Large-Scale Face Recognition

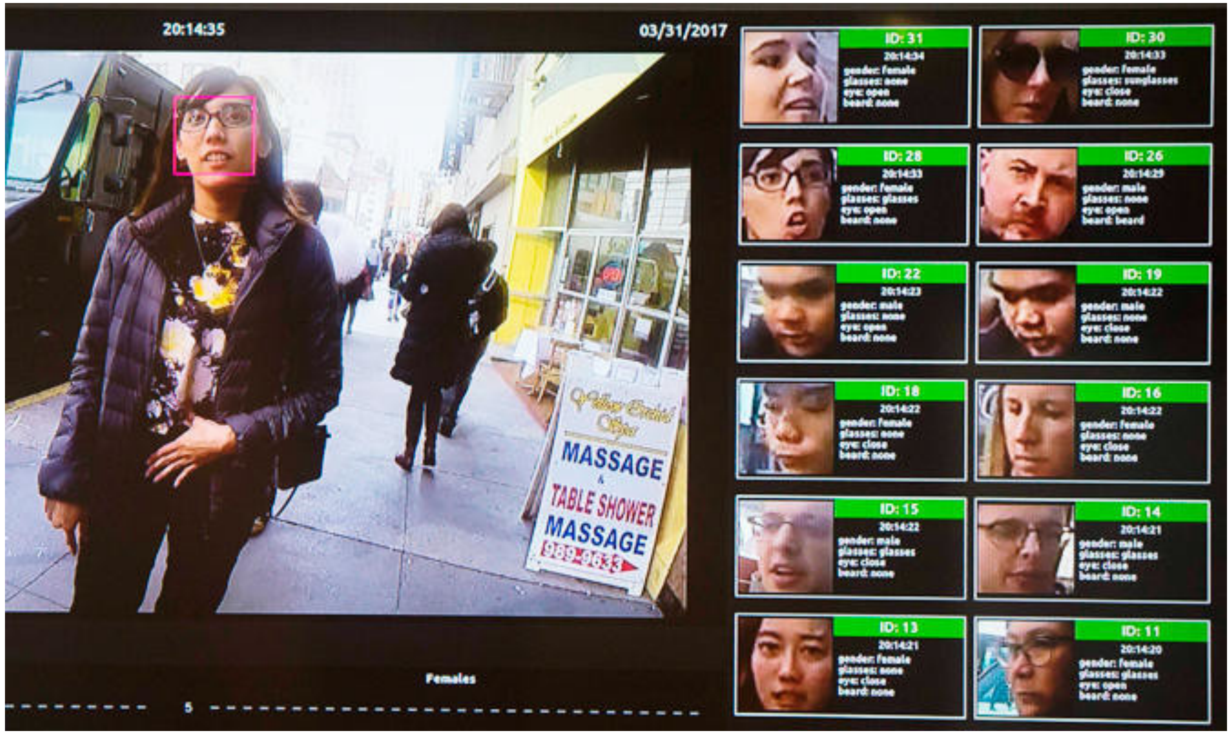
Lei Zhang

Microsoft

October 2, 2018

Face Recognition

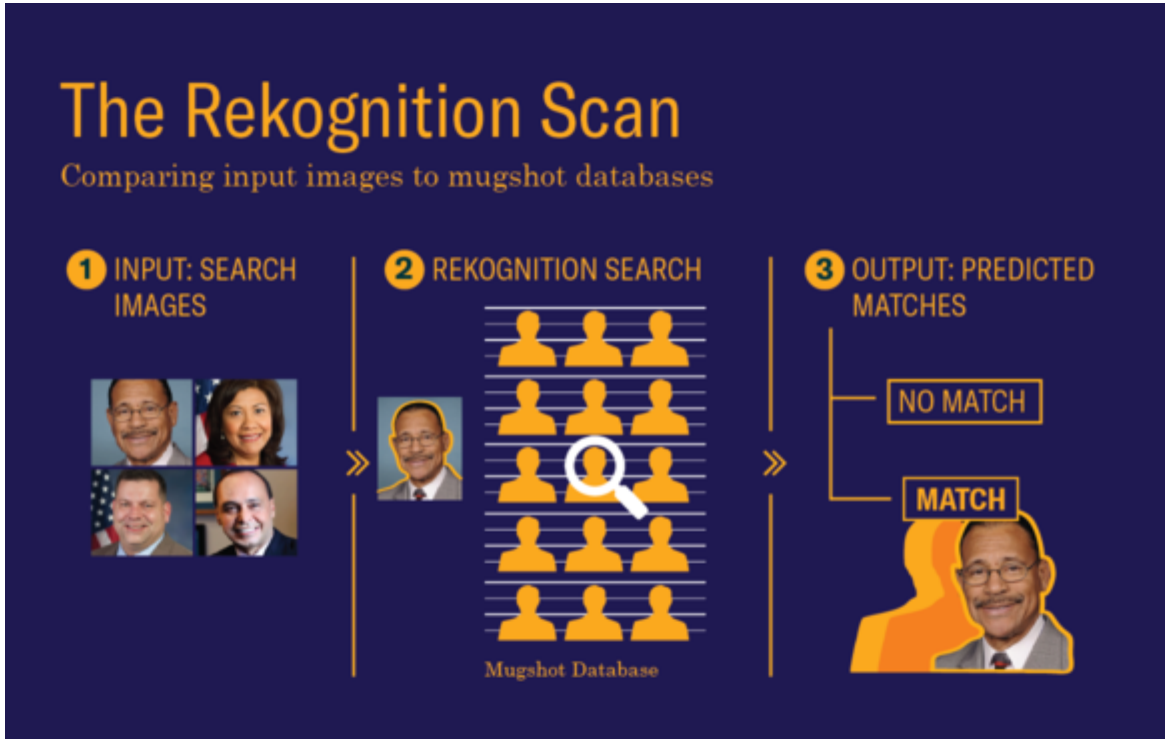
- Face recognition has been greatly advanced in recent years due to the breakthrough in deep learning
- Many real applications
    - Security & law enforcement
    - Financial authentication
    - Airports
    - Brands & PR agencies
    - Targeted advertising
    - …

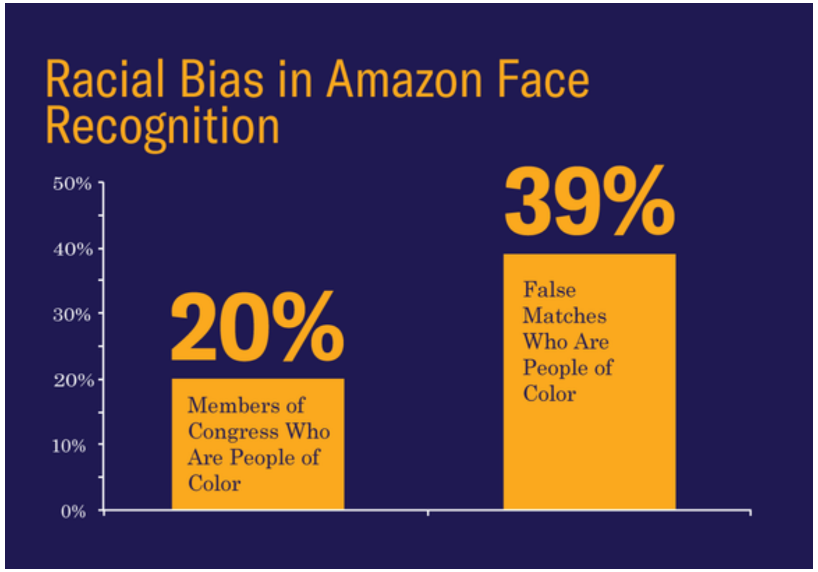


http://www.arabnews.com/node/1339101/science-technology

[Chinese park installs facial recognition software to stop toilet paper thieves](), 03/2017

[Amazon's Face Recognition Falsely Matched 28 Members of Congress With Mugshots](), 07/2018

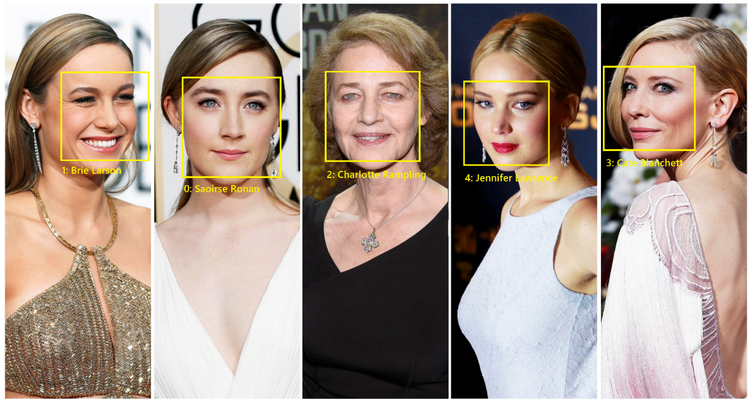




Face Recognition – Problem Definition

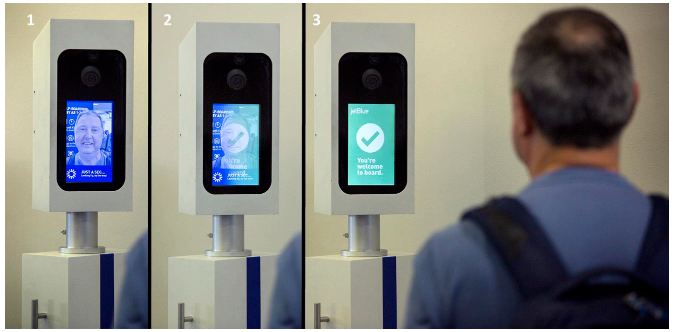
- Face verification – 1 vs. 1
  - Given two faces, answer if they are the same person or not
  - Example application: Phone unlocking

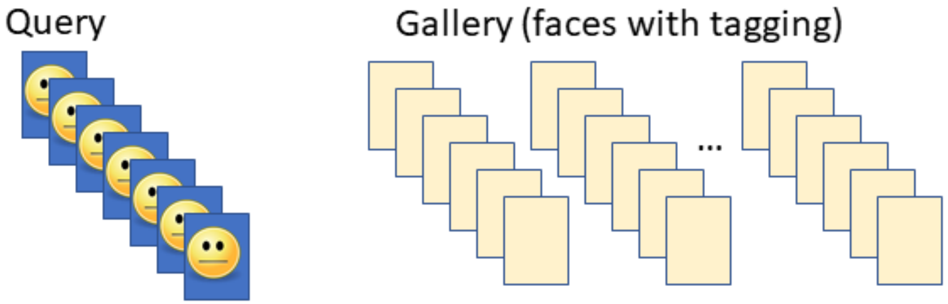
- Face identification – 1 vs. N
  - Given one face, answer whom he/she is among N people, or reject
  - Example application: Celebrity recognition



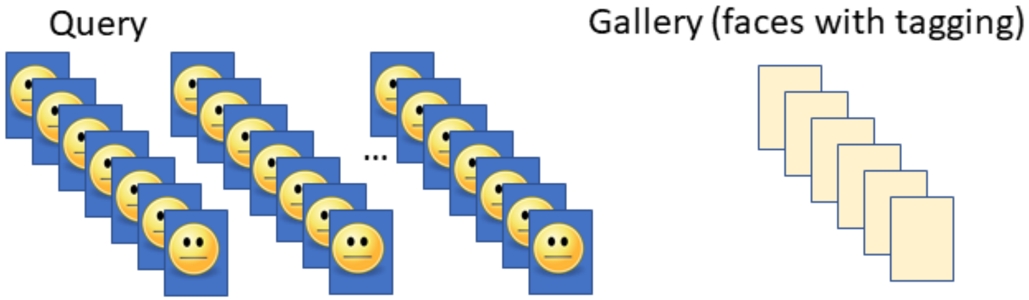
Face Recognition – Problem Definition

- Face Identification – M vs. N (M << N)

