The PASCAL Visual Object Classes Challenge 2010 (VOC2010)

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

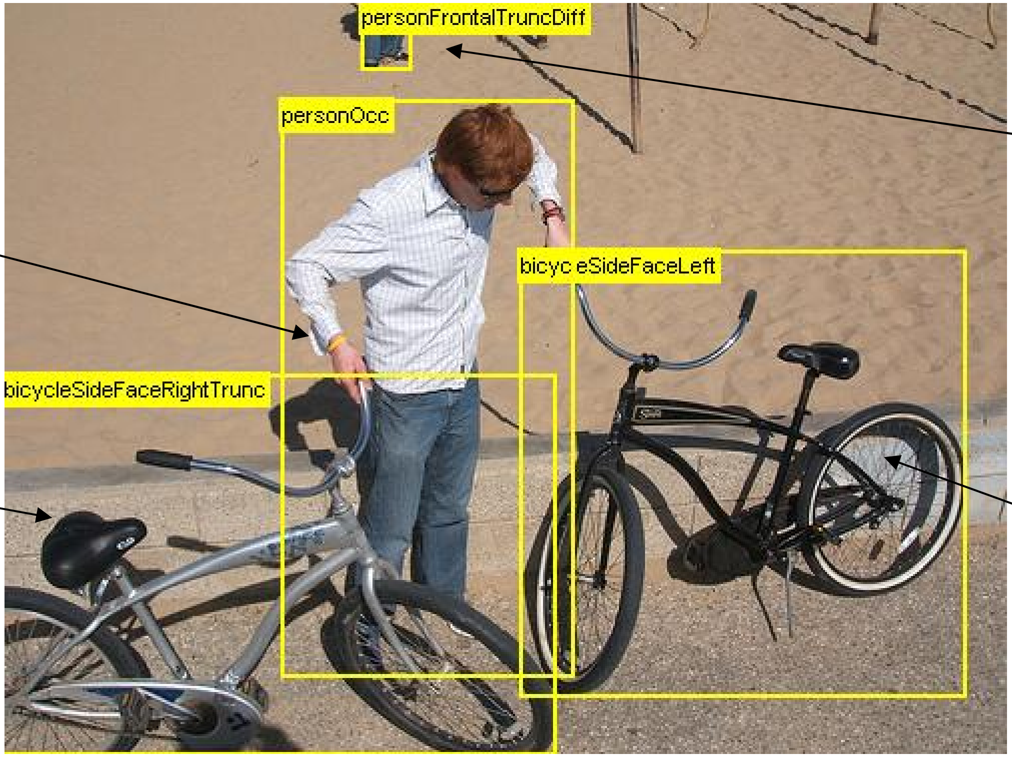
- Images downloaded from **flickr**
  - 500,000 images downloaded and random subset selected for annotation
  - Queries
    - Keyword e.g. “car”, “vehicle”, “street”, “downtown”
    - Date of capture e.g. “taken 21-July”
      - Removes “recency” bias in flickr results
    - Images selected from random page of results
      - Reduces bias toward particular flickr users
- 2008/9 datasets retained as subset of 2010
  - 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

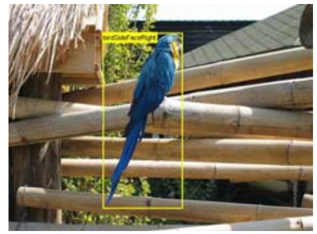
1. Amazon Mechanical Turk
   - Qualification task
   - Images labelled with presence/absence of object categories
   - Bounding boxes labelled for subsets of object categories e.g. bicycle/bus/car/motorbike
2. Experienced Annotators
   - Web-based tool, written guidelines
   - Annotation corrected and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

Examples

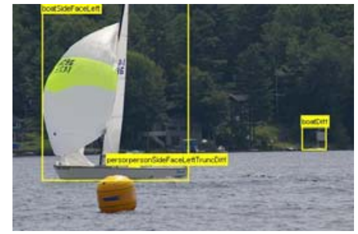
Aeroplane

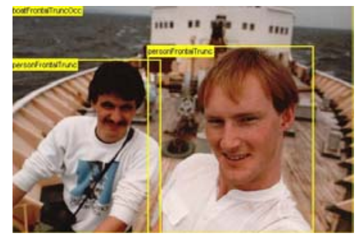
Bicycle

Bird

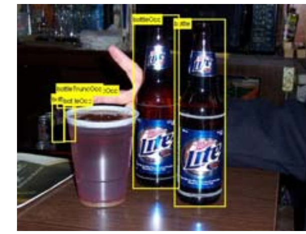


Boat





Bottle



Bus

Car

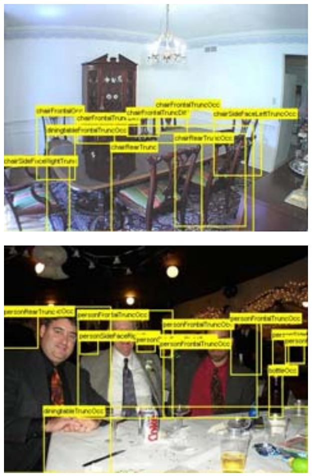
Cat

Chair

Cow

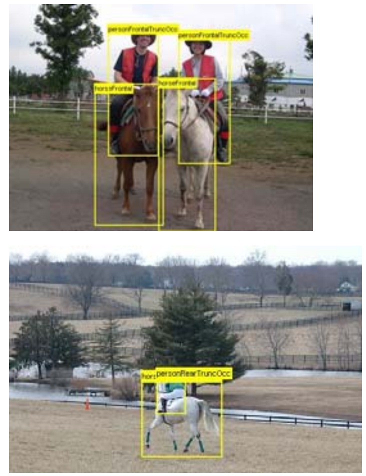
Examples

Dining Table

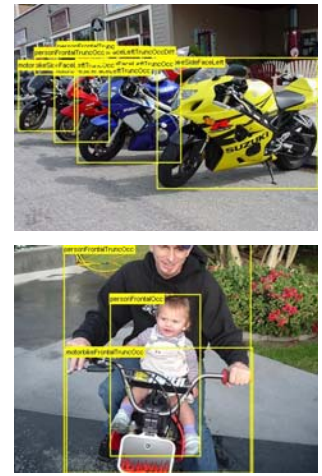


Dog

Horse



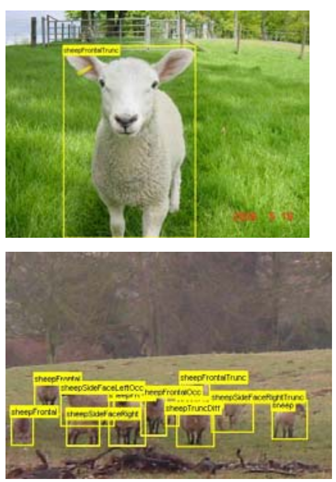
Motorbike



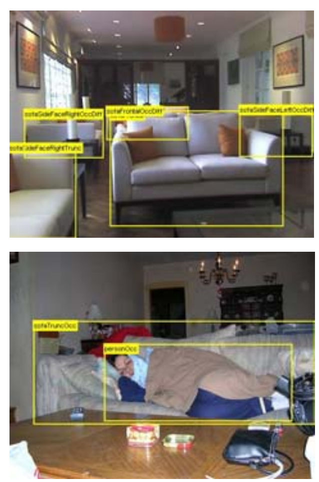
Person

Potted Plant

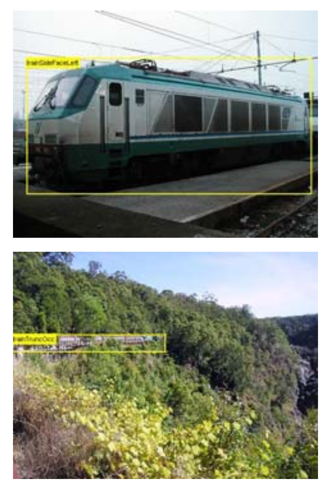
Sheep



Sofa



Train



TV/Monitor

Dataset Statistics

- Around 40% increase in size over VOC2009

	Training		Testing	
Images	10,103	(7,054)	9,637	(6,650)
Objects	23,374	(17,218)	22,992	(16,829)

VOC2009 counts shown in brackets

- Minimum ~500 training objects per category
  - ~1700 cars, 1500 dogs, 7000 people
- Approximately equal distribution across training and test sets

Best Practice

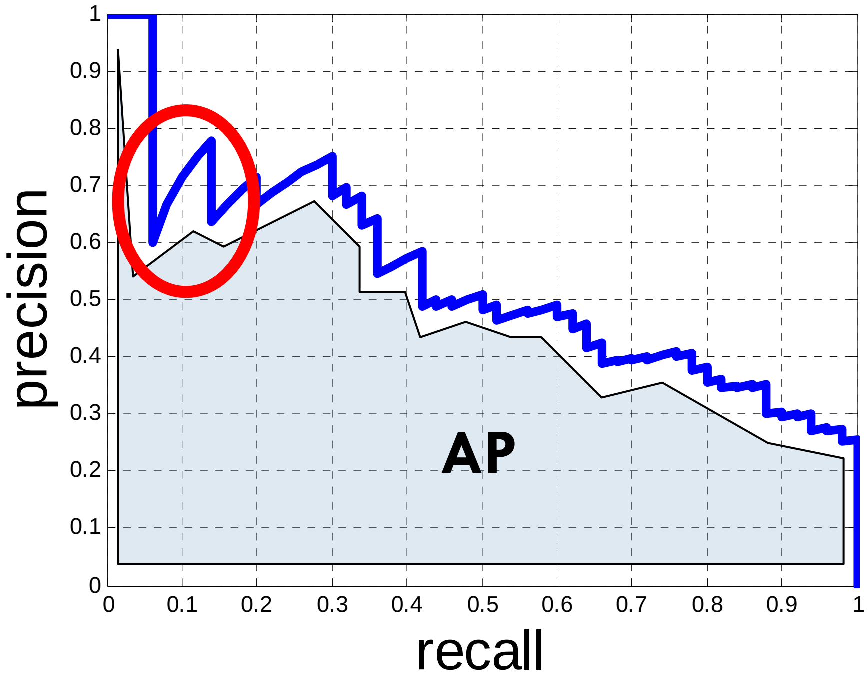
- If using the provided training data ("trainval"), **<u>all</u>** feature selection, parameter tuning, choice of classifier architecture, etc. should be done using the training data alone
  - Use suggested training/validation split
  - Use cross-validation
- Do report results on the most recent dataset i.e. 2010
- 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?

