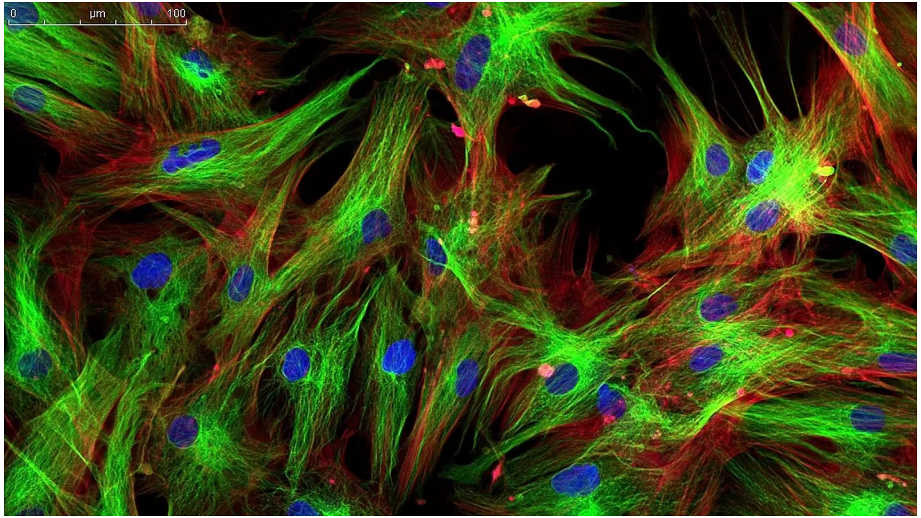
Image Analysis

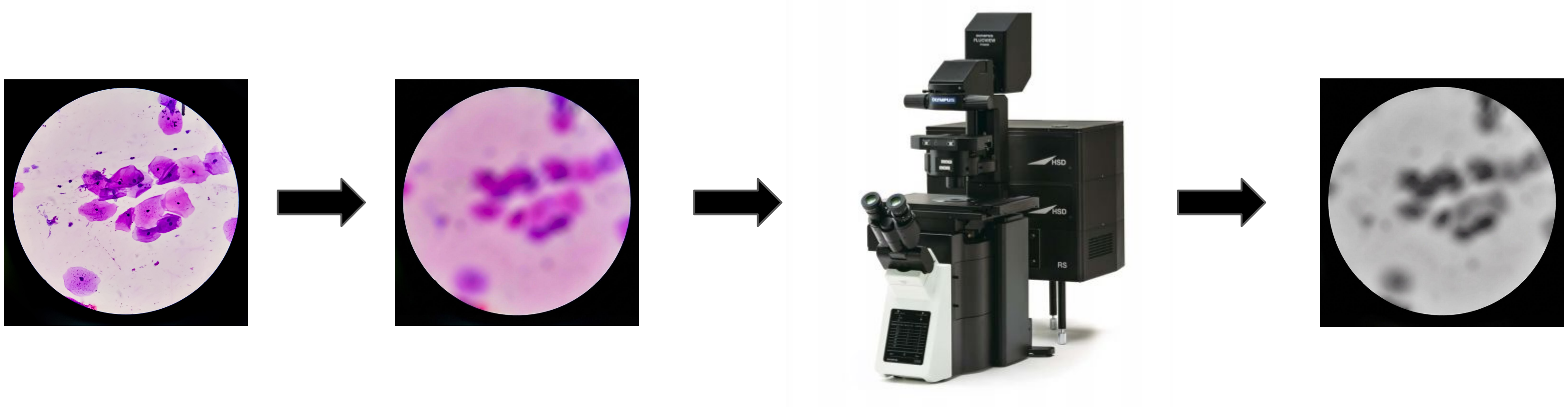
Primer

In RNA-seq we turn reads into counts, then normalize, then do stats. What’s the imaging analogue?



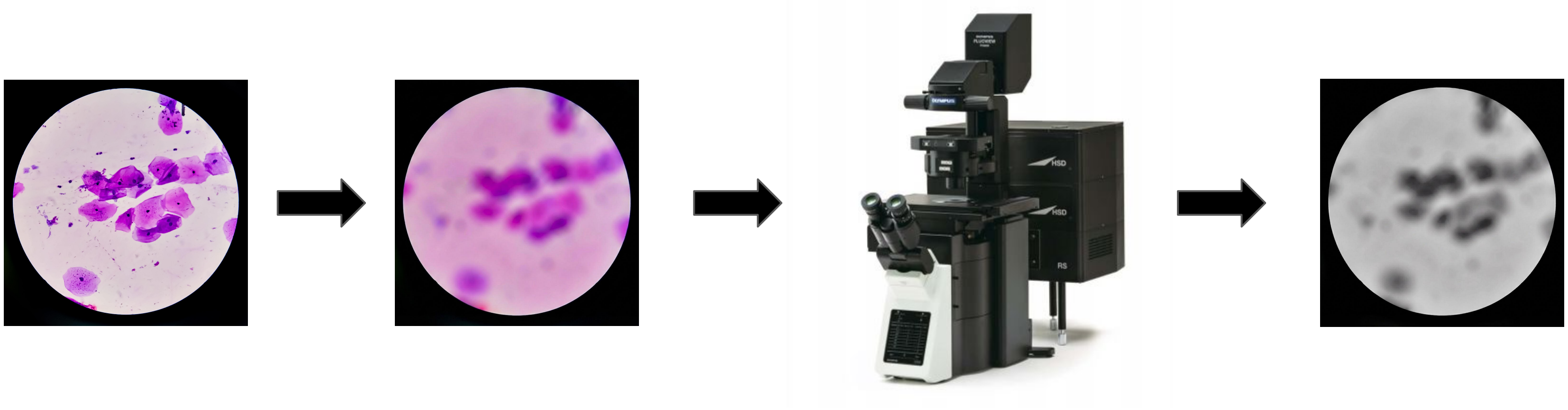
Measurement model: What is an image

An image is a function of intensity I on a grid of pixels (2D) or voxels (3D)



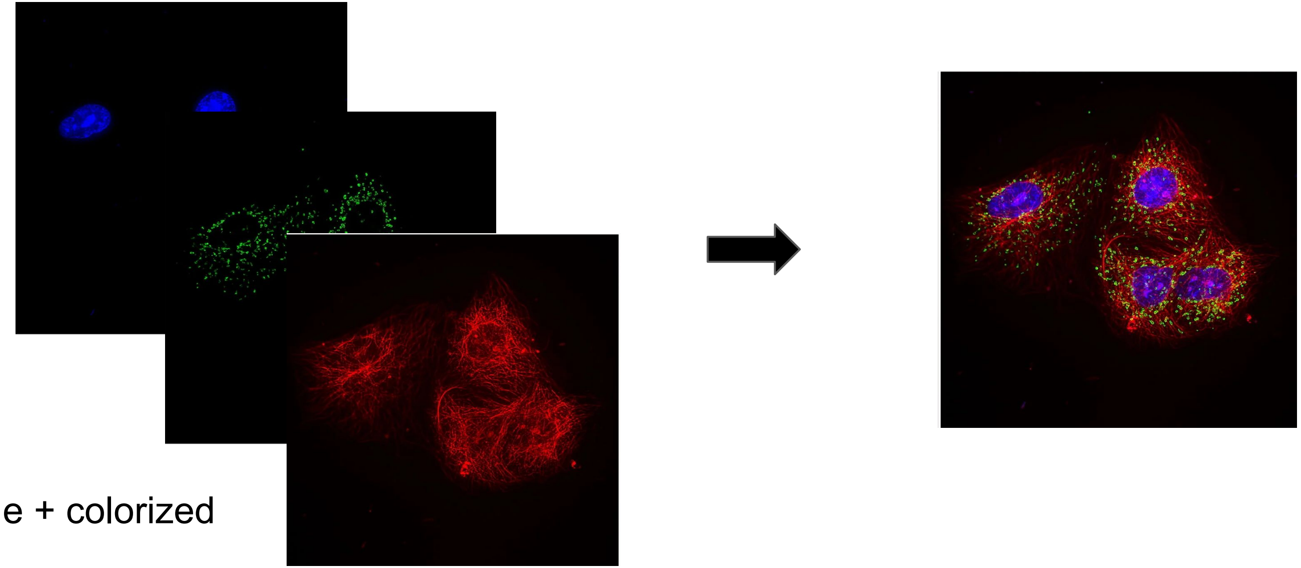
Measurement model: What is an image

$$I = (S * PSF) + B + \varepsilon$$



Color Spaces and Channels

Intensity can have multiple channels: RGB (3 channels), **grayscale** (1 channel)



Grayscale + colorized

Image Enhancement and Preprocessing

- Images are Noisy
- Illumination is rarely uniform

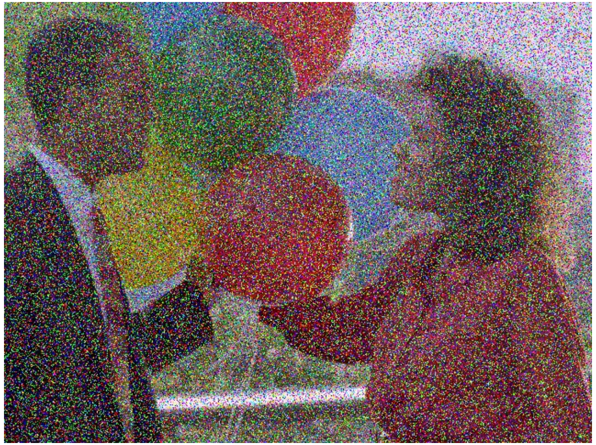
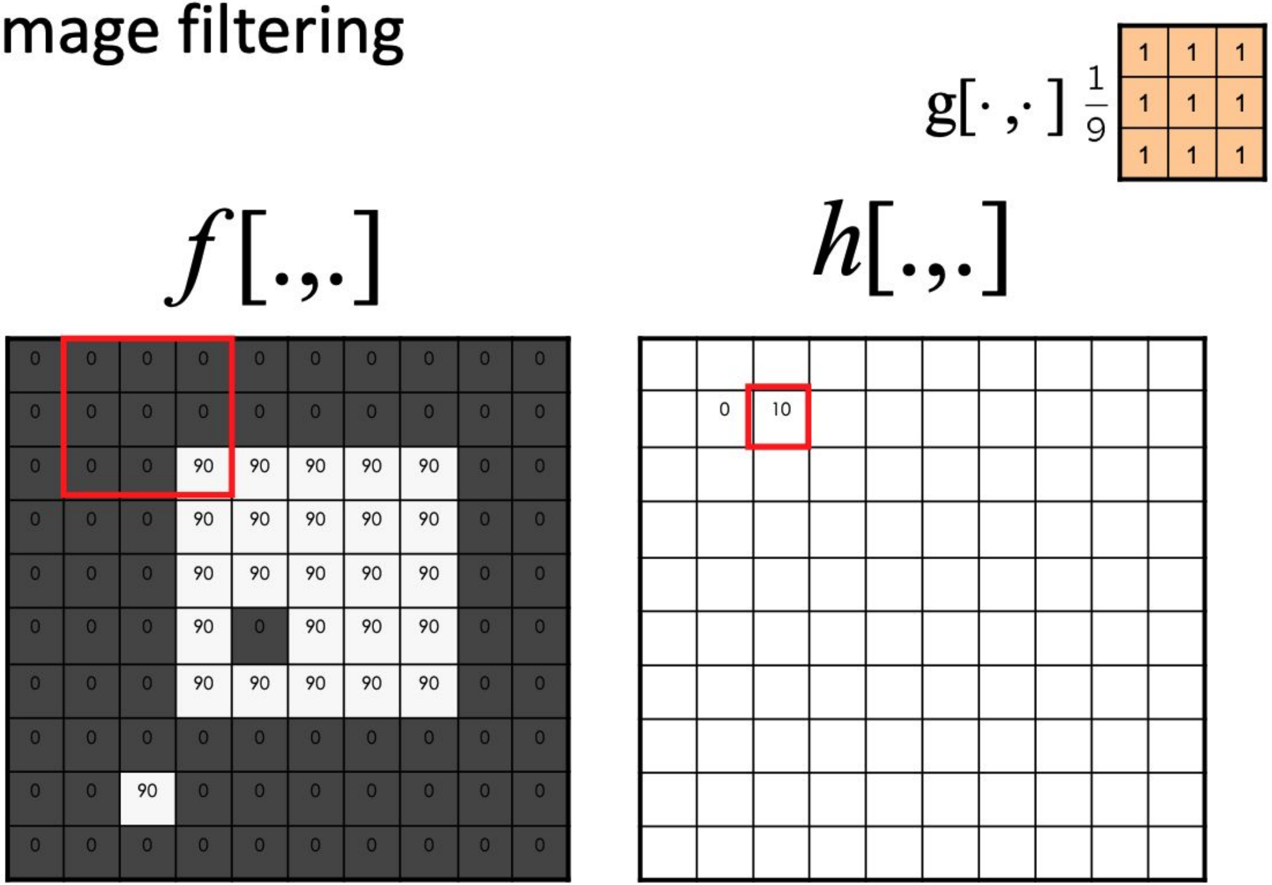