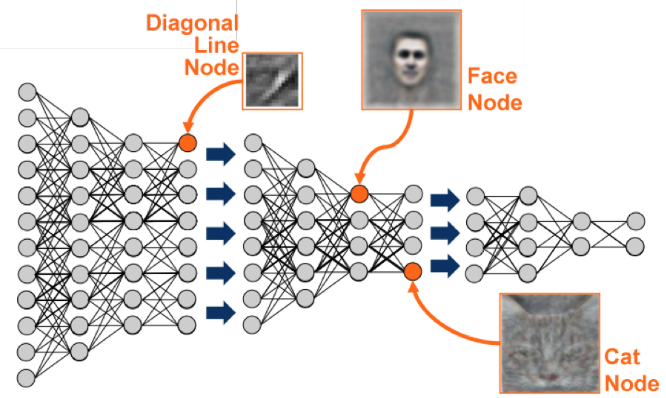


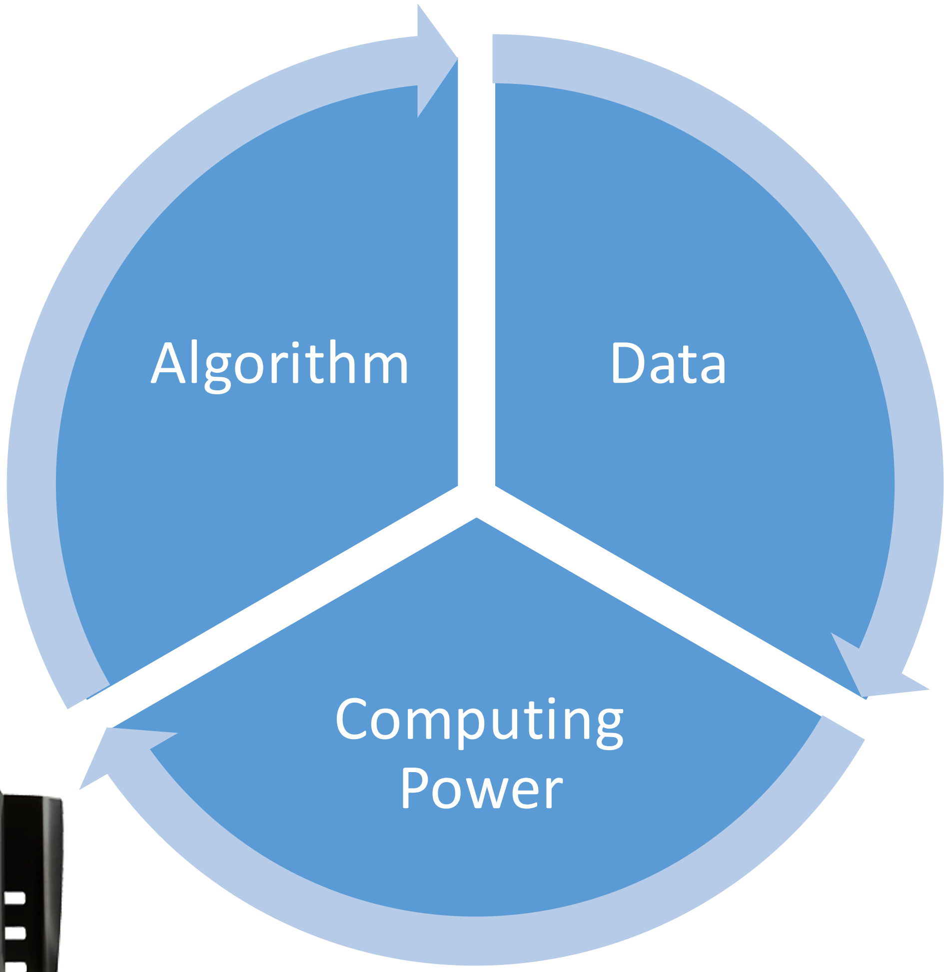


- Face Identification – M vs. N (M >> N)

Driving Forces Behind Face Recognition





?

Public (and Private) Face Datasets

Dataset	Public?	# of People	# of Faces
LFW	public	5k	13k
YFD	public	1.5k	3.4 k videos
CelebFaces	public	10k	202k
CASIA-WebFace	public	10k	500k
MS-Celeb-1M	public	100k	About 8,456k
Facebook	private	4k	4,400k
Google	private	8000k	100-200m

Mei Wang and Weihong Deng. "Deep Face Recognition: A Survey." *arXiv preprint arXiv:1804.06655* (2018). v7, 9/28/2018

TABLE IV
THE ACCURACY OF DIFFERENT VERIFICATION METHODS ON THE LFW DATASET.

Method	Public. Time	Loss	Architecture	Number of Networks	Training Set	Accuracy±Std(%)
DeepFace [160]	2014	softmax	Alexnet	3	Facebook (4.4M,4K)	97.35±0.25
DeepID2 [152]	2014	contrastive loss	Alexnet	25	CelebFaces+ (0.2M,10K)	99.15±0.13
DeepID3 [153]	2015	contrastive loss	VGGNet-10	50	CelebFaces+ (0.2M,10K)	99.53±0.10
FaceNet [144]	2015	triplet loss	GoogleNet-24	1	Google (500M,10M)	99.63±0.09
Baidu [105]	2015	triplet loss	CNN-9	10	Baidu (1.2M,18K)	99.77
VGGface [123]	2015	triplet loss	VGGNet-16	1	VGGface (2.6M,2.6K)	98.95
light-CNN [188]	2015	softmax	light CNN	1	MS-Celeb-1M (8.4M,100K)	98.8
Center Loss [181]	2016	center loss	Lenet+-7	1	CASIA-WebFace, CACD2000, Celebrity+ (0.7M,17K)	99.28
L-softmax [107]	2016	L-softmax	VGGNet-18	1	CASIA-WebFace (0.49M,10K)	98.71
Range Loss [224]	2016	range loss	VGGNet-16	1	MS-Celeb-1M, CASIA-WebFace (5M,100K)	99.52
L2-softmax [129]	2017	L2-softmax	ResNet-101	1	MS-Celeb-1M (3.7M,58K)	99.78
Normface [171]	2017	contrastive loss	ResNet-28	1	CASIA-WebFace (0.49M,10K)	99.19
CoCo loss [111]	2017	CoCo loss	-	1	MS-Celeb-1M (3M,80K)	99.86
vMF loss [62]	2017	vMF loss	ResNet-27	1	MS-Celeb-1M (4.6M,60K)	99.58
Marginal Loss [39]	2017	marginal loss	ResNet-27	1	MS-Celeb-1M (4M,80K)	99.48
SphereFace [106]	2017	A-softmax	ResNet-64	1	CASIA-WebFace (0.49M,10K)	99.42
CCL [128]	2018	center invariant loss	ResNet-27	1	CASIA-WebFace (0.49M,10K)	99.12
AMS loss [170]	2018	AMS loss	ResNet-20	1	CASIA-WebFace (0.49M,10K)	99.12
Cosface [172]	2018	cosface	ResNet-64	1	CASIA-WebFace (0.49M,10K)	99.33
Arcface [38]	2018	arcface	ResNet-100	1	MS-Celeb-1M (3.8M,85K)	99.83
Ring loss [235]	2018	Ring loss	ResNet-64	1	MS-Celeb-1M (3.5M,31K)	99.50

The Story Behind MS-Celeb-1M

A Grand Challenge in Search Engine

– Can We Recognize As Many As Possible Entities on the Web?

  – *How Many? And How Accurate?*

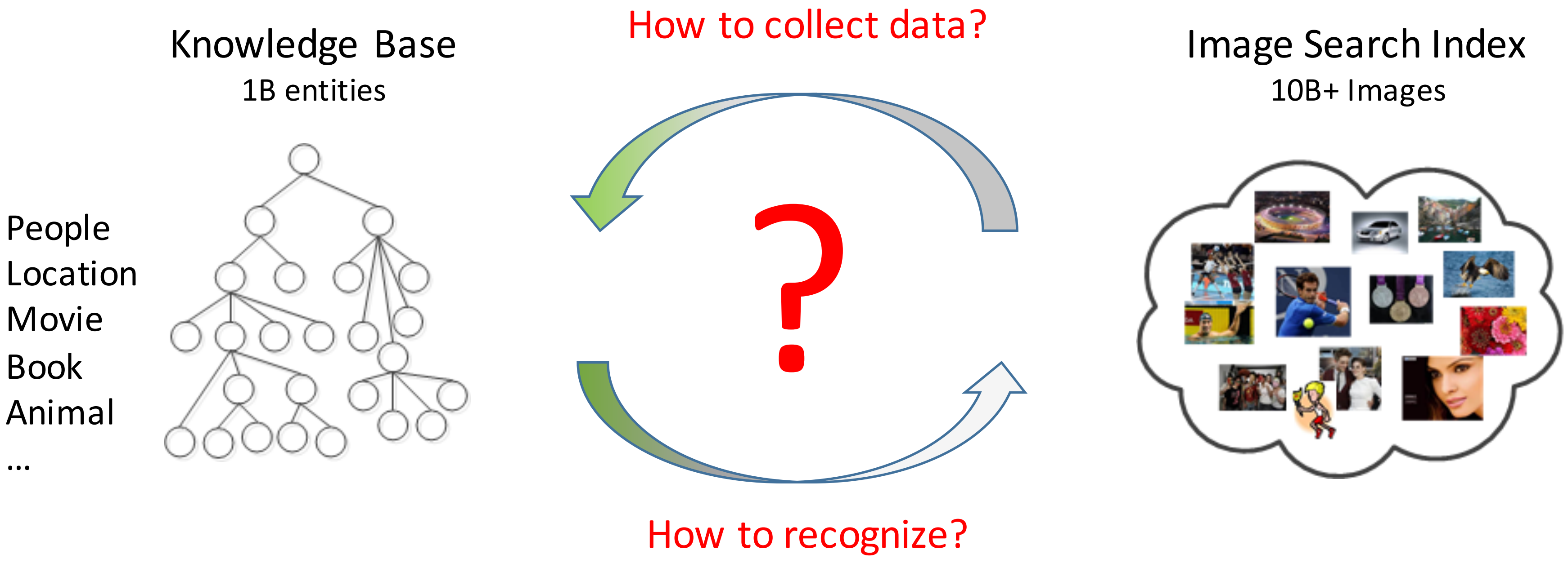
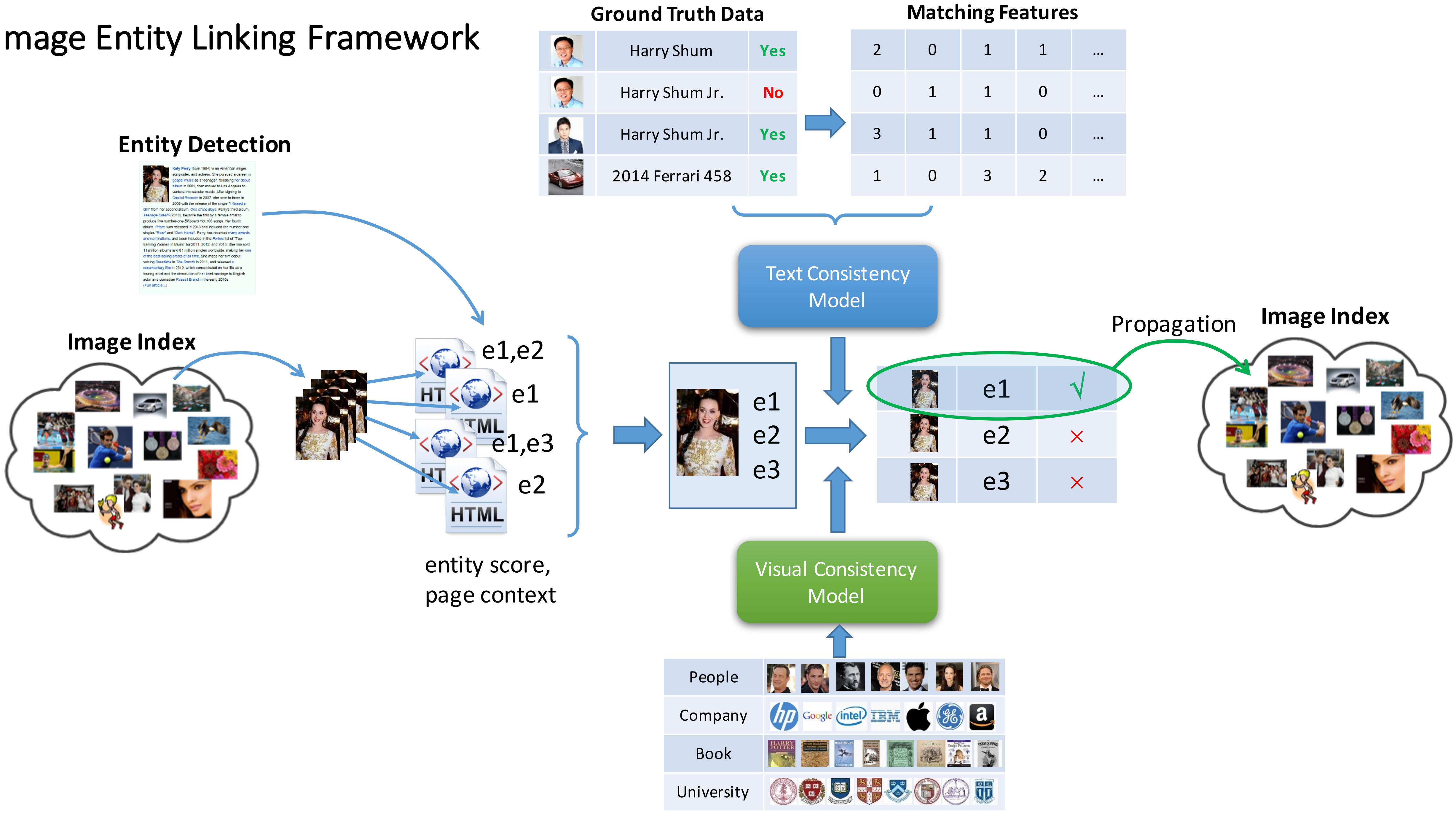


