The PASCAL Visual Object Classes Challenge 2007 (VOC2007)

Part 1 – Challenge & Classification Task

Mark Everingham
Luc Van Gool
Chris Williams
John Winn
Andrew Zisserman

Dataset

- Images downloaded from **flickr**
  - Collected January 2007 (some Christmas bias)
  - 500,000 images downloaded and random subset selected for annotation
- Annotation in one session with written guidelines
  - 20 classes
  - Bounding box
  - Viewpoint: front, rear, left, right, unspecified
  - "Truncated" flag: Bounding box ≠ object extent
  - "Difficult" flag: Objects ignored in challenge

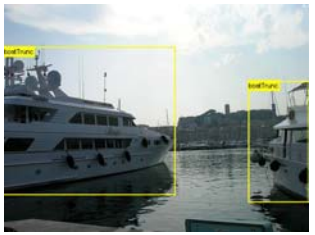
Examples

Aeroplane

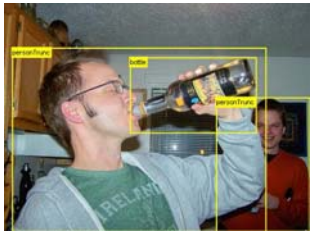
Bicycle

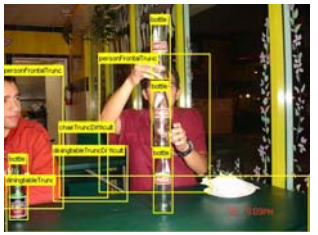
Bird

Boat



Bottle

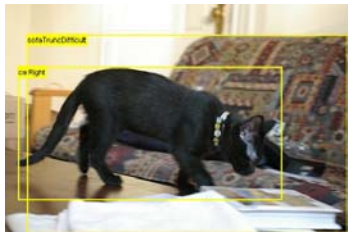




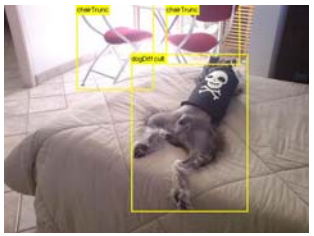
Bus

Car

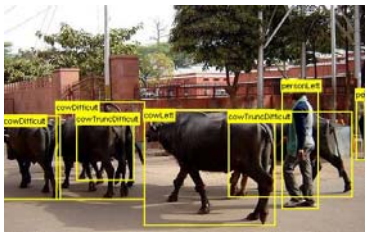
Cat



Chair

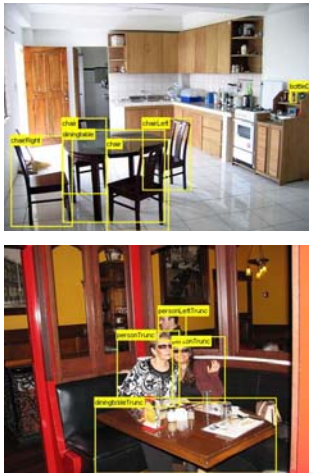


Cow

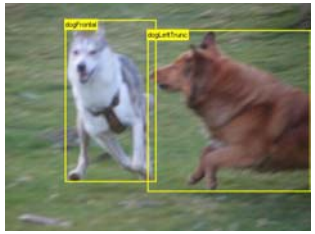


Examples

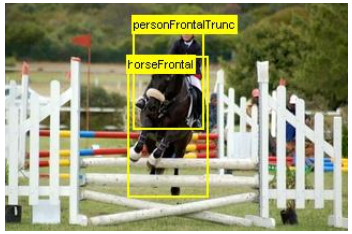
Dining Table

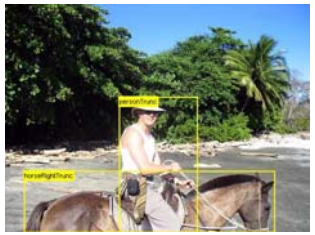


Dog



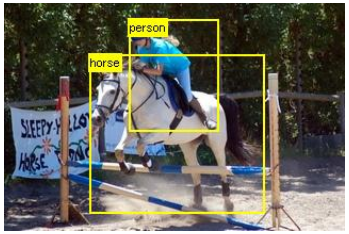
Horse

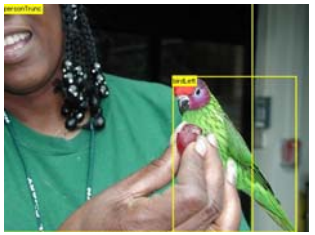




Motorbike

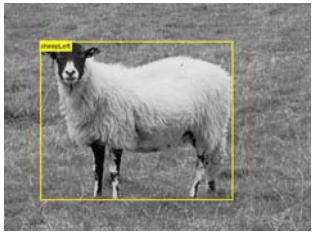
Person

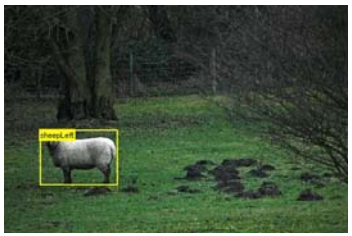




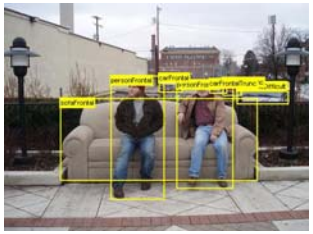
Potted Plant

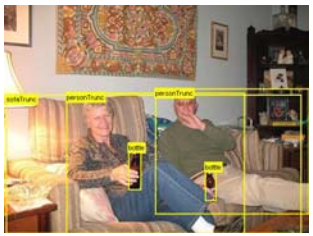
Sheep





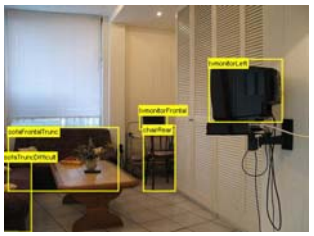
Sofa





Train

TV/Monitor



Dataset Statistics

	train		val		trainval		test	
	Images	Objects	Images	Objects	Images	Objects	Images	Objects
Aeroplane	112	151	126	155	238	306	204	285
Bicycle	116	176	127	177	243	353	239	337
Bird	180	243	150	243	330	486	282	459
Boat	81	140	100	150	181	290	172	263
Bottle	139	253	105	252	244	505	212	469
Bus	97	115	89	114	186	229	174	213
Car	376	625	337	625	713	1,250	721	1,201
Cat	163	186	174	190	337	376	322	358
Chair	224	400	221	398	445	798	417	756
Cow	69	136	72	123	141	259	127	244
Diningtable	97	103	103	112	200	215	190	206
Dog	203	253	218	257	421	510	418	489
Horse	139	182	148	180	287	362	274	348
Motorbike	120	167	125	172	245	339	222	325
Person	1,025	2,358	983	2,332	2,008	4,690	2,007	4,528
Pottedplant	133	248	112	266	245	514	224	480
Sheep	48	130	48	127	96	257	97	242
Sofa	111	124	118	124	229	248	223	239
Train	127	145	134	152	261	297	259	282
Tvmonitor	128	166	128	158	256	324	229	308
Total	2,501	6,301	2,510	6,307	5,011	12,608	4,952	12,032

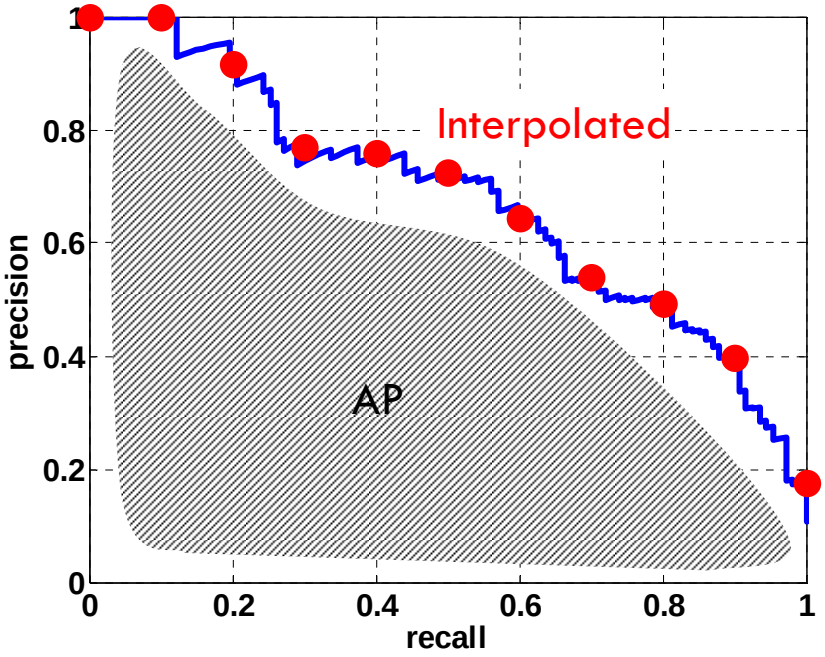
Classification Challenge

- Predict whether at least one object of a given class is present in an image

- Competition 1: Train on the supplied data
  - Which methods perform best given specified training data?
- Competition 2: Train on any (non-test) data
  - How well do state-of-the-art methods perform on these problems?
  - No results submitted

Evaluation

- Average Precision [TREC] averages precision over the entire range of recall
  - Curve interpolated to reduce influence of "outliers"



- A good score requires both high recall and high precision
- Application-independent

