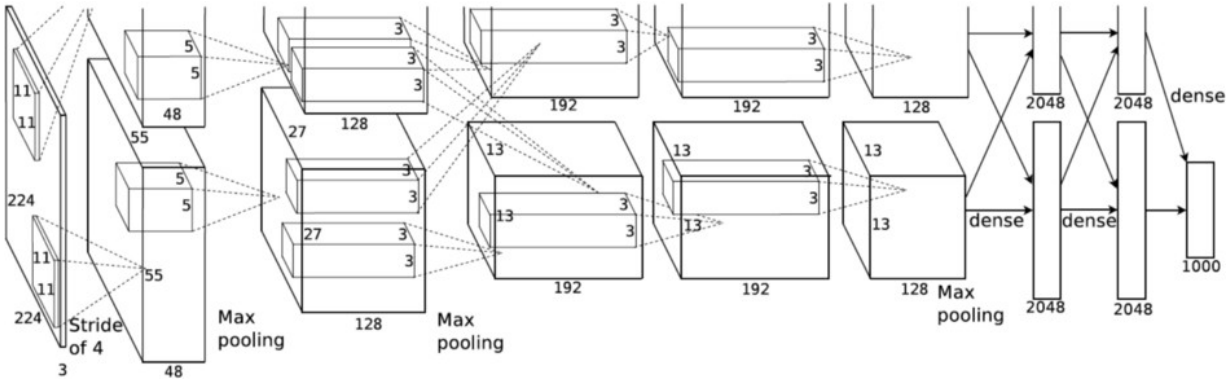


COMP9517: Computer Vision

Deep Learning Part 2

Outlines

Recap: CNNs for supervised image classification

- Beyond classification
- Beyond single image input
- Beyond strong supervision

Vision Beyond Classification

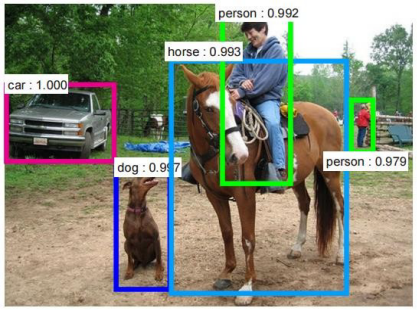
- An image is worth a thousand words
- Classification models learn only a few
- Resnet-50: bicycle, garden

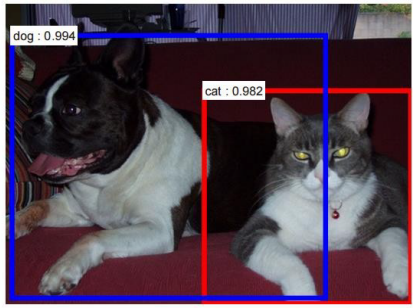
- Holy grail

A model that achieves human level scene understanding

Vision Beyond Classification

Object Detection





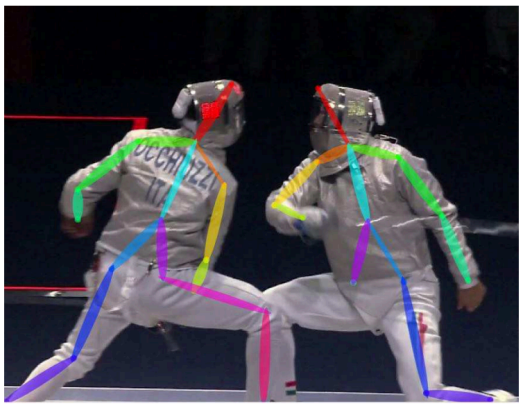
Semantic Segmentation

Scene Understanding

two young girls are playing with lego toy.

boy is doing backflip on wakeboard.

Pose Estimation

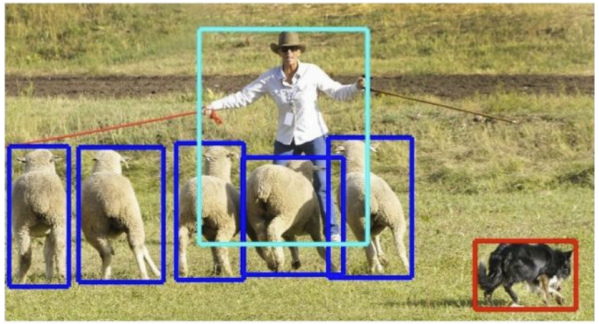


References and further reading: https://github.com/kjw0612/awesome-deep-vision

Vision Beyond Classification

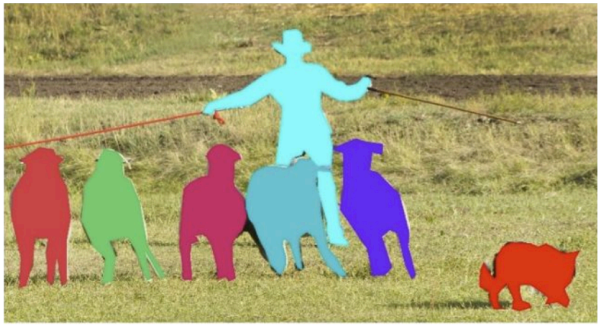
Identified Tasks

Object Detection



Semantic Segmentation

Instance Segmentation



Figures from [Microsoft COCO: Common Objects in Context, Lin et al, 2014]

Object Detection

- Multi-task

  classification + localization
- Input

  an RGB image
- Targets

  class label + bounding box

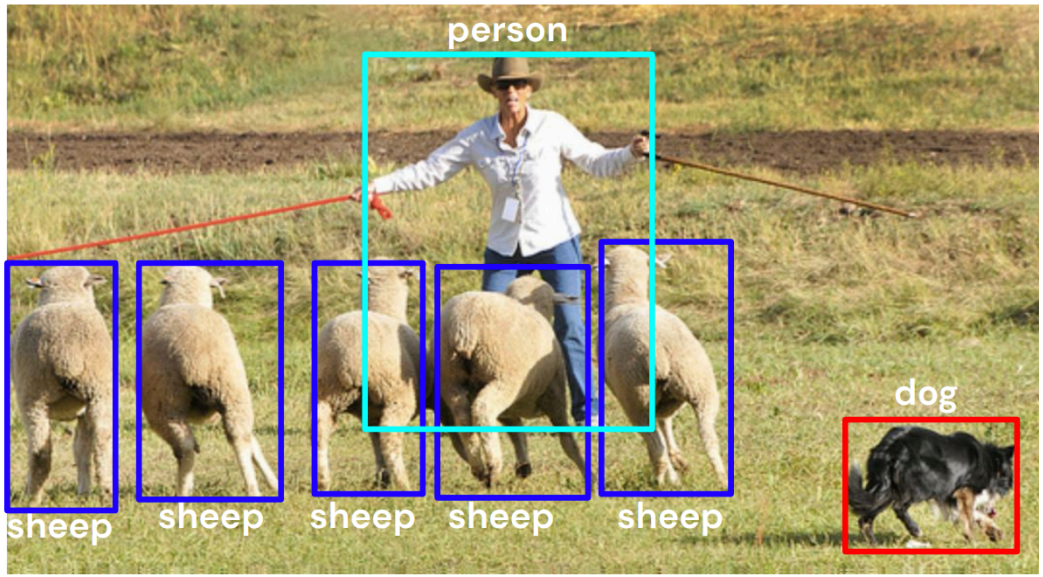
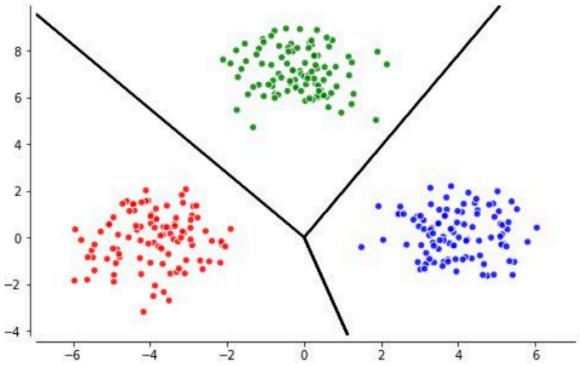


Image from COCO dataset - Microsoft COCO: Common Objects in Context, Lin et al, 2014

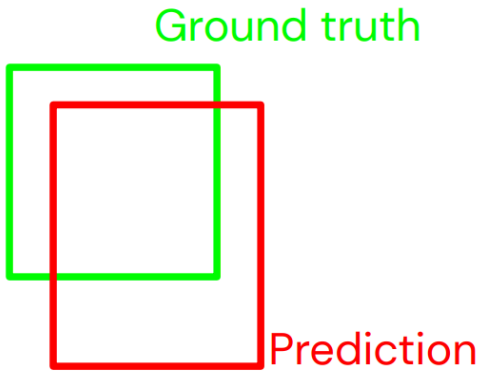
Object Detection

How to learn to predict class label + bounding box?