Image filtering



$g[\cdot\,,\cdot\,]\ \frac{1}{9}$

1	1	1
1	1	1
1	1	1

$f[.,.]$

0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0
0	0	0	90	90	90	90	90	0	0
0	0	0	90	90	90	90	90	0	0
0	0	0	90	90	90	90	90	0	0
0	0	0	90	0	90	90	90	0	0
0	0	0	90	90	90	90	90	0	0
0	0	0	0	0	0	0	0	0	0
0	0	90	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0

$h[.,.]$

	0	10							

$$h[m,n] = \sum_{k,l} g[k,l]\, f[m+k,n+l]$$

Credit: S. Seitz

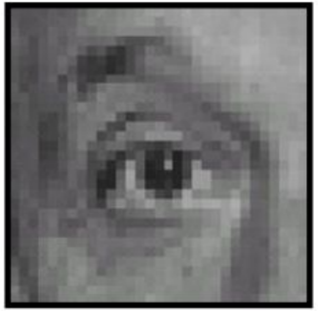
Practice with linear filters

Original

0	0	0
0	1	0
0	0	0

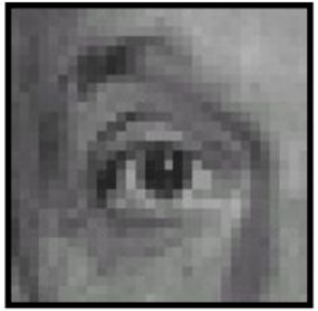
Source: D. Lowe

Practice with linear filters



Original

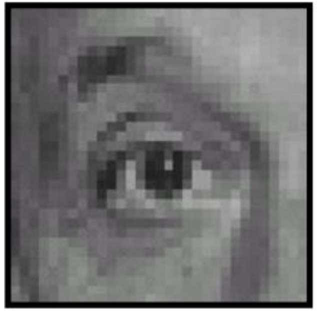
0	0	0
0	1	0
0	0	0



Filtered
(no change)

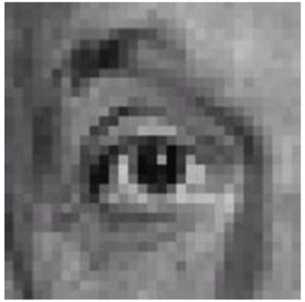
Source: D. Lowe

Practice with linear filters



Original

$$\begin{bmatrix} 0 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 0 \end{bmatrix} - \frac{1}{9}\begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix}$$

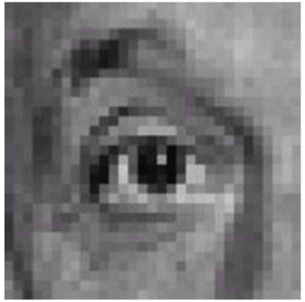


Source: D. Lowe

Practice with linear filters

Original

$$\begin{bmatrix}0&0&0\\0&2&0\\0&0&0\end{bmatrix} - \frac{1}{9}\begin{bmatrix}1&1&1\\1&1&1\\1&1&1\end{bmatrix}$$

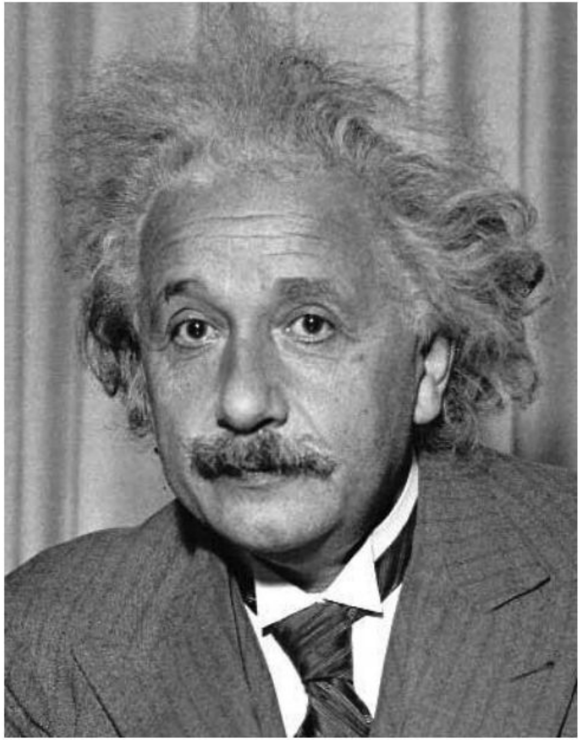


Sharpening filter

- Accentuates differences with local average

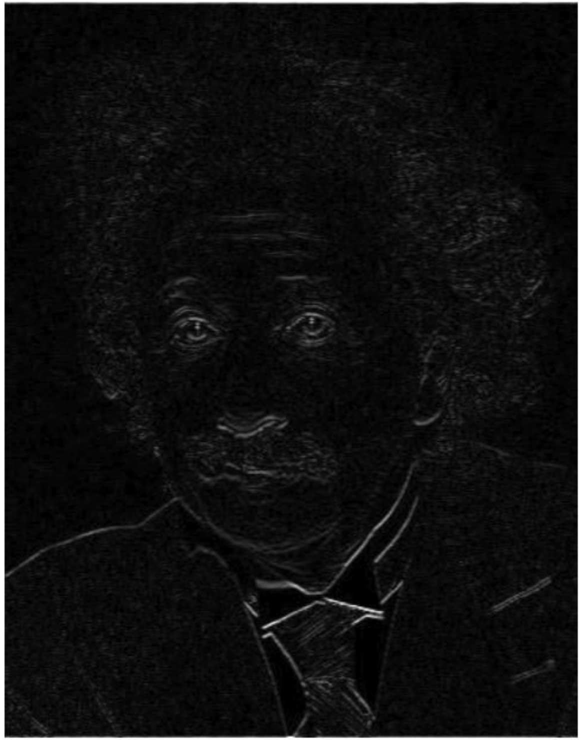
Source: D. Lowe

Other filters



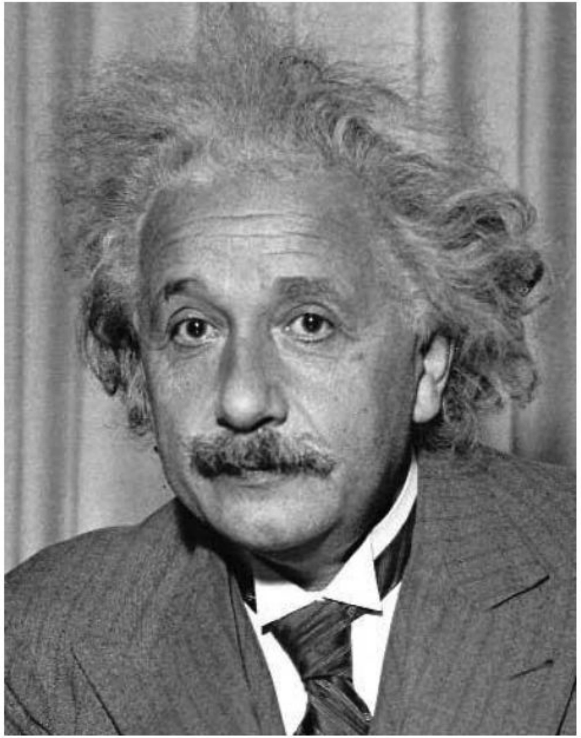
1	2	1
0	0	0
-1	-2	-1

Sobel



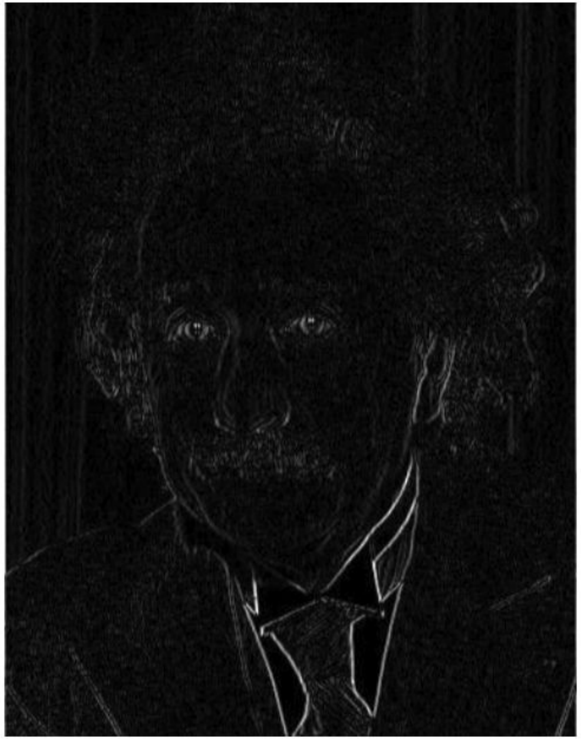
Horizontal Edge
(absolute value)

Other filters



1	0	-1
2	0	-2
1	0	-1

Sobel



Vertical Edge
(absolute value)

Smoothing with Gaussian filter

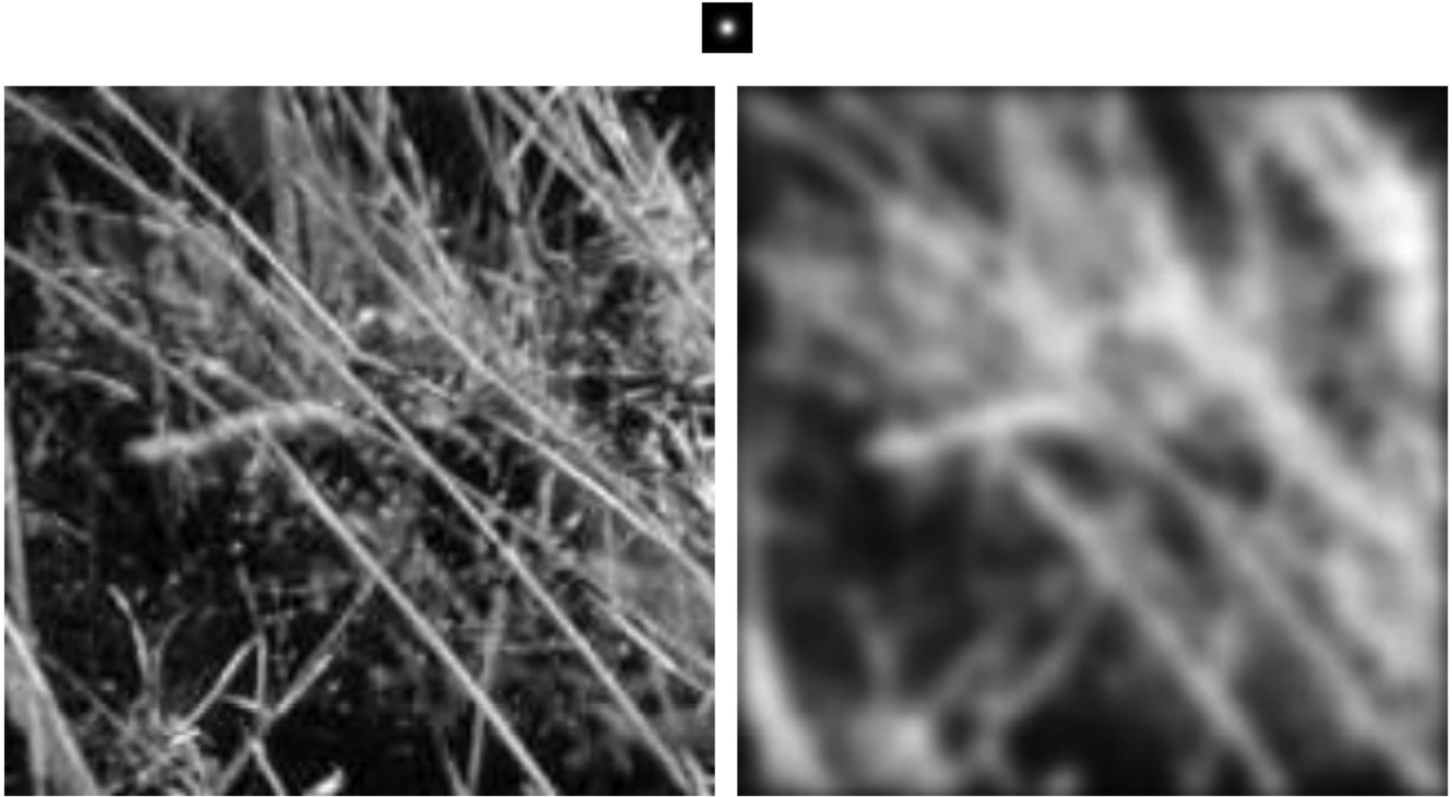
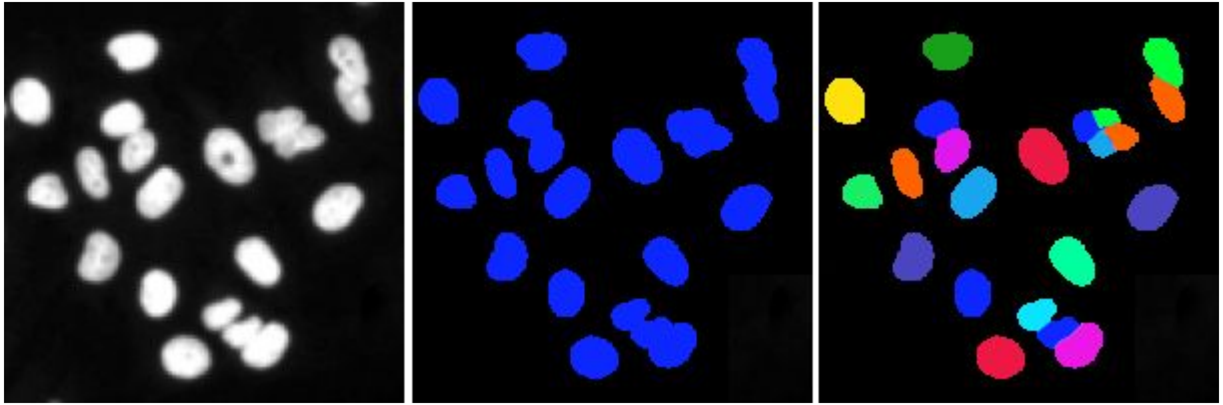