Precision/Recall vs. ROC

- VOC2006 used ROC for classification task
- Why not continue using ROC?
  - Area under curve (AUC) on ROC appeared to have "saturated"
    - Many methods giving AUC >95%
  - Difficult to interpret ROC
    - Is AUC of 95% useful?
- Why use precision/recall?
  - More intuitive for an "image retrieval" application
  - "Early" errors more visible on curve
  - AP empirically more sensitive and in general agreement with AUC in terms of ranking

Methods

- **"Bag of visual words and beyond"**
  - Sparse vs. dense interest points
  - Multiple feature types/classifiers
  - Spatial/non-spatial histograms
- Interest Operators
  - LoG, Harris-Laplacian
  - Color Harris
  - Edgels
  - Dense grids
  - Segmented regions

Methods: Bag of visual words and beyond

- Features
  - SIFT
  - Color histogram
  - Pairs of adjacent edge segments (PAS)
  - Textons
  - “Wiccest”
- Codebooks
  - K-Means
  - Random clustering forests
  - Soft per-image clustering (GMM)

Methods: Bag of visual words and beyond

- Histograms
  - No spatial information
  - Spatial pyramid
  - Bigrams of neighbouring features on a grid
- Feature Fusion
  - Concatenation of histograms
  - Voting of classifier per feature type
  - Linear SVM
  - Learnt kernel weighting feature type

Methods: Bag of visual words and beyond

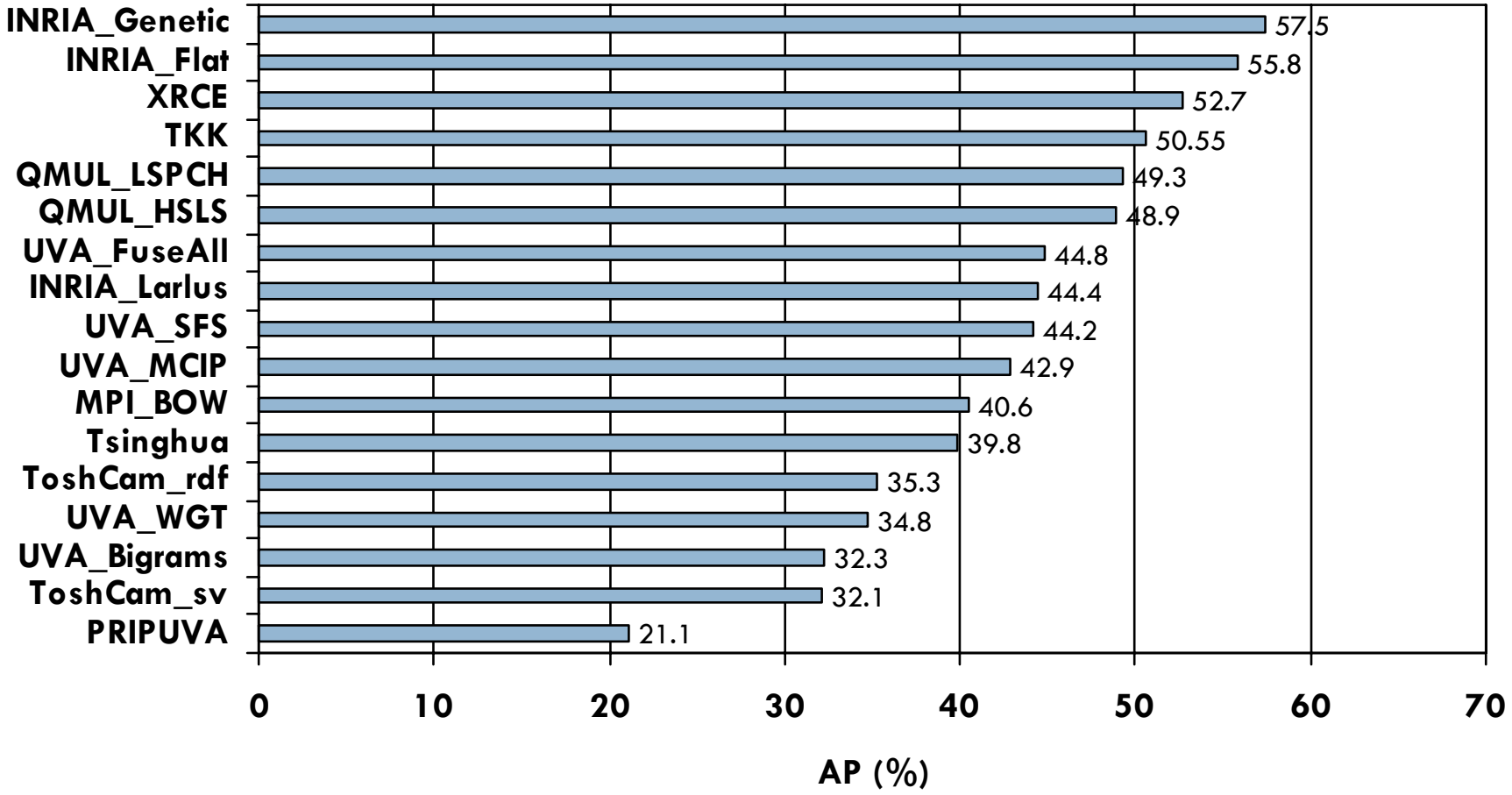
- Classifiers
  - SVM – chi-squared, Earth Mover's Distance, learnt RBF
  - Randomized decision forest
  - Fisher kernel logistic regression
  - RankBoost
- Other
  - "Semantic" features: sky, vegetation, etc.
- No attempts at "classification by detection"

Results: AP by Method and Class

	aero	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	mbike	pers	plant	sheep	sofa	train	tv
INRIA_Larlus	62.6	54.0	32.8	47.5	17.8	46.4	69.6	44.2	44.6	26.0	38.1	34.0	66.0	55.1	77.2	13.1	29.1	36.7	62.7	43.3
INRIA_Flat	74.8	62.5	51.2	69.4	29.2	60.4	76.3	57.6	53.1	41.1	54.0	42.8	76.5	62.3	84.5	35.3	41.3	50.1	77.6	49.3
INRIA_Genetic	77.5	63.6	56.1	71.9	33.1	60.6	78.0	58.8	53.5	42.6	54.9	45.8	77.5	64.0	85.9	36.3	44.7	50.6	79.2	53.2
MPI_BOW	58.9	46.0	31.3	59.0	16.9	40.5	67.2	40.2	44.3	28.3	31.9	34.4	63.6	53.5	75.7	22.3	26.6	35.4	60.6	40.6
PRIPUVA	48.6	20.9	21.3	17.2	6.4	14.2	45.0	31.4	27.4	12.3	14.3	23.7	30.1	13.3	62.0	10.0	12.4	13.3	26.7	26.2
QMUL_HSLS	70.6	54.8	35.7	64.5	27.8	51.1	71.4	54.0	46.6	36.6	34.4	39.9	71.5	55.4	80.6	15.8	35.8	41.5	73.1	45.5
QMUL_LSPCH	71.6	55.0	41.1	65.5	27.2	51.1	72.2	55.1	47.4	35.9	37.4	41.5	71.5	57.9	80.8	15.6	33.3	41.9	76.5	45.9
TKK	71.4	51.7	48.5	63.4	27.3	49.9	70.1	51.2	51.7	32.3	46.3	41.5	72.6	60.2	82.2	31.7	30.1	39.2	71.1	41.0
ToshCam_rdf	59.9	36.8	29.9	40.0	23.6	33.3	60.2	33.0	41.0	17.8	33.2	33.7	63.9	53.1	77.9	29.0	27.3	31.2	50.1	37.6
ToshCam_svm	54.0	27.1	30.3	35.6	17.0	22.3	58.0	34.6	38.0	19.0	27.5	32.4	48.0	40.7	78.1	23.4	21.8	28.0	45.5	31.8
Tsinghua	62.9	42.4	33.9	49.7	23.7	40.7	62.0	35.2	42.7	21.0	38.9	34.7	65.0	48.1	76.9	16.9	30.8	32.8	58.9	33.1
UVA_Bigrams	61.2	33.2	29.4	45.0	16.5	37.6	54.6	31.3	39.9	17.2	31.4	30.6	61.6	42.4	74.6	14.5	20.9	23.5	49.9	30.0
UVA_FuseAll	67.1	48.1	43.3	58.1	19.9	46.3	61.8	41.9	48.4	27.8	41.9	38.5	69.8	51.4	79.4	32.5	31.9	36.0	66.2	40.3
UVA_MCIP	66.5	47.9	41.0	58.0	16.8	44.0	61.2	40.5	48.5	27.8	41.7	37.1	66.4	50.1	78.6	31.2	32.3	31.9	66.6	40.3
UVA_SFS	66.3	49.7	43.5	60.7	18.8	44.9	64.8	41.9	46.8	24.9	42.3	33.9	71.5	53.4	80.4	29.7	31.2	31.8	67.4	43.5
UVA_WGT	59.7	33.7	34.9	44.5	22.2	32.9	55.9	36.3	36.8	20.6	25.2	34.7	65.1	40.1	74.2	26.4	26.9	25.1	50.7	29.7
XRCE	72.3	57.5	53.2	68.9	28.5	57.5	75.4	50.3	52.2	39.0	46.8	45.3	75.7	58.5	84.0	32.6	39.7	50.9	75.1	49.5