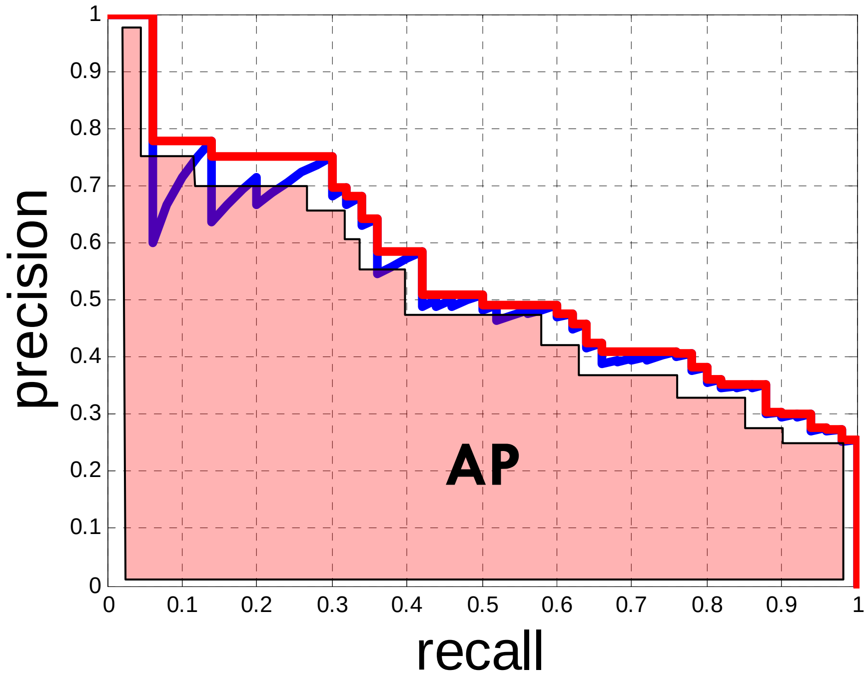
Average Precision



- 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

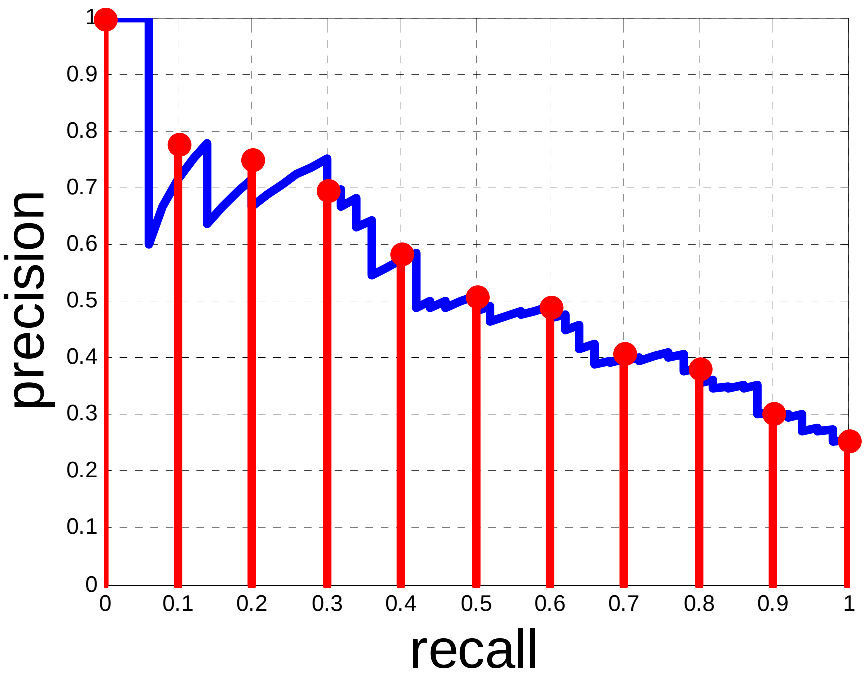
Average Precision: VOC2010



- 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

Average Precision: VOC2009 (Obsolete)



- From TREC challenge
- Interpolate precision as maximum precision at greater or equal recall
- Average interpolated precision at 11 recall levels 0, 0.1, 0.2, …
- Sawtooth shape is ignored
- Poor measurement accuracy
  - e.g. AP ~9% by one high ranked true positive

Methods

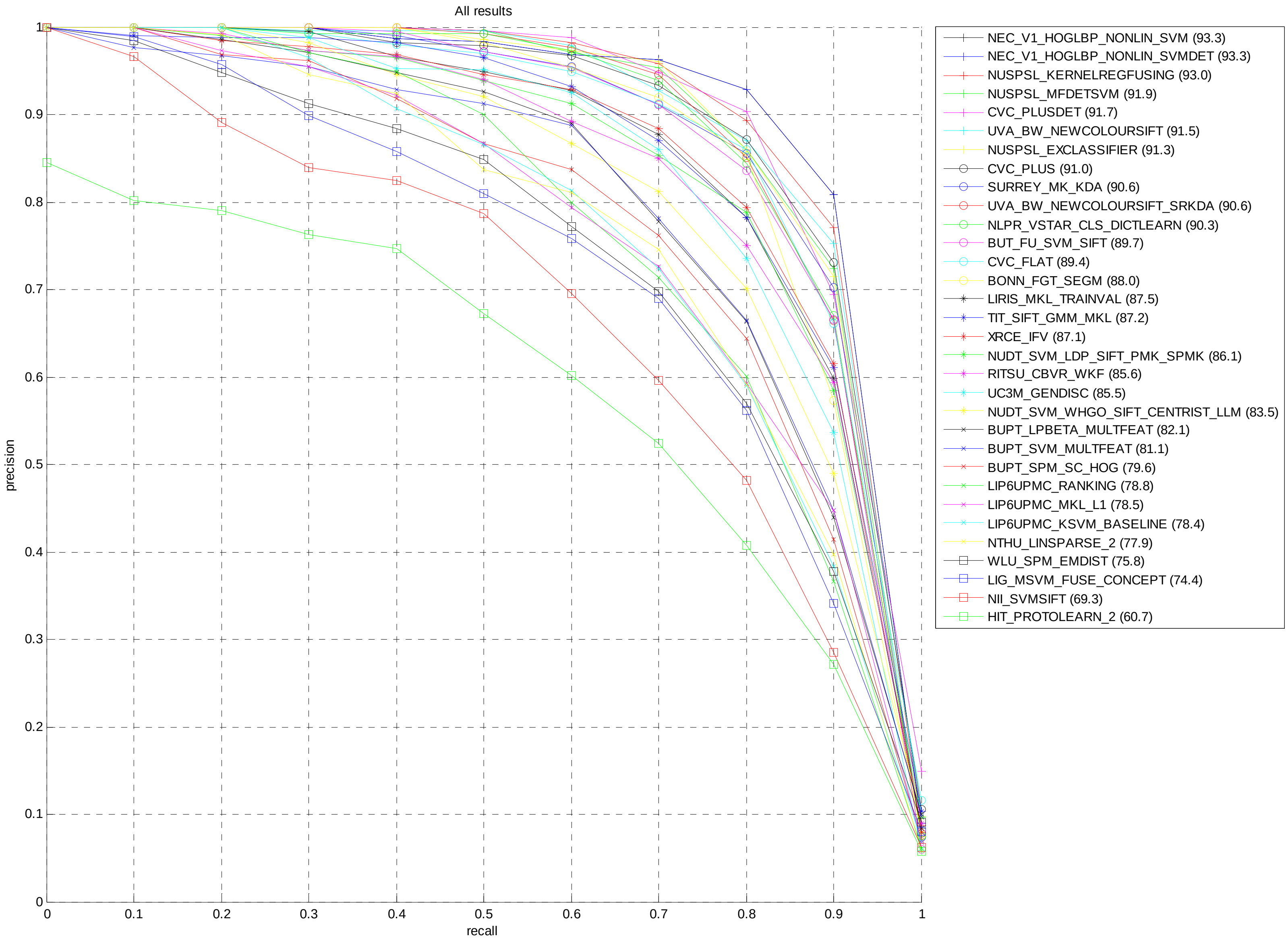
- 33 Methods, 22 Groups
  - VOC2009: 48 “Methods”, 20 Groups
- Basic recipe
  - bag of visual words and/or spatial pyramid
  - multiple features: interest points/dense/saliency, SIFT, HOG, color SIFT, LBP, gist, etc.
  - SVM classifier
  - feature combination by MKL or voting
- Additional ingredients
  - Combine detection scores, segmentation

AP by Class/Method

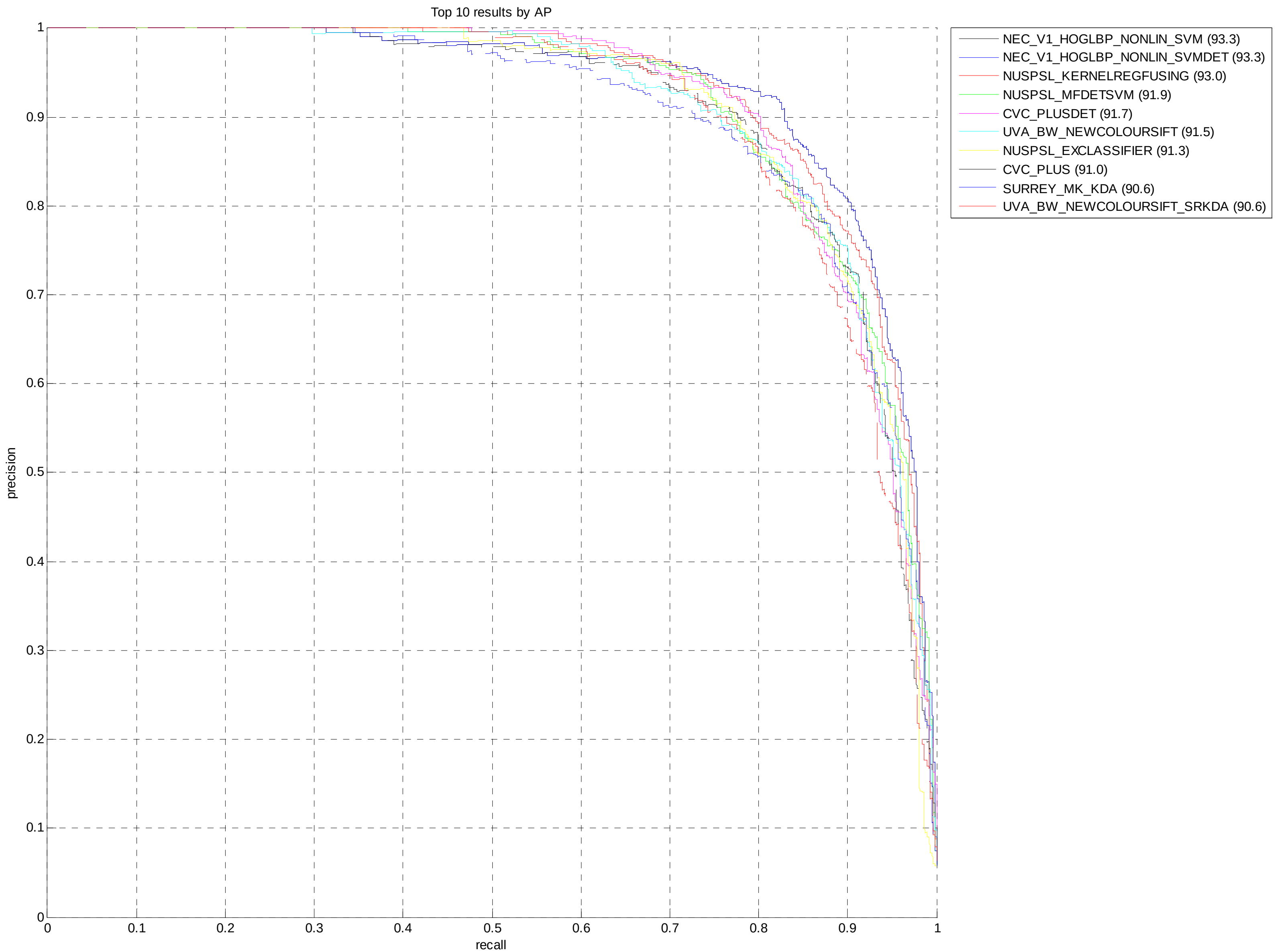
(1st, 2nd, 3rd place)

