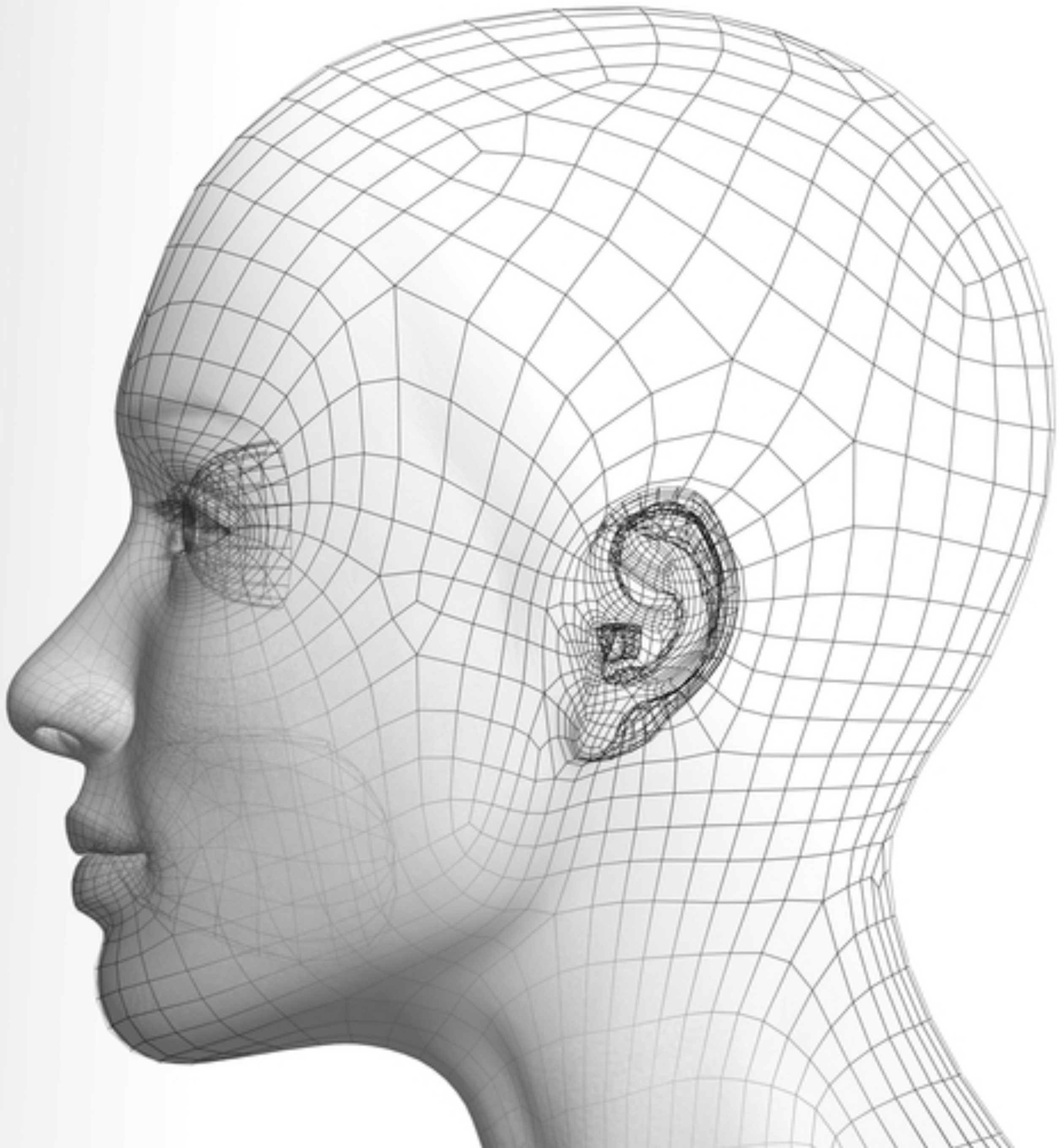
Deep Learning in Face Analysis

Chen-Change LOY

MMLAB

The Chinese University of Hong Kong

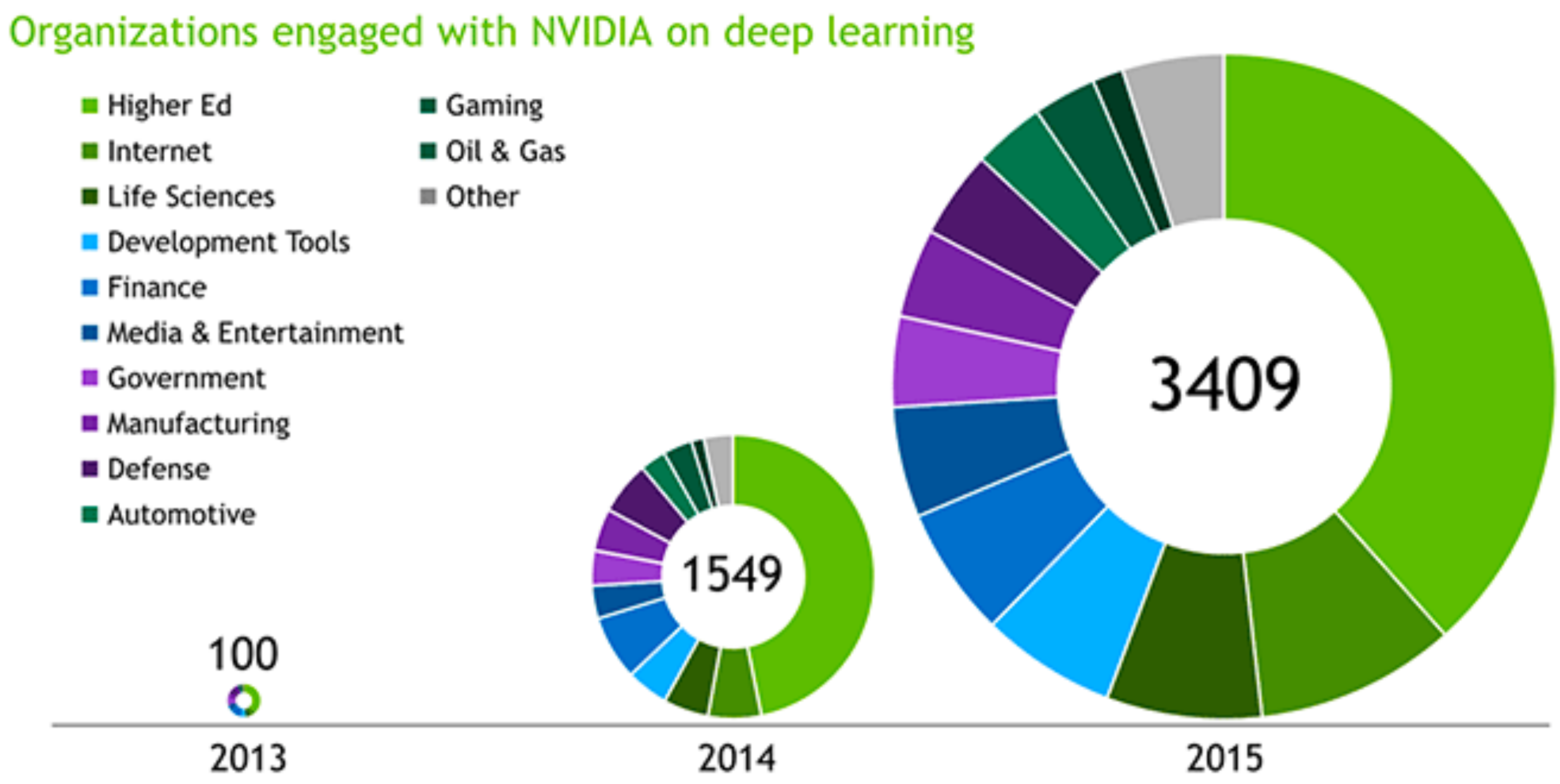
Homepage: http://personal.ie.cuhk.edu.hk/~ccloy/



Outline

- Overview
- Face detection
- Face attribute recognition
- Face hallucination

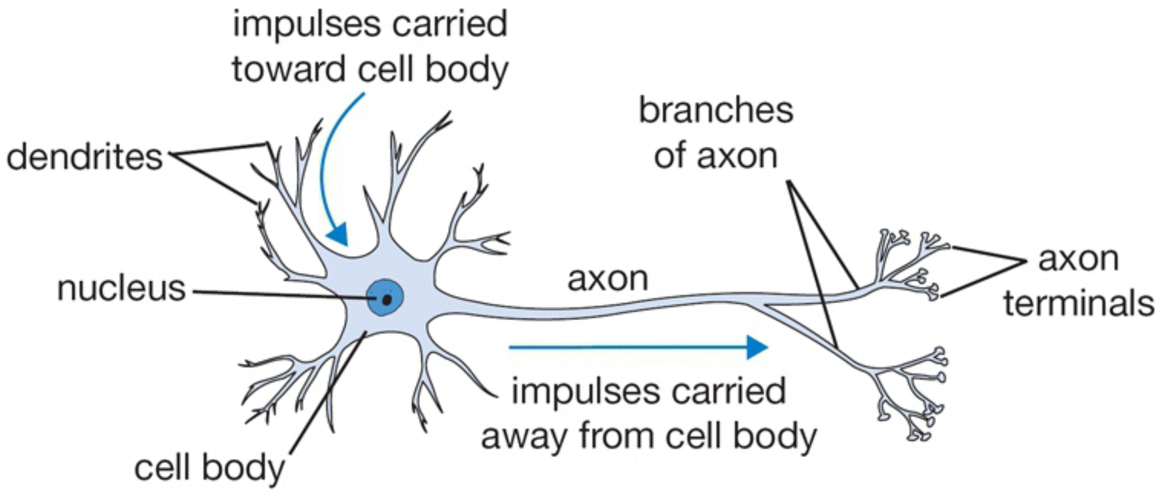
Every industry wants intelligence



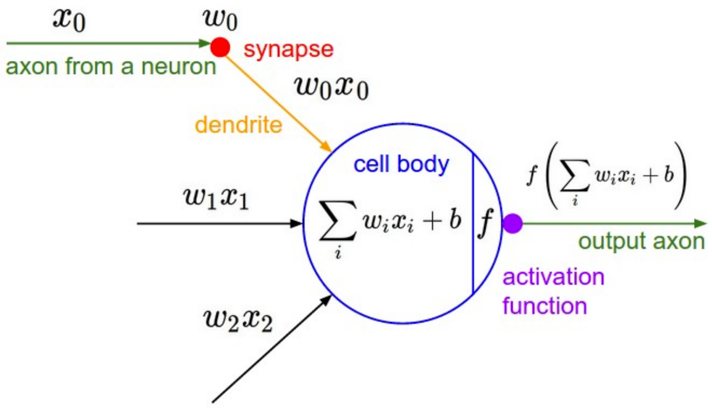
- Has shown impressive results in voice and image recognition
- Finding new applications, from fashion to finance

Credit: https://blogs.nvidia.com/blog/2016/01/12/accelerating-ai-artificial-intelligence-gpus/

Deep learning is not new

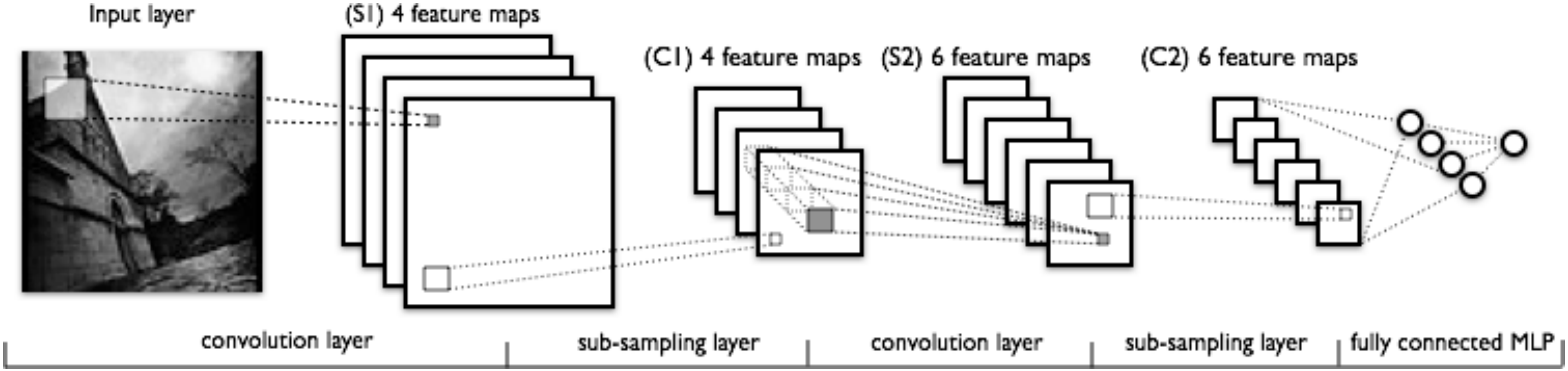


Biological inspiration



- Early works on learning neural networks
  - Frank Rosenblatt (1958) created the perceptron, an algorithm for pattern recognition based on a two-layer computer learning network using simple addition and subtraction
- Backpropagation was developed in several steps since 1960

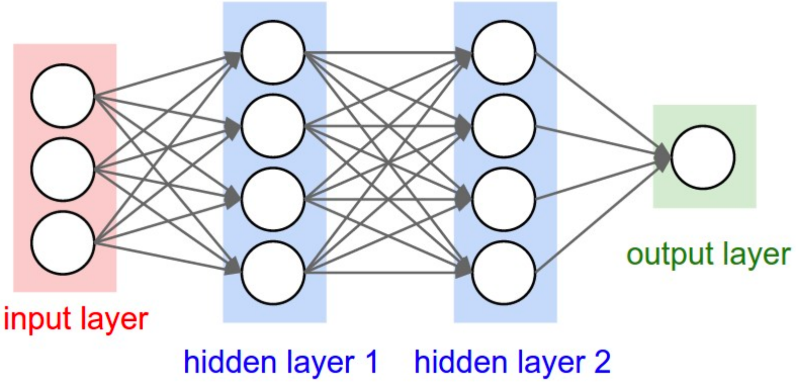
Convolutional Network



Deep learning timeline

Rumelhart, Hinton, and Williams, Nature 1986
Neural network back propagation

1986



- Solve general learning problems
- Tied with biological system

But it was given up …

- Hard to train
- Insufficient computational resources
- Small training sets
- Does not work well

Deep learning timeline

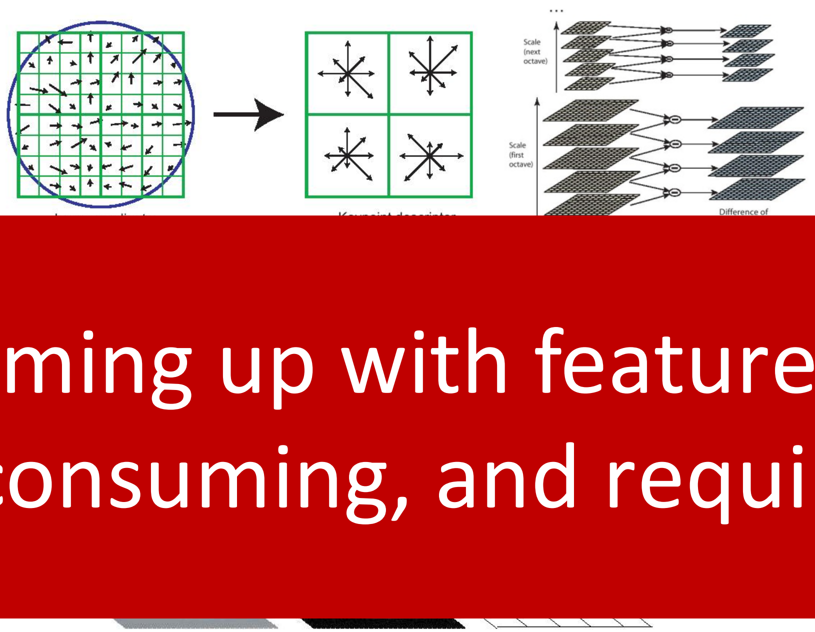
Rumelhart, Hinton, and Williams, Nature 1986
Neural network back propagation

SVM, Boosting, Decision tree, KNN

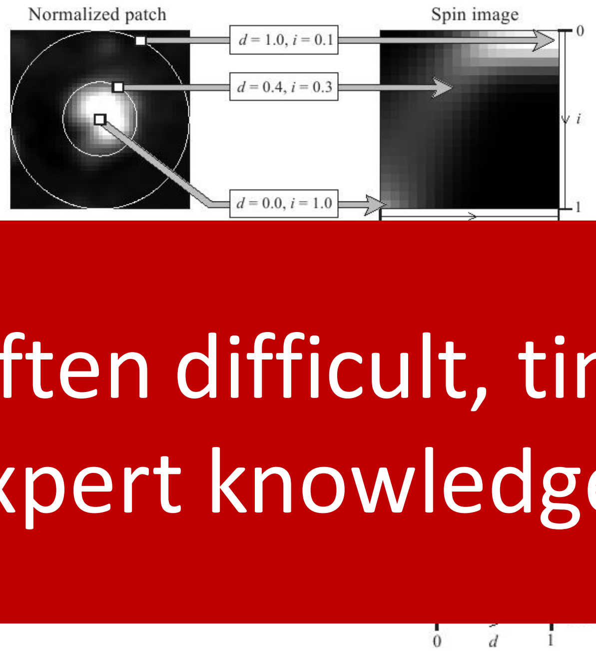
1986 | Dark Age of Neural Network | 2006

- Loose tie with biological systems
- Flat structures
- Specific methods for specific tasks
  - Hand-crafted features (GMM-HMM, SIFT, LBP, HOG)

Hand-crafted features

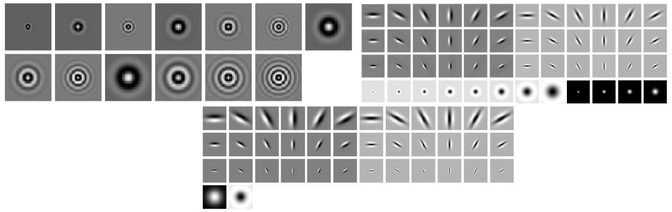


HoG

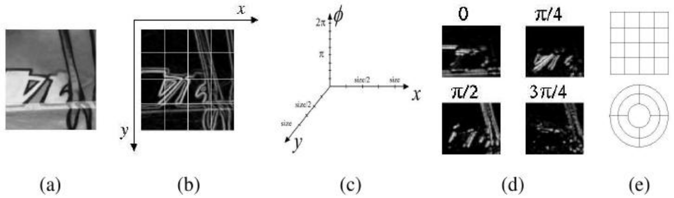


RIFT

Coming up with features is often difficult, time-consuming, and requires expert knowledge.



Textons



GLOH

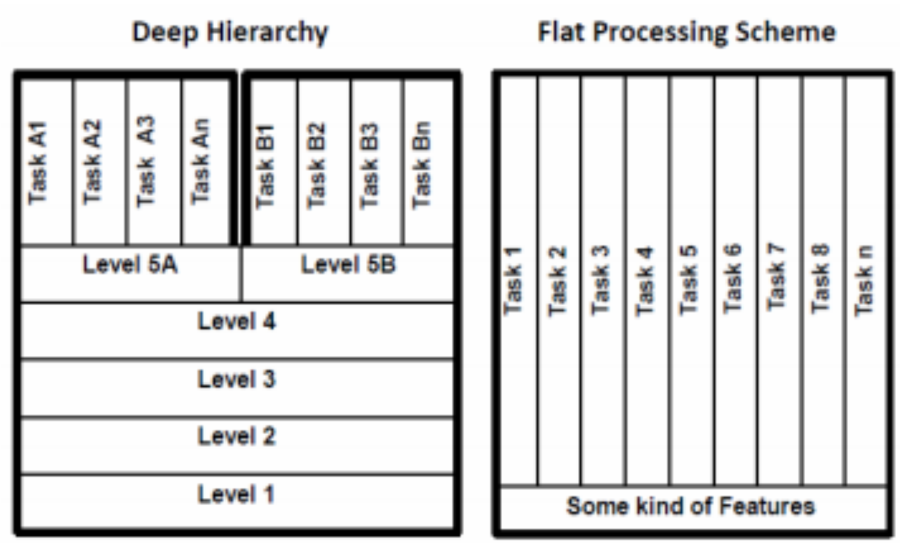
Deep learning timeline

Rumelhart, Hinton, and Williams, Nature 1986

Neural network back propagation

SVM, Boosting, Decision tree, KNN

1986 | Dark Age of Neural Network | 2006



Kruger TPAMI'13

- Loose tie with biological systems
- Flat structures
- Specific methods for specific tasks
    - Hand-crafted features (GMM-HMM, SIFT, LBP, HOG)

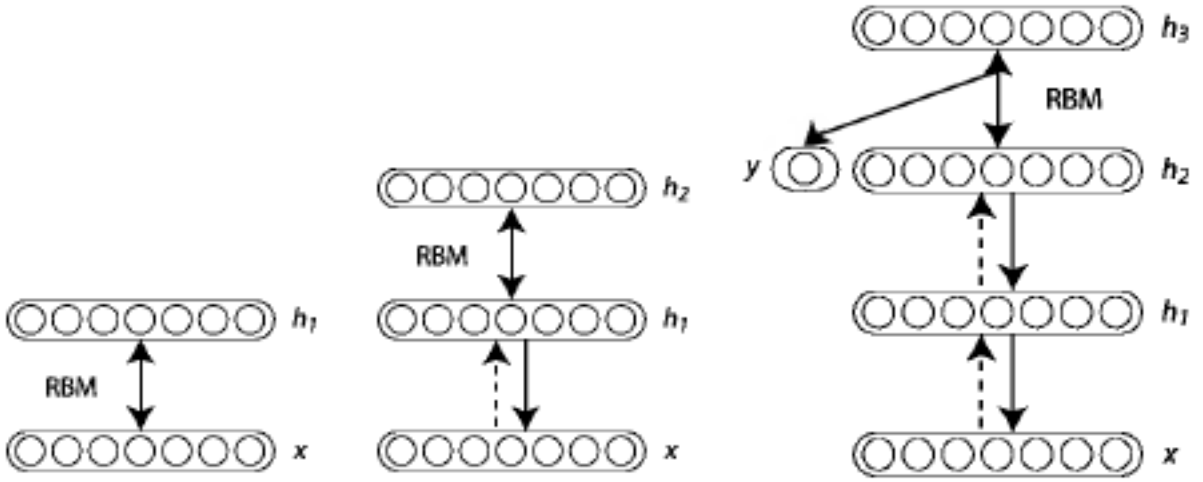
Deep learning timeline

Rumelhart, Hinton, and Williams, *Nature* 1986
Neural network back propagation

SVM, Boosting, Decision tree, KNN

Hinton et al, *Neural Computation* 2006
Deep belief net

1986 | Dark Age of Neural Network | 2006



(a) Train RBM for $\mathbf{x}$ (b) Train RBM for $\mathbf{h}^1$ (c) Train RBM for $\mathbf{h}^2$ and y

- Stacking many hidden layers
- Better learning algorithms
  - Unsupervised and layer-wised pre-training
  - Dropout to prevent overfitting
  - …

Deep learning timeline

Rumelhart, Hinton, and Williams, *Nature* 1986
Neural network back propagation

SVM, Boosting, Decision tree, KNN

Hinton et al, *Neural Computation* 2006
Deep belief net

Speech

Breakthrough in computer vision!

1986 | Dark Age of Neural Network | **2006** | **2011** | **2012**

deep learning results

task	hours of training data	DNN-HMM	GMM-HMM with same data
Switchboard (test set 1)	309	18.5	27.4
Switchboard (test set 2)	309	16.1	23.6
English Broadcast News	50	17.5	18.8
Bing Voice Search (Sentence error rates)	24	30.4	36.2
Google Voice Input	5,870	12.3	
Youtube	1,400	47.6	52.3

Deep Networks Advance State of Art in Speech

Deep Learning leads to breakthrough in speech recognition at MSR.

Microsoft

What made CV again respect neural nets?

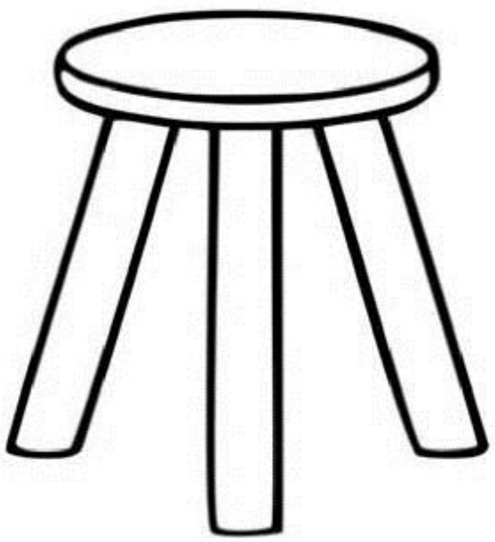
- Completely destroying non-deep learning methods on a modern competitive benchmark
  - ImageNet benchmark by Fei-Fei Li et al.

- Feature learned from large-scale dataset can be well generalized to other tasks and datasets!

What leads to the breakthrough?

- So, why indeed, did purely supervised learning with backpropagation not work well in the past? Geoffrey Hinton summarized the findings up to today in these four points:
    1. Our labeled datasets were thousands of times too small.
    2. Our computers were millions of times too slow.
    3. We initialized the weights in a stupid way.
    4. We used the wrong type of non-linearity.

What leads to the breakthrough?



Li Fei-Fei

Data

**ImageNet with
1 million images and
labels**

GPU

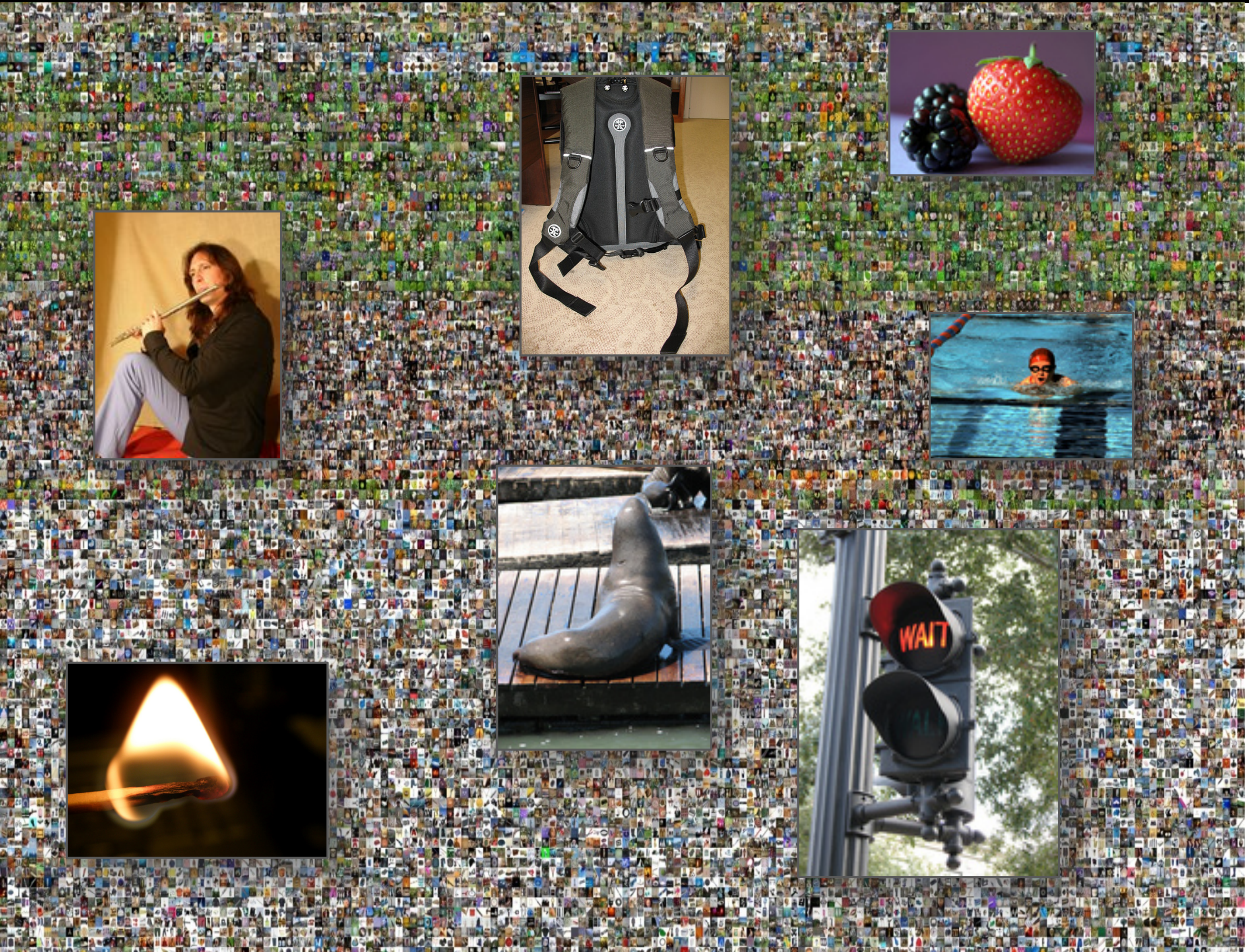
**1 Titan X is 20x faster than
16-core Xeon CPUs**

Geoffrey Hinton

Algorithms

**Network structure design
New training strategies**

ImageNet Large Scale Visual Recognition Challenge (ILSVRC)



- The most famous AI contest in the world
- Represent the state-of-the- art of computer vision
- 1,200,000 Training Images
- 100,000 Testing Images
- 1000 Classes

ImageNet Image Classification Challenge 2012

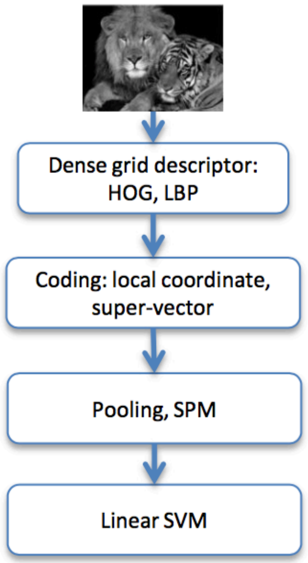
Rank	Name	Error rate	Description
1	**U. Toronto**	0.15315	Deep learning
2	U. Tokyo	0.26172	Hand-crafted features and learning models. Bottleneck.
3	U. Oxford	0.26979	
4	Xerox/INRIA	0.27058	

Object recognition over 1,000,000 images and 1,000 categories (2 GPUs)

Deep networks for ImageNet

Year 2010

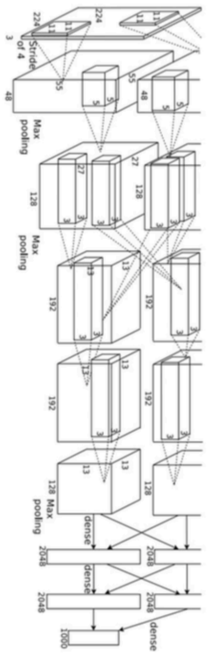
NEC-UIUC



[Lin CVPR 2011]