Image Segmentation

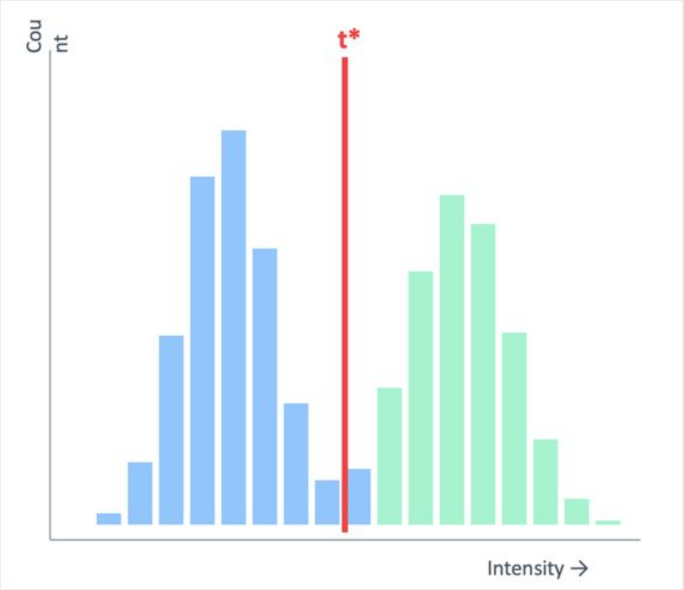
Partitioning a digital image into multiple segments



Global Thresholding

One threshold value for the entire image

- Set a single intensity value to separate objects (foreground) from background.
- Works well when illumination is consistent and foreground/background intensities are separable.
- Otsu’s Method: automatically finds an optimal threshold by minimizing intra-class variance (or equivalently maximizing between-class variance).

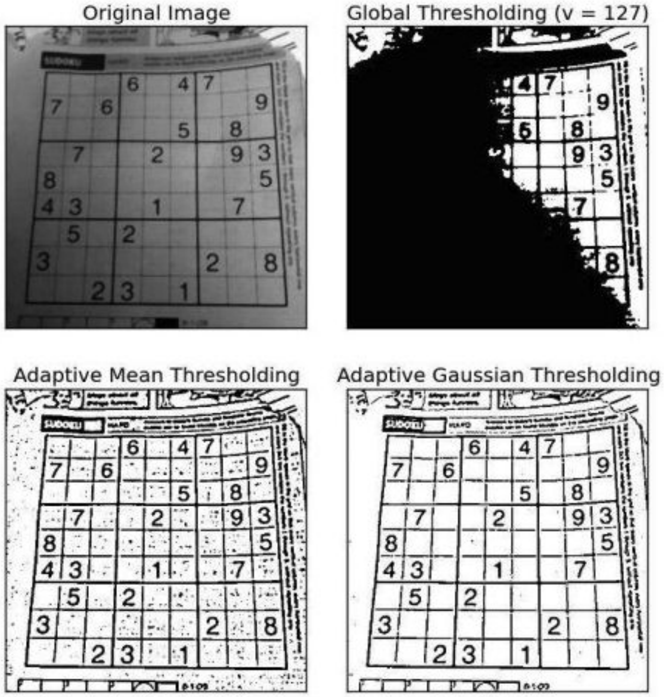


Histogram + single threshold

Adaptive/Local Thresholding

The threshold is mean of neighborhood

The threshold is the gaussian weighted sum of the neighborhood values


Original Image
Global Thresholding (v = 127)
Adaptive Mean Thresholding
Adaptive Gaussian Thresholding

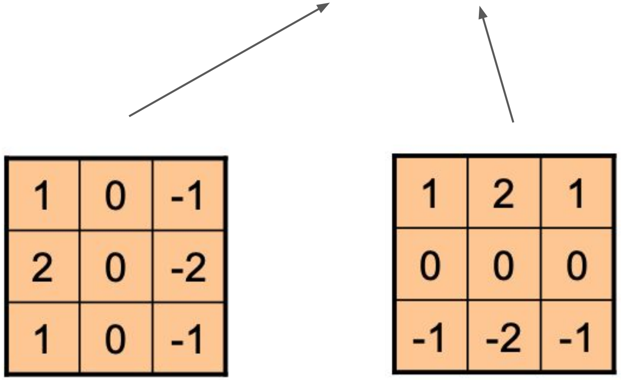
image

Edge-based Segmentation

- Find edges first, then find objects.
  - Idea: Object boundaries involve sharp changes in color or intensity.

$$\nabla I = (I_x, I_y)$$

How do we get I_x, I_y ?



1	0	-1
2	0	-2
1	0	-1

Sobel

1	2	1
0	0	0
-1	-2	-1

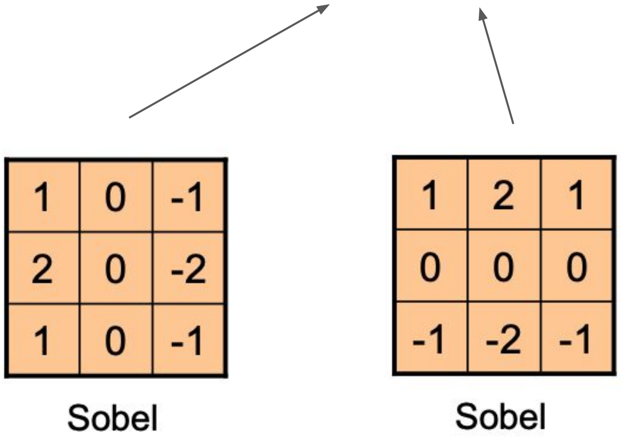
Sobel

Edge-based Segmentation

- Find edges first, then find objects.
  - Idea: Object boundaries involve sharp changes in color or intensity.

$$\nabla I = (I_x, I_y)$$

How do we get I_x, I_y ?



1	0	-1
2	0	-2
1	0	-1

Sobel

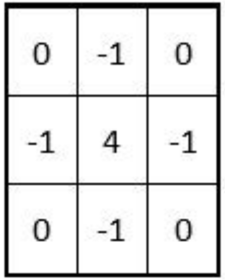
1	2	1
0	0	0
-1	-2	-1

Sobel

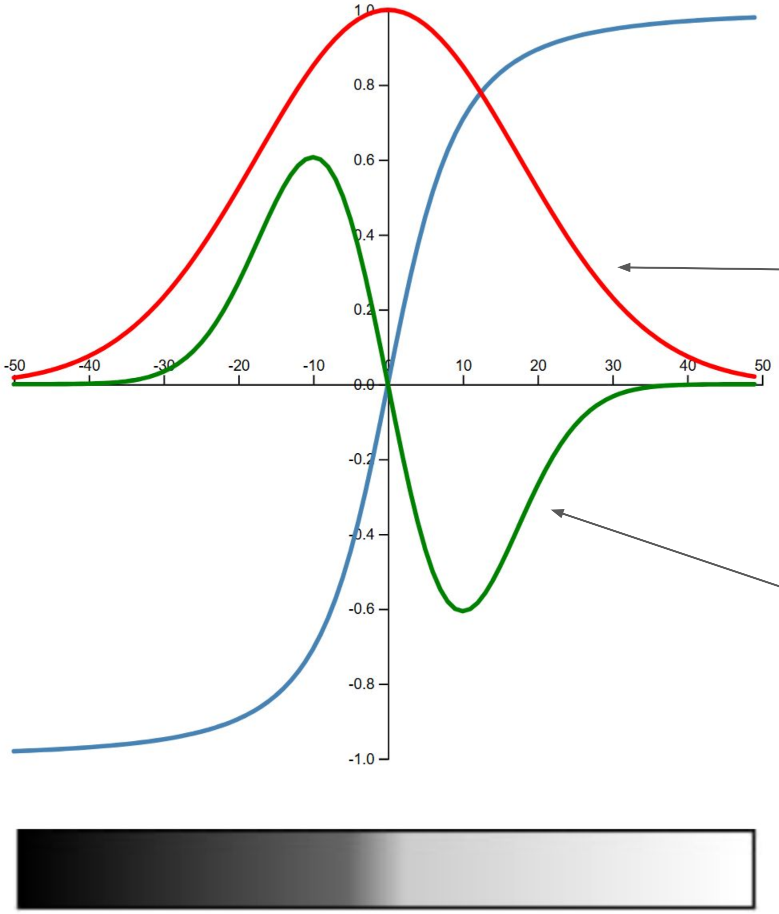
Edge-based Segmentation

Can also use second derivatives:

$$\nabla^2 I = \frac{\partial^2}{\partial x^2} I + \frac{\partial^2}{\partial y^2} I$$



0	-1	0
-1	4	-1
0	-1	0



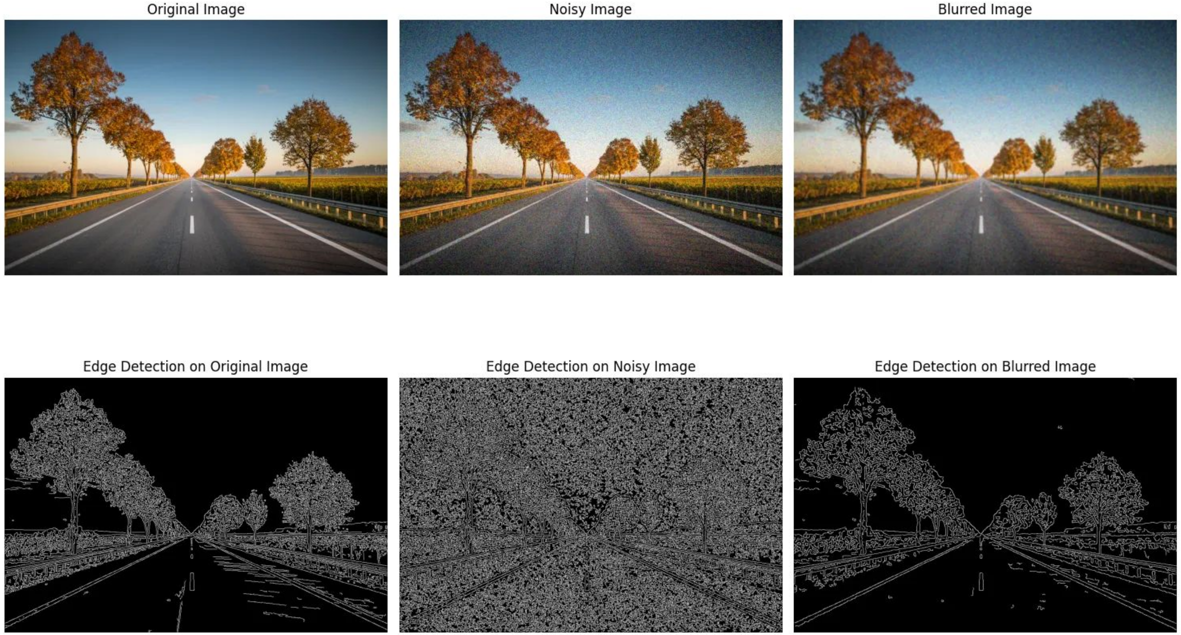
Intensity

Gradient (derivative)

Laplacian (second derivative)

Handling Noise

- Derivatives are very sensitive to noise -> apply a blur first



Original Image

Noisy Image

Blurred Image

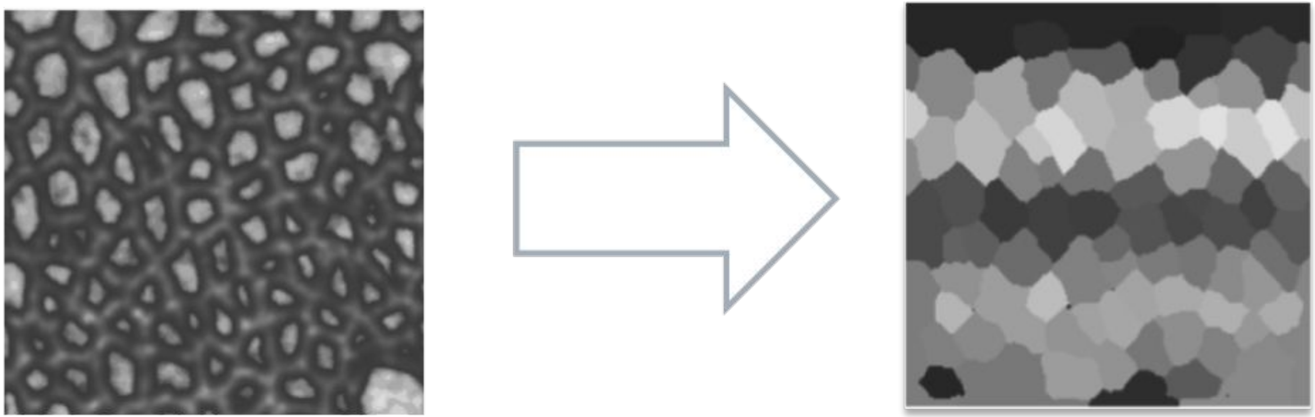
Edge Detection on Original Image

Edge Detection on Noisy Image

Edge Detection on Blurred Image

Region-based Segmentation

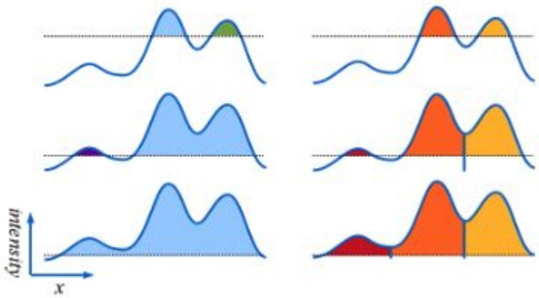
- Group similar pixels into regions starting from seeds (think paint-bucket tool).



