# 联合人脸检测和对准使用多任务级联卷积神经网络（mtcnn）

Kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li, Senior Member, IEEE, and Yu Qiao, Senior Member, IEEE

**摘要**：在不受约束的环境中人脸检测和对准由于各种姿势，照明和遮挡具有挑战性。近期研究表明深度学习的方法在这两个方面达到很好的效果。在这个篇文章中，我们提出一个深度级联多任务框架，利用固有的相关性检测和校准提高他们的性能。特别是我们的框架利用一个级联结构的三个阶段，详细设计深度卷积网络由粗到细的方式预测人脸和定位地标位置。此外，我们建议采用新的在线易分错样本挖掘策略，进一步改进在实践中的表现。我们的方法在挑战FDDB、WIDER FACE基准面检测和AFLW基准面对准上达到很高的准确性，同时保持实时的性能。

**关键字**：人脸检测，人脸对准，级联卷积神经网络

## 一、介绍

人脸检测和对准是大多数人脸应用所必要的，比如人脸识别和面部表情分析。然而，当人脸在视觉方面比较大的变化，例如遮挡，姿势变化和极端光照等，在现实世界应用中这些都属于巨大的挑战。

Viola and Jones[2]在级联的人脸检测器中提出利用Haar-Like特性和AdaBoost训练级联分类器，实现了良好的实时性。然而，相当多的作品[1,3,4]表明这种检测器的准确率可能在现实世界中显著降低，特别是应用具有较大的人脸视觉差异，或者具有更多的特征和分类时。除了级联结构之外，[5,6,7]引入了用于人脸检测的可变形部件模型（DPM），并且获得了显着的性能，然而他们计算代价太大并且在训练阶段需使用大量的注释。最近，卷积神经网络（CNN）在各种计算机视觉任务中取得了显着的成绩如图像分类[9]和面部识别[10]。Yang et al.[11]训练深卷积神经网络，研究利用深度CNNs进行人脸检测，启发了计算机视觉中深度学习方法的重大成功。

用于面部特征识别以获得面部的高特性，进一步产生面部候选窗口的区域。然而，由于其复杂的CNN结构，这种方法在实践中比较花费时间。Li et al.[19]使用级联CNNs进行人脸检测，但它需要边框校准，人脸检测具有额外的计算成本并且需忽略面部地标定位之间的固有相关性和边框回归。

人脸对准吸引广泛的研究兴趣，这一领域的研究可以大致分为两类，基于回归的方法[12,13,16]和模板拟合方法[14,15,7]。 最近，Zhang et al.[22]提出使用面部特征识别作为辅助方法，使用深卷积神经网络增强人脸对准性能。

然而，以前大多数的人脸检测和人脸对准方法忽略了这两者之间的固有相关性。 尽管有几个现有的工作试图联合解决他们，但这些作品仍有局限性。例如Chen et al. [18]采用像素值差异的特征联合进行校准和检测随机森林。但这些手工特征限制了它很多性能。 Zhang et al.[20]使用多任务CNN提高多视点人脸检测准确度，但检测窗口由弱人脸检测器产生，检测回归由初始化限制。

另一方面，在训练中挖掘易分错样本对于加强检测器的功效是至关重要的。但是，传统易分错样本挖掘通常以线下方式操作，显着增加了手动操作。理想的是设计了一种人脸检测的在线易分错样本挖掘方法，使其自动适应当前训练状态。

在本文中，我们提出了一个新的框架，使用统一的级联CNNs通过多任务学习整合这两个任务。提出的CNN包括三个阶段。在第一阶段，它通过浅CNN快速产生候选窗口。然后，它通过更复杂的CNN拒绝大量的非人脸窗口来优化窗口。最后，它使用更强大的CNN来细化结果并输出五个面部标记位置。感谢这个多任务学习框架，可以显着提高算法的性能。代码已在项目中发布。