Recap:



Classification



Regression

- Softmax + cross entropy for classification

- Quadratic loss for regression

$$l_2(x, y) = ||y - x||^2$$

Object Detection

Two categories of deep learning based methods

- Two-stage methods:
  - R-CNN
  - Fast R-CNN
  - **Faster R-CNN**

- One-stage methods:
  - YOLO
  - SSD
  - **RetinaNet**

Object Detection

Faster R-CNN

- Two-stage detector
  i. Identify bboxs
  ii. Classify and refine
- Architecture
  - RPN
  - Fast R-CNN

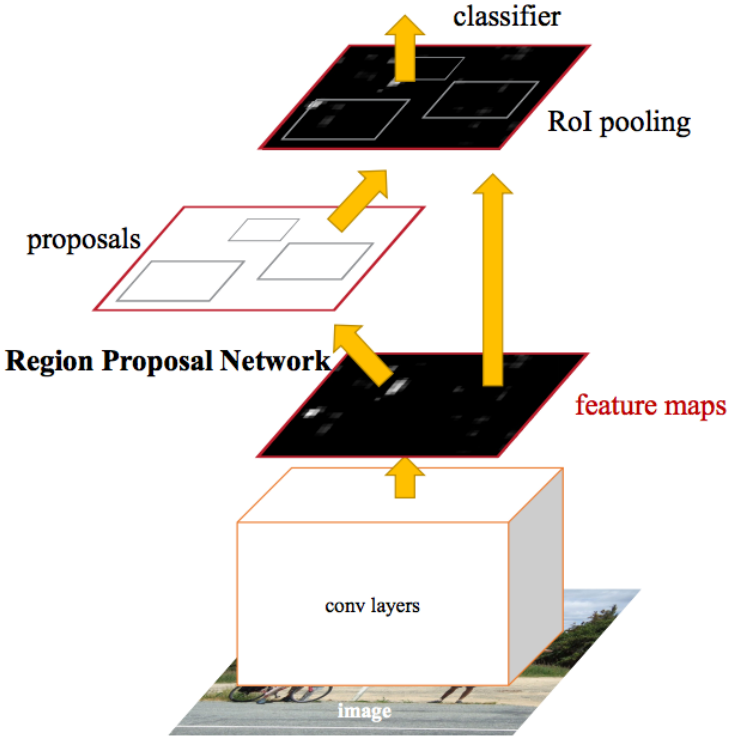
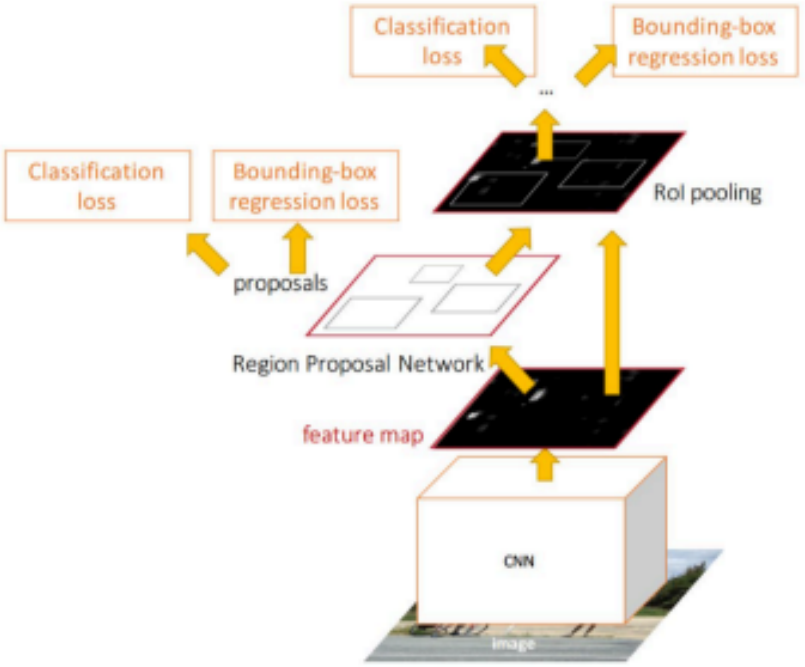


Figure from Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks, Ren et al, 2016

Object Detection

Faster R-CNN

- Region Proposal Network (RPN)



Object Detection

Faster R-CNN

- Region Proposal Network (RPN)

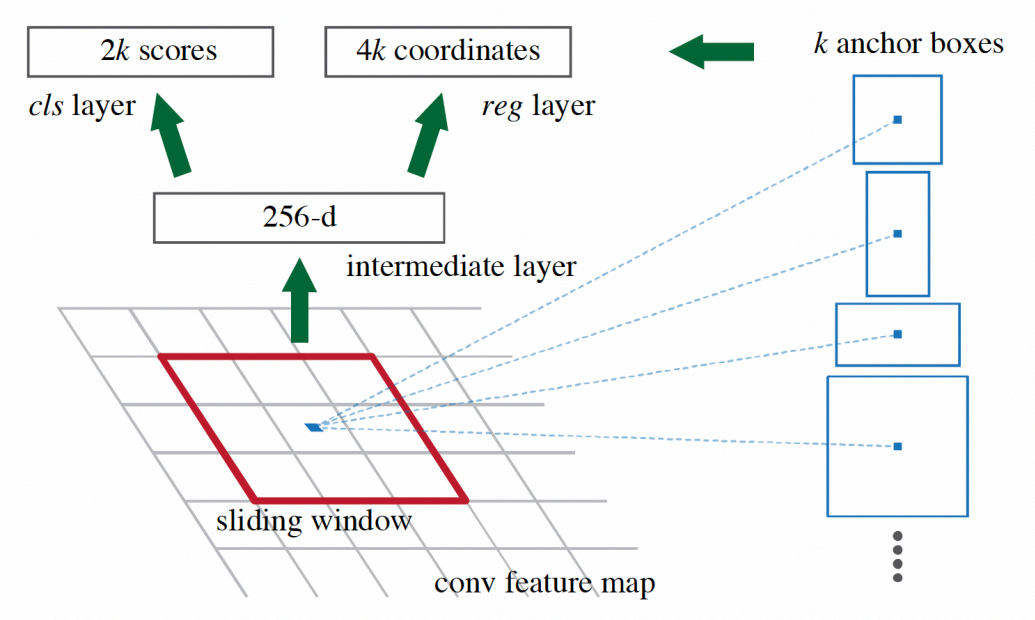
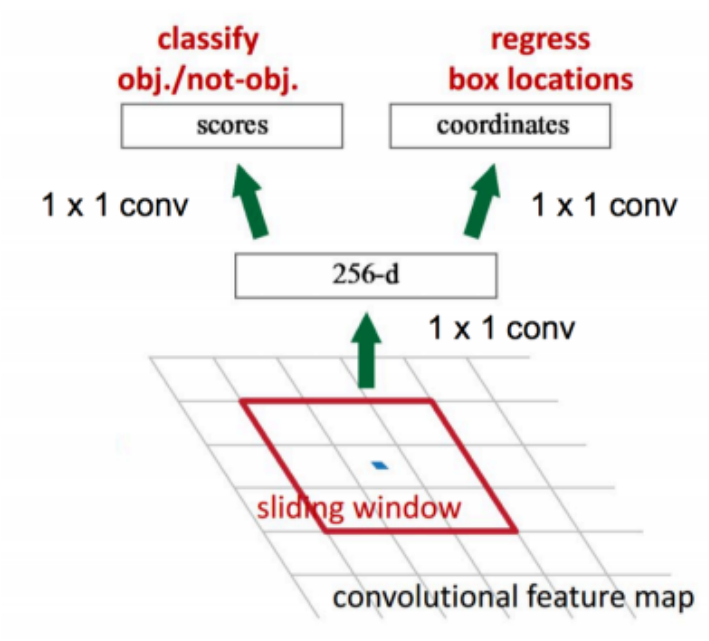


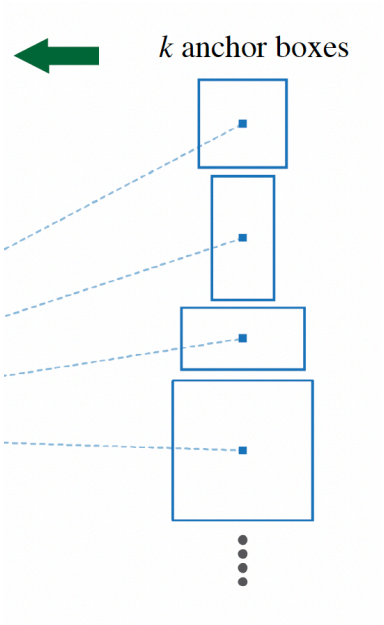
Figure from [Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks, Ren et al, 2016]

Object Detection

Faster R-CNN

- Region Proposal Network (RPN)





