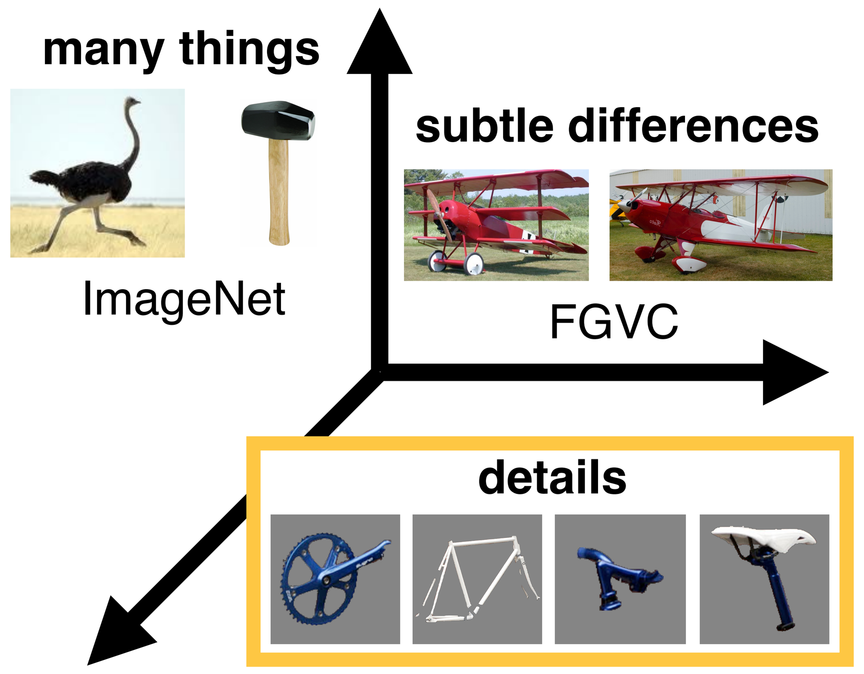
# Understanding Objects in Detail with Fine-grained Attributes

Andrea Vedaldi Siddharth Mahendran Stavros Tsogkas Subhransu Maji Ross Girshick Juho Kannala Esa Rahtu
Iasonas Kokkinos Matthew B. Blaschko Ben Taskar David Weiss Karen Simonyan Naomi Saphra Sammy Mohamed

## Detailed object understanding

Driving challenges in object recognition:




### Contributions

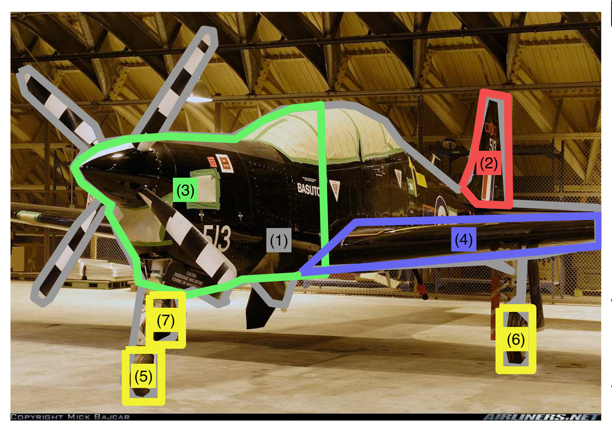
**Describing object in details:** parts and corresponding fine-grained attributes.

- A direct evaluation of fine-grained part detection and description.
- A supporting **Object in Detail (OID) dataset**.
- Efficient coarse-to-fine detailed part matching.

## The OID challenge and data

Goal: **directly evaluate detailed image understanding tasks**.

- ~7,500 aircraft images, 100 years of aviation
- 5 parts with 18 attributes



1 **aeroplane** facing-direction: SW; is-airliner: no; is-cargo-plane: no; is-glider: no; is-military-plane: yes; is-propellor-plane: yes; is-seaplane: no; plane-location: on ground/water; plane-size: medium plane; wing-type: single wing plane; undercarriage-arrangement: one-front-two-back; airline: UK–Air Force; model: Short S-312 Tucano T1 2 **vertical stabilizer** tail-has-engine: no-engine 3 **nose** has-engine-or-sensor: has-engine 4 **wing** wing-has-engine: no-engine 5 **undercarriage** cover-type: retractable; group-type: 1-wheel-1-axle; location: front-middle 6 **undercarriage** cover-type: retractable; group-type: 1-wheel-1-axle; location: back-left 7 **undercarriage** cover-type: retractable; group-type: 1-wheel-1-axle; location: back-right.

