Face detection without bells and whistles

Markus Mathias Rodrigo Benenson Marco Pedersoli Luc Van Gool

KU LEUVEN

Processing Speech & Images

mpii

10/09/2014

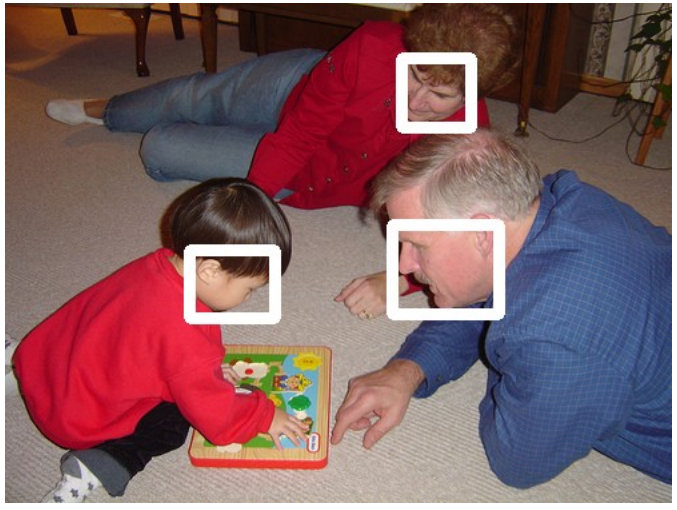
How to get a good face detector?

Main points

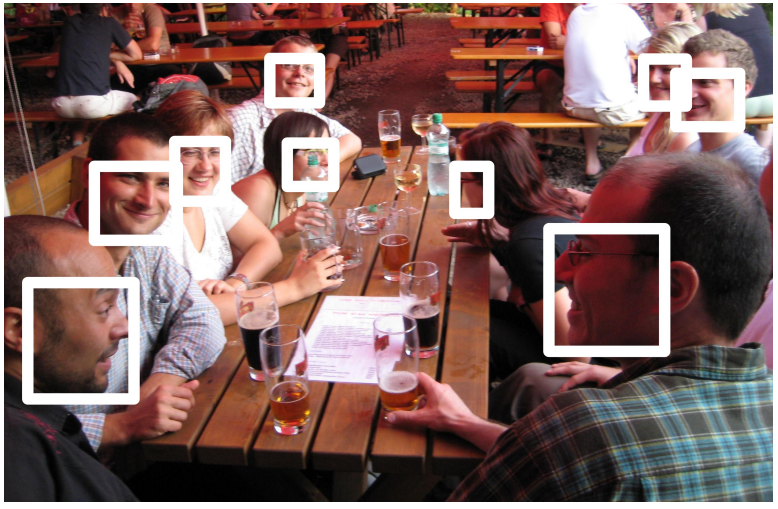
- Issues with existing face detection benchmarks
- Baseline methods can be surprisingly effective

Benchmark issues

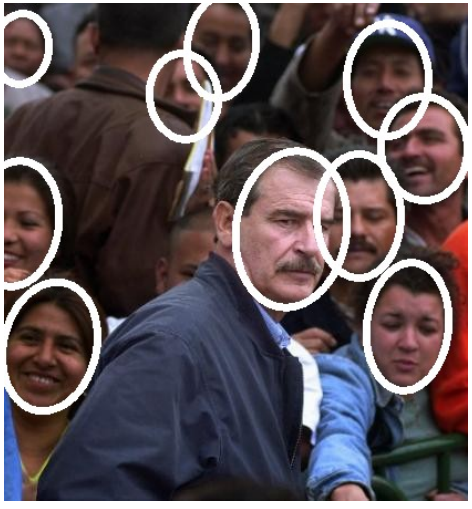
Most relevant detection benchmarks



Pascal Faces
[Yan et al. 2013]



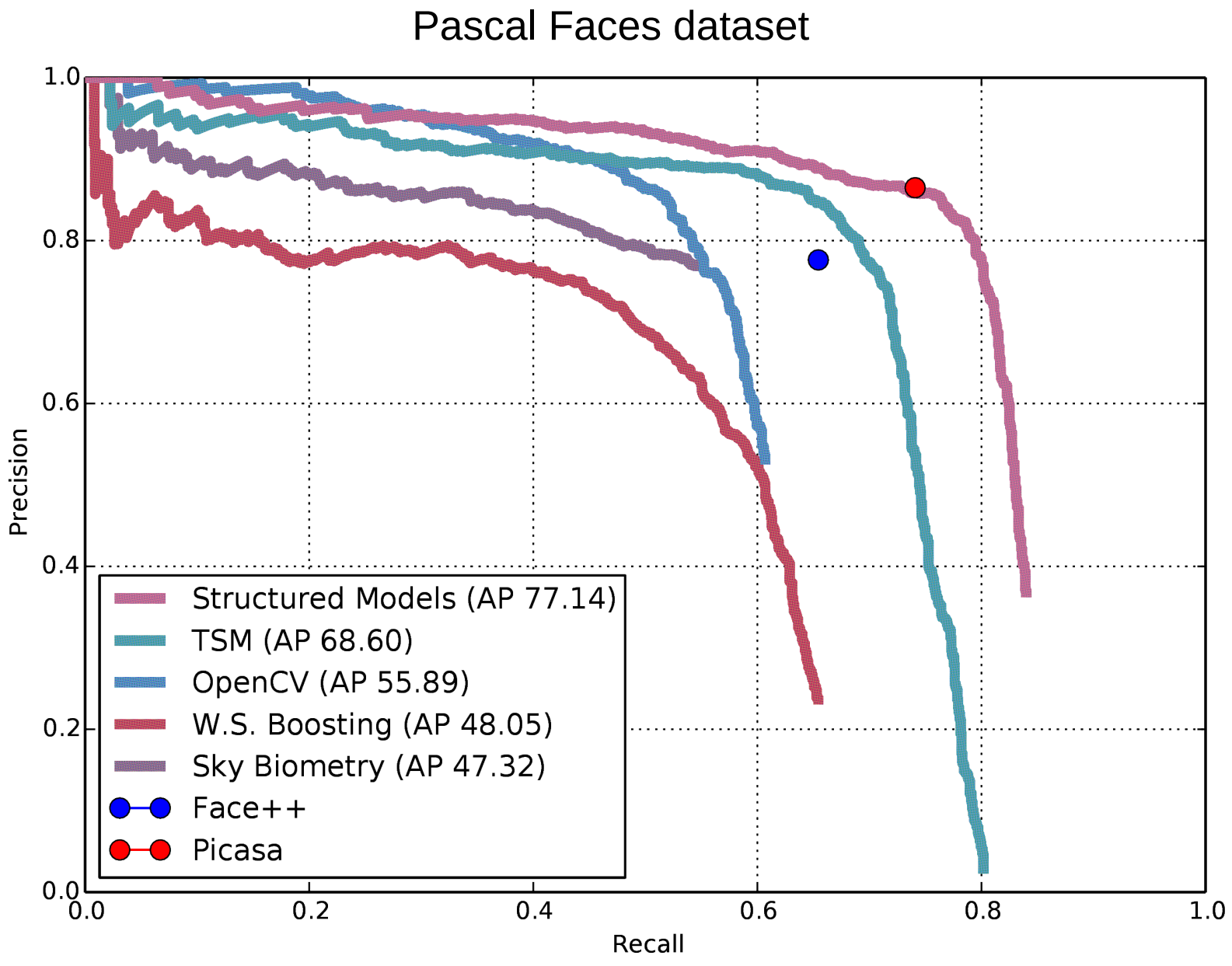
AFW
[Zhu and Ramanan CVPR 2012]



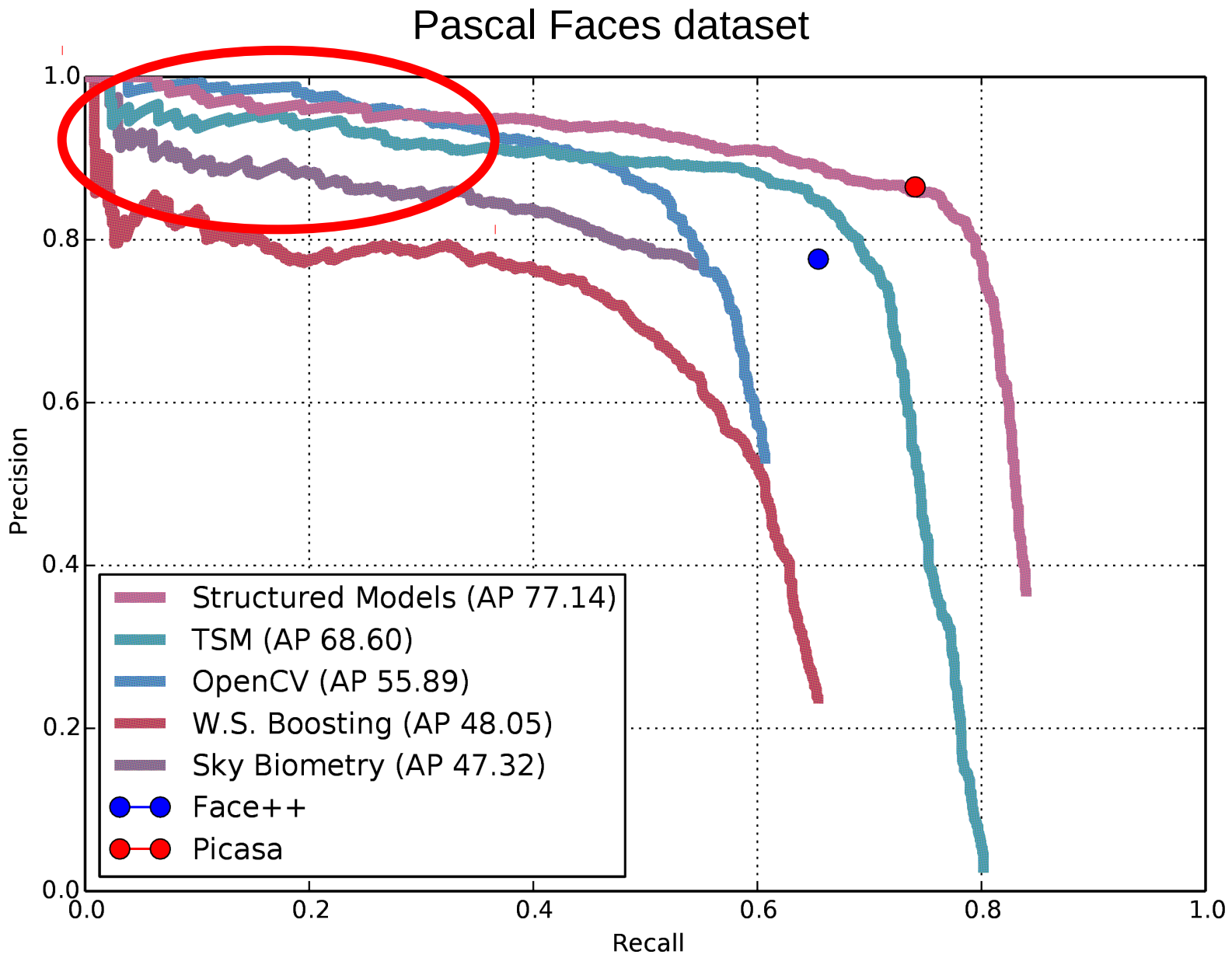
FDDB
[Vain et al. 2010]

