Vision

Machine Learning & Artificial Intelligence

Chen Yishuai

yschen@bjtu.edu.cn

School of Electronic Information Engineering, Beijing Jiaotong University

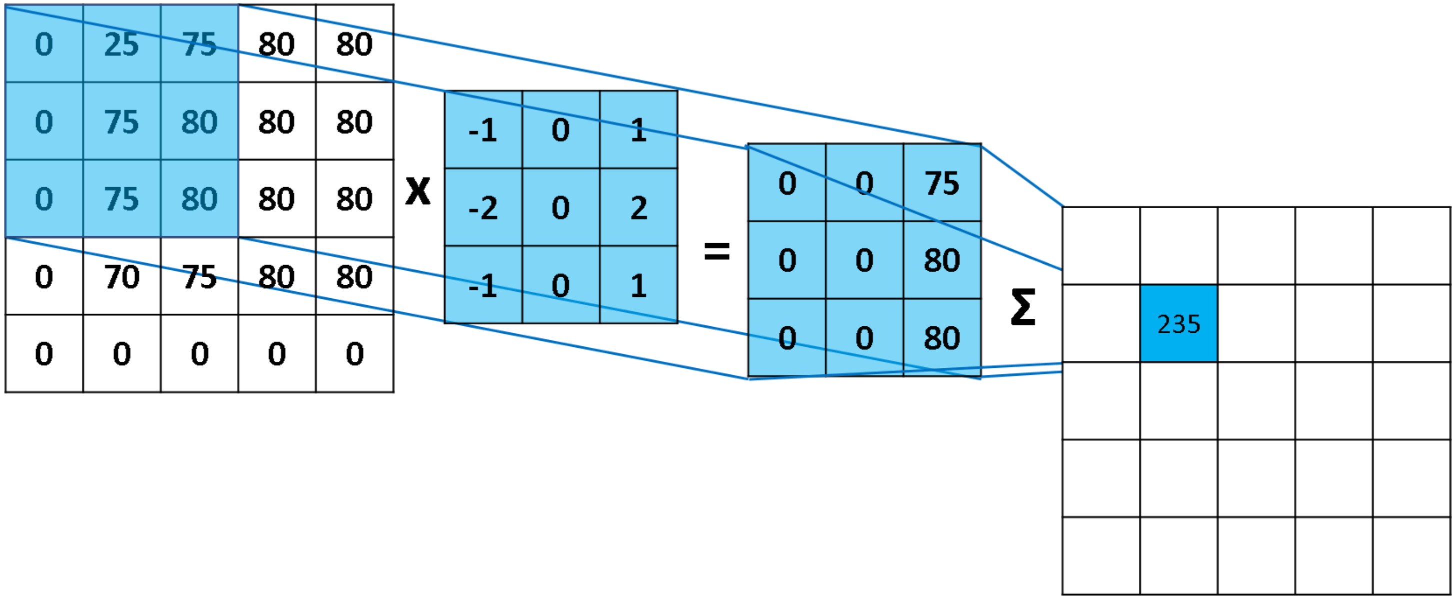
Content

- Convolution and filtering
- Classic methods
- Deep learning methods
- Object detection and recognition
- Applications

Convolution and Filtering

Convolution and Filtering

- 2D convolution
- Multiply the convolution kernel with the pixel value at the corresponding position of the graphic, and then add



0	25	75	80	80
0	75	80	80	80
0	75	80	80	80
0	70	75	80	80
0	0	0	0	0

x

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

=

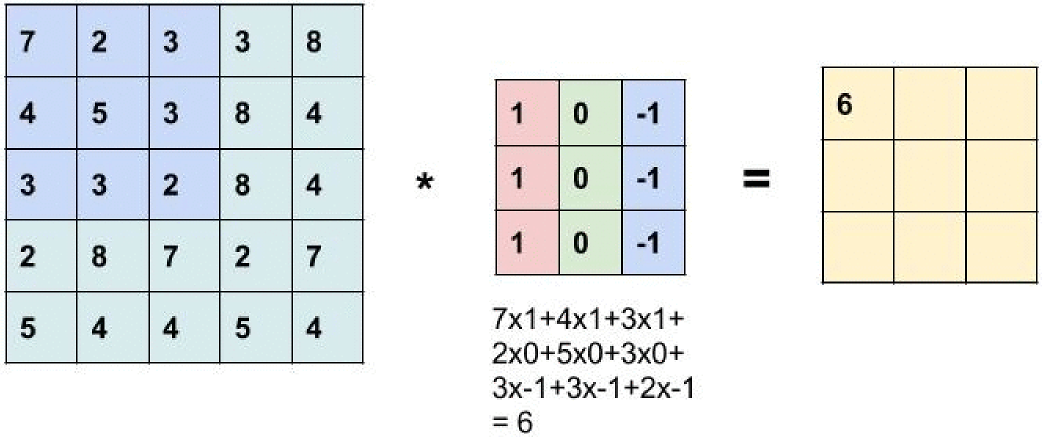
0	0	75
0	0	80
0	0	80

Σ

	235			

Image Convolution

- The convolution kernel slides on the picture to perform the convolution operation

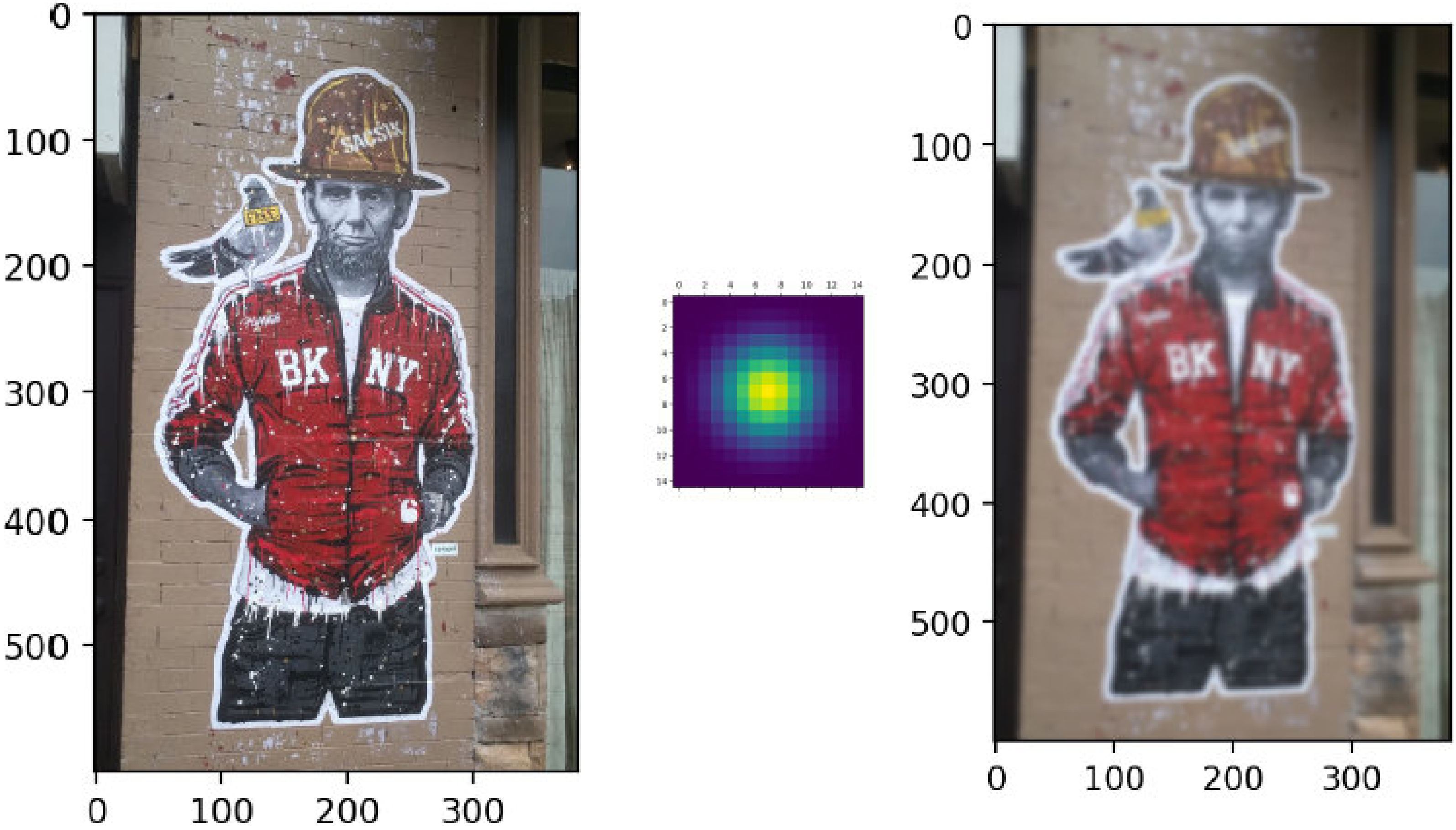


$$\begin{bmatrix} 7 & 2 & 3 & 3 & 8 \\ 4 & 5 & 3 & 8 & 4 \\ 3 & 3 & 2 & 8 & 4 \\ 2 & 8 & 7 & 2 & 7 \\ 5 & 4 & 4 & 5 & 4 \end{bmatrix} * \begin{bmatrix} 1 & 0 & -1 \\ 1 & 0 & -1 \\ 1 & 0 & -1 \end{bmatrix} = \begin{bmatrix} 6 & & \\ & & \\ & & \end{bmatrix}$$

7x1+4x1+3x1+
2x0+5x0+3x0+
3x-1+3x-1+2x-1
= 6

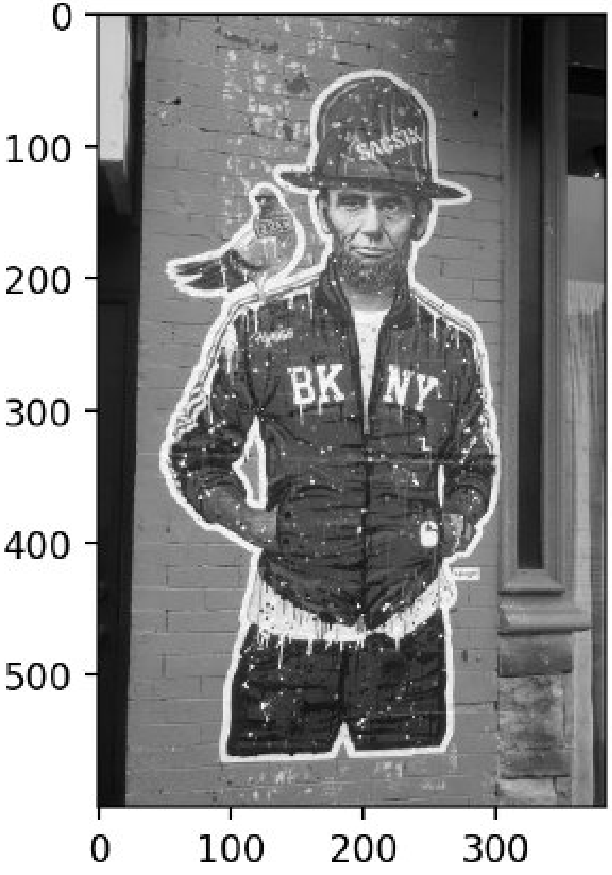
Convolution for Image Smoothing

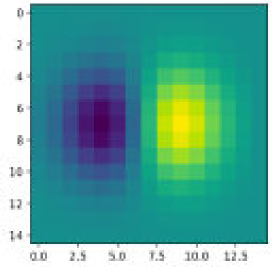
Obscured

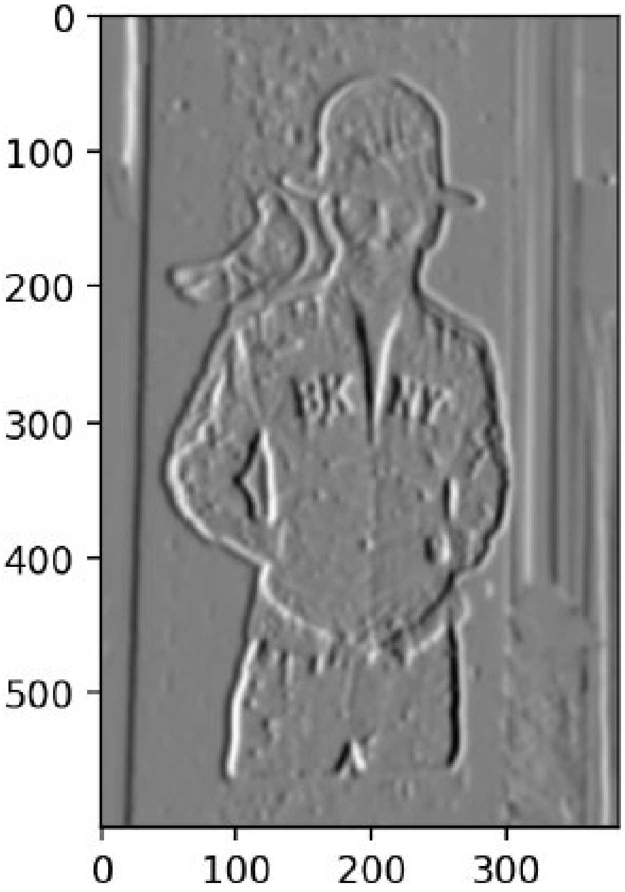


Convolution to Get Image Gradient

Extracting edges







Classic Methods

Feature Extraction

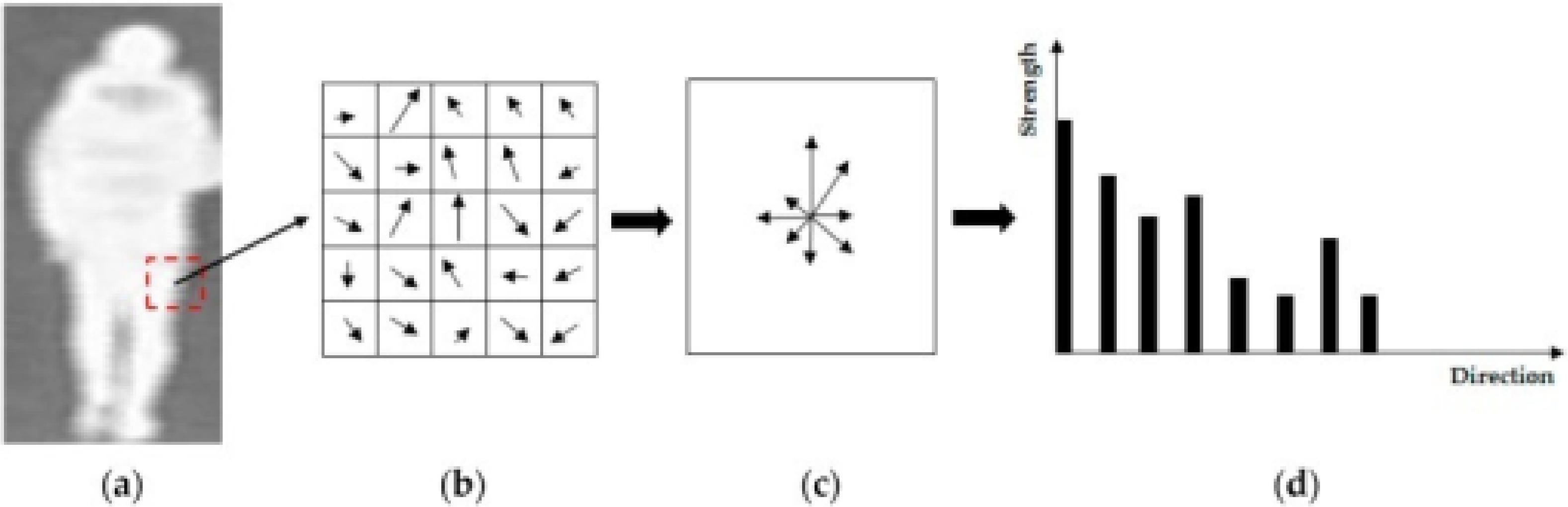
HOG、SIFT、Surf

Image HOG Feature

Histogram of oriented gradient

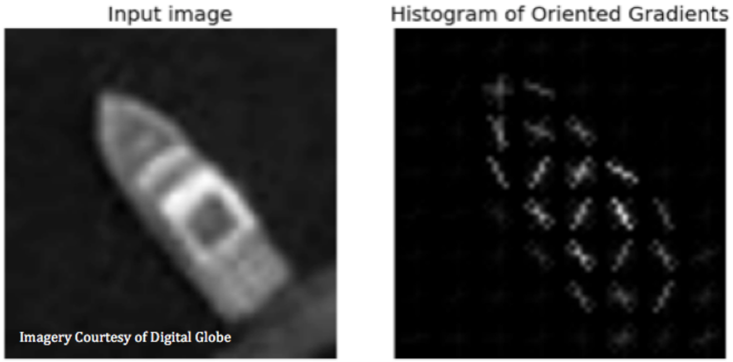
Image HOG feature

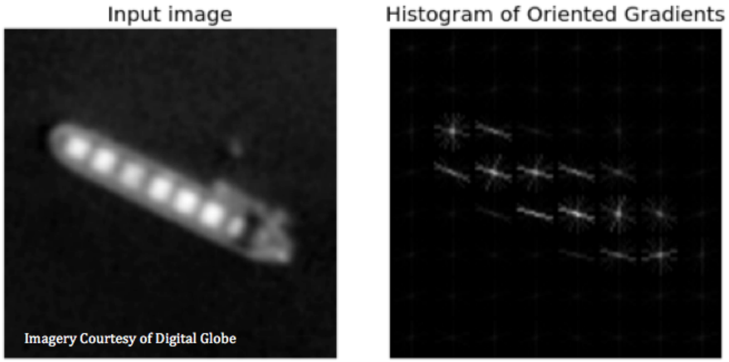
- 2005, Navneet Dalal & Bill Triggs， CVPR
- Suitable for pedestrian detection
- Pedestrians stand upright, subtle body movements do not affect detection results

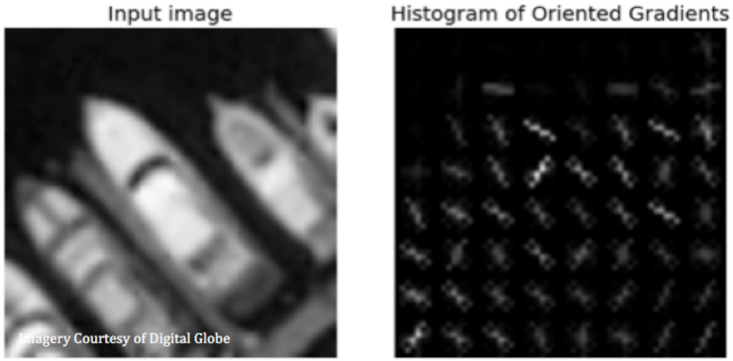


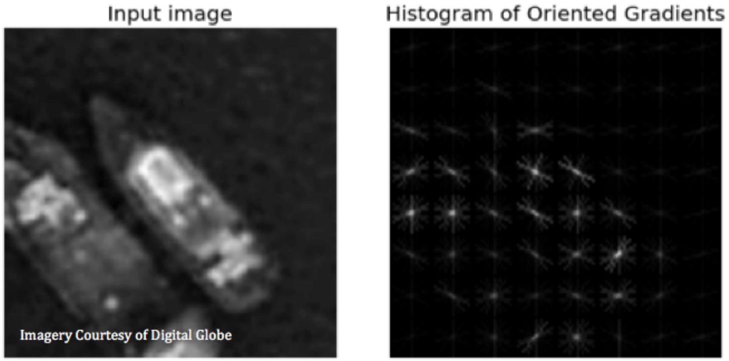
HOG Results

Describe the appearance and shape of objects in the image


Input image
Imagery Courtesy of Digital Globe
Histogram of Oriented Gradients

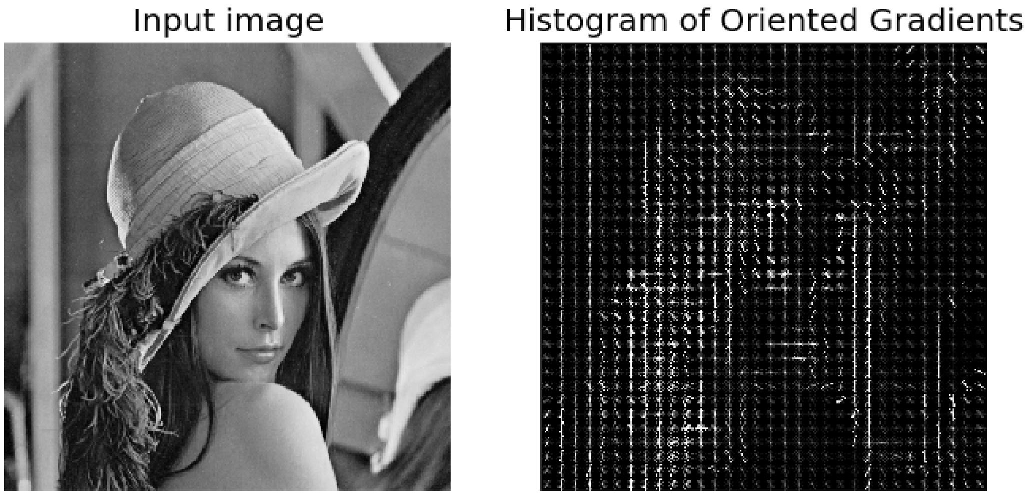

Input image
Imagery Courtesy of Digital Globe
Histogram of Oriented Gradients


Input image
agery Courtesy of Digital Globe
Histogram of Oriented Gradients


Input image
Imagery Courtesy of Digital Globe
Histogram of Oriented Gradients

Implementation

1. Divide the image into tiles and calculate the pixel gradient or edge direction in the tile
2. Use the statistic of histogram as features
3. Normalized to deal with light changes and shadows



SIFT Algorithm

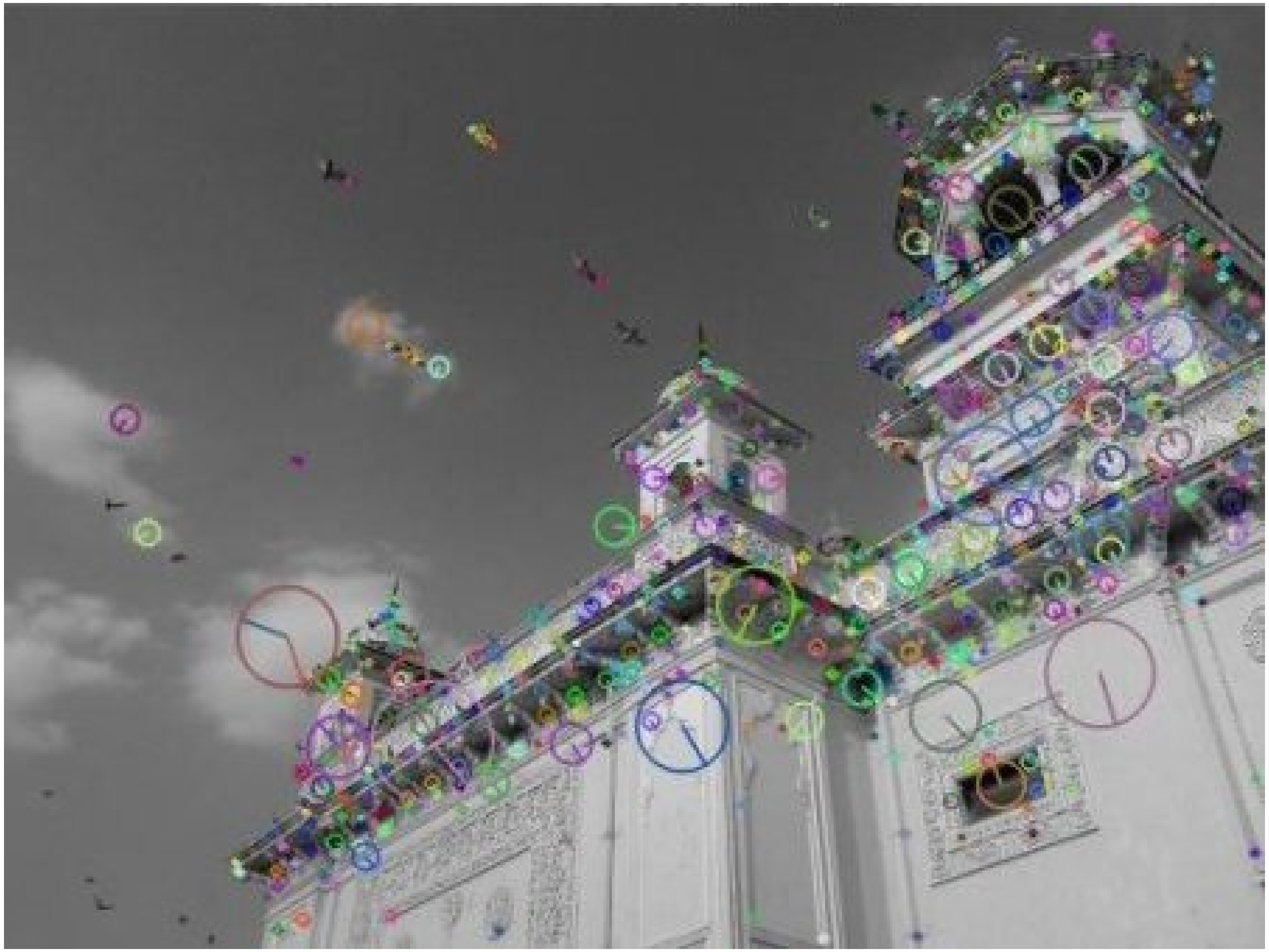
Key point detection and description

SIFT

- Scale-invariant feature transform
- Widely used in object recognition
- More than 3 SIFT features are sufficient to calculate the position and orientation of the target
- David Lowe, published in 1999, refined in 2004

Idea of SIFT

- Find the position, size, and direction of key points

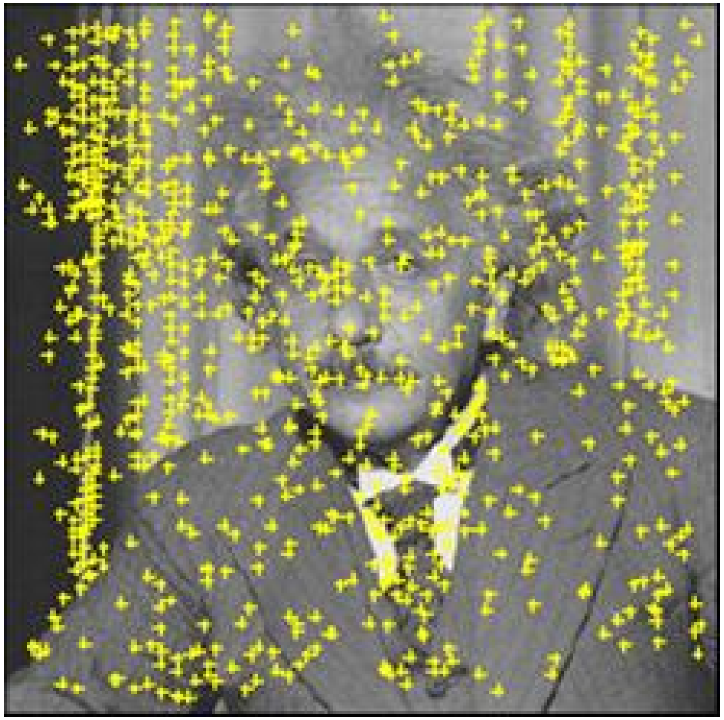


Key Point

- Extreme Value Detection
- Keypoint Positioning
- Key Point Description

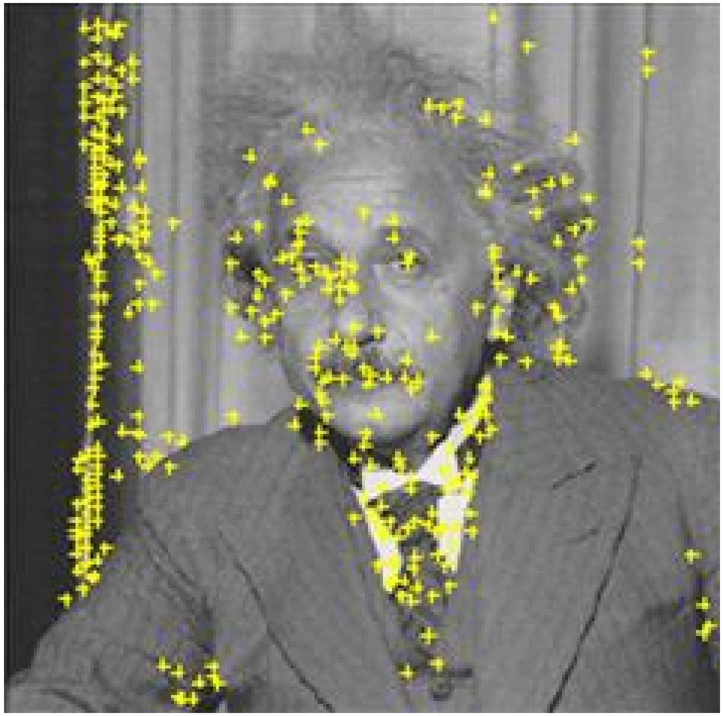
Extreme Value Detection

- Image convolution with Gaussian filtering at different scales
- Find key points using the differences in convolution results

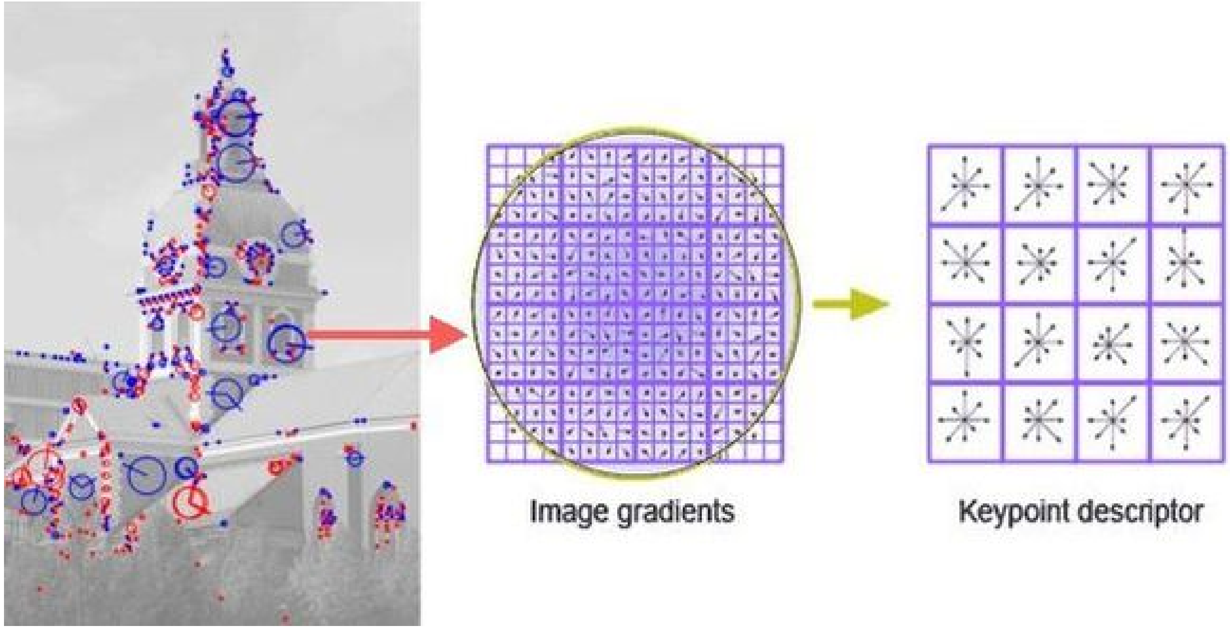


Keypoint Positioning

- From pixel information near key point, key point size, main curvature, screening key points
- Eliminate key points susceptible to noise