Object Detection

RetinaNet

- One-stage detector
- Architecture: ResNet + FPN + three subnets

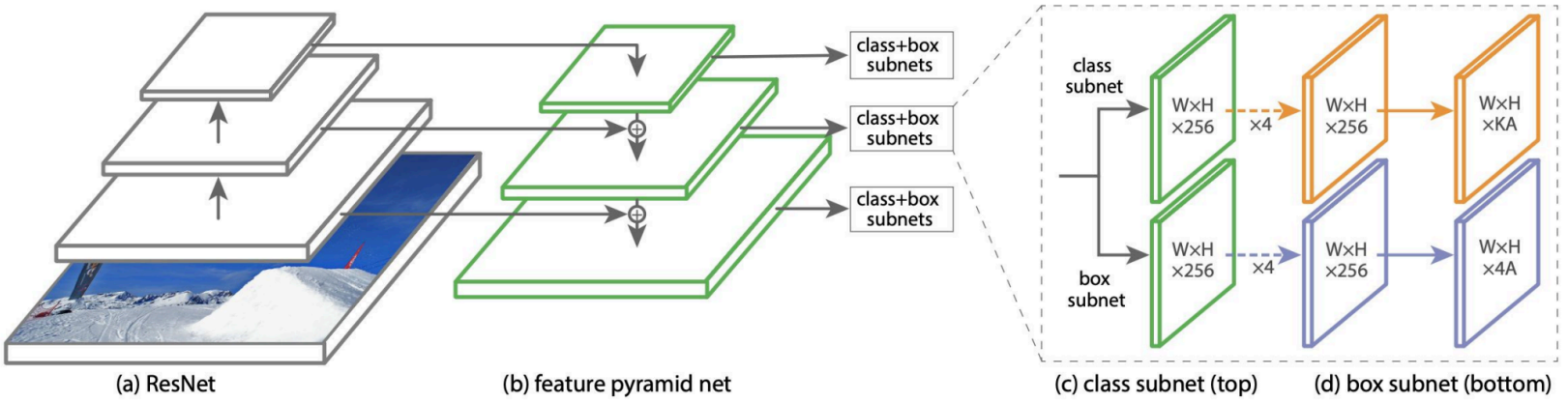


Figure from [Focal Loss for Dense Object Detection, Lin et al, 2017]

Object Detection

RetinaNet

- FPN (feature pyramid network)

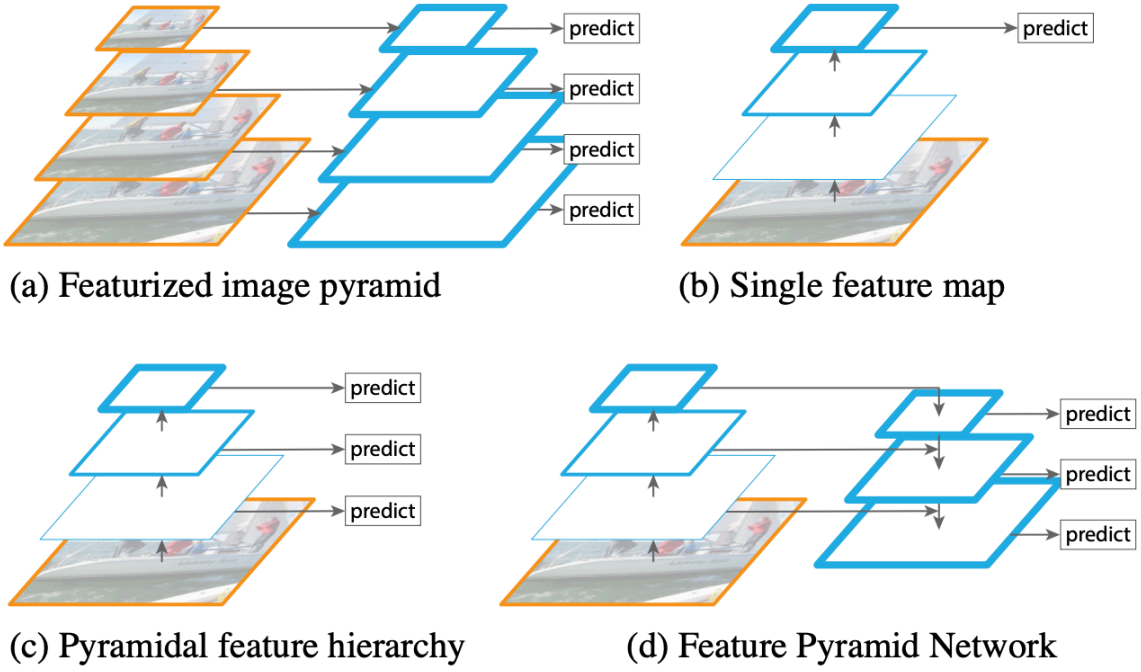


Figure from [Feature Pyramid Networks for Object Detection, Lin et al, 2017]

Object Detection

Challenges with one-stage detectors

- Most of the candidate bounding boxes are background

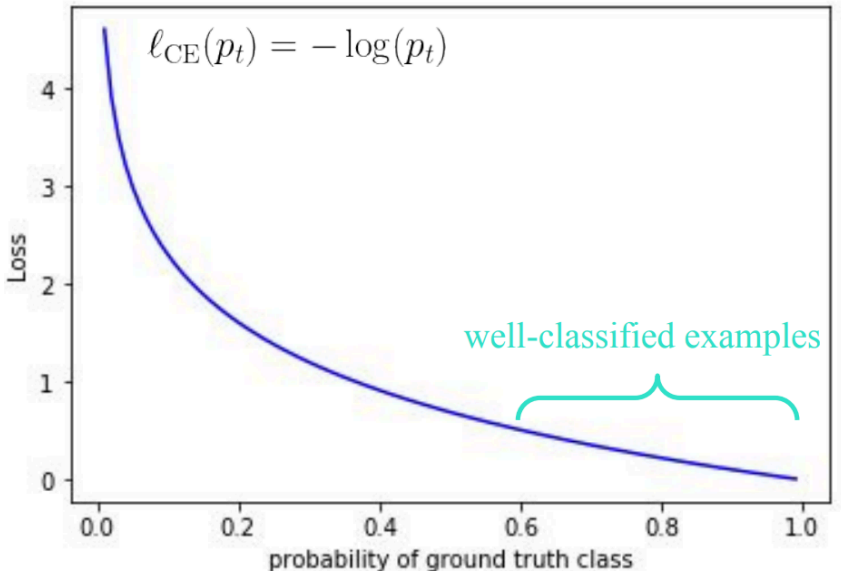


Figure from Focal Loss for Dense Object Detection, Lin et al, 2017

Object Detection

RetinaNet solution

- using Focal Loss (FL)

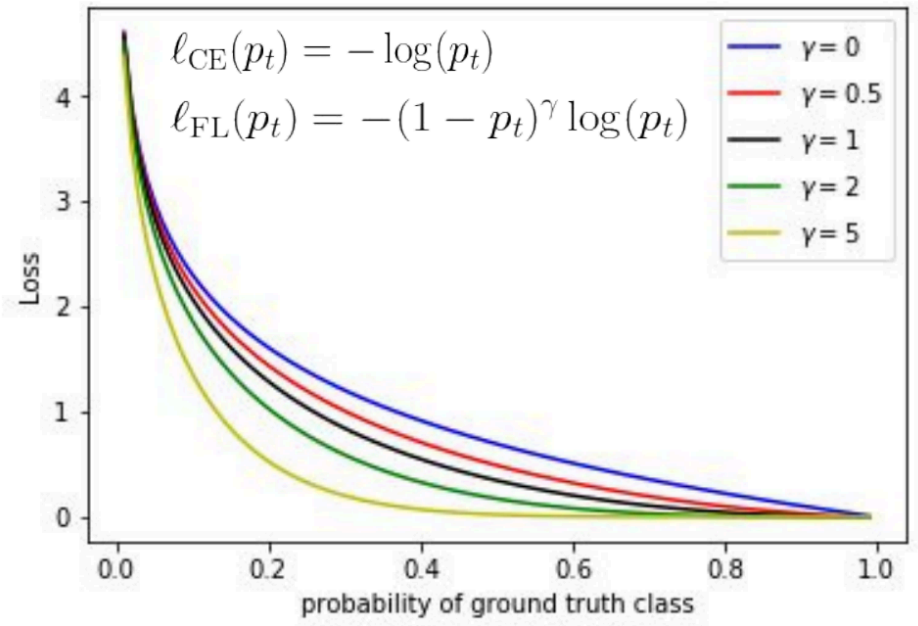


Figure from [Focal Loss for Dense Object Detection, Lin et al, 2017]

Semantic Segmentation

- Input

  an RGB image

- Targets

  class label for every pixel

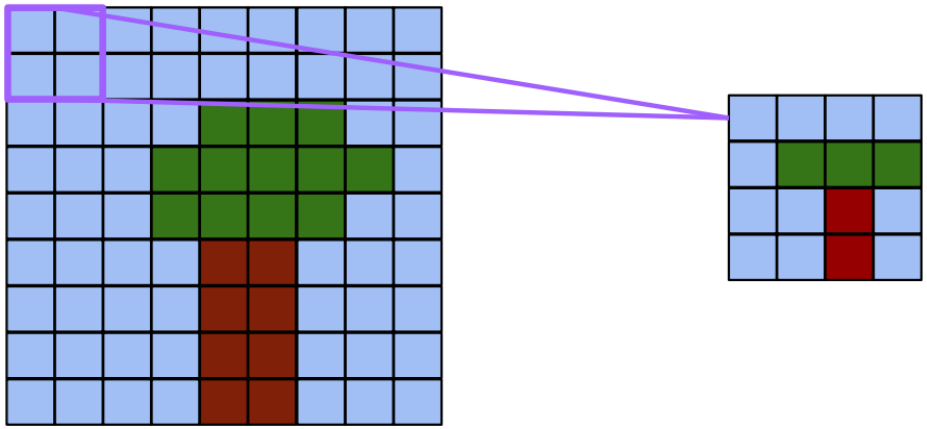
- Dense prediction problem

Image from COCO dataset - Microsoft COCO: Common Objects in Context, Lin et al, 2014

