Hierarchical Bilinear Network for High Performance Face Detection

Jiangjing Lv, Xiaohu Shao, Junliang Xing, Pengcheng Liu, Xiangdong Zhou, Xi Zhou

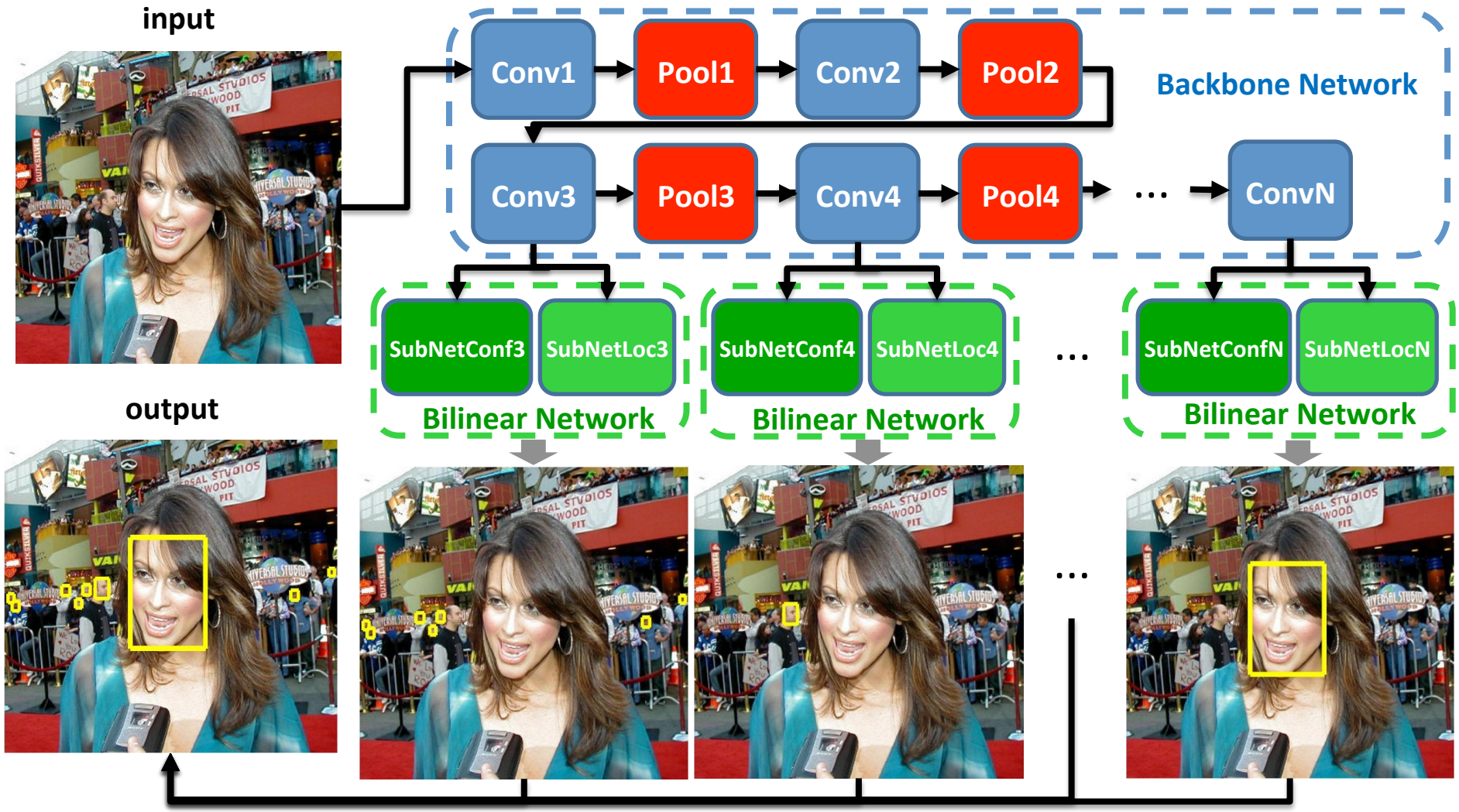
Background

- Traditional face detection
  - Real-time
  - Limited to severer view variations
- Deep Convolutional Network (DCN) based face detection
  - High performance
  - Suffer from high complexity and large size of model

