Convolutional Nets and Visual Concepts

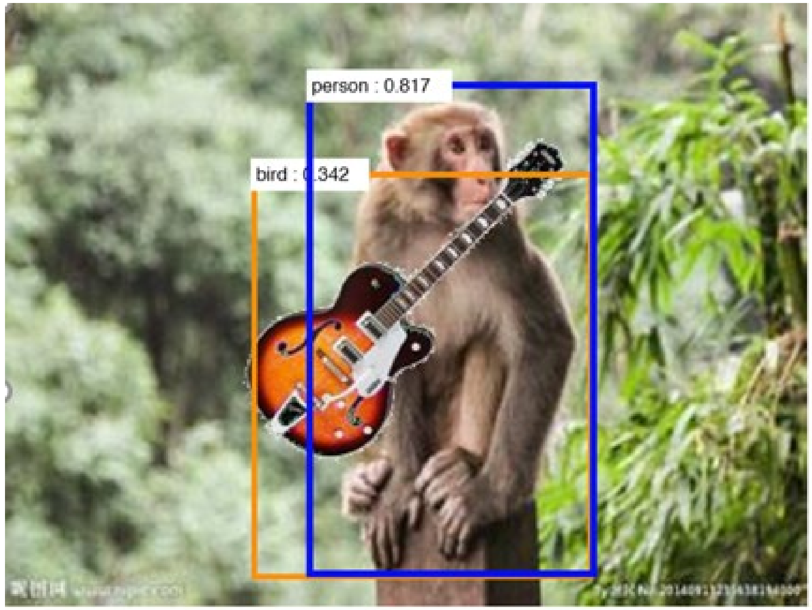
Alan Yuille

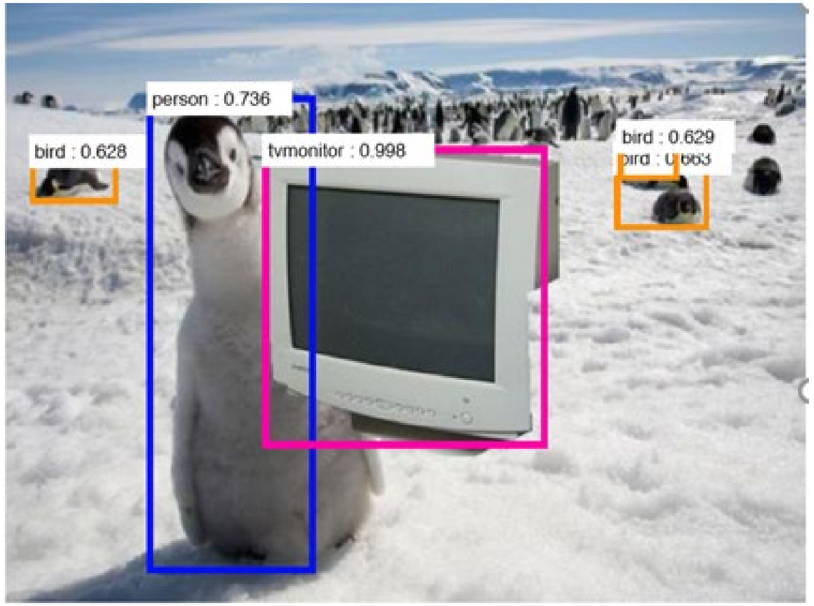
Dept. Cognitive Science and Computer Science

Johns Hopkins University

Background

- Deep Nets are hard to interpret and have unusual failure modes.
  In particular: they are sensitive to occlusion and context.





Jianyu Wang, Zhishuai Zhang, Cihang Xie, Yuyin Zhou, Vittal Premachandran, Jun Zhu, Lingxi Xie, and Alan Yuille. Visual concepts and compositional voting. *Annals of Mathematical Sciences and Applications,* 2018.
See also: A Rosenfield et al. The Elephant in the Room. Arxiv. 2018.