	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
BONN_FGT_SEGM	88.0	61.6	53.1	63.3	34.8	77.5	72.3	71.1	41.1	56.0	39.6	64.3	68.9	75.4	87.5	32.5	59.3	40.8	78.7	61.4
BUPT_LPBETA_MULTFEAT	82.1	38.6	39.5	46.5	15.5	55.0	46.4	46.5	39.9	21.3	31.2	37.6	45.8	41.4	75.5	15.6	41.7	25	62.5	44.3
BUPT_SPM_SC_HOG	79.6	47.0	42.9	52.3	21.3	66.6	50.1	58.7	44.3	21.8	32.7	46.0	49.7	51.7	72.4	13.2	44.1	28.1	61.5	48.8
BUPT_SVM_MULTFEAT	81.1	45.3	47.3	46.3	20.1	42.3	36.4	49.1	37.5	20.6	38.5	43.8	44.9	54.4	68.6	18.0	48.2	26.0	57.7	40.3
BUT_FU_SVM_SIFT	89.7	63.9	64.5	68.3	36.8	77.9	68.5	72.0	57.2	47.2	56.7	63.5	66.8	74.2	85.0	32.8	54.3	49.1	82.6	66.8
CVC_FLAT	89.4	57.6	63.0	68.5	32.0	76.7	64.7	66.9	51.5	48.4	50.0	54.8	63.1	69.9	83.5	33.6	54.8	46.1	82.2	65.9
CVC_PLUS	91.0	61.8	66.7	71.1	37.7	78.9	67.8	72.2	55.8	51.0	55.8	59.4	65.3	73.0	84.0	39.9	56.9	48.5	83.9	68.1
CVC_PLUSDET	91.7	70.0	66.8	71.3	49.0	81.4	77.5	71.2	60.0	52.6	55.7	61.0	70.9	76.7	88.4	43.2	59.7	53.8	84.7	71.3
HIT_PROTOLEARN_2	60.7	22.1	22.7	29.0	15.0	34.9	27.8	31.6	31.9	14.1	17.4	28.9	24.0	20.6	55.8	9.2	22.0	16.8	30.9	24.6
LIG_MSVM_FUSE_CONCEPT	74.4	43.0	37.5	50.4	22.0	60.7	47.1	46.8	47.5	22.2	35.0	42.1	42.9	48.4	73.8	15.6	31.8	28.9	63.8	46.6
LIP6UPMC_KSVM_BASELINE	78.4	54.1	49.9	61.1	24.6	68.3	58.0	59.9	50.7	35.7	42.5	55.0	60.8	63.1	71.1	25.9	51.5	39.9	74.1	59.6
LIP6UPMC_MKL_L1	78.5	55.9	54.6	62.5	25.0	69.3	59.5	60.0	51.3	37.9	46.7	54.0	60.5	64.0	72.8	32.8	52.6	38.5	72.7	61.1
LIP6UPMC_RANKING	78.8	51.3	46.1	58.2	19.5	68.6	55.6	59.4	46.8	30.7	36.0	49.3	52.3	60.0	76.3	17.8	49.1	35.3	66.3	56.6
LIRIS_MKL_TRAINVAL	87.5	57.0	61.7	68.2	29.9	76.6	61.9	67.5	56.9	35.1	50.6	55.1	62.2	69.3	83.6	35.9	52.9	42.7	79.8	66.3
NEC_V1_HOGLBP_NONLIN_SVM	93.3	71.7	69.9	76.9	42.0	85.3	77.4	79.3	60.0	55.8	60.6	71.1	75.7	77.7	86.8	33.5	61.5	55.8	87.5	69.9
NEC_V1_HOGLBP_NONLIN_SVMDET	93.3	72.9	69.9	77.2	47.9	85.6	79.7	79.4	61.7	56.6	61.1	71.1	76.7	79.3	86.8	38.1	63.9	55.8	87.5	72.9
NII_SVMSIFT	69.3	40.3	27.3	44.1	19.5	54.1	23.9	44.4	42.9	20.3	31.1	37.5	36.6	40.5	68.8	9.3	24.6	20.2	55.6	43.9
NLPR_VSTAR_CLS_DICTLEARN	90.3	77.0	65.3	75.0	53.7	85.9	80.4	74.6	62.9	66.2	54.1	66.8	76.1	81.7	89.9	41.6	66.3	57.0	85.0	74.3
NTHU_LINSPARSE_2	77.9	44.0	37.4	48.5	19.0	63.6	49.0	51.0	45.5	27.6	32.1	41.7	46.9	49.7	68.5	13.2	40.3	30.1	61.7	46.3
NUDT_SVM_LDP_SIFT_PMK_SPMK	86.1	59.3	60.2	68.7	28.7	74.8	63.5	68.0	52.5	41.4	47.1	57.5	60.9	68.2	81.5	29.4	52.1	44.5	79.1	4.7
NUDT_SVM_WHGO_SIFT_CENTRIST_LLM	83.5	54.2	55.2	66.8	28.5	72.1	65.4	64.2	51.9	36.1	49.3	55.6	58.0	66.5	82.1	25.3	48.1	41.7	78.4	59.5
NUSPSL_EXCLASSIFIER	91.3	77.0	70.0	75.6	50.7	83.2	77.1	75.4	62.5	62.6	62.7	64.6	77.9	81.8	91.1	44.8	64.2	53.2	86.3	77.1
NUSPSL_KERNELREGFUSING	93.0	79.0	71.6	77.8	54.3	85.2	78.6	78.8	64.5	64.0	62.7	69.6	82.0	84.4	91.6	48.6	64.9	59.6	89.4	76.4
NUSPSL_MFDETSVM	91.9	77.1	69.5	74.7	52.5	84.3	77.3	76.2	63.0	63.5	62.9	65.0	79.5	83.2	91.2	45.5	65.4	55.0	87.0	77.2
RITSU_CBVR_WKF	85.6	57.2	54.9	64.5	29.2	71.2	57.1	63.2	53.9	37.6	49.6	54.7	58.7	67.9	80.1	29.2	52.1	43.5	76.4	60.9
SURREY_MK_KDA	90.6	66.1	67.2	70.6	36.0	79.7	69.8	73.4	58.4	50.7	60.1	65.2	69.8	76.9	87.0	42.5	59.6	49.9	85.2	71.3
TIT_SIFT_GMM_MKL	87.2	56.6	59.6	66	32.6	72.7	63.1	64.8	54.6	41.2	49.3	58.8	59.1	68.2	82.9	31.2	49.2	43.2	75.0	63.4
UC3M_GENDISC	85.5	51.6	55.4	64.8	25.9	74.4	60.6	66.0	51.0	45.9	43.9	55.0	59.0	65.2	80.3	24.0	51.4	47.0	76.4	58.6
UVA_BW_NEWCOLOURSIFT	91.5	71.0	67.3	69.9	43.9	80.6	75.3	73.4	59.3	57.8	60.8	64.0	70.6	80.0	88.6	50.8	65.6	56.1	83.0	76.2
UVA_BW_NEWCOLOURSIFT_SRKDA	90.6	66.9	63.4	70.2	49.4	81.8	76.7	70.9	60.0	57.1	60.5	64.5	67.4	79.1	90.2	53.3	63.5	58.0	81.9	74.4
WLU_SPM_EMDIST	75.8	48.9	36.8	44.3	21.2	65.8	52.1	52.1	45.4	28.2	35.0	45.3	47.8	54.2	71.0	14.7	39.8	32.7	62.2	48.0
XRCE_IFV	87.1	59.6	59.9	69.7	31.3	76.4	62.9	64.3	52.5	42.4	55.1	59.7	64.3	70.4	83.9	32.6	53.3	50.4	80.0	67.6

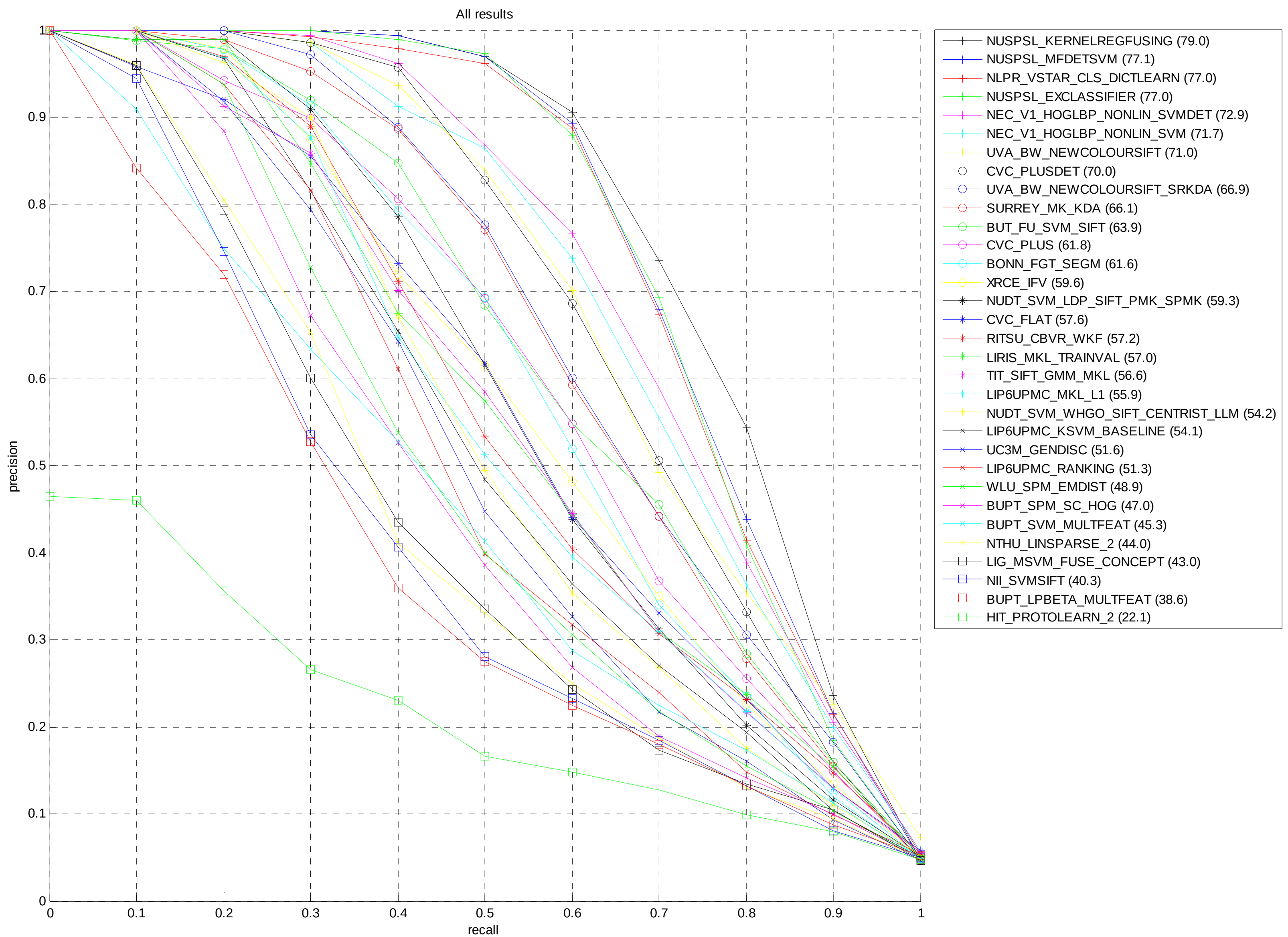
Precision/Recall: Aeroplane (All)