Image Entity Linking Framework



Overall Results

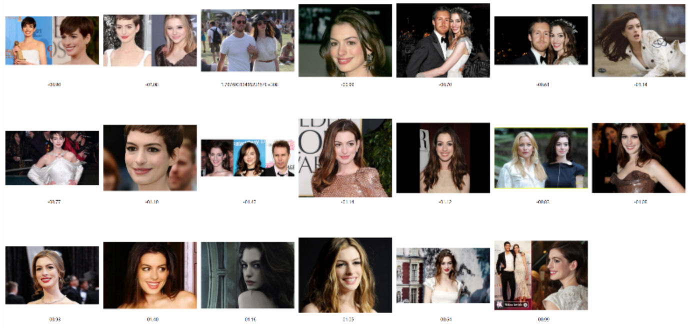
- People Segment

	Coverage (# Image)	# Entity	Precision*
V2 (Text + Visual)	**93M (+70%)**	**300K**	**98.5%**
V1 (Text)	54M	300K	98.6%

* Measured on 2.5K name queries and their top 10 resulting images

- More segments (ongoing):
  - Location/attraction entities
  - Movie entities
  - Animal/dog breed/cat breed
  - Plant/flower
  - …

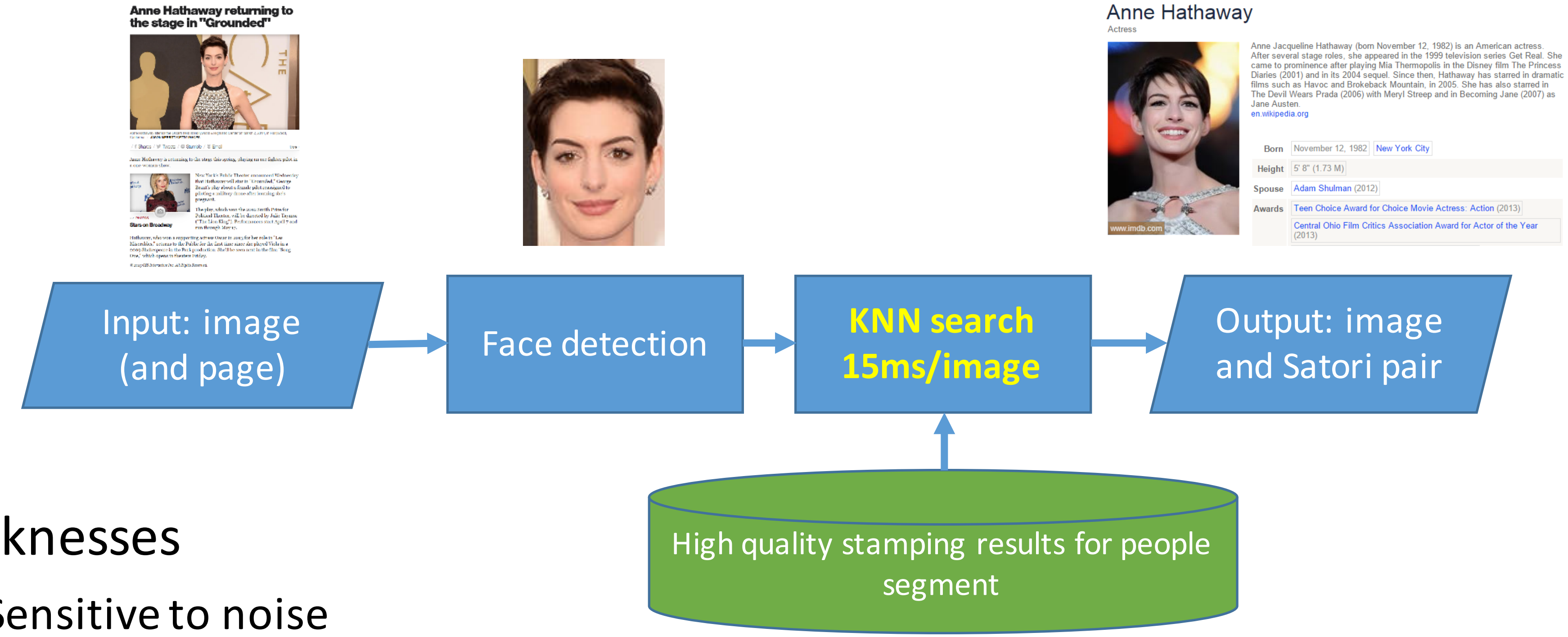
Anne Hathaway (22K images)



Justine Bieber (133K images)
Selena Gomez (128K images)
Miley Cyrus (111K images)
…

Instance-based KNN Search

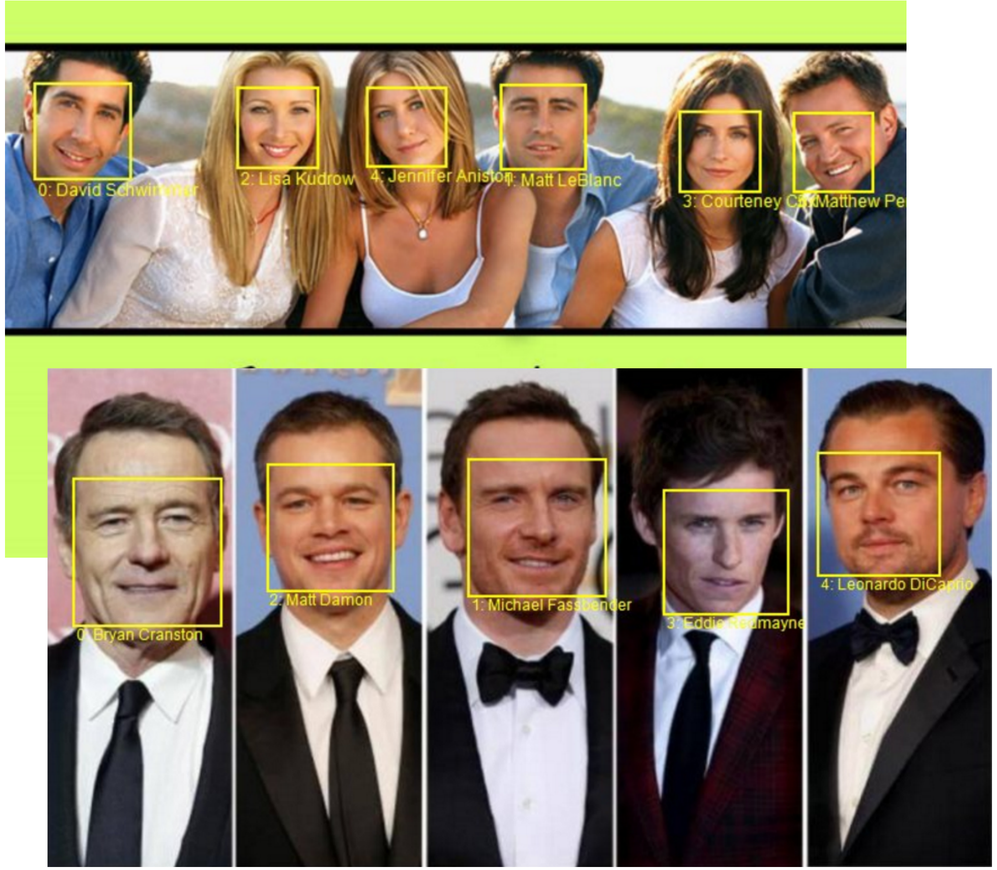
- Key Idea
  - Based on the high quality stamping results, build a high precision celebrity recognition engine



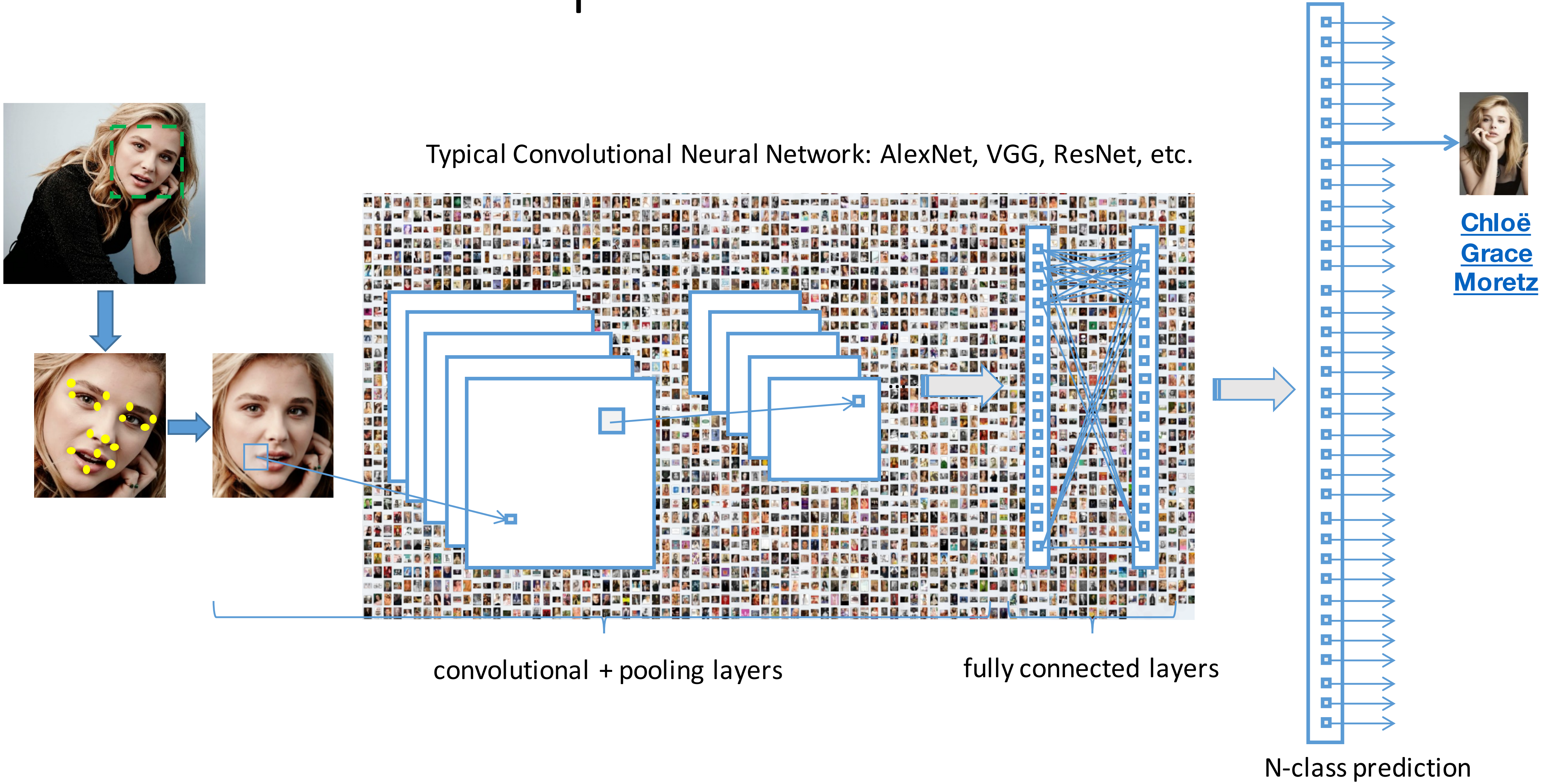
- Weaknesses
  - Sensitive to noise
  - Limited generalization ability

One Step Further – Celebrity Recognition

- Can we recognize people purely based on image pixels?



