Object Detection – Part 1

Danna Gurari
University of Colorado Boulder
Fall 2023

Review

- Last lecture:
  - Scene Classification Problem and Applications
  - Scene Classification: Datasets and Evaluation Metrics
  - Scene Classification Models: Deep Features and Fine-Tuning
  - Attribute Classification: Problem, Applications, and Datasets
  - Student-led Lecture Overview

- Assignments (Canvas)
  - Reading assignment was due earlier today
  - Next reading assignments due Wednesday and next Monday

- Questions?

Object Detection: Today’s Topics

- Problem
- Applications
- Datasets
- Evaluation metric
- Overview of object detection algorithms and baseline (R-CNN)

Object Detection: Today's Topics

- Problem
- Applications
- Datasets
- Evaluation metric
- Overview of object detection algorithms and baseline (R-CNN)

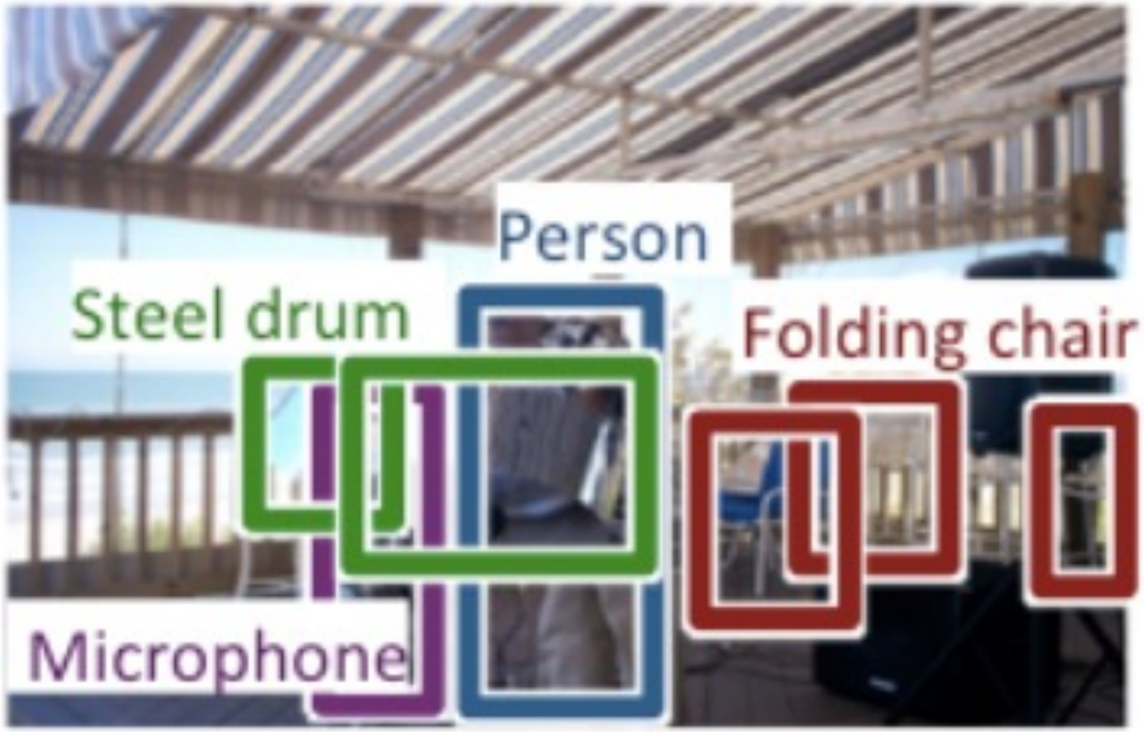
Problem Definition

- Localize with a bounding box object(s) of interest

Focus for today's lecture

Problem: Semantic Object Detection

- Localize with a bounding box every instance of an object from pre-specified categories



[Russakovsky et al; IJCV 2015]

A reasonably solved problem

Problem: **Salient** Object Detection

- Localize with a bounding box the salient object(s)

[Liu et al; CVPR 2007]

Object Detection vs Object Recognition

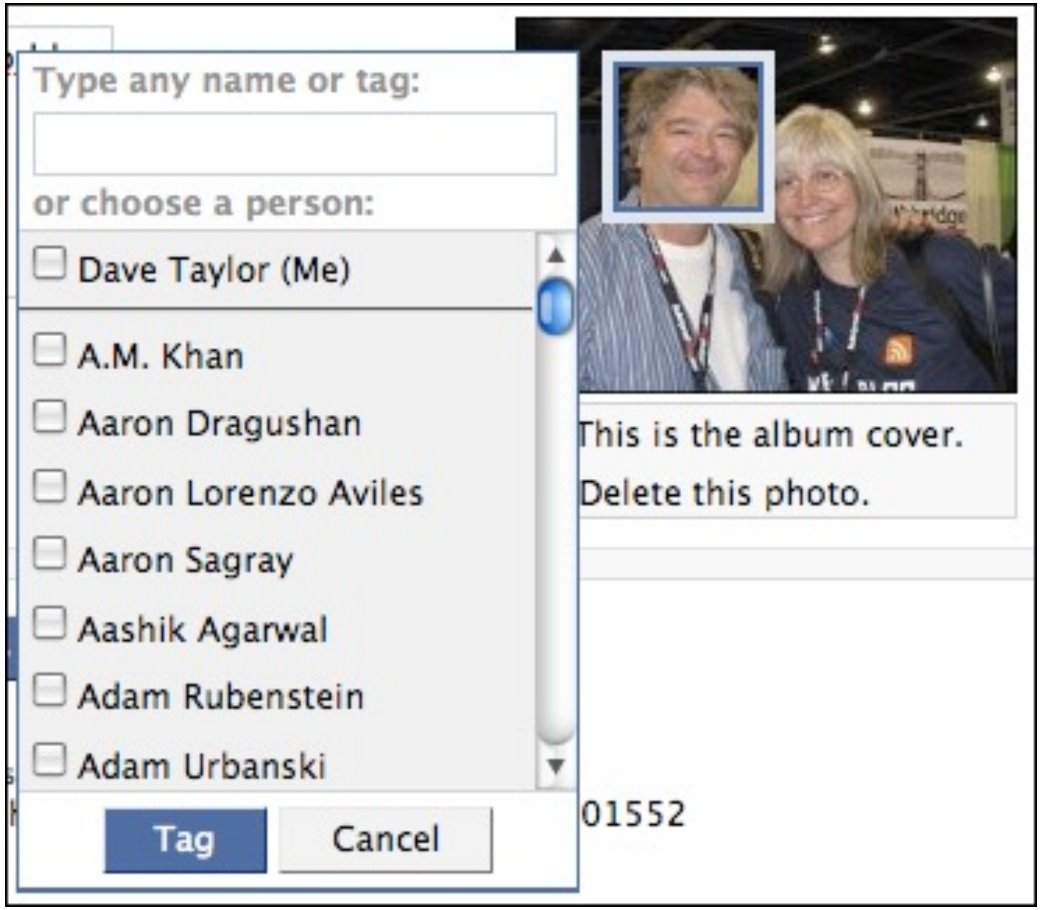
"What is the difference between (semantic) object detection and object recognition?"

- Must learn appearance of object rather than only its image context; e.g., giraffe

Object Detection: Today's Topics

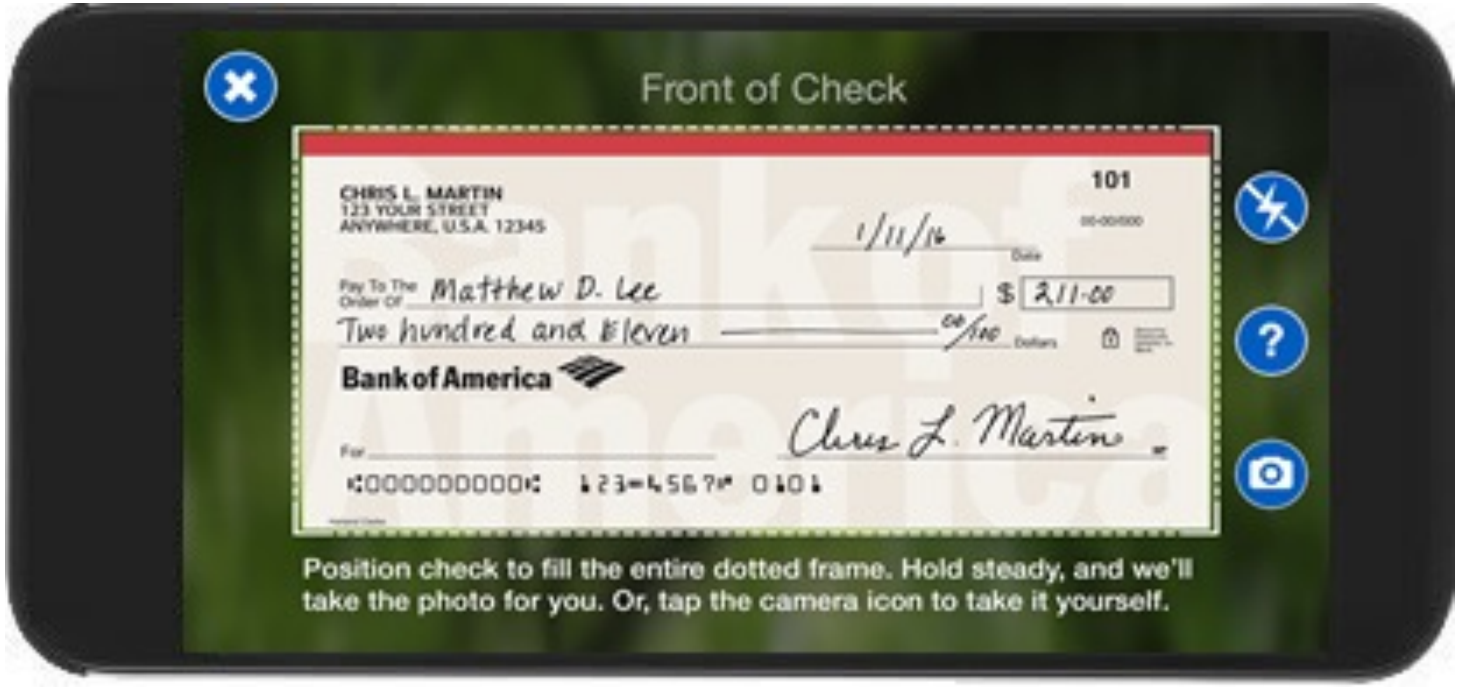
- Problem
- Applications
- Datasets
- Evaluation metric
- Overview of object detection algorithms and baseline (R-CNN)

Social Media



Face detection
(e.g., Facebook)

Banking

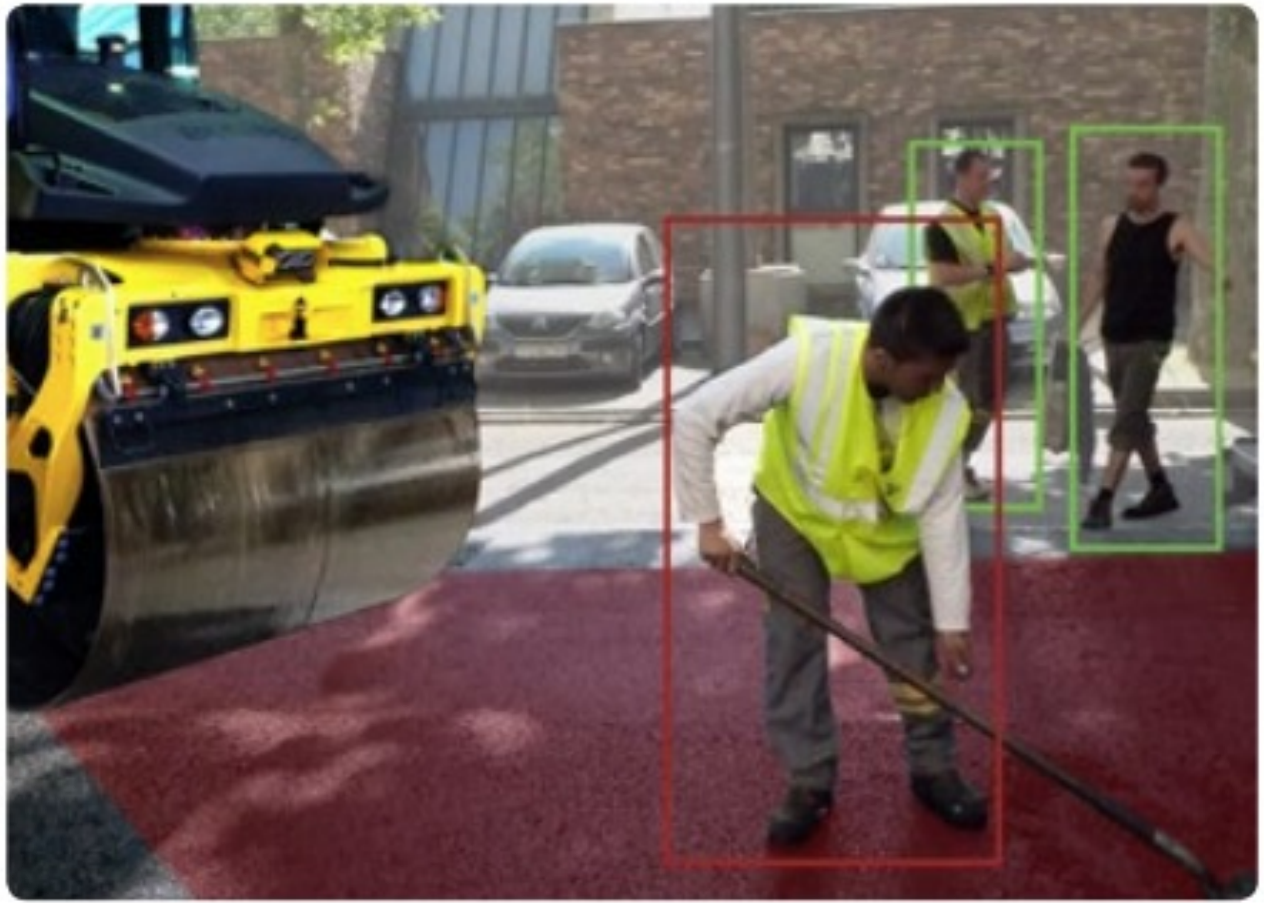


Mobile check deposit
(e.g., Bank of America)

Transportation

License Plate Detection (e.g., AllGoVision)

Construction Safety



Pedestrian Detection
(e.g., Blaxtair)
http://media.brintex.com/Occurrence/121/Brochure/3435/brochure.pdf

Counting

Counting Fish (e.g., SalmonSoft)
http://www.wecountfish.com/?page_id=143

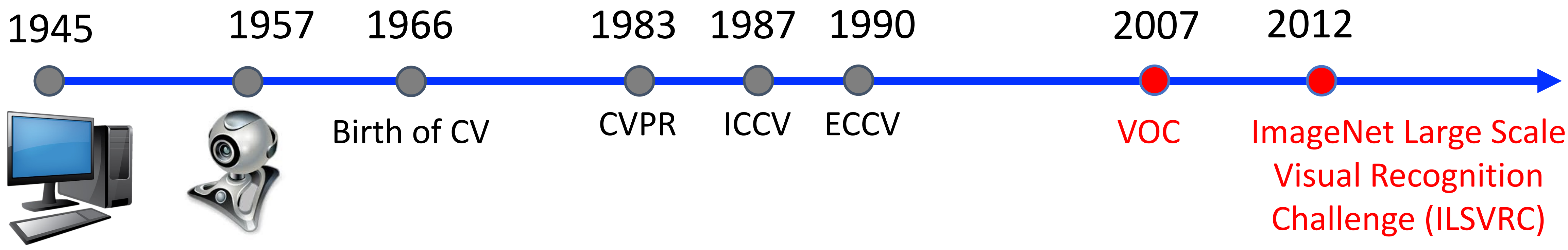
Business Traffic Analytics

Can you think of any other potential applications?

Object Detection: Today's Topics

- Problem
- Applications
- Datasets
- Evaluation metric
- Overview of object detection algorithms and baseline (R-CNN)

Object Detection Datasets



VOC

1. Category Selection

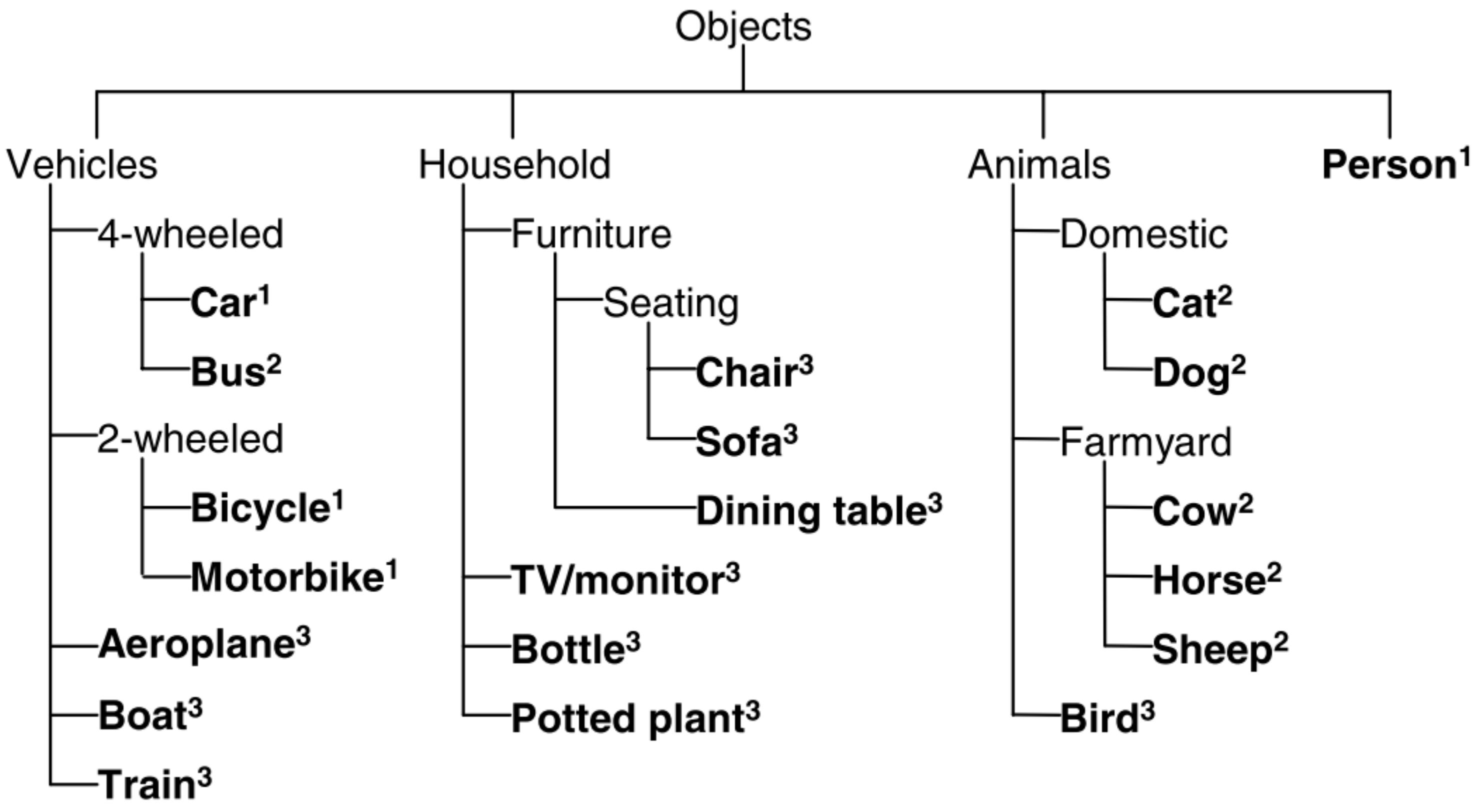
- 20 categories chosen:

1) Initial 4 categories stem from existing dataset

2) 2006: added 6 classes

3) 2007: added 10 classes

- Additional categories provide a broader domain and finer-grained categories, including visually similar things



(superscript indicates year of inclusion in the challenge: 2005[1], 2006[2], 2007[3])

Mark Everingham, Luc Van Gool, Christopher K. I. Williams, John Winn, and Andrew Zisserman. The PASCAL Visual Object Classes (VOC) Challenge. IJCV 2010.