Semantic Segmentation

UpSampling

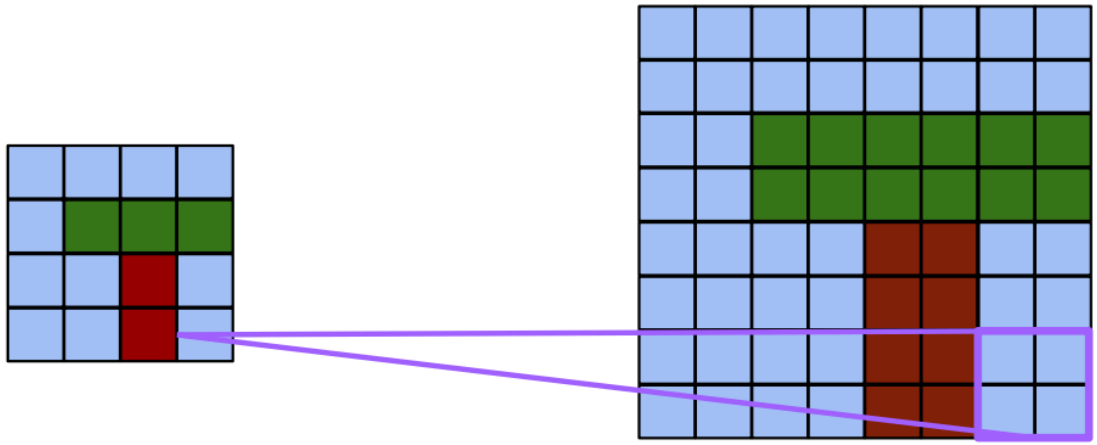
Recap Pooling: compute mean or max over small windows to reduce resolution.



Semantic Segmentation

UpSampling – Unpooling

Upsample to increase resolution; here 2x2 kernel.



Semantic Segmentation

U-Net

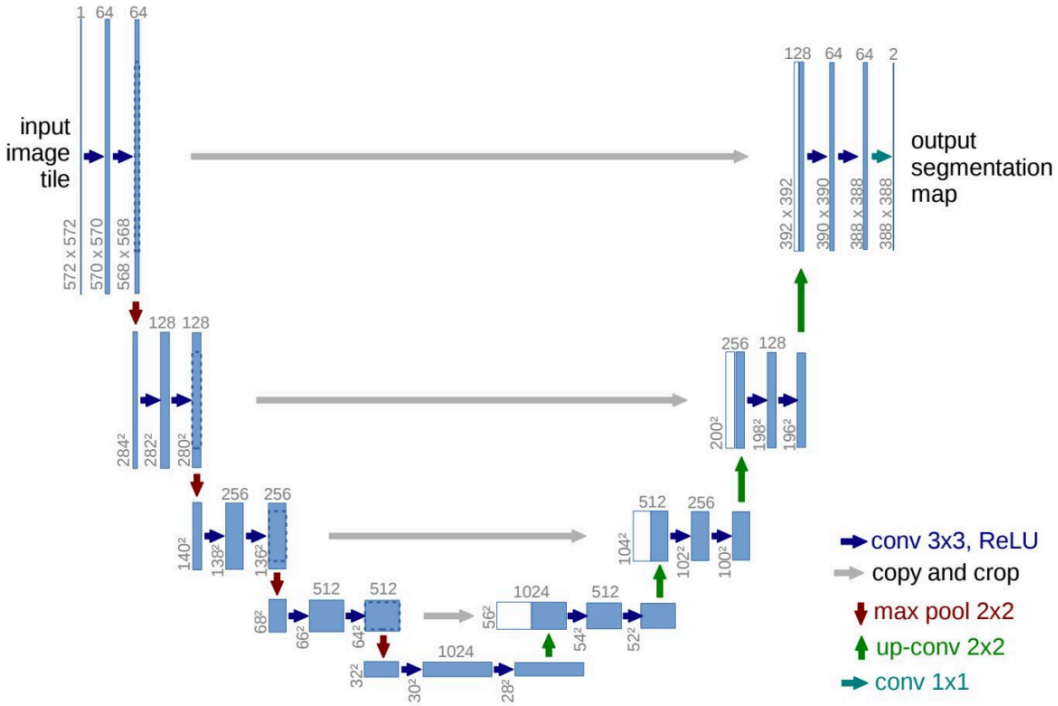


Figure from U-Net: Convolutional Networks for Biomedical Image Segmentation, Ronneberger et al, 2015

Semantic Segmentation

U-Net

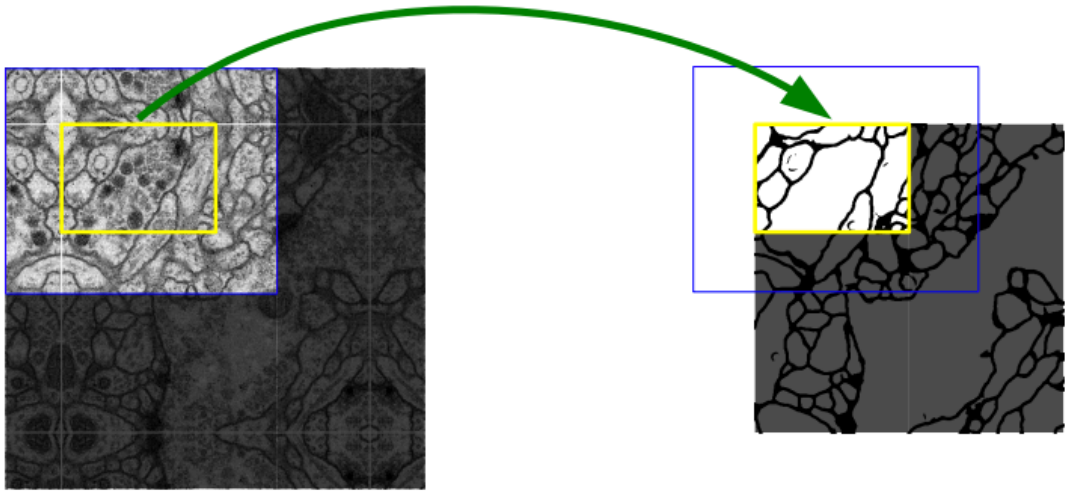


Figure from U-Net: Convolutional Networks for Biomedical Image Segmentation, Ronneberger et al, 2015

Instance Segmentation

- Input

  an RGB image

- Targets

  class label for every instance

- Object detection + segmentation

Image from COCO dataset - Microsoft COCO: Common Objects in Context, Lin et al, 2014

Instance Segmentation

Two categories of methods

- Two-stage methods
  - Top-Down ('detect-then-segment')
    Mask R-CNN
  - Bottom-Up
    Semantic segmentation + instance embedding
- Single stage methods

  YOLACT
  SOLO
  PloarMask
  AdaptIS

Instance Segmentation

Mask R-CNN

- Faster R-CNN + mask head
- ROIAlign instead of ROI pooling in Faster R-CNN

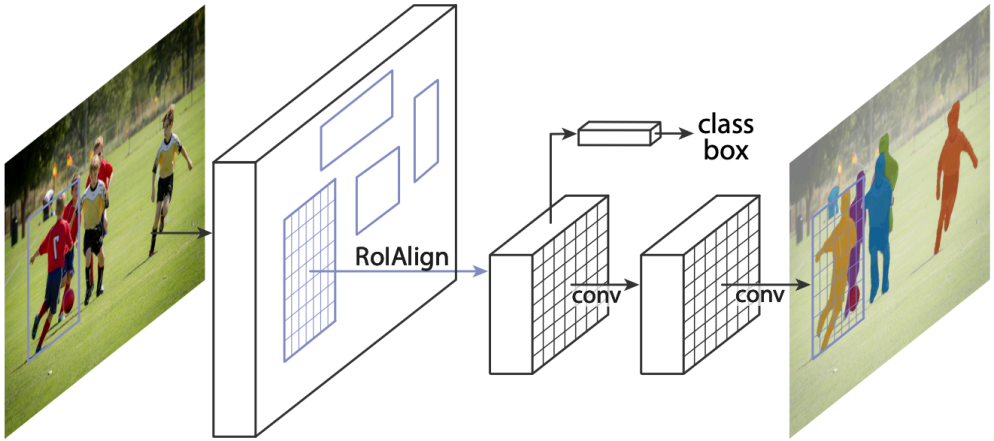
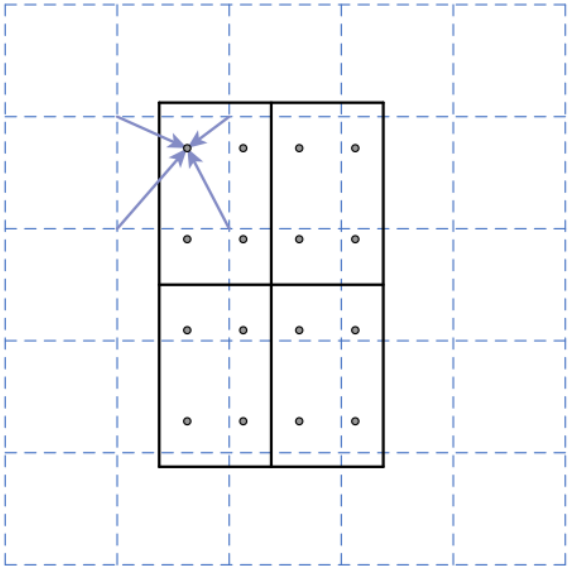