Visual Concepts: Internal Representations

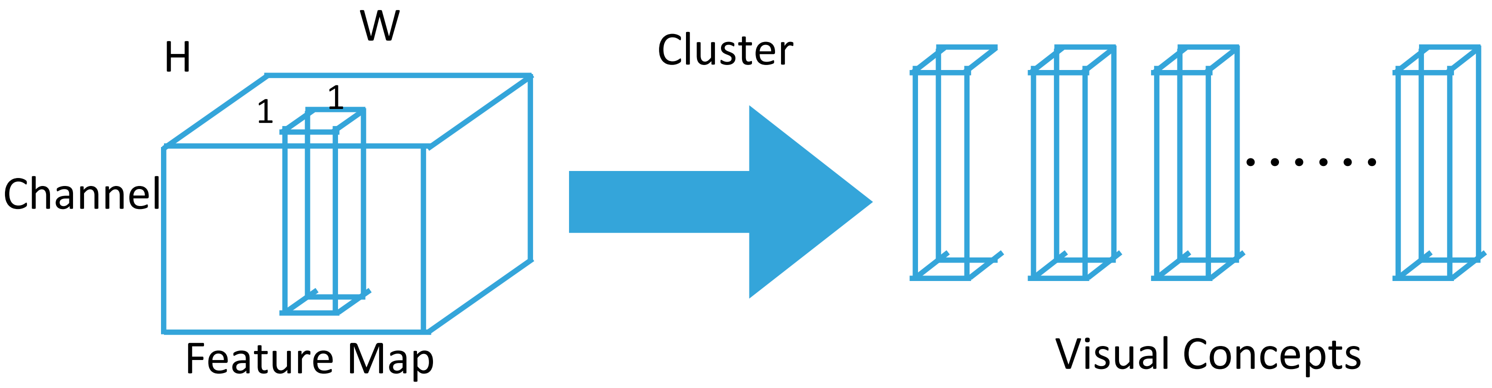
- We study internal representations within Deep Nets.
- We restrict ourselves to study vehicles at fixed scale from the Pascal3D+ dataset.
- We showed that visual concepts, encoded by feature populations, represented subparts of the vehicles.
- We quantified the visual concepts for a series of tasks including semantic part detection under occlusion.
- *J. Wang et al. Visual concepts and compositional voting. Annals of Mathematical Sciences and Applications, 2018.*
- *J. Wang et al. Detecting Semantic Parts on Partly Occluded Objects. BMVC. 2017.*

Background

- It has been shown (e.g., B. Zhou et al. ICLR 2015) that deep nets contain internal representations represented by neural features. The findings included:
- (I) If Deep Nets are trained to perform scene recognition, then the internal representations correspond to objects.
- (II) If Deep Nets are trained to perform object recognition, then the internal representations correspond to object parts.

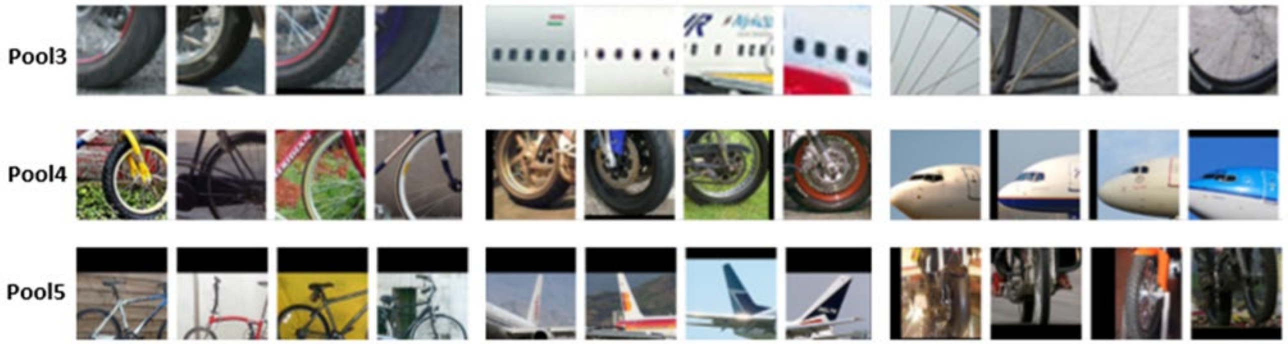
Visual Concepts

- We conjectured that subparts of objects are encoded by populations of feature vectors – instead of by features themselves.
- These *visual concepts* were found by clustering the feature vectors. We restricted ourselves to vehicles from Pascal3D+ and fixed the scale of the objects.