VOC

1. Category Selection

- 20 categories chosen:

1) Initial 4 categories stem from existing dataset

2) 2006: added 6 classes

3) 2007: added 10 classes

- Additional categories provide a broader domain and finer-grained categories, including visually similar things

→

2. Image Collection

- 500,000 images retrieved from Flickr by querying with a number of keywords

(many query terms per category)

- **aeroplane**, airplane, plane, biplane, monoplane, aviator, bomber, hydroplane, airliner, aircraft, fighter, airport, hangar, jet, boeing, fuselage, wing, propellor, flying
- **bicycle**, bike, cycle, cyclist, pedal, tandem, saddle, wheel, cycling, ride, wheelie
- **bird**, birdie, birdwatching, nest, sea, aviary, birdcage, bird feeder, bird table
- **boat** ship, barge, ferry, canoe, boating, craft, liner, cruise, sailing, rowing, watercraft, regatta, racing, marina, beach, water, canal, river, stream, lake, yacht
- **bottle**, cork, wine, beer, champagne, ketchup, squash, soda, coke, lemonade, dinner, lunch, breakfast
- **bus**, omnibus, coach, shuttle, jitney, double-decker, motorbus, school bus, depot, terminal, station, terminus, passenger, route
- **car**, automobile, cruiser, motorcar, vehicle, hatchback, saloon, convertible, limousine, motor, race, traffic, trip, rally, city, street, road, lane, village, town, centre, shopping, downtown, suburban
- **cat**, feline, pussy, mew, kitten, tabby, tortoiseshell, ginger, stray
- **chair**, seat, rocker, rocking, deck, swivel, camp, chaise, office, studio, armchair, recliner, sitting, lounge, living room, sitting room
- **cow**, beef, heifer, moo, dairy, milk, milking, farm
- **dog**, hound, bark, kennel, heel, bitch, canine, puppy, hunter, collar, leash
- **horse**, gallop, jump, buck, equine, foal, cavalry, saddle, canter, buggy, mare, neigh, dressage, trial, racehorse, steeplechase, thoroughbred, cart, equestrian, paddock, stable, farrier
- **motorbike**, motorcycle, minibike, moped, dirt, pillion, biker, trials, motorcycling, motorcyclist, engine, motocross, scramble, sidecar, scooter, trail
- **person**, people, family, father, mother, brother, sister, aunt, uncle, grandmother, grandma, grandfather, grandpa, grandson, granddaughter, niece, nephew, cousin
- **sheep**, ram, fold, fleece, shear, baa, bleat, lamb, ewe, wool, flock
- **sofa**, chesterfield, settee, divan, couch, bolster
- **table**, dining, cafe, restaurant, kitchen, banquet, party, meal
- **potted plant**, pot plant, plant, patio, windowsill, window sill, yard, greenhouse, glass house, basket, cutting, pot, cooking, grow
- **train**, express, locomotive, freight, commuter, platform, subway, underground, steam, railway, railroad, rail, tube, underground, track, carriage, coach, metro, sleeper, railcar, buffet, cabin, level crossing
- **tv/monitor**, television, plasma, flatscreen, flat screen, lcd, crt, watching, dvd, desktop, computer, computer monitor, PC, console, game

Mark Everingham, Luc Van Gool, Christopher K. I. Williams, John Winn, and Andrew Zisserman. The PASCAL Visual Object Classes (VOC) Challenge. IJCV 2010.

VOC

1. Category Selection

- 20 categories chosen:

1) Initial 4 categories stem from existing dataset

2) 2006: added 6 classes

3) 2007: added 10 classes

- Additional categories provide a broader domain and finer-grained categories, including visually similar things

2. Image Collection

- 500,000 images retrieved from Flickr by querying with a number of keywords

3. Image Verification + Image Annotation

- University of Leeds annotation party to recruit annotators

- Annotation guidelines & real-time assistance

- Review of every annotation

- Annotate only "minority" classes at end of party to increase the count of them

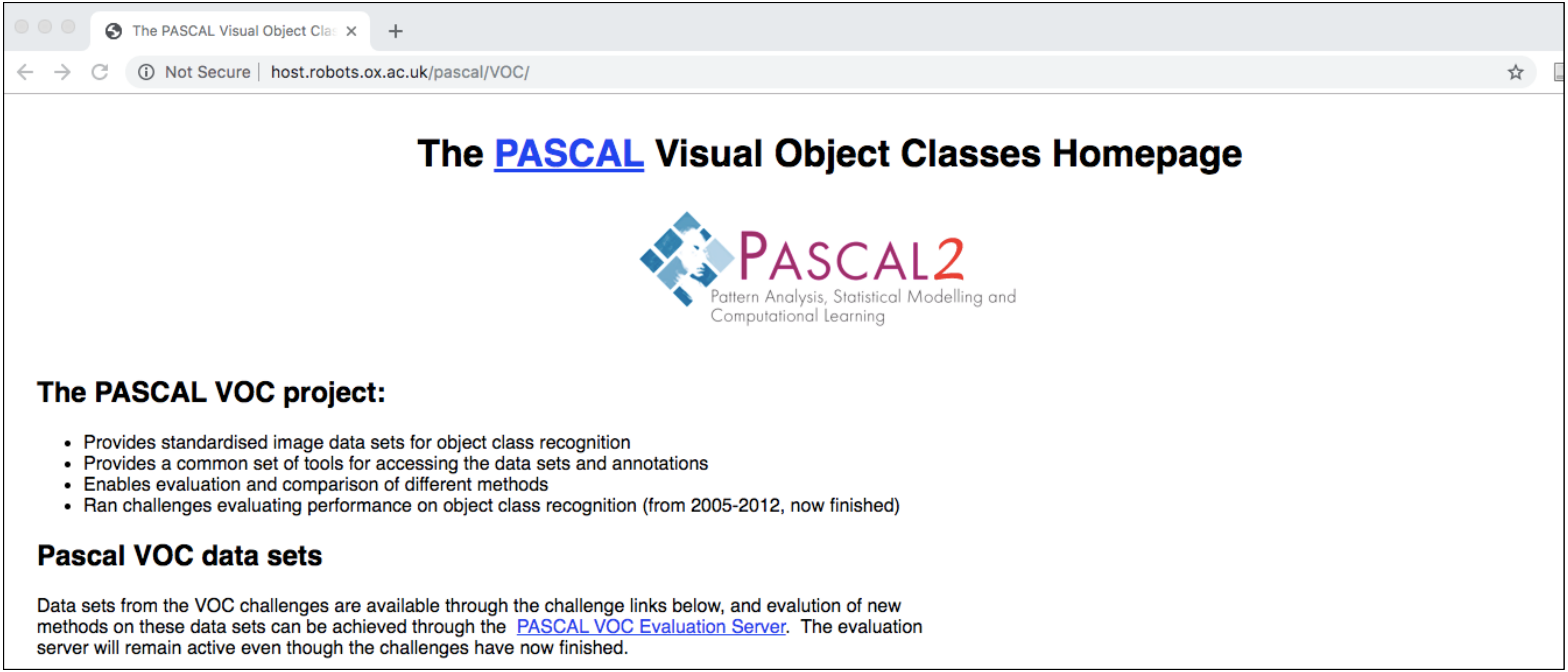
Mark Everingham, Luc Van Gool, Christopher K. I. Williams, John Winn, and Andrew Zisserman. The PASCAL Visual Object Classes (VOC) Challenge. IJCV 2010.

VOC Guidelines:

What are potential limitations of this task design for resulting datasets (and so algorithms developed with such datasets)?

What to label	*All objects of the defined categories*, unless: • you are unsure what the object is. • the object is very small (at your discretion). • less than 10-20% of the object is visible, *such that you cannot be sure what class it is*. e.g. if only a tyre is visible it may belong to car or truck so cannot be labelled car, but feet/faces can only belong to a person. If this is not possible because too many objects, mark image as bad.
Viewpoint	Record the viewpoint of the 'bulk' of the object e.g. the body rather than the head. Allow viewpoints within 10-20 degrees. If ambiguous, leave as 'Unspecified'. Unusually rotated objects e.g. upside-down people should be left as 'Unspecified'.
Bounding box	Mark the bounding box of the visible area of the object (*not* the estimated total extent of the object). Bounding box should contain all visible pixels, except where the bounding box would have to be made excessively large to include a few additional pixels (<5%) e.g. a car aerial.
Truncation	If more than 15-20% of the object lies outside the bounding box mark as Truncated. The flag indicates that the bounding box does not cover the total extent of the object.
Occlusion	If more than 5% of the object is occluded within the bounding box, mark as Occluded. The flag indicates that the object is not totally visible within the bounding box.
Image quality/ illumination	Images which are poor quality (e.g. excessive motion blur) should be marked bad. However, poor illumination (e.g. objects in silhouette) should not count as poor quality unless objects cannot be recognised. Images made up of multiple images (e.g. collages) should be marked bad.
Clothing/mud/ snow etc.	If an object is 'occluded' by a close-fitting occluder e.g. clothing, mud, snow etc., then the occluder should be treated as part of the object.
Transparency	Do label objects visible through glass, but treat reflections on the glass as occlusion.
Mirrors	Do label objects in mirrors.
Pictures	Label objects in pictures/posters/signs only if they are photorealistic but not if cartoons, symbols etc.

VOC Annual Workshop



The PASCAL Visual Object Clas

Not Secure | host.robots.ox.ac.uk/pascal/VOC/

The PASCAL Visual Object Classes Homepage

PASCAL2
Pattern Analysis, Statistical Modelling and Computational Learning

The PASCAL VOC project:

- Provides standardised image data sets for object class recognition
- Provides a common set of tools for accessing the data sets and annotations
- Enables evaluation and comparison of different methods
- Ran challenges evaluating performance on object class recognition (from 2005-2012, now finished)

Pascal VOC data sets

Data sets from the VOC challenges are available through the challenge links below, and evalution of new methods on these data sets can be achieved through the PASCAL VOC Evaluation Server. The evaluation server will remain active even though the challenges have now finished.

http://host.robots.ox.ac.uk/pascal/VOC/

VOC: Datasets Evolved

The table below gives a brief summary of the main stages of the VOC development.

Year	Statistics	New developments	Notes
2005	Only 4 classes: bicycles, cars, motorbikes, people. Train/validation/test: 1578 images containing 2209 annotated objects.	Two competitions: classification and detection	Images were largely taken from exising public datasets, and were not as challenging as the flickr images subsequently used. This dataset is obsolete.
2006	10 classes: bicycle, bus, car, cat, cow, dog, horse, motorbike, person, sheep. Train/validation/test: 2618 images containing 4754 annotated objects.	Images from flickr and from Microsoft Research Cambridge (MSRC) dataset	The MSRC images were easier than flickr as the photos often concentrated on the object of interest. This dataset is obsolete.

http://host.robots.ox.ac.uk/pascal/VOC/

ILSVRC

"ILSVRC follows in the footsteps of the PASCAL VOC challenge… which set the precedent for standardized evaluation of recognition algorithms in the form of yearly competitions."

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, Li Fei-Fei , IJCV 2015

ILSVRC

1. Category Selection

- 200 ImageNet classes which:
1) exclude synset overlap
2) exclude object classes too “big” in the image
3) are basic-level categories
4) backward compatible: VOC

Class name in PASCAL VOC (20 classes)	Closest class in ILSVRC-DET (200 classes)
aeroplane	airplane
bicycle	bicycle
bird	bird
boat	*watercraft*
bottle	*wine bottle*
bus	bus
car	car
cat	domestic cat
chair	chair
cow	*cattle*
dining table	*table*
dog	dog
horse	horse
motorbike	motorcyle
person	person
potted plant	*flower pot*
sheep	sheep
sofa	sofa
train	train
tv/monitor	tv or monitor

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, Li Fei-Fei , IJCV 2015

ILSVRC

1. Category Selection

- 200 ImageNet classes which:
1) exclude synset overlap
2) exclude object classes too “big” in the image
3) are basic-level categories
4) backward compatible: VOC

→

2. Image Collection

- Subset of images from ImageNet

- Additional images from Flickr

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, Li Fei-Fei , IJCV 2015

ILSVRC

1. Category Selection

- 200 ImageNet classes which:
1) exclude synset overlap
2) exclude object classes too “big” in the image
3) are basic-level categories
4) backward compatible: VOC

2. Image Collection

- Subset of images from ImageNet

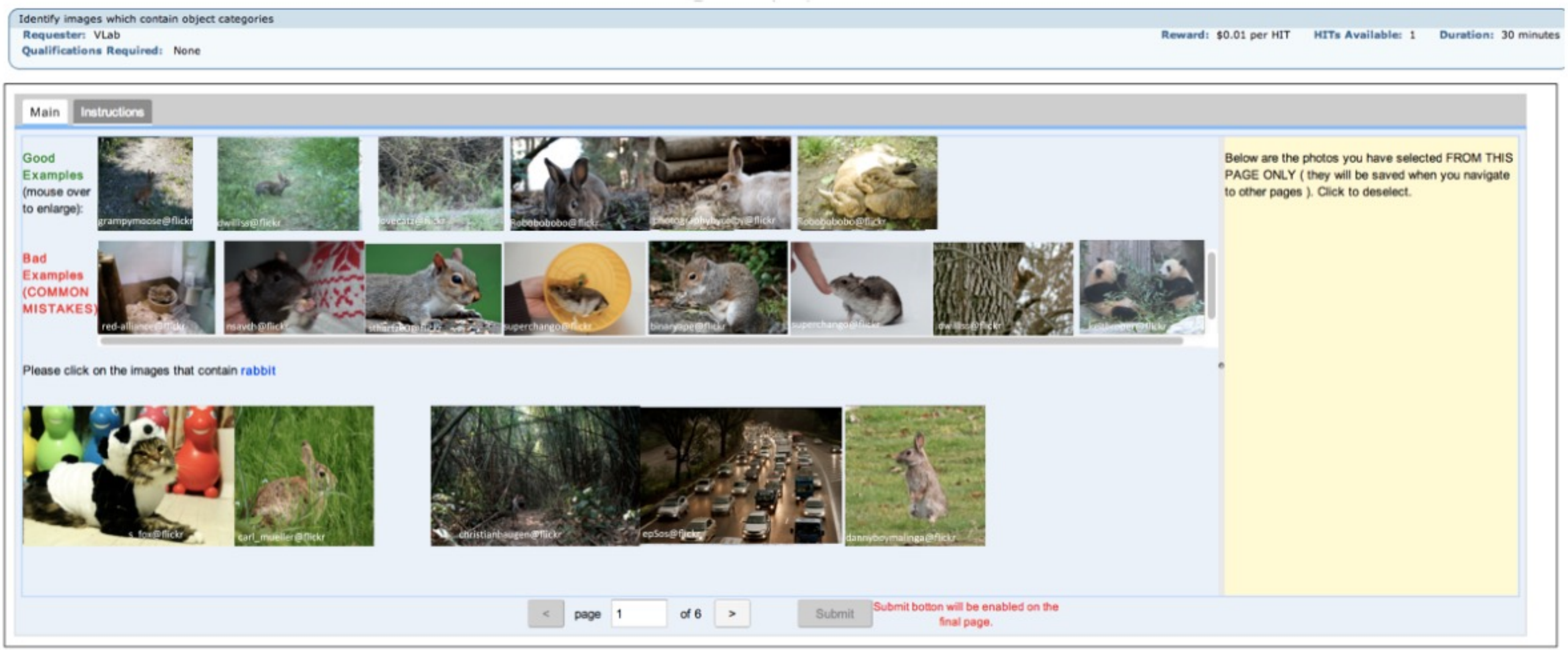
- Additional images from Flickr

3. Image Verification

- Crowdsource assigning all relevant categories from 200 object categories to each image

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, Li Fei-Fei , IJCV 2015

Recall from ImageNet: Object Presence Labeling



Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, Li Fei-Fei , IJCV 2015

ILSVRC

1. Category Selection

- 200 ImageNet classes which:
1) exclude synset overlap
2) exclude object classes too “big” in the image
3) are basic-level categories
4) backward compatible: VOC

2. Image Collection

- Subset of images from ImageNet
- Additional images from Flickr

3. Image Verification

- Crowdsource assigning all relevant categories from 200 object categories to each image

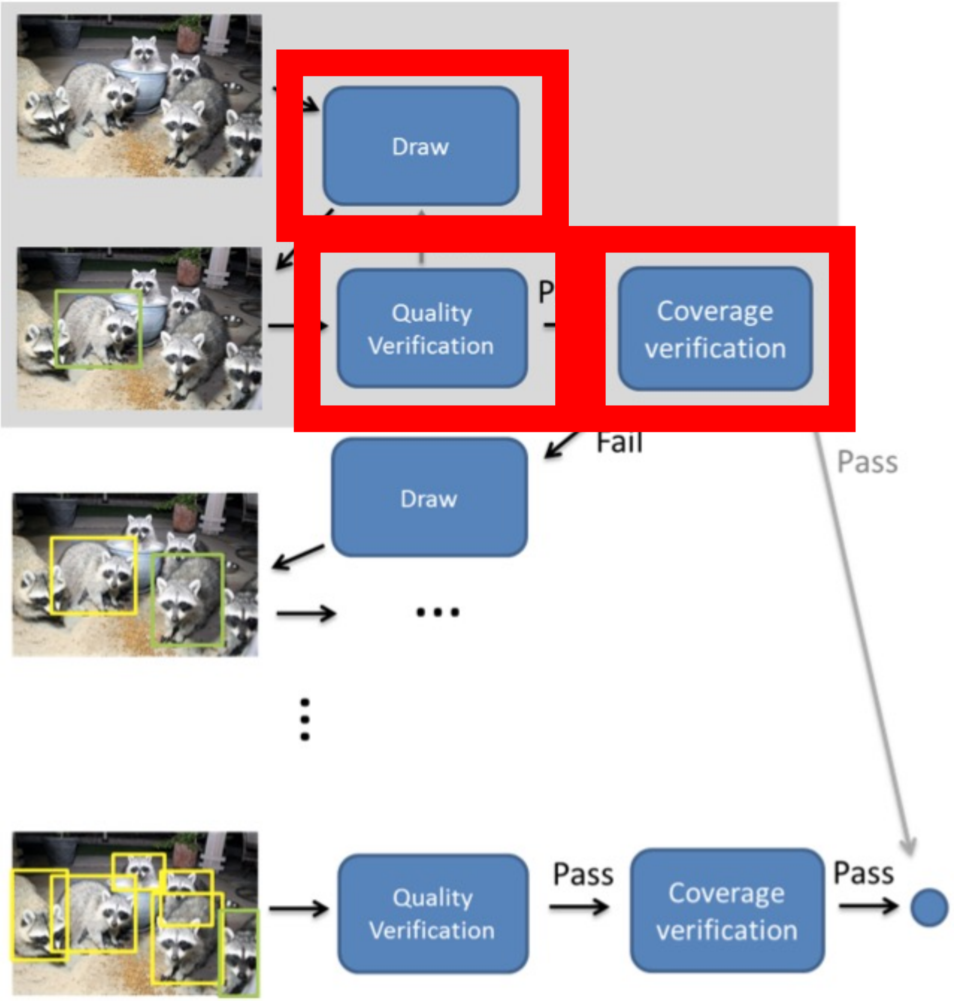
4. Image Annotation

- Crowdsource demarcating a bounding box around EVERY instance of every object category

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, Li Fei-Fei , IJCV 2015

ILSVRC: Efficient Object Localization

- 3 Tasks:

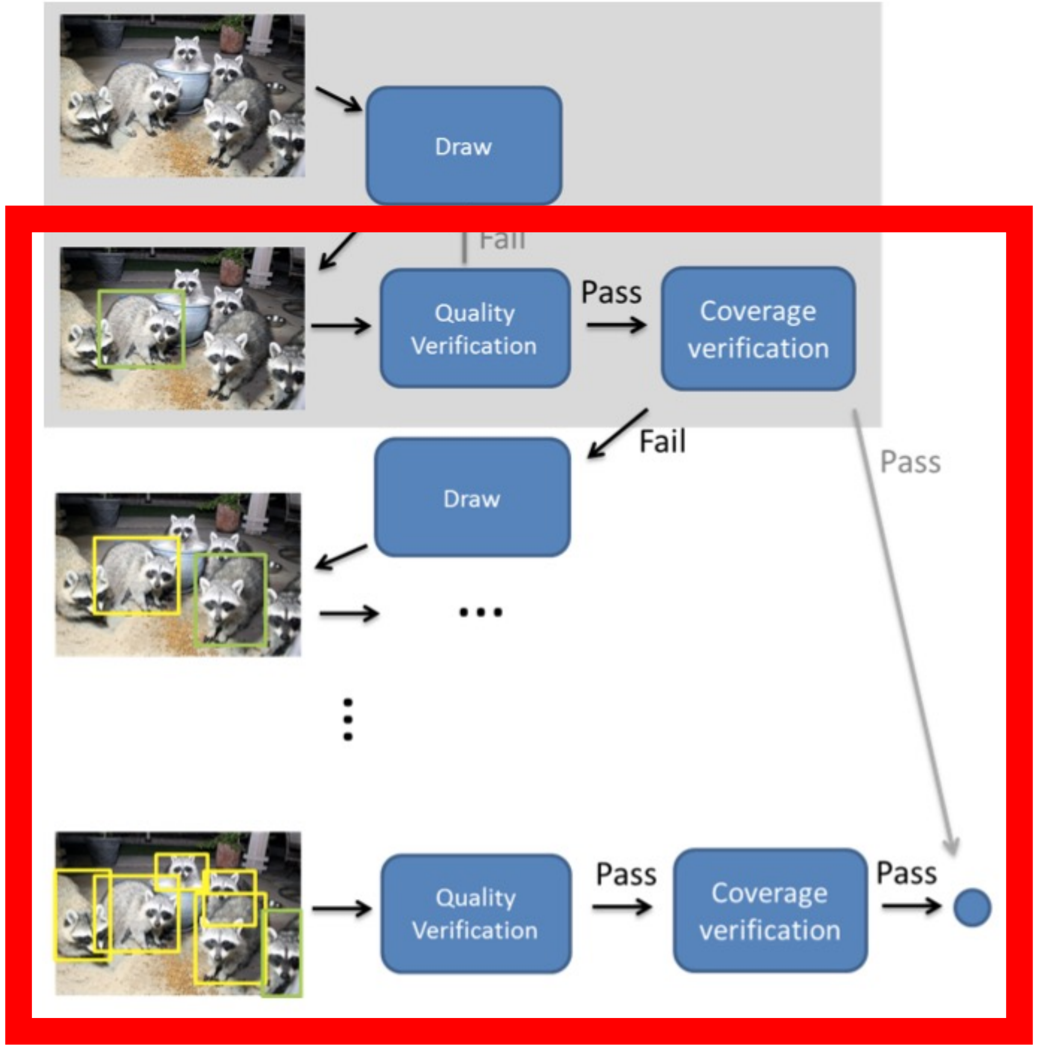


Idea: each task has fixed and predictable amount of work

Hao Su, Jia Deng, and Li Fei-Fei. Crowdsourcing Annotations for Visual Object Detection. AAAI 2012.

