



# Phát hiện đối tượng

1

## Nội dung buổi học

- Giới thiệu về bài toán phát hiện đối tượng
- Một số kỹ thuật truyền thống
- Các kỹ thuật học sâu

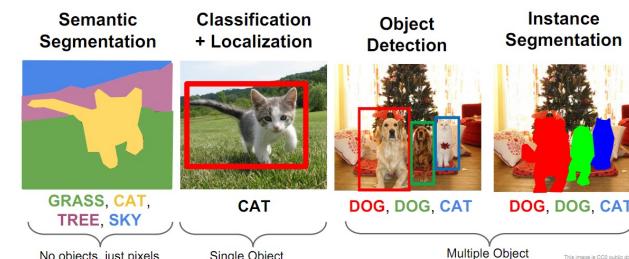


## Giới thiệu bài toán phát hiện đối tượng



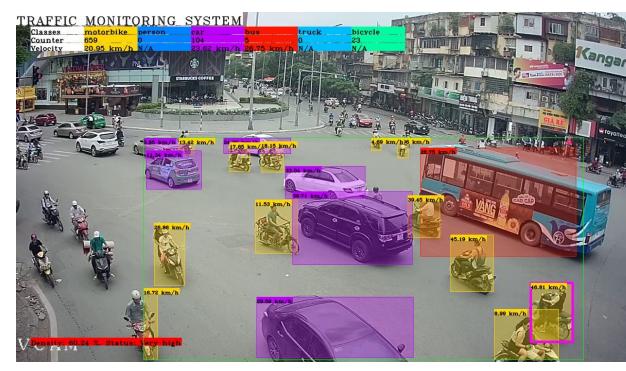
3

## Các bài toán thị giác máy



## Một số ứng dụng bài toán phát hiện đối tượng

- Giao thông thông minh



5

## Một số ứng dụng bài toán phát hiện đối tượng

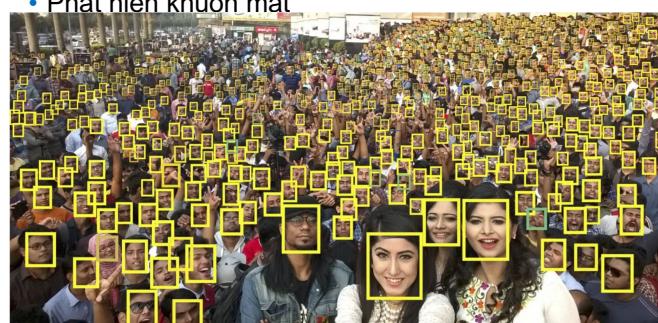
- Điểm bán hàng



6

## Một số ứng dụng bài toán phát hiện đối tượng

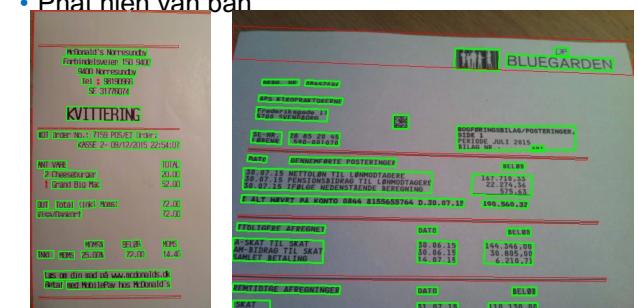
- Phát hiện khuôn mặt



7

## Một số ứng dụng bài toán phát hiện đối tượng

- Phát hiện văn bản



8

### Một số ứng dụng bài toán phát hiện đối tượng

- Robot tự động hái dâu.



SCHOOL OF INFORMATION AND COMMUNICATION TECHNOLOGY

9

9

### Các kỹ thuật truyền thống



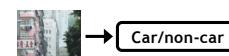
SCHOOL OF INFORMATION AND COMMUNICATION TECHNOLOGY

10

10

### Cửa sổ trượt

- Huấn luyện một bộ phân loại nhị phân



SCHOOL OF INFORMATION AND COMMUNICATION TECHNOLOGY

Slide: Kristen Grauman

11

11

### Cửa sổ trượt

- Sinh các vùng cửa sổ và đánh giá điểm từng cửa sổ



SCHOOL OF INFORMATION AND COMMUNICATION TECHNOLOGY

Slide: Kristen Grauman

12

12

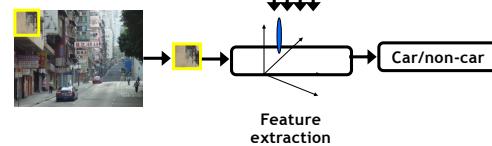
## Cửa sổ trượt

### Huấn luyện:

- Thu thập dữ liệu
  - Lựa chọn đặc trưng
  - Xây dựng bộ phân loại
- Cho một ảnh mới:**
- Trượt cửa sổ
  - Đánh giá bằng bộ phân loại



Training examples



Feature  
extraction

Car/non-car

Slide: Kristen Grauman



SCHOOL OF INFORMATION AND COMMUNICATION TECHNOLOGY

13

13

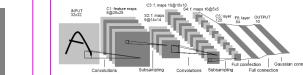
## Các bộ phân loại khác nhau

### Nearest neighbor



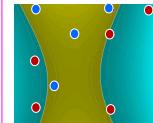
10 examples

### Neural networks



ReLU, Sigmoid, Softmax, ...