- Comparison of many systems
- Both, research methods and commercial products

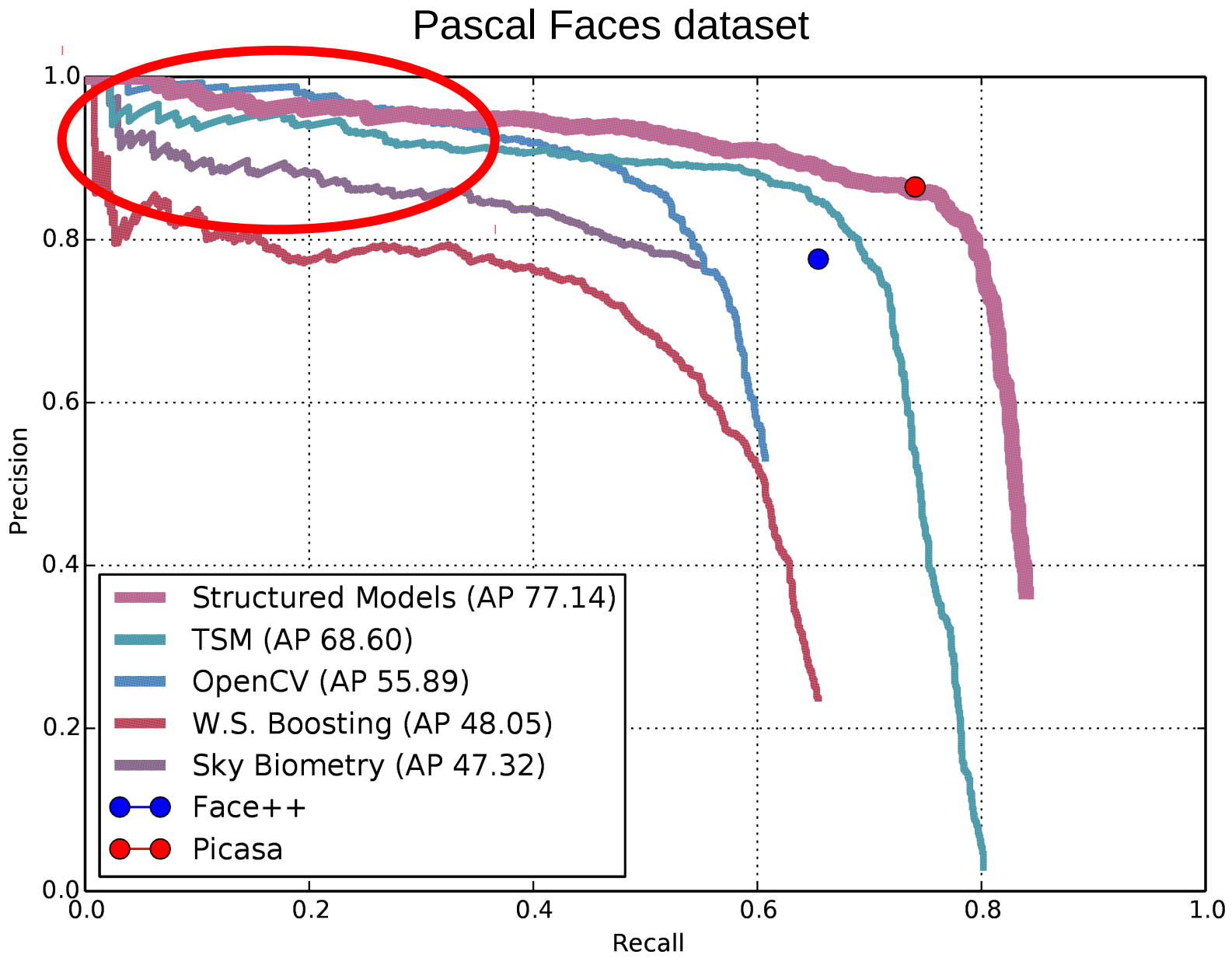
Suspicious curves



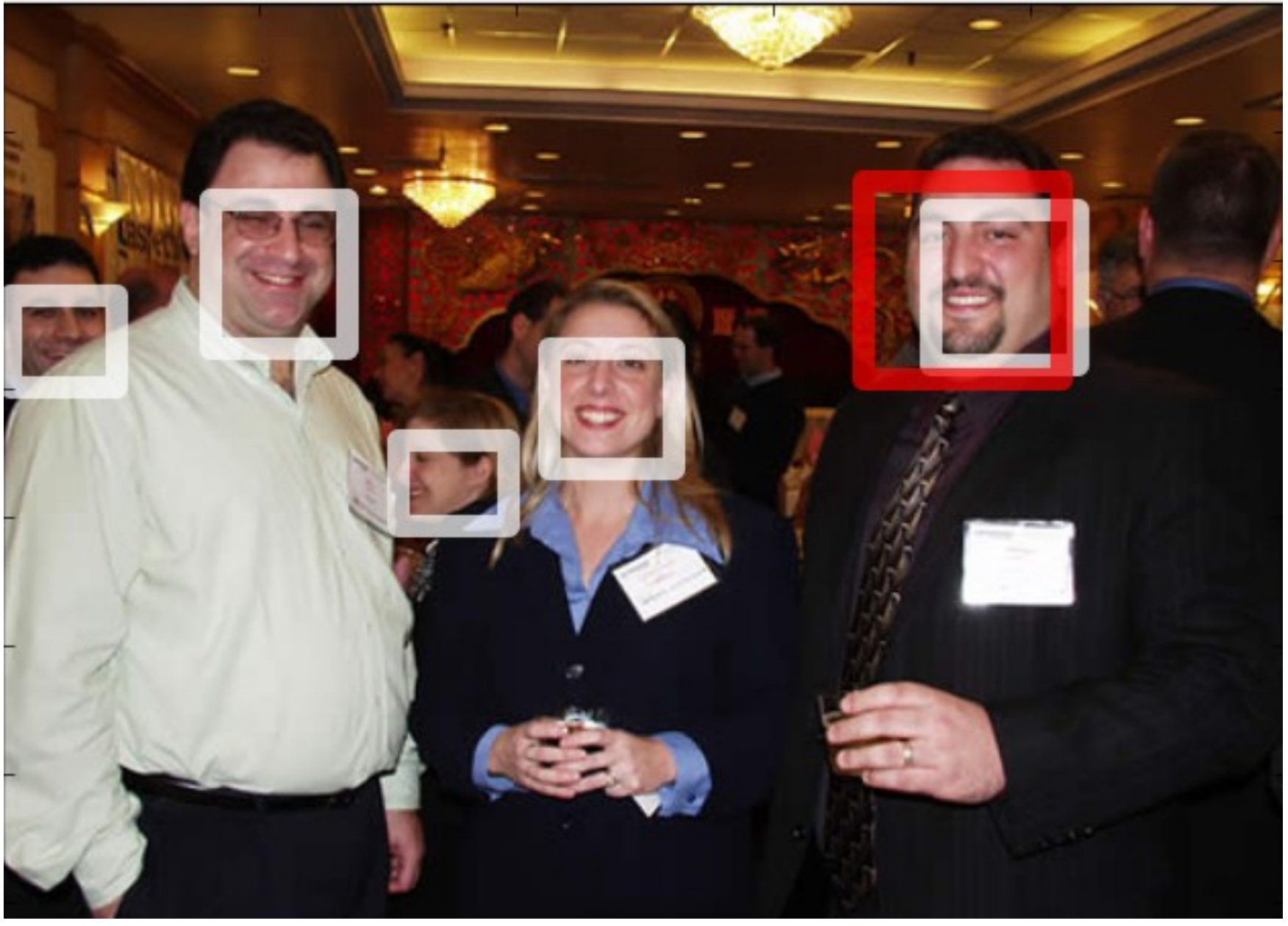
Suspicious curves



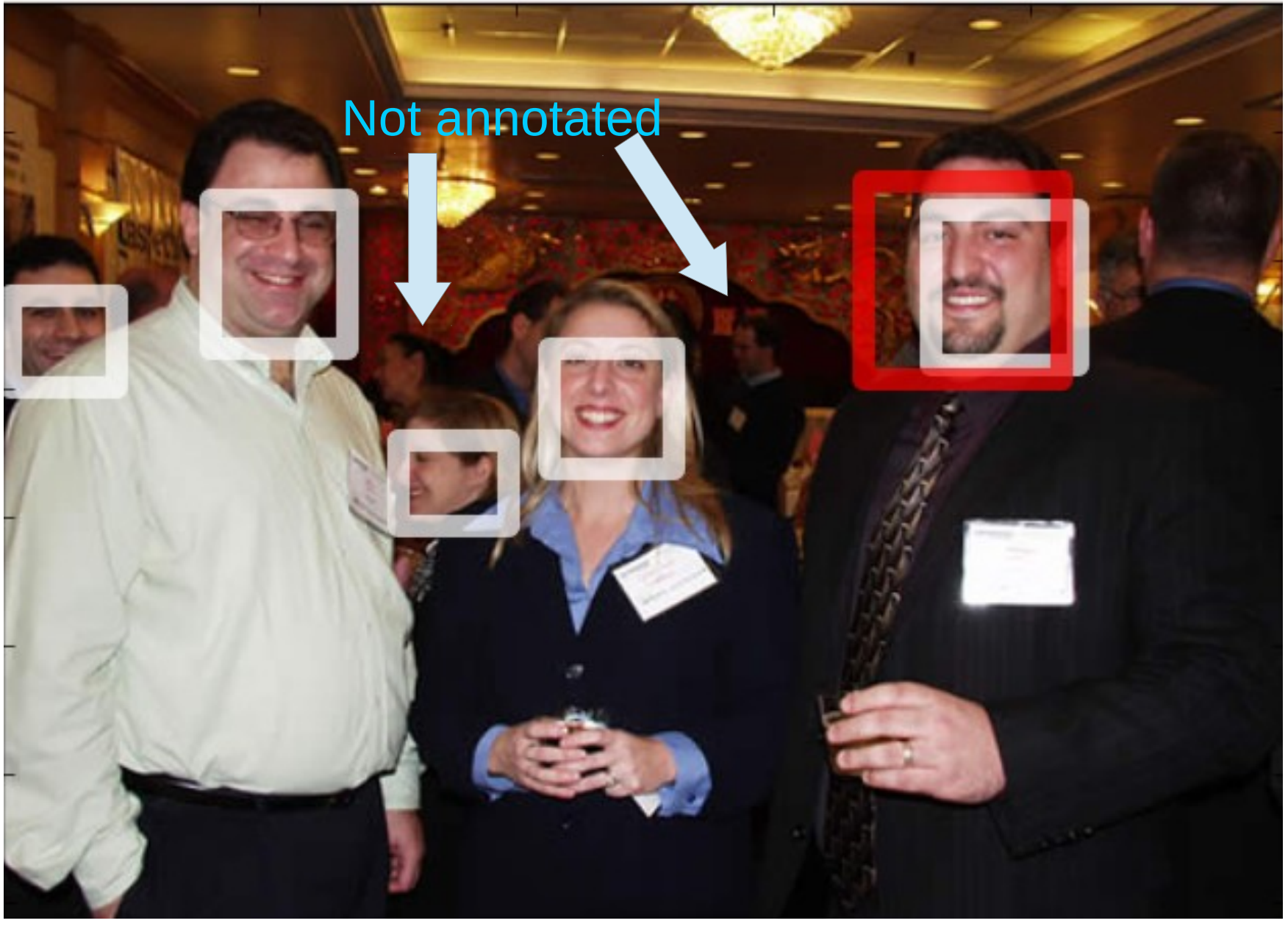
Suspicious curves



Top scoring false positives

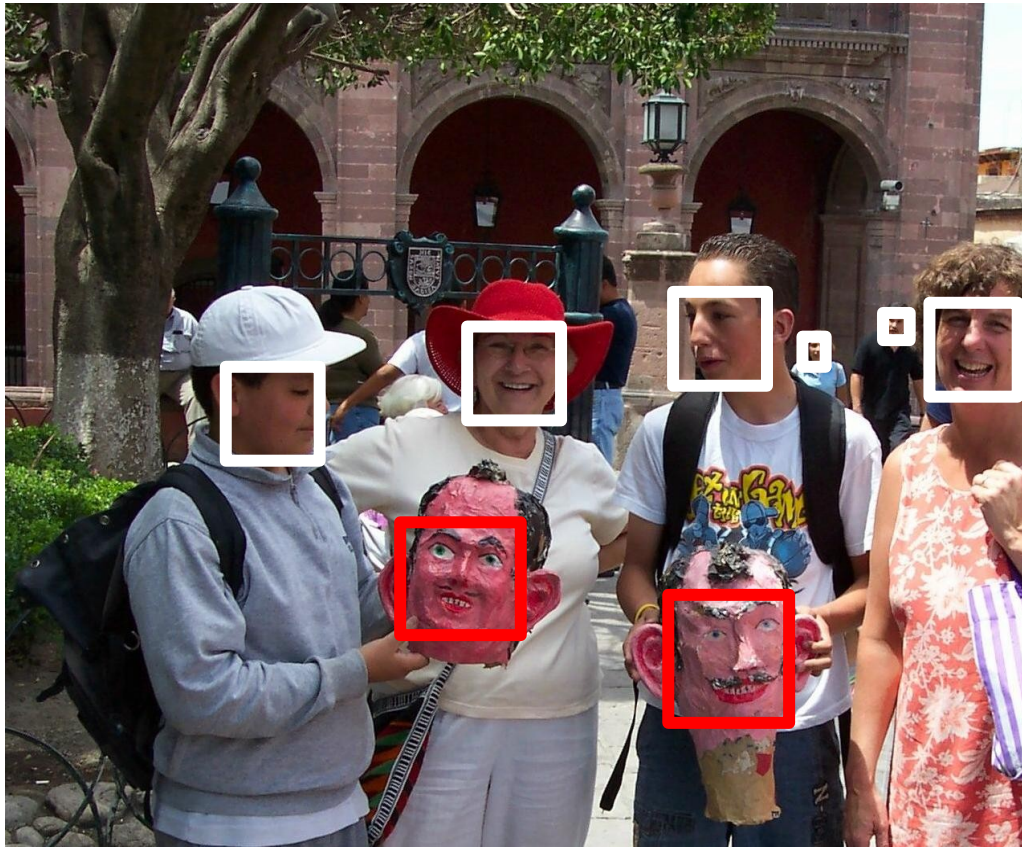


Top scoring false positives



Issues with the current benchmarks

- What constitutes a face?
  - What is the minimal annotated face size?
  - Which annotation policy is used?



Issues with the current benchmarks

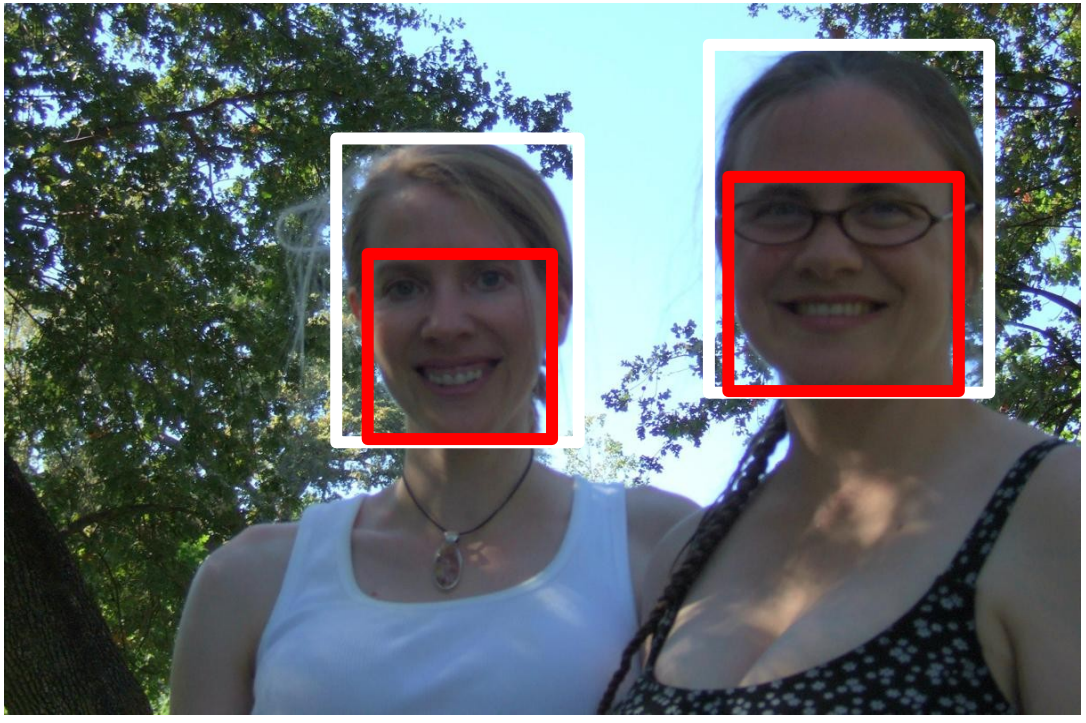
- What constitutes a face?
- What is the minimal annotated face size?
- Which annotation policy is used?

- Boundary effects have severe impact on overall detector quality

Issues with the current benchmarks

- What constitutes a face?
- What is the minimal annotated face size?
- Which annotation policy is used?

Within the dataset



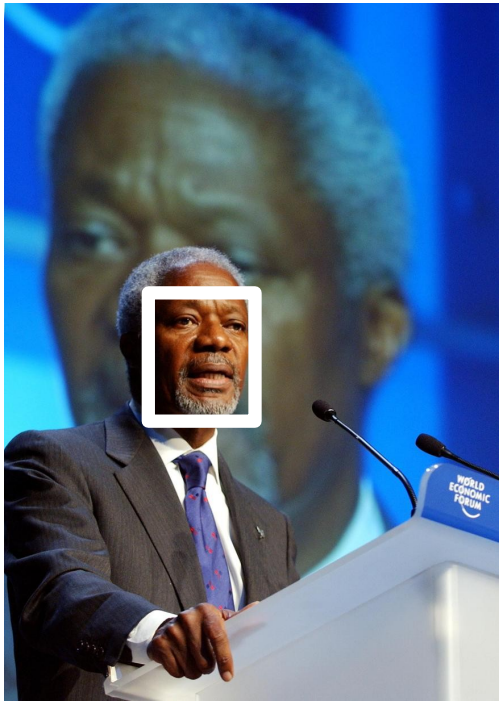
across datasets

Issues with the current benchmarks

We want a **fair** and **meaningful** comparison between methods

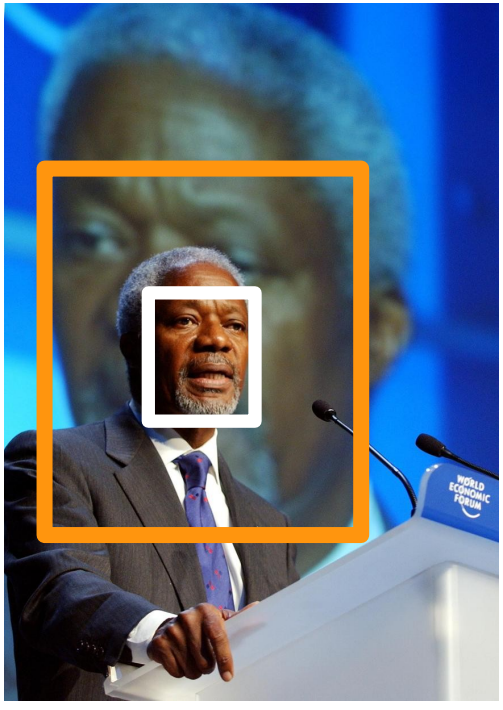
Solution: Improved annotations

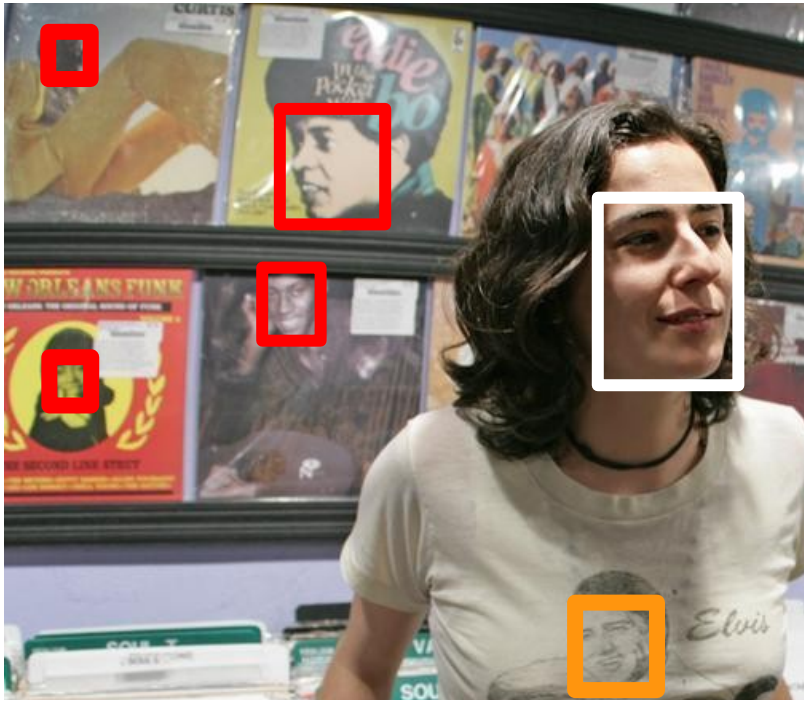
- We modified bounding boxes to ensure a consistent policy and minimal size

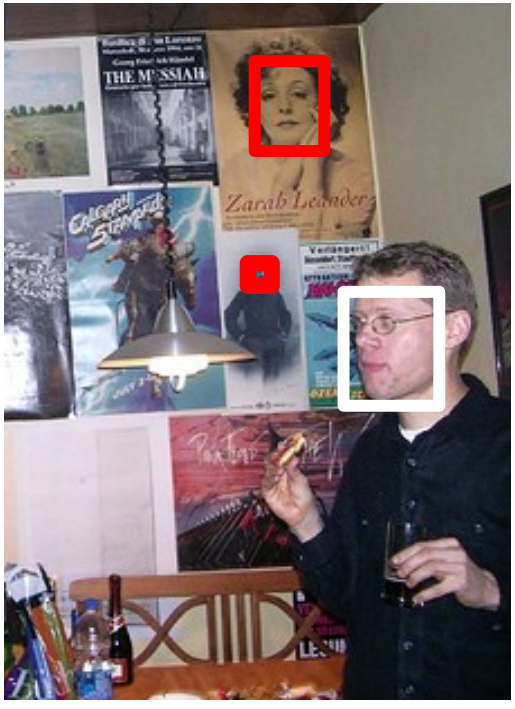


Solution: Improved annotations

- We modified bounding boxes to ensure a consistent policy and minimal size
- We add more bounding boxes (ignore labels when unclear)

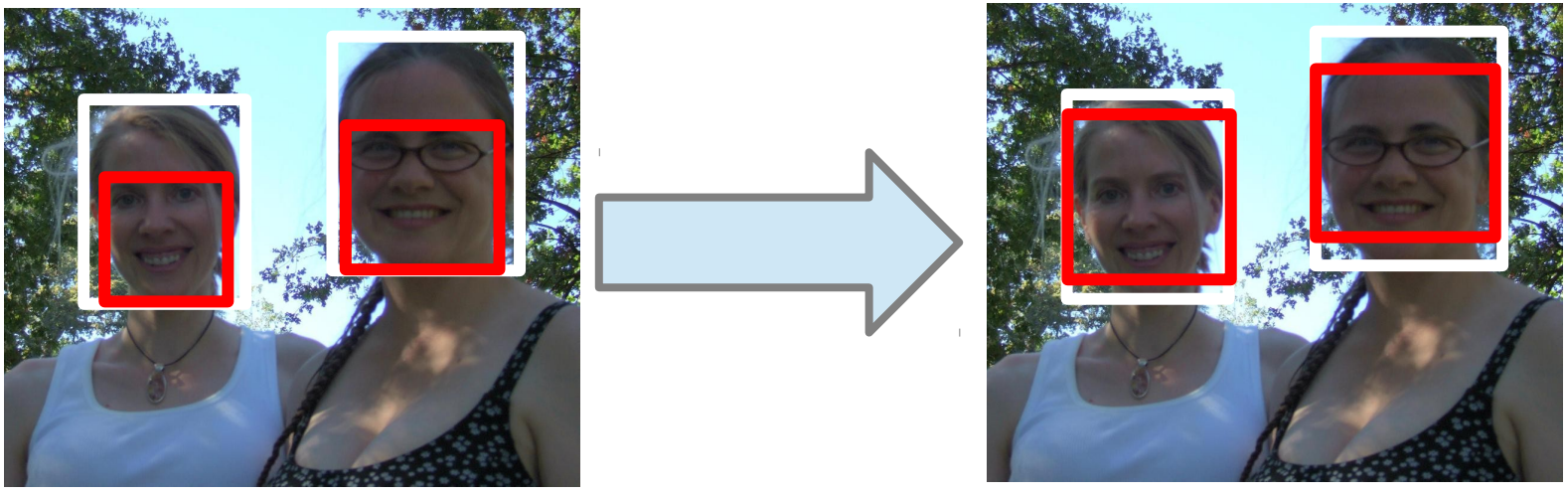






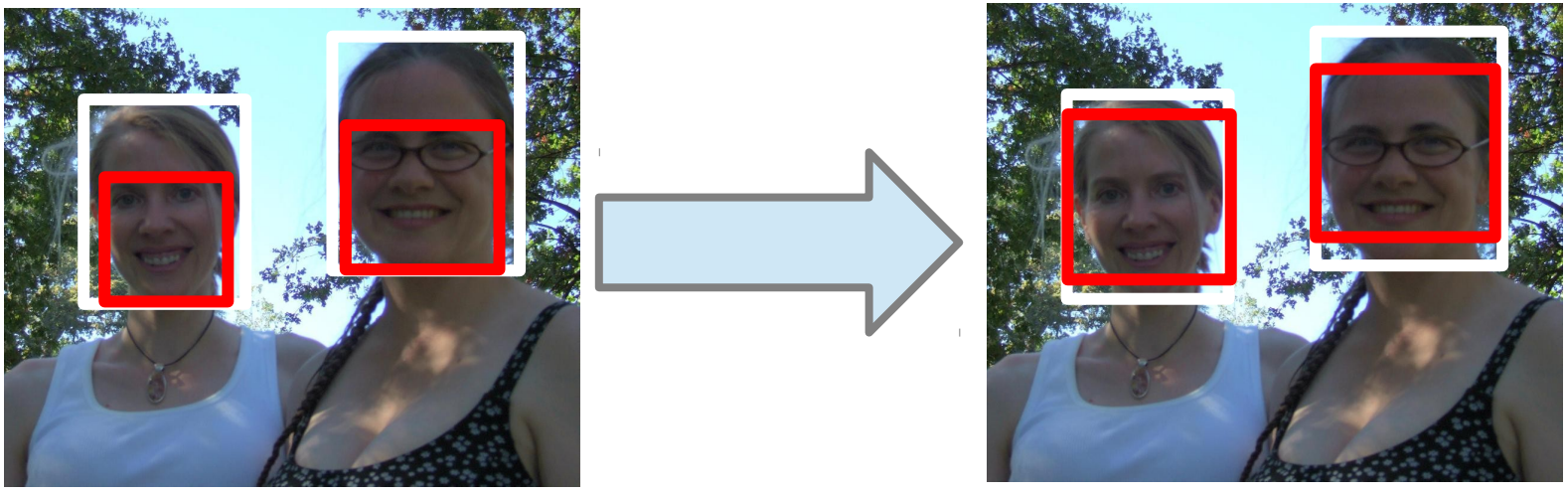
Solution: Handling different policies

- Estimate global rigid transform for each method
  - Translation and scaling
  - Maximize detection/annotation overlap



