The PASCAL Visual Object Classes Challenge 2011 (VOC2011)

Part 1 – Challenge & Classification Task

Mark Everingham
Luc Van Gool
Chris Williams
John Winn
Andrew Zisserman

PASCAL2
Pattern Analysis, Statistical Modelling and Computational Learning

Dataset Collection

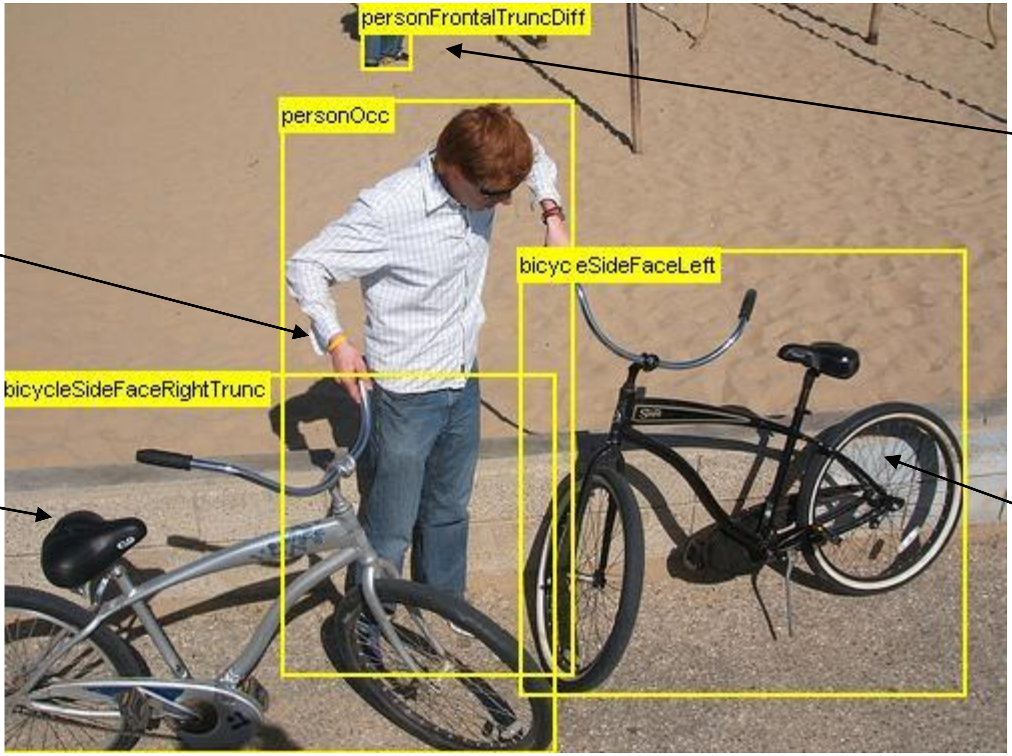
- Images downloaded from **flickr**
  - 500,000 images downloaded and random subset selected for annotation
  - Queries
    - Keyword e.g. "car", "vehicle", "street", "downtown"
    - Date of capture e.g. "taken 21-July"
      - Removes "recency" bias in flickr results
    - Images selected from random page of results
      - Reduces bias toward particular flickr users
- 2008-2010 datasets retained as subset of 2011
  - Assignments to training/test sets maintained

Annotation

- Complete annotation of all objects from 20 categories

Occluded
Object is significantly occluded within BB

Truncated
Object extends beyond BB



Difficult
Not scored in evaluation

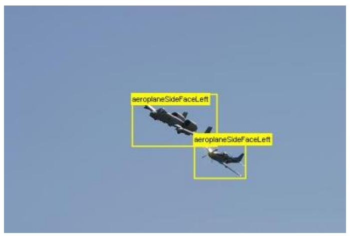
Pose
Facing left

Annotation Procedure

1. Amazon Mechanical Turk
   - Qualification task
   - Images labelled with presence/absence of object categories
   - Bounding boxes labelled for subsets of object categories e.g. bicycle/bus/car/motorbike
2. Experienced Annotators
   - Web-based tool, written guidelines
   - Annotation corrected and refined
   - Annotation checked by second annotator

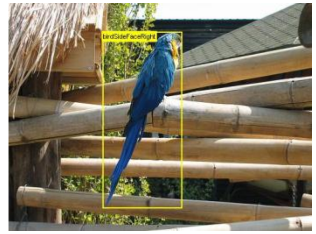
Examples

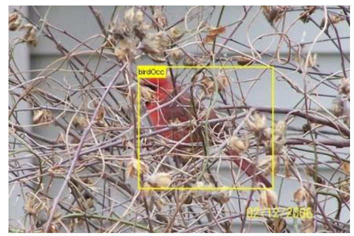
Aeroplane



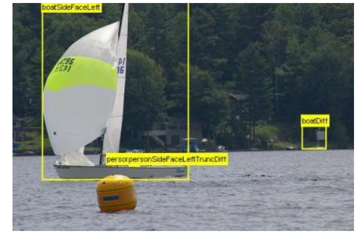
Bicycle

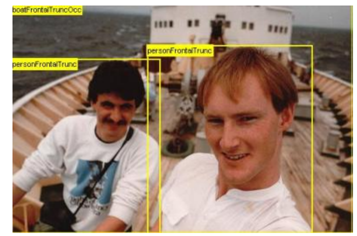
Bird



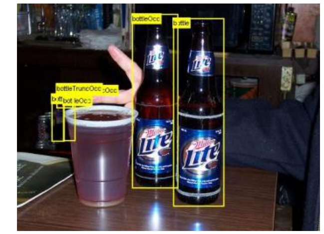


Boat



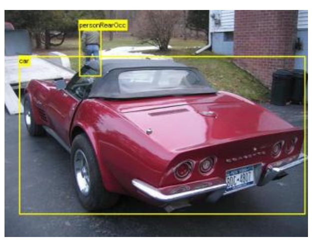


Bottle



Bus

Car



Cat

Chair

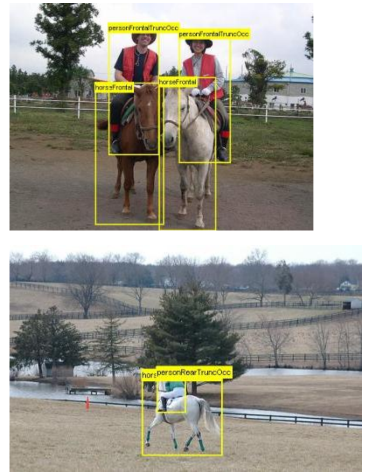
Cow

Examples

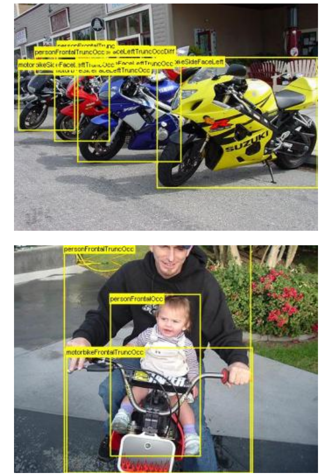
Dining Table

Dog

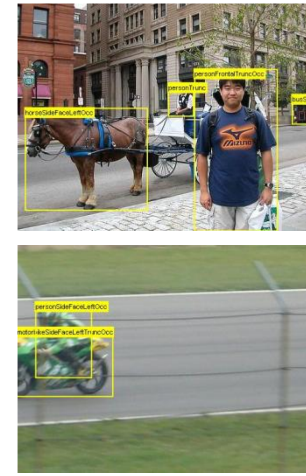
Horse



Motorbike

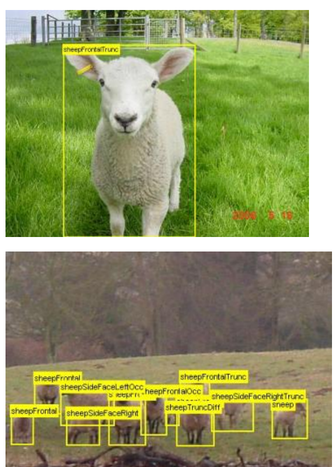


Person

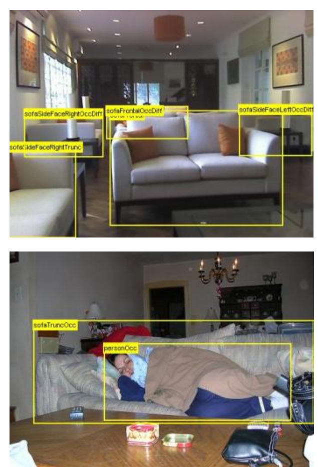


Potted Plant

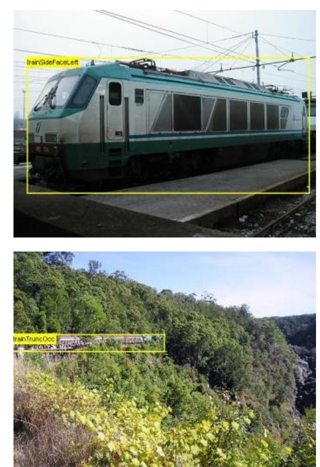
Sheep



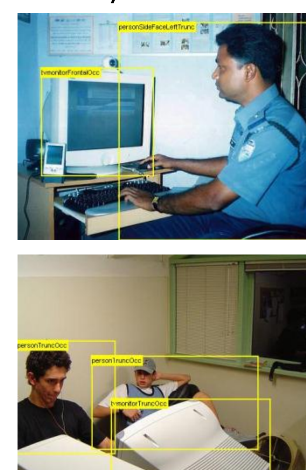
Sofa



Train



TV/Monitor



Dataset Statistics

- Around 15% increase in size over VOC2010

	Training		Testing	
Images	11,540	(10,103)	10,994	(9,637)
Objects	27,450	(23,374)	27,078	(22,992)

VOC2010 counts shown in brackets

- Minimum ~600 training objects per category
  - ~2,000 cars, 1,500 dogs, 8,500 people
- Approximately equal distribution across training and test sets

Best Practice

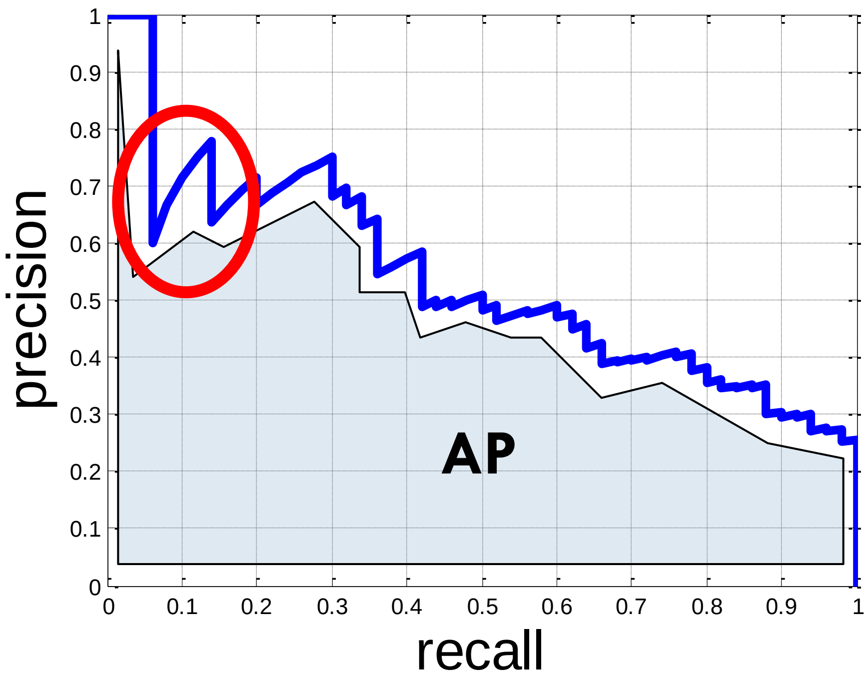
- If using the provided training data ("trainval"), **<u>all</u>** feature selection, parameter tuning, choice of classifier architecture, etc. should be done using the training data alone
  - Use suggested training/validation split
  - Use cross-validation
- Do report results on the most recent dataset **<u>(2010)</u>**
- Results on the test set should be generated **<u>infrequently</u>** to avoid optimization on test data
  - To compare features etc. use either cross-validation or the VOC2007 dataset (test annotation available)

- Do cite us please! PASCAL VOC costs money and time…

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?