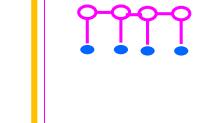
### Support Vector Machines



### Boosting



### Conditional Random Fields

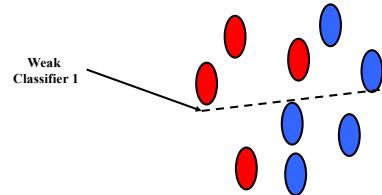


SCHOOL OF INFORMATION AND COMMUNICATION TECHNOLOGY

14

14

## Minh họa Boosting



Slide credit: Paul Viola

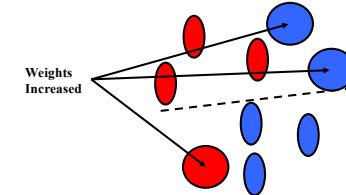


SCHOOL OF INFORMATION AND COMMUNICATION TECHNOLOGY

15

15

## Minh họa Boosting

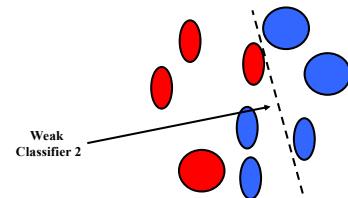


SCHOOL OF INFORMATION AND COMMUNICATION TECHNOLOGY

16

16

## Minh họa Boosting

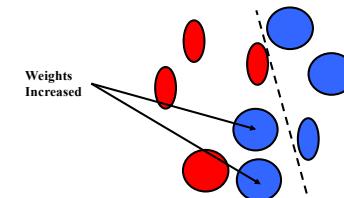


SOICT

SCHOOL OF INFORMATION AND COMMUNICATION TECHNOLOGY

17

## Minh họa Boosting

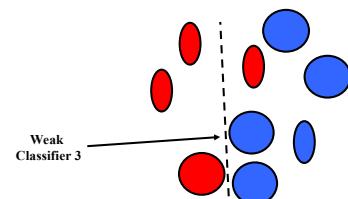


SOICT

SCHOOL OF INFORMATION AND COMMUNICATION TECHNOLOGY

18

## Minh họa Boosting

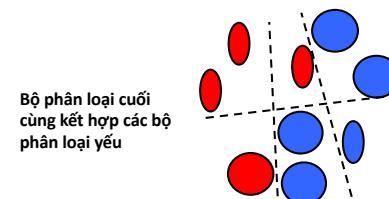


SOICT

SCHOOL OF INFORMATION AND COMMUNICATION TECHNOLOGY

19

## Minh họa Boosting



SOICT

SCHOOL OF INFORMATION AND COMMUNICATION TECHNOLOGY

20

19

## Viola-Jones face detector

ACCEPTED CONFERENCE ON COMPUTER VISION AND PATTERN RECOGNITION 2001

### Rapid Object Detection using a Boosted Cascade of Simple Features

Paul Viola  
viola@merl.com  
Mitsubishi Electric Research Labs  
201 Broadway, 8th FL  
Cambridge, MA 02139

Michael Jones  
mjones@crl.dec.com  
Compaq CRL  
One Cambridge Center  
Cambridge, MA 02142

#### Abstract

This paper describes a machine learning approach for vi-

tested at 15 frames per second on a conventional 700 MHz Intel Pentium III. In other face detection systems, auxiliary information, such as image differences in video sequences,

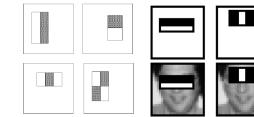


SCHOOL OF INFORMATION AND COMMUNICATION TECHNOLOGY

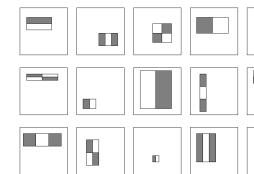
21

21

## Viola-Jones detector: đặc trưng



**Bộ lọc “hình chữ nhật”**  
Độ chênh lệch giữa vùng đen và trắng  
Tính toán rất nhanh nhờ ánh tích phân



Giá trị tại  $(x, y)$  là tổng các điểm ảnh ở trên và bên trái  $(x, y)$   
Ánh tích phân



SCHOOL OF INFORMATION AND COMMUNICATION TECHNOLOGY

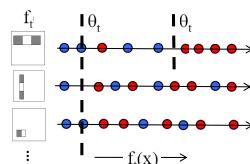
Slide: Kristen Grauman

22

22

## Viola-Jones detector: AdaBoost

- Muốn lựa chọn hình chữ nhật đặc trưng và ngưỡng để phân chia tốt nhất mẫu dương tính (faces) âm tính (non-faces), nghĩa là cực tiểu hóa weighted error.



Giá trị đặc trưng trên tập huấn luyện faces và non-faces.

Chọn ra bộ phân loại yếu:

$$h_t(x) = \begin{cases} +1 & \text{if } f_t(x) > \theta_t \\ -1 & \text{otherwise} \end{cases}$$

Bước tiếp theo, đánh trọng số lại các mẫu dựa theo sai sót, và chọn cặp đặc trưng/ngưỡng khác



SCHOOL OF INFORMATION AND COMMUNICATION TECHNOLOGY

Slide: Kristen Grauman

23

23

- Given example images  $(x_1, y_1), \dots, (x_n, y_n)$  where  $y_i = 0, 1$  for negative and positive examples respectively.

- Initialize weights  $w_{1,i} = \frac{1}{2m} + \frac{1}{2l}$  for  $y_i = 0, 1$  respectively, where  $m$  and  $l$  are the number of negatives and positives respectively.

- For  $t = 1, \dots, T$ :

- Normalize the weights,

$$w_{t,i} \leftarrow \frac{w_{t,i}}{\sum_{j=1}^n w_{t,j}}$$

so that  $w_t$  is a probability distribution.

- For each feature,  $j$ , train a classifier  $h_j$  which is restricted to using a single feature. The error is evaluated with respect to  $w_t$ ,  $\epsilon_j = \sum_i w_t |h_j(x_i) - y_i|$ .

- Choose the classifier,  $h_{t,i}$ , with the lowest error  $\epsilon_t$ .

- Update the weights:

$$w_{t+1,i} = w_{t,i} \beta_t^{1-\epsilon_t}$$

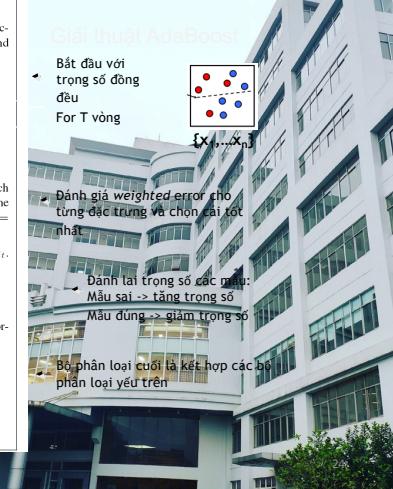
where  $\epsilon_t = 0$  if example  $x_i$  is classified correctly,  $\epsilon_t = 1$  otherwise, and  $\beta_t = \frac{\epsilon_t}{1-\epsilon_t}$ .