Year 2012

AlexNet



[Krizhevsky NIPS 2012]

Year 2014

GoogLeNet

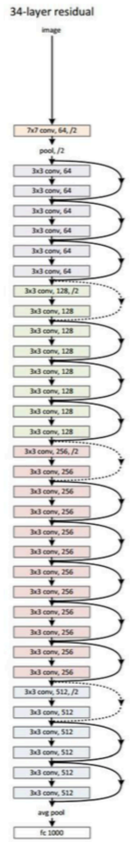
[Szegedy arxiv 2014]

VGG

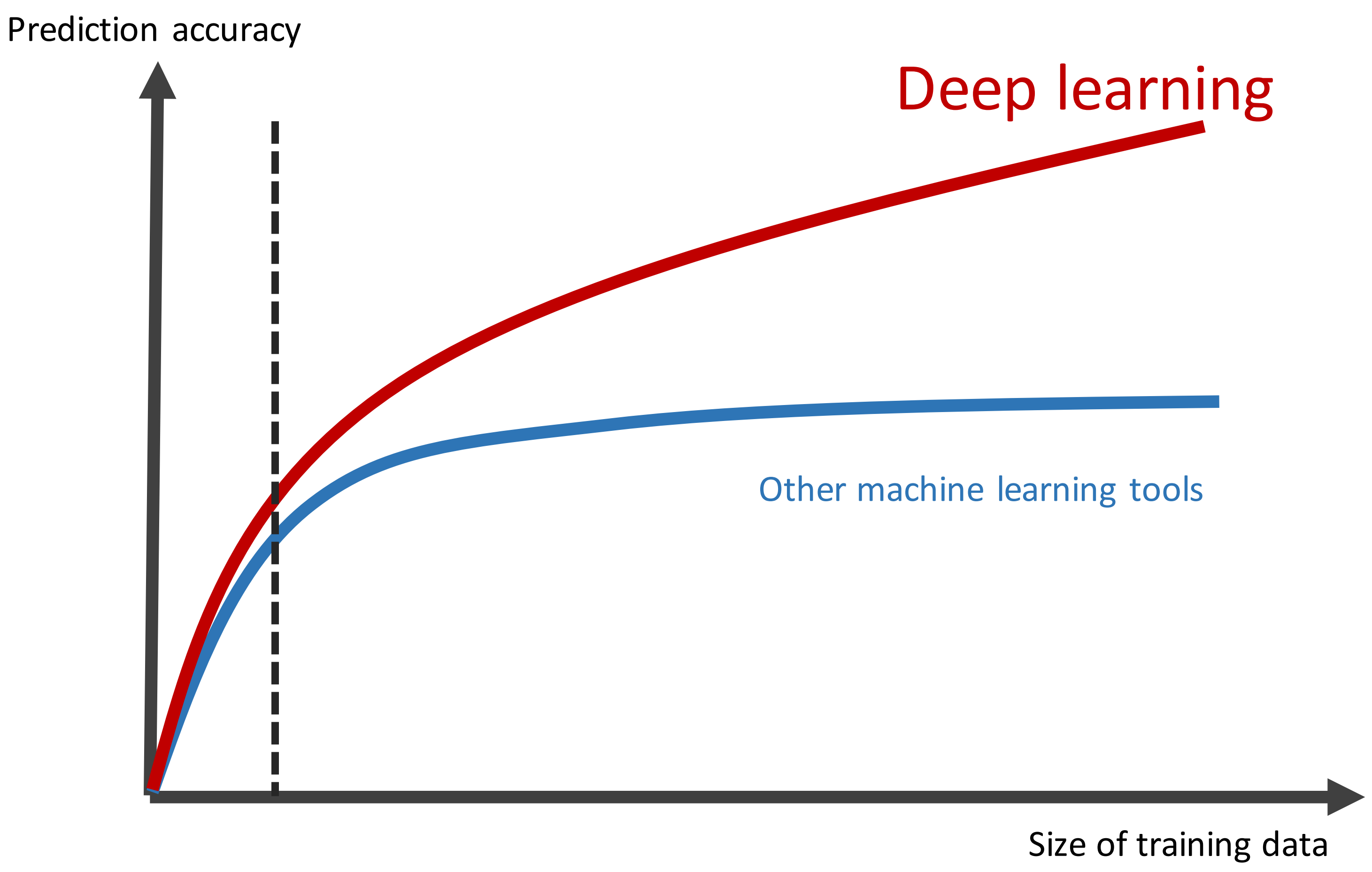
[Simonyan arxiv 2014]

Year 2015

MSRA ResNet



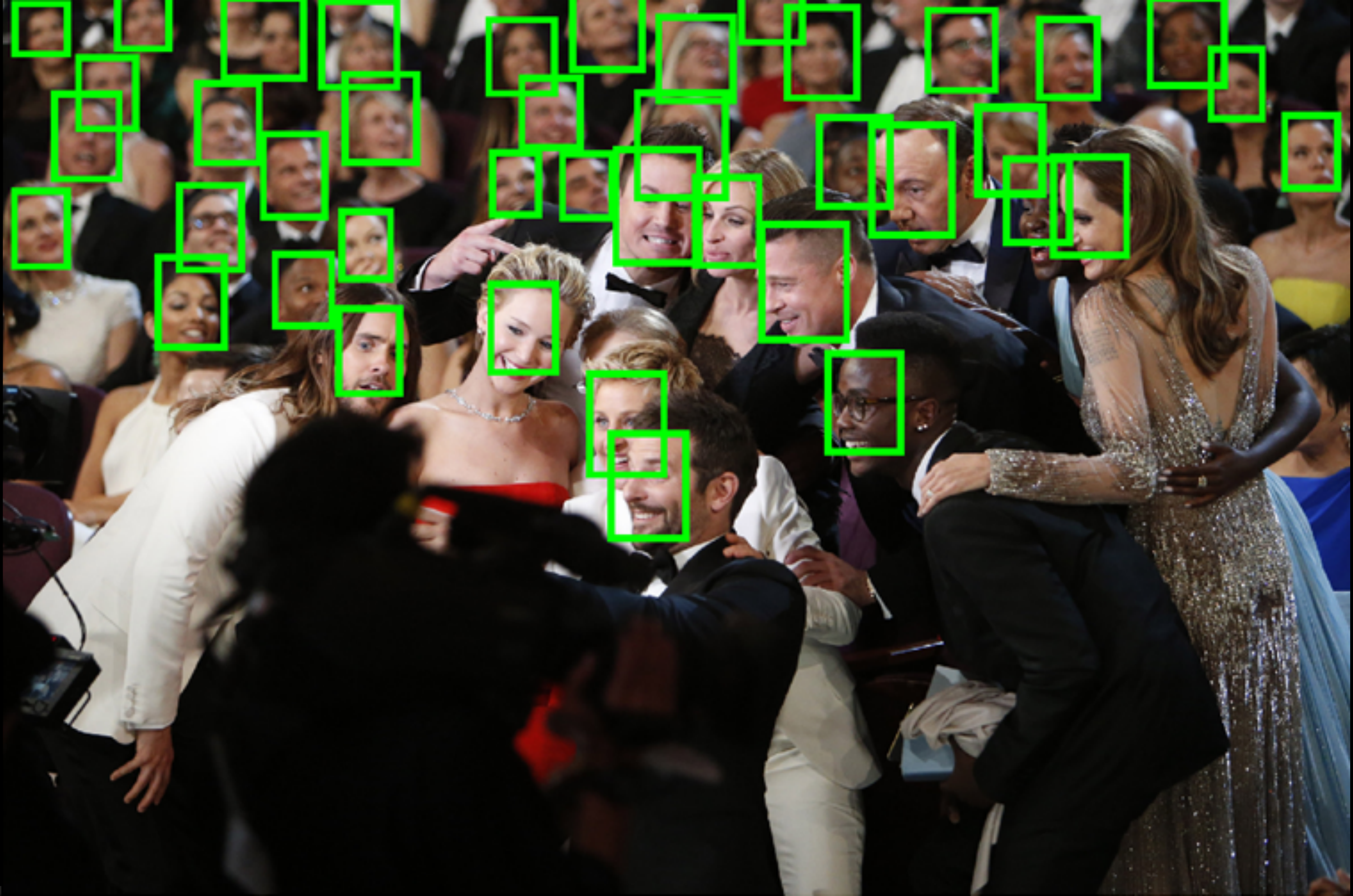
Some observations

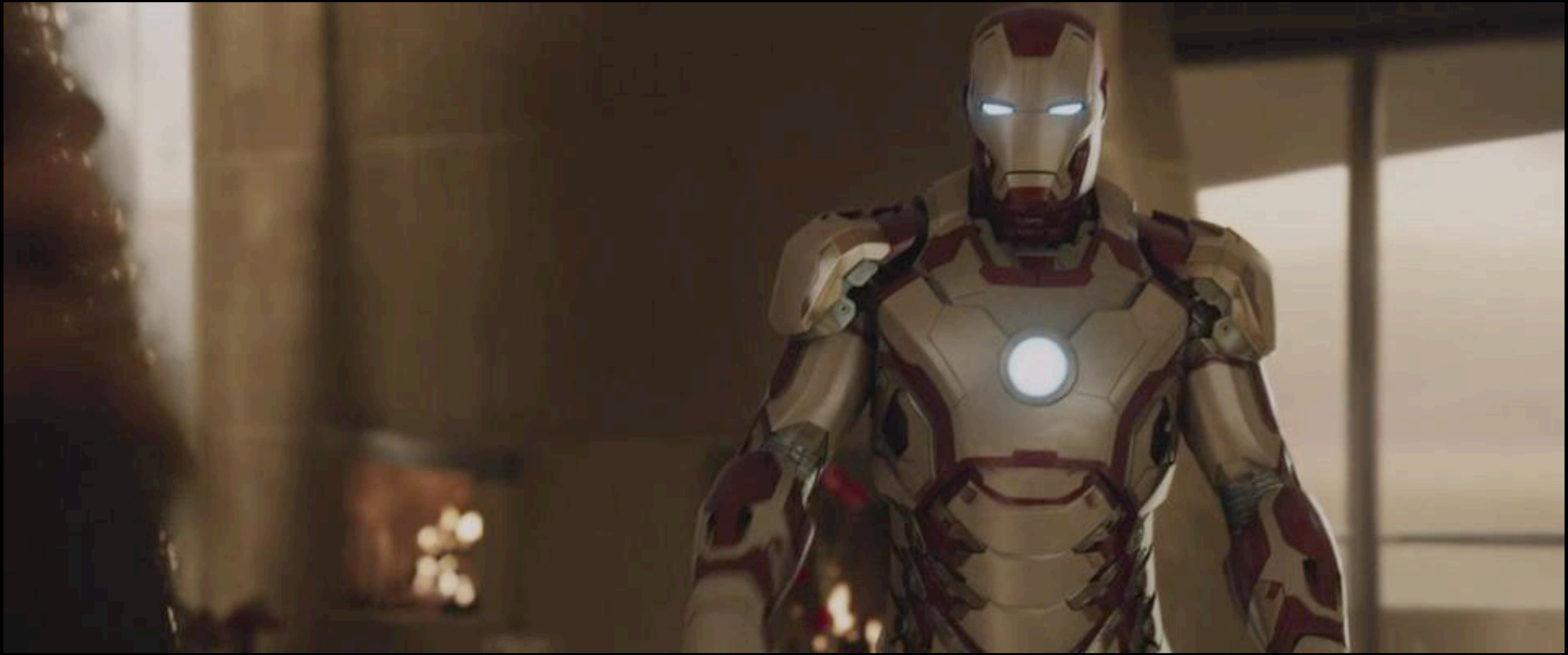


Why deep learning works so well?

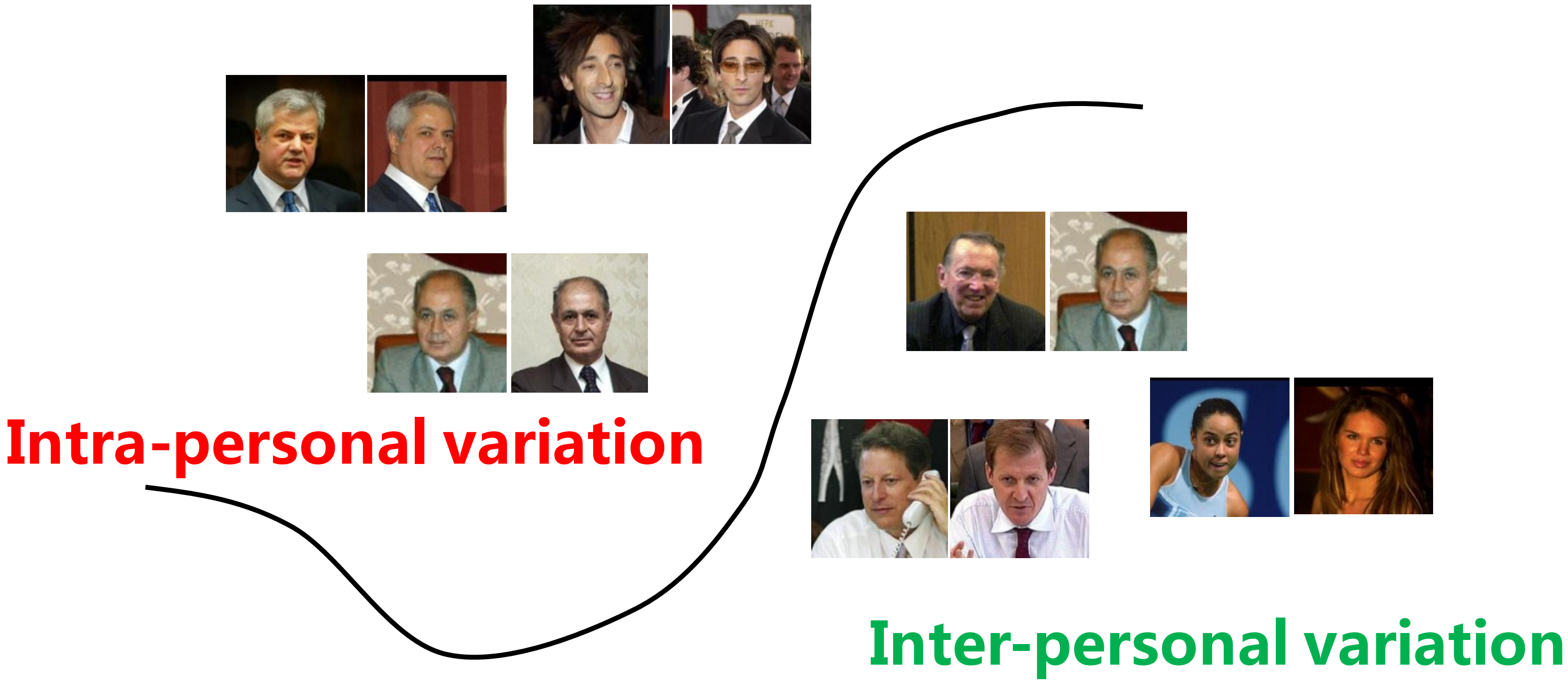
- **Local minima do not arise in very high dimensional space**, so greedy-search gradient optimization is not trapped in a "box”
- With distributed representations, it is possible to represent exponential number of regions with a linear number of parameters. **Multiple layers help to implement complex functions more concisely.**

Bengio et al., Identifying and attacking the saddle point problem in high-dimensional non-convex optimization, 2014
LeCun et. al., The Loss Surfaces of Multilayer Networks, 2015
Goodfellow et al., Qualitatively characterizing neural network optimization problems, 2015



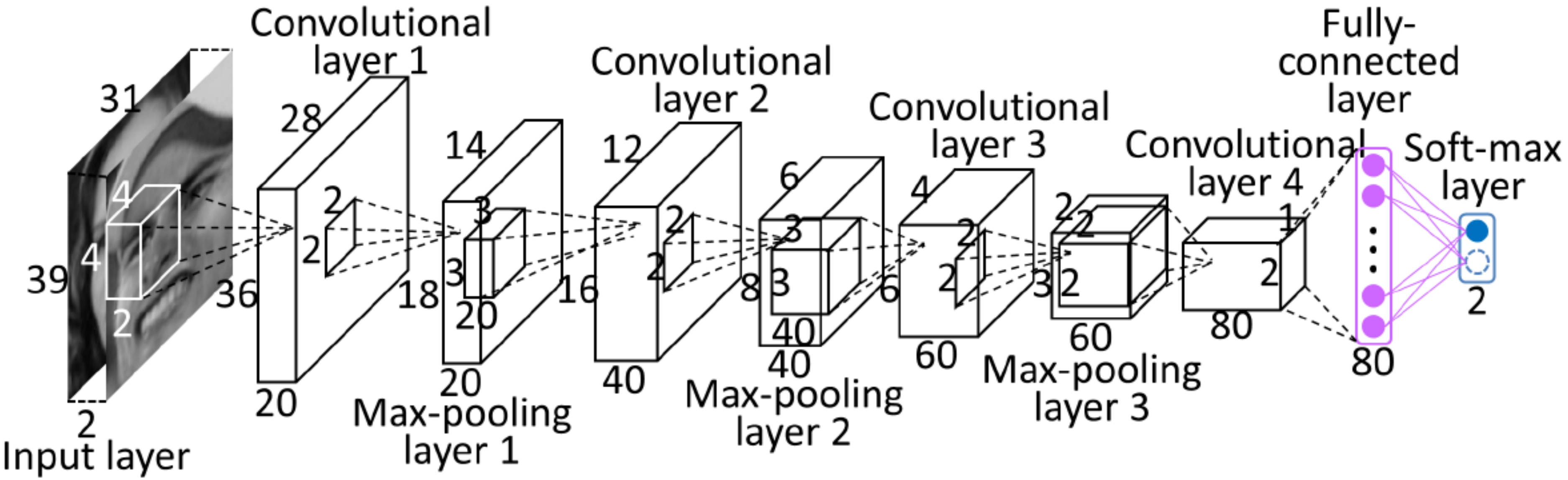


Eternal topic on face recognition



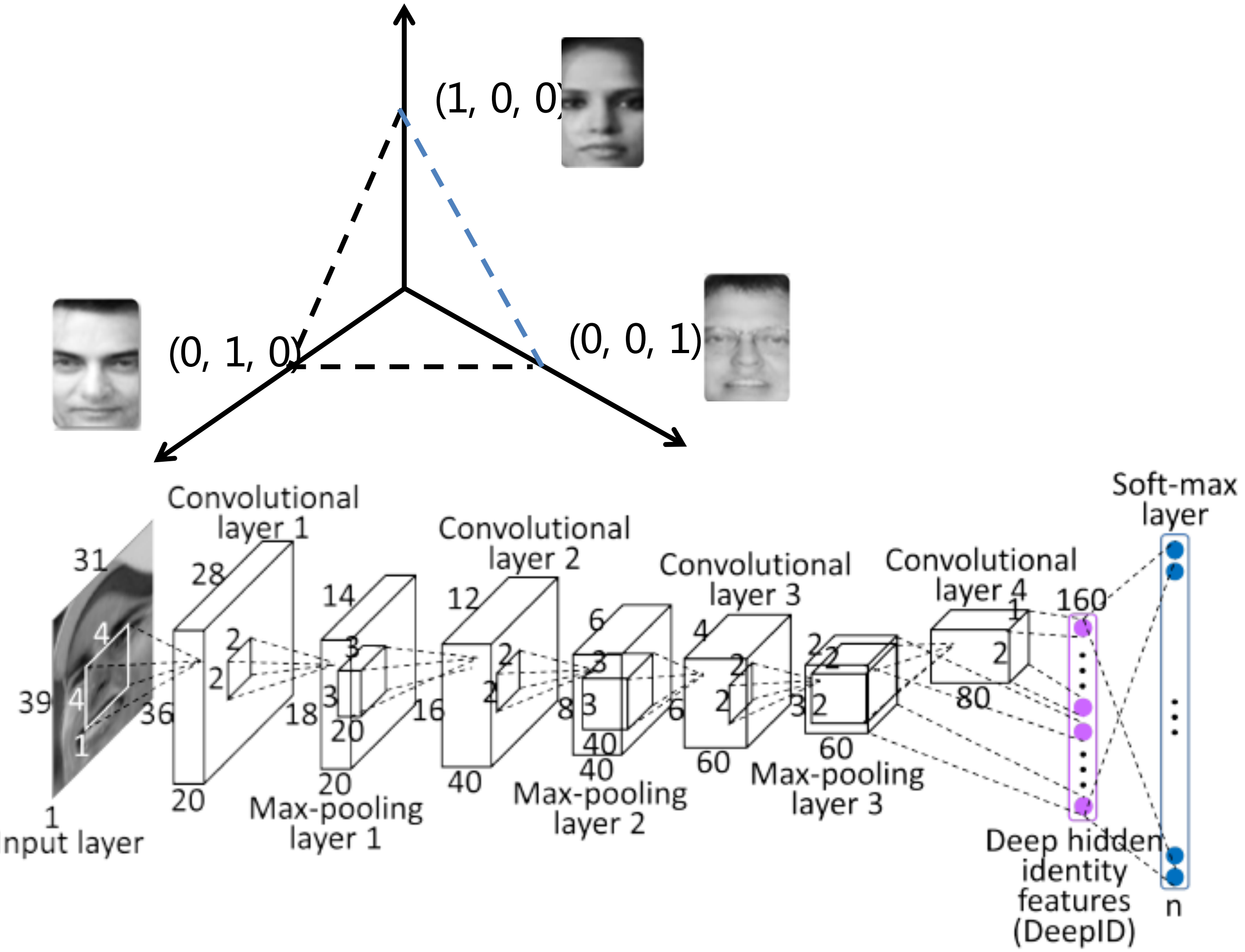
How to separate the two types of variations?

Learn identity features with verification signal

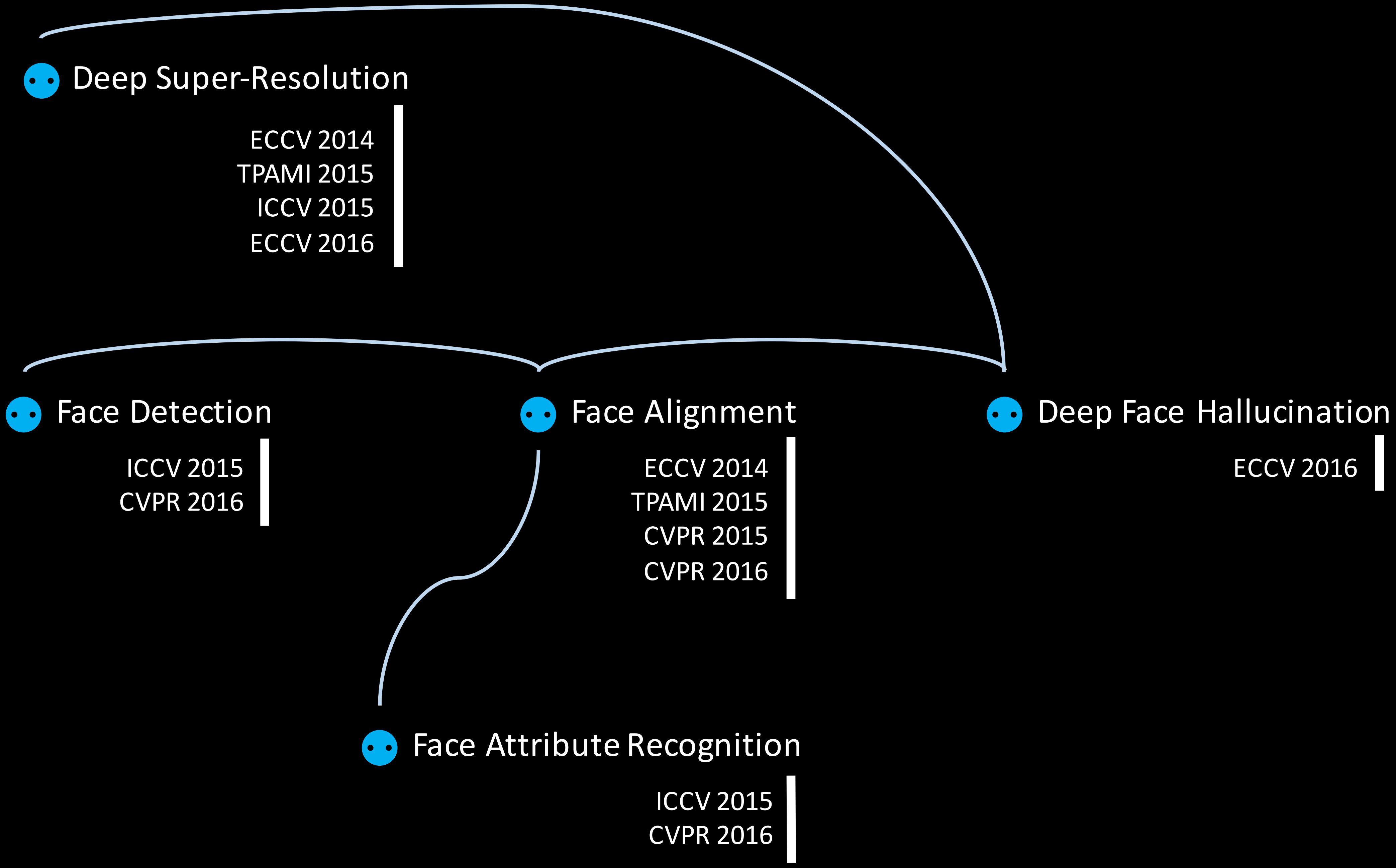


Y. Sun, X. Wang, and X. Tang, "Hybrid Deep Learning for Computing Face Similarities," Proc. ICCV, 2013.

DeepID: Learn identity features with identification signal



Y. Sun, X. Wang, and X. Tang, "Deep Learning Face Representation from Predicting 10,000 classes," Proc. CVPR, 2014.

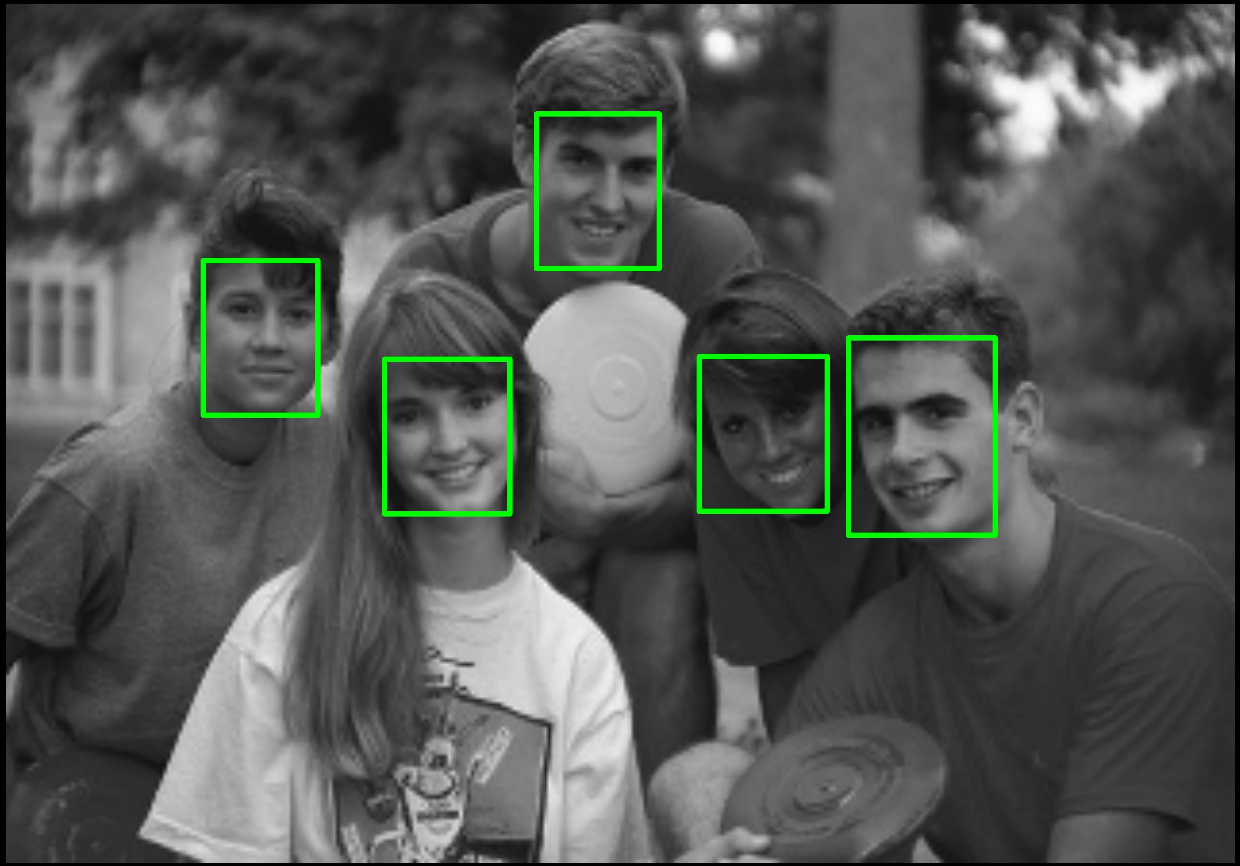

Deep Super-Resolution
ECCV 2014
TPAMI 2015
ICCV 2015
ECCV 2016
Face Detection
ICCV 2015
CVPR 2016
Face Alignment
ECCV 2014
TPAMI 2015
CVPR 2015
CVPR 2016
Deep Face Hallucination
ECCV 2016
Face Attribute Recognition
ICCV 2015
CVPR 2016

Face Detection

WIDER FACE: A Face Detection Benchmark
S. Yang, P. Luo, C. C. Loy, X. Tang
in Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2016


WIDER FACE

Face Detection Dataset



- **Basic information**
  - 130 images
  - 507 faces
- **Characteristic**
  - Gray-scale, mostly frontal
- **Methods**
  - Viola-Jones detector. IJCV 2001.
  - Assembly of part detector. In ECCV 2004.

1998

MIT+CMU

Face Detection Dataset

- **Basic information**
  - 2845 images
  - 5171 faces
- **Characteristic**
  - Mostly celebrity face.
- **Methods**
  - Domain Adaptation of a Cascade of Classifiers. CVPR 2011.
  - Detecting and Aligning Faces by Image Retrieval. CVPR 2013.