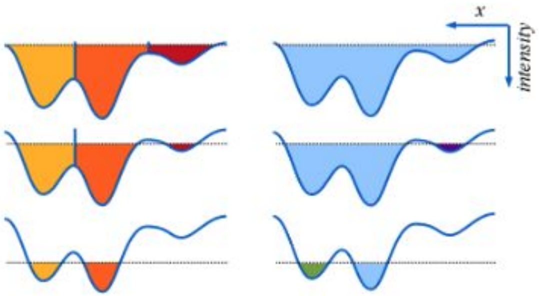
2D Image

Segmentation

Watershed Algorithm

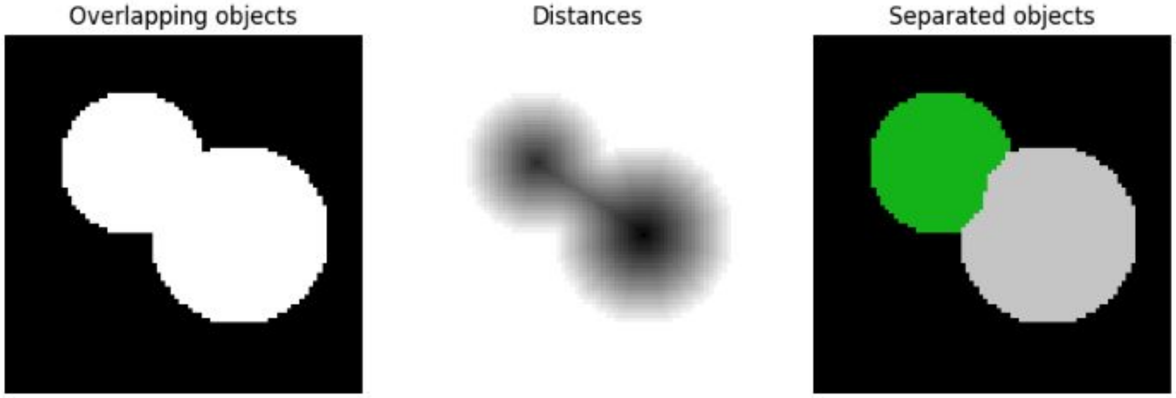


Schematic overview of the flooding algorithm for the watershed method



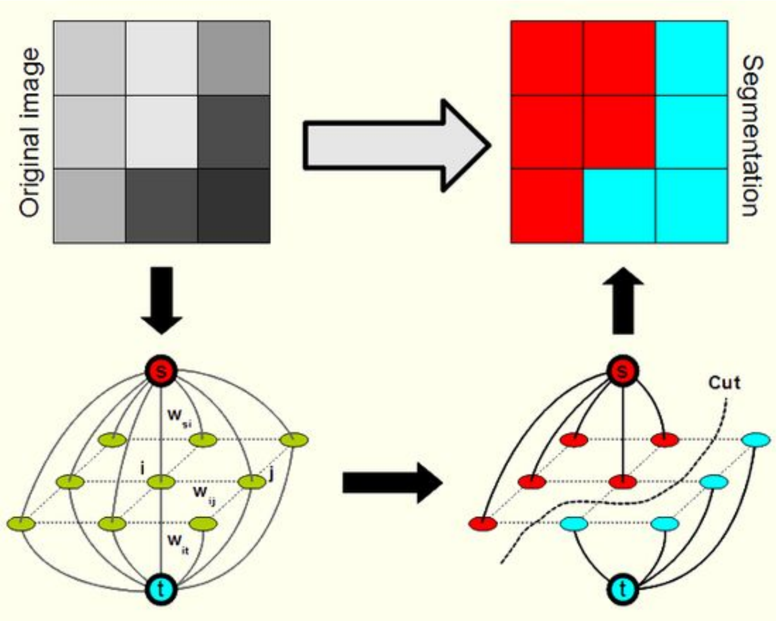
Inverted the schematic overview to emphasize the flooding.

Starting marker is distance from background



Graph-based segmentation

- Idea: treat pixels (or super pixels) as nodes. Define edges in local patches or across entire image.



Evaluating Segmentation Accuracy

Compare predictions to a “ground truth” mask

How do we know the algorithm worked?

Ground truth: human-annotated masks.

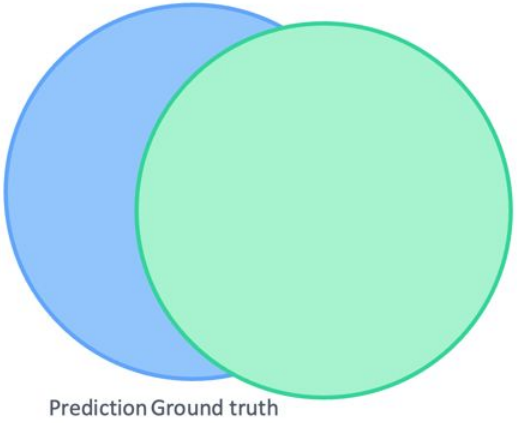
Metrics:

IoU (Intersection over Union) = overlap / union

Dice coefficient = 2 × overlap / (|A| + |B|)

$$\text{IoU} = |A \cap B| \,/\, |A \cup B|$$

$$\text{Dice} = 2|A \cap B| \,/\, (|A| + |B|)$$



Overlap matters

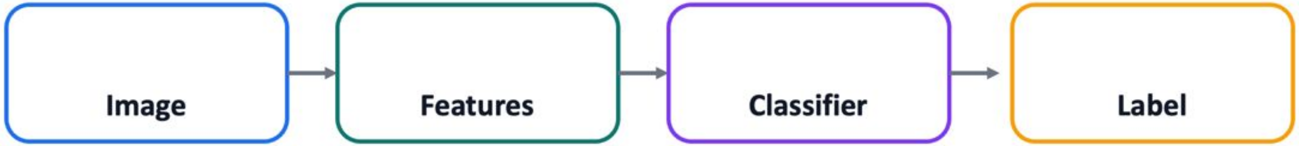
Higher is better (0 → 1)

Traditional Machine Learning in Imaging

- Requires manual feature engineering (shape, texture, intensity).
- Feed features into classifiers (Random Forests, SVMs) to categorize cells (e.g., “Healthy” vs. “Apoptotic”).
- Limitation: Performance plateaus and depends heavily on human intuition for feature selection.

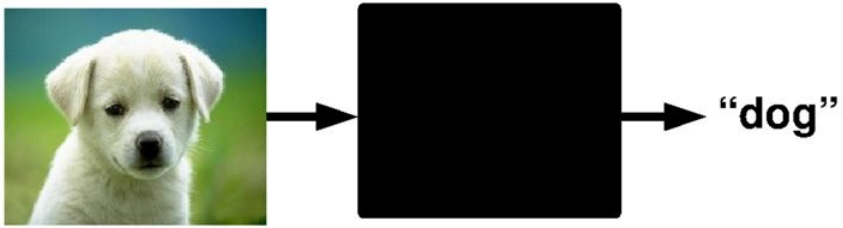
Key idea

Humans design features
(shape, texture, intensity)
→ model learns decision boundary.



Supervised Learning: Examples

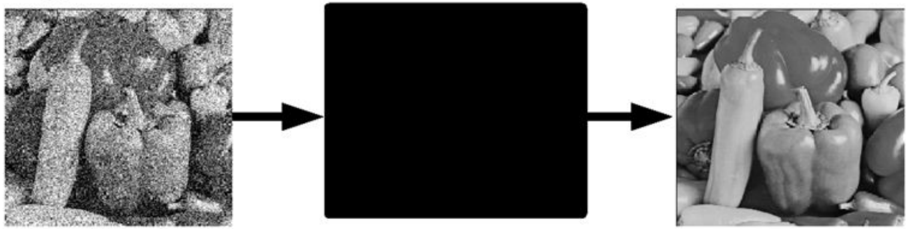
Classification



“dog”

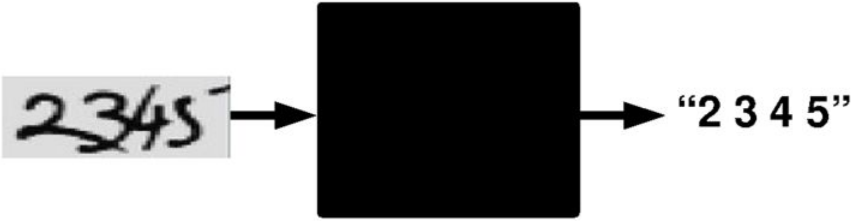
classification

Denoising



regression

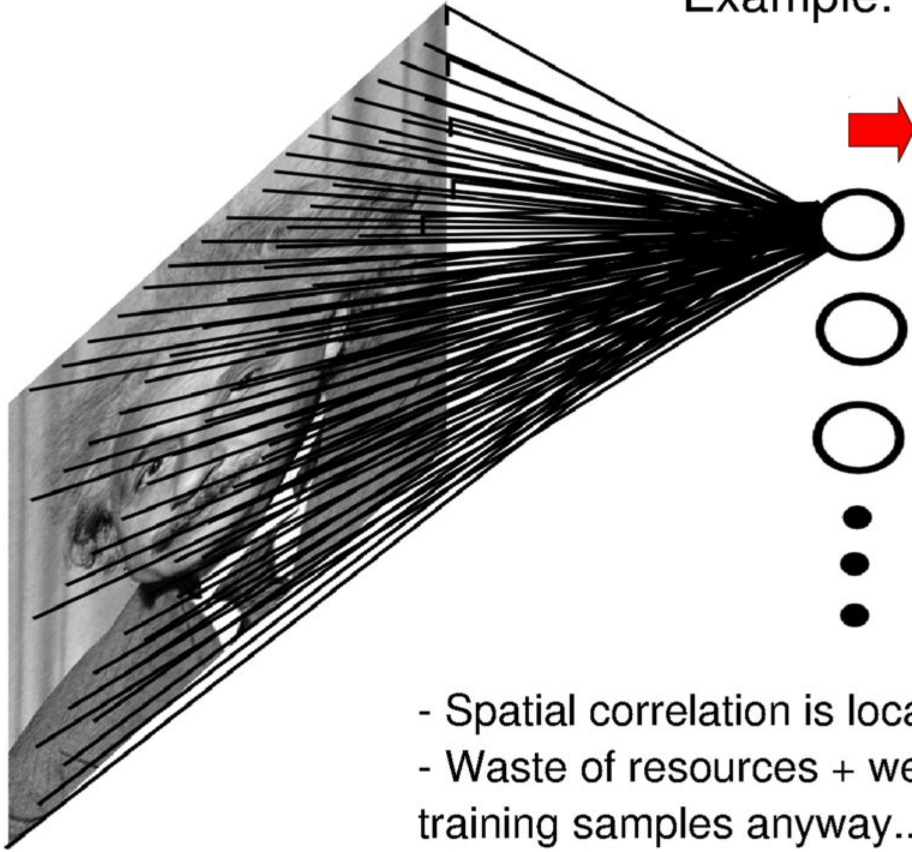
OCR



“2 3 4 5”

structured prediction

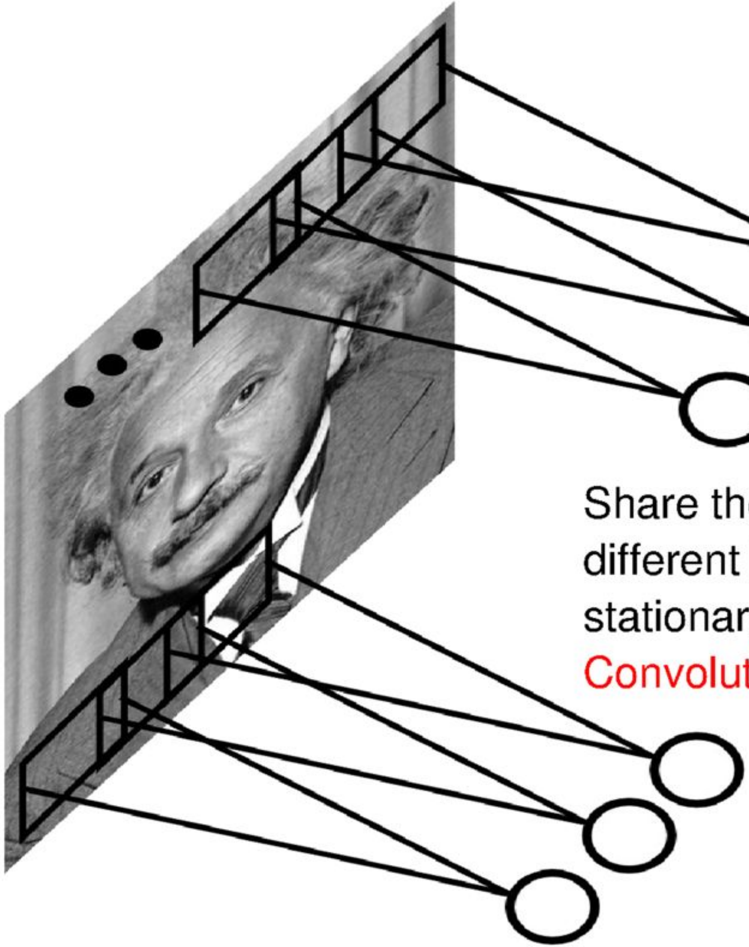
Fully Connected Layer



Example: 200x200 image
40K hidden units
→ **~2B parameters**!!!