Model-based People Identification



In Microsoft Cognitive Service

Microsoft

Cognitive Services


```
"detail": {
  "celebrities": [
    {
      "name": "Harry Shum",
      "faceRectangle": {
        "left": 253,
        "top": 116,
        "width": 70,
        "height": 70
      },
      "confidence": 0.9997298
    }
  ]
}
}
],
"tags": [
  {
    "name": "person",
```


Recognize celebrities

The Celebrity Model is an example of Domain Specific Models. Our new celebrity recognition model recognizes 200K celebrities from business, politics, sports and entertainment around the World.

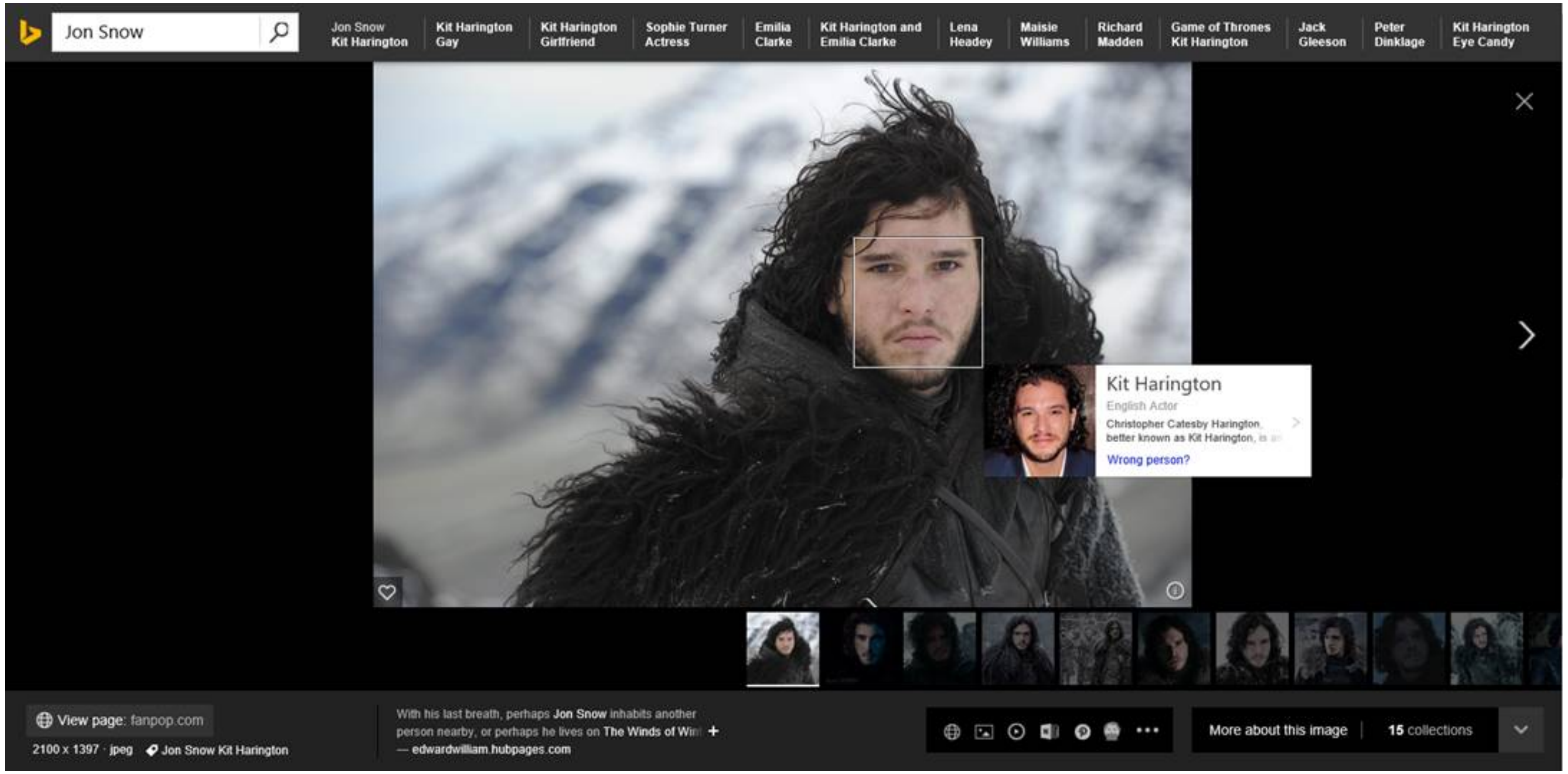
In Image Caption (captionbot.ai)

Kenneth Tran, et al, CVPRW 2016

Sasha Obama, Malia Obama, Michelle Obama, Peng Liyuan et al. posing for a picture with Forbidden City in the background.

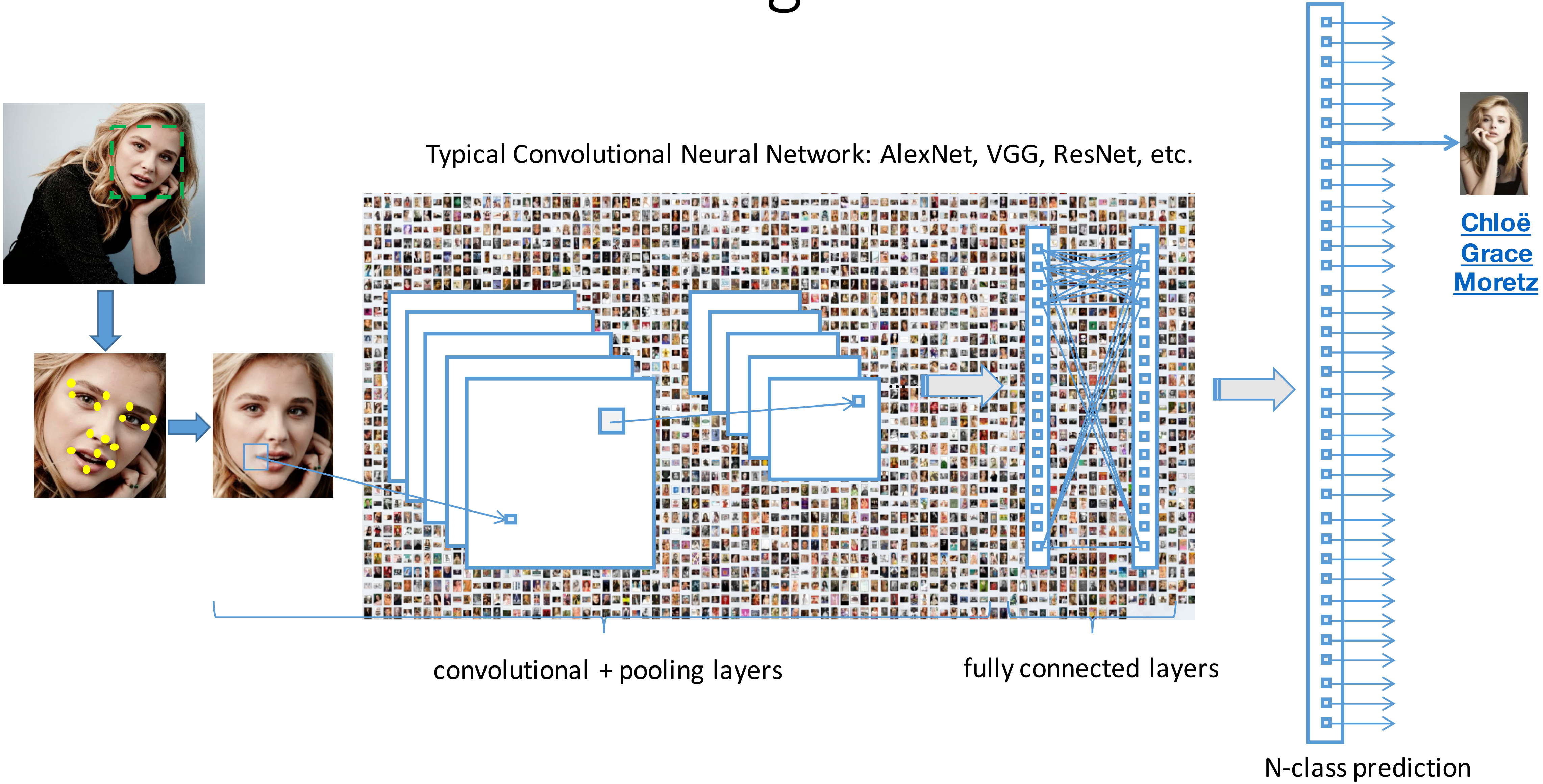
In XiaoIce (小冰)

In Bing Image Search

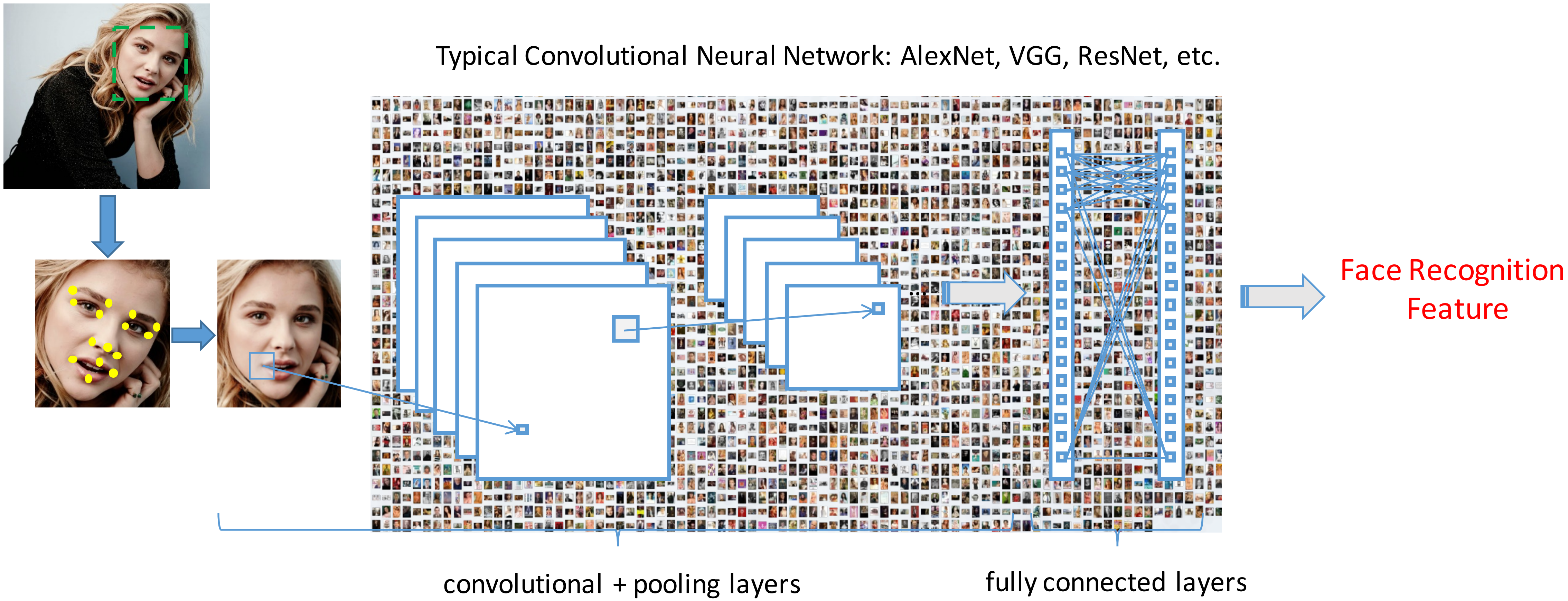


- More examples: [steve jobs actor](), [friends]()