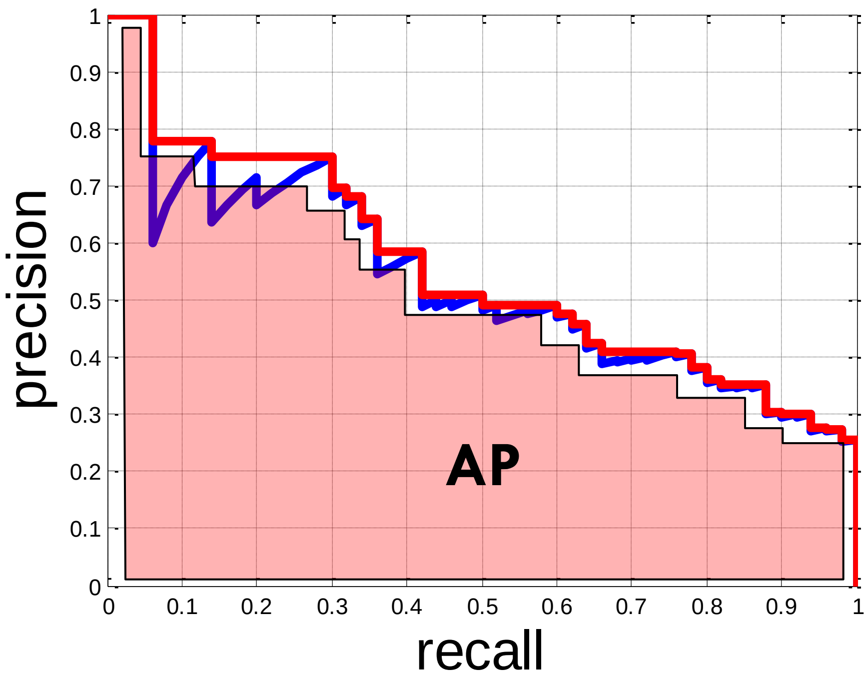
Average Precision



- Average Precision (AP) measures area under precision/recall curve
- Application independent
- A good score requires both high recall and high precision

- "Sawtooth" shape is irrelevant: can obtain both higher recall **and** precision by changing threshold

Average Precision: VOC2010-2011



- Interpolate curve to create version for which the precision is monotonically non-increasing
- Measure area under interpolated curve

- Sawtooth shape is ignored
- Area is measured with maximum accuracy

Methods

- 19 Methods, 11 Groups
  - VOC2010: 33 "Methods", 22 Groups
- Basic recipe
  - Bag of visual words and/or spatial pyramid
  - Multiple features: interest points/dense/saliency, SIFT, HOG, color SIFT, LBP, gist, etc.
  - Vector quantization, histogram representation
  - Linear/non-linear/Fisher kernels
  - SVM classifier
  - Feature/classifier combination by MKL or voting

Methods

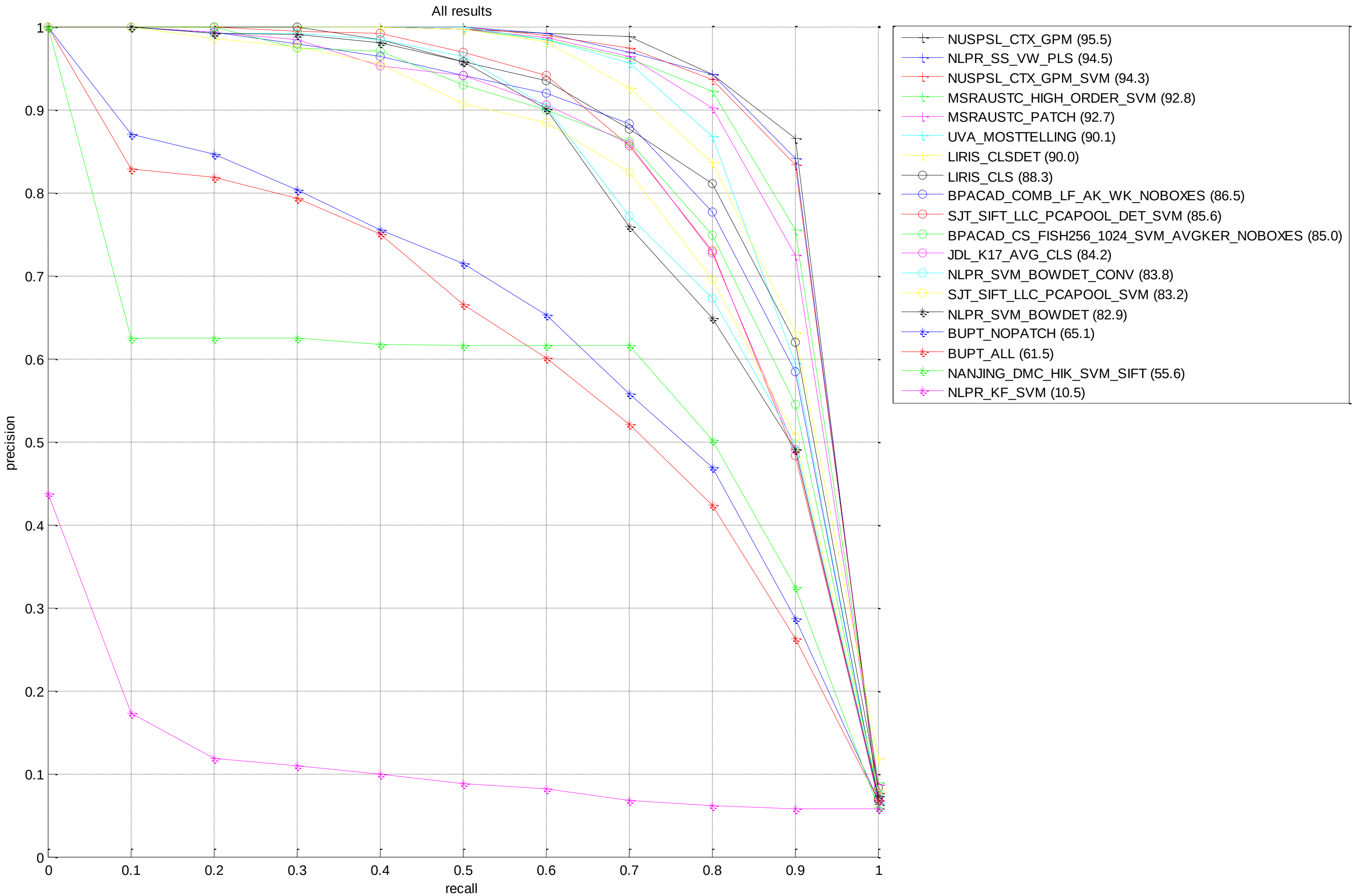
- Additional ingredients
  - Inclusion of detection scores (from latent-SVM model)
  - Partial least squares dimensionality reduction
  - Sparse coding, max pooling
  - Context-aware features
  - Segmentation as selective search
  - Text features (from nearest neighbour images)

AP by Class/Method

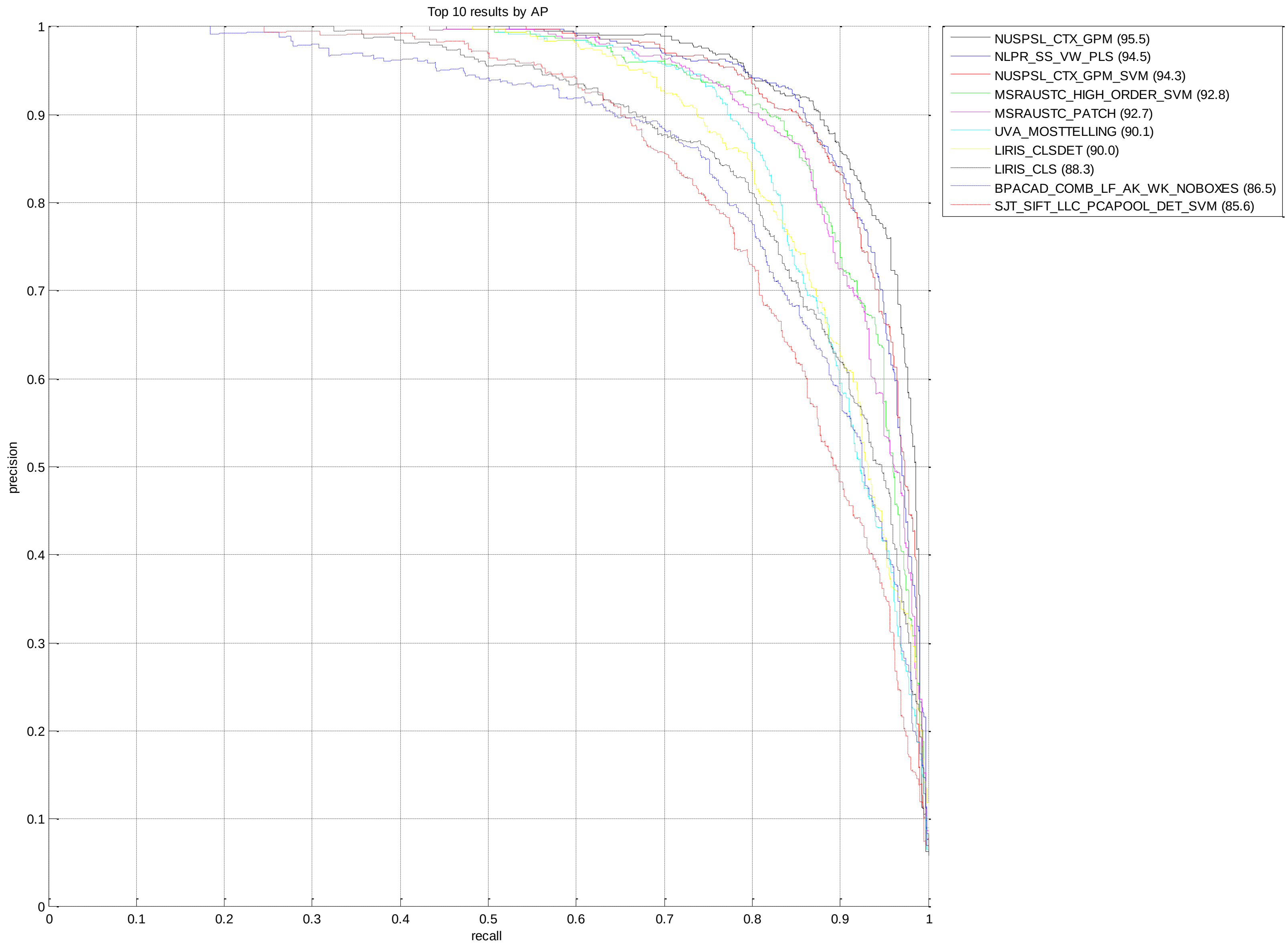
(1st, 2nd, 3rd place)