- Spatial correlation is local
- Waste of resources + we have not enough training samples anyway..

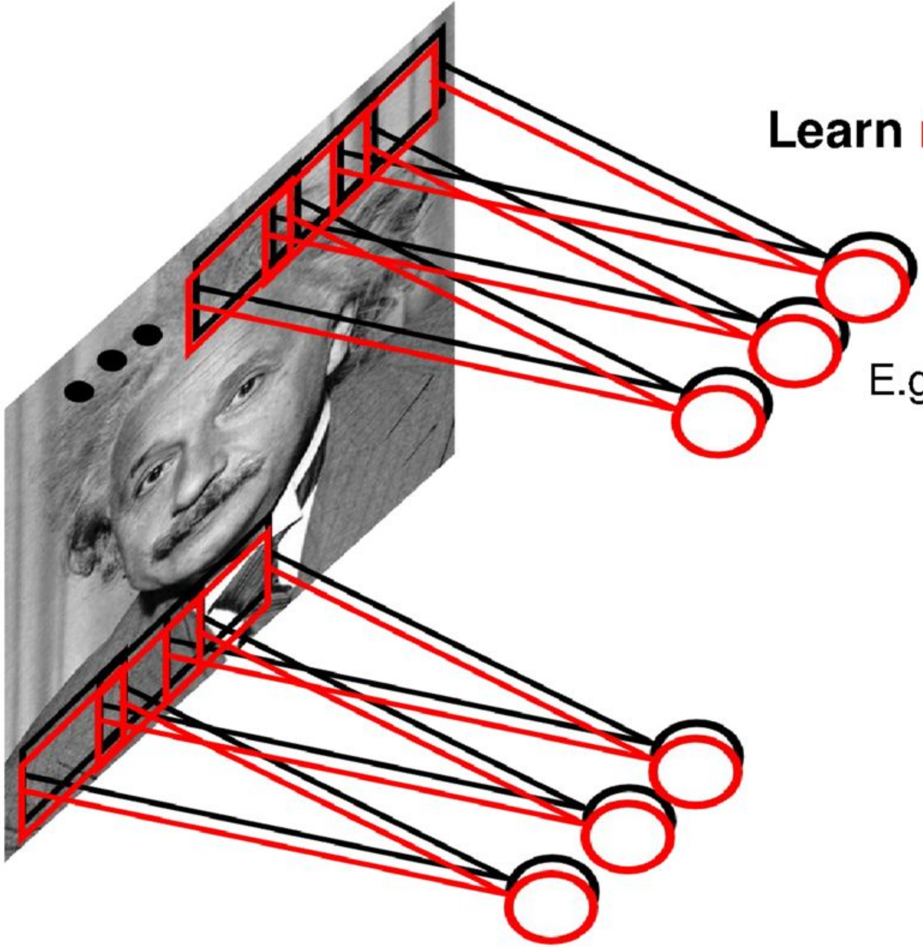
Convolutional Layer



Share the same parameters across different locations (assuming input is stationary):

Convolutions with learned kernels

Convolutional Layer



Learn multiple filters.

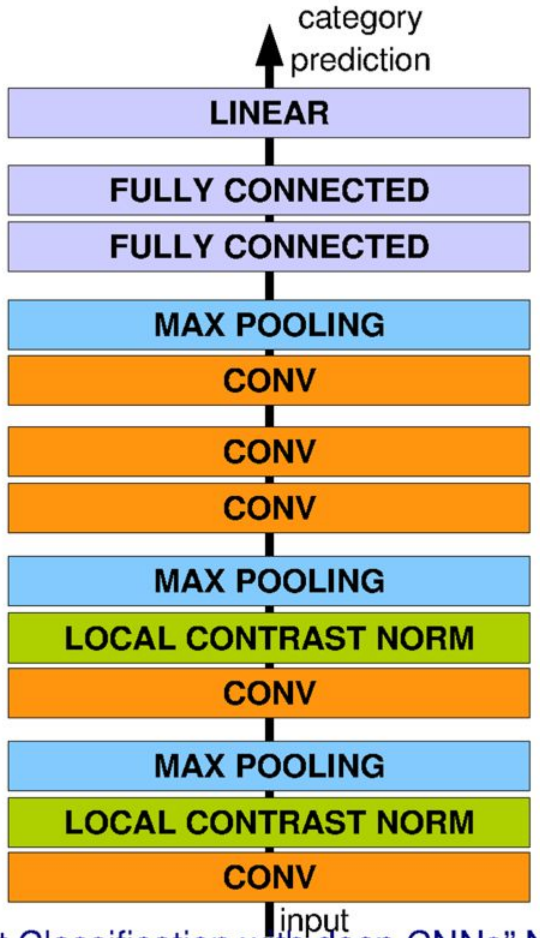
E.g.: 200x200 image
100 Filters
Filter size: 10x10
10K parameters

Ranzato

ImageNet

Examples of hammer:

Architecture for Classification



Krizhevsky et al. “ImageNet Classification with deep CNNs” NIPS 2012

Deep Residual Learning for Image Recognition

Kaiming He, Xiangyu Zhang, Shaoqing Ren, Jian Sun

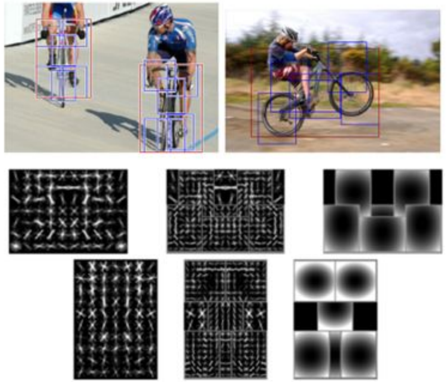
Cited by 290k papers as of 10/14/2025

Nearing the most cited paper *ever*

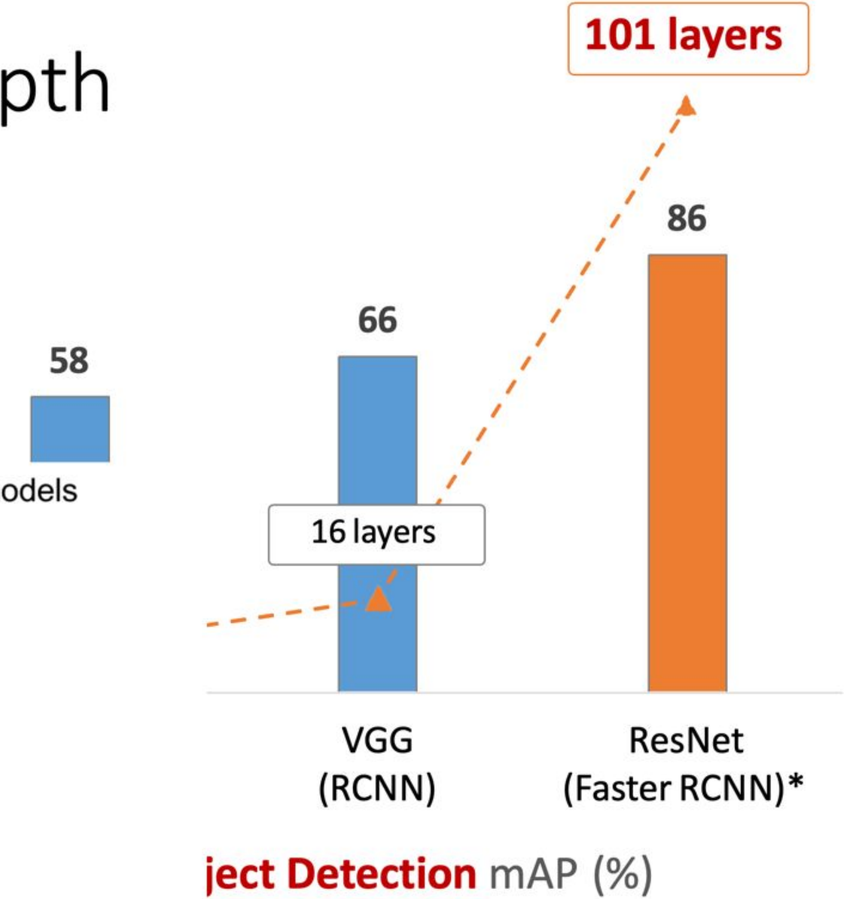
Revolution of Depth

Engines of
visual recognition

Discriminatively trained part-based models



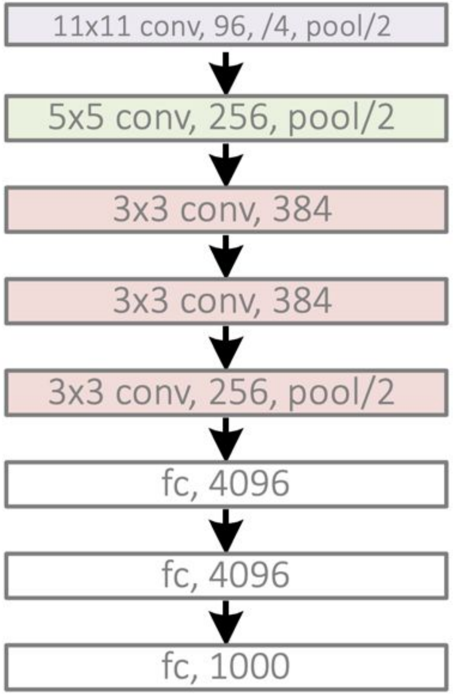
P. Felzenszwalb, R. Girshick, D. McAllester, D. Ramanan, "Object Detection with Discriminatively Trained Part-Based Models," PAMI 2009



*w/ other improvements & more data

Revolution of Depth

AlexNet, 8 layers
(ILSVRC 2012)



11x11 conv, 96, /4, pool/2

5x5 conv, 256, pool/2

3x3 conv, 384

3x3 conv, 384

3x3 conv, 256, pool/2

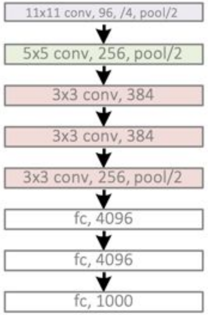
fc, 4096

fc, 4096

fc, 1000

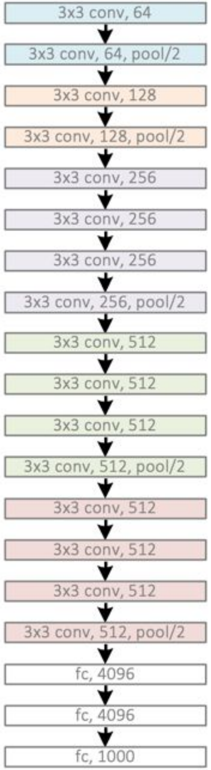
Revolution of Depth

AlexNet, 8 layers
(ILSVRC 2012)



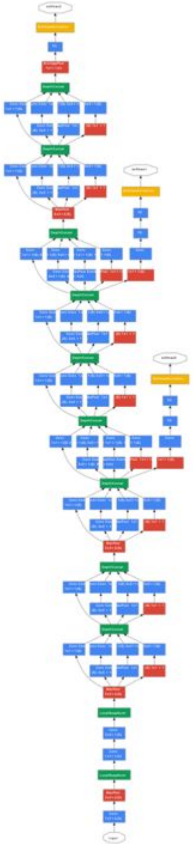
11x11 conv, 96, /4, pool/2
5x5 conv, 256, pool/2
3x3 conv, 384
3x3 conv, 384
3x3 conv, 256, pool/2
fc, 4096
fc, 4096
fc, 1000

VGG, 19 layers
(ILSVRC 2014)



3x3 conv, 64
3x3 conv, 64, pool/2
3x3 conv, 128
3x3 conv, 128, pool/2
3x3 conv, 256
3x3 conv, 256
3x3 conv, 256
3x3 conv, 256, pool/2
3x3 conv, 512
3x3 conv, 512
3x3 conv, 512
3x3 conv, 512, pool/2
3x3 conv, 512
3x3 conv, 512
3x3 conv, 512
3x3 conv, 512, pool/2
fc, 4096
fc, 4096
fc, 1000

GoogleNet, 22 layers
(ILSVRC 2014)