1998
MIT+CMU

2010
FDDB

Face Detection Dataset

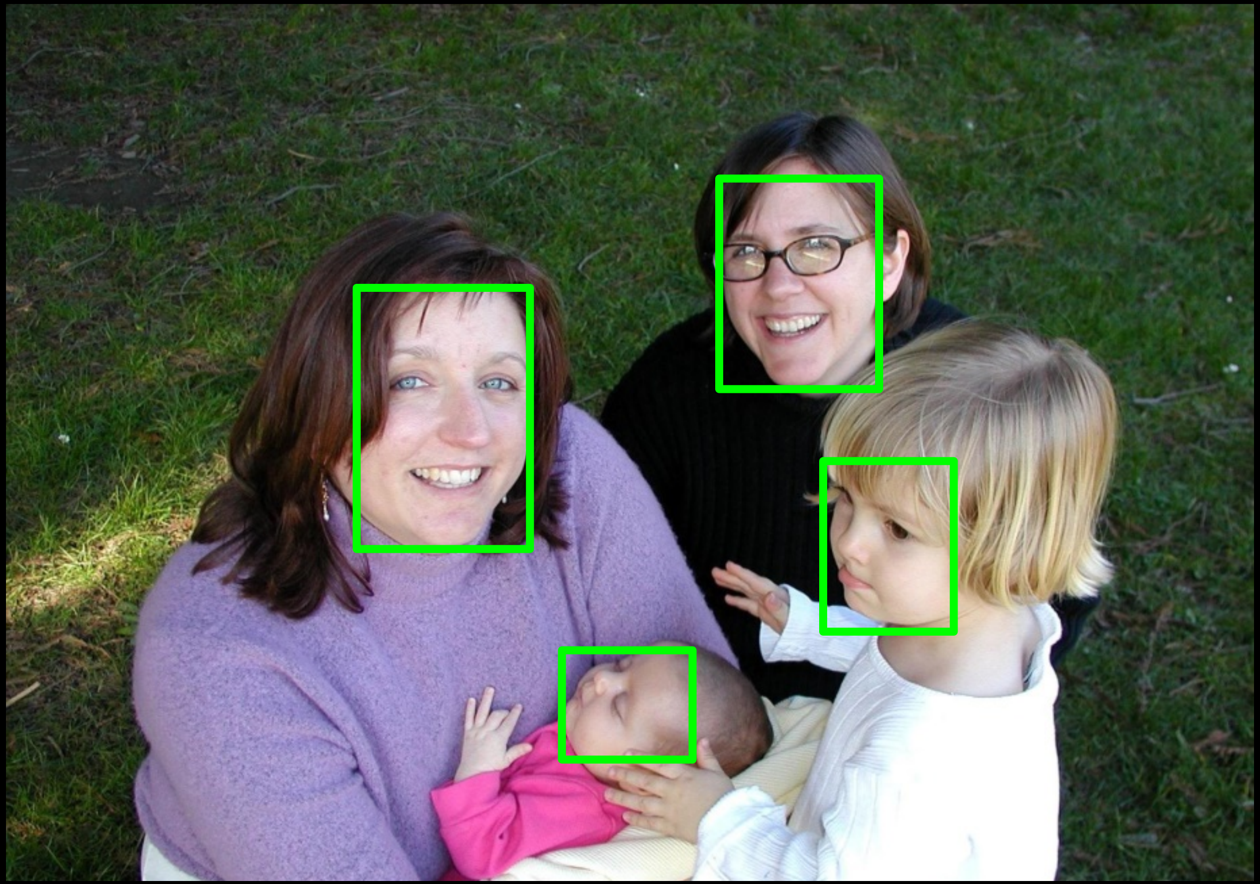
- **Basic information**
  - 851 images
  - 1,335 faces
- **Characteristic**
  - Most of image has only one face.
- **Methods**
  - Tree Parts Model. CVPR, 2012.

1998 — MIT+CMU

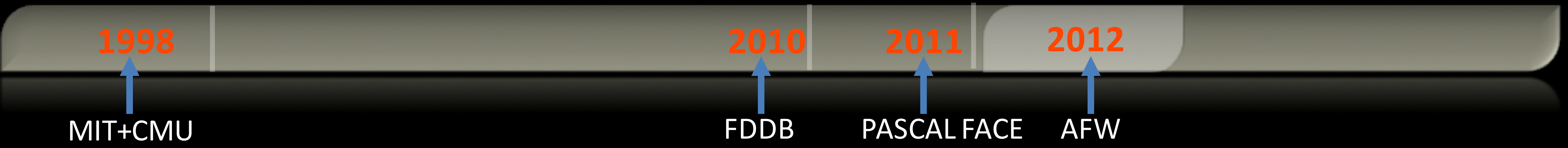
2010 — FDDB

2011 — PASCAL FACE

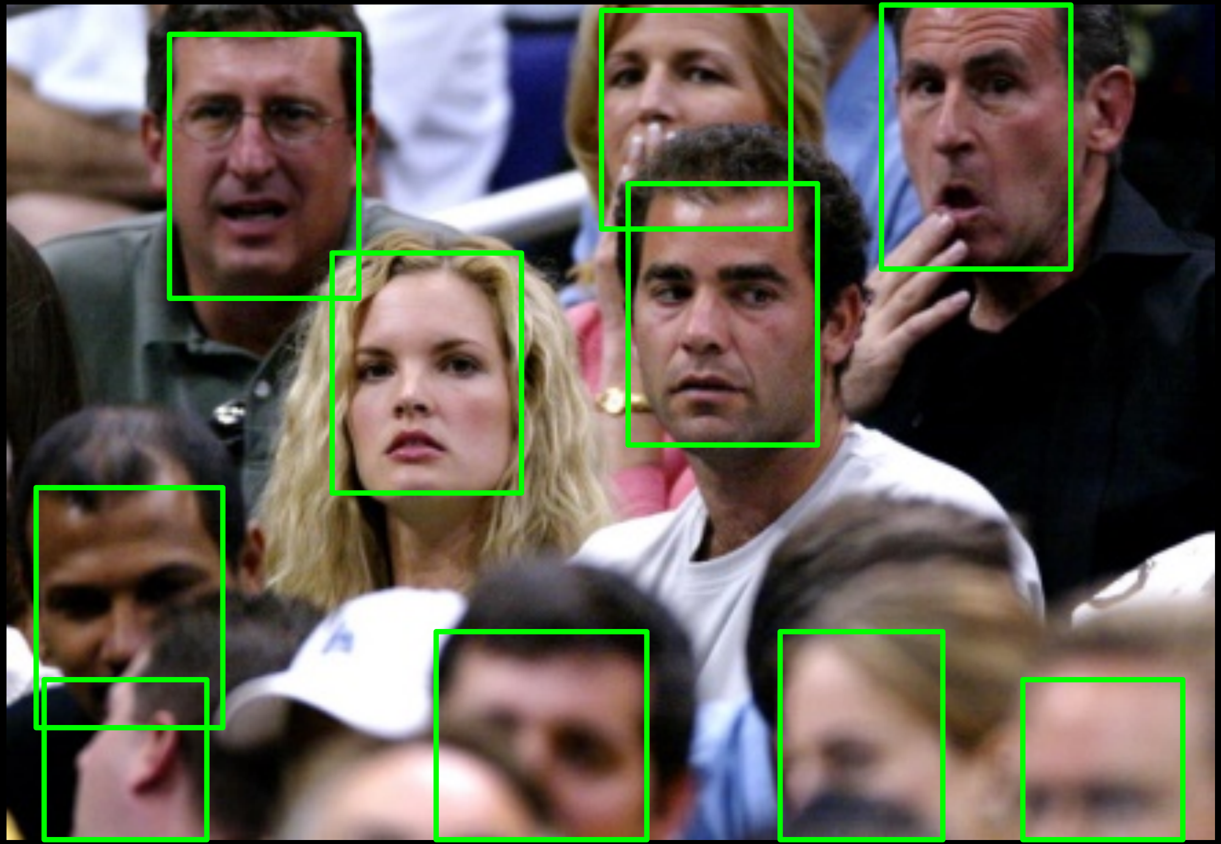
Face Detection Dataset



- **Basic information**
  - 205 images
  - 468 faces
- **Characteristic**
  - Background is less clutter.
- **Methods**
  - Boosted Exemplar. CVPR, 2014.
  - Joint Cascade. ECCV. 2014.



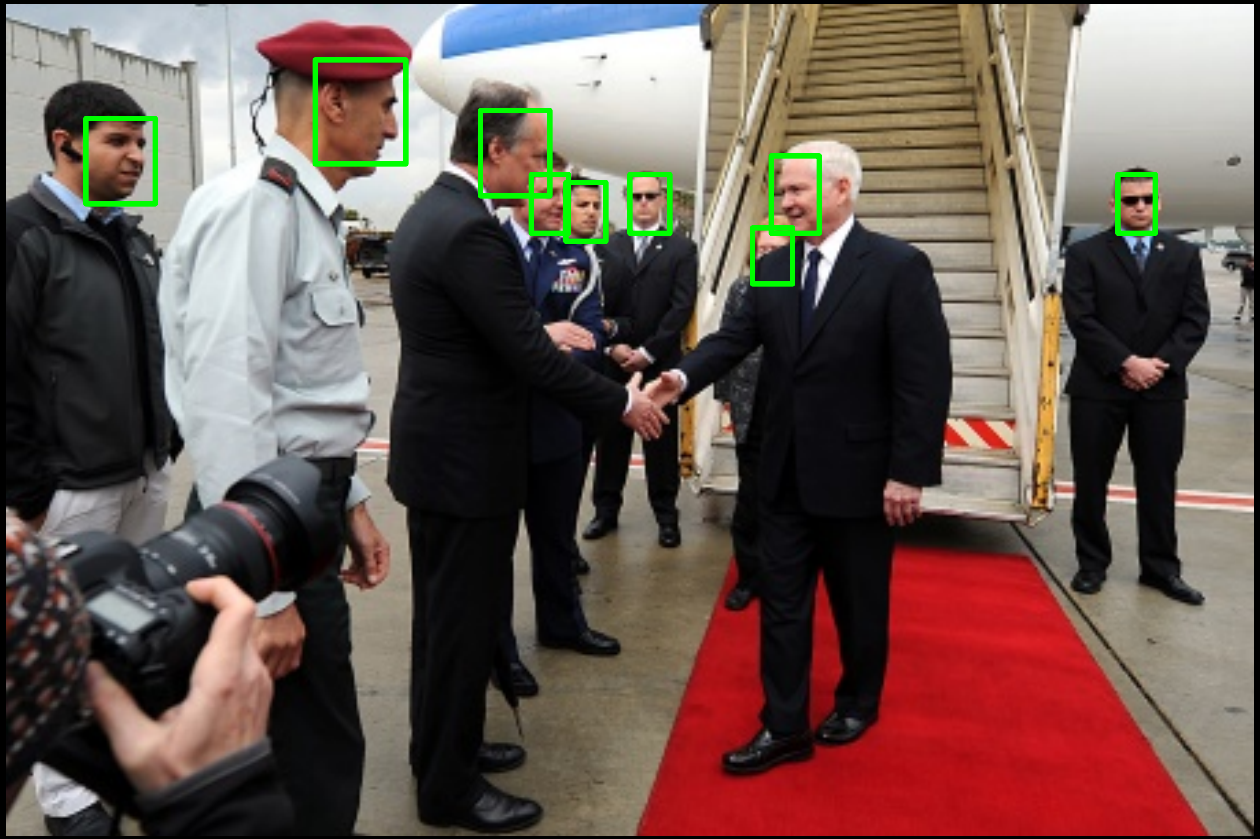
Face Detection Dataset



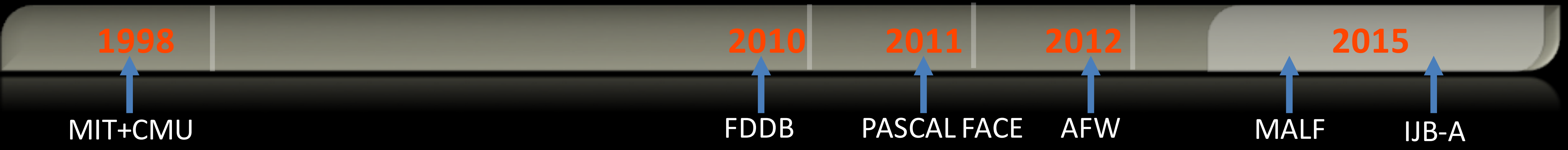
- **Basic information**
  - 5,250 images
  - 11,931 faces
- **Characteristic**
  - Most of faces in large or medium scale.
- **Methods**
  - HeadHunter. ECCV. 2014.
  - Multi-view CNN. ICMR, 2015.

1998 — MIT+CMU
2010 — FDDB
2011 — PASCAL FACE
2012 — AFW
2015 — MALF

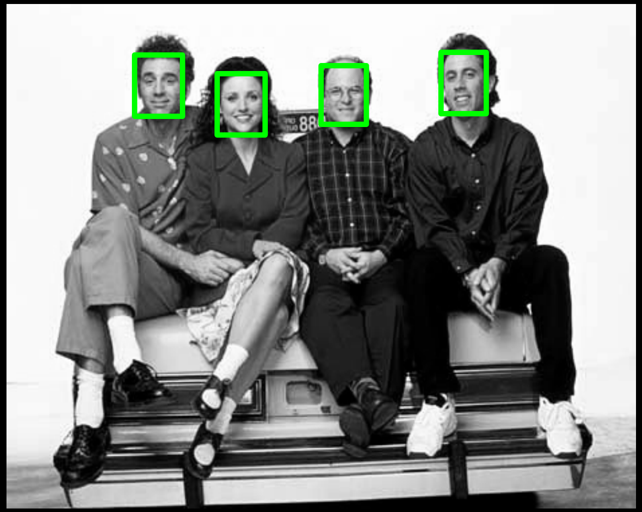
Face Detection Dataset

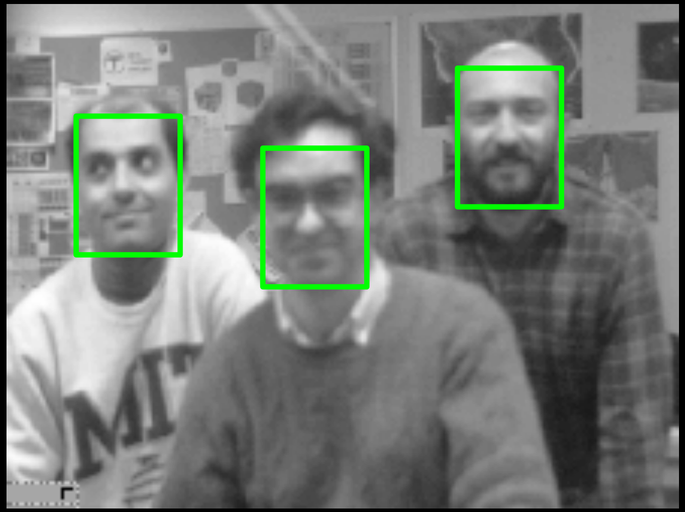


- **Basic information**
  - 24,327 images
  - 49,759 faces
- **Characteristic**
  - Large number of video frames, highly redundant.
- **Methods**
  - Compact Cascade CNN. arXiv. 2015
  - Faster R-CNN. arXiv. 2016

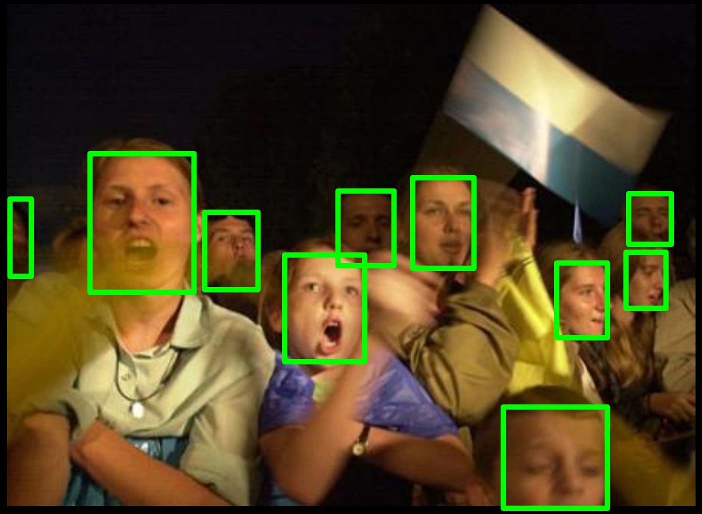


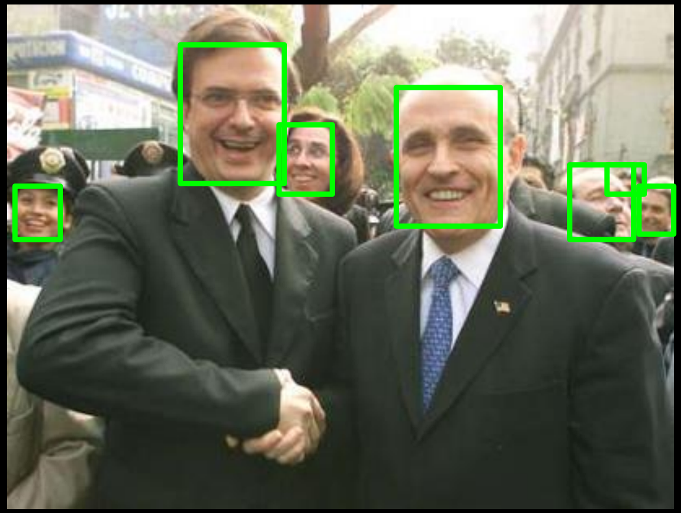
Diversity



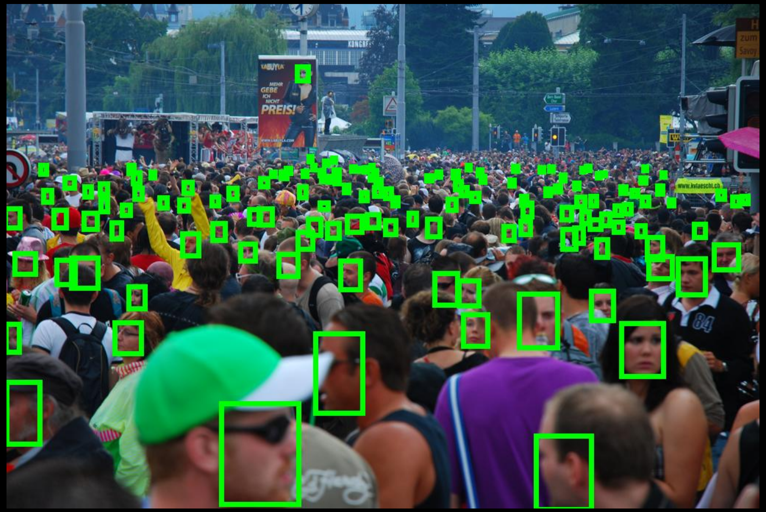


MIT+CMU



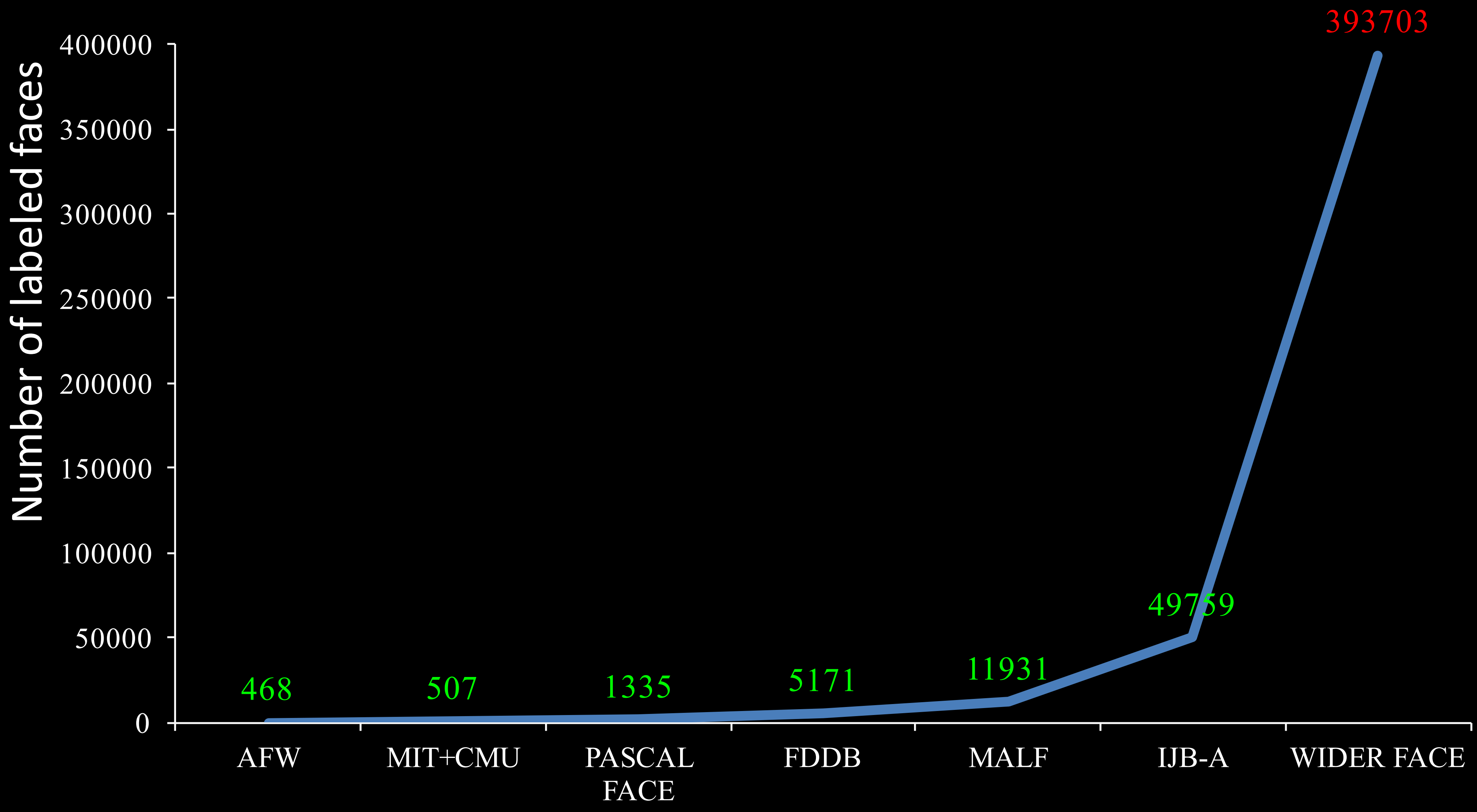


FDDB



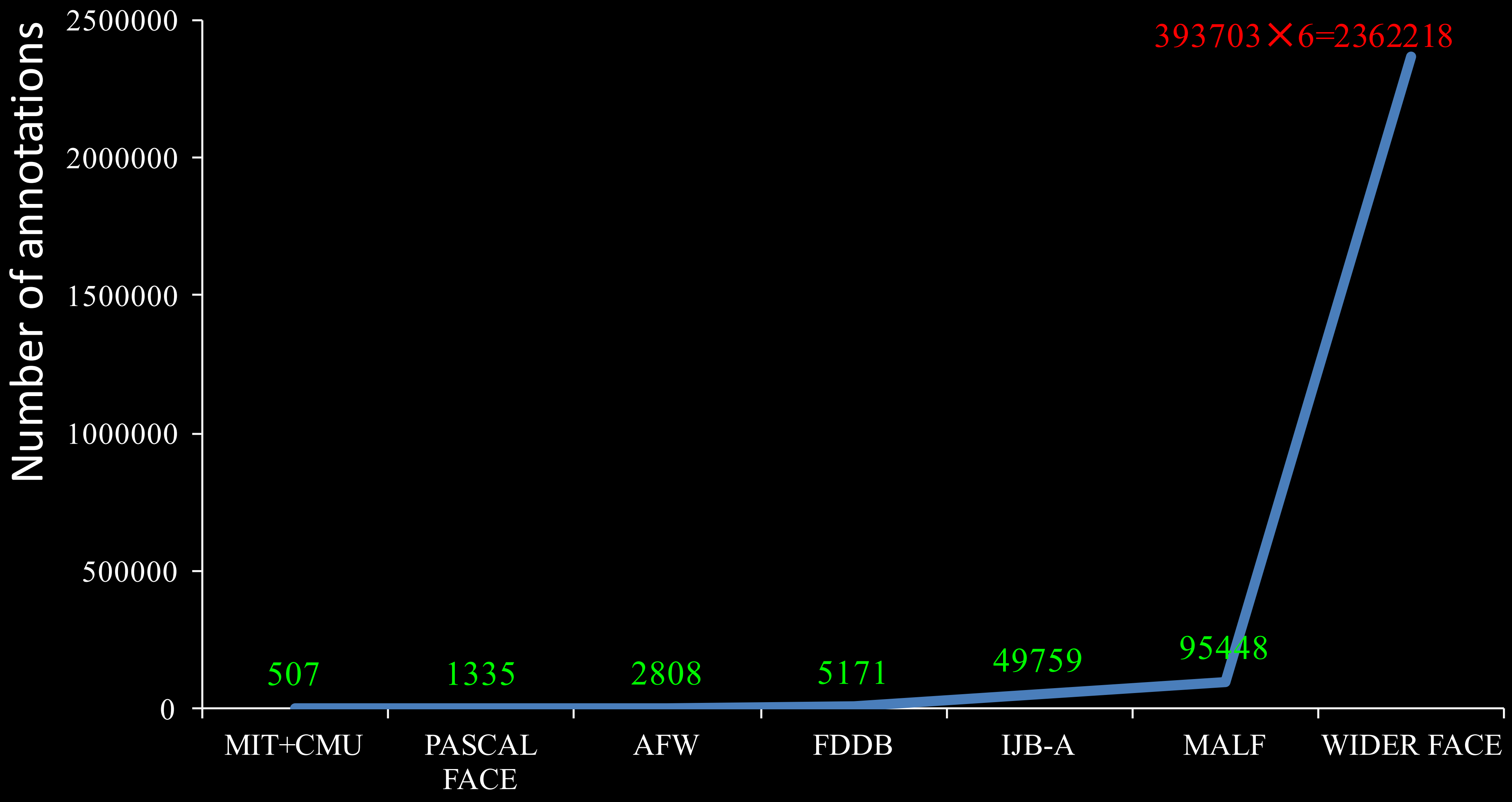
WIDER FACE

Data scale

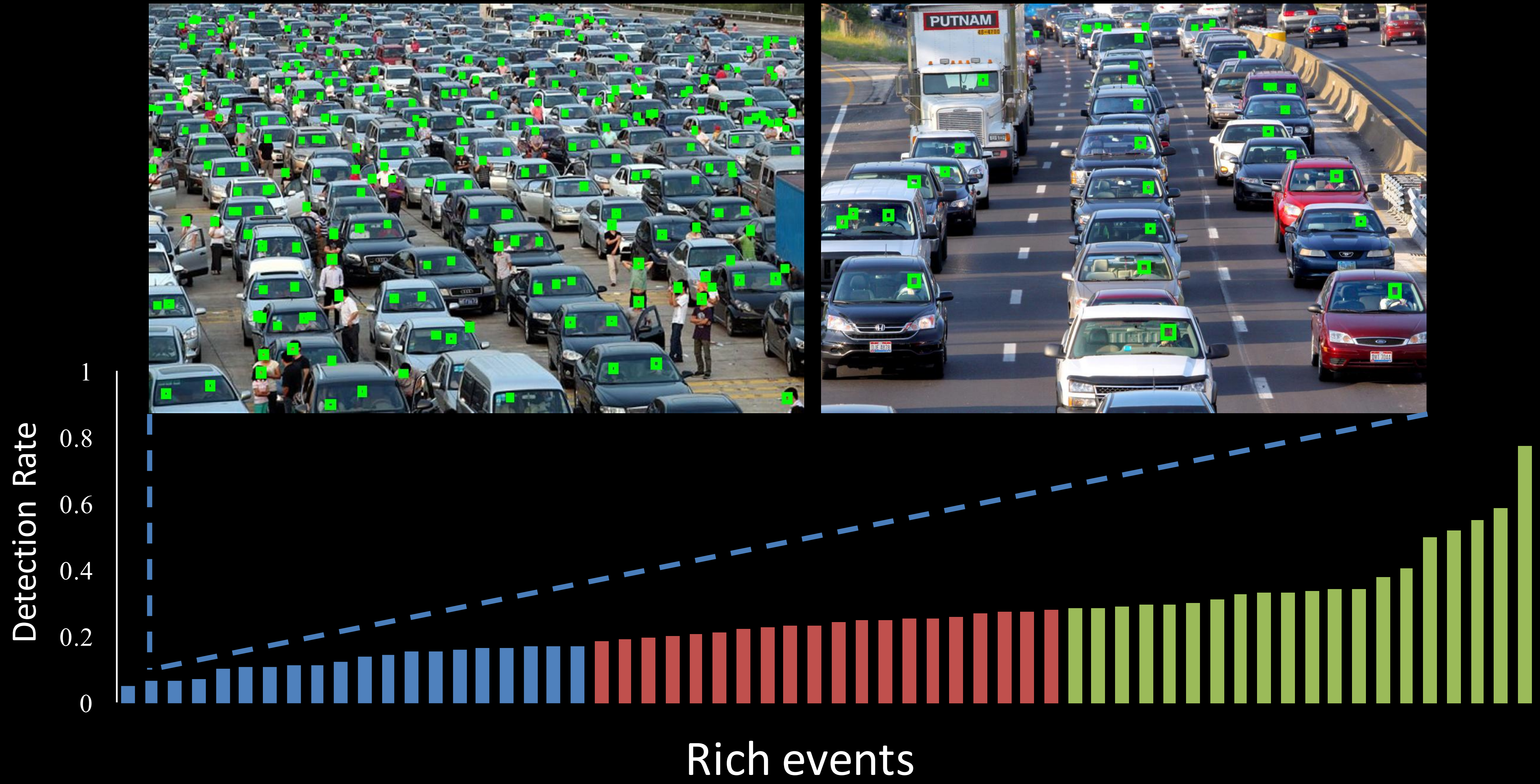


Number of labeled faces
400000
350000
300000
250000
200000
150000
100000
50000
0
393703
49759
11931
5171
1335
507
468
AFW
MIT+CMU
PASCAL FACE
FDDB
MALF
IJB-A
WIDER FACE