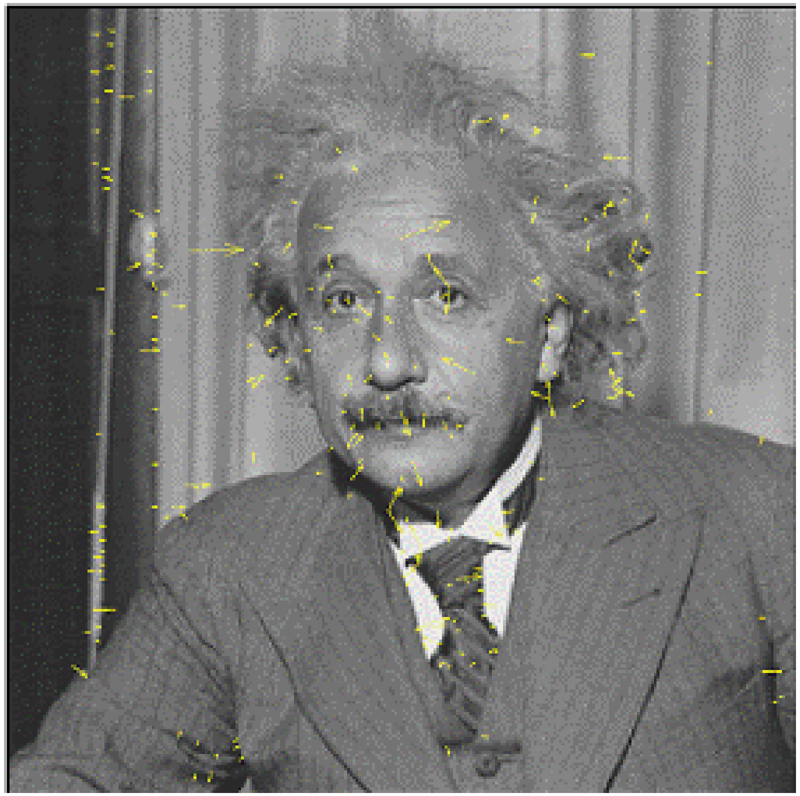
Key Point Description



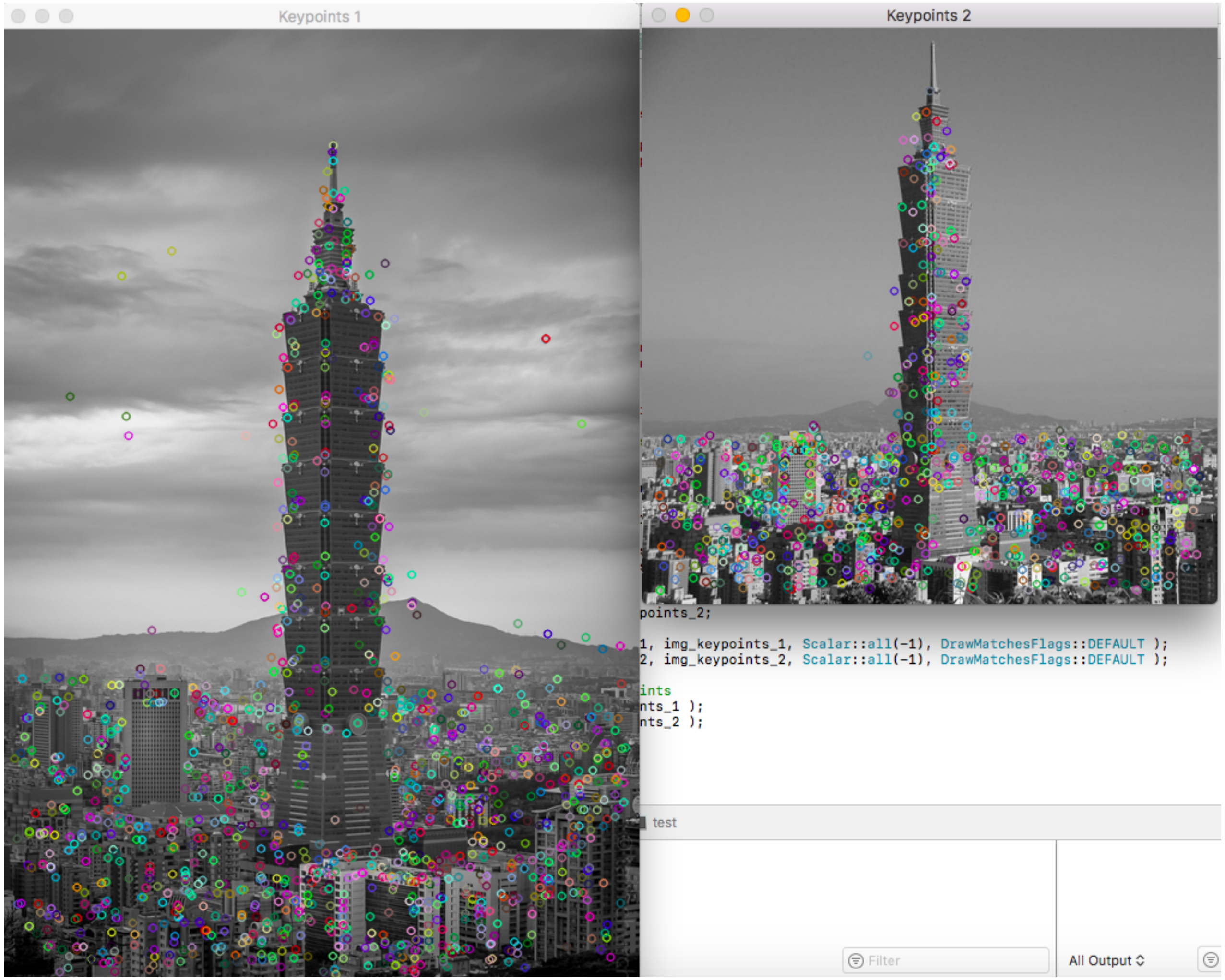
A 500*500 image, get about 2000 features

Key Point Descriptor

- Based on histogram, so it stays the same under different light and viewing angles



Keypoint Extraction Results

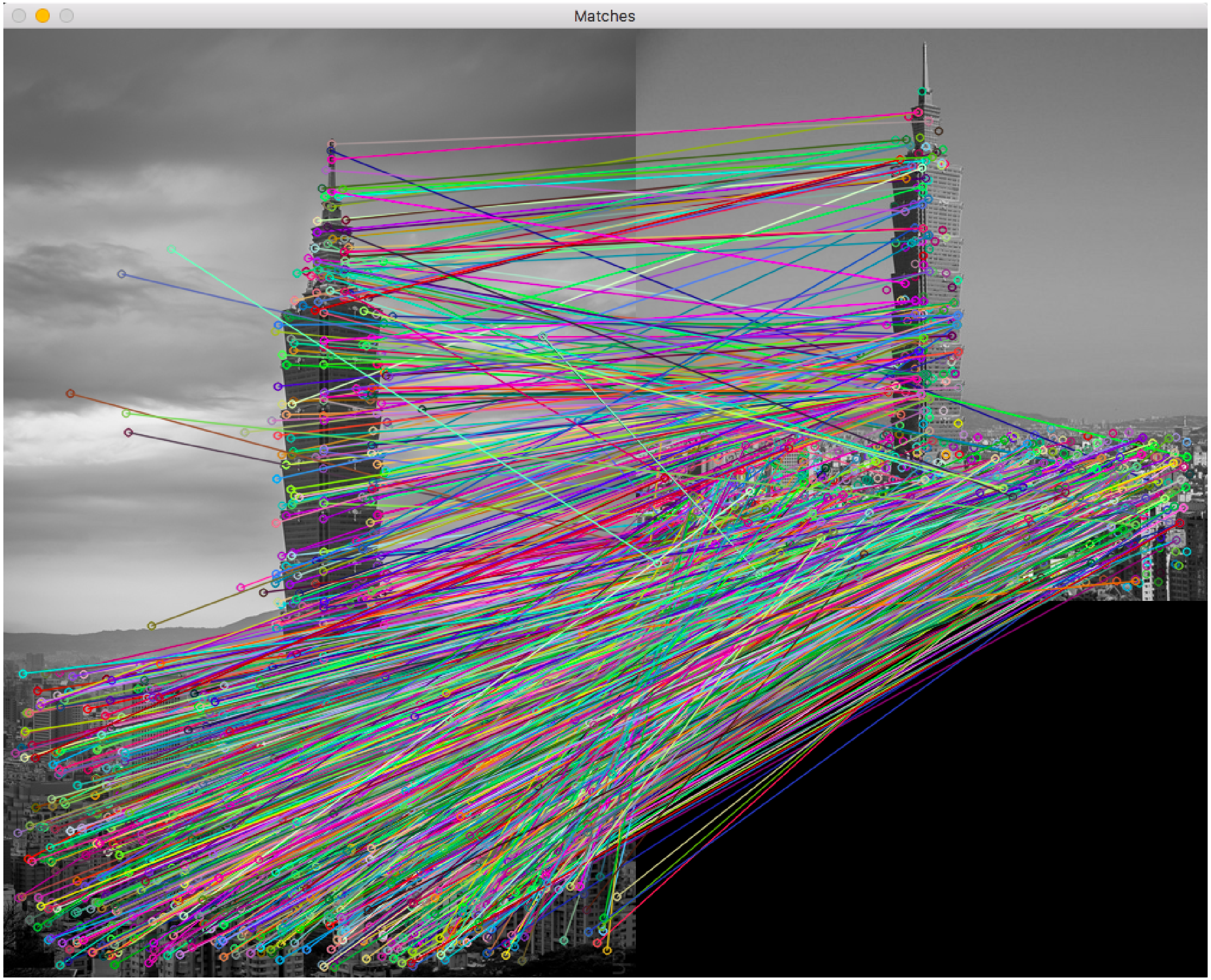


```
points_2;

1, img_keypoints_1, Scalar::all(-1), DrawMatchesFlags::DEFAULT );
2, img_keypoints_2, Scalar::all(-1), DrawMatchesFlags::DEFAULT );

ints
nts_1 );
nts_2 );
```

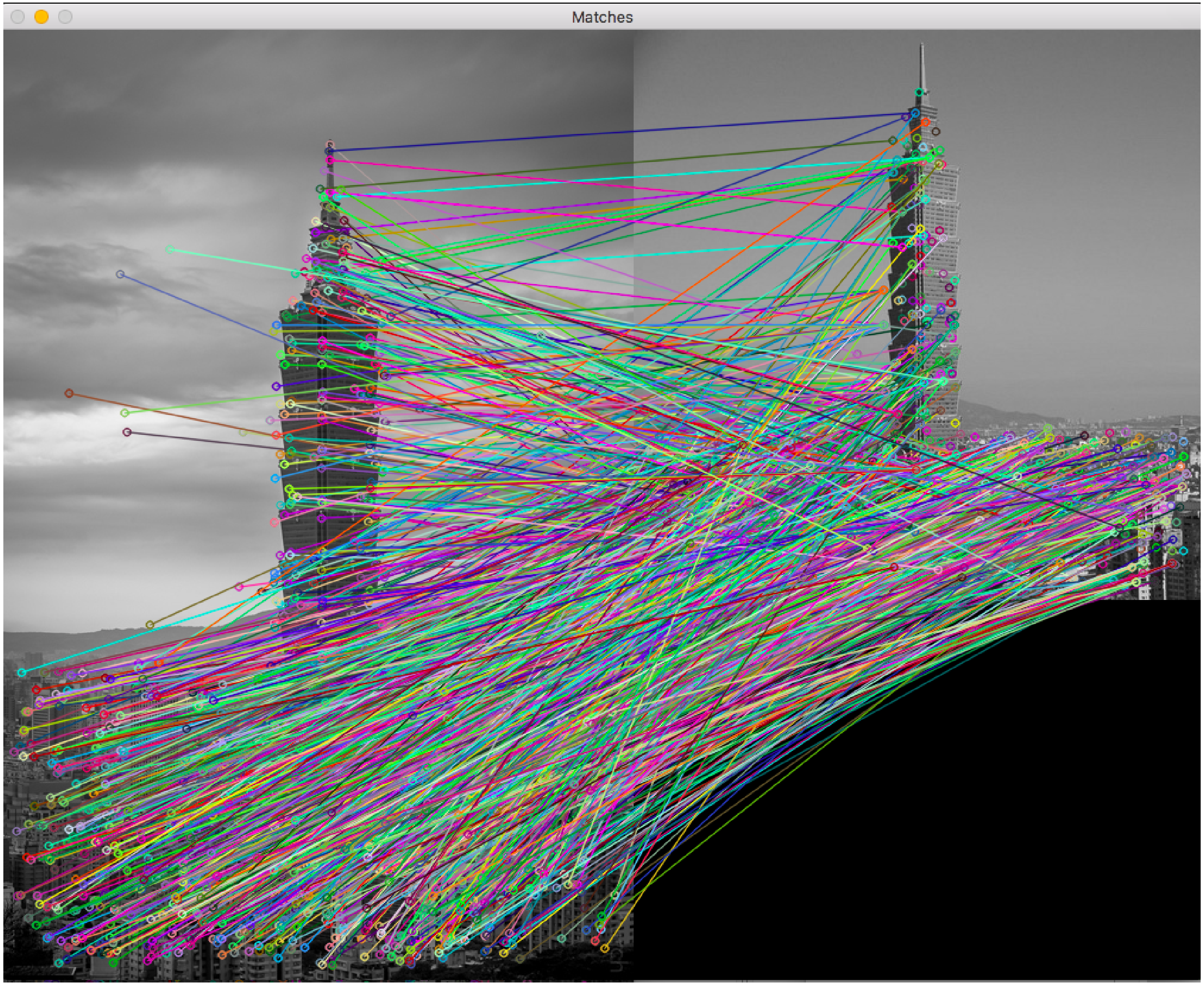
Matching



SURF

- Speeded Up Robust Features
- 2006, ECCV
- Inspired by SIFT, similar, faster, more stable performance
    - Feature point detection and description
    - Descriptor pairing

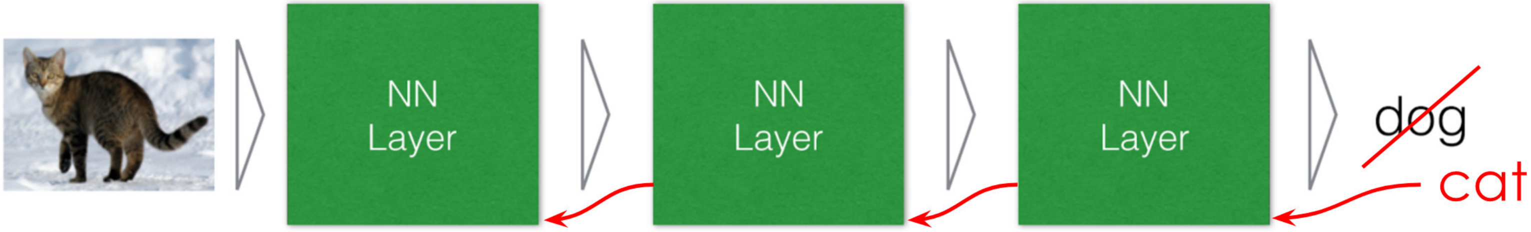
SURF Algorithm Results



Deep Learning Methods

Deep CNN

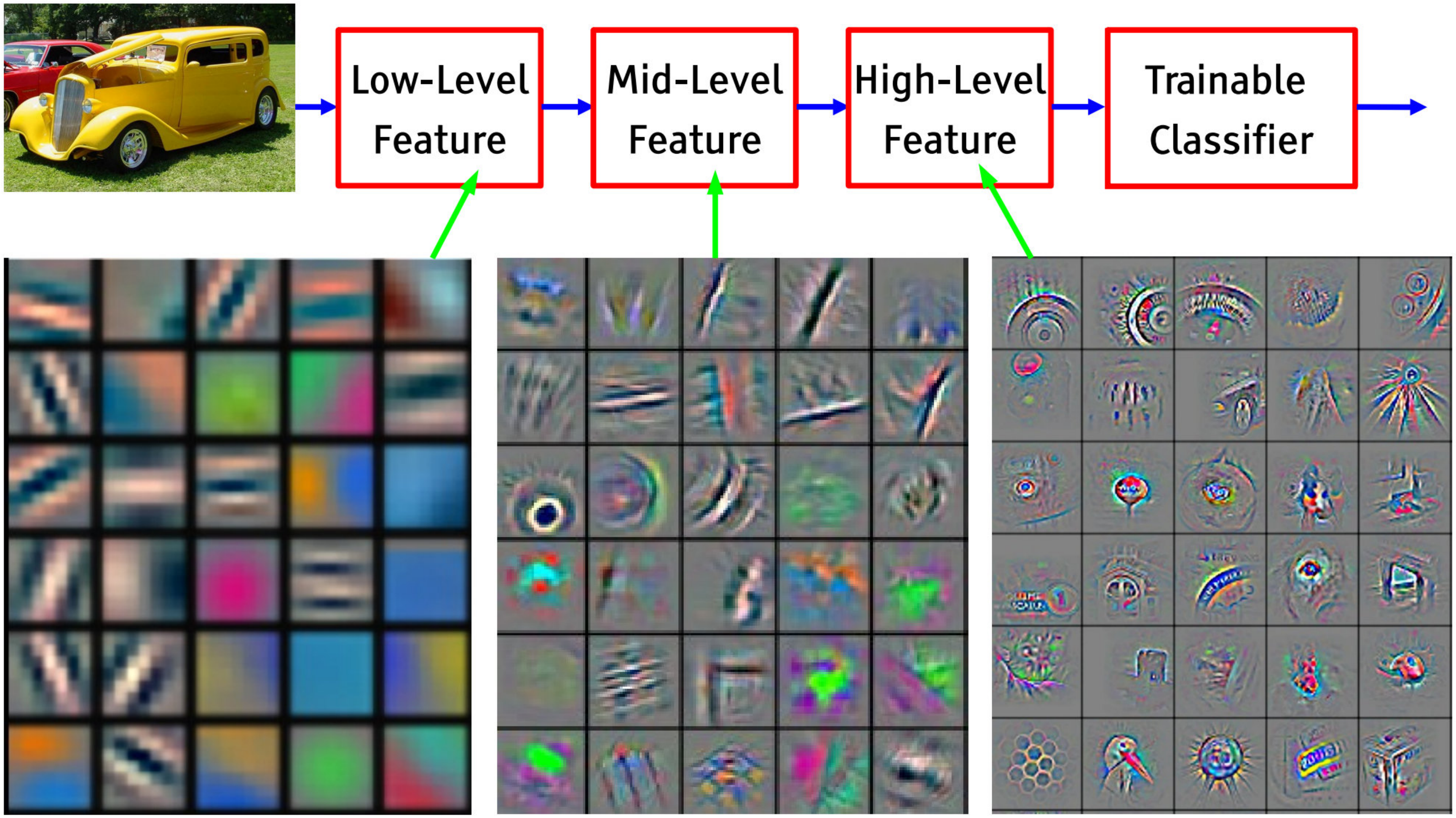
- Send the raw data directly to the multilayer neural network for learning
- Multiple convolution and pooling layers
- An error occurred, adjusting the convolution kernel all the way



Supervised Learning

Deep CNN

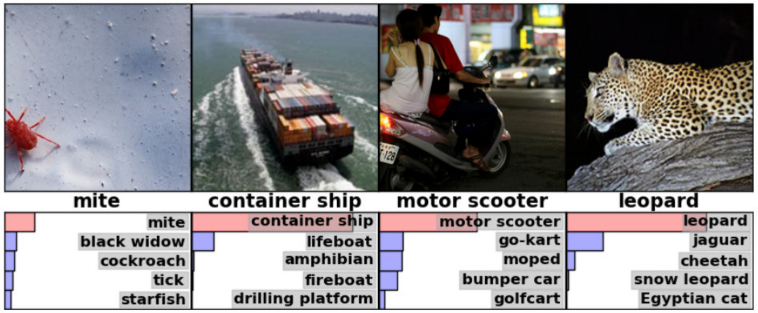
- Extract simple features at the bottom and complex features at the high level



Feature visualization of convolutional net trained on ImageNet from [Zeiler & Fergus 2013]

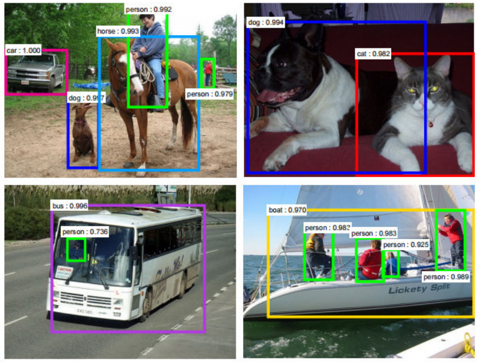
Application

Object detection, recognition, handwriting recognition, object segmentation

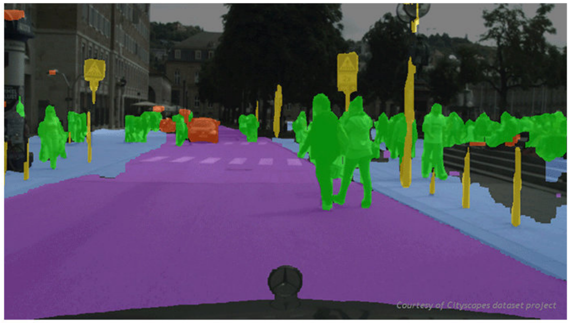


[Krizhevsky 2012]

[Ciresan et al. 2013]



[Faster R-CNN - Ren 2015]

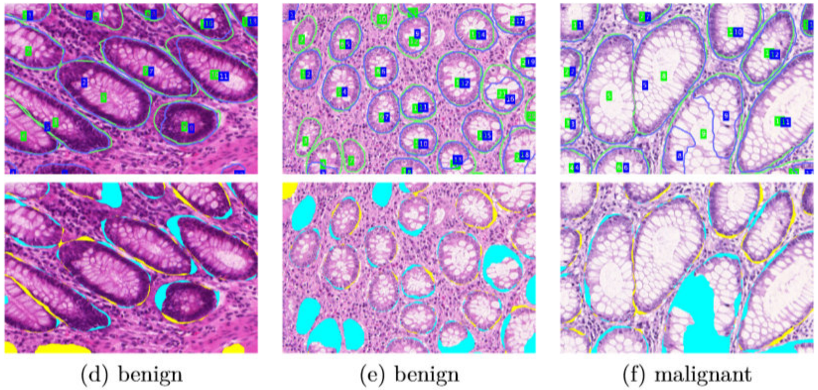


[NVIDIA dev blog]

Application

Disease recognition, face recognition, facial element recognition

[Stanford 2017]



(d) benign (e) benign (f) malignant

[Nvidia Dev Blog 2017]

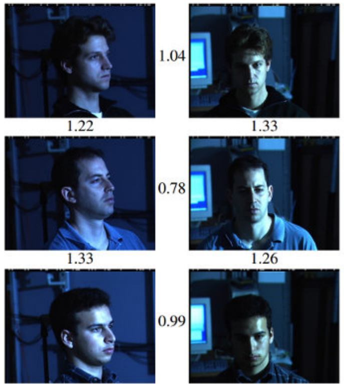
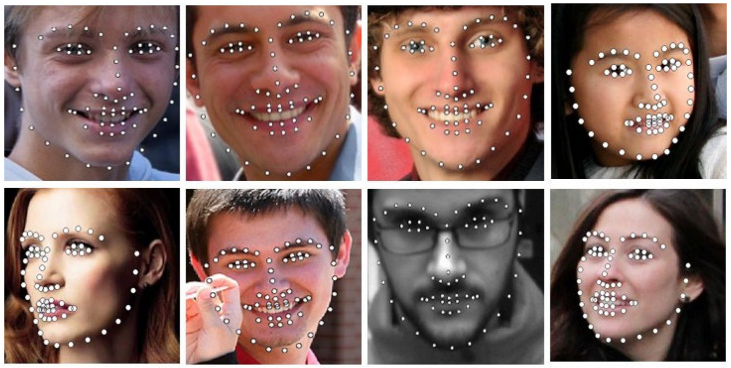


Figure 1. **Illumination and Pose invariance.**

[FaceNet - Google 2015]

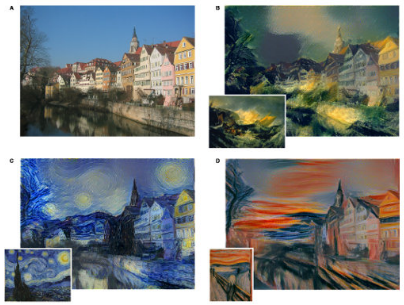


[Facial landmark detection CUHK 2014]

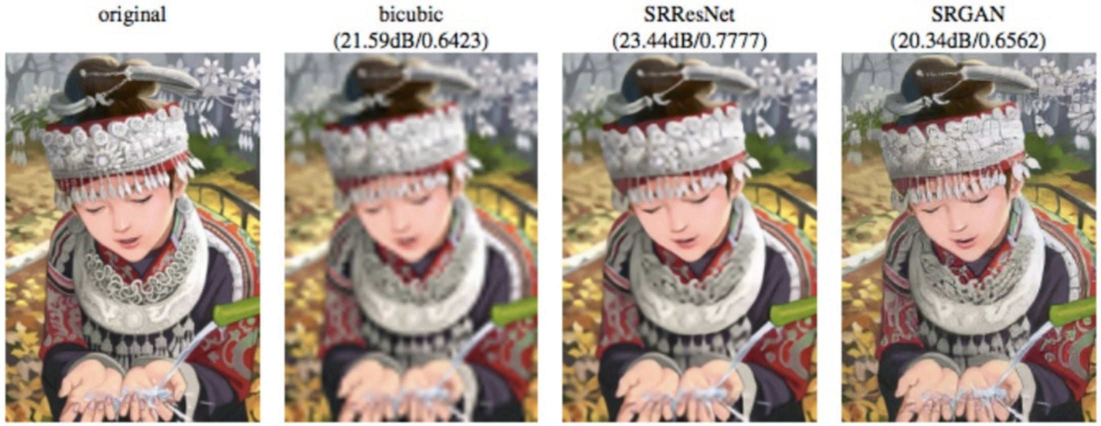
Application

Painting, image style conversion, sharpness enhancement

[DeepDream 2015]



[Gatys 2015]



[Ledig 2016]

Denoising

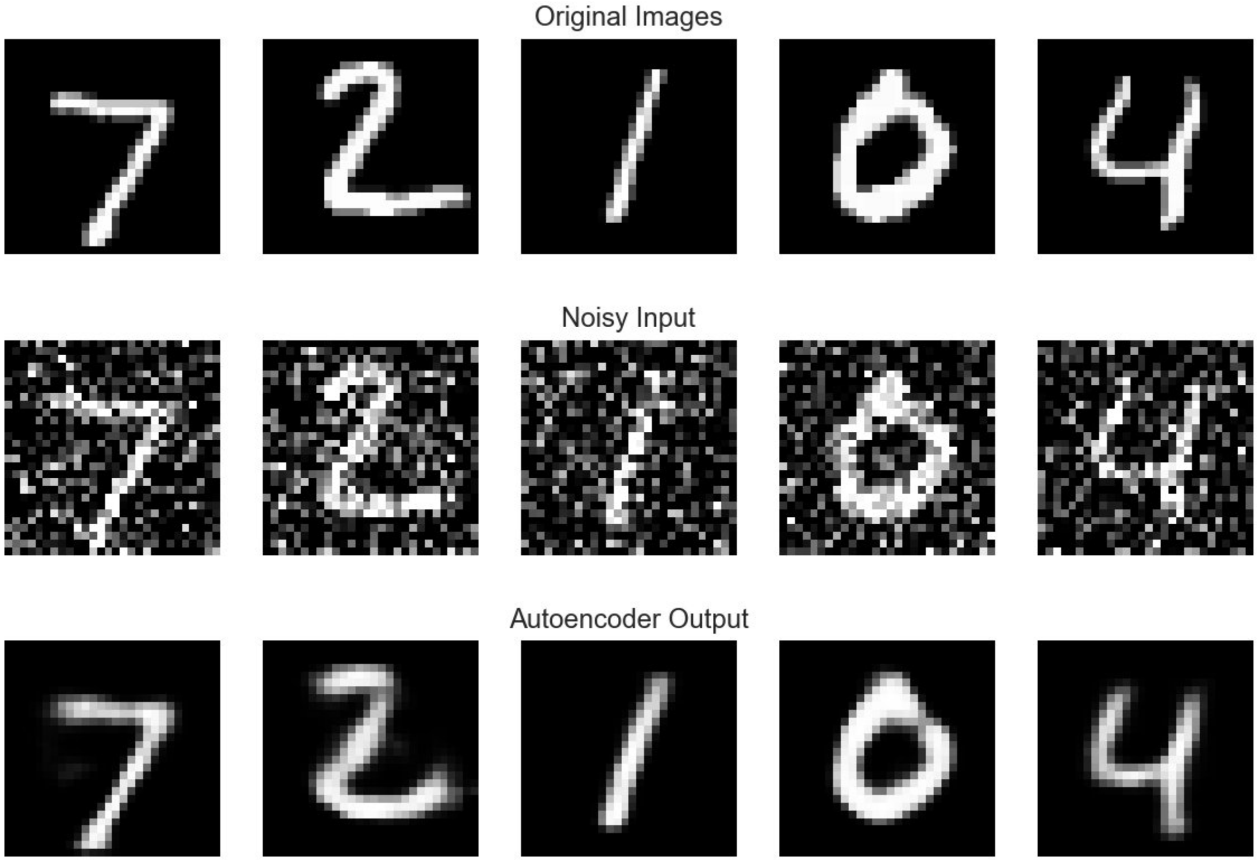


Image Conversion

- Image restoration, rendering, coloring
- Map extraction, scene conversion

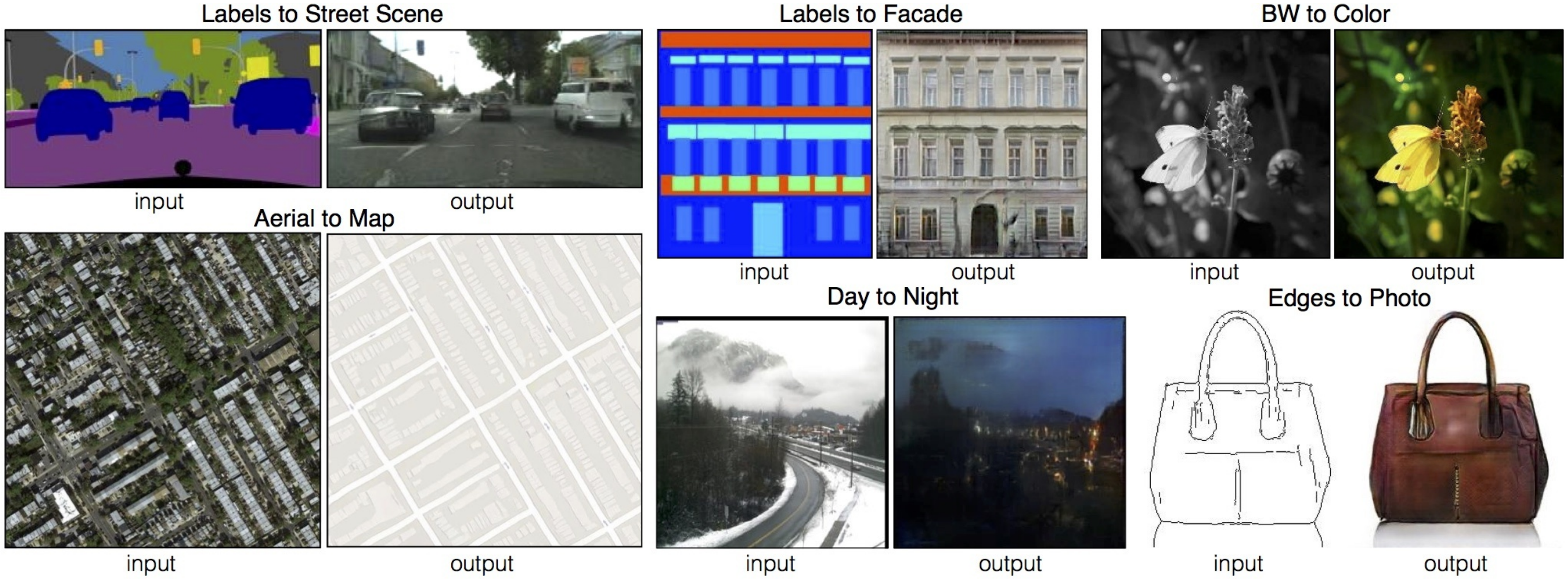
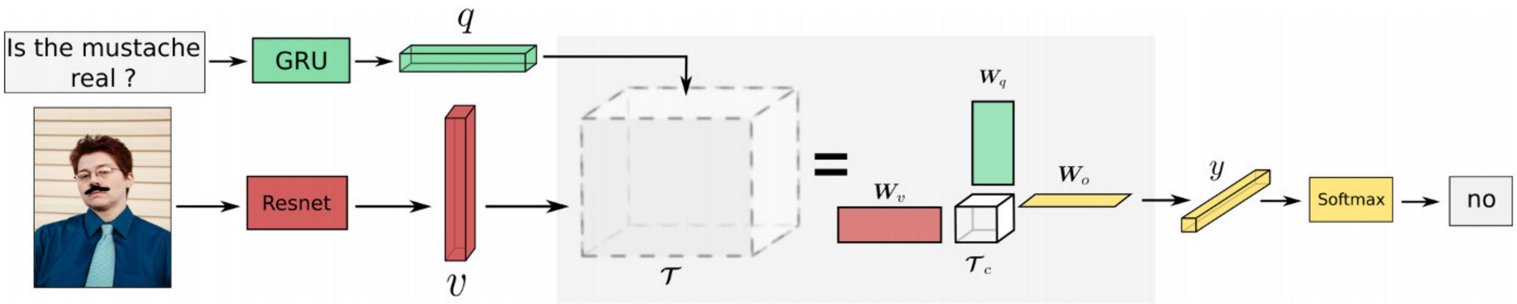


Image understanding

Image - Q&A - Text description



[VQA - Mutan 2017]

"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."