ILSVRC: Efficient Object Localization

- 3 Tasks:

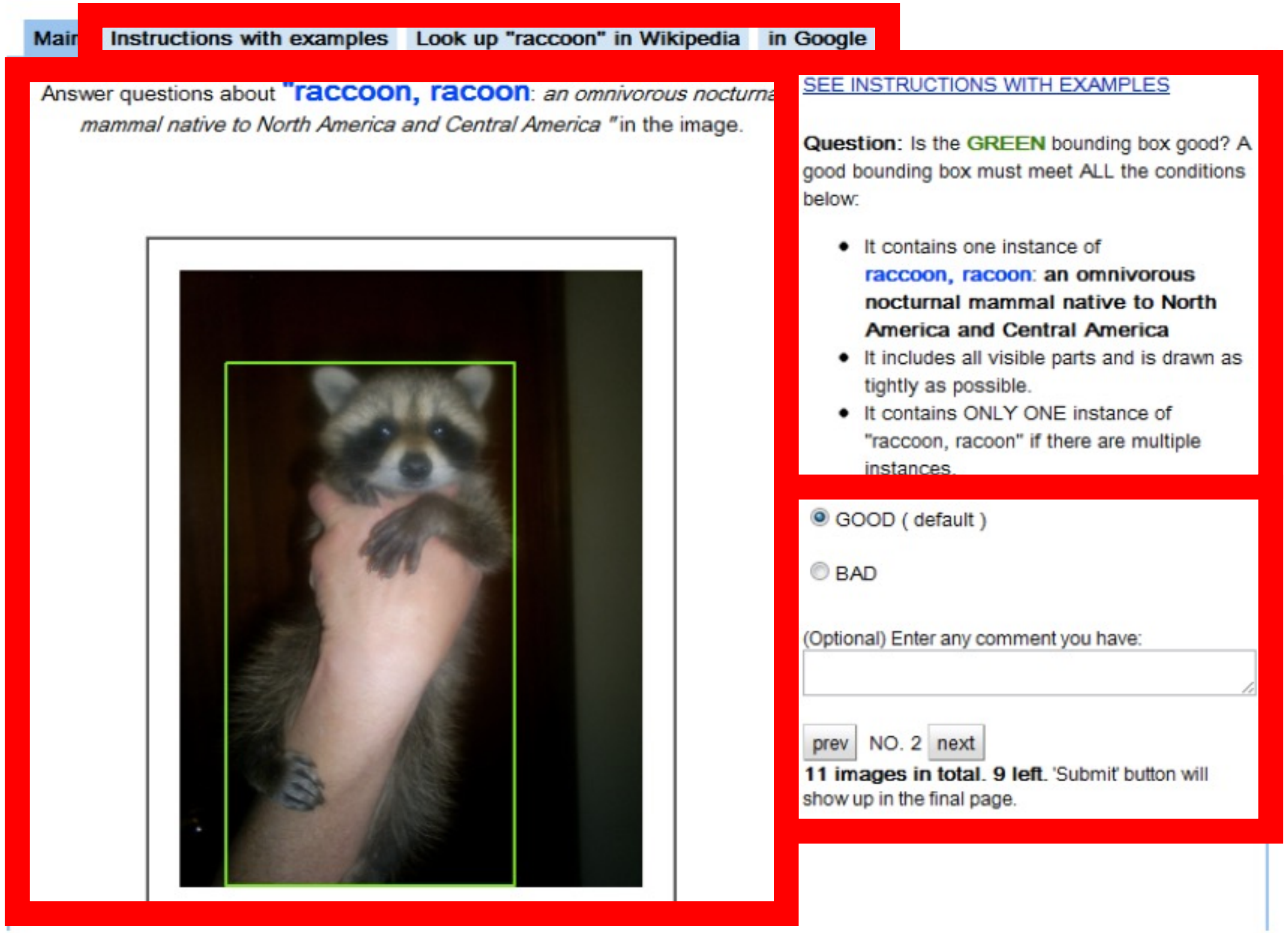


Hao Su, Jia Deng, and Li Fei-Fei. Crowdsourcing Annotations for Visual Object Detection. AAAI 2012.

ILSVRC: Drawing Task

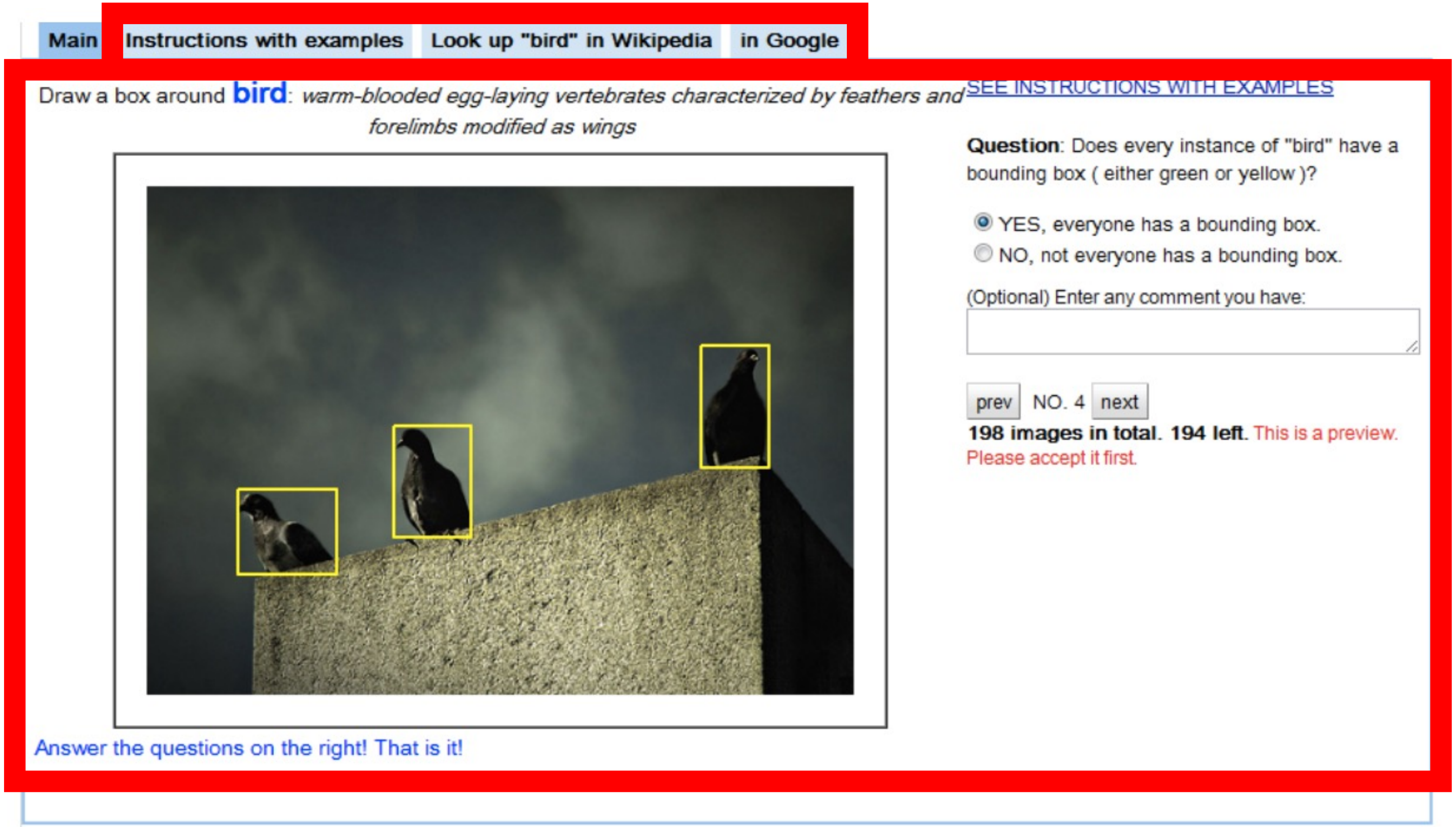
Hao Su, Jia Deng, and Li Fei-Fei. Crowdsourcing Annotations for Visual Object Detection. AAAI 2012.

ILSVRC: Quality Verification Task



Hao Su, Jia Deng, and Li Fei-Fei. Crowdsourcing Annotations for Visual Object Detection. AAAI 2012.

ILSVRC: Coverage Verification Task



Hao Su, Jia Deng, and Li Fei-Fei. Crowdsourcing Annotations for Visual Object Detection. AAAI 2012.

ILSVRC

1. Category Selection

- 200 ImageNet classes which:
1) exclude synset overlap
2) exclude object classes too “big” in the image
3) are basic-level categories
4) backward compatible: VOC

2. Image Collection

- Train dataset: 3 sources

- Val & test datasets: 2 sources

3. Object presence labeling

- Crowdsource assigning all relevant categories from 200 object categories to each image

4. Object localization

- Crowdsource demarcating a bounding box around EVERY instance of every object category

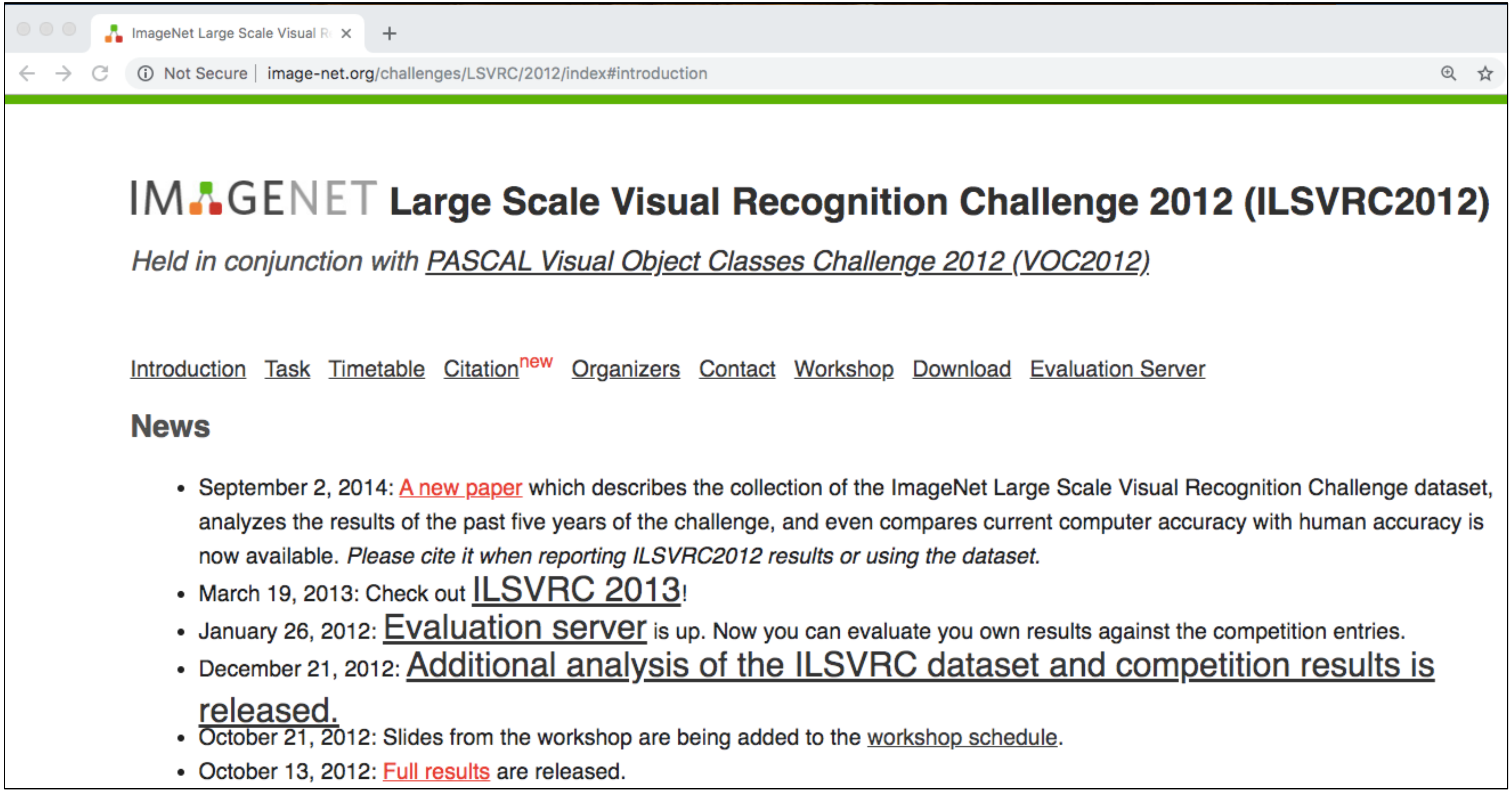
5. Author Review

- Ambiguous objects: BB for two categories with large overlap(~3%).

- Duplicates: >50% overlap for same object (~1%).

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, Li Fei-Fei , IJCV 2015

Object Detection: ILSVRC Annual Workshop

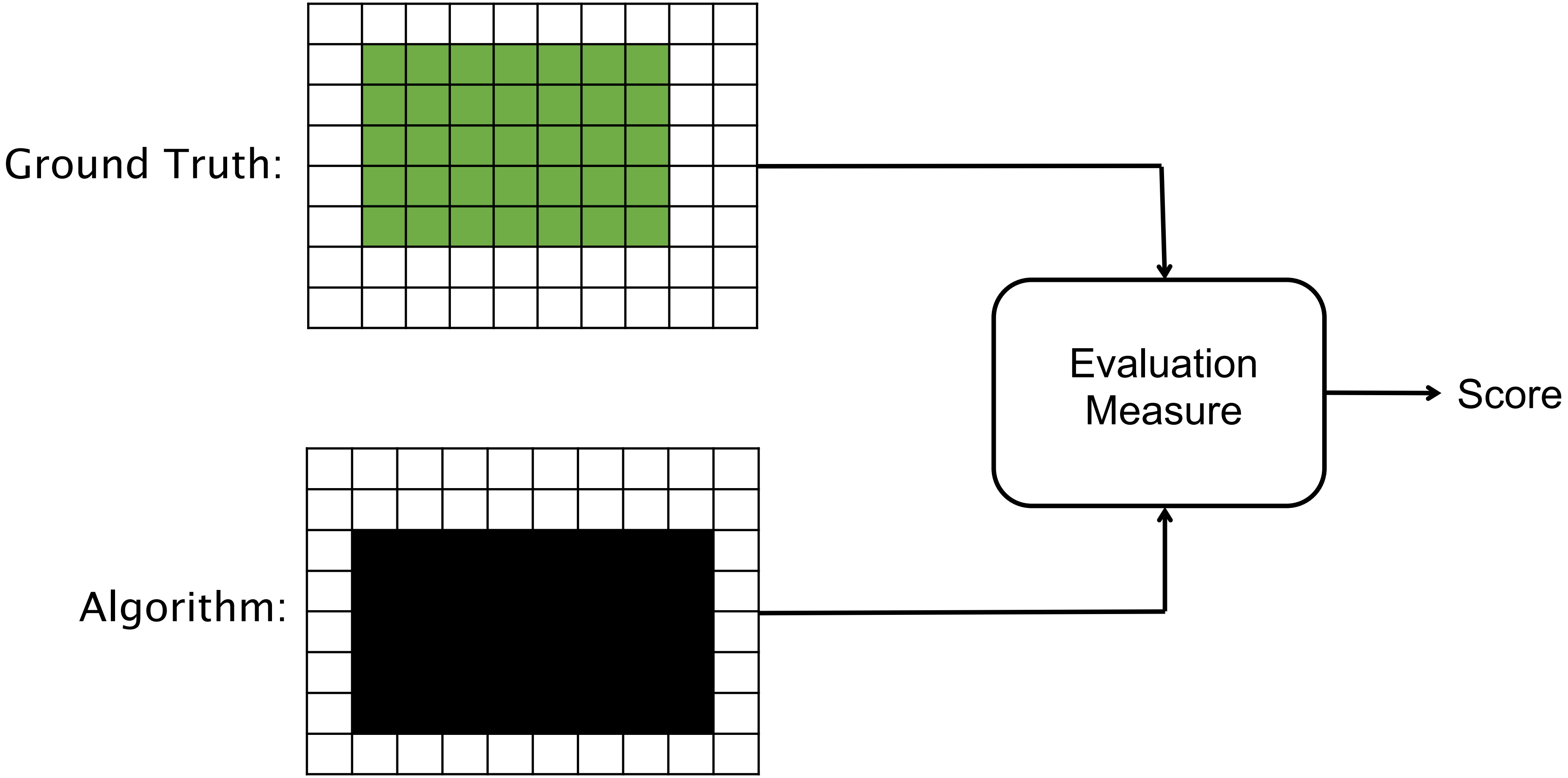


http://image-net.org/challenges/LSVRC/2012/index#introduction

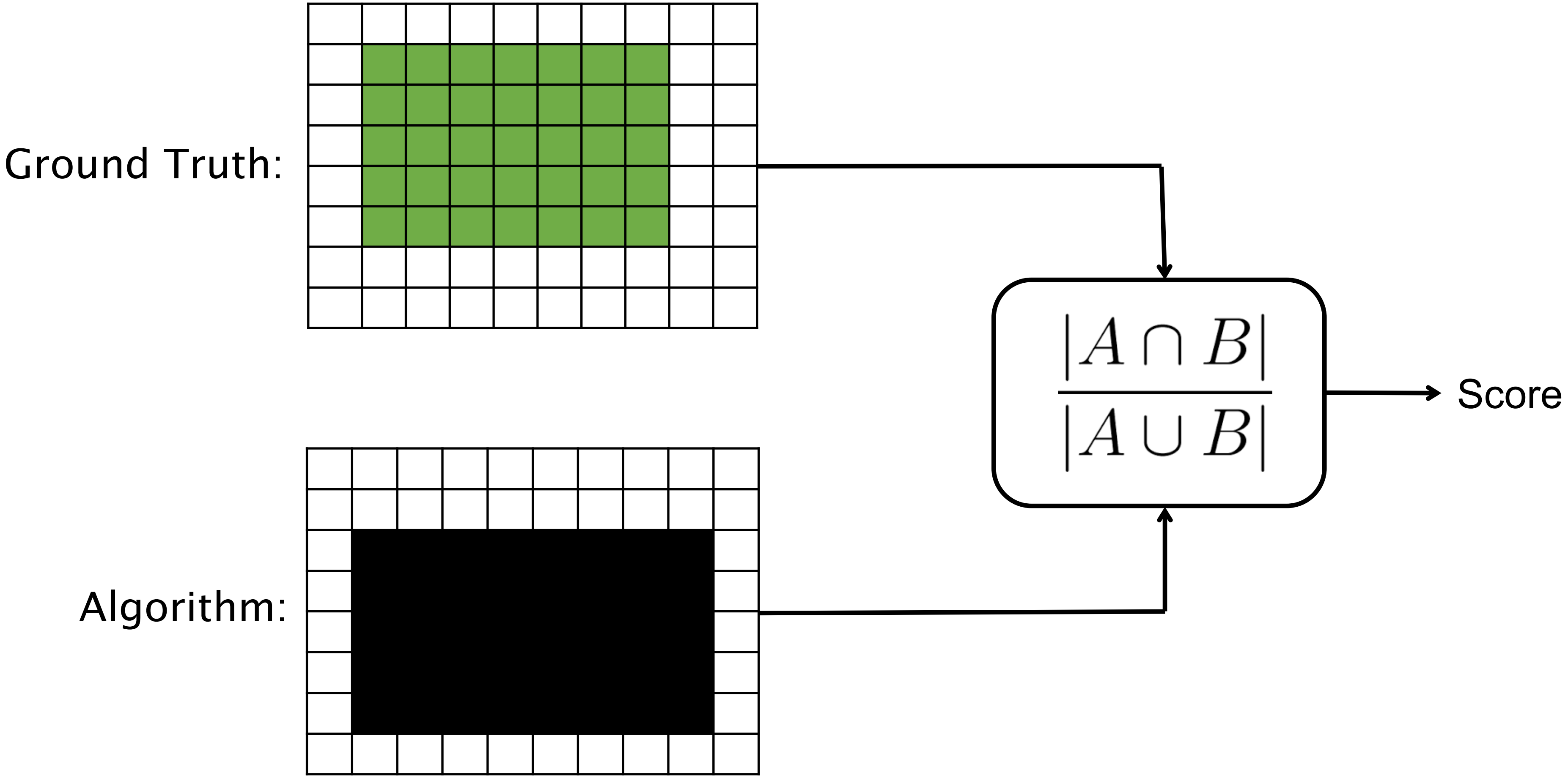
Object Detection: Today’s Topics

- Problem
- Applications
- Datasets
- Evaluation metric
- Overview of object detection algorithms and baseline (R-CNN)