### Data definition and construction

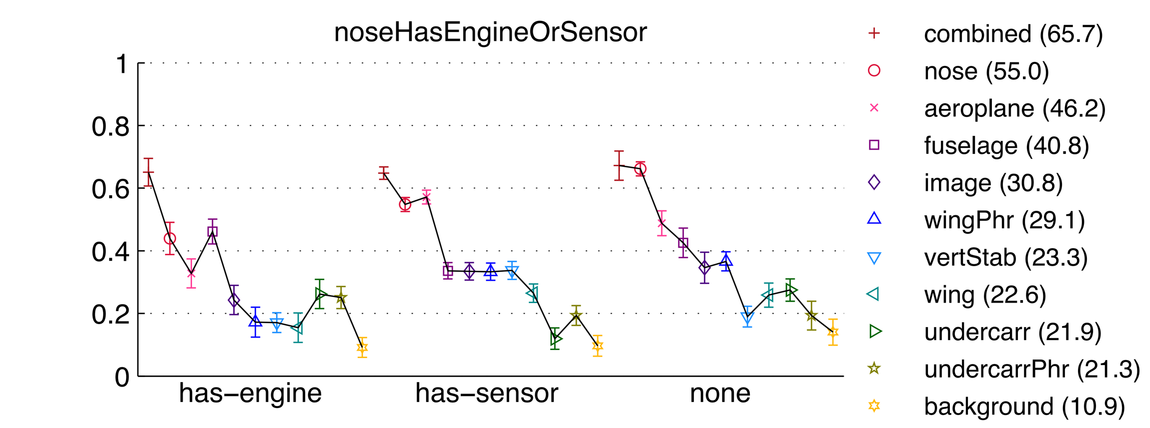
- Part and attribute definitions were extracted from human descriptions of objects.
- Amazon Mechanical Turk for part segmentation and attribute collection.
- Three weeks of intense work of several researchers in the CLSP Summer Workshop.

## Local vs global attribute modeling

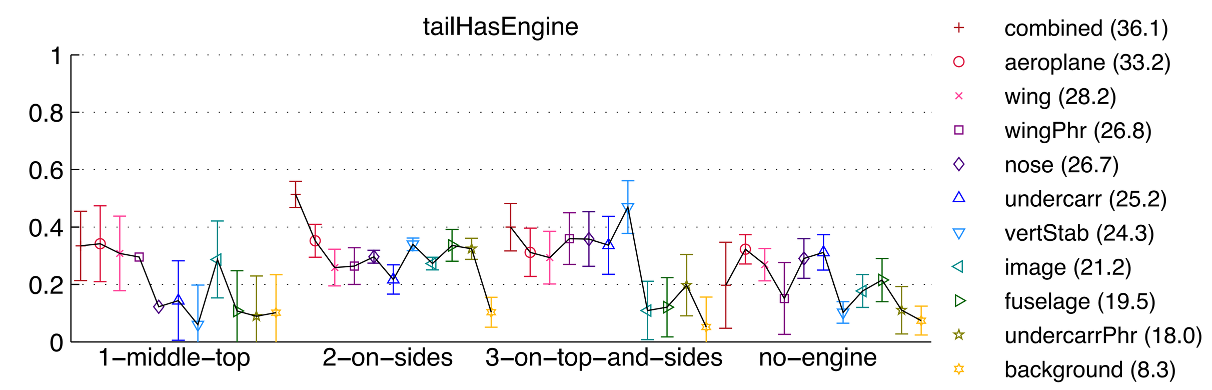
- **Locality of meaning.** Part attributes are semantically local (e.g. round nose). A modular and transferable visual model should be based on the part appearance only.
- **Globality of evidence.** Local evidence is often weak and part attributes are often best predicted by global cues.