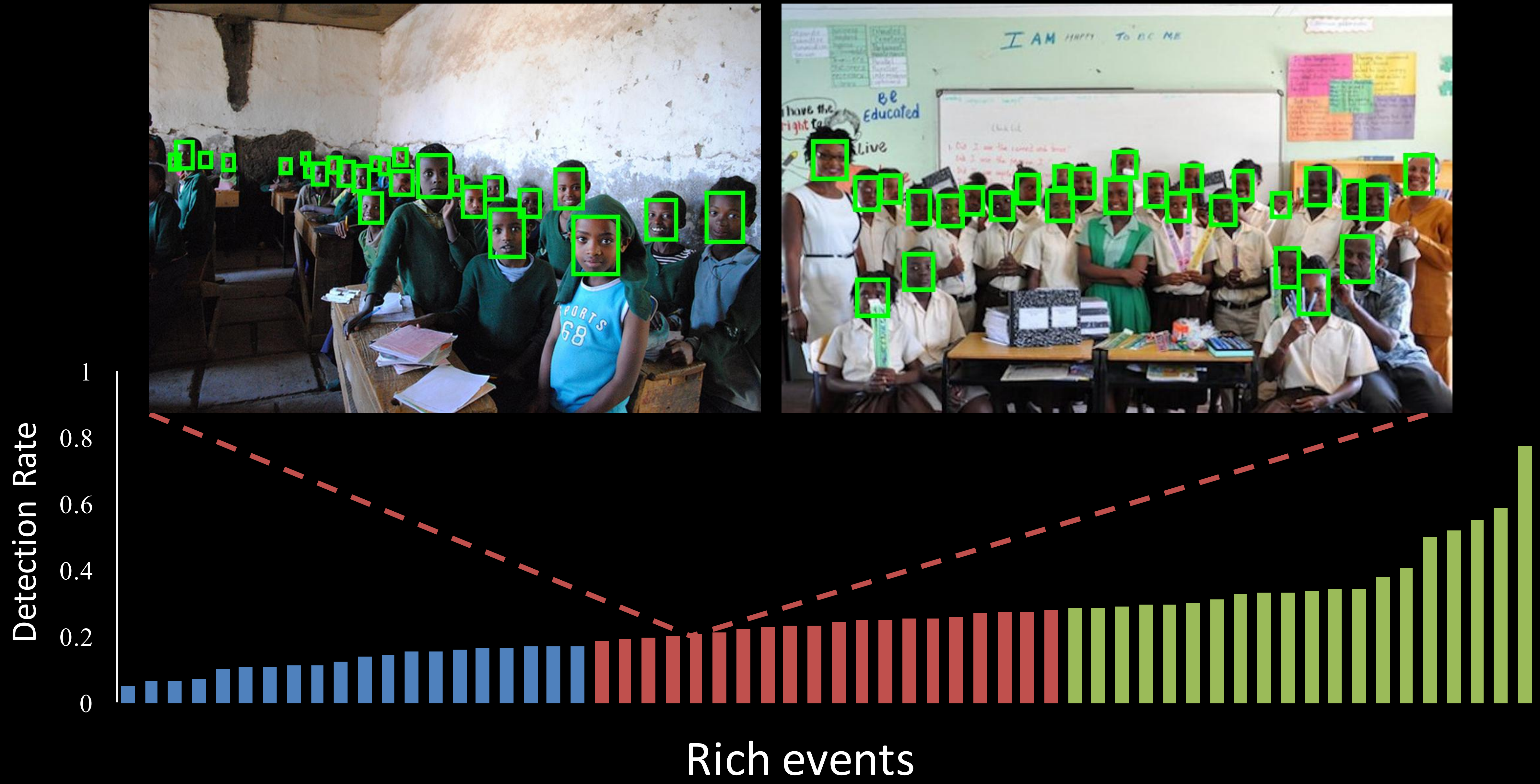
Richer annotations



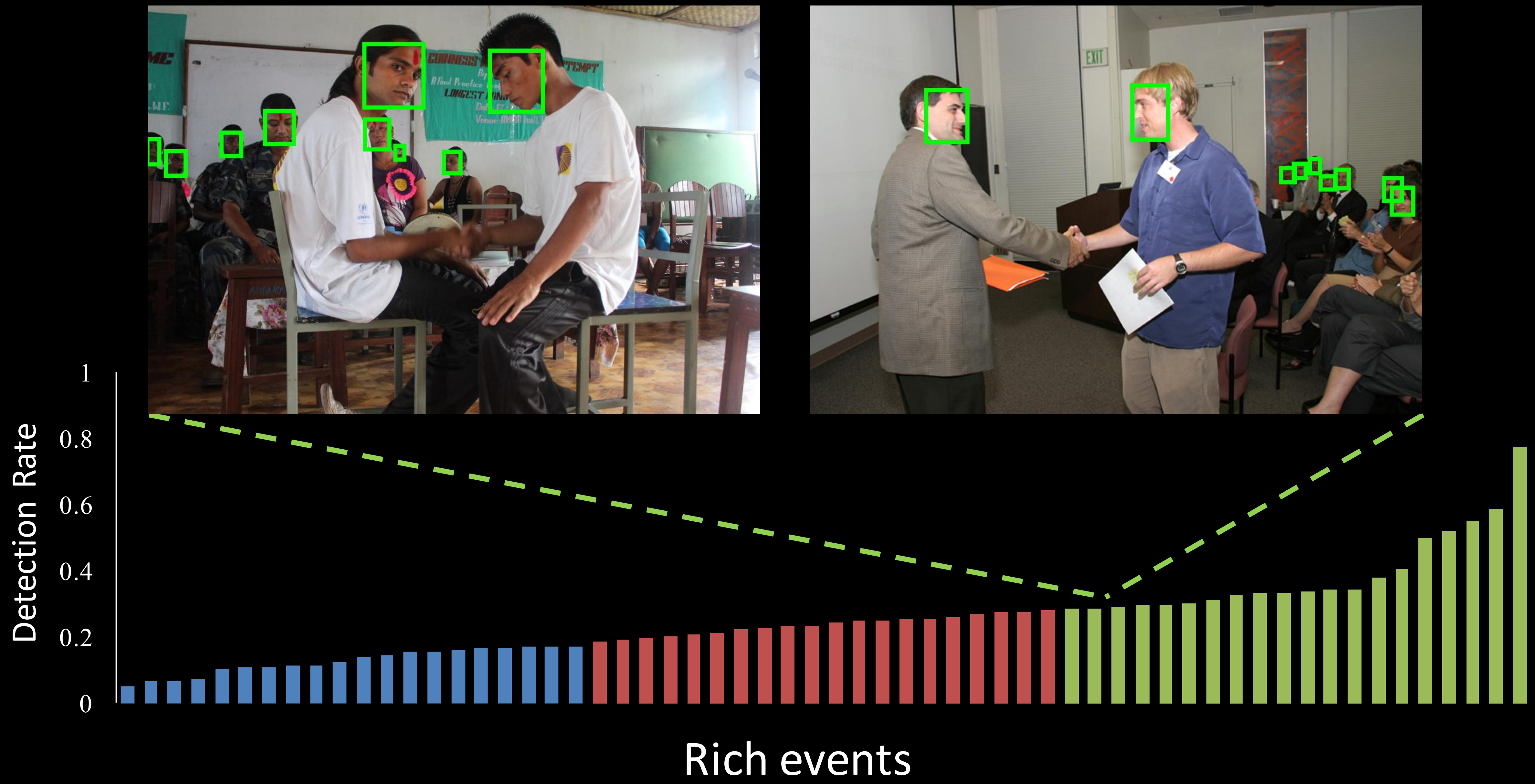
Traffic



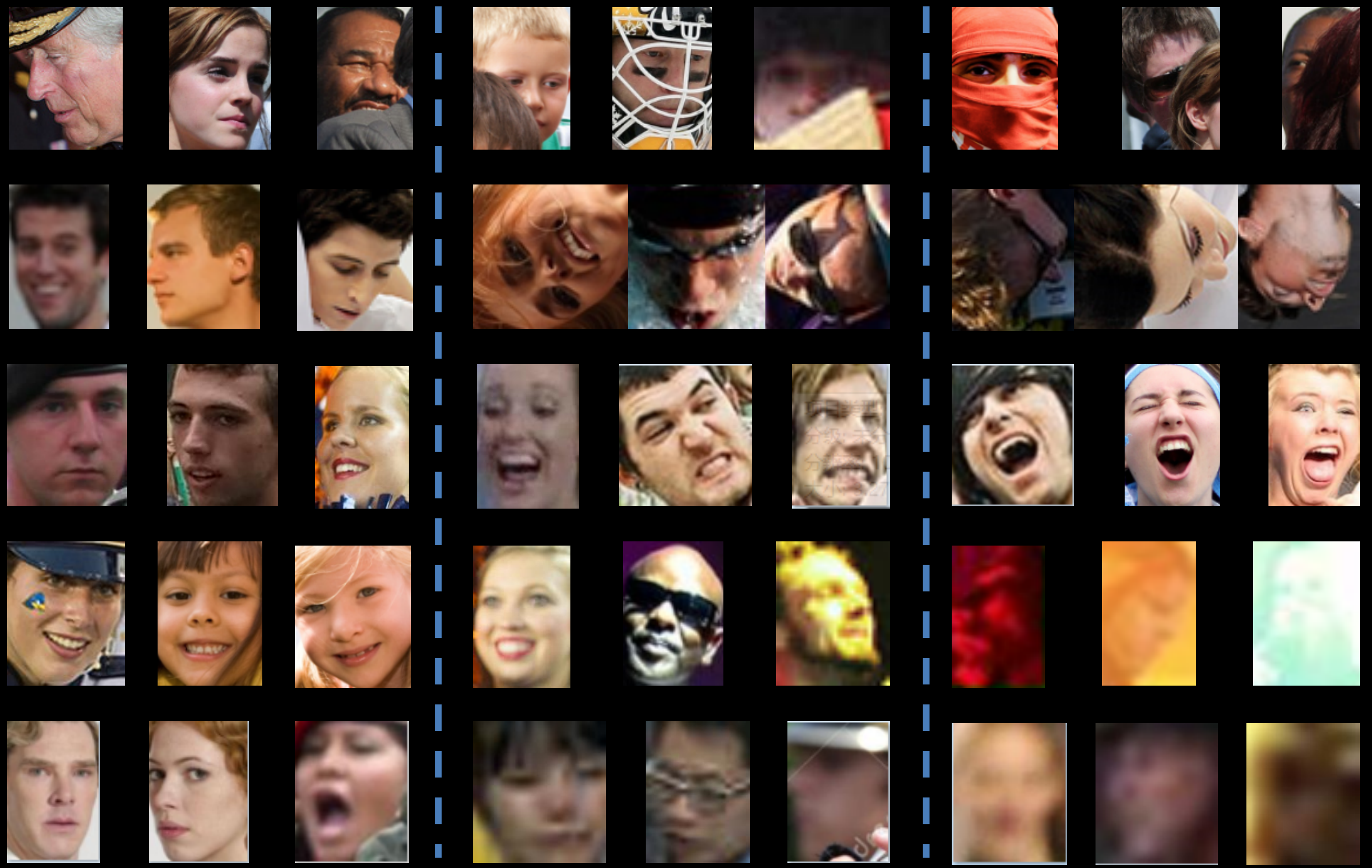
Students Schoolkids



Handshaking



Rich label annotations



Occlusion

Pose

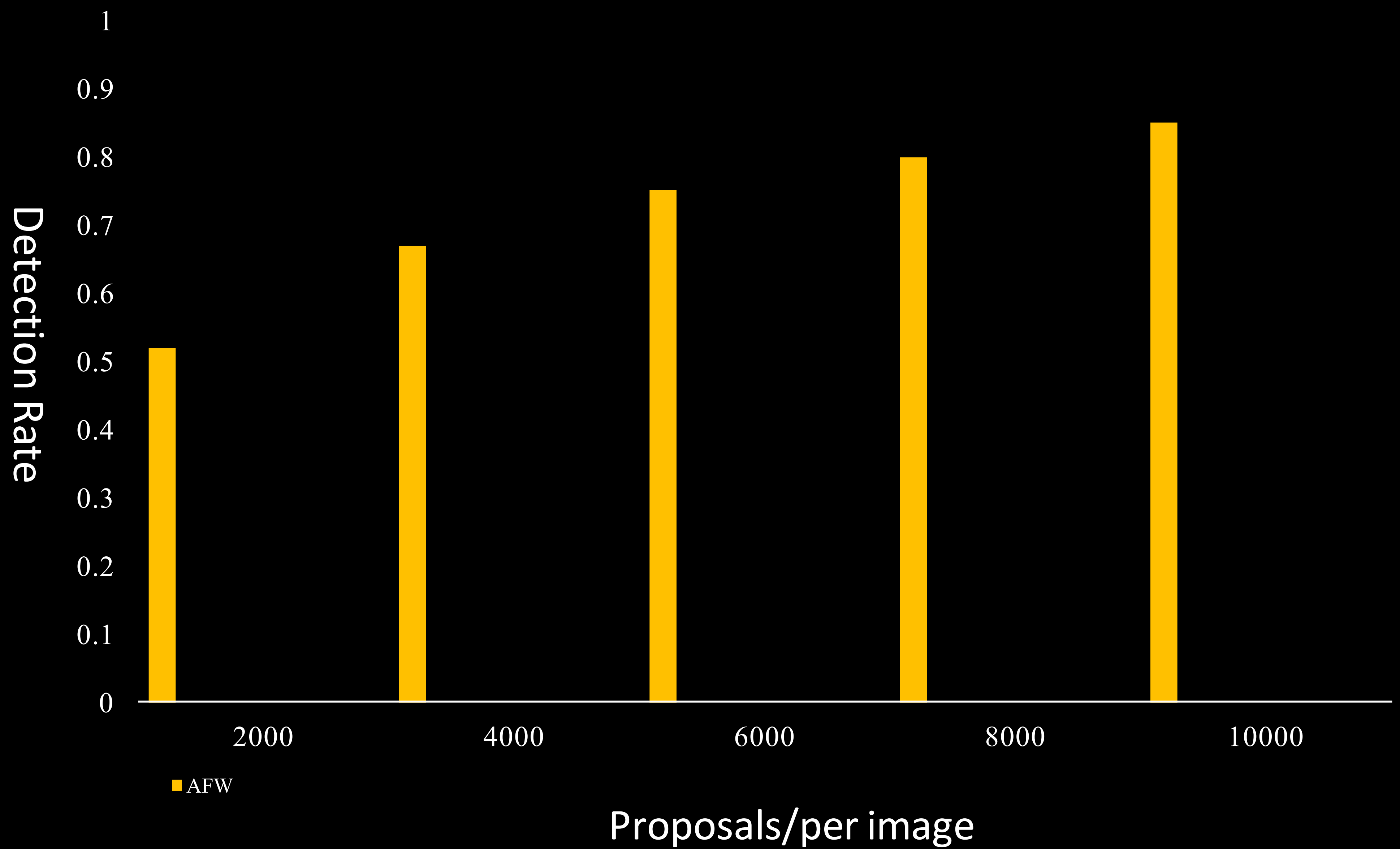
Expression

Illumination

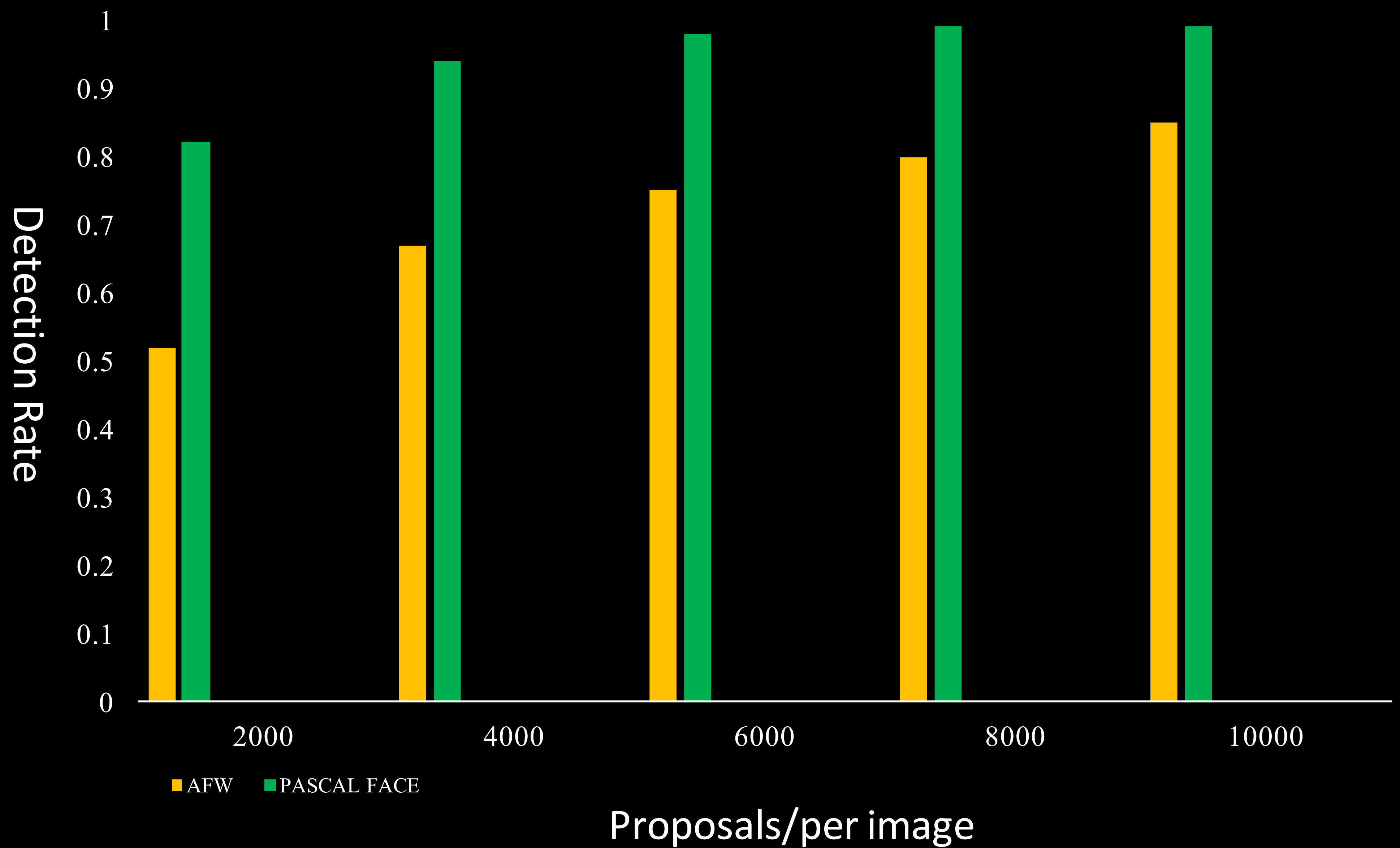
Blur

Normal | Intermediate | Extreme

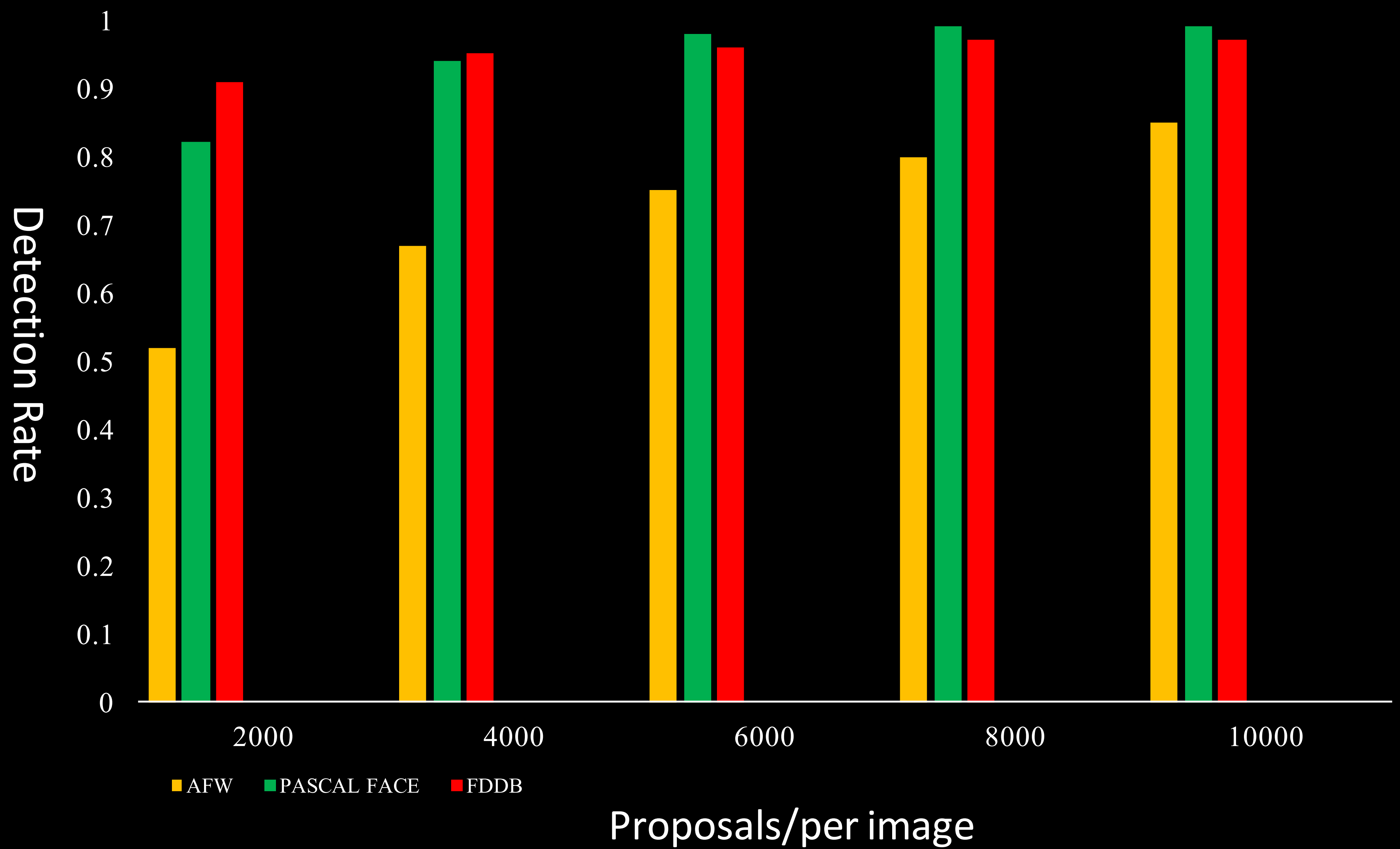
WIDER FACE is more challenging



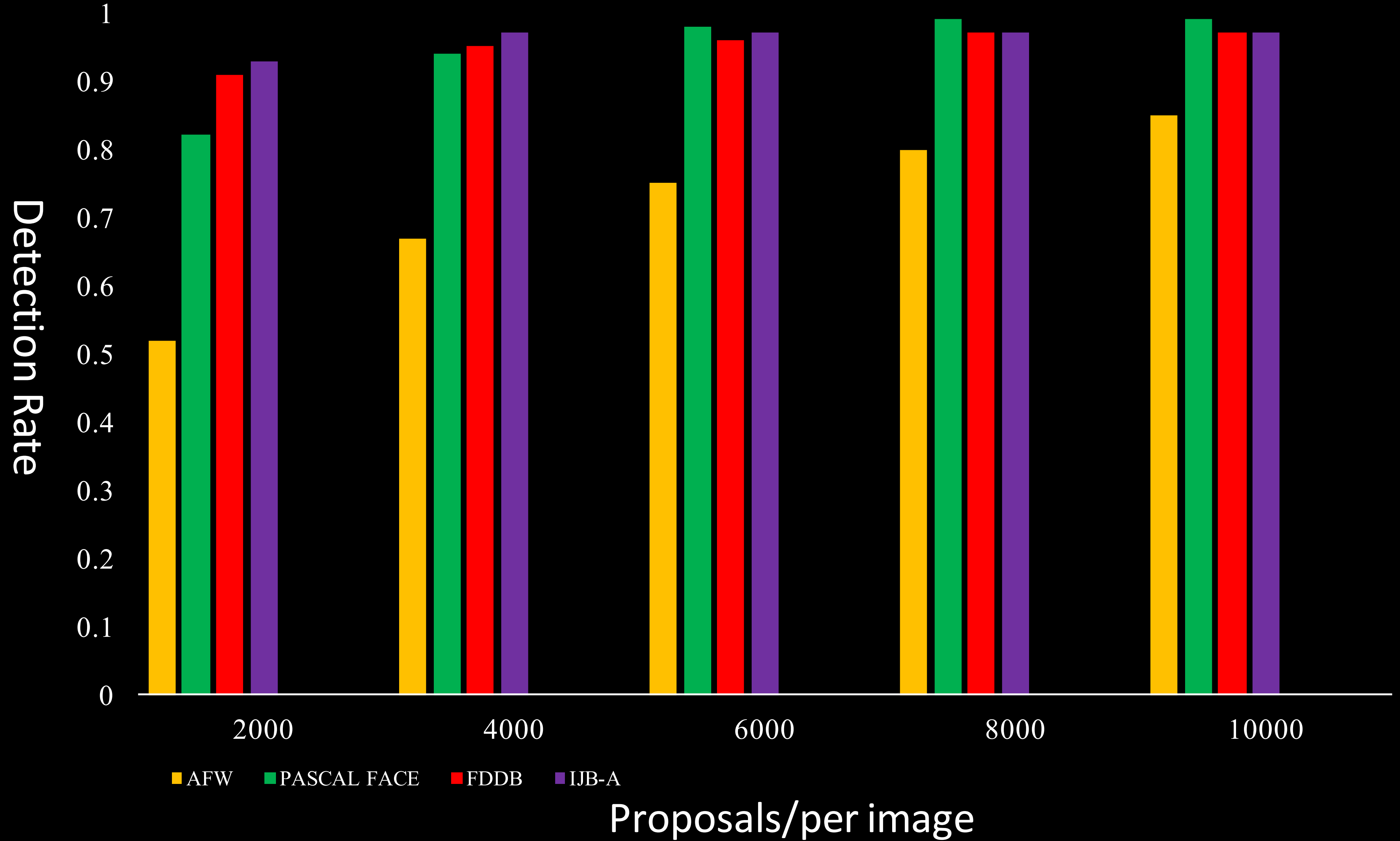
WIDER FACE is more challenging



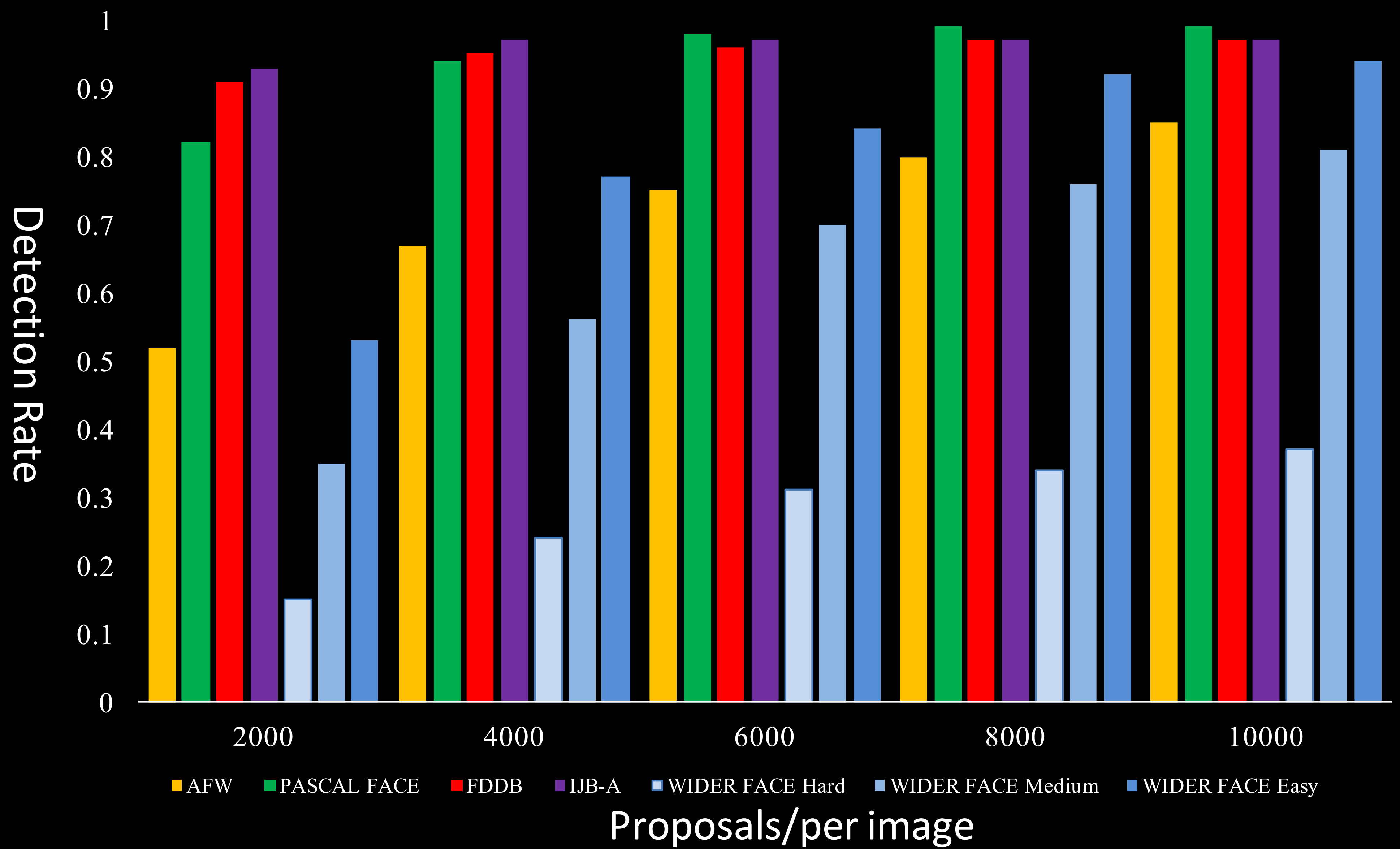
WIDER FACE is more challenging

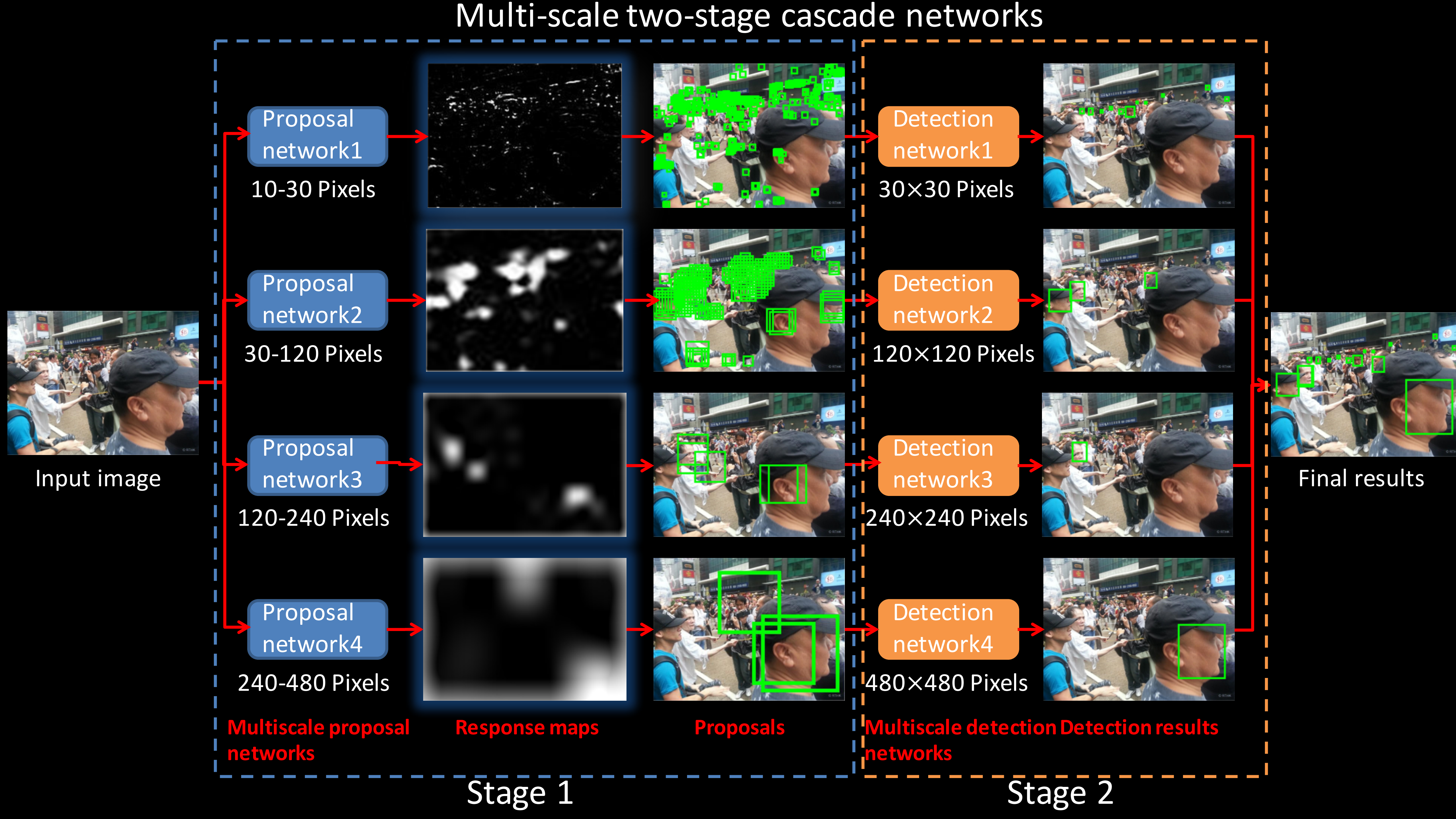


WIDER FACE is more challenging



WIDER FACE is more challenging



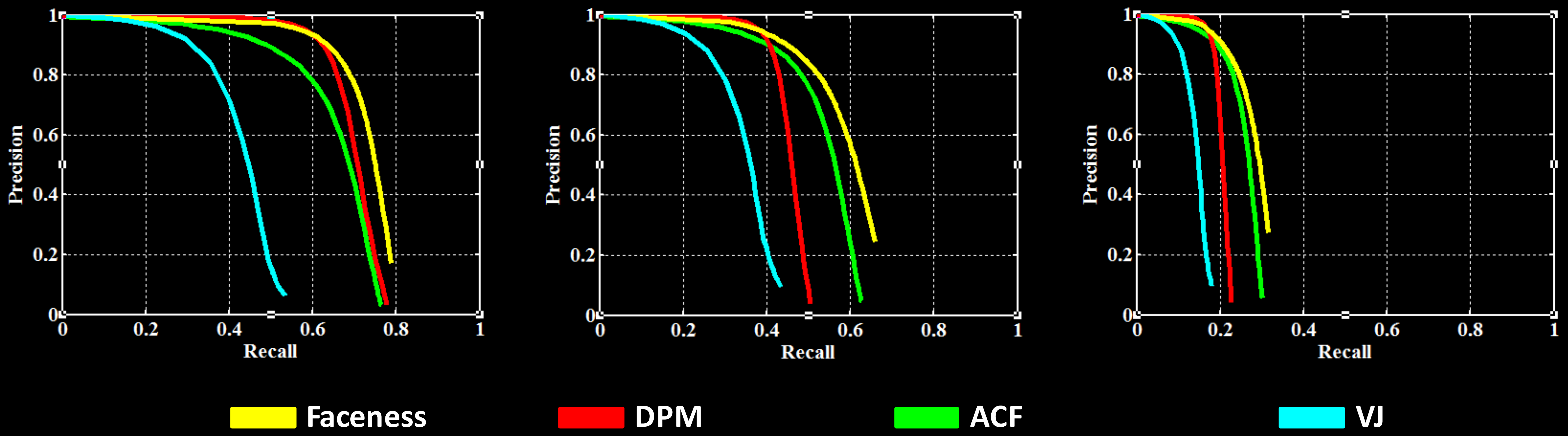


Multi-scale two-stage cascade networks
Input image
Proposal network1
10-30 Pixels
Proposal network2
30-120 Pixels
Proposal network3
120-240 Pixels
Proposal network4
240-480 Pixels
Multiscale proposal networks
Response maps
Proposals
Detection network1
30×30 Pixels
Detection network2
120×120 Pixels
Detection network3
240×240 Pixels
Detection network4
480×480 Pixels
Multiscale detection networks
Detection results
Final results
Stage 1
Stage 2

