Towards Understanding Deep Networks

Alan Yuille

Plan of the talk

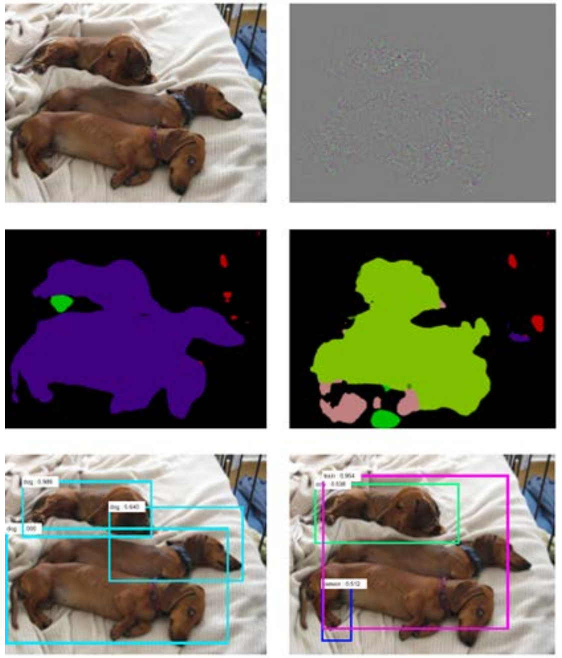
- (I) Project 0: Adversarial Noise. Attacking Deep Nets.
- (II) Project I: Parts, Voting, and Occlusion.
- (III) Project 2: Perceptual Similarity Learning, including Tufa's.

Adversarial Noise (AN)

- Imperceptible amounts of noise can drastically alter performance of deep nets for object classification – (C. Szegedy et al. 2013).
- Adversarial noise also applies to object detection and semantic segmentation. (C. Xie et al. Arxiv. 2017). Adversaries can be transferred across networks and even some tasks.

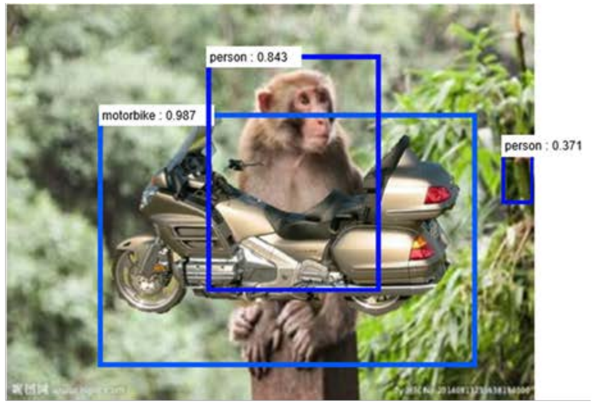
AN for Semantic Segmentation and Detection

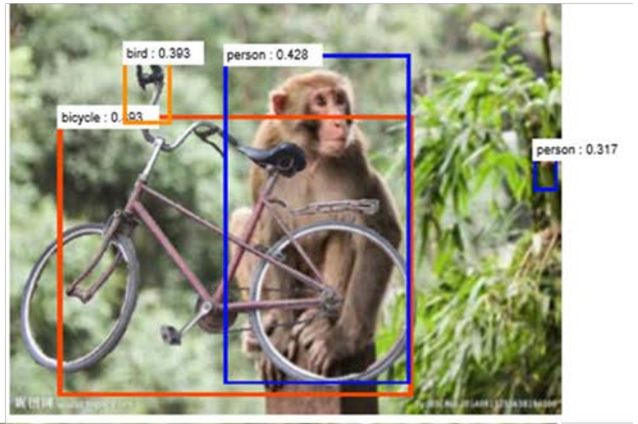
- AN can turn: Dogs into Cows,
  Train into an Airplane with shape ICCV
  Blank Image into a Bus with shape 2017