"boy is doing backflip on wakeboard."

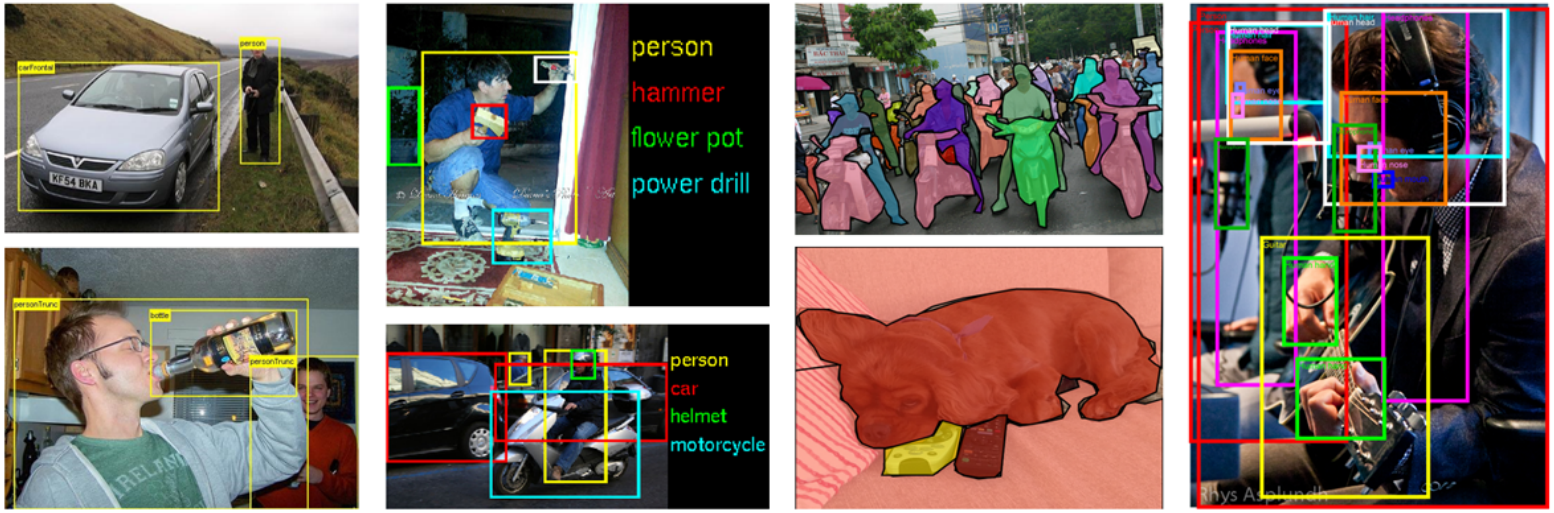
[Karpathy 2015]

Generating Image Descriptions (2015)

Object Detection & Recognition

Problems

Object detection, segmentation, recognition



(a) (b) (c) (d)

Difficulty

Masking, interference, noise

Mop dog

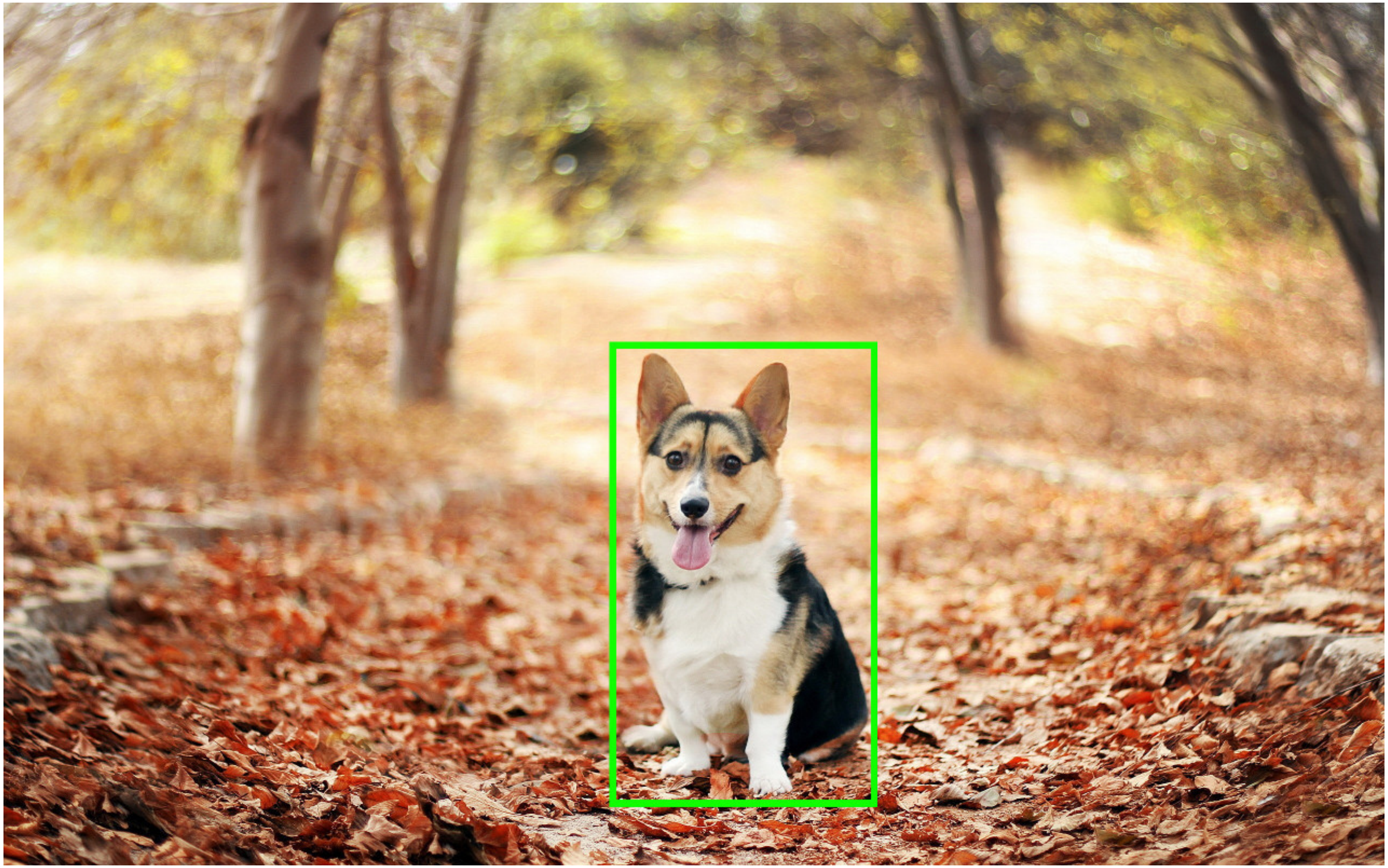
Cake dog

Clouds

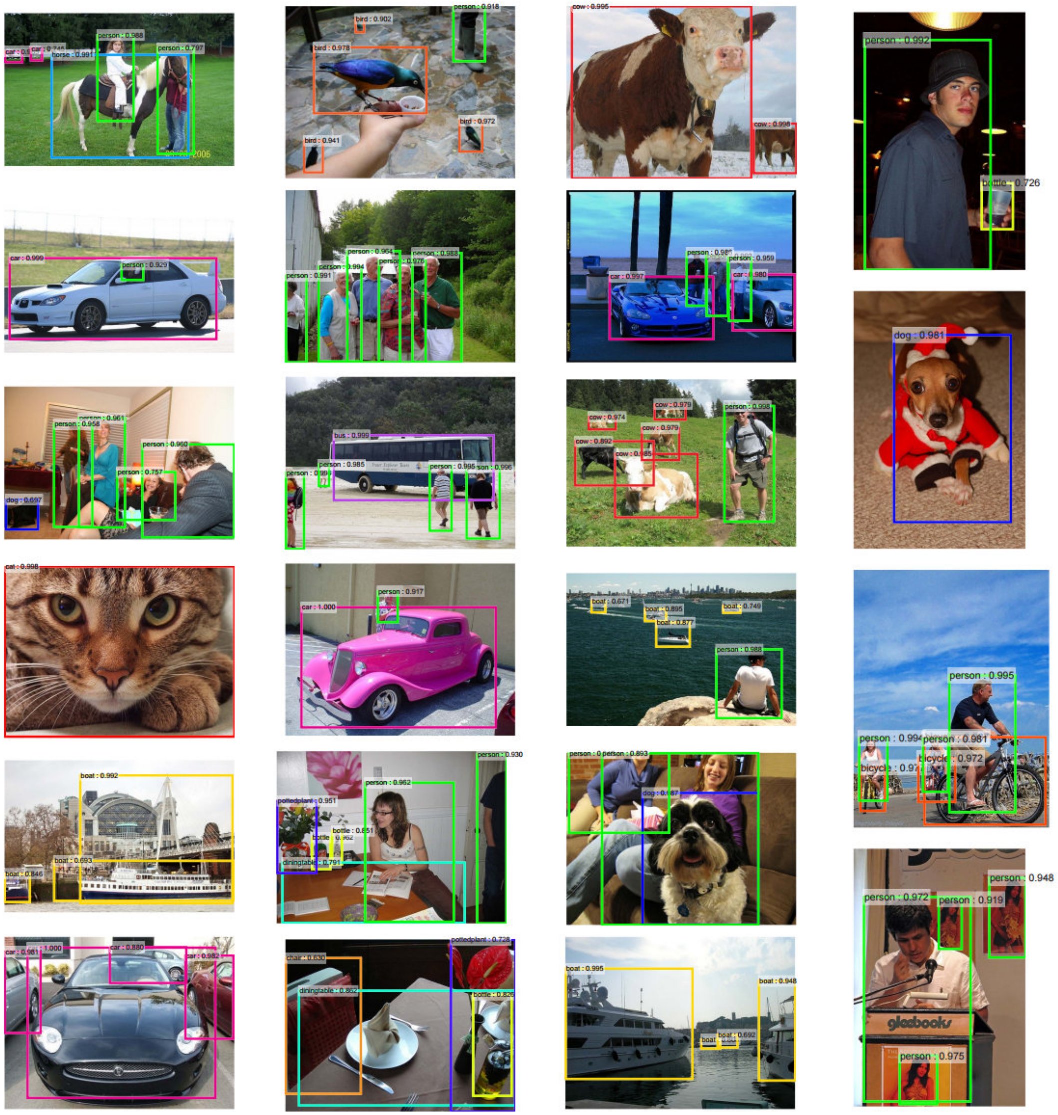
Object Detection

Image

Object Detection



Object Detection

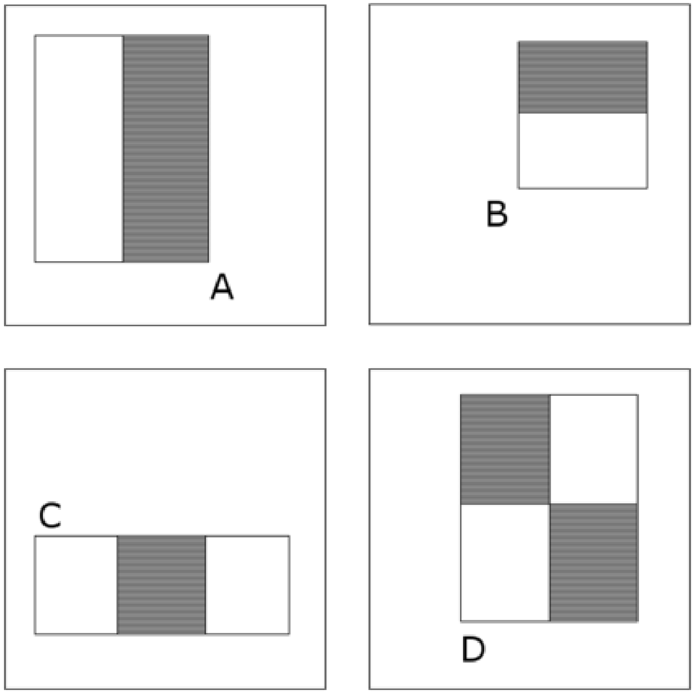


1) Traditional method

- V-J Detection
- HOG Detection
- DPM Algorithm

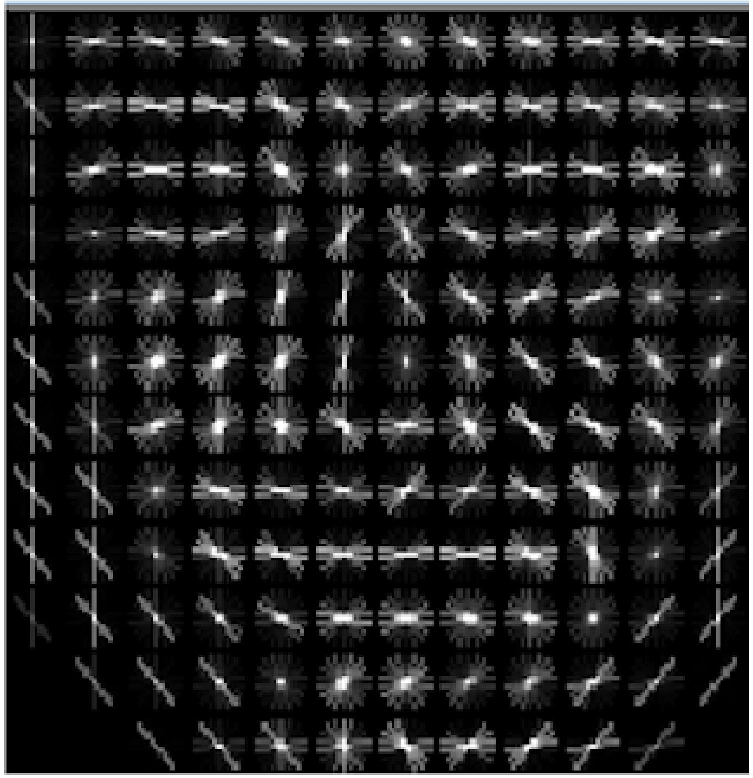
V-J Detection

- 2001，Paul Viola，Michael Jones
- Human face detection
- Haar feature



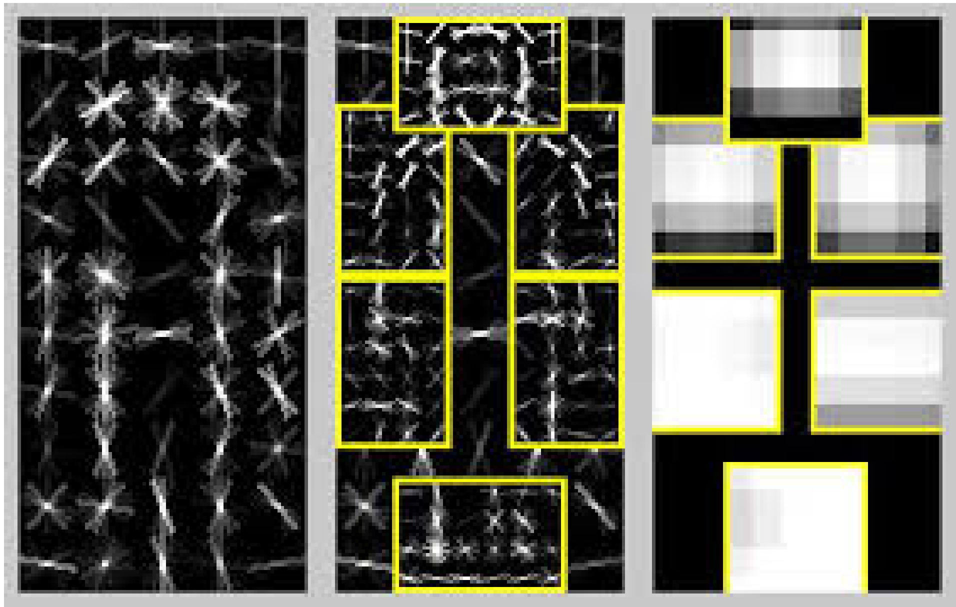
HOG Detection

Pixel gradient



DPM Algorithm

- Deformable Part-Based Model
- Each part has its own classifier (eg: eyes, mouth)
- The position of each part should be reasonable (eg: eyes above mouth)

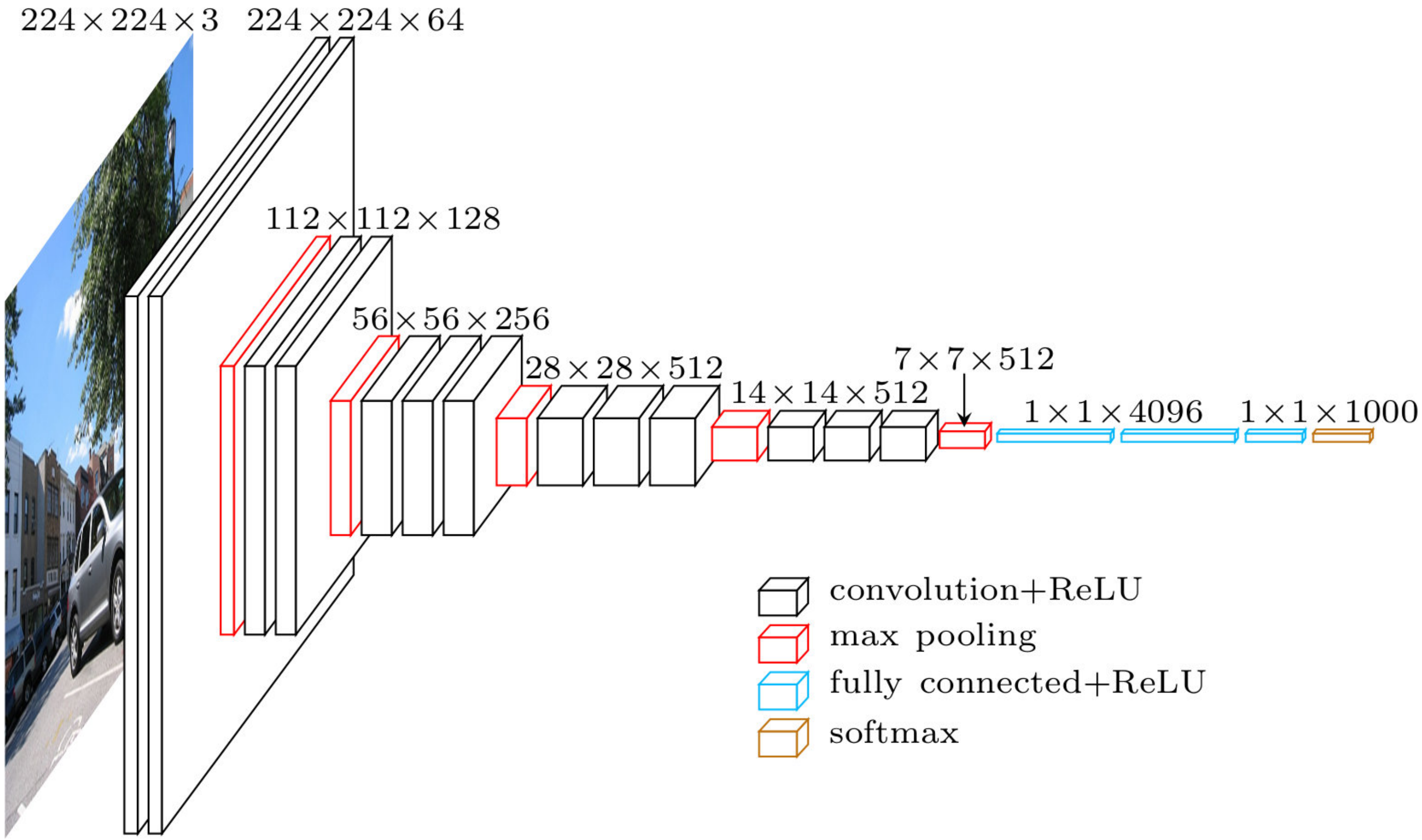


2) Deep Learning Methods

- 2012，AlexNet
- Two-stage detector
  - Find the area before identifying the target
  - RCNN、Pyramid Networks
- Single-stage detector
  - Identify the target without finding the area
  - YOLO、SSD、Retina-Net
- Evaluation mAP
  - VOC 83% （2018)，COCO（69% 2019)

VGG16

- CNN objection Recognition
- Oxford university， K. Simonyan， A. Zisserman， 2014

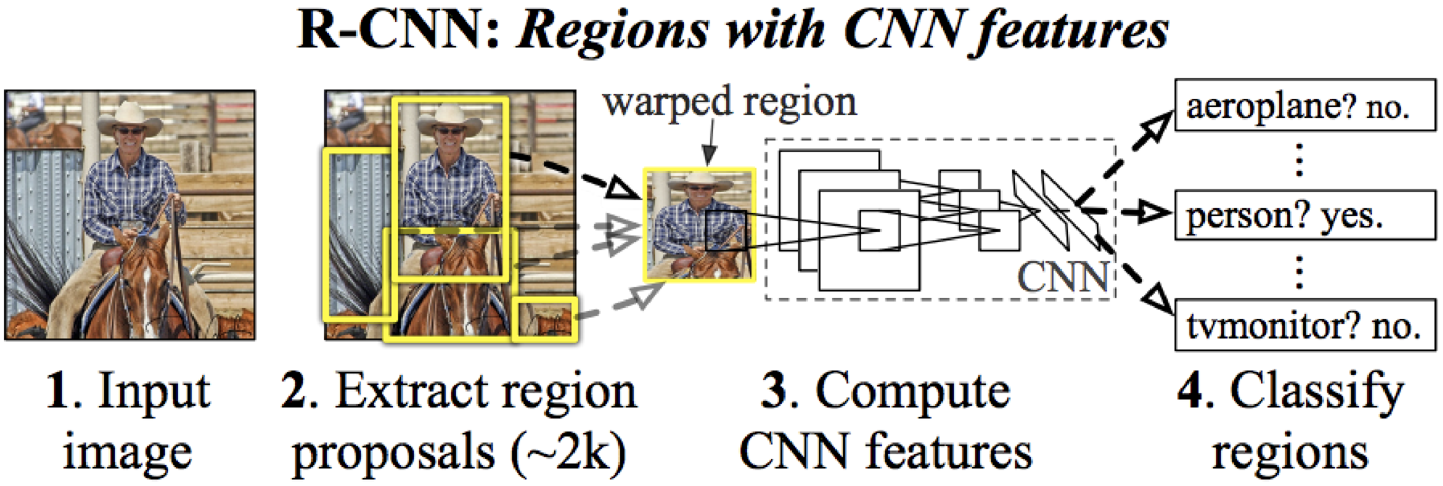


Two-stage detector

Find the area before identifying the target

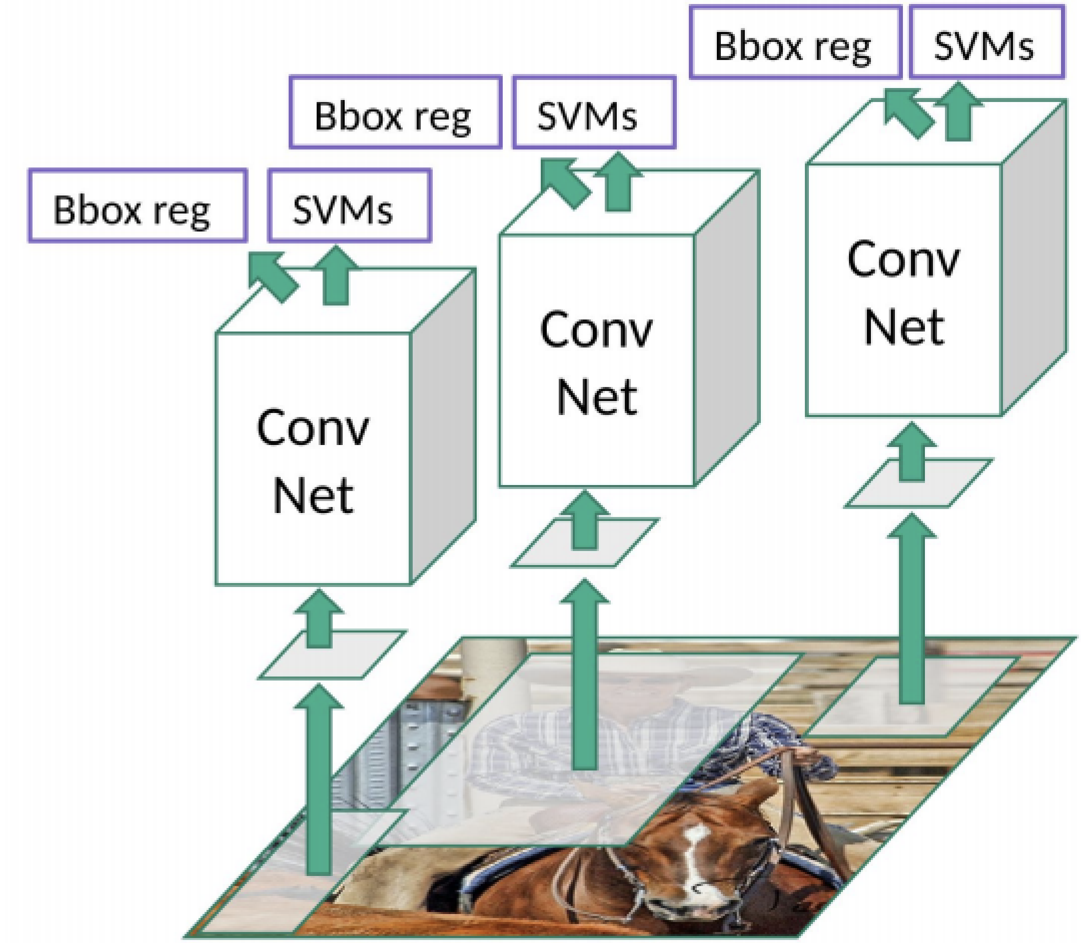
RCNN

- Initialize small areas
- Greedy algorithm merges regions
- Finally selected 2000 possible regions


R-CNN: *Regions with CNN features*
warped region
CNN
aeroplane? no.
person? yes.
tvmonitor? no.
1. Input image
2. Extract region proposals (~2k)
3. Compute CNN features
4. Classify regions