### Examples

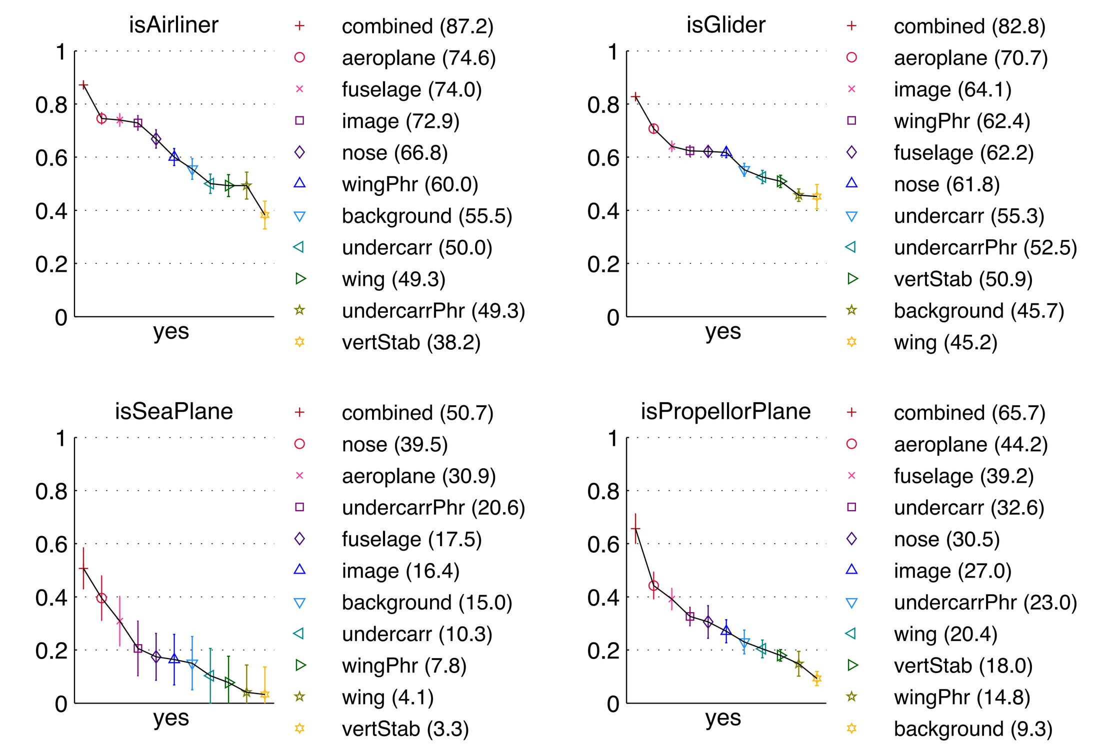
1) This nose attribute is well predicted by the nose appearance, but adding context is better still.




2) This tail attribute is not well predicted by the appearance of the tail (vertical stabilizer).