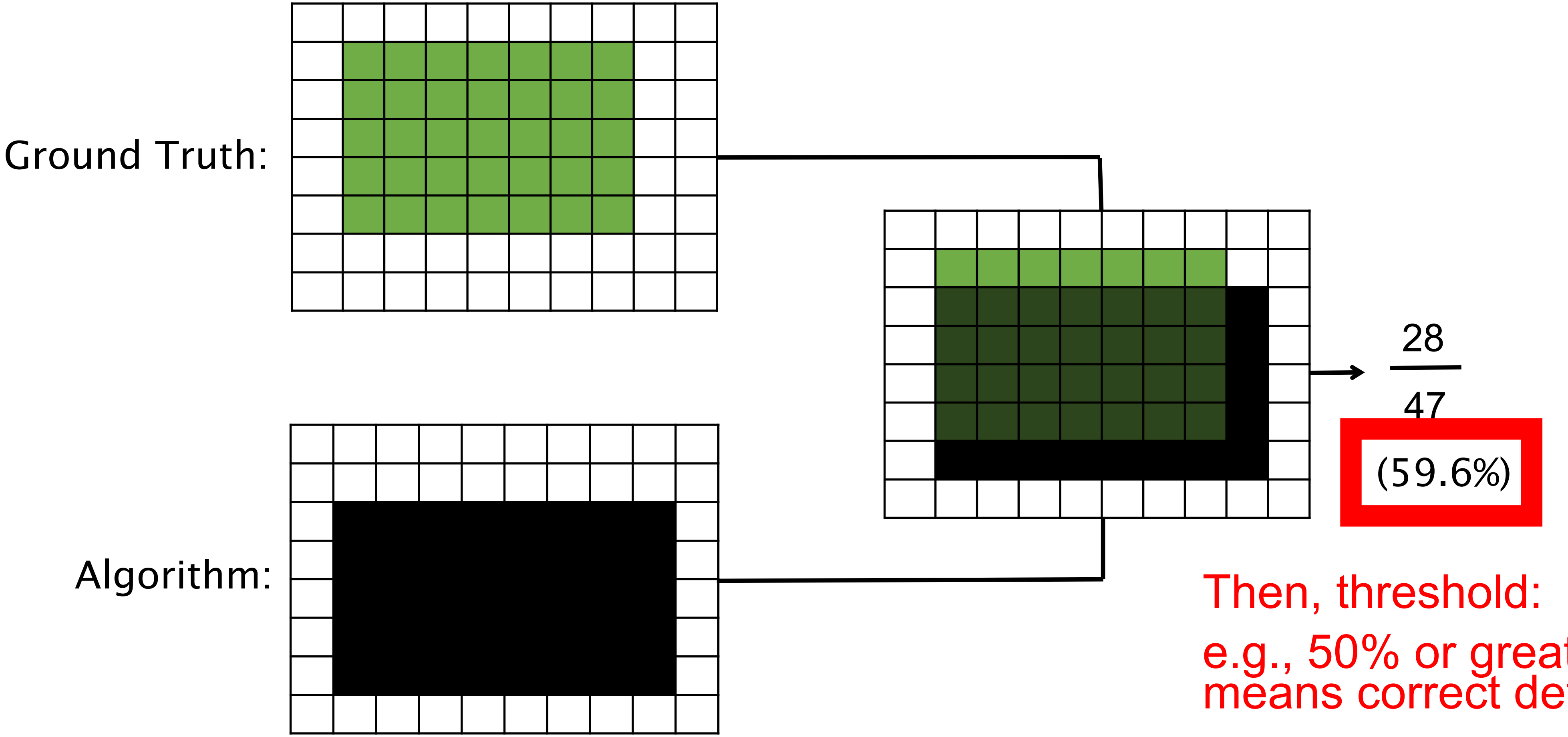
Single Object



Single Object: IoU (Intersection Over Union)



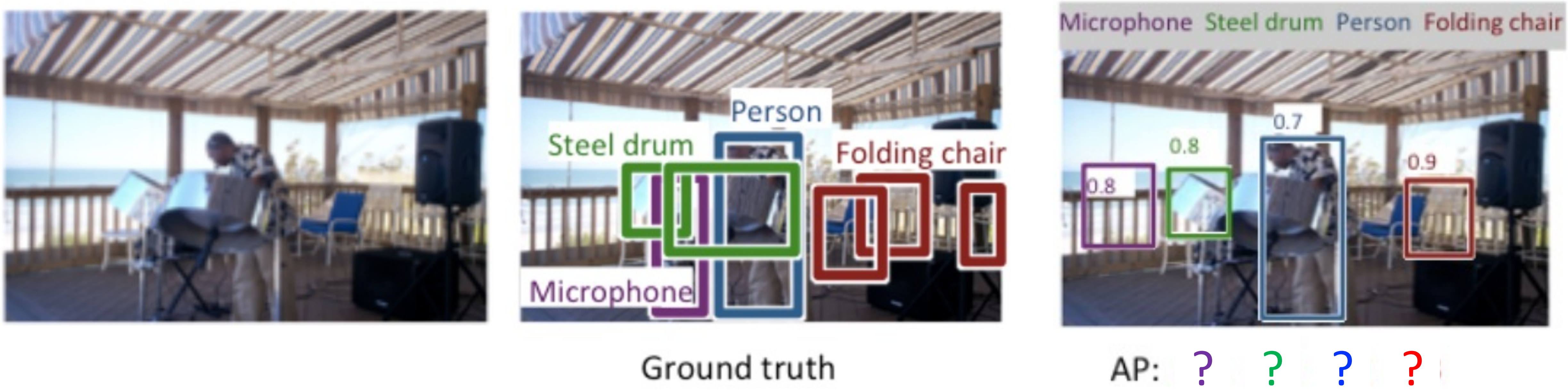
Single Object: IoU (Intersection Over Union)



Then, threshold:
e.g., 50% or greater means correct detection!

Evaluation Metric Basics: Precision

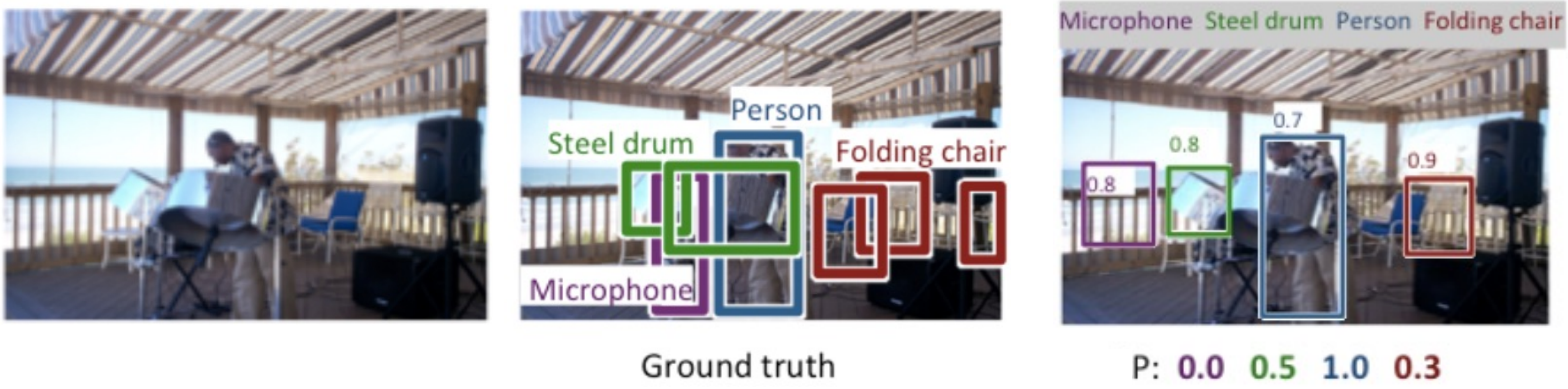
- For each object class (e.g., cat, dog, …), compute precision: fraction of correct detections from all detections using 0.5 IoU threshold



[Russakovsky et al; IJCV 2015]
https://jonathan-hui.medium.com/map-mean-average-precision-for-object-detection-45c121a31173

Evaluation Metric Basics: Precision

- For each object class (e.g., cat, dog, …), compute precision: fraction of correct detections from all detections using IoU threshold (e.g., 0.5)



[Russakovsky et al; IJCV 2015]
https://jonathan-hui.medium.com/map-mean-average-precision-for-object-detection-45c121a31173

Evaluation Metric: mAP

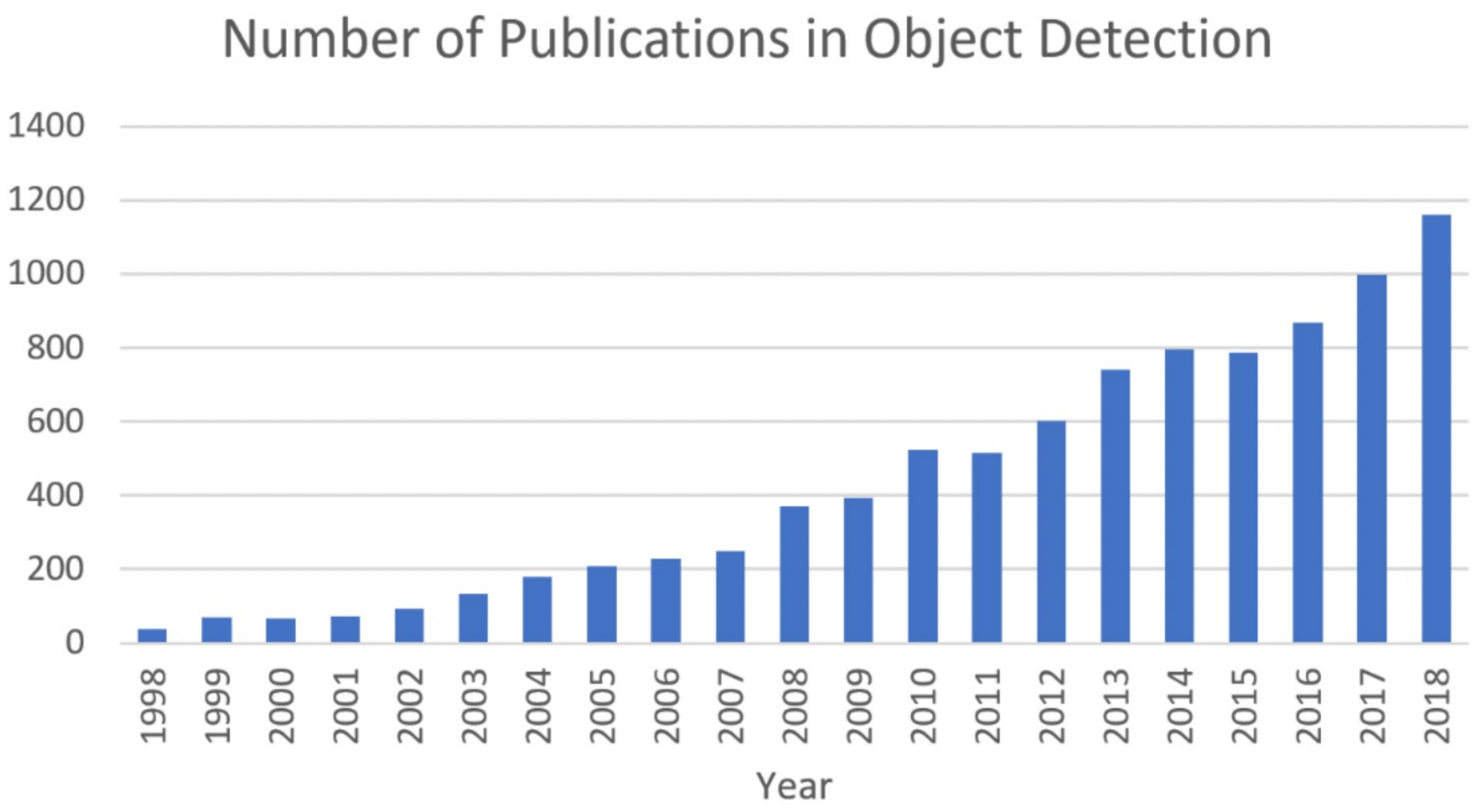
- For each object class (e.g., cat, dog, …), compute Average Precision (AP)
  - Vary IoU threshold in order to create a precision-recall curve, and then computer compute area under the curve
- Then, compute mean AP across all classes

[Russakovsky et al; IJCV 2015]
https://jonathan-hui.medium.com/map-mean-average-precision-for-object-detection-45c121a31173

Object Detection: Today’s Topics

- Problem
- Applications
- Datasets
- Evaluation metric
- Overview of object detection algorithms and baseline (R-CNN)

Community Research Engagement



"Data from Google scholar advanced search: allintitle: 'object detection' AND 'detecting objects'"

Zhou et al. Object Detection in 20 Years: A Survey. arXiv 2019.

Naïve Solution: Sliding Window Approach

Person?
Person?
Person?
Person?
Person?
Person?
Person?
Person?
Person?

Image Source: https://yourboulder.com/boulder-neighborhood-downtown/

Naïve Solution: Sliding Window Approach

Car?
Car?
Car?
Car?
Car?
Car?
Car?
Car?
Car?

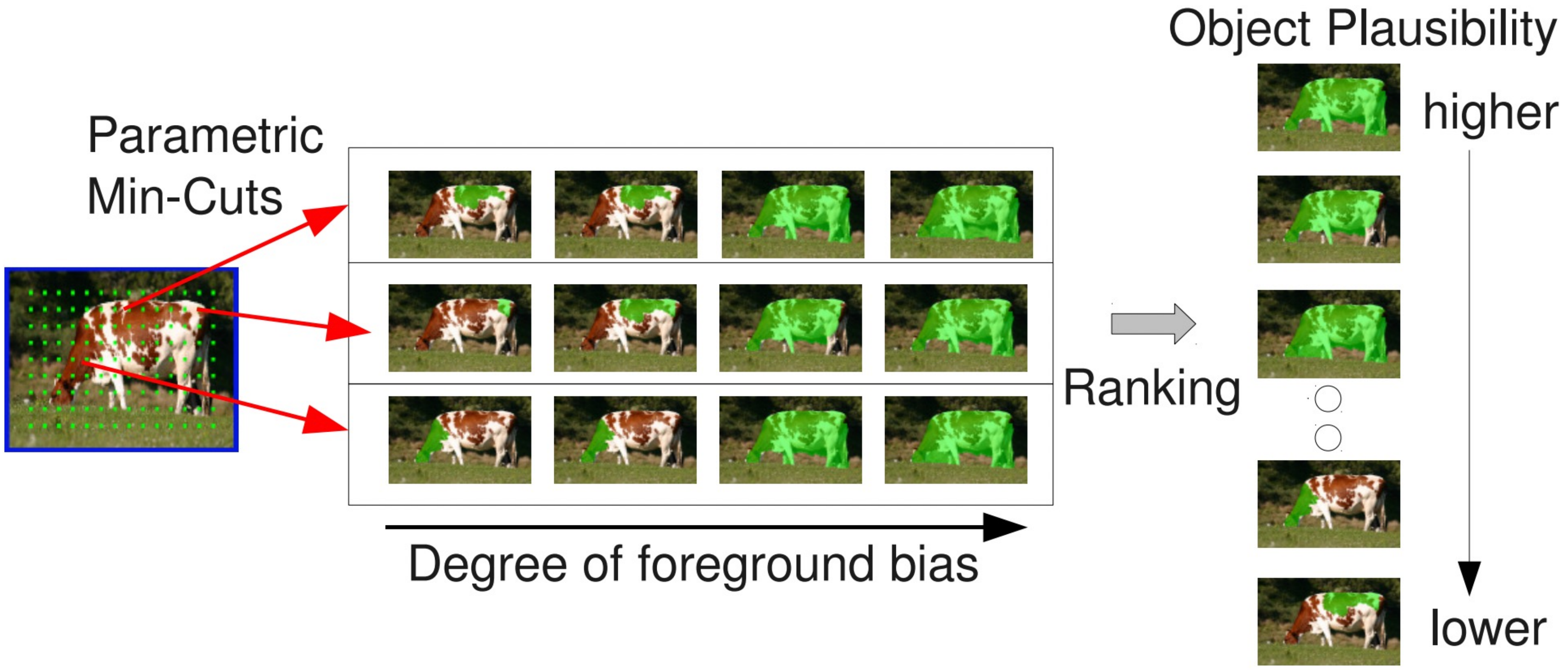
Image Source: https://yourboulder.com/boulder-neighborhood-downtown/

Naïve Solution: Sliding Window Approach

- Sliding window approach: must test different locations at…
  - Different scales
  - Different aspect ratios (e.g., for person vs car or car viewed at different angles)
- Number of regions to test? (e.g., 1920 x 1080 image)
  - Easily can explode to hundreds of thousands or millions of windows
- Key limitation
  - Very slow!

A Less Naïve Solution: Region Proposals

- Replace sliding window approach with *region proposals* (bounding boxes around "object"-like regions) found by grouping similar pixels based on low-/mid-level features; e.g., CPMC



Carreira and Sminchisescu. Constrained Parametric Min-Cuts for Automatic Object Segmentation. CVPR 2010.

