RetinaFace: Single-stage Dense Face Localisation in the Wild

团队：Datawhale 深度学习团队

汇报人： 苏静静

2019.11.30

目录

相关介绍

相关介绍

RetinaFace 由 InsightFace 和帝国理工大学联合提出的one-stage人脸检测算法，InsightFace 为目前针对 2D 与 3D 人脸分析（含检测、识别、对齐、属性识别等）最知名和开发者最活跃的开源库，是目前开源的人脸检测算法中效果最好的算法。

作者提供了三种基础网络，基于ResNet的ResNet50和ResNet152版本能提供更好的精度，以及基于mobilenet（0.25）的轻量版本mnet，检测速度更快。

https://github.com/deepinsight/insightface

相关介绍

图像金字塔 vs .特征金字塔

随着特征金字塔的出现，多尺度特征图上的滑动anchor迅速主导了人脸检测。

两阶段vs .单阶段

两阶段方法(如Faster)和单阶段方法(如SSD和RetinaNet)。
两阶段方法采用了一种具有高定位精度的“proposal与细化”机制。
与两阶段方法相比，单阶段方法效率更高，召回率更高，但存在假阳性率更高和定位准确性降低的风险。

上下文建模

利用特征金字塔的上下文模块，增强模型对微小人脸的上下文推理能力，扩大欧几里德网格的感受野。

论文摘要

论文摘要

提出了**一种鲁棒的single stage人脸检测器RetinaFace**，它利用**联合的额外监督和自监督多任务学习**的优点，对不同尺度的人脸进行像素级定位。