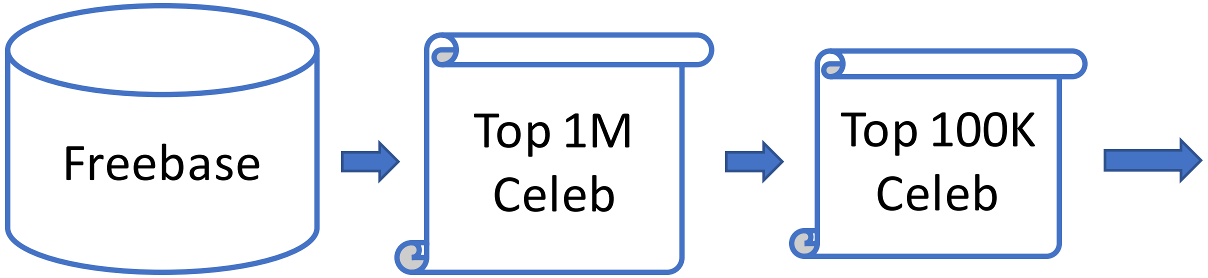
Towards Best Face Recognition Feature

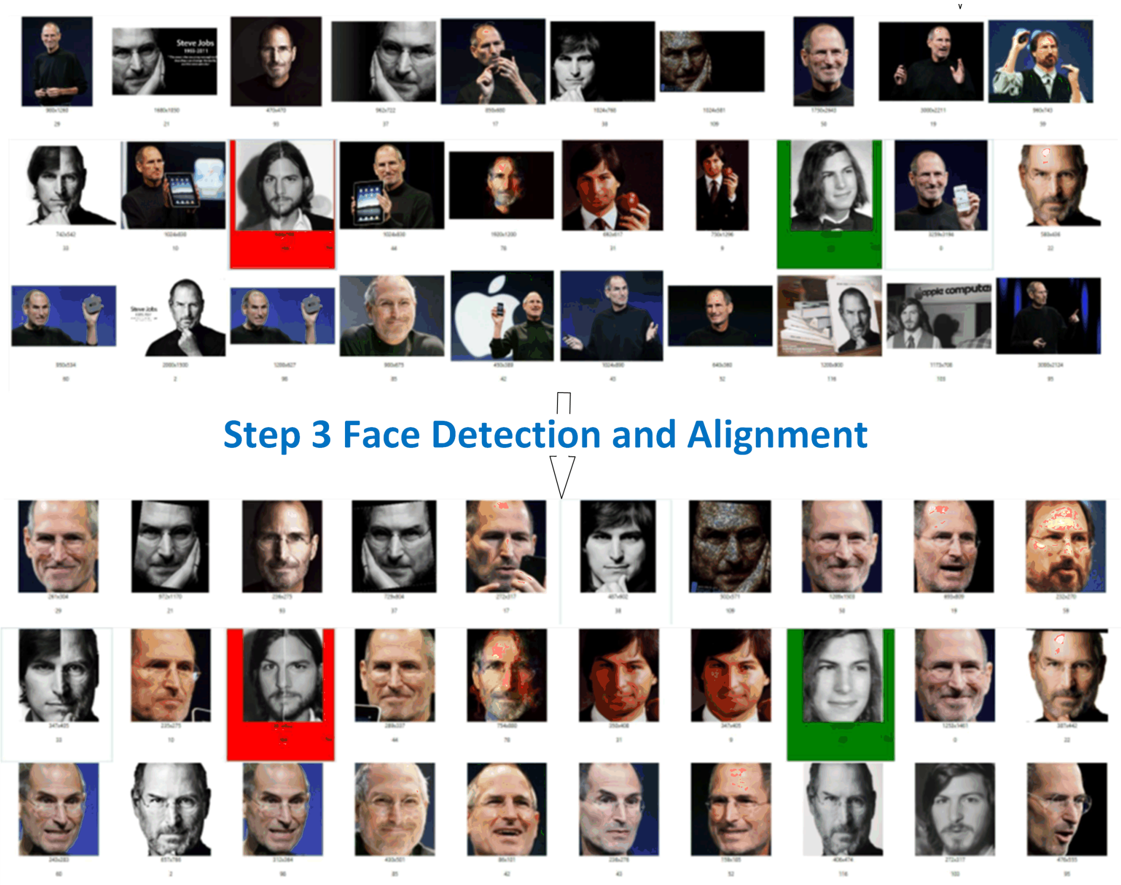


Towards Best Face Recognition Feature



Making Data Public – **Training Data**

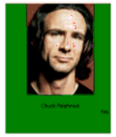




- Top 100K celebrity
- About 10M images
- Noisy label
- Cropped/Aligned versions

Making Data Public – **Measurement Data**

Step 3 Remove wrong faces, Select Two images per celebrity

Random: 

Hard: 

		# of Images	GT Published
Development Set	Random (Easy)	500	Yes
	Hard	500	Yes
Measurement Set	Random (Easy)	1000	No
	Hard	1000	No

Download links

- Training data
  https://www.msceleb.org/download/cropped
  https://www.msceleb.org/download/aligned
- Development data
  https://www.msceleb.org/download/devset

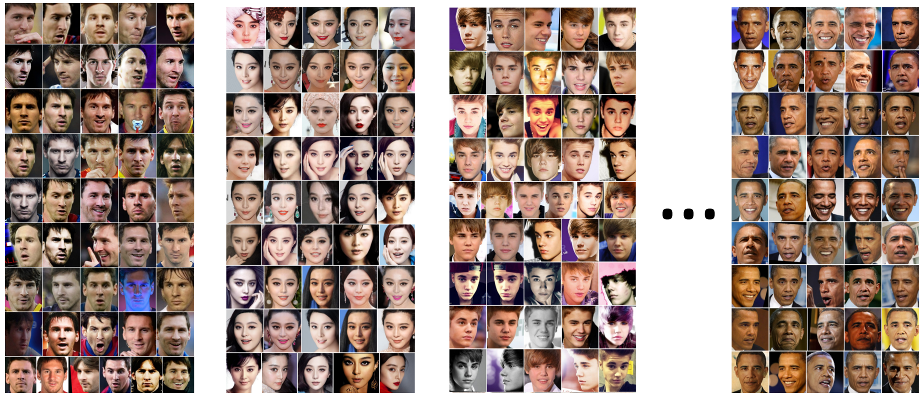
Yandong Guo, Lei Zhang, Yuxiao Hu, Xiaodong Guo, Jianfeng Gao. MS-Celeb-1M: A dataset and benchmark for large-scale face recognition. ECCV 2016.

One-Shot Face Recognition

- Learning best face representation
- Dealing with imbalanced data

Know you at **One** Glance

- Problem to Solve
  - **Limited number of training images** for some persons, in the scenario of large-scale face recognition



base set: many persons, many images per person

low-shot set : only one image per person

As shown, the training image could be faces with occlusion, drawings, or low resolution images

- Great value to study one-shot visual recognition
  - Naturally happens when the number of persons to be recognized is very large

Benchmark Task – MS-Celeb-1M Challenge #2

- To study this problem, we design and publish[1] the following task

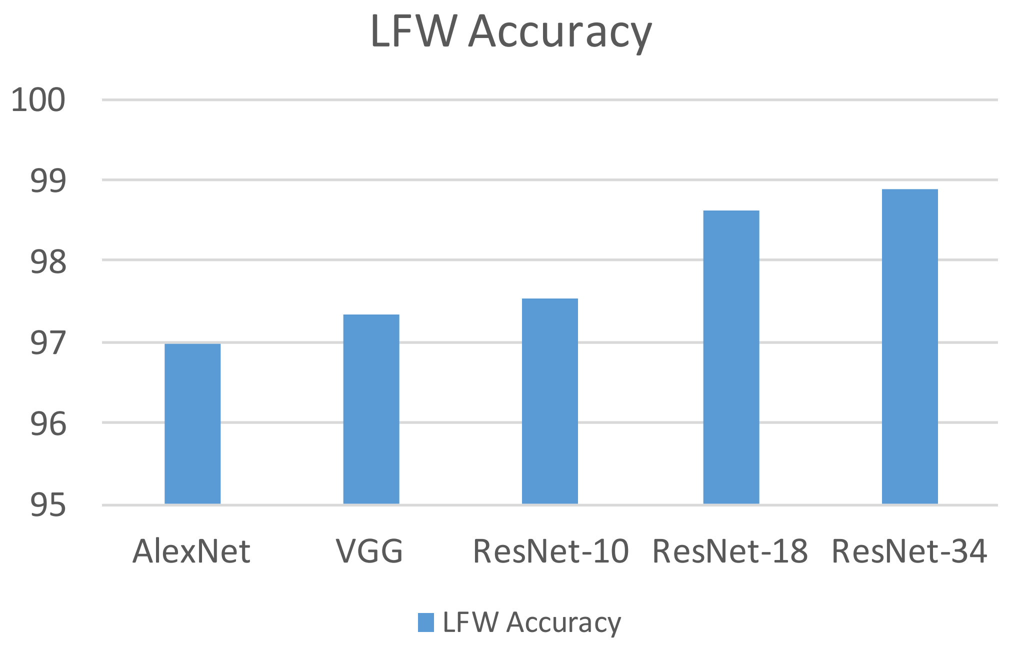
	Training	Testing
Base set: 20K persons	50-100 images/person	5 images/person
Low-shot set: 1K persons	**one** image/person	20 images/person

- **Goal**
  - Build a 21K-class classifier to recognize all the persons (in total 21K) in both the base and low-shot sets
- **Metric**
  - Mainly focus on the performance for persons in the low-shot set (coverage@high precision)
  - Keep good performance for persons in the base set

[1] http://www.msceleb.org/

Challenge One: Face Representation Learning

- Objective: to find face representations for the low-shot classes
- Solution: using the base set to train face representation model with **good generalization capability**
  - Train **deep** CNN model with **large-scale** training data
  - Add **additional loss** for better feature
- Evaluation on the LFW verification task
  - Our base set excludes celebrities in LFW by design => good generalization capability (human 97%)



Improve Face Feature with Additional Loss

- Many loss terms developed
  - Triplet Loss, Center Loss, Marginal Loss, SphereFace, Range Loss, Ring Loss, Cosine Loss
- Key Ideas Behind
  - Reduce intra-class variance while increasing inter-class variance

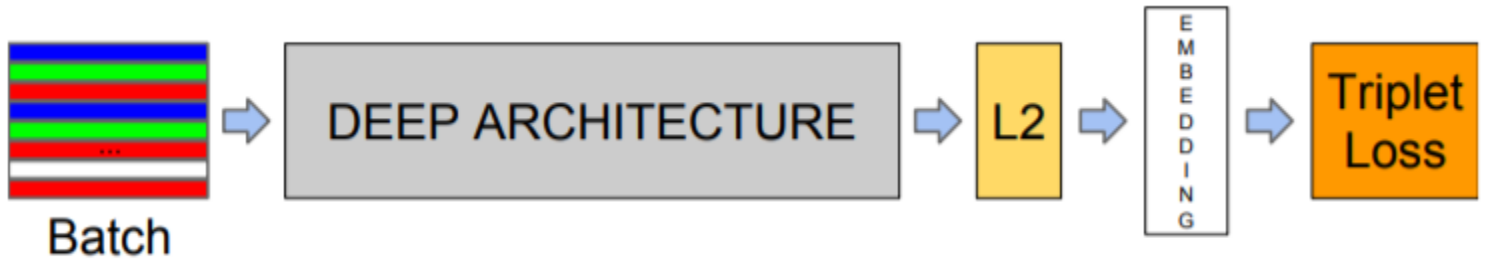
Mei Wang and Weihong Deng. "Deep Face Recognition: A Survey." *arXiv preprint arXiv:1804.06655* (2018). v7, 9/28/2018