A Less Naïve Solution: Region Proposals

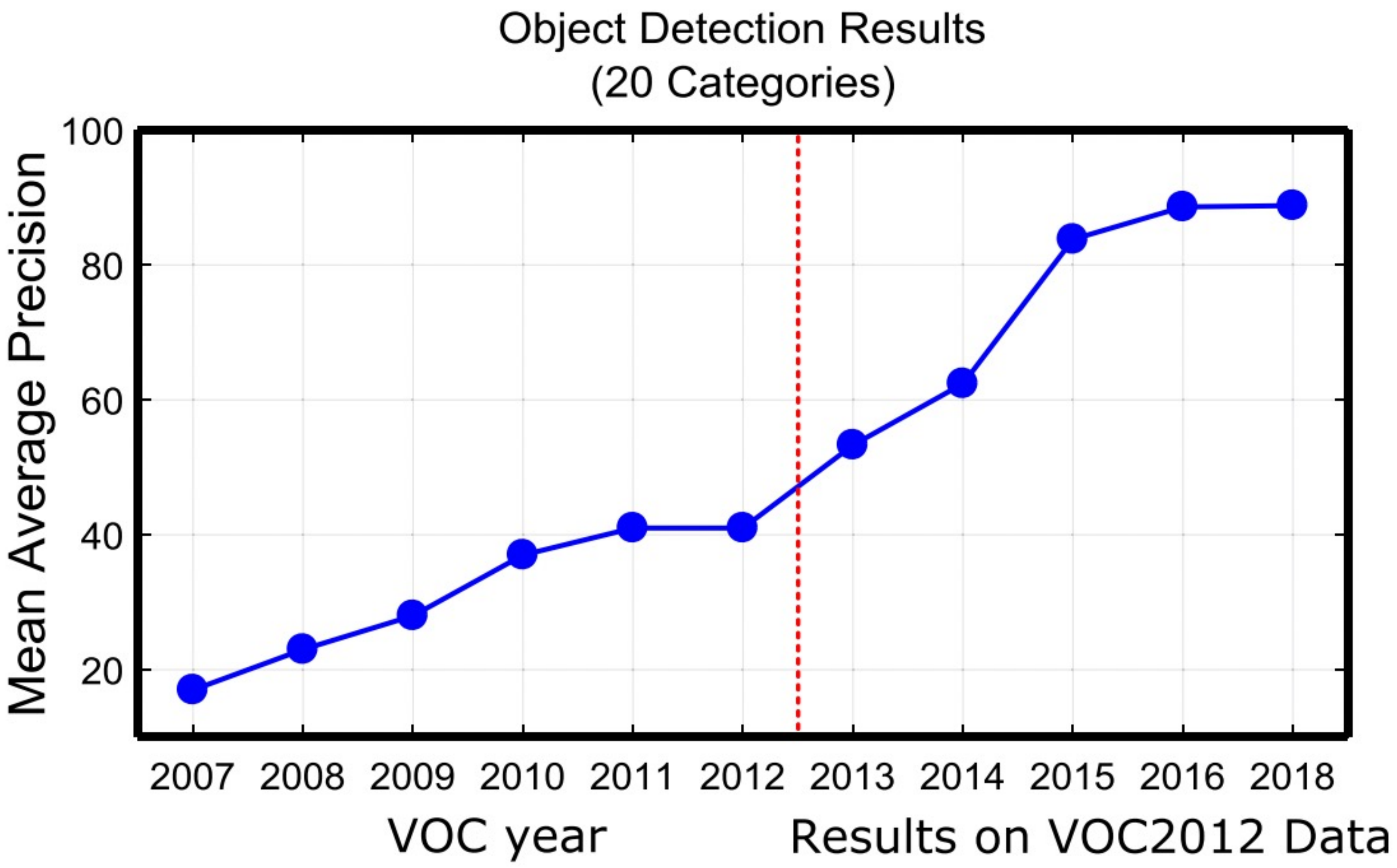
- Replace sliding window approach with *region proposals (*bounding boxes around "object"-like regions)*;* e.g., CPMC

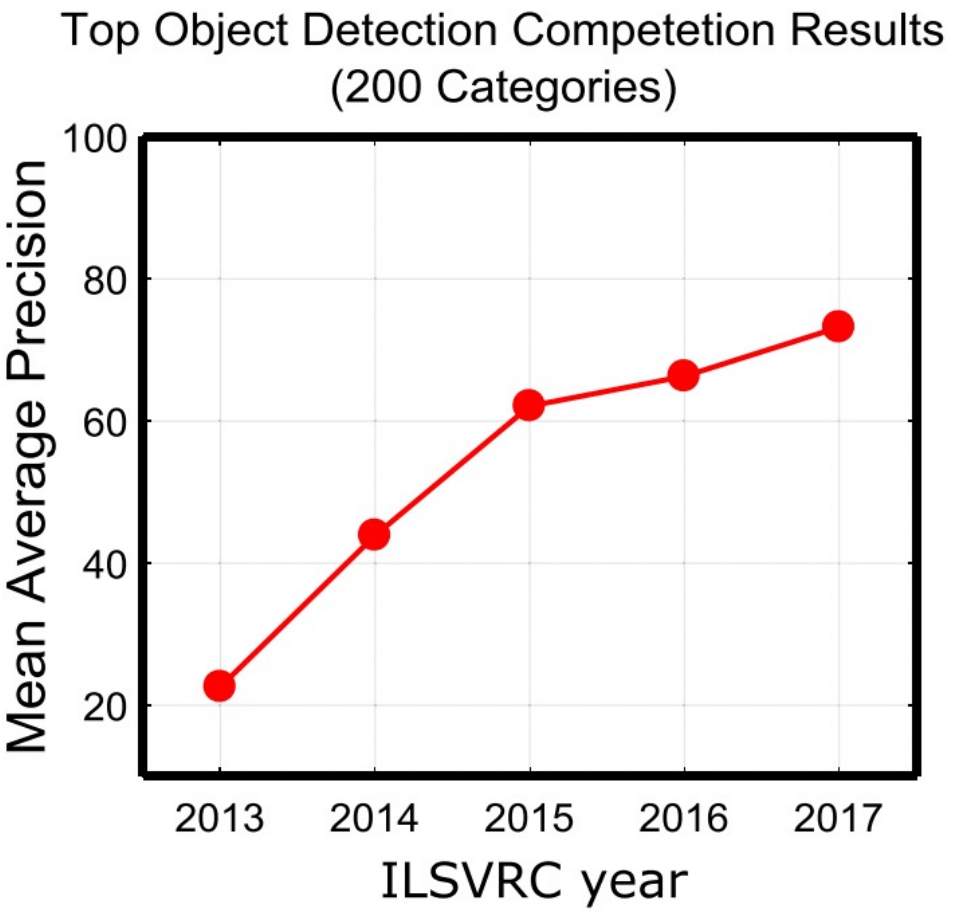
- **Advantage**: considerably fewer regions than needed in a naïve sliding window approach, with belief they will include the objects of interest (i.e., high recall)

- **Many options**: CPMC, Category Independent Object Proposals, Randomized Prim, Selective Search, and more

- A good start…

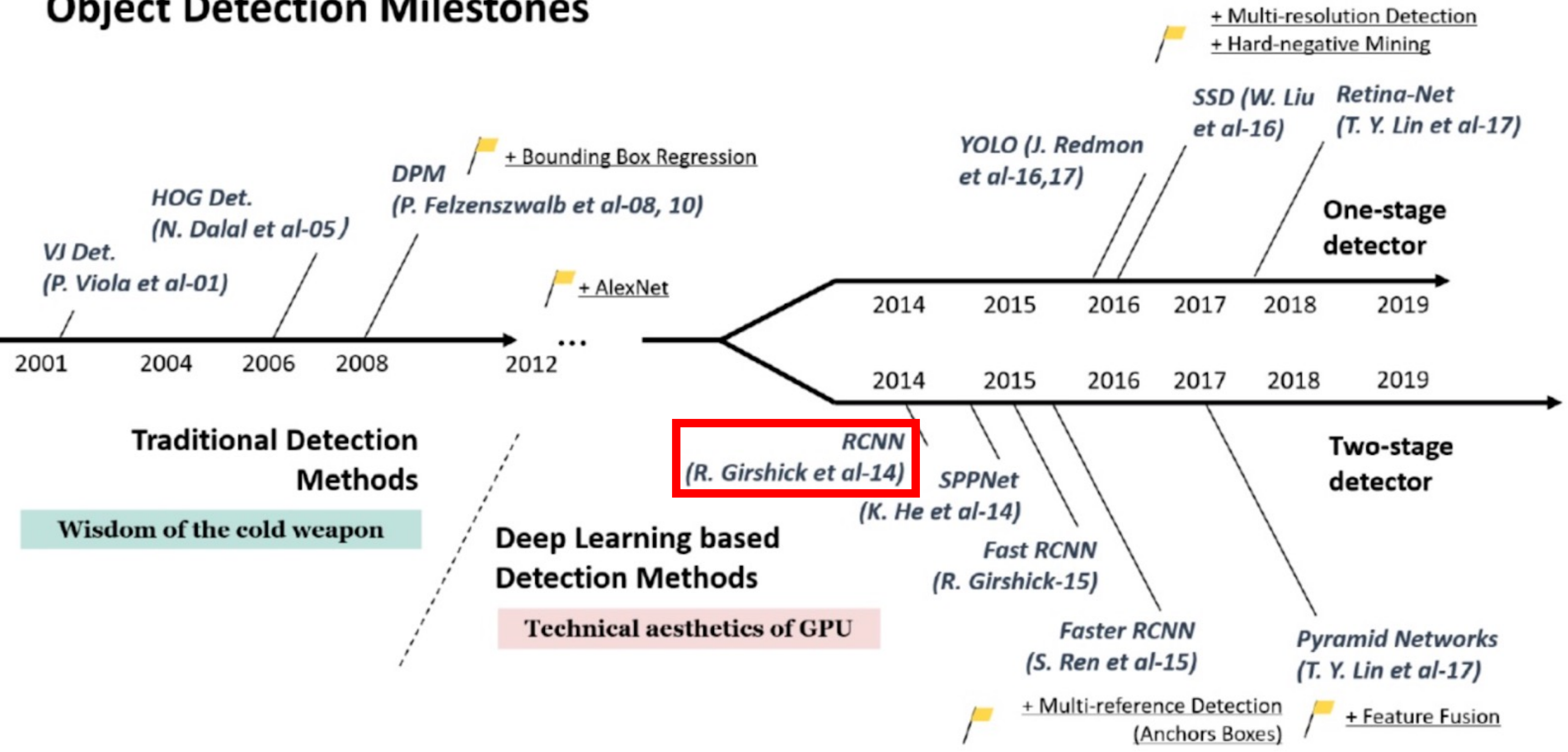
Turning Point (2012): Deep Learning Solutions





Li Liu et al. “Deep Learning for Generic Object Detection: A Survey.” IJCV 2019

Object Detection Milestones



Zhou et al. Object Detection in 20 Years: A Survey. arXiv 2019.

Why R-CNN?

Named after the proposed technique: use **R**egion proposals with **CNN** features

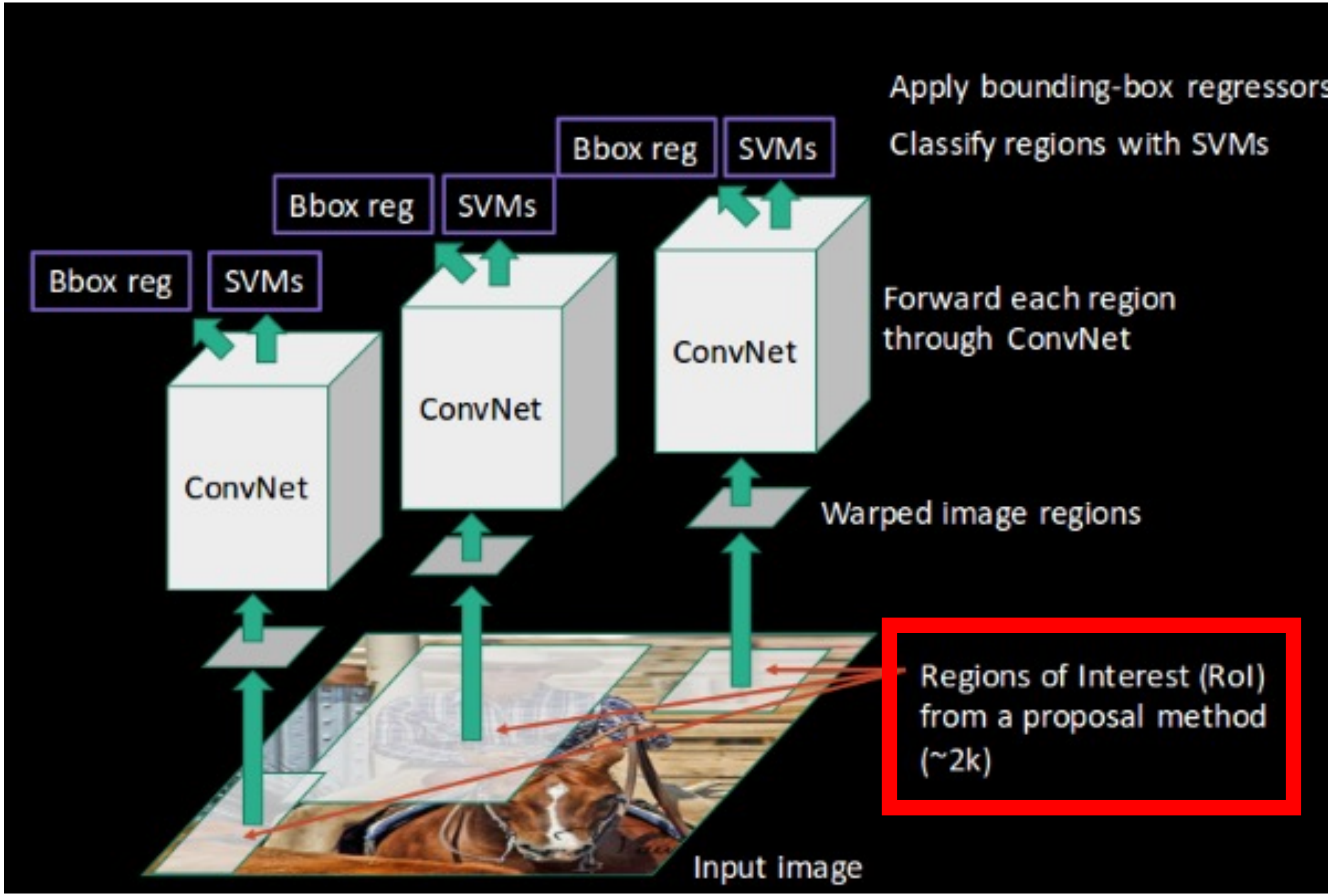
Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. “Rich feature hierarchies for accurate object detection and semantic segmentation.” CVPR 2014.

Key Contributions of R-CNN

1. Demonstrate how to accurately localize objects with a neural network (NN)
   - First time a CNN outperformed hand-crafted features on VOC, achieving mAP of 54% compared to 33% for previous HOG based model (VOC 2010)

2. Demonstrate how train an accurate (high-capacity) NN with a scarce amount of annotated detection data

Architecture



Selective Search used to enable comparison with prior work; creates ~2000 regions based on color, texture, size and shape

Figure Source: https://www.analyticsvidhya.com/blog/2018/10/a-step-by-step-introduction-to-the-basic-object-detection-algorithms-part-1/

Architecture

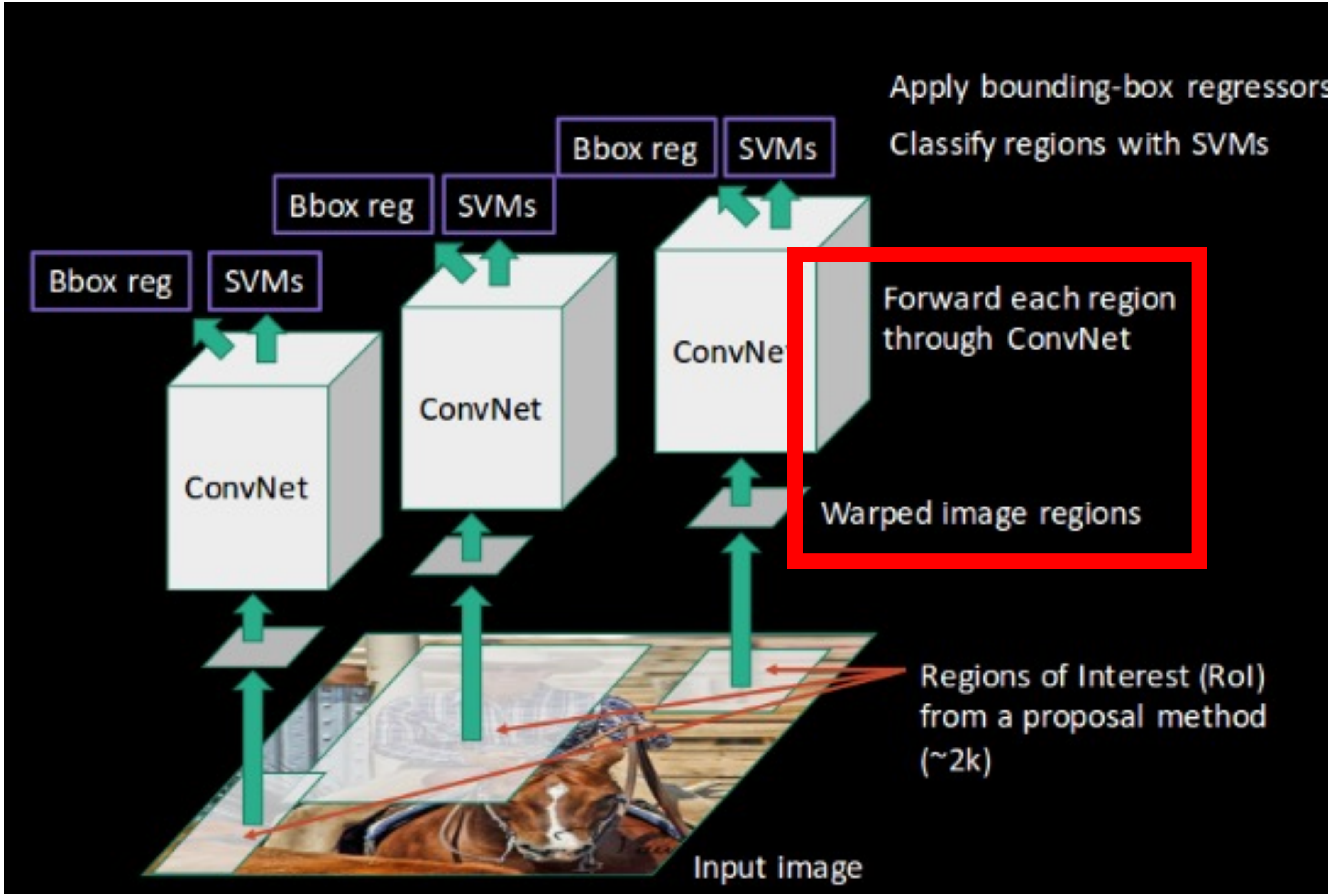


Figure Source: https://www.analyticsvidhya.com/blog/2018/10/a-step-by-step-introduction-to-the-basic-object-detection-algorithms-part-1/

Describe Each Region with Fixed-length Vector

Key idea: Given scarce amount of training data in detection datasets, devise good feature by fine-tuning model that is pre-trained on a large dataset

- Replace final layer of AlexNet (trained on ImageNet) with # of categories in (VOC) detection dataset

How many classes should be predicted?

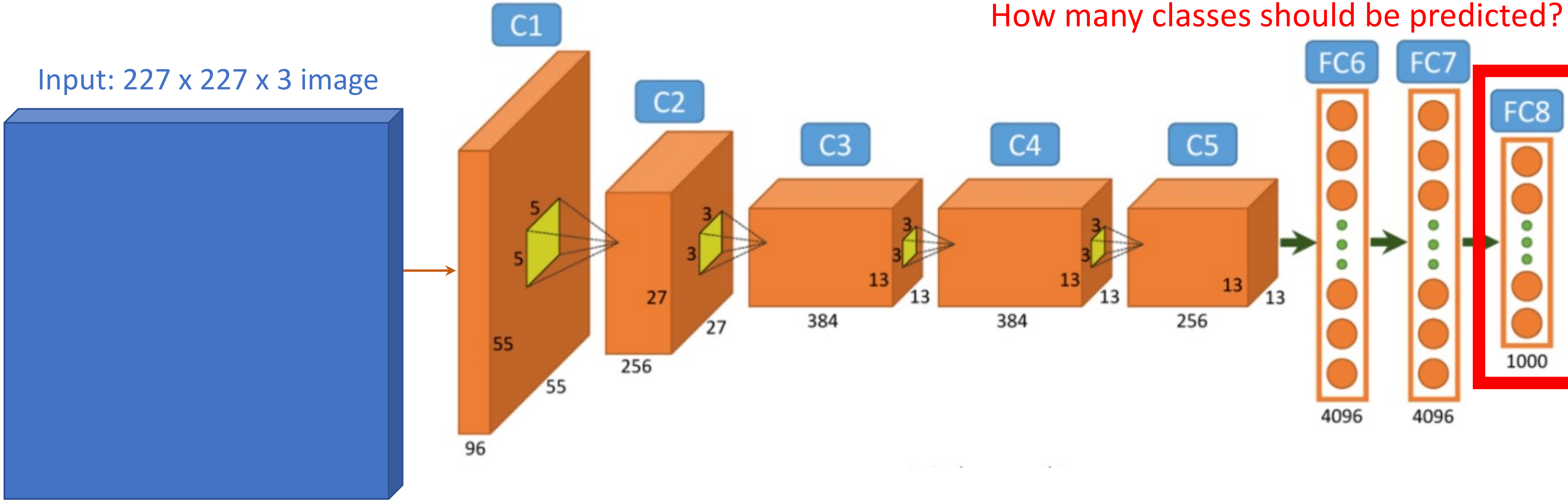


Image Source: https://www.researchgate.net/figure/Architecture-of-Alexnet-From-left-to-right-input-to-output-five-convolutional-layers_fig2_312303454

Describe Each Region with Fixed-length Vector

Key idea: Given scarce amount of training data in detection datasets, devise good feature by fine-tuning model that is pre-trained on a large dataset

- Replace final layer of AlexNet (trained on ImageNet) with # of categories in (ILSVRC) detection dataset

How many classes should be predicted?