- INRIA_Genetic is best for 19/20 classes, INRIA_Flat and XRCE close 2nd and 3rd

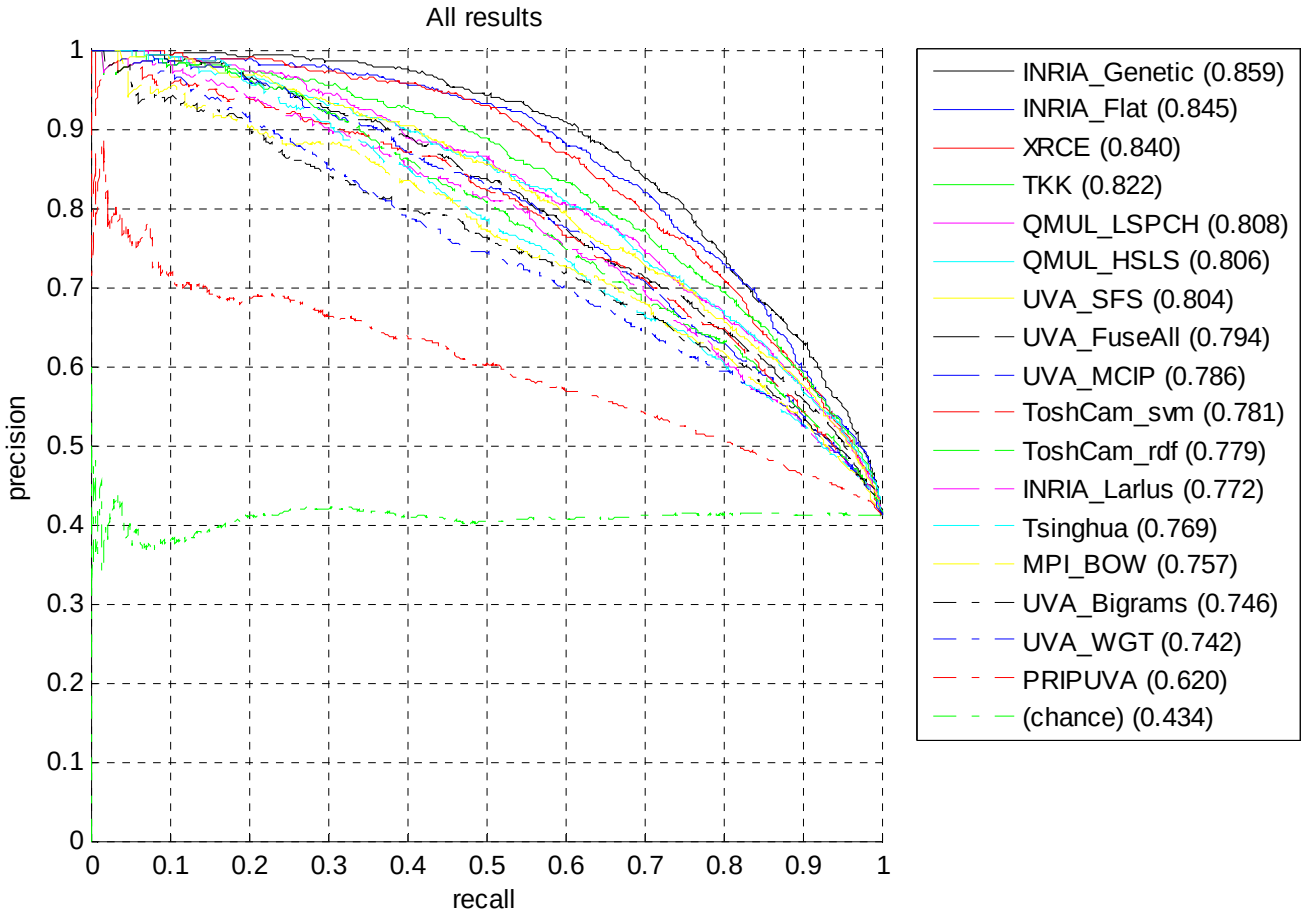
Median AP by Method



- Small differences between leading methods
- Convincing improvement over best method in 2006?