	aero plane	bicycle	bird	boat	bottle	bus	car	cat	chair	cow	dining table	dog	horse	motor bike	person	potted plant	sheep	sofa	train	tv/ monitor
BPACAD_COMB_LF_AK_WK...	86.5	58.3	59.7	67.4	33.2	74.2	64.0	65.5	58.5	44.8	53.5	57.0	60.7	70.8	84.6	39.4	55.4	50.5	80.7	63.1
BPACAD_CS_FISH256_1024...	85.0	57.0	57.7	65.9	30.7	75.0	62.4	64.4	56.9	42.2	50.9	55.3	59.1	69.1	84.2	39.3	52.3	46.7	78.9	61.8
BUPT_ALL	61.5	11.9	12.4	29.7	8.7	30.6	18.4	23.6	21.6	5.8	14.8	18.5	7.1	12.3	47.7	7.2	15.0	9.8	18.8	19.2
BUPT_NOPATCH	65.1	23.8	17.3	36.0	12.6	40.5	31.1	35.4	27.2	10.4	20.8	31.3	13.6	29.5	54.9	10.7	19.1	19.2	42.1	30.8
JDL_K17_AVG_CLS	84.2	52.0	54.5	63.2	25.3	71.2	58.0	61.1	50.2	33.3	44.3	49.7	57.9	65.1	79.9	20.9	47.4	43.0	77.7	56.7
LIRIS_CLS	88.3	56.2	59.3	68.6	33.2	76.6	62.2	64.5	55.3	42.6	55.1	56.2	61.9	70.0	82.5	37.3	56.4	48.3	79.6	64.7
LIRIS_CLSDET	90.0	66.2	63.3	70.9	47.0	80.9	73.9	63.9	61.1	52.7	57.9	56.9	69.6	73.8	88.4	46.3	65.3	54.2	81.3	72.7
MSRAUSTC_HIGH_ORDER_SVM	92.8	74.8	69.6	76.1	47.3	83.5	76.4	76.9	59.8	54.5	63.5	67.0	75.1	78.8	90.4	43.1	63.1	60.4	85.6	71.1
MSRAUSTC_PATCH	92.7	74.5	69.4	75.4	45.7	83.4	76.5	76.6	59.6	54.5	63.4	67.4	74.8	78.6	90.3	43.0	63.1	58.6	85.2	71.3
NANJING_DMC_HIK_SVM_SIFT	55.6	25.5	31.0	36.5	15.8	41.4	40.0	40.6	30.0	17.8	21.1	34.0	27.0	31.0	57.9	11.9	20.7	22.6	48.4	35.7
NLPR_KF_SVM	10.5	9.1	10.7	6.0	6.5	7.2	13.3	12.2	11.5	9.5	5.6	16.7	8.6	6.6	38.9	5.3	15.0	5.0	8.3	5.4
NLPR_SS_VW_PLS	94.5	82.6	79.4	80.7	57.8	87.8	85.5	83.9	66.6	74.2	69.4	75.2	83.0	88.1	93.5	56.2	75.5	64.1	90.0	76.6
NLPR_SVM_BOWDET	82.9	69.4	45.4	60.1	46.0	80.0	75.1	59.9	54.9	50.7	43.3	49.9	63.4	72.2	88.1	36.1	57.1	37.7	75.2	58.5
NLPR_SVM_BOWDET_CONV	83.8	69.8	47.8	60.5	45.4	80.5	74.6	60.4	54.0	51.3	45.3	51.5	64.5	72.6	87.7	35.9	57.7	39.8	75.8	62.7
NUSPSL_CTX_GPM	95.5	81.1	79.4	82.5	58.2	87.7	84.1	83.1	68.5	72.8	68.5	76.4	83.3	87.5	92.8	56.5	77.7	67.0	91.2	77.5
NUSPSL_CTX_GPM_SVM	94.3	78.5	76.4	80.0	57.0	86.3	82.1	81.5	65.6	74.7	66.5	73.4	81.9	85.3	91.9	53.2	73.9	65.1	89.5	76.0
SJT_SIFT_LLC_PCAPOOL_DET_SVM	85.6	66.5	51.9	60.3	45.4	76.8	70.3	65.1	56.4	34.3	49.6	52.4	63.1	71.5	86.8	26.1	56.9	47.9	75.5	65.6
SJT_SIFT_LLC_PCAPOOL_SVM	83.2	52.5	49.3	59.6	26.0	73.5	58.2	64.4	52.1	36.6	44.9	52.1	57.8	63.8	78.1	19.1	52.8	44.1	72.0	57.4
UVA_MOSTTELLING	90.1	74.1	66.5	76.0	57.0	85.6	81.2	74.5	63.5	62.7	64.5	66.6	76.5	81.2	90.8	58.7	69.3	66.3	84.7	77.2

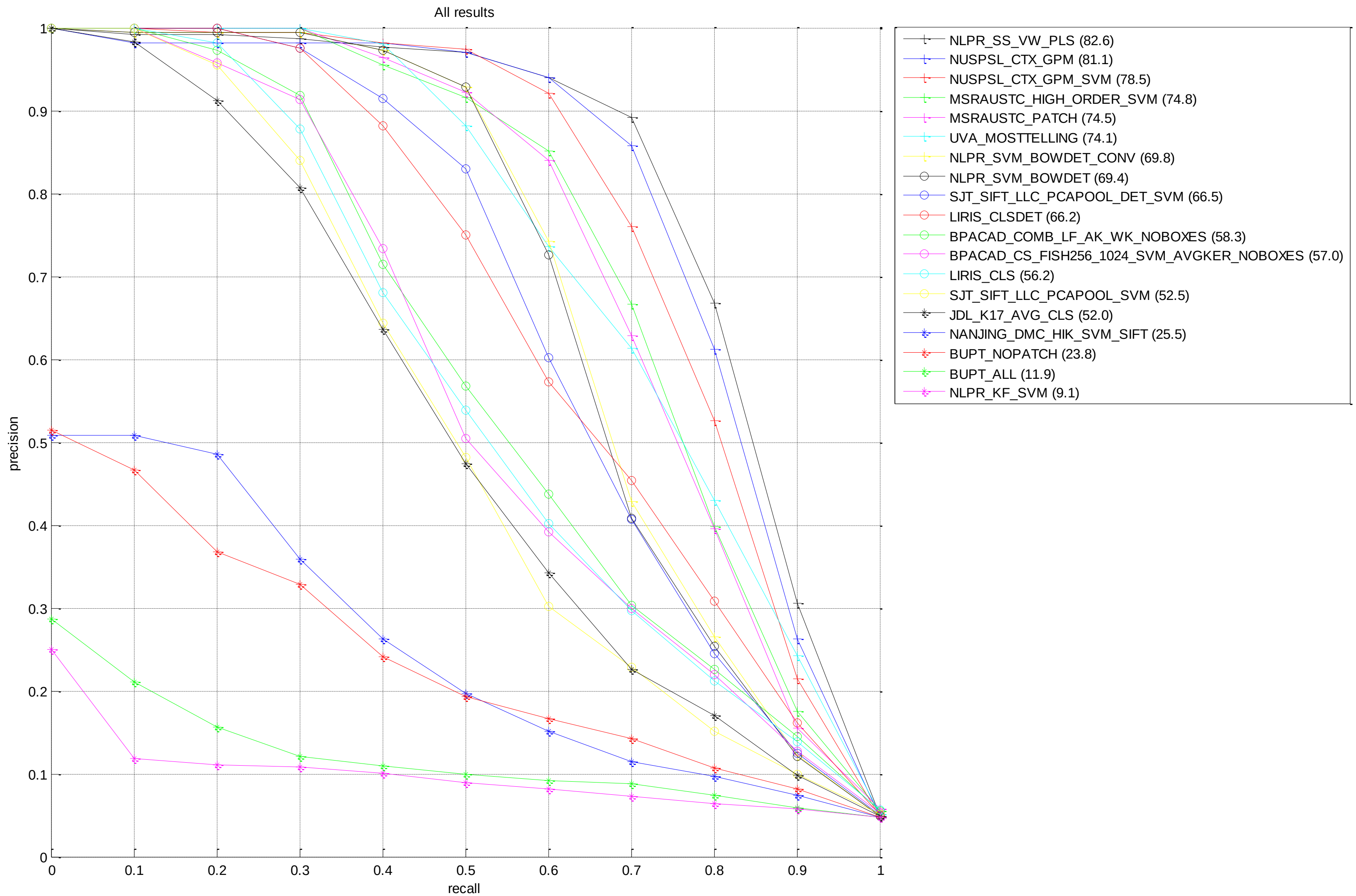
Precision/Recall: Aeroplane (All)



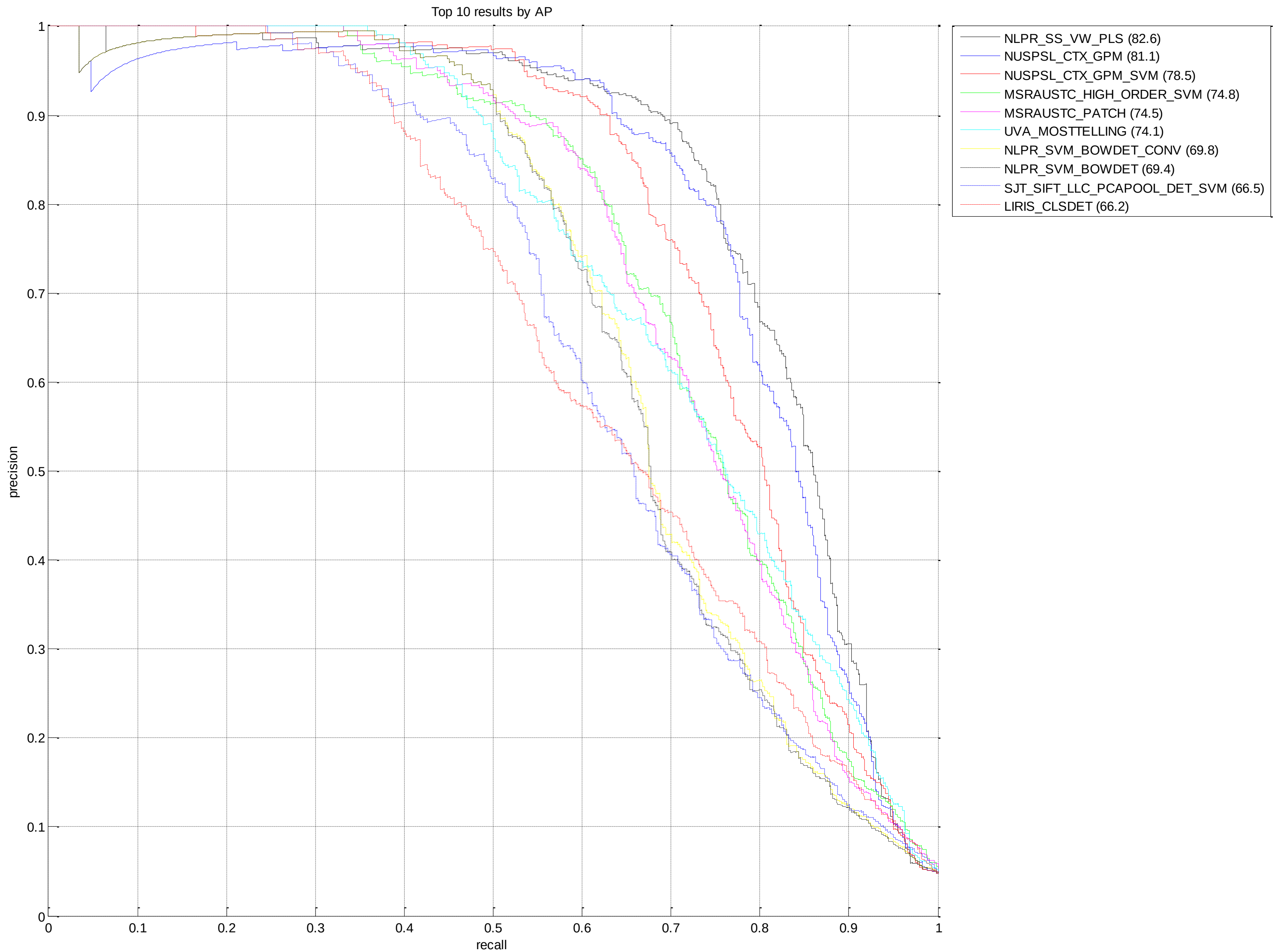
Precision/Recall: Aeroplane (Top 10 by AP)