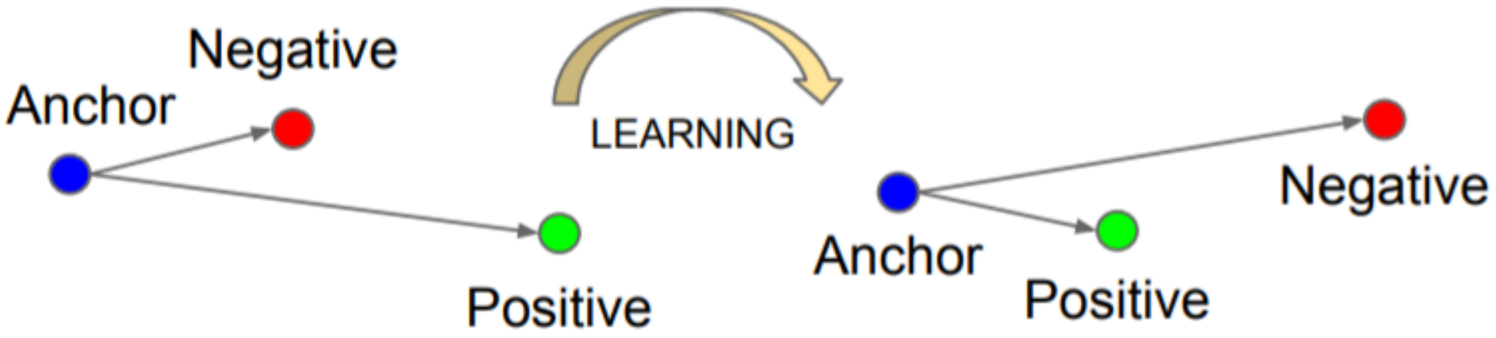
TABLE IV
THE ACCURACY OF DIFFERENT VERIFICATION METHODS ON THE LFW DATASET.

Method	Public. Time	Loss	Architecture	Number of Networks	Training Set	Accuracy$\pm$Std(%)
DeepFace [160]	2014	softmax	Alexnet	3	Facebook (4.4M,4K)	97.35$\pm$0.25
DeepID2 [152]	2014	contrastive loss	Alexnet	25	CelebFaces+ (0.2M,10K)	99.15$\pm$0.13
DeepID3 [153]	2015	contrastive loss	VGGNet-10	50	CelebFaces+ (0.2M,10K)	99.53$\pm$0.10
FaceNet [144]	2015	triplet loss	GoogleNet-24	1	Google (500M,10M)	99.63$\pm$0.09
Baidu [105]	2015	triplet loss	CNN-9	10	Baidu (1.2M,18K)	99.77
VGGface [123]	2015	triplet loss	VGGNet-16	1	VGGface (2.6M,2.6K)	98.95
light-CNN [188]	2015	softmax	light CNN	1	MS-Celeb-1M (8.4M,100K)	98.8
Center Loss [181]	2016	center loss	Lenet+-7	1	CASIA-WebFace, CACD2000, Celebrity+ (0.7M,17K)	99.28
L-softmax [107]	2016	L-softmax	VGGNet-18	1	CASIA-WebFace (0.49M,10K)	98.71
Range Loss [224]	2016	range loss	VGGNet-16	1	MS-Celeb-1M, CASIA-WebFace (5M,100K)	99.52
L2-softmax [129]	2017	L2-softmax	ResNet-101	1	MS-Celeb-1M (3.7M,58K)	99.78
Normface [171]	2017	contrastive loss	ResNet-28	1	CASIA-WebFace (0.49M,10K)	99.19
CoCo loss [111]	2017	CoCo loss	-	1	MS-Celeb-1M (3M,80K)	99.86
vMF loss [62]	2017	vMF loss	ResNet-27	1	MS-Celeb-1M (4.6M,60K)	99.58
Marginal Loss [39]	2017	marginal loss	ResNet-27	1	MS-Celeb-1M (4M,80K)	99.48
SphereFace [106]	2017	A-softmax	ResNet-64	1	CASIA-WebFace (0.49M,10K)	99.42
CCL [128]	2018	center invariant loss	ResNet-27	1	CASIA-WebFace (0.49M,10K)	99.12
AMS loss [170]	2018	AMS loss	ResNet-20	1	CASIA-WebFace (0.49M,10K)	99.12
Cosface [172]	2018	cosface	ResNet-64	1	CASIA-WebFace (0.49M,10K)	99.33
Arcface [38]	2018	arcface	ResNet-100	1	MS-Celeb-1M (3.8M,85K)	99.83
Ring loss [235]	2018	Ring loss	ResNet-64	1	MS-Celeb-1M (3.5M,31K)	99.50

Triplet Loss

Schroff, Florian, Dmitry Kalenichenko, and James Philbin. "Facenet: A unified embedding for face recognition and clustering." CVPR 2015





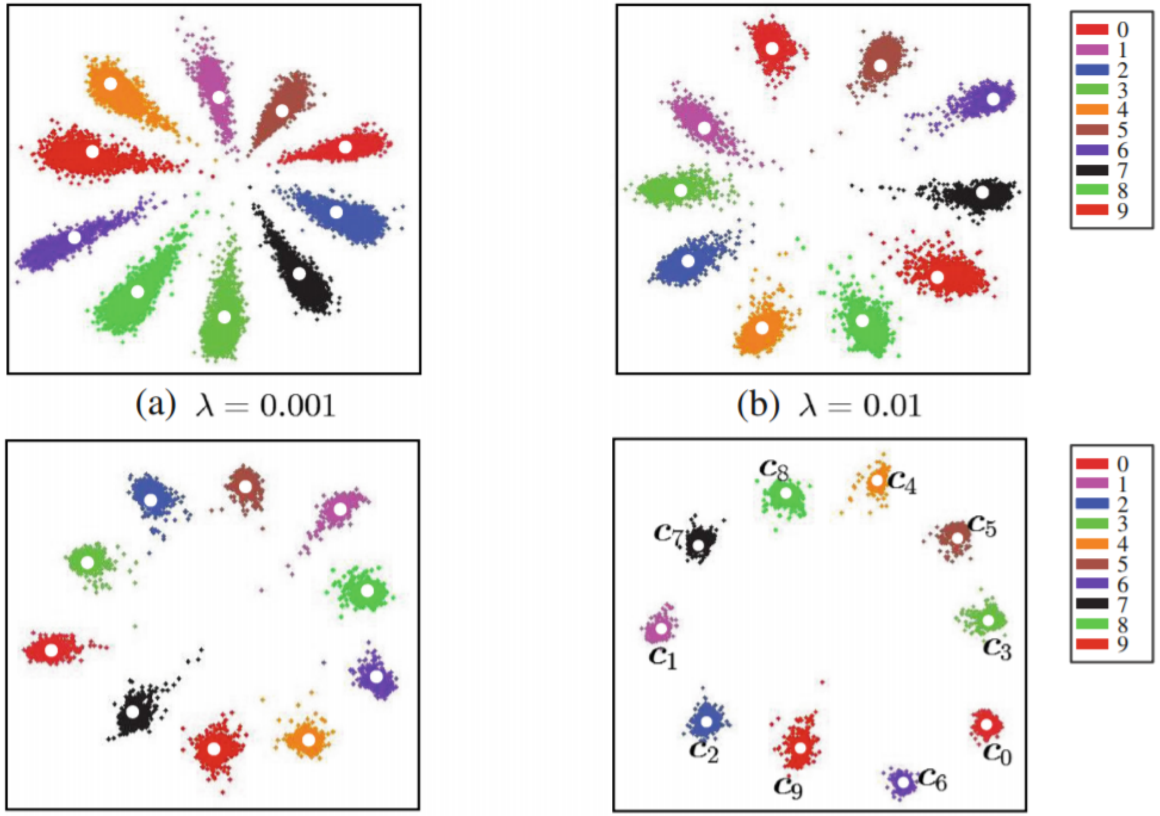
$$L = \sum_{i}^{N} \left[\|f(x_i^a) - f(x_i^p)\|_2^2 - \|f(x_i^a) - f(x_i^n)\|_2^2 + \alpha \right]_+$$

Center Loss

Wen, Yandong, Kaipeng Zhang, Zhifeng Li, and Yu Qiao. "A discriminative feature learning approach for deep face recognition." ECCV 2016.

$$\mathcal{L} = \mathcal{L}_S + \lambda \mathcal{L}_C$$

$$= -\sum_{i=1}^{m} \log \frac{e^{W_{y_i}^T \boldsymbol{x}_i + b_{y_i}}}{\sum_{j=1}^{n} e^{W_j^T \boldsymbol{x}_i + b_j}} + \frac{\lambda}{2} \sum_{i=1}^{m} \|\boldsymbol{x}_i - \boldsymbol{c}_{y_i}\|_2^2$$



(a) $\lambda = 0.001$ (b) $\lambda = 0.01$ (c) $\lambda = 0.1$ (d) $\lambda = 1$

Cosine Similarity Loss

Yandong Guo and Lei Zhang. "One-shot face recognition by promoting underrepresented classes." *arXiv preprint arXiv:1707.05574* (2017).

- Classification vector-centered Cosine Similarity (CCS)

$$\mathcal{L} = \mathcal{L}_s + \lambda \mathcal{L}_a$$

$$\mathcal{L}_s = -\sum_n \sum_k t_{k,n} \log p_k(x_n)$$

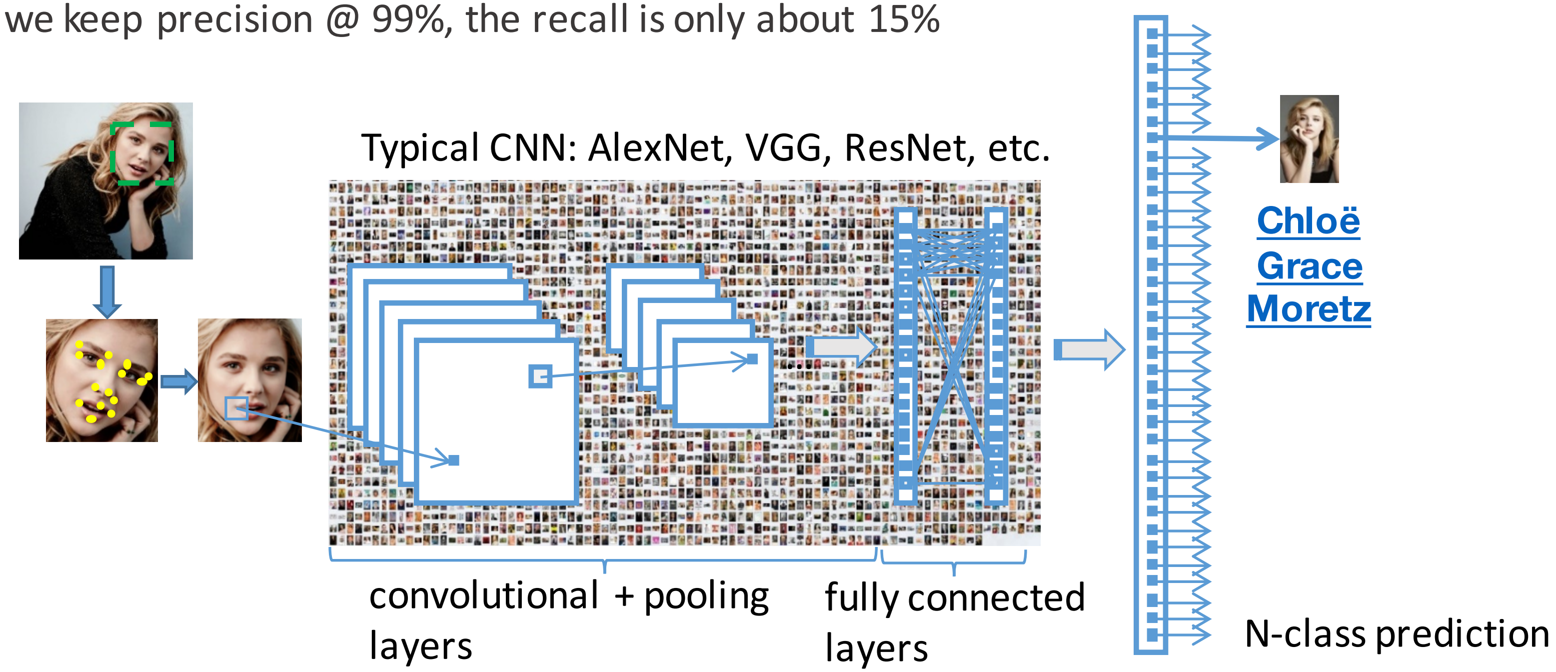
$$\mathbf{w}'_k \leftarrow \mathbf{w}_k$$

$$\mathcal{L}_a = -\sum_k \sum_{i \in C_k} \frac{\mathbf{w}_k'^T \phi(x_i)}{\|\mathbf{w}'\|_2 \|\phi(x_i)\|_2}$$