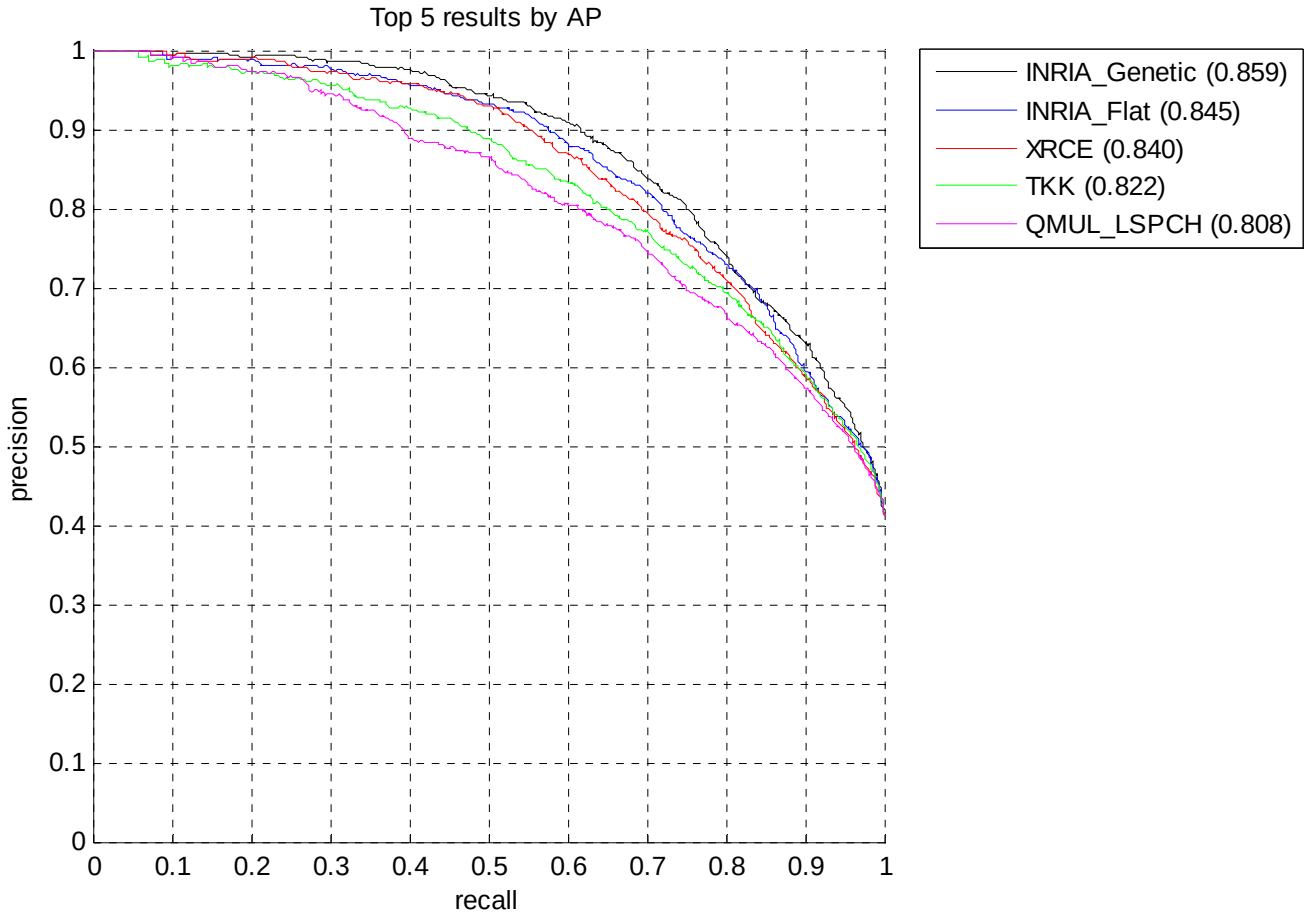
Example Precision/Recall: Person

- All methods



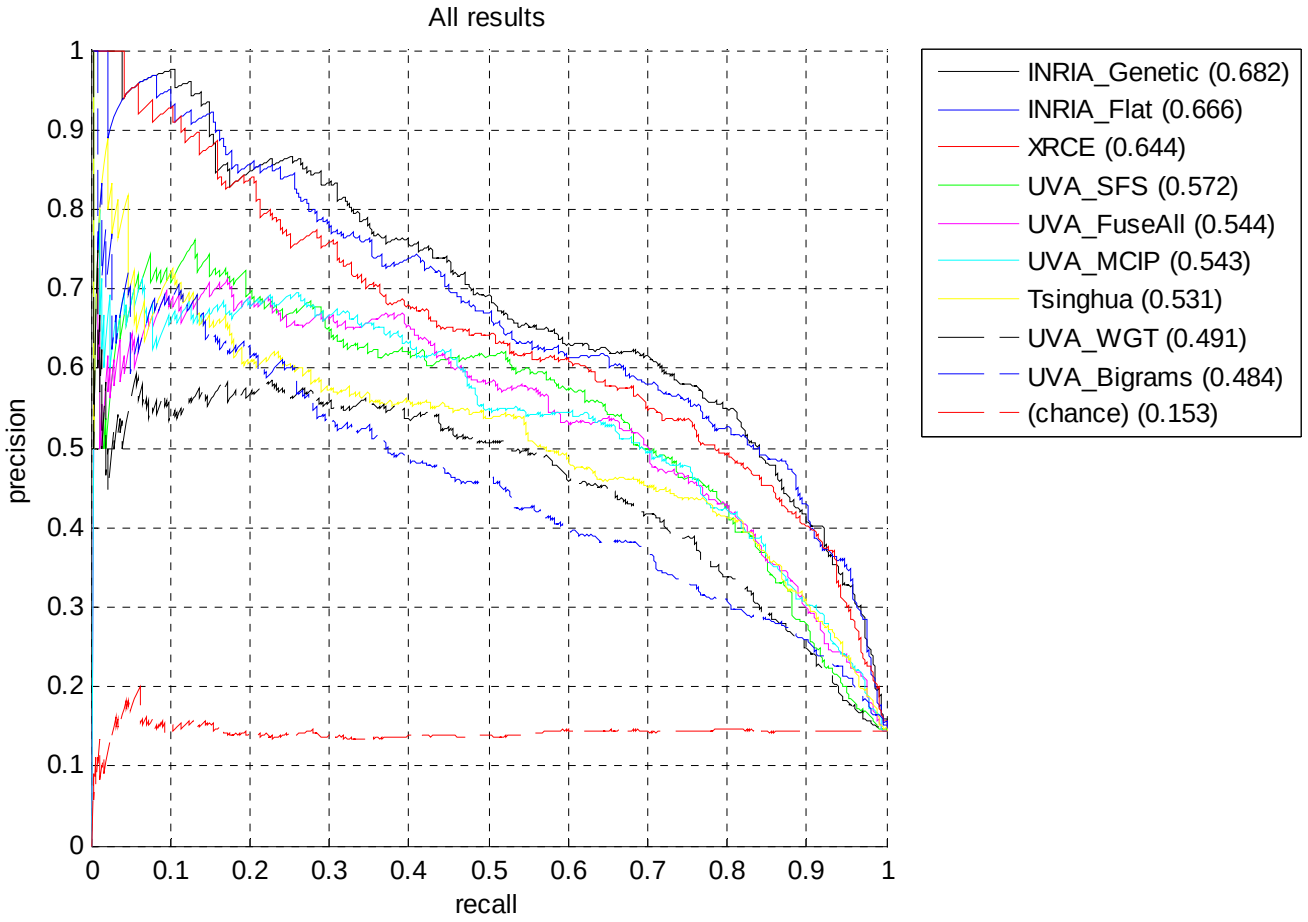
Example Precision/Recall: Person

- Top 5



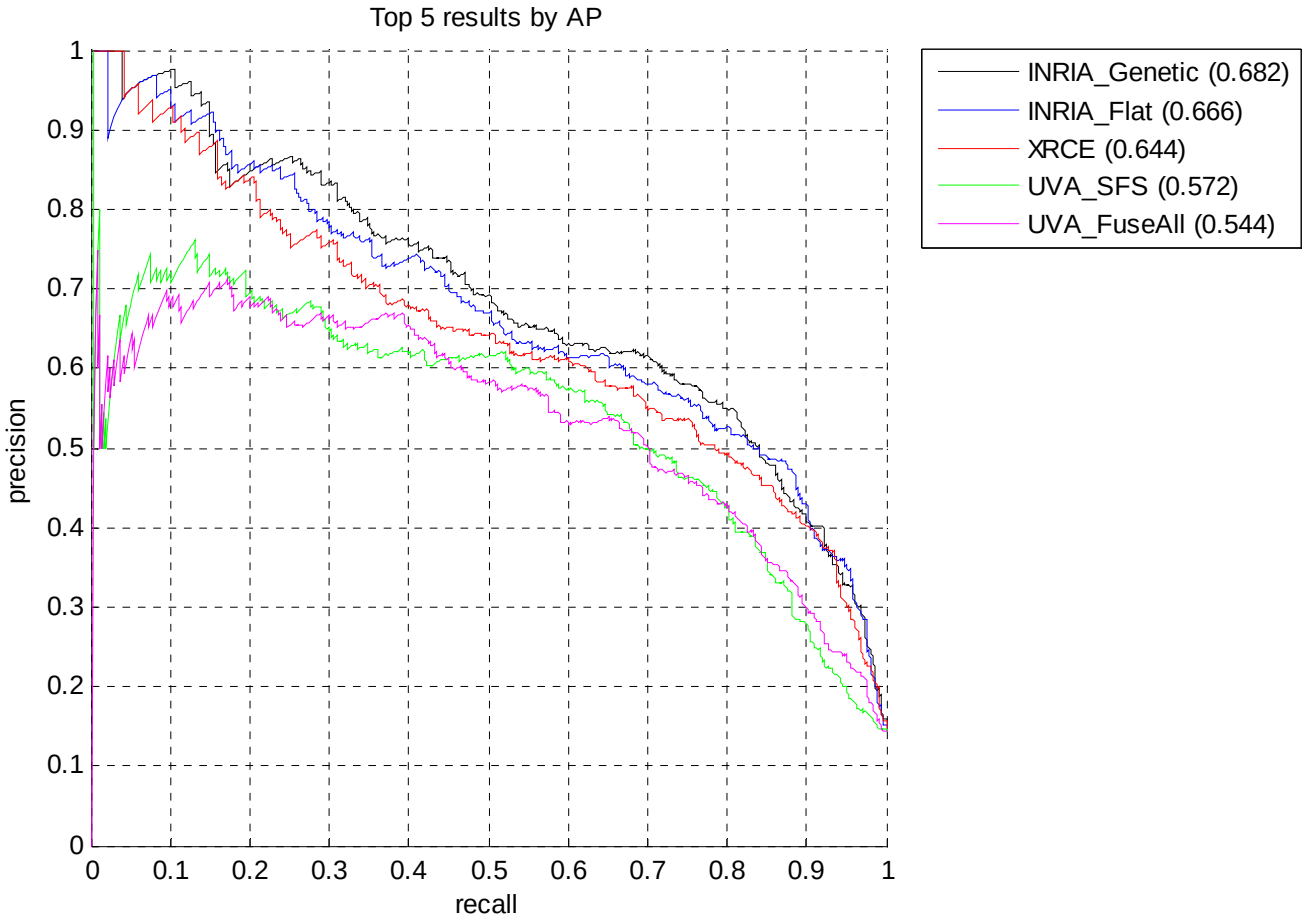
Example Precision/Recall: Cat

- All methods



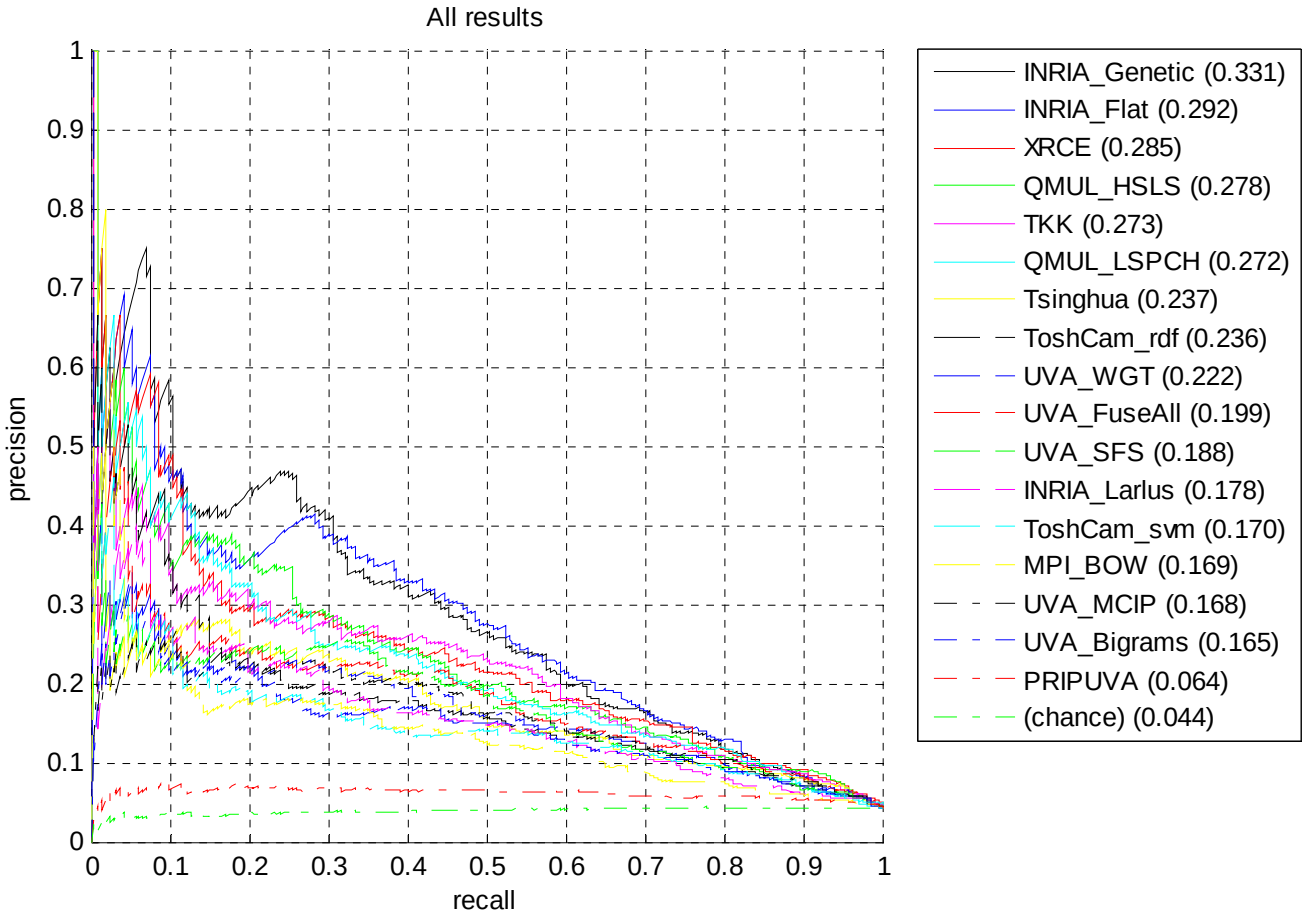