Image from Mask R-CNN, He et al, 2018

Instance Segmentation

Mask R-CNN

- ROIAlign



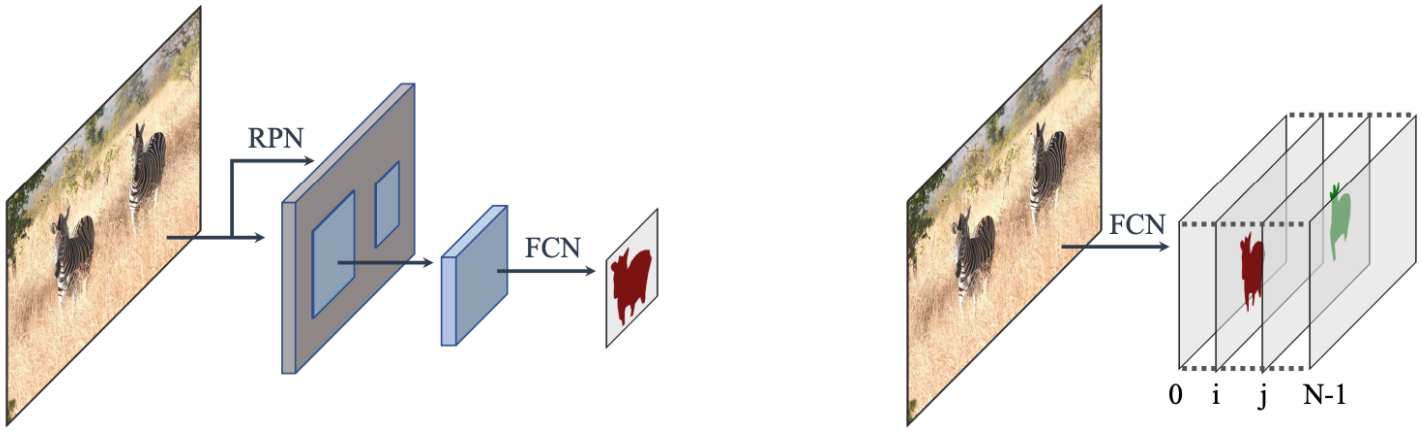
The dashed grid represents a feature map, the solid lines an RoI (with 2×2 bins in this example), and the dots the 4 sampling points in each bin. RoIAlign computes the value of each sampling point by bilinear interpolation from the nearby grid points on the feature map. No quantization is performed on any coordinates involved in the RoI, its bins, or the sampling points.

Image from Mask R-CNN, He et al, 2018

Instance Segmentation

SOLO (**s**egment **o**bjects by **lo**cations)

- Box-free
- the notion of "instance categories", i.e., the quantized center locations and object sizes.



(a) Mask R-CNN (b) SOLO

Image from SOLO: Segmenting Objects by Locations, Wang et al, 2020

Instance Segmentation

SOLO (**s**egment **o**bjects by **lo**cations)

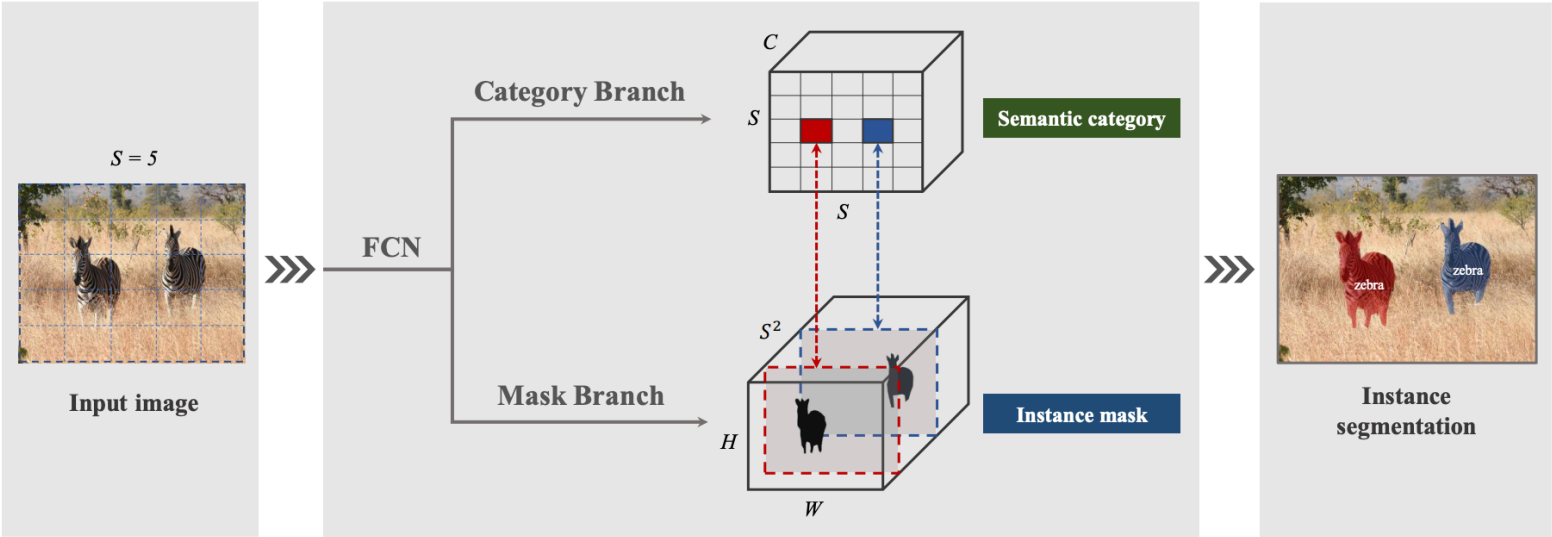


Image from [SOLO: Segmenting Objects by Locations, Wang et al, 2020]

Evaluation Metrics

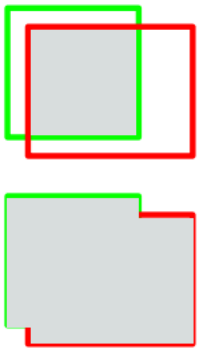
Classification

- Accuracy: percentage of correct predictions

Object detection & segmentation

– Recall image segmentation lecture in week 5

- Intersection-over-union (IoU)

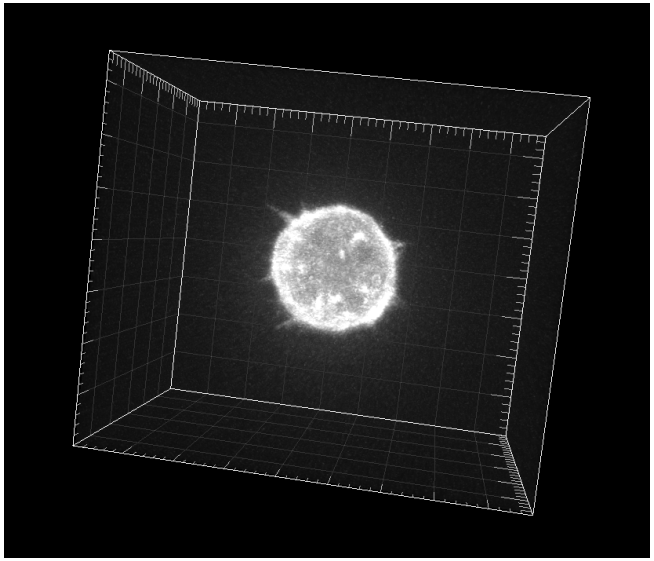


$$\mathcal{J}(\mathbf{P},\mathbf{T}) = \frac{\mathbf{P} \bigcap \mathbf{T}}{\mathbf{P} \bigcup \mathbf{T}}$$

- IoU non-differentiable: used only for evaluation

Beyond single image input

Motion helps object recognition when learning to see.