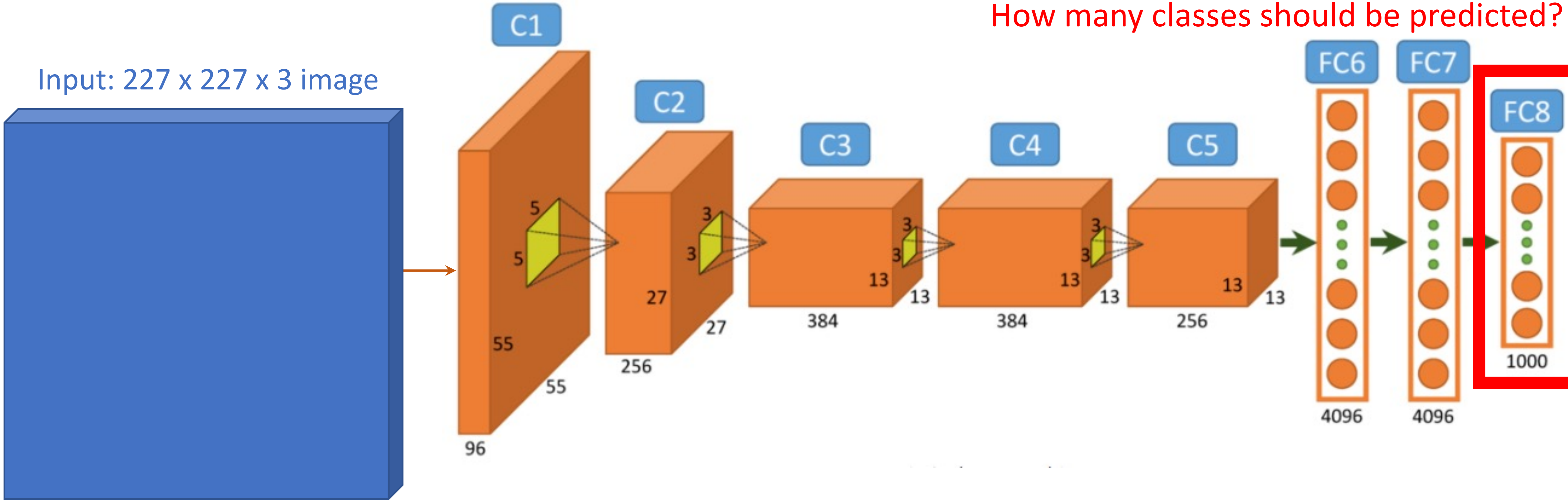


Image Source: https://www.researchgate.net/figure/Architecture-of-Alexnet-From-left-to-right-input-to-output-five-convolutional-layers_fig2_312303454

Describe Each Region with Fixed-length Vector

Key idea: Given scarce amount of training data in detection datasets, devise good feature by fine-tuning model that is pre-trained on a large dataset

- Replace final layer of AlexNet and use FC7 feature from fine-tuned model

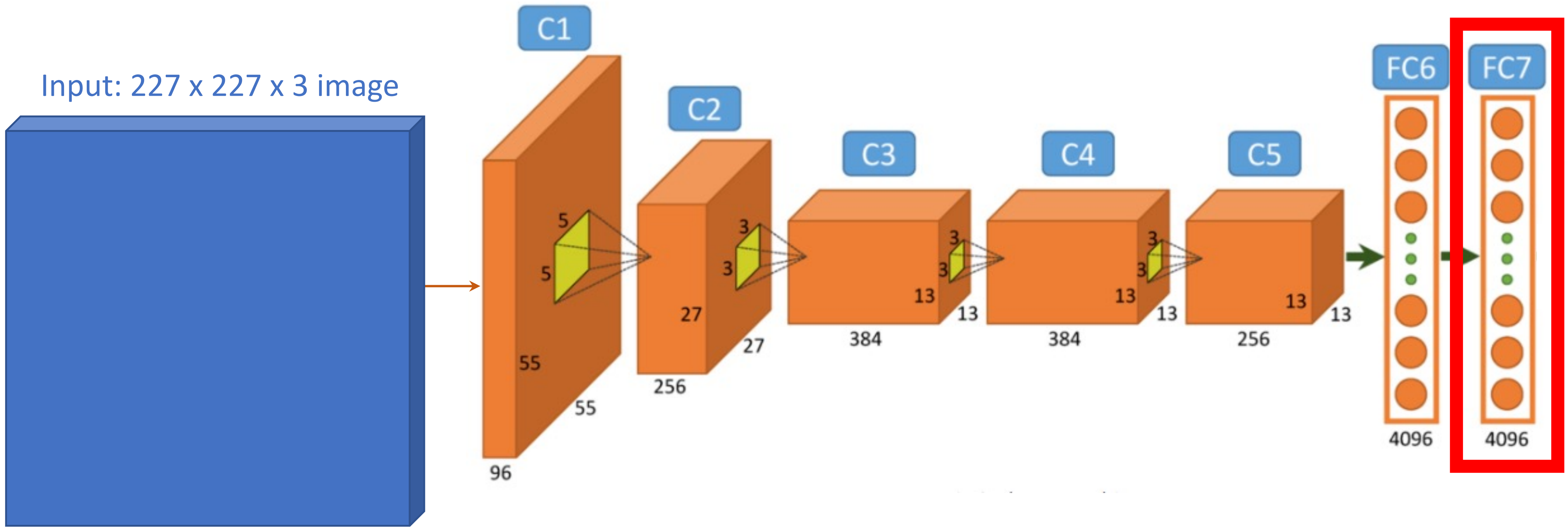


Image Source: https://www.researchgate.net/figure/Architecture-of-Alexnet-From-left-to-right-input-to-output-five-convolutional-layers_fig2_312303454

Describe Each Region with Fixed-length Vector

- Benefits of these features:
  - can be learned for a dataset instead of handcrafted (e.g., HOG, SIFT)
  - ~2 orders of magnitude smaller than traditional features (e.g., HOG, SIFT)

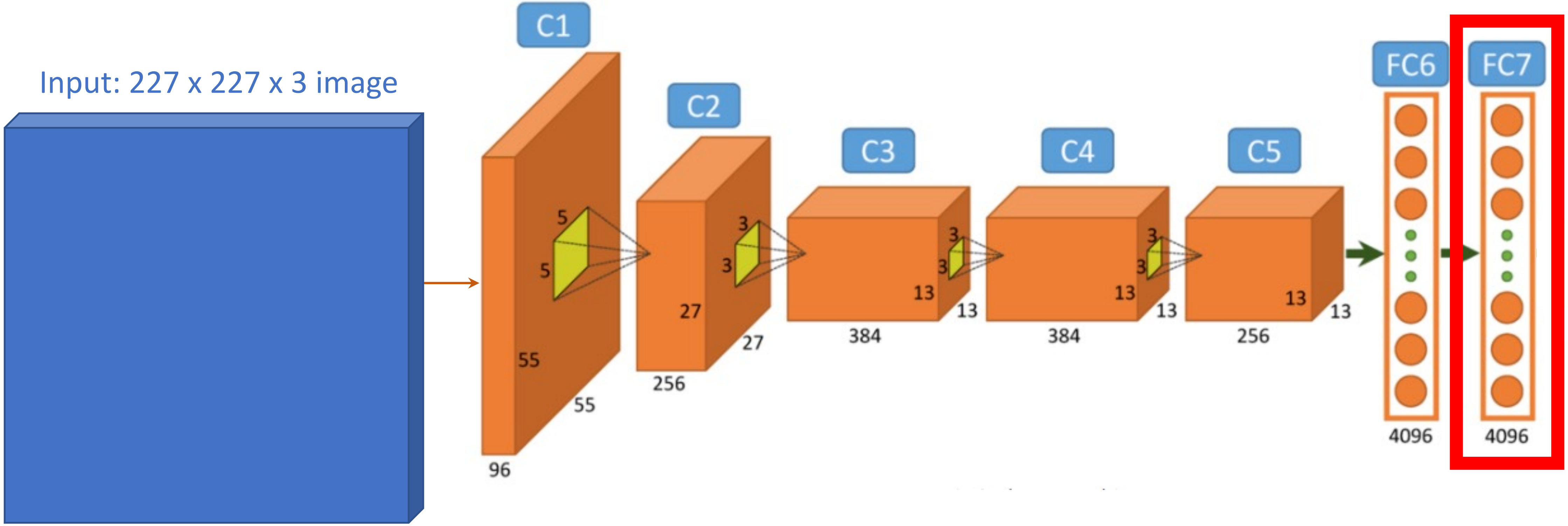
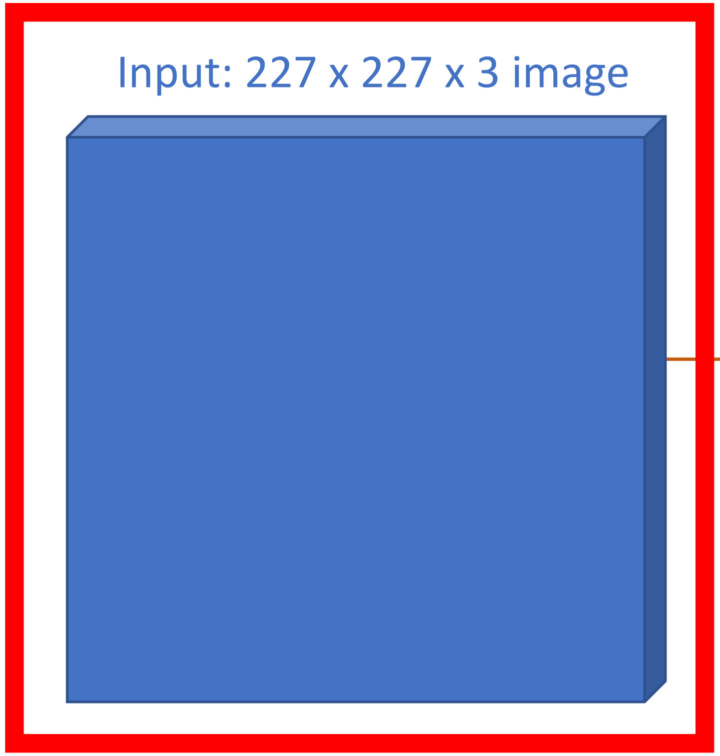


Image Source: https://www.researchgate.net/figure/Architecture-of-Alexnet-From-left-to-right-input-to-output-five-convolutional-layers_fig2_312303454

Describe Each Region with Fixed-length Vector

Challenge: how to resize a proposed region to the required size?



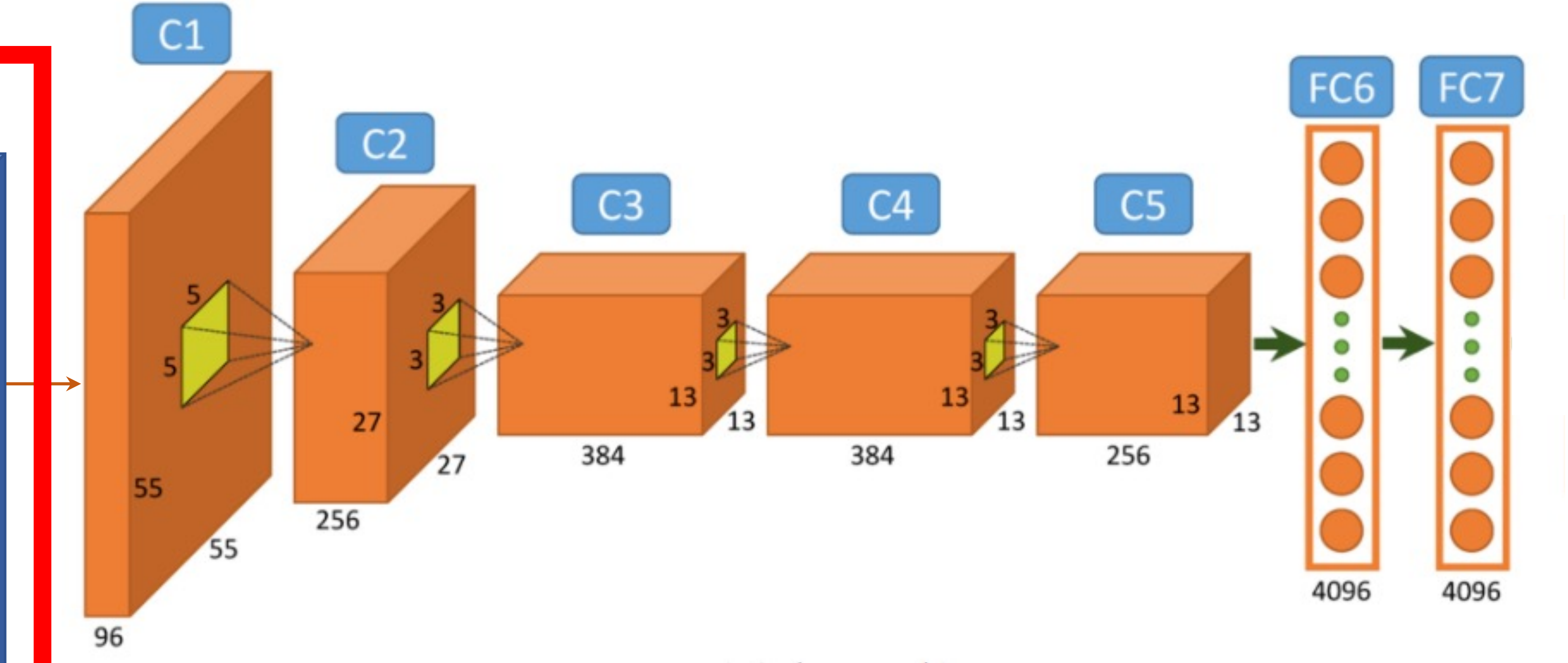


Image Source: https://www.researchgate.net/figure/Architecture-of-Alexnet-From-left-to-right-input-to-output-five-convolutional-layers_fig2_312303454

Describe Each Region with Fixed-length Vector

Challenge: how to resize a proposed region to the required size?

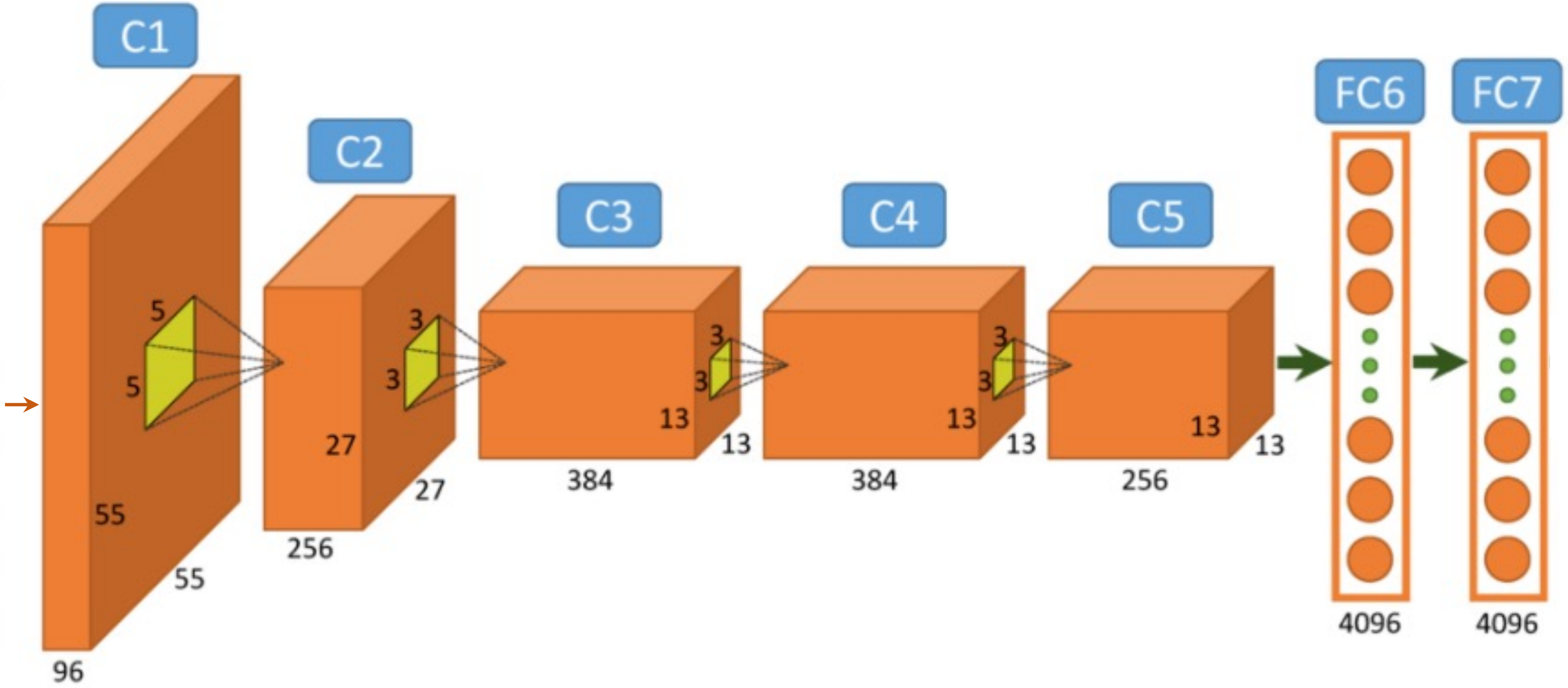


Image Source: https://www.researchgate.net/figure/Architecture-of-Alexnet-From-left-to-right-input-to-output-five-convolutional-layers_fig2_312303454

Describe Each Region with Fixed-length Vector

Challenge: how to resize a proposed region to the required size?