Solution: Handling different policies

- Estimate global rigid transform for each method
  - Translation and scaling
  - Maximize detection/annotation overlap

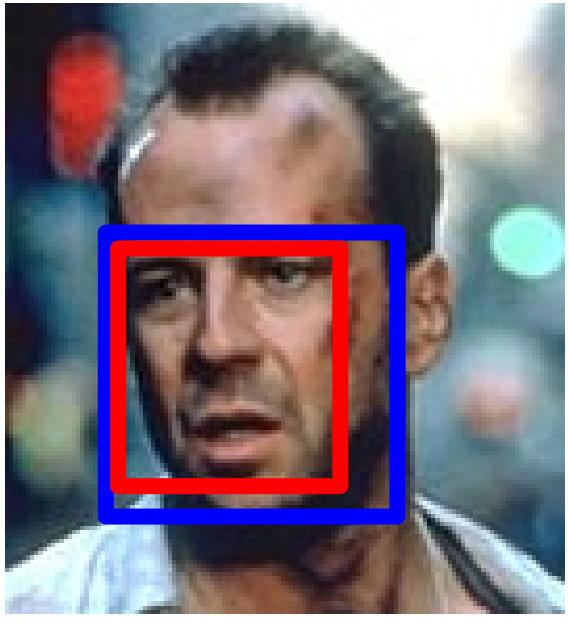


Bounding boxes adaptation applied for each method
⇨ Part of the evaluation protocol.
⇨ No advantage for any specific method.

Solution: Handling different scale ranges

- We want to evaluate the detectors for faces with α >= 30 pixel
- Assume annotation of faces >= 15 pixel

Annotation size = α
Detection size < α

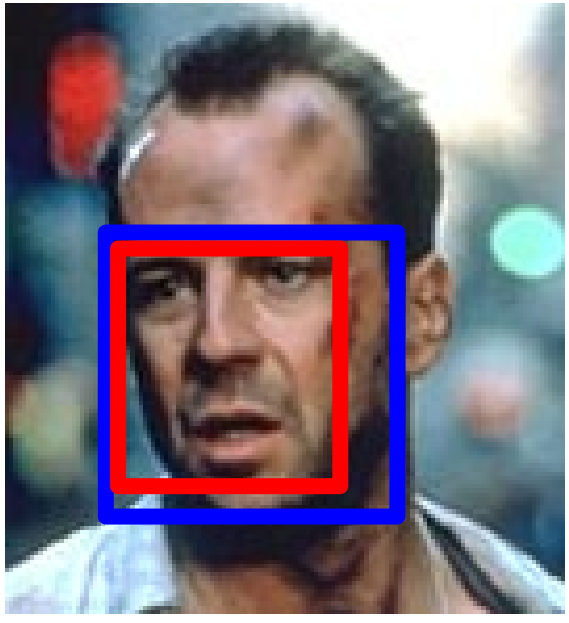


Case 1

Solution: Handling different scale ranges

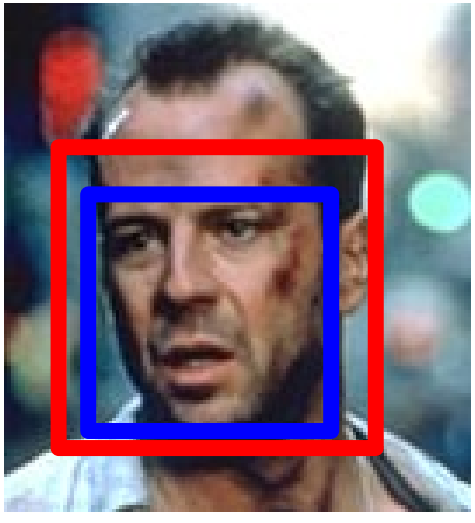
- We want to evaluate the detectors for faces with α >= 30 pixel
- Assume annotation of faces >= 15 pixel

Annotation size = α
Detection size < α



Case 1

Annotation size < α
Detection size = α

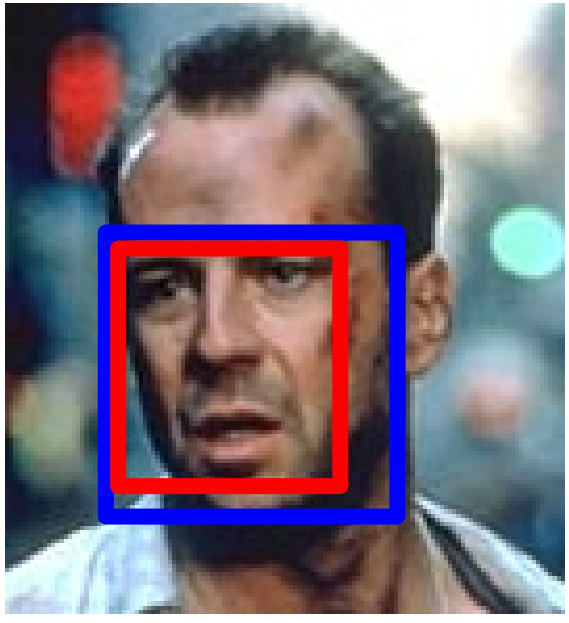


Case 2

Solution: Handling different scale ranges

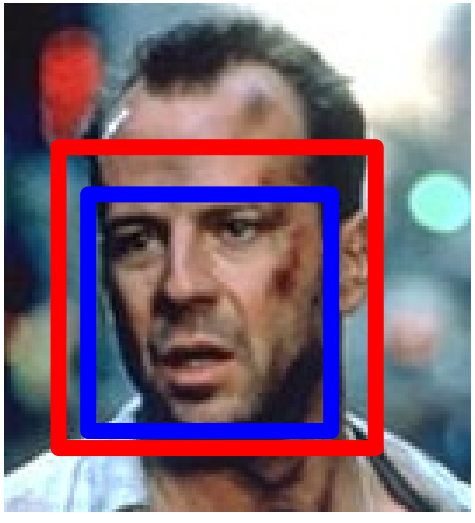
- We want to evaluate the detectors for faces with α >= 30 pixel
- Assume annotation of faces >= 15 pixel

Annotation size = α
Detection size < α



Case 1

Annotation size < α
Detection size = α



Case 2

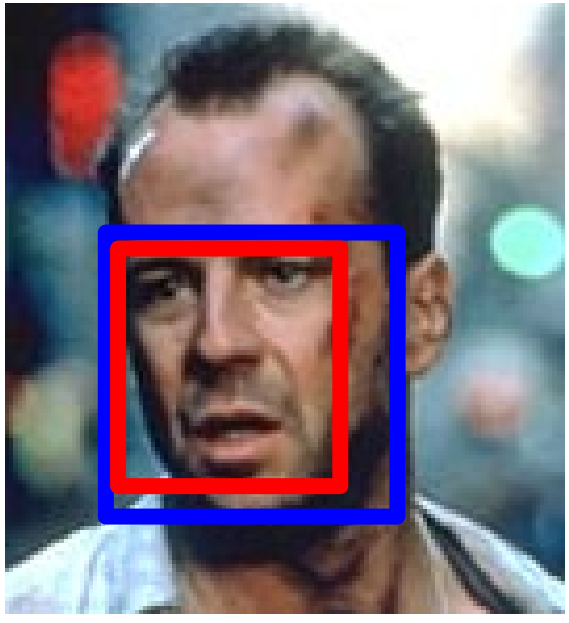
Detection < 15 pixel

Case 3

Solution: Handling different scale ranges

- We want to evaluate the detectors for faces with α >= 30 pixel
- Assume annotation of faces >= 15 pixel

Annotation size = α
Detection size < α



Annotation size < α
Detection size = α

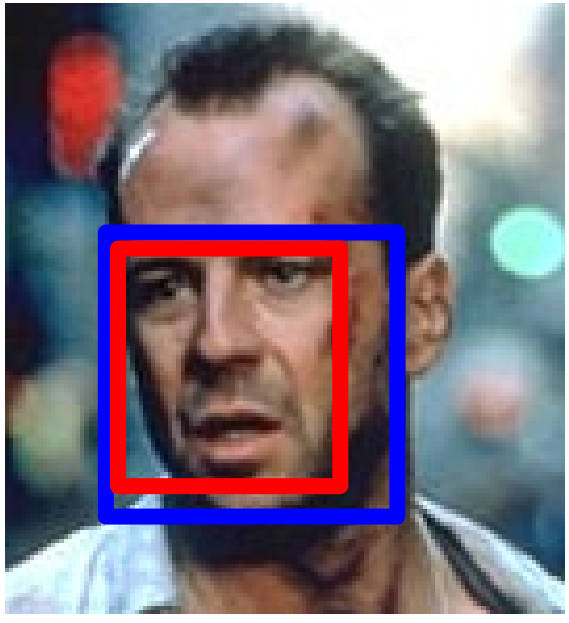
Detection < 15 pixel

Idea 1: Delete annotations smaller than α

Solution: Handling different scale ranges

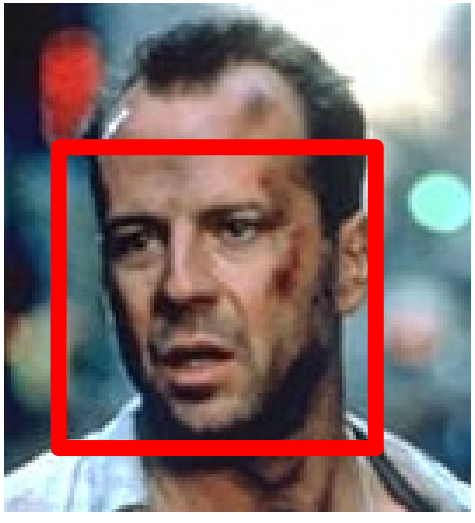
- We want to evaluate the detectors for faces with α >= 30 pixel
- Assume annotation of faces >= 15 pixel

Annotation size = α
Detection size < α



✓

Annotation size < α
Detection size = α



X (fp)

Detection < 15 pixel

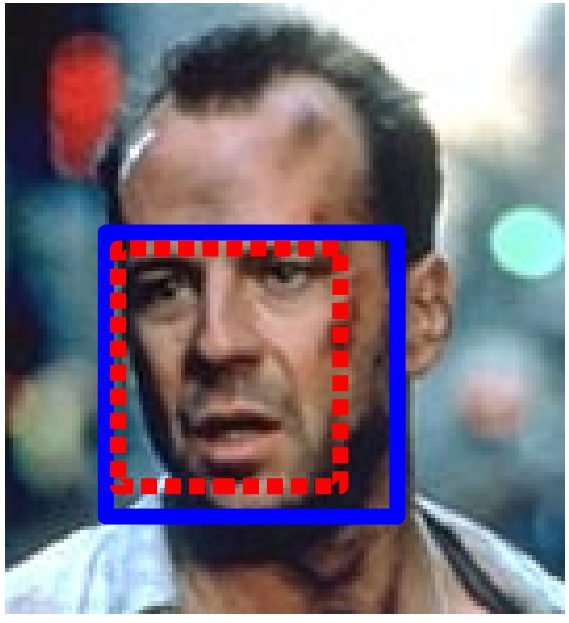
X (fp)

Idea 1: Delete annotations smaller than α

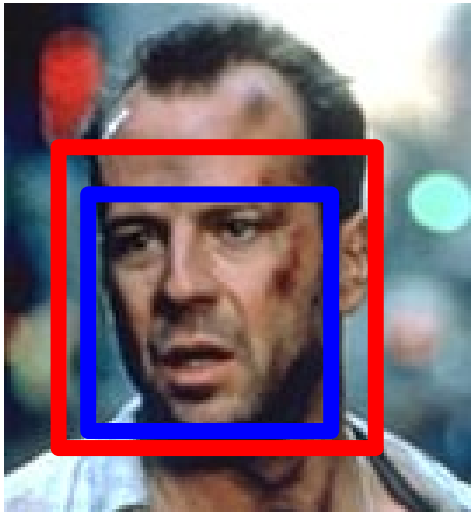
Solution: Handling different scale ranges

- We want to evaluate the detectors for faces with α >= 30 pixel
- Assume annotation of faces >= 15 pixel

Annotation size = α
Detection size < α



Annotation size < α
Detection size = α



Detection < 15 pixel

Idea 2: Delete detections smaller than α

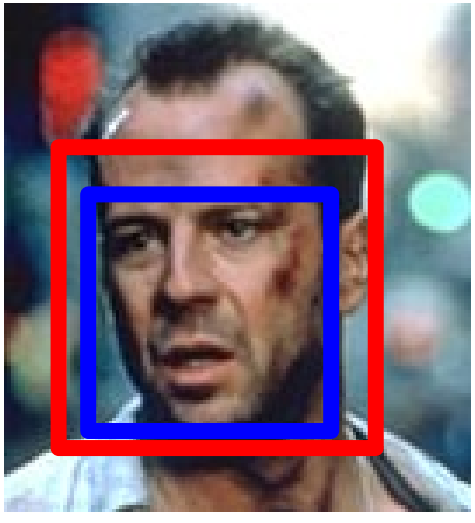
Solution: Handling different scale ranges

- We want to evaluate the detectors for faces with α >= 30 pixel
- Assume annotation of faces >= 15 pixel

Annotation size = α
Detection size < α

X (fn)

Annotation size < α
Detection size = α



✓

Detection < 15 pixel

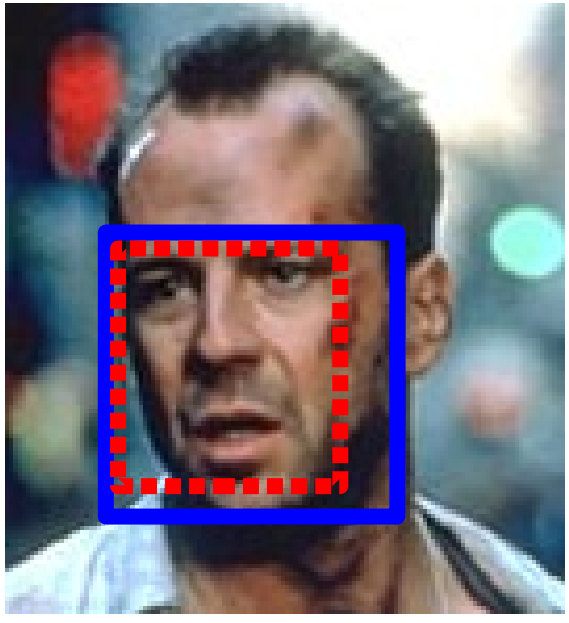
✓

Idea 2: Delete detections smaller than α

Solution: Handling different scale ranges

- We want to evaluate the detectors for faces with α >= 30 pixel
- Assume annotation of faces >= 15 pixel

Annotation size = α
Detection size < α



Annotation size < α
Detection size = α

Detection < 15 pixel

Idea 3: Delete annotations and detections smaller than α

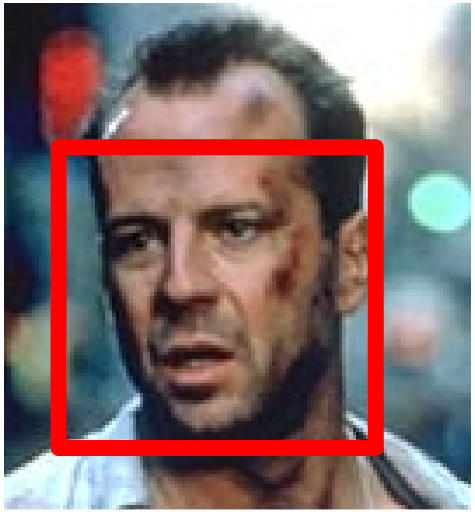
Solution: Handling different scale ranges

- We want to evaluate the detectors for faces with α >= 30 pixel
- Assume annotation of faces >= 15 pixel

Annotation size = α
Detection size < α

X (fn)

Annotation size < α
Detection size = α



X (fp)

