COMPUTER VISION PROBLEM ZOO

Olga Razvenskaya, Senior Research Engineer, NNRC Huawei

Computer vision

Computer Vision has a dual goal.

From the *biological science* point of view, computer vision aims to come up with computational models of the human visual system.

From the *engineering* point of view, computer vision aims to build autonomous systems which could perform some of the tasks which the human visual system can perform (and even surpass it in many cases).

Many vision tasks are related to the extraction of 3D and temporal information from time-varying 2D data such as obtained by one or more television cameras, and more generally the understanding of such dynamic scenes.

[12]

Overview

- Image Classification
- Semantic Segmentation
- Object Detection
- Image Generation
- Data Augmentation
- Super-Resolution
- Anomaly Detection
- Autonomous Driving
- …and many more

Image Classification

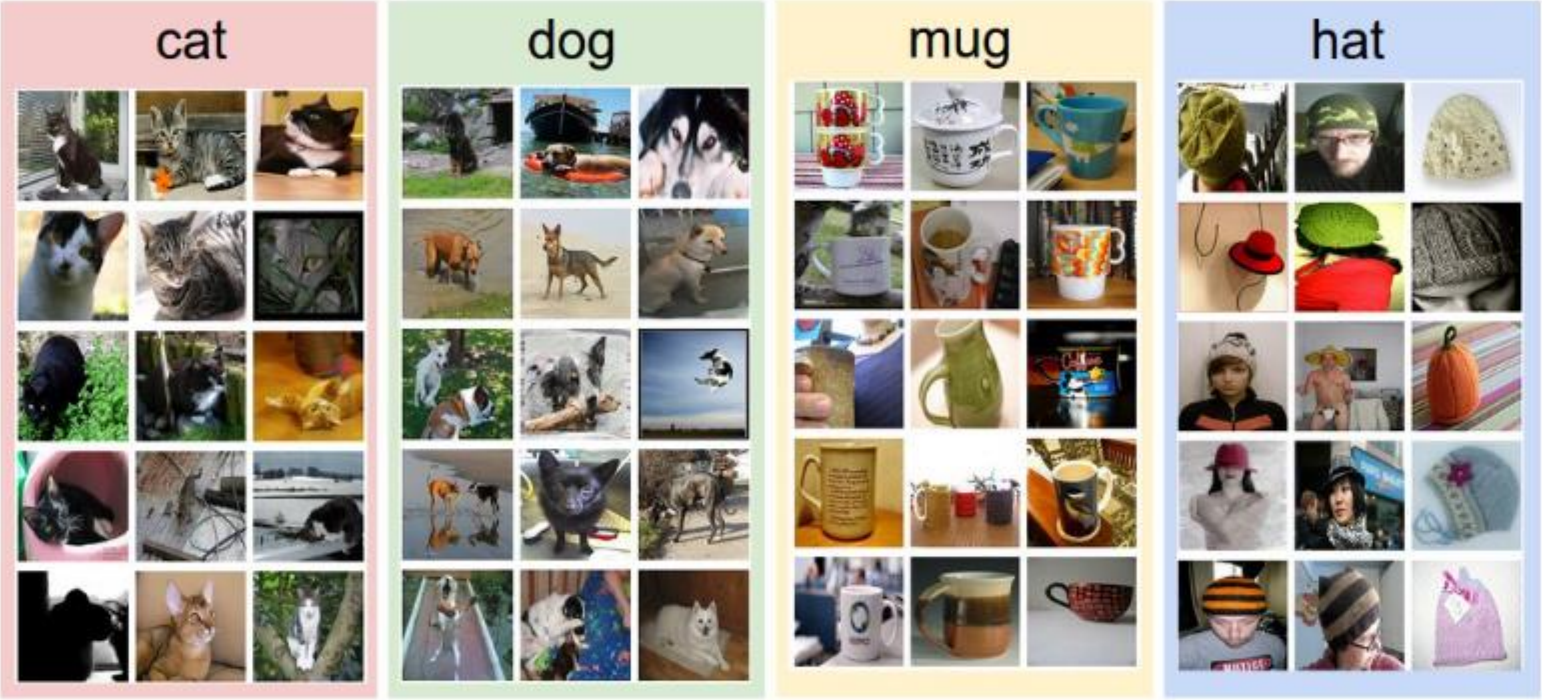


Image Classification - difficulties

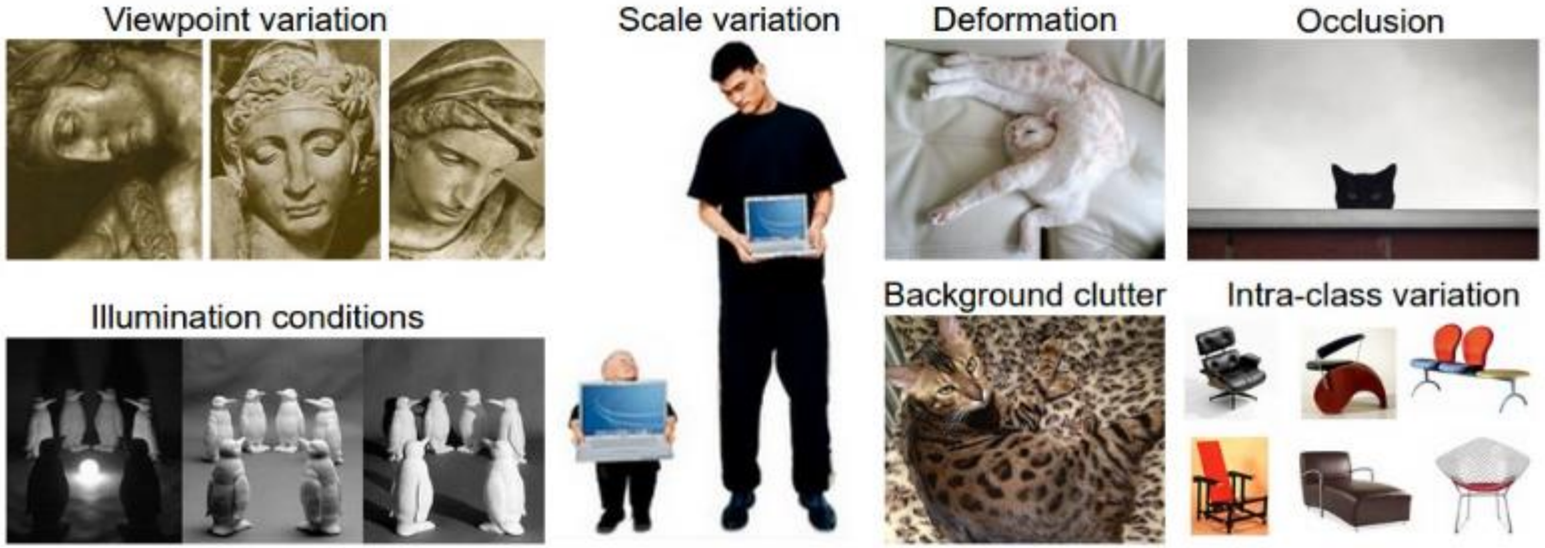
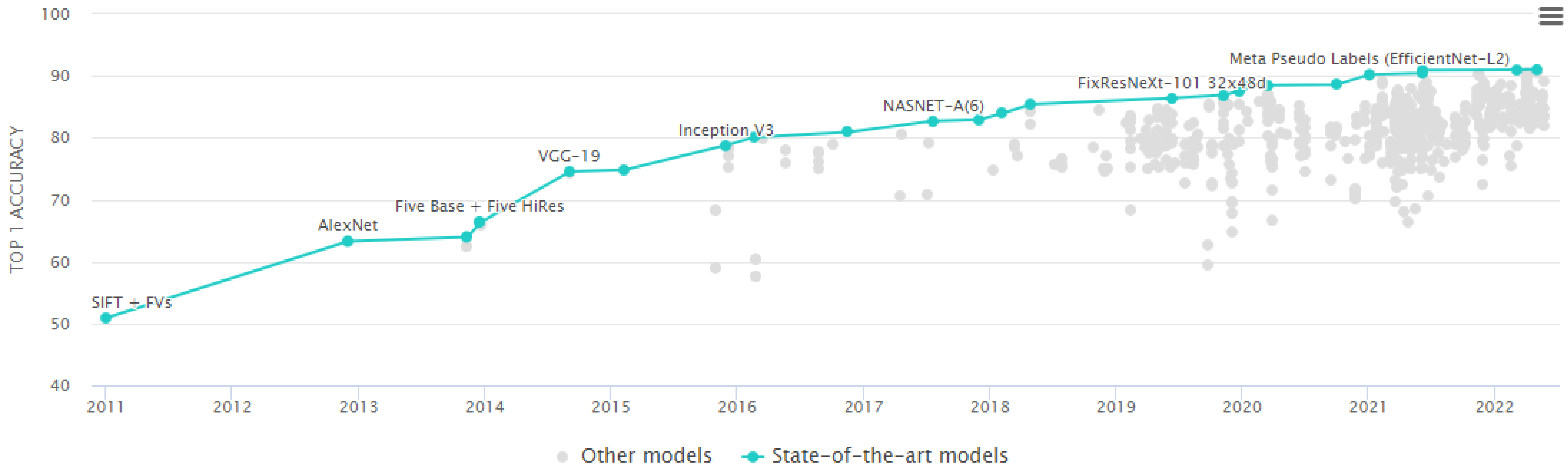
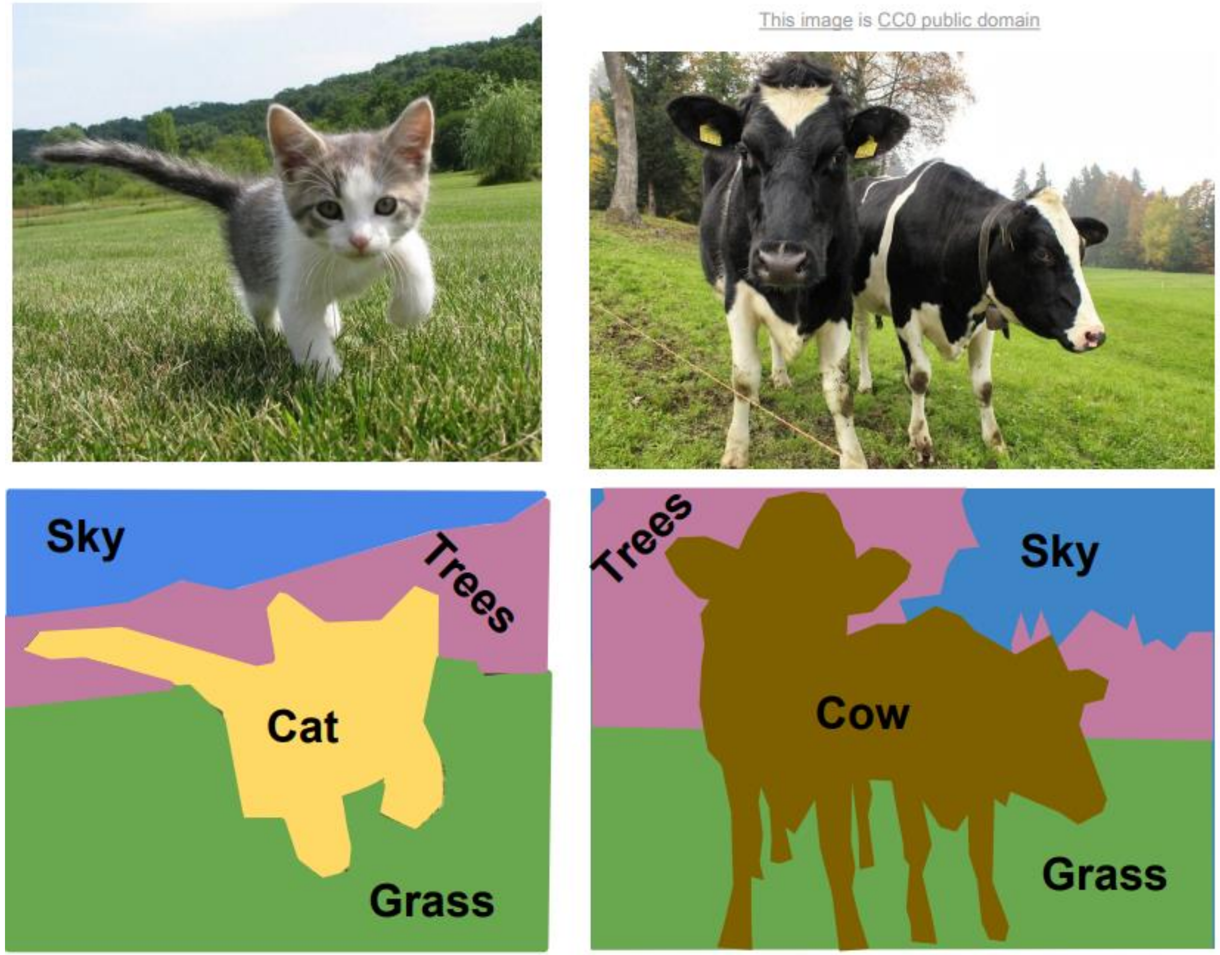


Image Classification – current state



The **ImageNet** dataset contains 14,197,122 annotated images according to the WordNet hierarchy. Since 2010 the dataset is used in the ImageNet Large Scale Visual Recognition Challenge (ILSVRC), a benchmark in image classification and object detection.
[8]

Semantic Segmentation



This image is CC0 public domain

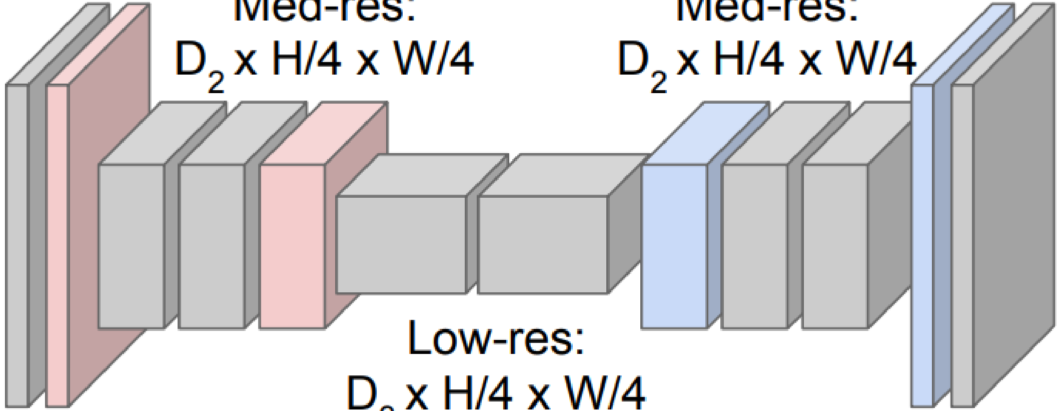
Semantic Segmentation

Downsampling: Pooling, strided convolution

Design network as a bunch of convolutional layers, with **downsampling** and **upsampling** inside the network!

Upsampling: ???