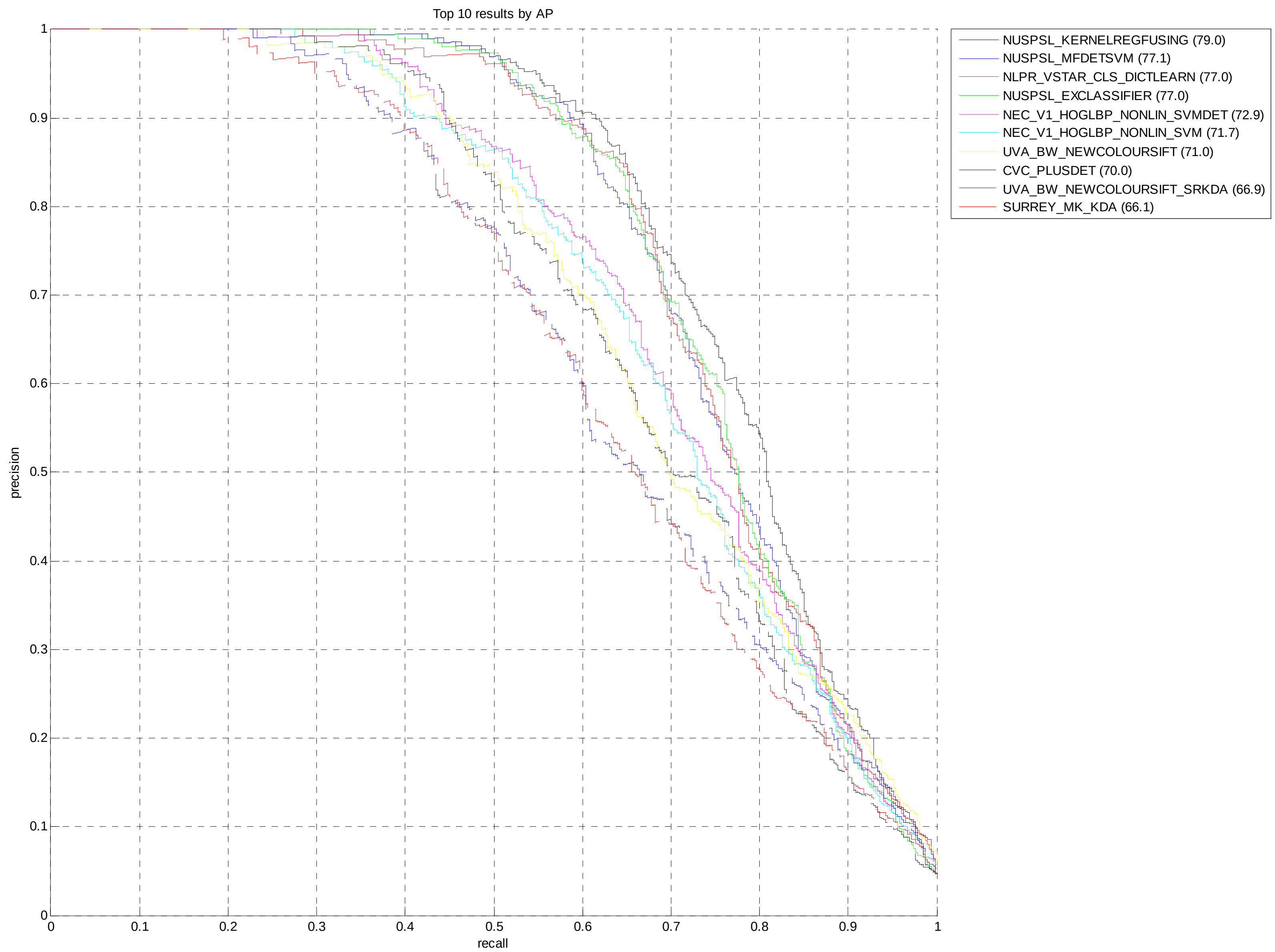
Precision/Recall: Aeroplane (Top 10 by AP)



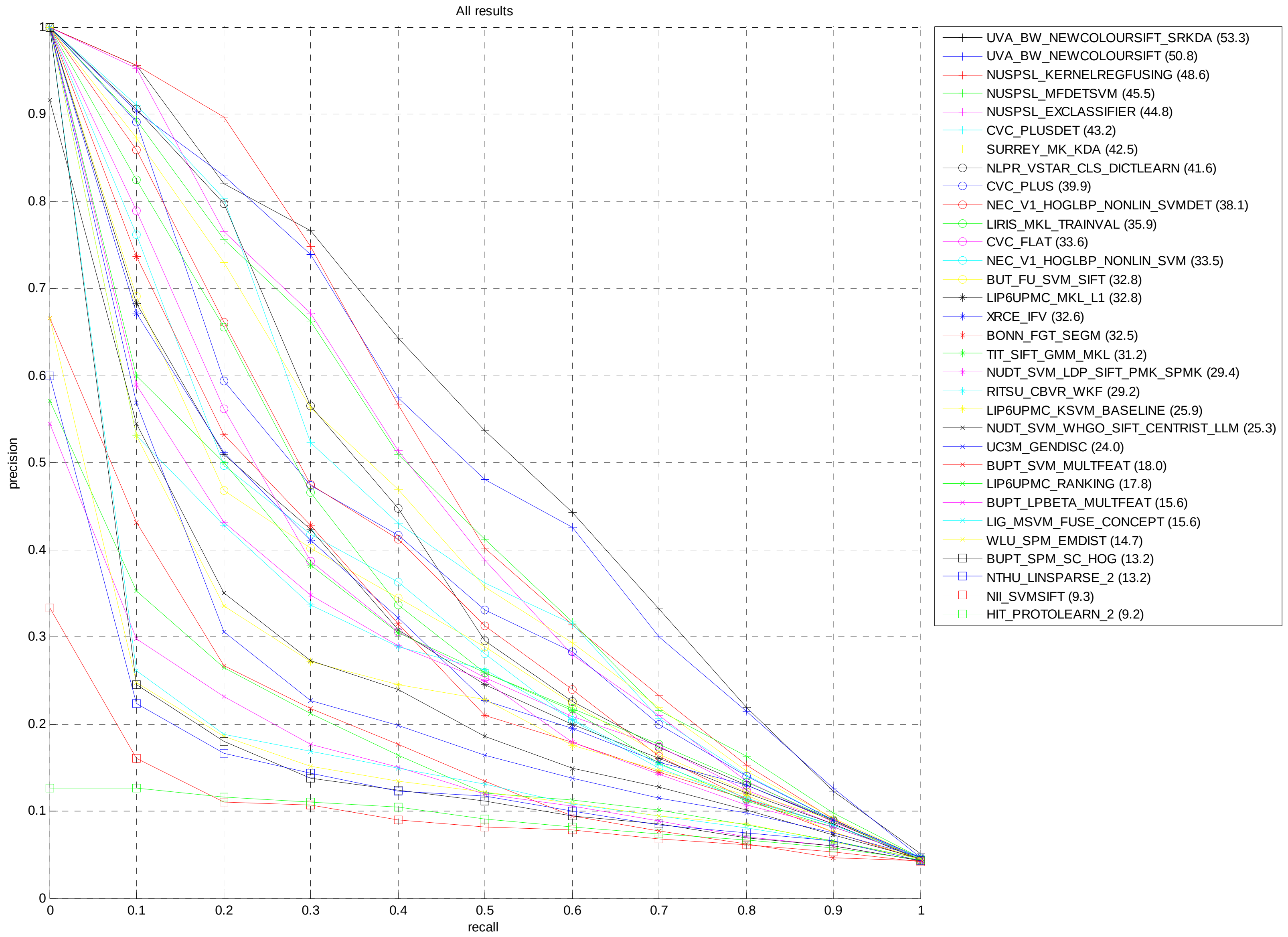
Precision/Recall: Bicycle (All)