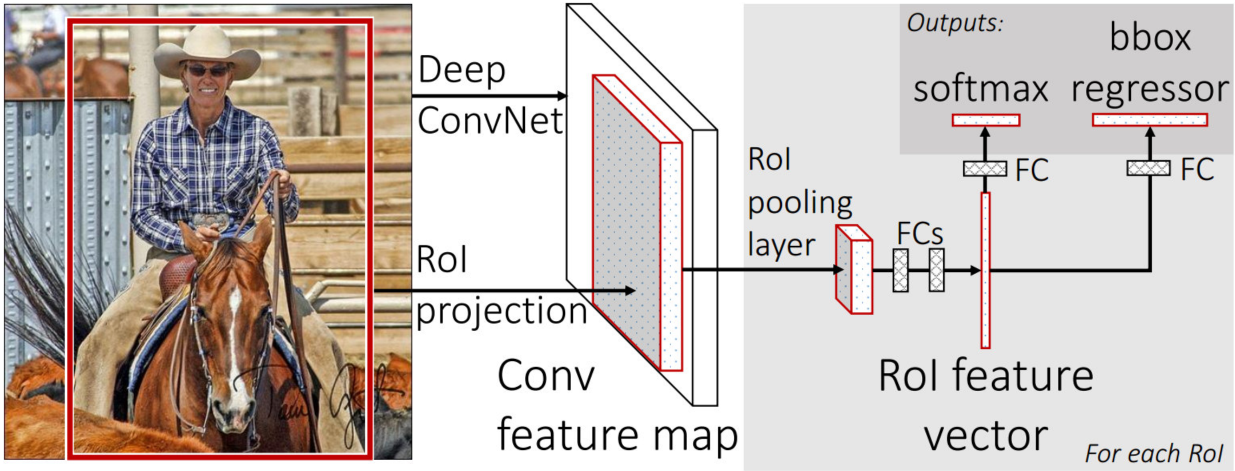
RCNN

- CNN: In addition to object recognition, it also recommends to adjust the area



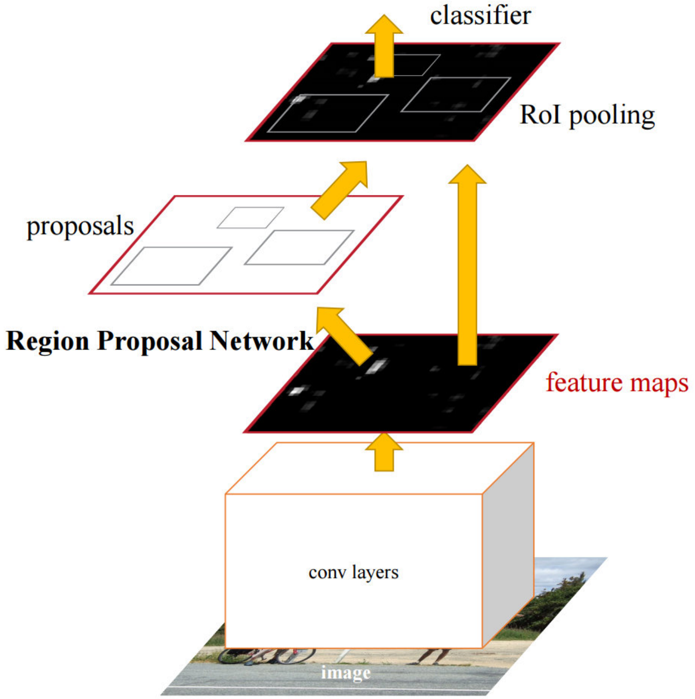
Fast R-CNN

- R-CNN: CNN on every area. Totally 2000 areas
- Improvement
    - CNN once for all images
    - Select the possible areas on the obtained feature map
- Dozens of times faster



Faster R-CNN

- Remove the time-consuming work of selective search for possible areas, use another network to predict areas where objects may appear
- 10 times faster

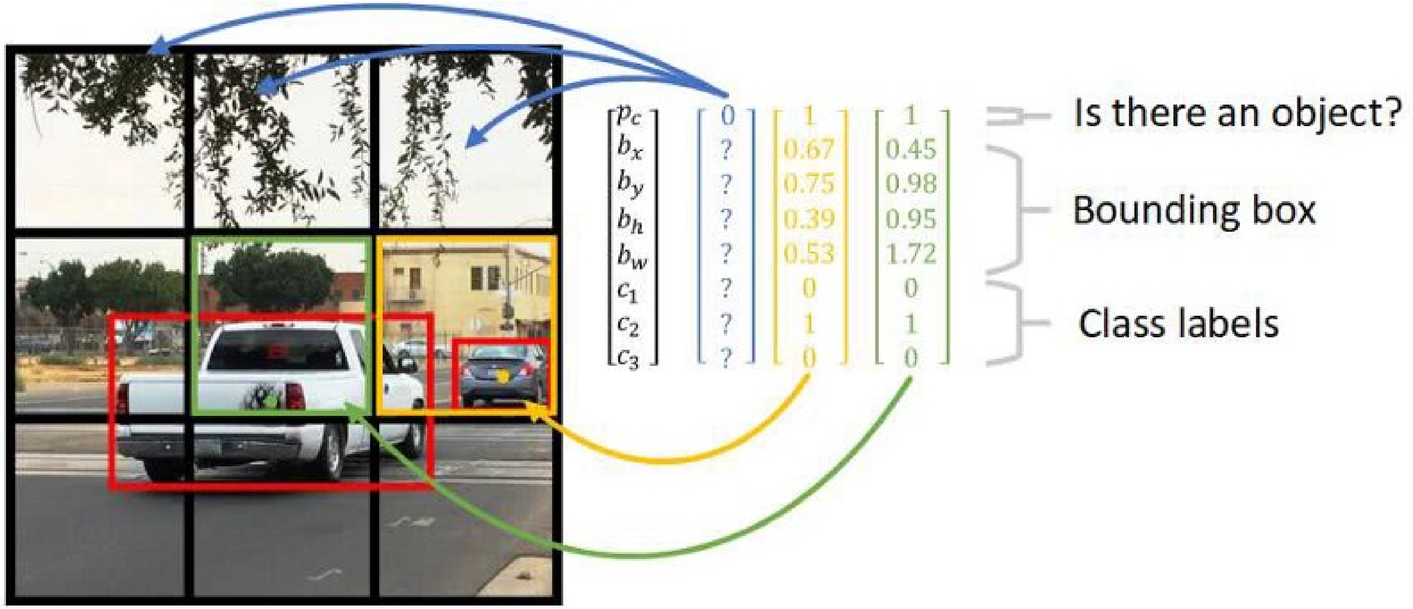


Single-Stage Detector

Identify the target without finding the area

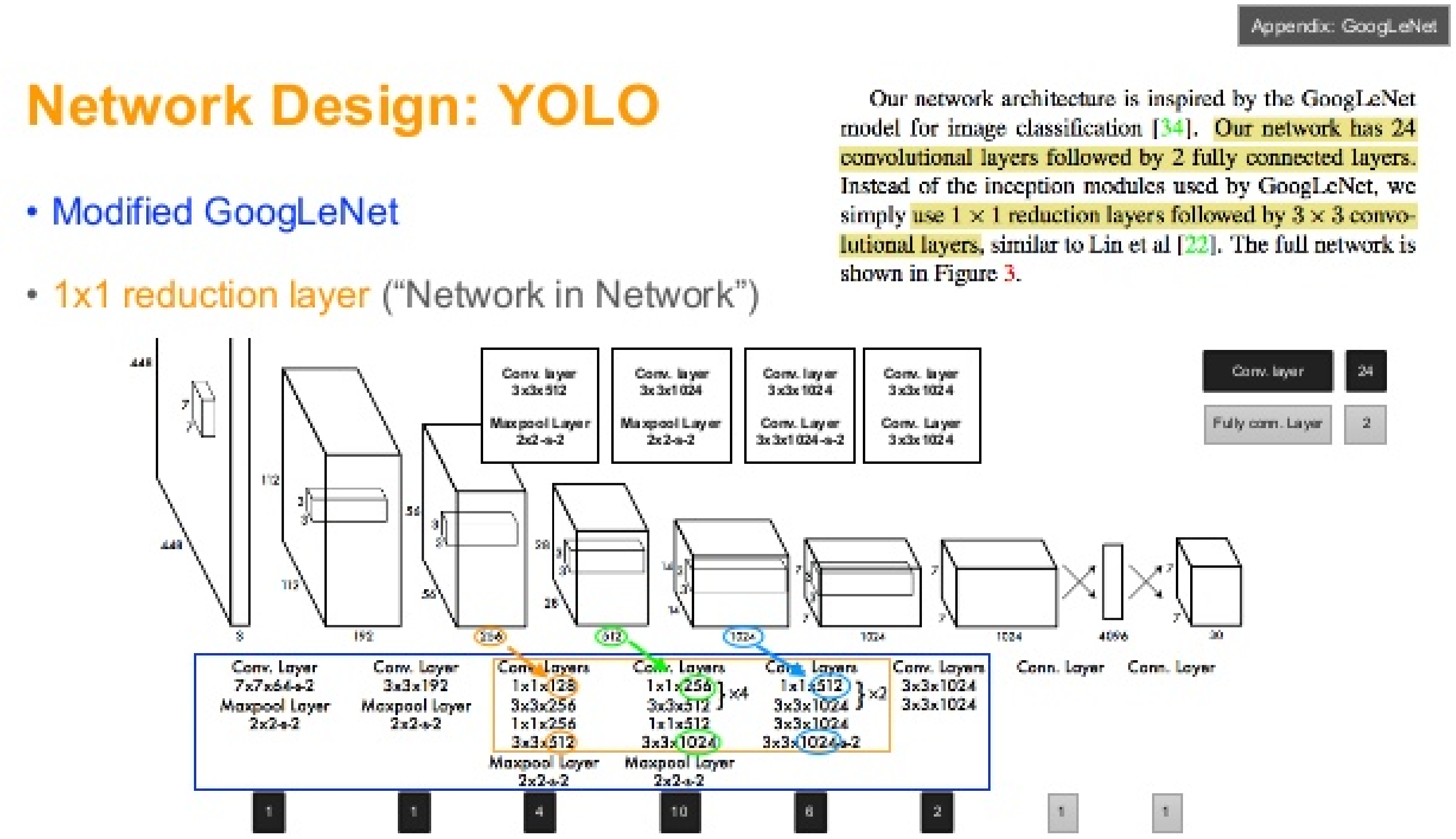
YOLO

- You Only Look Once， 2015
- Image divided into small blocks. Multiple possible object areas selected for each block.
- For each region, CNN gives its offset recommendation and object type judgment



YOLO

- Network: GoogleNet
- Faster, no problem at 45 frames per second



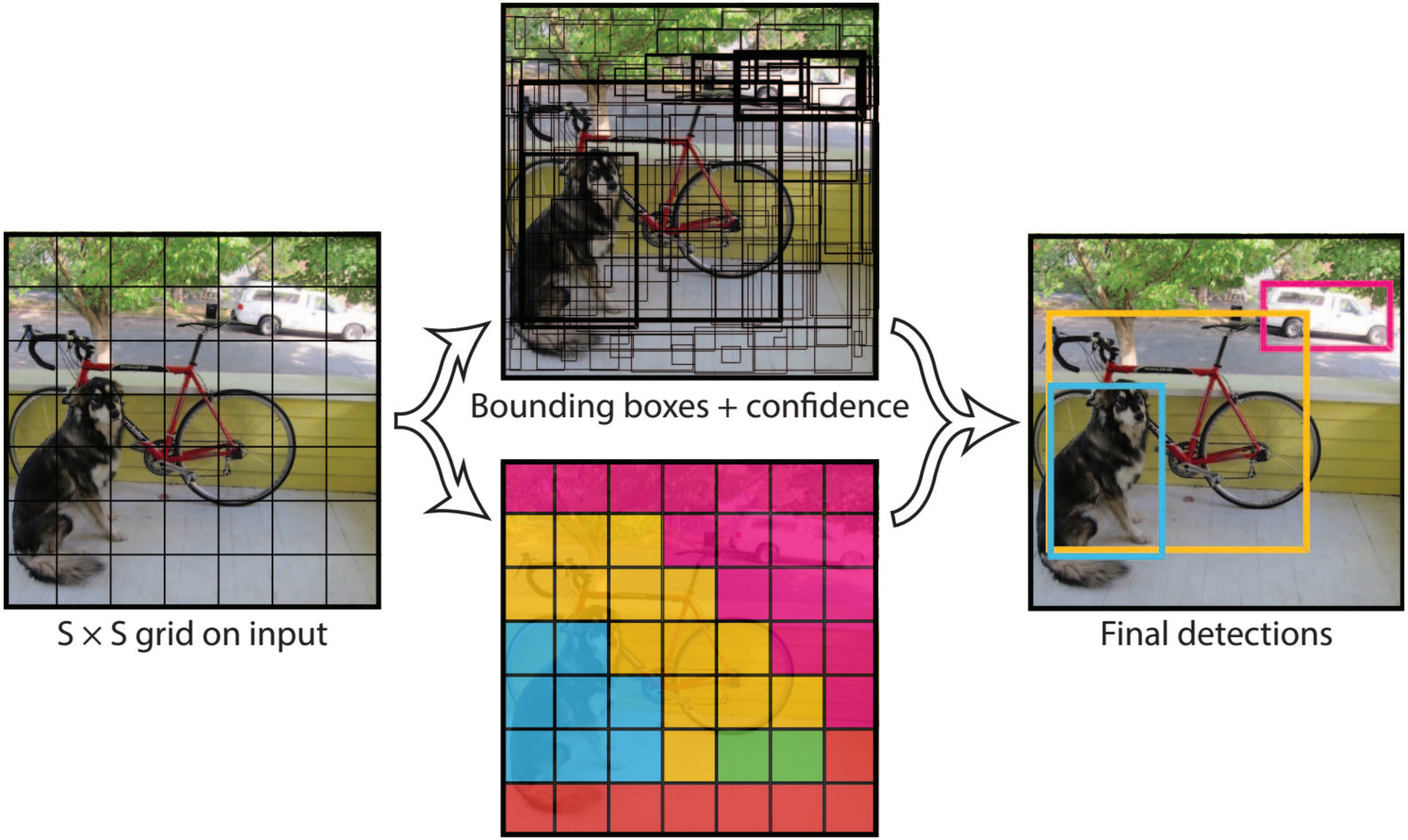
Appendix: GoogLeNet

Network Design: YOLO

- Modified GoogLeNet
- 1x1 reduction layer ("Network in Network")

Our network architecture is inspired by the GoogLeNet model for image classification [34]. Our network has 24 convolutional layers followed by 2 fully connected layers. Instead of the inception modules used by GoogLeNet, we simply use 1 × 1 reduction layers followed by 3 × 3 convolutional layers, similar to Lin et al [22]. The full network is shown in Figure 3.

Results



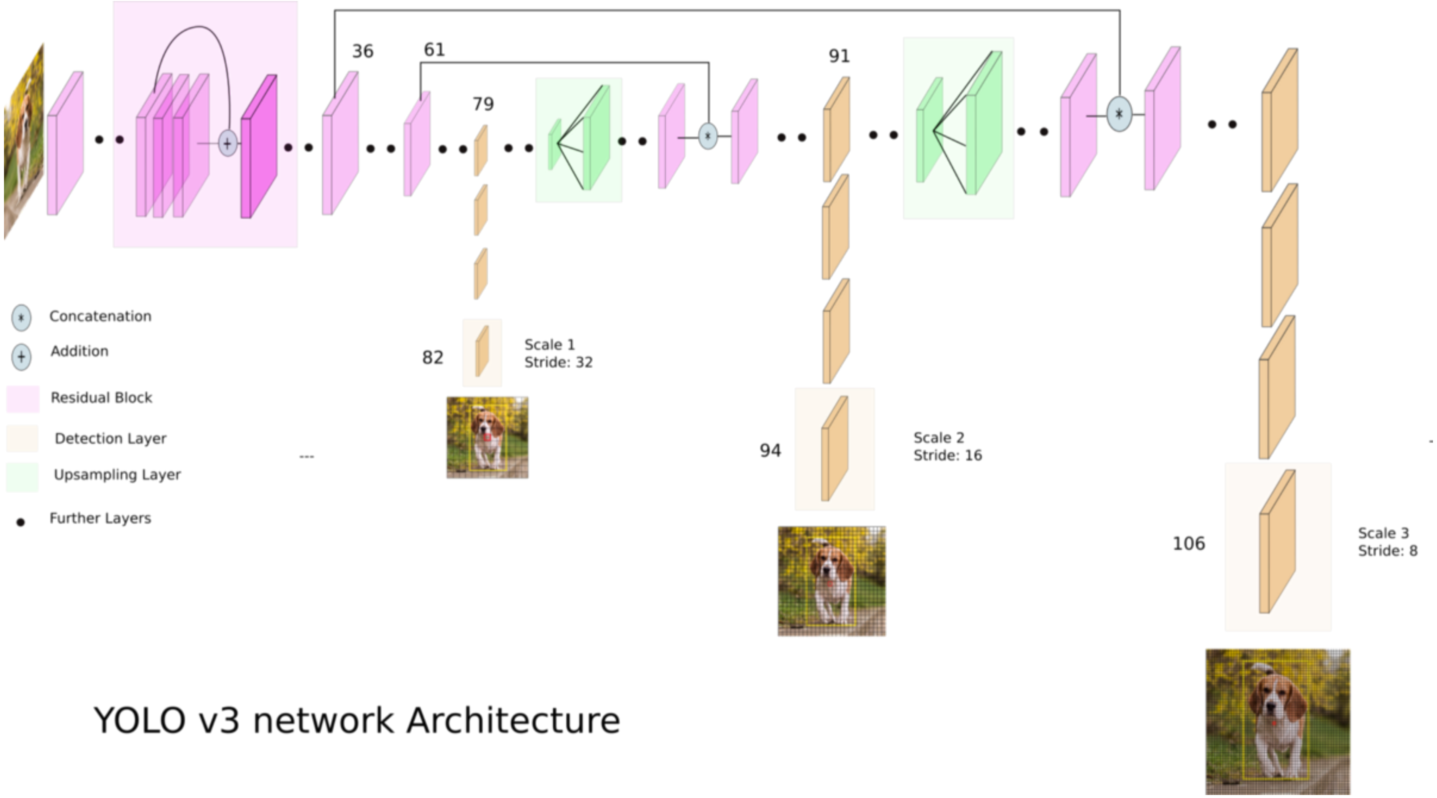
S × S grid on input

Bounding boxes + confidence

Class probability map

Final detections

YOLO v3

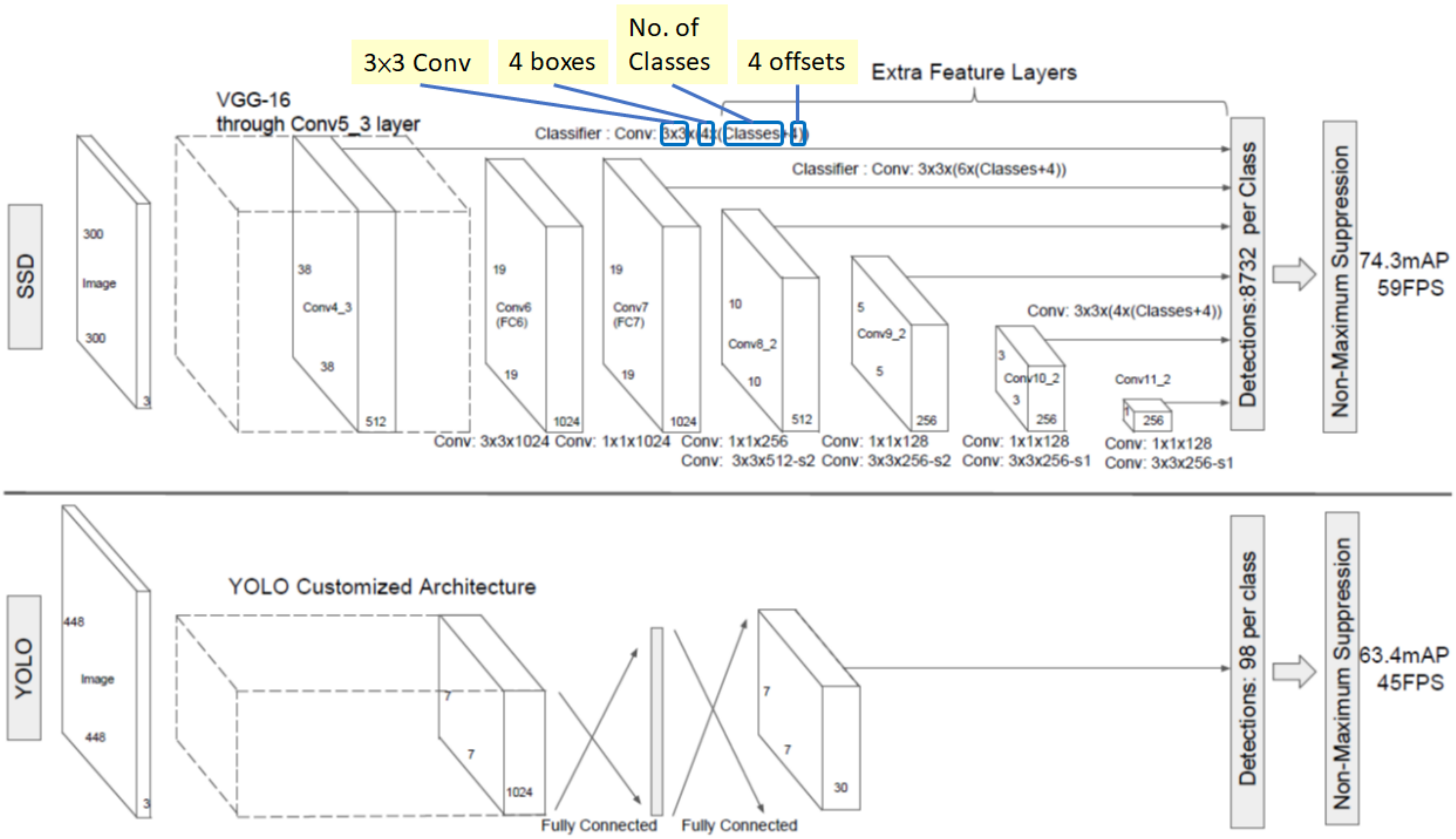


YOLO v3 network Architecture

- More acurrate
- Disadvantages: small object recognition is difficult, such as bird swarms

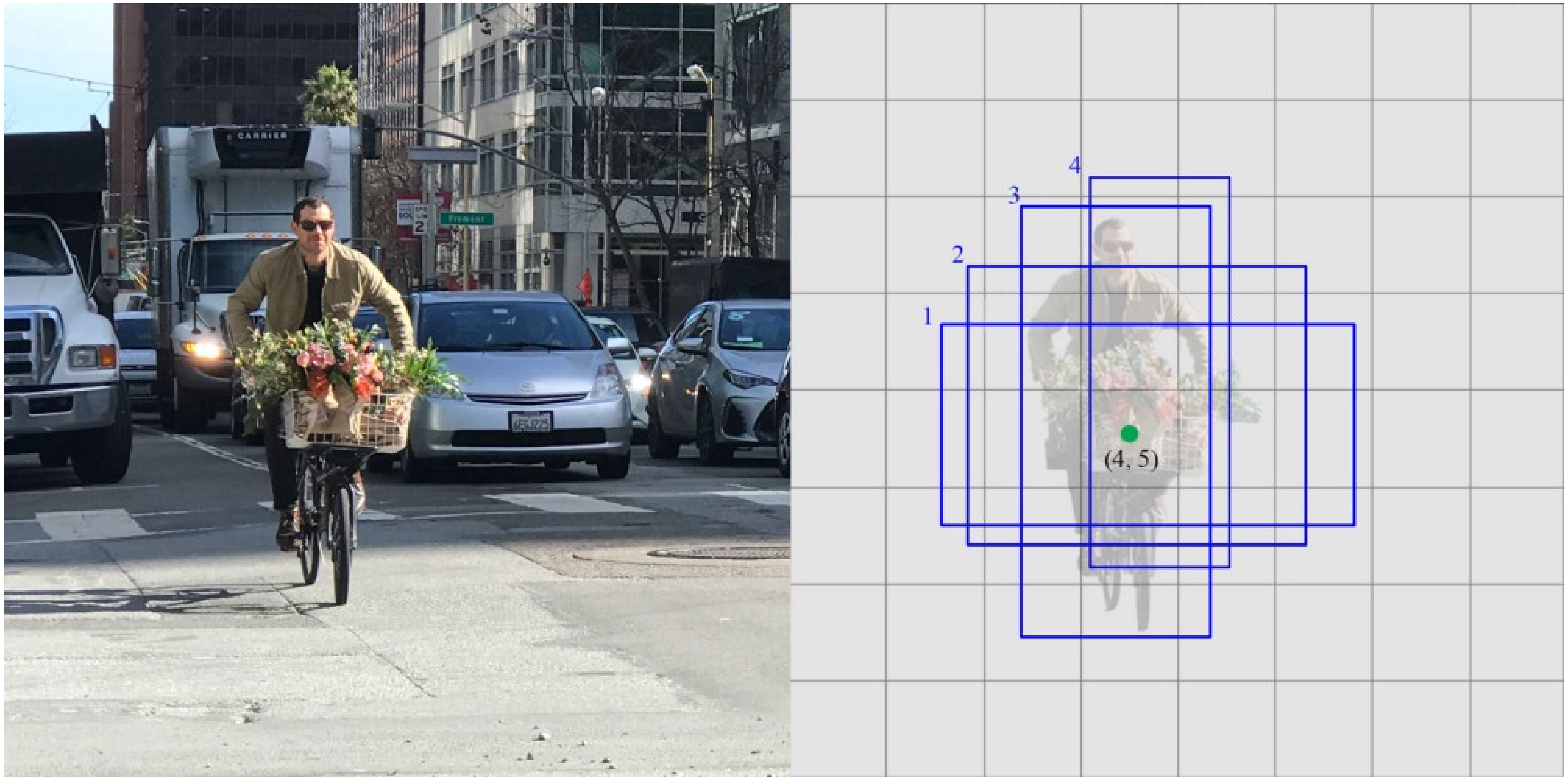
SSD

- Single Shot MultiBox Detector
- 2016, ECCV



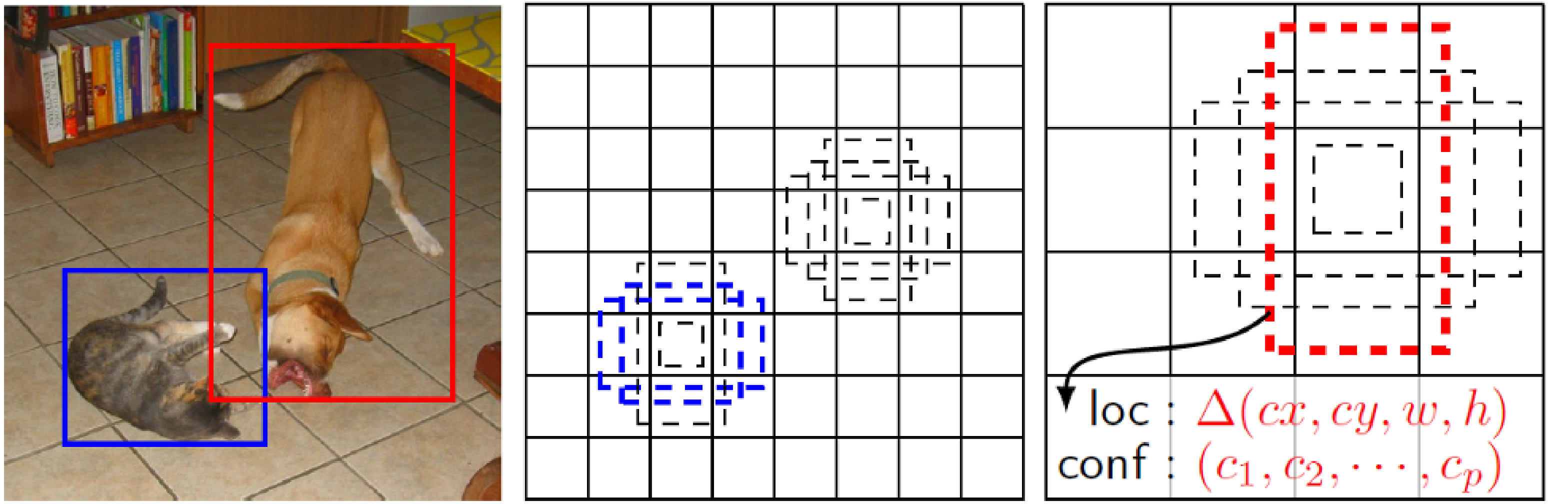
Default Object Box Shape

- Cars, people have specific shapes
- Manual selection of initial four default boxes