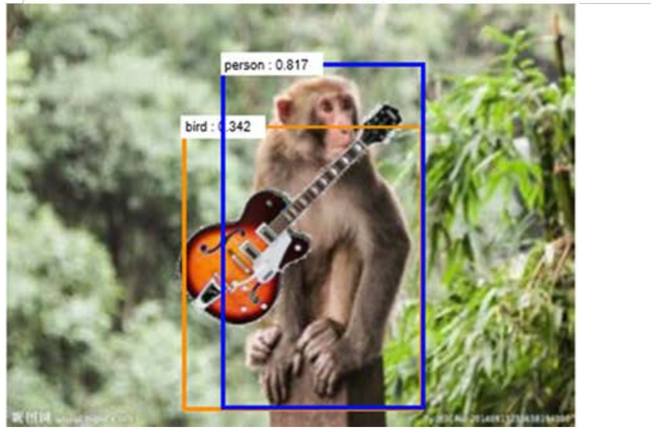


Adversarial Context

- A motorbike turns a monkey into a human.
- A bike turns a monkey into a human &
  the jungle turns the bike handle into a bird.
- A guitar turns the monkey into a human &
  the jungle turns the guitar into a bird.







Project I: Parts, Voting, and Occlusion

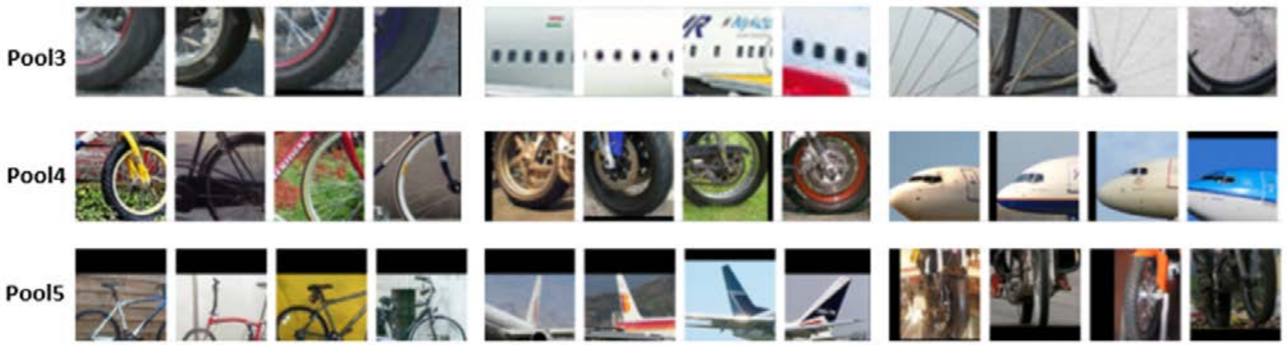
- Can we learn part models in a weakly supervised manner and use them to outperform supervised methods for part detection?
- Not yet. But how far can we get?
- Make this more interesting by adding occlusion.
- Why do this?
- (A) Supervised labeling of object parts is expensive and time-consuming.
- (B) Humans require little supervision.
- (C) Gives insight into Deep Nets. Develop new deep architectures based on compositionality.

Deep Nets and Parts.

- Deep Nets seem to represent parts of objects.
- This was first demonstrated by visualization studies of single filters/neurons (M. D. Zeiler and R. Fergus. ECCV. 2014).
- It was shown quantitatively in (B. Zhou et al. ICLR. 2015).

- We studied population encoding of parts in Deep Nets to obtain unsupervised part detectors.
- We compared them to single filter detectors and SVM supervised methods. (J. Wang et al. arxiv. 2015).

