计算机视觉课程报告

学号：	10185102144
姓名：	董辰尧

专业名称：	计算机科学与技术
学生年级：	2018 级本科生
课程性质：	专业选修
研修时间：	2020～2021 学年第 2 学期

计算机科学与技术学院

2021 年 6 月

课程内容统计

- 请自评你的项目完成情况，在表中相应位置划√。

课程学习自我评价

内容\评价	阅读文献 0—5 篇	阅读文献 5—10 篇	阅读文献 10 篇以上	代码 实现
第 12 章 目标识别	√			

总体课程学习情况自我评价

完成情况	尚未 完成	基本 完成	较好 完成	圆满 完成
总体情况		√		

第 11 章 目标识别

一、这一章学习中你的工作

二、查阅文献清单

格式：

[1] Topic modeling: Beyond bag-of-words[C]// Machine Learning, Proceedings of the Twenty-Third International Conference (ICML 2006), Pittsburgh, Pennsylvania, USA, June 25-29, 2006. 2006.

[2] Wang M , Deng W . Deep Face Recognition: A Survey[J]. arXiv, 2018.

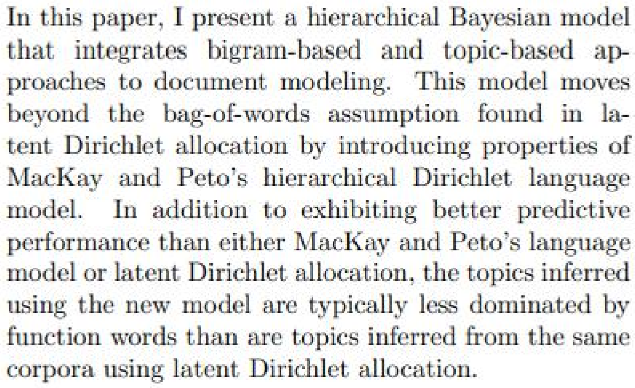
[3] Zhang K , Zhang Z , Li Z , et al. Joint Face Detection and Alignment Using Multitask Cascaded Convolutional Networks[J]. IEEE Signal Processing Letters, 2016, 23(10):1499-1503.

三、文献解读

1. 文献 1

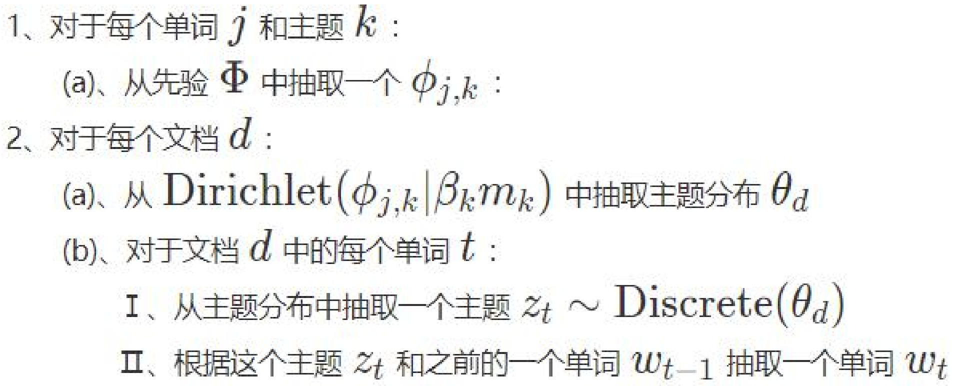
（a）文献名：Topic modeling: Beyond bag-of-words[C]// Machine Learning, Proceedings of the Twenty-Third International Conference (ICML 2006), Pittsburgh, Pennsylvania, USA, June 25-29, 2006. 2006.

（b）主要创新思想



In this paper, I present a hierarchical Bayesian model that integrates bigram-based and topic-based approaches to document modeling. This model moves beyond the bag-of-words assumption found in latent Dirichlet allocation by introducing properties of MacKay and Peto's hierarchical Dirichlet language model. In addition to exhibiting better predictive performance than either MacKay and Peto's language model or latent Dirichlet allocation, the topics inferred using the new model are typically less dominated by function words than are topics inferred from the same corpora using latent Dirichlet allocation.