WIDER FACE for testing

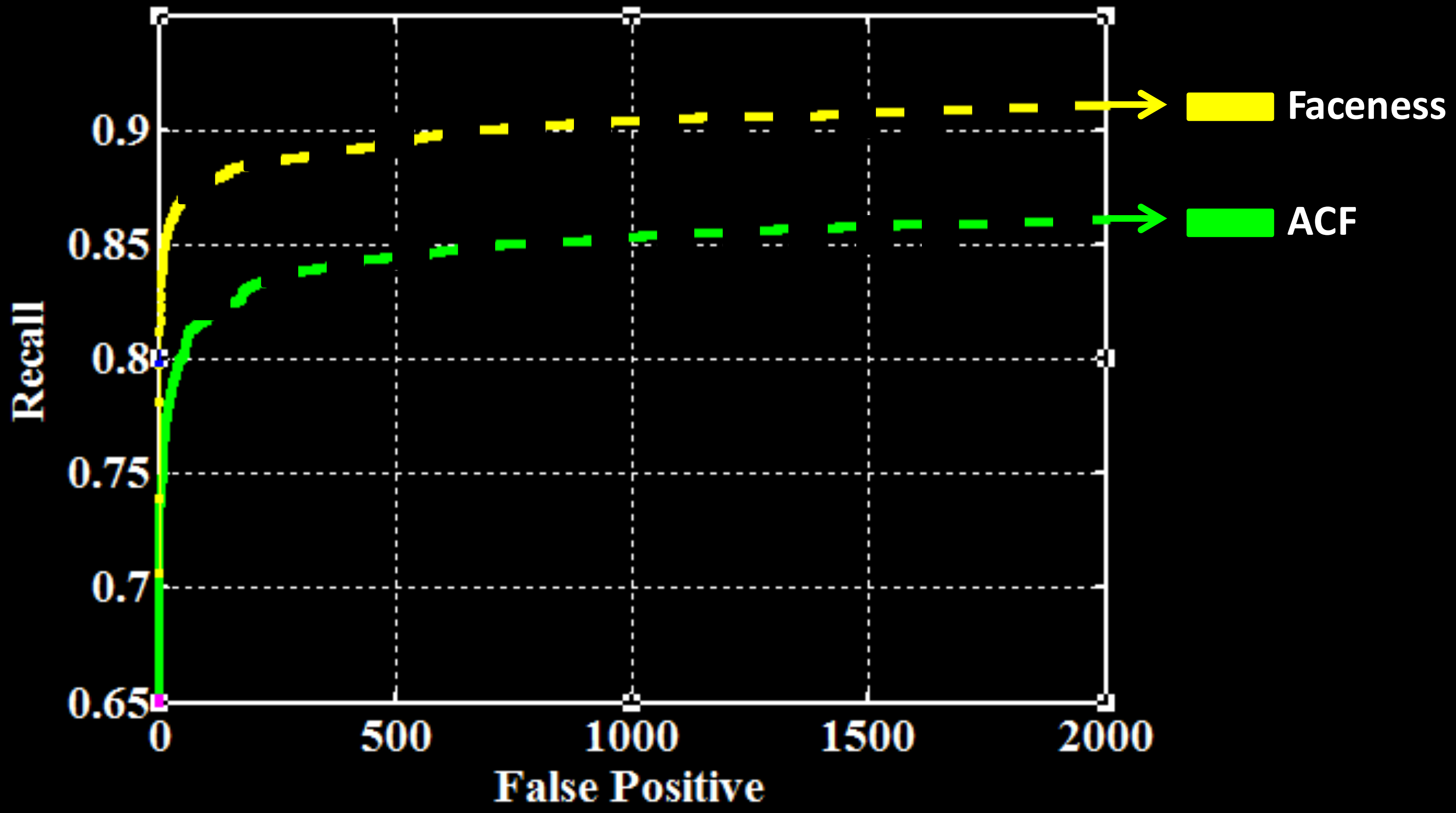
A face detector is trained using external data, and tested on the WIDER FACE test partition.

(a) Easy set **(b) Medium set** **(c) Hard set**



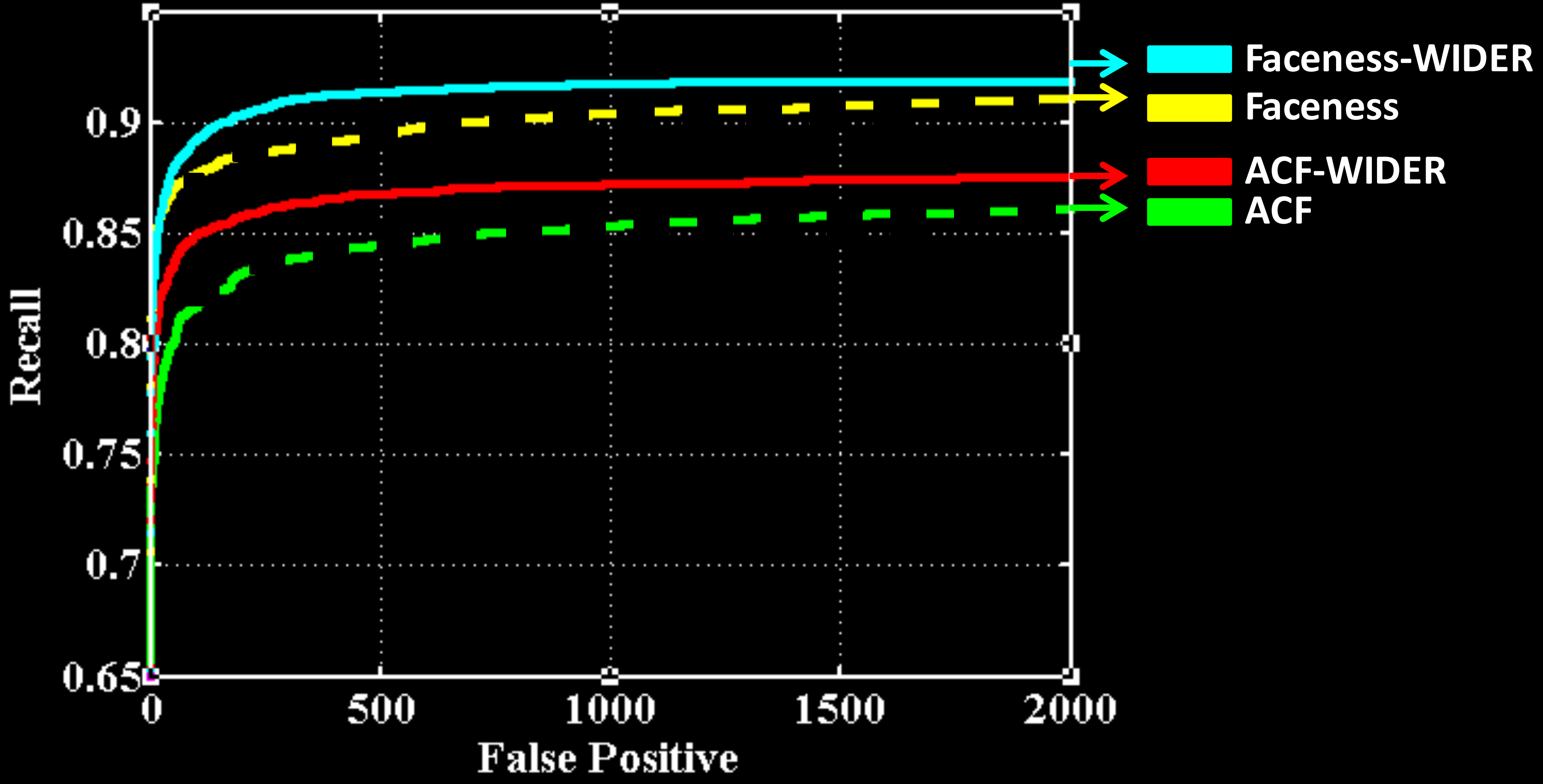
WIDER FACE for training

A face detector is trained using WIDER FACE training/validation partitions, and tested on FDDB dataset.



WIDER FACE for training

A face detector is trained using WIDER FACE training/validation partitions, and tested on FDDB dataset.



WIDER FACE: A Face Detection Benchmark

Multimedia Laboratory, Department of Information Engineering, The Chinese University of Hong Kong

HOME RESULTS

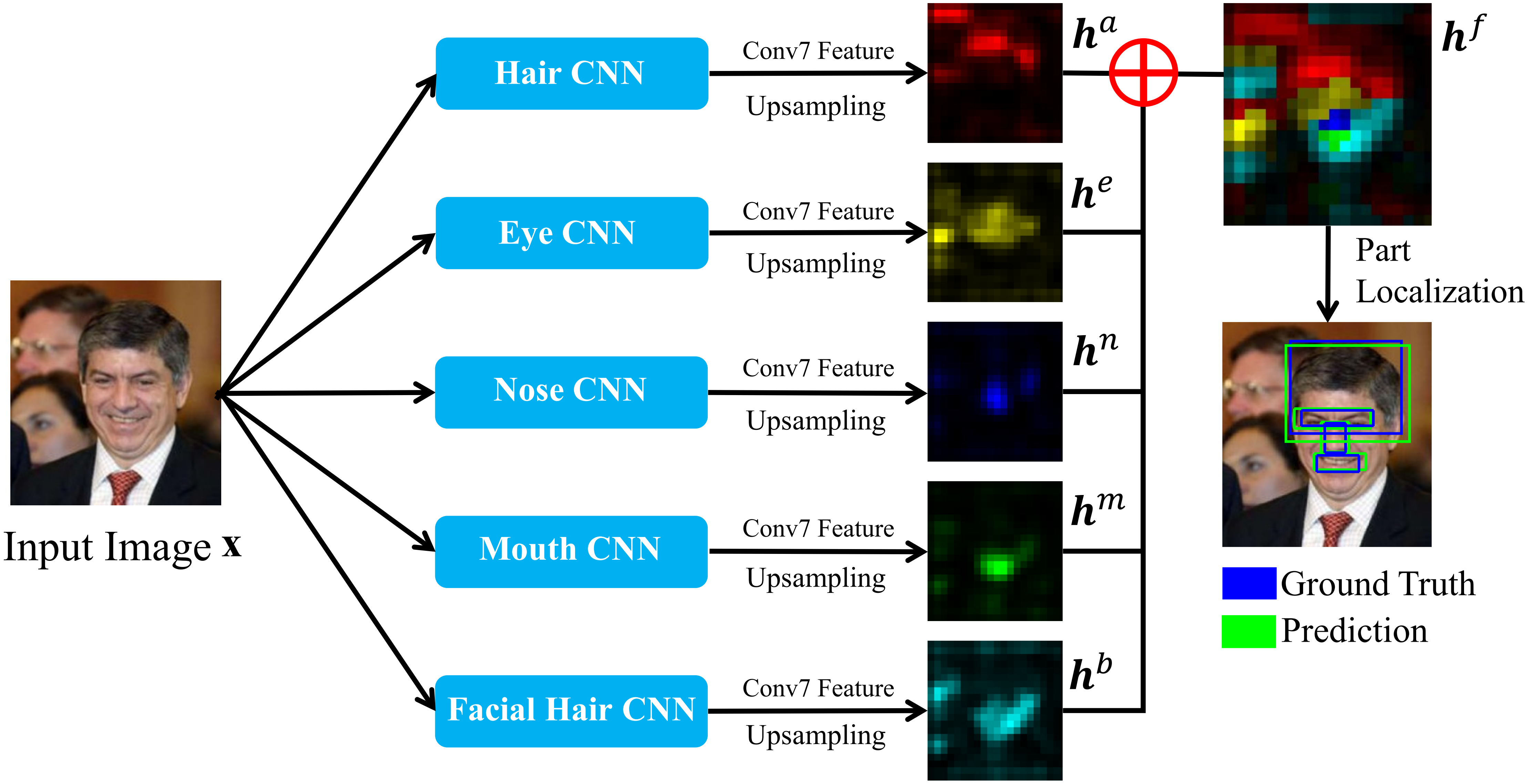
Scale | Pose | Occlusion | Expression | Makeup | Illumination

News

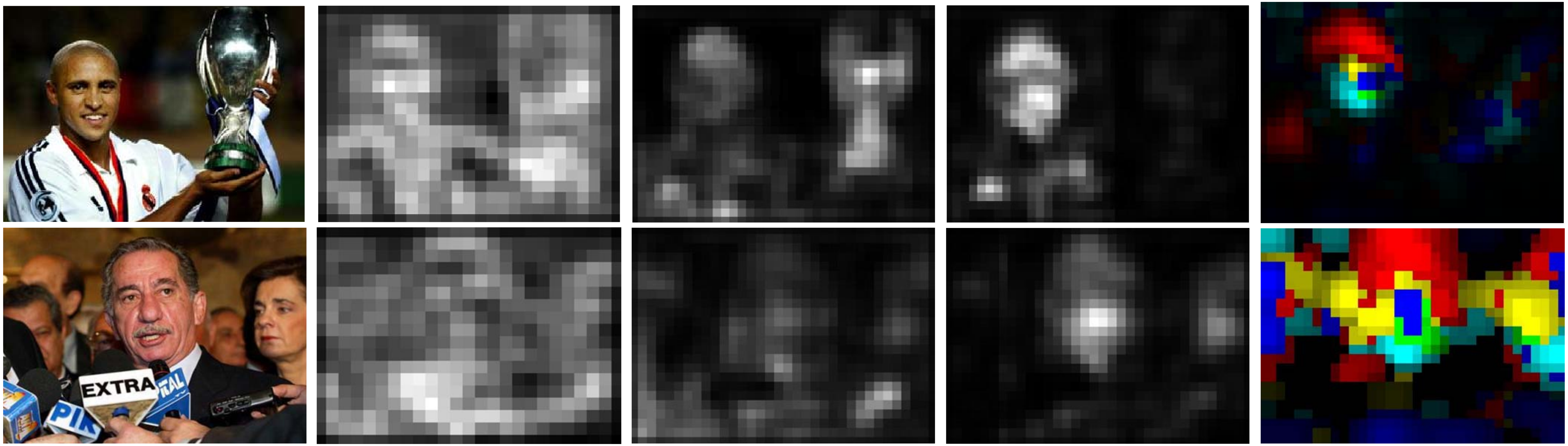
- **2016-04-17** The face attribute labels i.e. pose and occlusion are available. NEW!
- **2015-11-19** Results of four baseline methods: ACF, Faceness, Multiscale Cascade CNN, and Two-stage CNN are released.
- **2015-11-19** WIDER FACE v1.0 is released with images, face bounding box annotations, and event category annotations.

Webpage: http://mmlab.ie.cuhk.edu.hk/projects/WIDERFace/

FacenessNet [ICCV'15]



Why using attributes?



(a) Original image

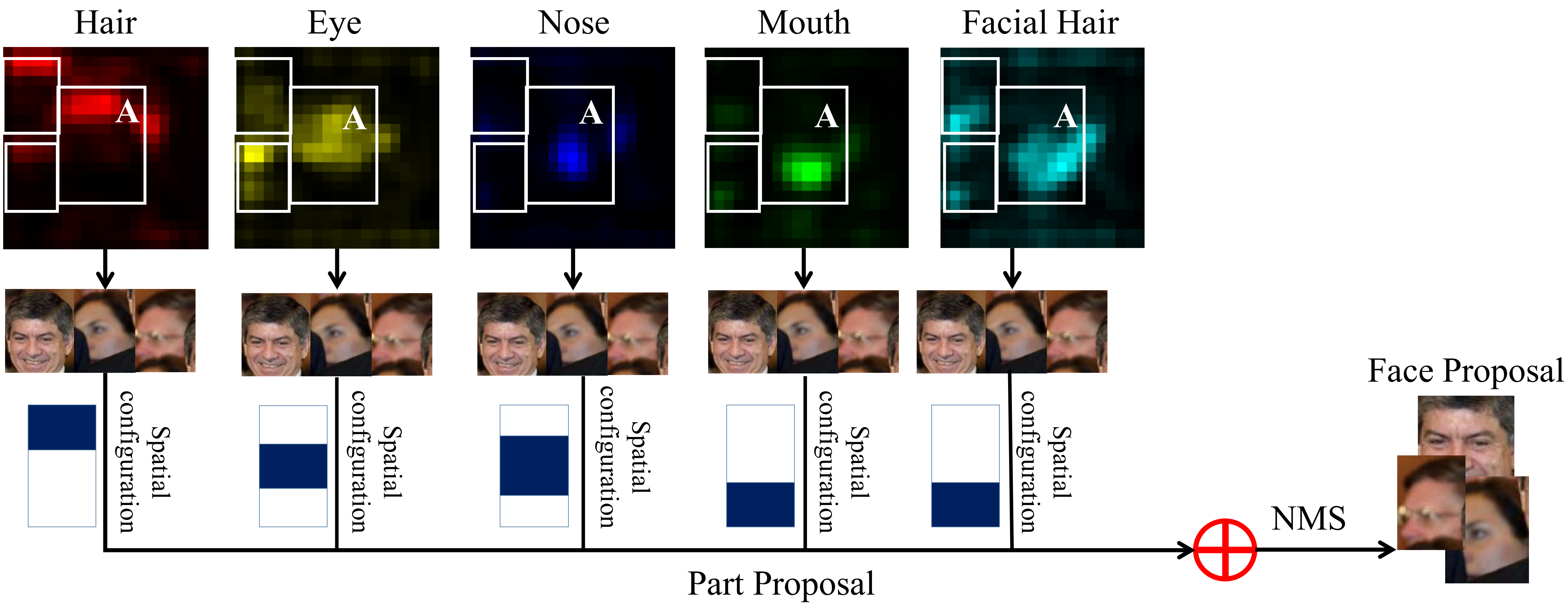
(b) Without pre-trained model

(c) Fine-tuned by classifying face vs. non-face images

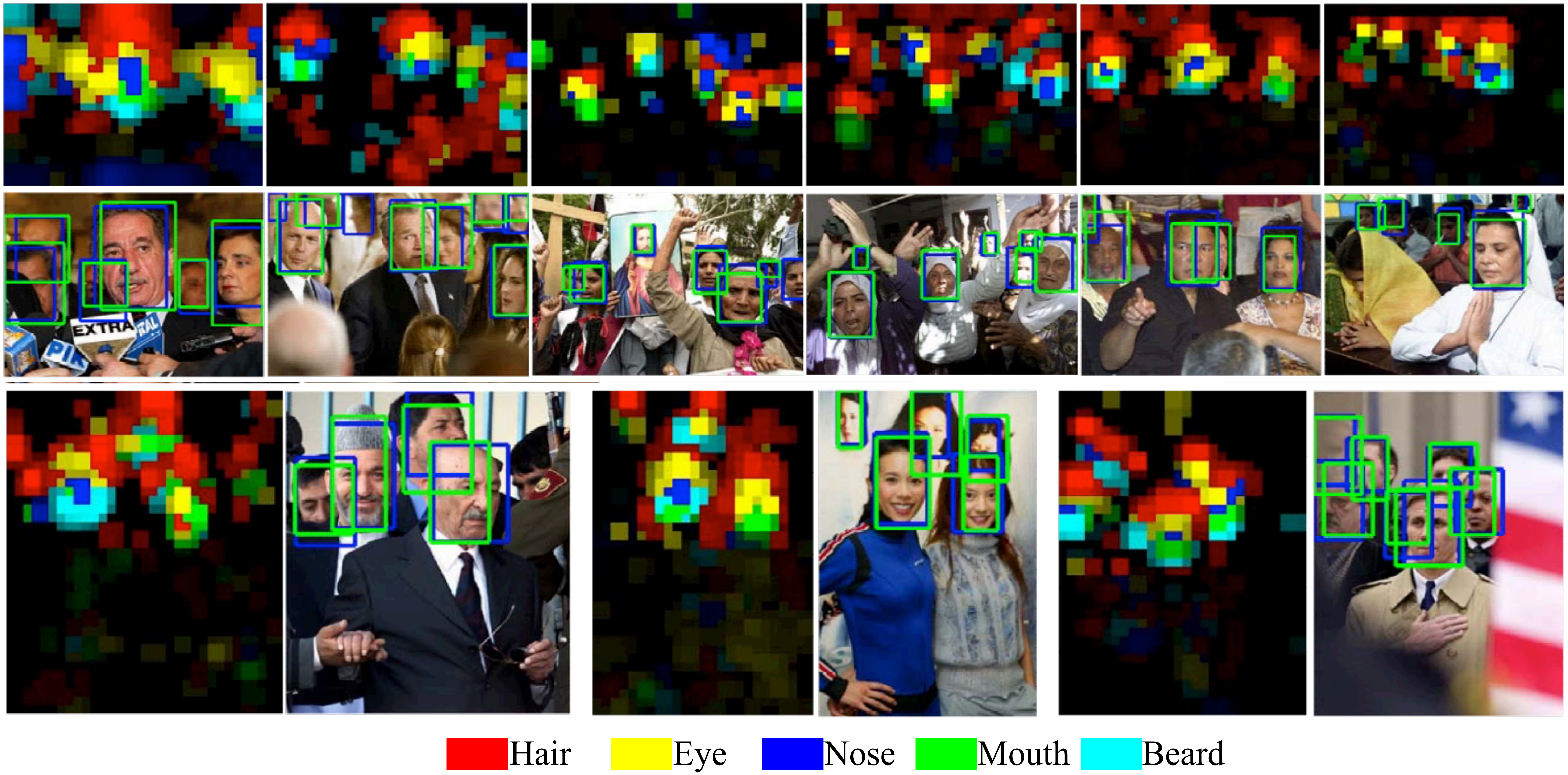
(d) Fine-tuned by classifying 25 face attributes

(e) Fine-tuned by classifying grouped face attributes w.r.t facial parts

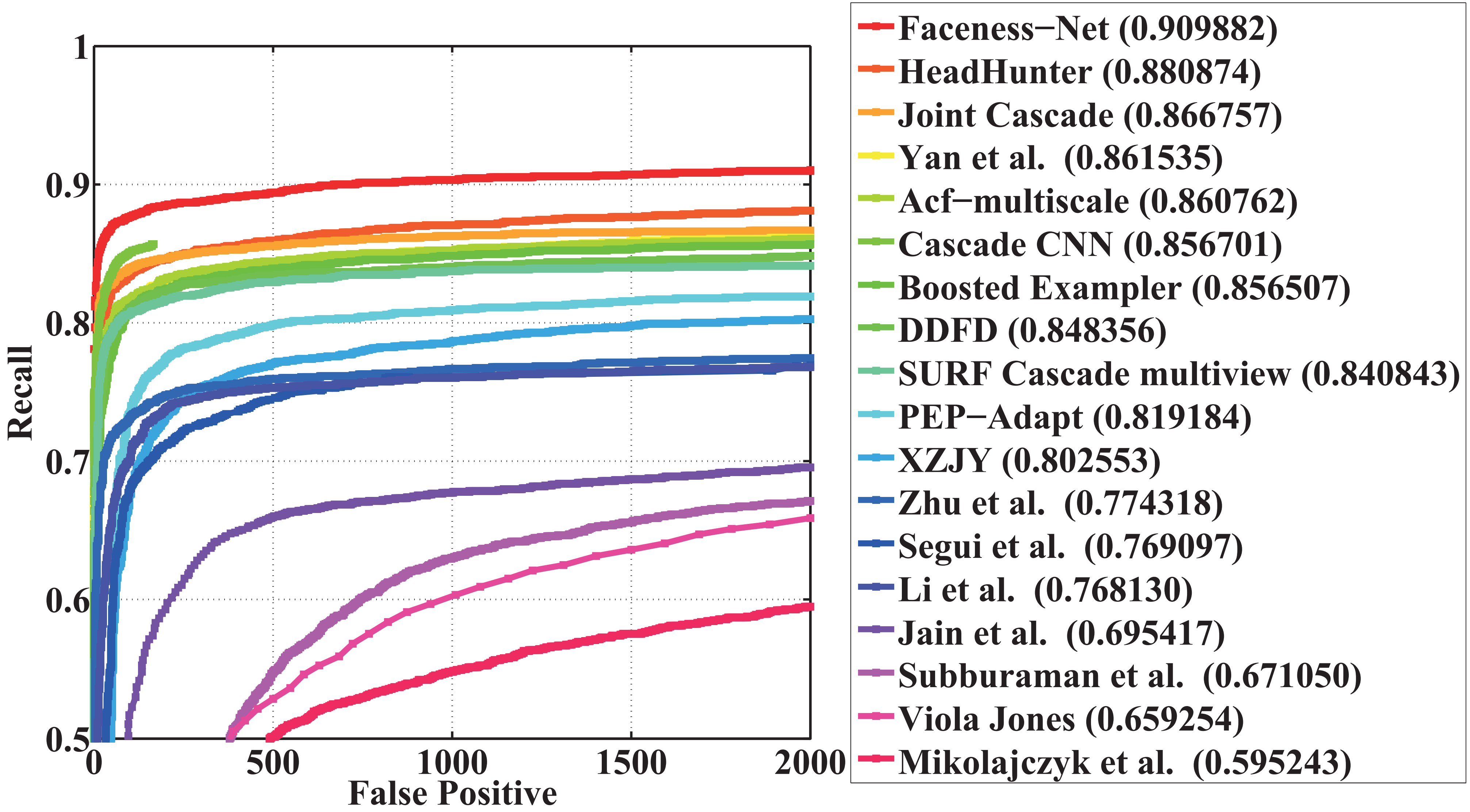
Generating face proposal



Partness Map



Results on FDDB



Face Attribute Recognition

Learning Deep Representation for Imbalanced Classification
C. Huang, Y. Li, C. C. Loy, X. Tang
in Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2016