Motivation

- Backbone Network
  - Arbitrary resolution as input
  - Feature pyramid for multi-scale face detection
- Bilinear Network
  - Confidence sub-network predicts face confidence
  - Localization sub-network regresses the face bounding boxes
  - Weights sharing (tiny model size)

The framework of HBN



Backbone Network

Layer Name	Filter Size	Stride	Pad	Parameter Number
Conv1	16 × 5 × 5	2	2	1.2K
Pool1	2×2	2	0	0
Conv2	24 × 3 × 3	1	1	3.4K
Pool2	2×2	2	0	0
Conv3	CH × 3 × 3	1	1	10.1K
Pool3	2×2	2	0	0
Conv4	CH × 3 × 3	1	1	10.1K
Pool4	2×2	2	0	0
Conv5	CH × 3 × 3	1	1	10.1K
Pool5	2×2	2	0	0
Conv6	CH × 3 × 3	1	1	10.1K

- **CH** is channel number, default set to 48.
- Total **45K** parameters.
- Spatial resolution is gradually reduced.

Backbone Network

Layer	Conv3		Conv4		Conv5		Conv6	
Scale	12^2	24^2	48^2	96^2	144^2	192^2	288^2	384^2
Proposal	12×12	24×24	48×48	96×96	144×144	192×192	288×288	384×384
	8×17	17×34	34×68	68×136	102×204	136×272	204×407	272×543
	17×8	34×17	68×34	136×68	204×102	272×136	407×204	543×272

- Reference bounding boxes
  - aspect ratios {1, 1/2, 2}
- Face Size:
  - From 8 pixels to 543 pixels

Bilinear Network

- **Confidence sub-networks**
  - Classification
  - ***Inception:*** Contextual information with different resolutions
  - SoftMax loss

- **Localization Sub-network**
  - 3×3 convolutional layer is used to predict offset information
  - Smooth L1 loss

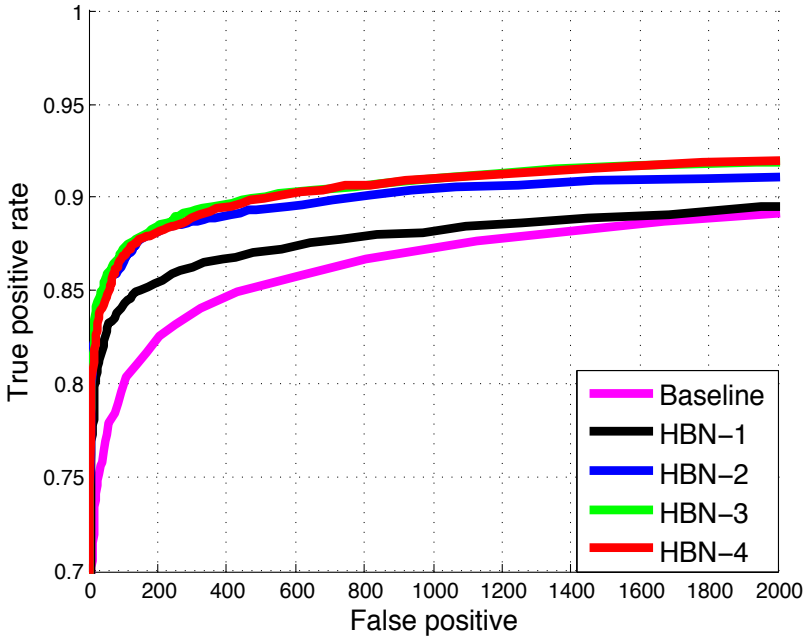
Experiments

- **Experimental settings**
  - Caffe platform
  - WIDER FACE training dataset
  - Hard negative mining
    - The ratio between the negatives and positives is **3:1**
  - Data augmentation
    - Scale, Blur, Noises, Mirror flip
  - etc.

