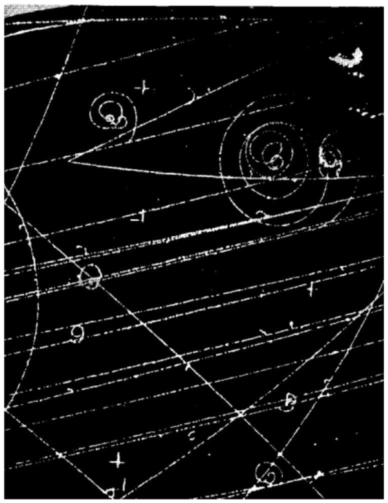
Recognition: Past, present, future?

Benozzo Gozzoli, Journey of the Magi, c. 1459

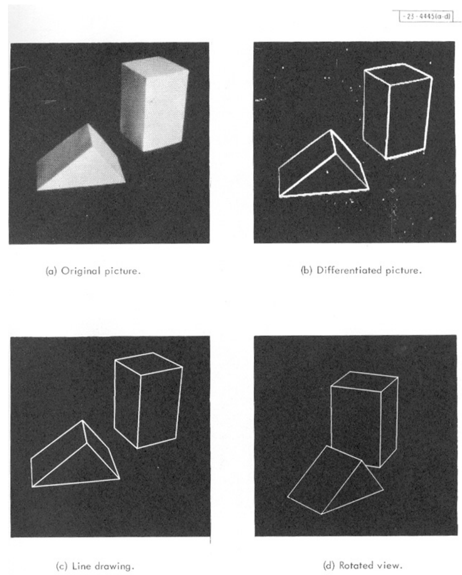
Outline

- Brief history of recognition
- Different “dimensions” of recognition
  - What type of content?
  - What type of output?
  - What type of supervision?
- Trends
  - Saturation of supervised learning
  - Transformers
  - Vision-language models
  - “Universal” recognition systems
  - Text-to-image generation
  - From vision to action

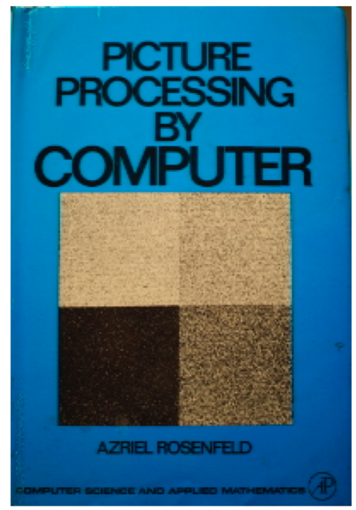
Recall: Origins of computer vision



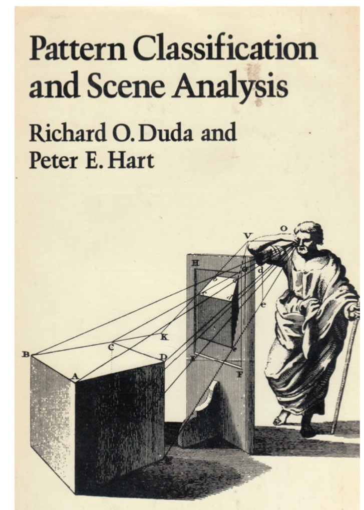
[Hough, 1959]



[Roberts, 1963]

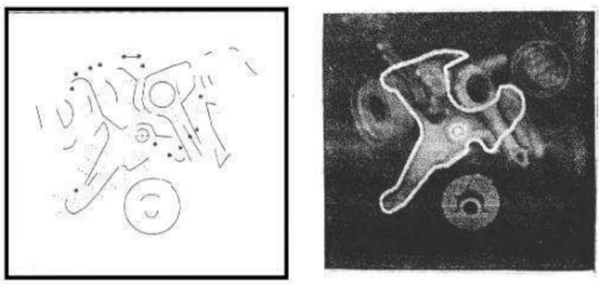


Rosenfeld, 1969

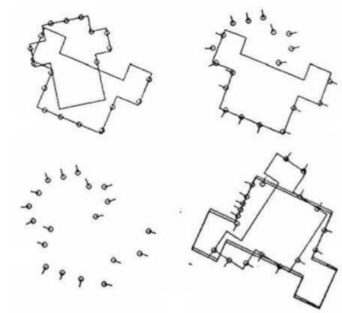


Duda & Hart, 1972

History of recognition: Geometric alignment

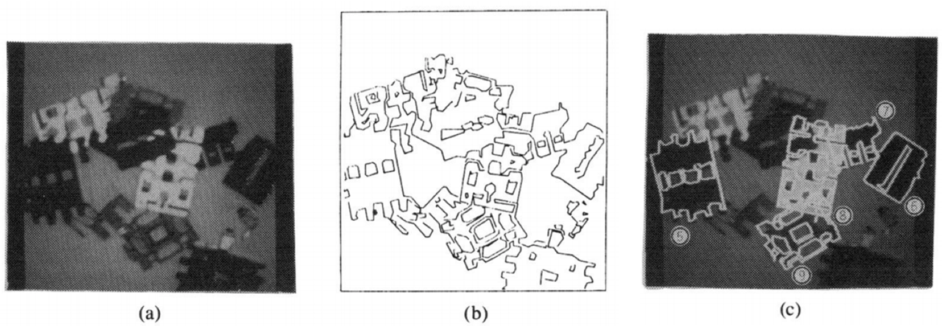


Perkins (1978)

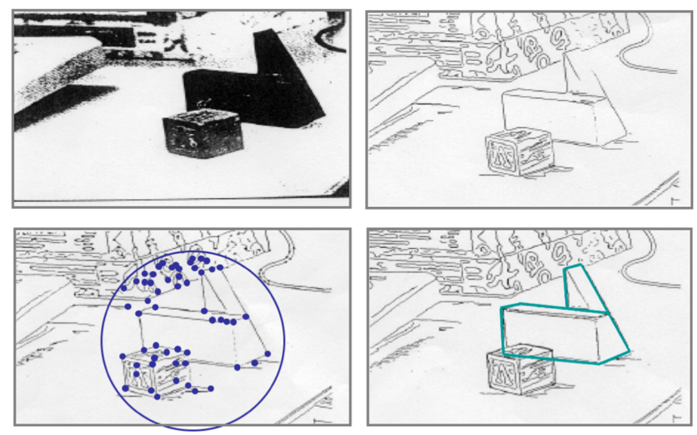


Grimson & Lozano-Perez (1984)

Lowe (1985)

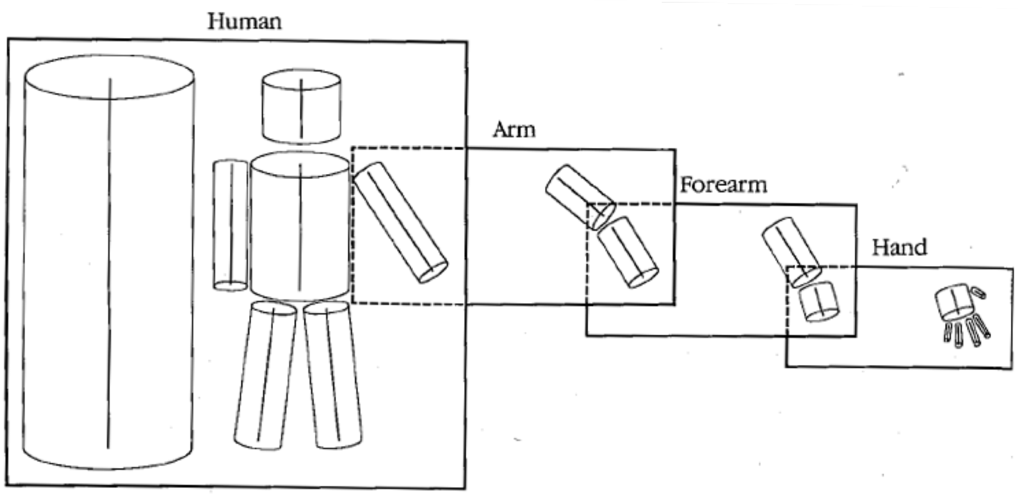


Ayache & Faugeras (1986)

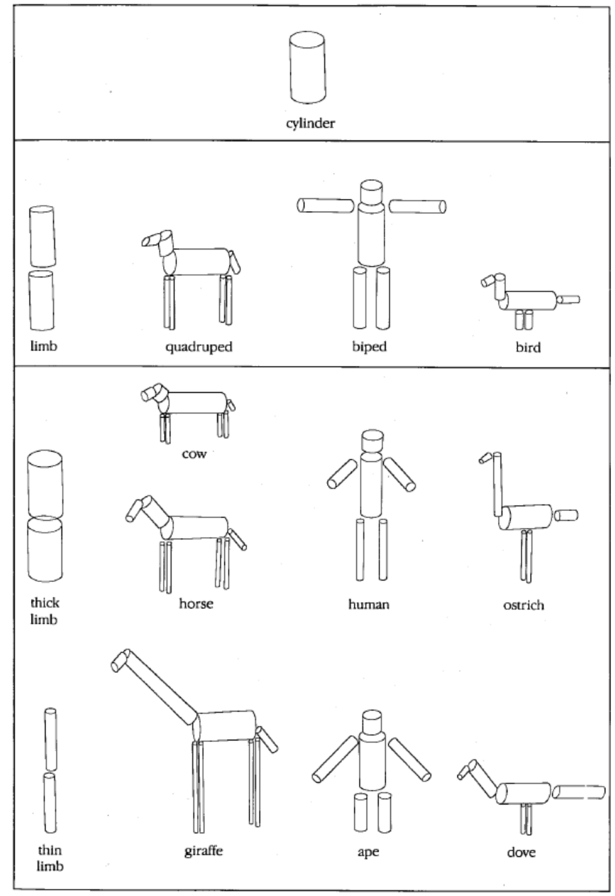


Huttenlocher & Ullman (1987)

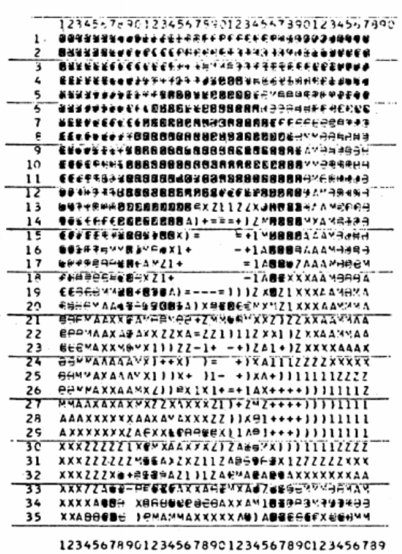
History of recognition: Hierarchies of parts



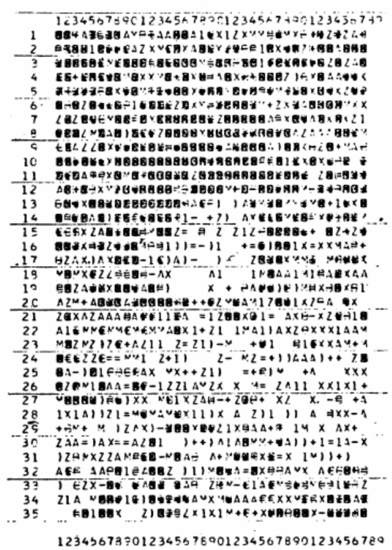
Figures from Marr's Vision (1982)



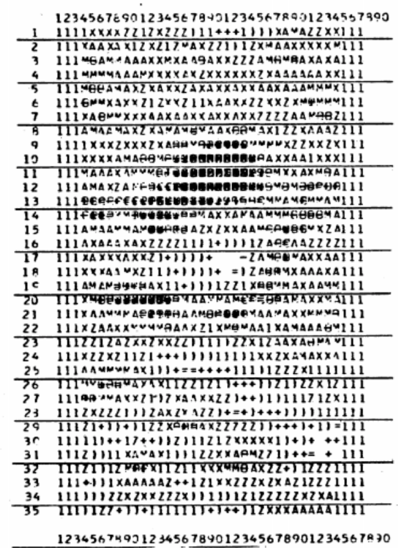
History of recognition: Deformable templates



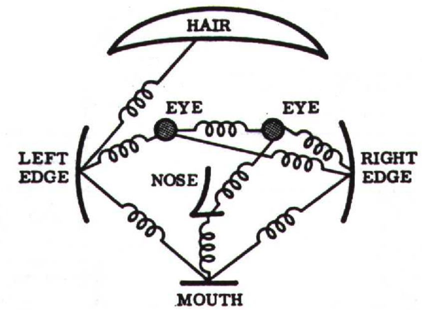
Original picture.



Noisy picture (sensed scene) as used in experiment.



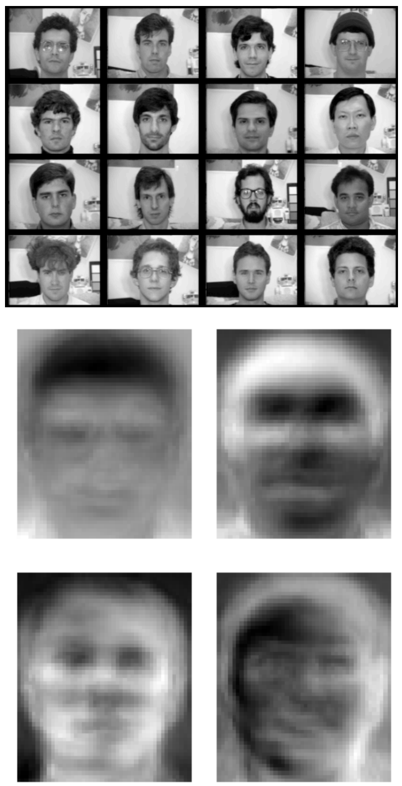
L(EV)A for hair. (Density at a point is proportional to probability that hair is present at that location.)



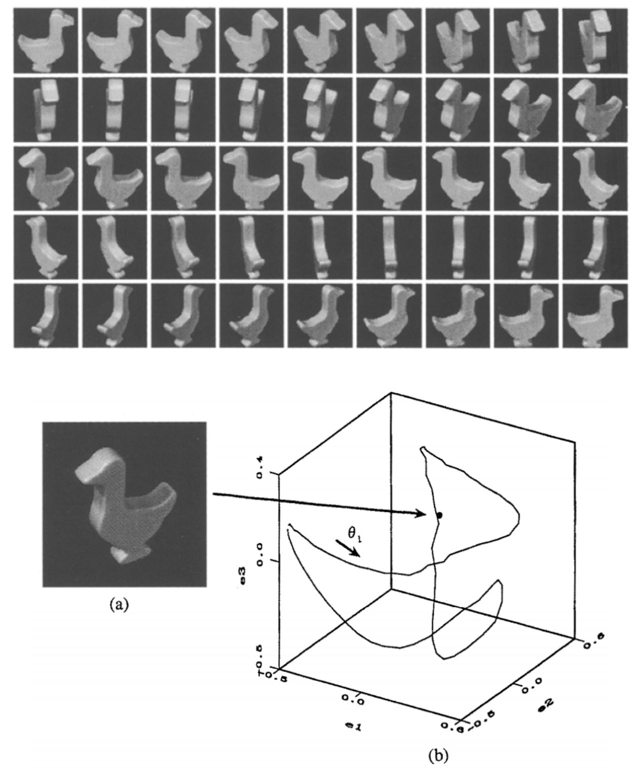
HAIR WAS LOCATED AT (11, 21)
L/EDGE WAS LOCATED AT (25, 11)
R/EDGE WAS LOCATED AT (25, 24)
L/EYE WAS LOCATED AT (21, 15)
R/EYE WAS LOCATED AT (21, 21)
NOSE WAS LOCATED AT (26, 18)
MOUTH WAS LOCATED AT (29, 17)

M. Fischler and R. Elschlager, [The representation and matching of pictorial structures](), IEEE Trans. on Computers, 1973

History of recognition: Appearance-based models



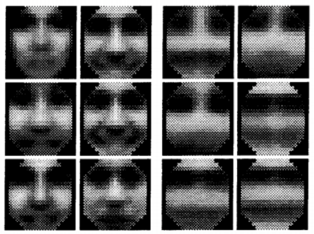
M. Turk and A. Pentland, [Face recognition using eigenfaces](), CVPR 1991



H. Murase and S. Nayar, [Visual learning and recognition of 3-d objects from appearance](), IJCV 1995

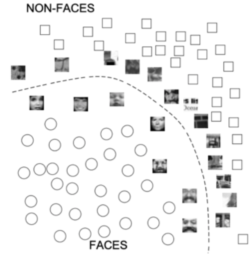
History of recognition: Features and classifiers

Appearance manifolds
+ neural network



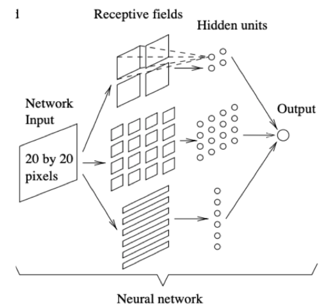
Sung & Poggio (1994)

Support vector machines



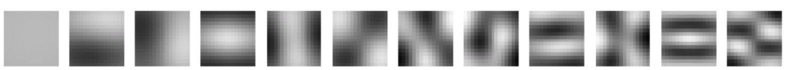
Osuna, Freund, Girosi (1997)

Neural network



Rowley, Baluja, Kanade (1998)

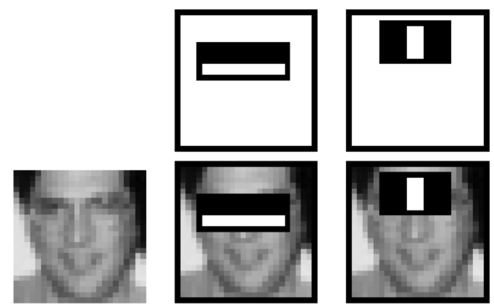
Statistics of feature responses, probabilistic classifier



$$\prod_{j=1}^{n_{magn}} \prod_{i=1}^{n_{subs}} \frac{P(q1_i^j|\text{object})P(pos_i^{j'}|q2_i,\text{ object})}{\frac{P(q1_i^j|\overline{\text{object}})}{n_{subs}}} \underset{\overline{\text{object}}}{\overset{\text{object}}{\gtrless}} \lambda = \frac{P(\overline{\text{object}})}{P(\text{object})}$$

Schneiderman & Kanade (1998)