Example Precision/Recall: Cat

- Top 5



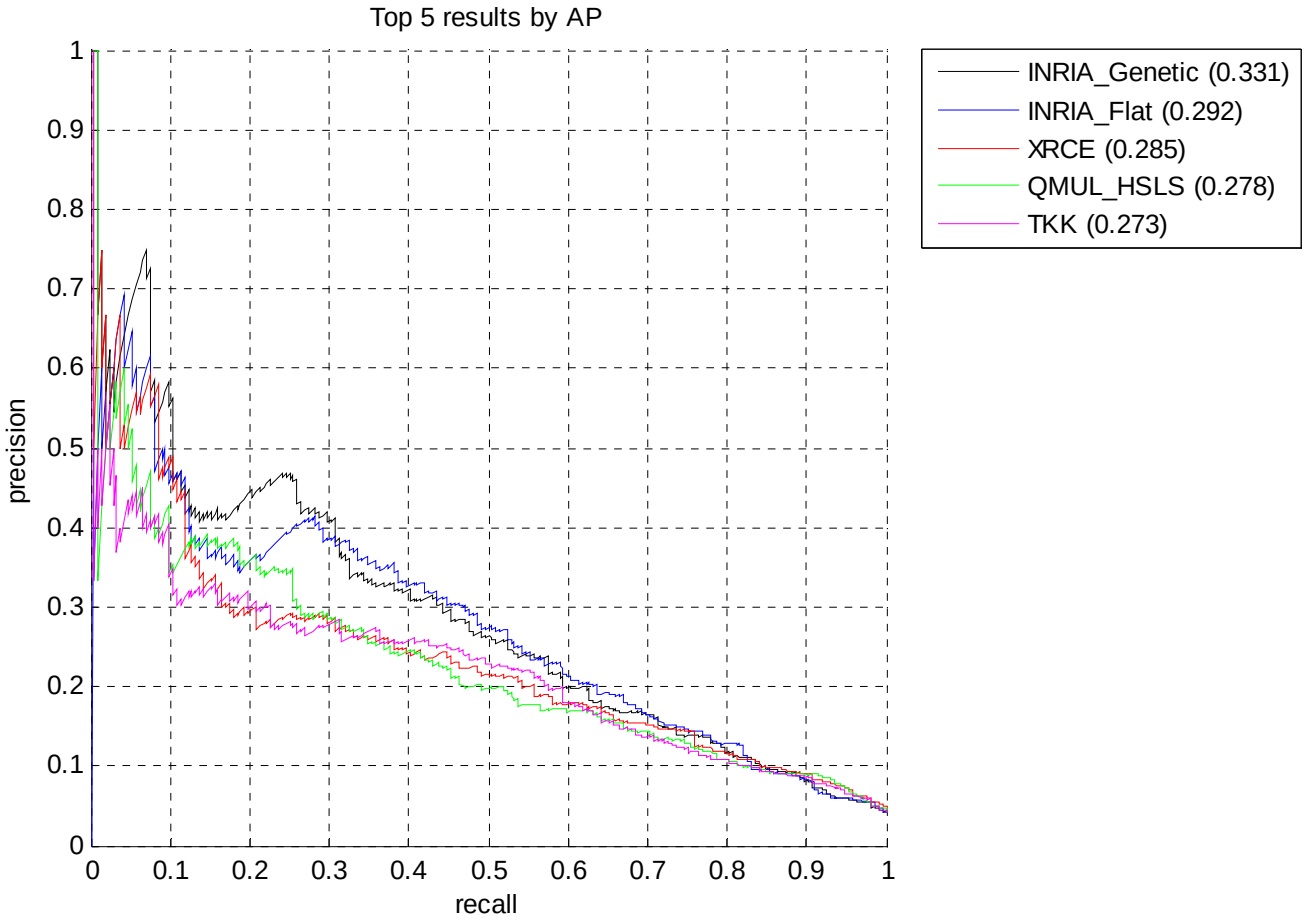
Example Precision/Recall: Bottle

- All methods

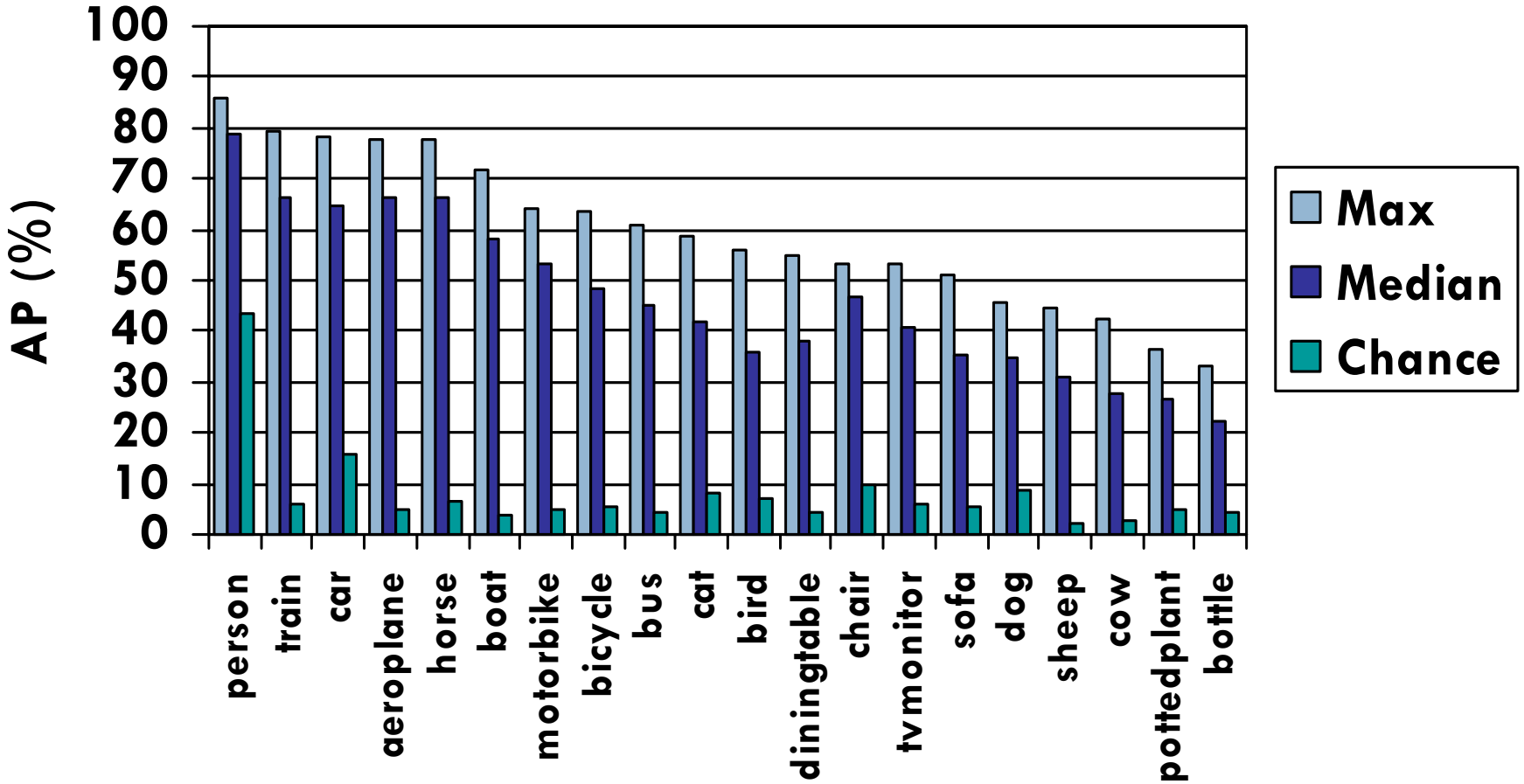


Example Precision/Recall: Bottle

- Top 5



AP by Class



- Good results on "person" due to prior?
- Classes indistinguishable by context prove difficult?