Video

- Motion – cues for object recognition during learning
- Natural data augmentation: translation, scale, 3D rotation, camera motion, light changes

Beyond single image input
identified Tasks

- Pairs of images input
  optical flow estimation

- Videos input
  Target tracking
  Action recognition

Optical Flow Estimation

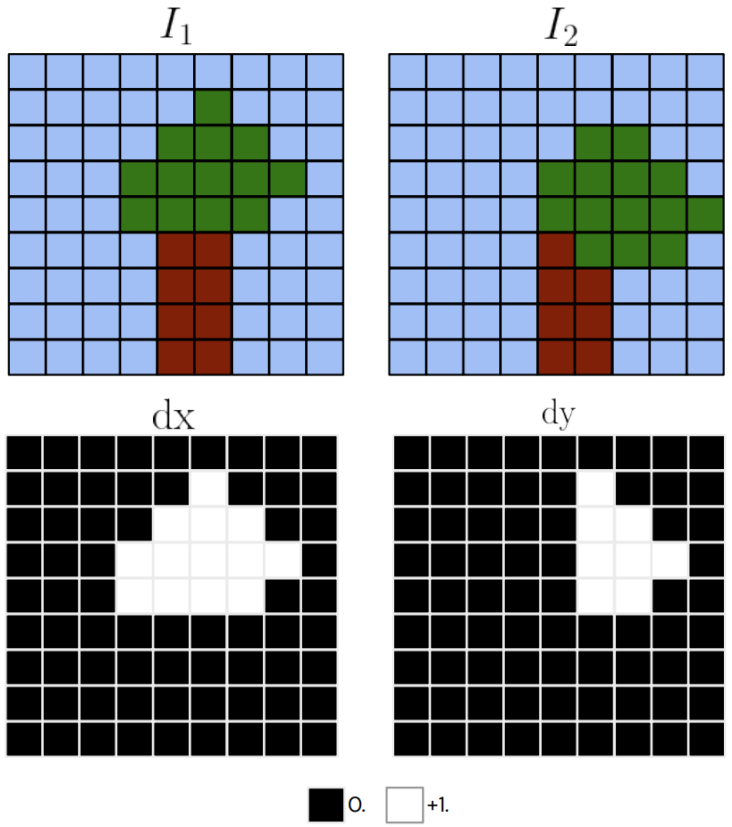
- Input

  A pair of RGB images

- Targets

  Dense flow map (real values)

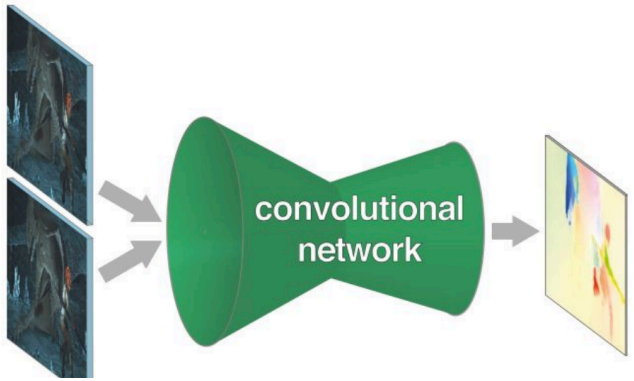
  2D translation displacements



Optical Flow Estimation

FlowNet

- Encoder-decoder architecture (similar to U-NET)
- Supervised training
- Loss: Euclidean distance



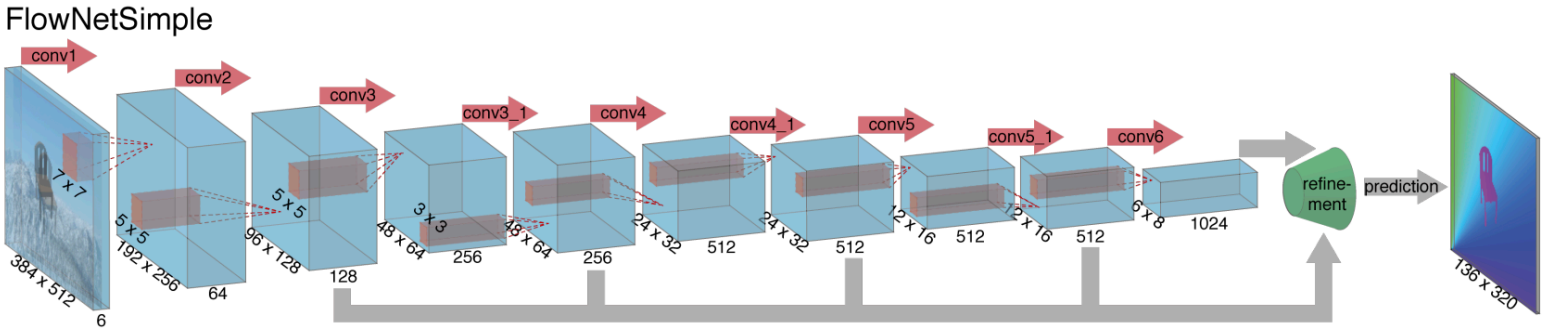
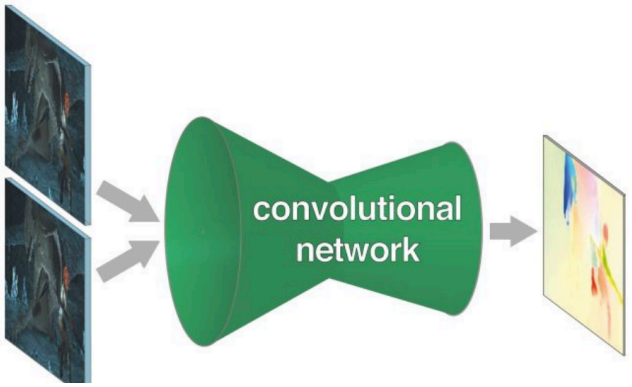


Image from [FlowNet: Learning optical flow with convolutional network, Wang et al, 2020]

Optical Flow Estimation

FlowNet

- Encoder-decoder architecture (similar to U-NET)
- Supervised training
- Loss: Euclidean distance



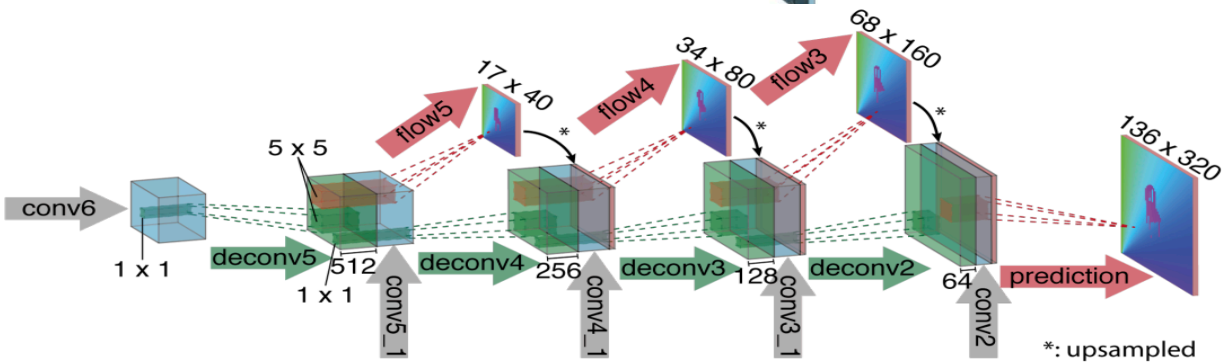


Image from [FlowNet: Learning optical flow with convolutional network, Wang et al, 2020]

“Motion Flow” Estimation

Estimating “motion flow” from a single image?

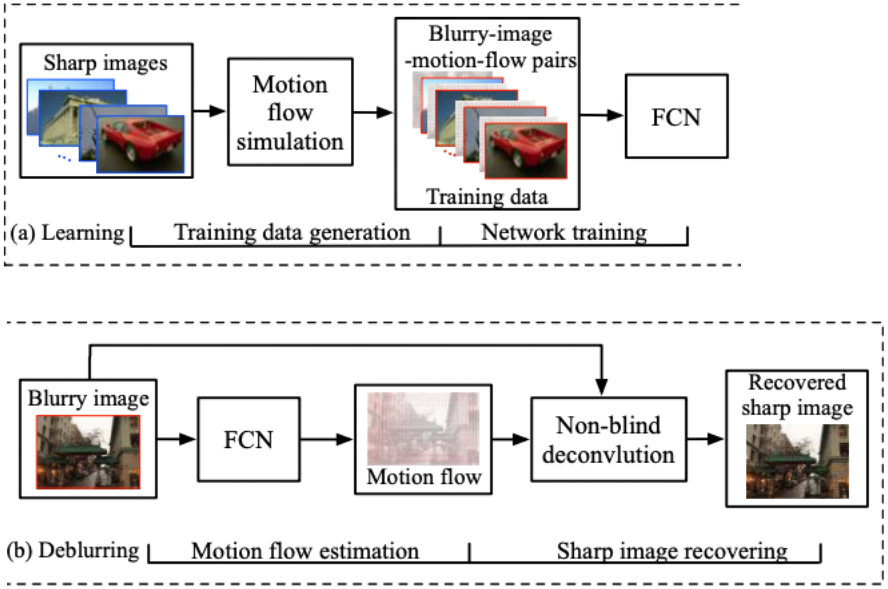
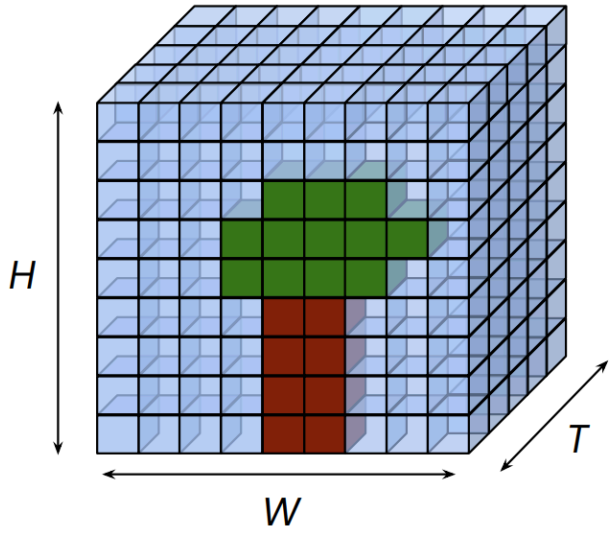