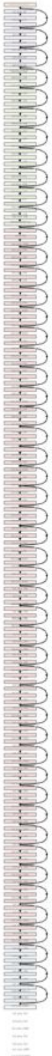


Revolution of Depth

AlexNet, 8 layers
(ILSVRC 2012)

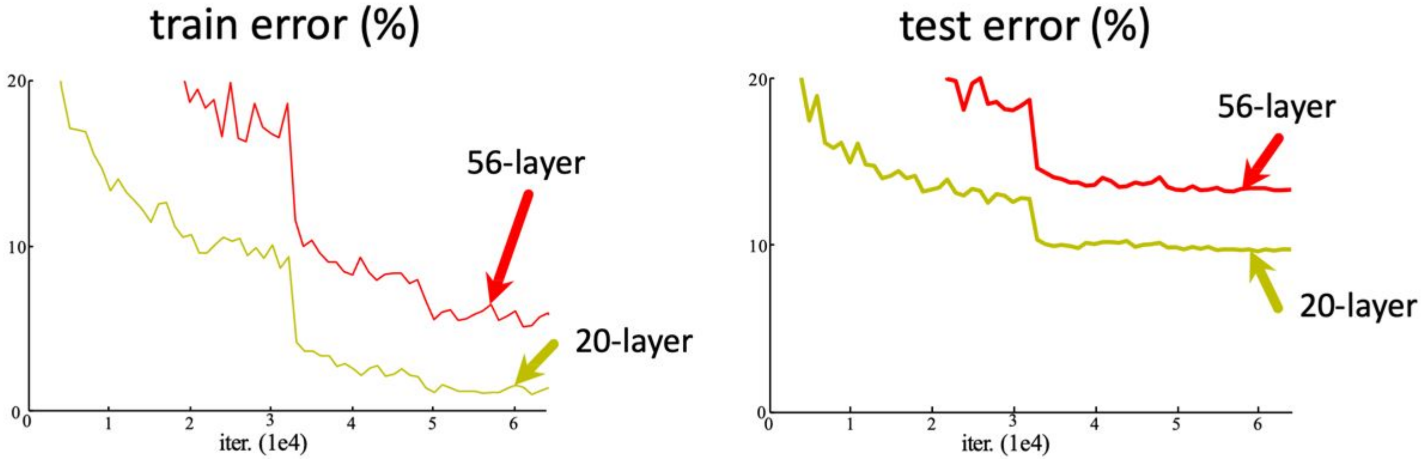
VGG, 19 layers
(ILSVRC 2014)

ResNet, 152 layers
(ILSVRC 2015)



Simply stacking layers?

CIFAR-10

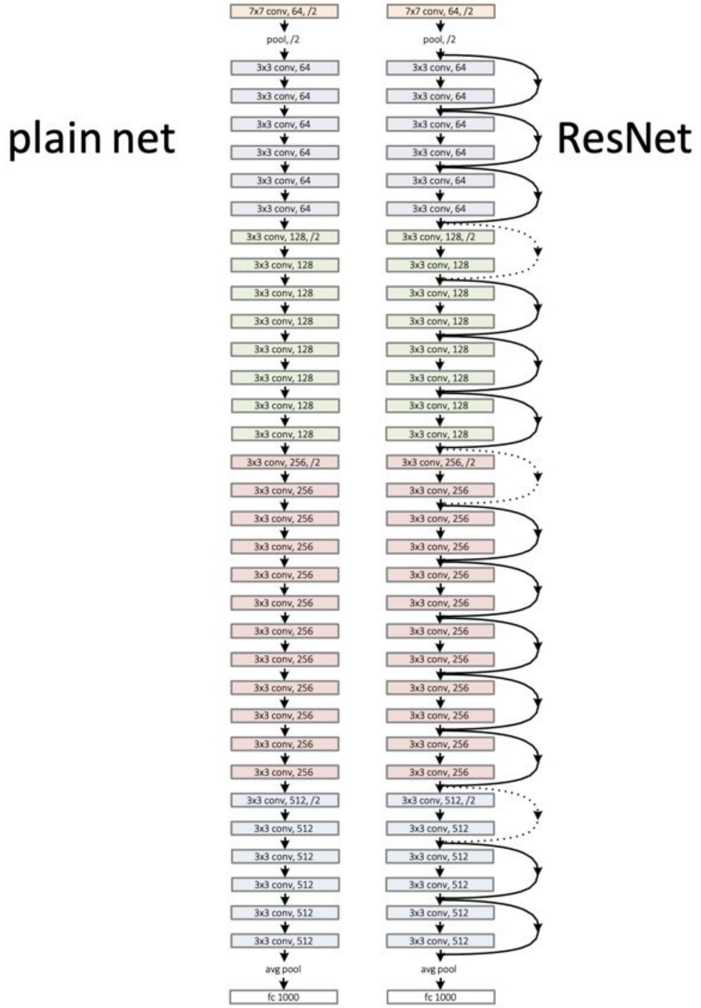


- *Plain* nets: stacking 3x3 conv layers…
- 56-layer net has **higher training error** and test error than 20-layer net

Network “Design”

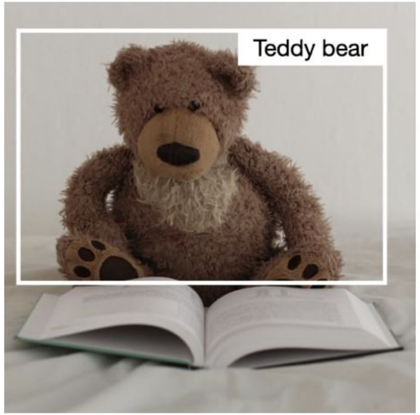
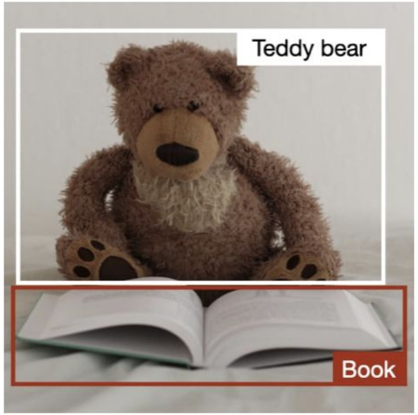
- Keep it simple

- Our basic design (VGG-style)
  - all 3x3 conv (almost)
  - spatial size /2 => # filters x2
  - Simple design; just deep!



Object detection

□ **Types of models** — There are 3 main types of object recognition algorithms, for which the nature of what is predicted is different. They are described in the table below:

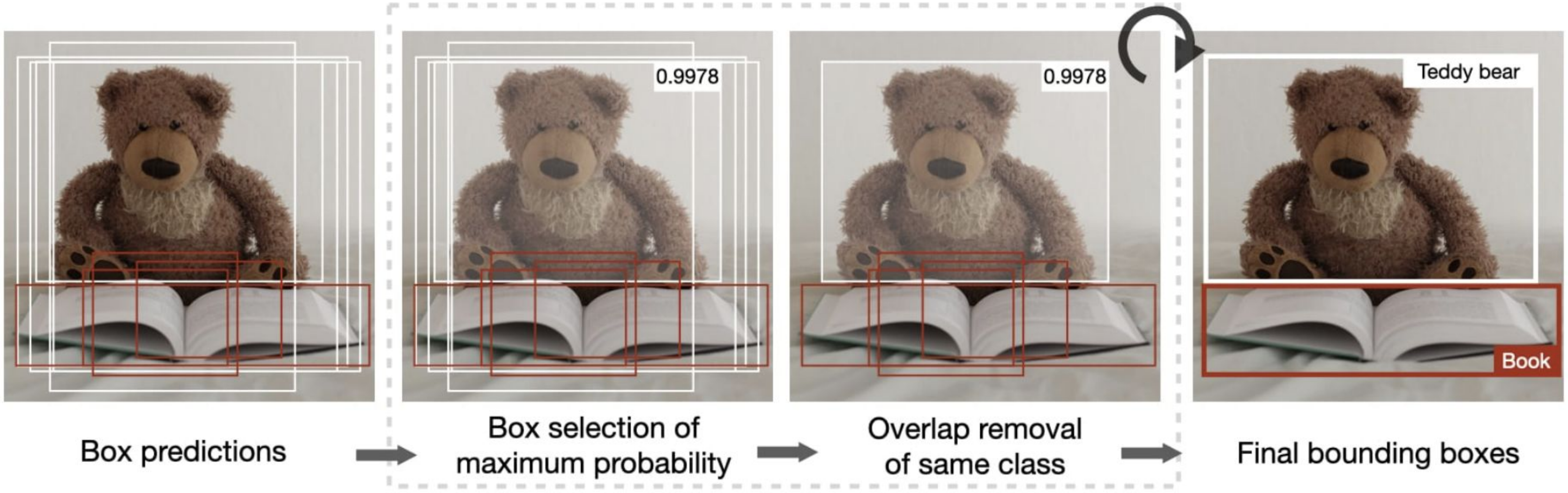
Image classification	Classification w. localization	Detection
		
• Classifies a picture • Predicts probability of object	• Detects an object in a picture • Predicts probability of object and where it is located	• Detects up to several objects in a picture • Predicts probabilities of objects and where they are located
Traditional CNN	Simplified YOLO, R-CNN	YOLO, R-CNN

Non-max Suppression (NMS)

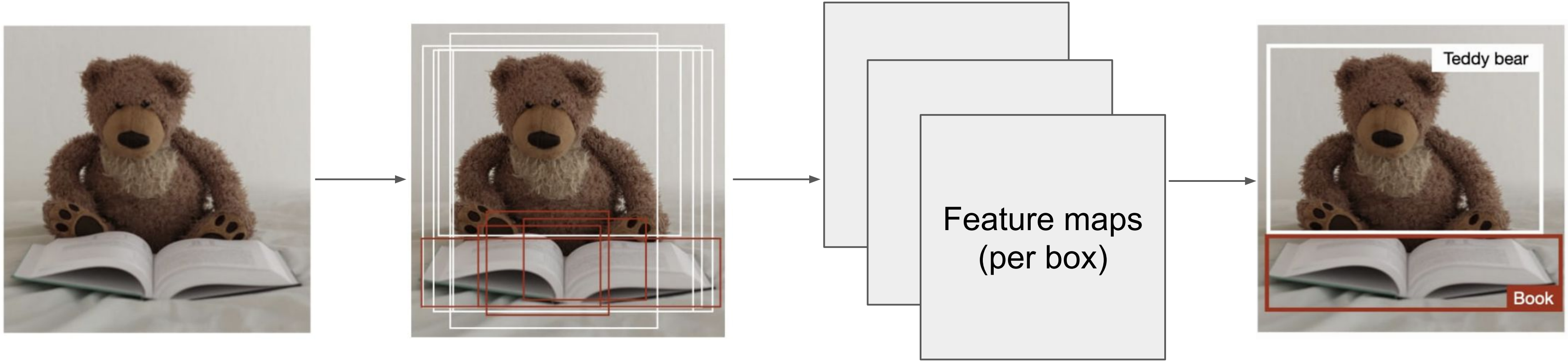
For a given class,

- Step 1: Pick the box with the largest prediction probability.

- Step 2: Discard any box having an $\mathrm{IoU} \geqslant 0.5$ with the previous box.


0.9978
0.9978
Teddy bear
Book
Box predictions
Box selection of maximum probability
Overlap removal of same class
Final bounding boxes

RCNN



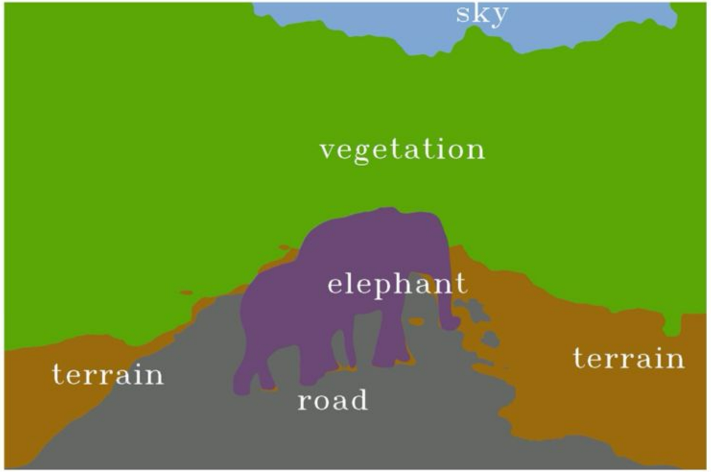
original image

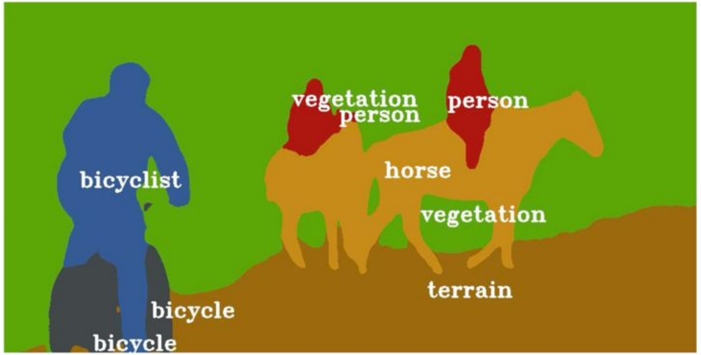
proposal bounding boxes

➡ crop and apply CNN

➡ classify, refine, and prune (NMS) boxes

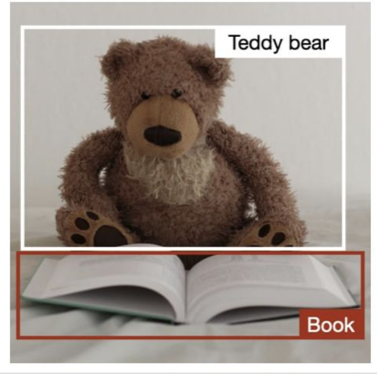
Semantic Segmentation



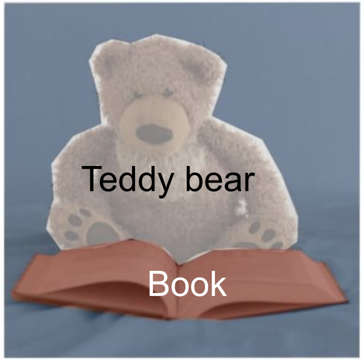


UNet - motivation

- Segmentation is harder than classification/detection (operates at different resolutions):
  - High-level context (what object is this pixel part of?)
  - Fine spatial detail (exact boundaries, small structures)