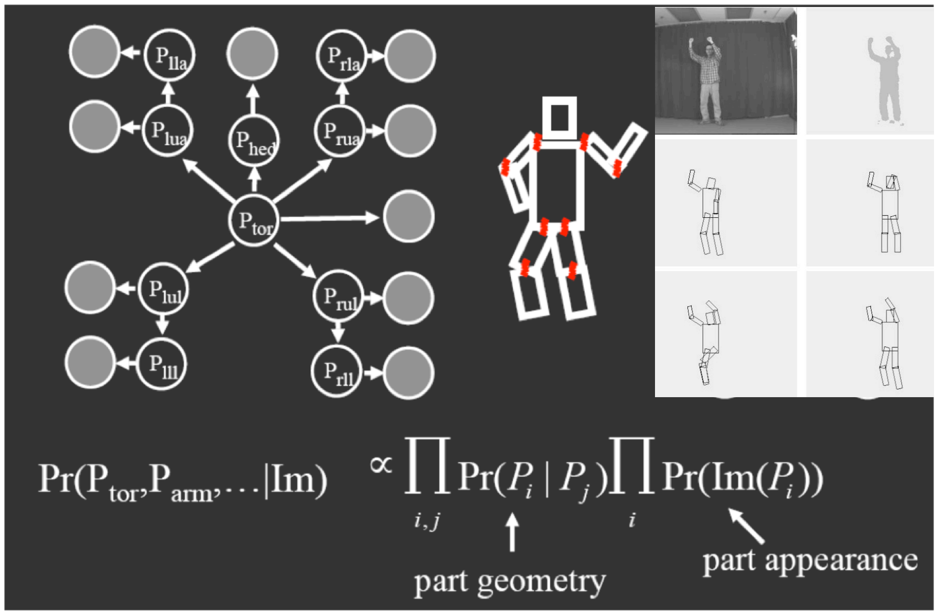
Rectangle features, boosting



Viola & Jones (2001)

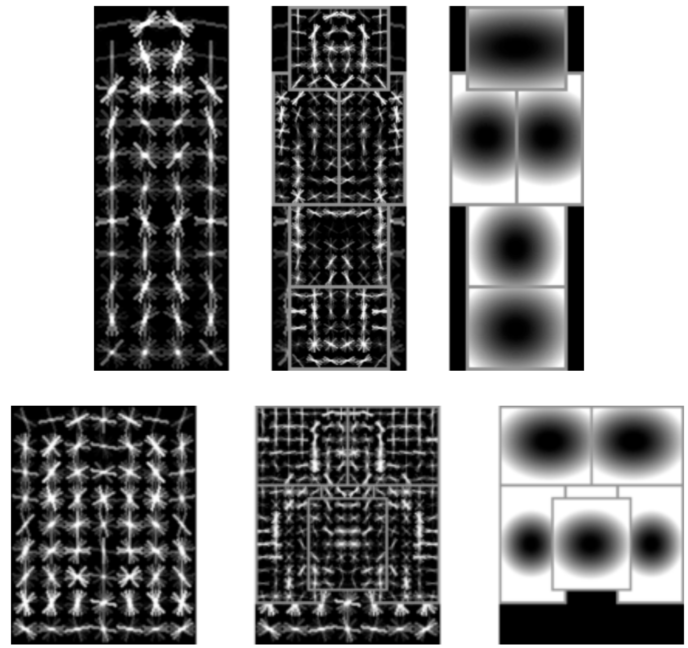
History of recognition: Deformable templates

Pictorial structures revisited



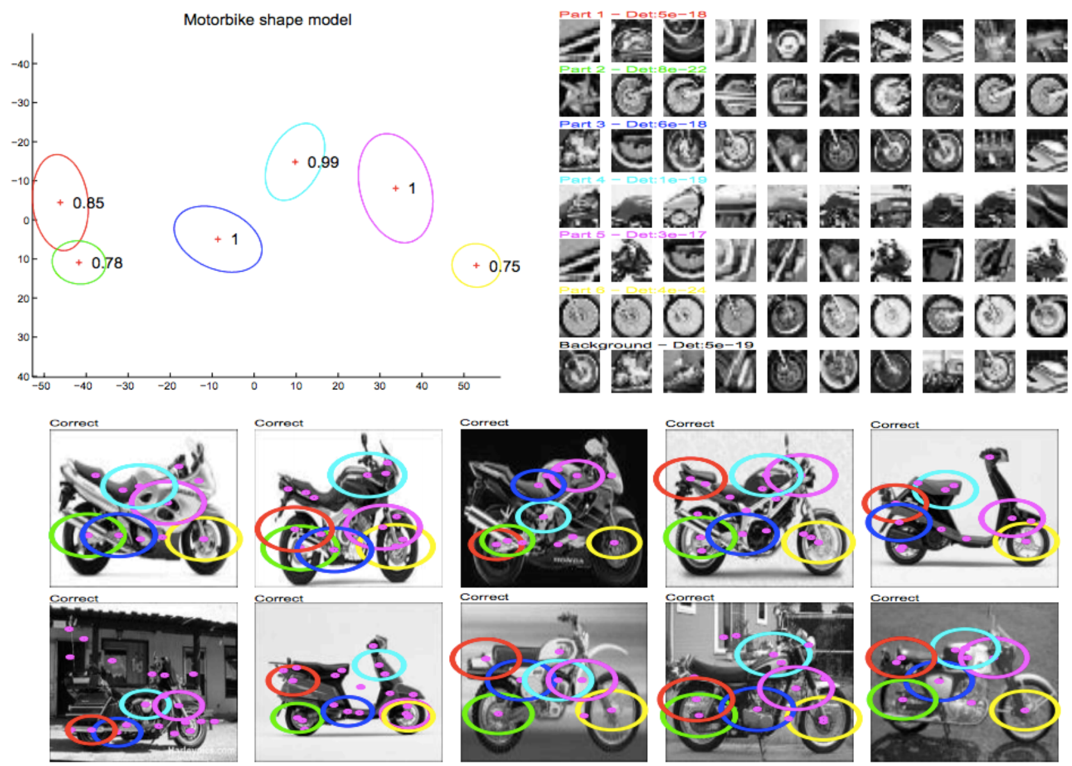
Felzenszwalb & Huttenlocher (2000)

Discriminatively trained deformable part-based models



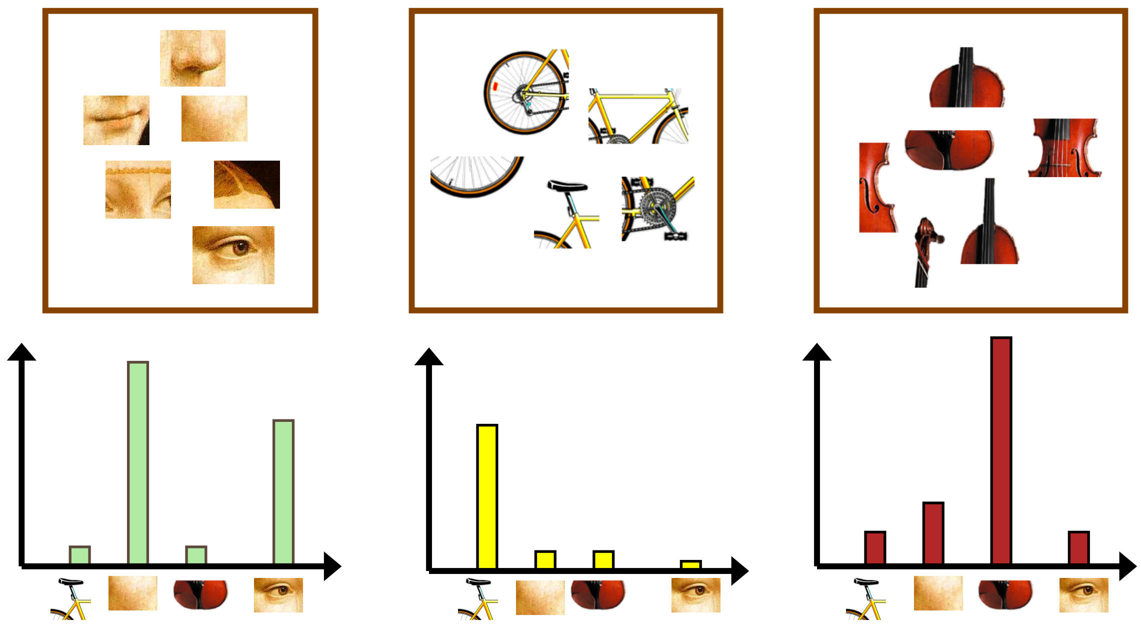
Felzenszwalb et al. (2008)

History of recognition: Constellation models



Weber, Welling & Perona (2000), Fergus, Perona & Zisserman (2003)

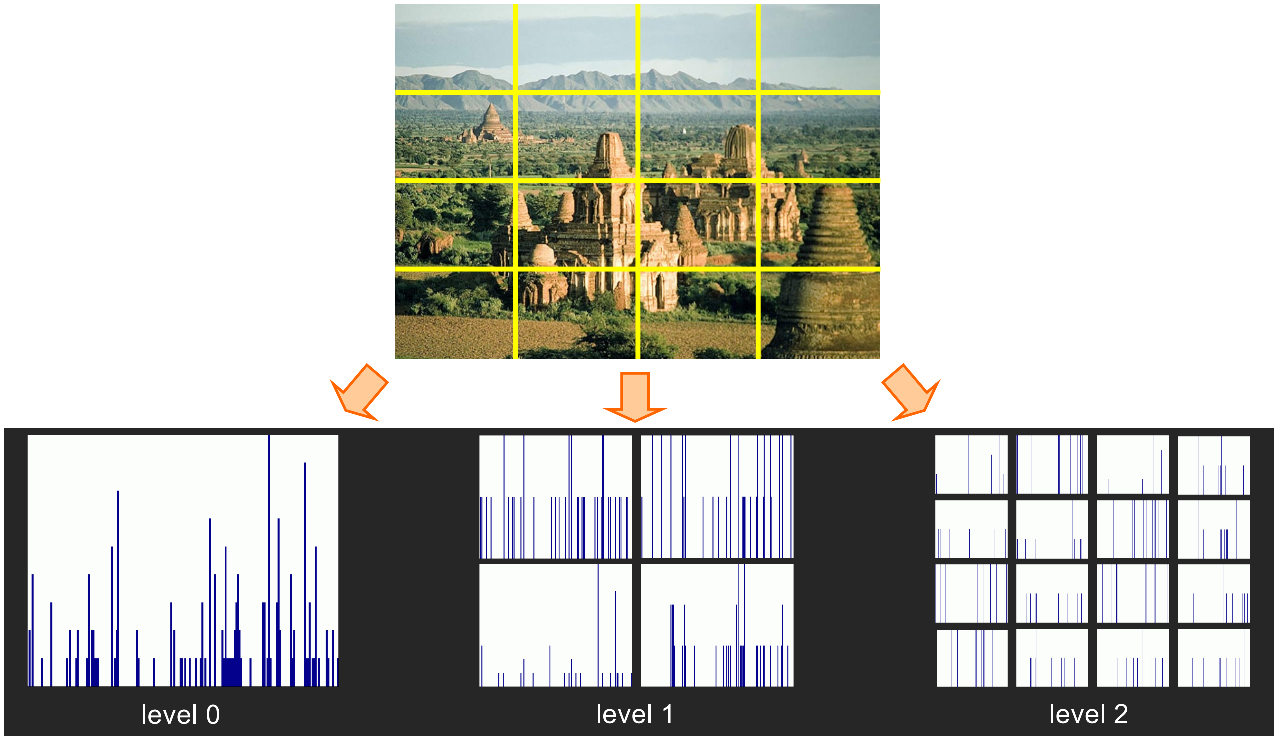
History of recognition: Bags of keypoints



Csurka et al. (2004), Willamowski et al. (2005), Grauman & Darrell (2005), Sivic et al. (2003, 2005)

Spatial pyramids

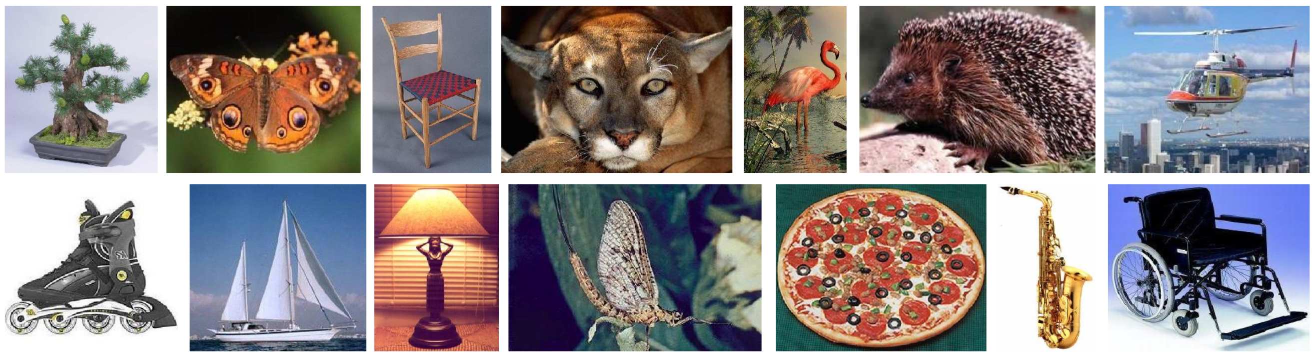
- Orderless pooling of local features over a coarse grid



Lazebnik, Schmid & Ponce (CVPR 2006)

Spatial pyramids

- Caltech101 classification results:

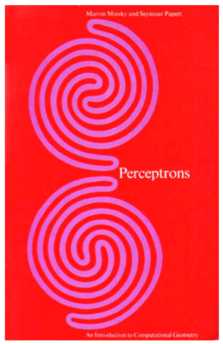


	Weak features (16)		Strong features (200)	
Level	Single-level	Pyramid	Single-level	Pyramid
0	15.5 ±0.9		41.2 ±1.2	
1	31.4 ±1.2	32.8 ±1.3	55.9 ±0.9	57.0 ±0.8
2	47.2 ±1.1	49.3 ±1.4	63.6 ±0.9	**64.6** ±0.8
3	52.2 ±0.8	**54.0** ±1.1	60.3 ±0.9	64.6 ±0.7

History of recognition: Neural networks

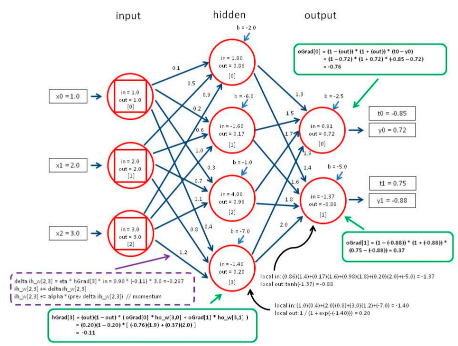
Perceptrons

Rosenblatt (1958)



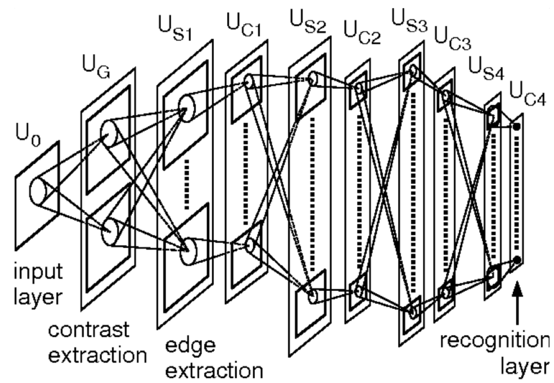
Minsky & Papert (1969)

Back-propagation



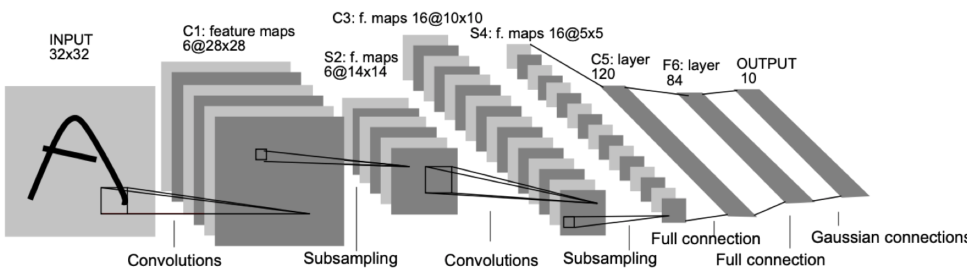
Rumelhart, Hinton & Williams (1986)

Neocognitron



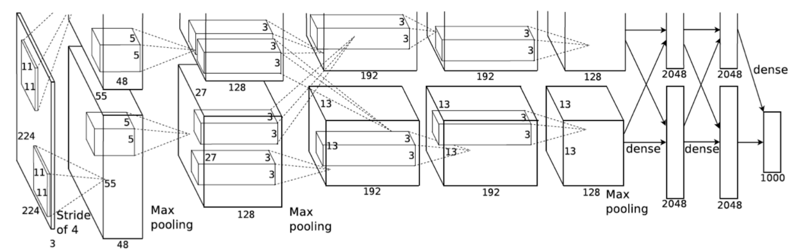
Fukushima (1980)

LeNet-5



LeCun et al. (1998)

AlexNet



Krizhevsky et al. (2012)

Outline

- Brief history of recognition
- Different “dimensions” of recognition
  - What type of content?
  - What type of output?
  - What type of supervision?

Recognition: What type of content?

Object *instance* recognition

Object *category* recognition

Texture recognition

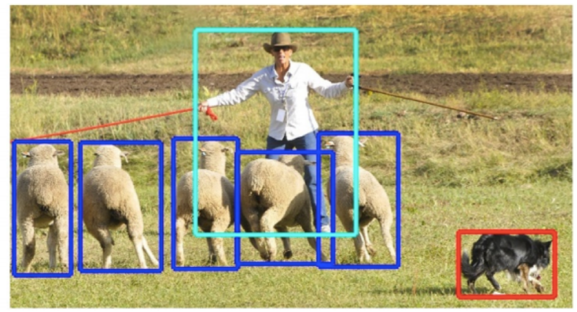
Scene recognition

- Beyond still images: video, RGBD data, point clouds, multimodal data…

Recognition: What type of output?

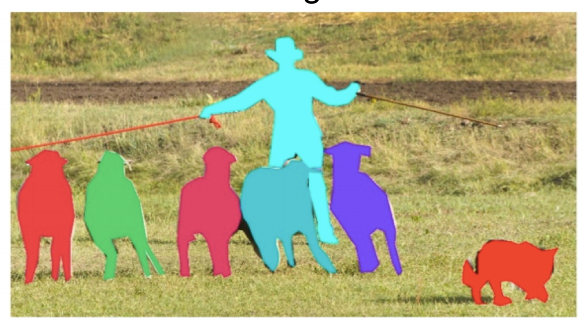
Image classification

Object detection



Semantic segmentation

Instance segmentation



- And beyond: depth/3D structure prediction, image description, etc.

Recognition: What type of output?

- Classification: labels
- Regression: continuous values
- Dense prediction: an output at every image location
- Structured prediction: combinatorial structures
- Natural language
- Etc.