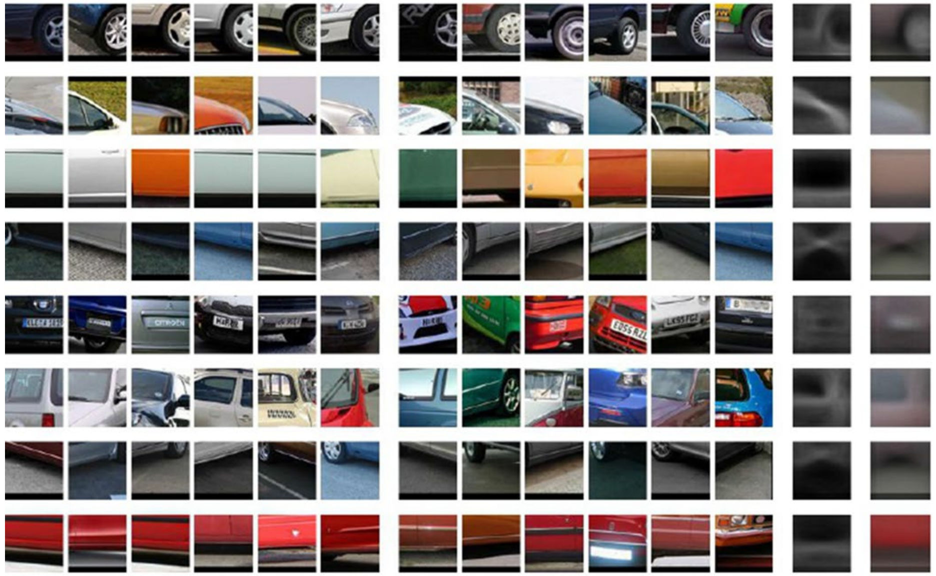
Visual Concepts: Clustering

- The clustering was done using k-means with k=200 (alternative clustering methods, and alternative values of k gave similar results).
- The clustering was done at different levels of the Deep Net. E.g., Pool3, Pool4, Pool5. Results were similar for AlexNet and VGG.
- Visual Concepts correspond to parts of objects. VCs at higher layers correspond to larger parts (e.g., Pool4 wheel, Pool3 wheel-part).



Visual Concepts: Perceptually Tight

- *Findings 1: The visual concepts were perceptually tight. Image patches corresponding to the same visual concept are very similar.*
- We show the closest 6 image patches (left), *a random sample of 6 patches from the top 500 image patches* (center), and *the mean of the edge map and of the patches of the top 500 patches* (right).



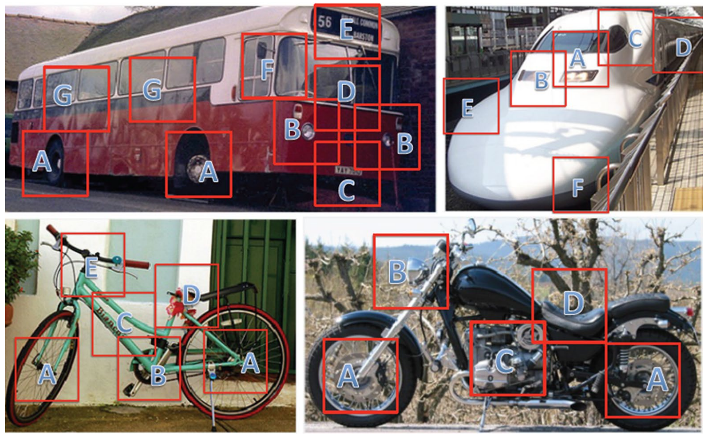
Visual Concepts: Coverage of the Object

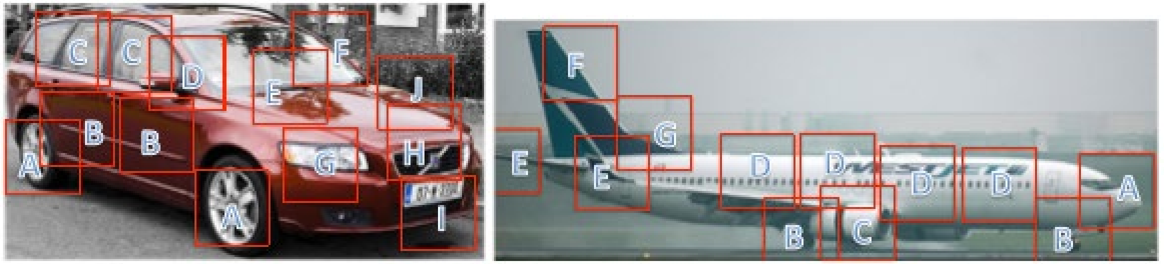
- *Visual Concepts respond to (cover) almost all parts of the object.*
- Here are 44 (out of 170) VCs for cars.
- This can be quantified, by showing that the objects could be represented in terms of VCs by binary encoding (see later).

To Explore: We Annotate Semantic Parts.