(c) 主要原理剖析及说明



1、对于每个单词 j 和主题 k：

(a)、从先验 Φ 中抽取一个 $\phi_{j,k}$：

2、对于每个文档 d：

(a)、从 $\mathrm{Dirichlet}(\phi_{j,k}|\beta_k m_k)$ 中抽取主题分布 θ_d

(b)、对于文档 d 中的每个单词 t：

Ⅰ、从主题分布中抽取一个主题 $z_t \sim \mathrm{Discrete}(\theta_d)$

Ⅱ、根据这个主题 z_t 和之前的一个单词 w_{t-1} 抽取一个单词 w_t

(d) 主要实验结果（现有原文章中的）

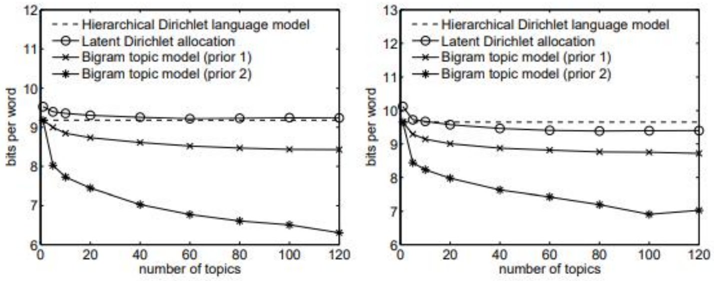


Figure 1. Information rates of the test data, measured in bits per word, under the different models versus number of topics. Left: Psychological Review Abstracts data. Right: 20 Newsgroups data.

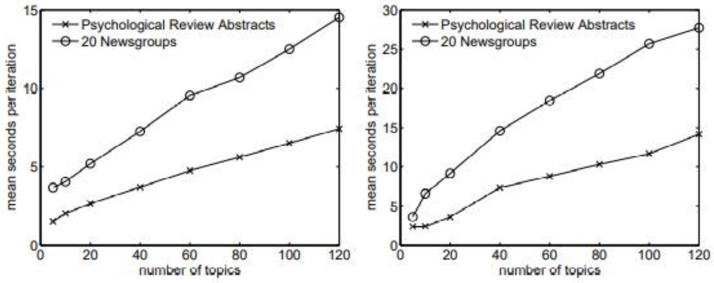
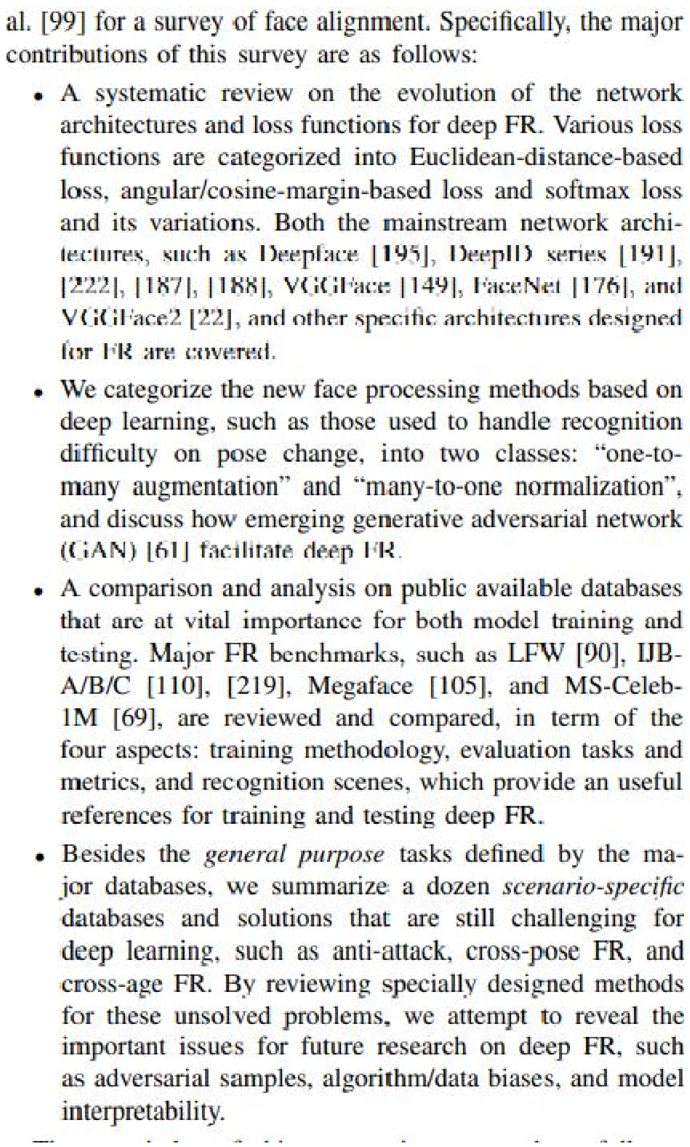


Figure 2. Mean time taken to perform a single iteration of the Gibbs EM algorithm described in section 4 as a function of the number of topics for both variants of the new model. Left: prior 1. Right: prior 2.

2. 文献 2

（a）文献名：Wang M， Deng W . Deep Face Recognition: A Survey[J]. arXiv, 2018.