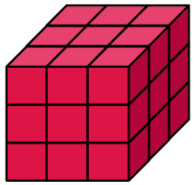
Image from: From Motion Blur to Motion Flow: a Deep Learning Solution for Removing Heterogeneous Motion Blur, Gong et al, 2017

Video input

Video models using 3D convolutions

- Stack frames TxHxWx3
- A volume



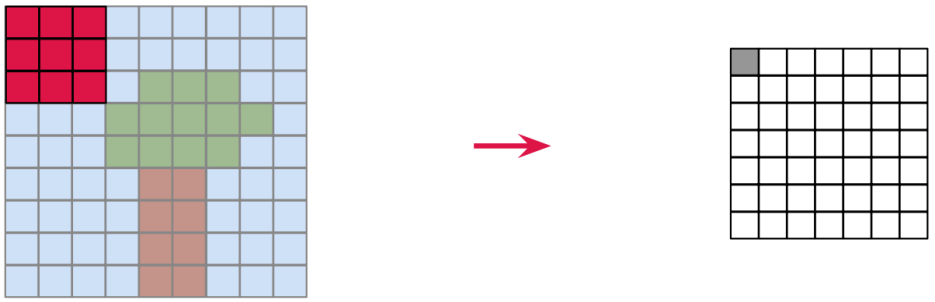


$$y = \sum_{i \in 3 \times 3 \times 3} \mathbf{w}_i \mathbf{x}_i + b$$

Video input

Recap 2D convolution operation

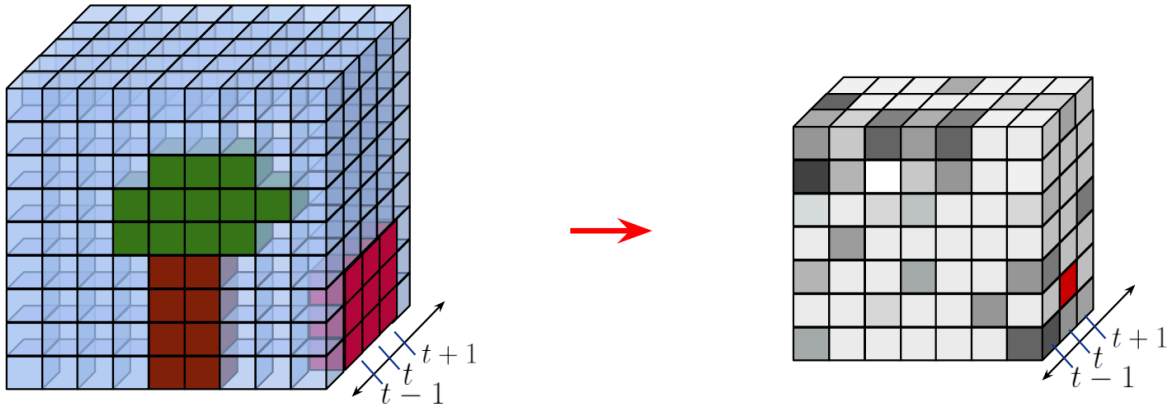
- The kernel slides across spatial dimensions.



Video input

3D convolution operation

- The kernel slides across spatial and time to generate spatio-temporal feature maps.
- Strided, dilated, and padded convolutions also apply in 3D.
- Usually used for video and 3D voxels



Action Recognition

- Input

  RGB video (optional + flow map)

- Targets

  Action label one_hot classes

  e.g. cricket shot

cricket shot

Video from Kinetics dataset, Carreira et al, 2017

Action Recognition

- SlowFast

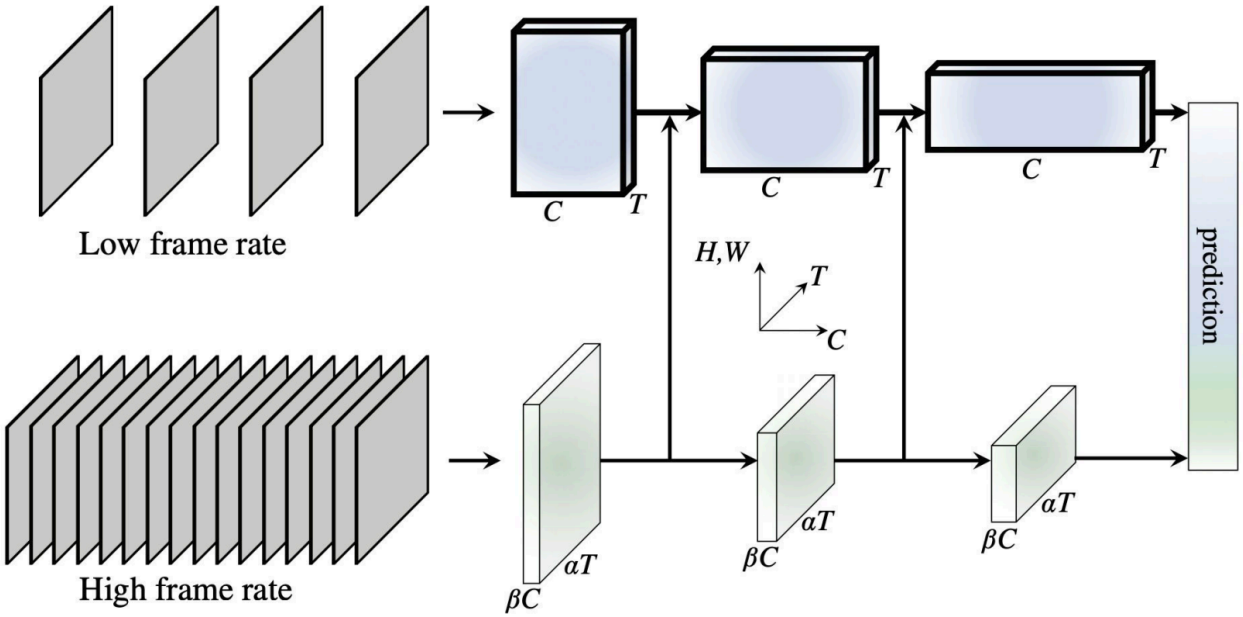
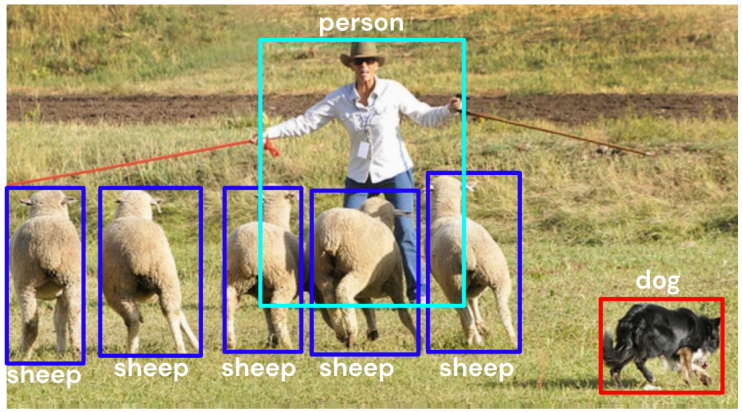
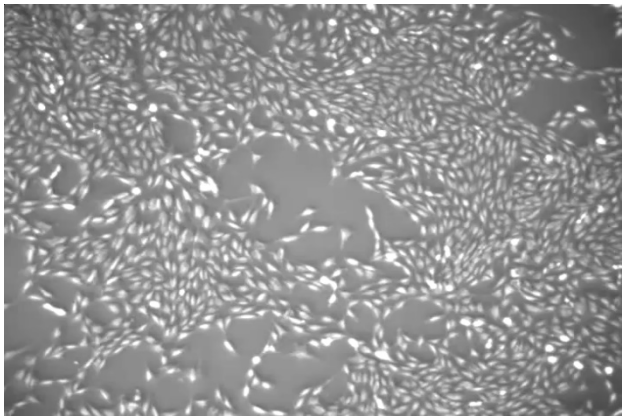


Image from SlowFast Networks for Video Recognition, Feichtenhofer et al, 2019

Beyond Strong Supervision

- Why? – Labelling is tedious.





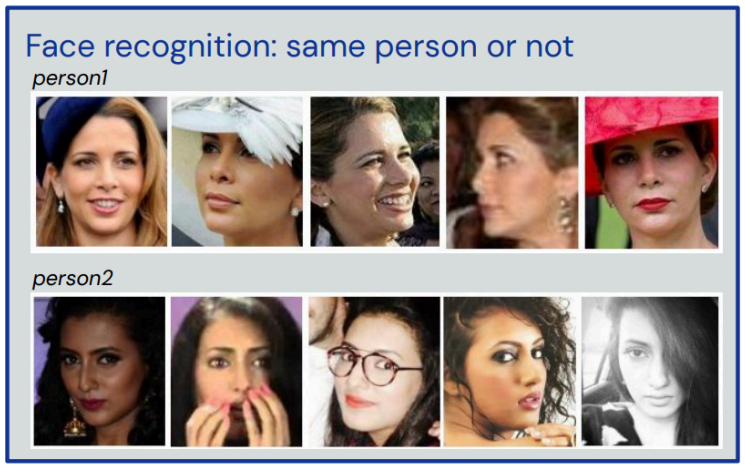
- Self-supervision – Metric learning

Image from COCO dataset and CTC dataset, respectively.