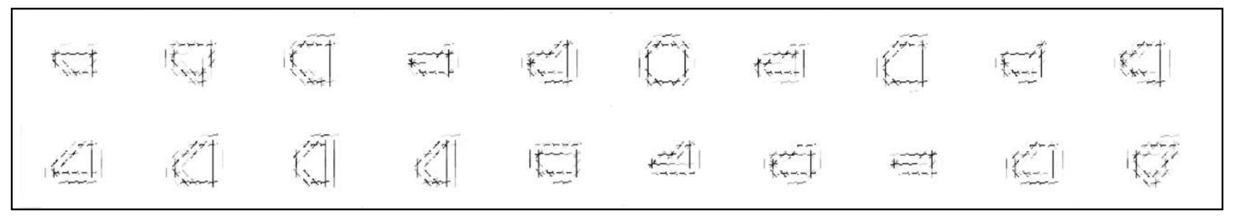


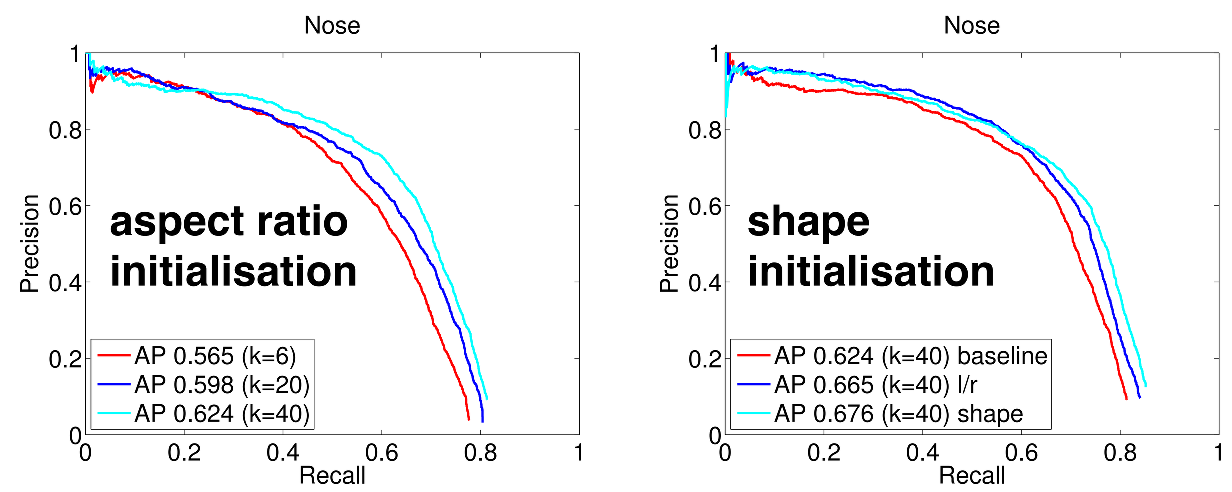

3) Global attributes are often best predicted by the overall appearance of the plane, but in some cases parts are better when considered in isolation.




## The richness of part appearances

- Large mixture models in DPMs perform best in detecting detailed parts.
- Detailed annotations can be exploited in initializing part templates.





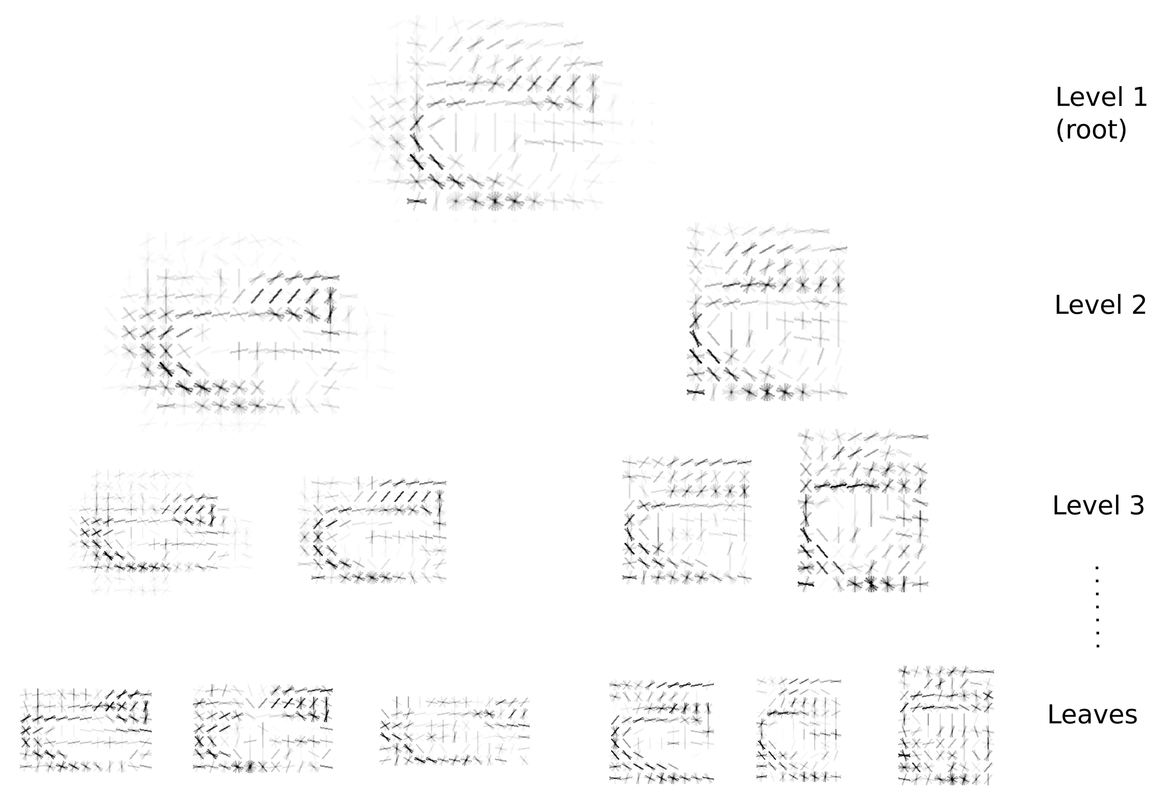

## Coarse-to-Fine Template Hierarchy

We introduce a fast algorithm to accelerate detection with many detailed templates.