Multi-Scale Feature Map

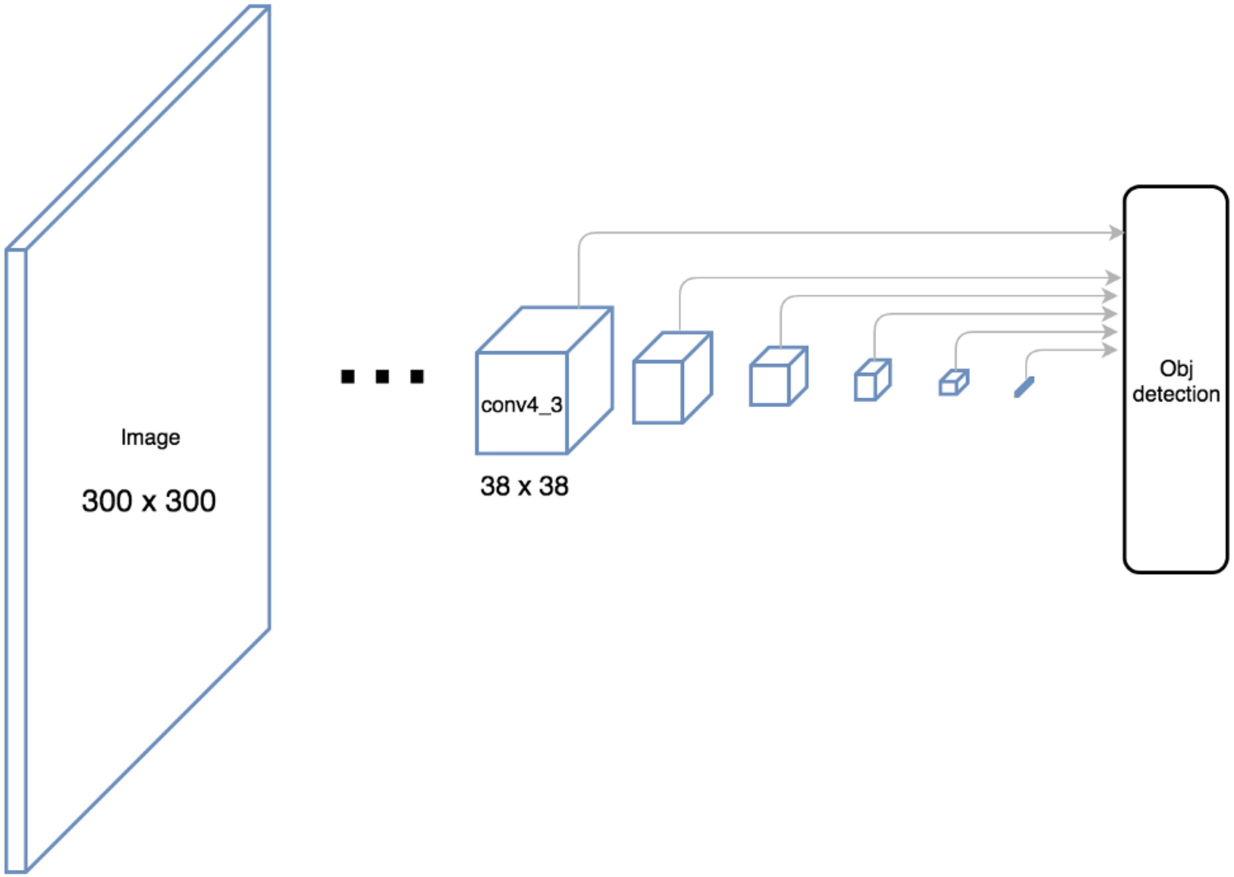
- Use blocks of different scales to detect objects of different scales


loc : $\Delta(cx, cy, w, h)$
conf : $(c_1, c_2, \cdots, c_p)$

(a) Image with GT boxes (b) 8×8 feature map (c) 4×4 feature map

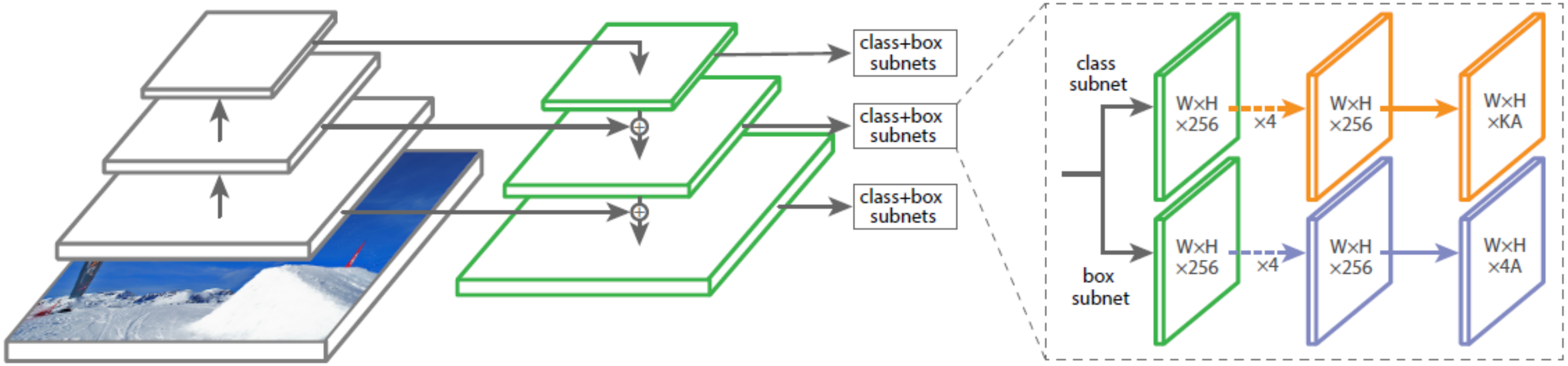
Multi-Resolution CNN

- Add 6 CNNs after VGG with different resolutions
- High-resolution CNN helps identify small targets



RetinaNet

- 2017 ICCV
- Backbone network: ResNet + Feature Pyramid Net （FPN)
  - Different levels of pyramid have different resolution
- Task network
  - Objection recognition + Bounding box discovery



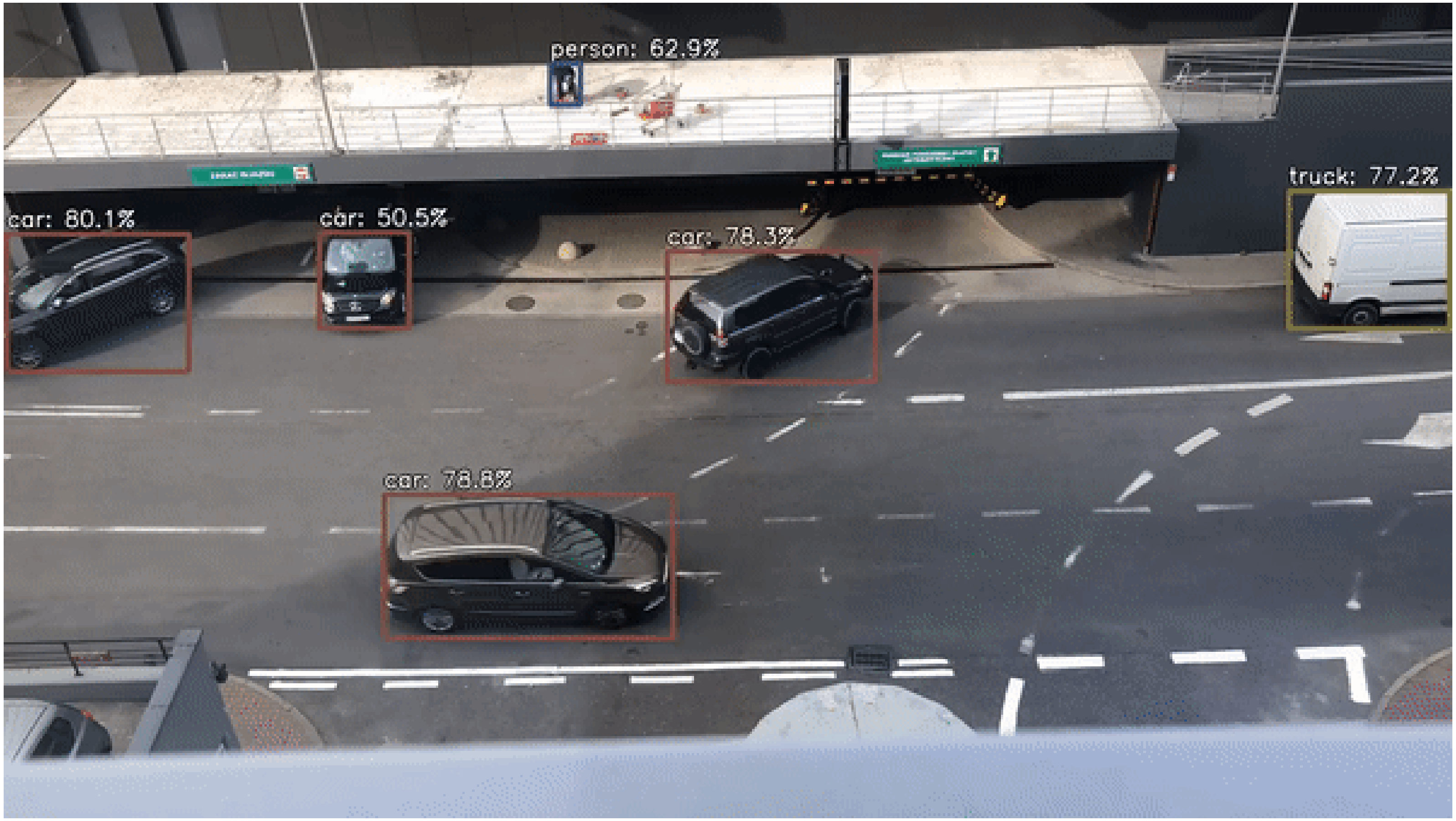
(a) ResNet (b) feature pyramid net (c) class subnet (top) (d) box subnet (bottom)

Focal Loss

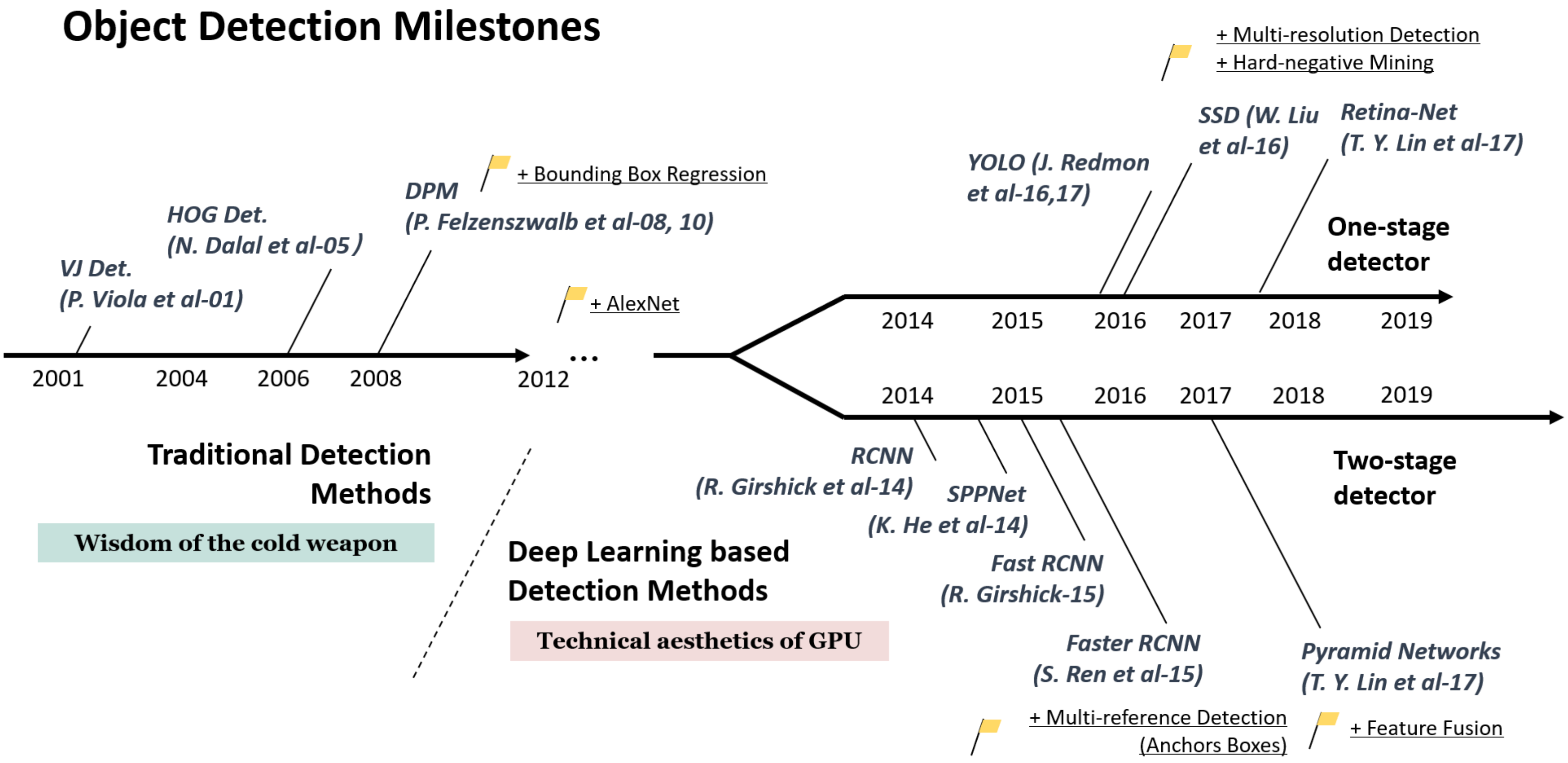
- The most important contribution is this Loss
- Use this Loss to replace cross entropy, greatly improving accuracy
- Reduce the weight of those easily identifiable classes in Loss and increase those that are difficult to classify
- : accurate prediction probability

$$(\quad) = -(1 - \quad) \qquad (\quad)$$

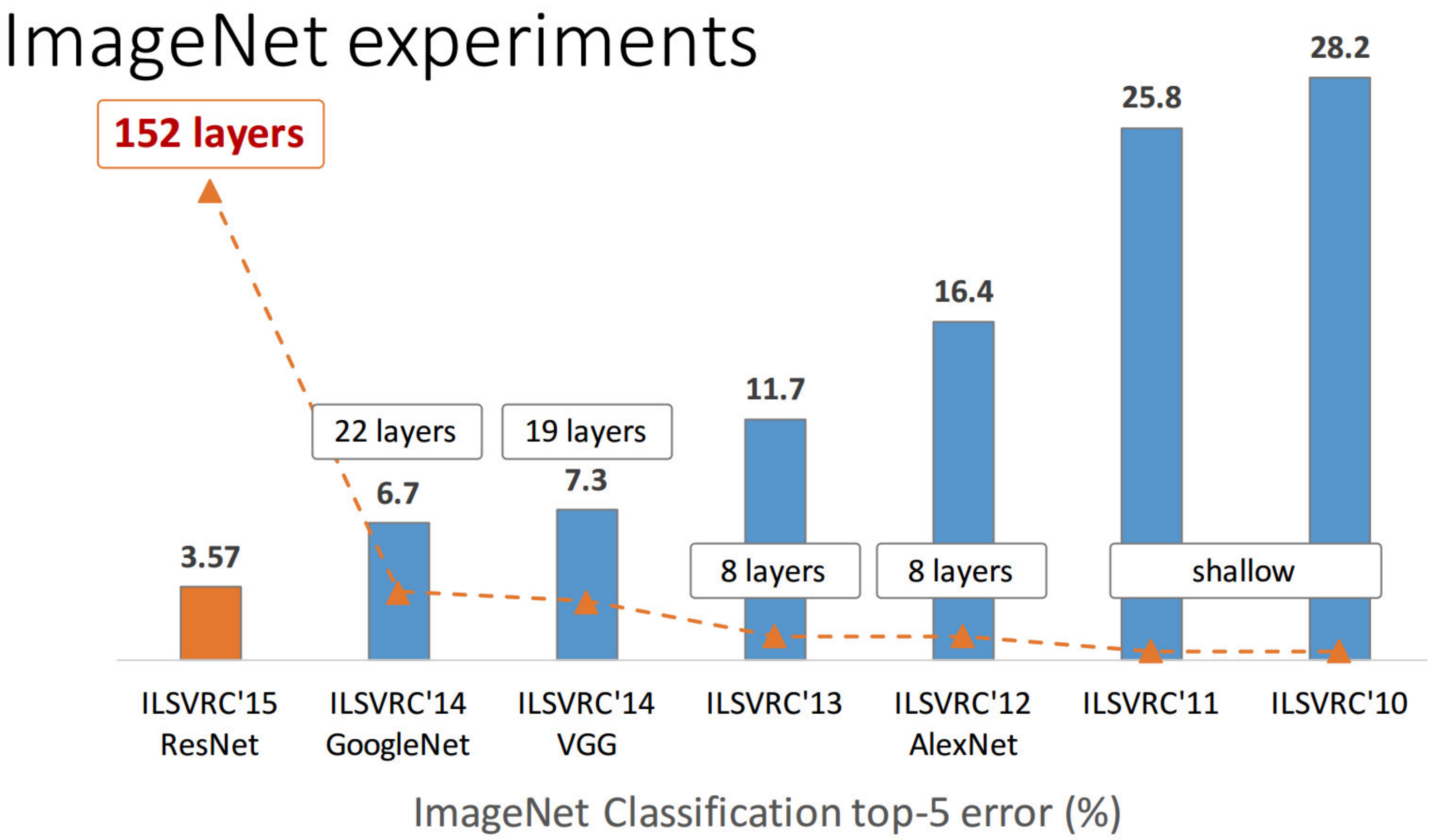
RetinaNet Results

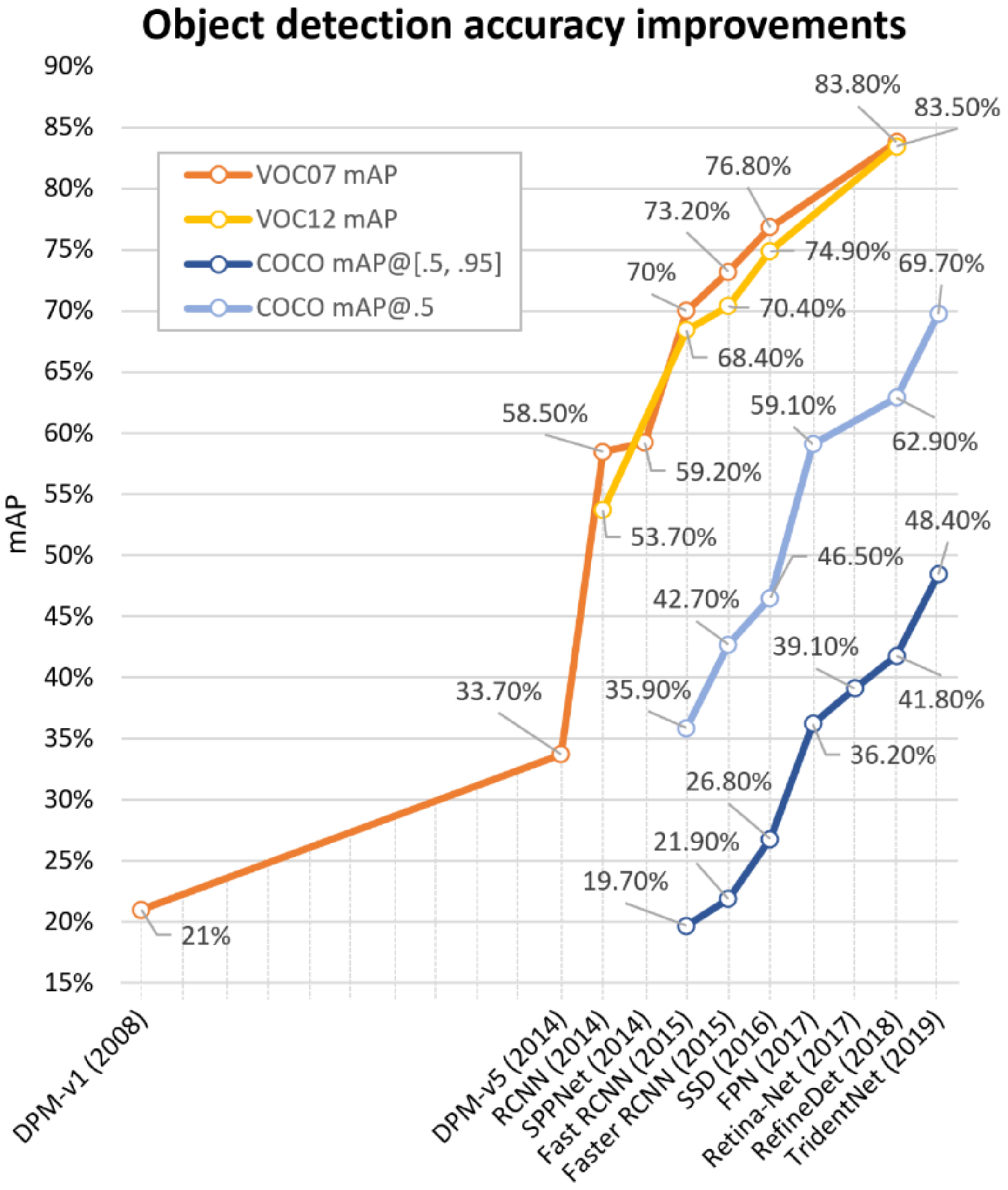


Summary


Object Detection Milestones
VJ Det.
(P. Viola et al-01)
HOG Det.
(N. Dalal et al-05)
DPM
(P. Felzenszwalb et al-08, 10)
+ Bounding Box Regression
+ AlexNet
2001
2004
2006
2008
2012
...
Traditional Detection Methods
Wisdom of the cold weapon
Deep Learning based Detection Methods
Technical aesthetics of GPU
+ Multi-resolution Detection
+ Hard-negative Mining
YOLO (J. Redmon et al-16,17)
SSD (W. Liu et al-16)
Retina-Net (T. Y. Lin et al-17)
One-stage detector
2014
2015
2016
2017
2018
2019
2014
2015
2016
2017
2018
2019
RCNN
(R. Girshick et al-14)
SPPNet
(K. He et al-14)
Fast RCNN
(R. Girshick-15)
Faster RCNN
(S. Ren et al-15)
Pyramid Networks
(T. Y. Lin et al-17)
Two-stage detector
+ Multi-reference Detection (Anchors Boxes)
+ Feature Fusion

Performance

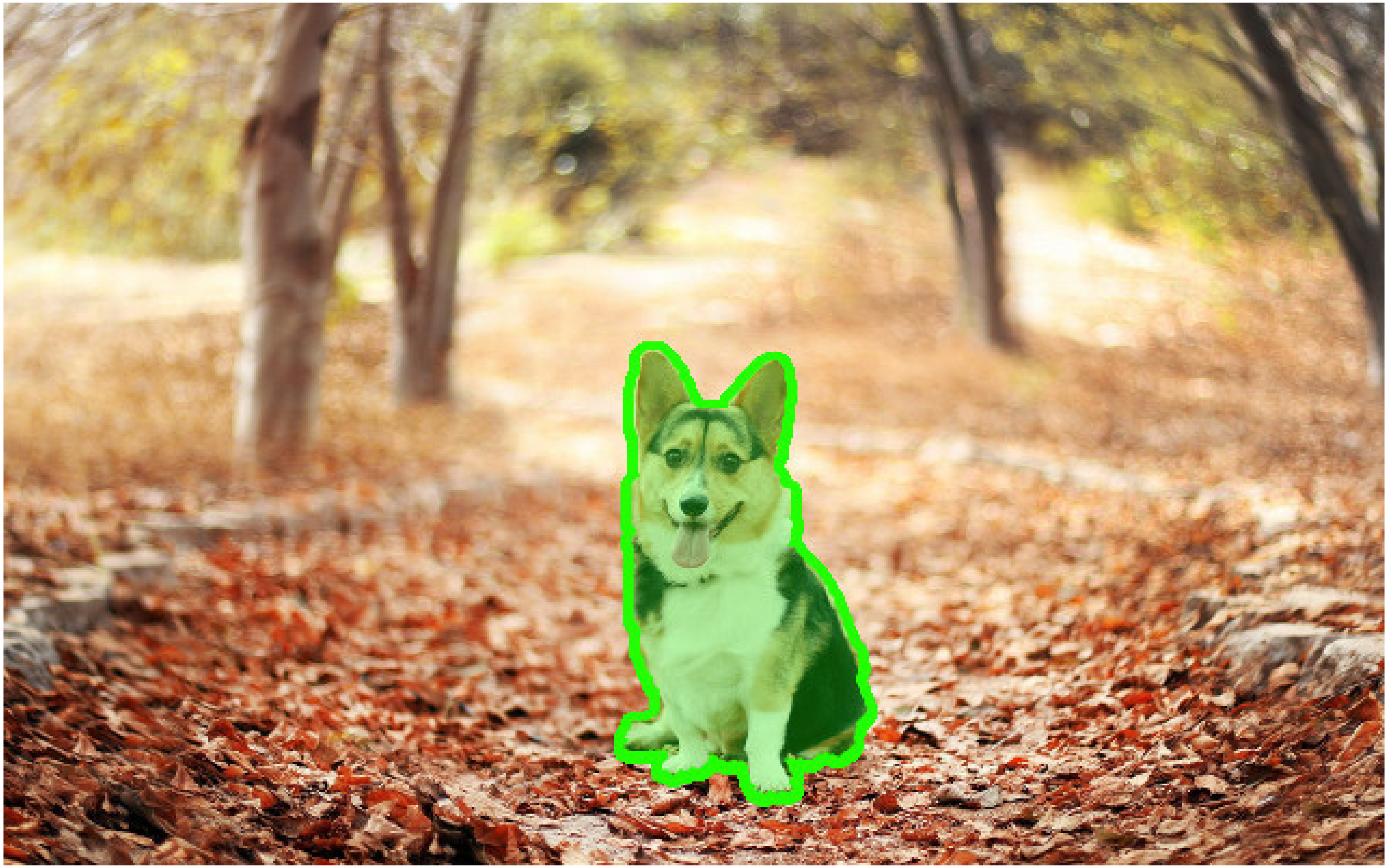



Object detection accuracy improvements
mAP
90%
85%
80%
75%
70%
65%
60%
55%
50%
45%
40%
35%
30%
25%
20%
15%
VOC07 mAP
VOC12 mAP
COCO mAP@[.5, .95]
COCO mAP@.5
21%
33.70%
58.50%
59.20%
53.70%
70%
68.40%
73.20%
70.40%
76.80%
74.90%
83.80%
83.50%
35.90%
42.70%
46.50%
59.10%
62.90%
69.70%
19.70%
21.90%
26.80%
36.20%
39.10%
41.80%
48.40%
DPM-v1 (2008)
DPM-v5 (2014)
RCNN (2014)
SPPNet (2014)
Fast RCNN (2015)
Faster RCNN (2015)
SSD (2016)
FPN (2017)
Retina-Net (2017)
RefineDet (2018)
TridentNet (2019)