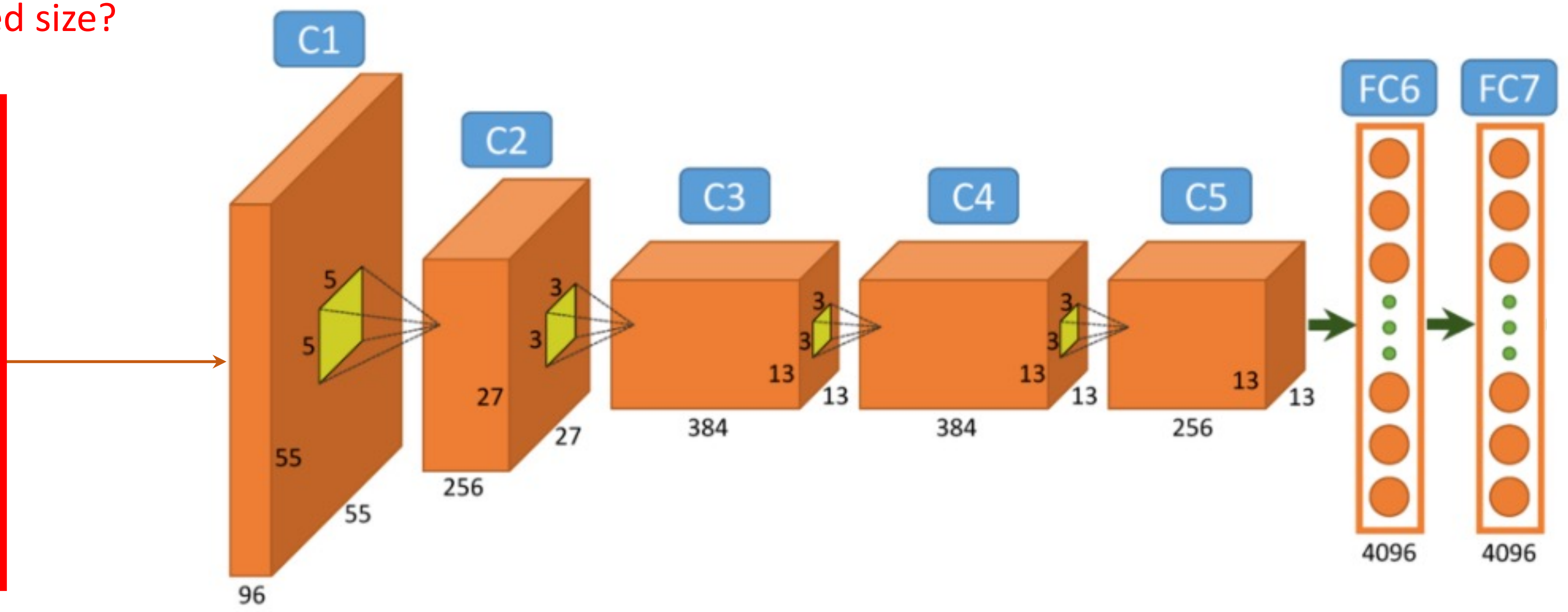


Image Source: https://www.researchgate.net/figure/Architecture-of-Alexnet-From-left-to-right-input-to-output-five-convolutional-layers_fig2_312303454

Region Resizing

As exemplified, region proposals come in different sizes and aspect ratios

Region Resizing

(1) (2) (3)

Scale region and then pad with original image context or constant

Anisotropic scaling

Chosen because experimentally shown to lead to the best results

Many ways to convert a region into a fixed input size of 227 x 227 x 3

Describe Each Region with Fixed-length Vector

Input: 227 x 227 x 3 image

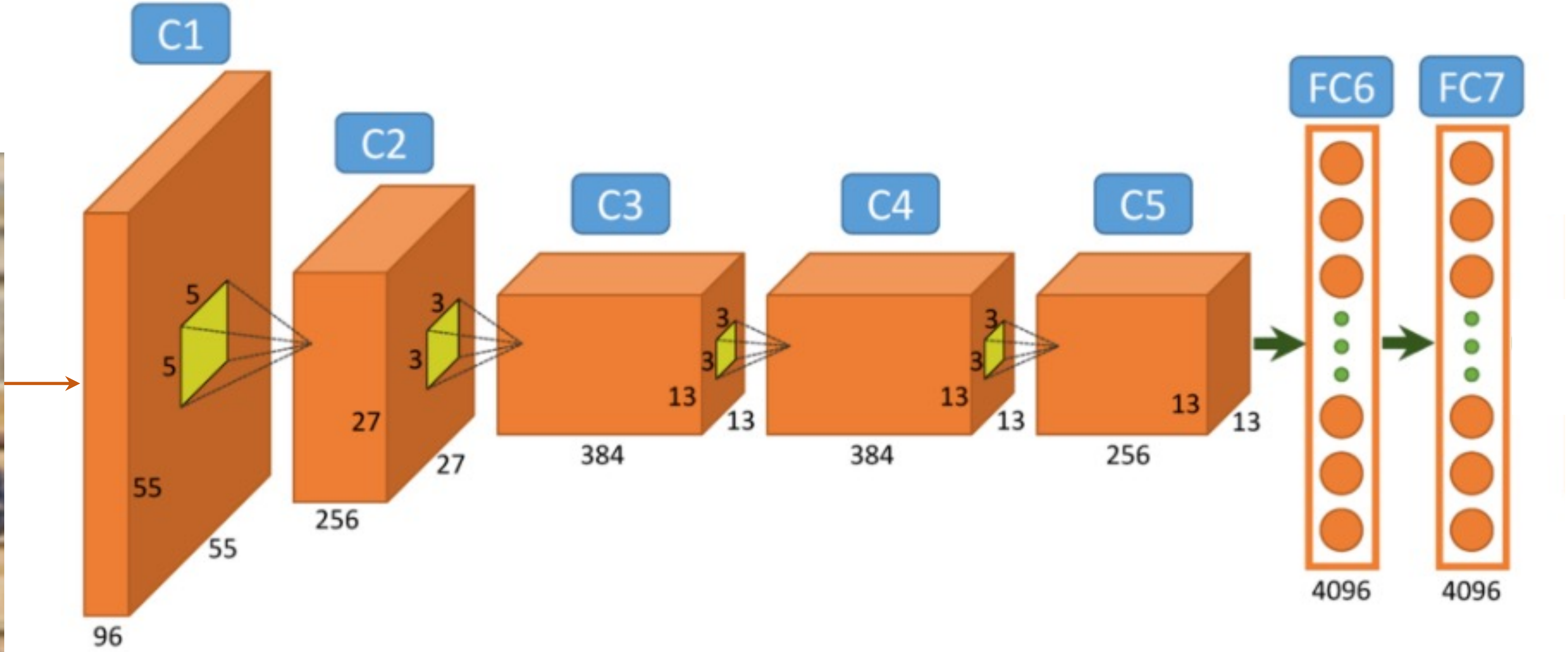


Image Source: https://www.researchgate.net/figure/Architecture-of-Alexnet-From-left-to-right-input-to-output-five-convolutional-layers_fig2_312303454

Architecture

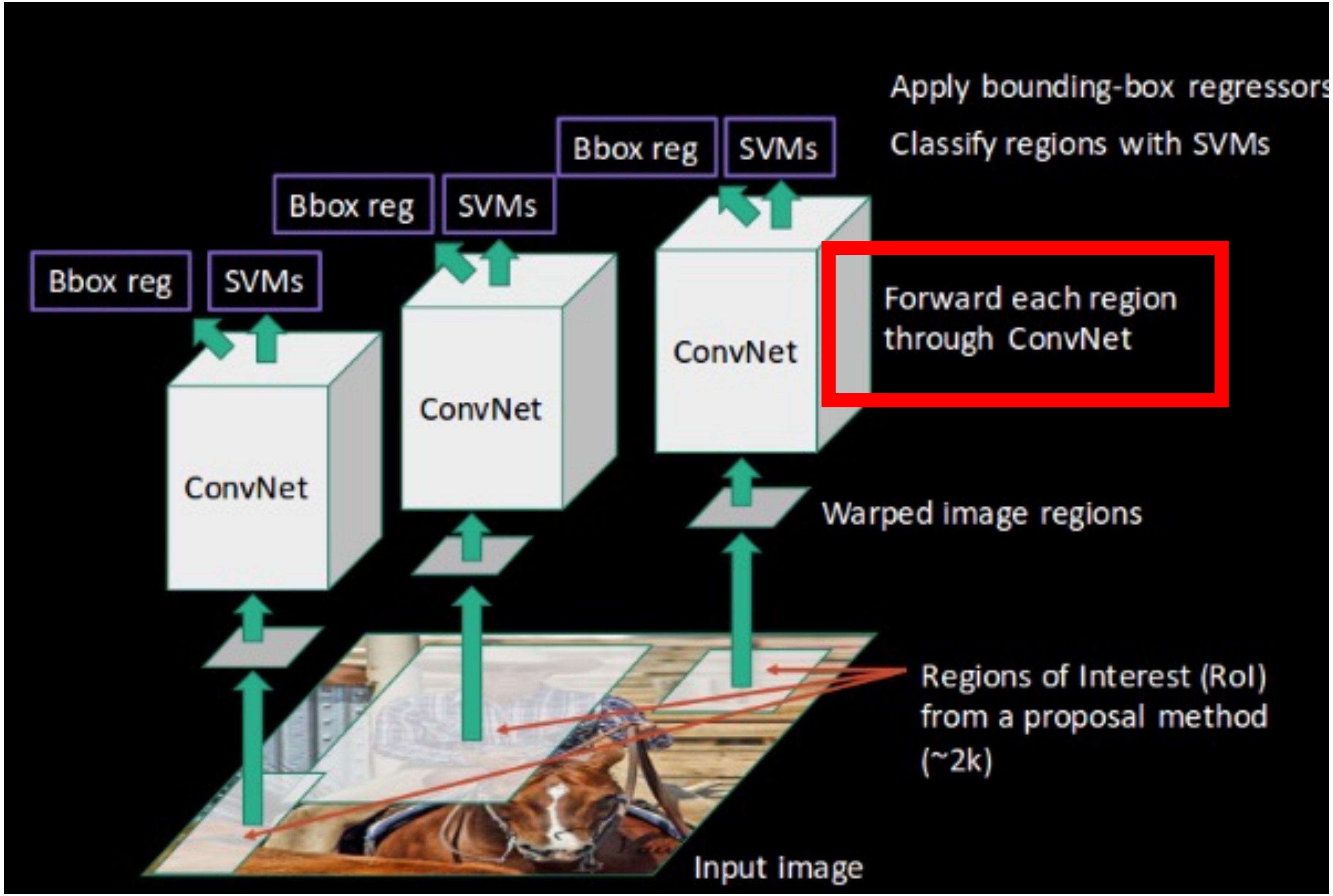


Figure Source: https://www.analyticsvidhya.com/blog/2018/10/a-step-by-step-introduction-to-the-basic-object-detection-algorithms-part-1/

Architecture

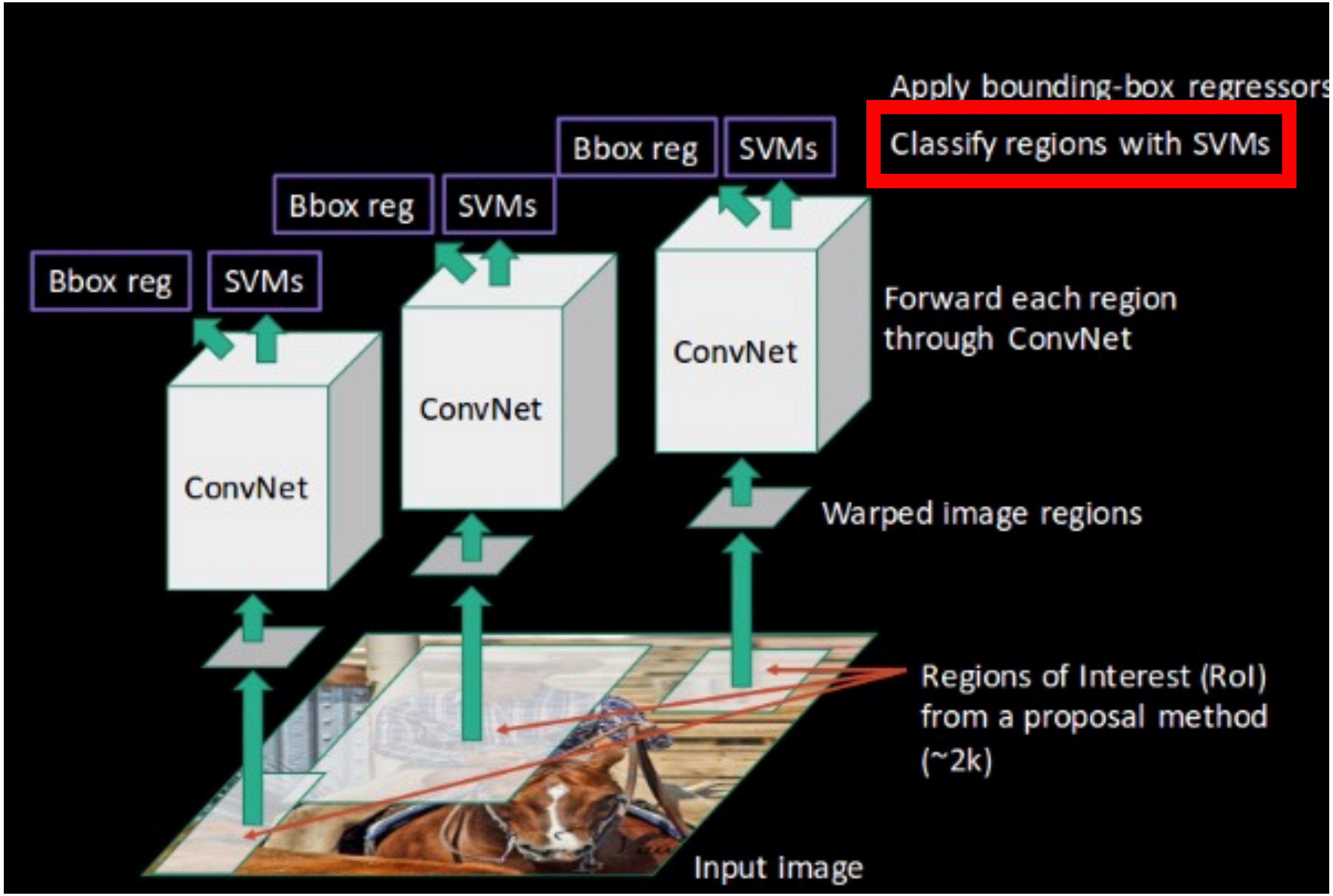
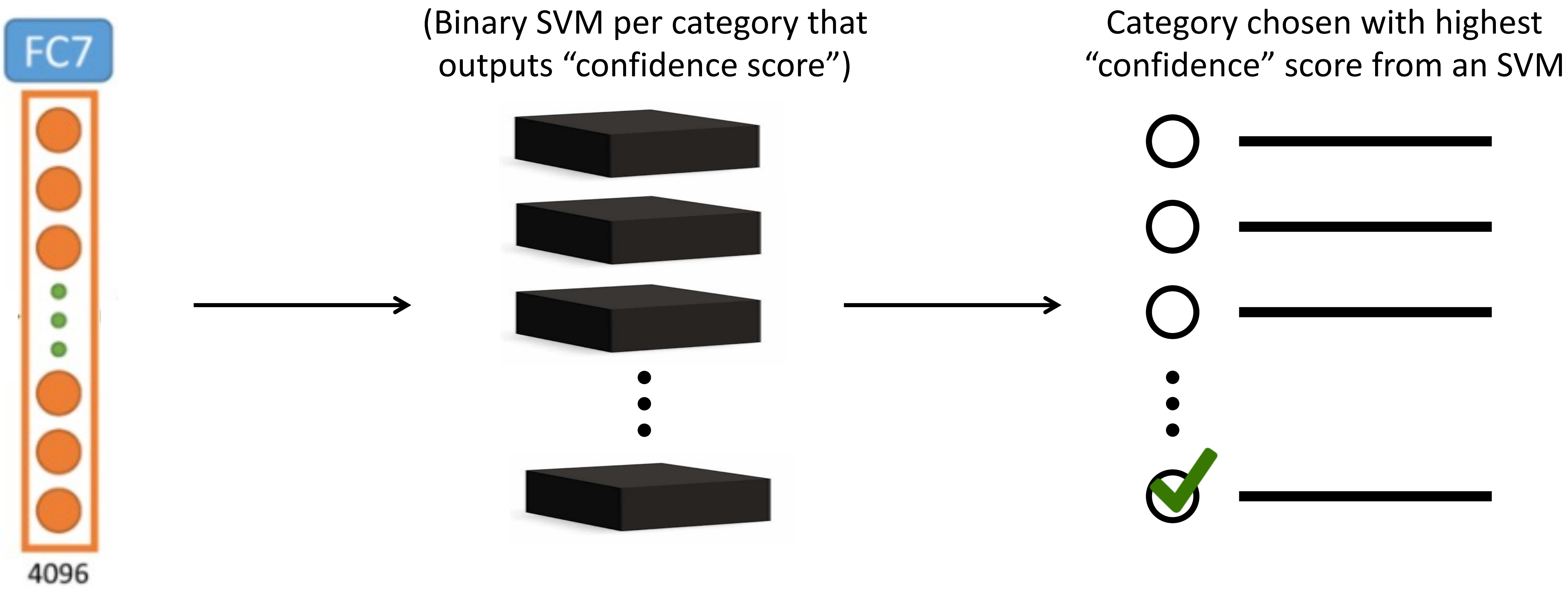


Figure Source: https://www.analyticsvidhya.com/blog/2018/10/a-step-by-step-introduction-to-the-basic-object-detection-algorithms-part-1/

Region Classification

- Assign each feature descriptor that characterizes a region a label from a pre-defined set of categories (i.e., multiple choice)



Architecture

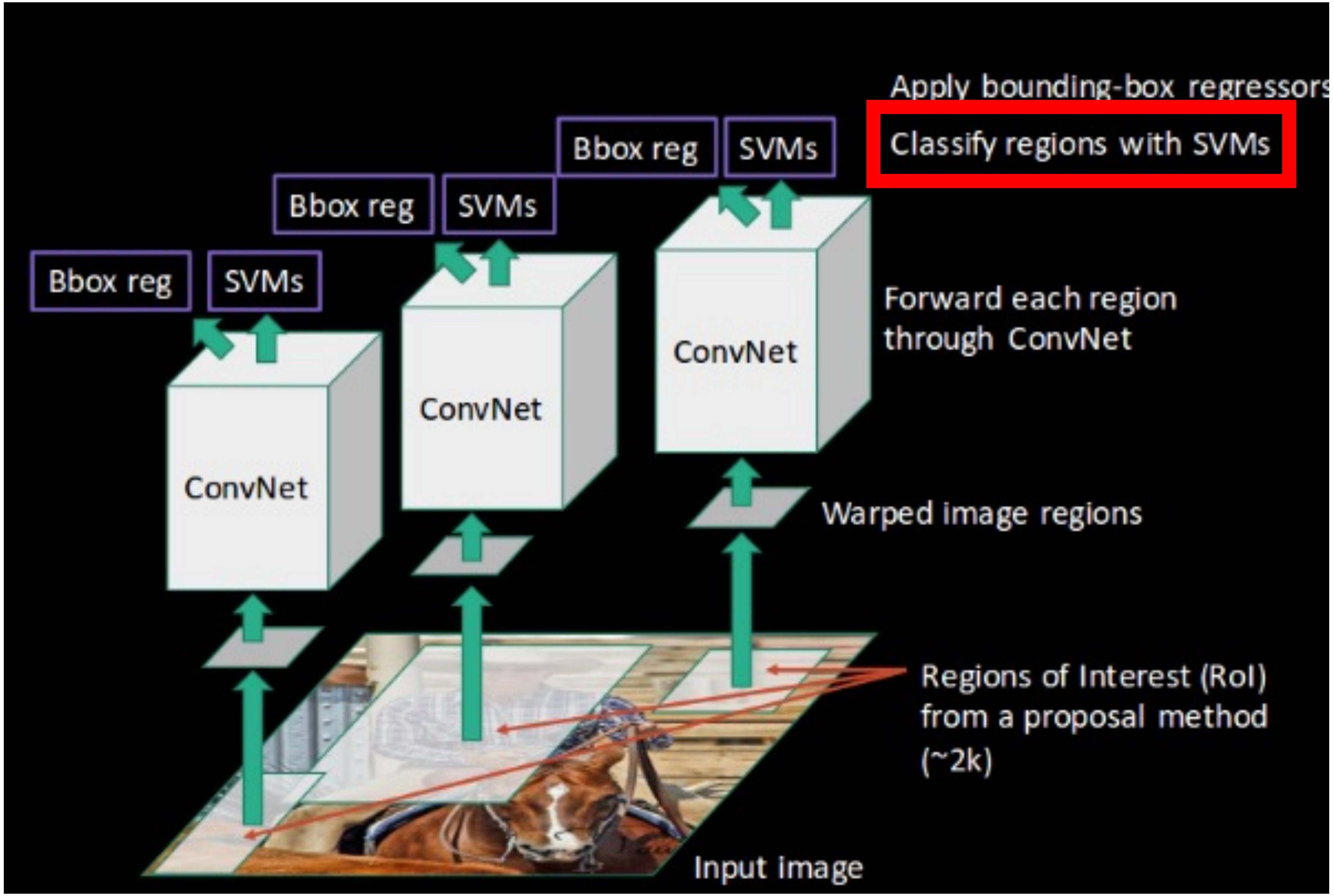


Figure Source: https://www.analyticsvidhya.com/blog/2018/10/a-step-by-step-introduction-to-the-basic-object-detection-algorithms-part-1/