- We annotated the vehicles in PASCAL 3D+.

To create the *Vehicle Semantic Part dataset.*





VCs as Key-Point, Semantic Part Detectors.

- VCs were fairly good for detecting key-points and semantic parts of the Vehicles. But much worse than supervised models.

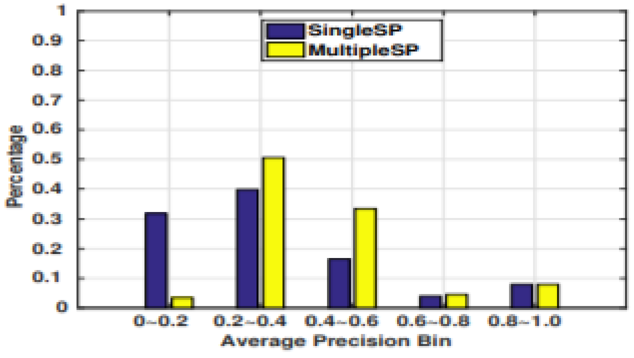
- Key-Points.

13 K-Ps for Bike.

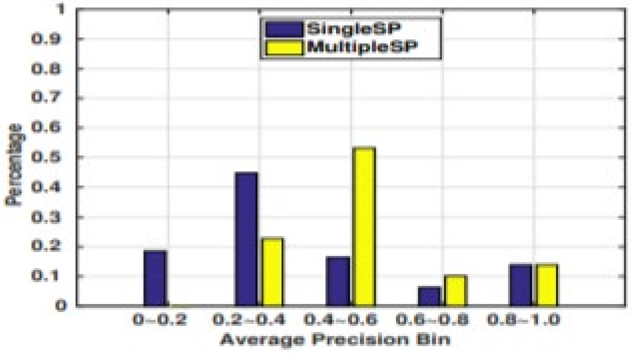
Bike	1	2	3	4	5	6	7	8	9	10	11	12	13	mAP
SF	.77	.84	.89	.91	.94	.92	.94	.91	.91	.56	.53	.15	.40	.75
VC	.91	.95	.98	.96	.96	.96	.97	.96	.97	.73	.69	.19	.50	.83

- Semantic Parts.

Yellow bars show the best APs for each VC.