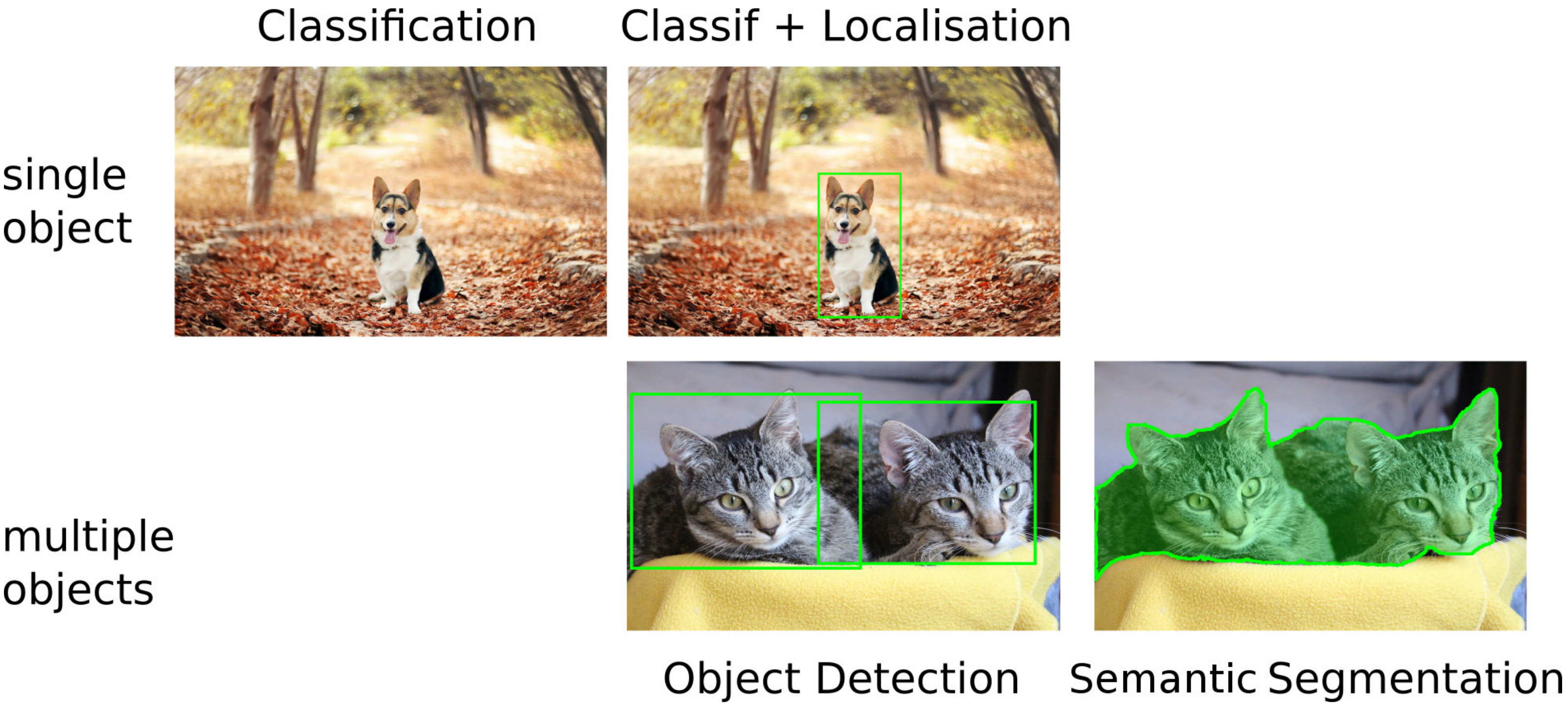
Object Segmentation

Extract the outline of an object from a image

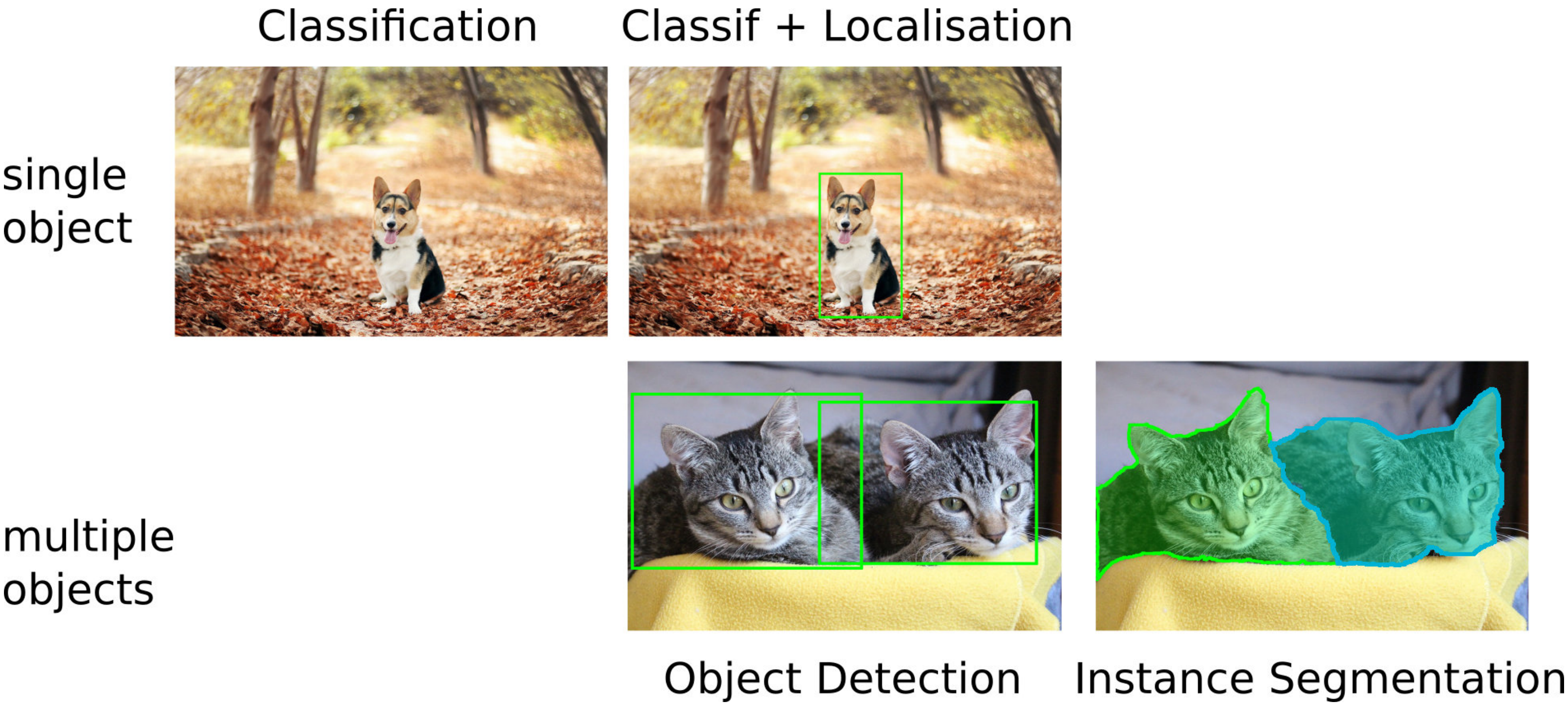
Image Segmentation



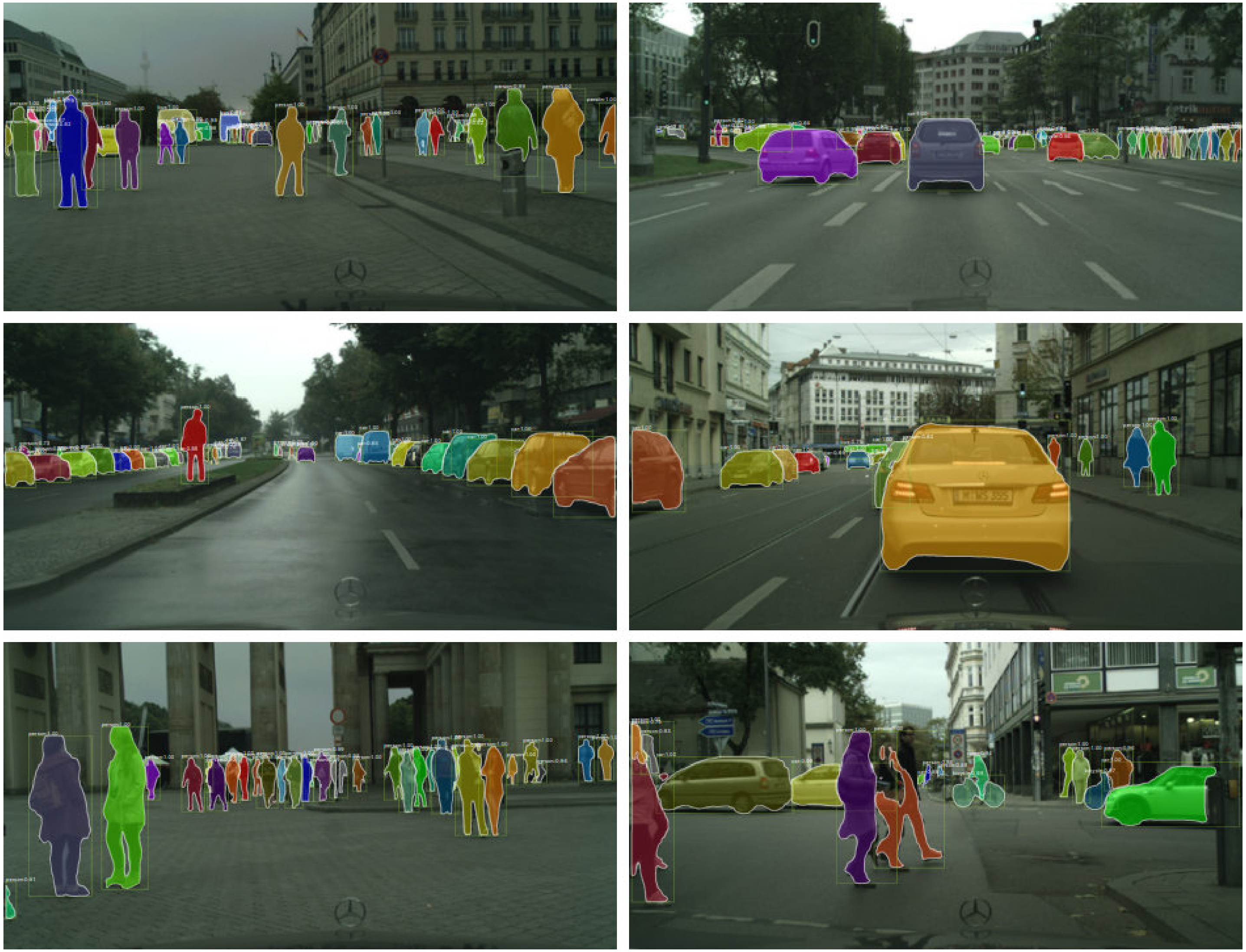
Semantic Segmentation



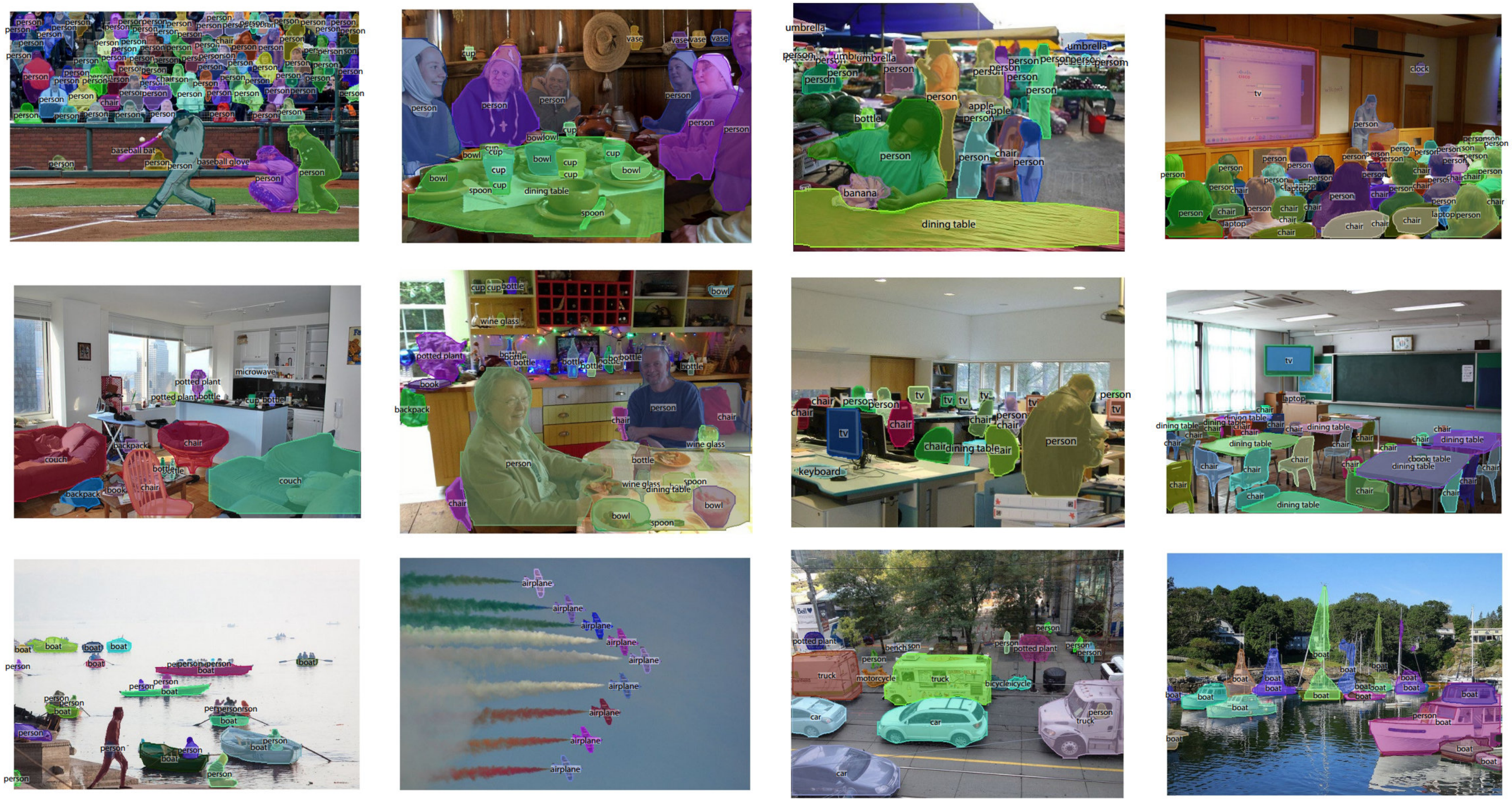
Instance Segmentation



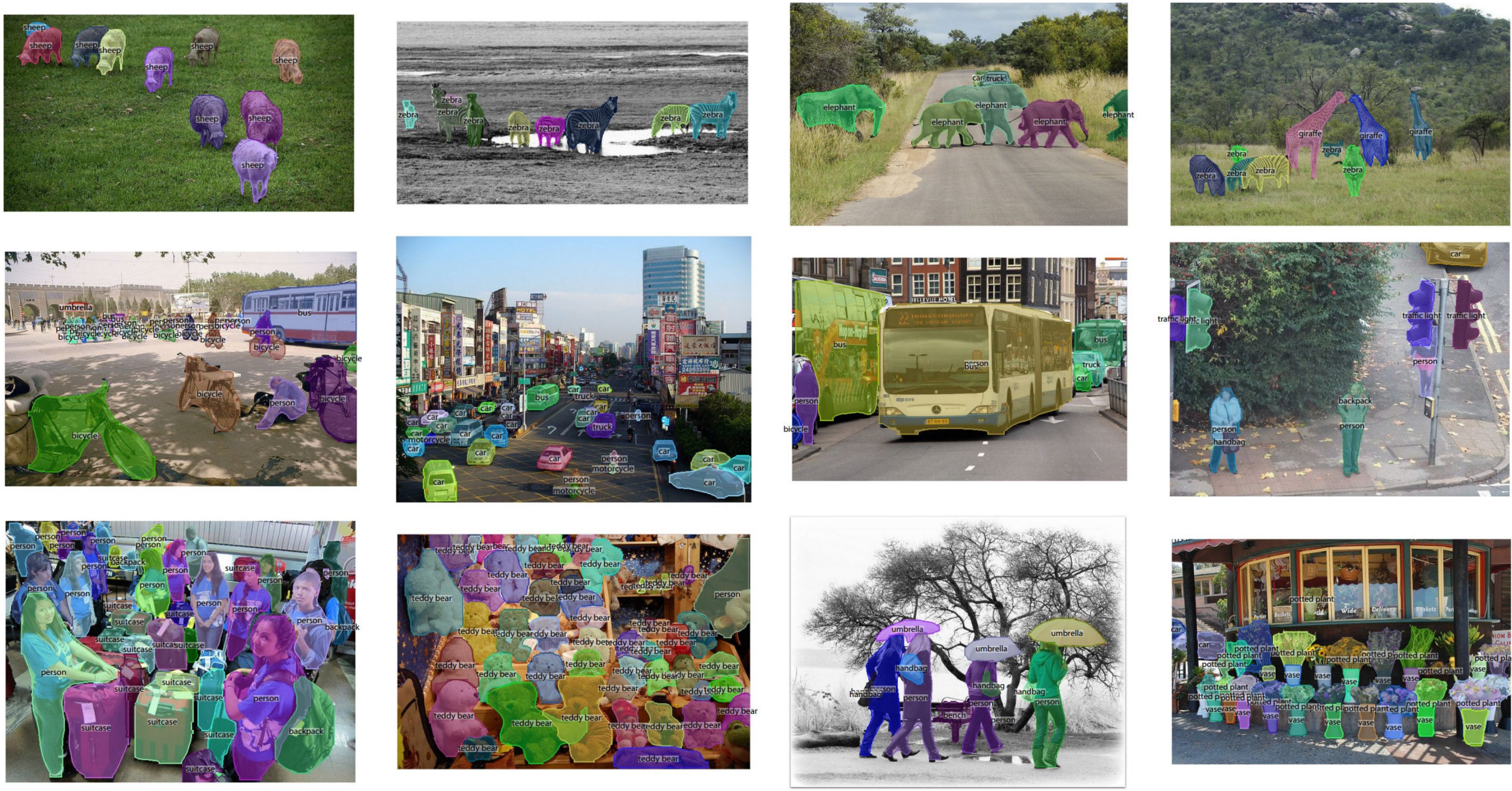
Instance Segmentation



Instance Segmentation



Instance Segmentation



Segmentation

1124 人在观看： [TSKS]我家的熊孩子.E84.180422 立即围观 >

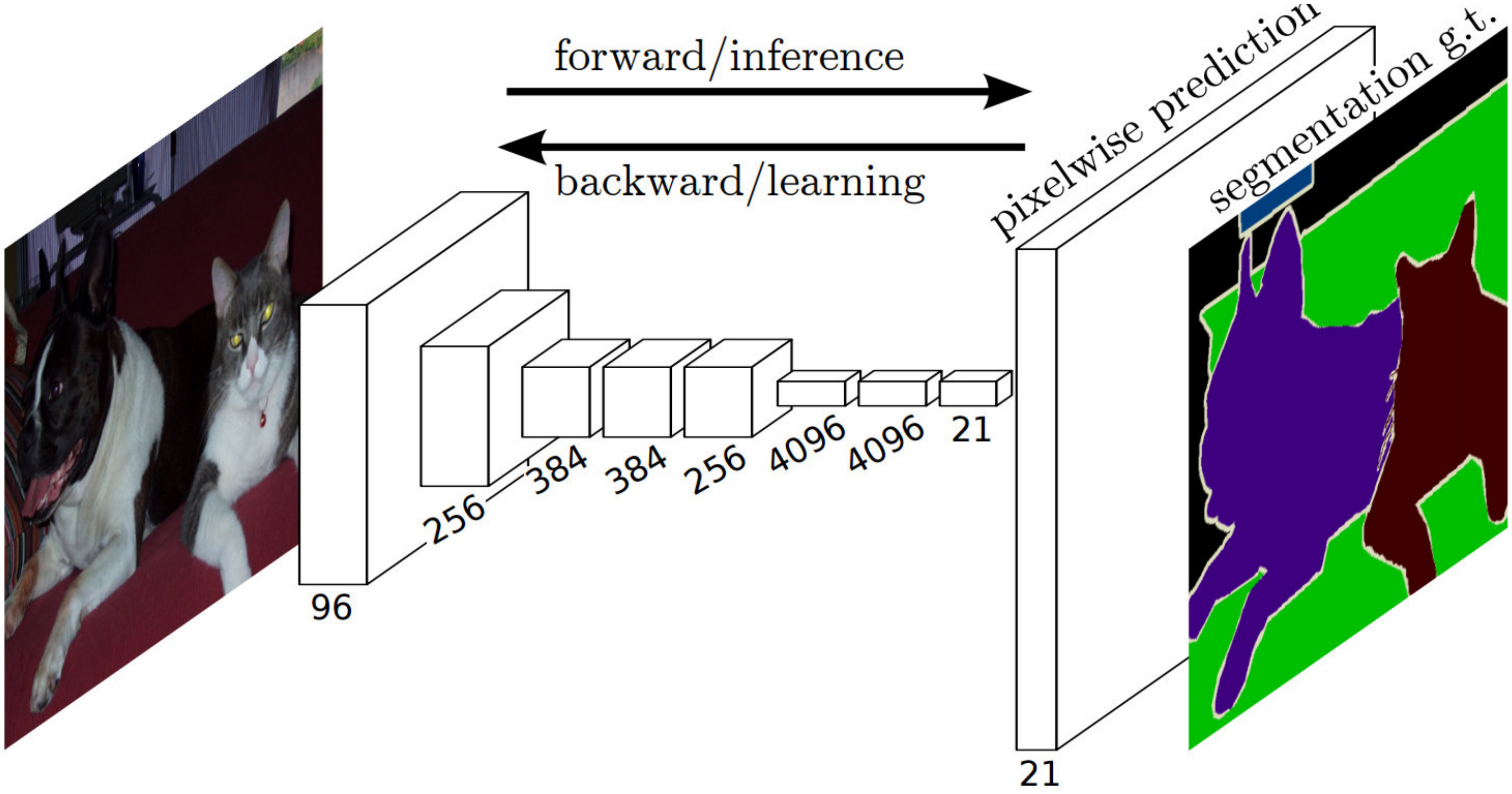
人工智能：图像分割

播放器初始化...[完成]
加载用户配置...[完成]
加载视频地址...[完成]
加载视频内容...

进入bilibili,一起发弹幕吐槽!

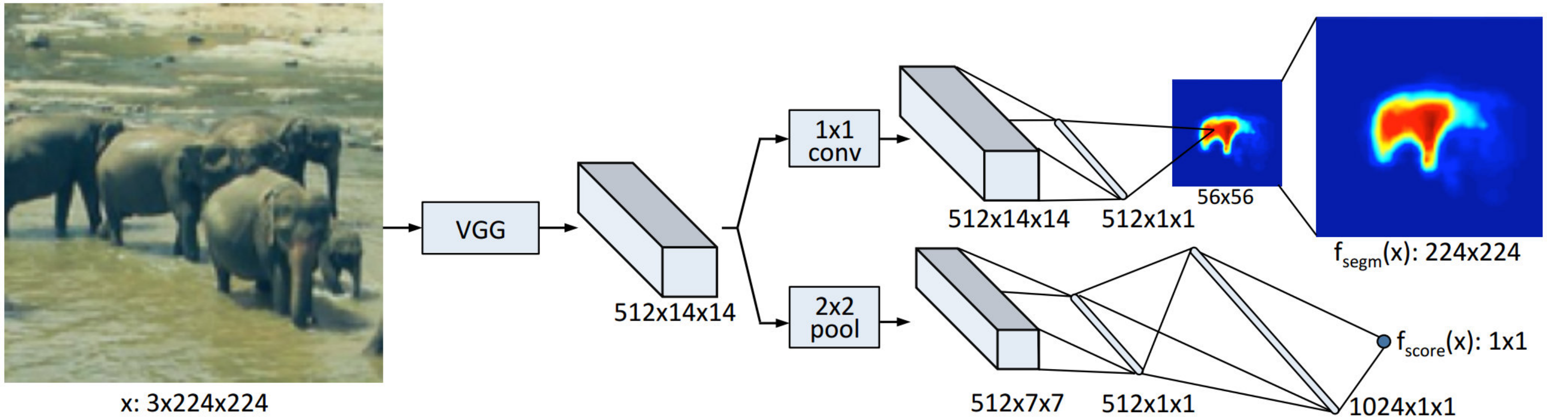
Instance Segmentation

Classify each pixel to get Mask



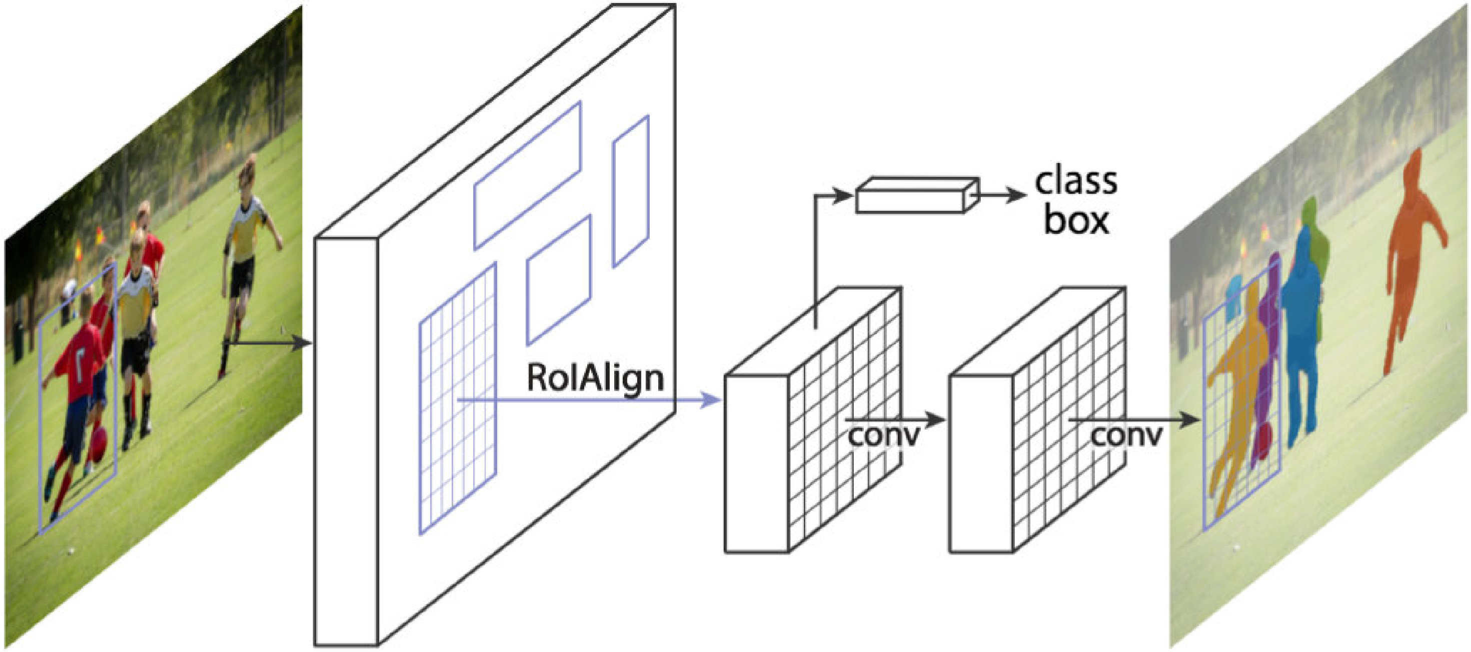
DeepMask

- Facebook， 2015 NIPS
- Two tasks after VGG
  - MASK
  - Object detection

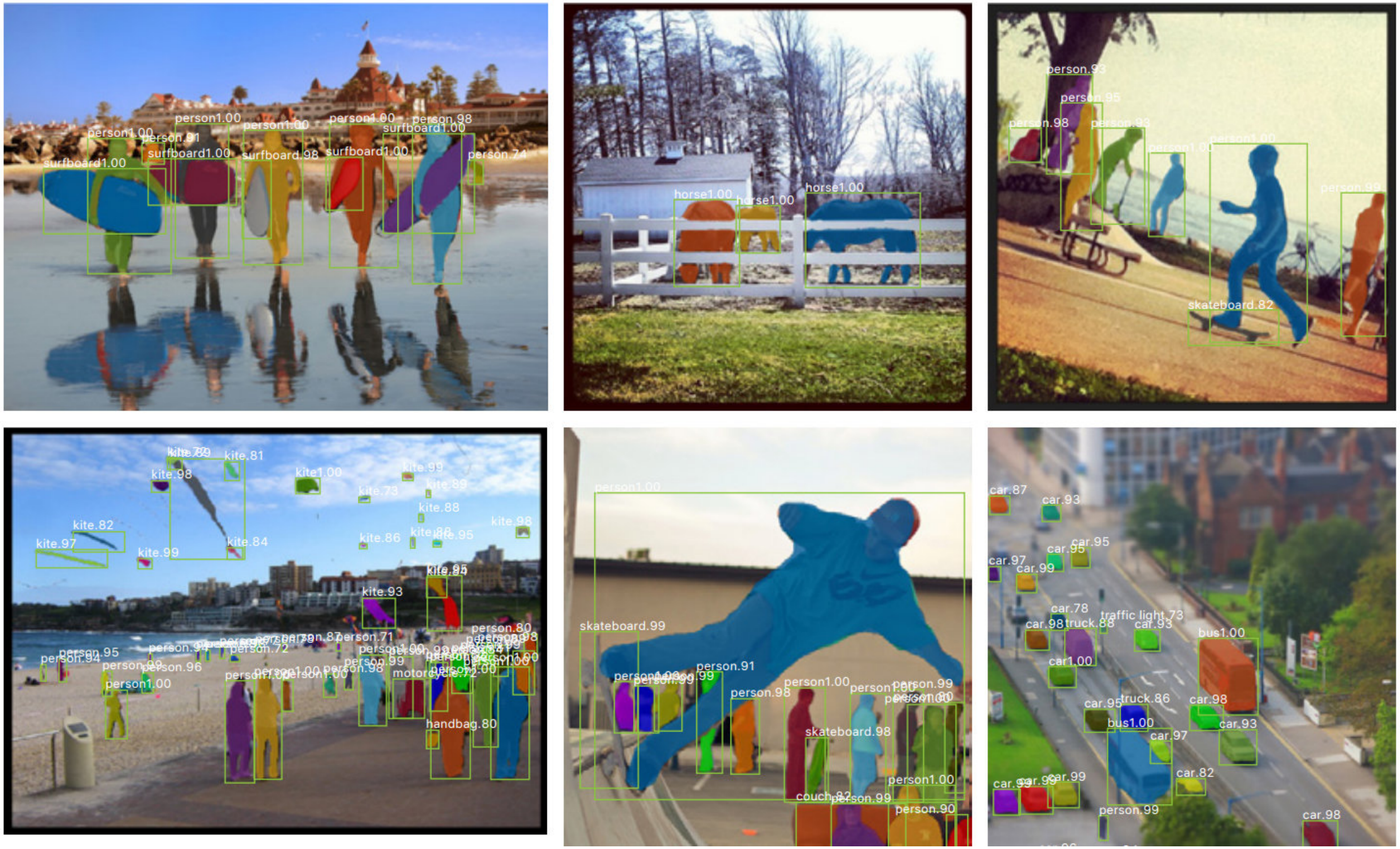


Mask RCNN

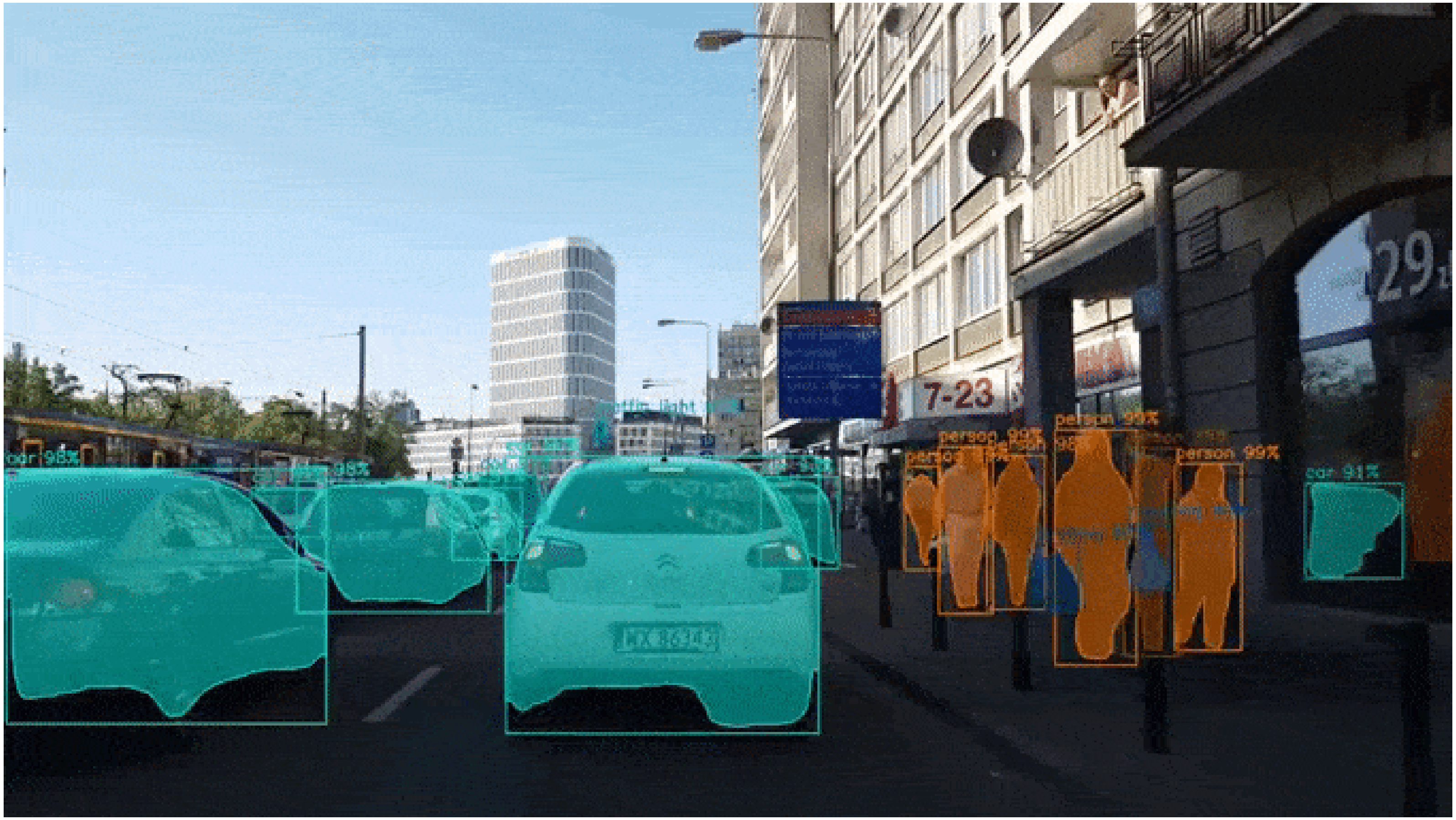
- 2017， Based on FPN (pyramid network) and ResNet