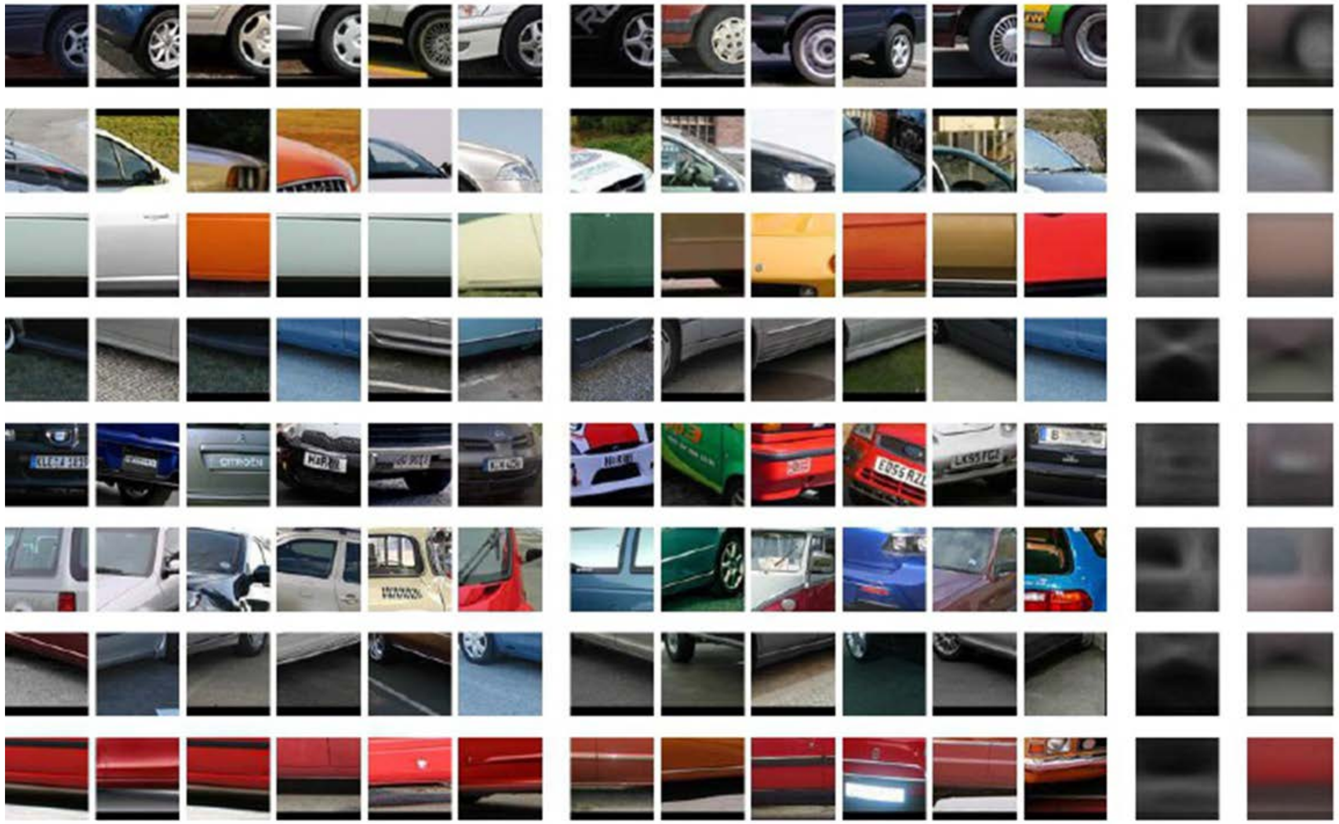
Methods



- Use Deep Nets trained for object classification on ImageNet.
- Observe feature responses of the Deep Nets applied to objects of fixed size from PASCAL 3D+ (Cars, Planes, Bikes,…).
- Cluster the features responses using k-means. Call the cluster centers “visual concepts”.
- Visualize the cluster centers by seeing which image patches correspond to them (those image patches whose feature vectors are assigned to the cluster). See top right.

Findings: Visualize tightness

- The clusters – visual concepts -- are extremely tight perceptually.
- Show best 6, random 6 from best 500, mean edge, mean intensity.



Findings: Visualize coverage

- The visual concepts (VCs) cover most of the object.
- Here are 44 (out of 170) VCs for cars.