Input:
$3 \times H \times W$



Med-res:
$D_2 \times H/4 \times W/4$

Med-res:
$D_2 \times H/4 \times W/4$

Low-res:
$D_3 \times H/4 \times W/4$

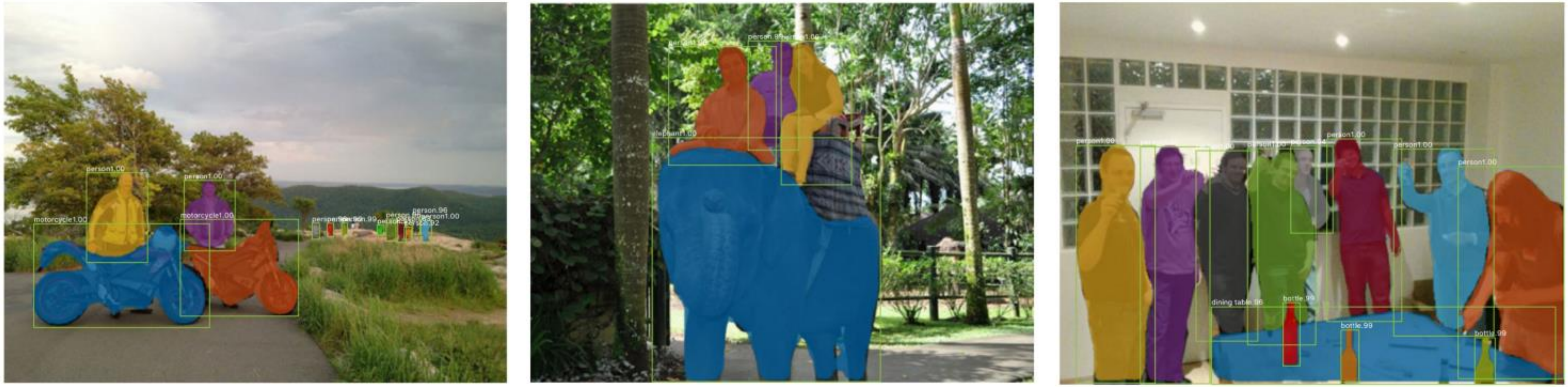
High-res:
$D_1 \times H/2 \times W/2$

High-res:
$D_1 \times H/2 \times W/2$

Predictions:
$H \times W$

[10]

Semantic Segmentation



Mask R-CNN paper: https://arxiv.org/pdf/1703.06870.pdf

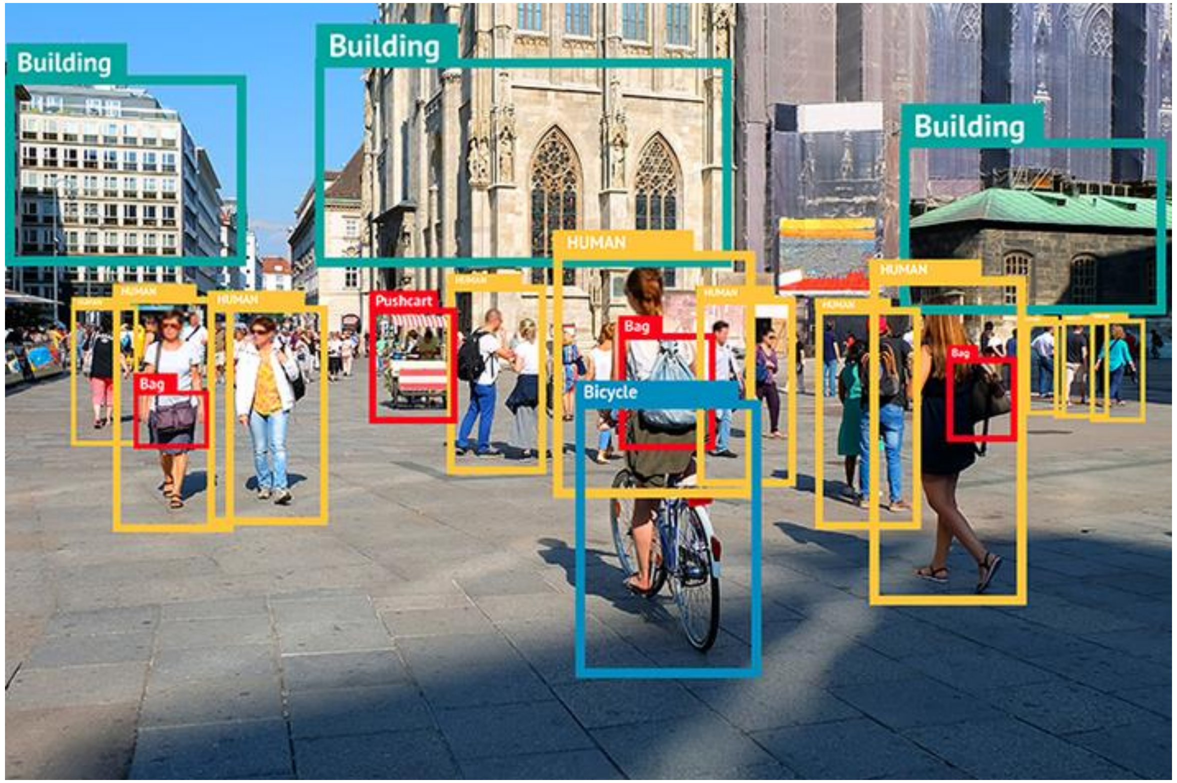
Object Detection

- **One-stage detectors**:

Object classification and bounding-box regression are done directly without using pre-generated region proposals (candidate object bounding-boxes)

- **Two-stage detectors**:

1. *Generation of region proposals*, e.g. by selective search as in R-CNN and Fast R-CNN, or by a Region Proposal Network (RPN) as in Faster R-CNN.
2. *Object classification for each region proposal*. Additionally other things can be done such as bounding-box regression for refining the region proposals, binary-mask prediction etc.



Object Detection

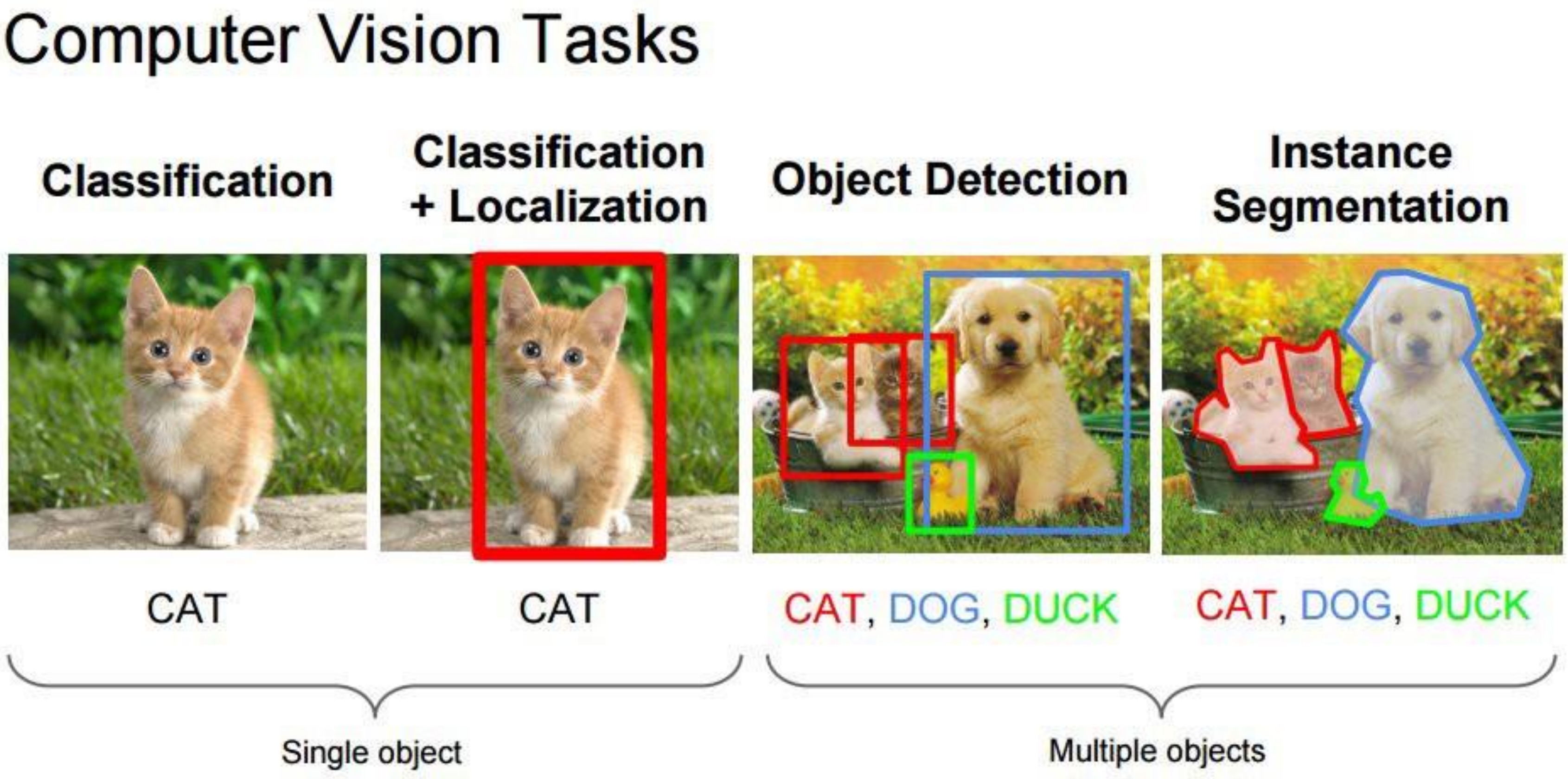
Most important **two-stage** object detection algorithms