Architecture

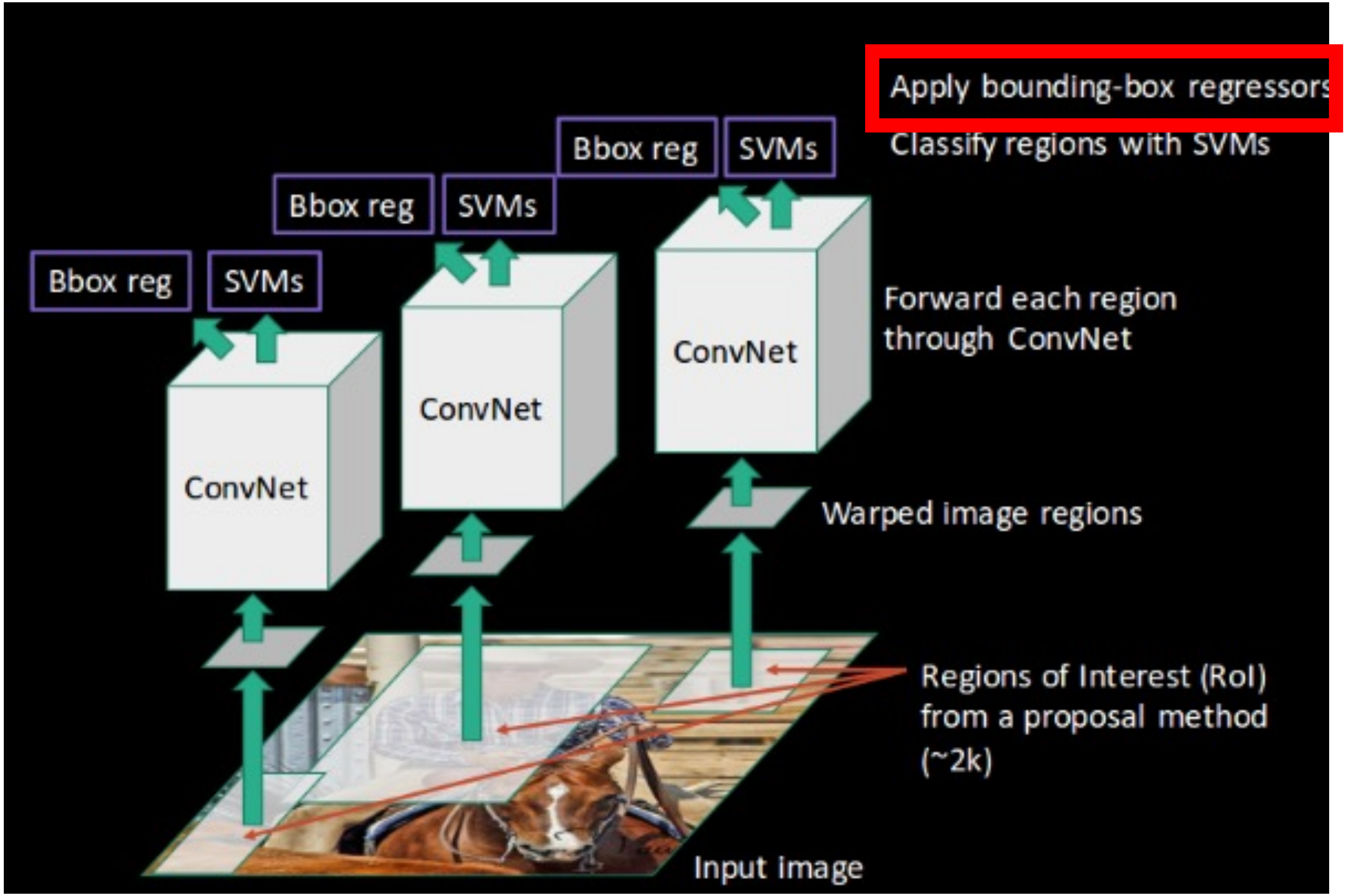


Figure Source: https://www.analyticsvidhya.com/blog/2018/10/a-step-by-step-introduction-to-the-basic-object-detection-algorithms-part-1/

Region Selection and Refinement

- Problem: ~2000 regions per image
- Solution: remove redundant regions through **non-maximum suppression;** for each class:
    1. Pick region with maximum score obtained from the SVM.
    2. Discard all regions belonging to that class nearby that chosen region (i.e., IoU score > 70%)
    3. Select next highest score region and then repeat steps 1 and 2
    4. Repeat step 3 until all regions are either discarded or kept

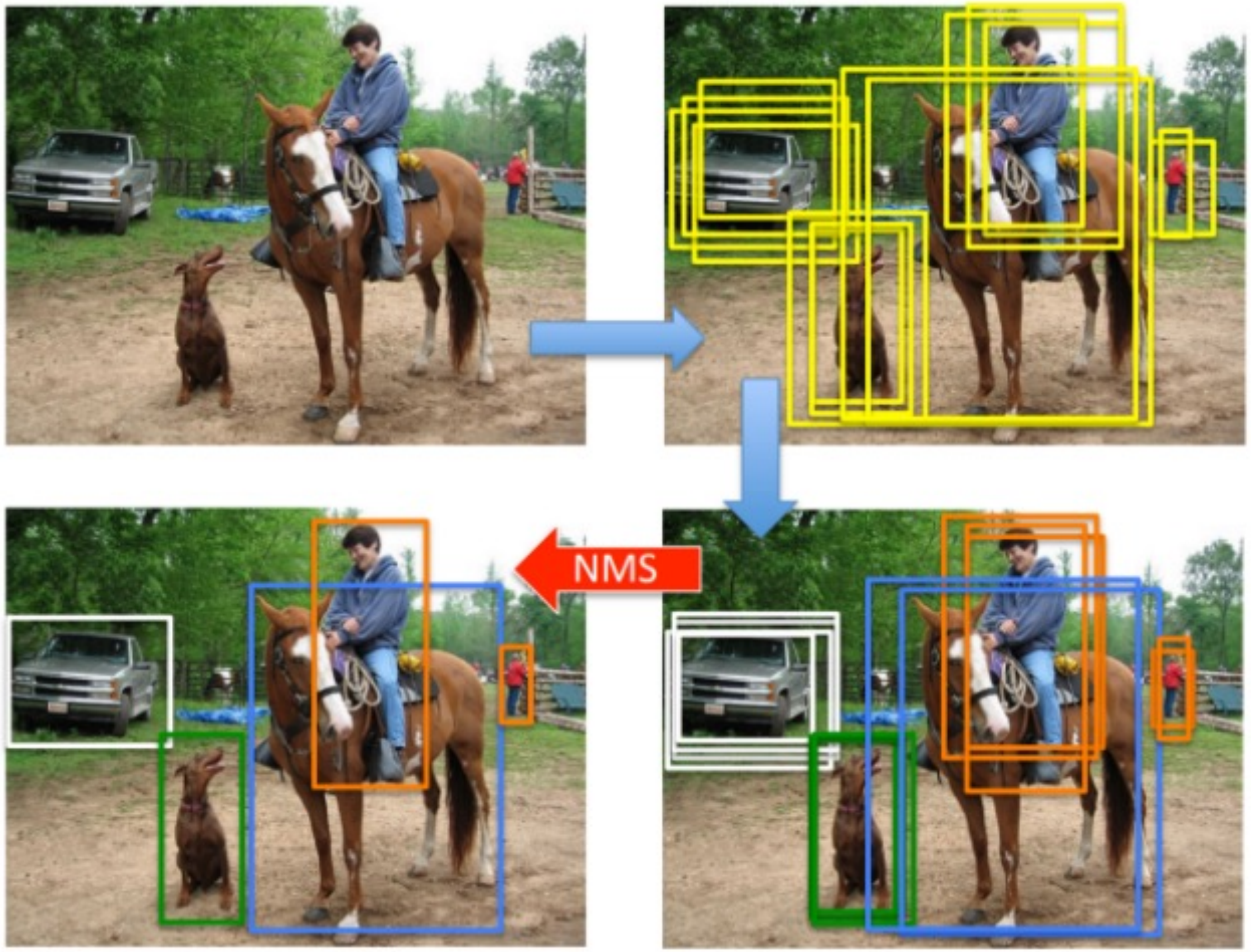


Image Source: https://towardsdatascience.com/deep-learning-method-for-object-detection-r-cnn-explained-ecdadd751d22

Region Selection and **Refinement**

Observing that a common issue was imperfect region proposals, transformations were learned to convert each region proposal to more closely match ground truth

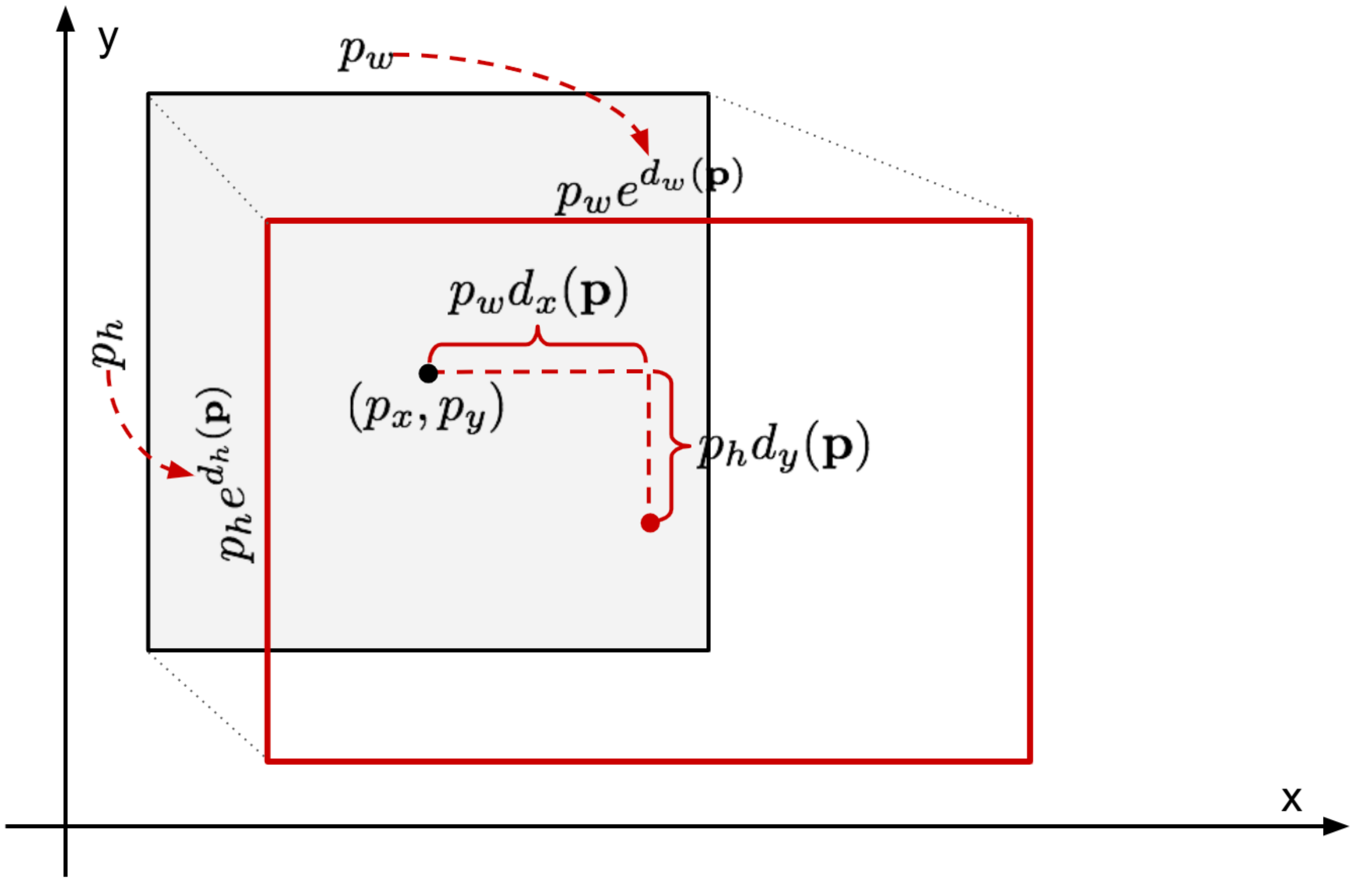


Image Source: https://lilianweng.github.io/lil-log/2017/12/31/object-recognition-for-dummies-part-3.html#bounding-box-regression

Region Selection and **Refinement**

- **Input**: original region location described by a center (p_x, p_y), width (p_w), and height (p_h)
- **Output**: four refinement functions: d_x, d_y, d_w, d_y
- Loss function for learning: SSE

$$\sum_{i \in \{x,y,w,h\}} (t_i - d_i(\mathbf{p}))^2$$

True location

Predicted location

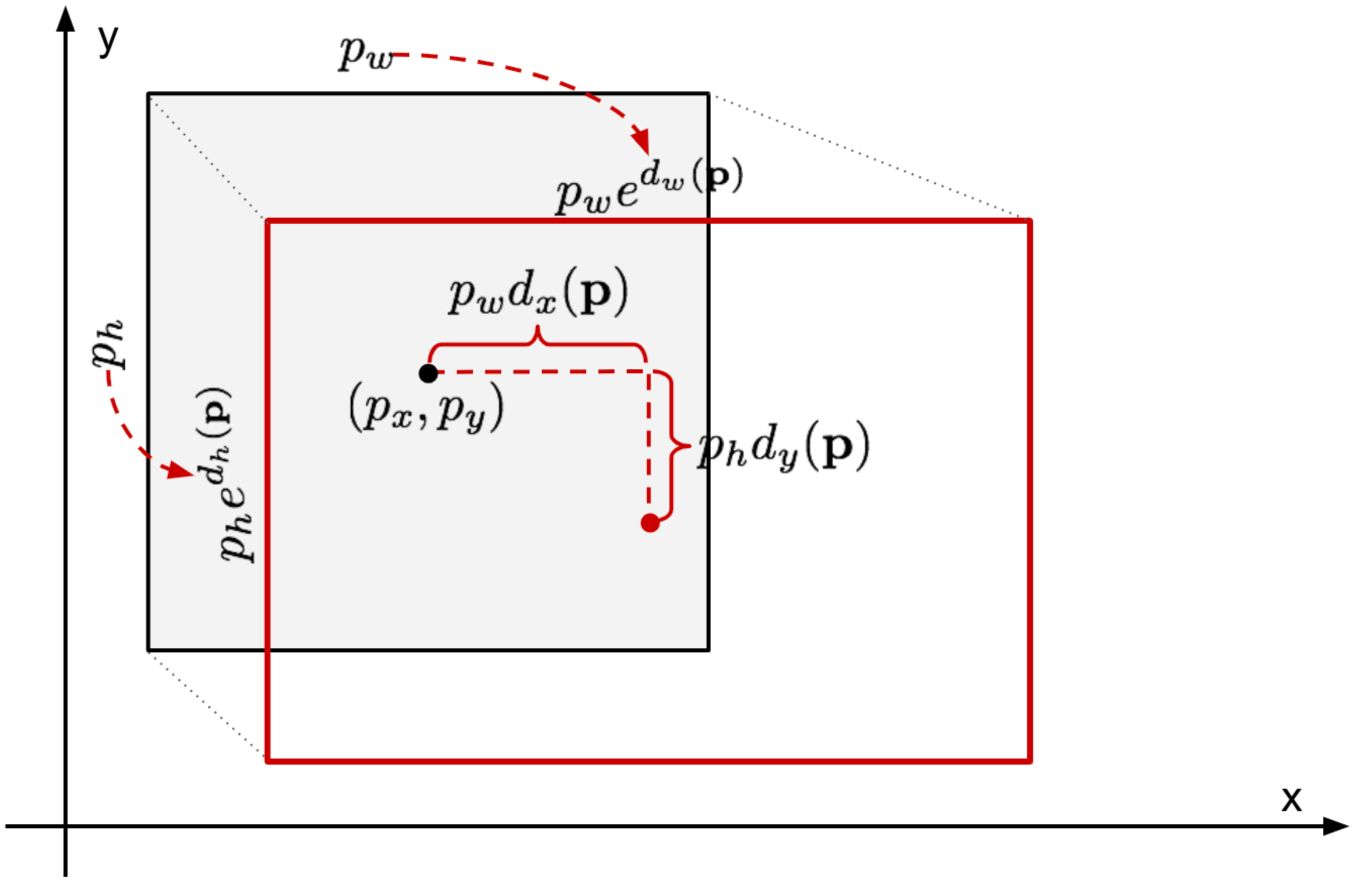
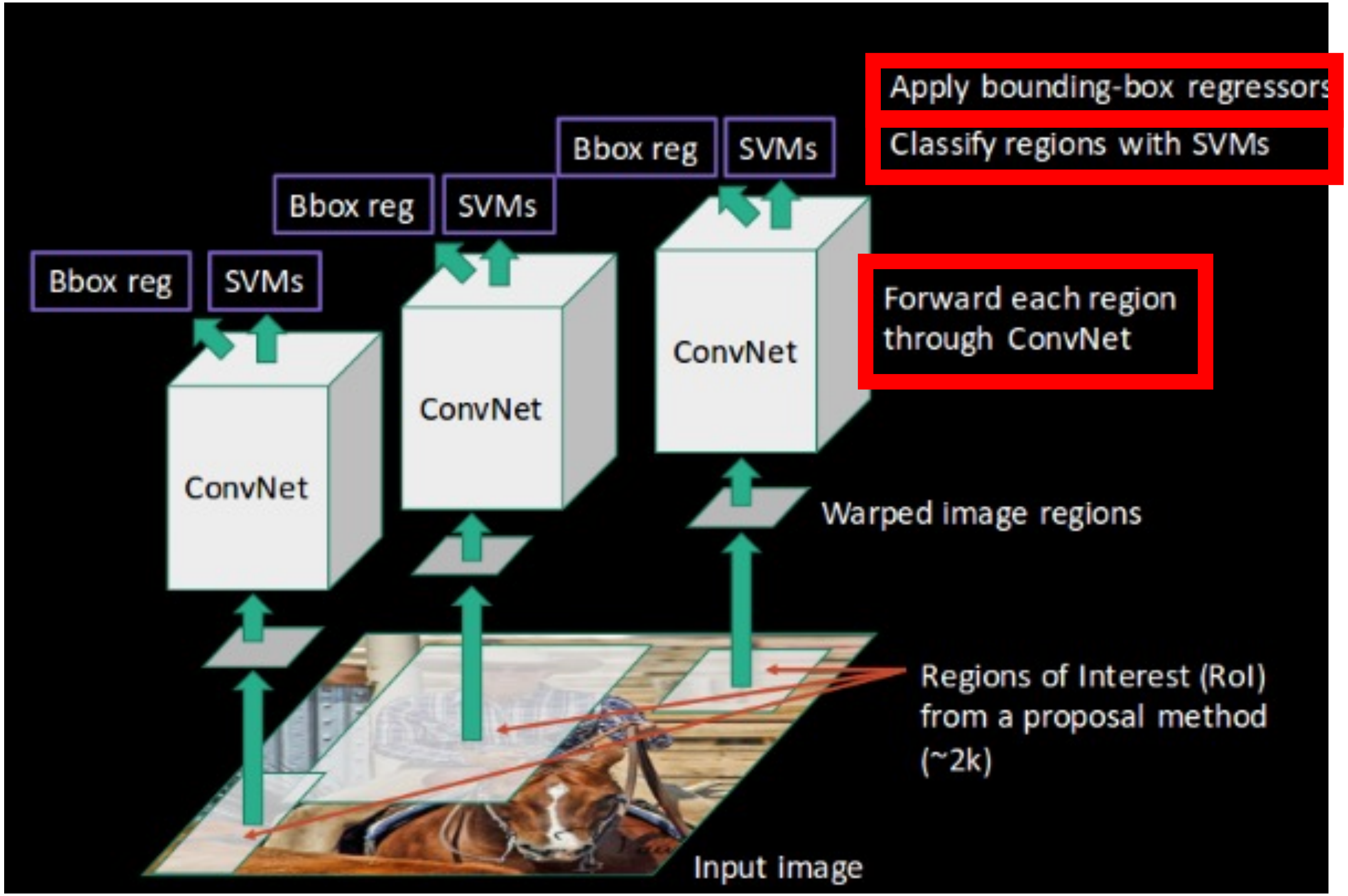


Image Source: https://lilianweng.github.io/lil-log/2017/12/31/object-recognition-for-dummies-part-3.html#bounding-box-regression

R-CNN Limitations



- Slow training procedure
  - Must train three models
- Slow at test time
  (~1 minute per image)
- Inefficient/complex architecture
  - Must store feature descriptor for each region proposal
  - Must refine initial region proposals

Figure Source: https://www.analyticsvidhya.com/blog/2018/10/a-step-by-step-introduction-to-the-basic-object-detection-algorithms-part-1/

Key Concluding Remarks

1. Deep CNN features for image subregions are valuable
(recall, deep CNN features for entire images were also deemed important for scene classification)

2. "We conjecture that the `supervised pre-training/domain-specific finetuning' paradigm will be highly effective for a variety of data-scarce vision problems."

Girshick et al. Rich feature hierarchies for accurate object detection and semantic segmentation. arXiv 2014

Object Detection: Today's Topics

- Problem
- Applications
- Datasets
- Evaluation metric
- Overview of object detection algorithms and baseline (R-CNN)


The End