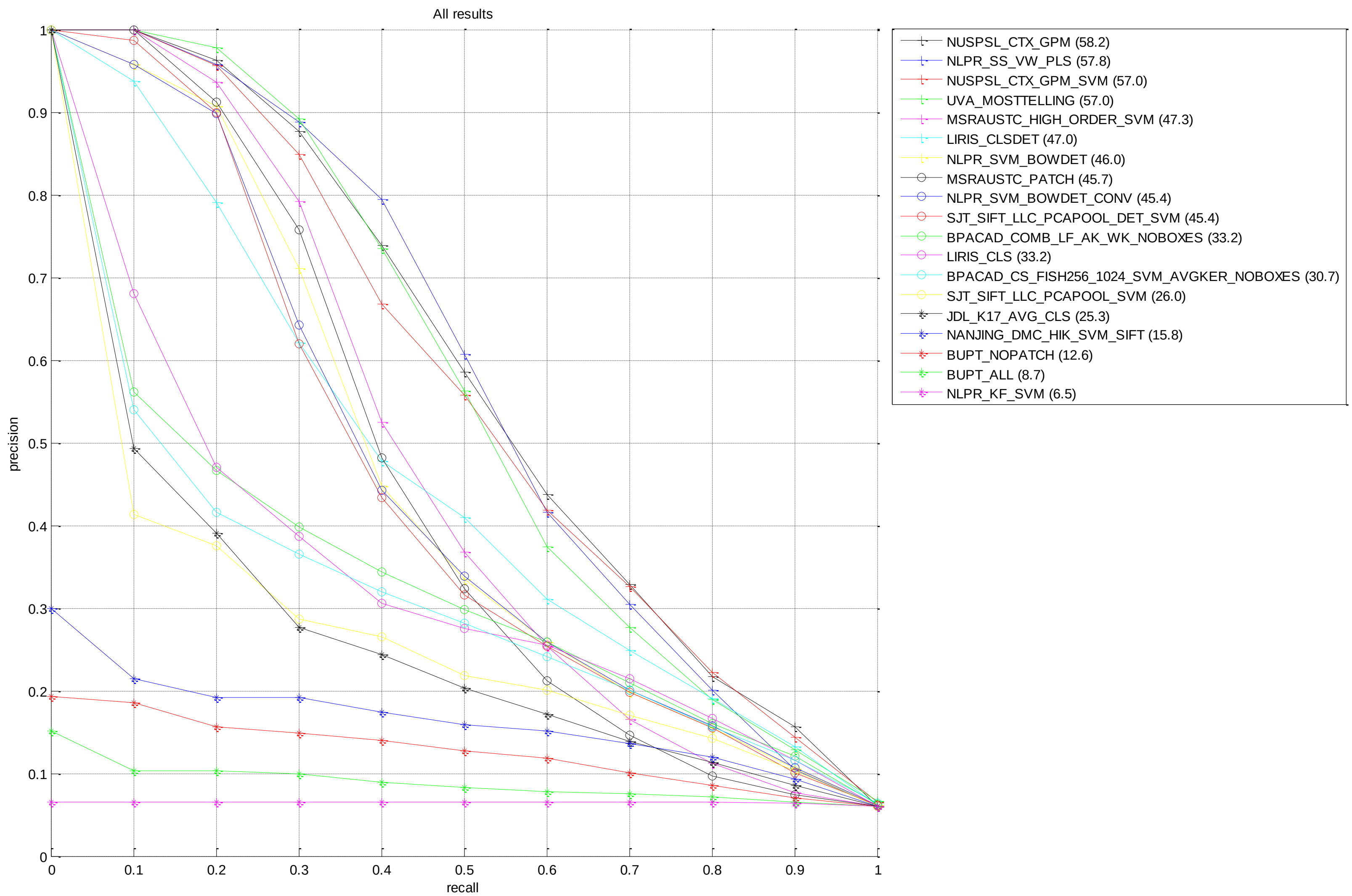
Precision/Recall: Bicycle (All)



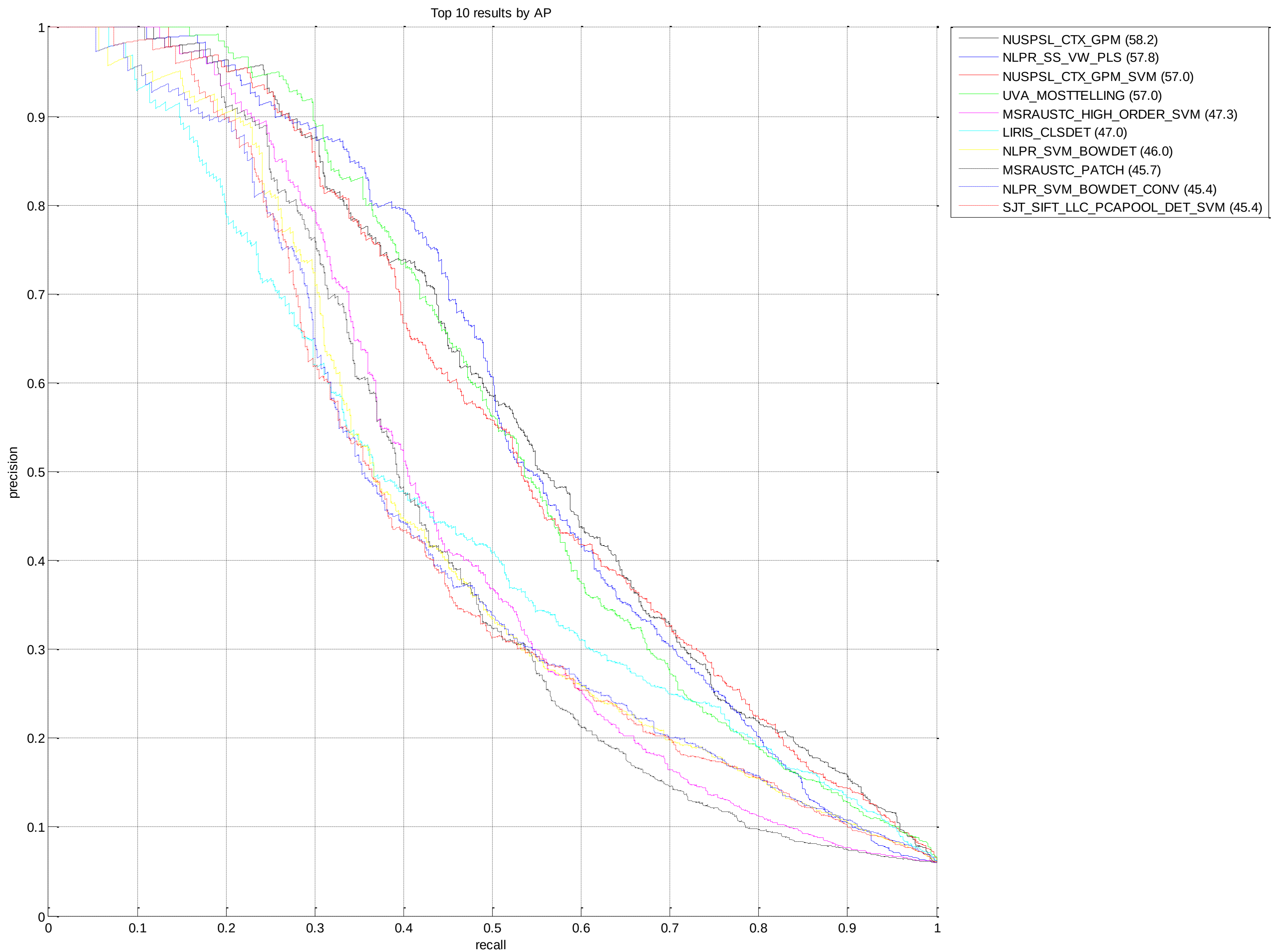
Precision/Recall: Bicycle (Top 10 by AP)