Detection < 15 pixel

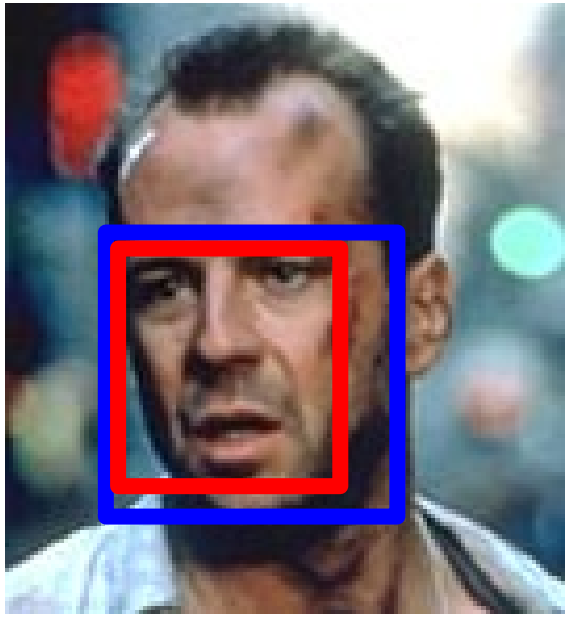
✓

Idea 3: Delete annotations and detections smaller than α

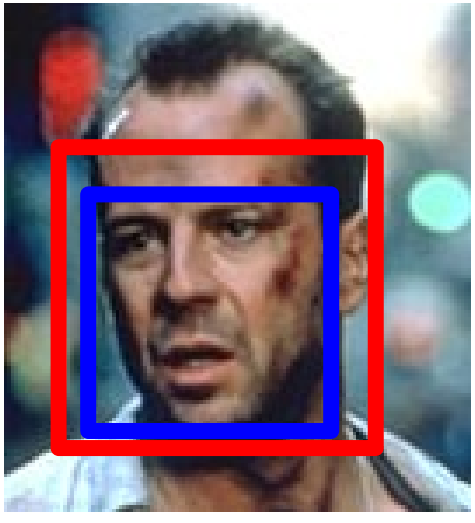
Solution: Handling different scale ranges

- Our solution:
  - Flag annotations < α = 30 pixel with “ignore” label
  - Delete detections <β, set $\beta = \sqrt{0.5 * \alpha^2}$

Annotation size = α
Detection size < α



Annotation size < α
Detection size = α

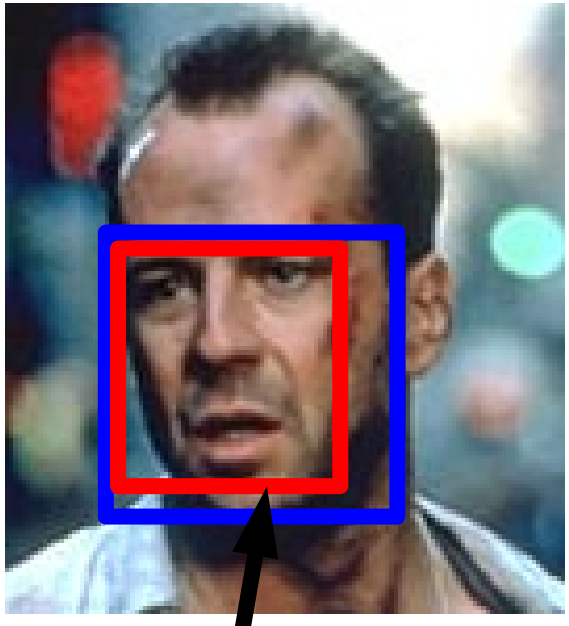


Detection < β = 21px

Solution: Handling different scale ranges

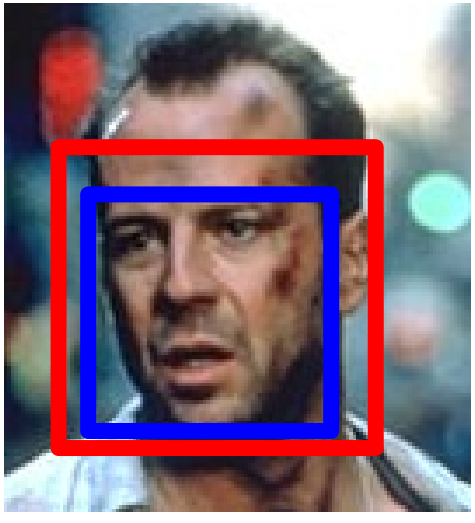
- Our solution:
  - Flag annotations < α = 30 pixel with “ignore” label
  - Delete detections <β, set $\beta = \sqrt{0.5 * \alpha^2}$

Annotation size = α
Detection size < α



Keep detection > β = 21px

Annotation size < α
Detection size = α

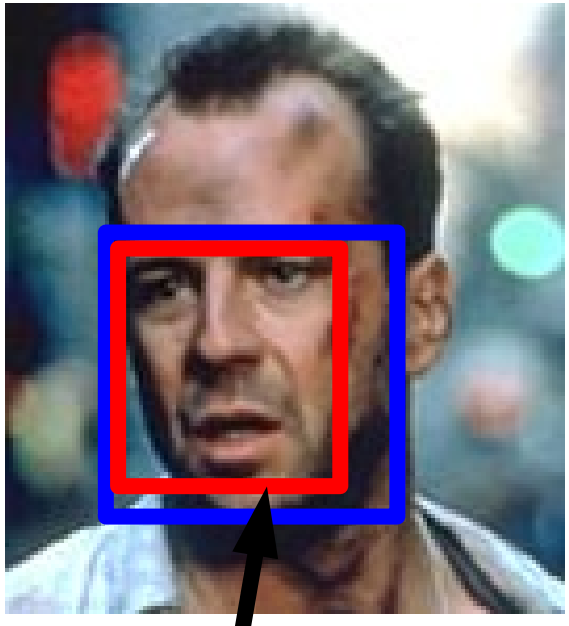


Detection < β = 21px

Solution: Handling different scale ranges

- Our solution:
  - Flag annotations < α = 30 pixel with "ignore" label
  - Delete detections <β, set $\beta = \sqrt{0.5 * \alpha^2}$

Annotation size = α
Detection size < α



Keep detection > β = 21px

Annotation size < α
Detection size = α

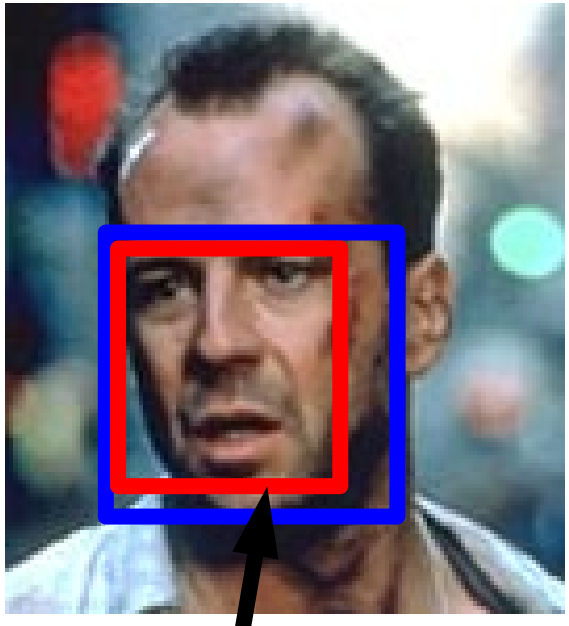
"Ignore" annotation

Detection < β = 21px

Solution: Handling different scale ranges

- Our solution:
  - Flag annotations < α = 30 pixel with "ignore" label
  - Delete detections <β, set $\beta = \sqrt{0.5 * \alpha^2}$

Annotation size = α
Detection size < α



Keep detection > β = 21px

Annotation size < α
Detection size = α

"Ignore" annotation

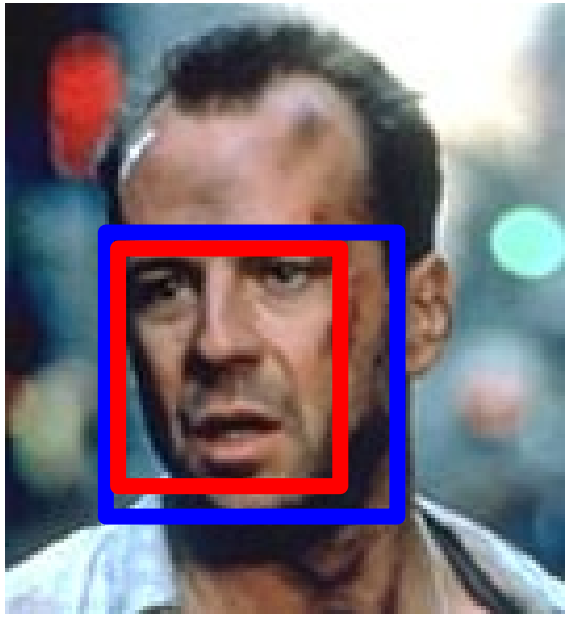
Detection < β = 21px

Delete detection

Solution: Handling different scale ranges

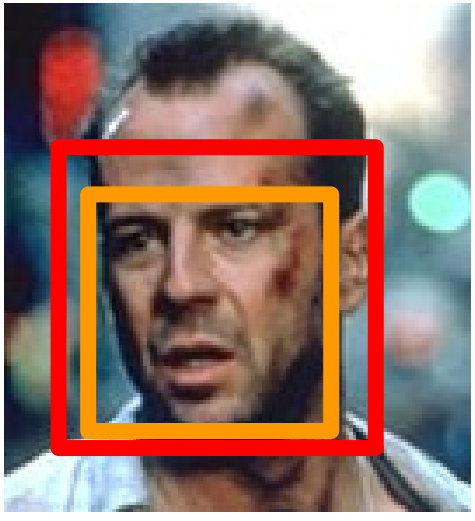
- Our solution:
  - Flag annotations < α = 30 pixel with "ignore" label
  - Delete detections <β, set $\beta = \sqrt{0.5 * \alpha^2}$