- The final strong classifier is:

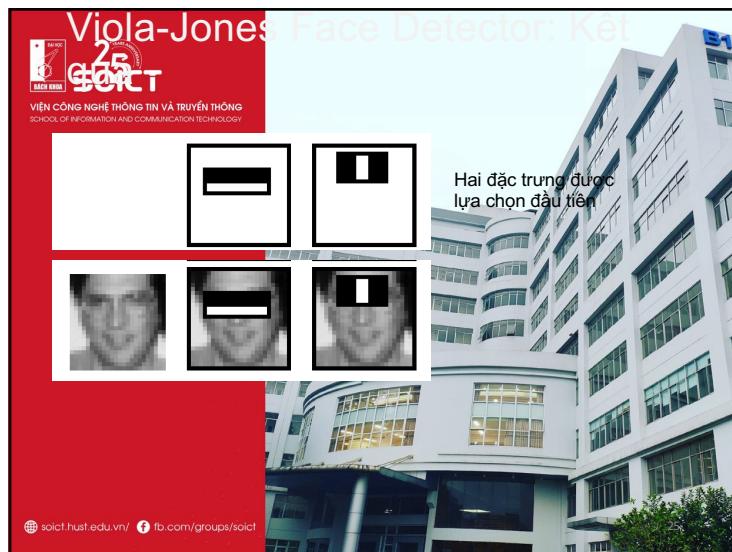
$$h(x) = \begin{cases} 1 & \sum_{t=1}^T \alpha_t h_t(x) \geq \frac{1}{2} \sum_{t=1}^T \alpha_t \\ 0 & \text{otherwise} \end{cases}$$

where  $\alpha_t = \log \frac{1}{\beta_t}$

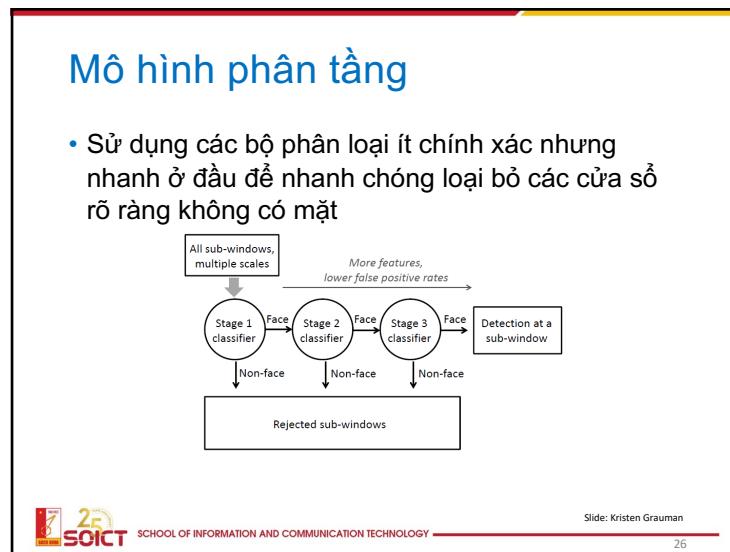
soict.hust.edu.vn/ fb.com/groups/soict



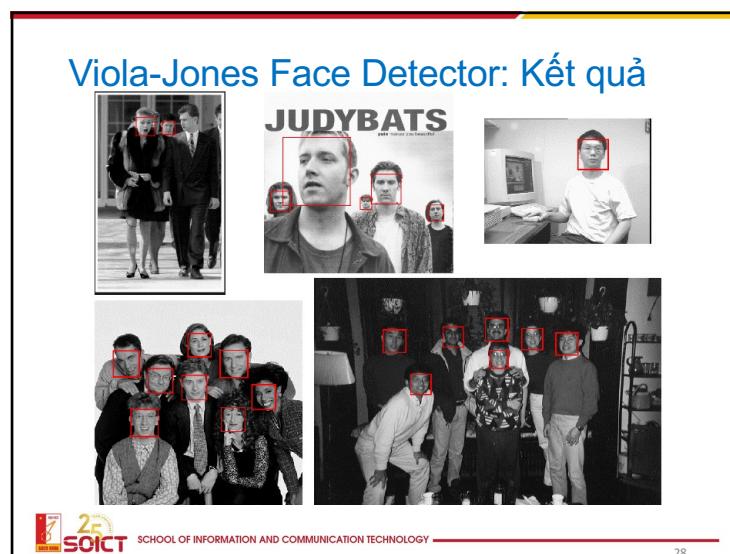
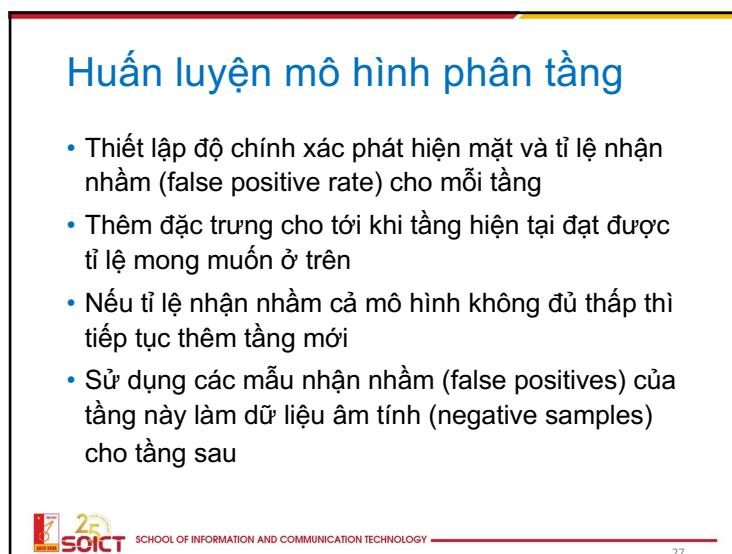
24