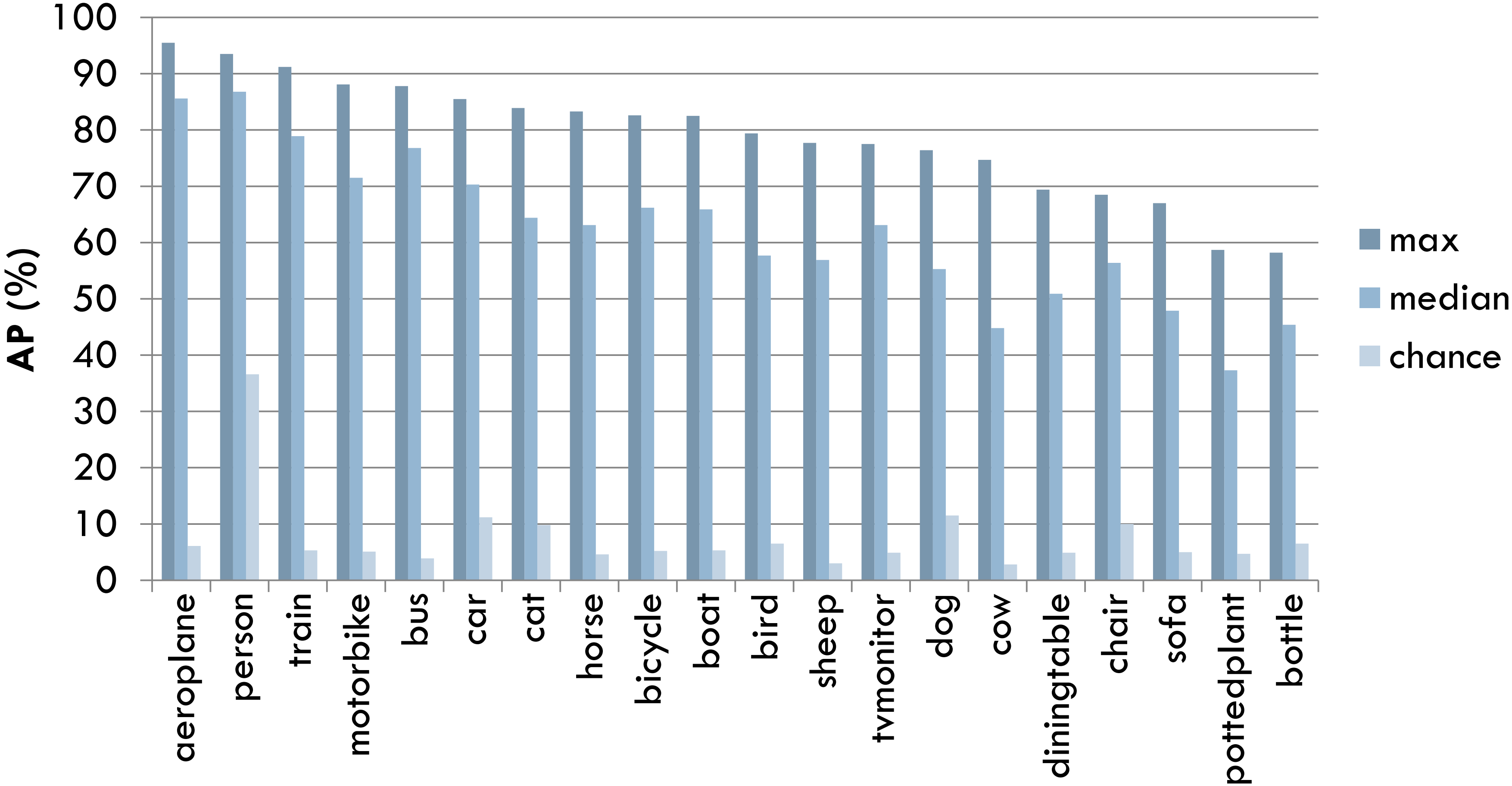
Precision/Recall: Bottle (All)



Precision/Recall: Bottle (Top 10 by AP)

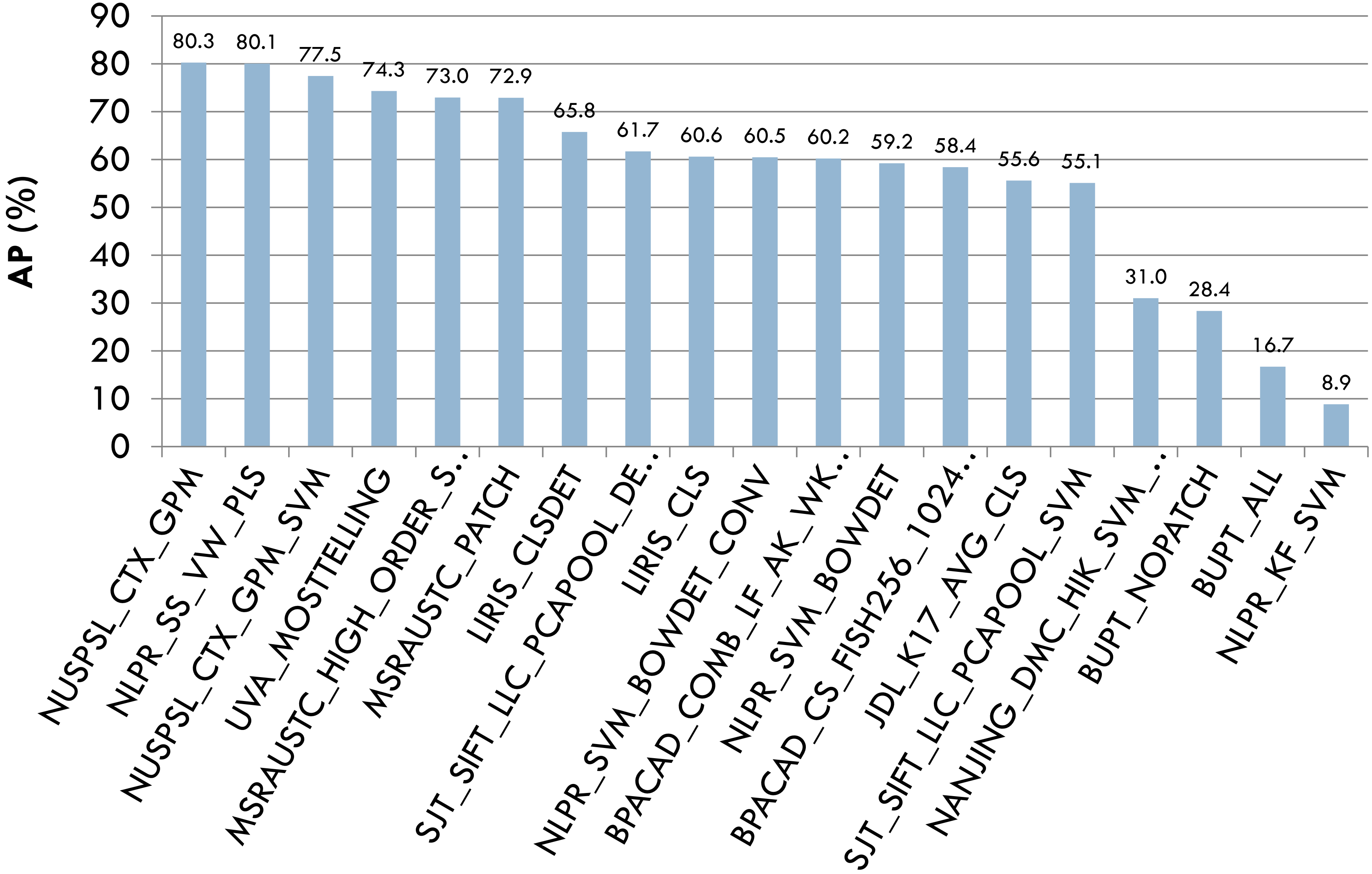


AP by Class



- Max AP: 95.5% (aeroplane) ... 58.2% (bottle)

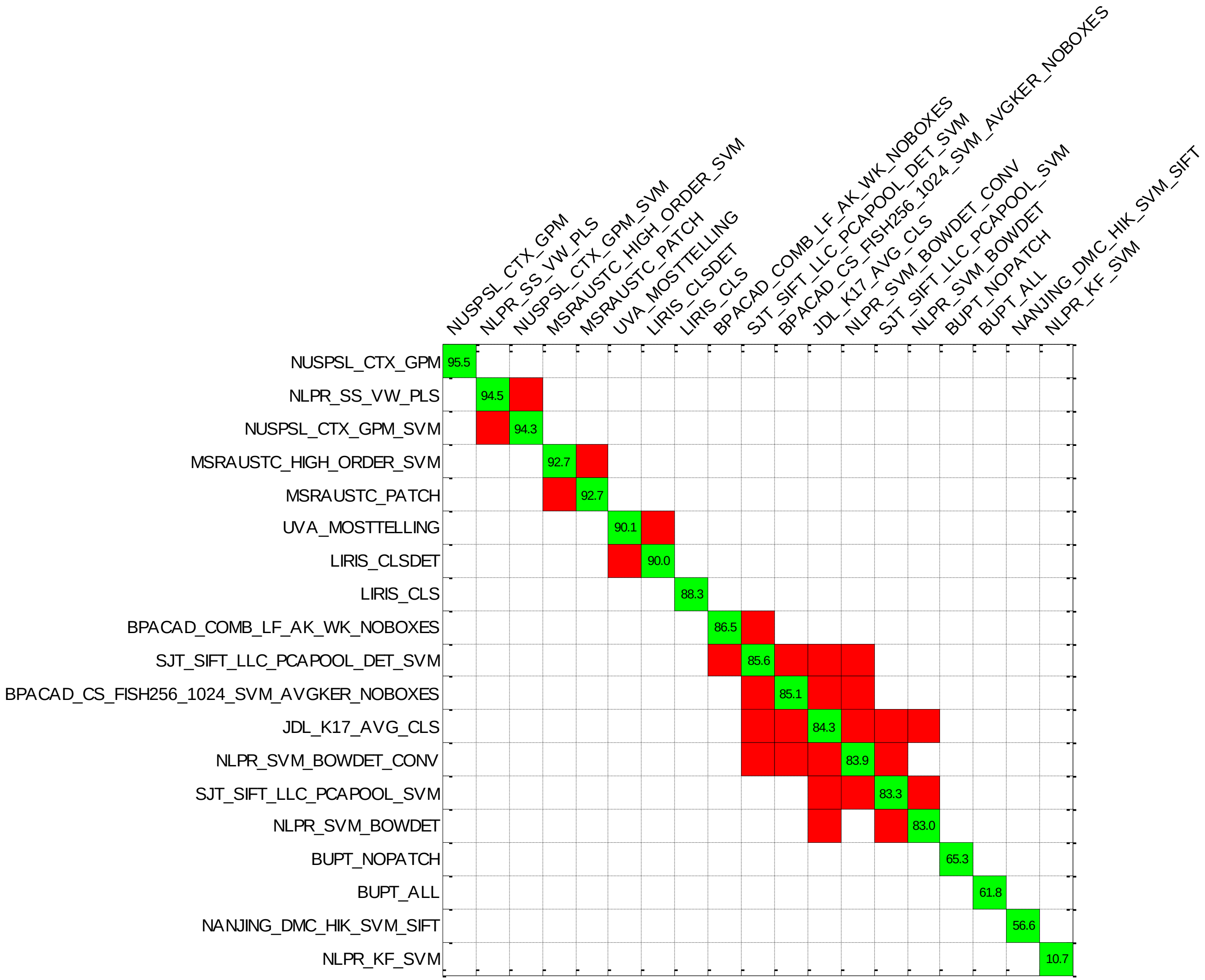
Median AP by Method



Statistical Significance (Preliminary)