Code available: http://mmlab.ie.cuhk.edu.hk/projects/LMLE.html

CelebA face attributes dataset

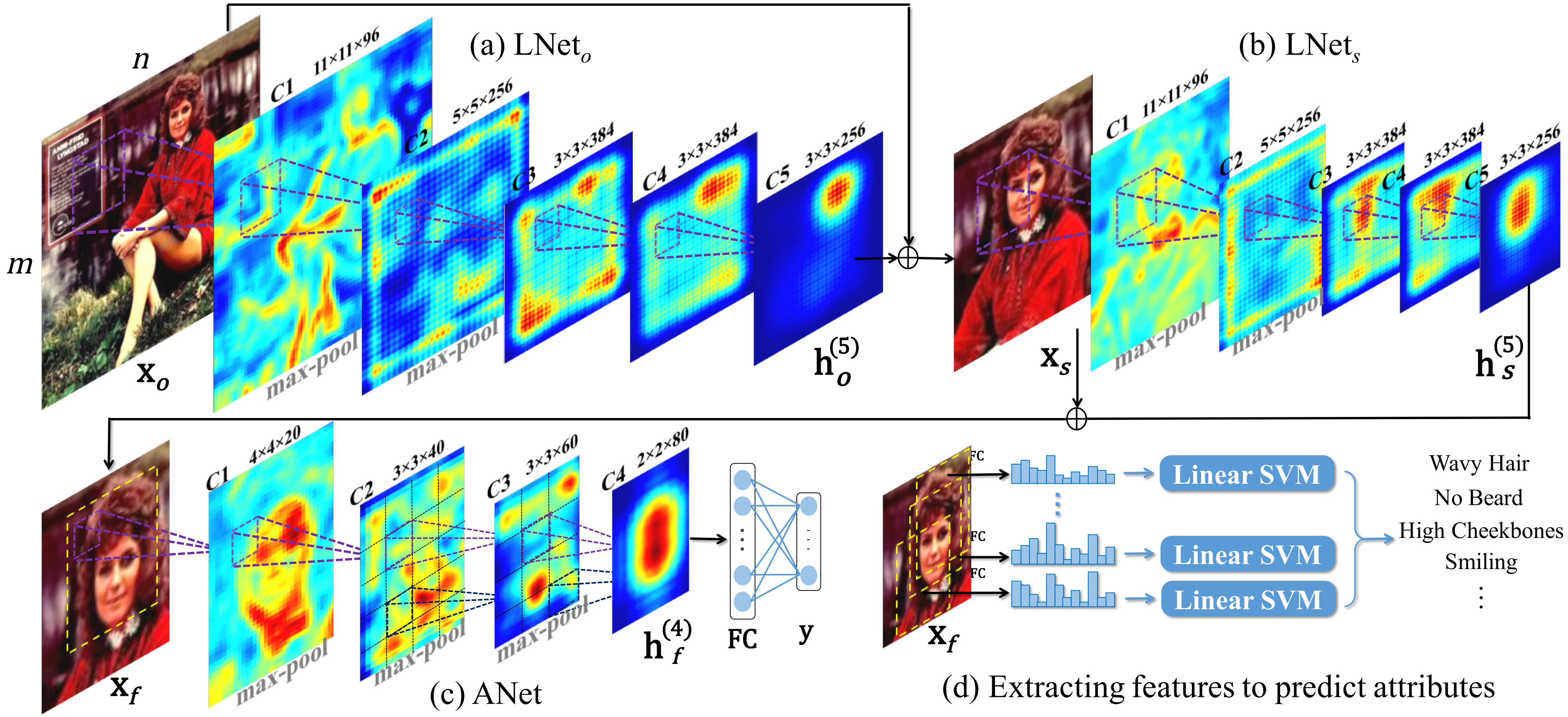
200K celebrity images, each with **40** attribute

Liu et al. "Deep Learning Face Attributes in the Wild", ICCV 2015

http://mmlab.ie.cuhk.edu.hk/projects/CelebA.html

CelebA face attributes dataset

Previous work



Liu et al. "Deep Learning Face Attributes in the Wild", ICCV 2015

Previous work

- *Classification accuracy* biased to the majority class
  - $accuracy \ = \left(\frac{tp + tn}{Np + Nn}\right)$

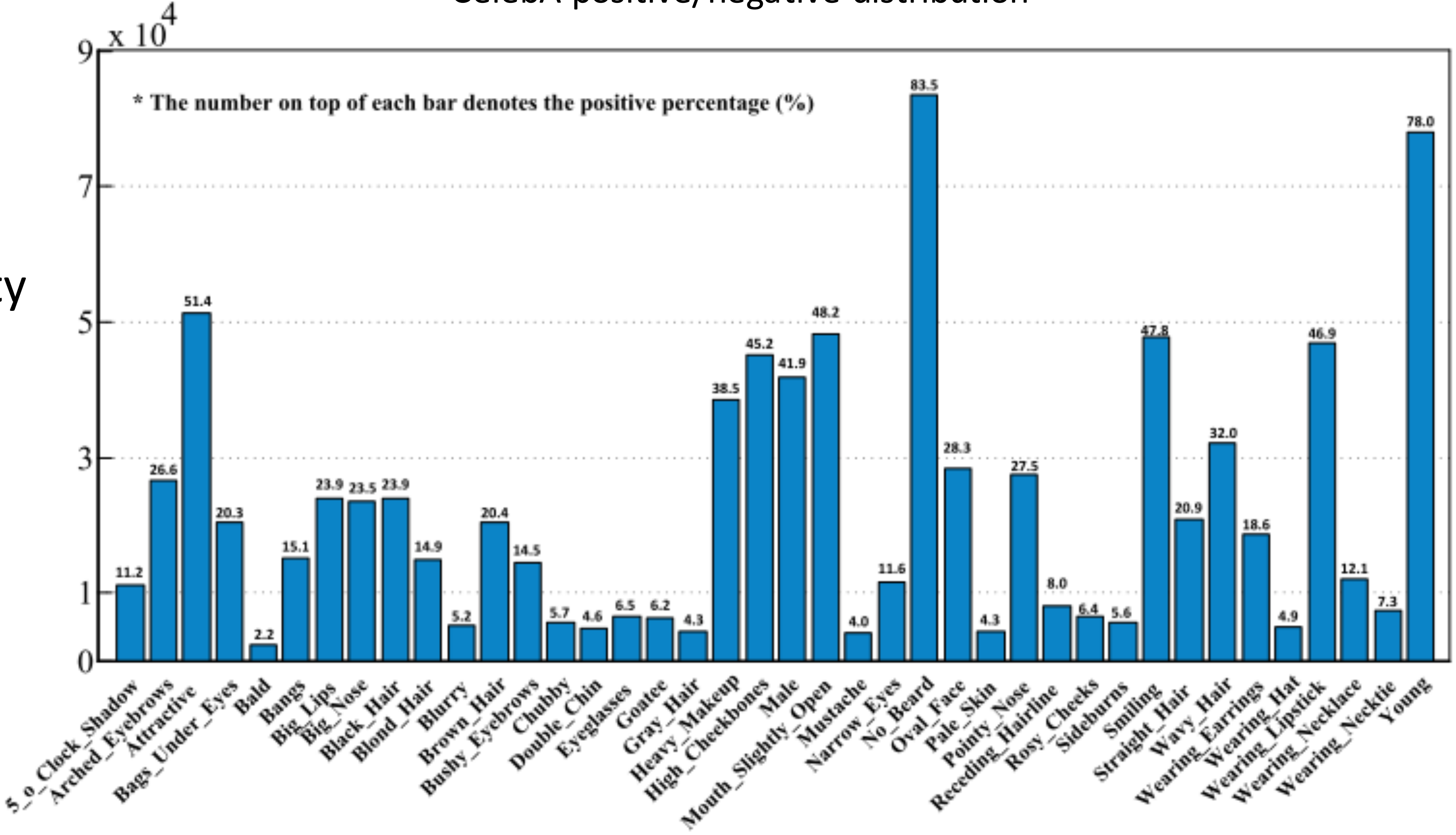
- We adopt a *balance accuracy*
  - $accuracy \ = \frac{1}{2}\left(\frac{tp}{Np} + \frac{tn}{Nn}\right)$

Np and Nn are the numbers of positive and negative samples, while tp and tn are the numbers of true positive and true negative.

A more fundamental problem

- Without handling imbalanced class issue
  - Prediction biases toward the majority class
  - Poor accuracy for the minority class

CelebA positive/negative distribution



Existing solutions

- Class re-sampling [Drummond & Holte, ICML’03]

  - Random under-sampling of majority class

    Remove valuable information

  - Random over-sampling of minority class

    Introduce artificial noise

- Cost-sensitive learning [Zadrozny et al., ICDM’03]

  - Assigns higher misclassification costs to the minority class

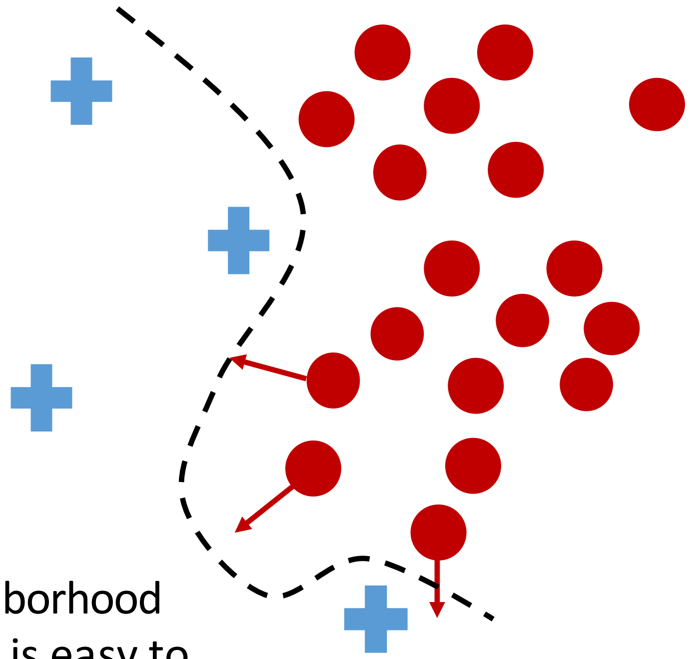
    How to design costs?

Motivation

- Is there a better way apart from sampling and cost learning?

Can we introduce tighter constrains to ameliorate such invasion?

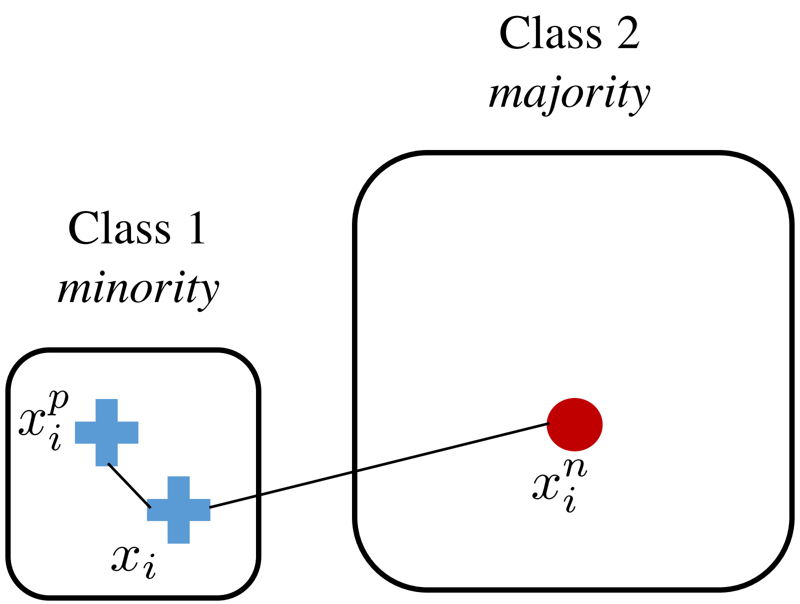
Minority class: very few instances with high degree of visual variability



The genuine neighborhood of these instances is easy to be invaded by other imposter nearest neighbors

Triplet loss helps to a certain extent

- Class-level constraint
  - x_i – an anchor
  - x_i^p – a positive instance (of the same class)
  - x_i^n – a negative instance (different class)

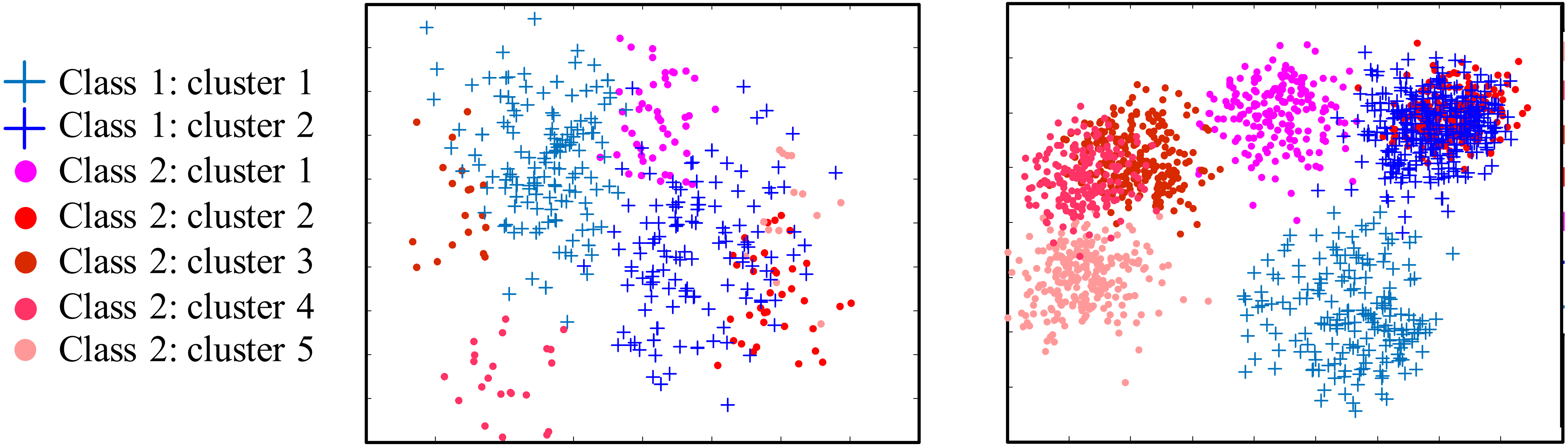


Wearing hat

Not wearing hat

Triplet loss helps to a certain extent

2D feature embedding of one imbalanced binary face attribute



Features extracted from DeepID2 model

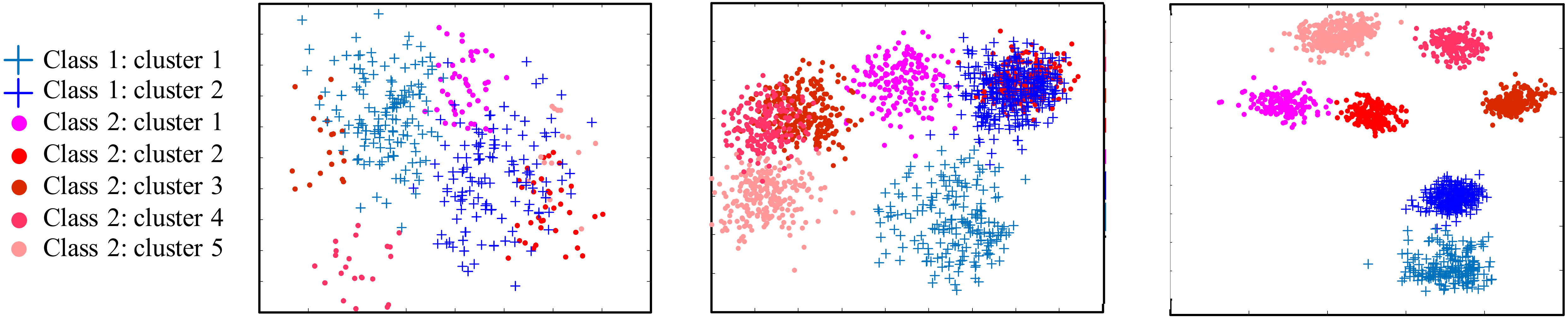
Triplet embedding

Contributions

- Learning deep feature embedding for imbalanced data classification
- A new method that preserves locality across clusters and discrimination between classes
- Large margin classification via fast cluster-wise kNN search

Our solution compared to triplet loss

2D feature embedding of one imbalanced binary face attribute



Features extracted from DeepID2 model

Triplet embedding

Our solution

Large Margin Local Embedding

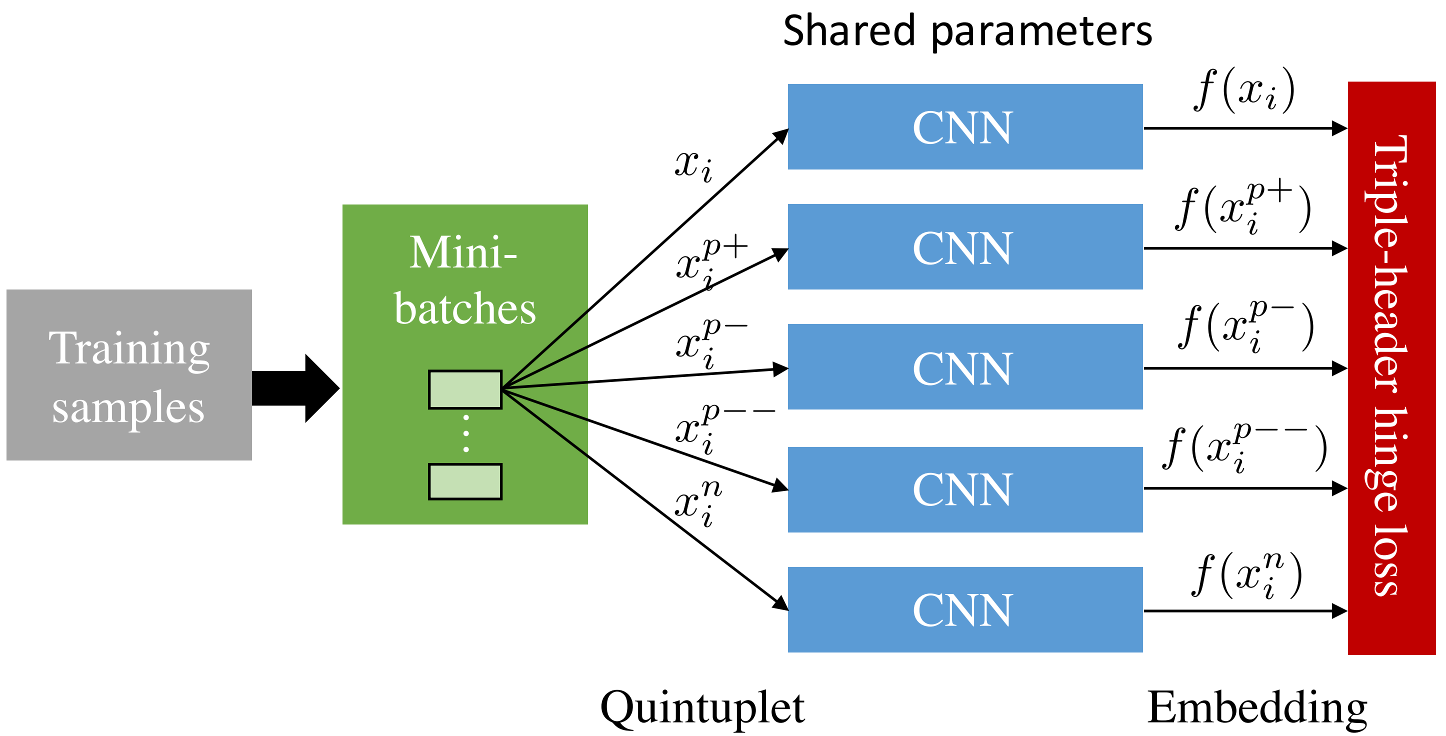
- Our goal:

  Learn a Euclidean embedding $f(x)$ from an image x into a feature space $\mathbb{R}^d$, such that the embedded features are discriminative with minimal possible local class imbalance.

- Main idea:

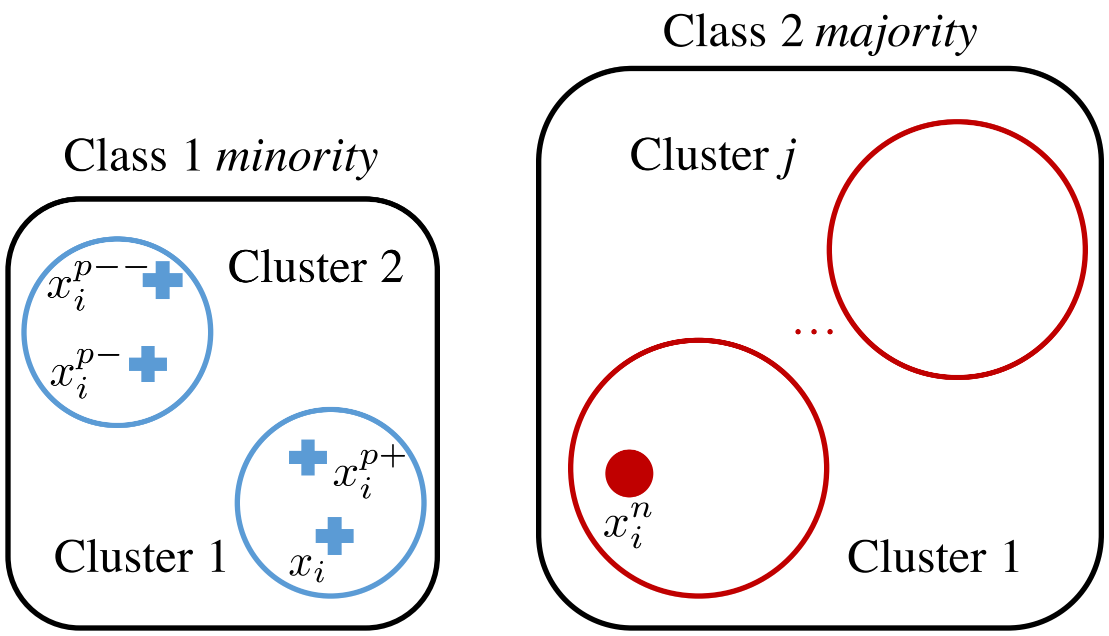
  1. Find patterns (clusters) in each class
  2. Draw classification boundary locally only between marginal clusters, so not depends on class size
  3. Learn deep features to reduce class imbalance in any local neighborhood

Large Margin Local Embedding



Quintuplet sampling

- Cluster- and class-level
  - x_i – an anchor
  - x_i^{p+} – the anchor's most distant within-cluster neighbor
  - x_i^{p-} – the nearest within-class neighbor of the anchor, but from a different cluster
  - x_i^{p--} – the most distant within-class neighbor of the anchor
  - x_i^{n} – the nearest between-class neighbor of the anchor



Quintuplet sampling

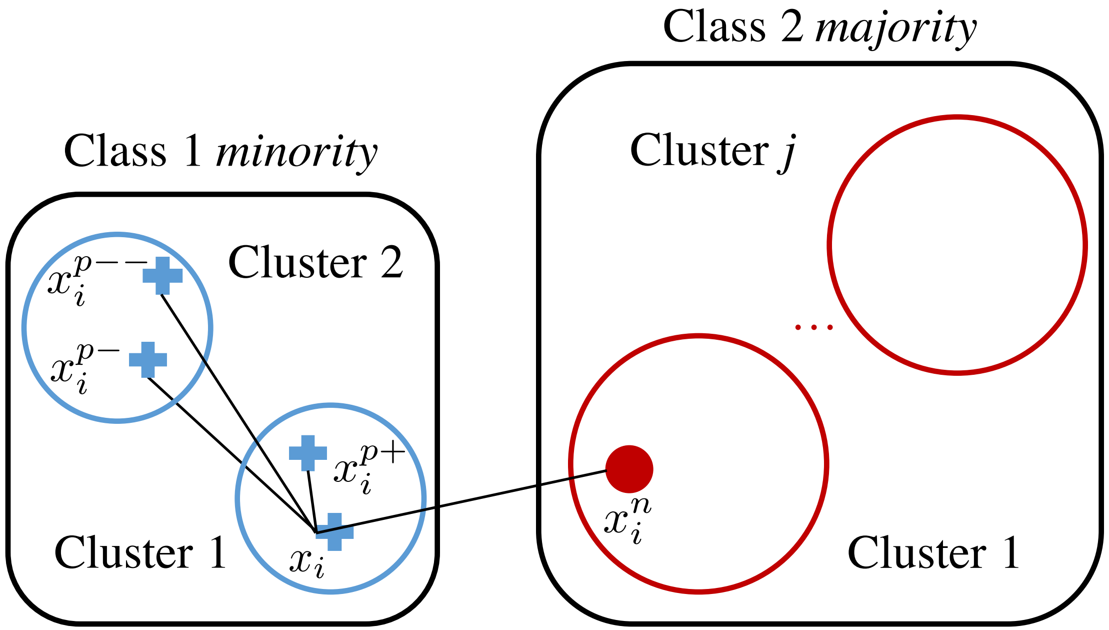
- Ensure the following relationship

$$D(f(x_i), f(x_i^n)) >$$

$$D(f(x_i), f(x_i^{p--})) >$$

$$D(f(x_i), f(x_i^{p-})) >$$

$$D(f(x_i), f(x_i^{p+}))$$



$D(f(x_i), f(x_j)) = \|f(x_i) - f(x_j)\|_2^2$ is the Euclidean distance

Advantages

- Richer information and a stronger constraint than the conventional class-level image similarity
- No information loss unlike under-sampling
- No artificial noise unlike over-sampling

How to obtain the clusters?

- Obtain the initial clusters for each class by applying *k*-means on some prior features

- Face attribute recognition, we use pre-trained DeepID2 features

- Alternating scheme
  - Refine the clusters using features extracted from the proposed model itself every n iterations

Triple-header hinge loss

- To constrain three margins between the four distances

$$\min \sum_i (\varepsilon_i + \tau_i + \sigma_i) + \lambda \|\mathbf{W}\|_2^2$$

s.t.:

$$\max\left(0, g_1 + D(f(x_i), f(x_i^{p+})) - D(f(x_i), f(x_i^{p-}))\right) \leq \varepsilon_i$$

$$\max\left(0, g_2 + D(f(x_i), f(x_i^{p-})) - D(f(x_i), f(x_i^{p--}))\right) \leq \tau_i$$

$$\max\left(0, g_3 + D(f(x_i), f(x_i^{p--})) - D(f(x_i), f(x_i^{n}))\right) \leq \sigma_i$$

$$\forall i,\ \varepsilon_i \geq 0,\ \tau_i \geq 0,\ \sigma_i \geq 0$$

Triple-header hinge loss

- To constrain three margins between the four distances

$$D(f(x_i), f(x_i^n)) > D(f(x_i), f(x_i^{p--})) > D(f(x_i), f(x_i^{p-})) > D(f(x_i), f(x_i^{p+}))$$

$$\min \sum_i (\varepsilon_i + \tau_i + \sigma_i) + \lambda \|\mathbf{W}\|_2^2$$

$$s.t.:$$

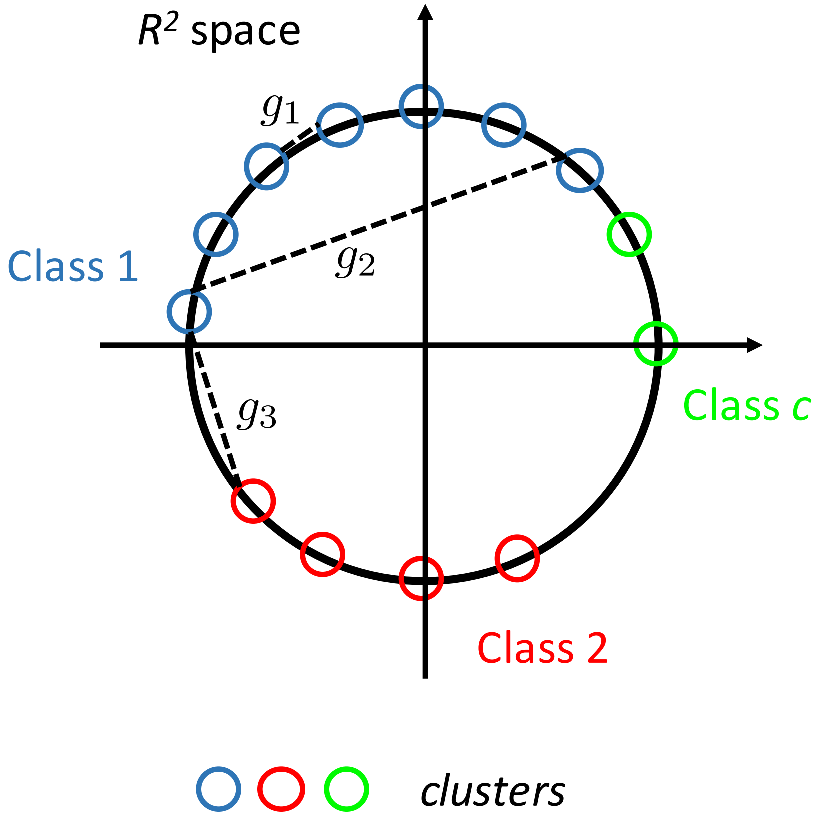
$$\max\left(0, g_1 + D(f(x_i), f(x_i^{p+})) - D(f(x_i), f(x_i^{p-}))\right) \leq \varepsilon_i$$

$$\max\left(0, g_2 + D(f(x_i), f(x_i^{p-})) - D(f(x_i), f(x_i^{p--}))\right) \leq \tau_i$$

$$\max\left(0, g_3 + D(f(x_i), f(x_i^{p--})) - D(f(x_i), f(x_i^{n}))\right) \leq \sigma_i$$

$$\forall i,\ \varepsilon_i \geq 0,\ \tau_i \geq 0,\ \sigma_i \geq 0$$

Triple-header hinge loss



$$\min \sum_i (\varepsilon_i + \tau_i + \sigma_i) + \lambda \|\mathbf{W}\|_2^2$$

s.t.:

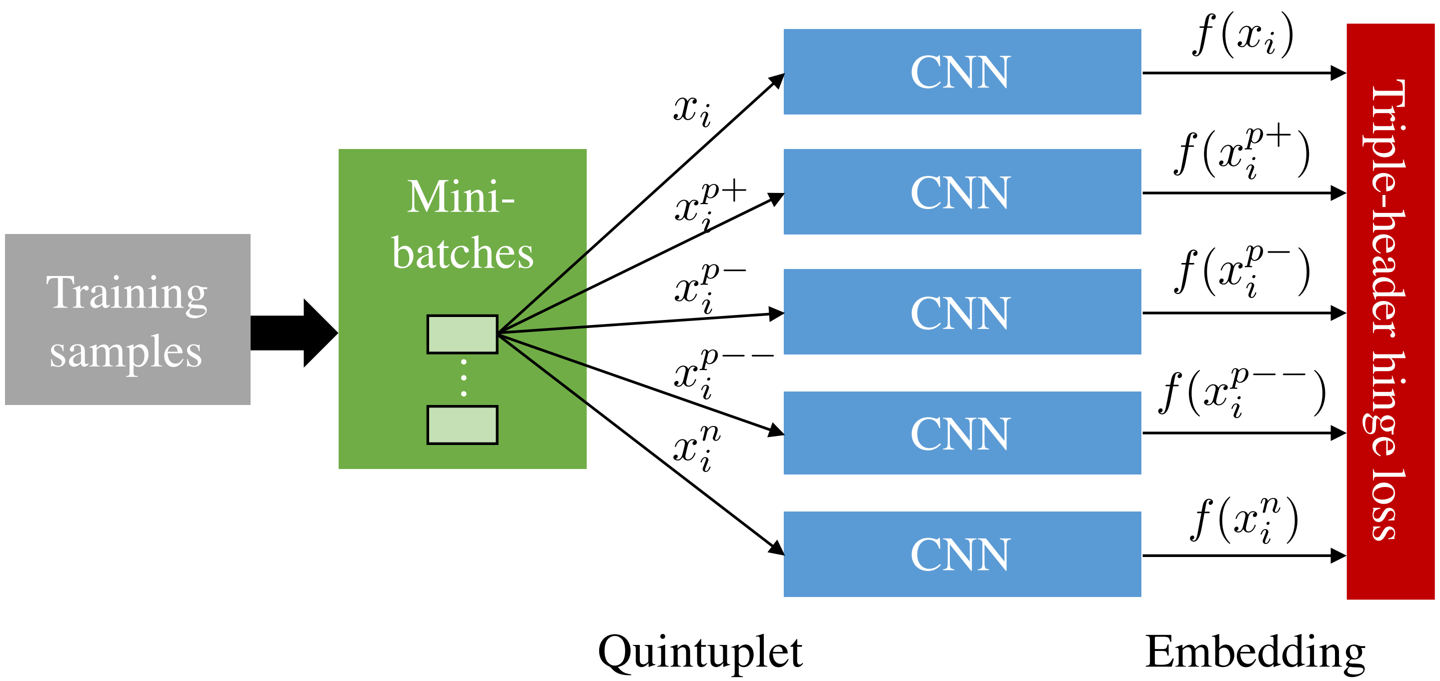
$$\max\left(0, g_1 + D(f(x_i), f(x_i^{p+})) - D(f(x_i), f(x_i^{p-}))\right) \leq \varepsilon_i$$

$$\max\left(0, g_2 + D(f(x_i), f(x_i^{p-})) - D(f(x_i), f(x_i^{p--}))\right) \leq \tau_i$$

$$\max\left(0, g_3 + D(f(x_i), f(x_i^{p--})) - D(f(x_i), f(x_i^{n}))\right) \leq \sigma_i$$

$$\forall i,\ \varepsilon_i \geq 0,\ \tau_i \geq 0,\ \sigma_i \geq 0$$

Network architecture (learning)



Summary of steps

Every 5000 iterations

Feature-based clustering

- Clustering by k-means
- Generate quintuplets from cluster & class membership

Feature learning/updating

- Re-sample batches equally from each class
- Forward their quintuplets to CNN to compute loss
- Back-propagation

Why is it effective?