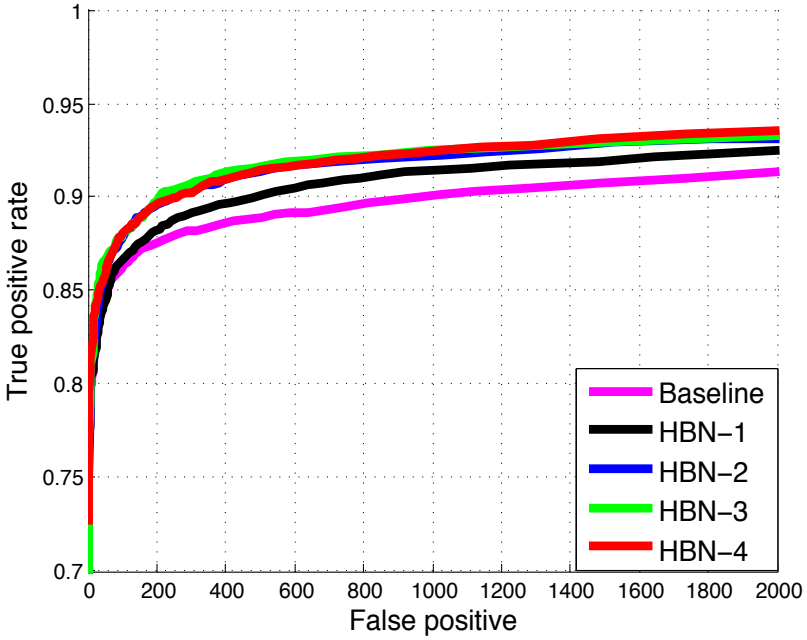
Model Analyses

Model	Conv3	Conv4	Conv5	Conv6
Baseline	C	C	C	C
HBN-1	I+C	C	C	C
HBN-2	I+C	I+C	C	C
HBN-3	I+C	I+C	I+C	C
HBN-4	I+C	I+C	I+C	I+C

The configuration of different confidence sub-networks, **I** and **C** represent the Inception module and the convolutional layer respectively