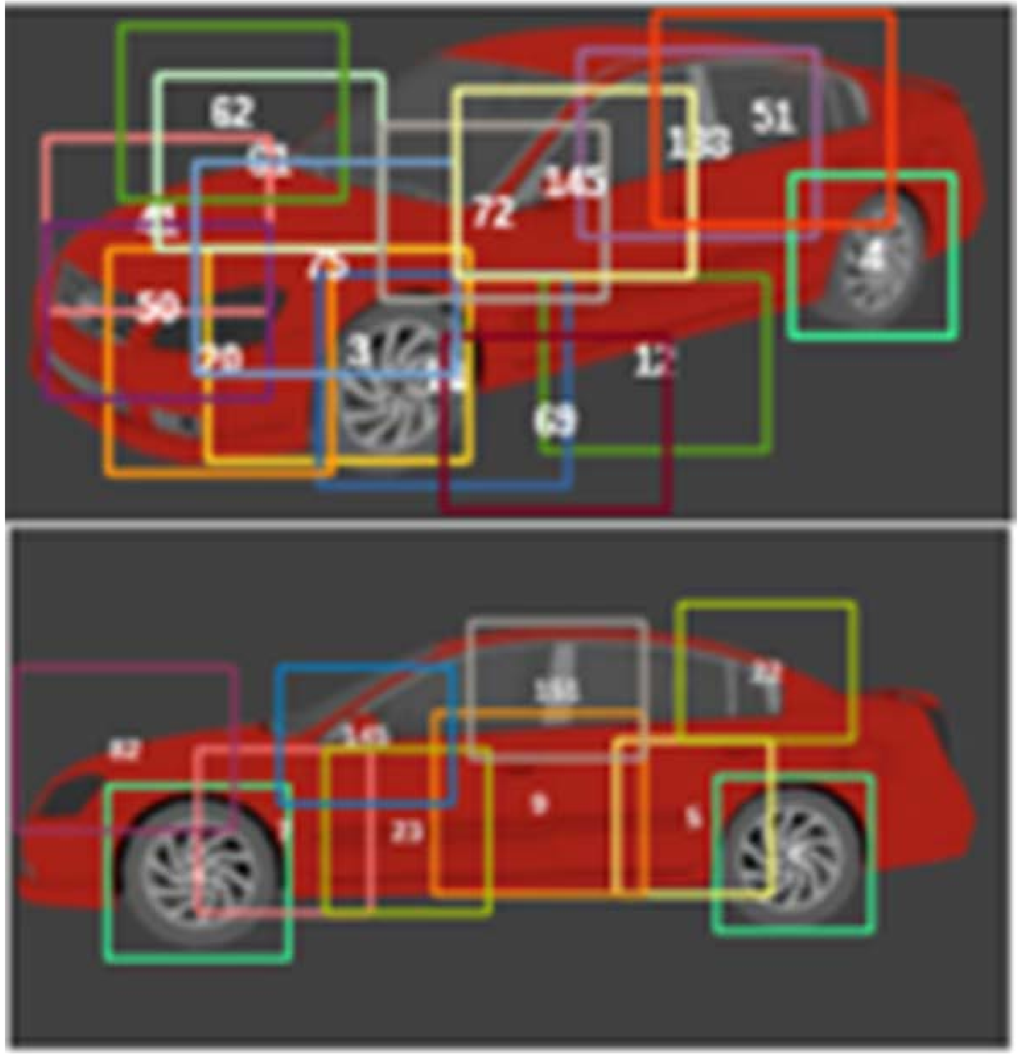
Visual Concepts as Part Detectors.

- Build a simple part detector – threshold the distance between Deep Net features vector and visual concept.
- Detect part if the population activity of deep network features is close to a visual concept.
- Compare to a detector based on single filters/neurons and with supervised methods (Support Vector Machine using Deep Net features).
- Correspondence problem – compare visual concepts with all parts on objects.
- Evaluate using datasets with ground truth.

.