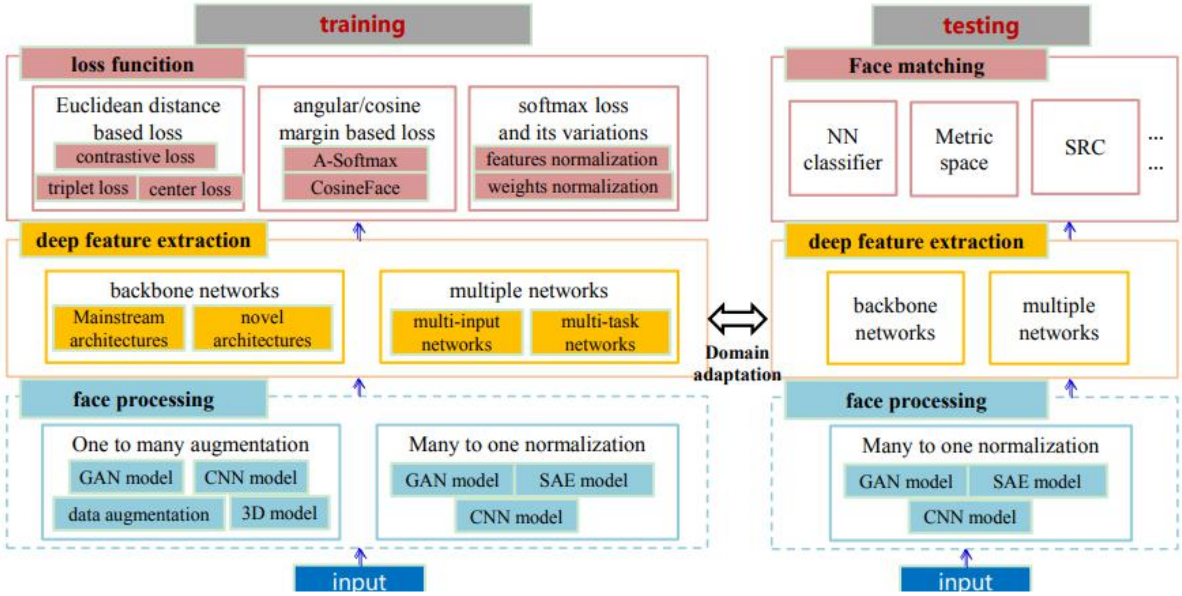
（b）主要创新思想



al. [99] for a survey of face alignment. Specifically, the major contributions of this survey are as follows:

- A systematic review on the evolution of the network architectures and loss functions for deep FR. Various loss functions are categorized into Euclidean-distance-based loss, angular/cosine-margin-based loss and softmax loss and its variations. Both the mainstream network architectures, such as Deepface [195], DeepID series [191], [222], [187], [188], VGGFace [149], FaceNet [176], and VGGFace2 [22], and other specific architectures designed for FR are covered.
- We categorize the new face processing methods based on deep learning, such as those used to handle recognition difficulty on pose change, into two classes: "one-to-many augmentation" and "many-to-one normalization", and discuss how emerging generative adversarial network (GAN) [61] facilitate deep FR.
- A comparison and analysis on public available databases that are at vital importance for both model training and testing. Major FR benchmarks, such as LFW [90], IJB-A/B/C [110], [219], Megaface [105], and MS-Celeb-1M [69], are reviewed and compared, in term of the four aspects: training methodology, evaluation tasks and metrics, and recognition scenes, which provide an useful references for training and testing deep FR.
- Besides the *general purpose* tasks defined by the major databases, we summarize a dozen *scenario-specific* databases and solutions that are still challenging for deep learning, such as anti-attack, cross-pose FR, and cross-age FR. By reviewing specially designed methods for these unsolved problems, we attempt to reveal the important issues for future research on deep FR, such as adversarial samples, algorithm/data biases, and model interpretability.

(c) 主要原理剖析及说明



1、人脸检测（Face Detection）

2、人脸对齐（Face Alignment）

3、人脸特征表征（Feature Representation）

目前进行人脸识别之前还会做活体检测

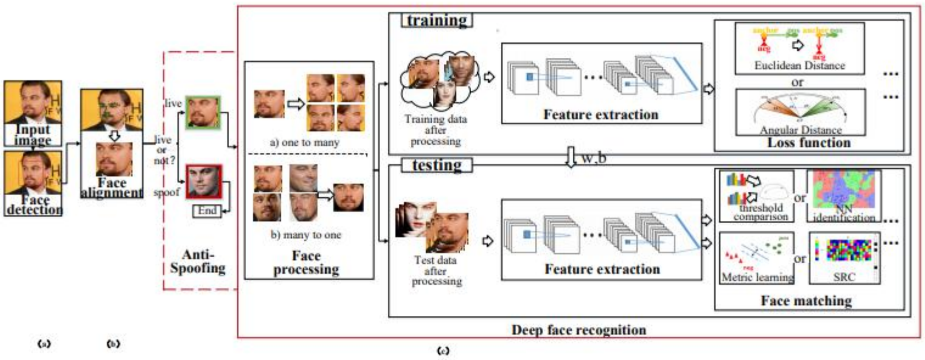


Fig. 3. Deep FR system with face detector and alignment. First, a face detector is used to localize faces. Second, the faces are aligned to normalized canonical coordinates. Third, the FR module is implemented. In FR module, face anti-spoofing recognizes whether the face is live or spoofed; face processing is used to handle recognition difficulty before training and testing; different architectures and loss functions are used to extract discriminative deep feature when training; face matching methods are used to do feature classification when the deep feature of testing data are extracted.