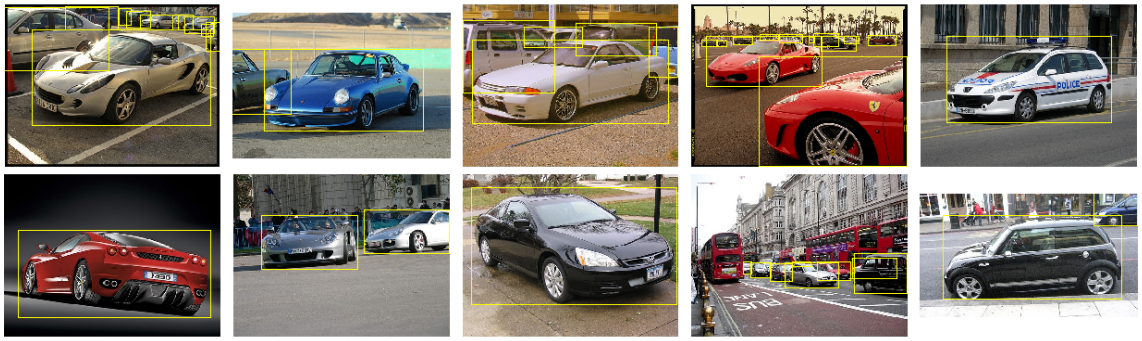
Statistical Significance

- Friedman/Nemenyi analysis of ranks

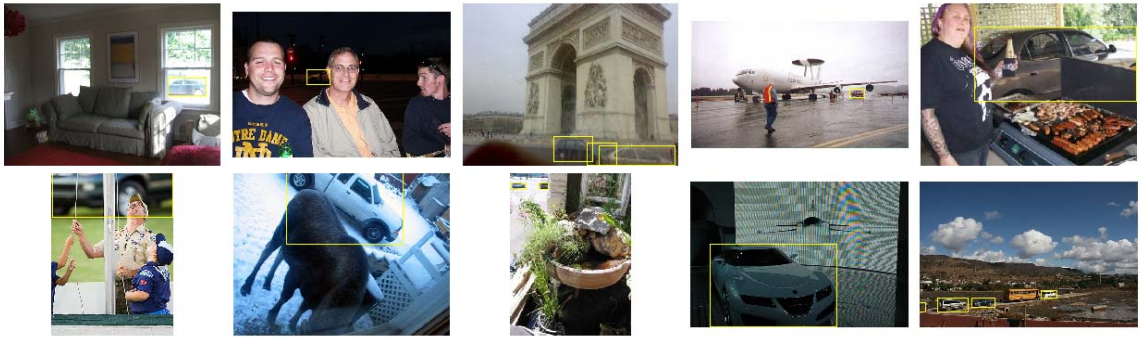
	INRIA_Genetic	INRIA_Flat	XRCE	TKK	QMUL_LSPCH	QMUL_HSLS	UVA_FuseAll	UVA_SFS	UVA_MCIP	INRIA_Larlus	Tsinghua	MPI_BOW	ToshCam_rdf	UVA_WGT	ToshCam_svm	UVA_Bigrams	PRIPUVA
INRIA_Genetic	-	1.2	2.0	4.4	4.5	5.0	7.1	7.2	8.1	8.7	10.0	10.2	11.6	12.3	13.4	13.9	15.9
INRIA_Flat	-1.2	-	0.8	3.3	3.3	3.9	6.0	6.1	6.9	7.6	8.8	9.0	10.5	11.2	12.2	12.8	14.8
XRCE	-2.0	-0.8	-	2.5	2.5	3.1	5.2	5.3	6.1	6.8	8.0	8.2	9.7	10.4	11.4	12.0	14.0
TKK	-4.4	-3.3	-2.5	-	0.1	0.6	2.7	2.8	3.7	4.3	5.6	5.8	7.2	7.9	9.0	9.5	11.5
QMUL_LSPCH	-4.5	-3.3	-2.5	-0.1	-	0.6	2.7	2.8	3.6	4.3	5.5	5.7	7.2	7.9	8.9	9.5	11.5
QMUL_HSLS	-5.0	-3.9	-3.1	-0.6	-0.6	-	2.1	2.2	3.1	3.7	5.0	5.2	6.6	7.3	8.4	8.9	10.9
UVA_FuseAll	-7.1	-6.0	-5.2	-2.7	-2.7	-2.1	-	0.1	1.0	1.6	2.9	3.1	4.5	5.2	6.3	6.8	8.8
UVA_SFS	-7.2	-6.1	-5.3	-2.8	-2.8	-2.2	-0.1	-	0.9	1.5	2.8	3.0	4.4	5.1	6.2	6.7	8.7
UVA_MCIP	-8.1	-6.9	-6.1	-3.7	-3.6	-3.1	-1.0	-0.9	-	0.7	1.9	2.1	3.6	4.3	5.3	5.9	7.9
INRIA_Larlus	-8.7	-7.6	-6.8	-4.3	-4.3	-3.7	-1.6	-1.5	-0.7	-	1.3	1.5	2.9	3.6	4.7	5.2	7.2
Tsinghua	-10.0	-8.8	-8.0	-5.6	-5.5	-5.0	-2.9	-2.8	-1.9	-1.3	-	0.2	1.7	2.4	3.4	4.0	6.0
MPI_BOW	-10.2	-9.0	-8.2	-5.8	-5.7	-5.2	-3.1	-3.0	-2.1	-1.5	-0.2	-	1.5	2.2	3.2	3.8	5.8
ToshCam_rdf	-11.6	-10.5	-9.7	-7.2	-7.2	-6.6	-4.5	-4.4	-3.6	-2.9	-1.7	-1.5	-	0.7	1.8	2.3	4.3
UVA_WGT	-12.3	-11.2	-10.4	-7.9	-7.9	-7.3	-5.2	-5.1	-4.3	-3.6	-2.4	-2.2	-0.7	-	1.1	1.6	3.6
ToshCam_svm	-13.4	-12.2	-11.4	-9.0	-8.9	-8.4	-6.3	-6.2	-5.3	-4.7	-3.4	-3.2	-1.8	-1.1	-	0.6	2.6
UVA_Bigrams	-13.9	-12.8	-12.0	-9.5	-9.5	-8.9	-6.8	-6.7	-5.9	-5.2	-4.0	-3.8	-2.3	-1.6	-0.6	-	2.0
PRIPUVA	-15.9	-14.8	-14.0	-11.5	-11.5	-10.9	-8.8	-8.7	-7.9	-7.2	-6.0	-5.8	-4.3	-3.6	-2.6	-2.0	-

Ranked Images: Car

- Class images:
  Highest ranked



- Class images:
  Lowest ranked



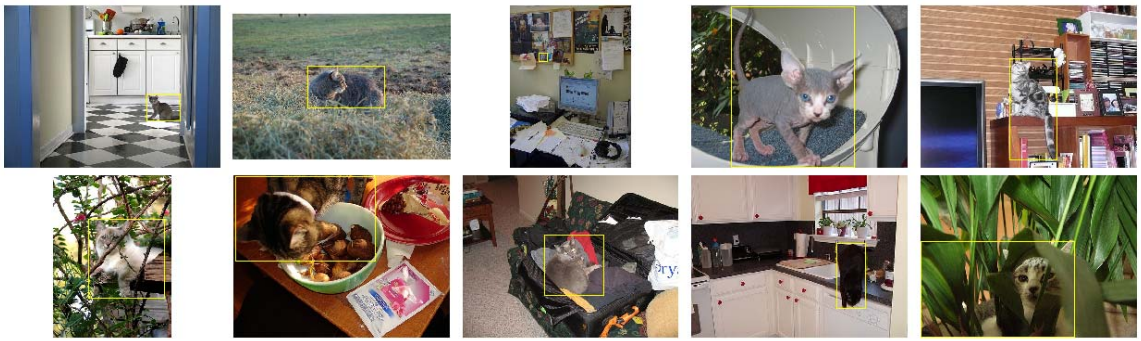
- Non-class images:
  Highest ranked

- Intuitive Confusion?

Ranked Images: Cat

- Class images:
  Highest ranked

- Class images:
  Lowest ranked

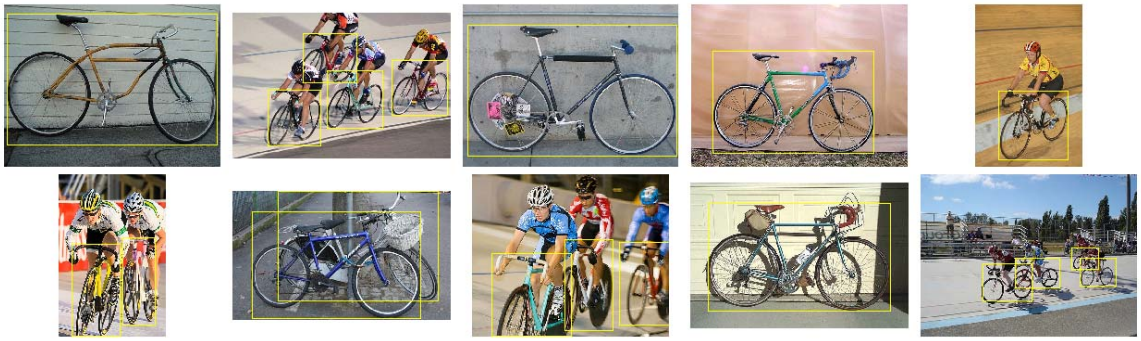


- Non-class images:
  Highest ranked

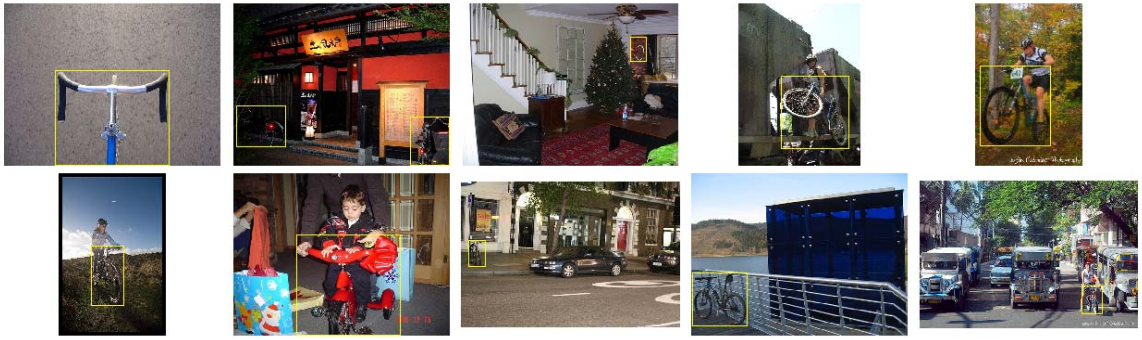
- Composition?