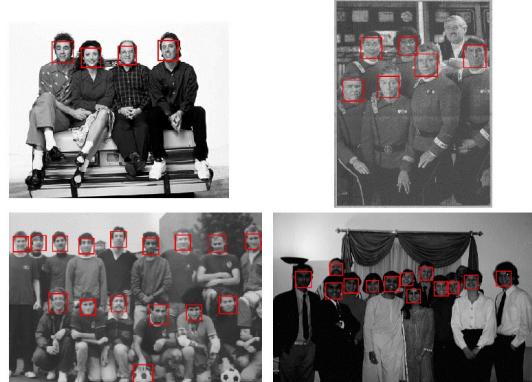
25



26



### Viola-Jones Face Detector: Kết quả



SCHOOL OF INFORMATION AND COMMUNICATION TECHNOLOGY

29

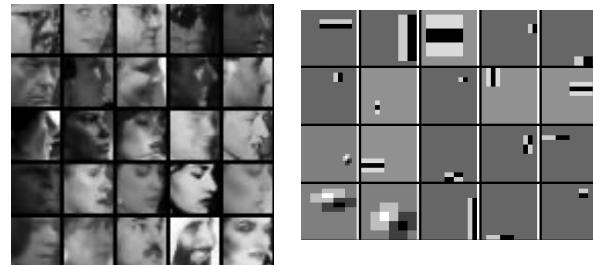
### Viola-Jones Face Detector: Kết quả



SCHOOL OF INFORMATION AND COMMUNICATION TECHNOLOGY

30

### Phát hiện mặt nghiêng



SCHOOL OF INFORMATION AND COMMUNICATION TECHNOLOGY

31

### Phát hiện mặt nghiêng



SCHOOL OF INFORMATION AND COMMUNICATION TECHNOLOGY

32

31

32

# Kỹ thuật học sâu

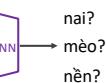


SCHOOL OF INFORMATION AND COMMUNICATION TECHNOLOGY

33

## Phương pháp cửa sổ trượt (sliding windows)

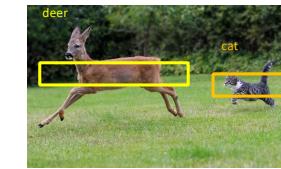
- Quét cửa sổ từ trái sang phải, từ trên xuống dưới. Tại mỗi vị trí thực hiện bài toán phân loại vùng cửa sổ hiện tại thành nhiều lớp đối tượng cộng thêm lớp nền.



SCHOOL OF INFORMATION AND COMMUNICATION TECHNOLOGY

35

## Phương pháp cửa sổ trượt (sliding windows)

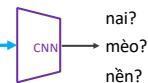
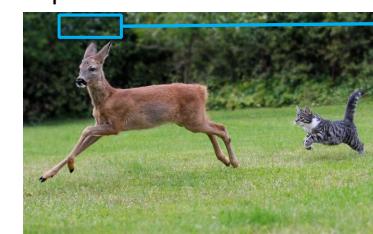


SCHOOL OF INFORMATION AND COMMUNICATION TECHNOLOGY

34

## Phương pháp cửa sổ trượt (sliding windows)

- Quét cửa sổ từ trái sang phải, từ trên xuống dưới. Tại mỗi vị trí thực hiện bài toán phân loại vùng cửa sổ hiện tại thành nhiều lớp đối tượng cộng thêm lớp nền.



SCHOOL OF INFORMATION AND COMMUNICATION TECHNOLOGY

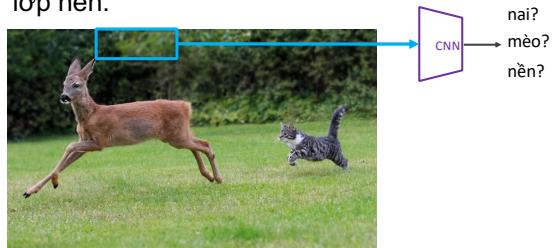
36

35

36

## Phương pháp cửa sổ trượt (sliding windows)

- Quét cửa sổ từ trái sang phải, từ trên xuống dưới.  
Tại mỗi vị trí thực hiện bài toán phân loại vùng cửa sổ hiện tại thành nhiều lớp đối tượng cộng thêm lớp nền.

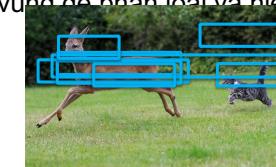


SCHOOL OF INFORMATION AND COMMUNICATION TECHNOLOGY

37

## Phương pháp dựa trên đề xuất vùng

- Thay vì quét tất cả vị trí (số lượng rất lớn!), chỉ phân tích để đề xuất ra một số vùng (box) có khả năng cao chứa đối tượng
- Các phương pháp này có hai giai đoạn (two-stage):
  - đề xuất vùng
  - xử lý từng vùng để phân loại và hiệu chỉnh tọa độ box

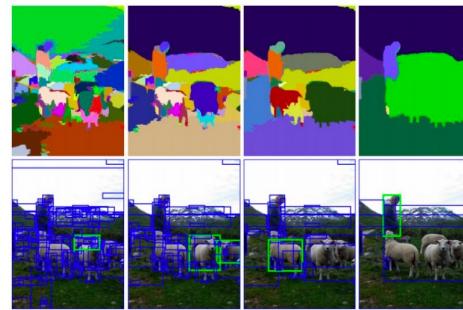


SCHOOL OF INFORMATION AND COMMUNICATION TECHNOLOGY

38

## SS: Selective Search

- Segmentation As Selective Search for Object Recognition. van de Sande et al. ICCV 2011

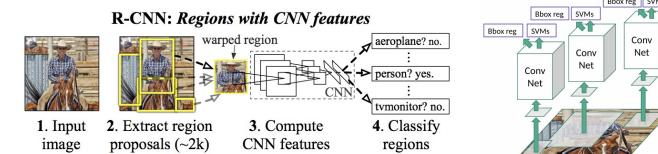


SCHOOL OF INFORMATION AND COMMUNICATION TECHNOLOGY

39

## R-CNN (Region-based ConvNet)

- Đề xuất một số vùng tiềm năng bằng thuật toán khác, chẳng hạn selective search
- Dùng mạng CNN trích xuất đặc trưng từng vùng rồi phân loại bằng SVM



SCHOOL OF INFORMATION AND COMMUNICATION TECHNOLOGY

40

39

40

## Fast-RCNN

- Đẩy tất cả các vùng (khoảng 2000) qua mạng trích xuất CNN cùng một lúc
- Crop thông tin ở lớp đầu ra của CNN thay vì crop vùng trên ảnh gốc như R-CNN
- Đẩy qua nhánh phân loại và nhánh hiệu chỉnh tọa độ

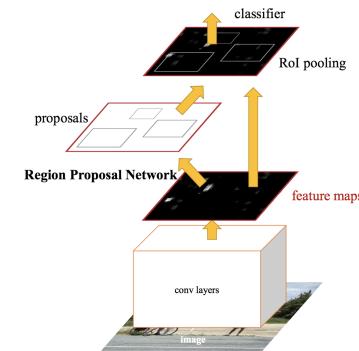


SCHOOL OF INFORMATION AND COMMUNICATION TECHNOLOGY

41

## Faster-RCNN

- Dùng một mạng riêng để đề xuất vùng thay cho selective search
- Còn gọi là phương pháp phát hiện đối tượng hai giai đoạn (two-stage object detector)