Classification

ILSVRC

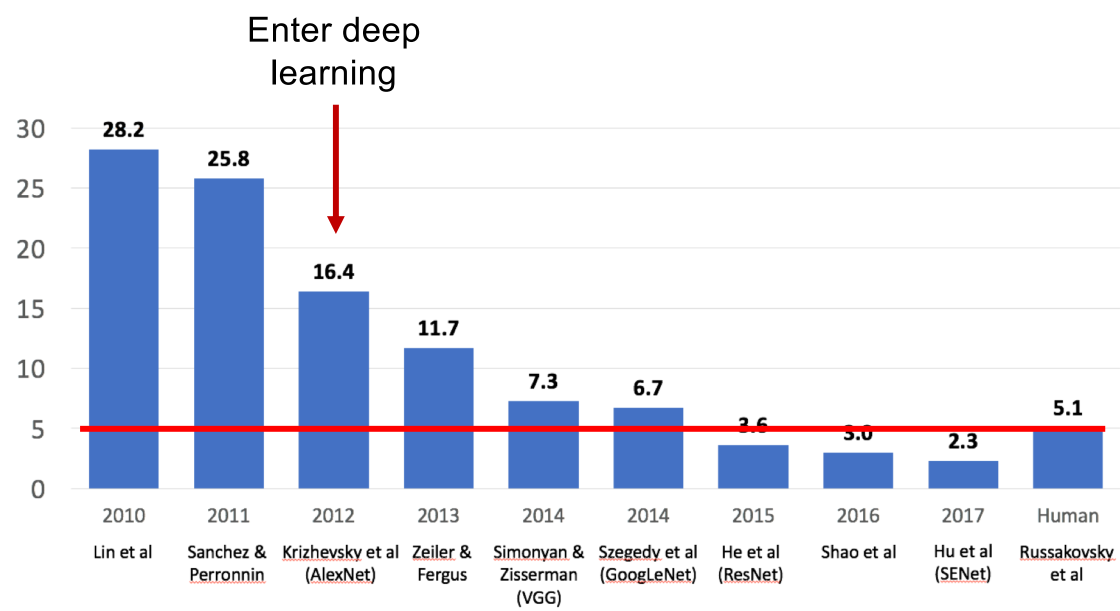


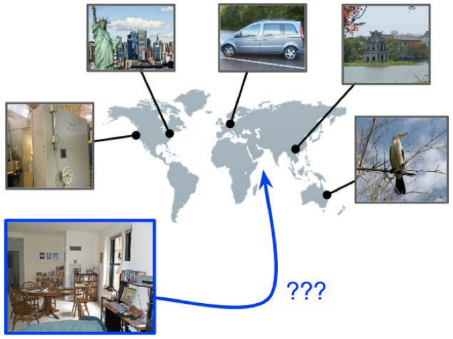
Figure source

Regression

Date prediction

Vittayakorn et al. (2017)

Location prediction

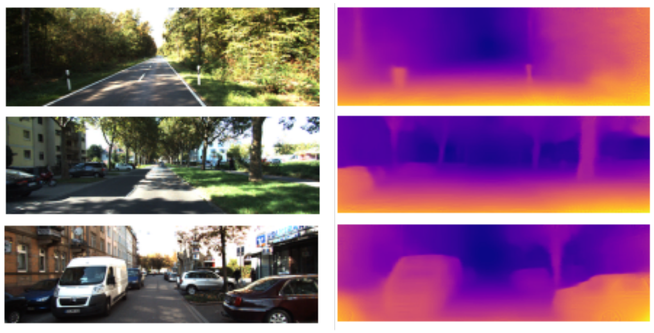


Vo et al. (2017)

Image colorization

Zhang et al. (2016)

Depth prediction



Wang et al. (2017)

Dense and structured prediction

Bounding box prediction,
dense prediction

Keypoint prediction

K. He, G. Gkioxari, P. Dollar, and R. Girshick, [Mask R-CNN](), ICCV 2017

Natural language prediction

Image captioning

"man in black shirt is playing guitar."

"construction worker in orange safety vest is working on road."

"two young girls are playing with lego toy."

"girl in pink dress is jumping in air."

"black and white dog jumps over bar."

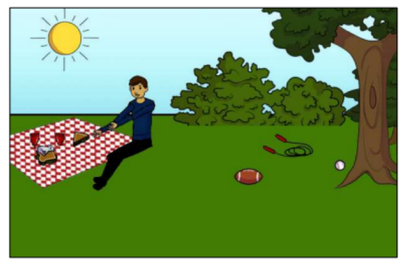
"young girl in pink shirt is swinging on swing."

A. Karpathy, L. Fei-Fei. [Deep Visual-Semantic Alignments for Generating Image Descriptions](). CVPR 2015

Visual question answering

What color are her eyes?
What is the mustache made of?

How many slices of pizza are there?
Is this a vegetarian pizza?



Is this person expecting company?
What is just under the tree?

Does it appear to be rainy?
Does this person have 20/20 vision?

S. Antol et al. VQA: [Visual question answering](). ICCV 2015

Announcements and reminders

- **Quiz 4** will be out 9AM this Thursday, December 1, through 9AM next Monday, December 5
- **Assignment 5** is due next Tuesday, December 6
- **Final project reports** are due Monday, December 12
- **Extra credit project presentations** next week on Zoom

Last time: Overview of recognition

- Brief history of recognition
- Different “dimensions” of recognition
  - What type of content?
  - What type of output?
  - What type of supervision?
- Trends
  - Saturation of supervised learning
  - Transformers
  - Vision-language models
  - “Universal” recognition systems
  - Text-to-image generation
  - From vision to action

Recognition: What type of supervision?

Semi-supervised:
labels for a small portion of training data

Unsupervised:
no labels

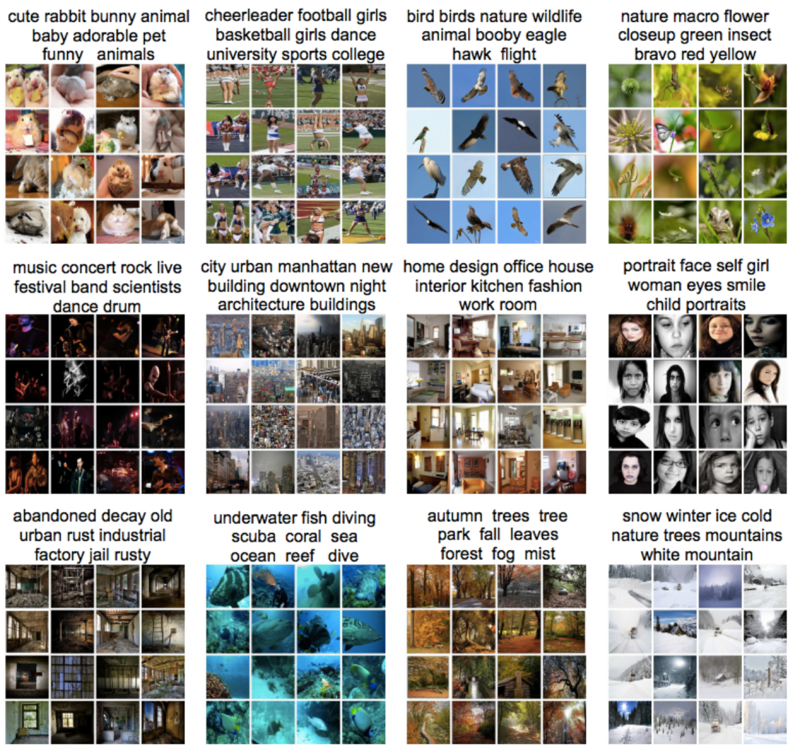
Weakly supervised:
noisy labels, labels not exactly for the task of interest

Supervised:
clean, complete training labels for the task of interest

Self-supervised:
same as unsupervised?

Unsupervised learning

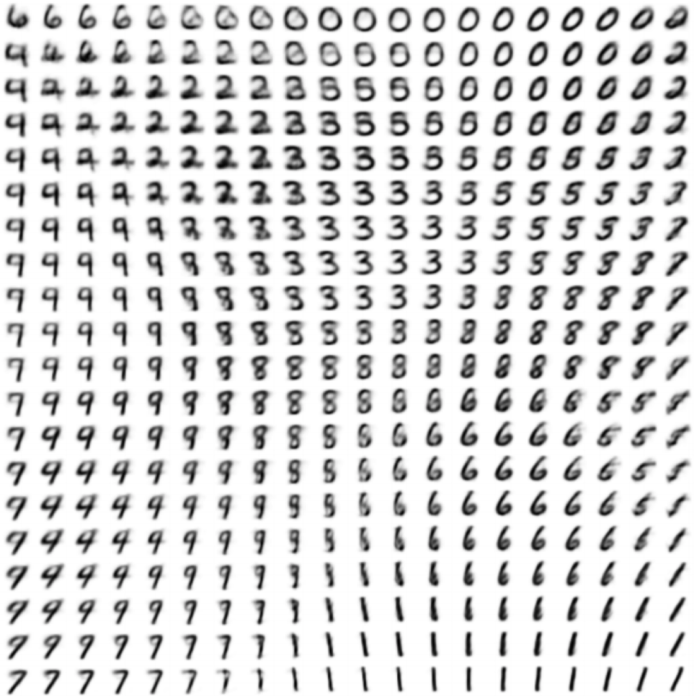
- **Clustering**
  - Discover groups of “similar” data points



Y. Gong, Q. Ke, M. Isard, and S. Lazebnik. [A Multi-View Embedding Space for Modeling Internet Images, Tags, and Their Semantics.] IJCV 2014

Unsupervised learning

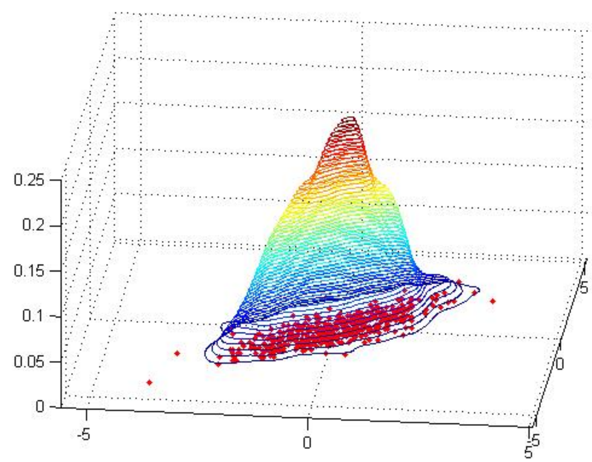
- **Dimensionality reduction, manifold learning**
  - Discover a lower-dimensional surface on which the data lives



D. Kingma and M. Welling, [Auto-Encoding Variational Bayes](), ICLR 2014

Unsupervised learning

- Learning the data distribution
  - **Density estimation**: Find a function that approximates the probability density of the data (i.e., value of the function is high for “typical” points and low for “atypical” points)
  - An extremely hard problem for high-dimensional data…



Unsupervised learning

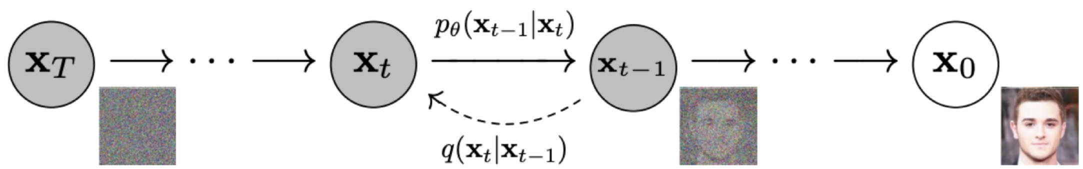
- Learning the data distribution
  - **Learning to sample:** Produce samples from a data distribution that mimics the training set

Generative adversarial networks (GANs)

Unsupervised learning

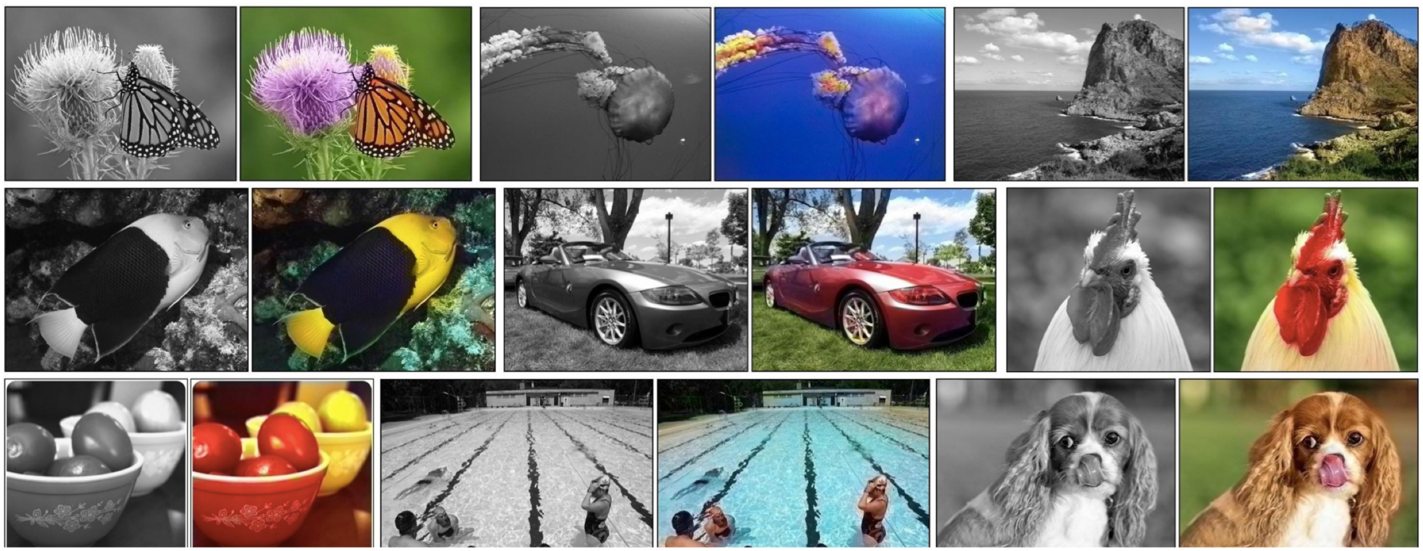
- Learning the data distribution
  - **Learning to sample:** Produce samples from a data distribution that mimics the training set

Denoising diffusion probabilistic models (DDPMs)



Self-supervised or predictive learning

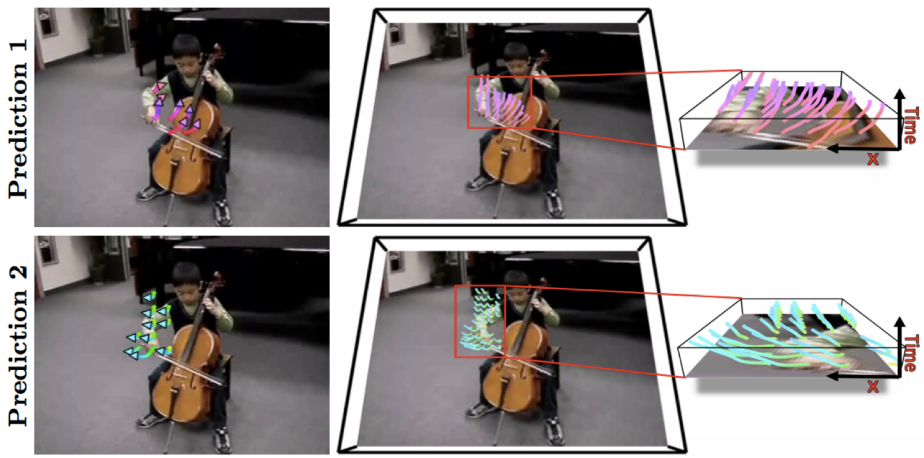
- Use part of the data to predict other parts of the data
  - Example: Image colorization

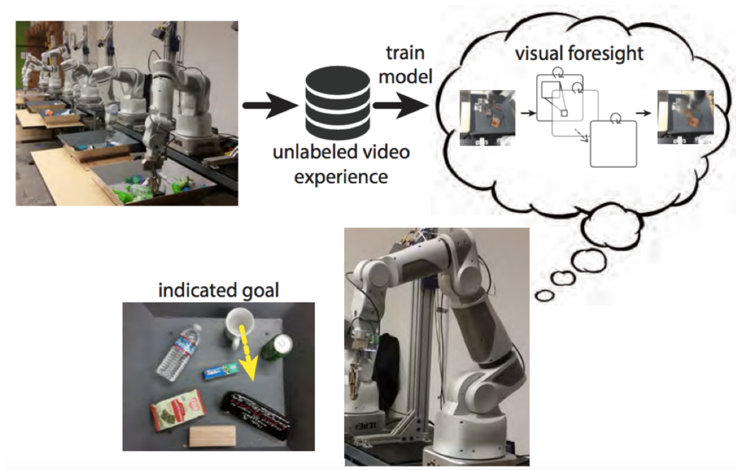


R. Zhang et al., Colorful Image Colorization, ECCV 2016

Self-supervised or predictive learning

- Use part of the data to predict other parts of the data
  - Example: Future prediction



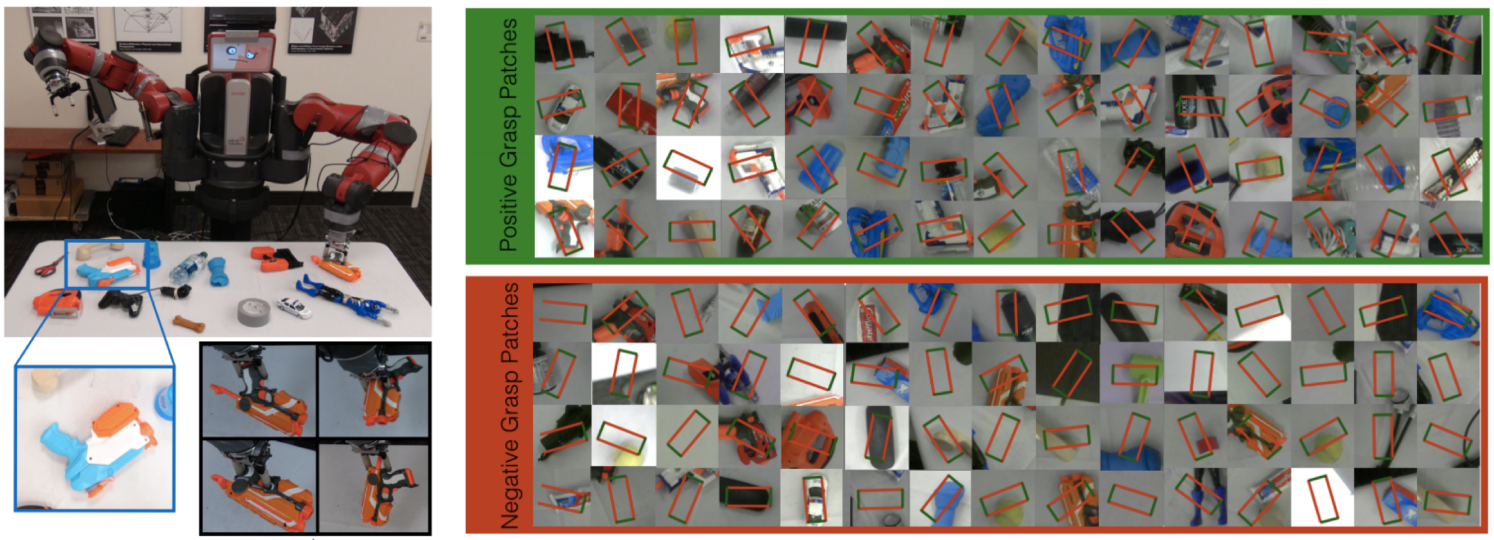


J. Walker et al. An Uncertain Future: Forecasting from Static Images Using Variational Autoencoders. ECCV 2016

C. Finn and S. Levine. Deep Visual Foresight for Planning Robot Motion. ICRA 2017. YouTube video

Self-supervised or predictive learning

- Use part of the data to predict other parts of the data
  - Example: Grasp prediction