Ranked Images: Bicycle

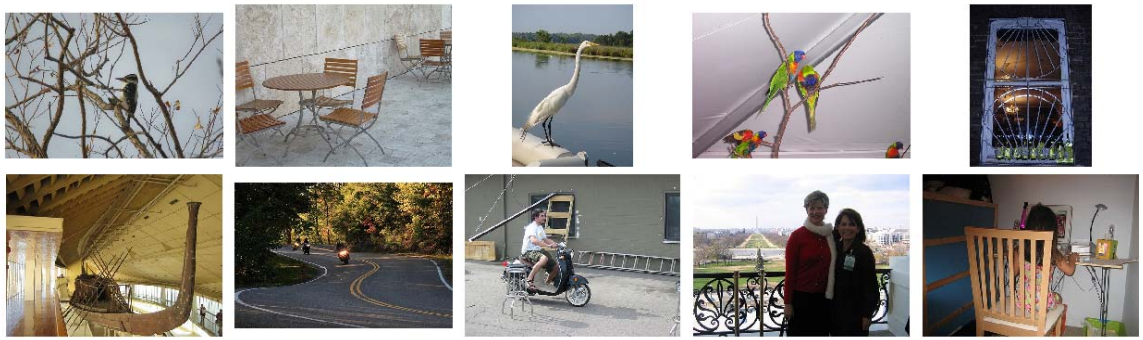
- Class images:
  Highest ranked



- Class images:
  Lowest ranked



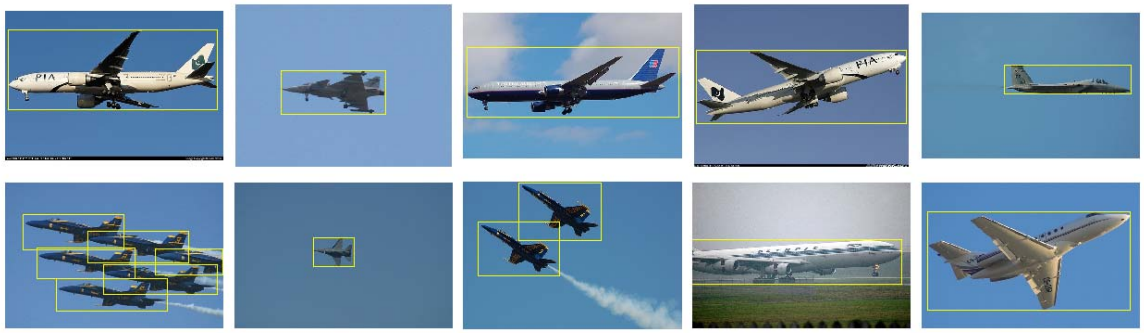
- Non-class images:
  Highest ranked



- "Structured" Texture?

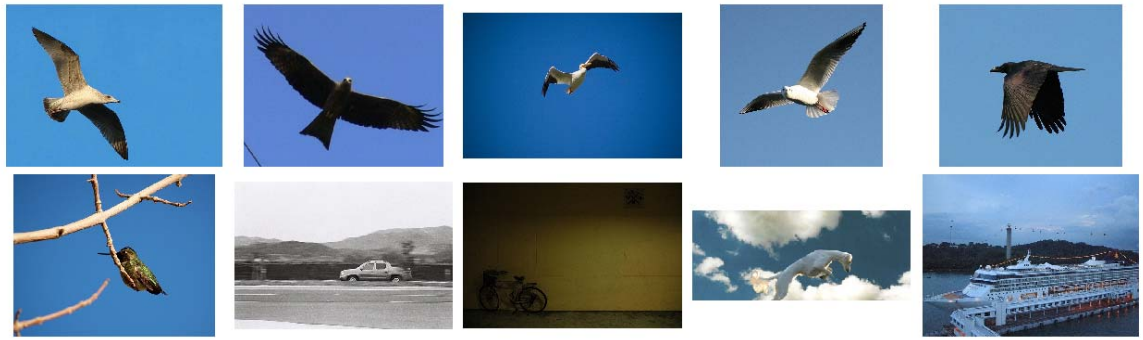
Ranked Images: Aeroplane

- Class images:
  Highest ranked



- Class images:
  Lowest ranked

- Non-class images:
  Highest ranked

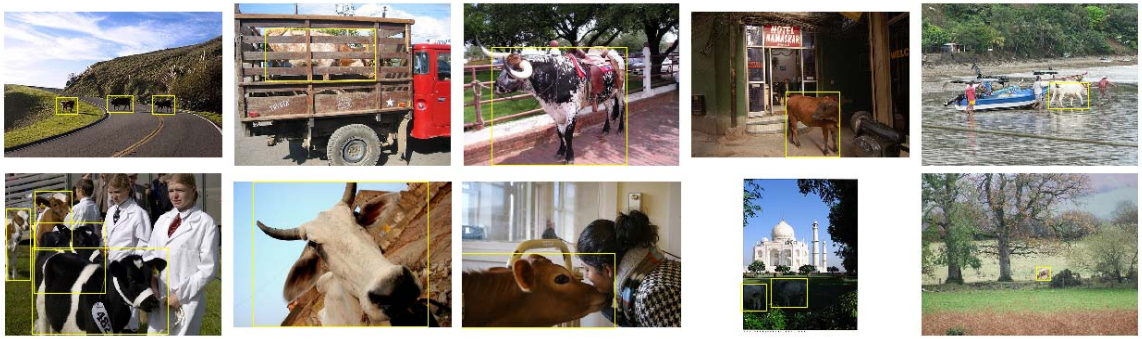


- Context?

Ranked Images: Cow

- Class images:
  Highest ranked

- Class images:
  Lowest ranked

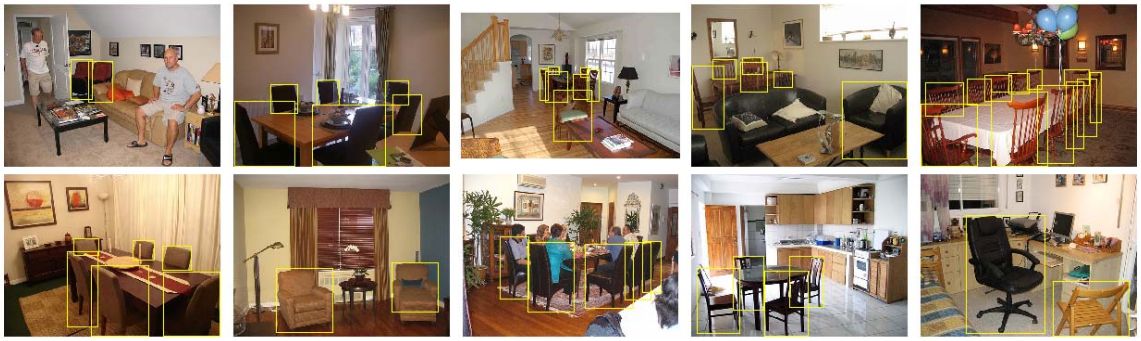


- Non-class images:
  Highest ranked

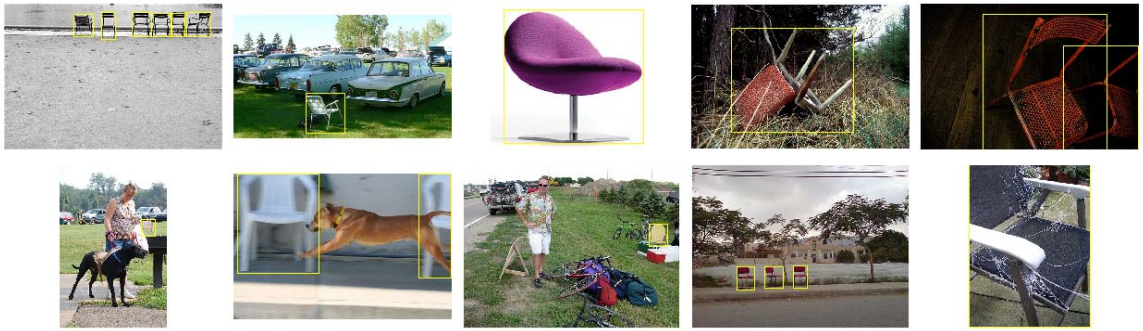
- Context?

Ranked Images: Chair

- Class images:
  Highest ranked



- Class images:
  Lowest ranked

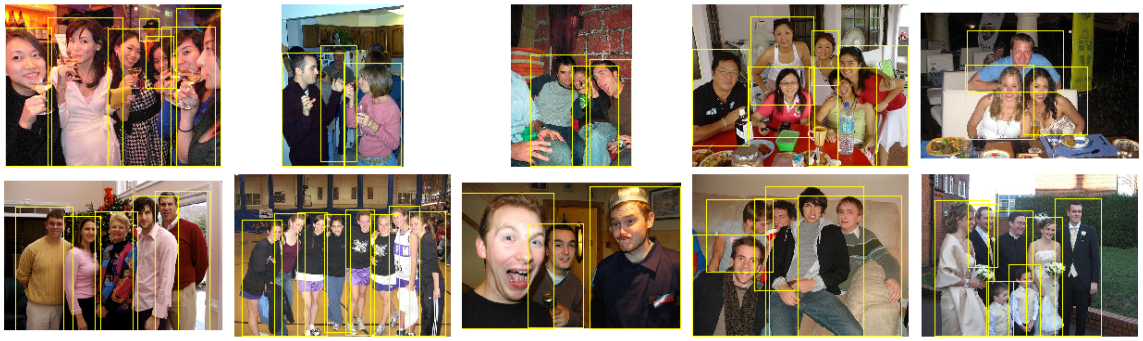


- Non-class images:
  Highest ranked

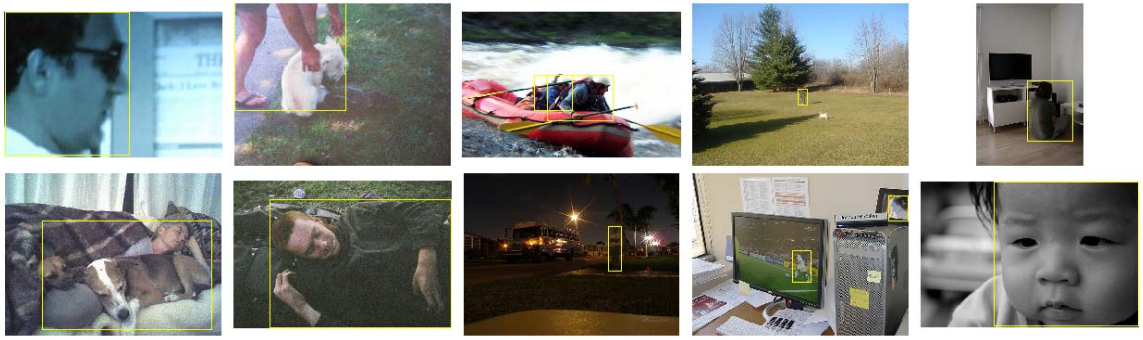
- Indoor scenes?