在以下五个方面做出了贡献：

(1)在WILDER FACE数据集中手工标注5个人脸Landmark，并在这个额外的监督信号帮助下，在hard face检测显著改善。

(2)进一步添加自监督网格解码器（mesh decoder）分支，与已有的监督分支并行预测像素级的3D形状人脸信息。

(3)在WILDER FACE hard测试集上，RetinaFace的平均精度(AP)比最先进的平均精度(AP)高出1.1%(达到AP = 91.4%)。

(4)在IJB-C测试集上，RetinaFace使现有人脸识别方法(ArcFace)结果得到提升(FAR=1e-6, TAR=89.59%)。

(5)采用轻量级backbone网络，RetinaFace可以在单个CPU核上实时运行vga分辨率图像。

论文思路

论文思路

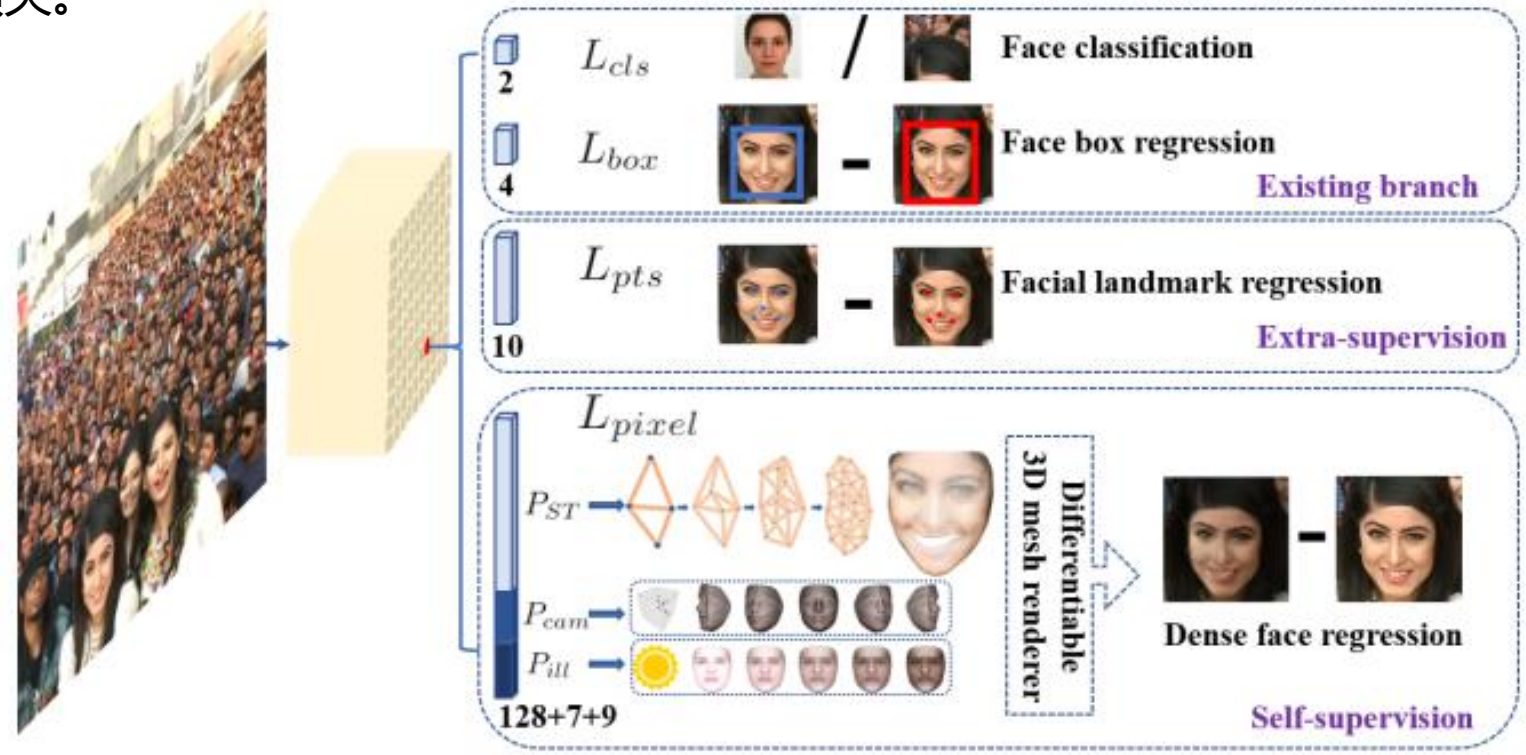
多任务学习

- 由于训练数据的限制，MTCNN和STN等未验证微小的人脸检测是否可以从5个人脸特征点的额外监督中获益。
- 在Mask R-CNN中，通过添加一个分支来与现有的分支并行预测一个对象掩码来进行Bbox的识别和回归，大大提高了检测性能。证实了密集的像素级注释也有助于改进检测。
- Mesh decoder 通过利用图形卷积在形状和纹理上实现了超实时的速度。

在这篇论文中就利用由5个人脸特征点构成的额外监督信号及其他任务验证是否对人脸检测效果有所改进。

论文思路

在目标检测这一块的变动不大，多加了人脸关键点的损失来辅助训练，同时增加了一个所谓自监督学习的分支，在这个分支上进行2D到3D的编码与解码，同时计算进行解码编码还原后的五个人脸关键点的损失。



论文思路

多任务损失函数

$$L = L_{cls}(p_i, p_i^*) + \lambda_1 p_i^* L_{box}(t_i, t_i^*) + \lambda_2 p_i^* L_{pts}(l_i, l_i^*) + \lambda_3 p_i^* L_{pixel}.$$

前面2项和以往的多任务人脸检测的损失是一样的，分类的损失，bbox的回归损失，第三项是五个人脸关键点的回归损失。最后一项损失是Dense Regression分支带来的损失，损失权重取值分别是0.25， 0.1和0.01，于带标签的检测和landamark的损失权重会更高， 而基于自监督的Dense Regression Branch的权重较少。

论文思路

Dense Regression Branch

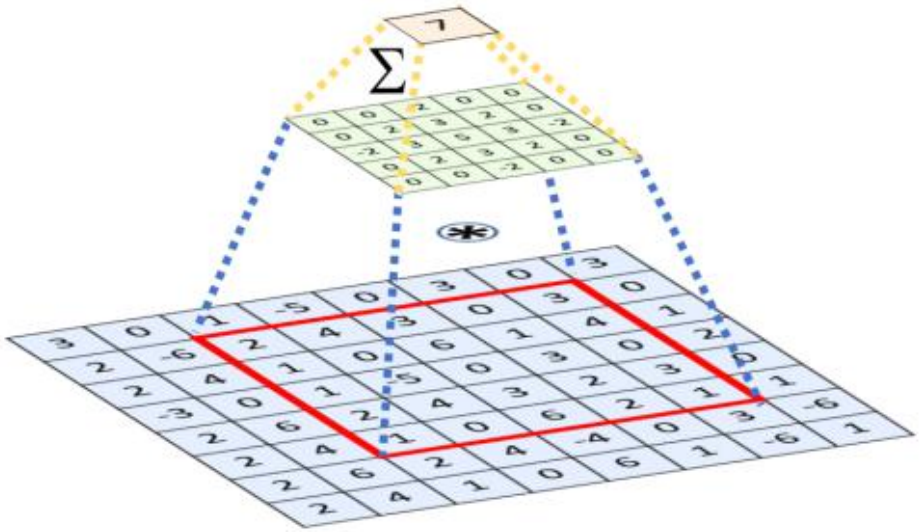
将2D的人脸映射到3D模型上，再将3D模型解码为2D图片，然后计算经过编解码的图片和原始图片的差别。Dense Regression Loss实际上就是经过编解码的图片和原始图片的五个人脸特征点的位置的差别。中间用到了图卷积。