本文的主要贡献总结如下：

（1）我们提出了一种新的级联基于CNNs的框架，用于联合人脸检测和对准，并仔细设计轻量级CNN架构以实现实时性能。

（2）我们提出了一种有效的在线实施采集样本的方法提高性能。

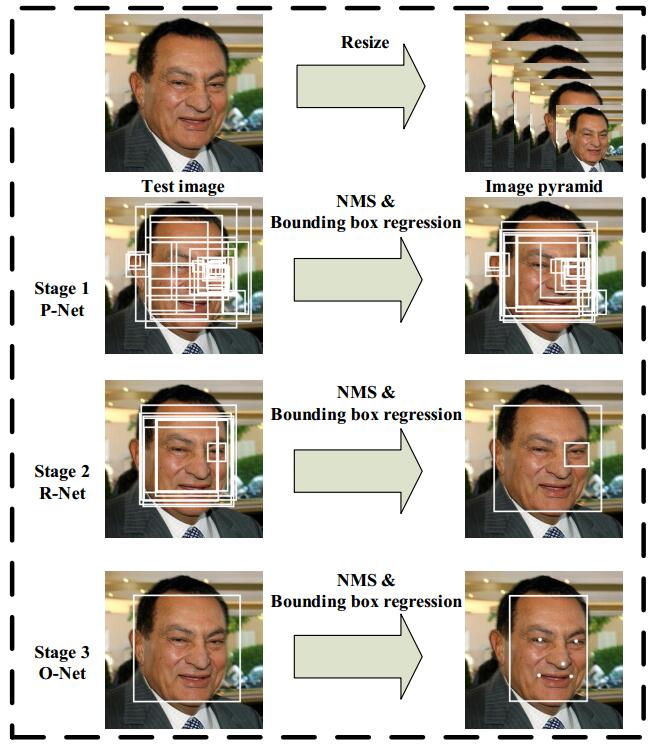
（3）对挑战基准进行广泛的实验，与人脸检测和人脸对准应用中的现有技术相比，以表现出提出方法的显着性能改进。

图1.我们的级联框架的管道包括三级多任务深卷积网络。首先，产生候选窗口通过快速建议网络（P-Net）。之后，我们完善这些候选人在下一阶段通过改进网络（R-Net）。在第三阶段，输出网络（O-Net）产生最终边框和面部地标位置。

## 二、方法

在本节中,我们将描述联合人脸检测和对准的方法。

### 1、总体框架

方法的整体如图1所示。给定一个图像，我们最初调整不同尺度来构建 一个图像金字塔,下面的输入 三级级联框架：

第一阶段：我们利用全卷积网络，称为投票网络（P-Net），以获得候选面部窗口及其边框回归向量。然后基于估计的边框回归向量校准候选。之后，我们采用非最大抑制（NMS）来合并高度重叠的候选。

第二阶段：所有候选窗口被馈送到另一CNN，称为精细网络（R-Net），其进一步拒绝大量假候选，使用边框回归执行校准，并且执行NMS。

第三阶段：这个阶段类似于第二阶段，但在这个阶段，我们的目标是确定更多监督区域。特别地，网络将输出五个面部地标位置。

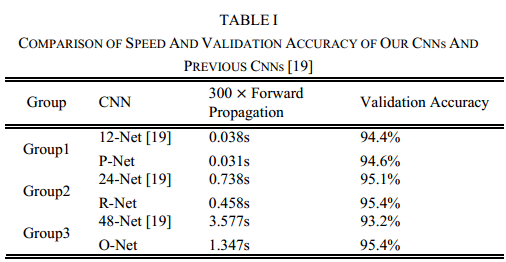


表1.

### 2、CNN 架构

在[19]中，多个CNN被设计用于人脸检测。然而，我们注意到它的性能可能受到以下事实的限制：

（1）卷积层中的一些滤波器缺乏限制其多样性的辨别能力。

（2）与其他多类异常检测和分类任务相比，人脸检测是一种具有挑战性的二进制分类任务，因此每层可能需要较少数量的过滤器。

为此，我们减少了滤波器的数量，并将5×5滤波器改为3×3滤波器以减少计算，同时增加深度以获得更好的性能。有了这些改进，与以前的架构[19]相比，我们可以获得更好的性能与更少的运行时间（训练阶段的结果显示在表I.为了公平比较，我们在每组中使用相同的训练和验证数据）。我们的CNN架构如图1所示。在卷积和全连接层之后，我们应用PReLU [30]作为非线性激活函数（输出层除外）。

### 3、训练

我们利用三个任务来训练我们的CNN检测器：人脸/非人脸分类，边框回归和面部地标定位。

人脸分类：学习目的是制定一个二分类器，对于每一个样本我们使用交叉熵损失：

(1)

是由网络结构表明样本是人脸的概率。表示真实的标签。

边框回归：对于每个候选窗口，我们预测它和最近的地标之间的偏移（即，边框的左边，顶部，高度和宽度）。学习目标是制定一个回归问题,我们使用每个样本的欧几里得损失:

(2)

是从网络结构预测的坐标；是真实的坐标；这儿有四个坐标,包括左前、高度和宽度，并且.