VS



UNet

- Idea: compute feature maps at multiple resolutions (down-sampling). Combine them iteratively (up-sampling).
    - High resolution (local features) + low resolution (context)

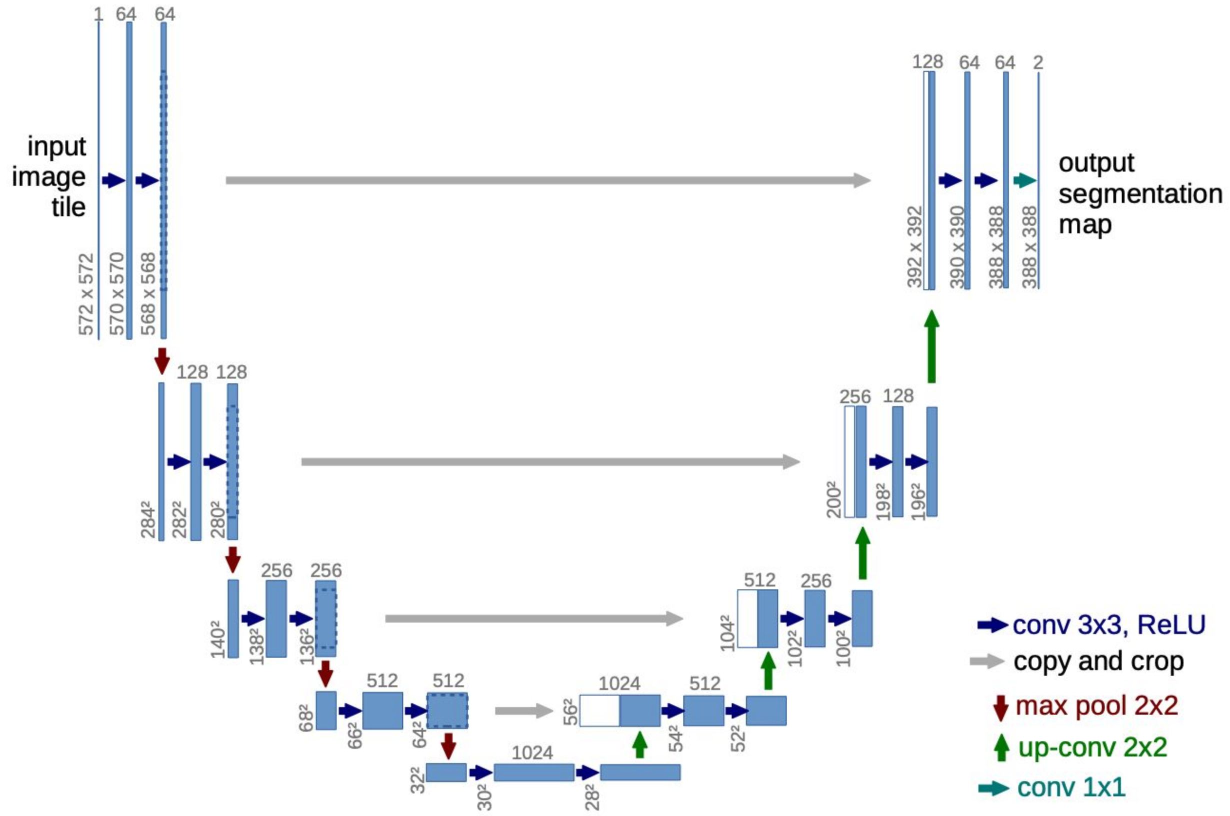
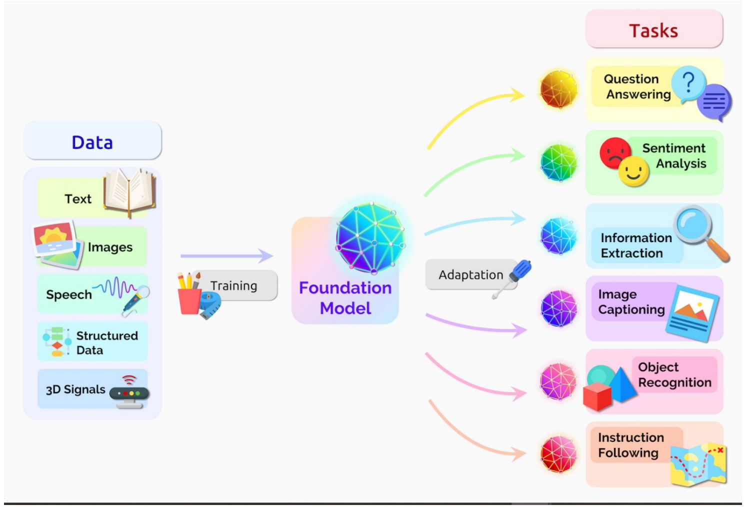
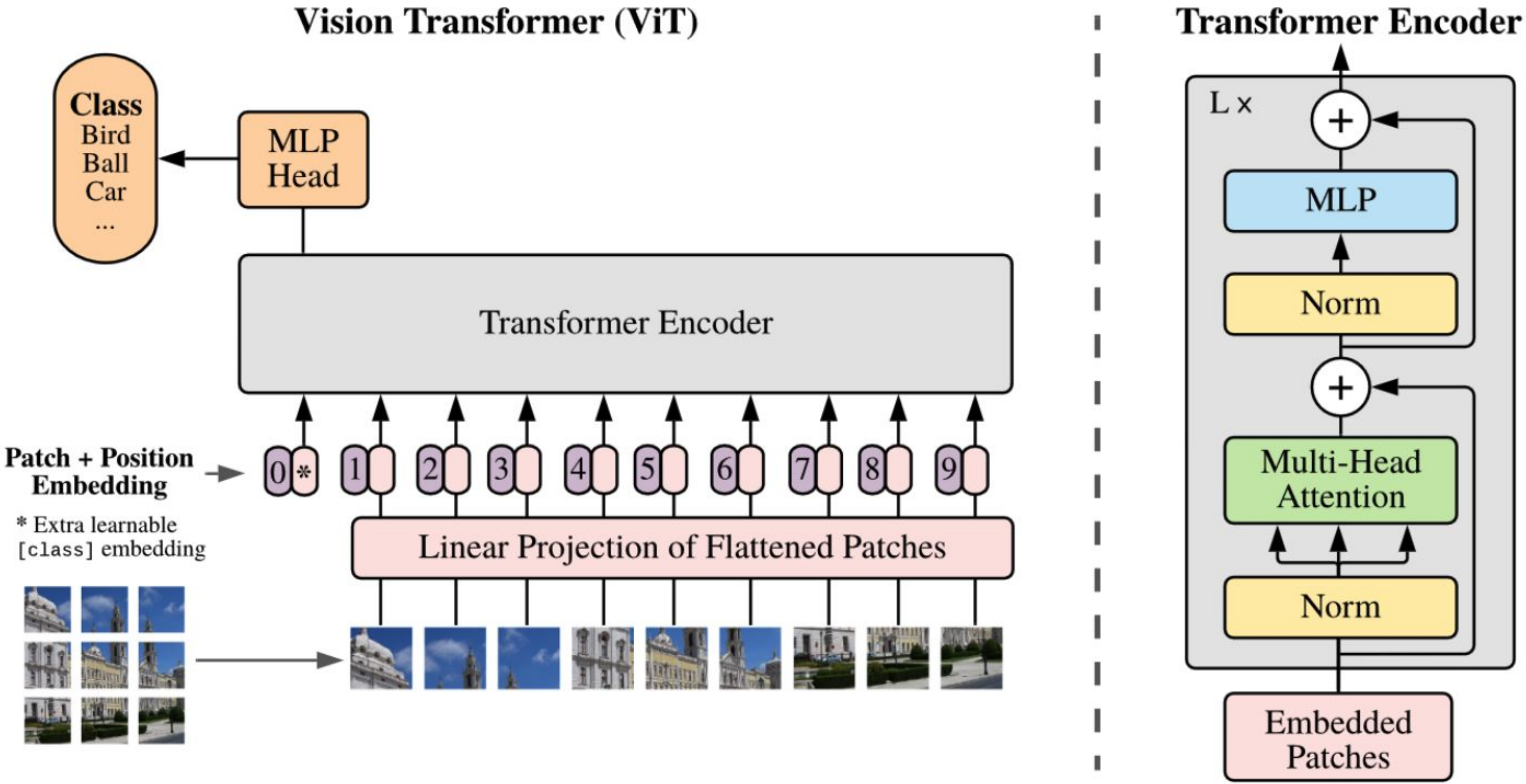


Fig. 1. U-net architecture (example for 32x32 pixels in the lowest resolution). Each blue box corresponds to a multi-channel feature map. The number of channels is denoted on top of the box. The x-y-size is provided at the lower left edge of the box. White boxes represent copied feature maps. The arrows denote the different operations.

Bottleneck of Traditional Methods

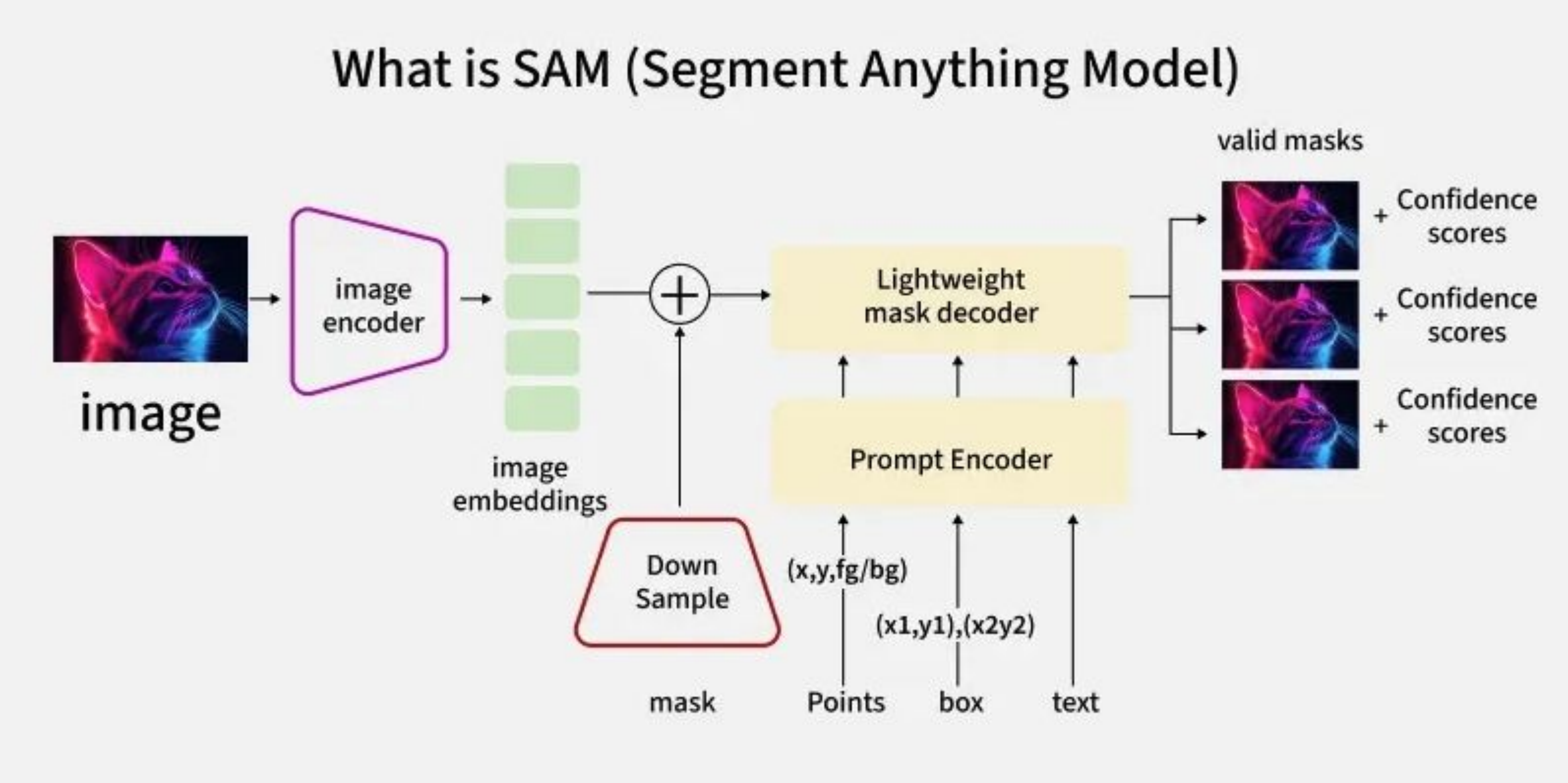
- Traditional models like (CNNs/U-Nets) require thousands of manually annotated biological images for every specific task
- Foundation models trained on massive unlabeled datasets can be adapted to downstream tasks