Annotation size = α
Detection size < α



✓

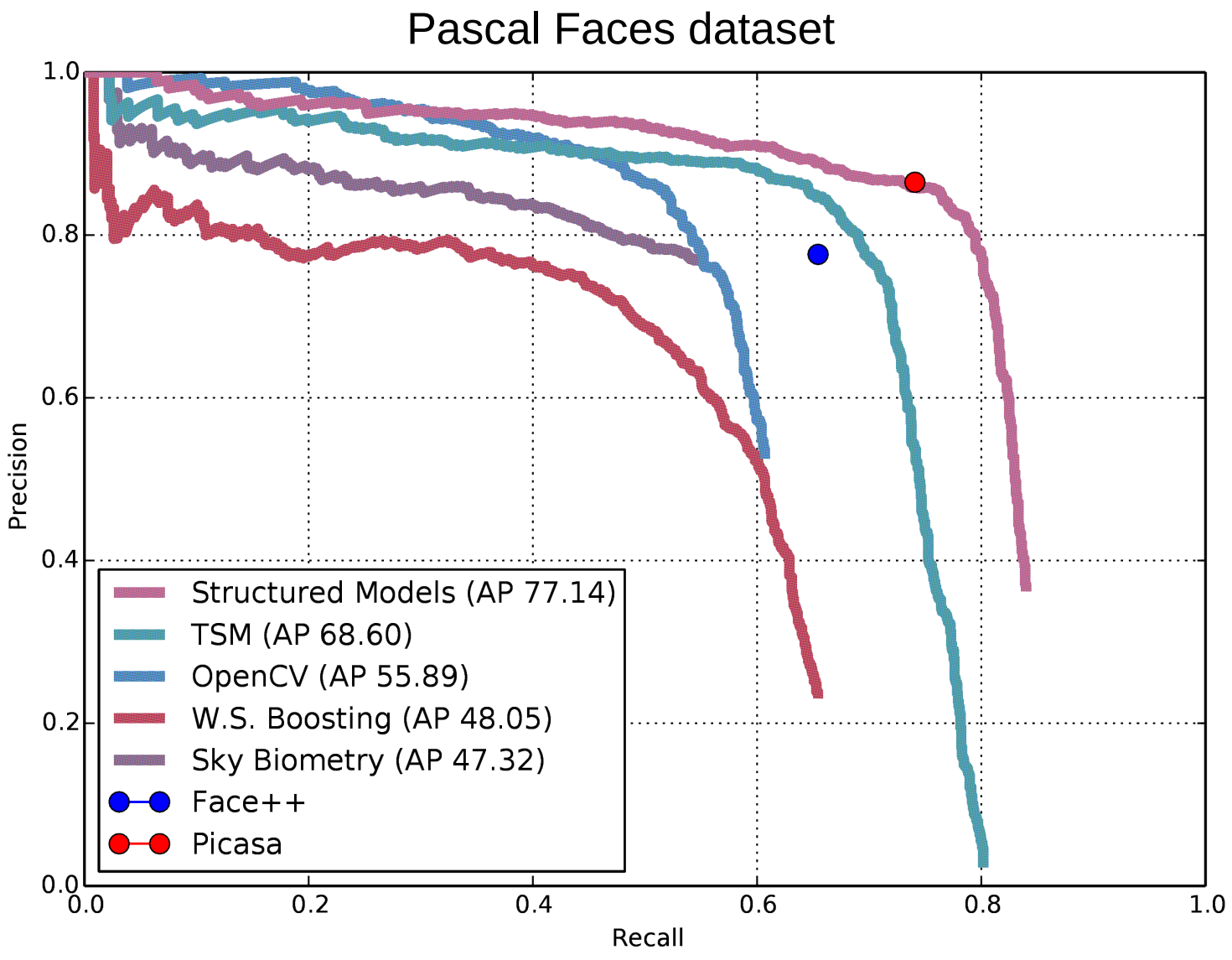
Annotation size < α
Detection size = α



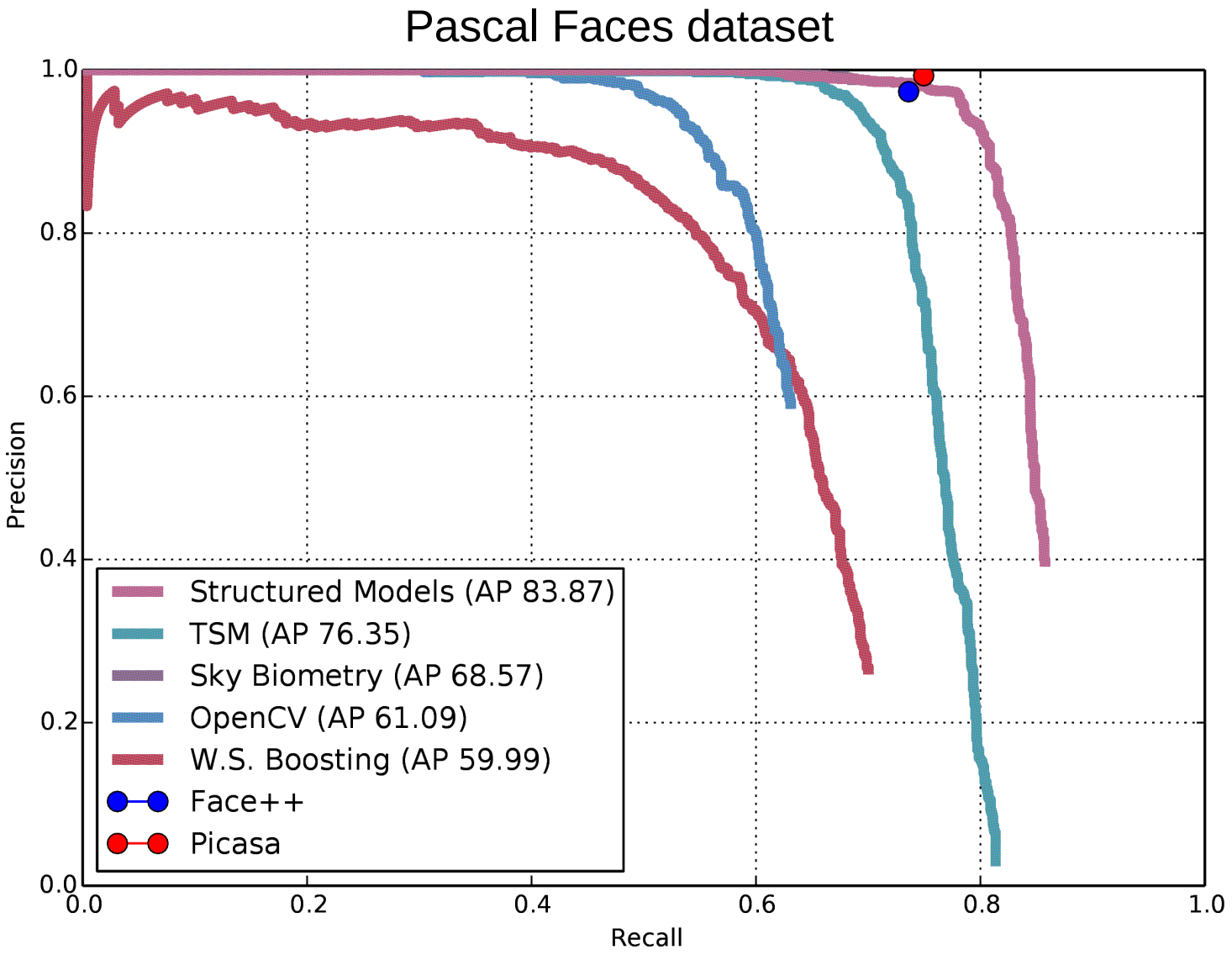
✓

✓

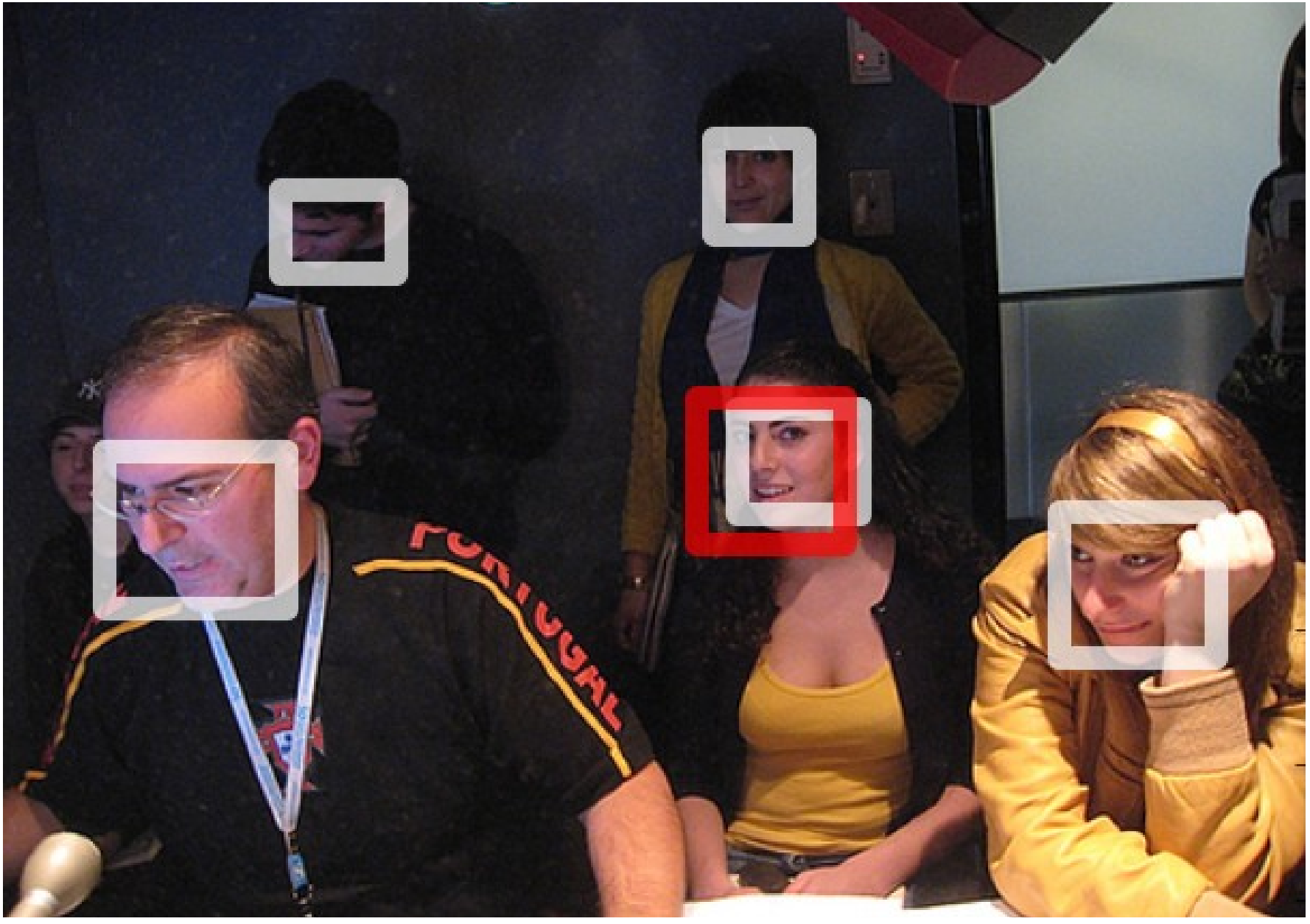
Previous evaluation



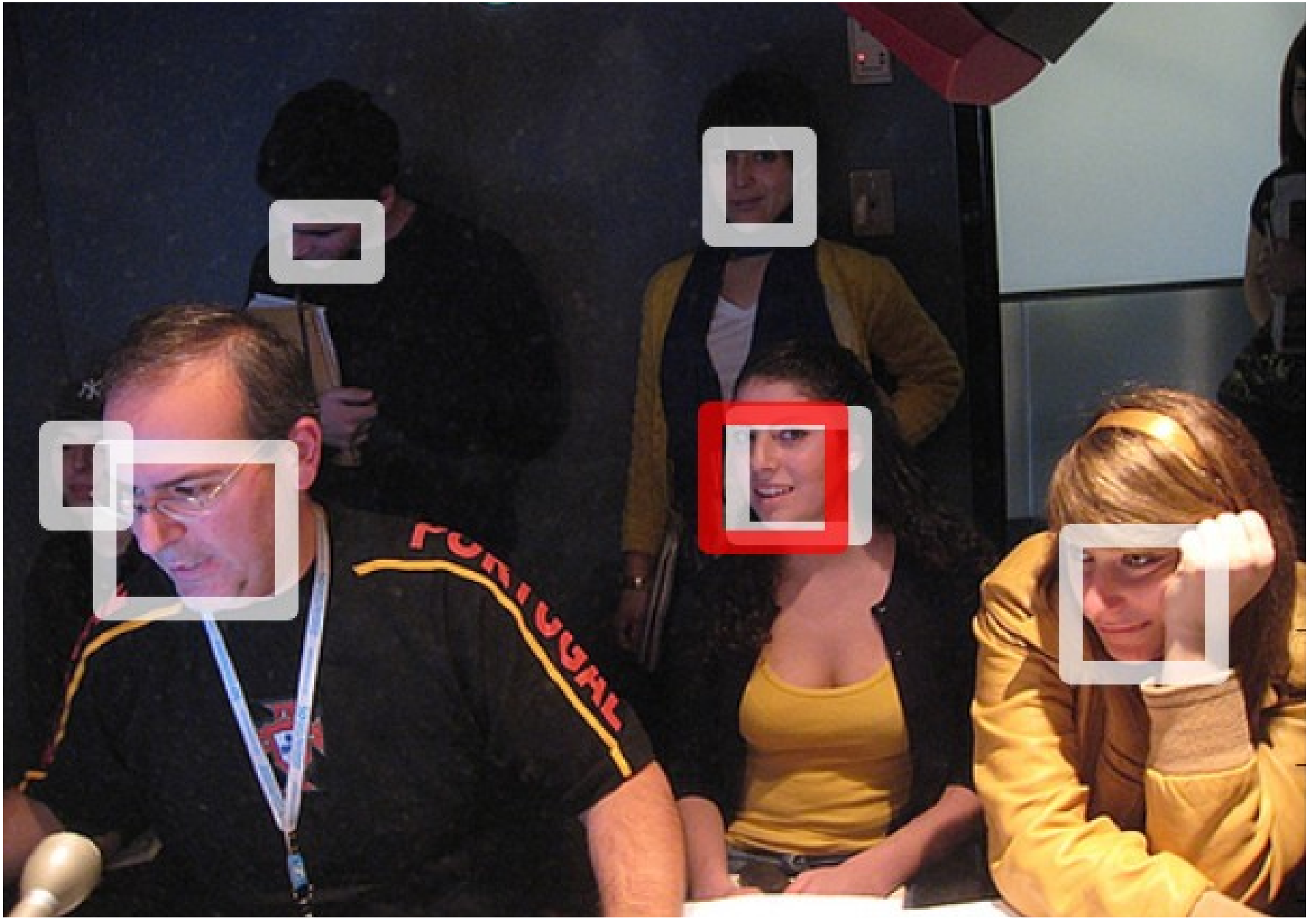
New evaluation + new annotations



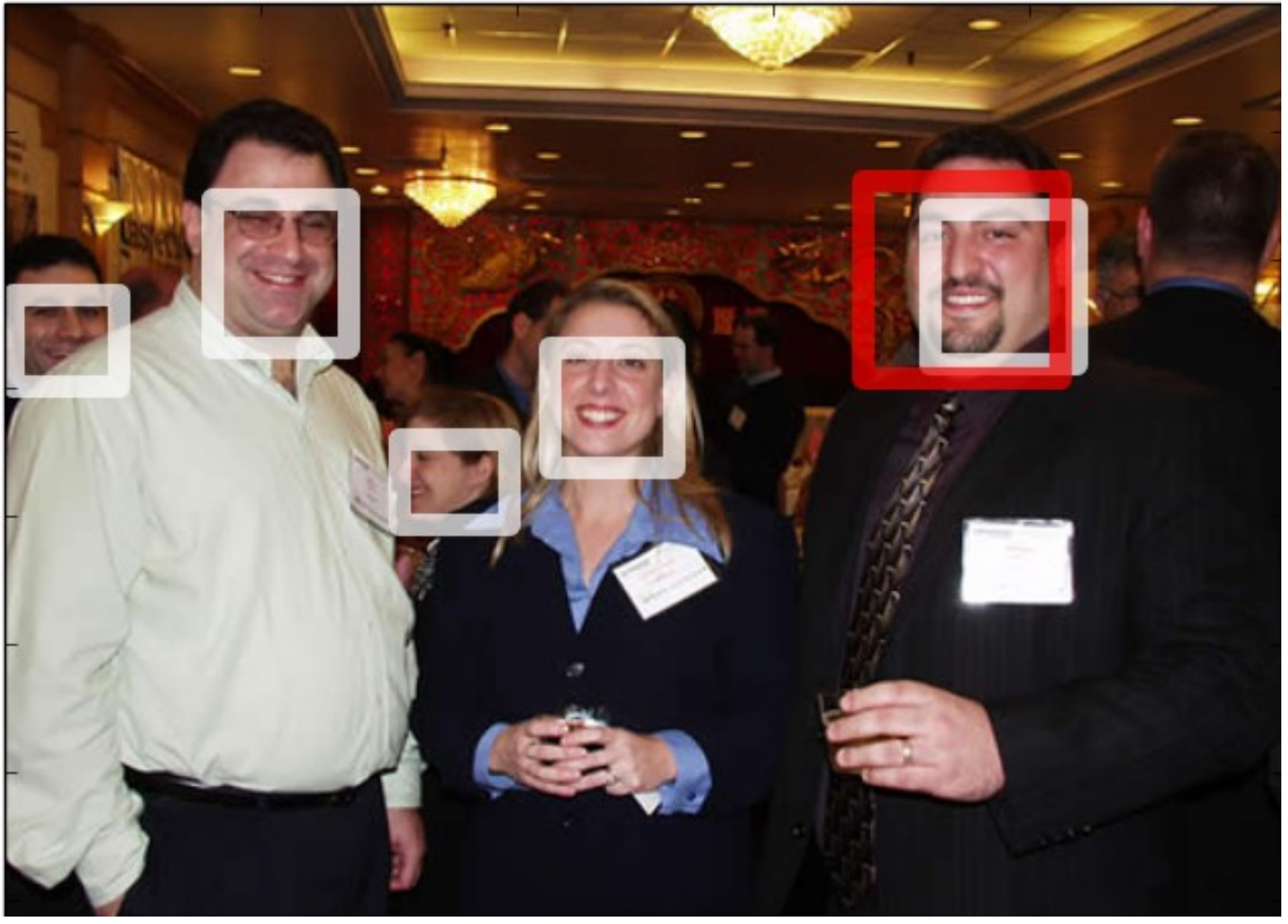
Before



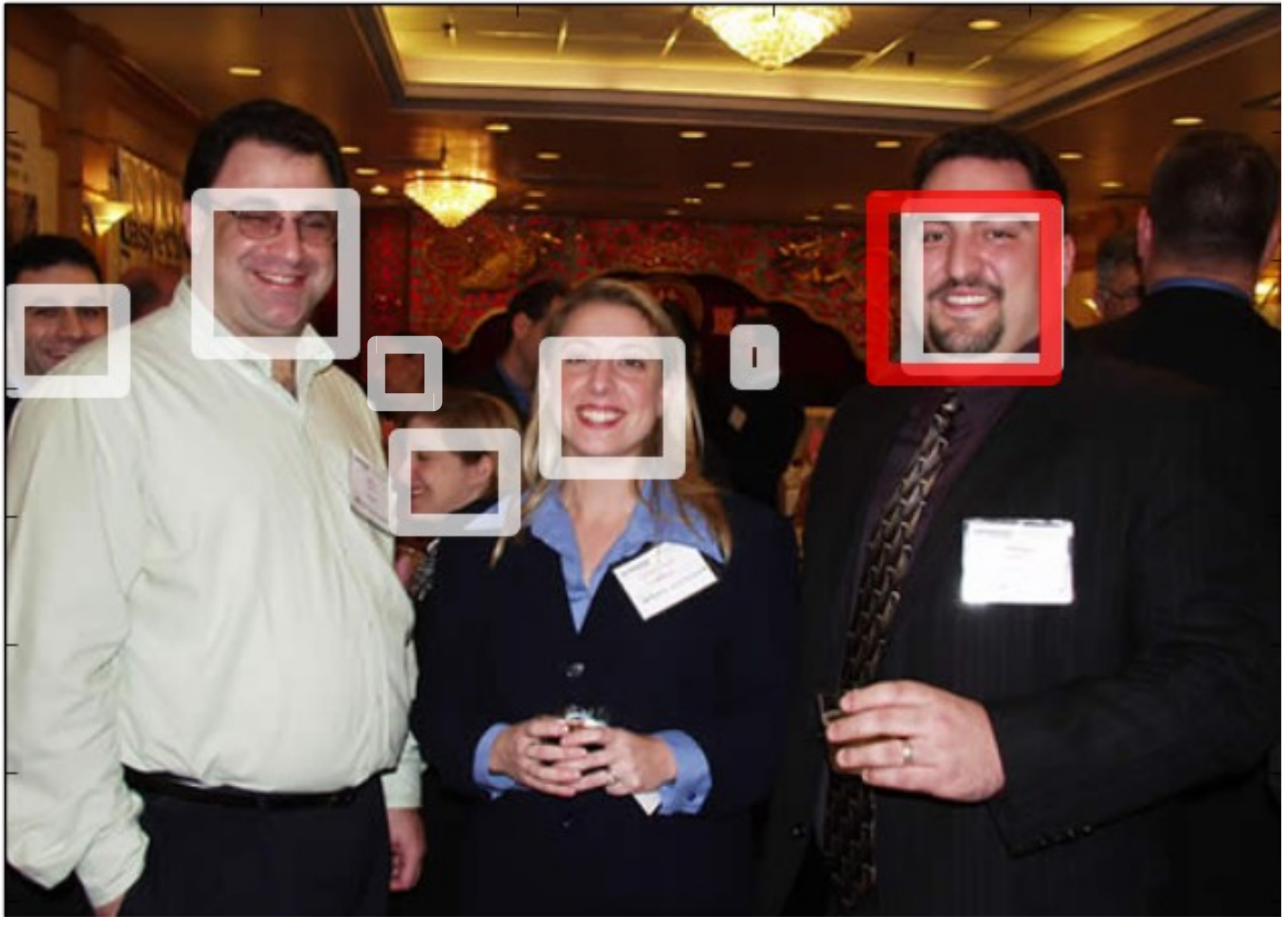
After



Before



After

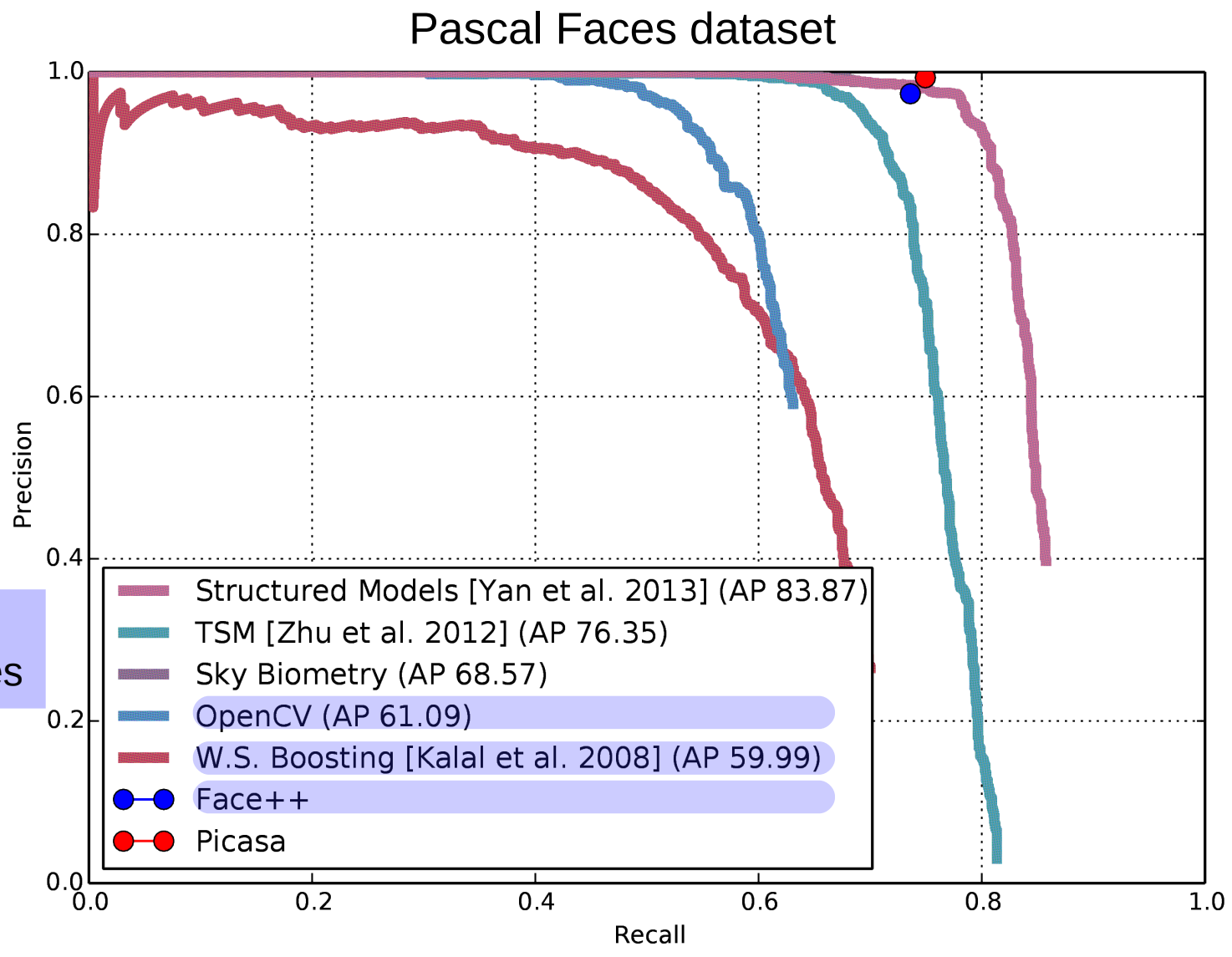


Before

After

Baselines

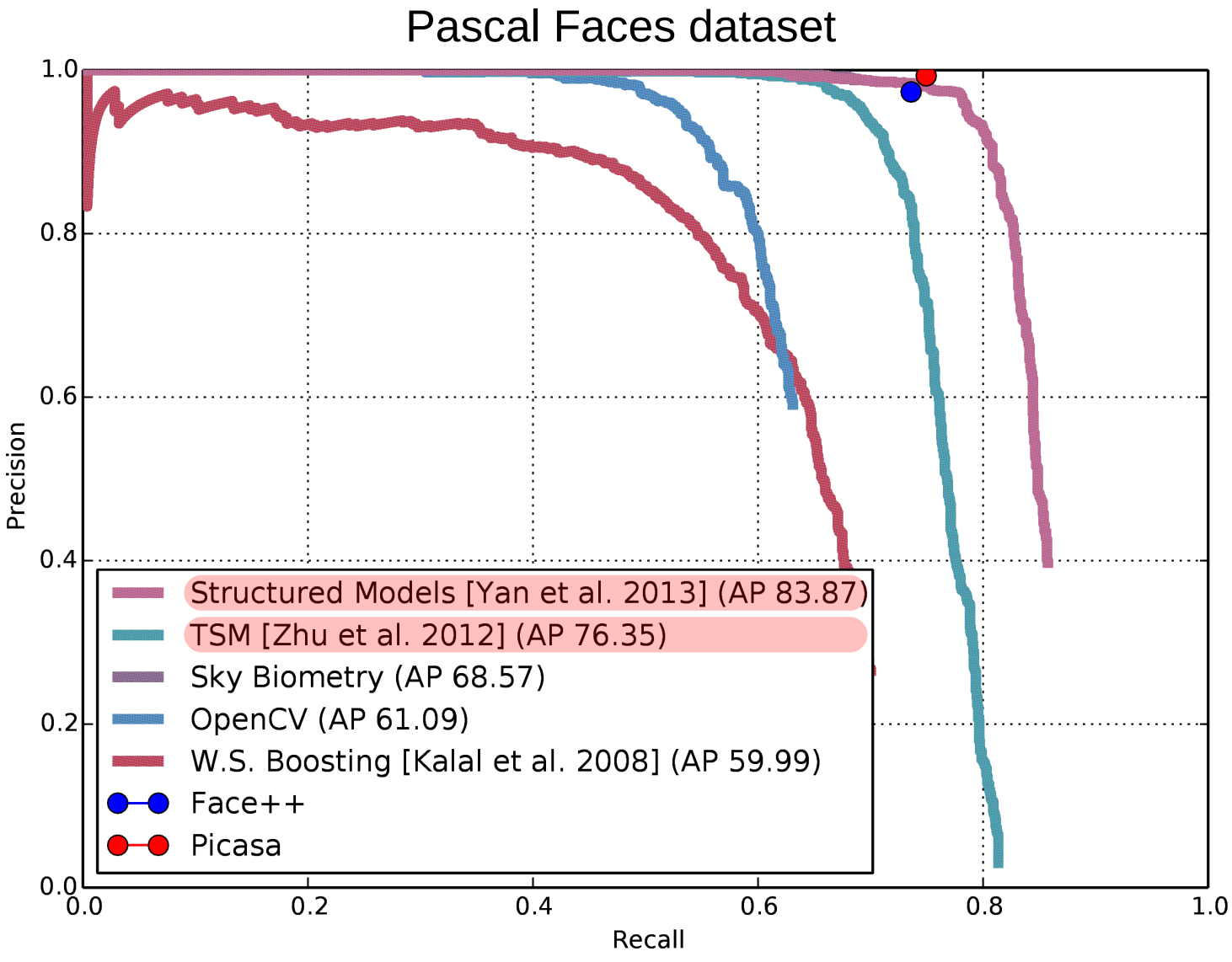
Some methods are based on Viola&Jones



Based on Viola&Jones

Some methods are based on DPM

Based on DPM