- Templates are greedily organized in a tree.
- Each parent filter is the average of its aligned children.
- The parent score **uniformly bounds** the children scores:

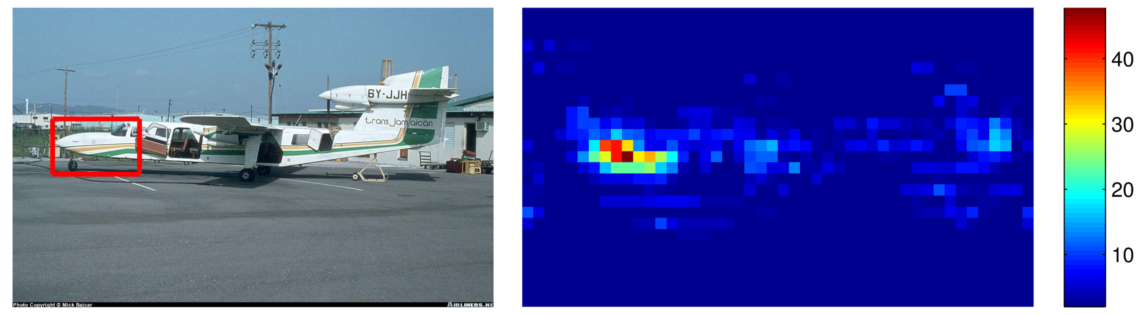
children filter score | parent (average) filter score | cheap bound

$$\max_{m\in\{1,\dots\}} \langle f_m, I\rangle \leq \langle \hat{f}, I\rangle + \sqrt{M\bar{E}(f_1,\dots,f_M,I)/p_e}$$

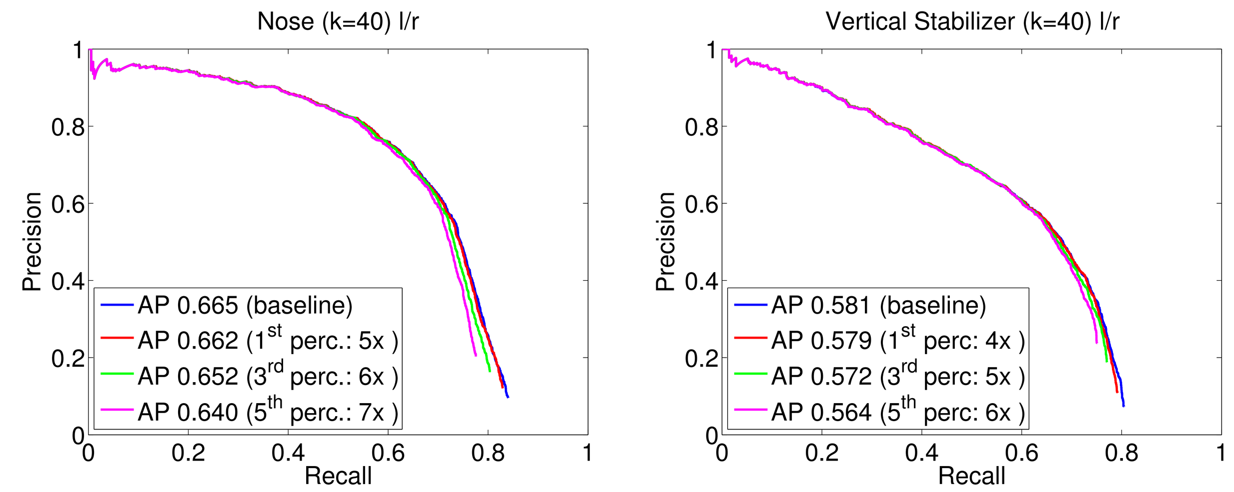



## CTF speedup

The bound can be used to cull search locations. Most templates are evaluated only where the part is likely to be found:



Probabilistic but tight bounds allow for a 5-7-fold speedup with negligible accuracy loss:




### CTF bound derivation

The probability of any of $M$ children scores to be larger than the parent score is no more than $M$ times the probability of an individual violation (union bound):

$$P[\exists m : \langle f_m, I\rangle > \langle \hat{f}, I\rangle] \leq M \sup_m P[\langle f_m, I\rangle > \langle \hat{f}, I\rangle]$$

Chebyshev's inequality $P[x > \alpha] \leq E[x^2]/\alpha$ allows bounding the individual terms by

$$P\left[\langle f_m, I\rangle > \langle \hat{f}, I\rangle + \sqrt{\frac{E[\langle f_m - \hat{f}, I\rangle^2]}{p_e}}\right] \leq p_e$$

Expected value of square residual:

$$E\left[\langle f_m - \hat{f}, I\rangle^2\right] = \sum_c V_{cm}\|I_c\|^2$$

$V_{cm}$ = 2nd-moment of HOG cell filter approximation residual.

## Acknowledgments

The bulk of this work was carried in the CLSP Summer Workshop 2012.

Ben Taskar's Memorial http://www.bentaskar.com