Beyond Strong Supervision

- Recap standard losses (e.g. cross-entropy, mean square error)
  - learn mapping between input(s) and output distribution / value(s)

- Metric learning
  - learn to predict distances between inputs given some similarity measure (e.g. same person or not)



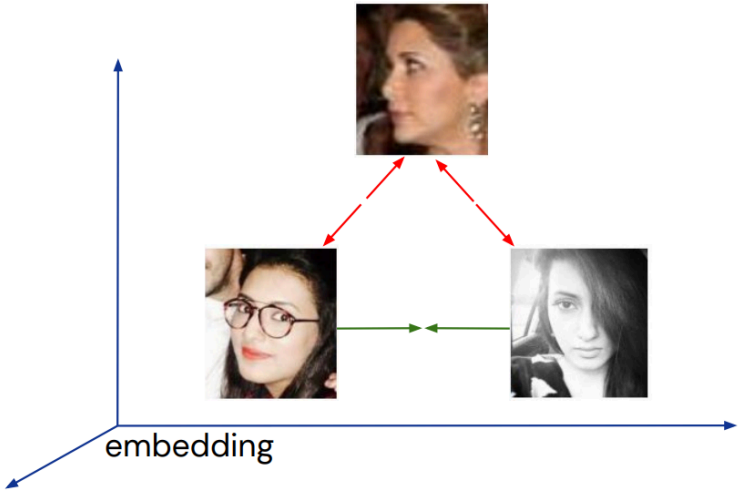


Image from [VGGFace2: A dataset for recognising faces across pose and age, Cao et al, 2018]

Beyond Strong Supervision

Metric Learning

- Contrastive loss (– margin loss)
  - Self-supervised representation, e.g. dimensionality reduction [1]
  - Difficult to choose the margin
- Triplet loss
  - Information retrieval [2]
  - Hard negative mining to select informative triplets
- State-of-the-art representation learning
  - Low-shot face recognition [3]

[1] Dimensionality reduction by learning an invariant mapping, Hadsell et al, 2006
[2] Learning to Learn from Web Data through Deep Semantic Embeddings, Gomez et al, 2018
[3] VGGFace2: A dataset for recognising faces across pose and age, Cao et al, 2018

Beyond Strong Supervision

State-of-the-art representation learning

- Composition of data augmentations
- Learnable non-linear transformation
- Larger mini-batches and longer training

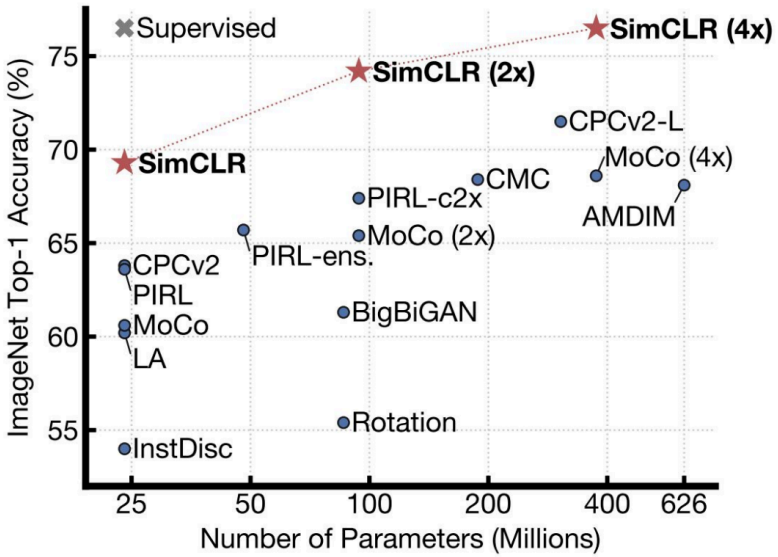


Image from A Simple Framework for Contrastive Learning of Visual Representations, Chen et al, 2020

Beyond Strong Supervision

Generative models: such as Generative adversarial networks (GAN)
Learning to generate images

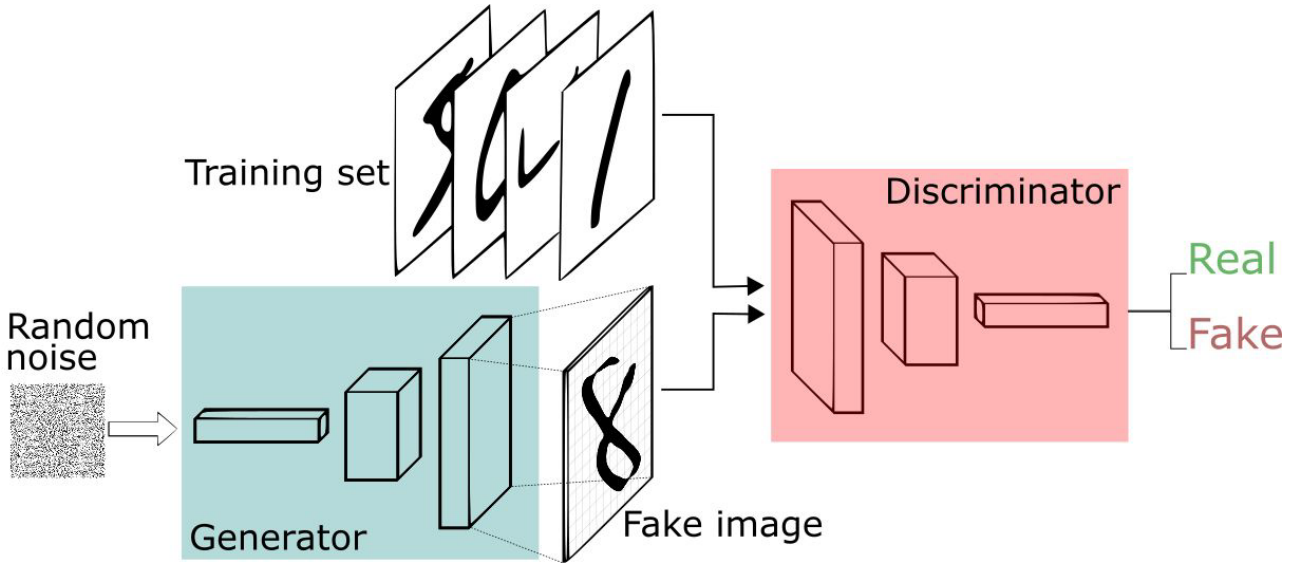


Image from https://wiki.pathmind.com/generative-adversarial-network-gan

Beyond Strong Supervision

Generative models: such as Generative adversarial networks (GAN)
Learning to generate images

Learn more?
Generative adversarial network: An overview of theory and applications, Aggarwal et al. 2021

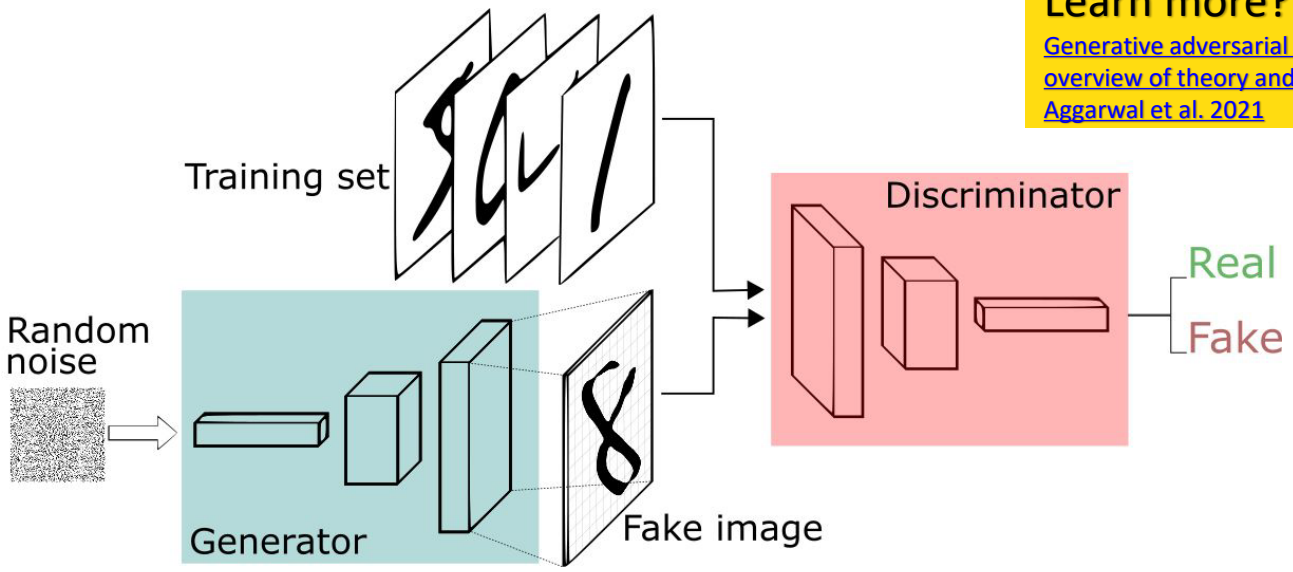


Image from https://wiki.pathmind.com/generative-adversarial-network-gan

References

- Some slides were adopted from the class notes of Stanford course cs231n
- Some slides were adopted from the DeepMind deep learning lecture series 2020