- RCNN and SPPNet (2014)
- Fast RCNN and Faster RCNN (2015)
- Mask R-CNN (2017)
- Pyramid Networks/FPN (2017)
- G-RCNN (2021)

Most important **one-stage** object detection algorithms

- YOLO (2016)
- SSD (2016)
- RetinaNet (2017)
- YOLOv3 (2018)
- YOLOv4 (2020)
- YOLOR (2021)

Image Recognition



[1], [2]

Image Generation

- Imagen, a text-to-image diffusion model with an unprecedented degree of photorealism and a deep level of language understanding. Imagen builds on the power of large transformer language models in understanding text and hinges on the strength of diffusion models in high-fidelity image generation.
- DALL·E 2 combines two approaches for the problem of text-conditional image generation. We first train a diffusion decoder to invert the CLIP image encoder. Our inverter is non-deterministic, and can produce multiple images corresponding to a given image embedding. The presence of an encoder and its approximate inverse (the decoder) allows for capabilities beyond text-to-image translation.

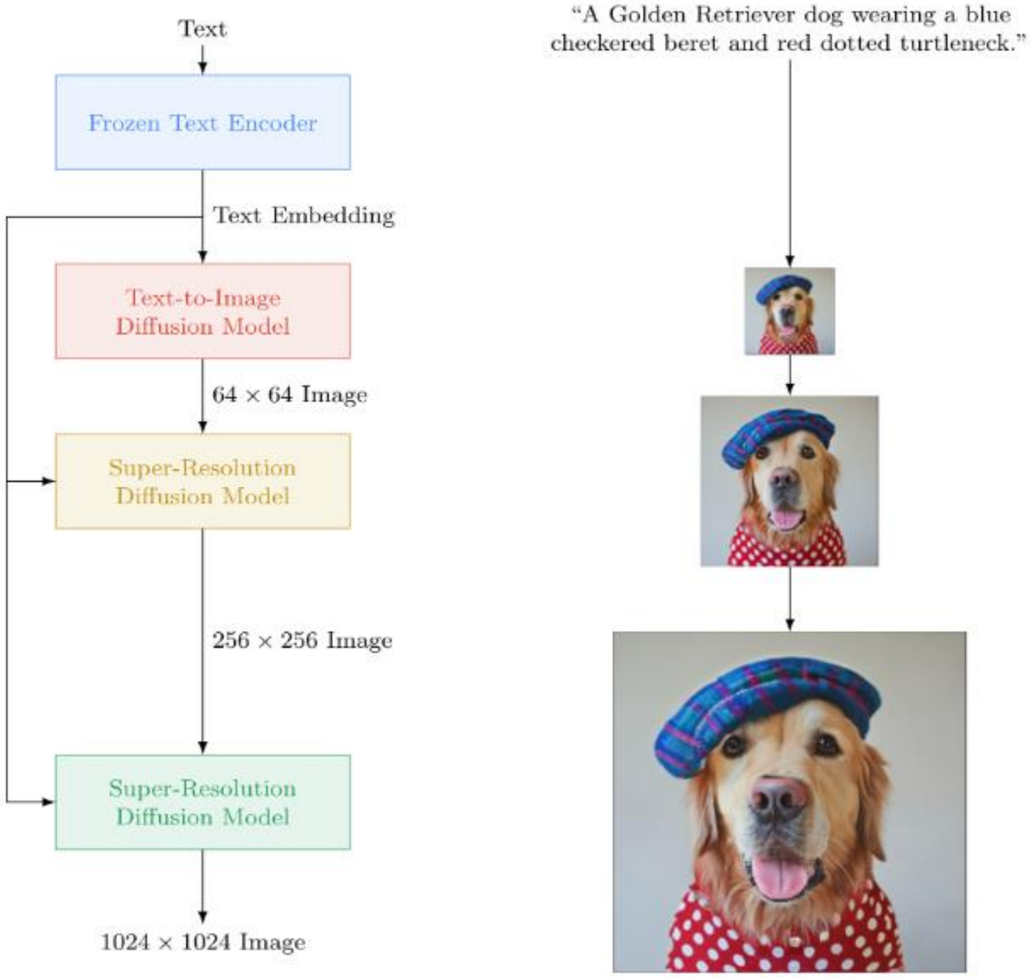
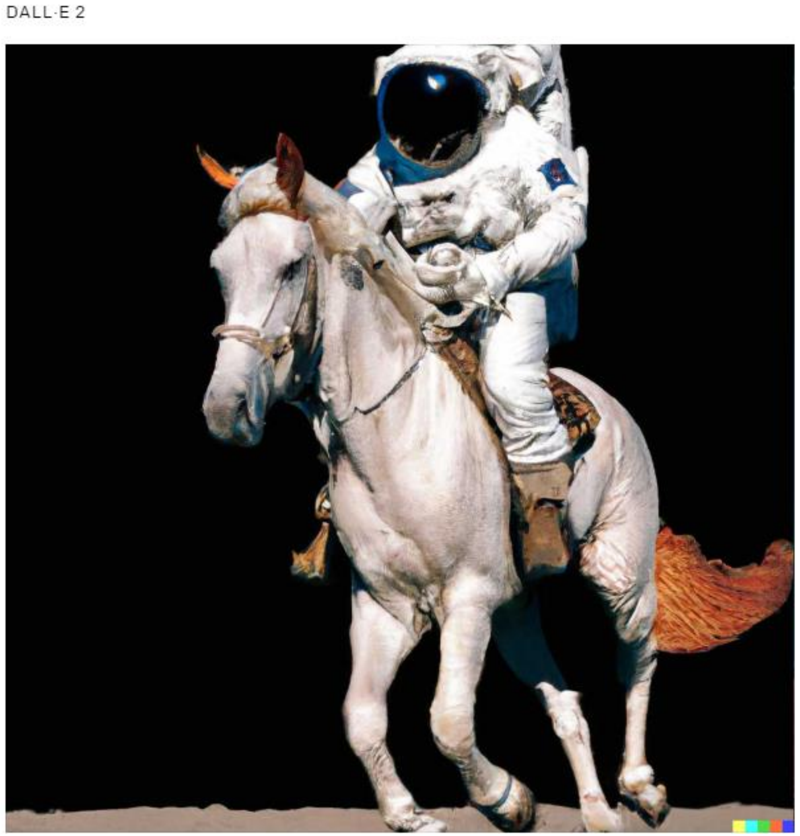

Text
Frozen Text Encoder
Text Embedding
Text-to-Image Diffusion Model
64 × 64 Image
Super-Resolution Diffusion Model
256 × 256 Image
Super-Resolution Diffusion Model
1024 × 1024 Image
"A Golden Retriever dog wearing a blue checkered beret and red dotted turtleneck."