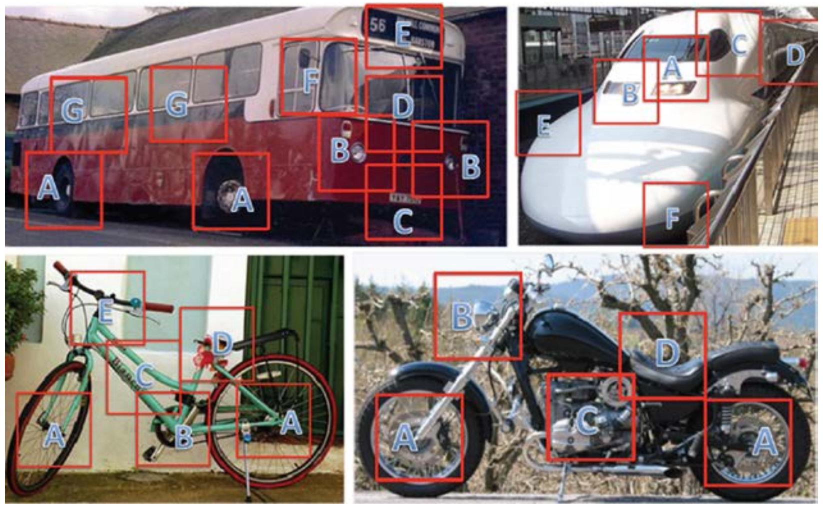
Dataset 1: Keypoints in PASCAL3D+

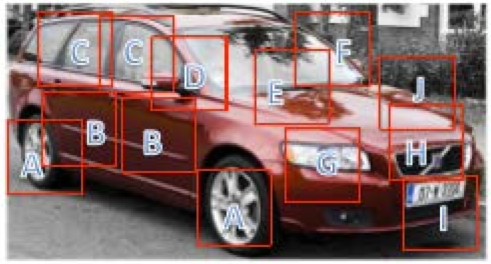
- Keypoints (10-15) in PASCAL3D+.
- Keypoints are colored circles (below).
- But keypoints are sparse and VC's give dense coverage (right).

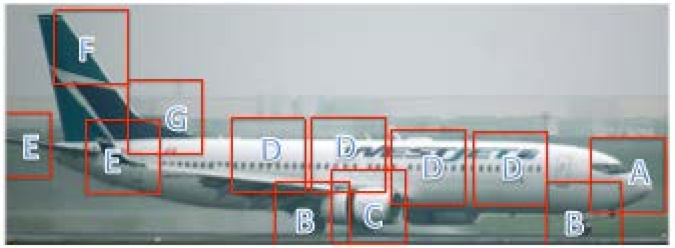


Dataset 2: Semantic Part Annotations.

- We labelled PASCAL 3D+ with semantic parts.







Findings: Visual Concepts as Detectors.

- Results for Keypoints and Semantic Parts in PASCAL3D+.
- (I) The visual concepts are better than single neurons.
- (II) the visual concepts do worse, but not too much worse, than supervised methods – Support Vector Machines (SVMs) using features from Deep Nets.