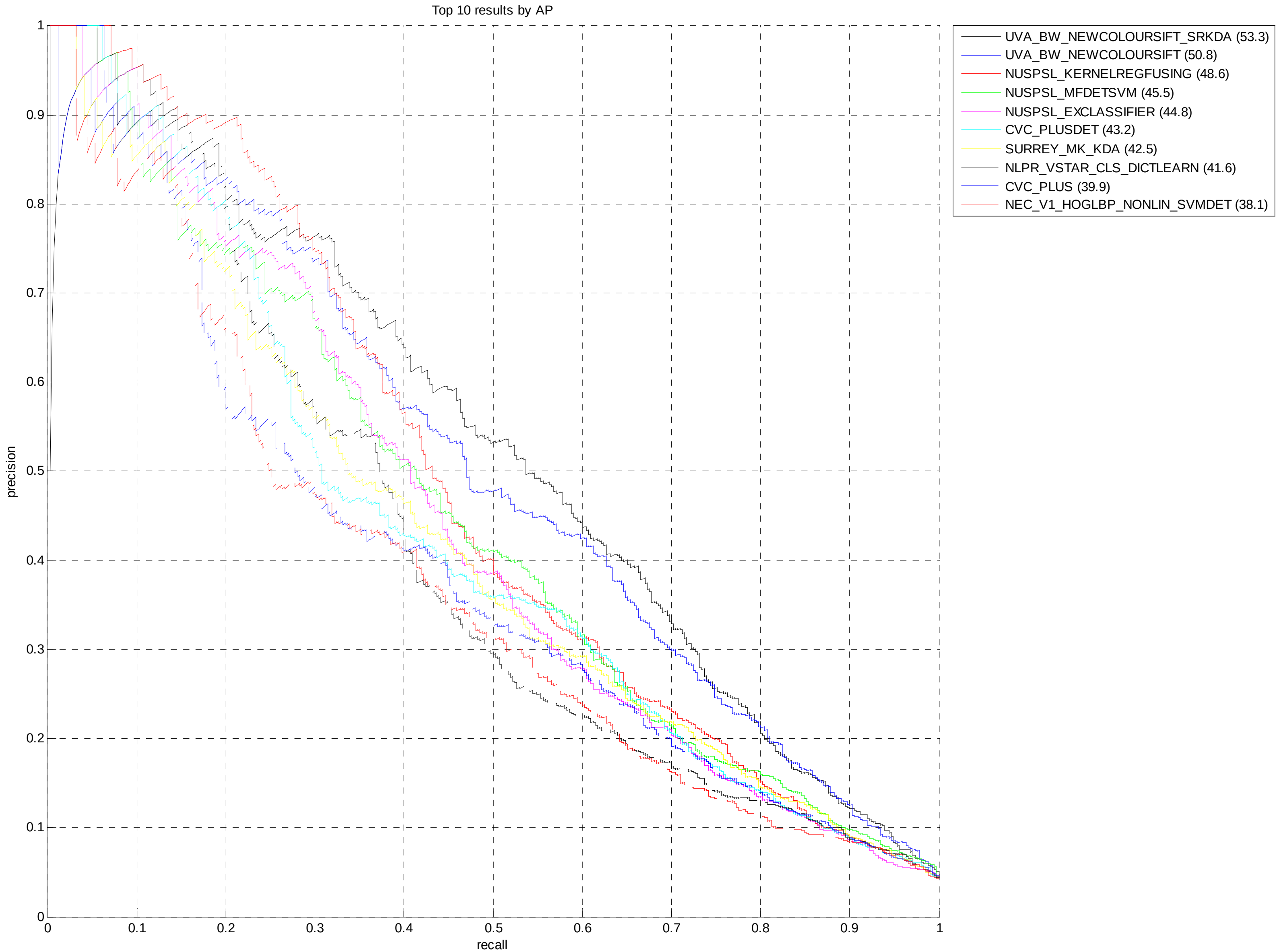
Precision/Recall: Bicycle (Top 10 by AP)



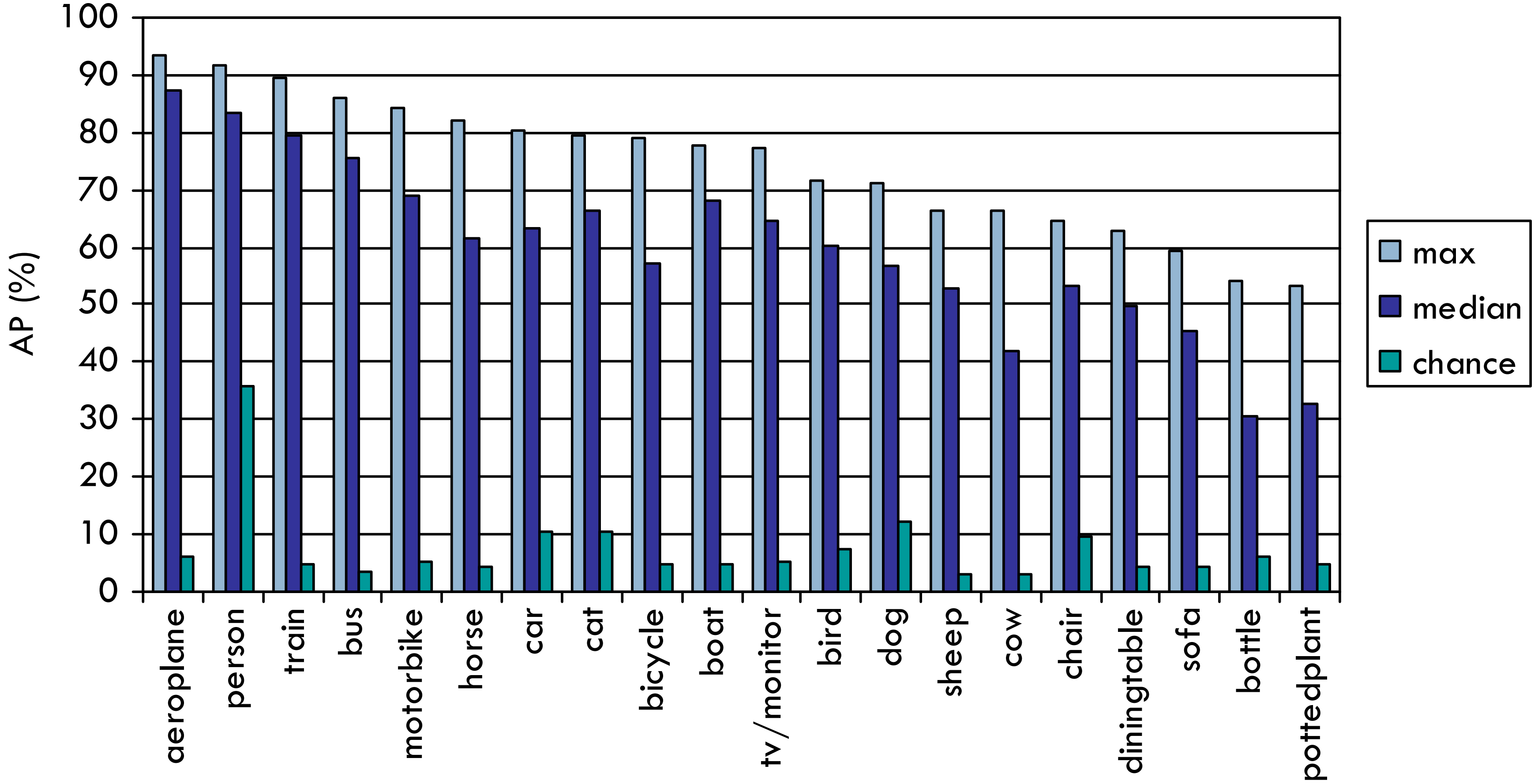
Precision/Recall: Potted plant (All)



Precision/Recall: Potted plant (Top 10 by AP)