Mesh Decoder

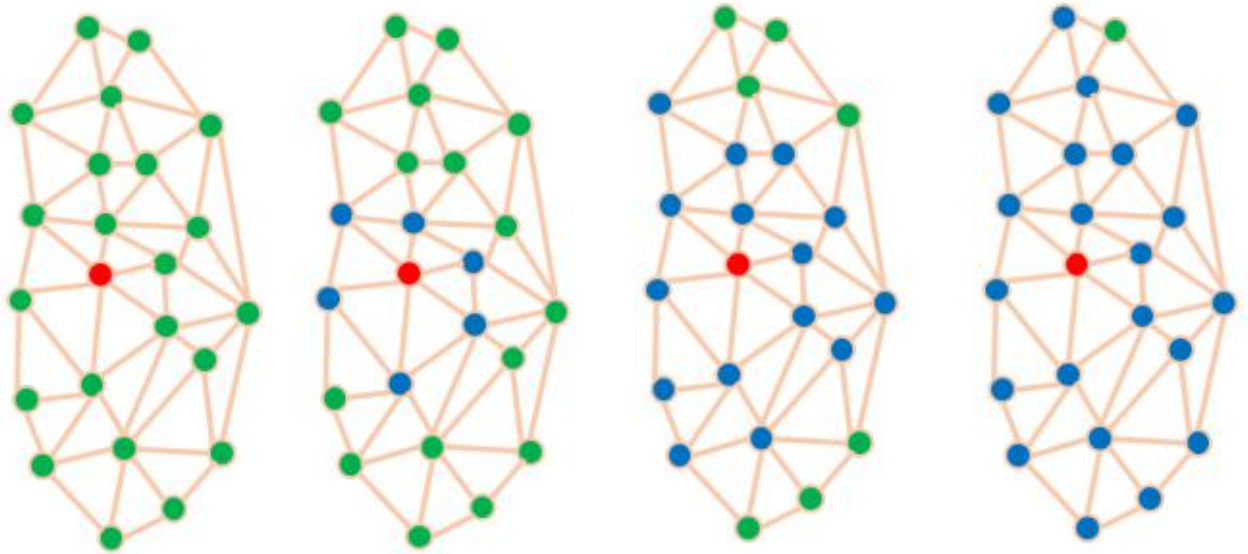
网格解码器是基于快速局部谱滤波（spectral filter）的图卷积（graph convolution）方法。参数量会比我们平时用的普通2D卷积计算量要少。

