**使用多任务级联卷积网络的面部检测与对齐**

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

**摘要**

由于各种姿势，光照以及遮挡，面部检测与对齐在无限制环境中具有一定挑战。最新的研究表明，深度学习方法可以在这两项任务上取得令人印象深刻的表现。在本论文中，我们提出了一个深度联级多任务框架，利用他们之间内在联系来提高他们的性能。特别地，我们的框架采用一个由三层精心设计的深度卷积网络组成的级联结构，可以以粗略到精准的方式预测面部和特征点定位。另外，在学习过程中，我们提出一个新的在线难例（hard sample）挖掘策略，可以自动提高性能，而无需手动选择样本。我们的方法通过最先进的技术在FDDB和WIDER FACE基准面部检测挑战，以及AFLW基准面部对齐挑战中取得良好的准确率，并同时保持实时的性能。

**索引词**

面部检测，面部对齐，联级卷积神经网络

**I.介绍**

面部检测和对其对很多面部应用来说是最基本的功能，比如面部识别和表情分析。然而，大的面部视觉变化，比如遮挡，大的姿势变化和极端的光线，对现实世界应用中的这些任务增加了巨大的挑战。

Viola和Jones[2]提出的级联面部检测器利用Haar-Like特征和AdaBoost来训练级联分类器，在实时效率方面实现了良好的性能。然而，相当多的研究工作[1, 3, 4]表明，即使用更高级的特征和分类器，该检测器也可能在真实世界应用中因更大的人脸视觉变化显著地降低表现。除了联级结构外， [5, 6, 7] 介绍了可变部件模型（DPM）用于面部检测并实现杰出的性能。然而，它们需要花费高计算量并在训练阶段通常需要昂贵的标注。近来，卷积神经网络（CNN）在各种计算机视觉任务中，比如图像分类[9]和面部识别[10]，取得显著进展。受到CNN在计算机视觉任务中的良好表现的启发，一些基于CNN的面部检测方法在近年被提出。Yang等人[11]训练了用于面部特征识别的深度卷积神经网络，以在面部区域获得高响应，而进一步产生面部的候选窗口。然而，由于它的复杂的CNN结构，这个方法在实践中是花费时间的。Li等人[19]使用联级的CNN进行面部检测，但是它需要在面部检测时用边界框校准，并有额外的计算花费，并且它忽略了面部定位和边界框回归之间的内在联系。

面部对齐也吸引了广泛的兴趣。基于回归的方法[12, 13, 16]和模板拟合方法[14, 15, 17]是两种流行的类别。最近，Zhang等人[22]提出使用面部特征识别作为辅助任务来增强使用深度卷积神经网络的面部对齐的性能。

然而，大部分有效的面部检测与面部对齐方法忽略了这两种任务的内在联系。虽然有几项研究工作试图共同解决他们，但是这些工作依旧有局限性。Chen等人[18]利用像素值的差异，用随机森林共同进行对齐和检测。但是，算法特性限制了其性能。Zhang等人[20]使用多任务CNN来提高多视角面部检测的准确性，但是检测的准确性被由很弱的面部检测器产生的初始检测窗口所限制。

另一方面，在训练过程中，挖掘训练中的难例是加强检测器能力的关键。然而，传统难例的挖掘通常采用离线方式，显著地增加手工操作。为面部检测和对其设计一个在线难例挖掘的方法，自动适应当前的训练过程，将会令人满意。

本篇论文，我们提供一个新的框架，通过多任务学习统一级联CNN来整合这两项任务。提供的CNN由三个阶段组成。在第一个阶段，他通过一个浅的CNN快速产生候选窗口。然后它通过一个更复杂的CNN来优化，拒绝大量的非面部窗口。最后提示用一个更加强大的CNN来优化结果并输出面部特征位置。由于这个多任务学习框架，算法的性能可以显著地提高。本论文的主要贡献总结如下：（1）我们为整合面部检测和对其提供了一种新的基于联级CNN的框架，并且为了实时性，精心设计了轻量的CNN架构。（2）我们提供一个有效的方法来进行在线难例挖掘，以提高性能。（3）对挑战基准测试进行了广泛的实验，以显示在面部检测和面部对齐任务中与最先进的技术相比，我们所提出的方法能够显著提高性能。