人脸特征点定位：与边框回归任务类似，面部标记检测被公式化为回归问题，并且我们最小化欧几里得损失：

(3)

是从网络结构预测的面部地标，是第i个样本的面部地标。有五个面部地标，包括左眼、右眼、鼻子、左嘴角和右嘴角，并且.

多源训练：由于我们在每个CNN中采用不同的任务，所以在学习过程中存在不同类型的训练图像，例如人脸，非人脸和部分对准的人脸。在这种情况下，不使用一些损耗函数（即，等式（1） - （3））。例如，对于样本背景区域我们只计算,其他的两个损失设置为0。这可以直接用样本类型指示符来实现。然后可以将整体学习目标表示为：

(4)

N是训练样本总数，表示为任务重要性。我们在P-Net和R-Net使用()，O-Net 使用()更准确定位面部地标。是样本分类标签。在这种情况下，自然地采用随机梯度下降来训练这些CNNs。

在线易分错样本挖掘：不同于对原始分类器进行训练后的传统易分错样本挖掘，我们在人脸/非人脸分类任务中进行在线易分错样本挖掘，这适合于训练过程。

特别地，在每个小批量中，我们对从所有样本在正向传播中计算的损耗进行排序，并将其中前70％选择为易分错样本。然后，我们只计算在反向传播中的这些分错样本的梯度。这意味着我们忽略在训练期间对加强检测器无用的简单样本。实验表明，这种策略产生更好的性能，无需手动选择样本。其有效性见第三节。

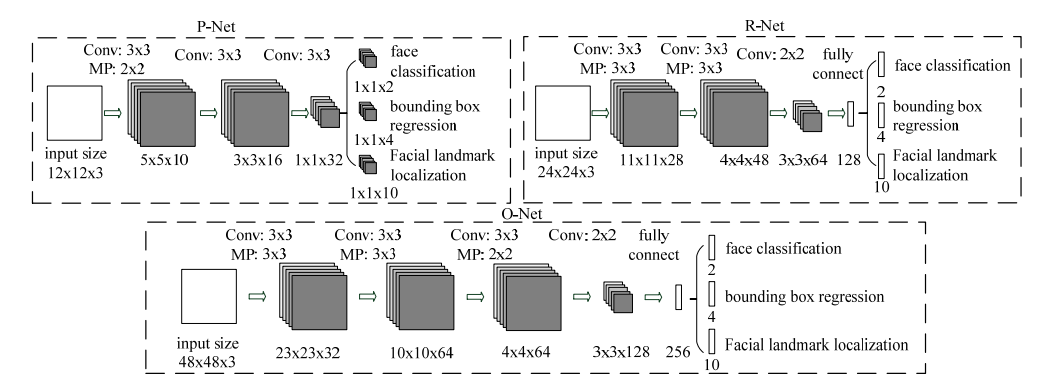


图2.P-Net，R-Net和O-Net的架构，其中“MP”表示最大池，“Conv”表示卷积。 卷积和合并中的步长分别为1和2。

## 三、实验

在本节中，我们首先评估提出的易分错样本挖掘策略的有效性。然后我们将人脸检测器和对准与Face Detection Data Set and Benchmark(FDDB)[25]，WIDER FACE [24]和Annotated Facial Landmarks（AFLW）中的最先进的方法进行比较[ 8]。 FDDB数据集包含5,171个人脸的注解，包含在2,845个图像的集合中。WIDER FACE数据集包括32,203个图像中的393,703个标记的面部边框，其中50％用于测试（根据图像的难度划分为三个子集），40％用于训练，剩余的用于验证.AFLW包含面部陆标注释 24,386个人脸，我们使用与[22]相同的测试子集。最后，我们评估我们的人脸检测器的计算效率。