损失函数

1、基于欧氏距离

压缩类内方差，增大类间方差

e.g. contrastive loss triplet loss Center loss

2、基于角度/cosine

使特征中间有较大的角度或者cosine距离

e.g. L-Softmax A-Softmax

angular/cosinemargin-based loss, which used to achieve a better result on a

clean dataset, is vulnerable to noise and becomes worse than

Center loss and Softmax in the high-noise region. ????

3、softmax及其变种

Normalization of feature or weight

影响算法性能的因素：

1、训练集：一般训练集类别数越多，图像数量越多，训练效果越好。此外训练集的收集和标注质量，不同类别的样本数量是否均衡，都对训练有影响。

2、CNN：一般CNN的容量越大，训练效果越好。CNN的模型容量参考ImageNet上的分类性能，与参数数量和运行速度并不是正比关系。

3、LOSS：这部分才是前面介绍的loss相关影响，特别注意，对比某个loss的性能提升，要综合考虑训练集和CNN，不能简单的看LFW上的识别率。

(d) 主要实验结果（现有原文章中的）

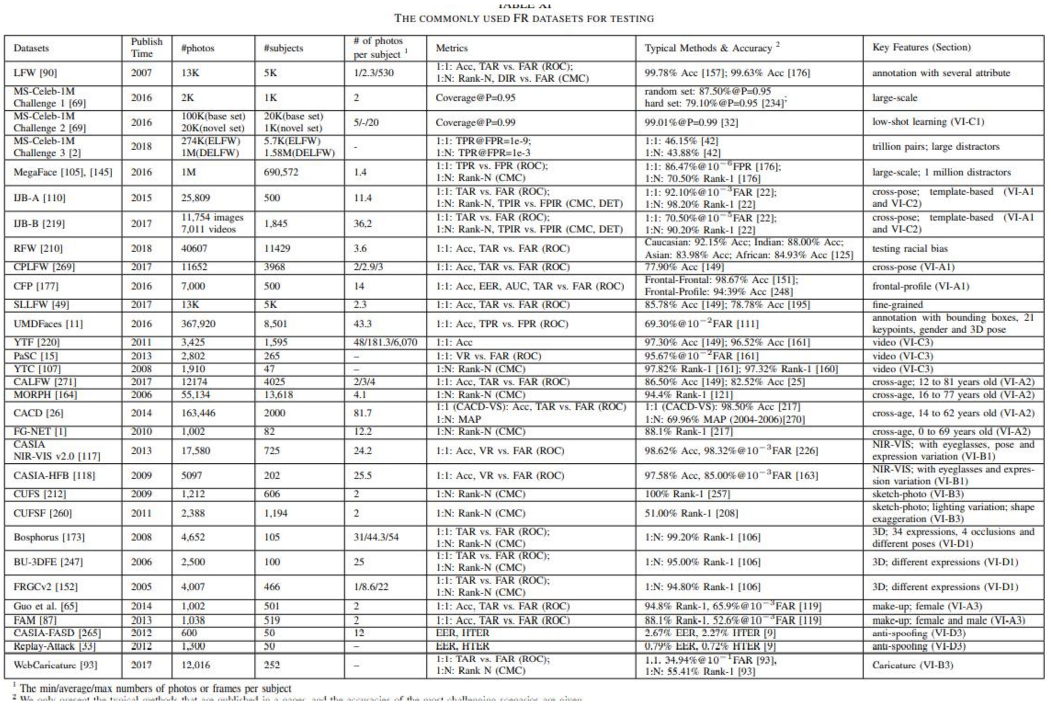


TABLE XI

THE COMMONLY USED FR DATASETS FOR TESTING

Datasets	Publish Time	#photos	#subjects	# of photos per subject [1]	Metrics	Typical Methods & Accuracy [2]	Key Features (Section)
LFW [90]	2007	13K	5K	1/2.3/530	1:1: Acc, TAR vs. FAR (ROC); 1:N: Rank-N, DIR vs. FAR (CMC)	99.78% Acc [157]; 99.63% Acc [176]	annotation with several attribute
MS-Celeb-1M Challenge 1 [69]	2016	2K	1K	2	Coverage@P=0.95	random set: 87.50%@P=0.95 hard set: 79.10%@P=0.95 [234];	large-scale