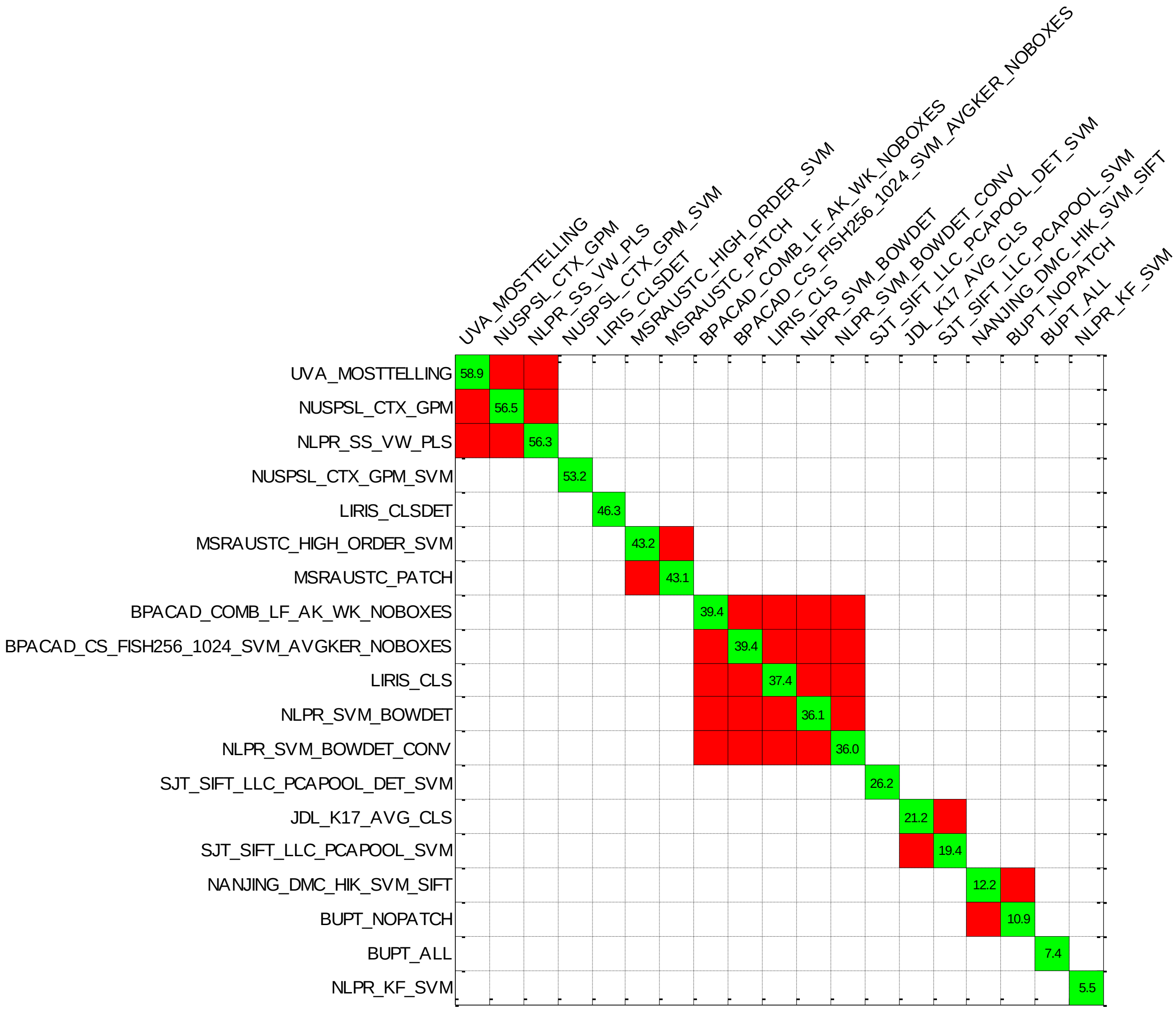
- Measure statistical significance of results with only a single test set
- Sample N=1000 test sets by bootstrap i.e. sample M images with replacement from original test set of size M
- To compare methods A and B:
  - Compute $AP_A(i)$ and $AP_B(i)$ for all sample test sets i
  - Compute paired differences $\delta i = AP_A(i) - AP_B(i)$
  - Test null hypothesis $\delta=0$ by computing percentiles of δ (p=0.9)
    - range does not contain $\delta = 0 \Rightarrow$ significant difference

Statistical Significance - Aeroplane

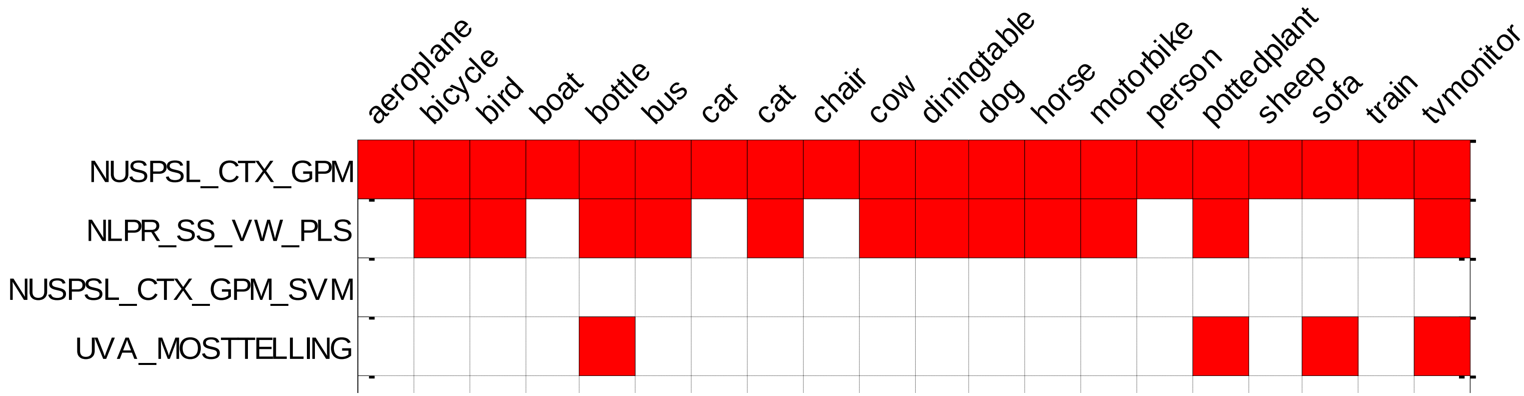


Statistical Significance - Bicycle

Statistical Significance – Potted Plant



Statistical Significance across Classes



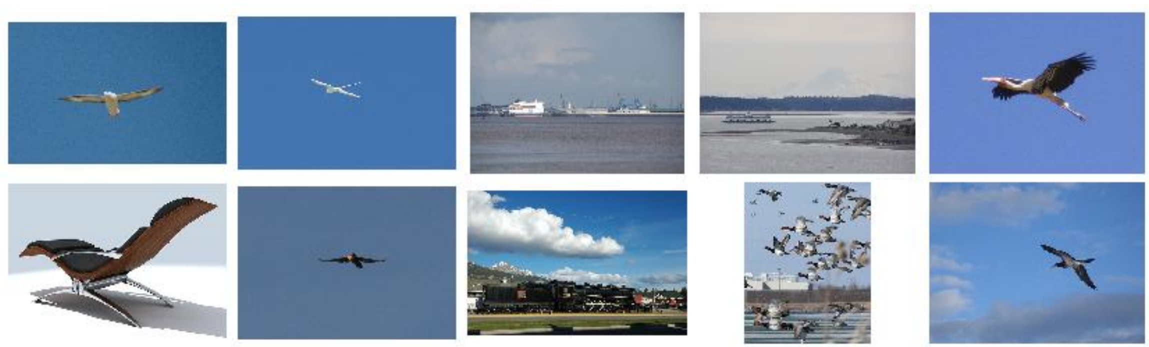
- **NUSPSL_CTX_GPM** gives best results for 11 classes
- Significantly better than all other methods for 7 classes
- Equivalent to **NLPR_SS_VW_PLS** for 12 classes
- Equivalent to **UVA_MOSTTELLING** for 4 classes

Ranked Images: Aeroplane

- Class images:
  Highest ranked

- Class images:
  Lowest ranked

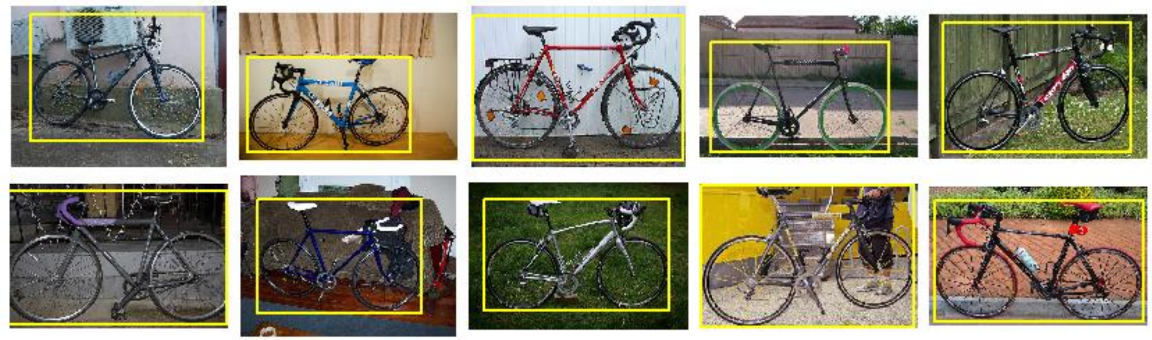
- Non-class images:
  Highest ranked



- Context?