SCHOOL OF INFORMATION AND COMMUNICATION TECHNOLOGY

42

## Đặc điểm các mạng không đề xuất vùng

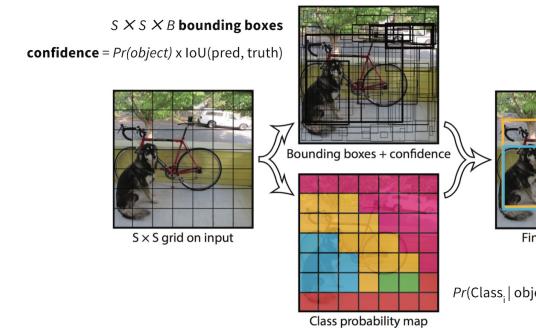
- Còn gọi là mạng một giai đoạn (one-stage)
- Các mạng này thường đề xuất một lưới box dày đặc trên ảnh ban đầu, thường có bước nhảy đều (stride)
- Từng box này sẽ được phân loại và hiệu chỉnh tọa độ (nếu box chứa đối tượng) bằng mạng CNN
- Các mạng một giai đoạn thường nhanh hơn và đơn giản hơn các mạng hai giai đoạn, nhưng độ chính xác có thể không cao bằng.



SCHOOL OF INFORMATION AND COMMUNICATION TECHNOLOGY

43

## YOLO- You Only Look Once

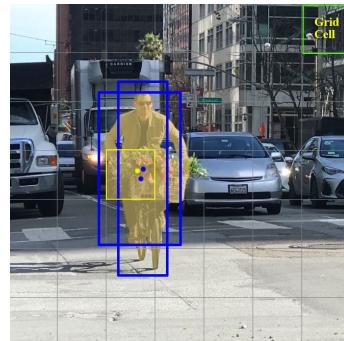


SCHOOL OF INFORMATION AND COMMUNICATION TECHNOLOGY

Redmon et al. CVPR 2016.

44

## YOLO- You Only Look Once

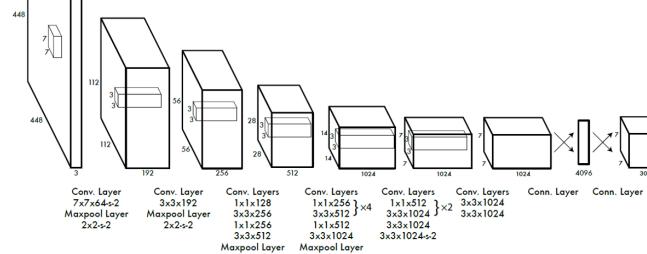


SOICT

SCHOOL OF INFORMATION AND COMMUNICATION TECHNOLOGY

45

## YOLO- You Only Look Once



SOICT

SCHOOL OF INFORMATION AND COMMUNICATION TECHNOLOGY

46

## YOLO- You Only Look Once

$$\begin{aligned}
 & \lambda_{\text{coord}} \sum_{i=0}^{S^2} \sum_{j=0}^B \mathbb{1}_{ij} \left[ (x_i - \hat{x}_i)^2 + (y_i - \hat{y}_i)^2 \right] \quad \text{1 when there is object, 0 when there is no object} \\
 & + \lambda_{\text{coord}} \sum_{i=0}^{S^2} \sum_{j=0}^B \mathbb{1}_{ij} \left[ (\sqrt{w_i} - \sqrt{\hat{w}_i})^2 + (\sqrt{h_i} - \sqrt{\hat{h}_i})^2 \right] \quad \text{Bounding Box size (w, h) when there is object} \\
 & + \sum_{i=0}^{S^2} \sum_{j=0}^B \mathbb{1}_{ij} (C_i - \hat{C}_i)^2 \quad \text{Confidence when there is object} \\
 & + \lambda_{\text{nobj}} \sum_{i=0}^{S^2} \sum_{j=0}^B \mathbb{1}_{i\text{nobj}} (C_i - \hat{C}_i)^2 \quad \text{1 when there is no object, 0 when there is object} \\
 & + \sum_{i=0}^{S^2} \mathbb{1}_i \sum_{c \in \text{classes}} (p_i(c) - \hat{p}_i(c))^2 \quad \text{Class probabilities when there is object}
 \end{aligned}$$

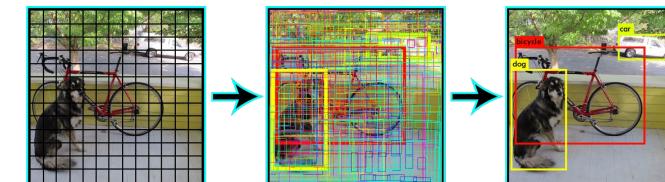
SOICT

SCHOOL OF INFORMATION AND COMMUNICATION TECHNOLOGY

47

## YOLO- You Only Look Once

- Non-maximal suppression: gom các box lại để đưa ra kết quả cuối cùng



SOICT

SCHOOL OF INFORMATION AND COMMUNICATION TECHNOLOGY

48

## YOLO

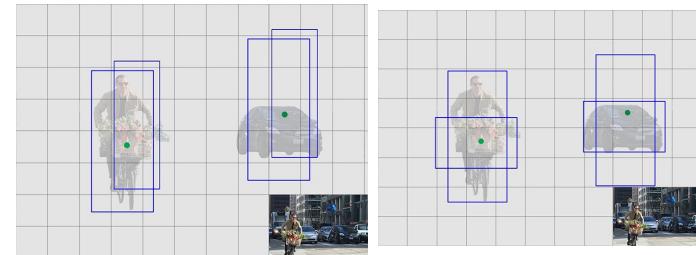
- OpenCV's People Choice Award ở CVPR 2016
- YOLOv2:
  - Thêm batch normalization
  - Fine-tuned để làm việc với độ phân giải cao hơn
- Yolov3:
  - Multi-scale training để nhận các đối tượng kích thước bé
- Yolov4:
  - Sử dụng backbone khác
  - Cải thiện v3 nhằm tăng độ chính xác
- **Yolov5,6,7,8,9,10**



