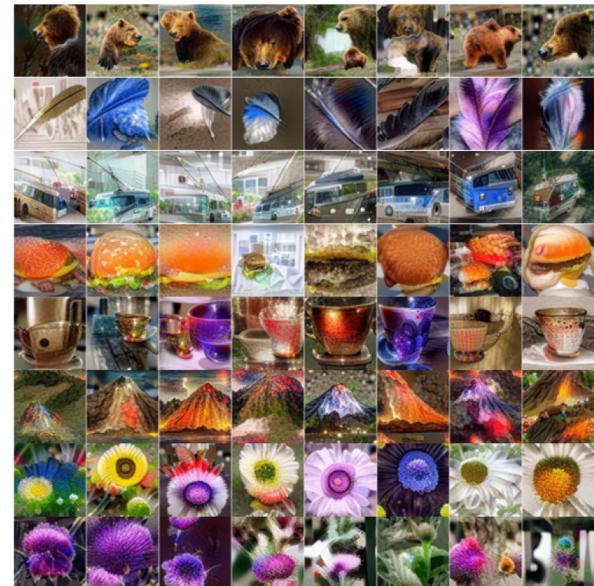
# Integrating discriminative and generative models?

- Can “hallucination” help with recognition?
  - Low-shot learning, incremental learning, transfer learning



Figure 1. Given a single image of a novel visual concept, such as a blue heron, a person can visualize what the heron would look like in other poses and different surroundings. If computer recognition systems could do such hallucination, they might be able to learn novel visual concepts from less data.

[Wang et al. \(2018\)](#)



High-quality images synthesized by “inverting” an ImageNet-pretrained network

[Yin et al. \(2019\)](#)

# Integrating discriminative and generative models?

- Can attribute-based manipulation of training examples help to train more accurate or less biased models?

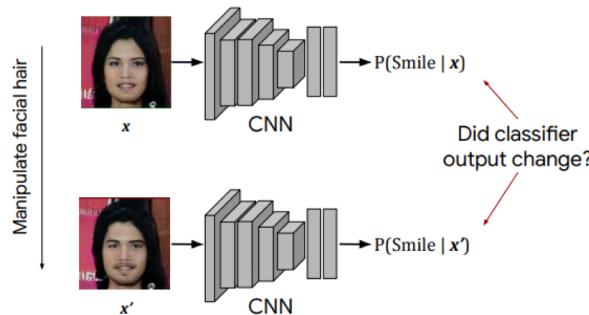


Figure 1: Counterfactual attribute sensitivity measures the effect of manipulating a specific property of an image on the output of a trained classifier. In this example, we consider the effect of adding facial hair on the output of a smiling classifier. If the classifier's output systematically changes as a result of adding or removing facial hair then a potentially undesirable bias has been detected since, all else being equal, facial hair should be irrelevant to the classification task.

[Denton et al. \(2019\)](#)



Figure 1: Our approach learns to *actively create* training images for fine-grained attribute comparisons, focusing attention on informative pairs of images unavailable in the real image training pool. Here we show two example pairs it generated to improve learning of *open* (left) and *masculine* (right).

[Yu and Grauman \(2019\)](#)

# Possible ways forward

- Develop the next generation of architectures
- Go beyond classification
  - “Rich” prediction tasks
  - Generation
- Work on integrating discriminative and generative models
- Move away from full supervision

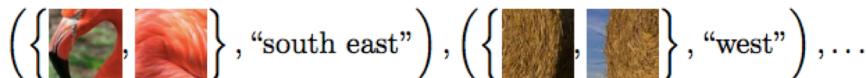
# Self-supervised learning

- For still image recognition tasks, self-supervision is more cumbersome (so far) as getting enough supervised training data
- However, it is very attractive (and possibly unavoidable) for video, audio, and sensorimotor learning

Ex. 1: **Inpainting** (remove patch and then predict it)



Ex. 2: **Context** (given two patches, predict their spatial relation)