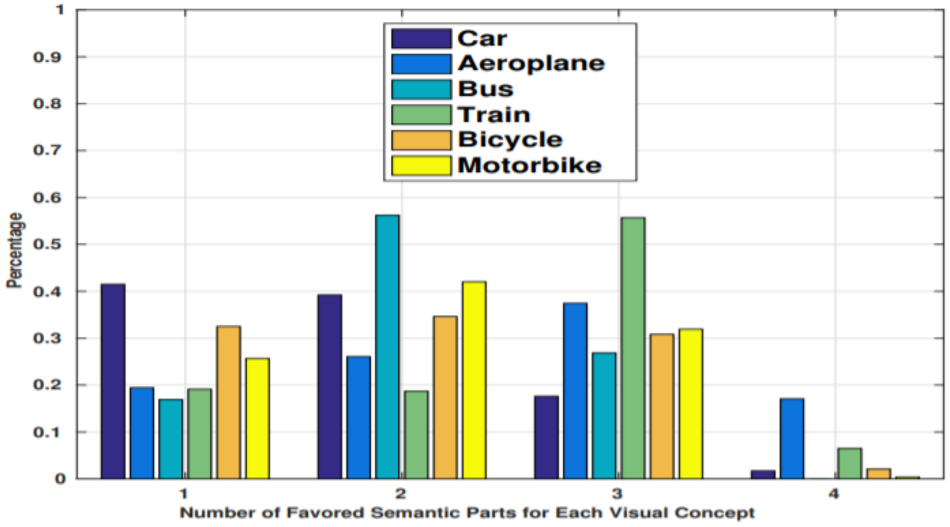
(a) car

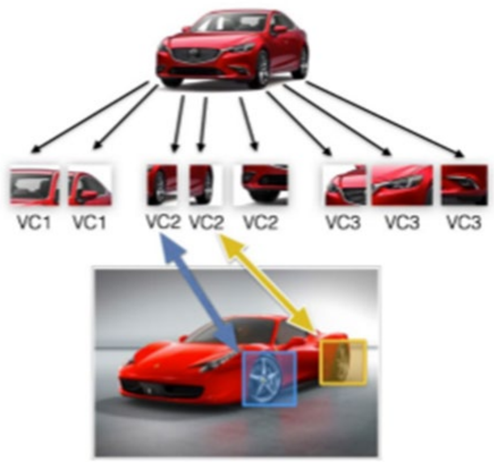


(e) bicycle

VCs detect subparts of Semantic Parts

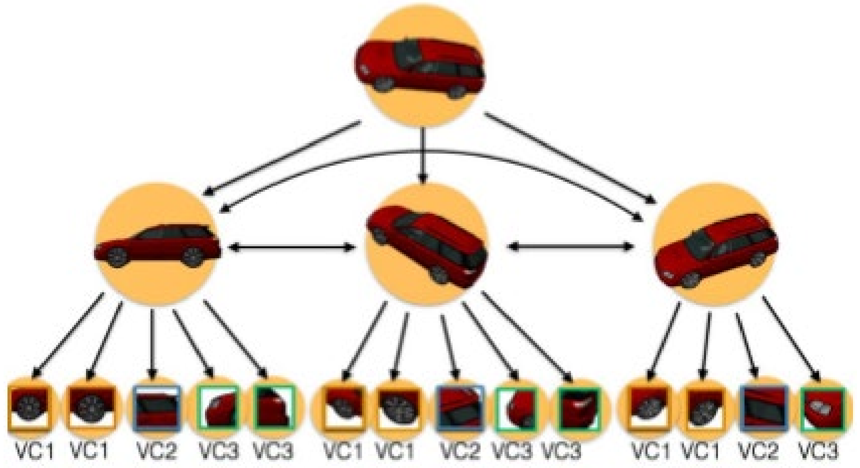
- VCs can act as unsupervised detectors for key-points and semantic-parts. Their Average Precisions (APs) are weaker than supervised methods.
- We observe that most VCs respond to several different semantic parts (typically 1-4). The VCs correspond to subparts of semantic parts (which are shared).



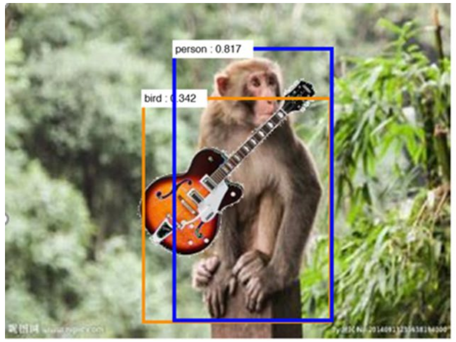


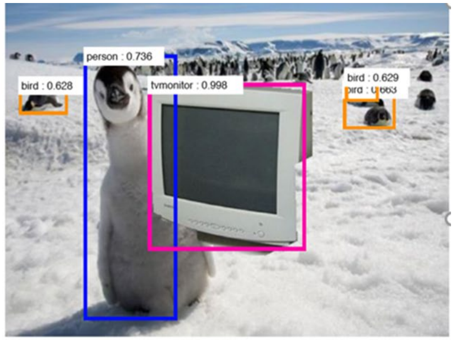
Combine VCs to detect Semantic Parts

- We design a compositional model for detecting semantic parts. Each model consists of a set of VCs which fire in different spatial positions. (Illustrated for object – car – instead of semantic part).
- Compositional Voting: each VC votes for the semantic part (depending on spatial position).



Semantic Part Detection with Occlusion





- We introduce occlusion to make semantic part detection more challenging. *Vehicle Occlusion Dataset.*
- Our intuition is that Deep Nets have difficulty with occlusion. *But compositional voting is likely to be most robust*. The occluded VC will not respond, but the un-occluded VCs will still vote.
- *Compositional voting* also includes context, image information outside the semantic part, because this is also robust.

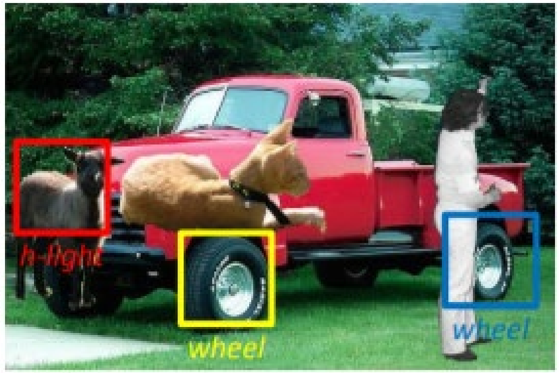
Detecting Semantic Parts with Occlusion

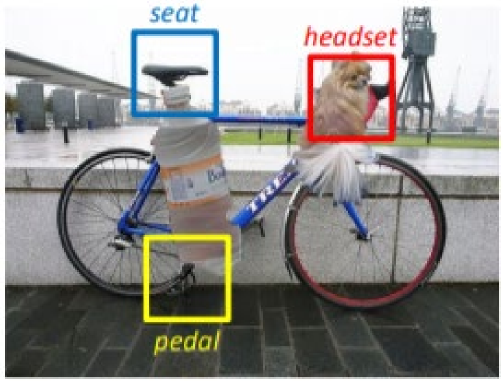
- In the occlusion dataset semantic parts

  can be: (i) fully occluded (red)