MS-Celeb-1M Challenge 2 [69]	2016	100K(base set)
		20K(novel set)
			20K(base set)
			1K(novel set)
				5/-/20	Coverage@P=0.99	99.01%@P=0.99 [32]	low-shot learning (VI-C1)

MS-Celeb-1M Challenge 3 [2]	2018	274K(ELFW)
		1M(DELFW)
			5.7K(ELFW)
			1.58M(DELFW)
				-	1:1: TPR@FPR=1e-9;
					1:N: TPR@FPR=1e-3
						1:1: 46.15% [42]
						1:N: 43.88% [42]
							trillion pairs; large distractors

MegaFace [105], [145]	2016	1M	690,572	1.4	1:1: TPR vs. FPR (ROC); 1:N: Rank-N (CMC)	1:1: 86.47%@10^{-6}FPR [176]; 1:N: 70.50% Rank-1 [176]	large-scale; 1 million distractors
IJB-A [110]	2015	25,809	500	11.4	1:1: TAR vs. FAR (ROC); 1:N: Rank-N, TPIR vs. FPIR (CMC, DET)	1:1: 92.10%@10^{-3}FAR [22]; 1:N: 98.20% Rank-1 [22]	cross-pose; template-based (VI-A1 and VI-C2)
IJB-B [219]	2017	11,754 images 7,011 videos	1,845	36,2	1:1: TAR vs. FAR (ROC); 1:N: Rank-N, TPIR vs. FPIR (CMC, DET)	1:1: 70.50%@10^{-5}FAR [22]; 1:N: 90.20% Rank-1 [22]	cross-pose; template-based (VI-A1 and VI-C2)
RFW [210]	2018	40607	11429	3.6	1:1: Acc, TAR vs. FAR (ROC)	Caucasian: 92.15% Acc; Indian: 88.00% Acc; Asian: 83.98% Acc; African: 84.93% Acc [125]	testing racial bias
CPLFW [269]	2017	11652	3968	2/2.9/3	1:1: Acc, TAR vs. FAR (ROC)	77.90% Acc [149]	cross-pose (VI-A1)
CFP [177]	2016	7,000	500	14	1:1: Acc, EER, AUC, TAR vs. FAR (ROC)	Frontal-Frontal: 98.67% Acc [151]; Frontal-Profile: 94:39% Acc [248]	frontal-profile (VI-A1)
SLLFW [49]	2017	13K	5K	2.3	1:1: Acc, TAR vs. FAR (ROC)	85.78% Acc [149]; 78.78% Acc [195]	fine-grained
UMDFaces [11]	2016	367,920	8,501	43.3	1:1: Acc, TPR vs. FPR (ROC)	69.30%@10^{-2}FAR [111]	annotation with bounding boxes, 21 keypoints, gender and 3D pose
YTF [220]	2011	3,425	1,595	48/181.3/6,070	1:1: Acc	97.30% Acc [149]; 96.52% Acc [161]	video (VI-C3)
PaSC [15]	2013	2,802	265	–	1:1: VR vs. FAR (ROC)	95.67%@10^{-2}FAR [161]	video (VI-C3)
YTC [107]	2008	1,910	47	–	1:N: Rank-N (CMC)	97.82% Rank-1 [161]; 97.32% Rank-1 [160]	video (VI-C3)
CALFW [271]	2017	12174	4025	2/3/4	1:1: Acc, TAR vs. FAR (ROC)	86.50% Acc [149]; 82.52% Acc [25]	cross-age; 12 to 81 years old (VI-A2)
MORPH [164]	2006	55,134	13,618	4.1	1:N: Rank-N (CMC)	94.4% Rank-1 [121]	cross-age, 16 to 77 years old (VI-A2)
CACD [26]	2014	163,446	2000	81.7	1:1 (CACD-VS): Acc, TAR vs. FAR (ROC) 1:N: MAP	1:1 (CACD-VS): 98.50% Acc [217] 1:N: 69.96% MAP (2004-2006)[270]	cross-age, 14 to 62 years old (VI-A2)
FG-NET [1]	2010	1,002	82	12.2	1:N: Rank-N (CMC)	88.1% Rank-1 [217]	cross-age, 0 to 69 years old (VI-A2)
CASIA NIR-VIS v2.0 [117]	2013	17,580	725	24.2	1:1: Acc, VR vs. FAR (ROC)	98.62% Acc, 98.32%@10^{-3}FAR [226]	NIR-VIS; with eyeglasses, pose and expression variation (VI-B1)
CASIA-HFB [118]	2009	5097	202	25.5	1:1: Acc, VR vs. FAR (ROC)	97.58% Acc, 85.00%@10^{-3}FAR [163]	NIR-VIS; with eyeglasses and expression variation (VI-B1)
CUFS [212]	2009	1,212	606	2	1:N: Rank-N (CMC)	100% Rank-1 [257]	sketch-photo (VI-B3)
CUFSF [260]	2011	2,388	1,194	2	1:N: Rank-N (CMC)	51.00% Rank-1 [208]	sketch-photo; lighting variation; shape exaggeration (VI-B3)
Bosphorus [173]	2008	4,652	105	31/44.3/54	1:1: TAR vs. FAR (ROC); 1:N: Rank-N (CMC)	1:N: 99.20% Rank-1 [106]	3D; 34 expressions, 4 occlusions and different poses (VI-D1)
BU-3DFE [247]	2006	2,500	100	25	1:1: TAR vs. FAR (ROC); 1:N: Rank-N (CMC)	1:N: 95.00% Rank-1 [106]	3D; different expressions (VI-D1)
FRGCv2 [152]	2005	4,007	466	1/8.6/22	1:1: TAR vs. FAR (ROC); 1:N: Rank-N (CMC)	1:N: 94.80% Rank-1 [106]	3D; different expressions (VI-D1)
Guo et al. [65]	2014	1,002	501	2	1:1: Acc, TAR vs. FAR (ROC)	94.8% Rank-1, 65.9%@10^{-3}FAR [119]	make-up; female (VI-A3)
FAM [87]	2013	1,038	519	2	1:1: Acc, TAR vs. FAR (ROC)	88.1% Rank-1, 52.6%@10^{-3}FAR [119]	make-up; female and male (VI-A3)
CASIA-FASD [265]	2012	600	50	12	EER, HTER	2.67% EER, 2.27% HTER [9]	anti-spoofing (VI-D3)
Replay-Attack [33]	2012	1,300	50	–	EER, HTER	0.79% EER, 0.72% HTER [9]	anti-spoofing (VI-D3)
WebCaricature [93]	2017	12,016	252	–	1:1: TAR vs. FAR (ROC); 1:N: Rank N (CMC)	1:1. 34.94%@10^{-1}FAR [93], 1:N: 55.41% Rank-1 [93]	Caricature (VI-B3)