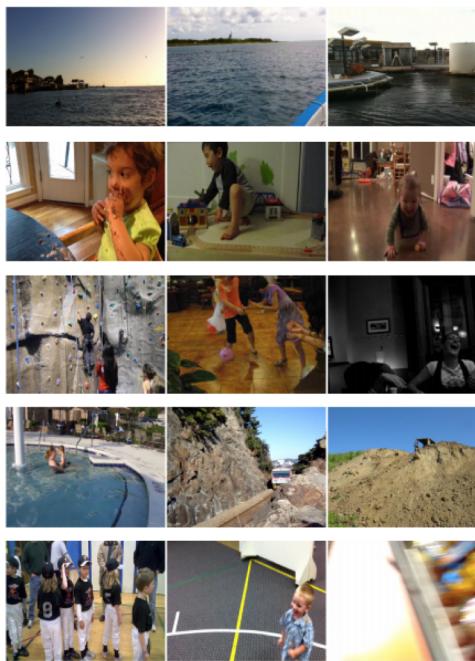


Ex. 3: **Colorization** (predict color given intensity)

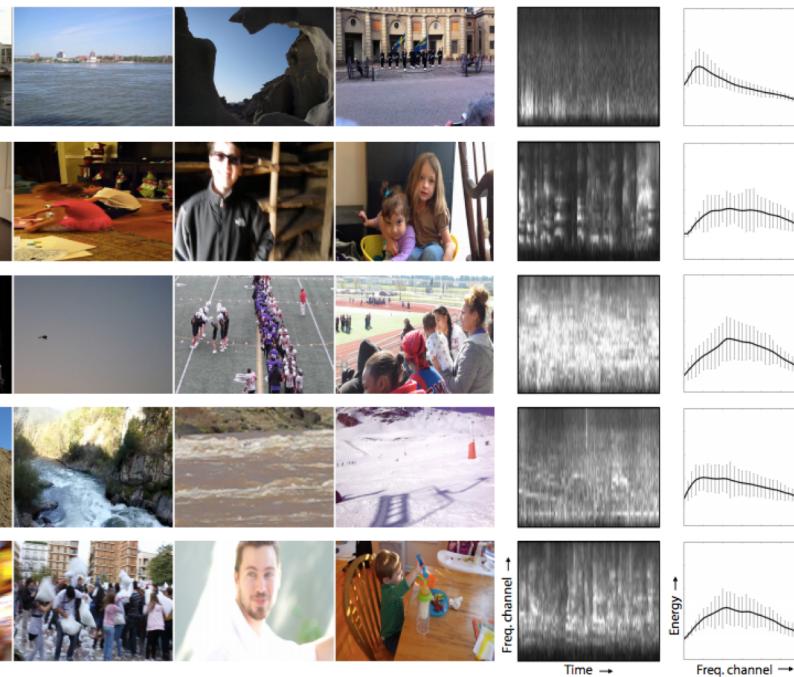


Figure source: G. Larsson, M. Maire, and G. Shakhnarovich, [Colorization as a Proxy Task for Visual Understanding](#), CVPR 2017