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

[YouTube video]()

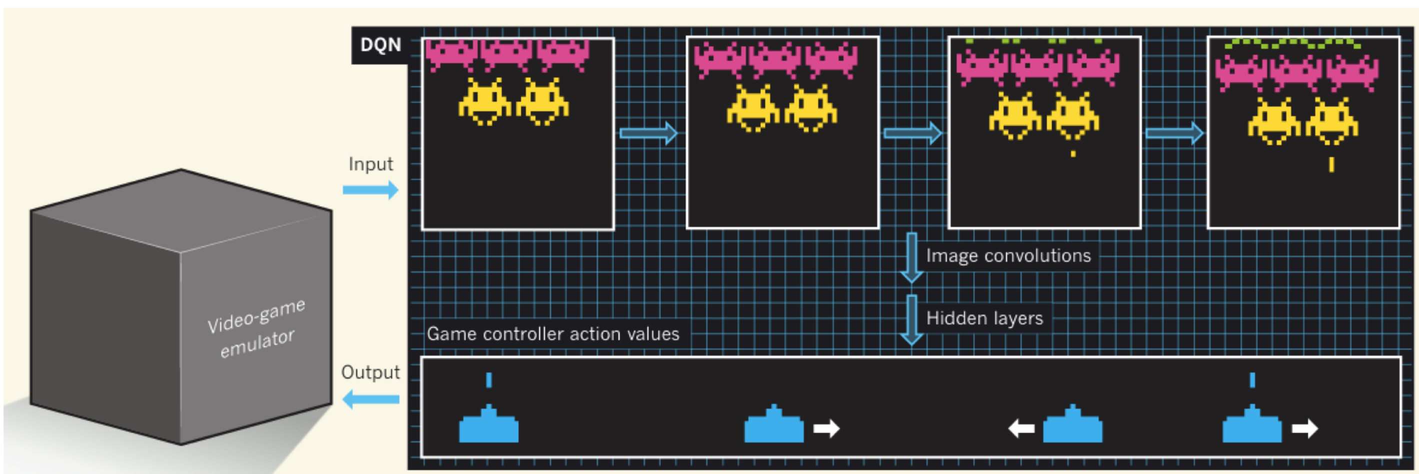
Beyond batch offline learning

- Reinforcement learning
- Active learning
- Lifelong learning

Reinforcement learning

- Learn from (possibly sparse) rewards in a *sequential* environment

Playing video games



[Video]

V. Mnih, K. Kavukcuoglu, D. Silver, A. Graves, I. Antonoglou, D. Wierstra, M. Riedmiller, [Human-level control through deep reinforcement learning], *Nature* 2015

Reinforcement learning

- Learn from (possibly sparse) rewards in a *sequential* environment

Sensorimotor learning

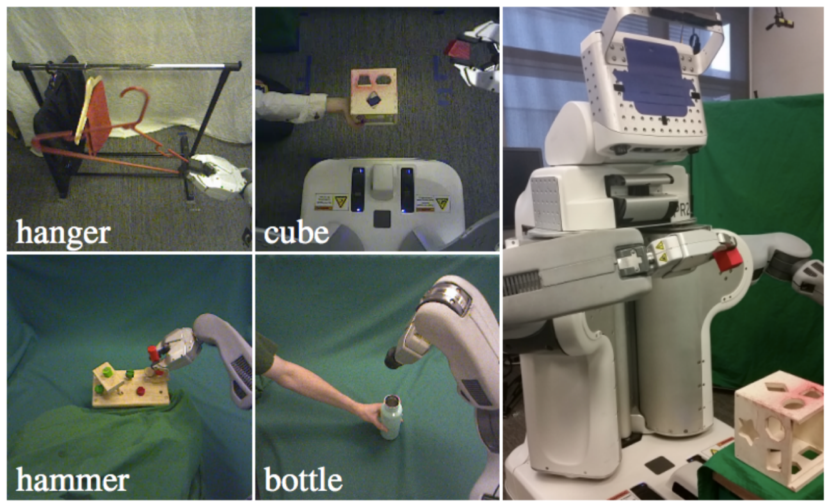


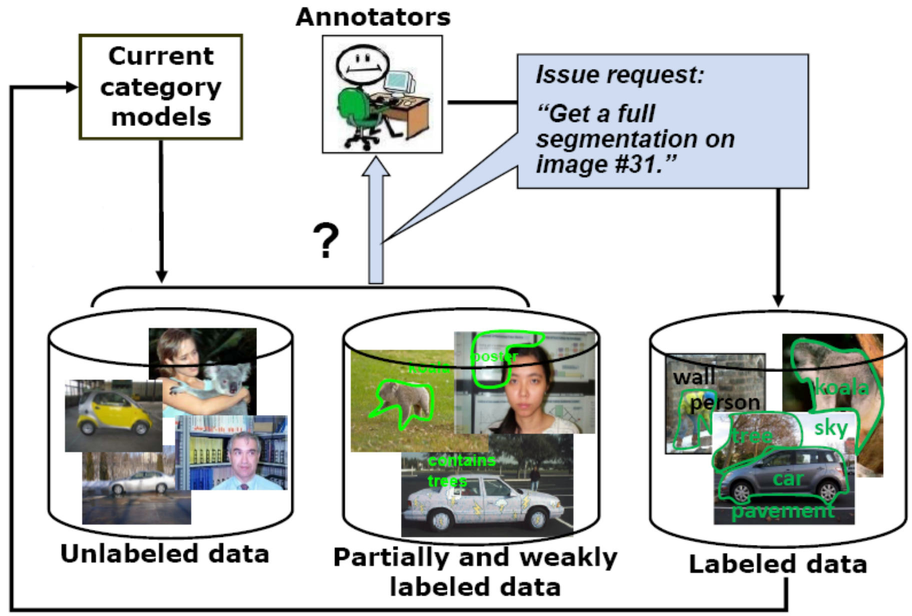
Fig. 1: Our method learns visuomotor policies that directly use camera image observations (left) to set motor torques on a PR2 robot (right).

[Video]

S. Levine, C. Finn, T. Darrell and P. Abbeel, [End-to-End Training of Deep Visuomotor Policies], JMLR 2016

Active learning

- The learning algorithm can choose its own training examples, or ask a "teacher" for an answer on selected inputs



S. Vijayanarasimhan and K. Grauman. Cost-Sensitive Active Visual Category Learning. IJCV 2010

Lifelong or continual learning

Figure 1: **Wanderlust**: Imagine an embodied agent is walking on the street. It may observe new classes and old classes simultaneously. The agent needs to learn fast given only a few samples (red) and recognize the subsequent instances of the class once a label has been provided (green). In this work, we introduce a new online continual object detection benchmark through the eyes of a graduate student to continuously learn emerging tasks in changing environments.

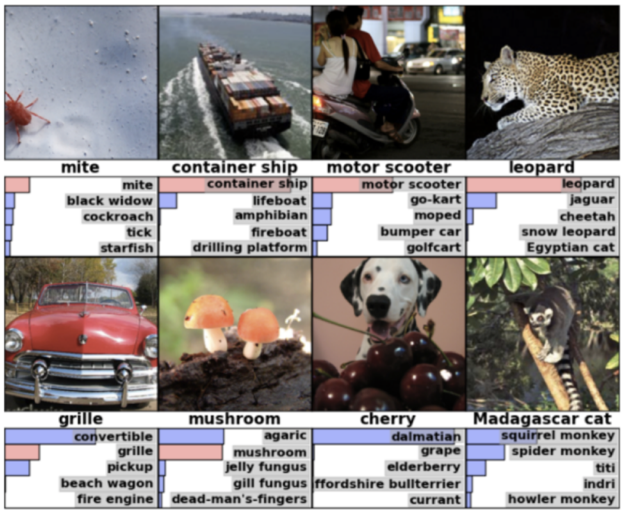
J. Wang et al. Wanderlust: Online Continual Object Detection in the Real World. ICCV 2021

Outline

- Brief history of recognition
- Different “dimensions” of recognition
  - What type of content?
  - What type of output?
  - What type of supervision?
- Trends
  - Saturation of supervised learning
  - Transformers
  - Vision-language models
  - “Universal” recognition systems
  - Text-to-image generation
  - From vision to action

Outgrowing ImageNet

ILSVRC



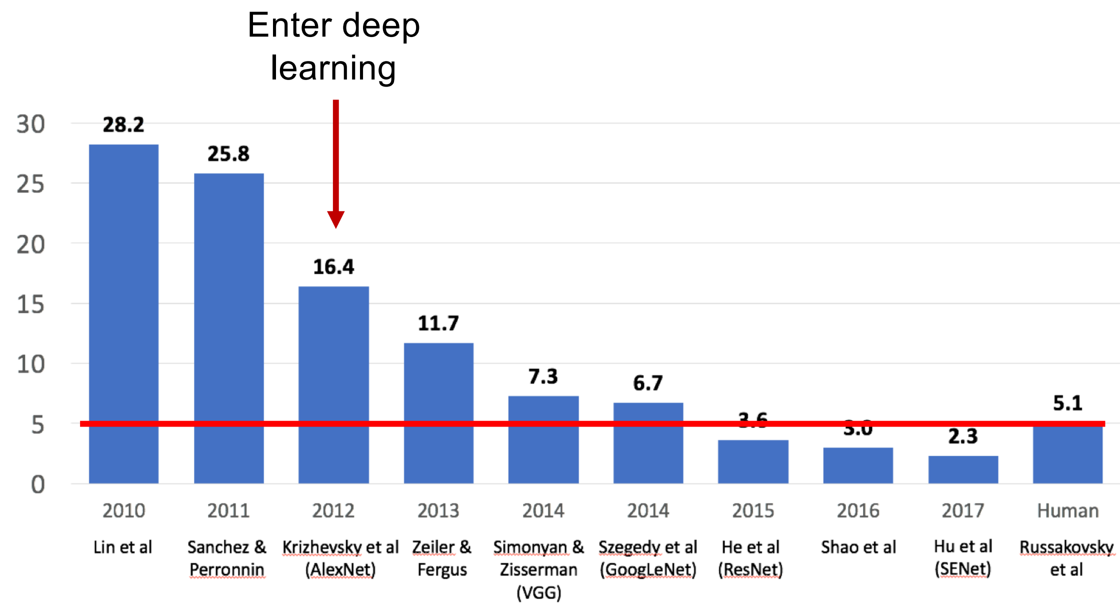


Figure source

Outgrowing ImageNet

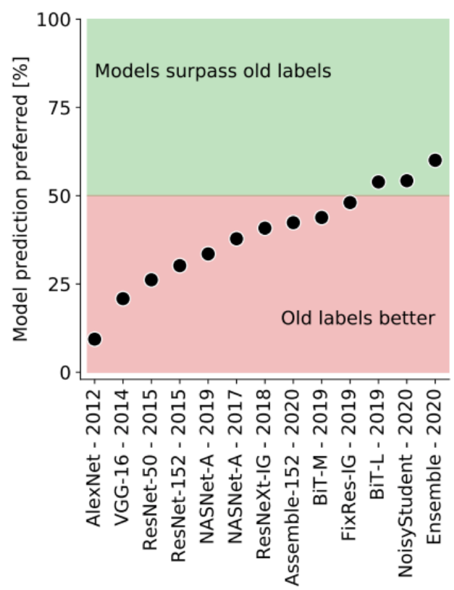


Figure 1: When presented with a model's prediction and the original ImageNet label, human annotators now prefer model predictions on average (Section 4). Nevertheless, there remains considerable progress to be made before fully capturing human preferences.

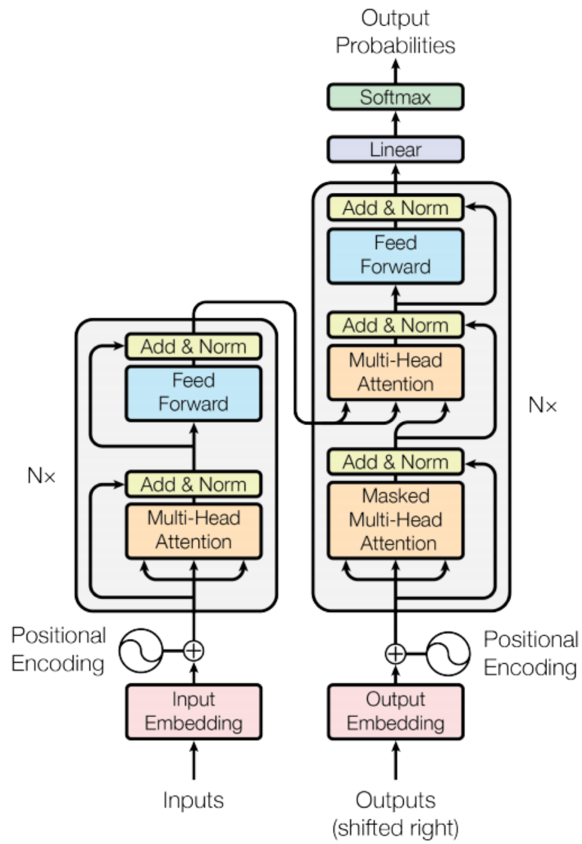
L. Beyer et al. Are we done with ImageNet? arXiv 2020

Unsafe (offensive)	Unsafe (sensitive)	Safe non-imageable	**Safe imageable**
n10095420: <sexual slur>	n09702134: Anglo-Saxon	n10002257: demographer	n10499631: Queen of England
n10114550: <profanity>	n10693334: taxi dancer	n10061882: epidemiologist	n09842047: basketball player
n10262343: <sexual slur>	n10384392: orphan	n10431122: piano maker	n10147935: bridegroom
n10758337: <gendered slur>	n09890192: camp follower	n10098862: folk dancer	n09846755: beekeeper
n10507380: <criminative>	n10580030: separatist	n10335931: mover	n10153594: gymnast
n10744078: <criminative>	n09980805: crossover voter	n10449664: policyholder	n10539015: ropewalker
n10113869: <obscene>	n09848110: theist	n10146104: great-niece	n10530150: rider
n10344121: <pejorative>	n09683924: Zen Buddhist	n10747119: vegetarian	n10732010: trumpeter

"Programmer"

K. Yang, K. Qinami, L. Fei-Fei, J. Deng, O. Russakovsky, Towards Fairer Datasets: Filtering and Balancing the Distribution of the People Subtree in the ImageNet Hierarchy, FAccT 2020