### 1、训练数据

由于我们联合执行人脸检测和对准，这里在我们的训练过程中使用四种不同类型的数据注释：

Negativies：重叠度（IoU）比率小于0.3的任何真实人脸；

Positives：IoU大于0.65的真实人脸；

Part faces：IoU在0.4到0.65之间的真实人脸；

Landmark faces：标记5个地标位置的人脸；

Part faces和Negativies之间存在不清楚的差异，并且不同人脸注解之间存在差异。因此，我们选择0.3到0.4之间的IoU区域。 Positives和Negativies用于人脸分类任务，Positives和Part faces用于边框回归，Landmark faces用于人脸地标定位。总训练数据由3：1：1：2（Negativies /Positives/ Part faces/ Landmark faces）数据组成。每个网络的训练数据收集描述如下：

1）P-Net：我们从WIDER FACE [24]中随机抽取几个数据包以收集Positives，Negativies和Part faces。然后，我们从CelebA [23]裁剪面孔作为地标面孔。

2）R-Net：我们框架的第一阶段来检测来自WIDER FACE [24]的面部以Positives，Negativies和Part faces，而从CelebA检测到界标面部[23]。

3）O-Net：类似于R-Net收集数据，使用我们的框架的前两个阶段来检测人脸和收集数据。

### 2、在线易分错样本挖掘的有效性

为了评估所提出的在线易分错样本挖掘策略的贡献，我们训练两个P-Nets（有和没有在线易错性样本挖掘），并比较它们在FDDB上的性能。图3.（a）表现出了来自FDDB上的两个不同P-Nets的结果。很明显，在线易错性样本挖掘是有利于提高性能。它可以使FDDB的整体性能提高1.5％。

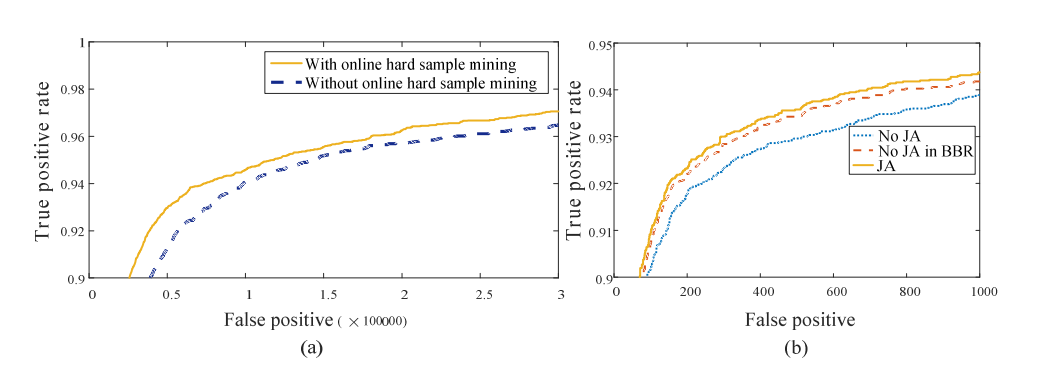


图3.（a）具有和不具有在线易分错样本挖掘的P-Net的检测性能。（b）“JA”表示O-Net中的人脸对准学习，而“No JA”表示不联合。“在BBR中没有JA”表示使用“No JA”O-Net进行边框回归。

### 3、联合检测和校准的有效性

为了评估联合检测和对准，我们在FDDB（具有相同的P-Net和R-Net）上评估两个不同的O-Nets（联合与不联合面部界标回归学习）的性能。我们比较了这两个O-Nets的边框回归的性能。图3.（b）表明联合地标本地化任务学习有助于增强面部分类和边框回归任务。

### 4、评估人脸检测

为了评估我们的人脸检测方法的性能，在FDDB比较我们的方法与现有技术的方法[1,5,6,11,18,19,26,27,28,29]，在WIDER FACE和最先进的方法[1，24，11]比较。图4.（a）- (d）表明，我们的方法在两个基准中都优于所有其他的方法。