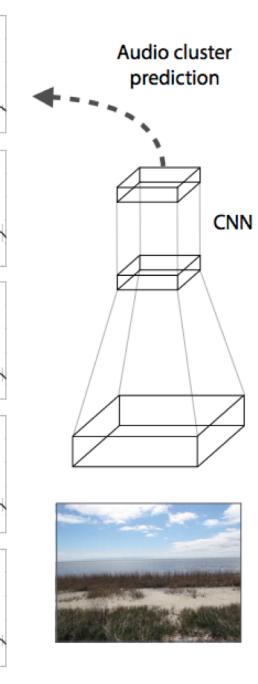
# Cross-modal self-supervision



(a) Images grouped by audio cluster

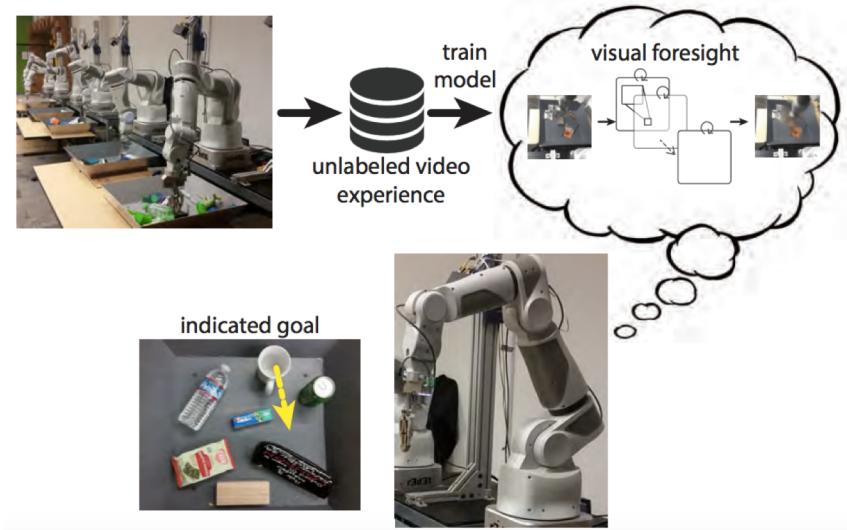
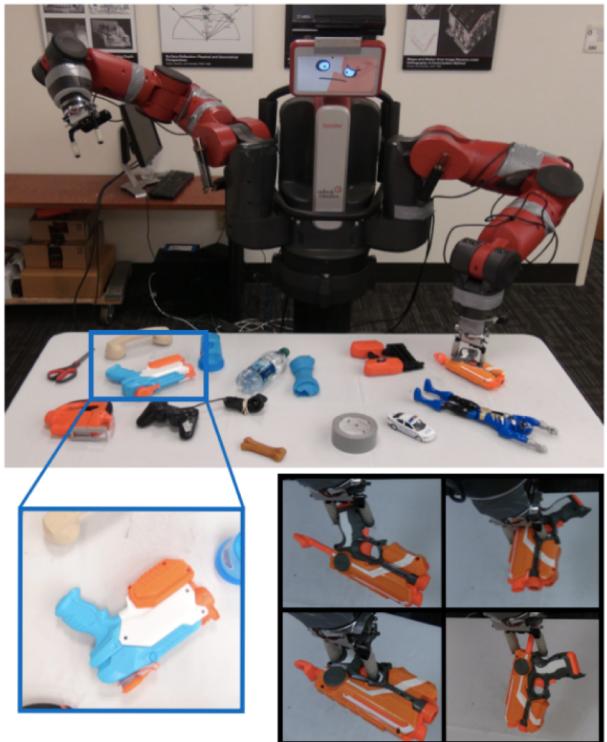


(b) Clustered audio stats. (c) CNN model



A. Owens, J. Wu, J. McDermott, W. Freeman, and A. Torralba, [Ambient Sound Provides Supervision for Visual Learning](#), ECCV 2016

# Self-supervision from future prediction



L. Pinto and A. Gupta, [Supersizing self-supervision: Learning to grasp from 50K tries and 700 robot hours](#), ICRA 2016

C. Finn and S. Levine. [Deep Visual Foresight for Planning Robot Motion](#). ICRA 2017

# Self-supervision from future prediction

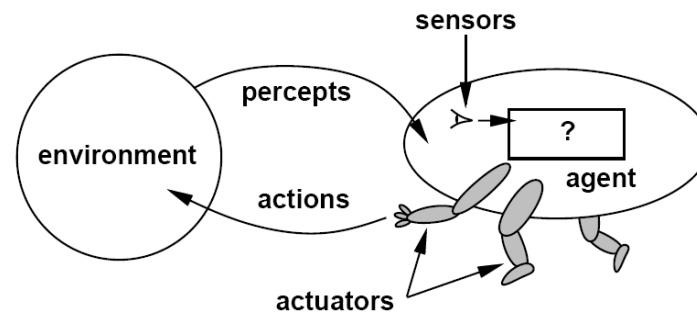
- Curiosity is also about predicting the effects of one's actions!



[Video](#)

# Possible ways forward

- Develop the next generation of architectures
- Go beyond classification
  - “Rich” prediction tasks
  - Generation
- Work on integrating discriminative and generative models
- Move away from full supervision
- Focus on embodied vision, sensorimotor learning



# Outline

- Vision
- Embodied learning (RL, robotics)

# Embodied platforms

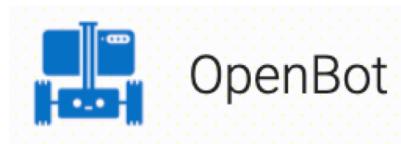
- Simulation: [AI2Thor](#), [Habitat](#)