**II.方法**

在本节中，我们将描述我们面向整合面部检测和对齐的方法。

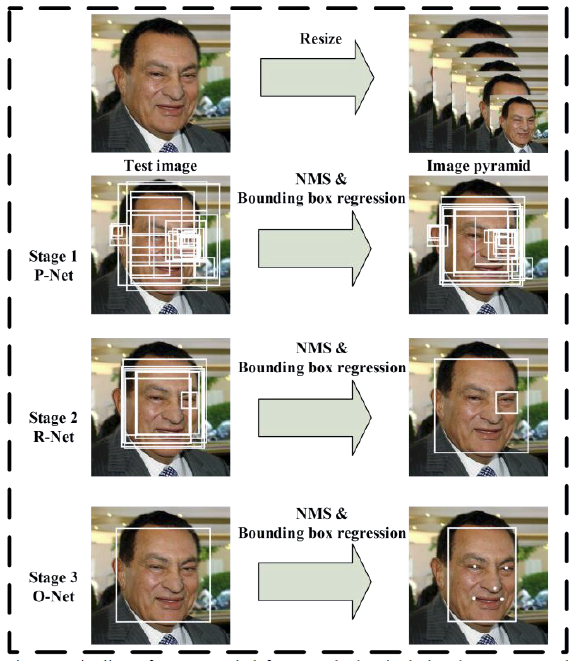
**A. 全部框架**

我们的方法的整体流水线如图1所示。给定一个图像，我们初始将它调整到不同比例，以构建一个图像金字塔。它是接下来三阶段级联框架的输入：

第一阶段：我们实现一个完全卷积的网络，叫做提案网络（P-Net），用以获取与[29]类似的候选窗口以及它们的边界框回归向量。然后我们用估计的边界框回归向量来校准候选边框向量。之后，我们采用非极大值抑制（NMS）来合并高度重叠的候选框。

第二阶段：所有候选边框都被送到另一个CNN，叫做提取网络（R-Net），进一步拒绝大量假候选框，用边界框回归执行校准，并用NMS候选框合并。

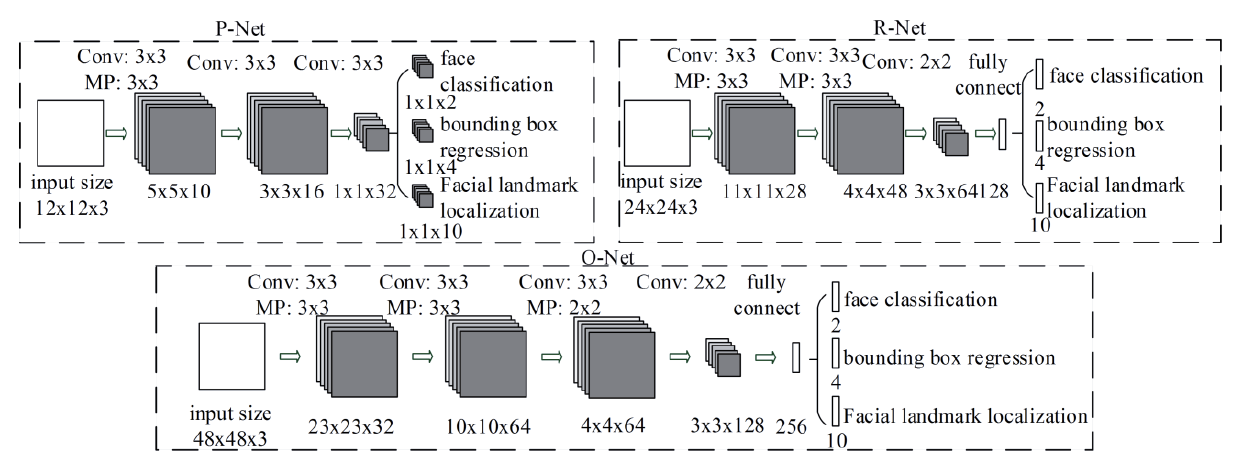
第三阶段：这个阶段跟第二阶段很相似，但是在这个阶段我们目标是更详细的描述面部。特别地，该网络将输出五个面部特征位置。



图一：我们的联级框架流水线，包括三个阶段的多任务深度卷机网络。首先，候选窗口是通过一个快速提案网络（P-Net）。之后，我们通过提取网络（R-Net）在下一个阶段细化这些候选框。在第三阶段，输出网络（O-Net）产生最后的边界框以及面部特征位置。

**B．CNN架构**

在[19]中，多个CNN已经被设计用于面部检测。但是，我们注意到，它的性能或许受到以下事实的限制：（1）一些过滤器缺少权重的多样性，这或许会他们产生歧义的描述。（2）与其他多分类对象检测和分类任务相比，面部检测是一个有挑战性的二分类任务，所以，它或许需要更少数量的过滤器，但是对它们做更多的区分。因此，我们减少过滤器的数量并且将5x5的过滤器改成3x3的过滤器以减少计算量同时增加深度的同时以获得更好的性能。有了这些改进，与之前[19]中的架构相比，我们可以获得更好的性能同时更少的运行时间（结果如表1所示。为了公平的比较，我们对这两种方法使用同样的数据）。我们的CNN架构如图2所示。