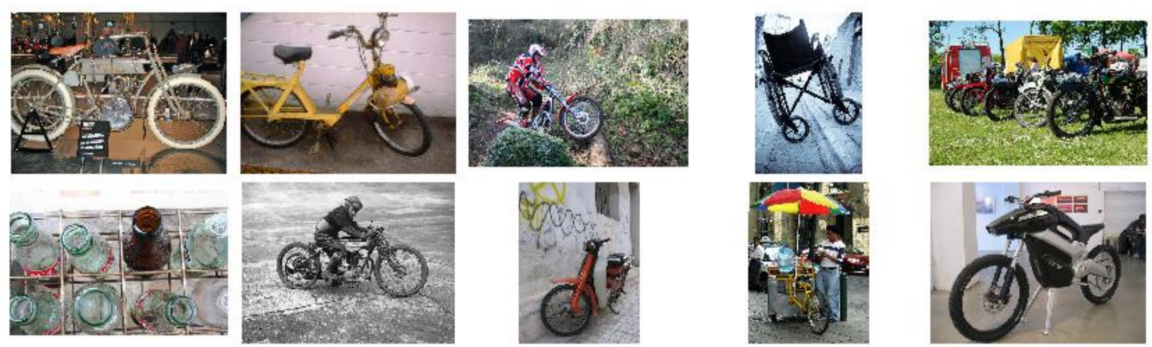
Ranked Images: Bicycle

- Class images: Highest ranked



- Class images: Lowest ranked

- Non-class images: Highest ranked

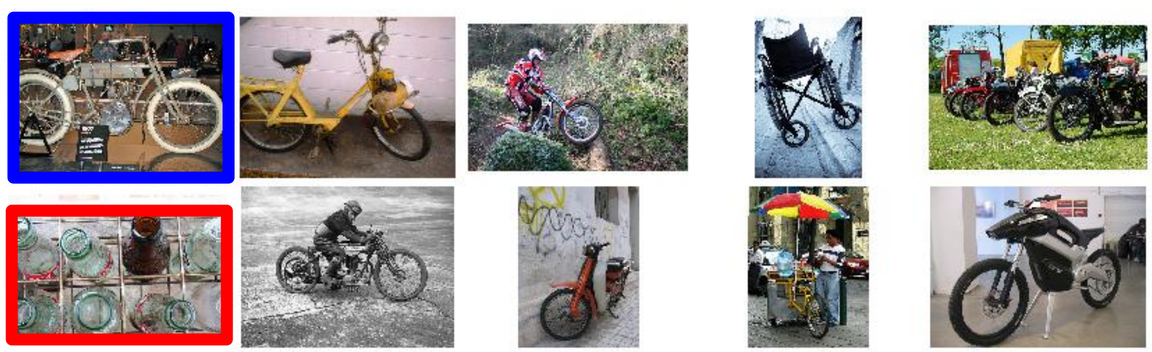


Non-bicycles 2009-2011

- 2009

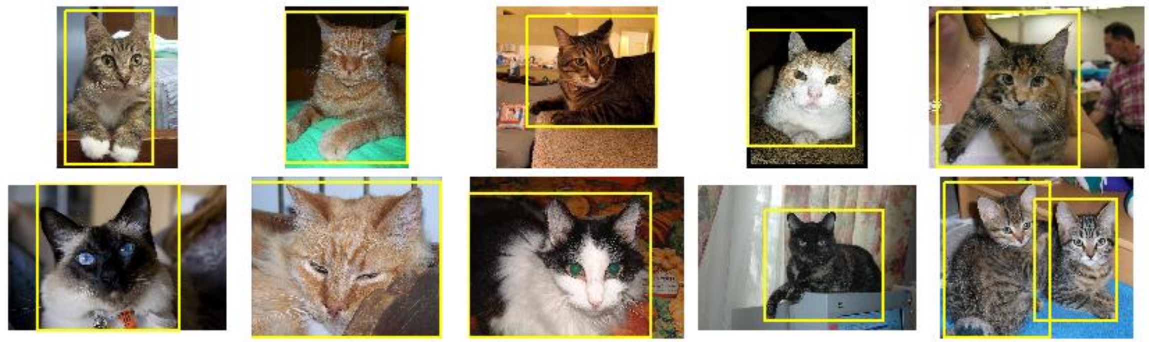
- 2010

- 2011

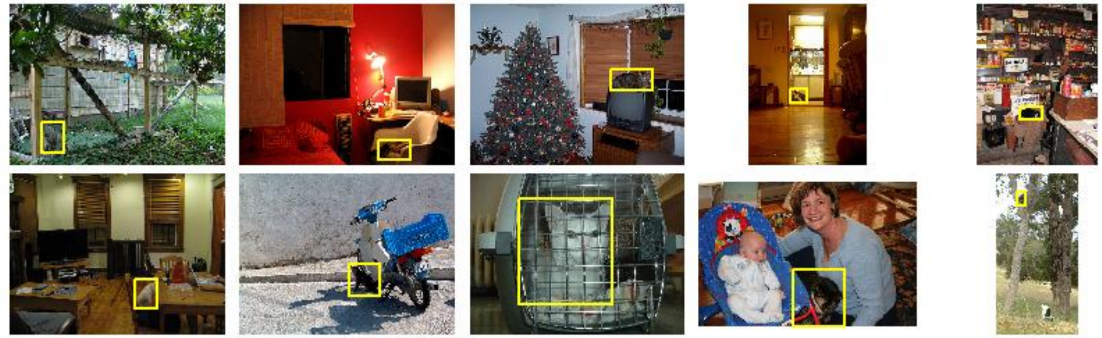


Ranked Images: Cat

- Class images: Highest ranked



- Class images: Lowest ranked

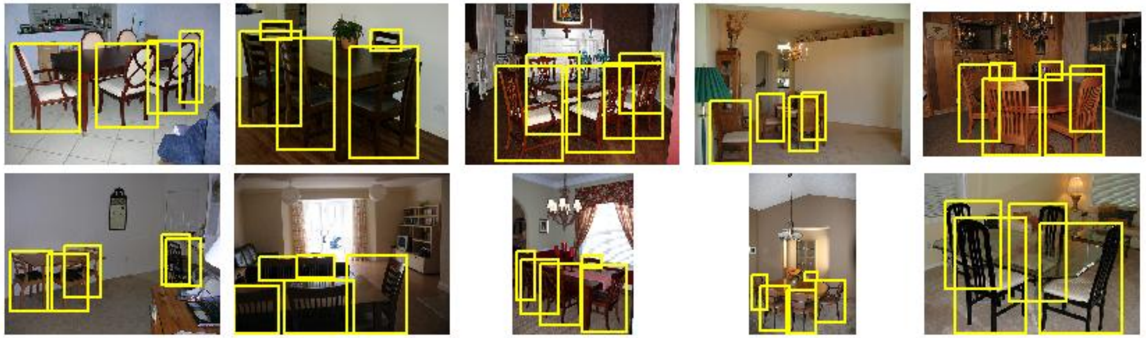


- Non-class images: Highest ranked

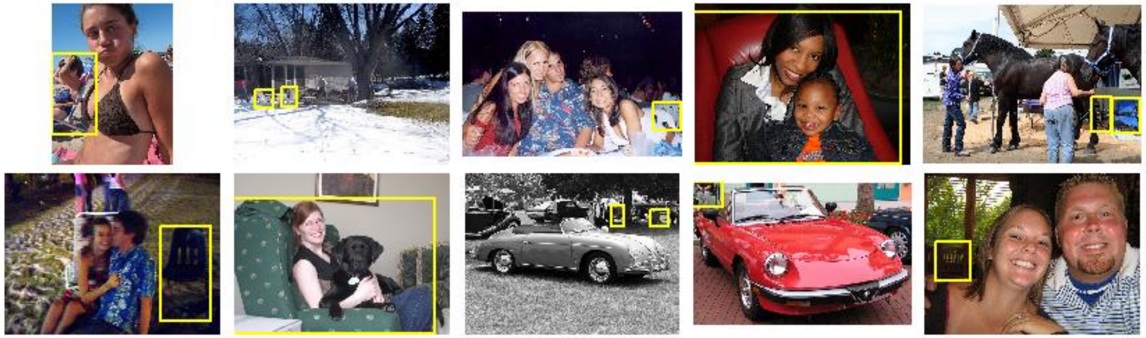
- "Composition"?

Ranked Images: Chair

- Class images:
  Highest ranked



- Class images:
  Lowest ranked

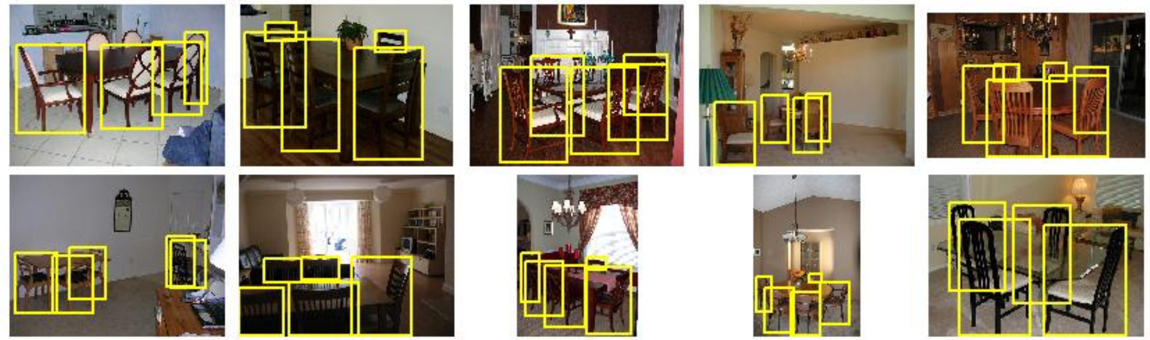


- Non-class images:
  Highest ranked

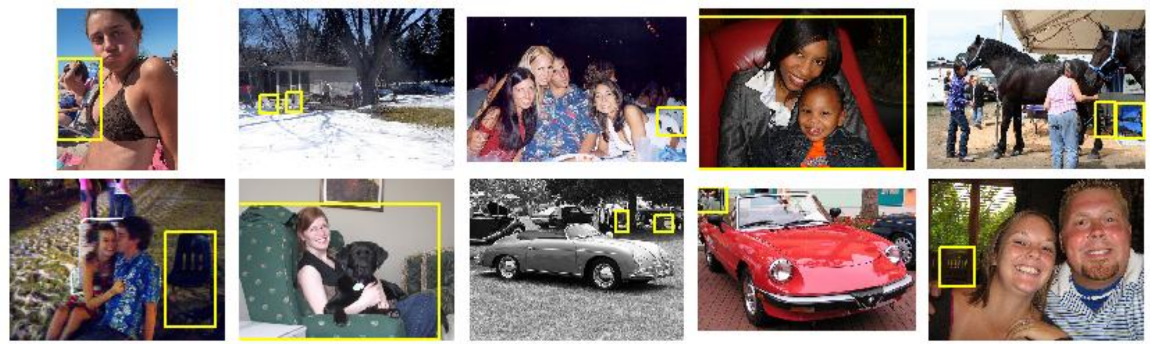
- Scene context? Sofa?