- Triplet loss
  - The similarity information is only extracted at the *class-level*
  - Homogeneously collapse each class irrespective of their different degrees of variation
  - When a class has high data variability, it is also hard to maintain the class-wise margin

- Triple-header hinge loss
  - Generates diverse quintuplets that differ in the membership of *both clusters and classes*
  - Captures the considerable data variability within each class
  - Can easily enforce the local margin

Nearest neighbor imbalanced classification

- We modified kNN in two ways:
  1. In the well-clustered embedding space LMLE, we treat each cluster as a class-specific exemplar, and perform a fast cluster-wise kNN search.
  2. Use a large margin decision

     Let $\phi(q)$ be query q's local neighborhood defined by its kNN cluster centroids $\{m_i\}_{i-1}^{k}$

$$y_q = \underset{c=1,\ldots,C}{\arg\max} \left(\min_{\substack{m_j \in \phi(q) \\ y_j \neq c}} D(f(q), f(m_j)) - \max_{\substack{m_i \in \phi(q) \\ y_i = c}} D(f(q), f(m_i)) \right)$$

CelebA dataset (100k train,10k test)

	Attractive	Mouth Open	Smiling	Wear Lipstick	High Cheekbones	Male	Heavy Makeup	Wavy Hair	Oval Face	Pointy Nose	Arched Eyebrows	Black Hair	Big Lips	Big Nose	Young	Straight Hair	Brown Hair	Bags Under Eyes	Wear Earrings	No Beard	Bangs
Imbalance level	1	2	2	3	5	8	11	18	22	22	23	26	26	27	28	29	30	30	31	33	35
Triplet-kNN [34]	83	92	92	91	86	91	88	77	61	61	73	82	55	68	75	63	76	63	69	82	81
PANDA [47]	85	93	98	97	89	99	95	78	66	67	77	84	56	72	78	66	85	67	77	87	92
ANet [29]	87	96	97	95	89	99	96	81	67	69	76	90	57	78	84	69	83	70	83	93	90
LMLE-kNN	88	96	99	99	92	99	98	83	68	72	79	92	60	80	87	73	87	73	83	96	98

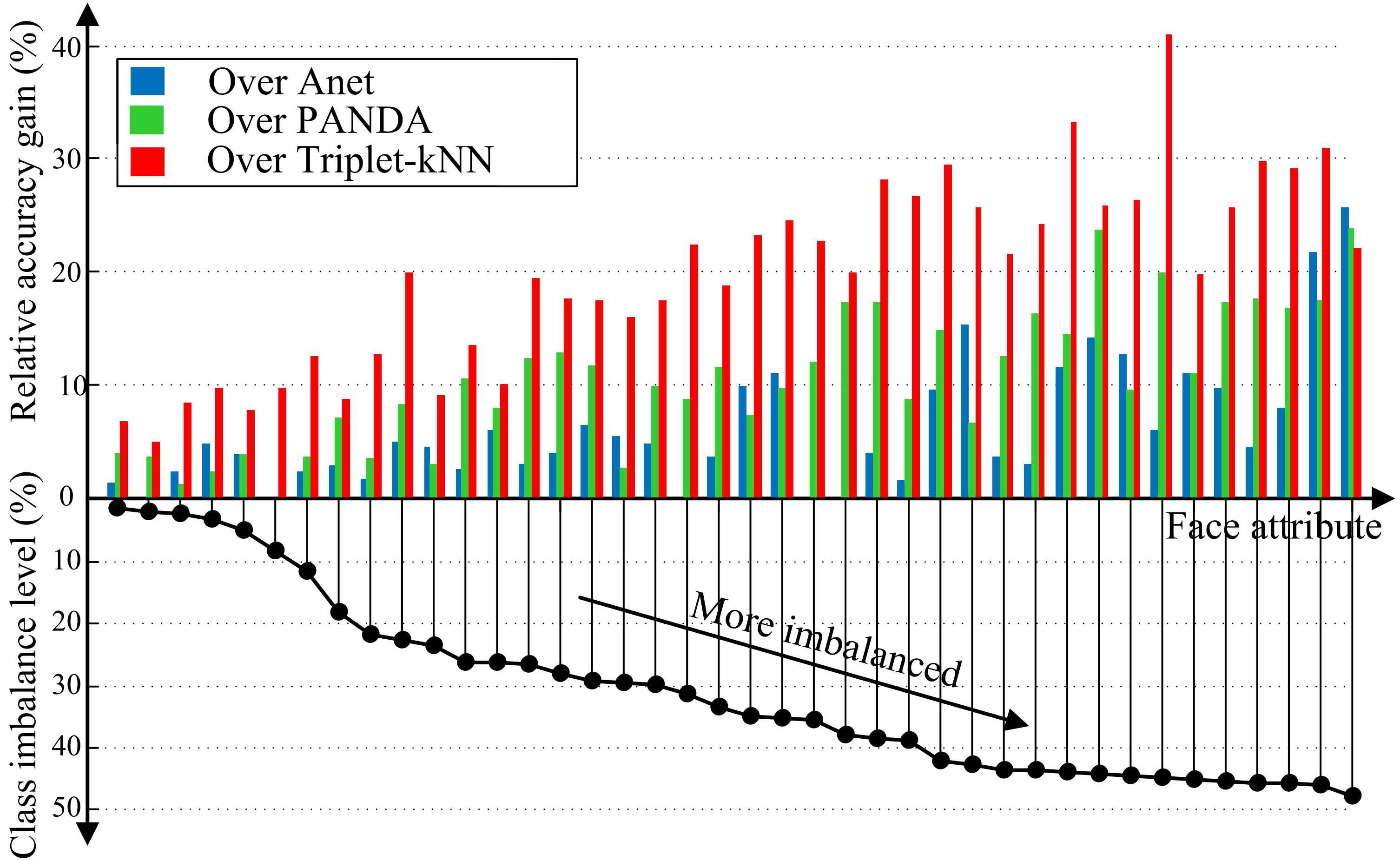
	Blond Hair	Bushy Eyebrows	Wear Necklace	Narrow Eyes	5 o'clock Shadow	Receding Hairline	Wear Necktie	Eyeglasses	Rosy Cheeks	Goatee	Chubby	Sideburns	Blurry	Wear Hat	Double Chin	Pale Skin	Gray Hair	Mustache	Bald		**Average**
Imbalance level	35	36	38	38	39	42	43	44	44	44	44	44	45	45	45	46	46	46	48		
Triplet-kNN [34]	81	68	50	47	66	60	73	82	64	73	64	71	43	84	60	63	72	57	75		**72**
PANDA [47]	91	74	51	51	76	67	85	88	68	84	65	81	50	90	64	69	79	63	74		**77**
ANet [29]	90	82	59	57	81	70	79	95	76	86	70	79	56	90	68	77	85	61	73		**80**
LMLE-kNN	99	82	59	59	82	76	90	98	78	95	79	88	59	99	74	80	91	73	90		**84**

Anet
classification accuracy = 87.24%,
balance accuracy = 80.02%

Ours
classification accuracy = **90.35**%,
balance accuracy = **84.25**%

Class imbalance level (= |positive class rate-50|%)

CelebA dataset (100k train,10k test)

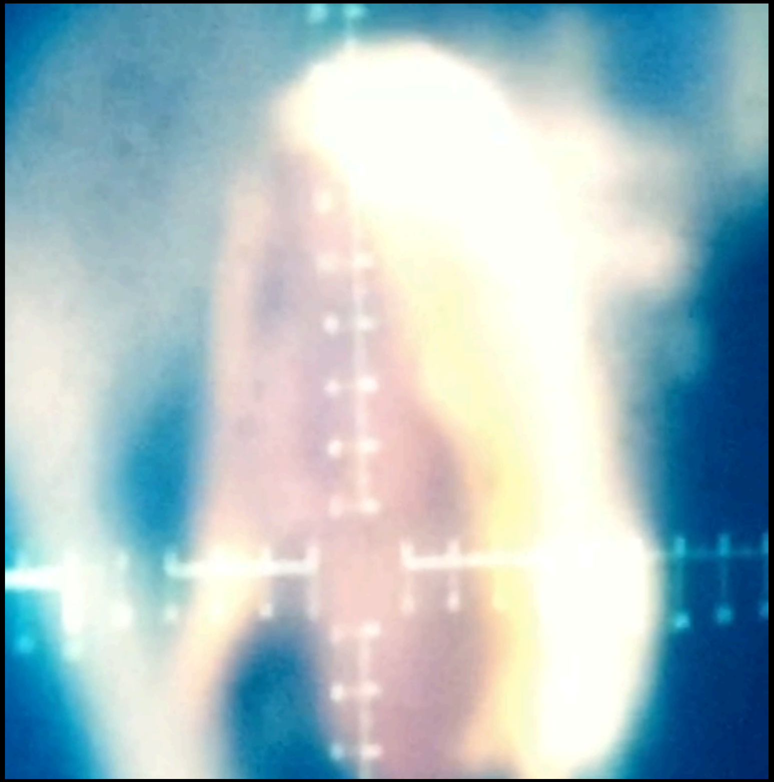


- Code available
- http://mmlab.ie.cuhk.edu.hk/projects/LMLE.html

Face Hallucination

Deep Cascaded Bi-Network for Face Hallucination
S. Zhu, S. Liu, C. C. Loy, X. Tang
in Proceedings of European Conference on Computer Vision, 2016

Code available: https://github.com/zhusz/ECCV16-CBN




Low Resolution
Input

Low Resolution Input


Deep Super-Resolution

ECCV 2014
TPAMI 2015
ICCV 2015
ECCV 2016

Face Detection

ICCV 2015
CVPR 2016

Face Alignment

ECCV 2014
TPAMI 2015
CVPR 2015
CVPR 2016

Deep Face Hallucination

ECCV 2016

Face Attribute Recognition

ICCV 2015
CVPR 2016

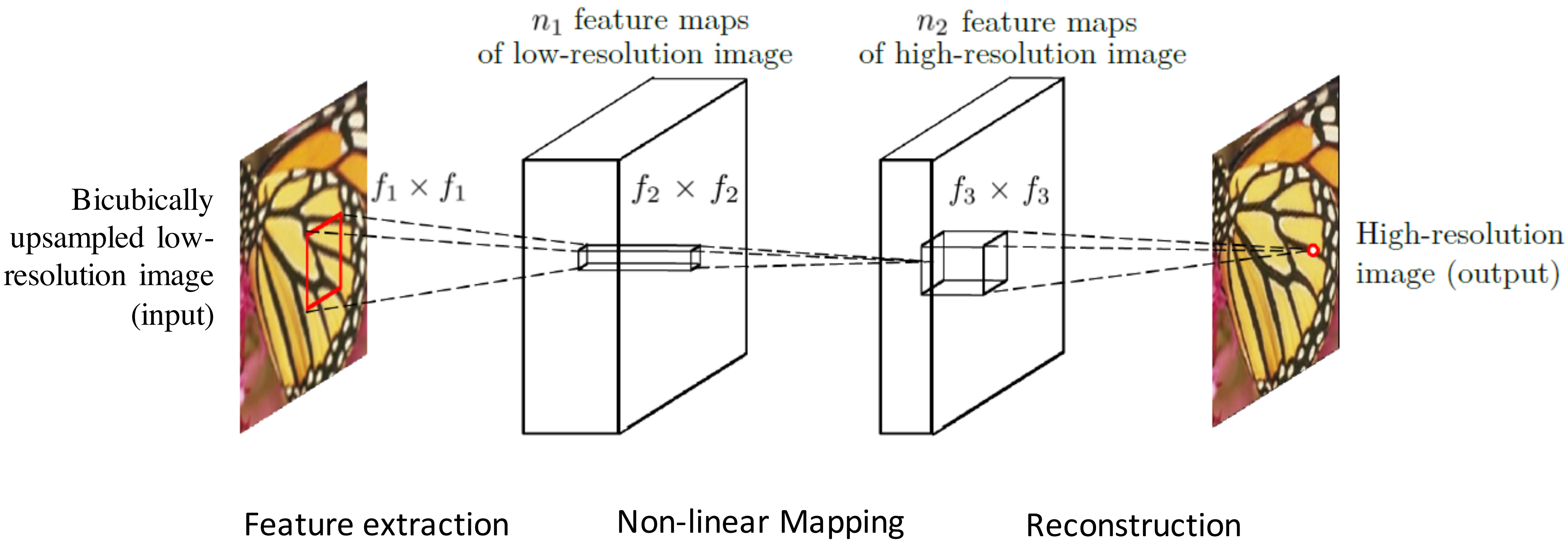
Face Alignment by Coarse-to-Fine Shape Searching

40 ms per-frame on MATLAB

Full version code available: https://github.com/zhusz/CVPR15-CFSS

S. Zhu, C. Li, C. C. Loy, X. Tang, Face Alignment by Coarse-to-Fine Shape Searching, CVPR 2015

Super-resolution CNN (SRCNN)

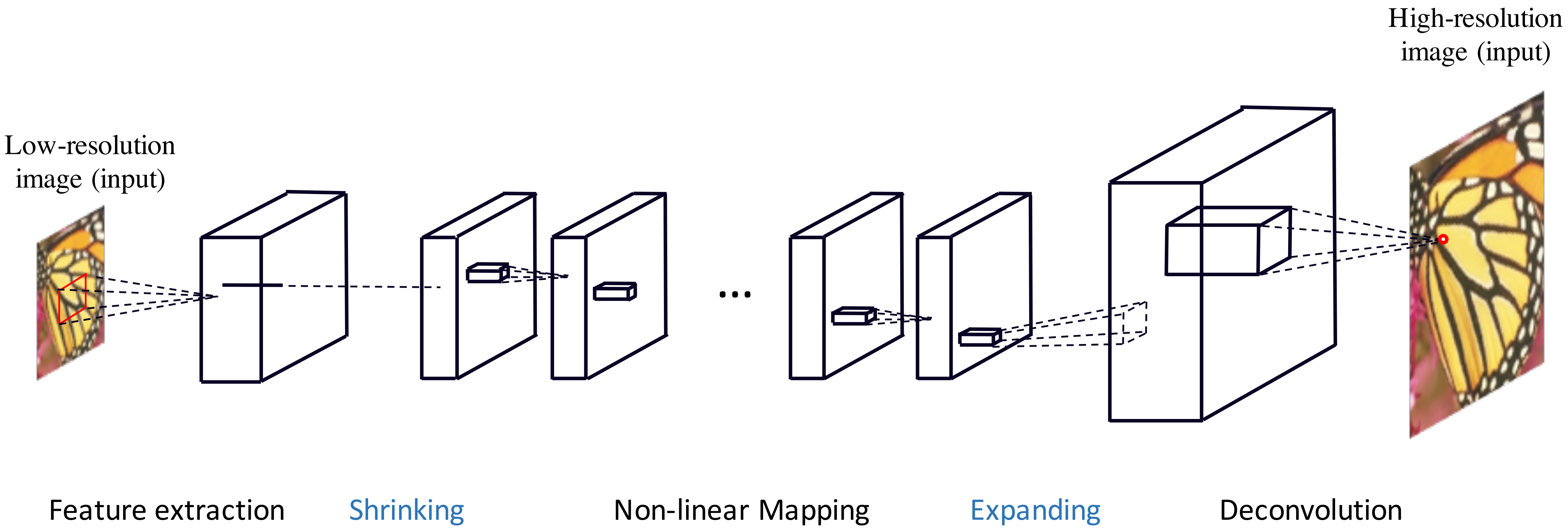


Put together operations that were traditionally treated individually

Full version code available: http://mmlab.ie.cuhk.edu.hk/projects/SRCNN.html

C. Dong, C. C. Loy, K. He, X. Tang, Image Super-Resolution Using Deep Convolutional Networks, TPAMI 2015

Fast Super-resolution CNN (FSRCNN)



40x faster than SRCNN, real-time on CPU, with no performance degradation

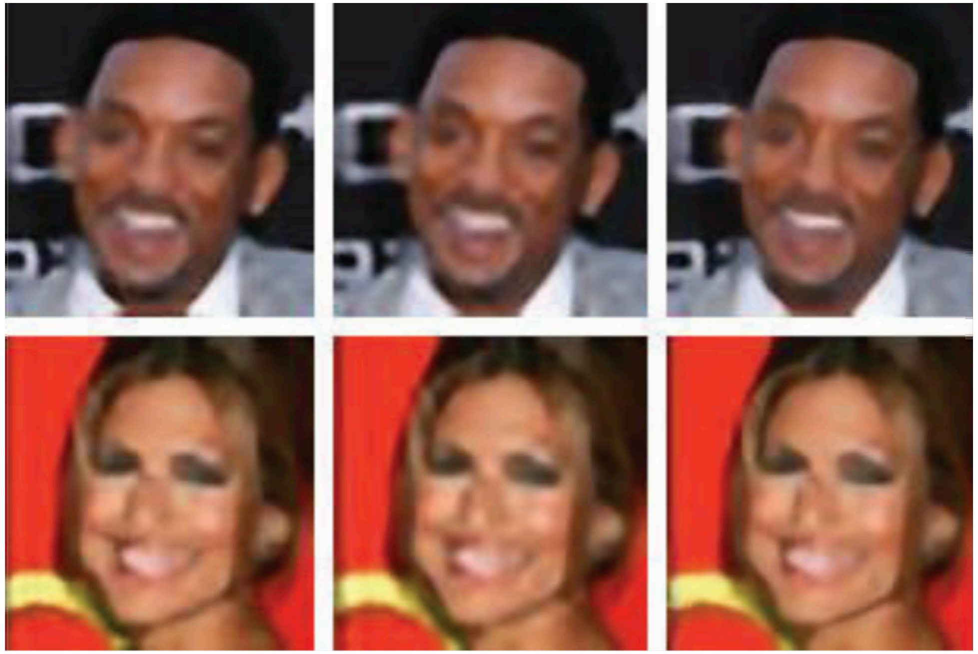
Full version code available: http://mmlab.ie.cuhk.edu.hk/projects/FSRCNN.html

C. Dong, C. C. Loy, X. Tang, Accelerating the Super-Resolution Convolutional Neural Network, ECCV 2016

Low Resolution Input | SRCNN | Existing Hallucination Methods | **OUR METHOD** | High-Resolution Image

General Super-Resolution

- Recovering without synthesizing
- Cannot cope with very low-resolution faces
- Not using face structural priors



Dong TPAMI'15 Salvador ICCV'15 Wang ICCV'15

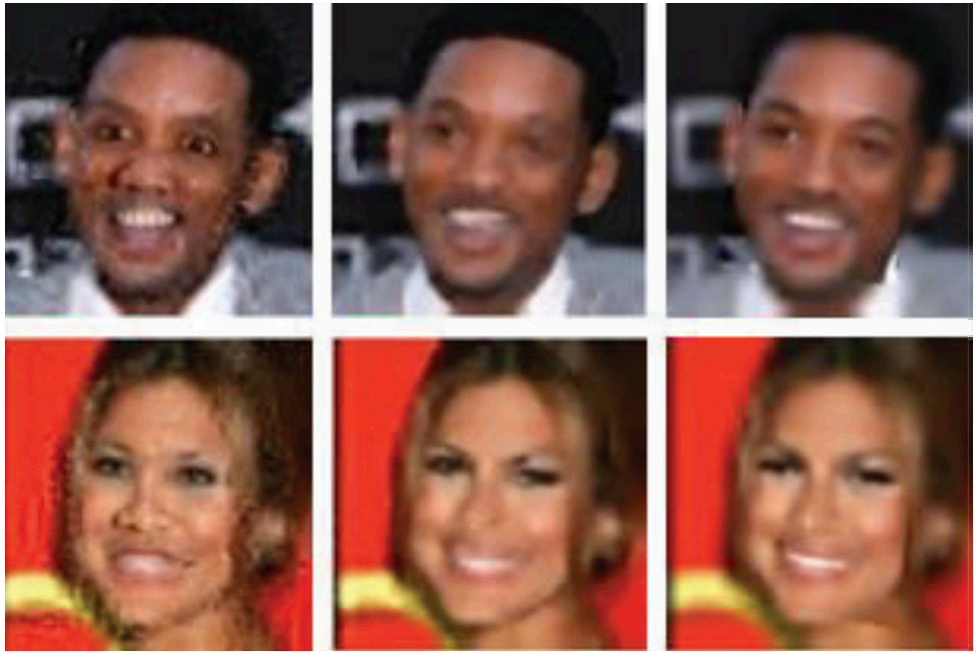
Dong, C., Loy, C.C., He, K., Tang, X.: Image super-resolution using deep convolutional networks. In: PAMI. (2015)

Salvador, J., Perez-Pellitero, E.: Naive bayes super-resolution forest. In: ICCV. (2015)

Wang, Z., Liu, D., Yang, J., Han, W., Huang, T.: Deep networks for image superresolution with sparse prior. In: ICCV. (2015)

Existing Face Hallucination Approaches

- Visually dissimilar
- Assumes correct alignments
- Exemplar based, slow



Yang CVPR'13 Tappen ECCV'12 Jin CVPR'15

Yang, C.Y., Liu, S., Yang, M.H.: Structured face hallucination. In: CVPR. (2013)

Tappen, M.F., Liu, C.: A bayesian approach to alignment-based image hallucination. In: ECCV. (2012)

Jin, Y., Bouganis, C.S.: Robust multi-image based blind face hallucination. In: CVPR. (2015)

The hallucination problem

Two desired capabilities

	Recovering	Synthesizing
Existing face hallucination approaches	**No**	**Yes**
General super-resolution approaches	**Yes**	**No**

Two information sources

	Original low-res	Spatial cues (face prior)
Existing face hallucination approaches	**Information not effectively used**	**Yes**
General super-resolution approaches	**Yes**	**Neglected**

How to enforce spatial cues?

- The chicken-and-egg dilemma
  - Face hallucination vs. dense face correspondence field

- Cascaded frameworks
  - General super-resolution
  - Face alignment

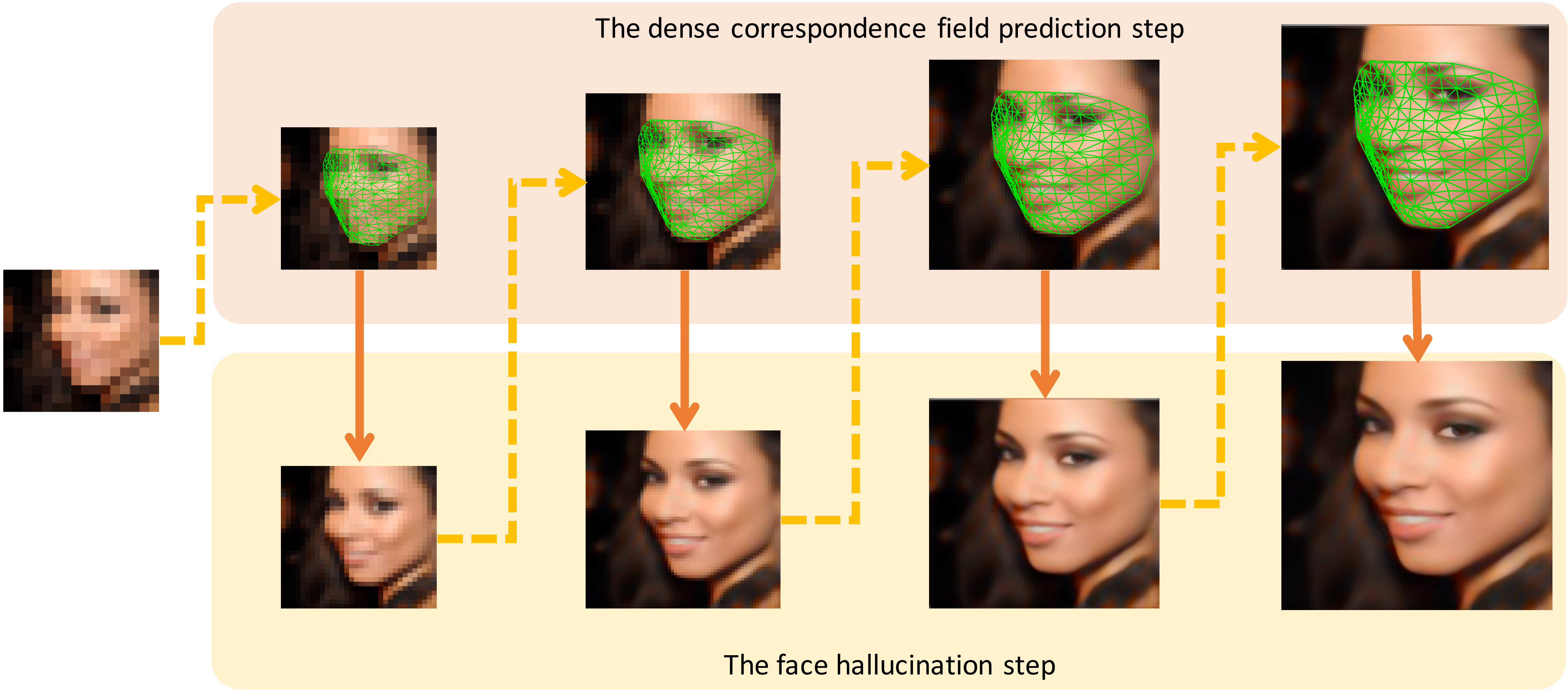
Contributions

- Task-alternating cascade framework
  - Between face hallucination or dense face correspondence field

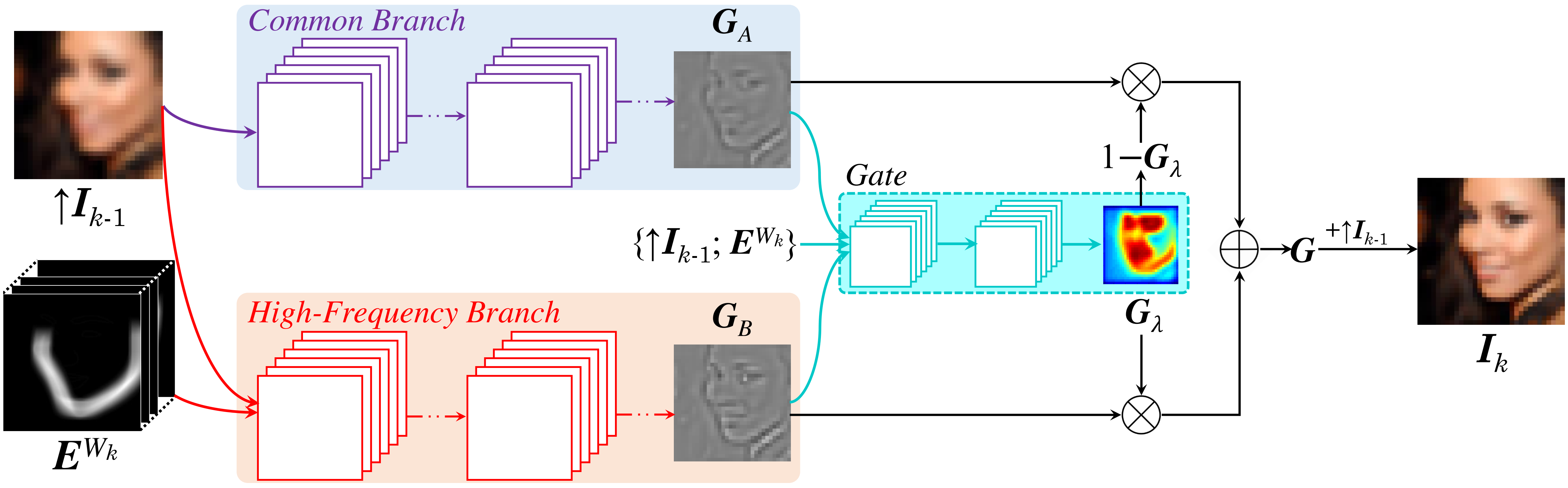
- A gated deep bi-network
  - Effectively exploits face spatial prior