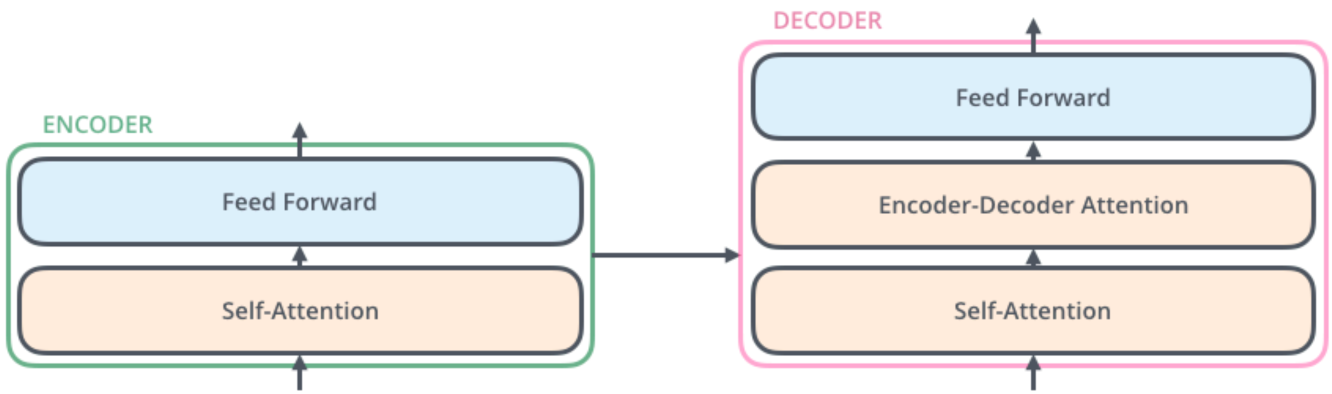
Transformers



A. Vaswani et al., Attention is all you need, NeurIPS 2017

Image source

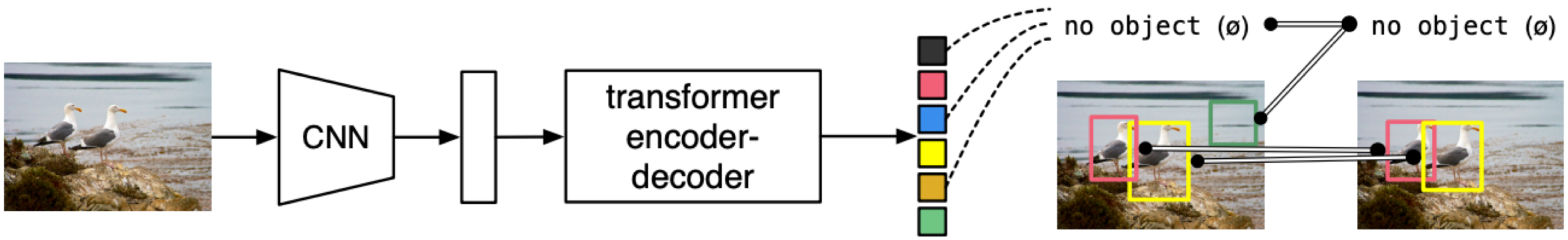
Transformers

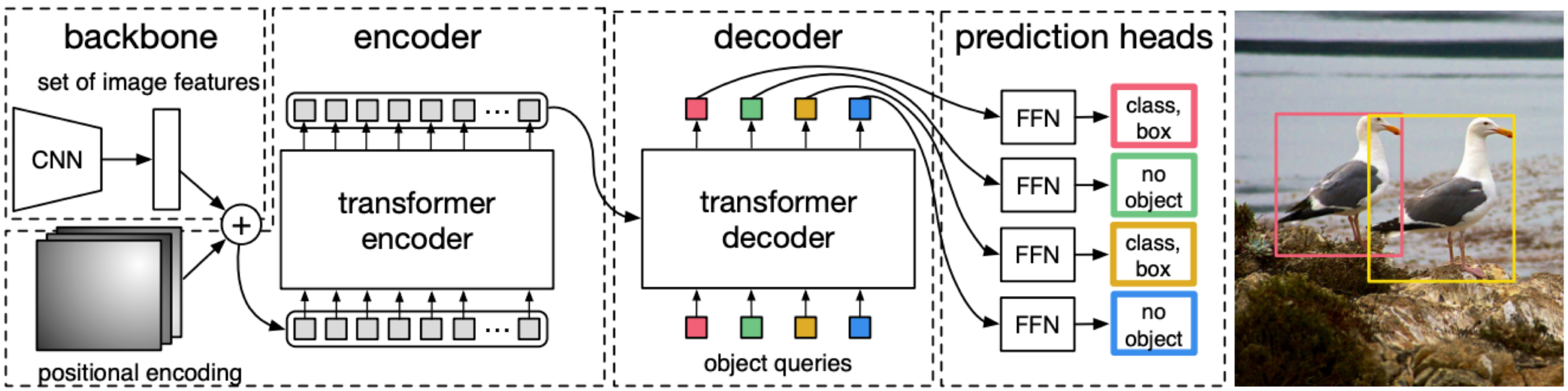


A. Vaswani et al., Attention is all you need, NeurIPS 2017

Image source

Transformers for everything: Detection transformer

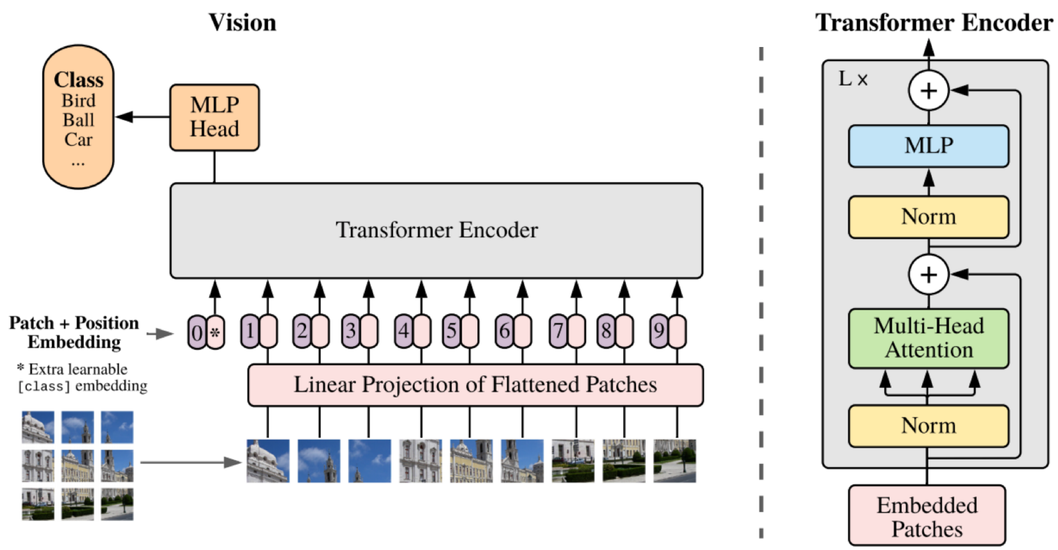




N. Carion et al. [End-to-end object detection with transformers](). ECCV 2020

Vision transformer (ViT) – Google

- Split an image into patches, feed linearly projected patches into standard transformer encoder
  - With patches of 14x14 pixels, you need 16x16=256 patches to represent 224x224 images
  - Self-supervised task: masked prediction (similar to BERT)



A. Dosovitskiy et al. [An image is worth 16x16 words: Transformers for image recognition at scale](). ICLR 2021

Vision transformer (ViT)

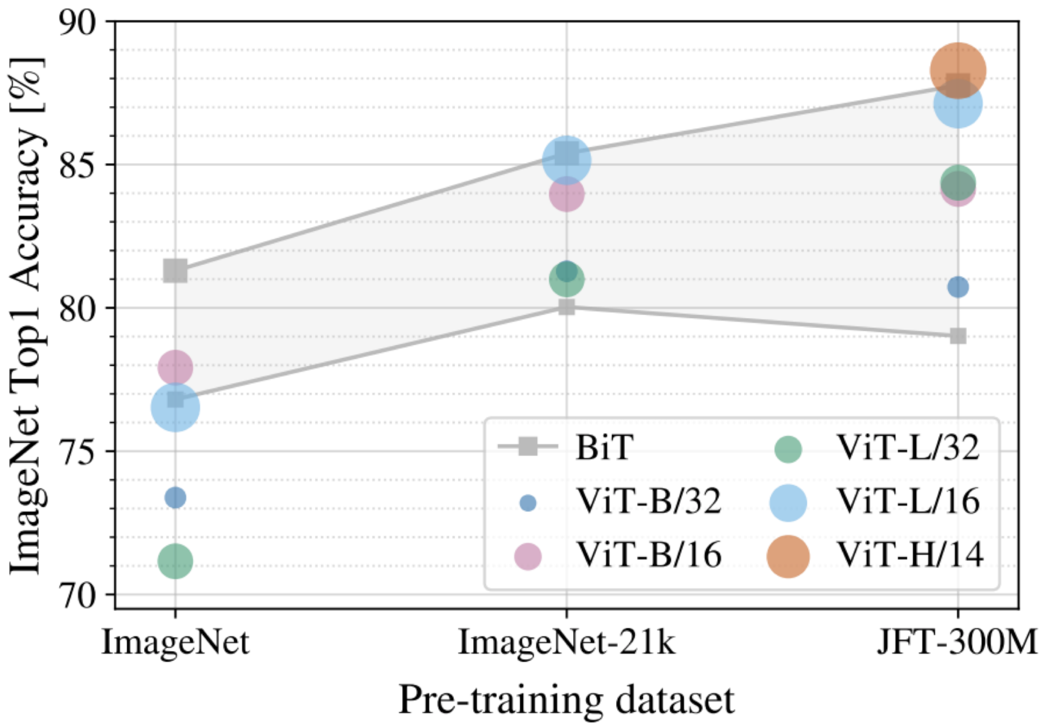


Figure 3: Transfer to ImageNet. While large ViT models perform worse than BiT ResNets (shaded area) when pre-trained on small datasets, they shine when pre-trained on larger datasets. Similarly, larger ViT variants overtake smaller ones as the dataset grows.

BiT: Big Transfer (ResNet)
ViT: Vision Transformer (Base/Large/Huge, patch size of 14x14, 16x16, or 32x32)

Internal Google dataset (not public)

A. Dosovitskiy et al. An image is worth 16x16 words: Transformers for image recognition at scale. ICLR 2021

Masked autoencoders

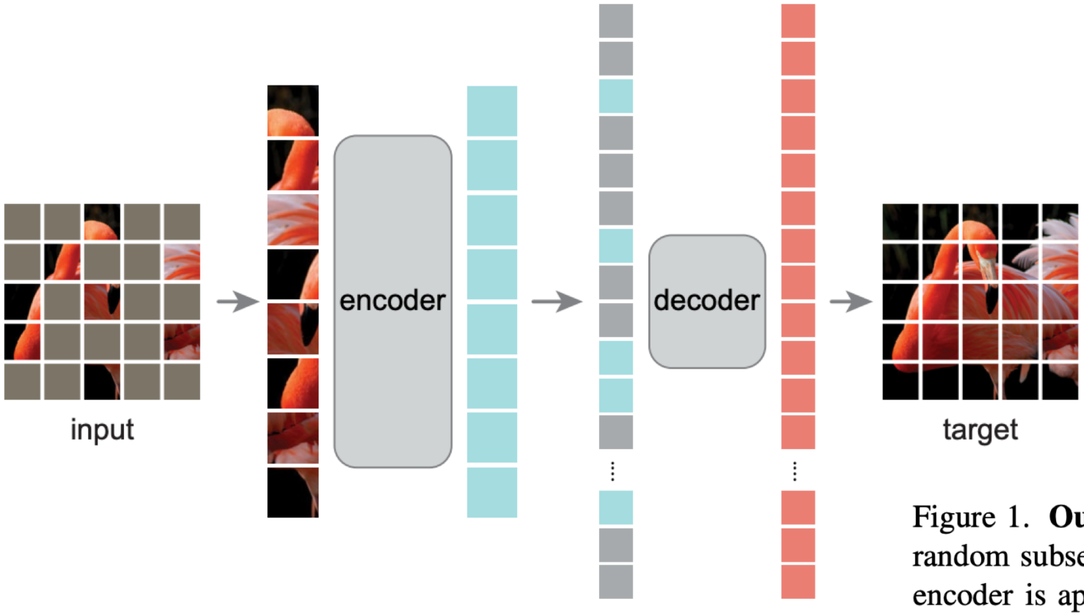


Figure 1. **Our MAE architecture**. During pre-training, a large random subset of image patches (*e.g.*, 75%) is masked out. The encoder is applied to the small subset of *visible patches*. Mask tokens are introduced *after* the encoder, and the full set of encoded patches and mask tokens is processed by a small decoder that reconstructs the original image in pixels. After pre-training, the decoder is discarded and the encoder is applied to uncorrupted images to produce representations for recognition tasks.

K. He et al. [Masked autoencoders are scalable vision learners](). CVPR 2022

Masked autoencoders

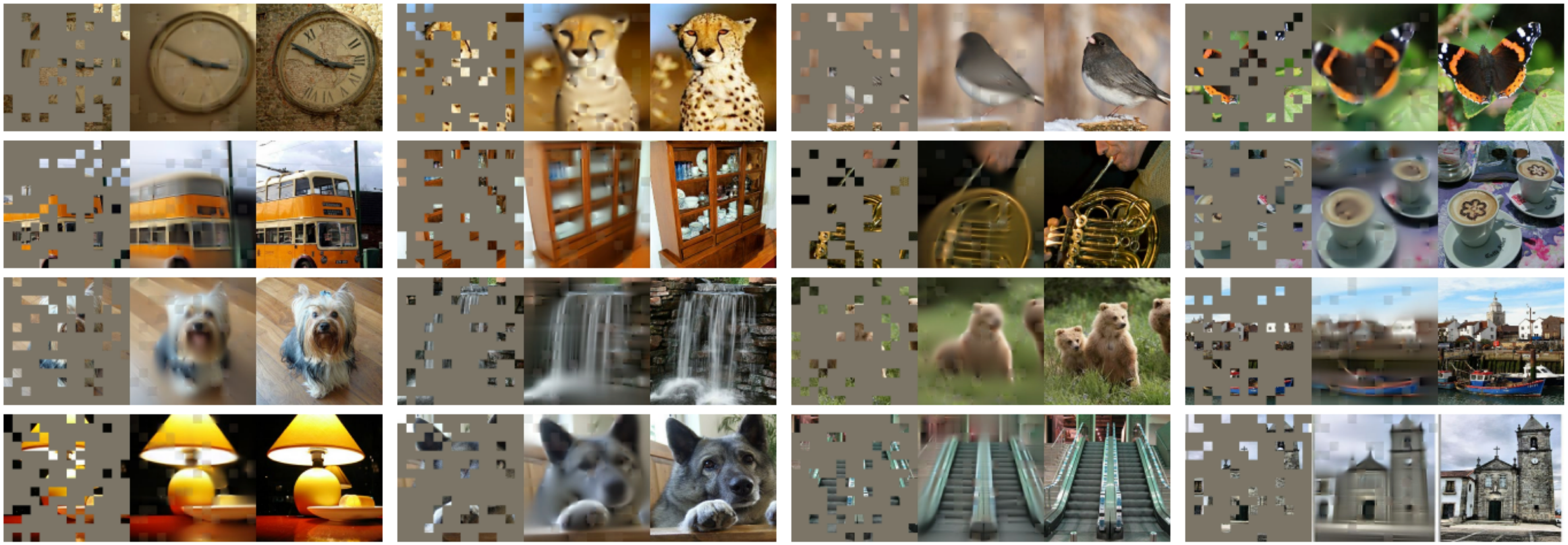


Figure 2. Example results on ImageNet *validation* images. For each triplet, we show the masked image (left), our MAE reconstruction† (middle), and the ground-truth (right). The masking ratio is 80%, leaving only 39 out of 196 patches. More examples are in the appendix. †*As no loss is computed on visible patches, the model output on visible patches is qualitatively worse. One can simply overlay the output with the visible patches to improve visual quality. We intentionally opt not to do this, so we can more comprehensively demonstrate the method's behavior.*

K. He et al. Masked autoencoders are scalable vision learners. CVPR 2022

Masked autoencoders

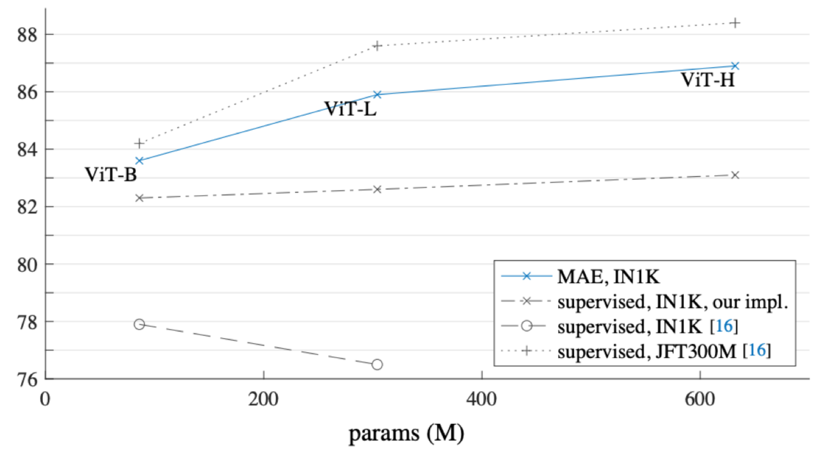


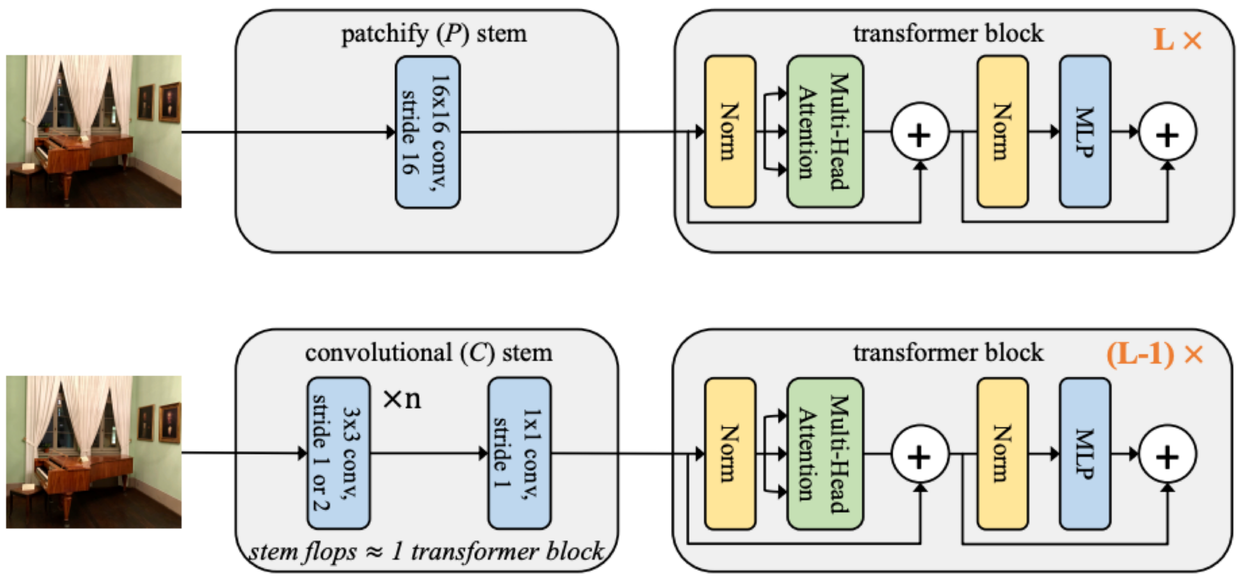
Figure 8. **MAE pre-training *vs.* supervised pre-training**, evaluated by fine-tuning in ImageNet-1K (224 size). We compare with the original ViT results [16] trained in IN1K or JFT300M.

		AP^{box}		AP^{mask}	
method	pre-train data	ViT-B	ViT-L	ViT-B	ViT-L
supervised	IN1K w/ labels	47.9	49.3	42.9	43.9
MoCo v3	IN1K	47.9	49.3	42.7	44.0
BEiT	IN1K+DALLE	49.8	**53.3**	44.4	47.1
MAE	IN1K	**50.3**	**53.3**	**44.9**	**47.2**

Table 4. **COCO object detection and segmentation** using a ViT Mask R-CNN baseline. All entries are based on our implementation. Self-supervised entries use IN1K data *without* labels. Mask AP follows a similar trend as box AP.

K. He et al. [Masked autoencoders are scalable vision learners](). CVPR 2022

Convolutional networks or transformers?



Original ViT (baseline, termed ViT$_P$):

- *Sensitive to lr and wd choice*
- *Converges slowly*
- *Works with AdamW, but not SGD*
- *Underperforms sota CNNs on ImageNet*

Ours (termed ViT$_C$, same runtime):

- ✓ *Robust to lr and wd choice*
- ✓ *Converges quickly*
- ✓ *Works with AdamW, and also SGD*
- ✓ *Outperforms sota CNNs on ImageNet*

T. Xiao et al. [Early convolutions help transformers see better](). NeurIPS 2021

Beyond transformers?