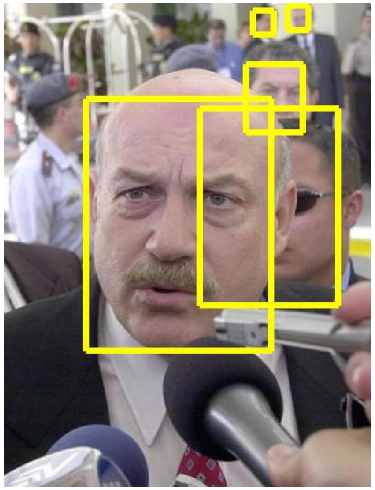


(a) CH=48

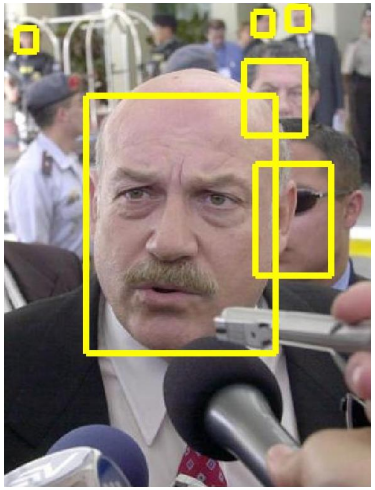


(b) CH=64

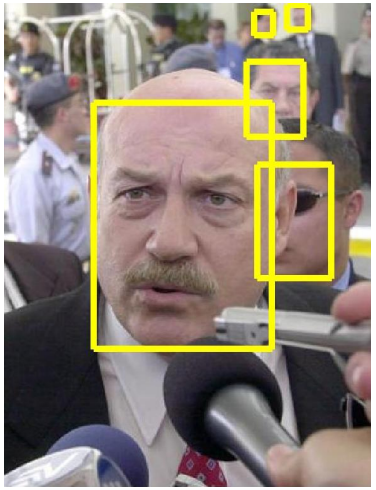
Baseline



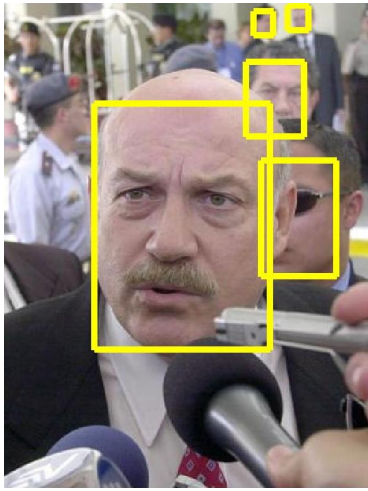
HBN-1



HBN-2



HBN-3



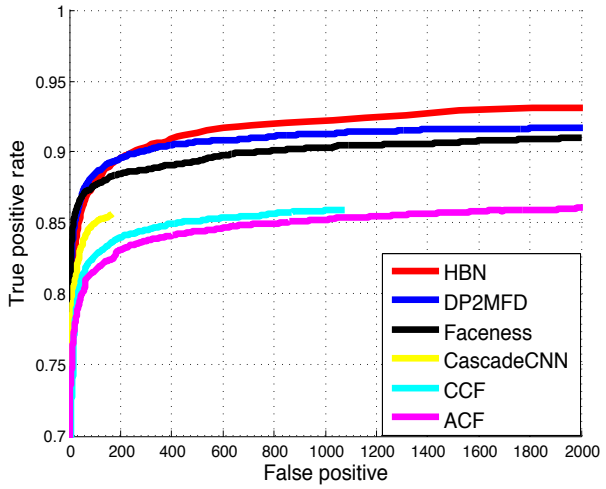
HBN-4

Model sizes and speeds

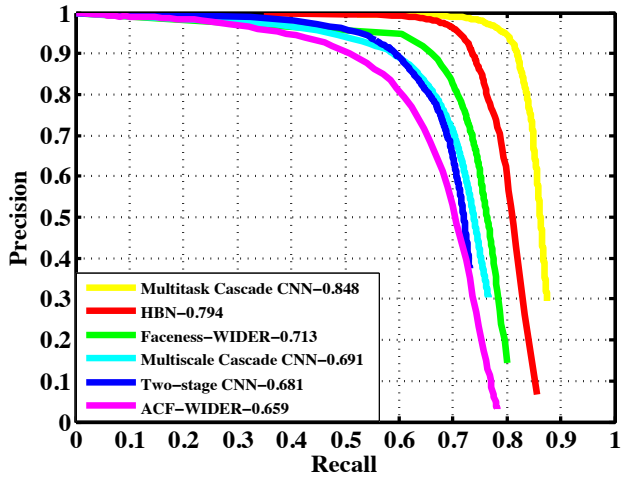
	Model	Baseline	HBN-1	HBN-2	HBN-3	HBN-4
CH=48	CS(MB)	0.6	1.3	**2.0**	2.7	3.7
	TS(MB)	0.4	1.1	**1.1**	1.1	1.1
	Speed(FPS)	110	79	**72**	67	62
CH=64	CS(MB)	0.8	1.5	**2.2**	3.0	3.7
	TS(MB)	0.6	1.3	**1.3**	1.3	1.3
	Speed(FPS)	97	71	**69**	65	58

CS and **TS** represent the size of model storaged by the Caffe platform and the theoretical size of model respectively.