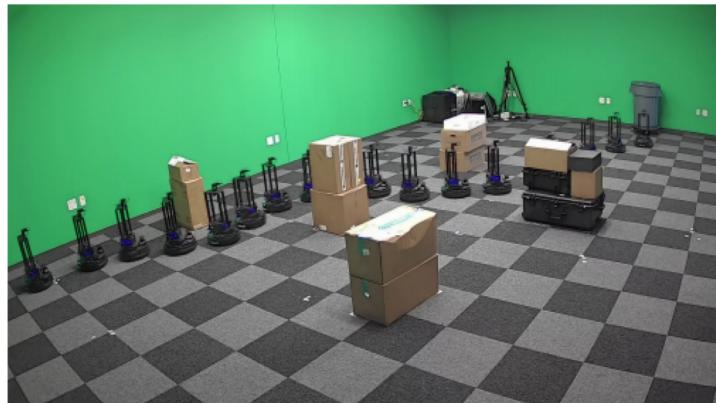
- Real robots: [PyRobot](#)



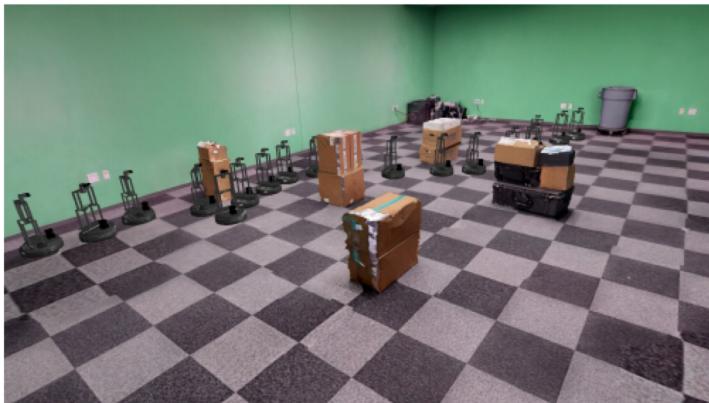
- Robot on your smartphone: [OpenBot](#)



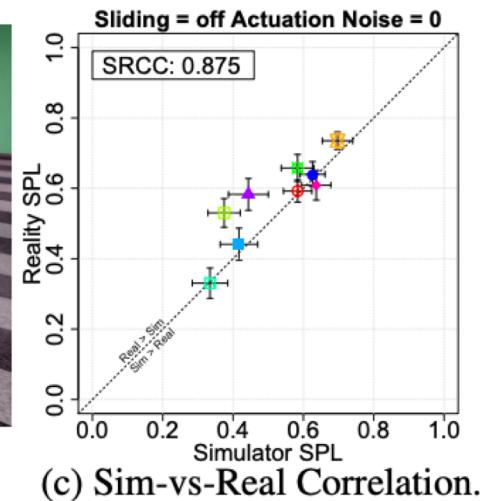
# Can we trust simulators?



(a) Reality.



(b) Simulation.



(c) Sim-vs-Real Correlation.

A. Kadian, J. Truong, A. Gokaslan, A. Clegg, E. Wijmans, S. Lee, M. Savva, S. Chernova, D. Batra, [Are We Making Real Progress in Simulated Environments? Measuring the Sim2Real Gap in Embodied Visual Navigation](#), arXiv 2019

# Can we scale up robot learning?



*Figure 1.* Our large-scale data collection setup, consisting of 14 robotic manipulators. We collected over 800,000 grasp attempts to train the CNN grasp prediction model.

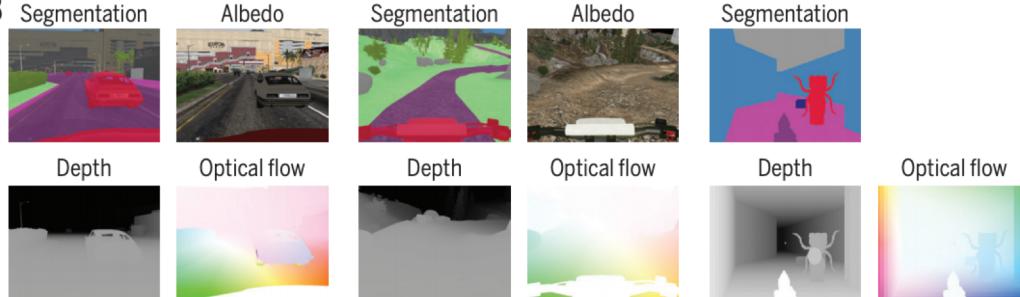
S. Levine et al., [Learning Hand-Eye Coordination for Robotic Grasping with Deep Learning and Large-Scale Data Collection](#), *International Symposium on Experimental Robotics*, 2017