论文思路

Dense Regression Branch



(a) 2D Convolution



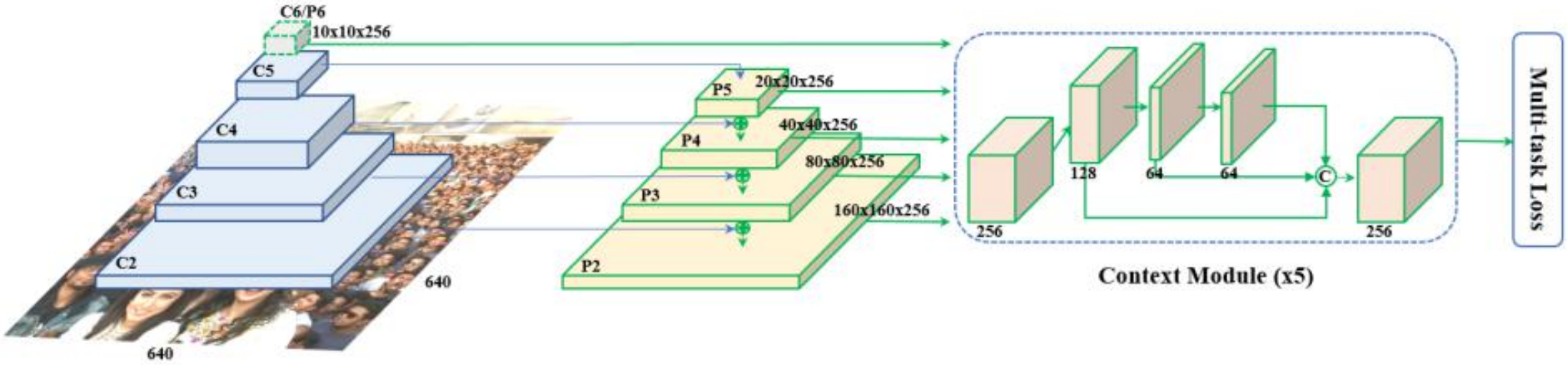
(b) Graph Convolution

$$Kernel_H * Kernel_w * Channel_{in} * Channel_{out}$$

$$K * Channel_{in} * Channel_{out}$$

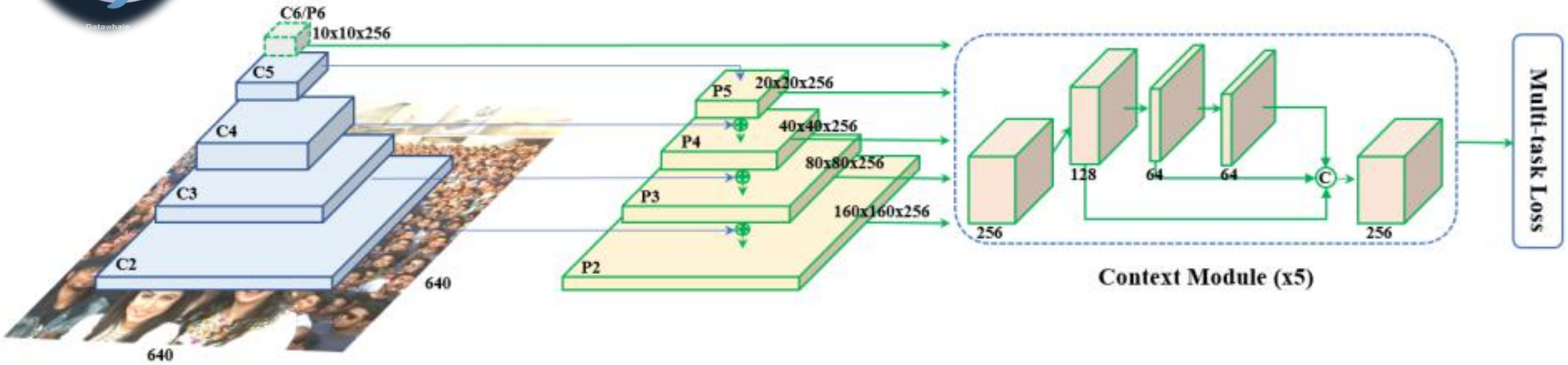
邻域距离是通过计算连接两个顶点的最小边数来计算的。

论文思路



- 作者采用 FPN 中的 Feature Pyramid 结构，并以 ResNet-152 作为 Backbone，其中， C2/C3/C4/C5 为 ResNet 中各个 Residual Block 所生成的 Feature Map，而 C6 由 C5 经过 3*3 的卷积层生成（步长为 2）。
- 设计了 Context Module 用于增加模型的感受野以及上下文信息，每层特征金字塔层都接一个独立的 Context Module，提高 感受野并增加网络背景建模能力。
- 通过可变形卷积网络（DCN）替换横向连接和上下文模块中的所有3x3卷积层，进一步加强了非刚性的上下文建模能力。

论文思路



Feature Pyramid	Stride	Anchor
P_2 (160 × 160 × 256)	4	16, 20.16, 25.40
P_3 (80 × 80 × 256)	8	32, 40.32, 50.80
P_4 (40 × 40 × 256)	16	64, 80.63, 101.59
P_5 (20 × 20 × 256)	32	128, 161.26, 203.19
P_6 (10 × 10 × 256)	64	256, 322.54, 406.37

- **损失头**：对于**负acnhors，只应用分类损失**。对于**正anchors，计算了多任务损失。**
- **Anchor 设置**：在特性金字塔levels(从P2到P6)上使用特定于尺度的anchor。

论文思路

训练样本

在训练过程中，IoU>0.5时为正样本，当IoU<0.3时为负样本。其他的在训练中被忽略。且采用标准OHEM来缓解正、负训练样本之间的显著不平衡。（根据损失值对负锚进行排序，并选择损失最大的anchors），负样本和正样本之间的比例至少为3:1。