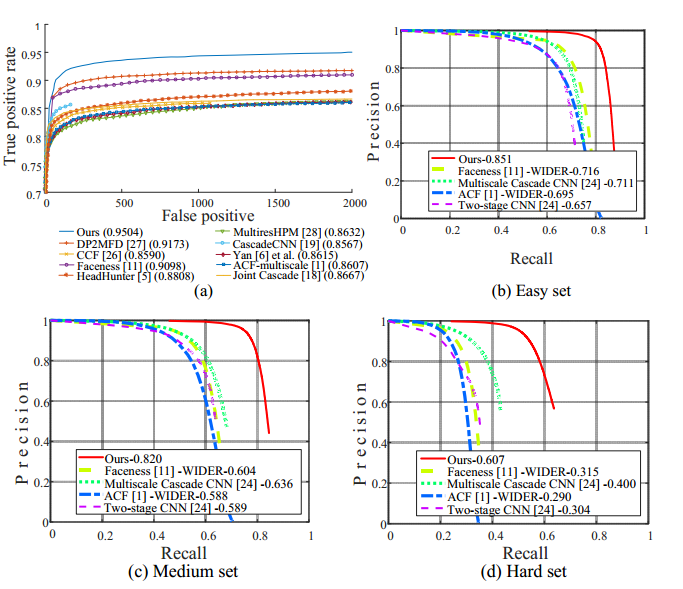


图4.（a）对FDDB的评价。（b-d）对WIDER FACE三个子集进行评估。该方法后的数字表示平均精度。

### 5、评估人脸对准

在这部分，我们比较以下方法的人脸对准性能：RCPR[12]，TSPM [7]，Luxand face SDK [17]，ESR [13]，CDM [15]，SDM [21] TCDCN [22]。平均误差由估计的地标和真实之间的距离测量，相对于视觉距离它是被标准化。图.5表明我们的方法优于所有最先进的方法。这也表明我们的方法在嘴角定位中显示较少的优势。这可能是由于在我们的训练数据中表达的小变化，其对嘴角位置具有显着影响。

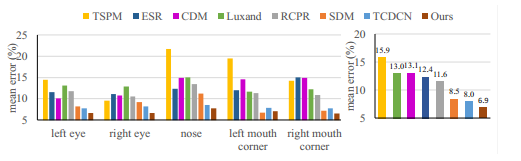
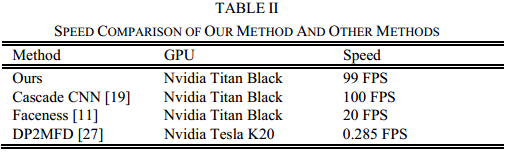


图5.评估AFLW脸对准

### 6、运行效率

给定级联结构，我们的方法可以实现高速联合人脸检测和对准。我们将我们的方法与GPU上的最先进技术进行比较，结果如表II所示。 注意，我们当前的实现是基于未优化的MATLAB代码。



表II

### 四、总结

在本文中，我们提出了一个多任务级联CNNs基于框架的联合人脸检测和对准。实验结果表明，我们的方法在几个具有挑战性的基准（包括人脸检测的FDDB和WIDER FACE基准，以及人脸对准的AFLW基准）中始终优于先进的方法，同时实现640x480 VGA图像的实时性能，具有20x20最小面部尺寸。性能改进的三个主要贡献是精心设计的级联CNNs架构，在线易分错样本挖掘策略和联合人脸对准学习。

### 五、参考文献

[1] B. Yang, J. Yan, Z. Lei, and S. Z. Li, “Aggregate channel features formulti-view face detection,” in IEEE International Joint Conference onBiometrics, 2014, pp. 1-8.

[2] P. Viola and M. J. Jones, “Robust real-time face detection. International journal of computer vision,” vol. 57, no. 2, pp. 137-154, 2004

[3] M. T. Pham, Y. Gao, V. D. D. Hoang, and T. J. Cham, “Fast polygonal integration and its application in extending haar-like features to improve object detection,” in IEEE Conference on Computer Vision and Pattern Recognition, 2010, pp. 942-949.

[4] Q. Zhu, M. C. Yeh, K. T. Cheng, and S. Avidan, “Fast human detection using a cascade of histograms of oriented gradients,” in IEEE Computer Conference on Computer Vision and Pattern Recognition, 2006, pp. 1491-1498.

[5] M. Mathias, R. Benenson, M. Pedersoli, and L. Van Gool, “Face detection without bells and whistles,” in European Conference on Computer Vision, 2014, pp. 720-735.

[6] J. Yan, Z. Lei, L. Wen, and S. Li, “The fastest deformable part model for object detection,” in IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 2497-2504.

[7] X. Zhu, and D. Ramanan, “Face detection, pose estimation, and landmark localization in the wild,” in IEEE Conference on Computer Vision and Pattern Recognition, 2012, pp.2879-2886.