Mask RCNN Results

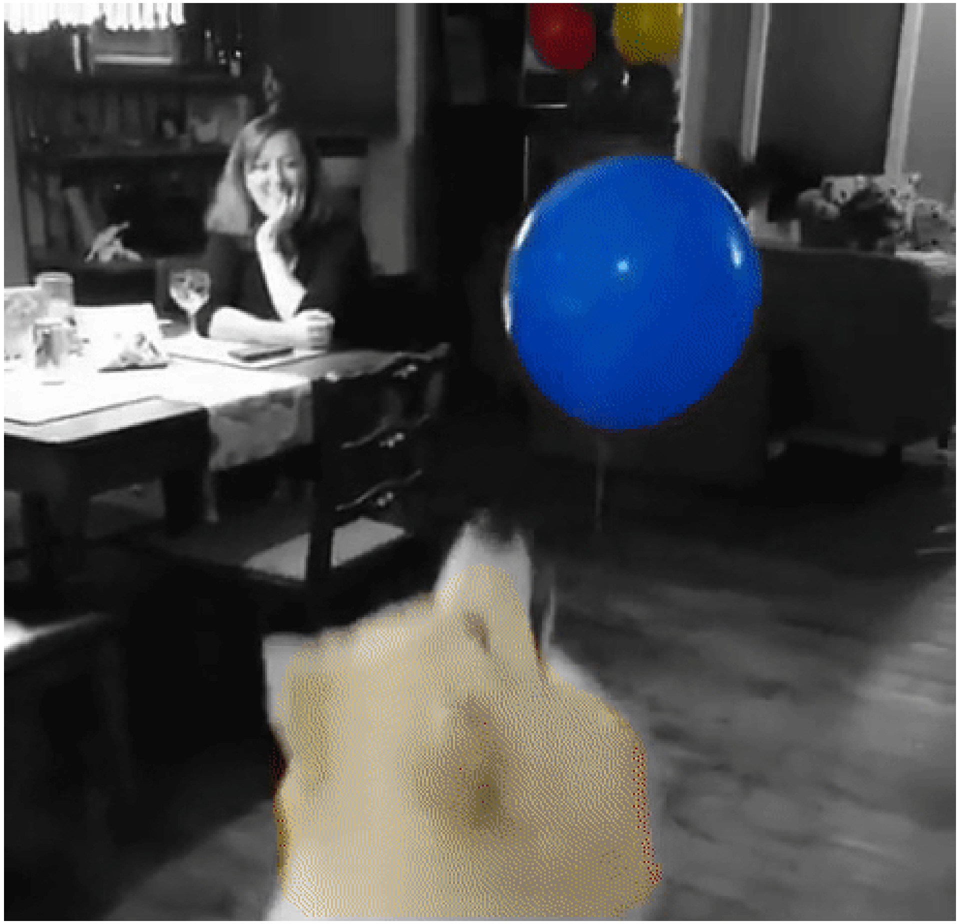


Mask RCNN Results

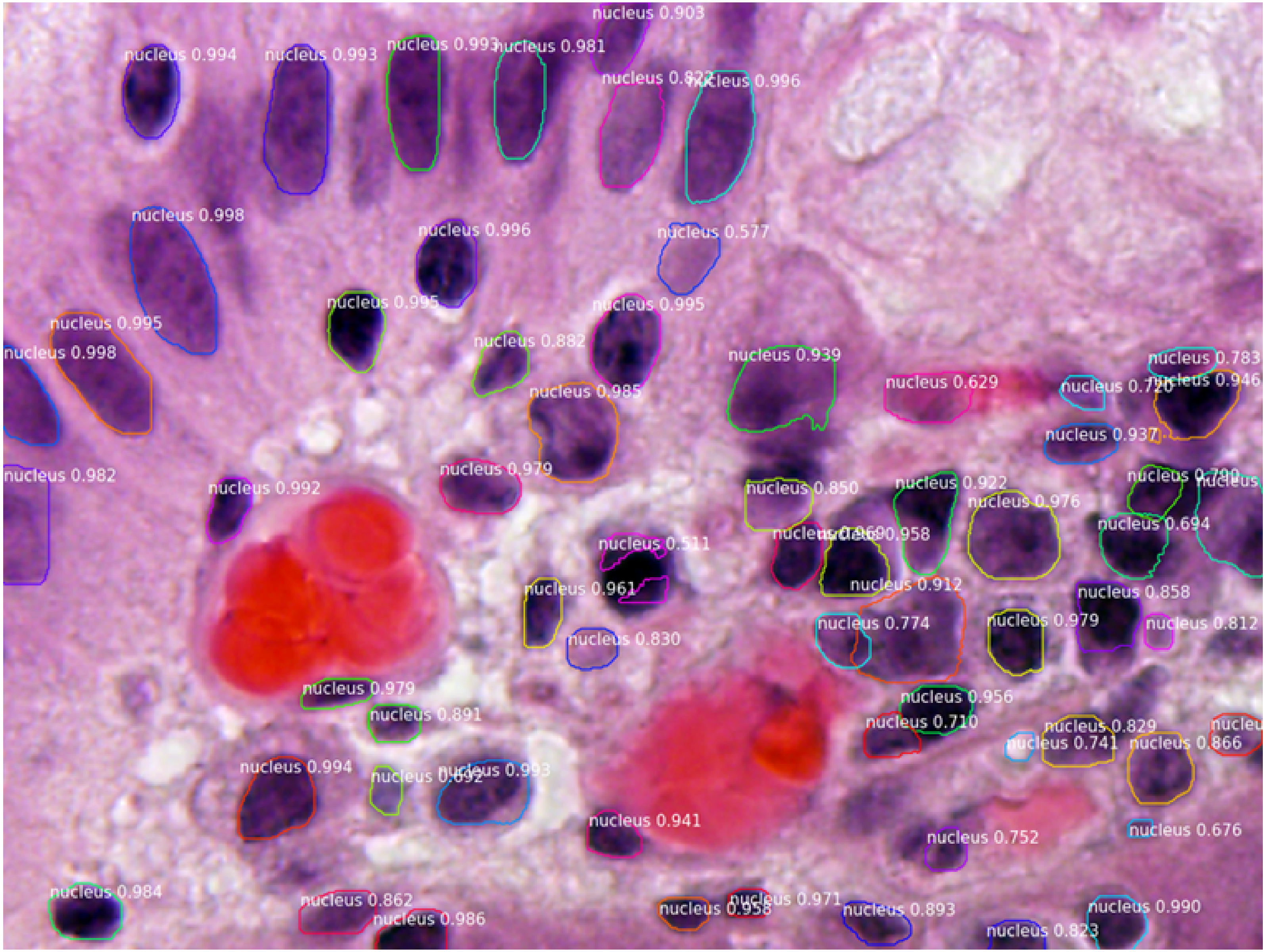


Application

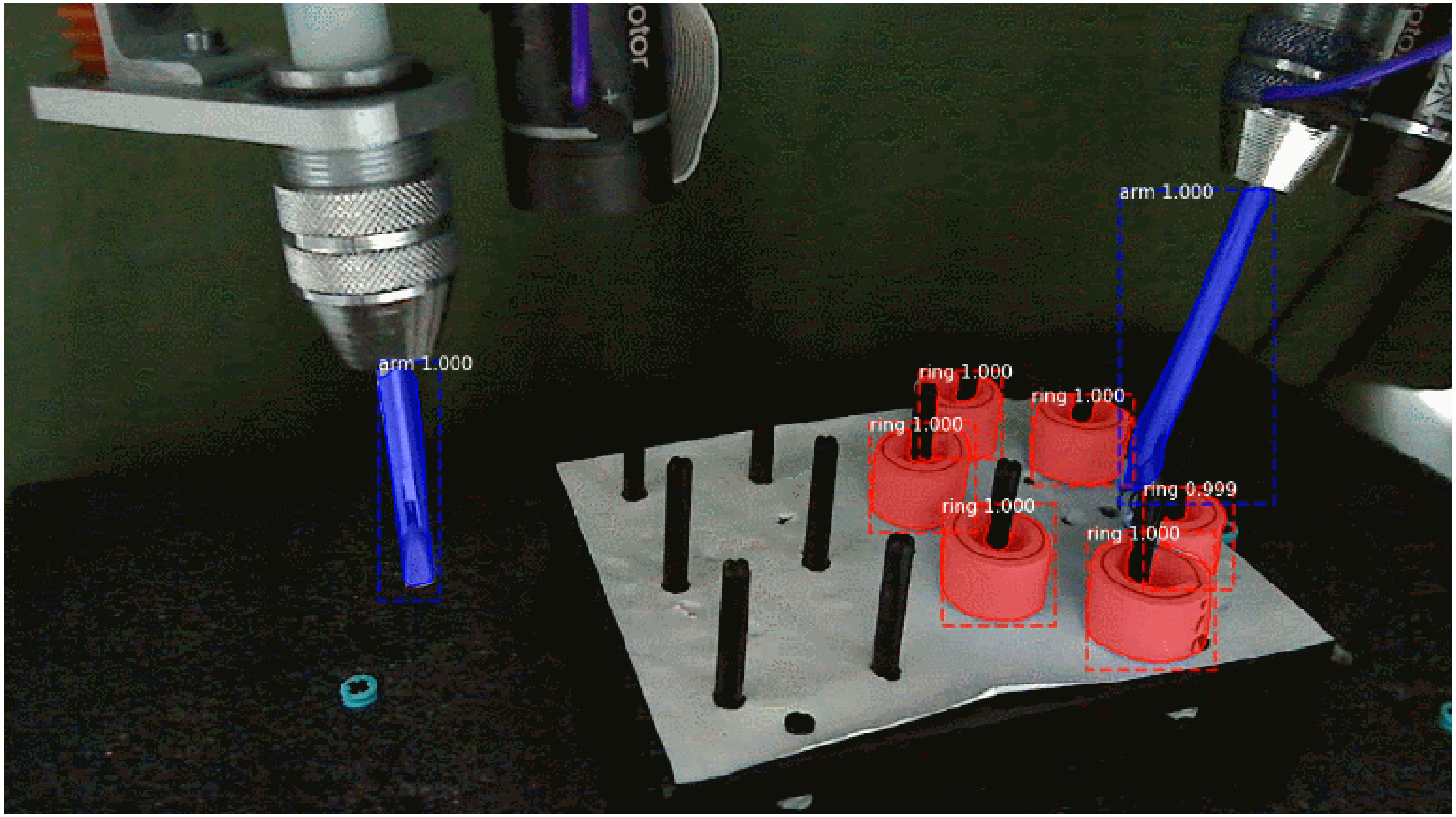
Tracking and Coloring Object



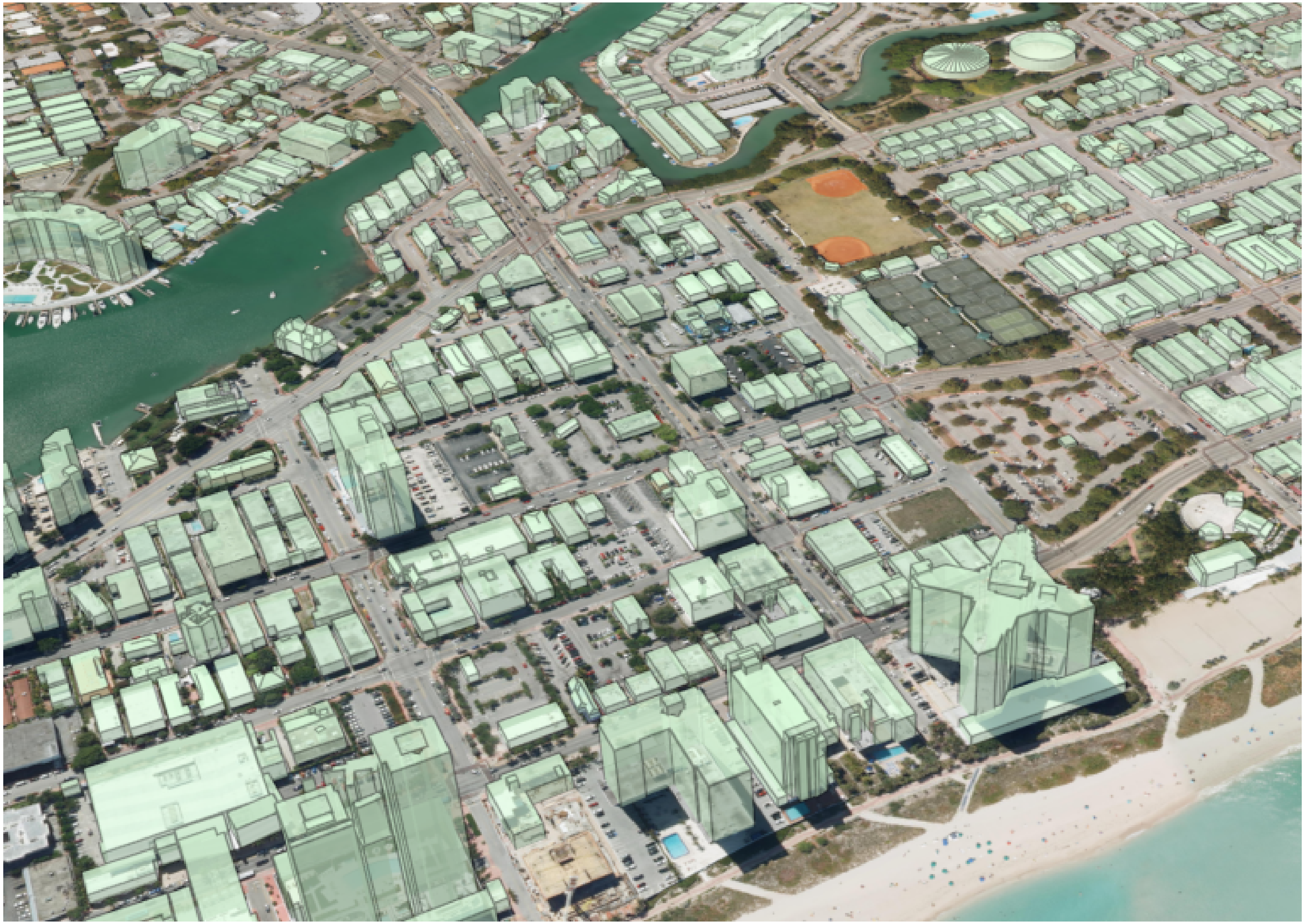
Nuclear Segmentation



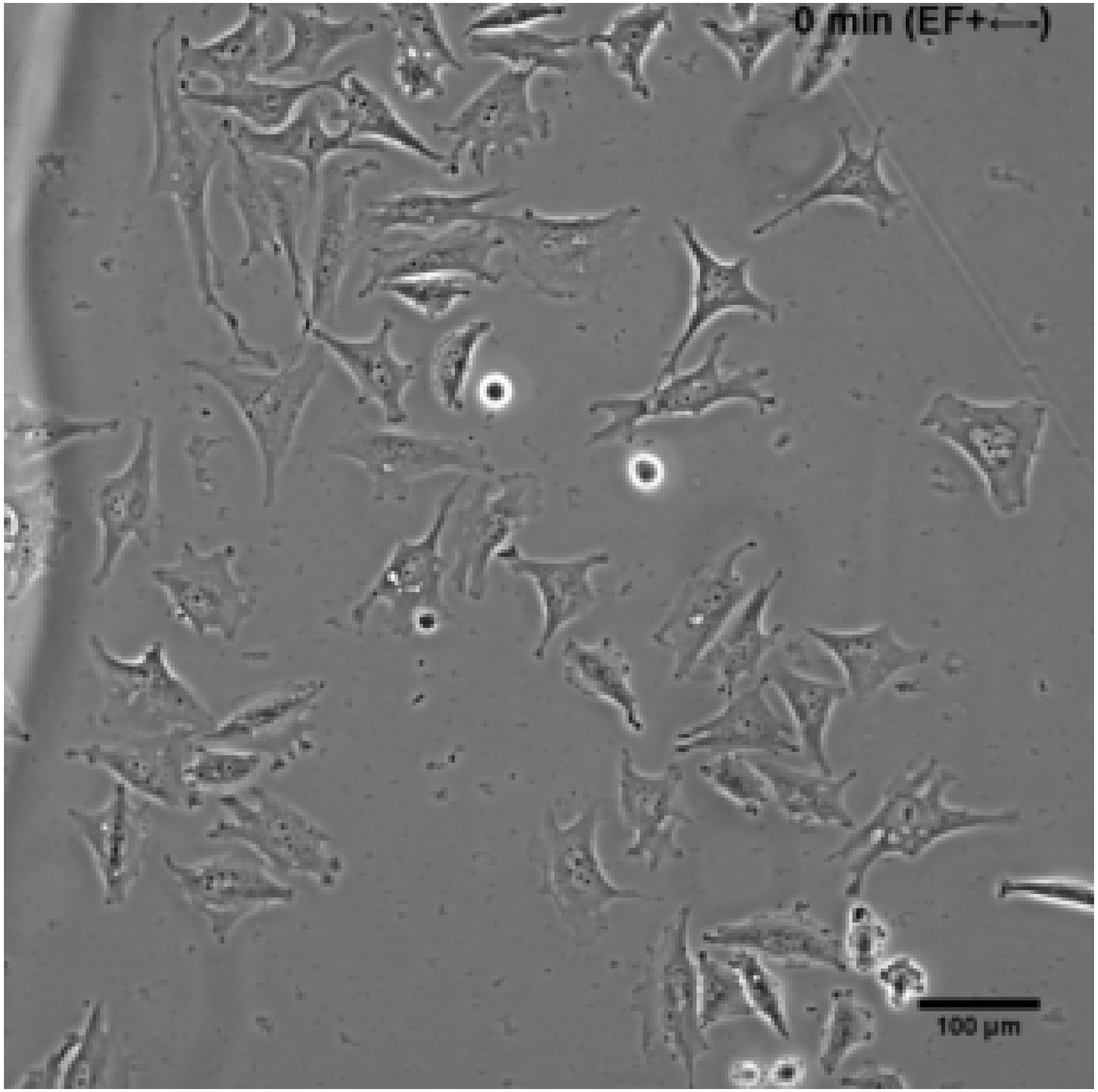
Industrial Robot



3D Buildings



Cell



Geographic Polygon

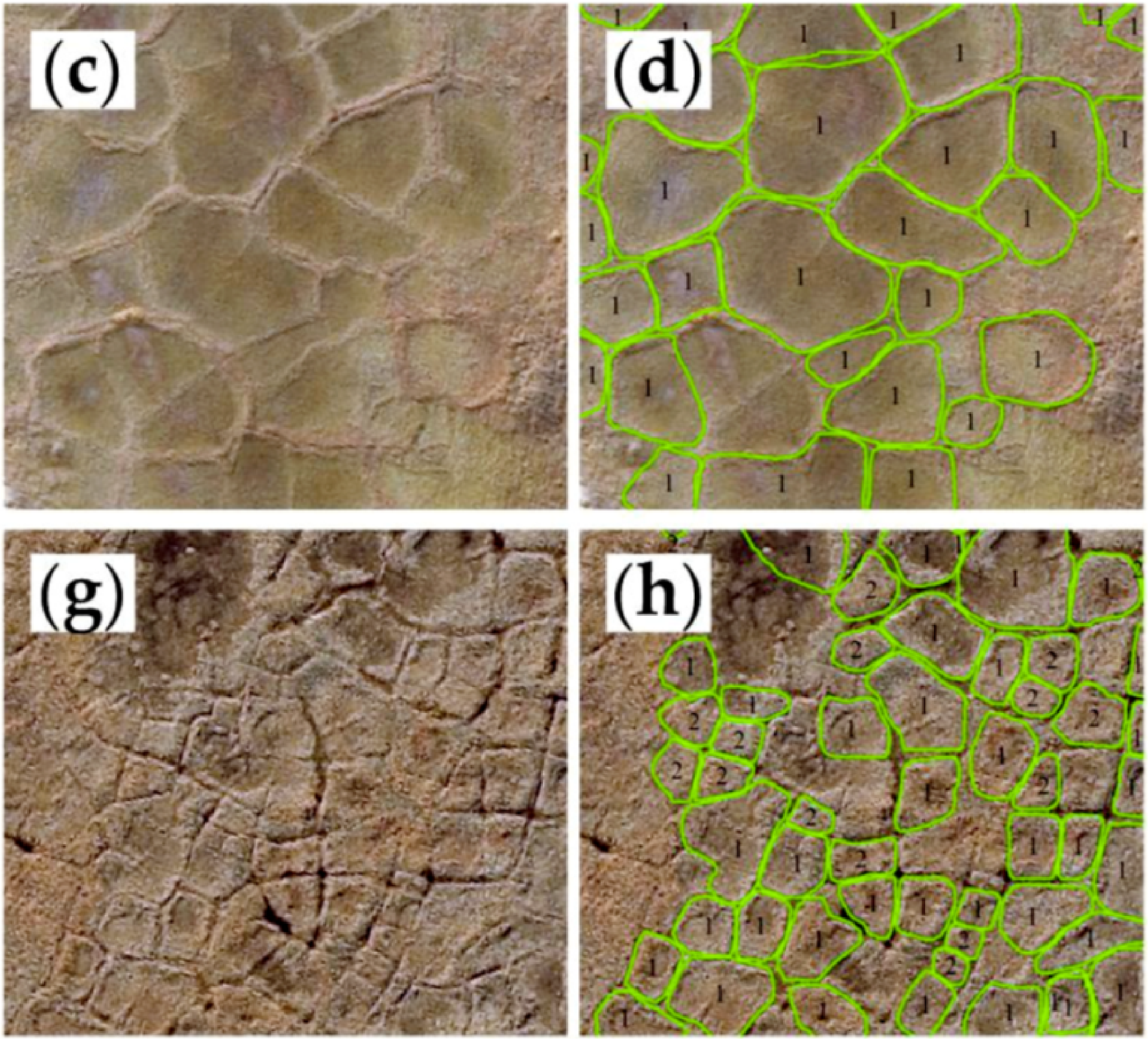
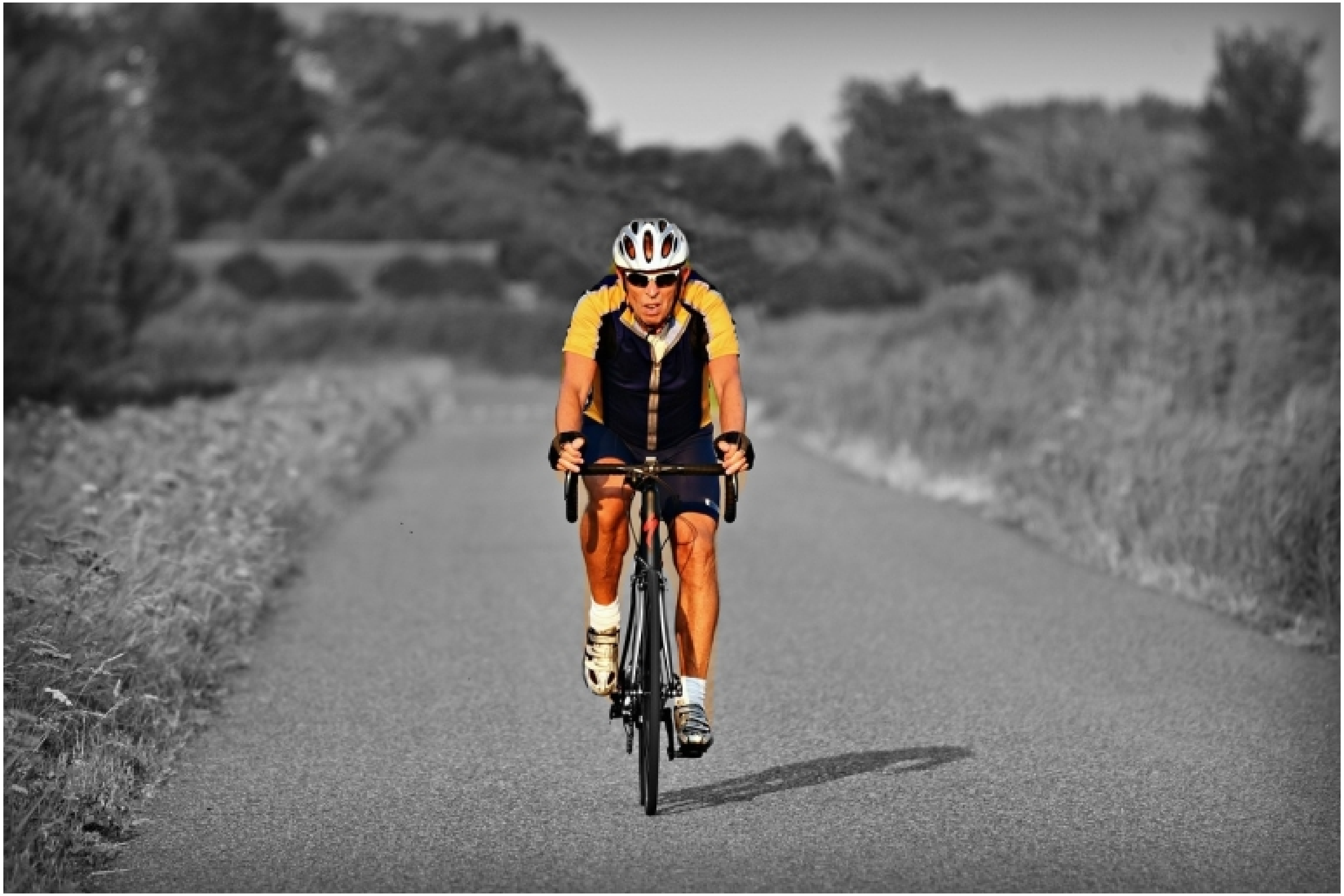
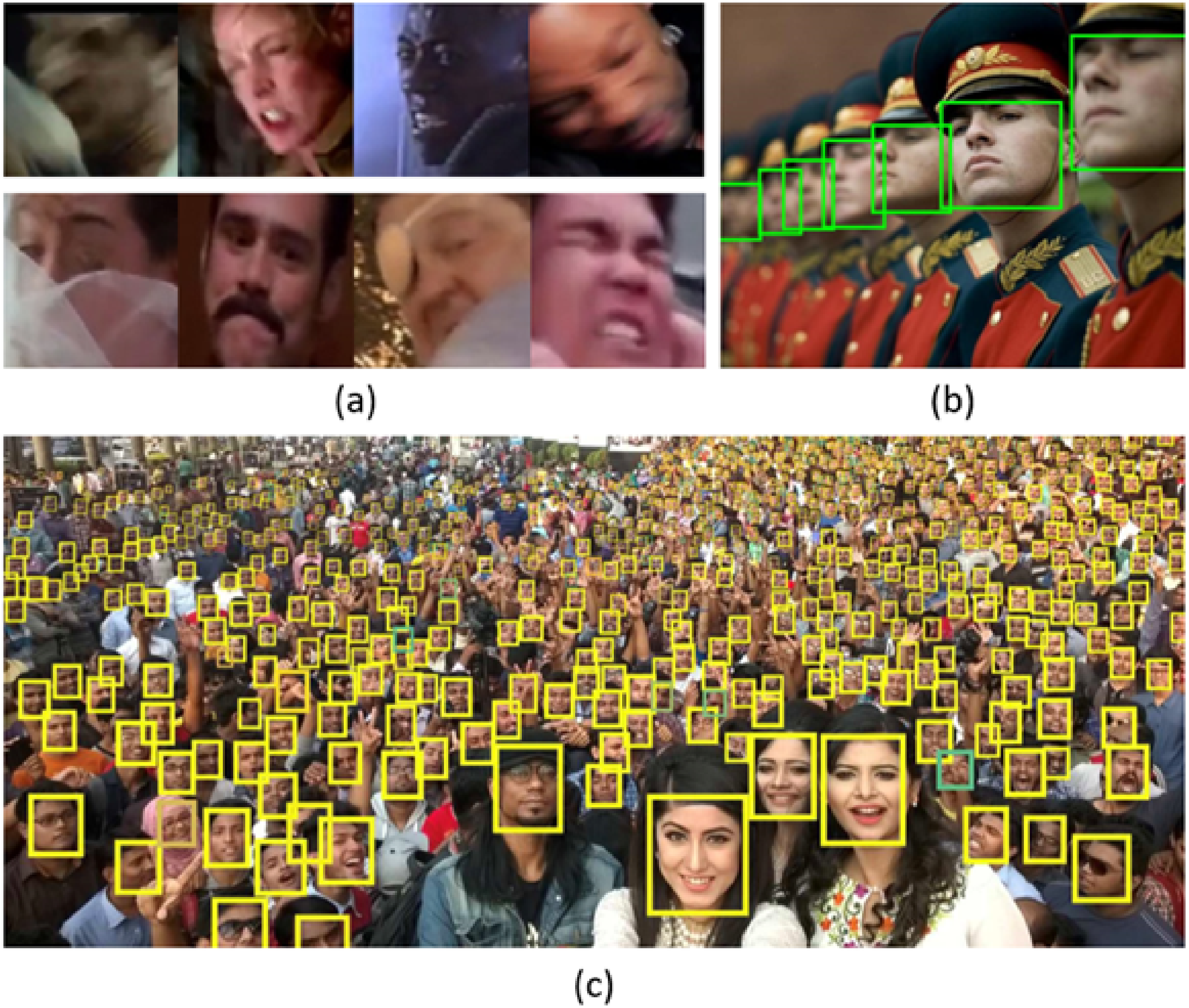


Photo Effects



Face Detection

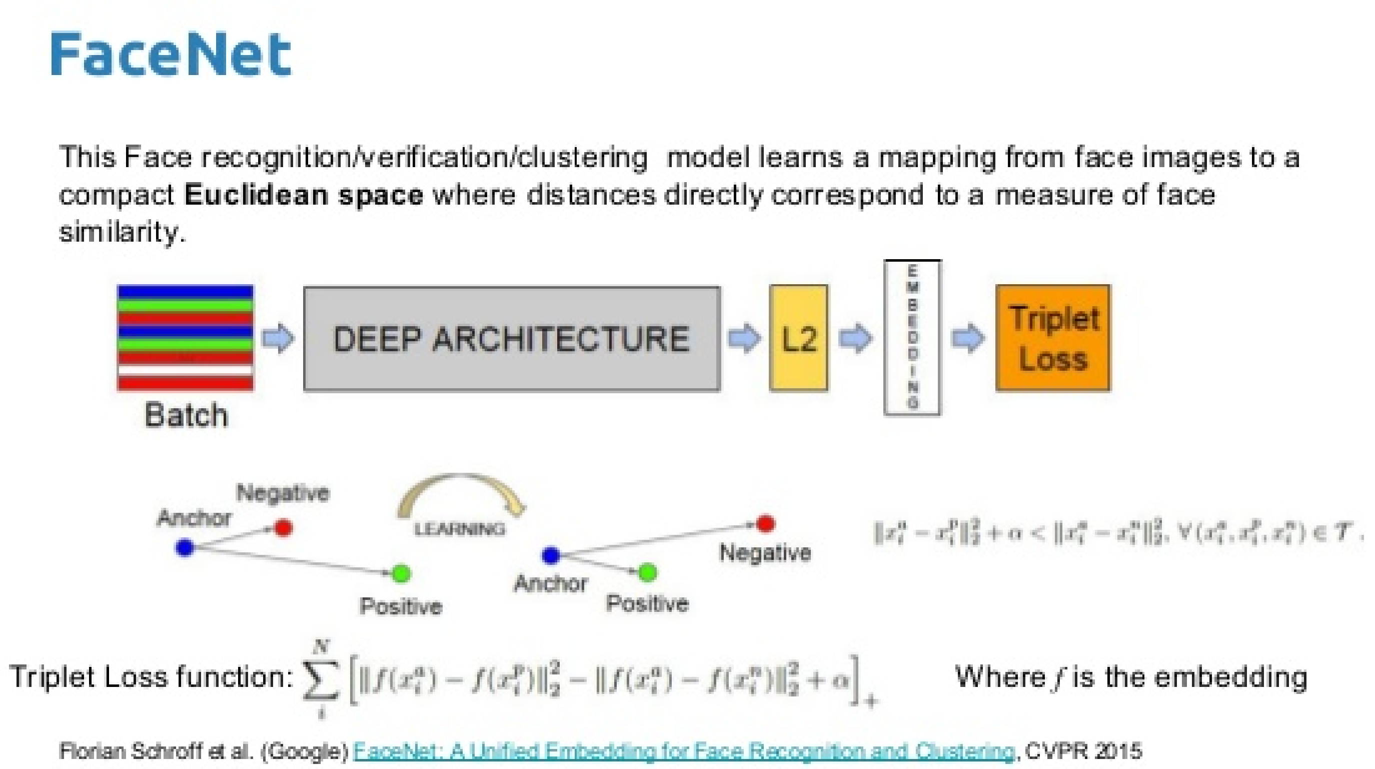


Face Recognition

- Face recognition technology finds suspect in Maryland shootings
- Pop star Taylor Swift, filtering fans and followers at concerts
- Shelter tracks use of shelters

FaceNet

In 2015 Google proposed



FaceNet

This Face recognition/verification/clustering model learns a mapping from face images to a compact **Euclidean space** where distances directly correspond to a measure of face similarity.