Methods	Dataset	Network	Accuracy
JB [2]	Public	–	96.33%
Human	–	–	97.53%
DeepFace[14]	Public	1	97.27%
DeepID2,3 [20, 22]	Public	200	99.53%
FaceNet [18]	Private	1	99.63%
Center face [24]	Private	1	99.28%
Center face [13]	Public	1	99.05%
Sphere face [13]	Public	1	99.42%
CCS face (ours)	Public	1	99.71%

Challenge Two: Classifier with Imbalanced Data

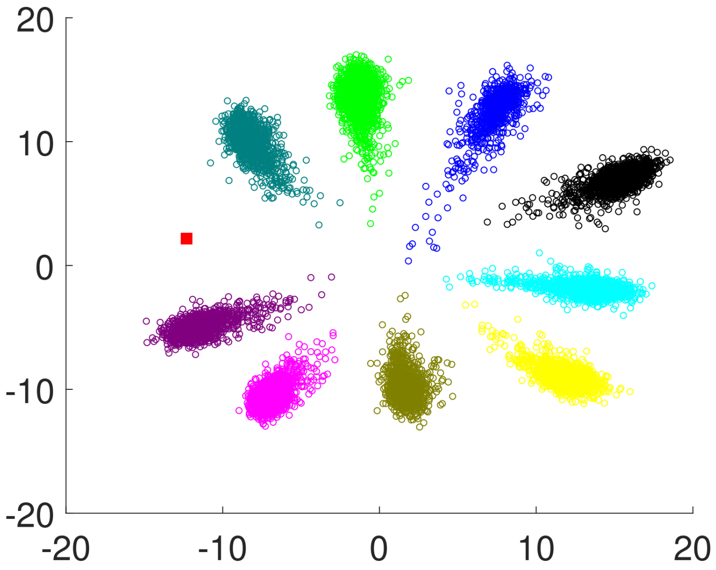
- Even with very good face representation model, classifier does not perform well
  - ResNet-34 trained on the base set
  - Final classifier trained on both the base set and the low-shot set
  - **99.8%** top-1 test accuracy on the base set
  - About **70%** top-1 test accuracy on the low-shot set, even when data boosting is applied
    - If we keep precision @ 99%, the recall is only about 15%

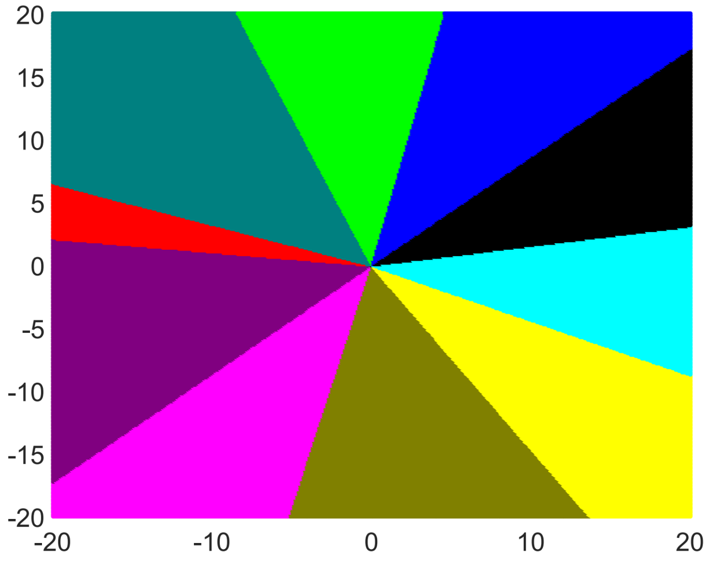


Why One-Shot Classes Perform So Bad?

- Logistic regression loss is additive

$$L = \Sigma_{i=1}^{N} cross_{\downarrow}entropy\left(p\left(\phi(x_i)\right), t_i\right)$$



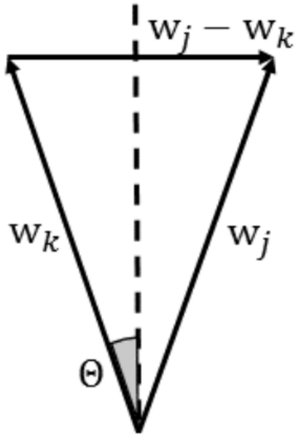


- You get what you provide

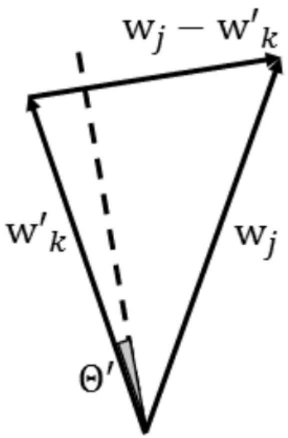
What Leads to Smaller Classification Space?

$$p_k(x_n) = \frac{\exp(\mathbf{w}_k^T \boldsymbol{\phi}(x_n))}{\sum_i \exp(\mathbf{w}_k^T \boldsymbol{\phi}(x_n))}$$

$$\frac{p_j(x)}{p_k(x)} = \frac{\exp(\mathbf{w}_j^T \boldsymbol{\phi}(x))}{\exp(\mathbf{w}_k^T \boldsymbol{\phi}(x))} = \exp[(\mathbf{w}_j - \mathbf{w}_k)^T \boldsymbol{\phi}(x)]$$



(a) $\|\mathbf{w}_k\|_2 = \|\mathbf{w}_j\|_2$

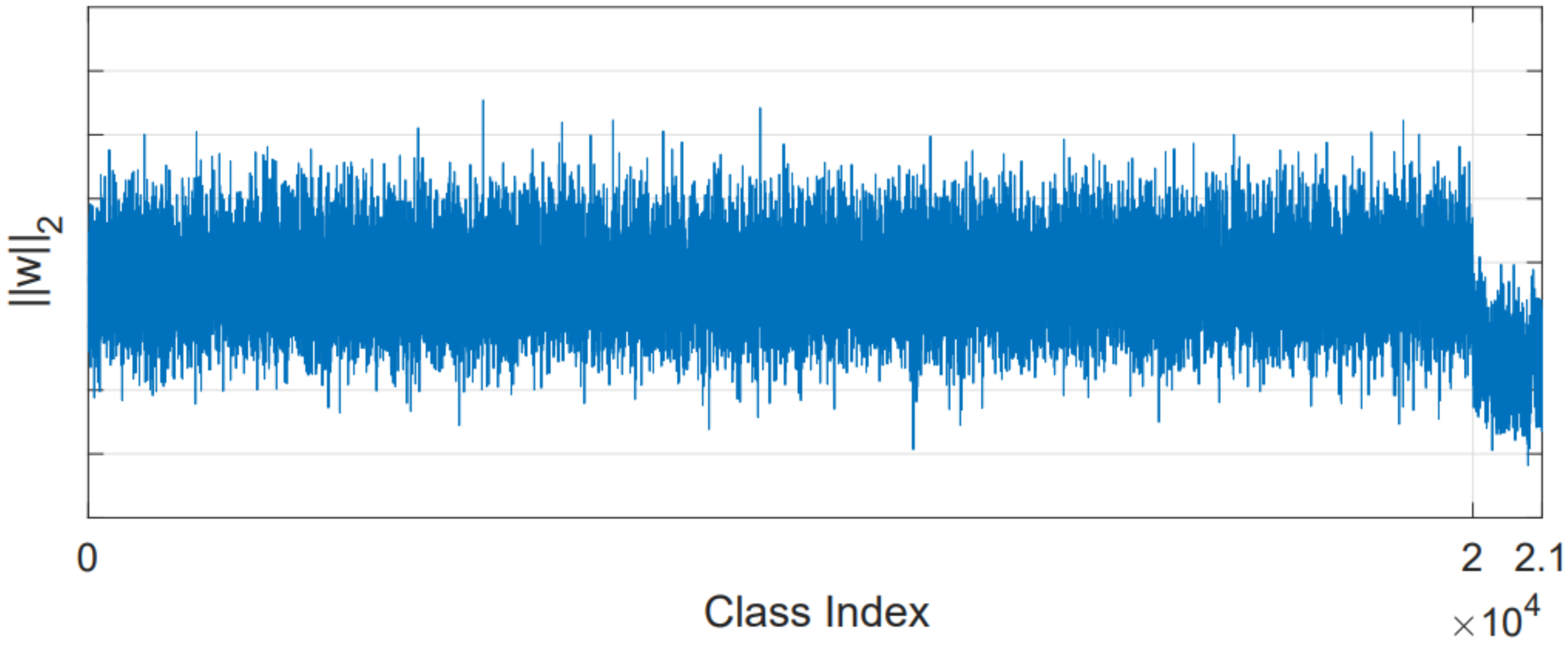


(b) $\|\mathbf{w}_k\|_2 < \|\mathbf{w}_j\|_2$

- Lack of samples introduces smaller classification space
- Accordingly, smaller classification space means smaller weighting vector norm for low-shot classes

* We removed the bias term to make the problem tractable.