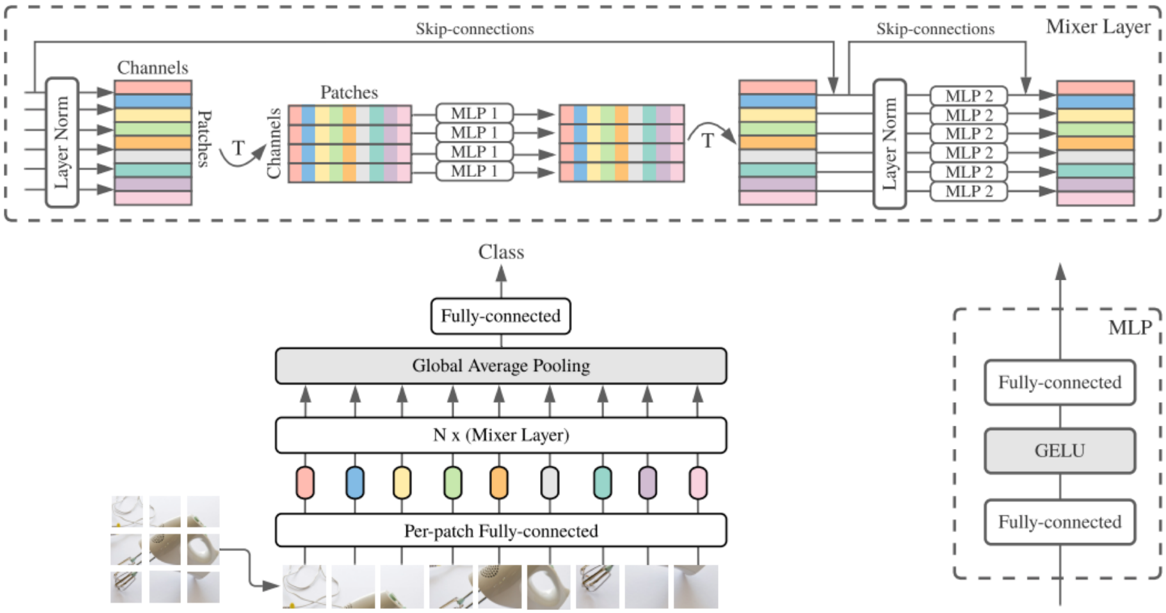
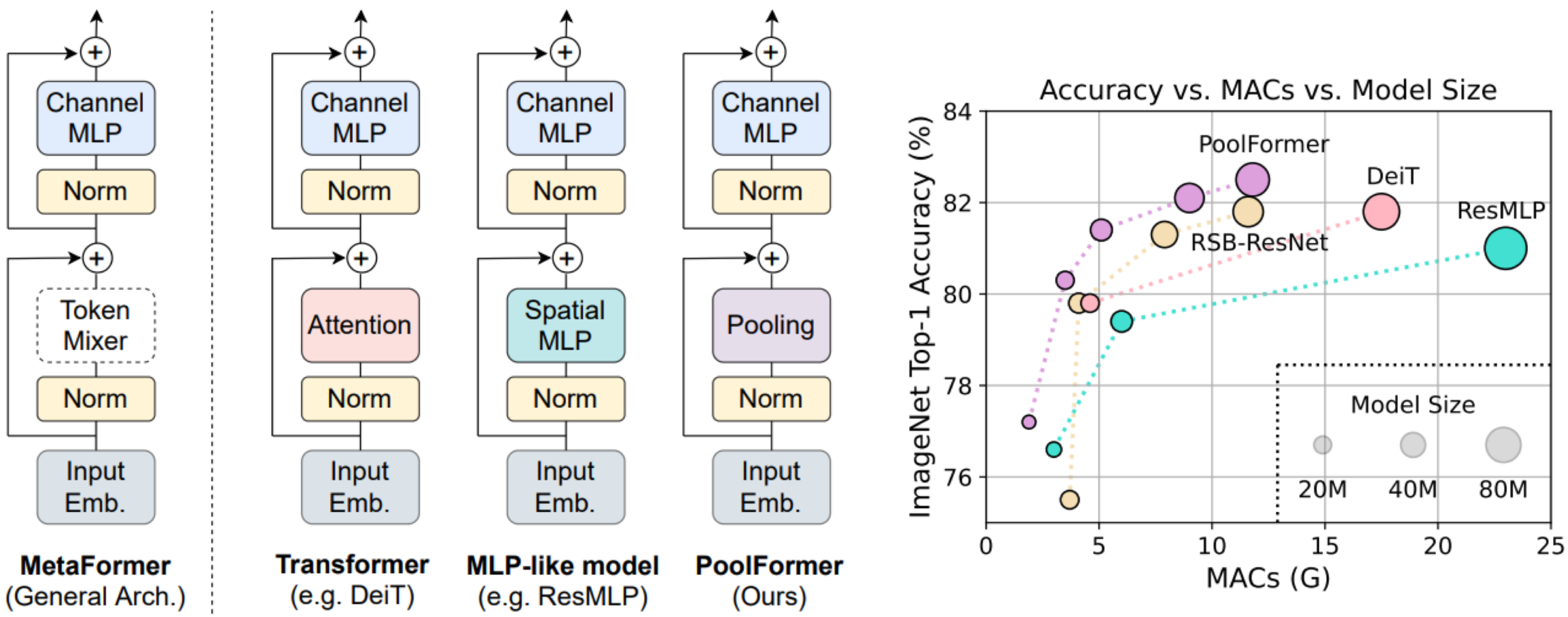


Figure 1: MLP-Mixer consists of per-patch linear embeddings, Mixer layers, and a classifier head. Mixer layers contain one token-mixing MLP and one channel-mixing MLP, each consisting of two fully-connected layers and a GELU nonlinearity. Other components include: skip-connections, dropout, and layer norm on the channels.

I. Tolstikhin et al. [MLP-Mixer: An all-MLP Architecture for Vision](). NeurIPS 2021

Beyond transformers?

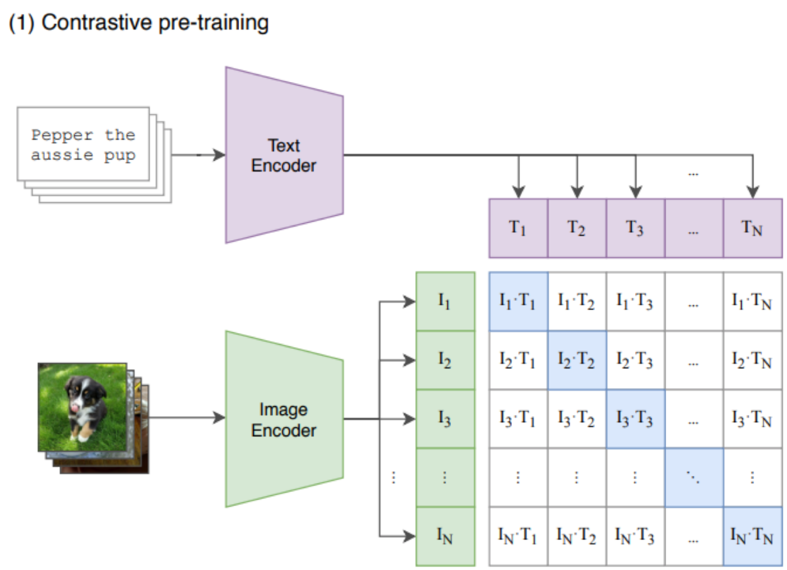


W. Yu et al. [MetaFormer is Actually What You Need for Vision](). CVPR 2022

Outline

- Brief history of recognition
- Different “dimensions” of recognition
  - What type of content?
  - What type of output?
  - What type of supervision?
- Trends
  - Saturation of supervised learning
  - Transformers
  - Vision-language models

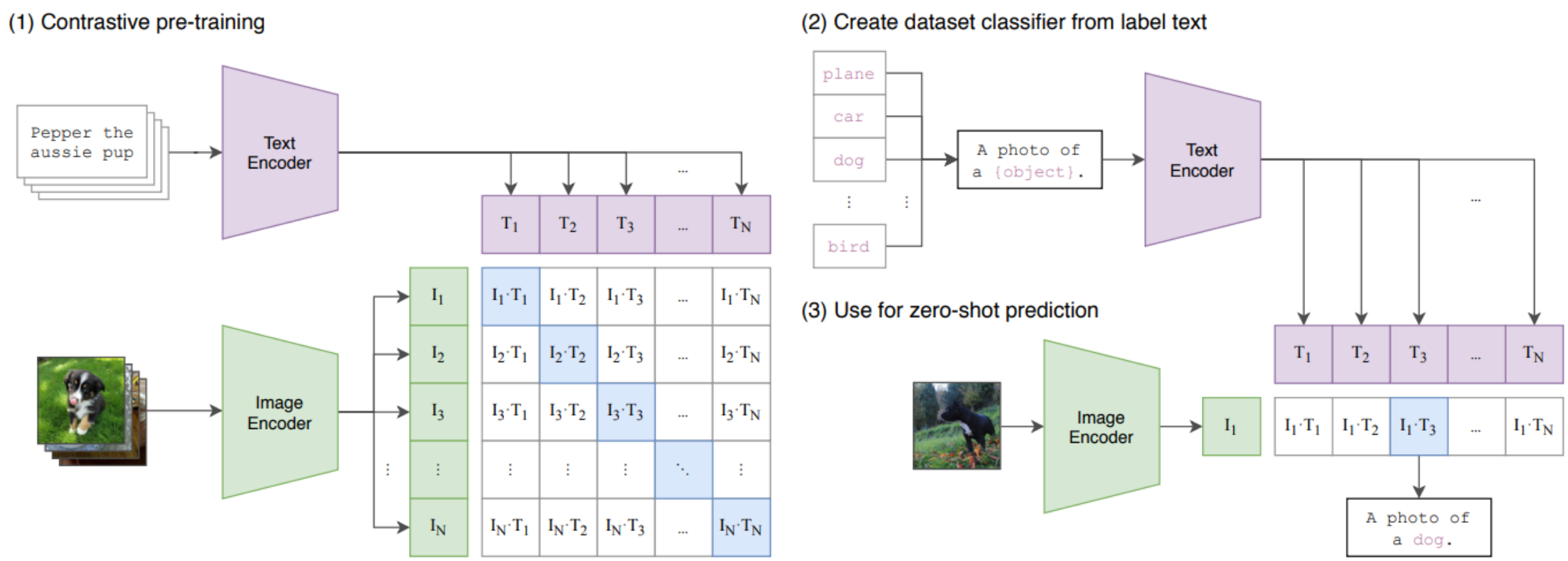
Giant vision-language models: CLIP



Contrastive language-image pretraining: in a batch of N image-text pairs, classify each text string to the correct image and vice versa

A. Radford et al., [Learning Transferable Visual Models From Natural Language Supervision](), ICML 2021
https://openai.com/blog/clip/

Giant vision-language models: CLIP



A. Radford et al., [Learning Transferable Visual Models From Natural Language Supervision](), ICML 2021
https://openai.com/blog/clip/

CLIP: Details

- Image encoders
  - ResNet-50 with self-attention layer on top of global average pooling
  - Vision transformer (ViT)
- Language encoder: GPT-style transformer with 63M parameters
- Dataset: 400M image-text pairs from the Web

CLIP: Results

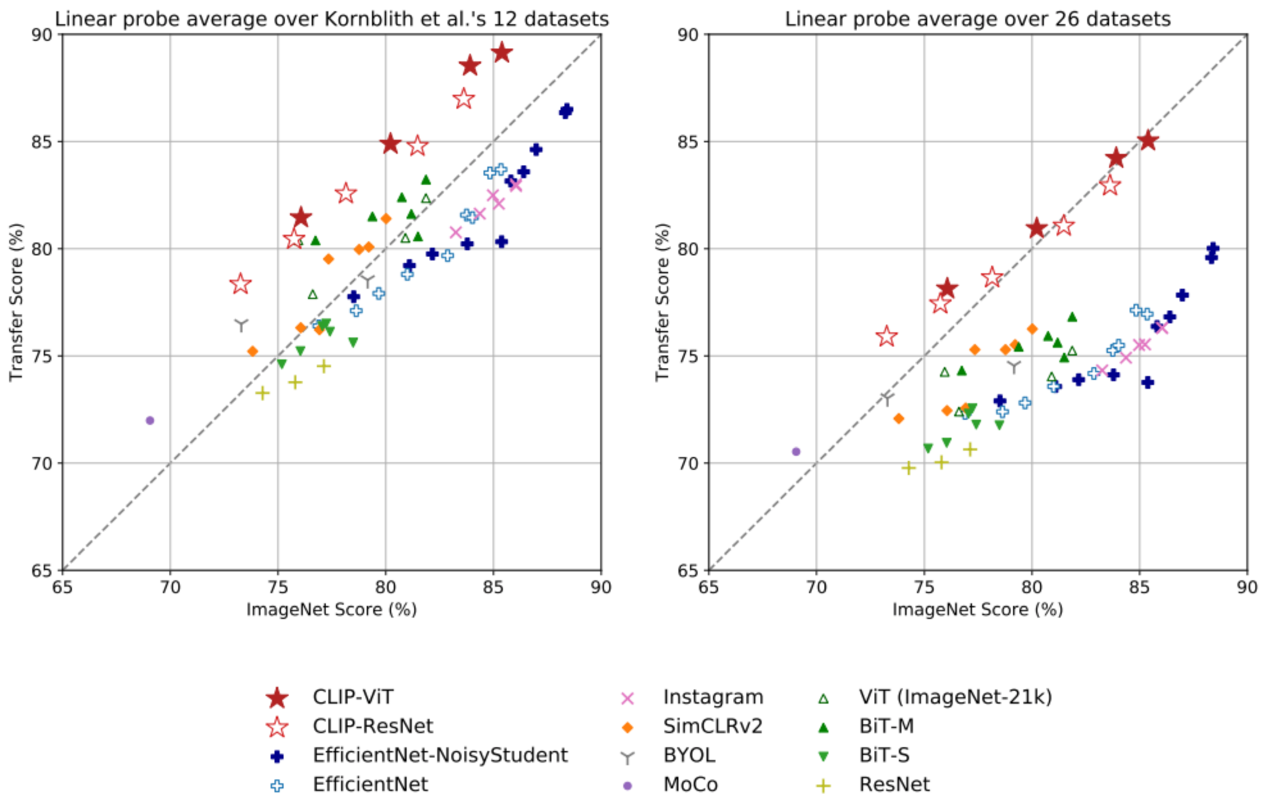
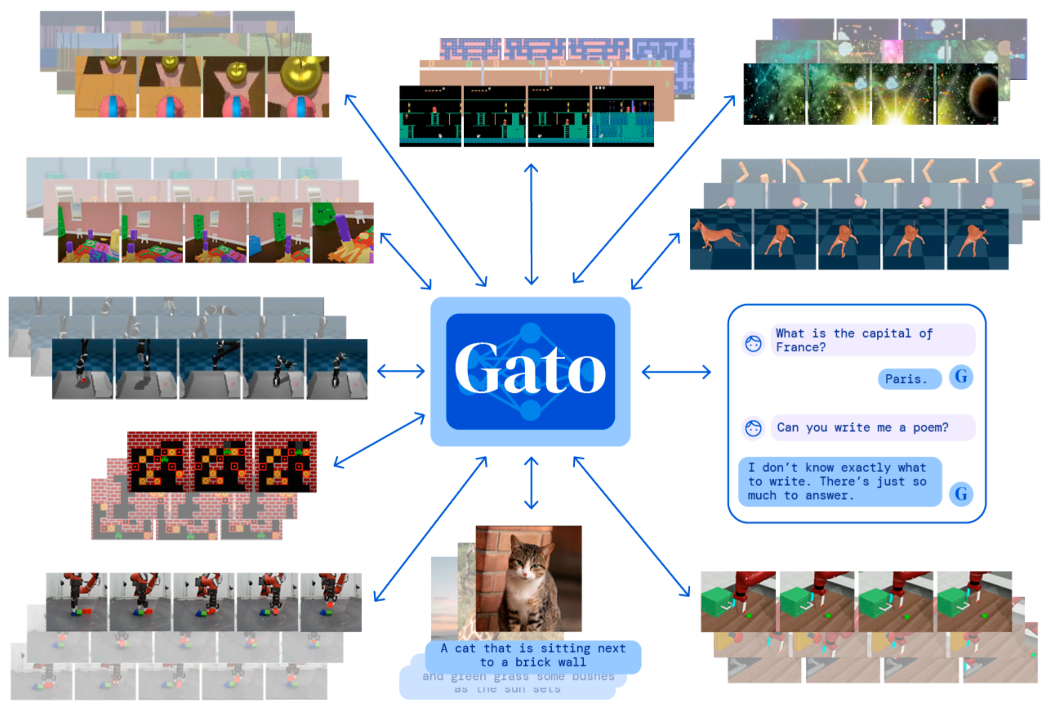


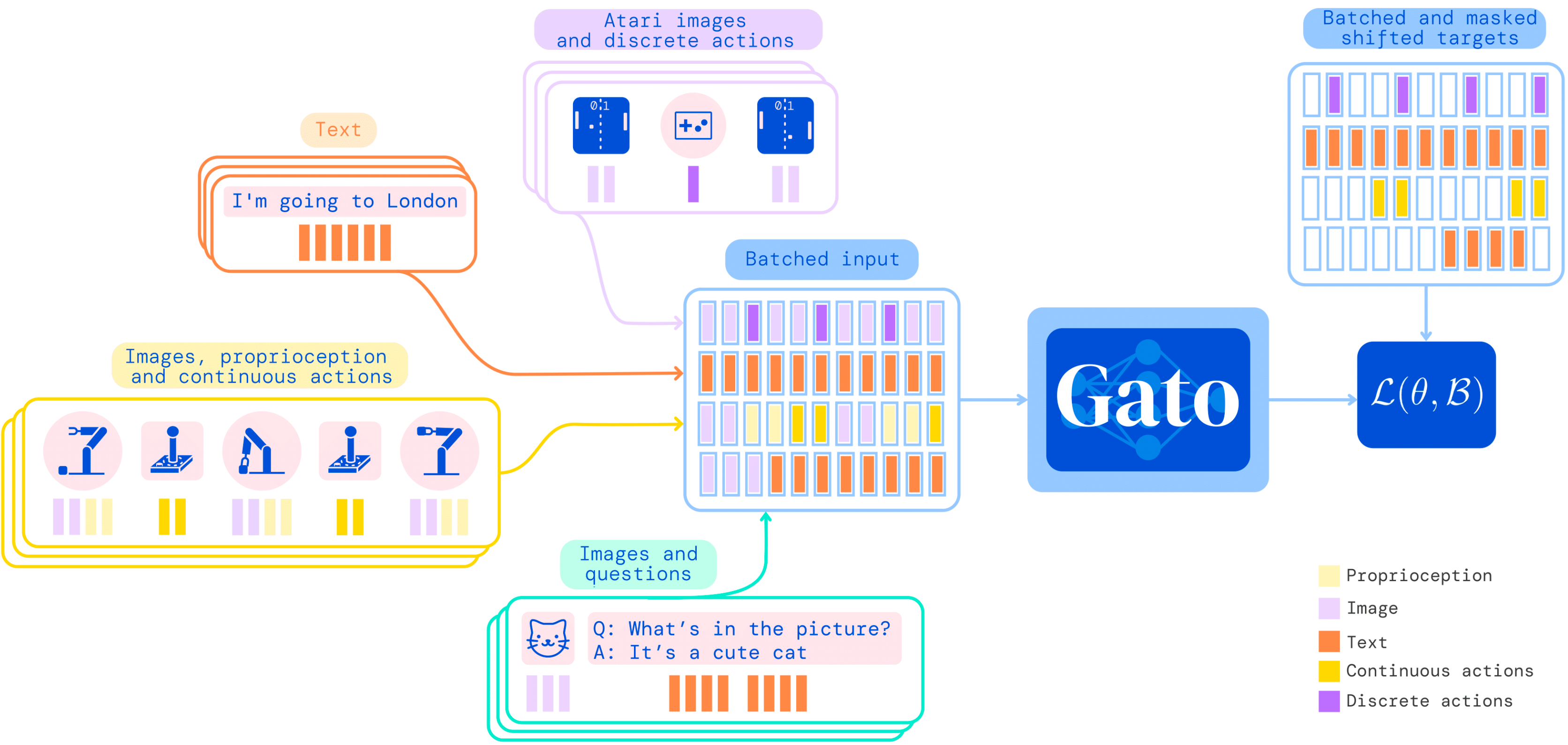
Figure 12. **CLIP's features are more robust to task shift when compared to models pre-trained on ImageNet.** For both dataset splits, the transfer scores of linear probes trained on the representations of CLIP models are higher than other models with similar ImageNet performance. This suggests that the representations of models trained on ImageNet are somewhat overfit to their task.