SCHOOL OF INFORMATION AND COMMUNICATION TECHNOLOGY

49

## YOLO v2

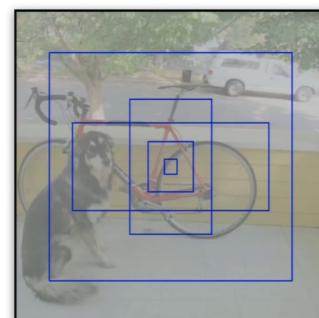


SCHOOL OF INFORMATION AND COMMUNICATION TECHNOLOGY

50

## YOLOv2

- Stack các tensor
- 5 boxes với mỗi cell
  - Kích cỡ khác nhau
- Dùng các boxes này là các prior
  - Predict offset từ các prior này
  - Kmean clustering của: training data bounding box

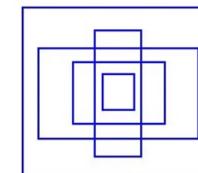


SCHOOL OF INFORMATION AND COMMUNICATION TECHNOLOGY

51

## YOLO v2

- Mỗi ô có 5 anchor box.  
Với mỗi anchor mạng sẽ đưa ra các thông tin:
  - offset của box: 4 số thực trong khoảng [0, 1]
  - Độ tin tưởng box đó có khả năng chứa đối tượng (objectness score).
  - Phân bố xác suất của đối tượng trong box đó ứng với các lớp đối tượng khác nhau (class scores).
- Tổng cộng mỗi ô có số đầu ra là:  $5 * (4 + 1 + 20) = 125$  số thực



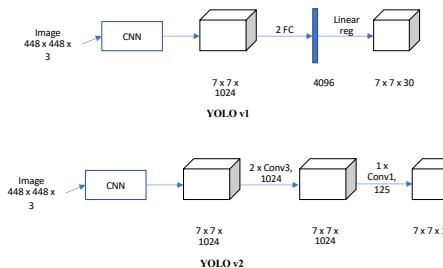
5 anchor box



SCHOOL OF INFORMATION AND COMMUNICATION TECHNOLOGY

52

## YOLO v2



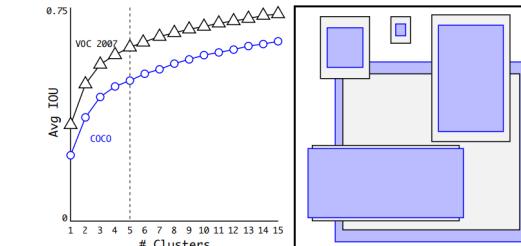
SCHOOL OF INFORMATION AND COMMUNICATION TECHNOLOGY

53

53

## YOLO v2

- Xác định kích thước mặc định của các anchor bằng cách áp dụng k-means trên tập box các đối tượng đã được đánh nhãn trong tập huấn luyện

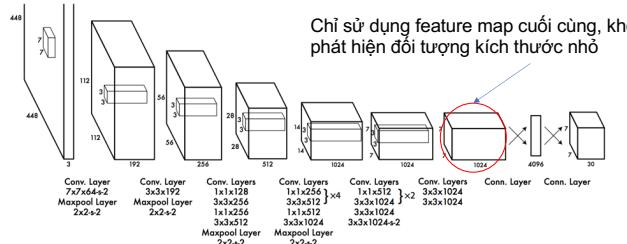


SCHOOL OF INFORMATION AND COMMUNICATION TECHNOLOGY

54

## YOLO v2

- Nhược điểm của YOLO v1 và v2:

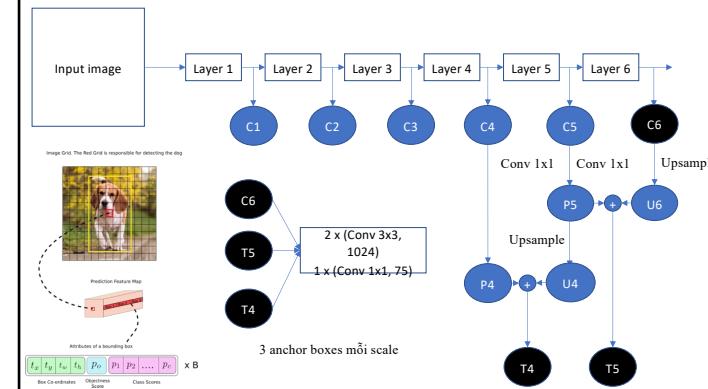


SCHOOL OF INFORMATION AND COMMUNICATION TECHNOLOGY

55

55

## YOLO v3

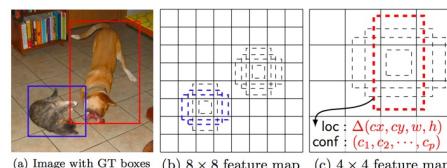


SCHOOL OF INFORMATION AND COMMUNICATION TECHNOLOGY

56

## SSD: Single Shot Detector

- Tương tự YOLO nhưng lưới box dày đặc hơn, có nhiều lưới với các kích thước box khác nhau
- Kiến trúc mạng backbone khác với YOLO
- Data augmentation + Hard negative mining



SCHOOL OF INFORMATION AND COMMUNICATION TECHNOLOGY

57

## SSD: Single Shot Detector

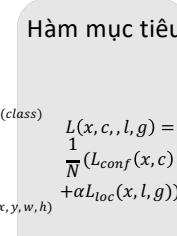
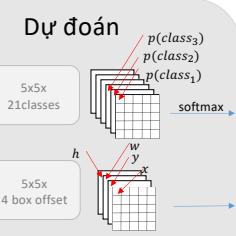
- Mạng backbone: VGG-16
- Thêm các lớp tích chập phụ phía sau các lớp của mạng backbone
- Phát hiện đối tượng ở nhiều mức khác nhau trong mạng (Multi-scale)



SCHOOL OF INFORMATION AND COMMUNICATION TECHNOLOGY

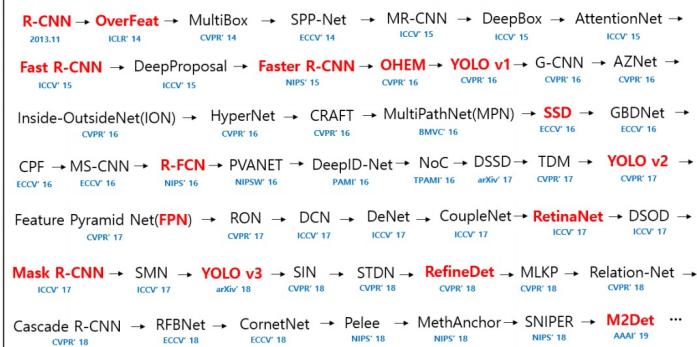
58

## SSD: Single Shot Detector



SCHOOL OF INFORMATION AND COMMUNICATION TECHNOLOGY

59



SCHOOL OF INFORMATION AND COMMUNICATION TECHNOLOGY

60

## One-stage vs two-stage

**Faster and simpler  
one-stage object detector  
(dense sampling of object  
locations, scales, and aspect ratios)**

YOLO YOLO-v2 YOLO-v3

SSD DSSD

MDCN SqueezeNet

## RetinaNet

CornetNet

EfficientDe

More accurate  
two-stage object detector  
(proposal-driven mechanism)