Code available: https://github.com/zhusz/ECCV16-CBN

Task-alternating cascade framework

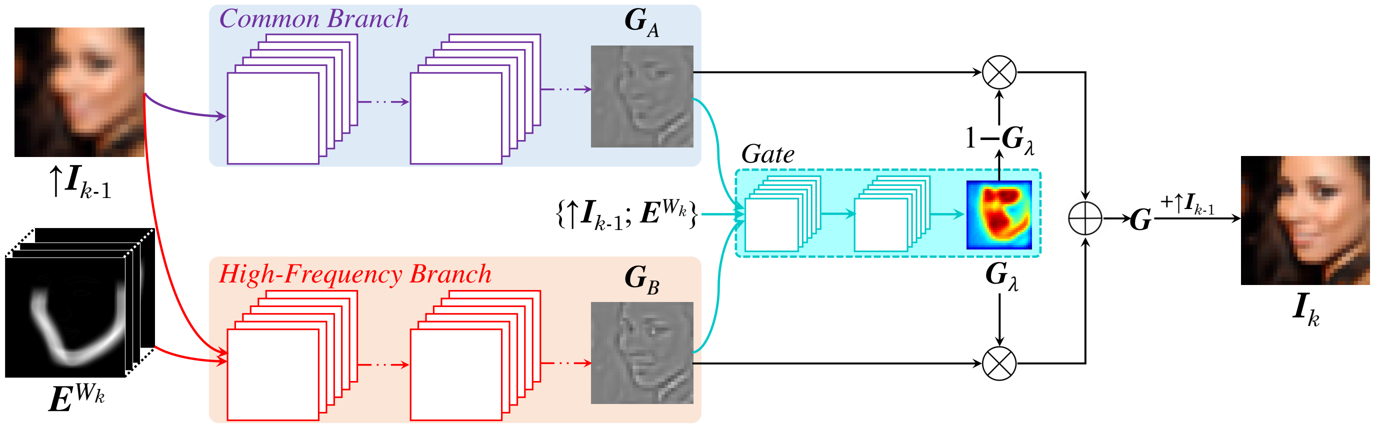


Gated Bi-Network



The face hallucination step

The response of each branch

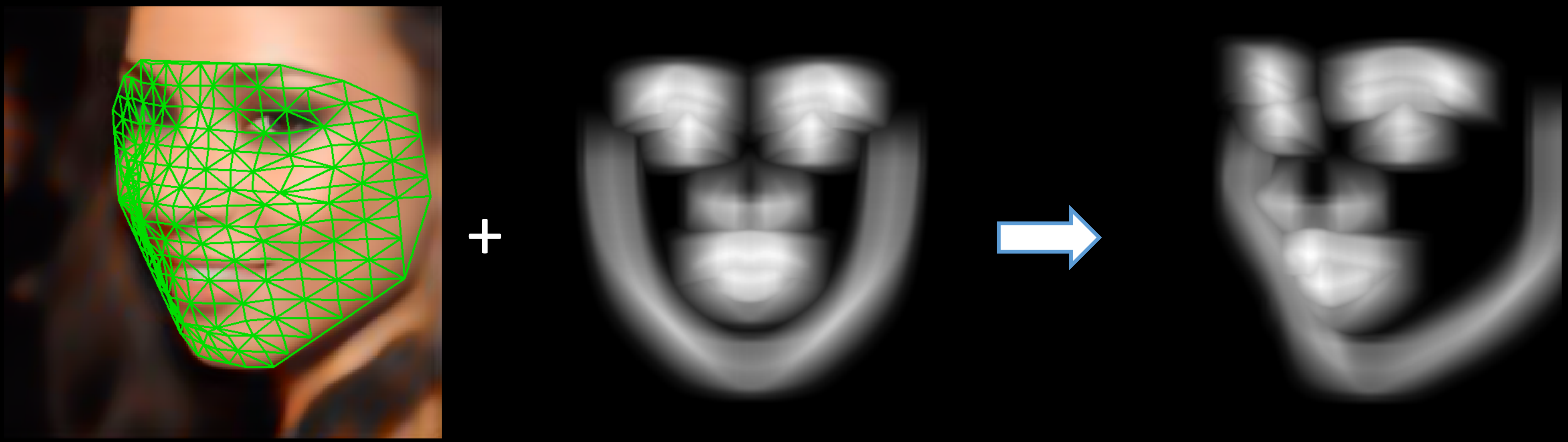


(a) Bicubic (b) Common (c) High-Freq. (d) **CBN** (e) Original

Face structural prior

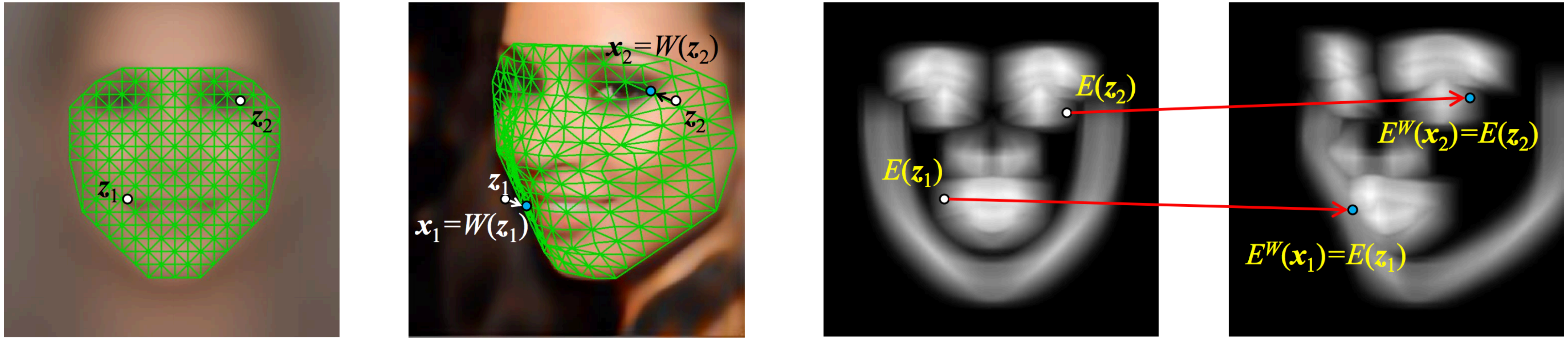
Synthesis guided by face structural prior

Face prior warping



High-frequency face prior

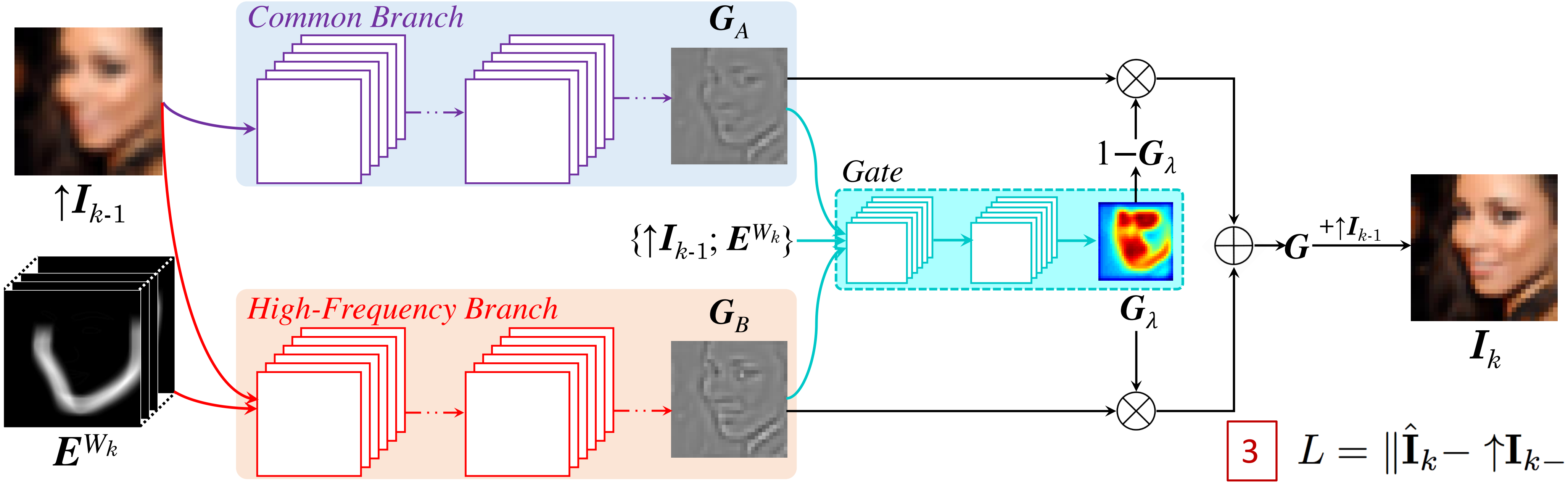
- Preliminary high-frequency map
  - Residual image between the original image and bicubic interpolation of low-res
  - Warp the residual map into the mean face template domain
  - Average the magnitude of the warped residual maps over all training images



(a) Mean Face M (b) Face image $\boldsymbol{I}$ (c) High-Frequency Prior $\boldsymbol{E}$ (d) Warped Prior $\boldsymbol{E}^W$

Learning the gated bi-network

1 $L_A = \|\hat{\mathbf{I}}_k - \uparrow\mathbf{I}_{k-1} - \mathbf{G}_A\|_F^2$



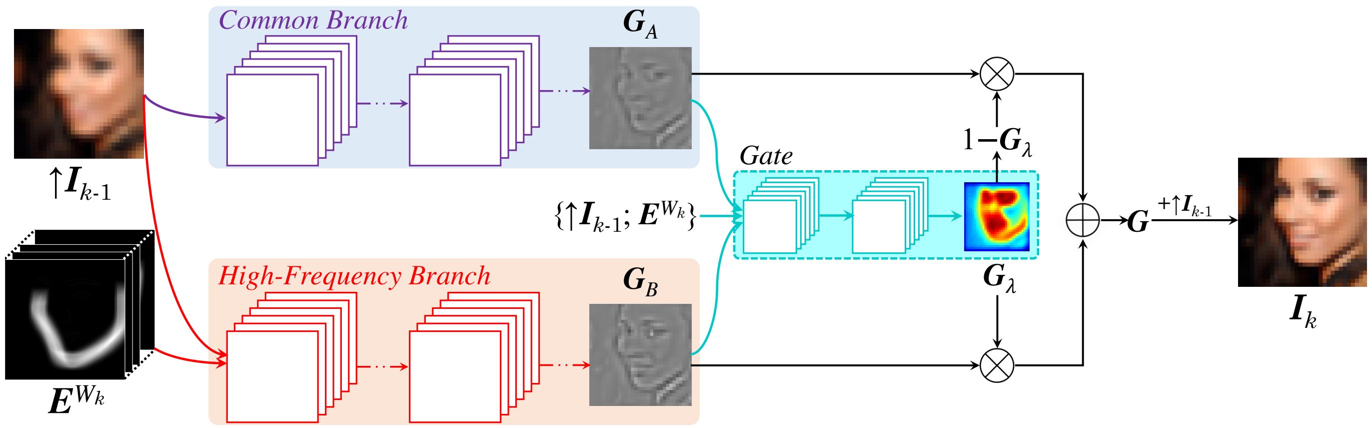
3 $L = \|\hat{\mathbf{I}}_k - \uparrow\mathbf{I}_{k-1} - \mathbf{G}\|_F^2$

2 $L_B = \sum_{c=1}^{C} \|(\mathbf{E}^{W_k})_c \otimes (\hat{\mathbf{I}}_k - \uparrow\mathbf{I}_{k-1} - \mathbf{G}_B)\|_F^2$

Quantitative results

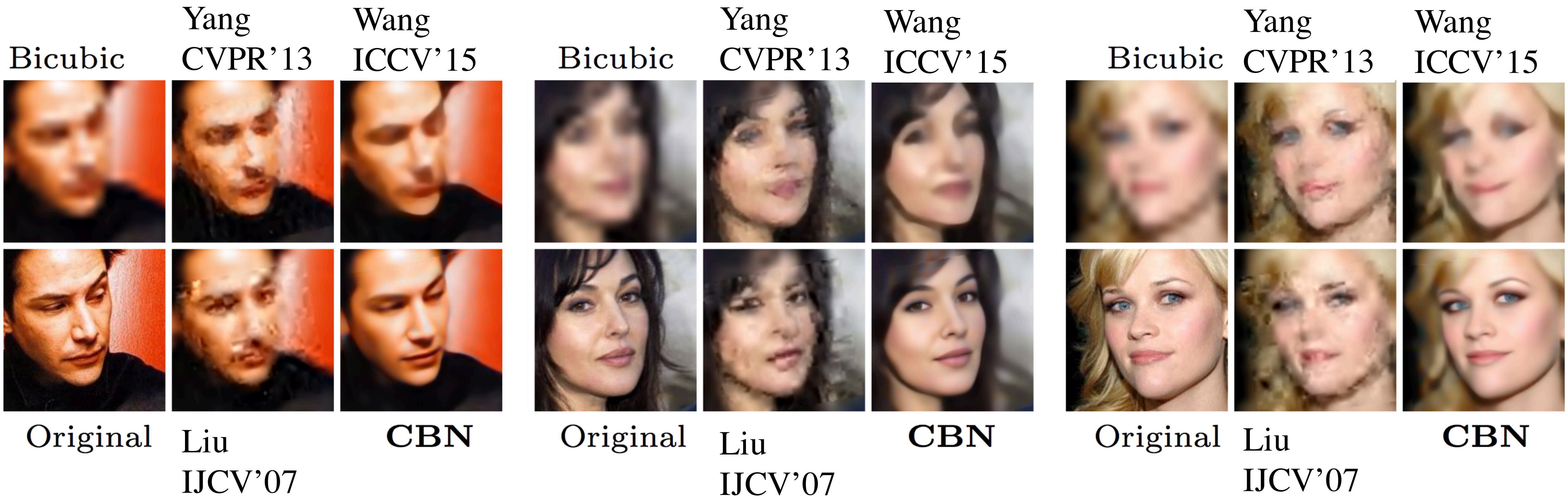
Dataset	Input Size	Bicubic	(I) General super-resolution				(II) Face hallucination			CBN
			A+ [50]	SRCNN [14]	CSCN [15]	NBF [19]	PCA [51, 12]	[52]	[8]	**CBN**
MultiPIE	4×	33.66 (.900)	34.53 (.910)	34.75 (.913)	35.10 (.920)	34.73 (.912)	33.98 (.904)	34.07 (.907)	34.31 (.903)	**35.65 (.926)**
PubFig	2×	34.78	35.89	36.12	36.47	35.98	-	-	-	**36.66**
	3×	31.52	32.02	32.13	32.88	32.09	-	-	-	**33.17**
	4×	29.61	30.02	30.15	30.79	30.16	-	-	-	**31.28**
HELEN	2×	41.96	42.77	42.95	43.37	43.01	-	-	-	**43.51**
	3×	38.52	38.89	39.10	39.57	39.15	-	-	-	**39.78**
	4×	36.59	36.81	36.87	37.61	36.89	-	-	-	**37.94**
MultiPIE	5px	25.39 (.752)	25.63 (.767)	25.72 (.771)	25.93 (.773)	25.75 (.769)	25.62 (.767)	25.83 (.774)	25.72 (.769)	**27.14 (.808)**
PubFig	8px	22.32	22.79	22.98	23.25	23.08	23.37	23.57	23.10	**26.83**
	5px	20.63	20.96	21.07	21.33	21.04	21.42	21.58	21.19	**25.31**
HELEN	8px	21.86	22.24	22.47	22.69	22.53	22.95	23.01	22.62	**26.36**
	5px	20.28	20.50	20.59	20.84	20.57	21.09	21.13	20.64	**25.09**

Ablation study



Dataset	1a. Only Common Branch i.e. Vanilla Cascaded CNN	1b. Only High-Freq. Branch	2. Fixed Correspondence	3. Single Cascade	**Full Model**
PubFig	23.76	24.66	23.85	22.09	**25.31**
HELEN	23.57	24.53	23.77	21.83	**25.09**
PubFig83	28.06	29.31	28.34	26.70	**29.83**

Qualitative results (large pose)



Yang, C.Y., Liu, S., Yang, M.H.: Structured face hallucination. In: CVPR. (2013)
Liu, C., Shum, H.Y., Freeman, W.T.: Face hallucination: Theory and practice. IJCV (2007)

Qualitative results

Original Bicubic General Super Resolution Existing Face Hallucination CBN

Bicubic Interpolation **Deep Face Hallucination**

Bicubic Interpolation **Deep Face Hallucination**

Bicubic Interpolation **Deep Face Hallucination**

Over-synthesis

Ghosting effect

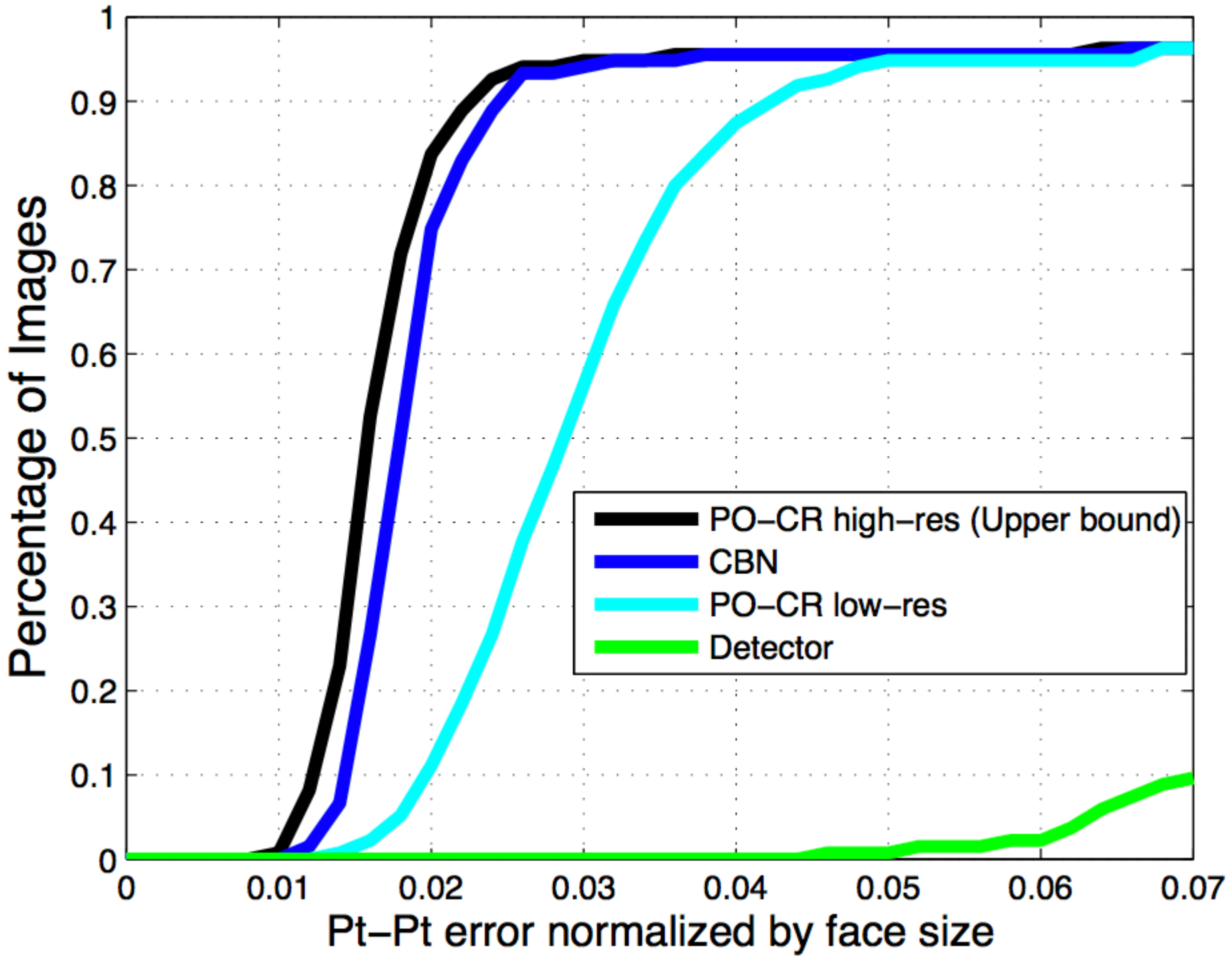
Inaccurate details

Lower bound

Original 3pxIOD 5pxIOD 8pxIOD 10pxIOD

Low-res face alignment

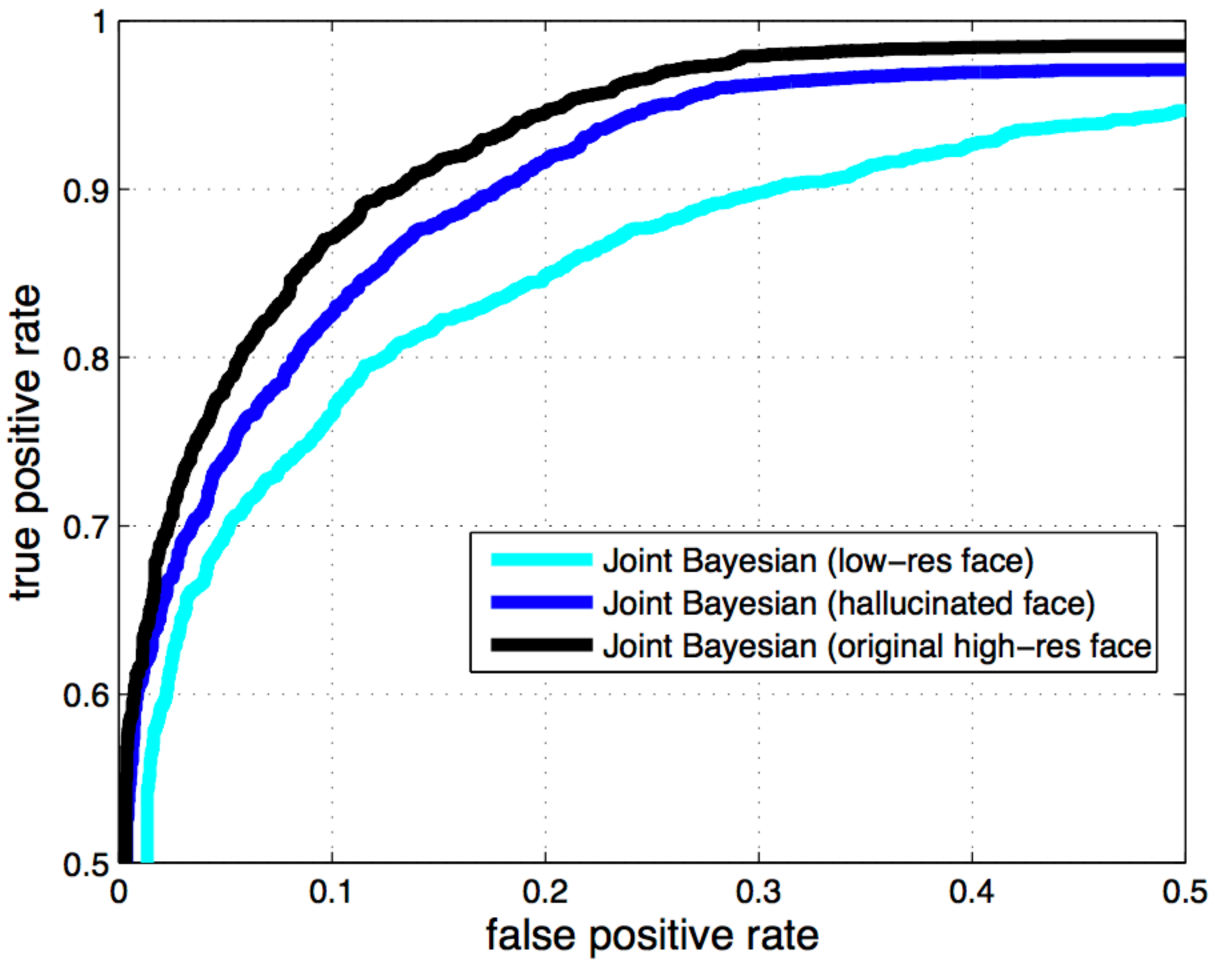
- Face alignment at low-res as by product
- iBUG dataset
- 5pxIOD input



Tzimiropoulos, G.: Project-out cascaded regression with an application to face alignment. In: CVPR. (2015)

Low-res face verification

- Evaluate identity preserving property
- Joint Bayesian approach retrained based on input resolution
- LFW with unrestricted protocol
- 5pxIOD input



Conclusion

- Hallucinating faces under substantial shape deformation and appearance variation
- Adaptively refine the dense correspondence field and hallucinate faces in an alternating manner
- Guided by the high-frequency prior, our framework can leverage spatial cues in the hallucination process

Low Resolution
Input

Low Resolution
Input

Thanks!

References

1. S. Zhu, S. Liu, C. C. Loy, X. Tang, " Deep cascaded bi-network for face hallucination," in Proceedings of European Conference on Computer Vision (ECCV), 2016
[PDF] [Technical Report][Project Page]

2. S. Yang, P. Luo, C. C. Loy, X. Tang, " WIDER FACE: A face detection benchmark," in Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016
[PDF] [Project Page]

3. C. Huang, Y. Li, C. C. Loy, X. Tang, " Learning deep representation for imbalanced classification," in Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016
[PDF][Project Page]

4. S. Zhu, C. Li, C. C. Loy, X. Tang, " Unconstrained face alignment via cascaded compositional learning," in Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016
[PDF][Project Page]

5. S. Yang, P. Luo, C. C. Loy, X. Tang, " From facial part responses to face detection: A deep learning approach," in Proceedings of IEEE International Conference on Computer Vision (ICCV), 2015
[PDF] [Project Page]

6. Z. Zhang, P. Luo, C. C. Loy, X. Tang, "Learning deep representation for face alignment with auxiliary attributes," IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI), vol. 38, no. 5, pp. 918–930, 2015
[Preprint] [Project Page]

7. C. Dong, C. C. Loy, K. He, and X. Tang, "Image super-resolution using deep convolutional networks," IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI), vol. 38, no. 2, pp. 295-307, 2015
[DOI] [Preprint] [Supplementary Material] [Project Page]

8. S. Zhu, C. Li, C. C. Loy, X. Tang, "Face alignment by coarse-to-fine shape searching," in Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015
[PDF] [Project Page]

9. C. Dong, C. C. Loy, K. He, X. Tang, "Learning a deep convolutional network for image super-resolution," in Proceedings of European Conference on Computer Vision (ECCV), 2014
[PDF] [Supplementary Material] [Technical Report] [Project Page]

10. Z. Zhang, P. Luo, C. C. Loy, X. Tang, "Facial landmark detection by deep multi-task learning," in Proceedings of European Conference on Computer Vision (ECCV), 2014
[PDF] [Technical Report] [Project Page]

Backup

Warping function

The *dense face correspondence field* defines a pixel-wise correspondence mapping from $M \subset \mathbb{R}^2$ (the 2D face region in the mean face template) to the face region in image $\mathbf{I}$. We represent the dense field with a warping function [38], $\mathbf{x} = W(\mathbf{z}) : M \to \mathbb{R}^2$, which maps the coordinates $\mathbf{z} \in M$ from the mean shape template domain to the target coordinates $\mathbf{x} \in \mathbb{R}^2$. See Fig. 3(a,b) for a clear illustration. Following [39], we model the warping residual $W(\mathbf{z}) - \mathbf{z}$ as a linear combination of the dense facial deformation bases, i.e.

$$W(\mathbf{z}) = \mathbf{z} + \mathbf{B}(\mathbf{z})\mathbf{p} \tag{1}$$

where $\mathbf{p} = [p_1 \dots p_N]^\top \in \mathbb{R}^{N\times 1}$ denotes the deformation coefficients and $\mathbf{B}(\mathbf{z}) = [\mathbf{b}_1(\mathbf{z}) \dots \mathbf{b}_N(\mathbf{z})] \in \mathbb{R}^{2\times N}$ denotes the deformation bases. The N bases are chosen in the AAMs manner [40], that 4 out of N correspond to the similarity transform and the remaining for non-rigid deformations. Note that the bases are pre-defined and shared by all samples. Hence the dense field is actually controlled by the deformation coefficients $\mathbf{p}$ for each sample. When $\mathbf{p} = \mathbf{0}$, the dense field equals to the mean face template.

[38]. Alabort-i Medina, J., Zafeiriou, S.: Unifying holistic and parts-based deformable model fitting. In: CVPR. (2015)
[39]. Snape, P., Roussos, A., Panagakis, Y., Zafeiriou, S.: Face flow. In: ICCV. (2015)

High-frequency prior

High-frequency prior. We define high-frequency prior as the indication for location with high-frequency details. In this work, we generate high-frequency prior maps to enforce spatial guidance for hallucination. The prior maps are obtained from the mean face template domain. More precisely, for each training image, we compute the residual image between the original image $\hat{\mathbf{I}}$ and the bicubic interpolation of $\mathbf{I}_0$, and then warp the residual map into the mean face template domain. We average the magnitude of the warped residual maps over all training images and form the preliminary high-frequency map. To suppress the noise and provide a semantically meaningful prior, we cluster the preliminary high-frequency map into C continuous contours (10 in our implementation). We form a C-channel maps, with each channel carrying one contour. We refer this C-channel maps as our high-frequency prior, and denote it as $E_k(\mathbf{z}) : M_k \to \mathbb{R}^C$. We use $\mathbf{E}_k$ to represent $E_k(\mathbf{z})$ for all $\mathbf{z} \in M_k$. An illustration of the prior is shown in Fig. 3(c).

Network

Table 1. The architecture of the bi-network in the first cascade.

Network	Layer Index (Depth)	Kernel Size	Stride	Pad	Output Channels	Rectifier	Learining Rate (Pre-train)	Learning Rate (End-to-end)
Common Sub-net (24 layers)	1-4	3×3	1	1	64	ReLU	10^{-4}	10^{-5}
	5-20	3×3	1	1	128	ReLU	10^{-4}	10^{-5}
	21-23	3×3	1	1	32	ReLU	10^{-4}	10^{-5}
	24	3×3	1	1	1	/	10^{-5}	10^{-6}
High-frequency Sub-net (24 layers)	1-4	3×3	1	1	64	ReLU	10^{-4}	10^{-5}
	5-20	3×3	1	1	128	ReLU	10^{-4}	10^{-5}
	21-23	3×3	1	1	32	ReLU	10^{-4}	10^{-5}
	24	3×3	1	1	1	/	10^{-5}	10^{-6}
Gate Network (6 layers)	1-5	3×3	1	1	64	ReLU	/	10^{-4}
	6	3×3	1	1	1	/	/	10^{-5}

Table 2. The architecture of the bi-network in the subsequent cascades.

Network	Layer Index (Depth)	Kernel Size	Stride	Pad	Output Channels	Rectifier	Learining Rate (Pre-train)	Learning Rate (End-to-end)
Common Sub-net (12 layers)	1-4	3×3	1	1	64	ReLU	10^{-5}	10^{-6}
	5-8	3×3	1	1	128	ReLU	10^{-5}	10^{-6}
	9-11	3×3	1	1	32	ReLU	10^{-5}	10^{-6}
	12	3×3	1	1	1	/	10^{-6}	10^{-7}
High-frequency Sub-net (12 layers)	1-4	3×3	1	1	64	ReLU	10^{-5}	10^{-6}
	5-8	3×3	1	1	128	ReLU	10^{-5}	10^{-6}
	9-11	3×3	1	1	32	ReLU	10^{-5}	10^{-6}
	12	3×3	1	1	1	/	10^{-6}	10^{-7}
Gate Network (6 layers)	1-5	3×3	1	1	64	ReLU	/	10^{-5}
	6	3×3	1	1	1	/	/	10^{-6}