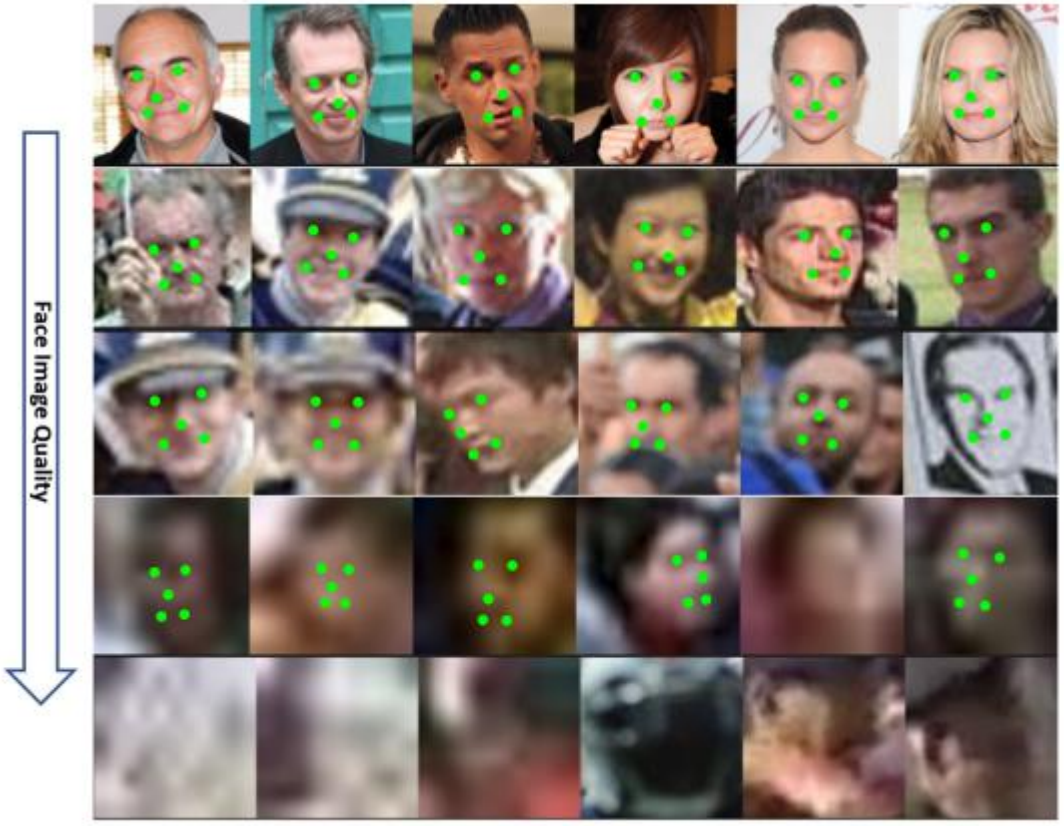
数据增强

随机裁剪修改为裁剪正方形区域，其边长为原图短边长乘以 [0.3, 1] 中的随机因子。除此以外，作者还采用随机翻转以及色彩抖动。

实验与结果

实验与结果

数据集



Level	Face Number	Criterion
1	4,127	indisputable 68 landmarks [44]
2	12,636	annotatable 68 landmarks [44]
3	38,140	indisputable 5 landmarks
4	50,024	annotatable 5 landmarks
5	94,095	distinguish by context

实验与结果

Method	Easy	Medium	Hard	mAP [33]
FPN+Context	95.532	95.134	90.714	50.842
+DCN	96.349	95.833	91.286	51.522
$+L_{pts}$	96.467	96.075	91.694	52.297
$+L_{pixel}$	96.413	95.864	91.276	51.492
$+L_{pts} + L_{pixel}$	**96.942**	**96.175**	**91.857**	**52.318**

增加多任务能够提高人脸检测性能。

实验与结果

Methods	LFW	CFP-FP	AgeDB-30
MTCNN+ArcFace [11]	99.83	98.37	98.15
RetinaFace+ArcFace	**99.86**	**99.49**	**98.60**

人脸检测及对齐的准确性会影响人脸识别的准确性。

实验与结果

Methods	LFW	CFP-FP	AgeDB-30
MTCNN+ArcFace [11]	99.83	98.37	98.15
RetinaFace+ArcFace	**99.86**	**99.49**	**98.60**