Baselines are trained using AFLW

We use 5 templates for the face class.

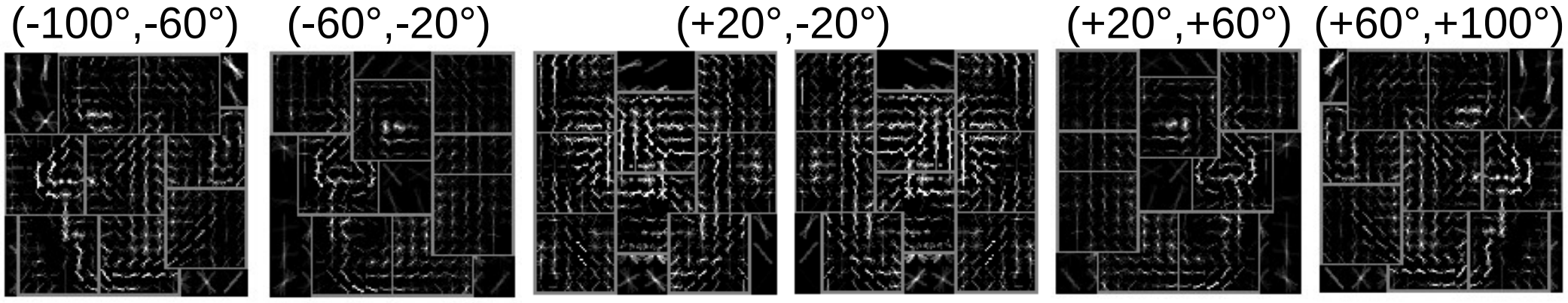
(-100°,-60°)	(-60°,-20°)	(+20°,-20°)	(+20°,+60°)	(+60°,+100°)
				
2544 samples	5810 samples	6752 samples	mirrored	mirrored

AFLW training data

[Koestinger et al. ICCV 2011]

DPM v5 baseline

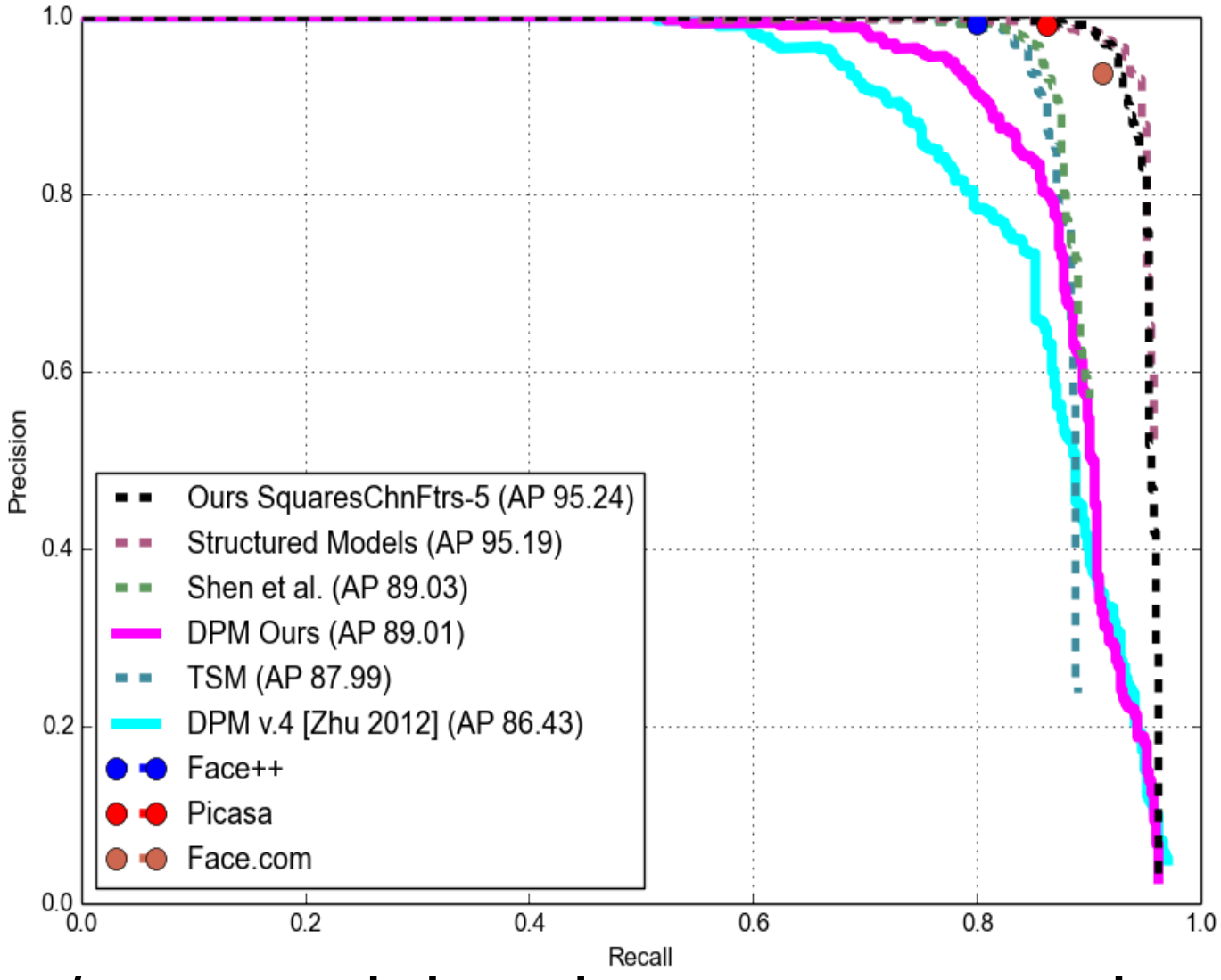
- Using default parameters, except initialization



AFLW training data

[Koestinger et al. ICCV 2011]

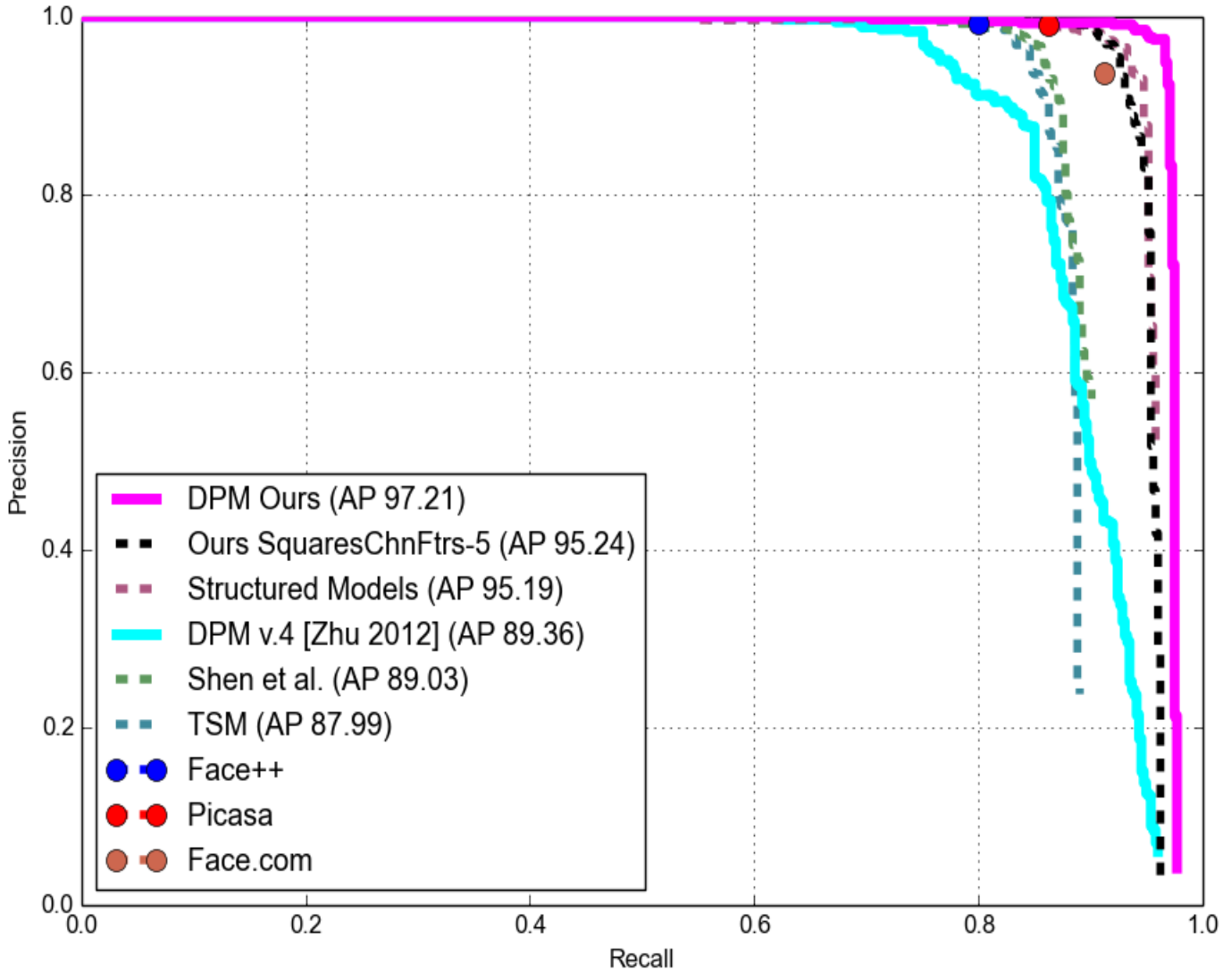
DPM on AFW dataset



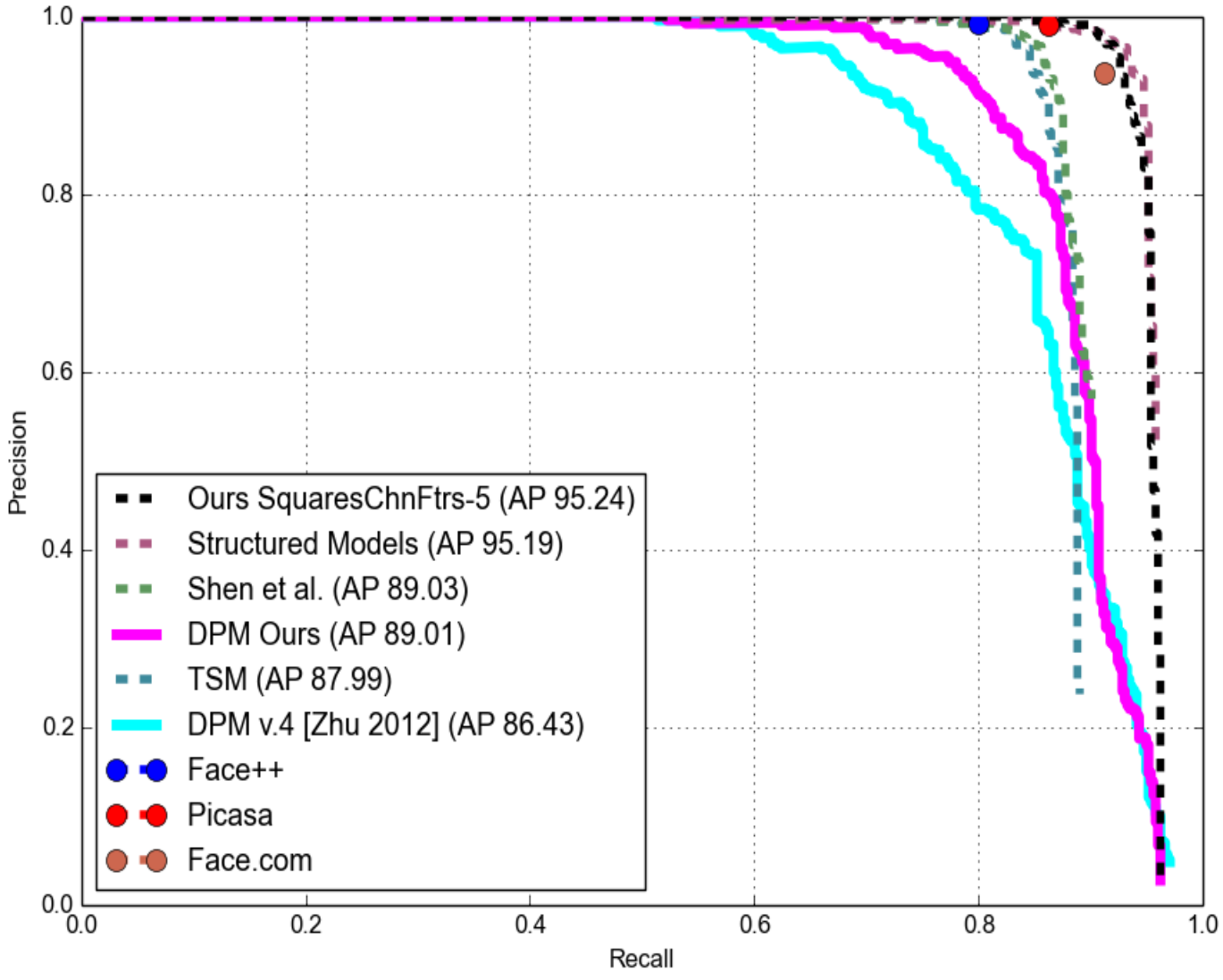
- Better/more training data, newer version

⇨ 2.5 percent points better than [Zhu et al. CVPR 2012]