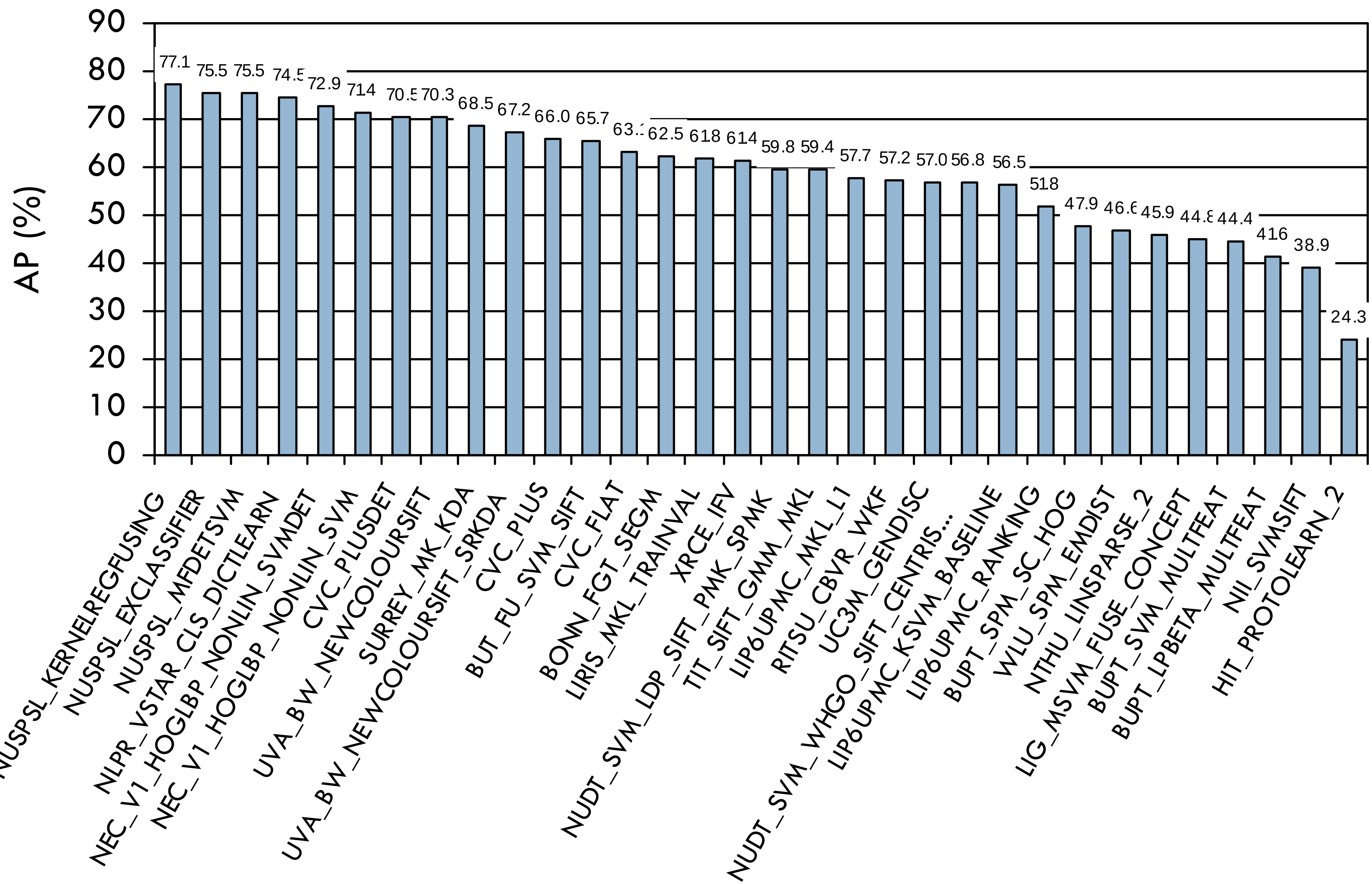


AP by Class



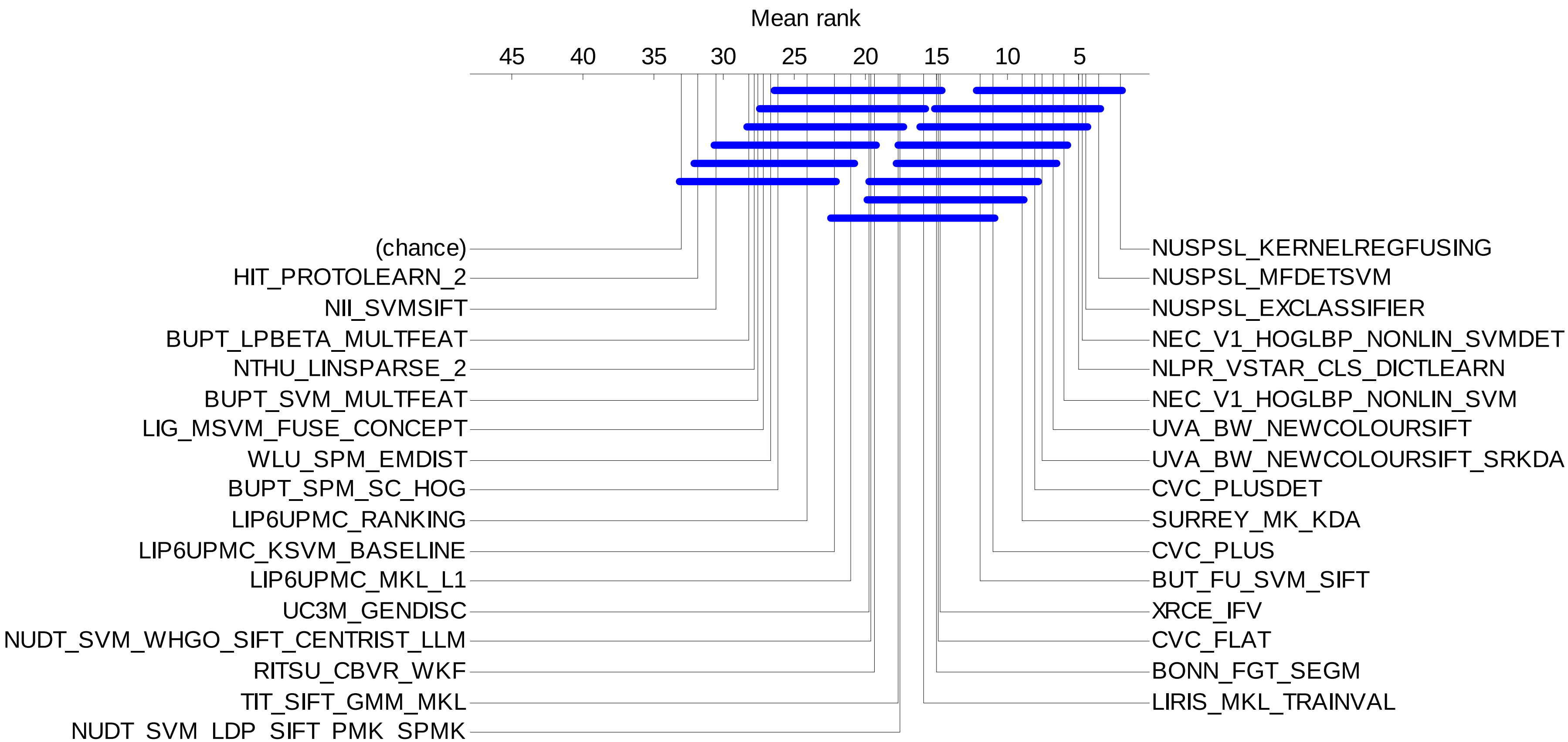
- Max AP: 93.3% (aeroplane) ... 53.3% (potted plant)

Median AP by Method



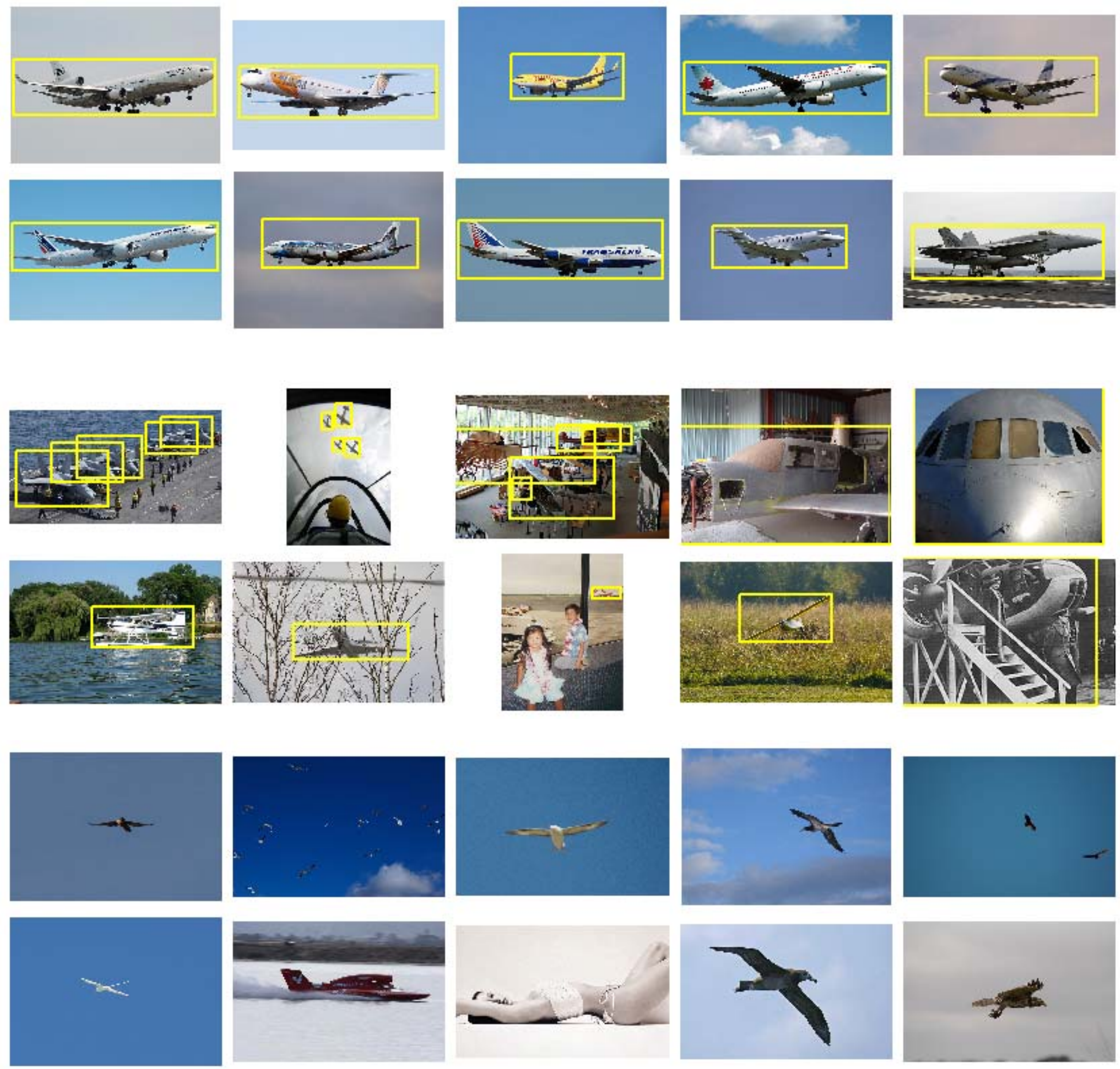
Statistical Significance

- Friedman/Nemenyi analysis
  - Compare differences in **mean rank** of methods over classes using non-parametric version of ANOVA
  - Mean rank must differ by at least 5.4 to be considered significant (p=0.05)



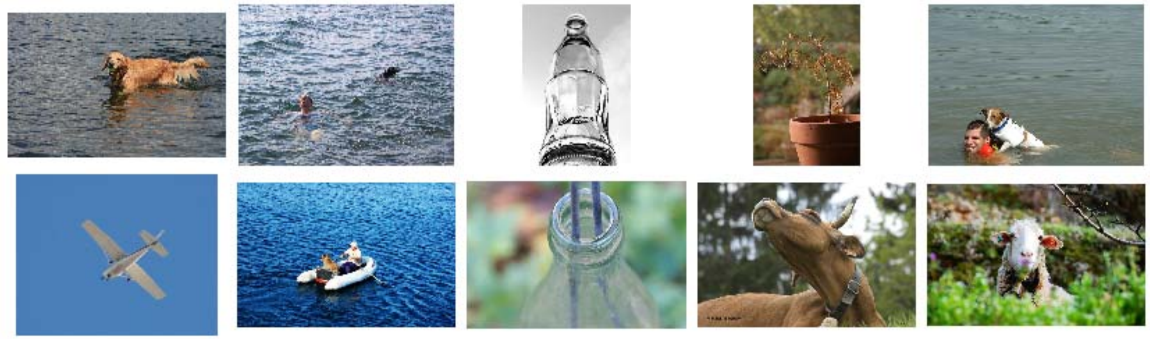
Ranked Images: Aeroplane

- Class images: Highest ranked
- Class images: Lowest ranked
- Non-class images: Highest ranked
- Context?



Non-Birds & Non-Boats

- Non-bird images: Highest ranked



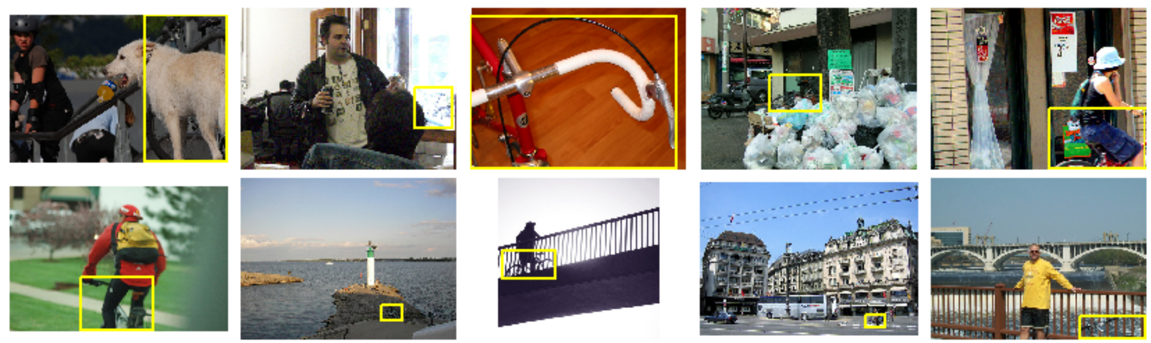
- Non-boat images: Highest ranked

- Water texture and scene composition?

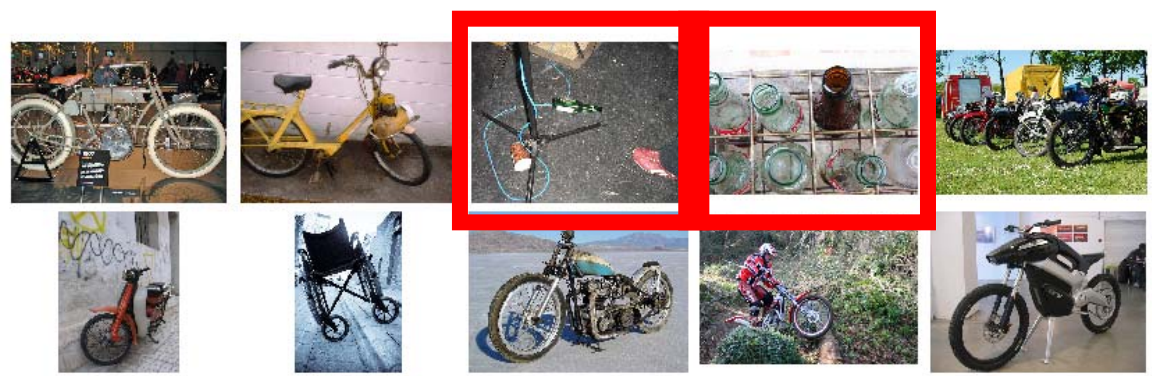
Ranked Images: Bicycle

- Class images: Highest ranked

- Class images: Lowest ranked



- Non-class images: Highest ranked



- “Texture”?