图二：P-Net，R-Net以及O-Net的架构。MP意思是最大池化，Conv意思是卷积。卷积和池化的步长分别是1和2。

**C．训练**

我们利用三个任务来训练我们的CNN检测器：面部/非面部分类器，边界框回归以及面部特征定位。

1）面部分类：学习目标被定义为一个二分类问题。对于每一个样本Xi，我们使用交叉熵损失函数：



pi是网络产生的指示一个样本是一个面部的概率。符号yidet∈{0,1}表示正确标注（ground-truth）的标签。

2）边界框回归：对于每个候选窗口，我们预测它与最接近真实值（比如，边界框的左上角坐标，高和宽）的偏移。学习目标被定义为一个回归问题，并且我们对每个样本Xi使用欧几里得损失函数：



box i是从网络获得的回归目标，并且ybox i是正确标注的坐标。有4个坐标，包括左上角，高，宽，任何ybox i∈R4。

3）面部特征定位：与边界框回归任务类似，面部特征检测被定义为一个回归问题并且我们最小化欧氏距离损失函数：



landmark i是从网络中获得的面部特征坐标，并且ylandmark i是真实坐标。有五个面部特征点，包括左眼，右眼，鼻子，左嘴角，右嘴角，任何ylandmark i∈R10

4）多源训练：由于我们在每个CNN中采用不同的任务，因此在学习过程中有不同类型的训练图像，比如，面部，非面部以及部分对齐的面部。在这种情况下，一些损失函数（比如公式（1）-（3））没有被使用。举个例子，对于样本的背景区域，我们仅计算，并且其它两个损失值被设置为0。这个可以直接用一个样本类型指示器来实现。那么整个学习目标就可以被定义为



N是训练样本的数量，aj表示任务的重要性。我们在P-Net和R-Net中使用（adet=1，abox=0.5，alandmark=0.5），O-Net为了获得更准确的面部特征定位，我们使用（adet=1，abox=0.5，alandmark=1），βj i∈{0,1}是样本类型指示器。在这种情况下，采用随机梯度下降来训练CNN是自然而然的。

5）在线难例挖掘：不同于在原始分类器被训练之后进行传统的难例挖掘，我们在面部分类任务中进行在线难例挖掘，以适应训练过程。

特别的，在每个小批量中，我们对所有样本在正向传播阶段计算损失并进行排序，选择前70%做为难例。然后我们在反向传播阶段仅从计算难例中的梯度。这意味着，我们忽略了训练过程中对检测器强化鲜有帮助的容易样本（easy samples）。实验表明，这种策略在没有手工样本选择的情况下产生更好的性能。它的效果在第三章展现。

**III.实验**

在本章节中，我们首先评估了提供难例挖掘策略的效果。然后我们比较FDDB[25]，WIDER FACE[24]以及AFLW[8]中的最新方法与我们的面部检测器和对齐的方法。FDDB数据集包含了在2845张图片中标注了5171张面部。WIDER FACE数据集包含了在32203张图片中标记了393703个边界框，其中50%的图片根据图片难度分成三个子集，40%用于训练，剩余的用于验证。AFLW包含了24386张面部的面部特征注释，我们使用与[22]相同的测试子集。最后，我们评估了我们面部检测的计算效率。

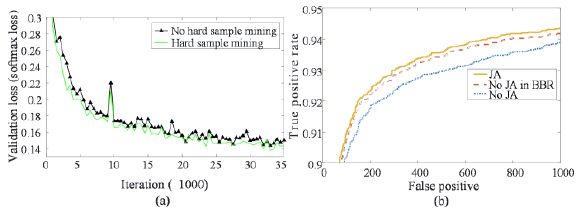
**A．训练数据**

由于我们联合执行面部检测和对齐，在我们的训练过程中，我们使用4种不同类型的数据标注：（i）负样本：特征数据的交并比（Intersection-over-Union，IoU）小于0.3；（ii）正样本：特征数据的IoU高于0.65；（iii）部分面部样本：特征数据的IoU介于0.4和0.65之间；（iv）特征面部样本：面部标上5个特征点。负样本和正样本用于面部分类任务，负样本和部分面部样本用于边界框回归，特征面部样本用于面部特征点定位。每个网络的训练数据描述如下：

1）P-Net：我们随机从WIDER FACE[24]数据集中切割几块，以收集正样本，负样本以及部分面部样本。然后，我们从CelebA[23]数据集中切割做为特征面部样本。

2）R-Net：我们使用我们框架第一阶段来从WIDER FACE[24]中检测面部，以收集正样本，负样本以及部分面部样本，同时特征面部样本从CelebA[23]中检测得到。