# What internal representations are needed for embodied tasks?

**A**

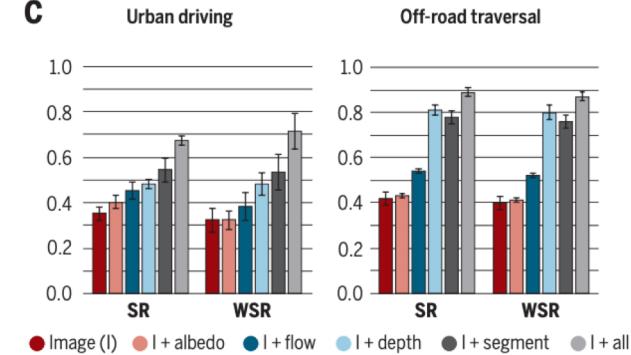


**B**

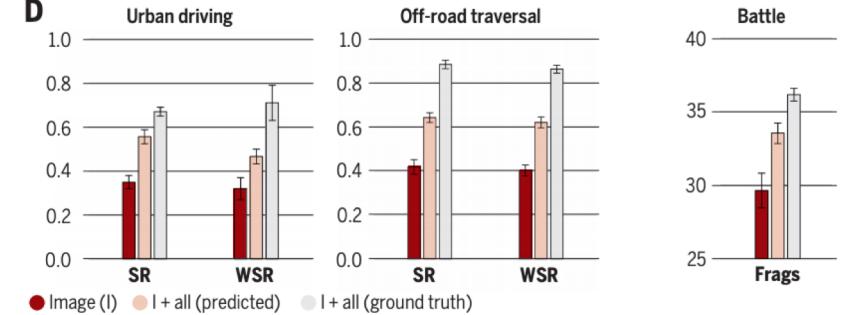


“Our main finding is that computer vision does matter. Models equipped with intermediate representations train faster, achieve higher task performance, and generalize better to previously unseen environments.”

**C**



**D**



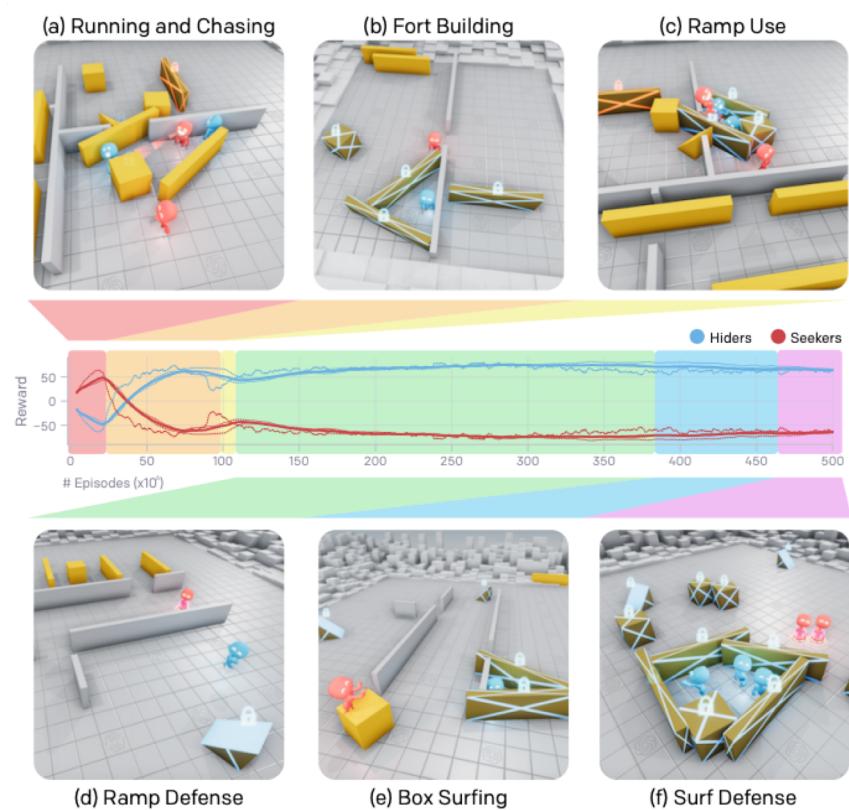
B. Zhou, P. Krähenbühl, and V. Koltun, [Does Computer Vision Matter for Action?](#) Science Robotics, 4(30), 2019 ([video](#))