[8] M. Köstinger, P. Wohlhart, P. M. Roth, and H. Bischof, “Annotated facial landmarks in the wild: A large-scale, real-world database for facial landmark localization,” in IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2011, pp. 2144-2151.

[9] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “Imagenet classification with deep convolutional neural networks,” in Advances in neural information processing systems, 2012, pp. 1097-1105.

[10] Y. Sun, Y. Chen, X. Wang, and X. Tang, “Deep learning face representation by joint identification-verification,” in Advances in Neural Information Processing Systems, 2014, pp. 1988-1996.

[11] S. Yang, P. Luo, C. C. Loy, and X. Tang, “From facial parts responses to face detection: A deep learning approach,” in IEEE International Conference on Computer Vision, 2015, pp. 3676-3684.

[12] X. P. Burgos-Artizzu, P. Perona, and P. Dollar, “Robust face landmark estimation under occlusion,” in IEEE International Conference on Computer Vision, 2013, pp. 1513-1520.

[13] X. Cao, Y. Wei, F. Wen, and J. Sun, “Face alignment by explicit shape regression,” International Journal of Computer Vision, vol 107, no. 2, pp. 177-190, 2012.

[14] T. F. Cootes, G. J. Edwards, and C. J. Taylor, “Active appearance models,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 23,no. 6, pp. 681-685, 2001.

[15] X. Yu, J. Huang, S. Zhang, W. Yan, and D. Metaxas, “Pose-free facial landmark fitting via optimized part mixtures and cascaded deformable shape model,” in IEEE International Conference on Computer Vision,2013, pp. 1944-1951.

[16] J. Zhang, S. Shan, M. Kan, and X. Chen, “Coarse-to-fine auto-encoder networks (CFAN) for real-time face alignment,” in European Conference on Computer Vision, 2014, pp. 1-16.

[17] Luxand Incorporated: Luxand face SDK, <http://www.luxand.com/>

[18] D. Chen, S. Ren, Y. Wei, X. Cao, and J. Sun, “Joint cascade face detection and alignment,” in European Conference on Computer Vision, 2014, pp.109-122.

[19] H. Li, Z. Lin, X. Shen, J. Brandt, and G. Hua, “A convolutional neural network cascade for face detection,” in IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 5325-5334.

[20] C. Zhang, and Z. Zhang, “Improving multiview face detection with multi-task deep convolutional neural networks,” IEEE Winter Conference on Applications of Computer Vision, 2014, pp. 1036-1041.

[21] X. Xiong, and F. Torre, “Supervised descent method and its applications to face alignment,” in IEEE Conference on Computer Vision and Pattern Recognition, 2013, pp. 532-539.

[22] Z. Zhang, P. Luo, C. C. Loy, and X. Tang, “Facial landmark detection by deep multi-task learning,” in European Conference on Computer Vision, 2014, pp. 94-108.

[23] Z. Liu, P. Luo, X. Wang, and X. Tang, “Deep learning face attributes in the wild,” in IEEE International Conference on Computer Vision, 2015, pp. 3730-3738.

[24] S. Yang, P. Luo, C. C. Loy, and X. Tang, “WIDER FACE: A Face Detection Benchmark”. arXiv preprint arXiv:1511.06523.

[25] V. Jain, and E. G. Learned-Miller, “FDDB: A benchmark for face detection in unconstrained settings,” Technical Report UMCS-2010-009, University of Massachusetts, Amherst, 2010.

[26] B. Yang, J. Yan, Z. Lei, and S. Z. Li, “Convolutional channel features,” in IEEE International Conference on Computer Vision, 2015, pp. 82-90.

[27] R. Ranjan, V. M. Patel, and R. Chellappa, “A deep pyramid deformable part model for face detection,” in IEEE International Conference on Biometrics Theory, Applications and Systems, 2015, pp. 1-8.

[28] G. Ghiasi, and C. C. Fowlkes, “Occlusion Coherence: Detecting andLocalizing Occluded Faces,” arXiv preprint arXiv:1506.08347.

[29] S. S. Farfade, M. J. Saberian, and L. J. Li, “Multi-view face detection using deep convolutional neural networks,” in ACM on International Conference on Multimedia Retrieval, 2015, pp. 643-650.

[30] K. He, X. Zhang, S. Ren, J. Sun, “Delving deep into rectifiers: Surpassing human-level performance on imagenet classification,” in IEEE International Conference on Computer Vision, 2015, pp. 1026-1034.