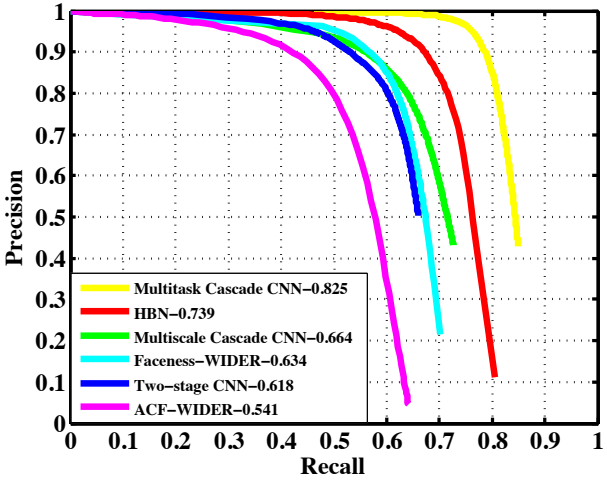
Comparison with other methods



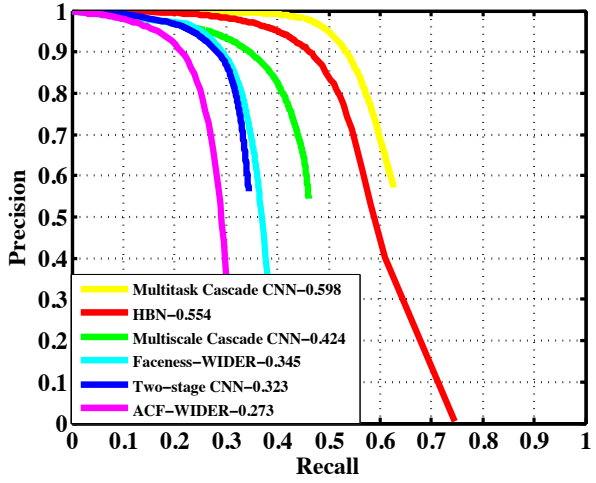
(a) FDDB



(b) Easy Set

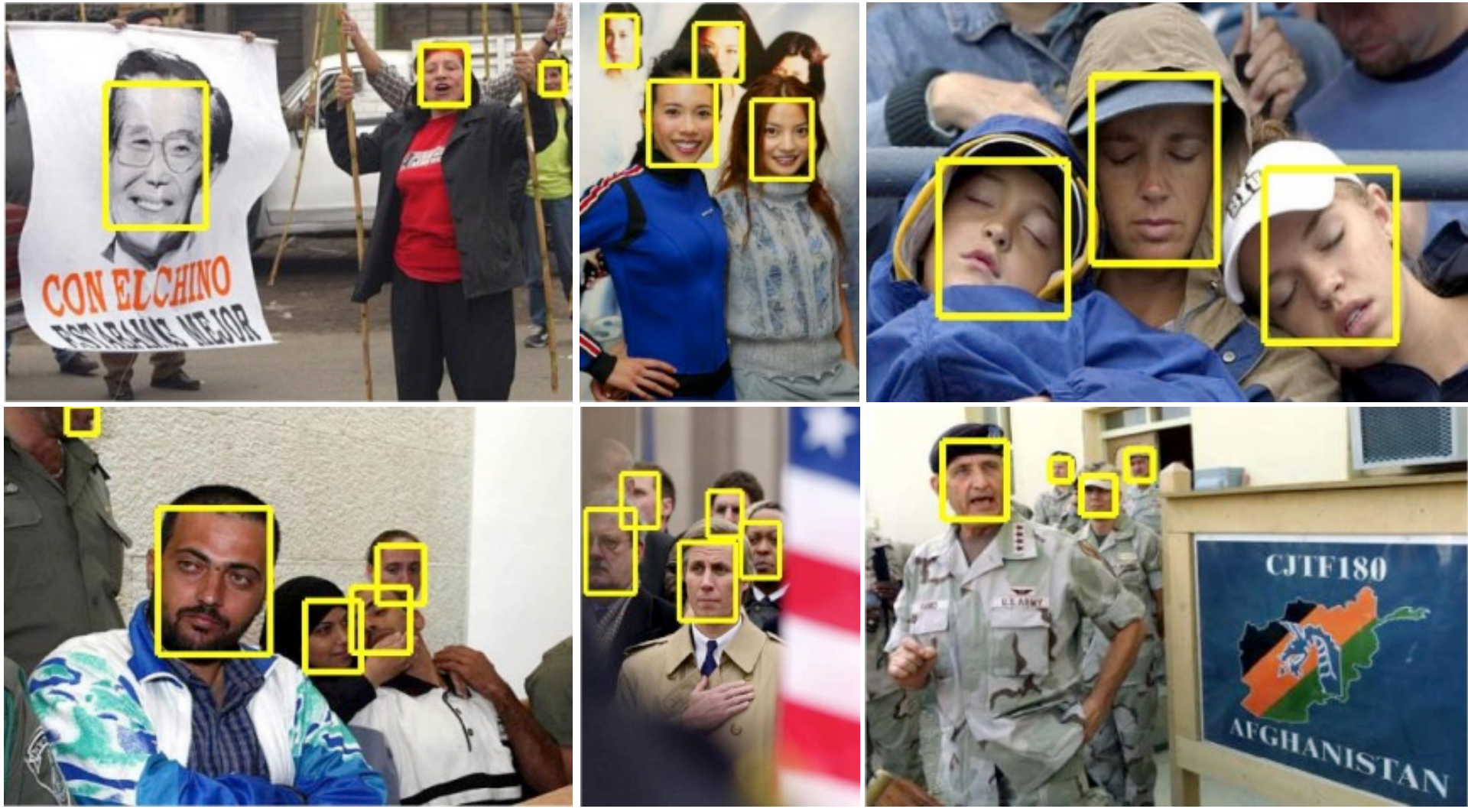


(c) Medium Set

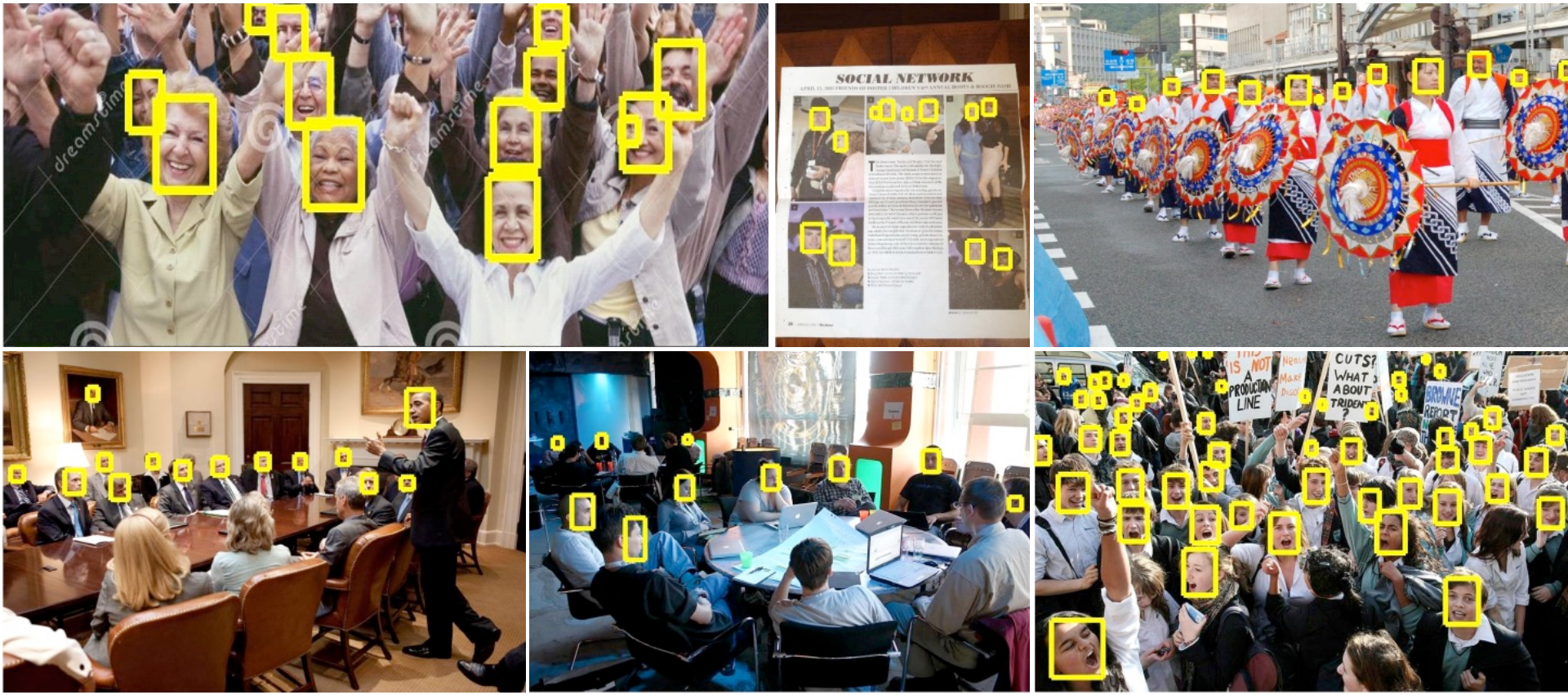


(d) Hard Set

Examples of our face detection method



Examples of our face detection method



Conclusion

- End-to-end face detection method for different scale faces.
- Tiny model size by weights sharing.
- Fast face detection.
- Promising results.

Thank You!

jiangjing.ljj@alibaba-inc.com