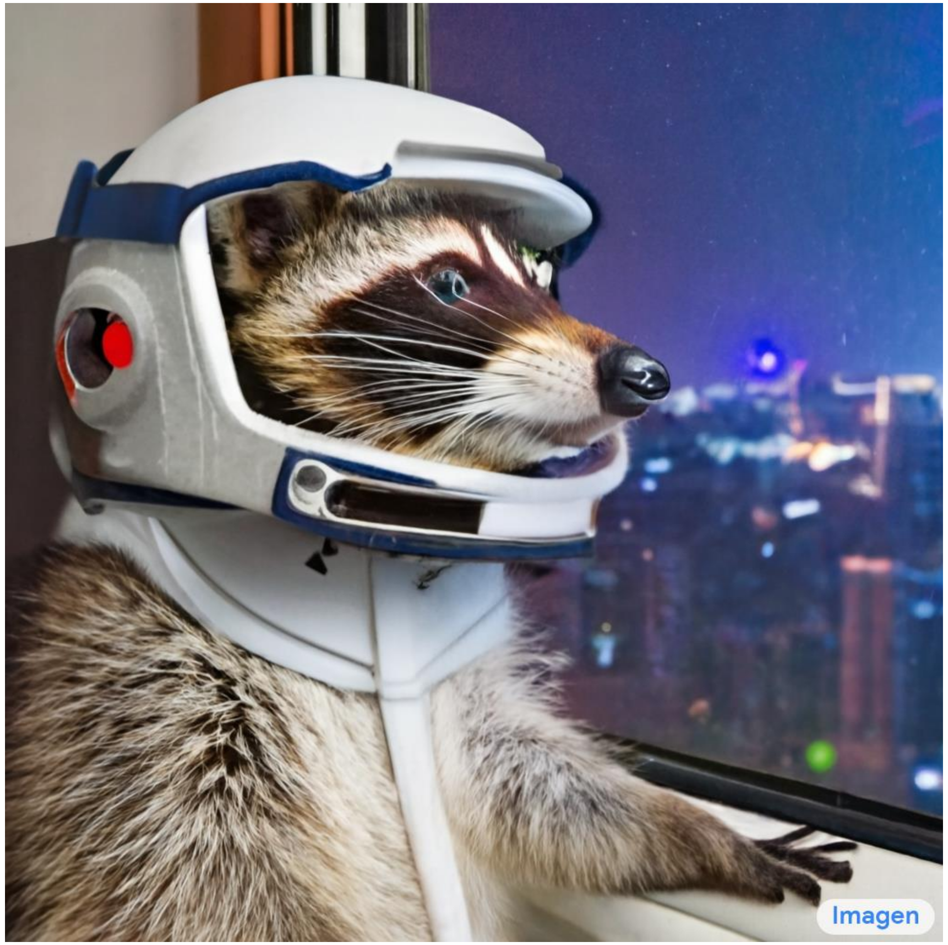
Imagen high-level flow

Image Generation

[6]



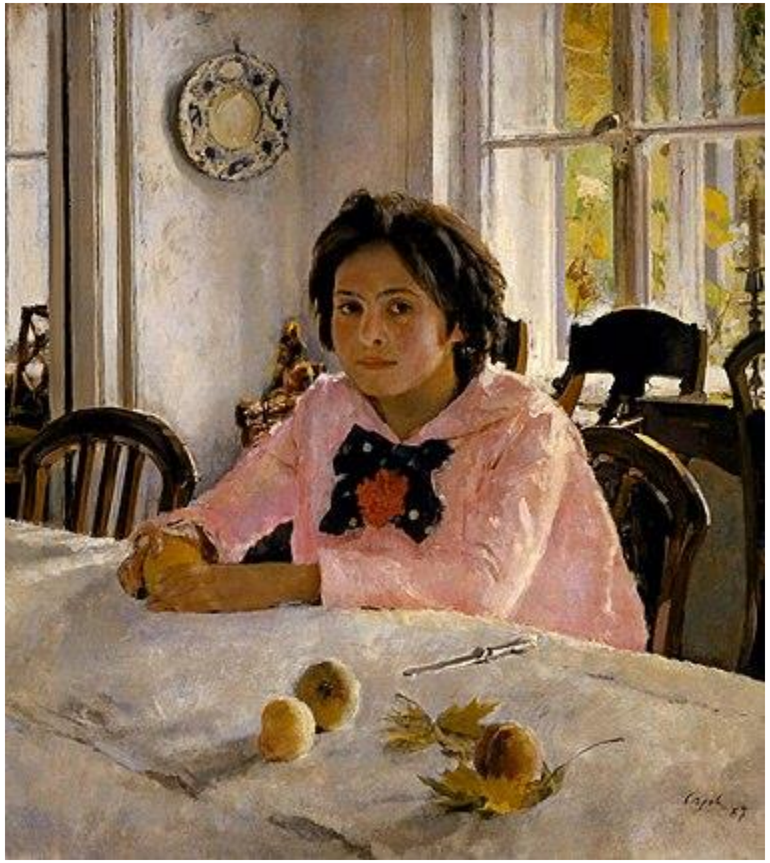
An astronaut riding a in space in a photorealistic style



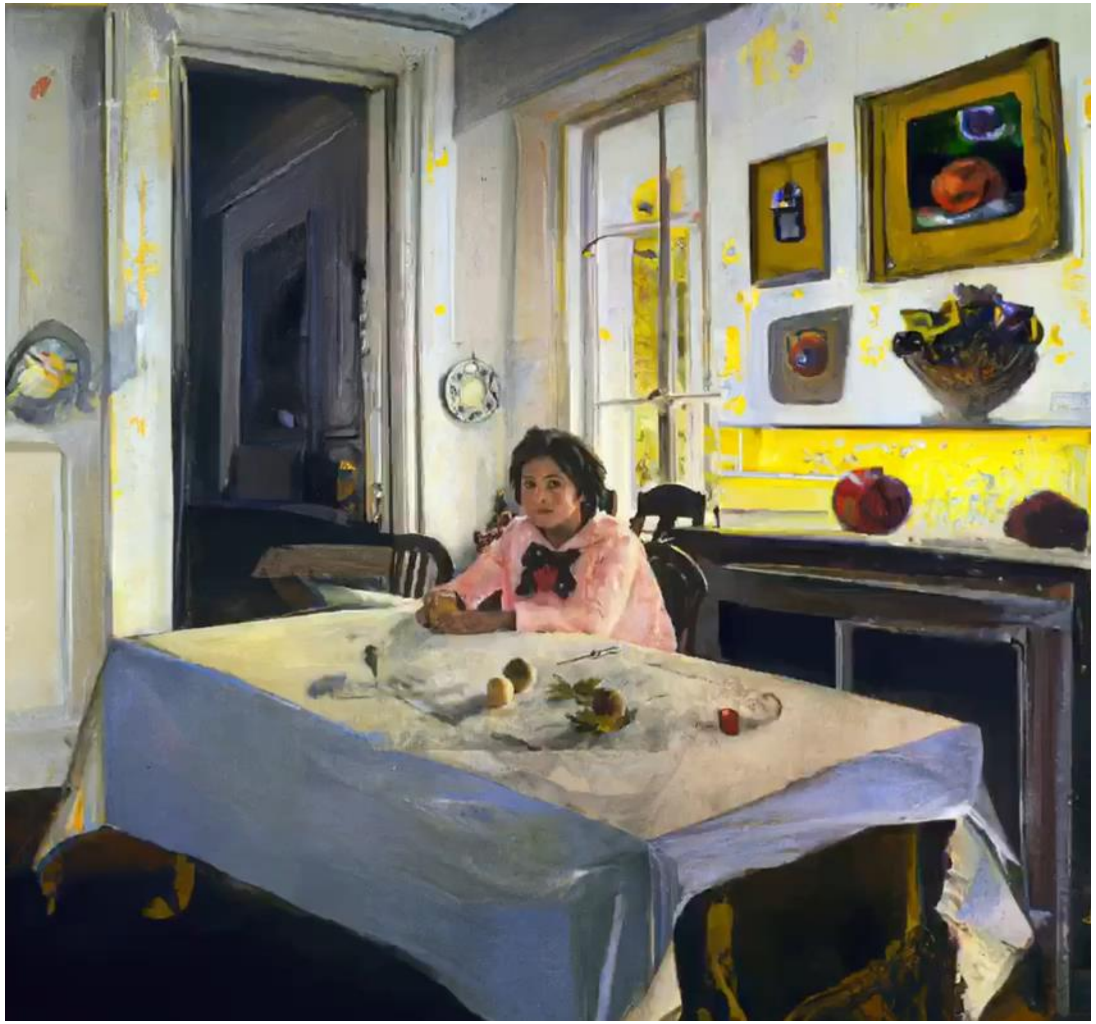
[7]