Weight Vector Norm Distribution



*We remove the bias term

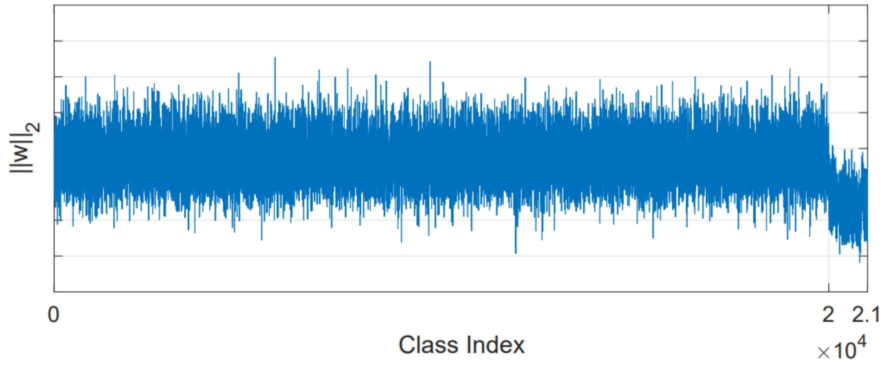
Underrepresented Classes Promotion (UP)

- Underrepresented Classes Promotion

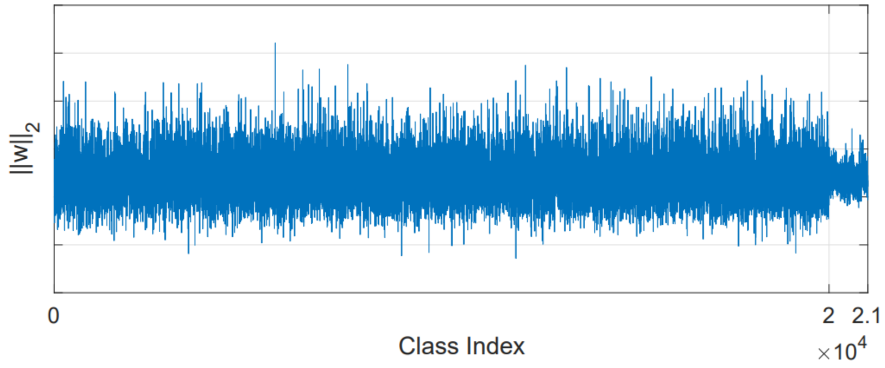
$$\mathcal{L}_{up} = \sum_{n} -t_{k,n} \log p_k(x_n) + \frac{1}{|C_n|} \sum_{k \in C_n} \| \|\mathbf{w}_k\|_2^2 - \alpha \|_2^2 ,$$

$$\alpha = \frac{1}{|C_b|} \sum_{k \in C_b} \|\mathbf{w}_k\|_2^2 .$$

Where C_b is the class set for the base classes, C_n is the class set for the low-shot classes



(a) Without UP Term



(b) With UP Term

Other Methods We Have Tried

- Shrink

$$\mathcal{L}_{l2} = \sum_{n} -t_{k,n} \log p_k(x_n) + \sum_{k} \|\mathbf{w}_k\|_2^2 \, .$$

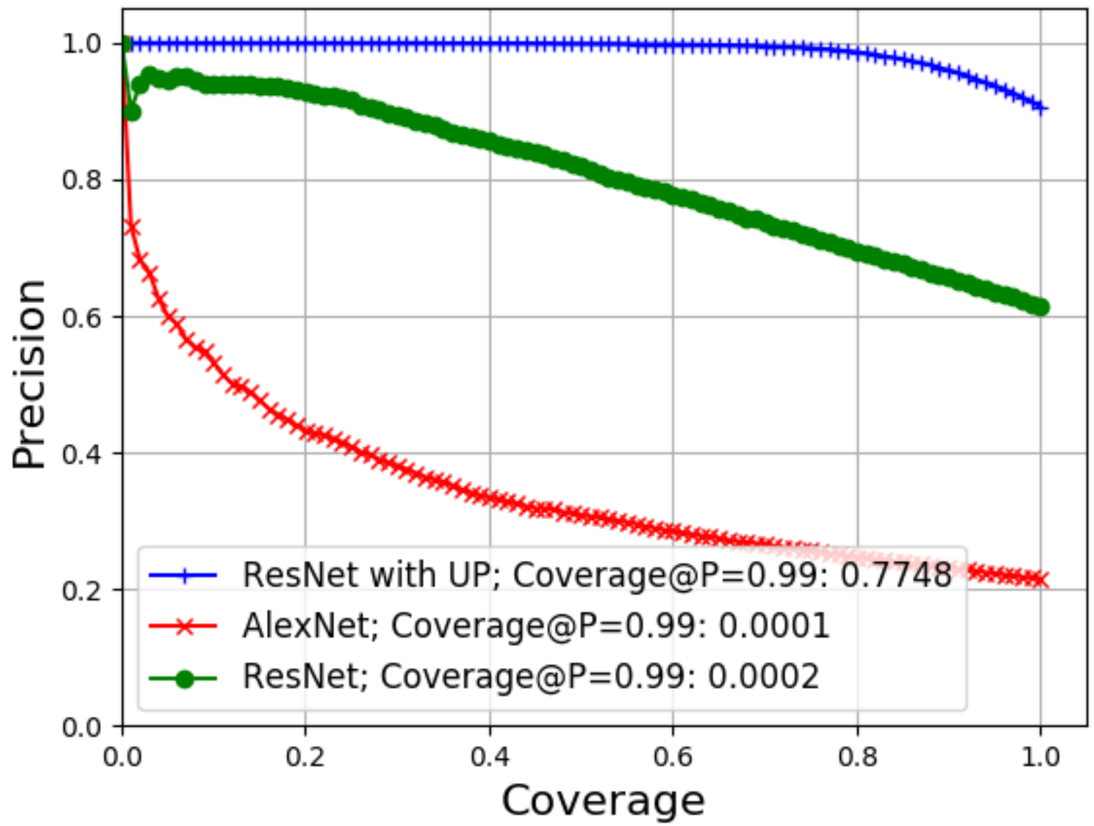
- Equal Norm

$$\mathcal{L}_{eq} = \sum_{n} -t_{k,n} \log p_k(x_n) + \sum_{k \in \{C_n \cup C_b\}} \| \|\mathbf{w}_k\|_2^2 - \beta\|_2^2 \, ,$$

$$\beta = \frac{1}{|\{C_n \cup C_b\}|} \sum_{k \in \{C_n \cup C_b\}} \|\mathbf{w}_k\|_2^2 .$$

Experimental Results on Our Benchmark Task

- Dataset Revisit
  - *Base set*: 20K celebrities, 50-100 images per celebrity
  - *Low-shot set*: 1K celebrities, **one image** per celebrity for training, 20 images per celebrity for testing
- Performance on **low-shot classes**



- Red-> Green: improvement by better CNN model (AlexNet -> ResNet-34)
- Green->Blue: improvement by the new loss term and data boosting

More Experimental Results

- Metric: Coverage at high precision, test on the low-shot classes, same data boosting applied (x100)

Method	C@99%	C@99.9%
Fixed Feature	25.65%	0.89%
SGM [8]	27.23%	4.24%
Update Feature	26.09%	0.97%
Direct Train	15.25%	0.84%
Shrink Norm (Eq.12)	32.58%	2.11%
Equal Norm (Eq.13)	32.56%	5.18%
UP Only (Eq.10)	77.48%	47.53%
CCS Only (Eq.4)	62.55%	11.13%
Our: CCS (4) plus UP (10)	**94.89%**	**83.60%**
Hybrid [28]	92.64%	N/A
Doppelganger [19]	73.86%	N/A
Generation-based [3]	61.21%	N/A

*"Low-shot Visual Object Recognition", Bharath Hariharan, Ross Girshick

Other Improvement – Generative Learning

- The UP prior acts as a regularizer and treats different classes indifferently
- How to take into account different intra person variance?
- Generate virtual samples to span the space for low shot classes
  - Key idea: *generate samples in feature space, rather than in image space*

Method	C@P=99%	C@P=99.9%
Fixed-Feature	25.65%	0.89%
SGM [8]	27.23%	4.24%
Update Feature	26.09%	0.97%
Direct Train	15.25%	0.84%
Shrink Norm[1]	32.58%	2.11%
Equal Norm[1]	32.56%	5.18%
Up Term [1]	77.48%	47.53%
Ours	$\boxed{94.84\%}$	83.82%

Zhengming Ding, Yandong Guo, Lei Zhang, Yun Fu. One-Shot Face Recognition via Generative Learning, *IEEE Conference on Automatic Face and Gesture Recognition* (FG), 2018

Summary

- Face recognition – great progress made in the past five years
  - Large-scale datasets developed and made publicly available
  - Better algorithms led to better face representation
- In real applications, many challenges still remain and desire for more studies
  - Large pose, large age variation, low resolution, etc.
  - Person re-identification in videos
  - Bias caused by improperly constructed datasets
  - Privacy concerns
  - …

Thanks!

leizhang@microsoft.com

MS-Celeb-1M (http://msceleb.org)

Backup Slides

Challenge Two: Classifier with Imbalanced Data

- Why a classifier is needed?
  - KNN has been widely adopted
  - If the feature extractor is PERFECT, KNN is the optimal solution, if not, **we need a classifier to describe the partition of the feature space**

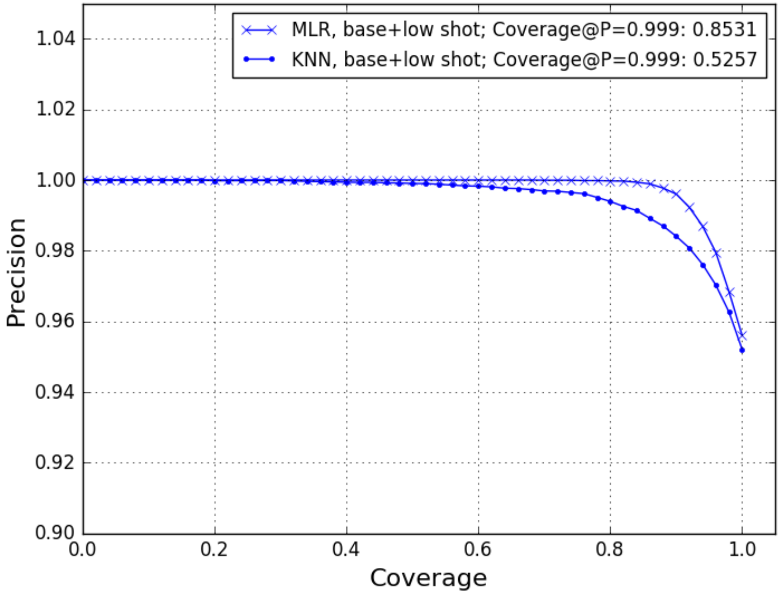
	K-Nearest Neighborhood (KNN)	Multinomial Logistic Regression (MLR)
Advantages	No additional training needed to add/remove persons	Better performance in the large-scale scenario if there are many images for each class[1,2] 1. Computing complexity is linear to the number of classes; 2. Weighting vectors in MLR is trained with global information;
Disadvantage	Not good for large scale 1. Not practical to keep all the face images for every person in the gallery; 2. If select a subset, what and how many images to select is still an open challenge; 3. The accuracy relies on the annotation accuracy;	Additional training needed*

- We train multinomial logistic regression as our classifier.

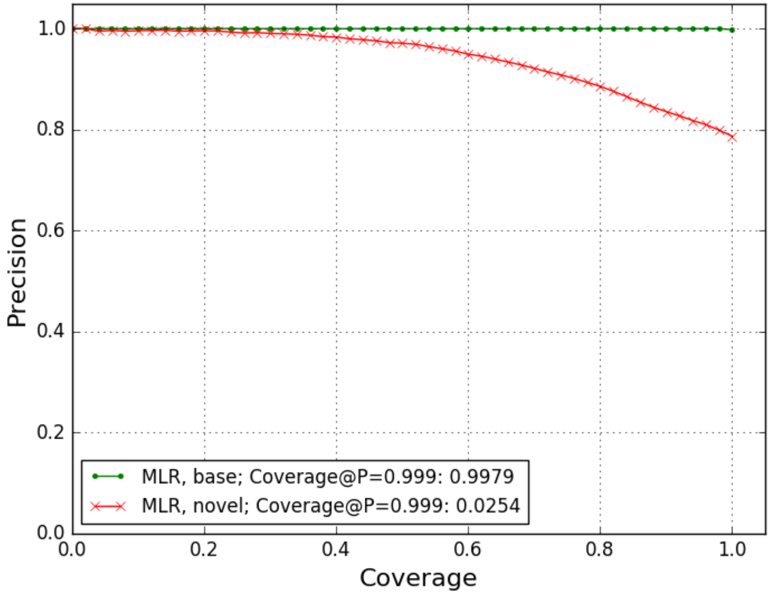
[1] Yue Wu, etc. “Low-shot Face Recognition with Hybrid Classifiers”.
[2] Yan Xu, etc. “High Performance Large Scale Face Recognition with Multi-Cognition Softmax and Feature Retrieval”.
[*] We patented technologies to train MLR very fast

Closer Look on KNN vs. MLR

- Both the methods were tested on the development set of low-shot learning track of MSCeleb-1M
- ResNet-34 trained with the all the training set of low-shot learning track of MSCeleb-1M (pool5 as feature)
- Results shown in Figure-a



a



b

- In Figure-a, we observe **much higher coverage** at high precision for MLR compared with KNN
- In Figure-b, we observe that with MLR, the performance on the low-shot classes is **much worse** than that of the base classes
- How to improve? Option A: Hybrid; Option B: Direct boosting