A Non-Bicycle Nemesis

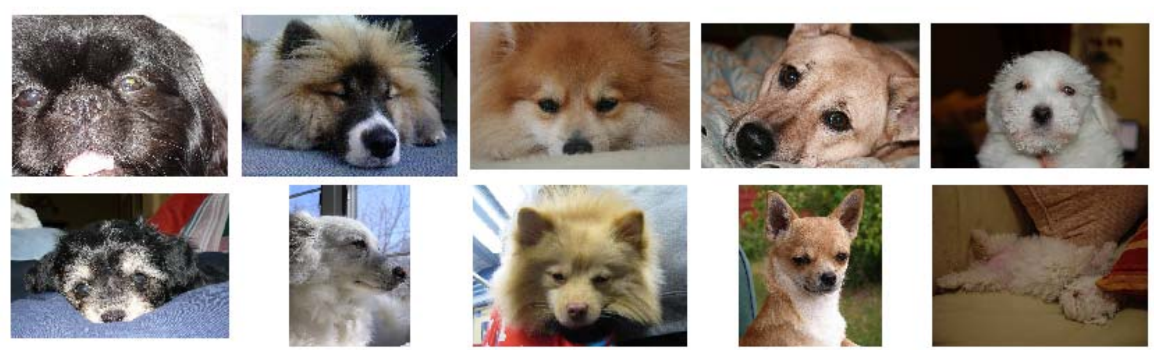
- 2008: Rank 1
- 2009: Rank 1
- 2010: Rank 3 (progress?)

Ranked Images: Cat

- Class images: Highest ranked

- Class images: Lowest ranked

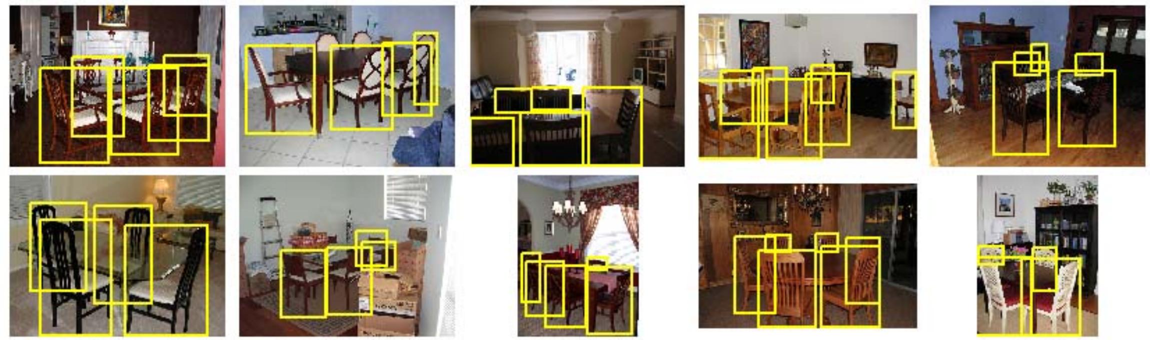
- Non-class images: Highest ranked



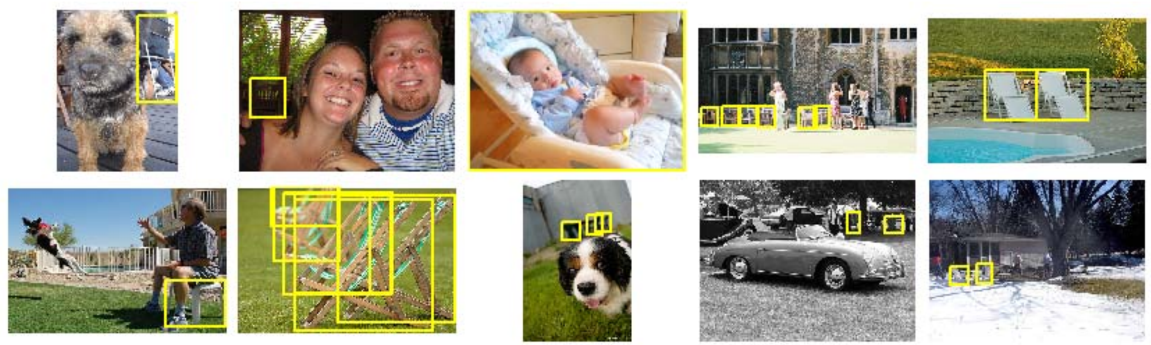
- "Composition"?

Ranked Images: Chair

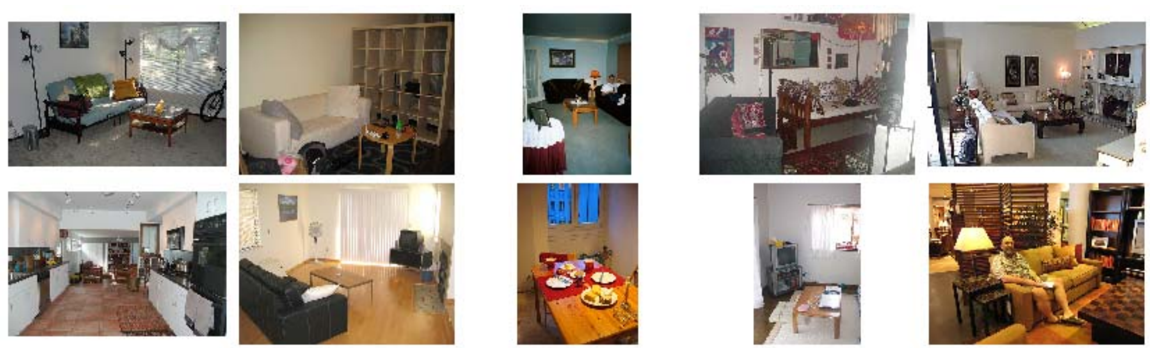
- Class images: Highest ranked



- Class images: Lowest ranked



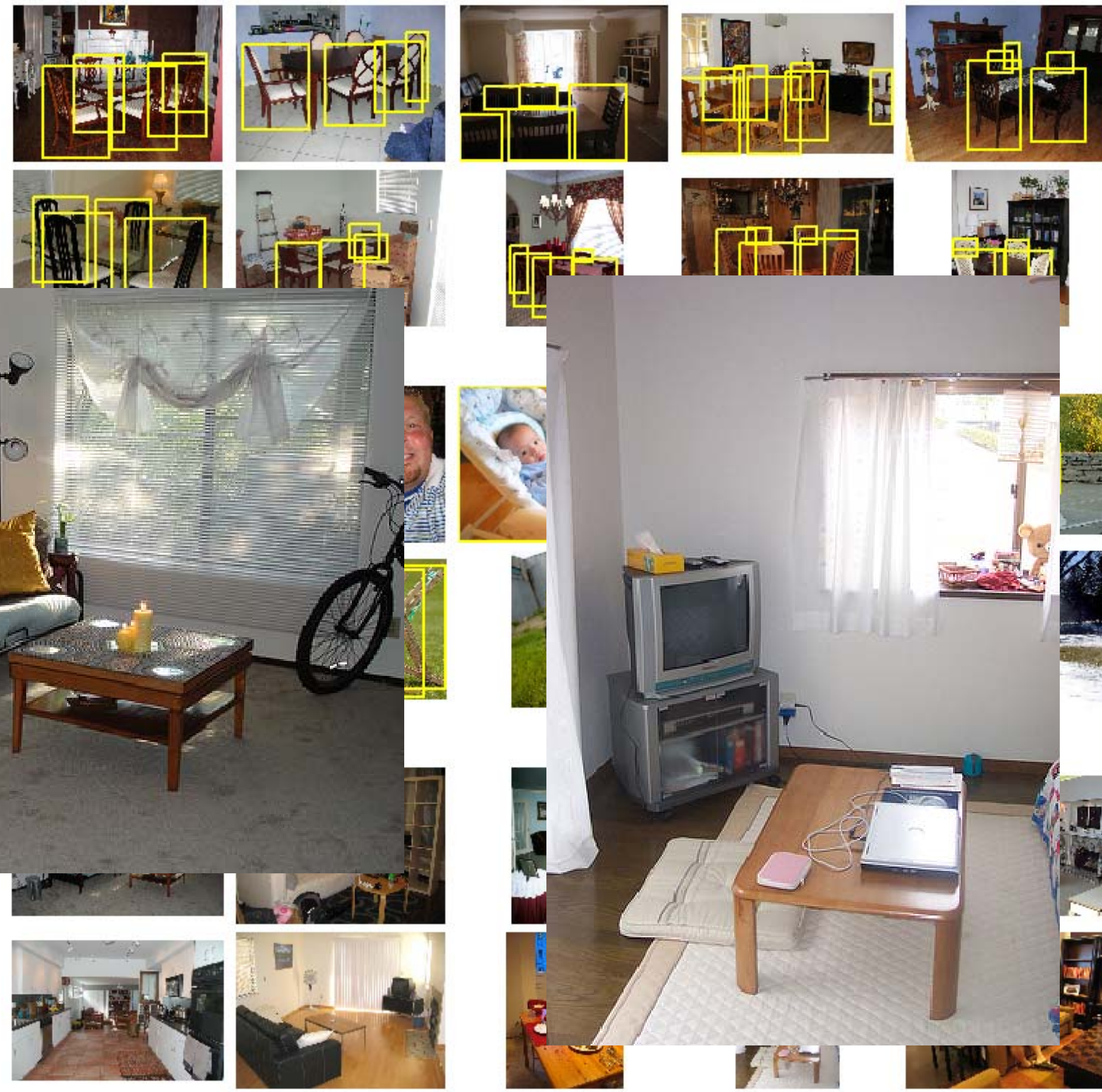
- Non-class images: Highest ranked



- Scene context?