[1] The min/average/max numbers of photos or frames per subject

[2] We only present the typical methods that are published in a paper, and the accuracies of the most challenging scenarios are given.

3. 文献 3

（a）文献名：Zhang K ， Zhang Z ， Li Z , et al. Joint Face Detection and Alignment Using Multitask Cascaded Convolutional Networks[J]. IEEE Signal Processing Letters, 2016, 23(10):1499-1503.

（b）主要创新思想

本文的贡献：

1） 提出一个新的基于 CNN 的级联型框架，用于联和（joint）人脸检测和对齐；还设计轻量级的 CNN 架构使得速度上可以达到实时。

2） 提出一个有效的 online hard sample mining 方法来提高表现能力

3） 在人脸检测和人脸对齐上提高了不少精度

(c) 主要原理剖析及说明

整体框架如下：

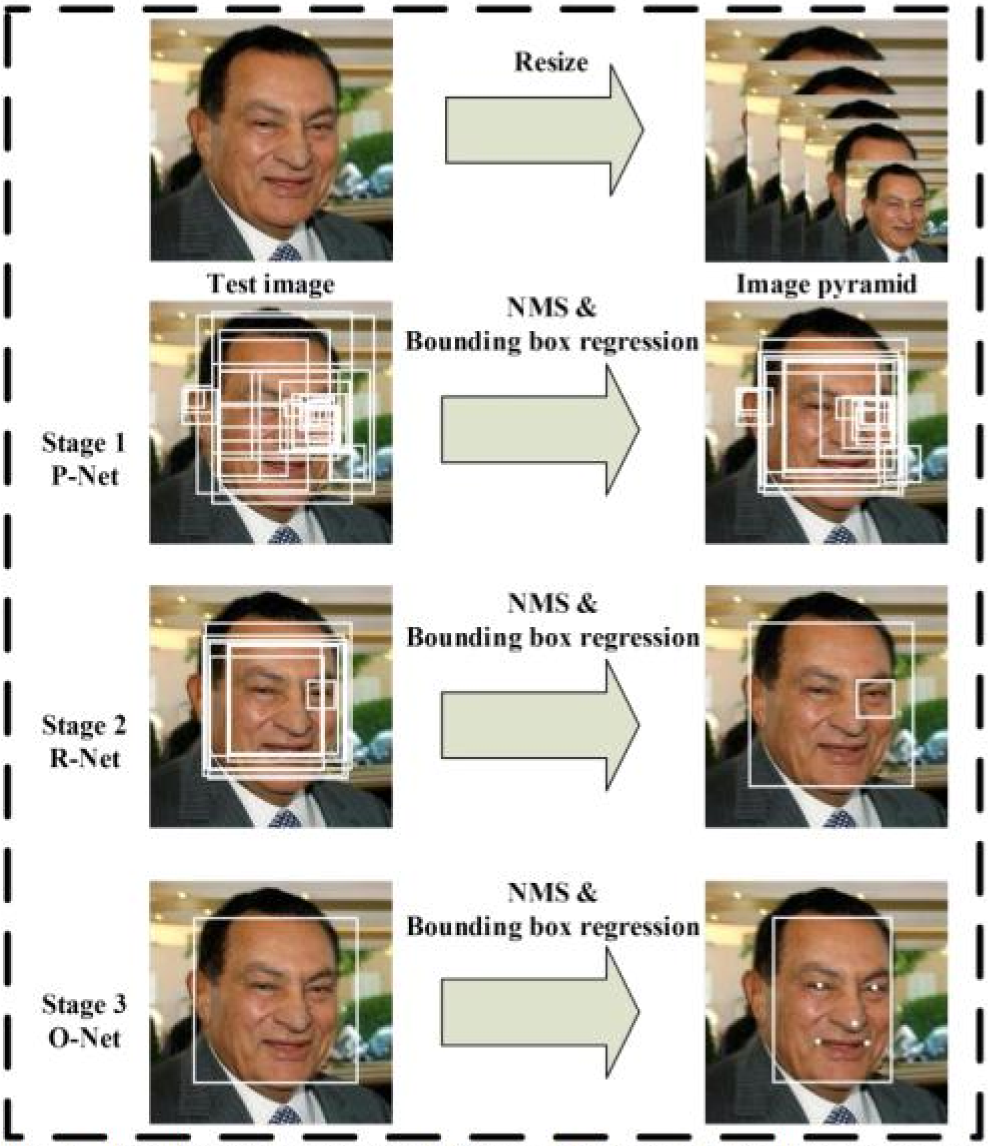


Fig. 1. Pipeline of our cascaded framework that includes three-stage multi-task deep convolutional networks. Firstly, candidate windows are produced through a fast Proposal Network (P-Net). After that, we refine these candidates in the next stage