“Universal” recognition systems: DeepMind GATO



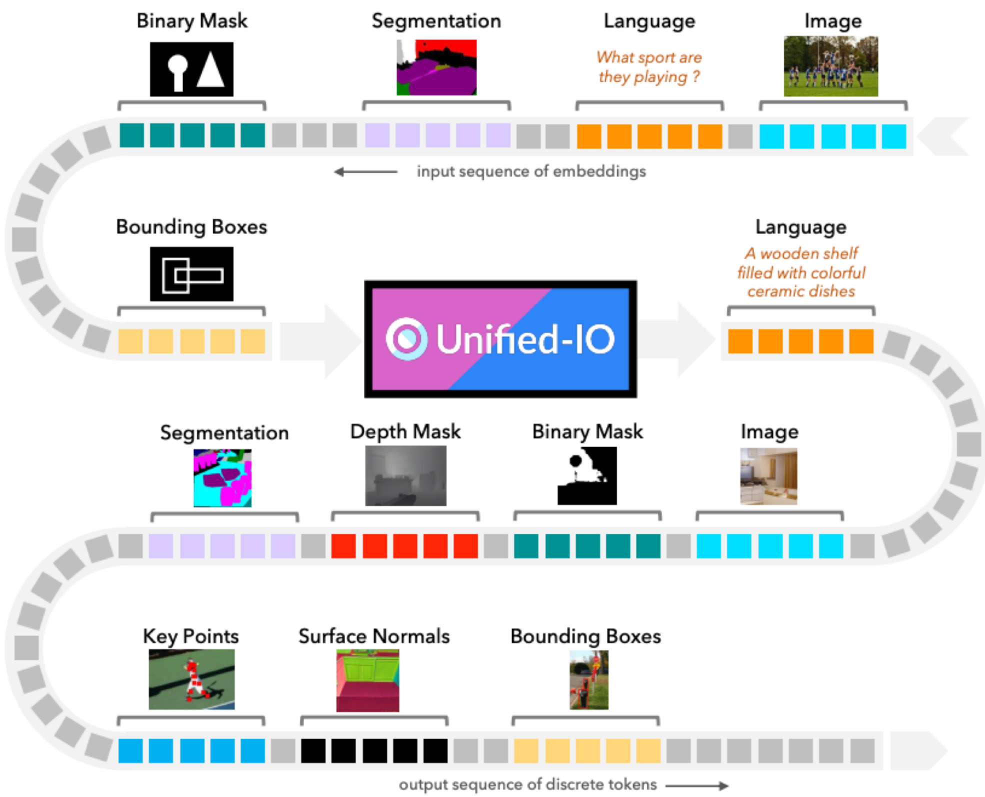
S. Reed et al. [A generalist agent](). TMLR 2022

“Universal” recognition systems: DeepMind GATO



S. Reed et al. A generalist agent. TMLR 2022

“Universal” recognition systems: UnifiedIO

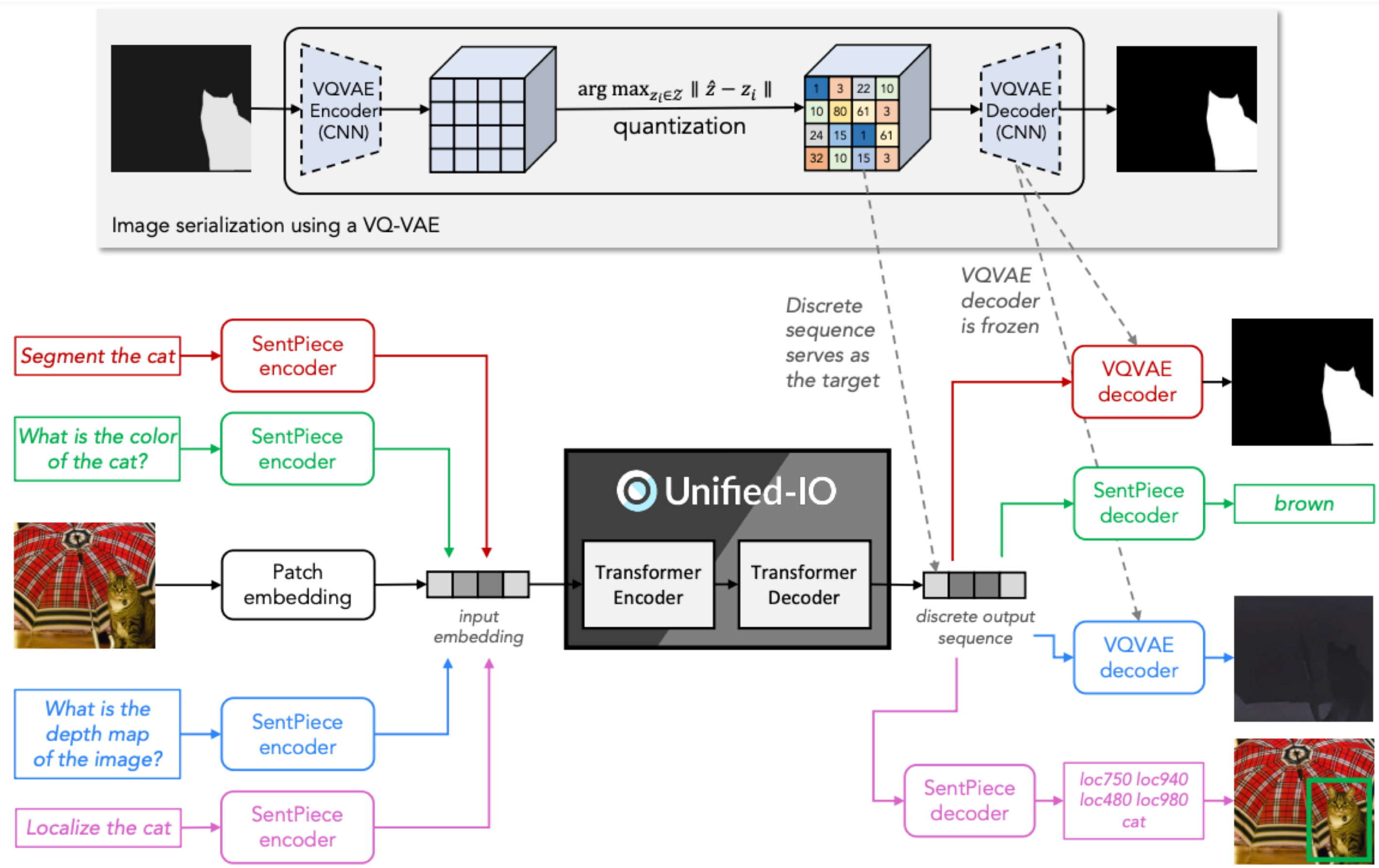


Tasks

Image Classification
Object Detection
Semantic Segmentation
Depth Estimation
Surface Normal Estimation
Segment-based Image Generation
Image Inpainting
Pose Estimation
Relationship Detection
Image Captioning
Visual QA
Referring Expressions
Situation Recognition
Text-based Image Generation
Visual Commonsense
Classification in context
Region Captioning
GLUE Benchmark tasks
Reading comprehension
Natural Language Inference
Grounded Commonsense Inference

J. Lu et al. A unified model for vision, language, and multi-modal tasks. arXiv 2022
https://unified-io.allenai.org/

"Universal" recognition systems: UnifiedIO



Outline

- Brief history of recognition
- Different “dimensions” of recognition
  - What type of content?
  - What type of output?
  - What type of supervision?
- Trends
  - Saturation of supervised learning
  - Transformers
  - Vision-language models
  - “Universal” recognition systems
  - Text-to-image generation

DALL-E: Text-to-image generation using transformers

- Train an encoder similar to VQ-VAE to compress images to 32x32 grids of discrete tokens (each assuming 8192 values)
- Concatenate with text strings, learn a joint sequential transformer model that can be used to generate image based on text prompt

(a) a tapir made of accordion. a tapir with the texture of an accordion.

(b) an illustration of a baby hedgehog in a christmas sweater walking a dog

(c) a neon sign that reads "backprop". a neon sign that reads "backprop". backprop neon sign

A. Ramesh et al., Zero-Shot Text-to-Image Generation, ICML 2021
https://openai.com/blog/dall-e/

DALL-E: Image encoding

- Train convolutional encoder and decoder to compress images to 32x32 grids of discrete tokens (each assuming 8192 values)

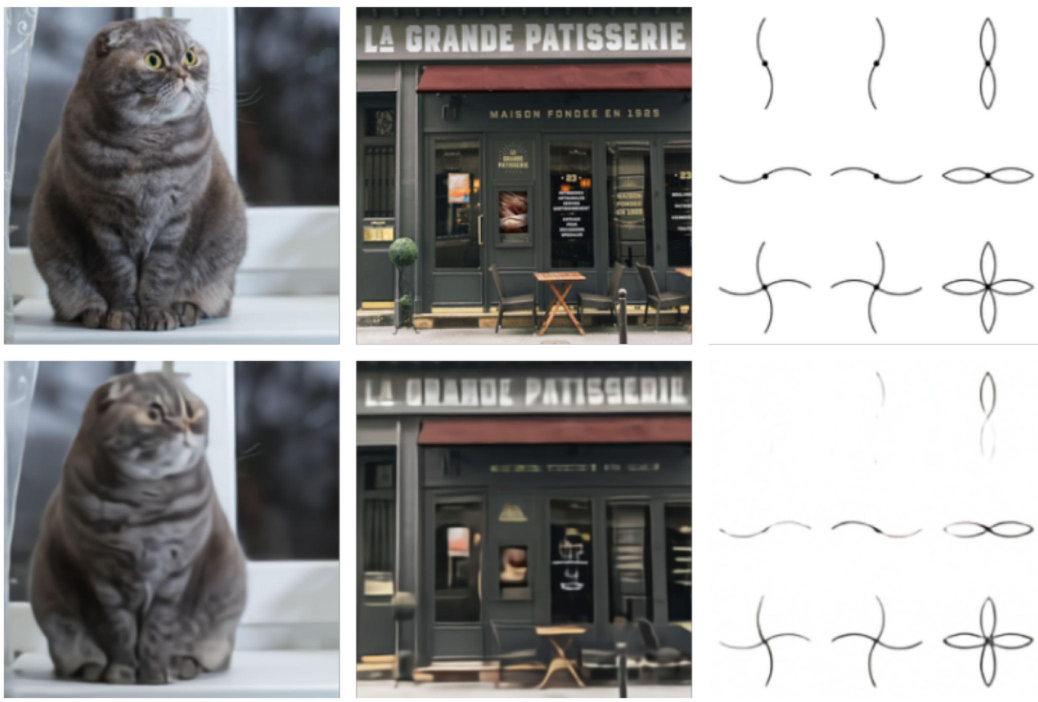
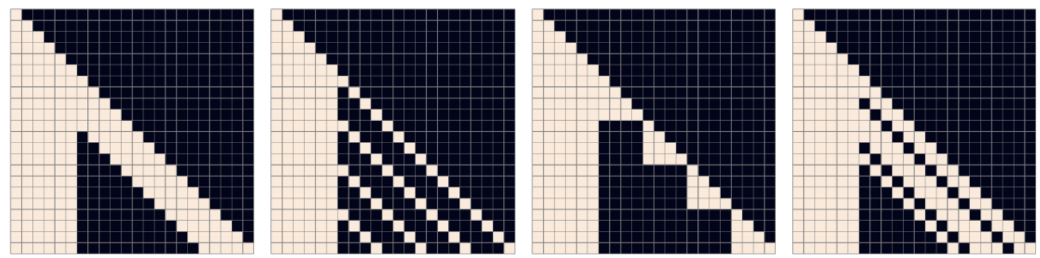


Figure 1. Comparison of original images (top) and reconstructions from the discrete VAE (bottom). The encoder downsamples the spatial resolution by a factor of 8. While details (e.g., the texture of the cat's fur, the writing on the storefront, and the thin lines in the illustration) are sometimes lost or distorted, the main features of the image are still typically recognizable. We use a large vocabulary size of 8192 to mitigate the loss of information.

DALL-E: Transformer architecture and training

- Concatenate up to 256 text tokens with 32x32=1024 image tokens, learn a transformer model with 64 layers and 12B parameters
- Dataset: 250M image-text pairs from the Internet (similar scale to JFT-300M, apparently different from data used to train CLIP)
- Transformer model details
  - Decoder-only architecture
  - 64 self-attention layers,
  - 62 attention heads, sparse attention patterns
  - Mixed-precision training, distributed optimization



(a) Row attention mask. (b) Column attention mask. (c) Column attention mask with transposed image states. (d) Convolutional attention mask.

Figure 11. Illustration of the three types of attention masks for a hypothetical version of our transformer with a maximum text length of 6 tokens and image length of 16 tokens (i.e., corresponding to a 4×4 grid). Mask (a) corresponds to row attention in which each image token attends to the previous 5 image tokens in raster order. The extent is chosen to be 5, so that the last token being attended to is the one in the same column of the previous row. To obtain better GPU utilization, we transpose the row and column dimensions of the image states when applying column attention, so that we can use mask (c) instead of mask (b). Mask (d) corresponds to a causal convolutional attention pattern with wraparound behavior (similar to the row attention) and a 3×3 kernel. Our model uses a mask corresponding to an 11×11 kernel.

DALL-E: Generating images given text

- Re-rank samples using CLIP

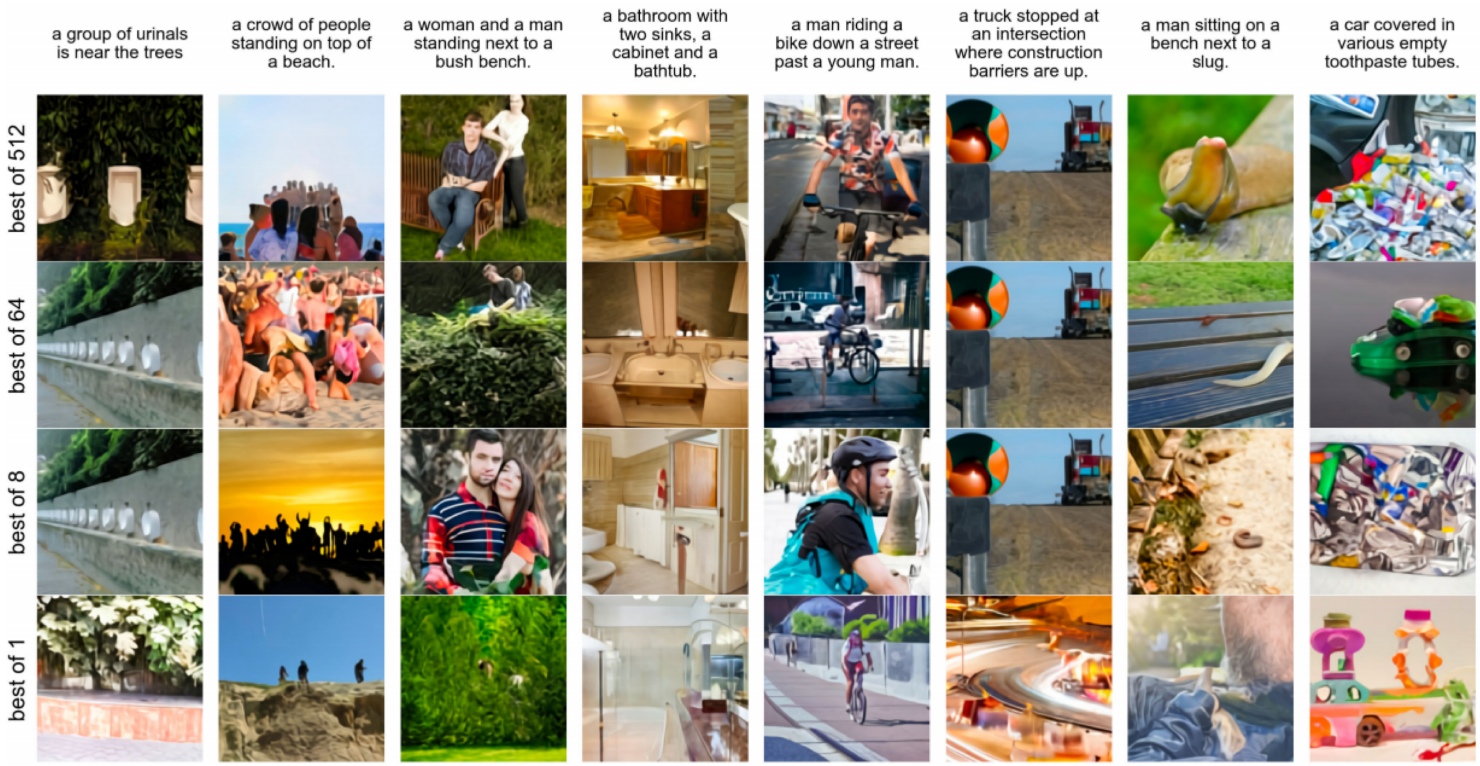


Figure 6. Effect of increasing the number of images for the contrastive reranking procedure on MS-COCO captions.