R-CM

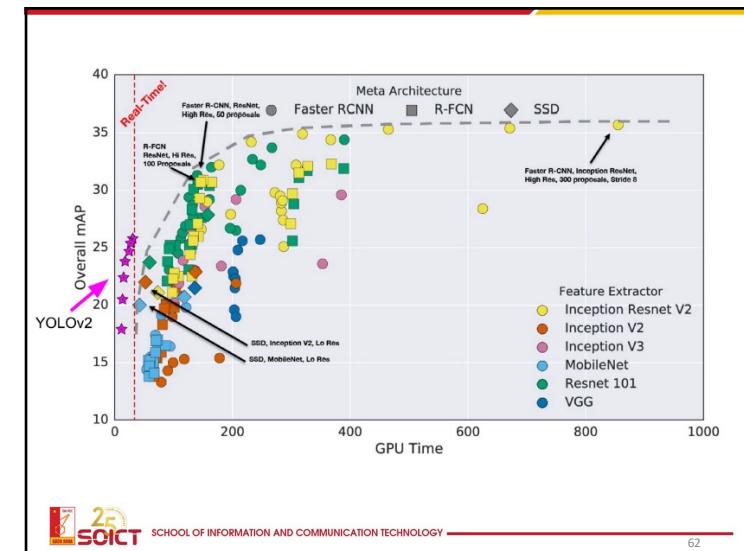
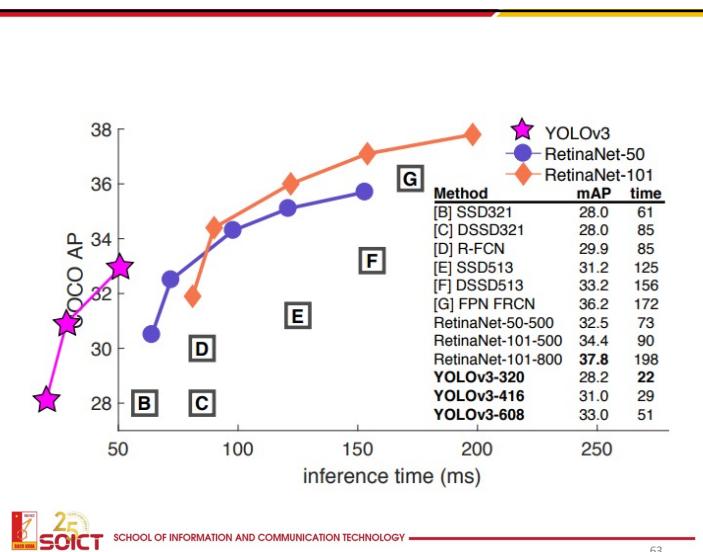
Fast R-CNN

CNN

## Mask R-CNN



61



62



63

## Các bộ dữ liệu chuẩn

- PASCAL VOC Challenge
- ImageNet Large Scale Visual Recognition Challenge (ILSVR)
  - 200 Categories for detection



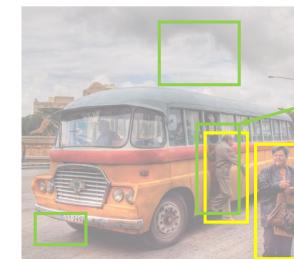
- Common Objects in Context (COCO)
  - 80 Object categories



SCHOOL OF INFORMATION AND COMMUNICATION TECHNOLOGY

65

## True Positive



— Dự đoán  
— ground truth

### True positive:

- IoU lớn hơn một ngưỡng nào đó (0.5)



SCHOOL OF INFORMATION AND COMMUNICATION TECHNOLOGY

66

## False positive



— Dự đoán  
— ground truth

### True positive:

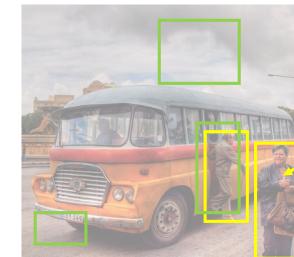
- False positive:
- IoU nhỏ hơn một ngưỡng nào đó (0.5)



SCHOOL OF INFORMATION AND COMMUNICATION TECHNOLOGY

67

## False negative



— Dự đoán  
— ground truth

### True positive:

### False positive:

### False negative:

- Đối tượng mà mô hình không đoán ra được



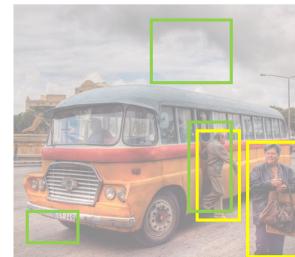
SCHOOL OF INFORMATION AND COMMUNICATION TECHNOLOGY

68

67

17

## True negative



— Dự đoán  
— ground truth

**True positive:**

**False positive:**

**False negative:**

- Đối tượng mà mô hình không  
đoán ra được

**True Negative** là gì?



SCHOOL OF INFORMATION AND COMMUNICATION TECHNOLOGY

69

## Các độ đo

		Predicted 1	Predicted 0
		True 1	false negative
True 0	True 1	true positive	false positive
	False 1	false negative	true negative

		Predicted 1	Predicted 0
		True 1	FN
True 0	True 1	TP	FN
	False 1	FP	TN

$$precision = \frac{TP}{TP + FP}$$

$$recall = \frac{TP}{TP + FN}$$



SCHOOL OF INFORMATION AND COMMUNICATION TECHNOLOGY

70

## Các độ đo

- Điểm (score) các dự đoán thường nằm trong  
khoảng từ 0 tới 1



— Dự đoán  
— ground truth

Đây là các box có **score > 0**.