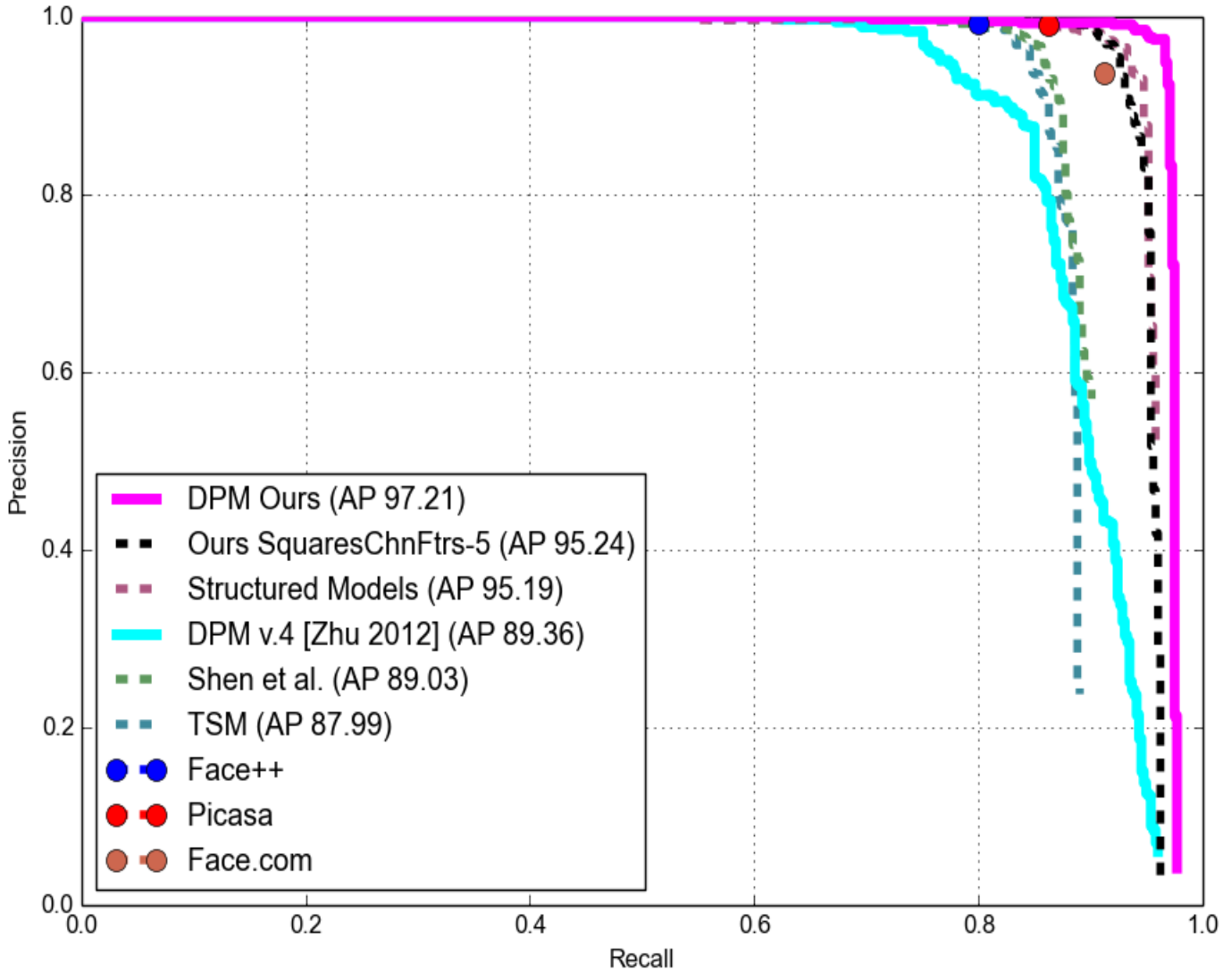
DPM on AFW dataset



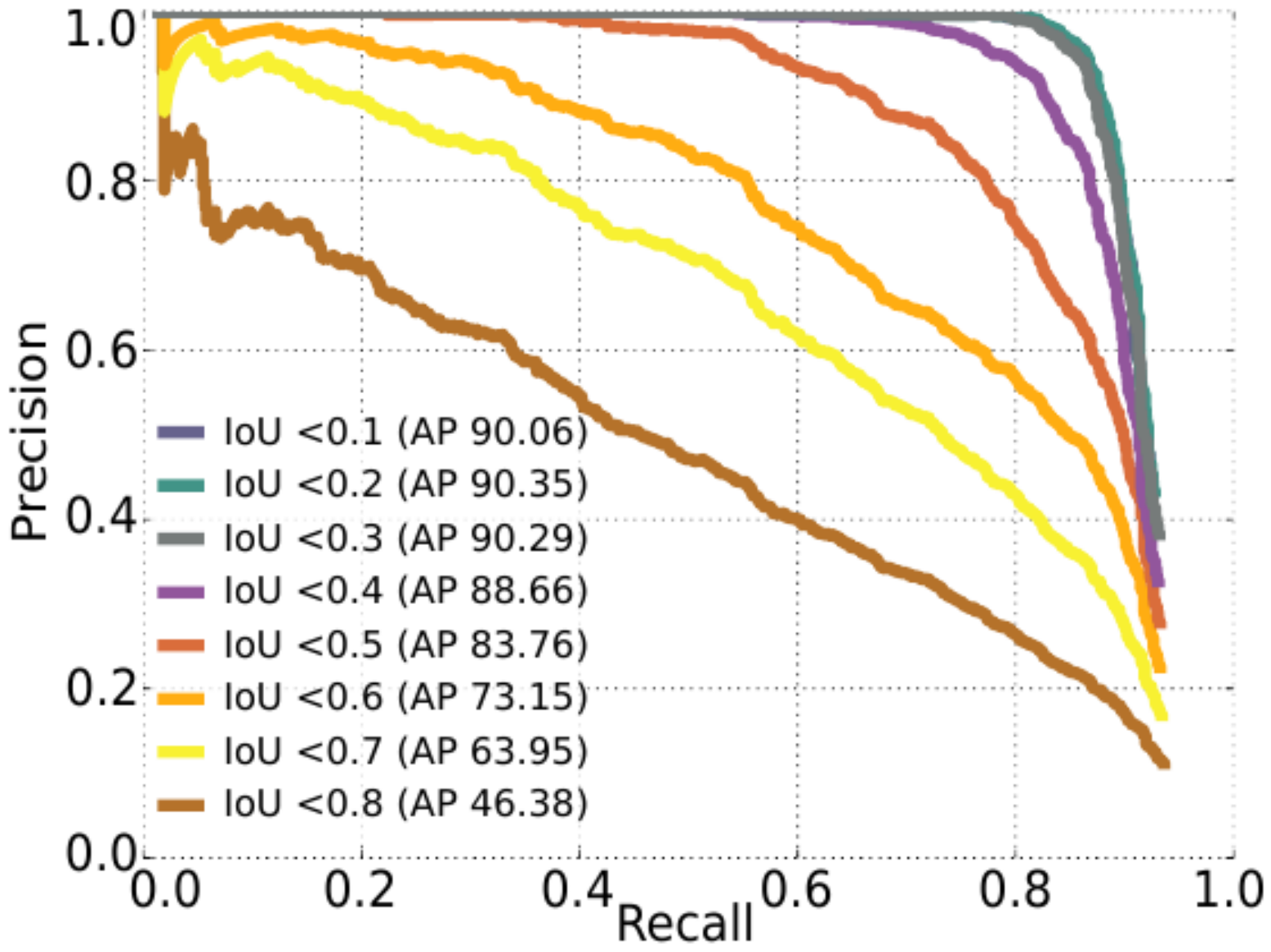
NMS threshold 0.5



NMS threshold 0.3



NMS matters a lot!

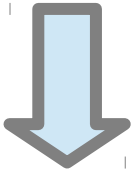


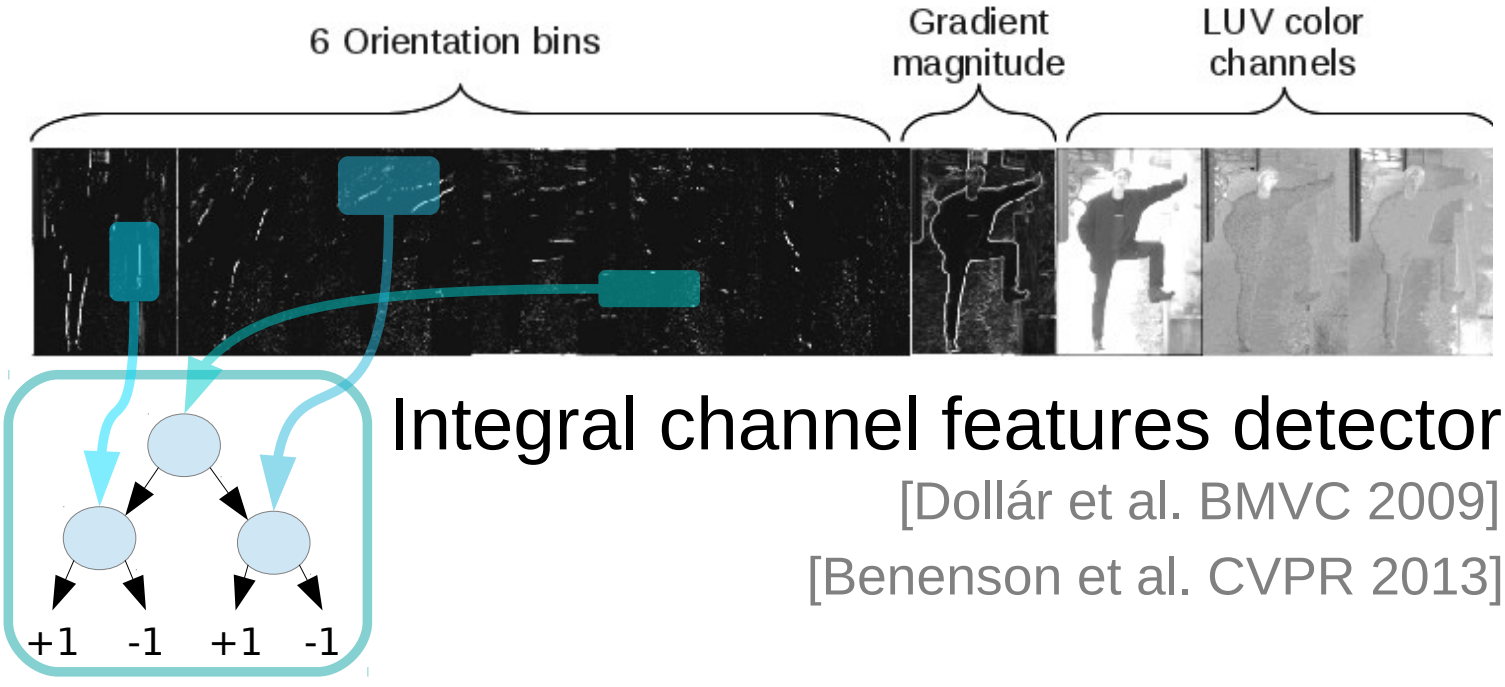
Overlapping DPM detections

Intersection/Union just smaller than 0.5

Viola&Jones baseline

[Viola and Jones IJCV 2004]





Integral channel features detector

[Dollár et al. BMVC 2009]

[Benenson et al. CVPR 2013]

Ours SquaresChnFtrs-5

We use 5 templates for the face class.

(-100°,-60°)	(-60°,-20°)	(+20°,-20°)	(+20°,+60°)	(+60°,+100°)
				
2544 samples	5810 samples	6752 samples	mirrored	mirrored

AFLW training data

[Koestinger et al. ICCV 2011]

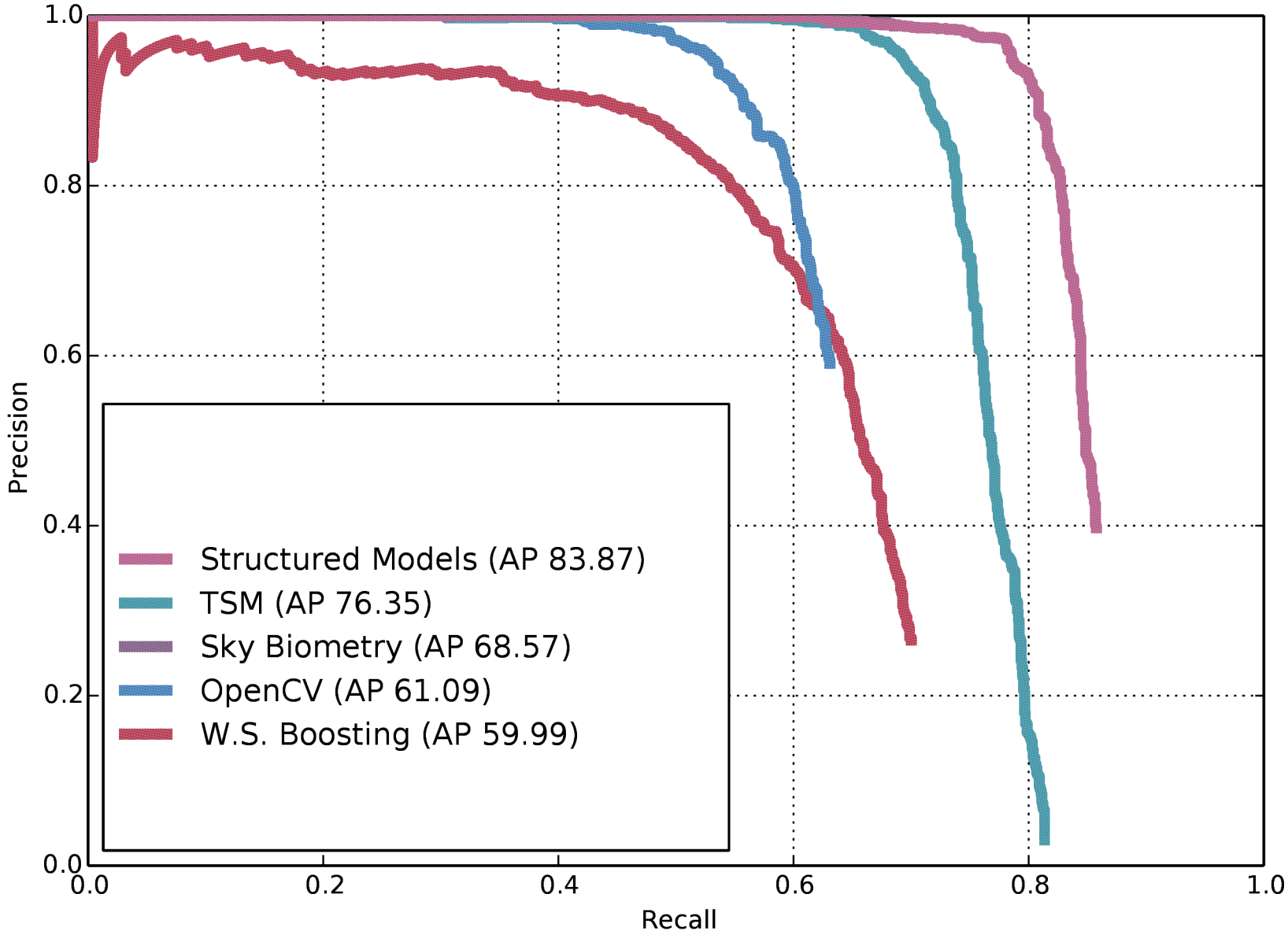
What makes a face detector (truly) tick?

- Number of training samples
- Number of templates (components)
- Influence of color channels
- Number of weak learners

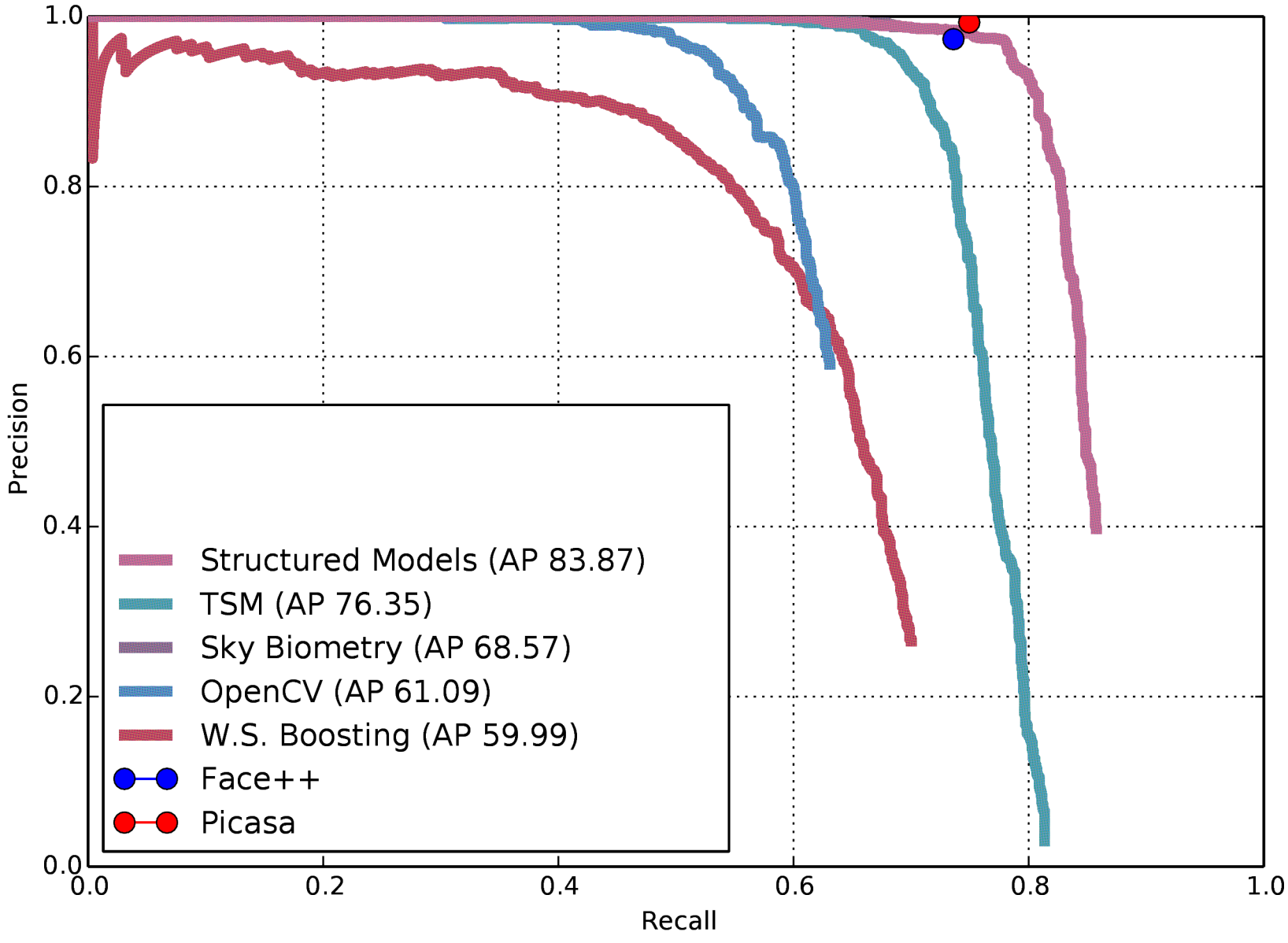
What makes a face detector (truly) tick?