# Multiple agents and communication



[Source](#)

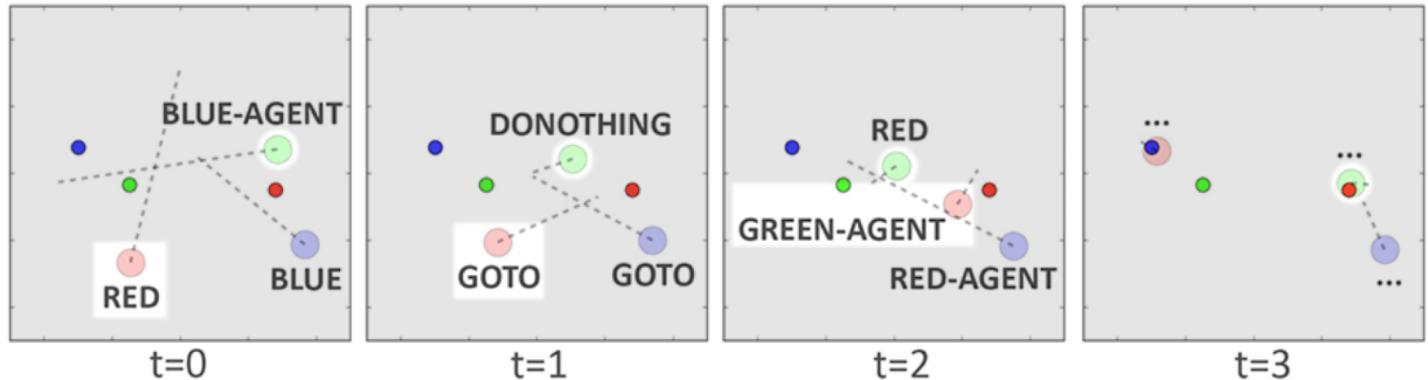
# Hide and seek



<https://openai.com/blog/emergent-tool-use/>

B. Baker et al., [Emergent tool use from multi-agent autocurricula](#), ICLR 2020

# Emergence of language



In the above step-by-step run, at t=0 the red agent says a word corresponding to the red landmark (center right), then at t=1 says a word that is equivalent to 'Goto', then in t=2 says 'green-agent'. The green-agent hears its instructions and immediately moves to the red landmark.

<https://openai.com/blog/learning-to-communicate/>  
Video

A. Mordatch and P. Abbeel, [Emergence of Grounded Compositional Language in Multi-Agent Populations](#), AAAI 2018

# RL for negotiation

Divide these objects between you and another Turker. Try hard to get as many points as you can!

Send a message now, or enter the agreed deal!

Items	Value	Number You Get
	8	<input type="button" value="1"/>
	1	<input type="button" value="1"/>
	0	<input type="button" value="0"/>