A photo of a raccoon wearing an astronaut helmet, looking out of the window at night.

Image Generation



Valentin Serov
«*Girl with Peaches*» (1887)

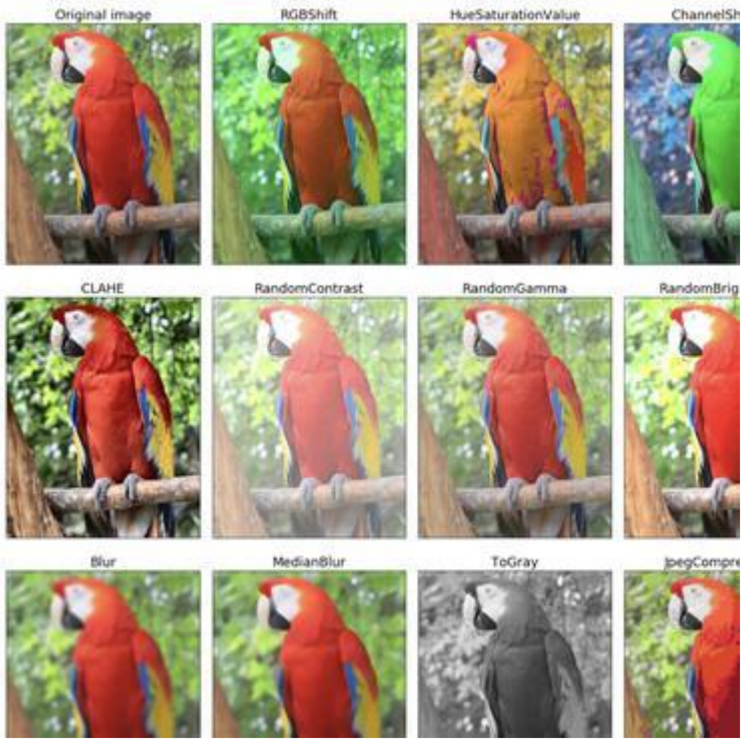


[4]

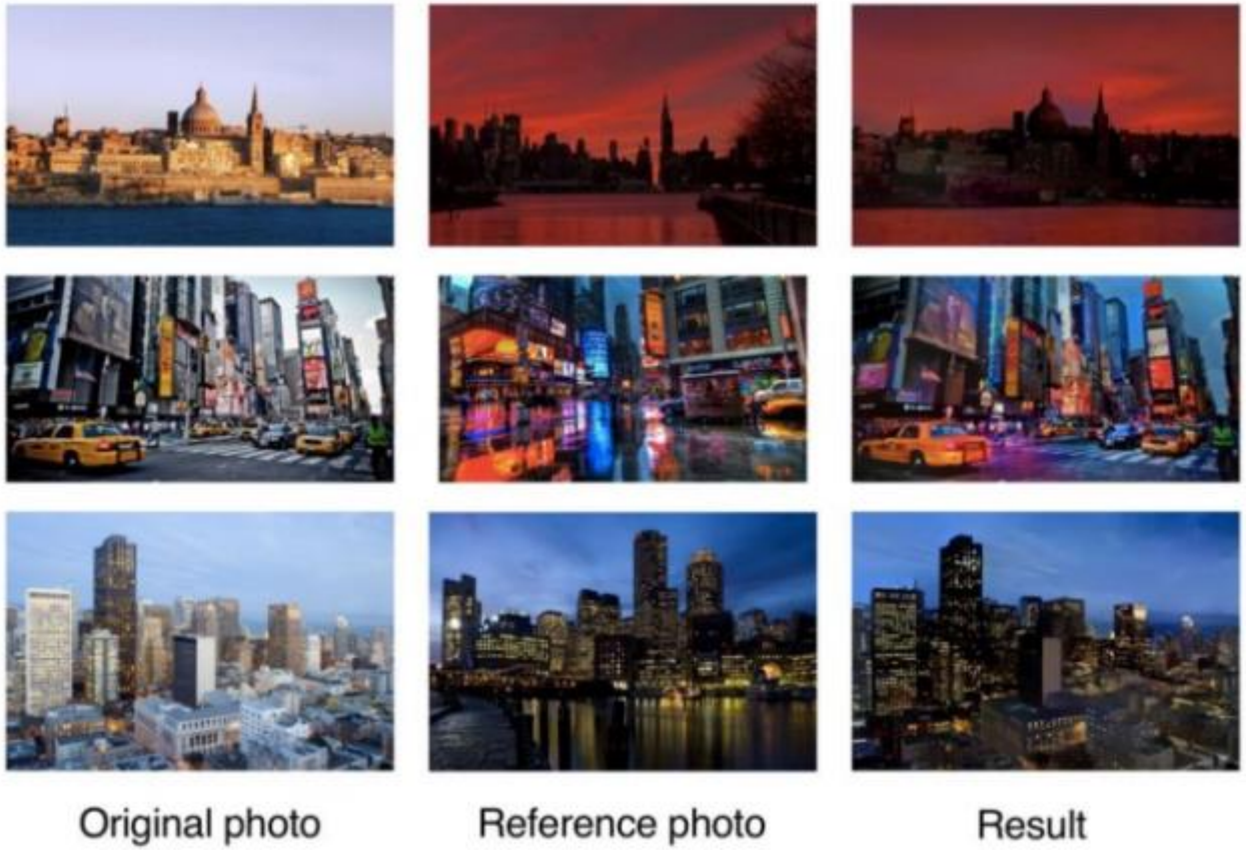
Data Augmentation

Data augmentation involves techniques used for increasing the amount of data, based on different modifications, to expand the amount of examples in the original dataset. Data augmentation not only helps to grow the dataset but it also increases the diversity of the dataset. When training machine learning models, data augmentation acts as a regularizer and helps to avoid overfitting.