- Why?
- (I) The SVMs have more information (i.e. supervision).
- (II) Some visual concepts respond well to several (1,2, or 3) semantic parts. The evaluation penalizes these as false positives.
- (II) Several visual concepts respond well to the same semantic part.

Summary of Visual Concepts as Detectors

- The visual concepts perform well as unsupervised part detectors.
- They are beaten by supervised methods, but not badly.

- They give some insight into part representations in Deep Nets.
- They are visually very tight.

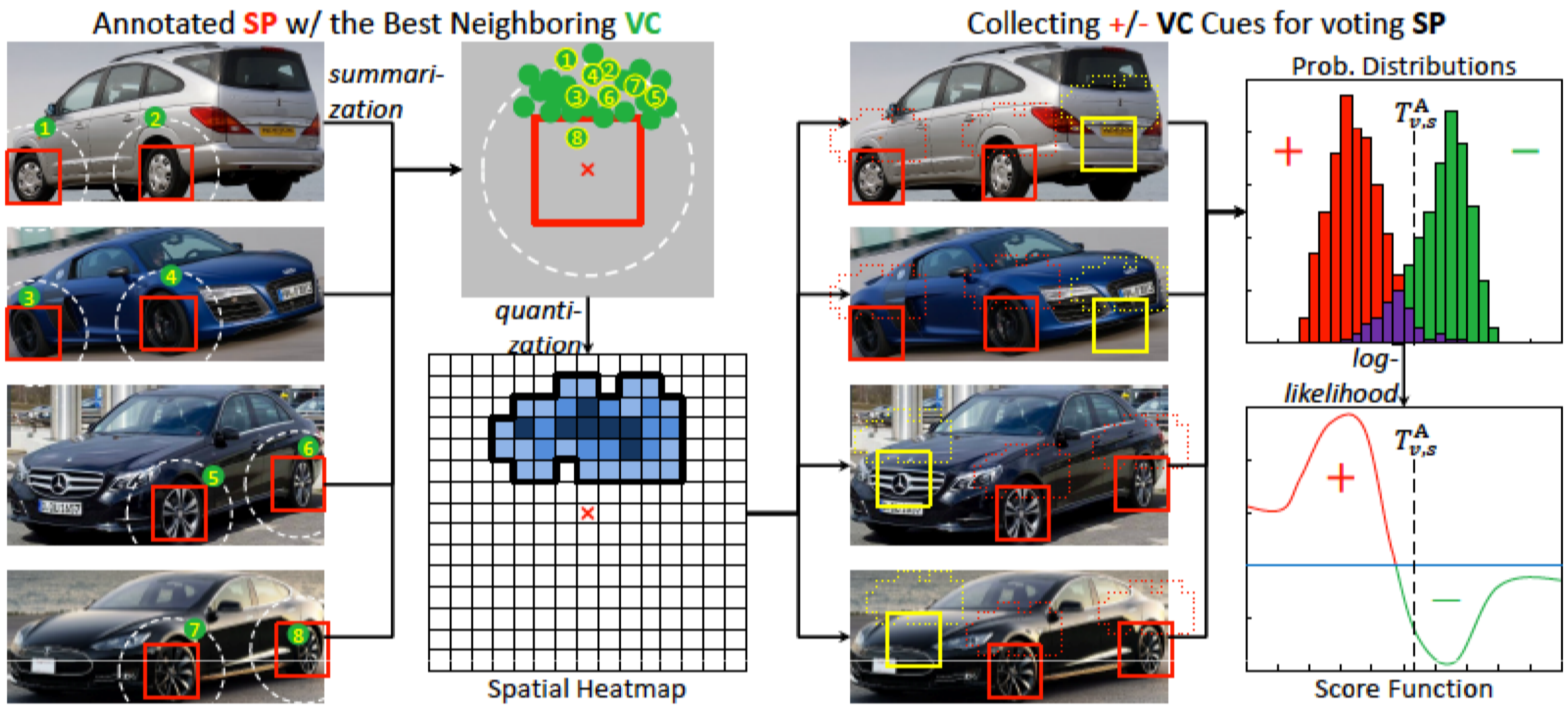
- But can we do better by combining them? Intuitively, visual concepts capture subparts of the parts.