**Mark Deal Agreed**

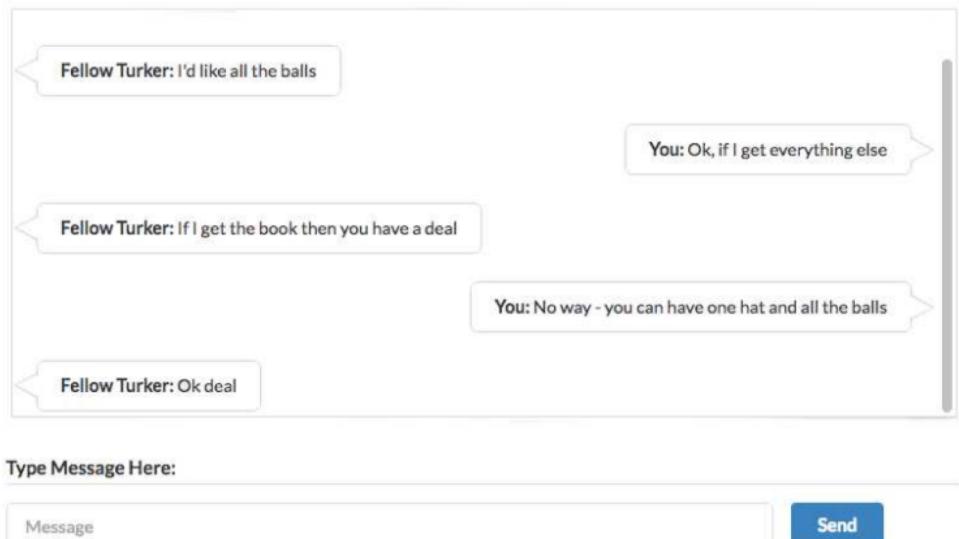


Figure 1: A dialogue in our Mechanical Turk interface, which we used to collect a negotiation dataset.

# RL for negotiation

Input		
Human	<b>1xbook value=3 1xhat value=1 3xball value=2</b>	
RL+ROLLOUTS	<b>1xbook value=6 1xhat value=4 3xball value=0</b>	
<hr/>		
RL+ROLLOUTS	You can have the balls if i can have the hat and book	
Human	Nope, give me the books and you can have everything else	
RL+ROLLOUTS	Ok. I will take the book and one hat	
Human	I get all the balls?	
RL+ROLLOUTS	Yes. I can do that.	
Human	Ok, deal	
Output		Reward
RL+ROLLOUTS	<b>1xbook 1xhat</b>	10/10
Human	<b>3xball</b>	6/10

Figure 5: Example of a model stubbornly holding out for maximum points in a dialogue.

Input		Reward
Human	RL+ROLLOUTS	
<b>3xbook value=2 3xhat value=1 1xball value=1</b>	<b>3xbook value=1 3xhat value=2 1xball value=1</b>	
Human	I'd like the books and the hats.	
RL+ROLLOUTS	I will take the books and hats	
Human	I really need the books and at least one other object.	
RL+ROLLOUTS	Ok, you can have one book and one ball	
Human	If I can have all the books, I can leave you the rest.	
RL+ROLLOUTS	Deal	
Output		Reward
RL+ROLLOUTS	<b>3xhat 1xball</b>	7/10
Human	<b>3xbook</b>	6/10

Figure 6: Example of model compromising.

Input			Reward
RL+ROLLOUTS	Human		
<b>4xbook value=0 2xhat value=5 1xball value=0</b>	<b>4xbook value=1 2xhat value=2 1xball value=2</b>		
RL+ROLLOUTS	You can have the ball if I can have everything else.		
Human	No can't do that the ball is not going to give me anything		
RL+ROLLOUTS	I need the hats and you can have the rest		
Human	Deal		
RL+ROLLOUTS	Great! Thank you!		
Output		Reward	
RL+ROLLOUTS	<b>2xhat</b>	10/10	
Human	<b>4xbook 1xball</b>	6/10	

Figure 7: Dialogue in which the model's initial interest in the valueless books allows it to compromise while achieving a maximum score.

# Scary AI?

Results are shown in Table 3, and show that the simple LIKELIHOOD model produces the most human-like responses, and the alternative training and decoding strategies cause a divergence from human language. Note however, that this divergence may not necessarily correspond to lower quality language—it may also indicate different strategic decisions about what to say. Results in §6.4 show all models could converse with humans.

Bob: I can i i everything else

Alice: balls have zero to me to me to me to me to me to me to

Bob: you i everything else

Alice: balls have a ball to me to me to me to me to me to me to me

GIZMODO

We come from the future



Shop



Subscribe



HOME LATEST REVIEWS EARTHER SCIENCE IO9 FIELD GUIDE VIDEO

FACEBOOK

## No, Facebook Did Not Panic and Shut Down an AI Program That Was Getting Dangerously Smart



Tom McKay

7/31/17 8:31PM • Filed to: SKYNET ▾



153



14



<https://gizmodo.com/no-facebook-did-not-panic-and-shut-down-an-ai-program-1797414922>  
See also: Dhruv Batra's response

# Outline

- Vision
- Embodied learning (RL, robotics)
- Language

# Large-scale language models

Model Alias	Org.	Article Reference
ULMfit	fast.ai	<i>Universal Language Model Fine-tuning for Text Classification</i> Howard and Ruder
 ELMo	AllenNLP	<i>Deep contextualized word representations</i> Peters et al.
OpenAI GPT	OpenAI	<i>Improving Language Understanding by Generative Pre-Training</i> Radford et al.
 BERT	Google	<i>BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding</i> Devlin et al.
XLM	Facebook	<i>Cross-lingual Language Model Pretraining</i> Lample and Conneau

[Image source](#)

# Should we be scared of GPT-3?

Opinion

The New York Times

## How Do You Know a Human Wrote This?

Machines are gaining the ability to write, and they are getting terrifyingly good at it.