Ranked Images: Person

- Class images:
  Highest ranked



- Class images:
  Lowest ranked

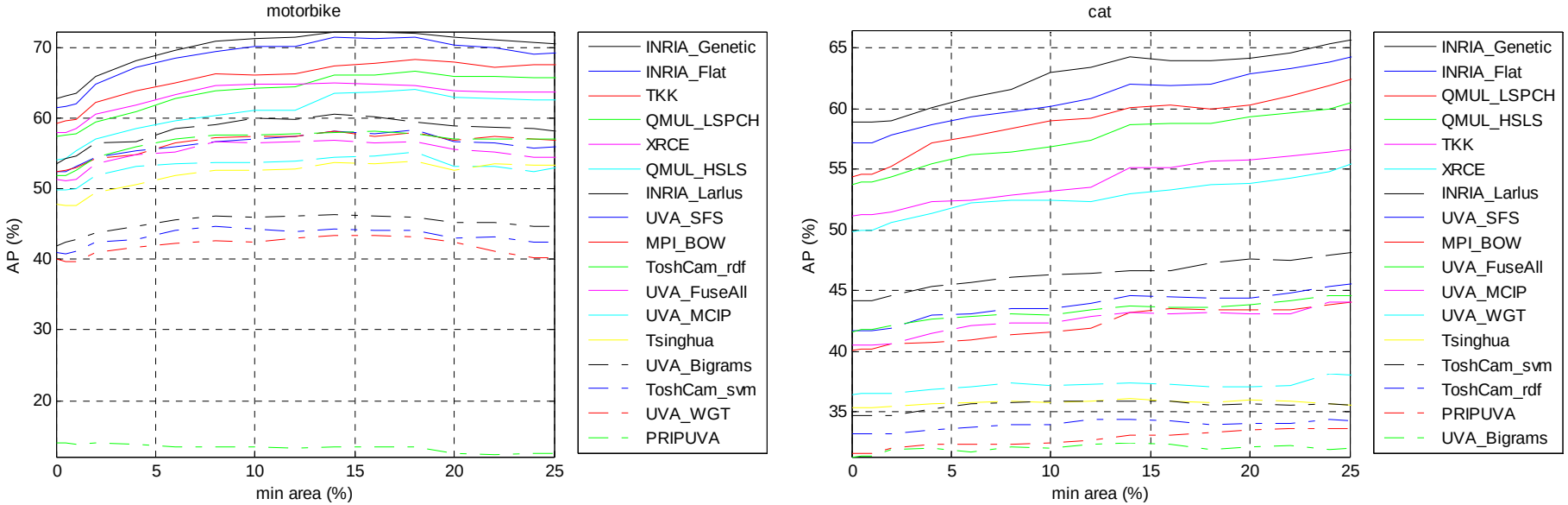


- Non-class images:
  Highest ranked

- People on motorbikes?

AP vs. Object Class Area

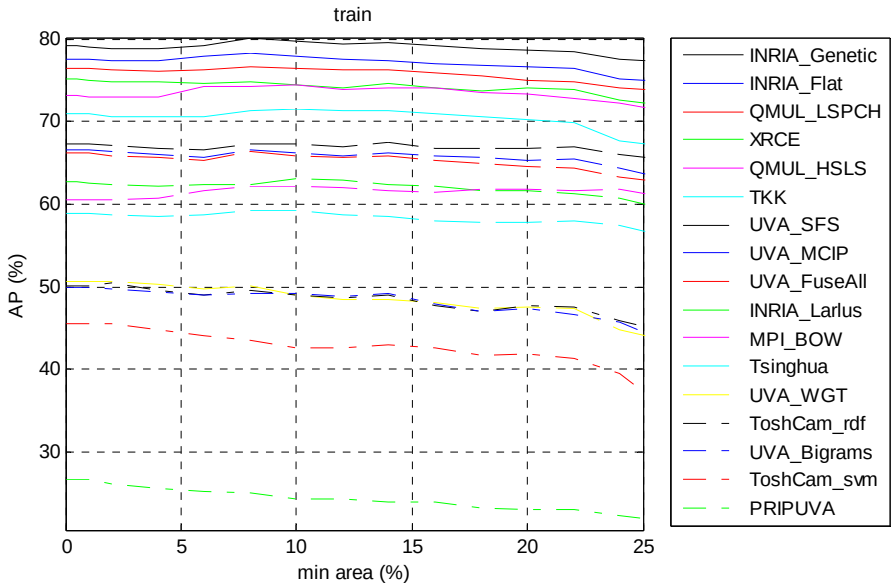
- Do these methods have a bias toward larger objects?

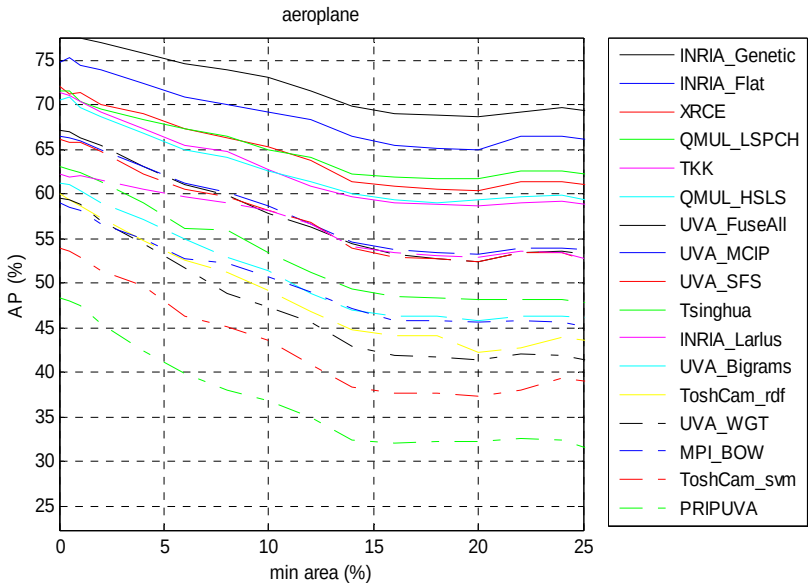


- Moderate evidence for several classes – car, cat, motorbike
- Performance drops off due to increasing chance of occlusion?

AP vs. Object Class Area

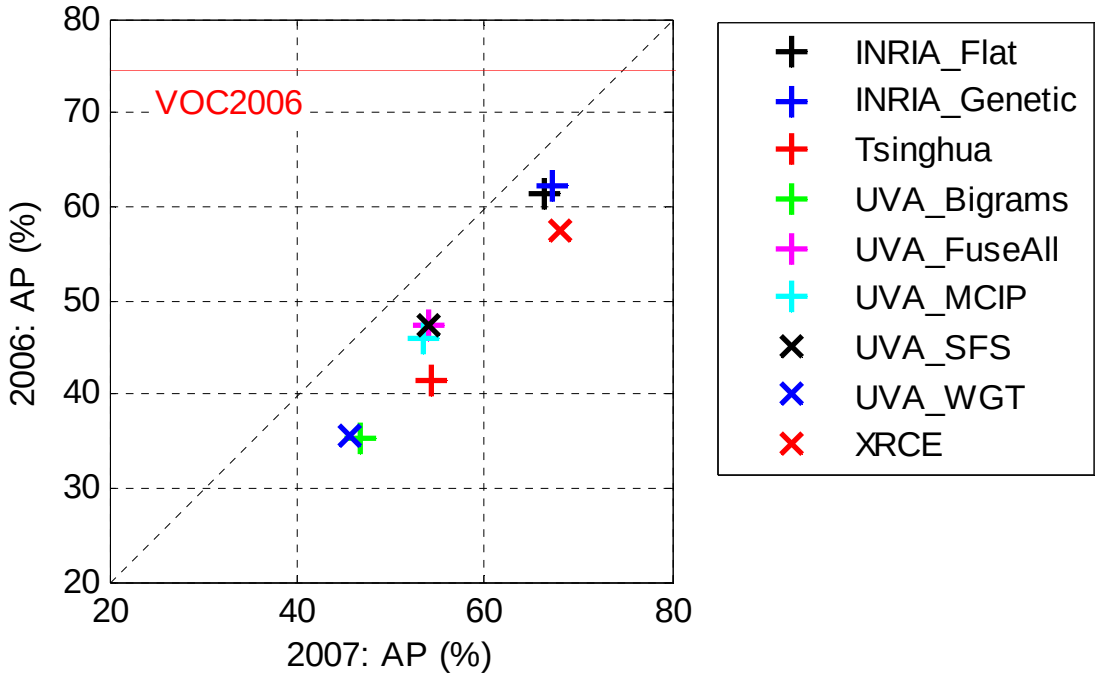
- For most classes, correlation with object class area is zero or negative





- Methods are learning more about context/scene appearance than object appearance?

VOC2006 vs. VOC2007 Test Data



- High correlation between results on 2007 and 2006 test data
- Some evidence of "over-fitting" – no method equalled results when trained on 2006 data