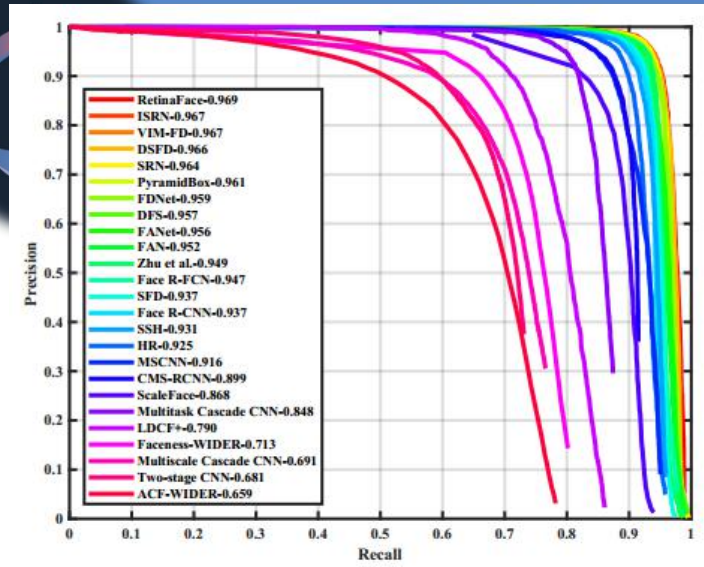
人脸检测及对齐的准确性会影响人脸识别的准确性。

实验与结果

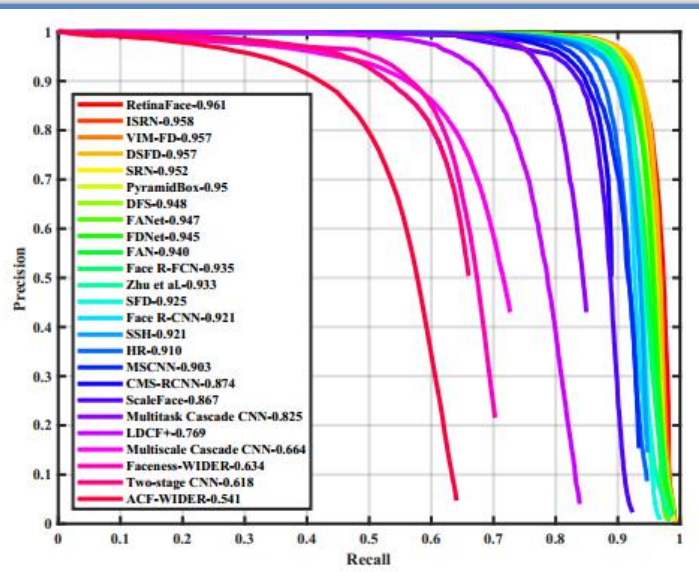
使用轻量级backbone网络

Backbones	VGA	HD	4K
ResNet-152 (GPU)	75.1	443.2	1742
MobileNet-0.25 (GPU)	1.4	6.1	25.6
MobileNet-0.25 (CPU-m)	5.5	50.3	-
MobileNet-0.25 (CPU-1)	17.2	130.4	-
MobileNet-0.25 (ARM)	61.2	434.3	-

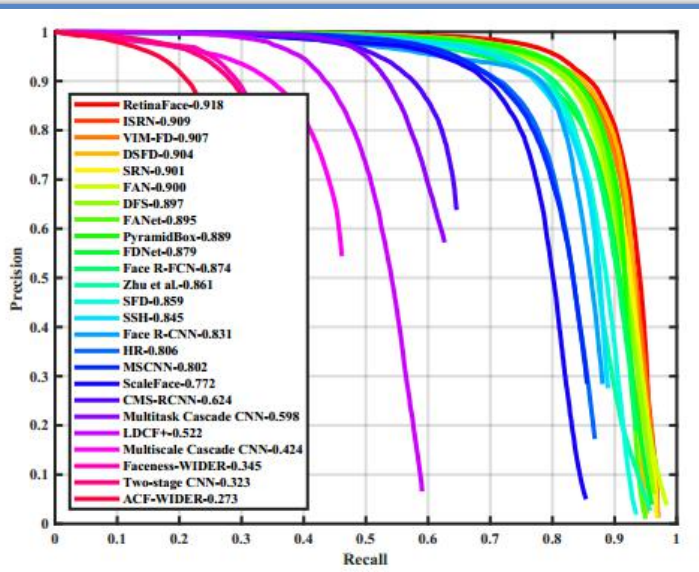
Table 5. Inference time (ms) of RetinaFace with different backbones (ResNet-152 and MobileNet-0.25) on different input sizes (VGA@640x480, HD@1920x1080 and 4K@4096x2160). "CPU-1" and "CPU-m" denote single-thread and multi-thread test on the Intel i7-6700K CPU, respectively. "GPU" refers to the NVIDIA Tesla P40 GPU and "ARM" platform is RK3399(A72x2).