3）O-Net：与R-Net类似的收集数据，但我们用我们框架的前两个阶段来检测面部。



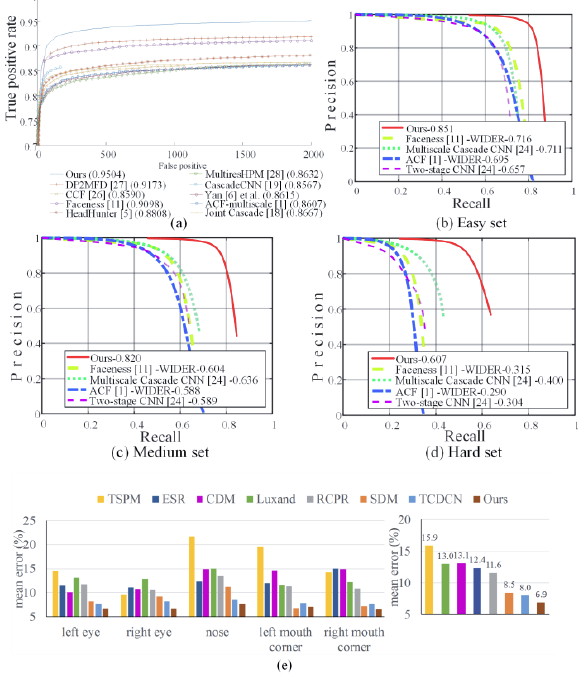
图三：（a）使用和不使用在线难例挖掘时O-Net的验证集损失。（b）“JA”表示联合面部对其学习，“No JA”表示不联合。“No JA in BBR”表示在训练CNN边界框回归时不用联合

**B.在线难例挖掘的效果**

为了评估在线难例挖掘策略的贡献，我们训练两个O-Net（有和没有在线难例挖掘）并且比较他们的损失曲线。为了更直观的比较，我们仅为面部分类任务训练O-Net。包括网络初始化在内的所有训练参数在这两个O-Net中都是相同的。为了更容易的比较它们，我们使用固定的学习率。图三（a）显示了这两种不同训练方式的损失曲线。非常明显的是，难例挖掘有利于性能提升。

**C.联合检测和对齐的效果**

为了评估联合检测和对齐的贡献，我们在FDDB数据集（为了公平起见，使用相同的P-Net和R-Net）上评估了两个不同O-Net（联合面部特征回归任务和不联合）。我们在这两个O-Net中也比较了边界框回归的性能。图3（b）表明了联合特征定位任务学习对面部分类和边界框回归是有意的。



图四：（a）FDDB上的评估。（b-d）WIDER FACE三个子集上的评估。方法后的数字只是平均准确率。（e）在AFLW上面部对齐的评估

**D.面部检测评估**

为了评估我们面部检测方法的性能，我们比较了我们的方法与FDDB上最新的方法[5, 6, 11, 18, 19, 26, 27, 28, 29]，以及WIDER FACE上最新的方法[1, 24, 11]。图四（a）-（d）显示了我们的方法在两个基准上始终优于所有之前的方法。我们也在一些挑战照片上评估了我们的方法。

**E.面部对齐评估**

在这部分，我们就面部对齐方法的性能与如下几个方法进行了对比：RCPR[12]，TSPM[7]，Luxand face SDK[17]，ESR[13]，CDM[15]，SDM[21]以及TCDCN[22]。在测试阶段，我们的方法面部检测失败的有13张图片。所以，我们剪裁掉这13张图片的中心区域，并视其为O-Net的输入。测量的平均误差是估计特征和实际特征的距离，并对眼间距进行归一化处理。图四（e）显示了我们的方法要优于所有最新方法。

**F.运行时效率**

给定的联级架构，我们的方法可以在联合面部检测和对齐实现非常快的速度。在2.6GHz的CPU上实现16fps，在GPU（NVIDIA Titan Black）上有99fps。我们目前的实现是基于未优化的MATLAB代码。

**IV.结论**

在本文中，我们提出一个基于多任务联级CNN的框架，用于联合面部检测和对齐。实验结果表明，我们的方法在多个挑战基准测试（包括用于面部检测的FDDB和WIDER FACE，面部对齐的AFLW基准）中始终优于最新方法，同时保持实时性性能。未来，我们将利用面部检测与其他面部分析任务之间的内在联系，进一步提高性能。

**参考文献**

[1] B. Yang, J. Yan, Z. Lei, and S. Z. Li, “Aggregate channel eatures for

multi-view face detection,” in IEEE International Joint Conference on Biometrics, 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 eatures,” 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 iometrics Theory, Applications and Systems, 2015, pp. 1-8.

[28] G. Ghiasi, and C. C. Fowlkes, “Occlusion Coherence: Detecting and Localizing 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.