through a Refinement Network (R-Net). In the third stage, The Output Network (O-Net) produces final bounding box and facial landmarks position.

详细流程如下：

给定一张图片，首先将该图片重新调整到不同尺度大小，得到一个图像金字塔，该图像金字塔就是后面三阶段级联结构的输入。

阶段 1：利用一个全卷积网络，称为 Proposal Network (P-Net)，来获得候选窗口和它们的 bounding box regression vectors，然后利用 bounding box regression vectors 来调整候选框。之后，利用非极大值抑制（non-maximum suppression，NMS）来合并那些高度重合的候选框。

阶段 2：第 1 阶段产生的所有候选框作为另一个 CNN 的输出，该 CNN 称为 Refine

Network (R-Net)。该阶段的作用是利用 bounding box regression 和 NMS 进一步排除掉大量错误的候选框。

阶段 3：与第 2 阶段类似，但这个阶段会输出 5 个人脸关键点位置。

(d) 主要实验结果（现有原文章中的）

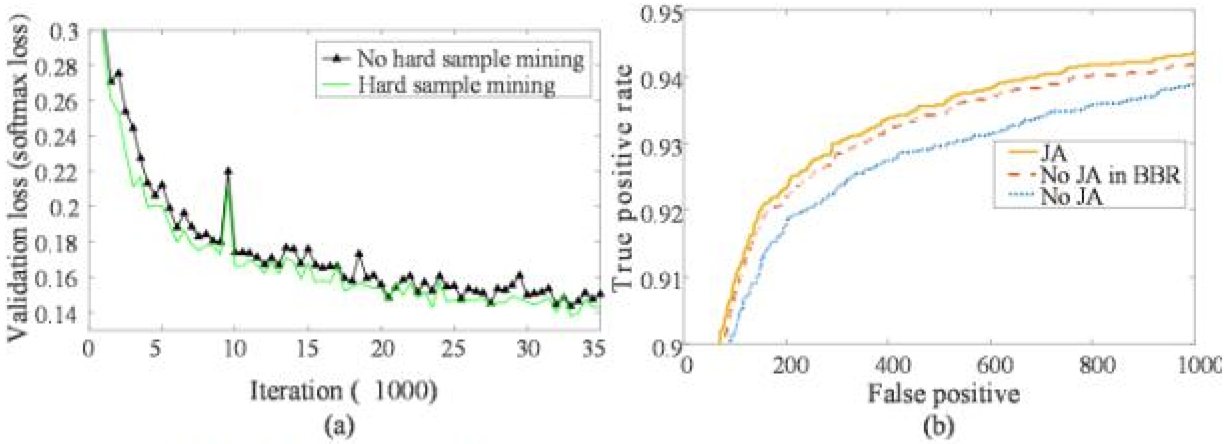


Fig. 3. (a) Validation loss of O-Net with and without hard sample mining. (b) "JA" denotes joint face alignment learning while "No JA" denotes do not joint it. "No JA in BBR" denotes do not joint it while training the CNN for bounding box regression.