By **Farhad Manjoo**  
Opinion Columnist

July 29, 2020

<https://www.nytimes.com/2020/07/29/opinion/gpt-3-ai-automation.html>

See also:

<https://www.nytimes.com/2020/11/24/science/artificial-intelligence-ai-gpt3.html>

MIT Technology Review

Opinion

## GPT-3, Bloviator: OpenAI's language generator has no idea what it's talking about

Tests show that the popular AI still has a poor grasp of reality.

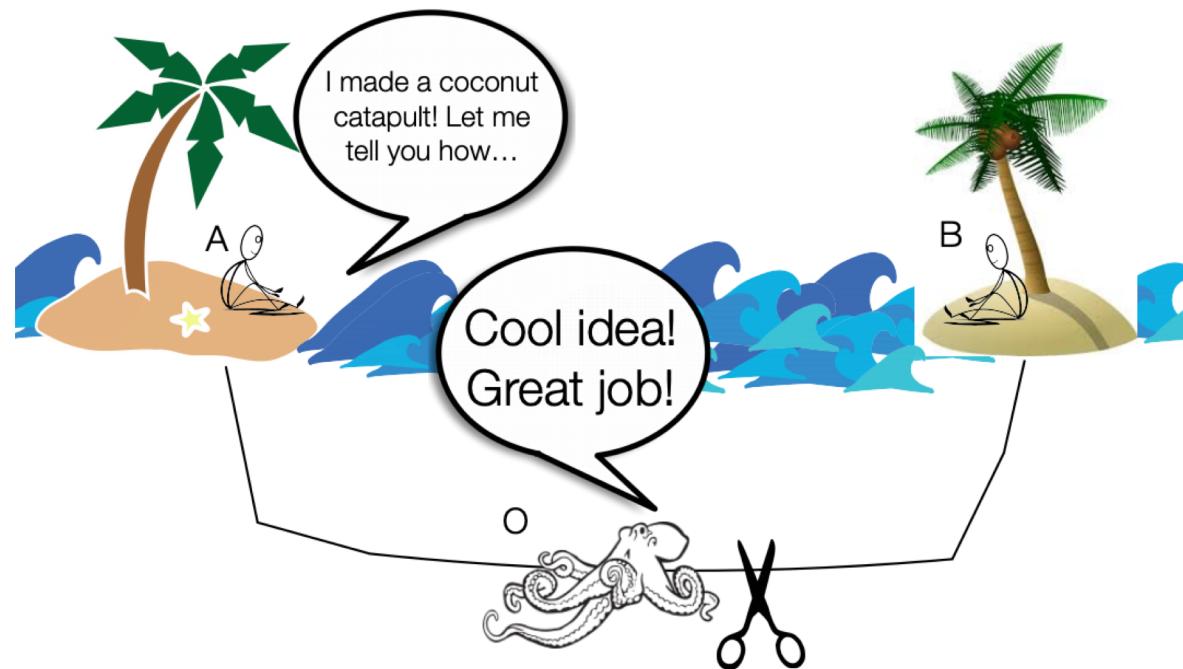
by **Gary Marcus** and **Ernest Davis**

August 22, 2020

<https://www.technologyreview.com/2020/08/22/1007539/gpt3-openai-language-generator-artificial-intelligence-ai-opinion/>

# Language: The importance of grounding

- *Meaning* cannot be learned from *form* alone – it requires knowing about the relationship between language and the outside world



[Image source](#)

# Further challenges

- Reasoning
- Memory
- Lifelong learning

# Parting thoughts

- **Embodied learning is the way to go! However...**
- The next breakthroughs are not likely to come cheaply
- Access to data, computation, and platforms will be key
- The next few years will make it clearer which problems have been truly solved and which ones have been underestimated
- The hard problems are getting into “AI-complete” territory