[13]



Data Augmentation



Original photo

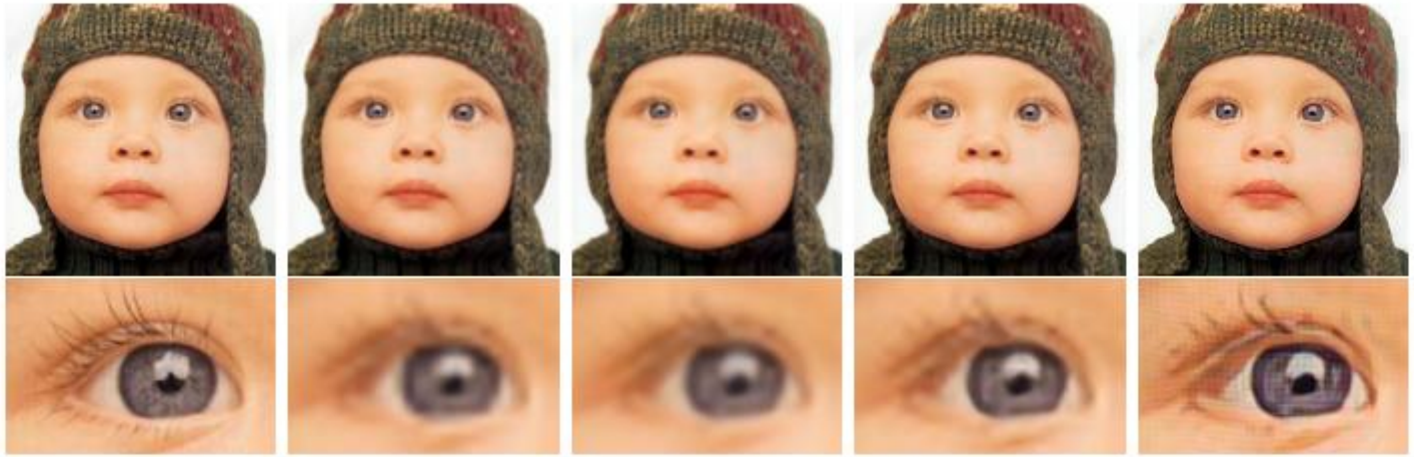
Reference photo

Result

Neural style transfer: It grabs the texture/ambiance/appearance of one image (aka, the “style”) and mixes it with the content of another. [11]

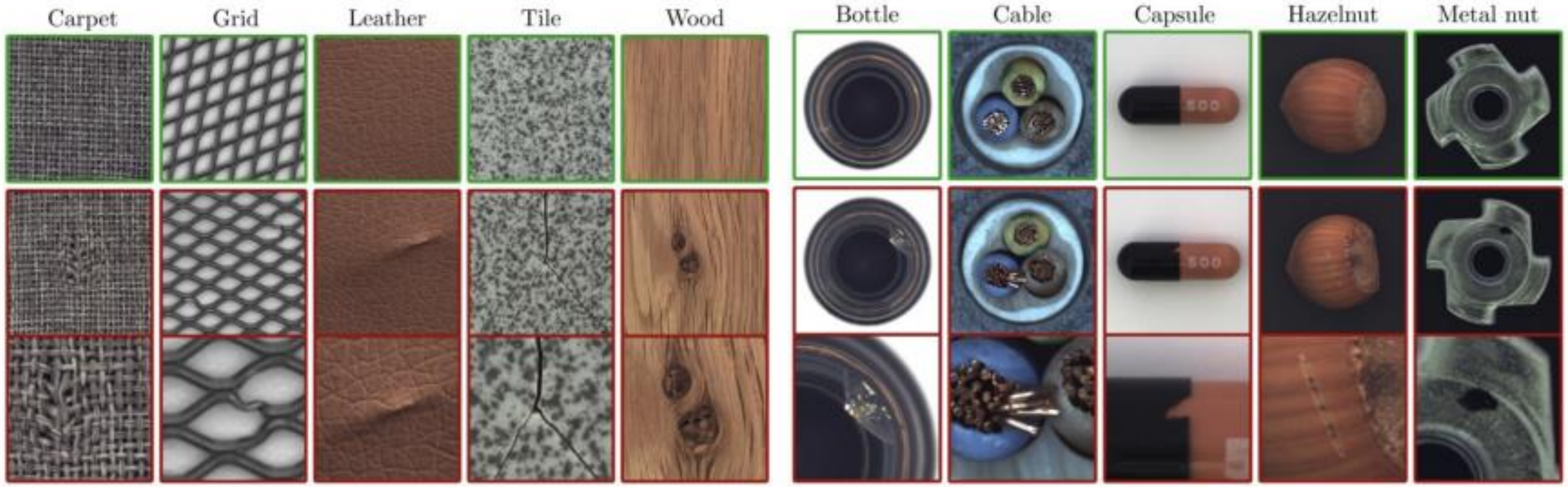
Super-Resolution

Super resolution is the task of taking an input of a low resolution (LR) and upscaling it to that of a high resolution.



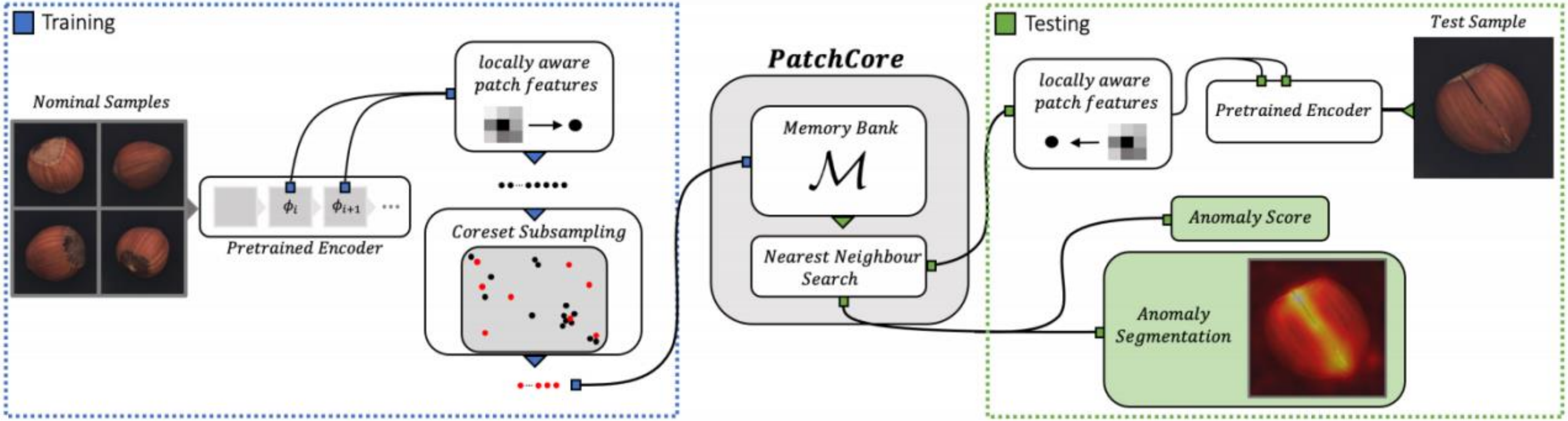
Ground Truth **Bicubic** **Ours** (ℓ_{pixel}) **SRCNN** [11] **Ours** (ℓ_{feat})