Ranked Images: Chair

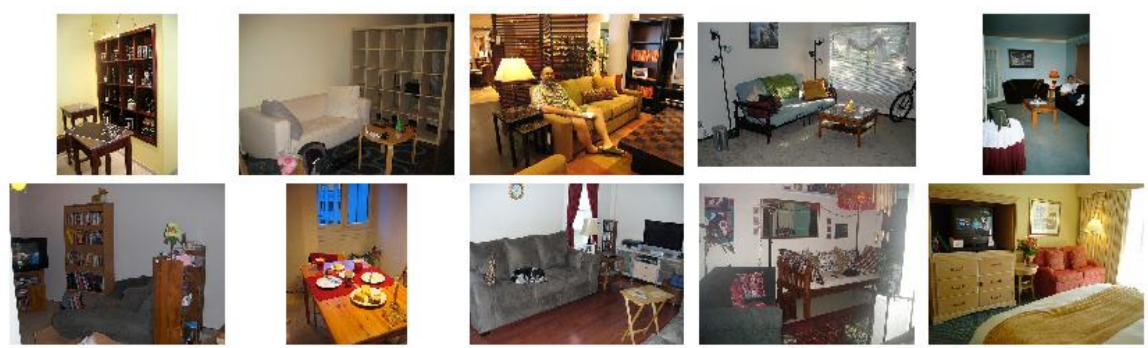
- Class images:
  Highest ranked



- Class images:
  Lowest ranked

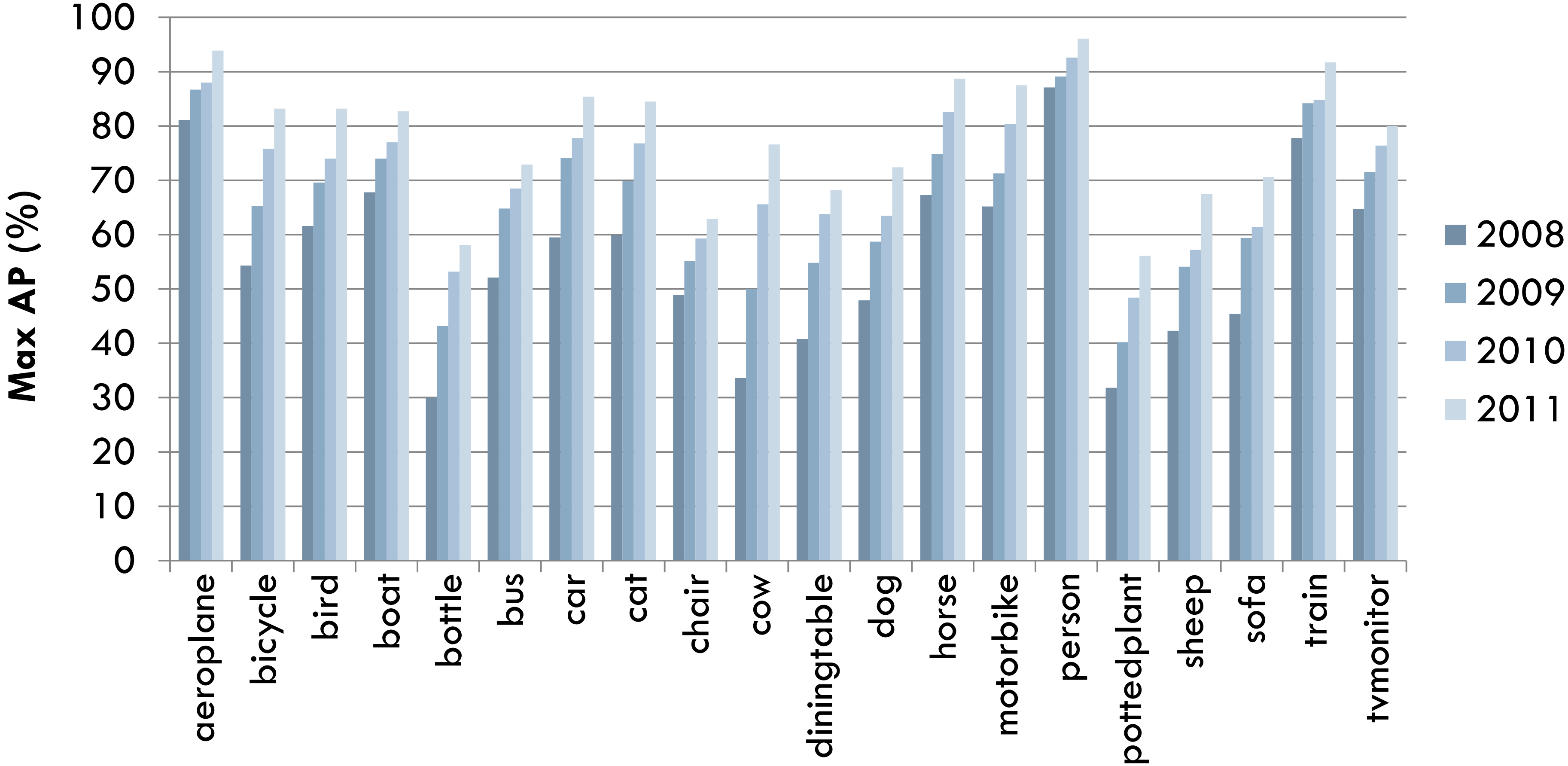


- Non-class images:
  Highest ranked



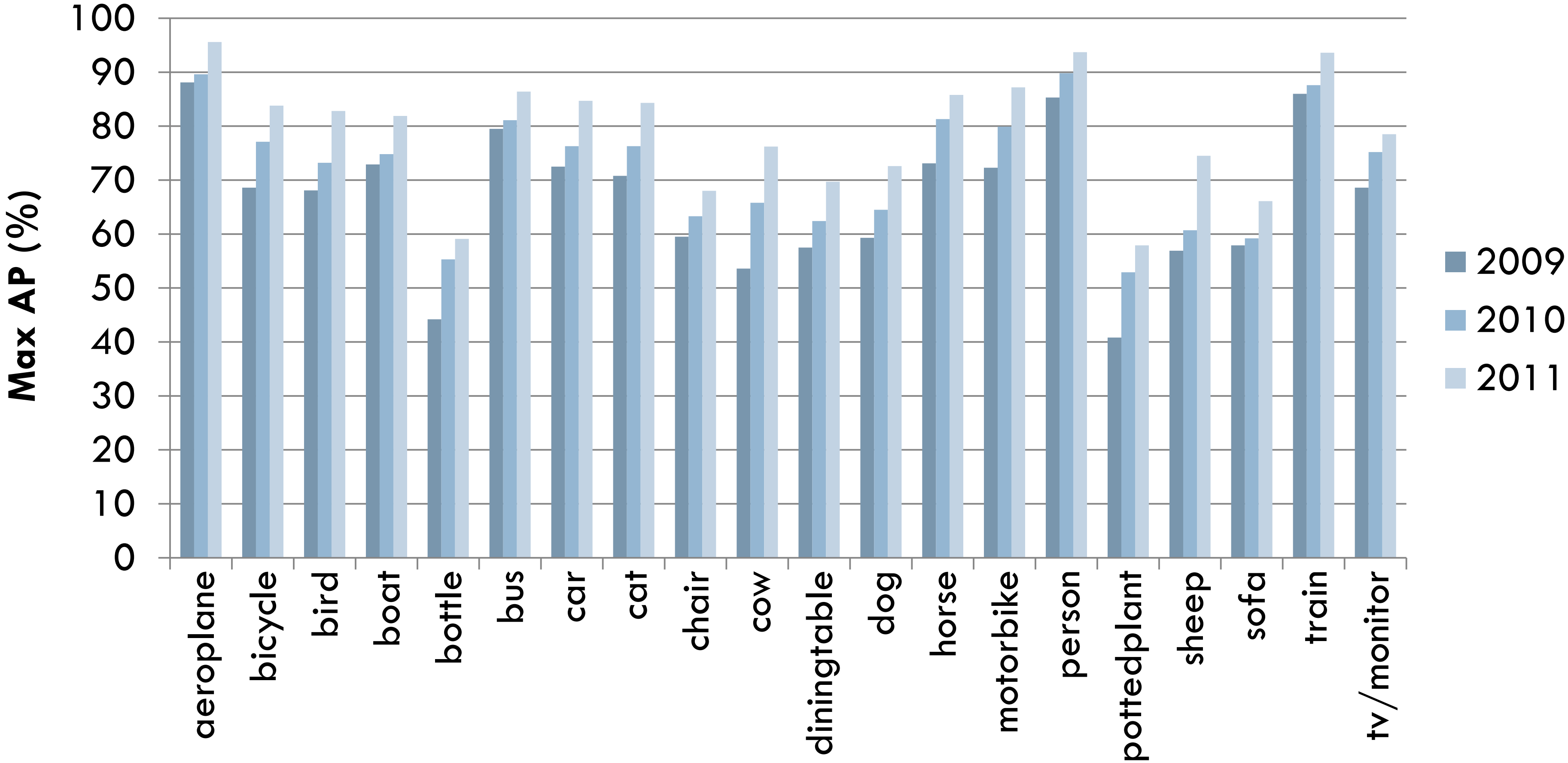
- Scene context? Sofa?

Progress 2008-2011



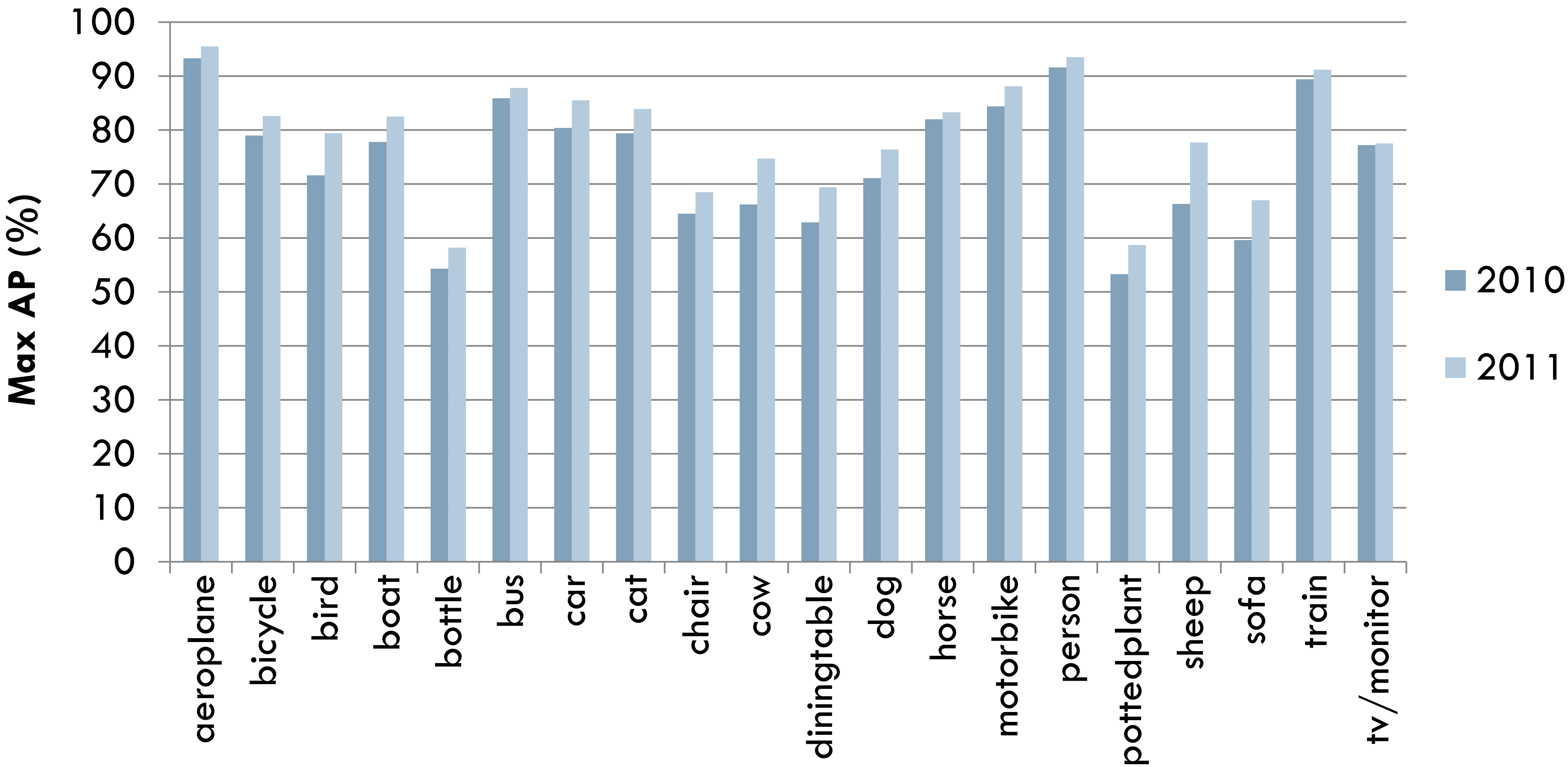
- Results on 2008 data improve for best methods 2009-2011 for all categories
  - Caveats: More training data + re-use of test data

Progress 2009-2011



- Results on 2009 data improve for best methods 2010-2011 for all categories
  - Caveats: More training data + re-use of test data

Progress 2010-2011



- Results on 2010 data improve for best 2011 methods for all categories
  - Caveats: More training data + re-use of test data

Prizes

- Winner:
  - **NUSPSL_CTX_GPM**
    Chen Qiang[1], Song Zheng[1], Yan Shuicheng[1], Hua Yang[2], Huang Zhongyang[2], Shen Shengmei[2]
    [1]*National University of Singapore*
    [2]*Panasonic Singapore Laboratories*
- Honourable Mentions:
  - **NLPR_SS_VW_PLS**
    Yinan Yu, Junge Zhang, Yongzhen Huang, Weiqiang Ren, Chong Wang, Jinchen Wu, Kaiqi Huang, Tieniu Tan
    National Laboratory of Pattern Recognition, Institute of Automation Chinese Academy of Sciences
  - **UVA_MOSTTELLING**
    Jasper Uijlings
    University of Amsterdam and University of Trento