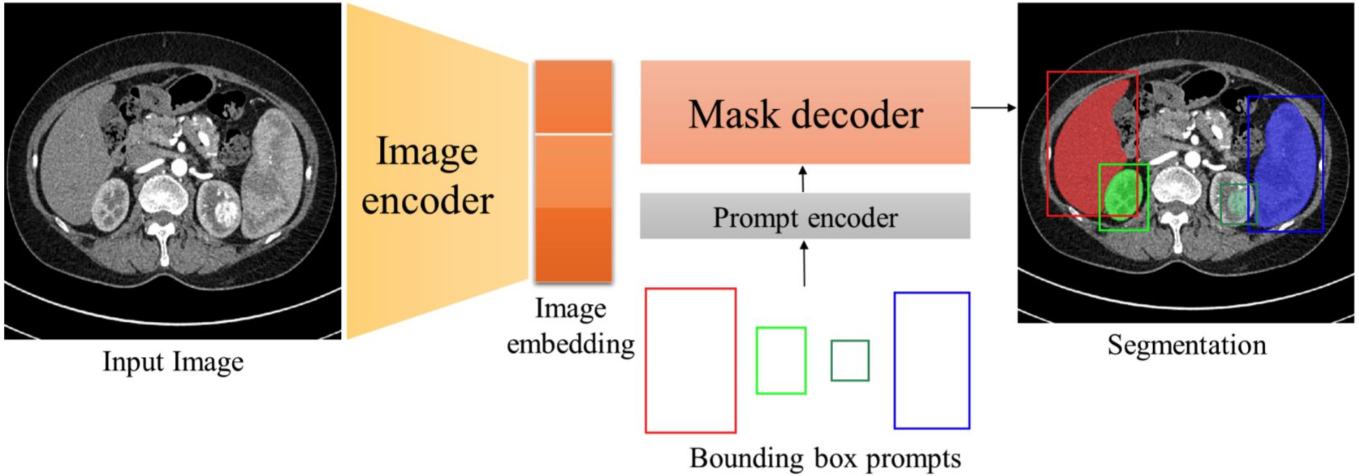
Supervised training on 300M labeled images

Segment Anything Model



Adapting SAM for biology

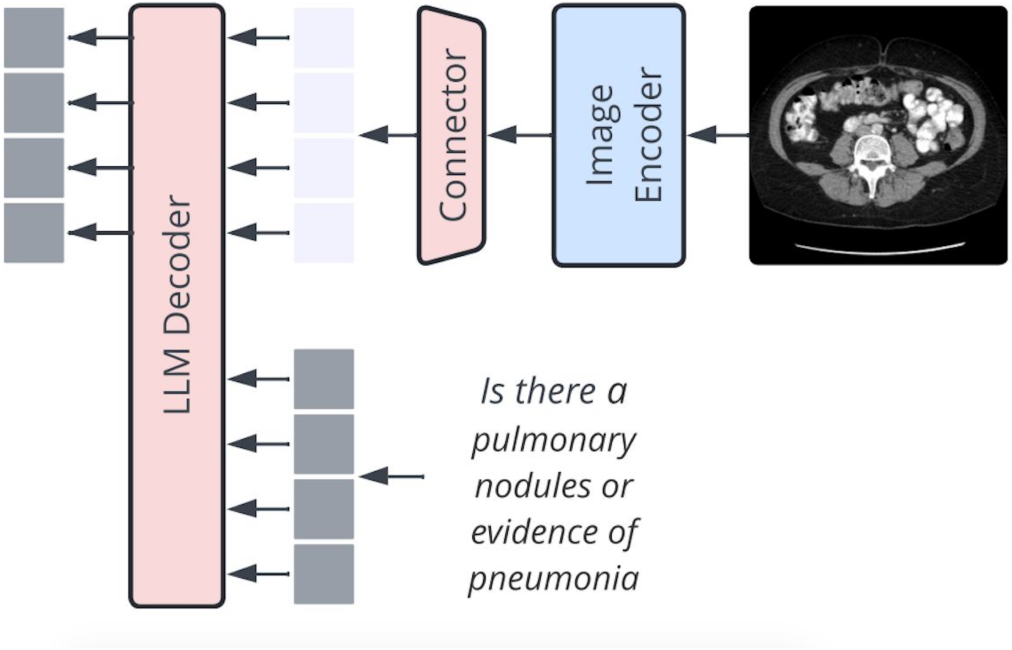
- Original model struggles with low-contrast, highly irregular biological structures
- MedSAM fine tunes the mask decoder on massive proprietary database of medical and microscopy images



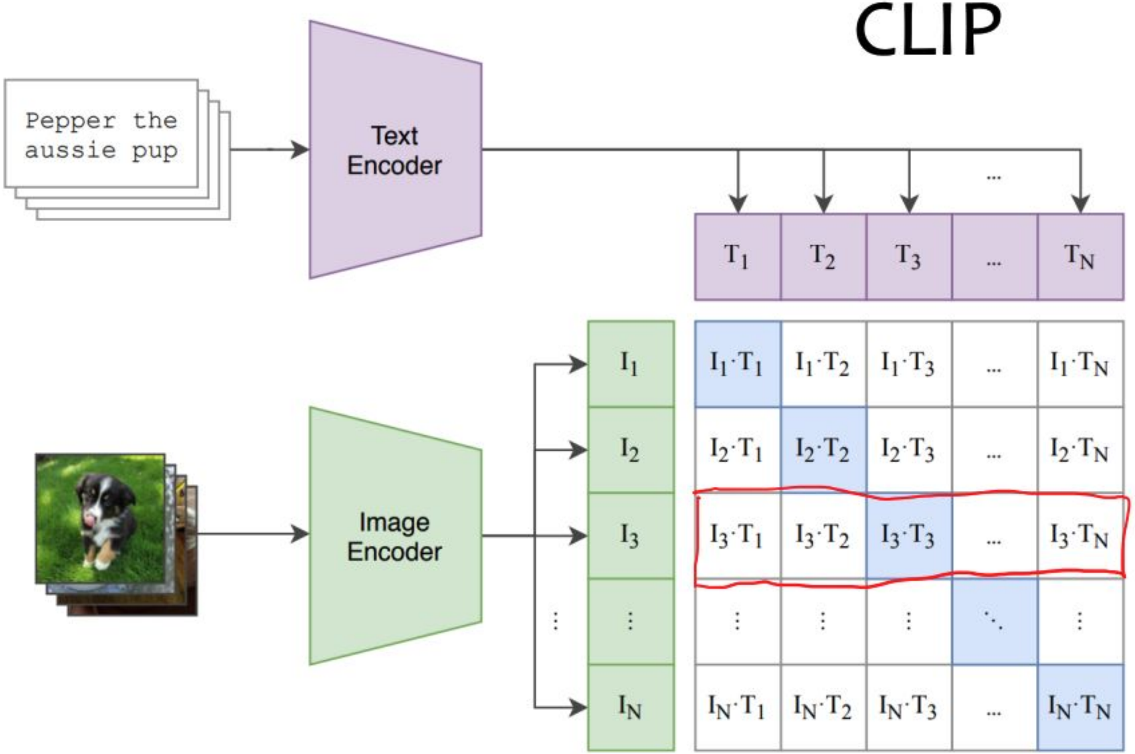
Vision Language Models

- VLM bridge visual and text modalities by mapping images and task into a shared latent space

- The model understands what the image looks like as well as the meaning of words describing it



CLIP


Pepper the aussie pup
Text Encoder
Image Encoder
T_1 T_2 T_3 ... T_N
I_1 I_2 I_3 ⋮ I_N
$I_1 \cdot T_1$ $I_1 \cdot T_2$ $I_1 \cdot T_3$ ... $I_1 \cdot T_N$
$I_2 \cdot T_1$ $I_2 \cdot T_2$ $I_2 \cdot T_3$ ... $I_2 \cdot T_N$
$I_3 \cdot T_1$ $I_3 \cdot T_2$ $I_3 \cdot T_3$ ... $I_3 \cdot T_N$
⋮ ⋮ ⋮ ⋱ ⋮
$I_N \cdot T_1$ $I_N \cdot T_2$ $I_N \cdot T_3$ ... $I_N \cdot T_N$

Correct objective function:

$$\max \sum_{i=1}^{N}\left[\log \frac{\exp\left(\frac{I_i^\top T_i}{\tau}\right)}{\sum_{j=1}^{N} \exp\left(\frac{I_i^\top T_j}{\tau}\right)} + \log \frac{\exp\left(\frac{I_i^\top T_i}{\tau}\right)}{\sum_{j=1}^{N} \exp\left(\frac{I_j^\top T_i}{\tau}\right)}\right]$$

Given a mini-batch of *N* (image, caption) pairs, both encoders are simultaneously pre-trained to maximize the cosine similarity of corresponding image-caption embedding vectors and minimize all other pairwise cosine similarities

Figure from Radford et al. (2021) https://arxiv.org/pdf/2103.00020

VQA in Bioimaging

- You can chat about your imaging data
- For example “Count the number of multi-nucleated cells”
- LLaVA-Med or specialized multimodal LLMs designed for biomedical reasoning

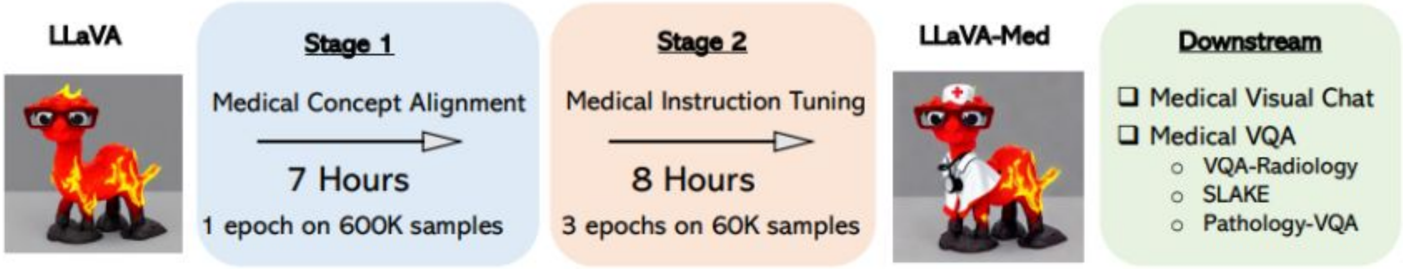
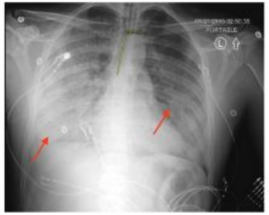


Figure 3: LLaVA-Med was initialized with the general-domain LLaVA and then continuously trained in a curriculum learning fashion (first biomedical concept alignment then full-blown instruction-tuning). We evaluated LLaVA-Med on standard visual conversation and question answering tasks.

Visual input example, Biomedical image from [6]



Caption: Chest x-ray (CXR) on Day 2 of admission post-intubation (yellow line showing the level of the endotracheal tube). Rapidly worsening ground-glass opacities are seen throughout the lungs with relative subpleural sparing (red arrows)

Source: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7398722/

User	What is shown in this image?
LLaVA-Med	The image is a chest X-ray (CXR) that shows bilateral patchy infiltrates, which are areas of increased opacity in the lungs. These infiltrates can be indicative of various lung conditions, such as infections, inflammation, or other lung diseases.
User	Are there any devices implanted in this patient?
LLaVA-Med	Yes, it appears that the patient has several devices implanted. These devices may include endotracheal tubes, central venous catheters, and/or pulmonary artery catheters. These devices are typically used for monitoring and managing various medical conditions, as well as for providing support during surgical procedures.
User	What is shown in this image?
LLaVA	The image shows a ribcage with wires coming out of it, indicating that it is likely a medical illustration or diagram.
User	Are there any devices implanted in this patient?
LLaVA	Yes, there are several devices protruding from the patient's chest, which are likely medical devices used for treatment or monitoring.
User	What is shown in this image?
Language GPT-4	The image shows a chest x-ray taken on Day 2 of a patient's admission after intubation. You can see the endotracheal tube marked by a yellow line, and rapidly worsening ground-glass opacities throughout the lungs. There is also relative subpleural sparing, which is indicated by red arrows.
User	Are there any devices implanted in this patient?
Language GPT-4	Yes, there is an endotracheal tube implanted in the patient, as indicated by the yellow line in the image.

Table 2: Example comparison of medical visual chat and reasoning capabilities. The language-only GPT-4 is considered as the performance upper bound, as the golden captions and inline mentions are fed into GPT-4 as the context, without requiring the model to understand the raw image.