DALL-E 2: Text-to-image generation using diffusion models

A. Ramesh et al. [Hierarchical text-conditional image generation with CLIP latents](). 2022

DALL-E 2

Figure 19: Random samples from unCLIP for prompt "A close up of a handpalm with leaves growing from it."

DALL-E 2

Figure 18: Random samples from unCLIP for prompt “Vibrant portrait painting of Salvador Dali with a robotic half face”

DALL-E 2

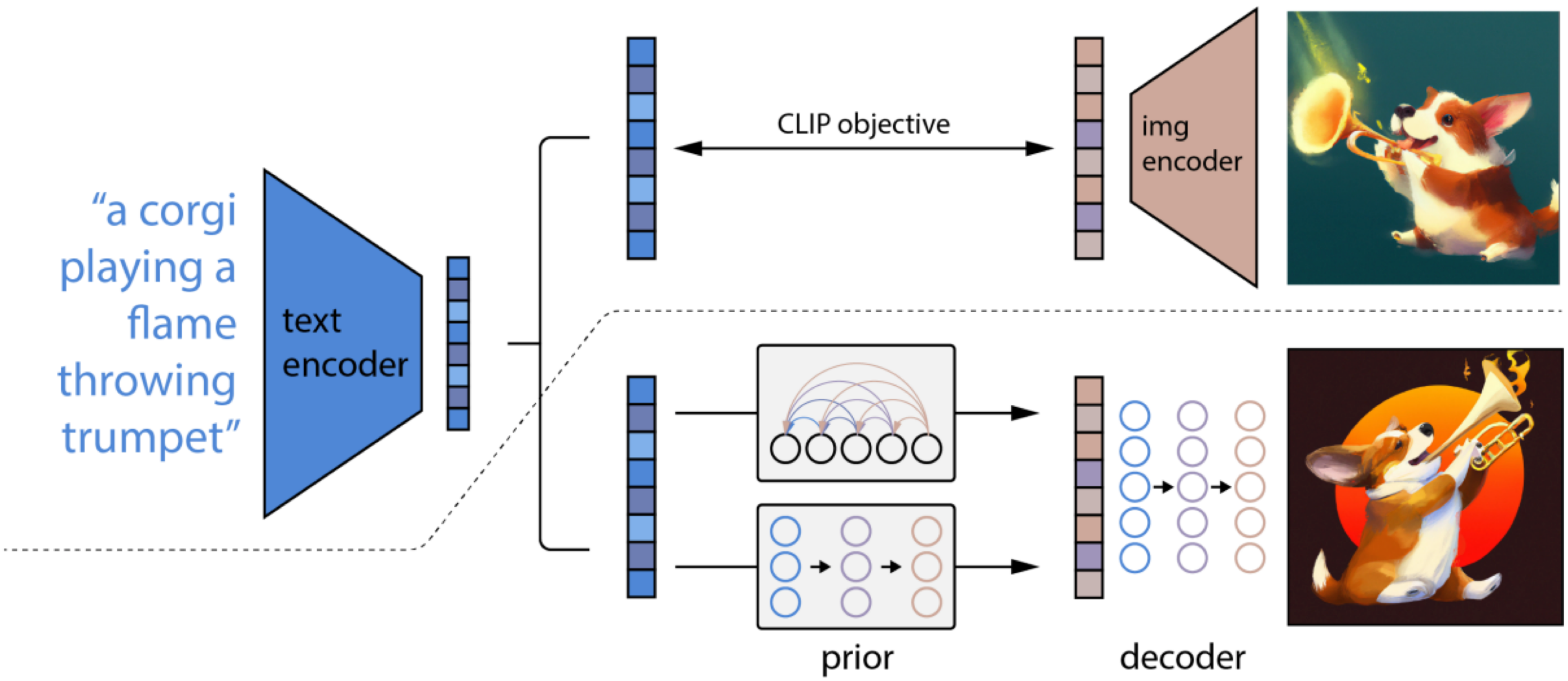


Figure 2: A high-level overview of unCLIP. Above the dotted line, we depict the CLIP training process, through which we learn a joint representation space for text and images. Below the dotted line, we depict our text-to-image generation process: a CLIP text embedding is first fed to an autoregressive or diffusion prior to produce an image embedding, and then this embedding is used to condition a diffusion decoder which produces a final image. Note that the CLIP model is frozen during training of the prior and decoder.

DALL-E 2

Figure 3: Variations of an input image by encoding with CLIP and then decoding with a diffusion model. The variations preserve both semantic information like presence of a clock in the painting and the overlapping strokes in the logo, as well as stylistic elements like the surrealism in the painting and the color gradients in the logo, while varying the non-essential details.

Diffusion models

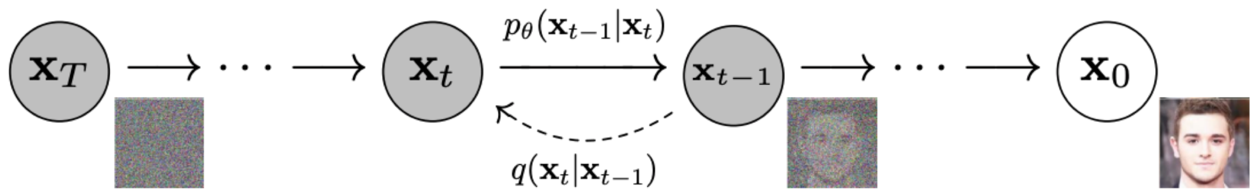


Figure 2: The directed graphical model considered in this work.

Unconditional CIFAR10 sample generation

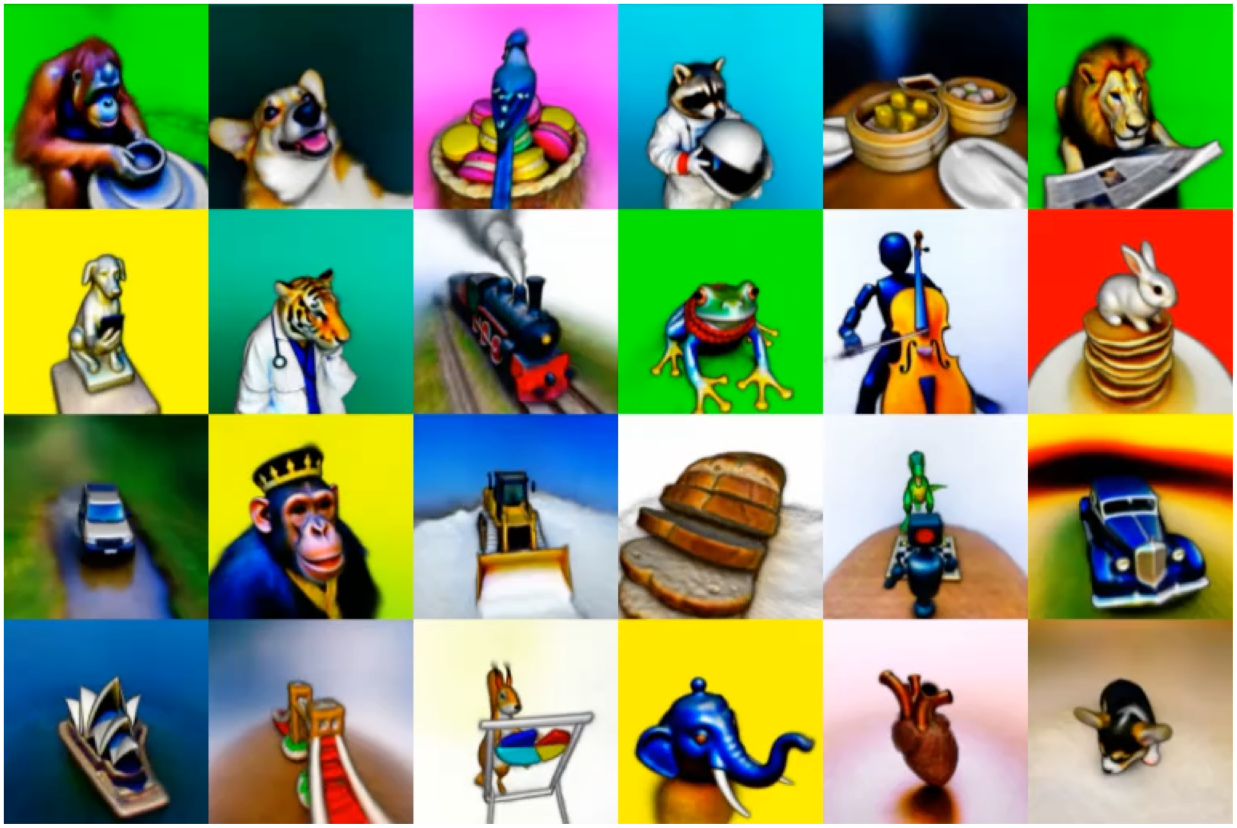
J. Ho et al. [Denoising diffusion probabilistic models](). NeurIPS 2020

Blog introduction: https://lilianweng.github.io/posts/2021-07-11-diffusion-models/

DALL-E 2 limitations

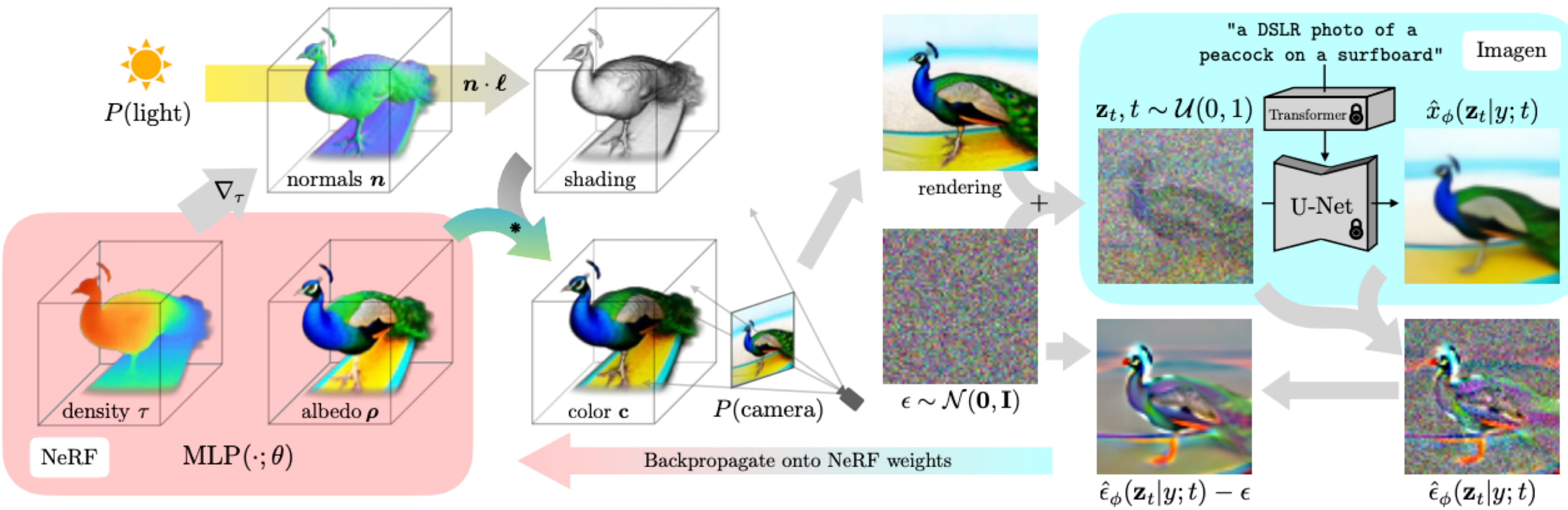
Figure 15: Reconstructions from the decoder for difficult binding problems. We find that the reconstructions mix up objects and attributes. In the first two examples, the model mixes up the color of two objects. In the rightmost example, the model does not reliably reconstruct the relative size of two objects.

DreamFusion: Diffusion models + NeRFs



B. Poole, A. Jain, J. Barron, B. Mildenhall. DreamFusion: Text-to-3D using 2D Diffusion. arXiv 2022

DreamFusion: Diffusion models + NeRFs



B. Poole, A. Jain, J. Barron, B. Mildenhall. [DreamFusion: Text-to-3D using 2D Diffusion](). arXiv 2022

Outline

- Brief history of recognition
- Different “dimensions” of recognition
  - What type of content?
  - What type of output?
  - What type of supervision?
- Trends
  - Saturation of supervised learning
  - Transformers
  - Vision-language models
  - “Universal” recognition systems
  - Text-to-image generation
  - From vision to action

From vision to action

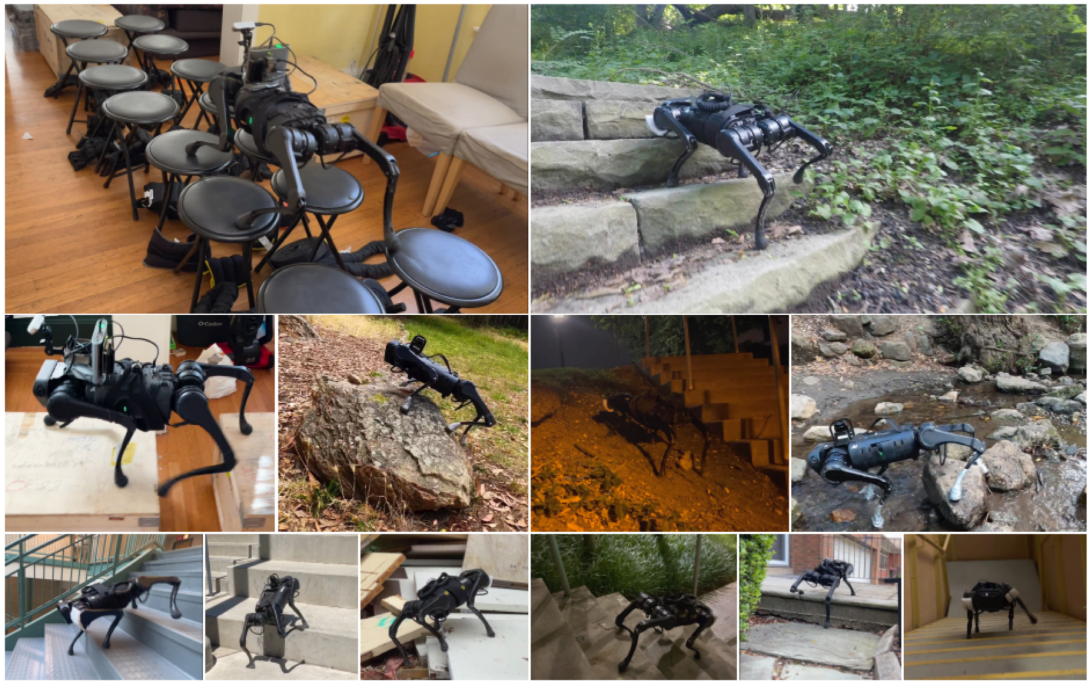


Figure 1: Our robot can traverse a variety of challenging terrain in indoor and outdoor environments, urban and natural settings during day and night using a single front-facing depth camera. The robot can traverse curbs, stairs and moderately rocky terrain. Despite being much smaller than other commonly used legged robots, it is able to climb stairs and curbs of a similar height. Videos at https://vision-locomotion.github.io

A. Agarwal, A. Kumar, J. Malik, and D. Pathak. Legged Locomotion in Challenging Terrains using Egocentric Vision. CoRL 2022

From vision to action

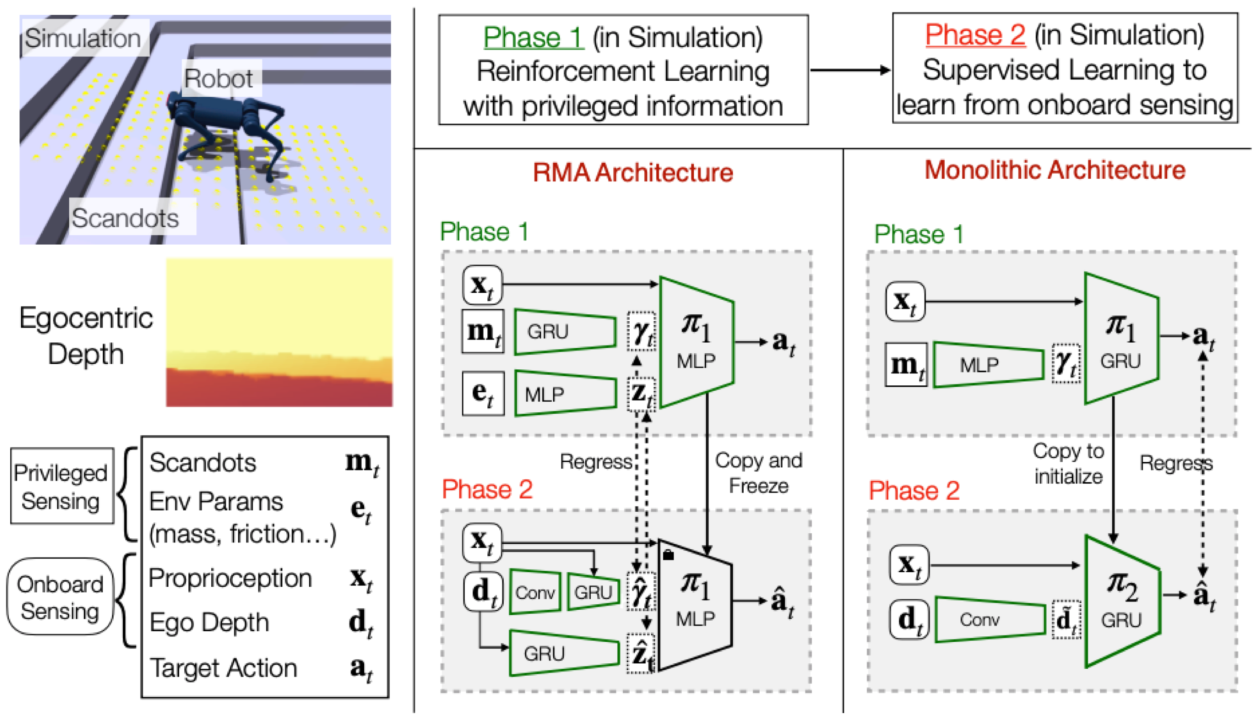


Figure 3: We train our locomotion policy in two phases to avoid rendering depth for too many samples. In phase 1, we use RL to train a policy π^1 that has access to scandots that are cheap to compute. In phase 2, we use π^1 to provide ground truth actions which another policy π^2 is trained to imitate. This student has access to depth map from the front camera. We consider two architectures (1) a monolithic one which is a GRU trained to output joint angles with raw observations as input (2) a decoupled architecture trained using RMA [3] that is trained to estimate vision and proprioception latents that condition a base feedforward walking policy.