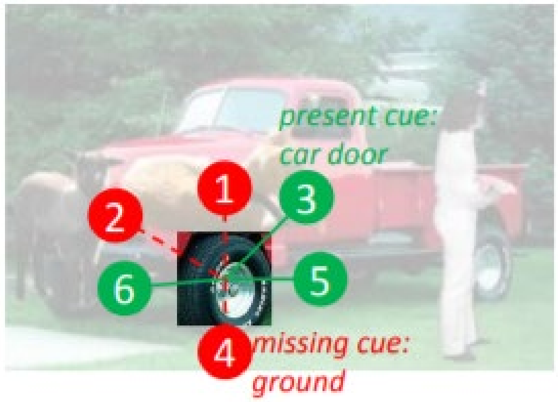
  (ii) partially occluded (blue)

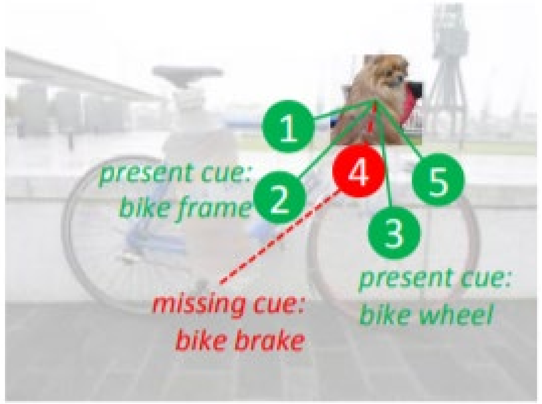
  (iii) un-occluded (yellow).





- Compositional voting uses VCs on and off the semantic parts. If a VC is detected (green) then it votes for the semantic part. If a VC is occluded (red) then it gives no vote.
- Note: a semantic part can be detected even if it is fully occluded.





Compositional Voting: Detect Semantic Parts

- The compositional voting method (VT) outperforms alternatives like Deep Nets if there is significant occlusion.
- *Main idea: explicit representation of subparts (by VC) enables the algorithm to switch them on and off automatically. This makes them robust to occlusion.*

Object	2 Occ's, $0.2 \leqslant r < 0.4$			3 Occ's, $0.4 \leqslant r < 0.6$			4 Occ's, $0.6 \leqslant r < 0.8$		
	SV	**FR**	**VT**	**SV**	**FR**	**VT**	**SV**	**FR**	**VT**
airplane	12.0	**26.8**	23.2	9.7	**20.5**	19.3	7.5	**15.8**	15.1
bicycle	44.6	65.7	**71.7**	33.7	54.2	**66.3**	15.6	37.7	**54.3**
bus	12.3	**41.3**	31.3	7.3	**32.5**	19.3	3.6	**21.4**	9.5
car	13.4	**35.9**	**35.9**	7.7	22.0	**23.6**	4.5	**14.2**	13.8
motorbike	11.4	35.9	**44.1**	7.9	28.8	**34.7**	5.0	19.1	**24.1**
train	4.6	20.0	**21.7**	3.4	**11.1**	8.4	2.0	**7.2**	3.7
mean	16.4	37.6	**38.0**	11.6	28.2	**28.6**	6.4	19.2	**20.1**

- *J. Wang et al. BMVC (2017). See also, Z. Zhang et al. CVPR. 2018.*

Visual Concepts: Summary

- The Deep Nets encode representations of the parts. These are stored by the activity patterns of the feature vectors (individual features were less successful – quantitatively). *Note: vehicles only (rigid classes) and fixed scale.*
- *Making this representation explicit – e.g., by compositional voting – enables us to detect semantic parts despite heavy occlusion.* The algorithm can automatically switch off subparts (VCs) if they are not detected in the correct locations.
- It is harder for Deep Nets to deal with occluders, because their representations are not explicit, so it is difficult to switch parts off.
- *Can we extend this too classify objects? 2D Compositional Networks.*