Ranked Images: Chair

- Class images:
  Highest ranked

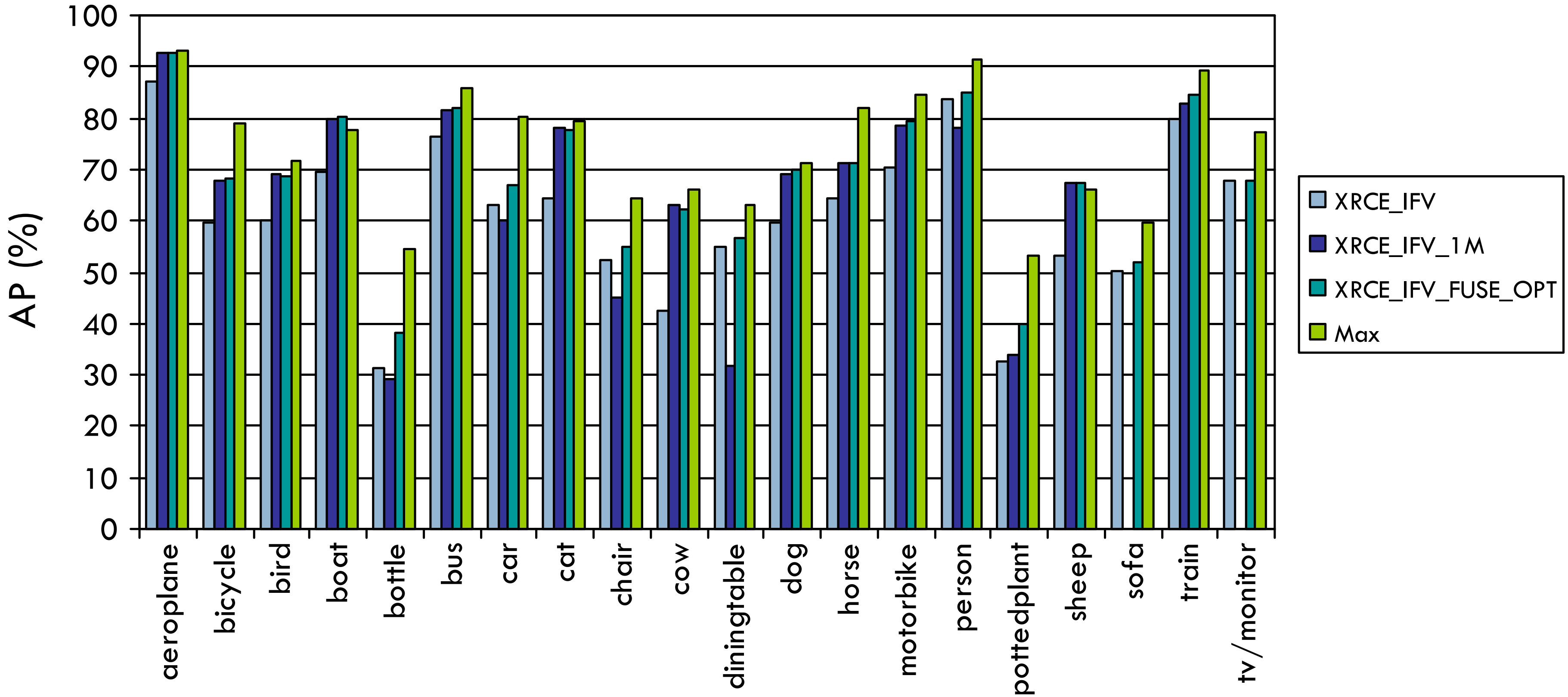


- Cl
  Lo

- N
  Highest ranked

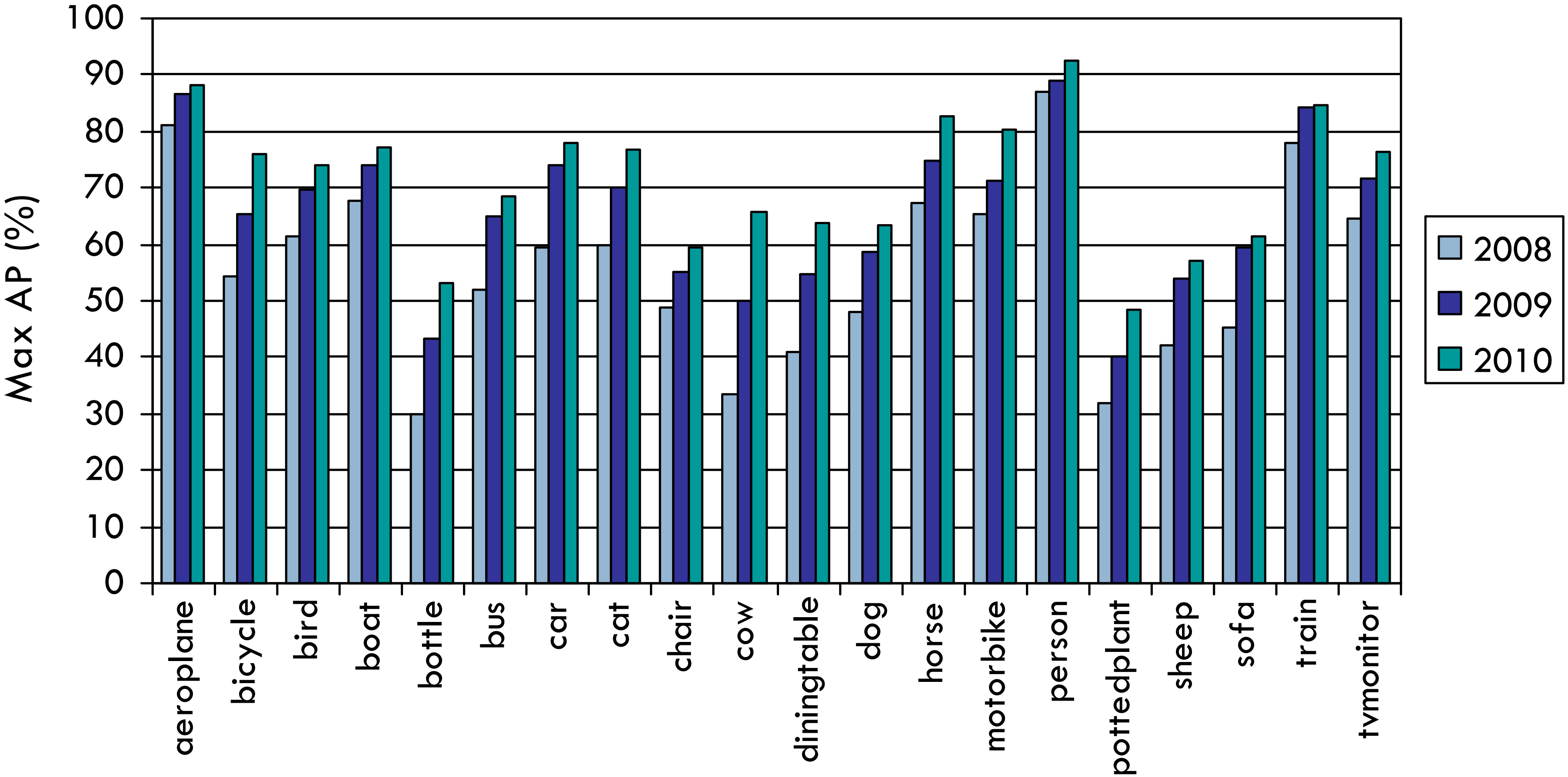
- Scene context?

Additional Training Data



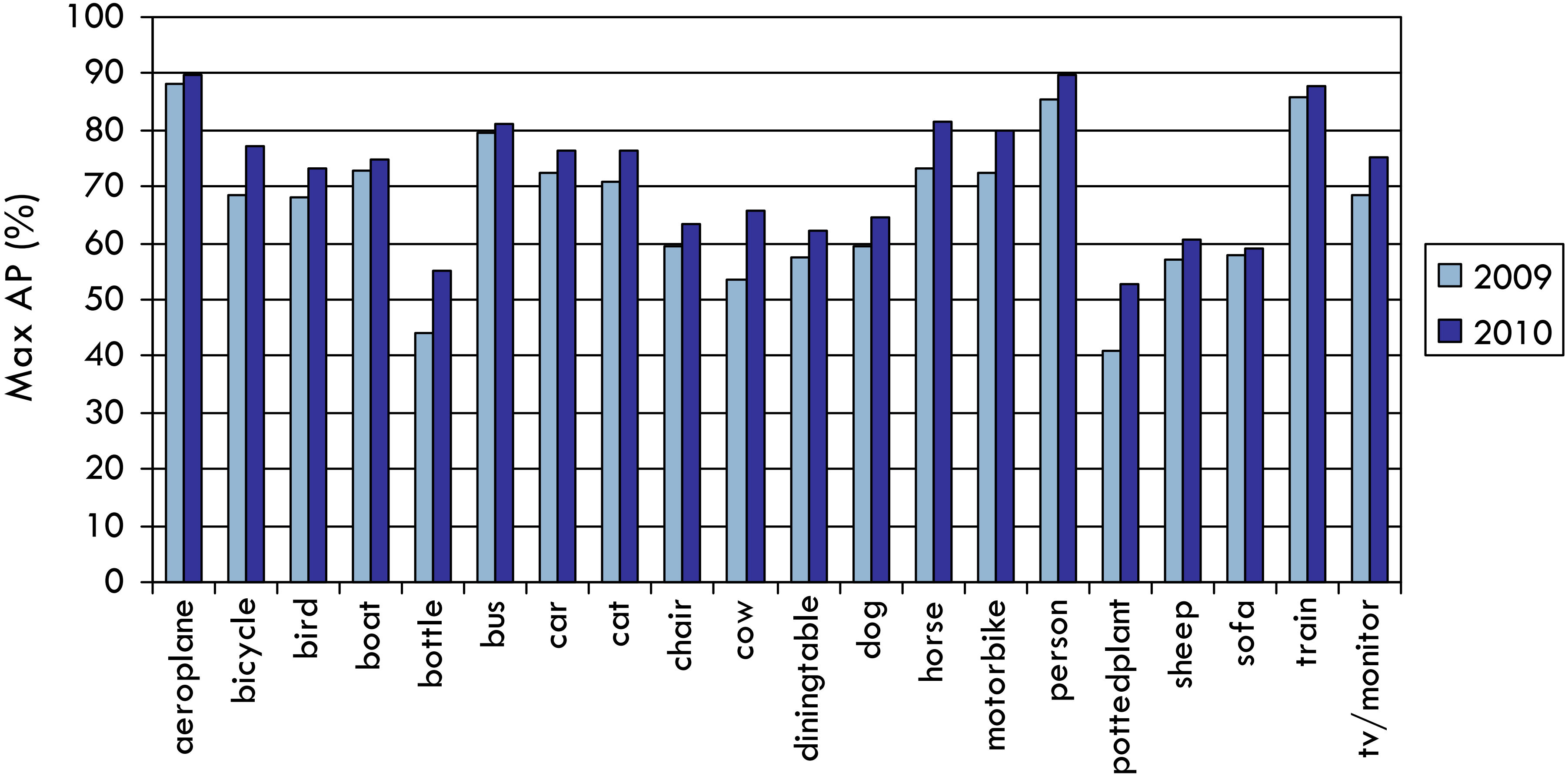
- XRCE collected around ~1M additional training images by finding a flickr group for each VOC category (no manual annotation)
- Results improve. For 2 classes (boat, sheep) better than best method trained only on VOC2010 data

Progress 2008-2010



- Results on 2008 data improve for best 2009 and 2010 methods for all categories, by over 100% for some categories
  - Caveat: Better methods or more training data?

Progress 2009-2010



- Best 2010 methods improve on 2009 results for all categories
  - Caveat: Better methods or more training data?

Prizes

- Winner:
  - **NUSPSL_KERNELREGFUSING**

Qiang Chen[1], Zheng Song[1], Si Liu[1], Xiangyu Chen[1], Tat-Seng Chua[1], Shuicheng Yan[1], Yang Hua[2], Zhongyang Huang[2], Shengmei Shen[2]
[1]*National University of Singapore*
[2]*Panasonic Singapore Laboratories*