Anomaly Detection



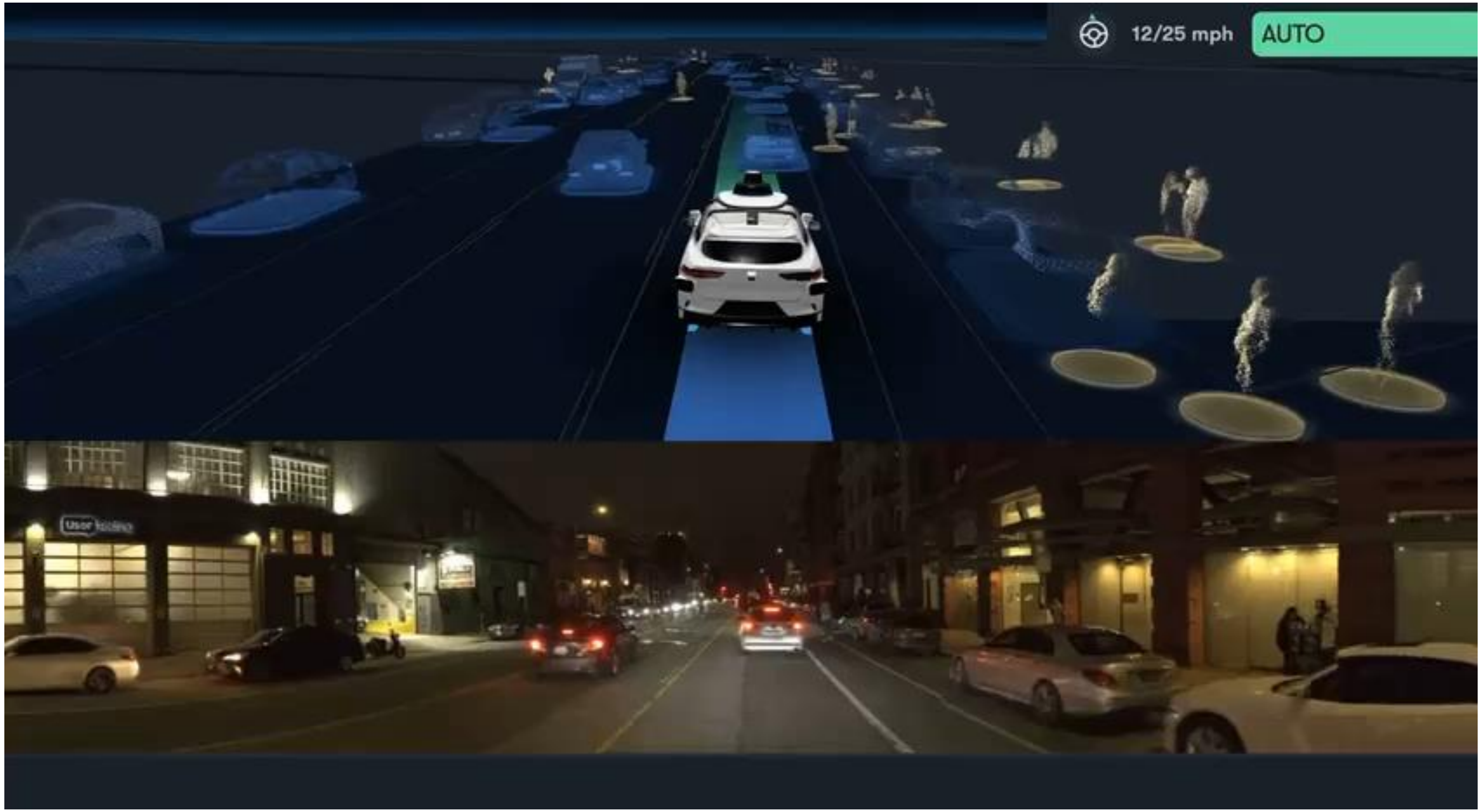
Example images of the MVTec AD dataset. For each category, the top row shows an anomaly-free image. The middle row shows an anomalous example. In the bottom row, a close-up view that highlights the anomalous region is provided. [14]

Anomaly Detection



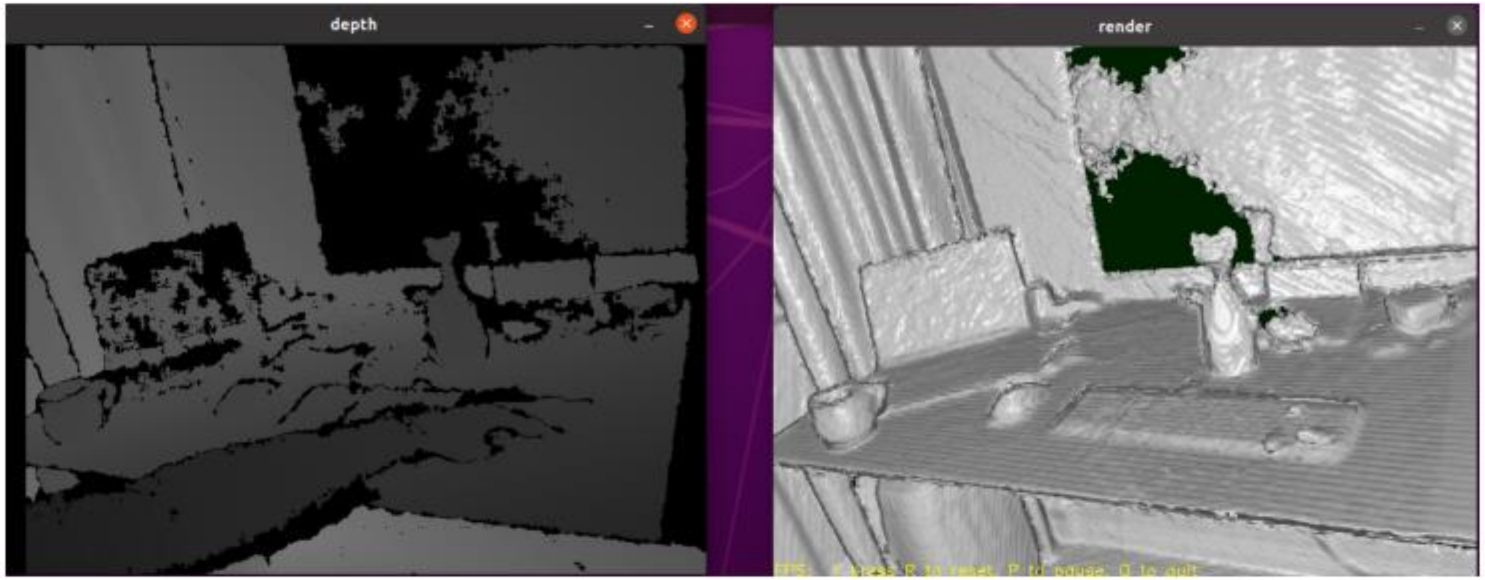
Overall framework of PatchCore. During training, normal samples are decomposed into a memory bank of neighbourhood-aware patch-level features. To reduce redundancy and inference time, this memory bank is downsampled via greedy coreset subsampling algorithm. During test time, images are classified as anomalies if at least one patch is anomalous, and pixel-level anomaly segmentation is generated by scoring each patch-feature.
[14]

Autonomous Driving



[5]

3D-model from depth camera



```
FPS: 8 Press R to reset, P to pause, Q to quit
```

…and many more

- 2D Human Pose Estimation
- 3D Face Animation
- Facial Recognition
- Depth Estimation
- Optical Character Recognition
- Medical Image Segmentation
- Video Classification
- Scene Parsing
- …

Q&A

References

1. https://twitter.com/MikeTamir
2. https://towardsdatascience.com/everything-you-ever-wanted-to-know-about-computer-vision-heres-a-look-why-it-s-so-awesome-e8a58dfb641e
3. Canziani, Alfredo, Adam Paszke, and Eugenio Culurciello. "An analysis of deep neural network models for practical applications." *arXiv preprint arXiv:1605.07678* (2016).
4. https://t.me/denissexy/5833
5. https://twitter.com/dmitri_dolgov/status/1535648456550670339
6. https://openai.com/dall-e-2/
7. https://imagen.research.google/

References

8. https://paperswithcode.com/dataset/imagenet
9. https://arxiv.org/pdf/1602.06541.pdf "A Survey of Semantic Segmentation" by Martin Thoma
10. http://cs231n.stanford.edu/slides/2017/cs231n_2017_lecture11.pdf
11. https://arxiv.org/abs/1703.07511 "Deep Photo Style Transfer" by Fujun Luan, Sylvain Paris, Eli Shechtman, Kavita Bala
12. http://cds.cern.ch/record/400313/files/p21.pdf Computer Vision: Evolution and Promise by T. S. Huang
13. https://paperswithcode.com/task/data-augmentation/codeless
14. https://arxiv.org/pdf/2204.11161.pdf "A Survey on Unsupervised Industrial Anomaly Detection Algorithms" by Yajie Cui et. al.

Thank you!