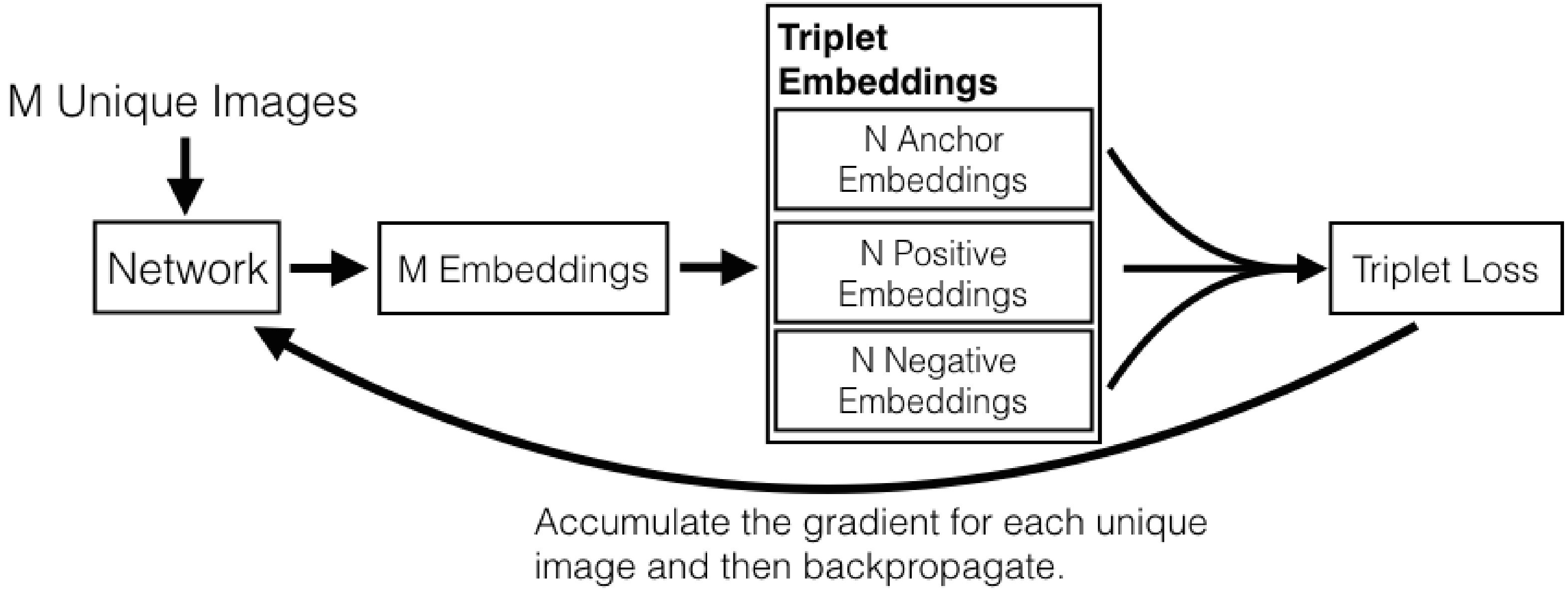
$$\|x_i^a - x_i^p\|_2^2 + \alpha < \|x_i^a - x_i^n\|_2^2, \ \forall (x_i^a, x_i^p, x_i^n) \in \mathcal{T} .$$

Triplet Loss function: $\sum_i^N \left[\|f(x_i^a) - f(x_i^p)\|_2^2 - \|f(x_i^a) - f(x_i^n)\|_2^2 + \alpha \right]_+$ Where f is the embedding

Florian Schroff et al. (Google) FaceNet: A Unified Embedding for Face Recognition and Clustering, CVPR 2015

FaceNet Architecture

Using Triple Loss to capture similarities and differences between different faces



FaceNet Design

Convert a human face into a 128-dimensional vector representation

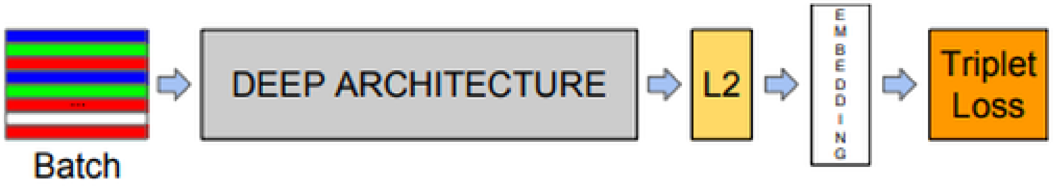


Figure 2. **Model structure.** Our network consists of a batch input layer and a deep CNN followed by L_2 normalization, which results in the face embedding. This is followed by the triplet loss during training.

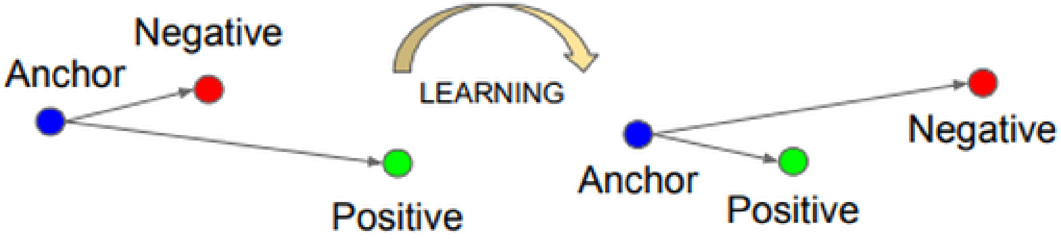
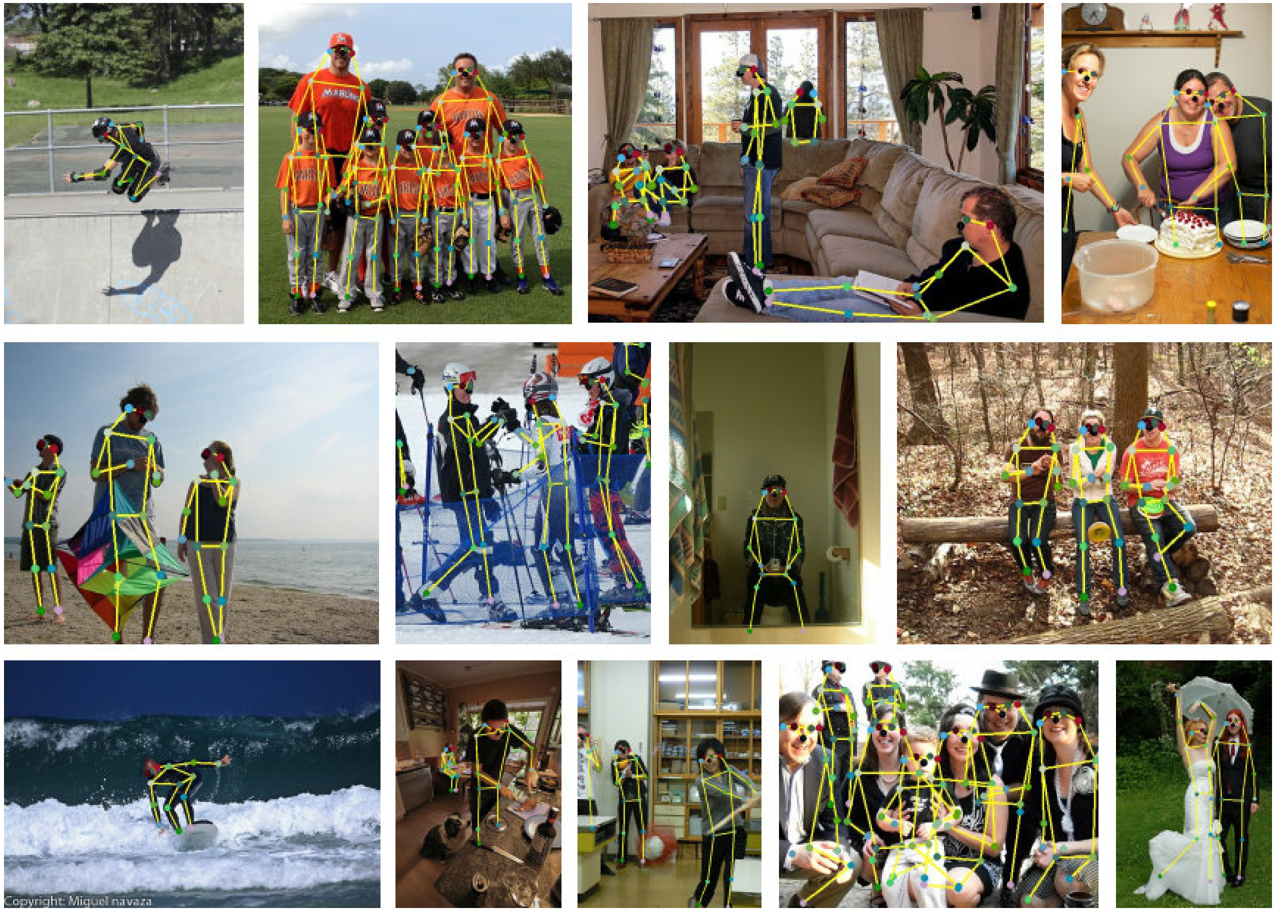
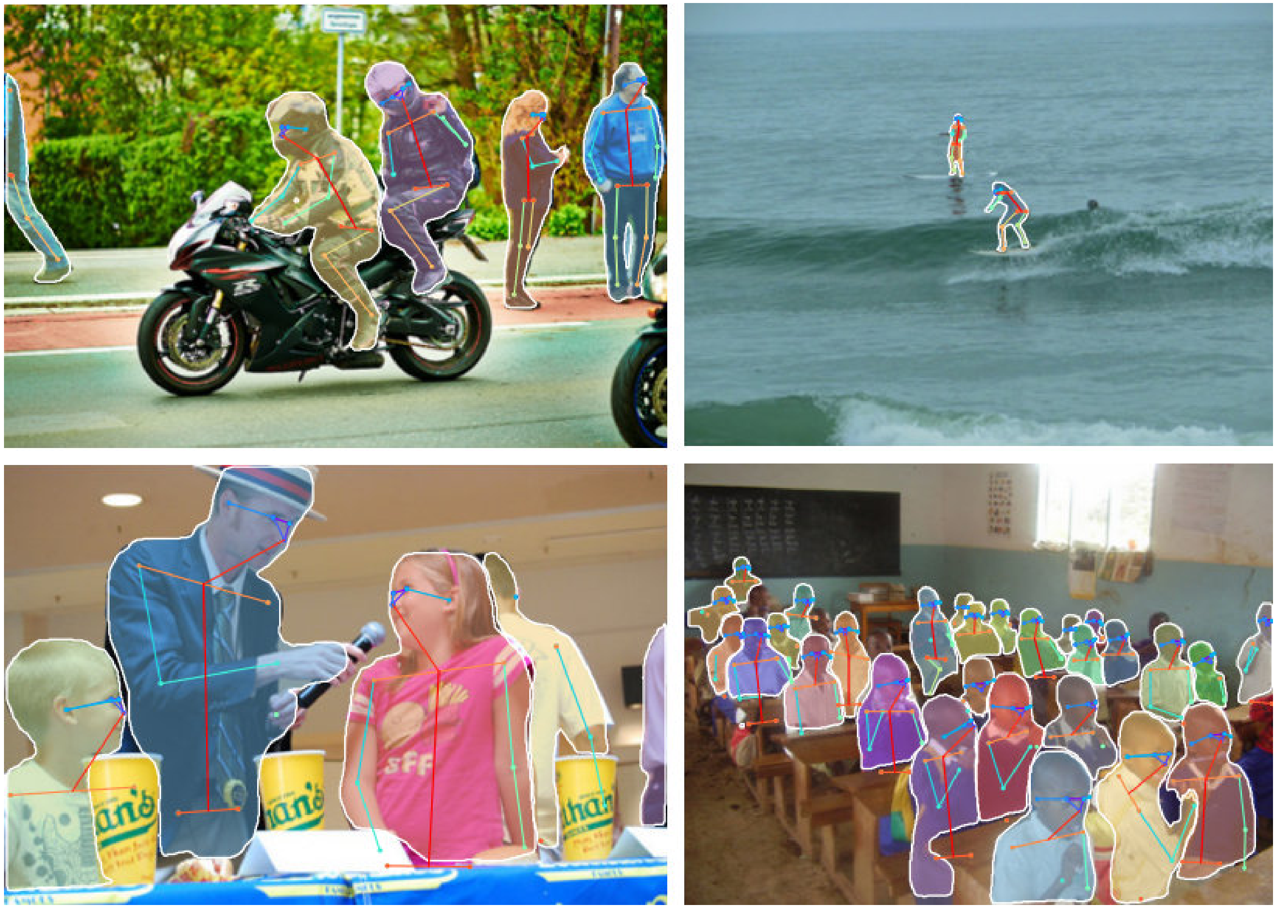


Figure 3. The **Triplet Loss** minimizes the distance between an *anchor* and a *positive*, both of which have the same identity, and maximizes the distance between the *anchor* and a *negative* of a different identity.

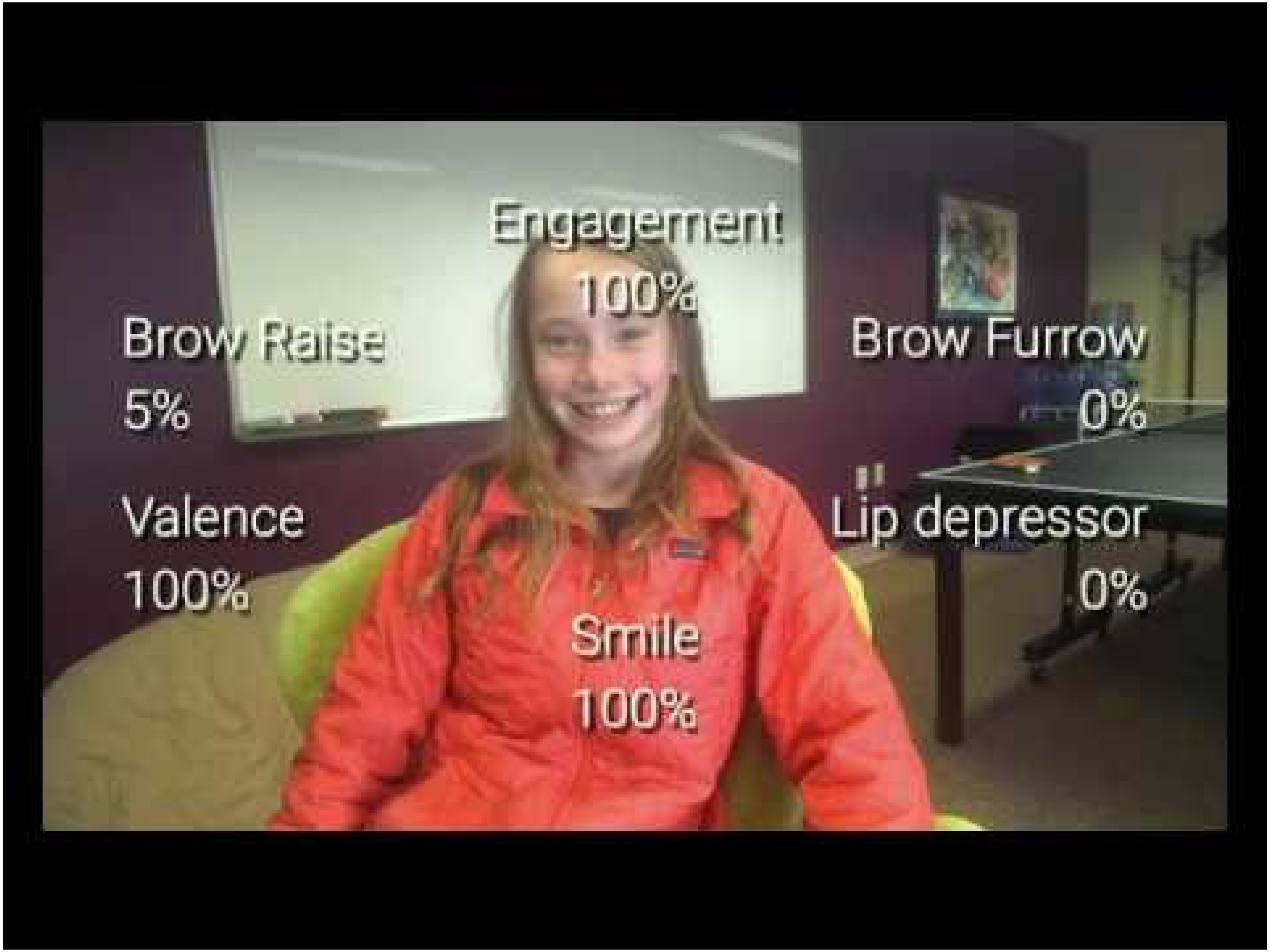
Pose Detection and Recognition



Pose Detection and Recognition



Emotion


Engagement
100%
Brow Raise
5%
Brow Furrow
0%
Valence
100%
Lip depressor
0%
Smile
100%

Traffic Flow Counting

741 人在观看： 【孝利家民宿2】合集 (已更新至E12.180422… 立即围观 >

人工智能：交通流量计数

播放器初始化...[完成]
加载用户配置...[完成]
加载视频地址...[完成]
加载视频内容...

进入bilibili,一起发弹幕吐槽!

Traffic Flow Counting

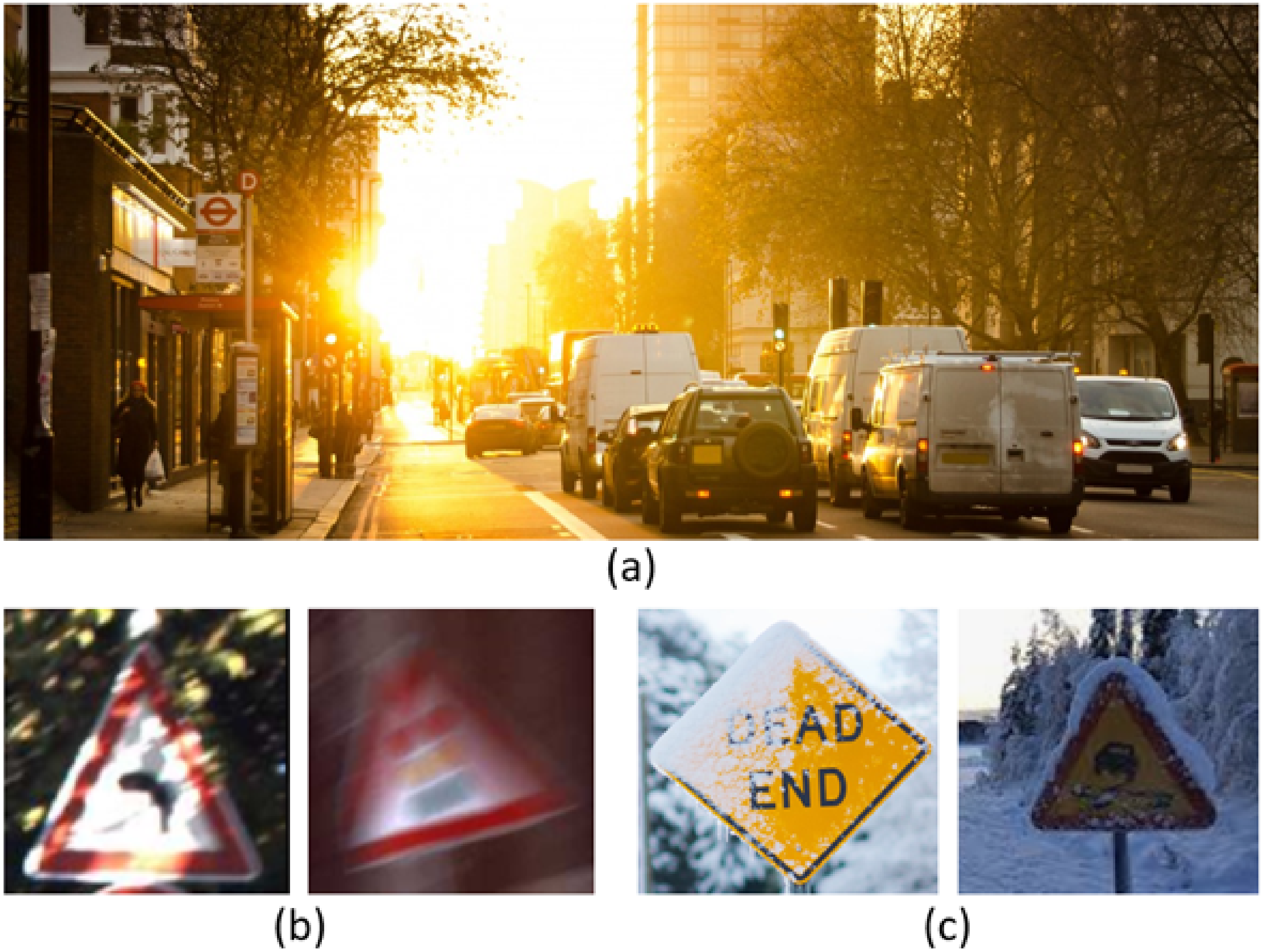
454 人在观看： 【木鱼微剧场】《东方快车谋杀案》阿加莎... 立即围观 >

人工智能：交通流量计数II

```
播放器初始化...[完成]
加载用户配置...[完成]
加载视频地址...[完成]
加载视频内容...
```

进入bilibili,一起发弹幕吐槽!

Traffic Signal Recognition



(a)

(b)

(c)

Rail Recognition

664 人在观看： 【QiTV】【战神4】纯剧情剪辑完结合集（1… 立即围观 >

人工智能：铁路信号检测

播放器初始化...[完成]
加载用户配置...[完成]
加载视频地址...

进入bilibili,一起发弹幕吐槽!

Crossing Monitoring

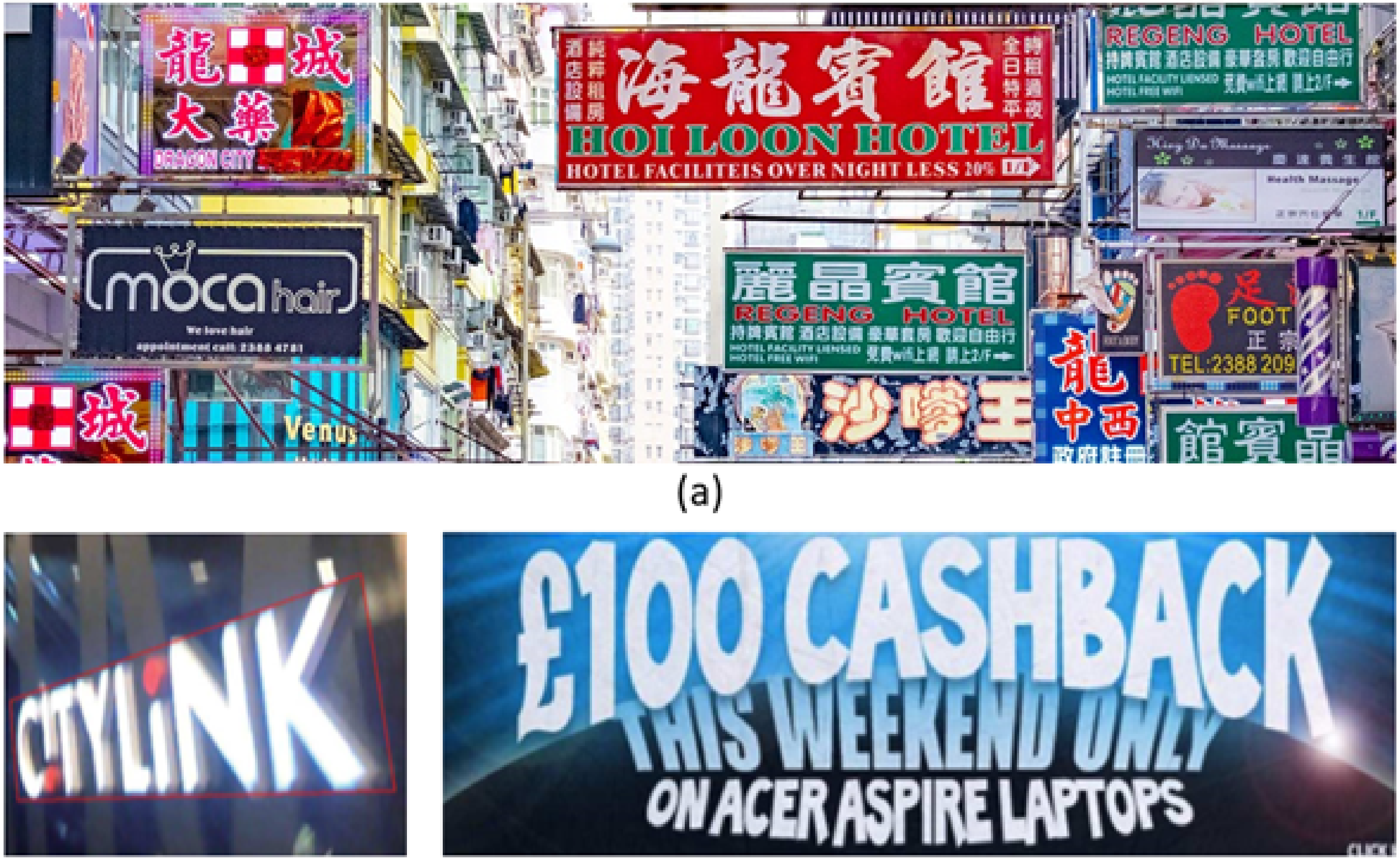
428 人在观看： [TSKS]孝利家的民宿2.E12.180422.中字 立即围观 >

人工智能：铁路信号检测

播放器初始化...[完成]
加载用户配置...[完成]
加载视频地址...[完成]
加载视频内容...

进入bilibili,一起发弹幕吐槽!

Text Recognition



(a)

(b)

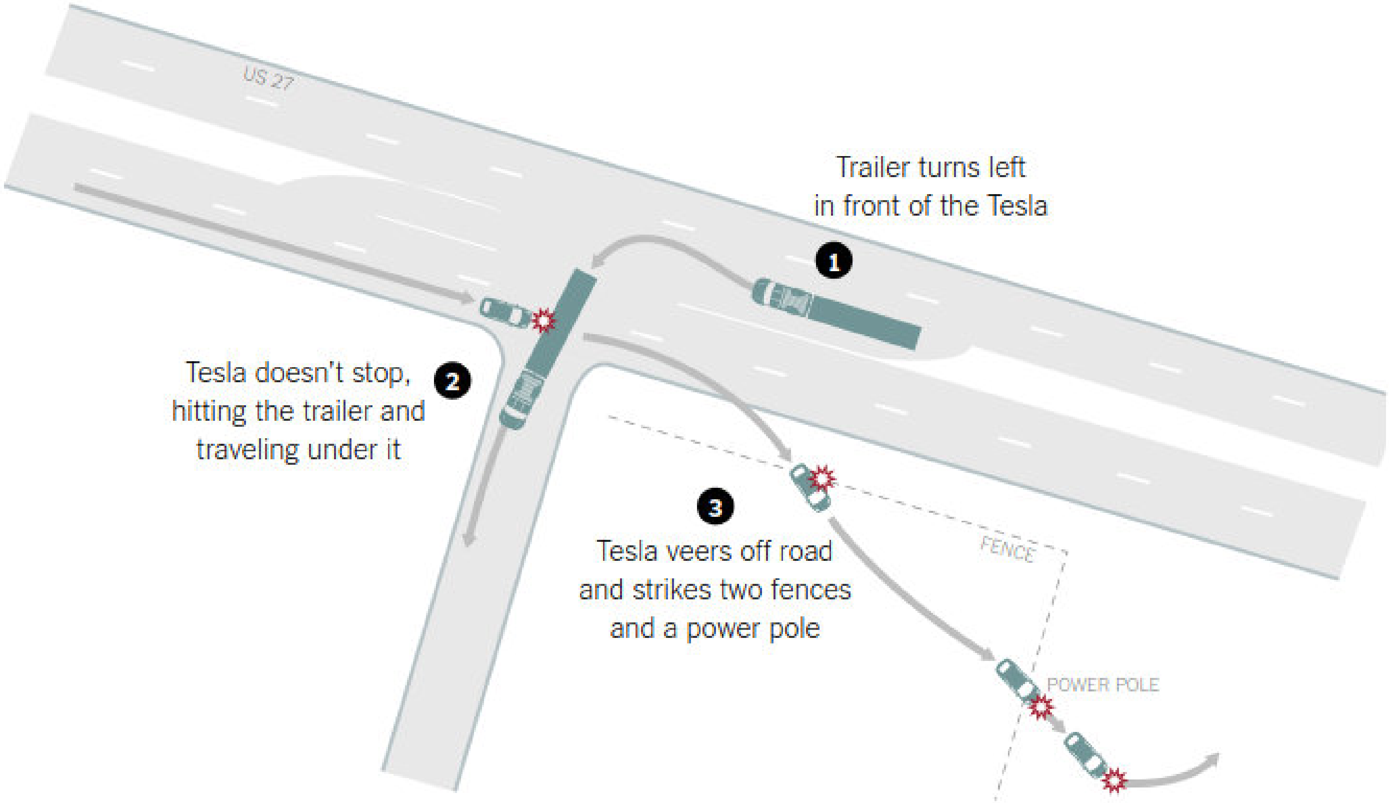
(c)

Problems

Accuracy, privacy protection, fairness

Accuracy Problem

- Tesla's autonomous driving system fails to identify white vans


US 27
1
Trailer turns left in front of the Tesla
2
Tesla doesn't stop, hitting the trailer and traveling under it
3
Tesla veers off road and strikes two fences and a power pole
FENCE
POWER POLE

Accuracy Problem

<!-- ---

Privacy Protection

- On May 14, 2019, the San Francisco City Supervisory Commission passed a decree by 8 votes to 1 to ban city workers from purchasing and using face recognition technology
- Face recognition technology tends to endanger civil rights and civil liberties far more than its claimed benefits. This technology will exacerbate racial inequalities and threaten our ability to live without long-term government surveillance.

Quiz

- What is instance segmentation for?
- In reality, what problems should be paid attention to in the application of computer vision technology?
- Give examples of computer vision applications you might need at work
- Deep learning brings major breakthroughs in the field of images, please give an example that impresses you