- Number of training samples
- **Number of templates (components)**
- Influence of color channels
- Number of weak learners

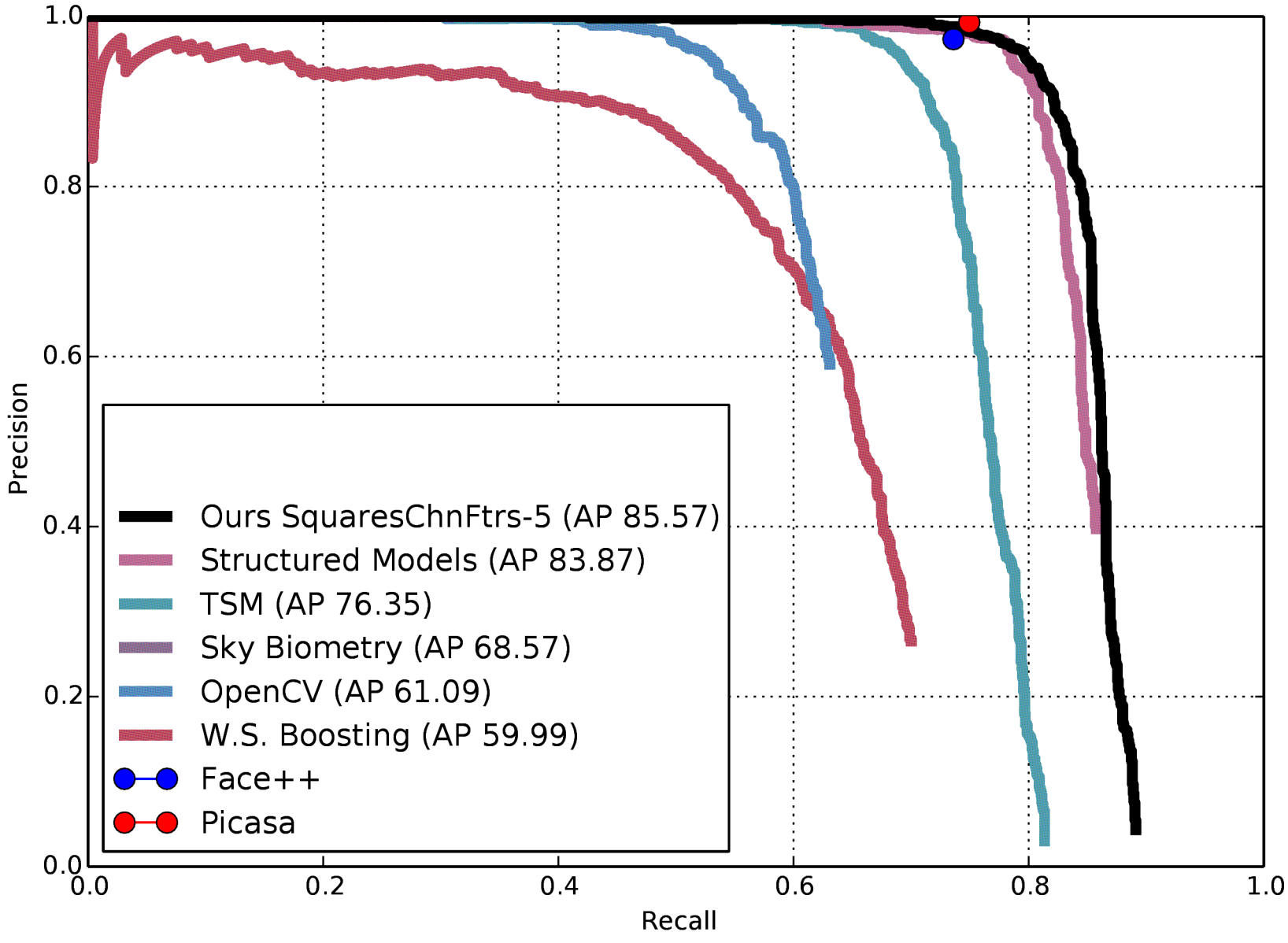
Results on Pascal Faces



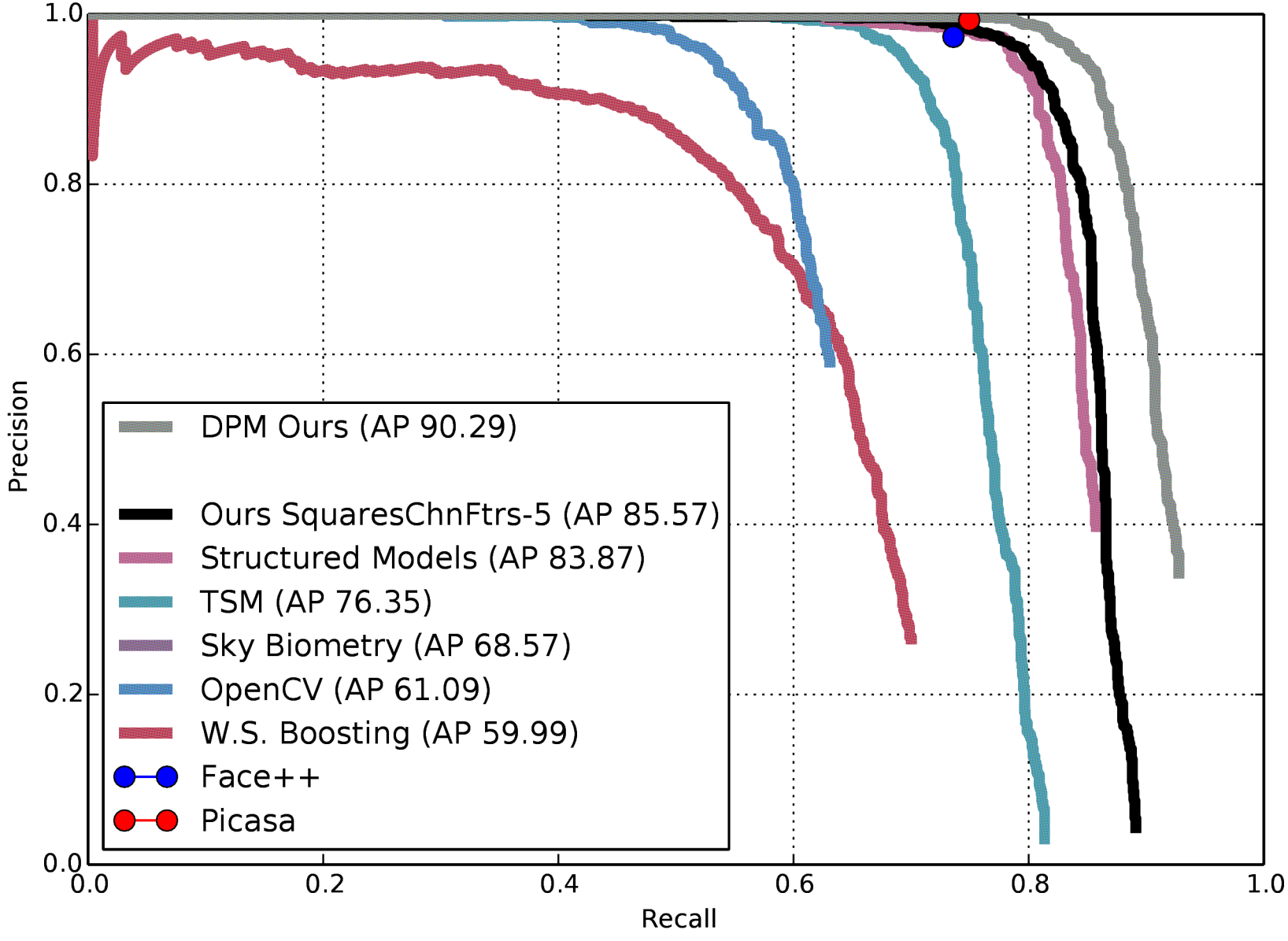
Results on Pascal Faces



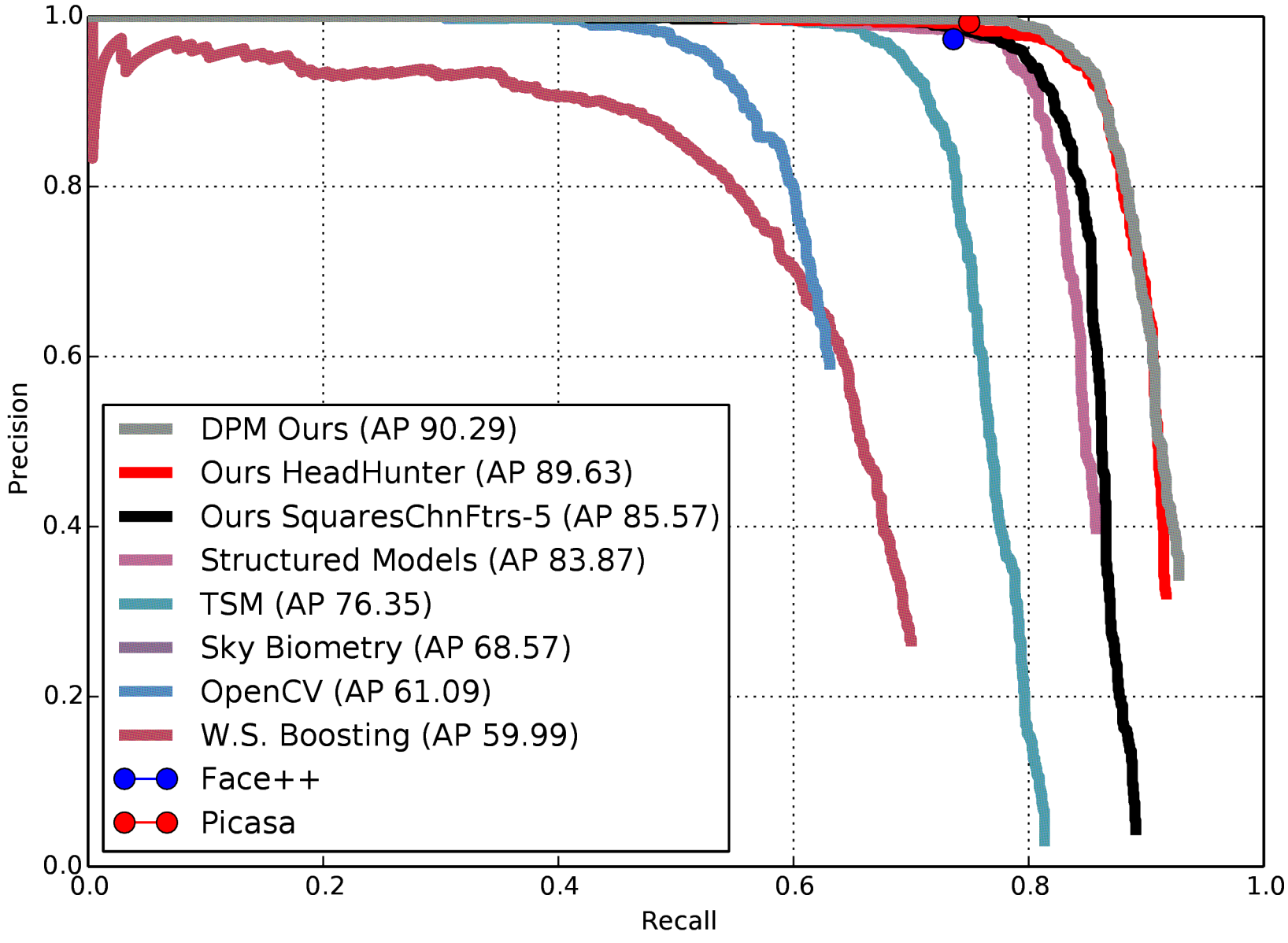
Results on Pascal Faces



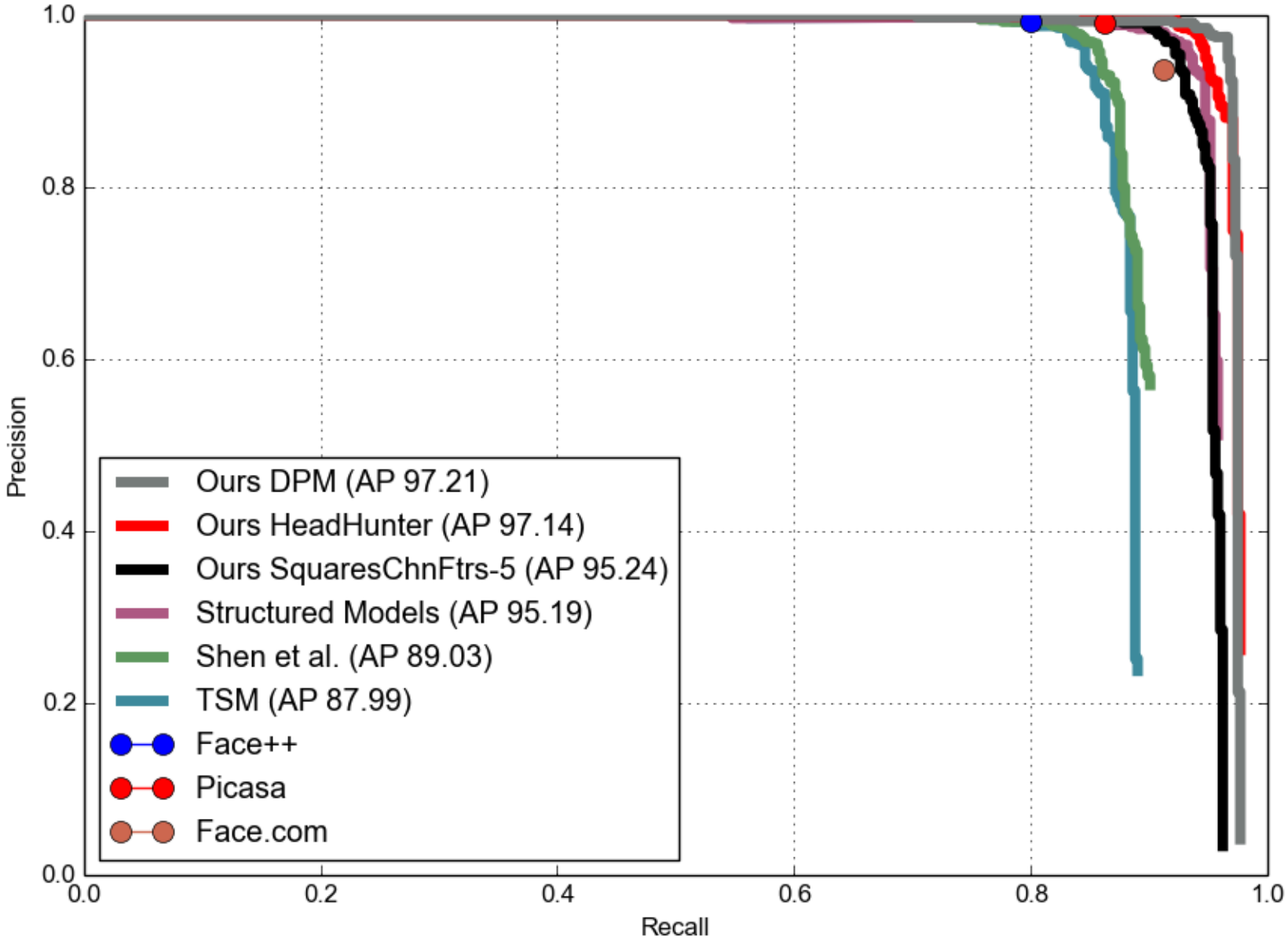
Results on Pascal Faces



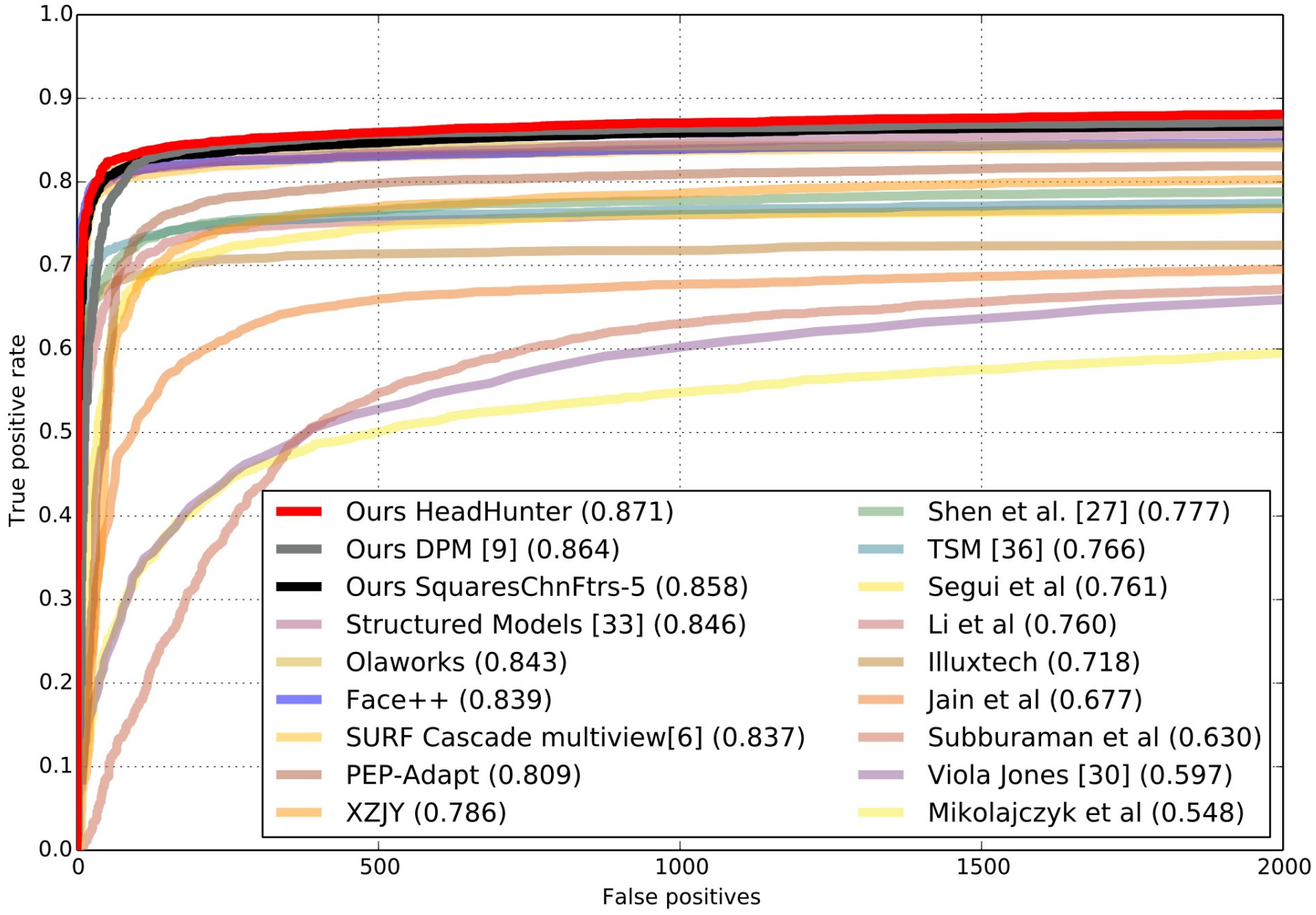
Results on Pascal Faces



Results on AFW



Results on FDDB



Contributions

- Release of a new, more principled, evaluation toolkit:
  - New evaluation toolbox
  - New annotations
- Research systems on par with commercial products
- Vanilla DPM and rigid templates reach top performance

Take home message

- Detection evaluation is non-trivial
- Baseline methods are surprisingly effective

Questions?

- Evaluation code, annotations, trained models at:
  http://markusmathias.de/face_detection/