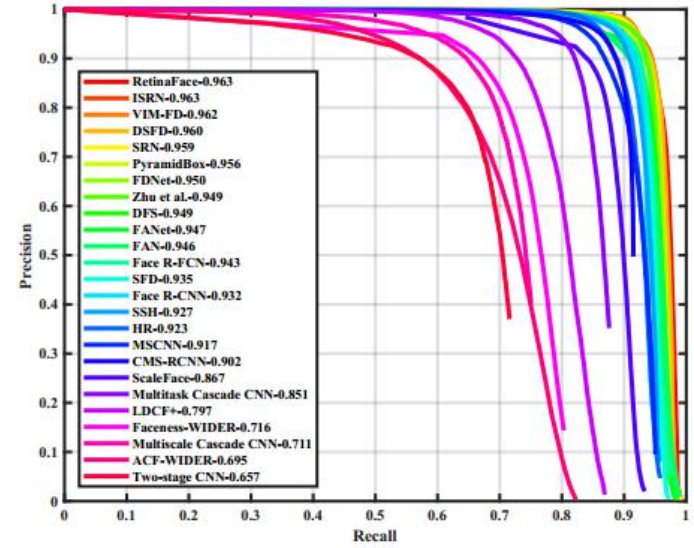
(a) Val: Easy



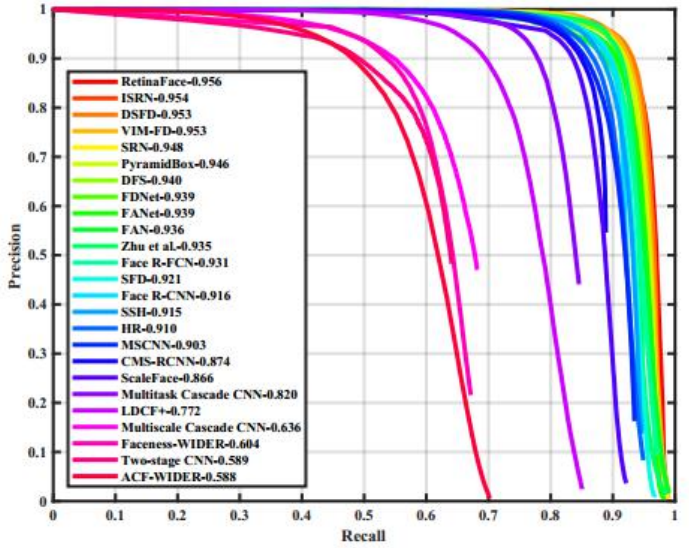
(b) Val: Medium



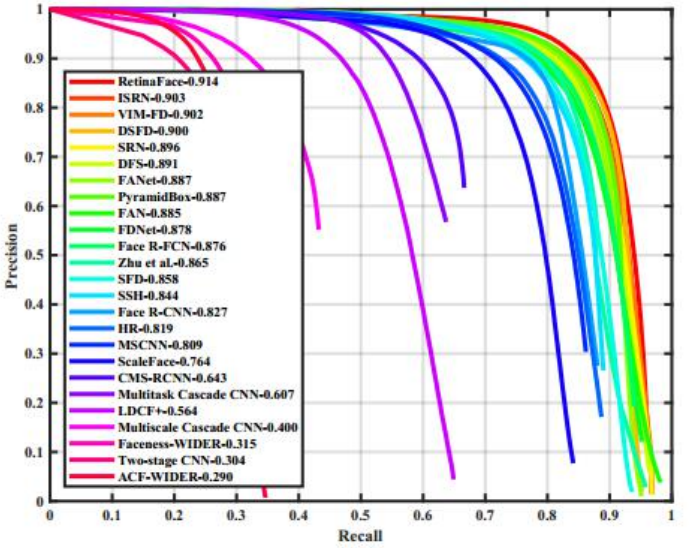
(c) Val: Hard



(d) Test: Easy

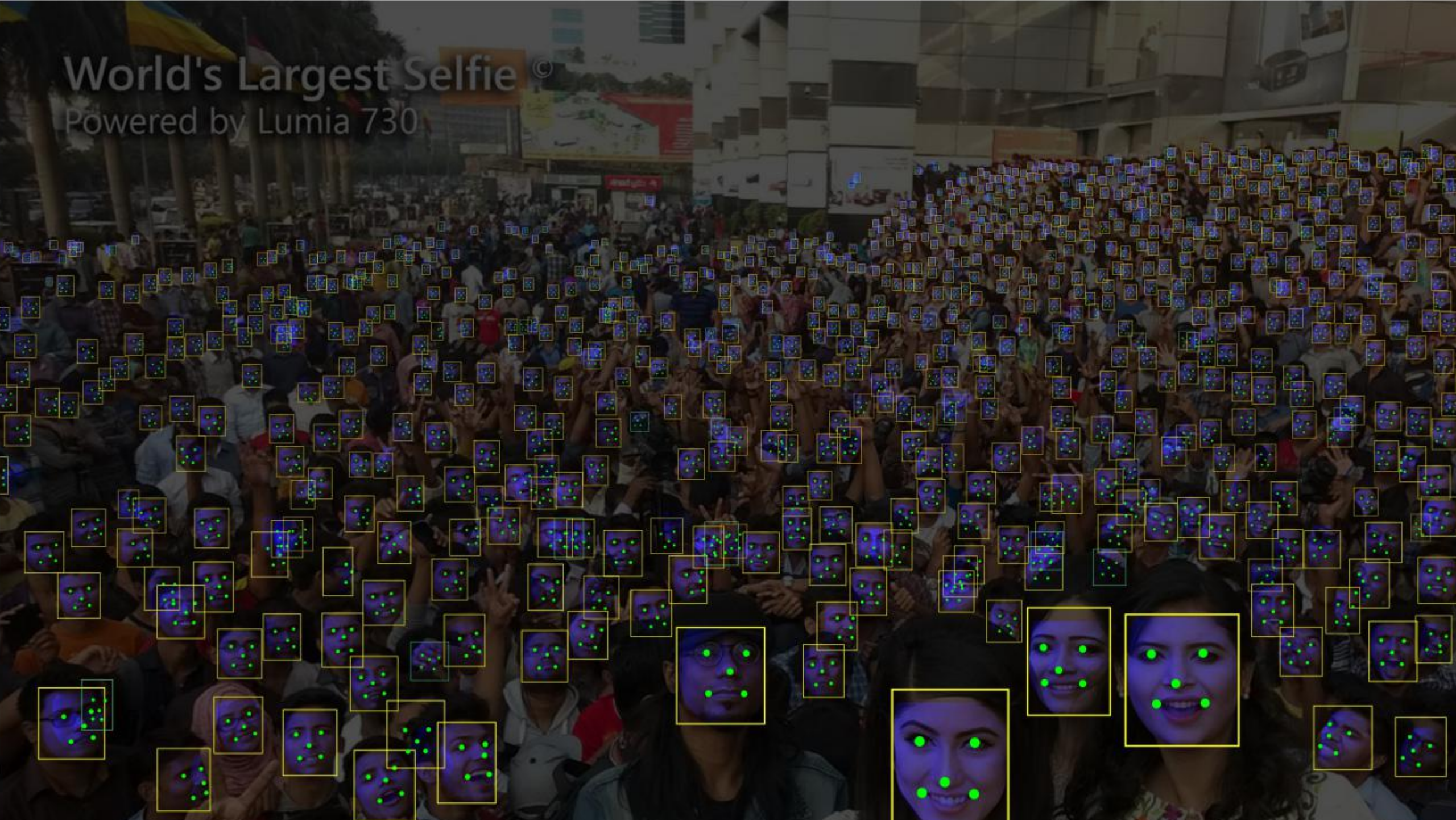


(e) Test: Medium



(f) Test: Hard

Figure 5. Precision-recall curves on the WIDER FACE validation and test subsets.


World's Largest Selfie ©
Powered by Lumia 730

总结

总结

Summary

总结 SUMMARY

- 特征金字塔的技术，实现了多尺度信息的融合，对检测小物体有重要的作用。
- 利用特征金字塔上的上下文模块，增强模型上下文推理能力，扩大欧几里德网格的感受野。
- 多任务学习策略，如增加人脸特征点等任务后能够提高人脸检测算法的效果。
- 使用轻量级backbone网络，RetinaFace可以实时运行。
- 人脸检测算准确性会影响人脸识别效果。

Thank you~