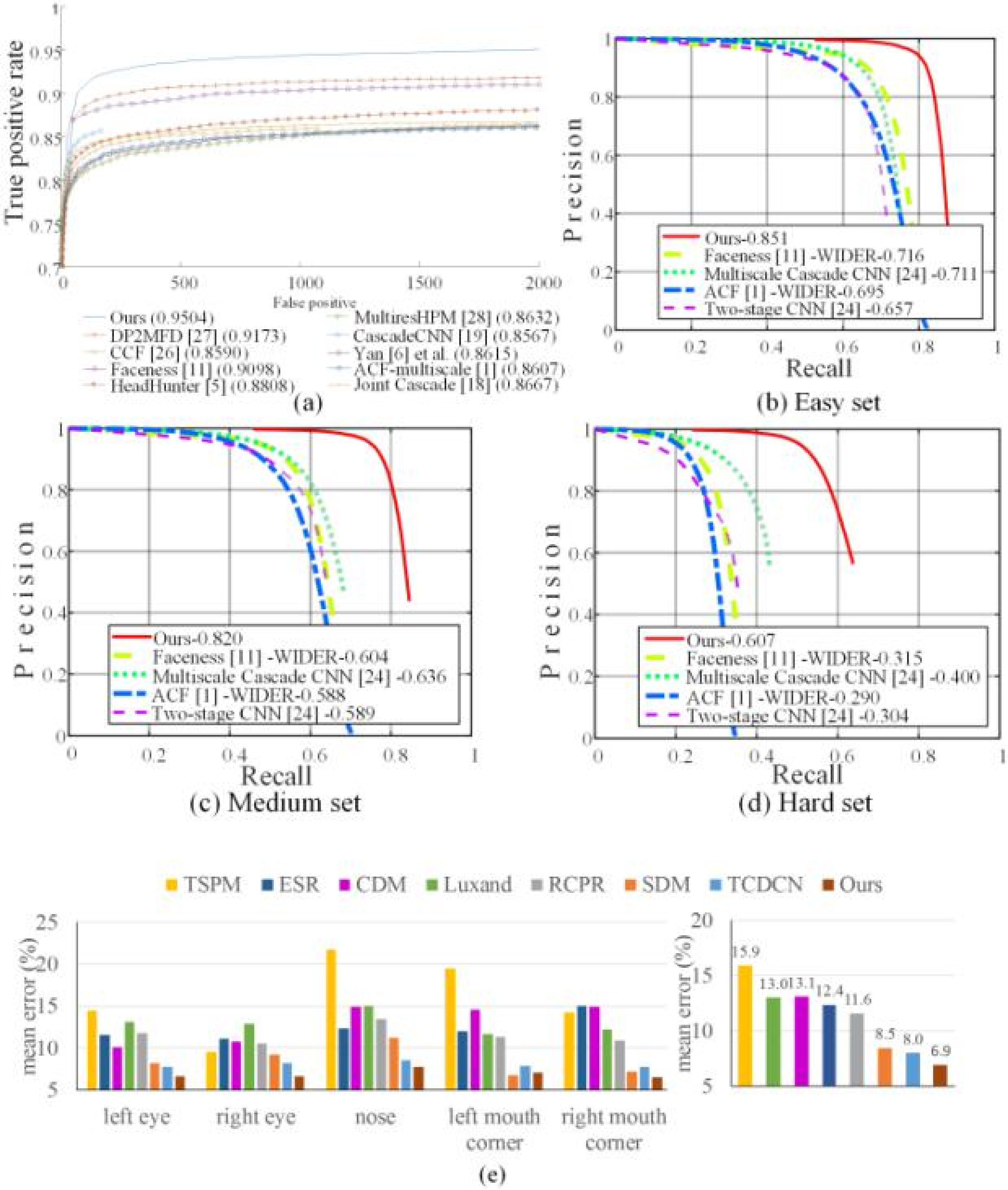


Fig. 4. (a) Evaluation on FDDB. (b-d) Evaluation on three subsets of WIDER FACE. The number following the method indicates the average accuracy. (e) Evaluation on AFLW for face alignment

三、本章学习小结

目标识别是指用计算机实现人的视觉功能,它的研究目标就是使计算机具有从一幅或多幅图像或者是视频中认知周围环境的能力(包括对客观世界三维环境的感知、识别与理解）。目标识别作为视觉技术的一个分支，就是对视场内的物体进行识别，如人或交通工具，先进行检测，检测完后进行识别，然后分析他们的行为。

通过本次学习，学到了许多和目标识别相关的知识，收获很多。