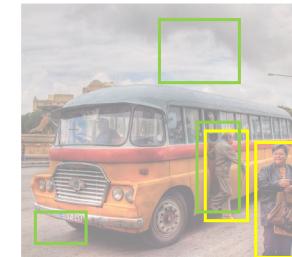
- **Recall hoàn hảo!**
- Nhưng **precision rất tệ!**



SCHOOL OF INFORMATION AND COMMUNICATION TECHNOLOGY

71

## How do we evaluate object detection?



— predictions  
— ground truth

Here are all the boxes that  
are predicted with **score > 0.5**

We are setting a **threshold** of  
0.5



SCHOOL OF INFORMATION AND COMMUNICATION TECHNOLOGY

72

71

18

## Mean average Precision - mAP

- Chạy object detector trên ảnh test
- Với mỗi loại, tính average precision
  - Sắp xếp từ giá trị cao xuống thấp
  - So sánh IoU với các bounding box khác, nếu tồn tại  $\text{IoU} > 0.5 \rightarrow$  đánh nhãn positive và loại bounding box được so sánh này
  - Nếu không đánh nhãn negative
  - Vẽ Precision-recall Curve
- Mean Average Precision (mAP) = trung bình với mỗi loại
  - Ví dụ:
    - Car AP = 0.5
    - Cat AP = 0.80
    - Dog AP = 0.86
    - $\rightarrow \text{mAP} = 0.77$
- Cách khác tính mAP với mỗi ngưỡng IoU ví dụ 0.6, 0.7, 0.9 rồi chia trung bình



SCHOOL OF INFORMATION AND COMMUNICATION TECHNOLOGY

73

73

## mAP (2)

- Tính AP với mỗi loại
  - Sắp xếp từ giá trị cao xuống thấp
  - So sánh IoU với các bounding box khác, nếu tồn tại  $\text{IoU} > 0.5 \rightarrow$  đánh nhãn positive và loại
  - Nếu không đánh nhãn negative
  - 2 cái đầu là true positive

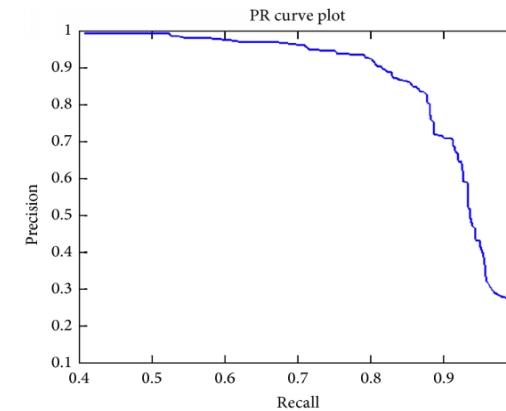


SCHOOL OF INFORMATION AND COMMUNICATION TECHNOLOGY

74

74

## Precision – recall curve (PR curve)

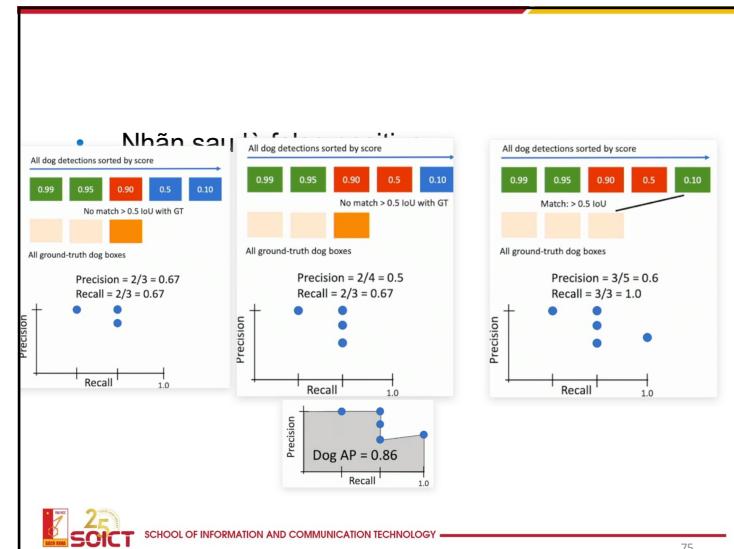


SCHOOL OF INFORMATION AND COMMUNICATION TECHNOLOGY

76

75

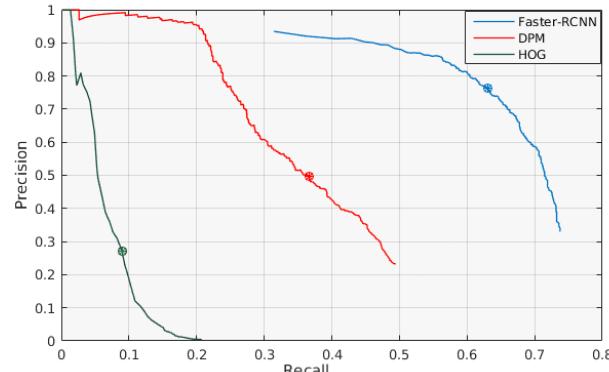
76



SCHOOL OF INFORMATION AND COMMUNICATION TECHNOLOGY

75

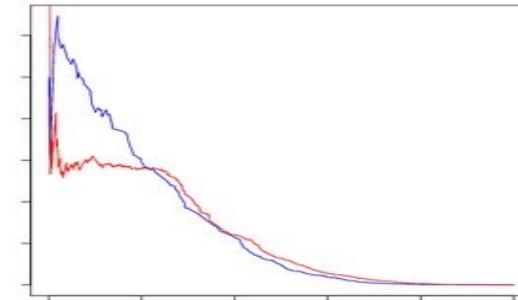
## Mô hình nào là tốt nhất?



SCHOOL OF INFORMATION AND COMMUNICATION TECHNOLOGY

77

## Mô hình nào là tốt nhất?



- Average precision (AP):** Giá trị precision trung bình (AP) của từng lớp đối với tất cả các ngưỡng IoU:
- mAP:** trung bình các AP



SCHOOL OF INFORMATION AND COMMUNICATION TECHNOLOGY

78

## Tài liệu tham khảo

- Kristen Grauman (CS 376: Computer Vision, Spring 2018, The University of Texas at Austin)



SCHOOL OF INFORMATION AND COMMUNICATION TECHNOLOGY

80

Thank  
you!

[www.soict.hust.edu.vn/](http://soict.hust.edu.vn/) [fb.com/groups/soict](https://fb.com/groups/soict)


81