Project 2. Combining visual concepts by voting

- VC-Voting: use a composition of visual concepts to vote for detecting parts.
- Each VC votes is based on: (i) the confidence that the VC has been detected (project 1), (ii) the relative spatial positions of the VC.
- VC-voting is not fully unsupervised because we specify which visual concepts can be used for each part (we are relaxing this cheat).
- But we now compare to the toughest opponent: Deep Nets trained directly for part detection.
- J. Wang et al. Arxiv. 2016.

VC-Voting: Visual Concepts for Wheel Detection

- Green circles denote visual concepts which are detected.
- Each visual concept has a vote (log-likelihood ratio), the spatial heatmap give the relative spatial locations.

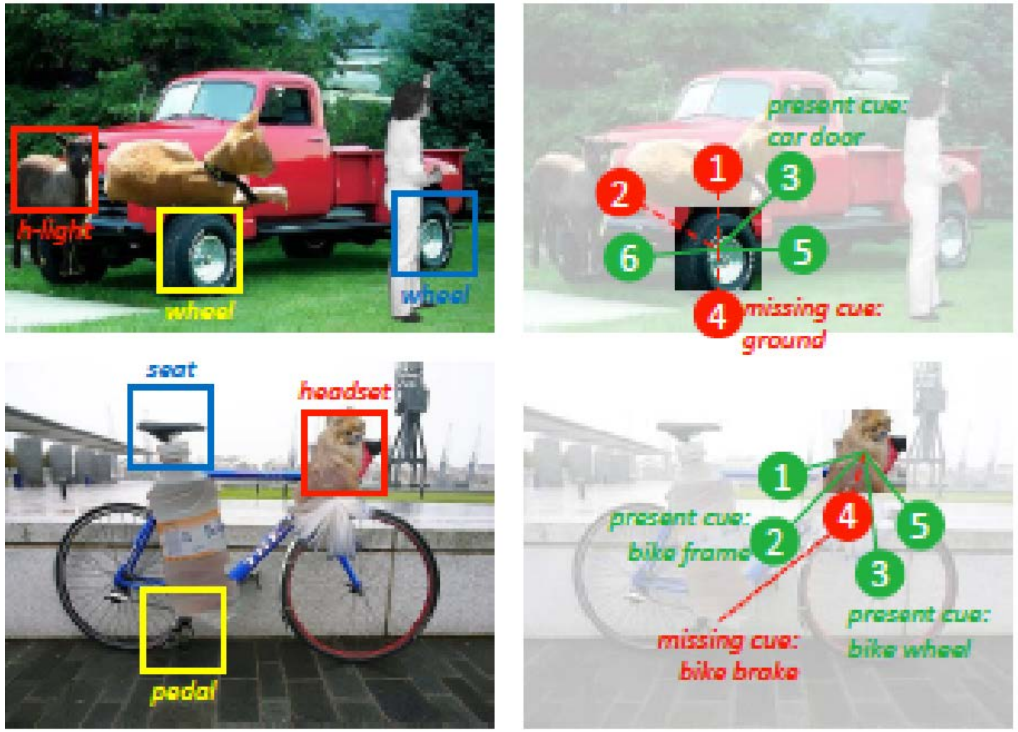


Occlusion makes the tasks more challenging

- Most real world objects are partly occluded.
- It can be shown – e.g., monkey with guitar -- that Deep Nets for object detection are sensitive to occlusion.
- Voting methods are less sensitive to occlusion because they are robust if some visual concepts are missing.
- Compare Deep Nets with VC-Voting.
- We do not use occlusion when training the Deep Nets or VC-Voting. We want to see how the methods adapt to stimuli that they have not been exposed to.
- Goal: Train on a few images, test on an infinite set.

The Occlusion Dataset

- Create dataset by introducing occlusions at random.
- Red, blue, and yellow boxes are fully-occluded, partially-occluded, and non-occluded respectively.
- Green and red circles indicate which visual concepts are detected or missing.
- Note: voting can detect a part from context even if the part itself is occluded.



Findings: detecting parts without occlusion.

- VC-Voting is slightly worse than Deep Nets trained for this task. Better on Bikes and Motor-Bikes., worse on Planes and Trains.
- VC-Voting is much better than SVM on deep features (project 1).
- Our method uses far less information – only uses a small part of the feature space.

Object	Natural Detection			
	S-VC	**SVM**	**FR**	**VT**
airplane	10.1	18.2	**45.3**	30.6
bicycle	48.0	58.1	75.9	**77.8**
bus	6.8	26.0	**58.9**	58.1
car	18.4	27.4	**66.4**	63.4
motorbike	10.0	18.6	45.6	**53.4**
train	1.7	7.2	**40.7**	35.5
mean	15.8	25.9	**55.5**	53.1

Findings: detecting parts with occlusion

- Our voting method outperforms Deep Nets as the amount of occlusion increases.

Object	2 Occluders, $0.2 \leqslant r < 0.4$				3 Occluders, $0.4 \leqslant r < 0.6$				4 Occluders, $0.6 \leqslant r < 0.8$			
	S-VC	**SVM**	**FR**	**VT**	**S-VC**	**SVM**	**FR**	**VT**	**S-VC**	**SVM**	**FR**	**VT**
airplane	6.6	12.0	**26.3**	23.2	5.0	9.7	**20.2**	19.3	3.8	7.5	**15.2**	15.1
bicycle	37.7	44.6	63.8	**71.7**	29.1	33.7	53.8	**66.3**	14.2	15.6	37.4	**54.3**
bus	2.7	12.3	**36.0**	31.3	1.2	7.3	**27.5**	19.3	0.5	3.6	**18.2**	9.5
car	7.4	13.4	32.9	**35.9**	3.7	7.7	19.2	**23.6**	1.9	4.5	11.9	**13.8**
motorbike	6.4	11.4	33.1	**44.1**	4.1	7.9	26.5	**34.7**	2.4	5.0	17.8	**24.1**
train	0.9	4.6	17.9	**21.7**	0.6	3.4	**10.0**	8.4	0.4	2.0	**7.7**	3.7
mean	10.3	16.4	35.0	**38.0**	7.3	11.6	26.2	**28.6**	3.9	6.4	18.0	**20.1**

- VC-Voting works very well for most parts, but fails badly on a few.
- Other technical issues, e.g., part proposals.

Project 1: Conclusion

- Claim: Simple intuitive methods based on composition can perform as well as Deep Nets for some tasks and be more adaptive to unforeseen factors like occlusion.
- Belief: this can help design much more effective Deep Architectures with Human-like capabilities.

- Human performance – preliminary psychophysical studies show that human performance on object/part detection is superior to Deep Nets and also to VC-Voting – so there is more to do.