# 暗光人脸检测

「随便」小组

信息科学技术学院

2022 年 6 月 9 日

1. 引入
2. 尝试的思路
3. 鲁棒性与可解释性

1 引入

2 尝试的思路

3 鲁棒性与可解释性

# 暗光人脸检测

- 人脸检测已经能做到很高的正确率，在一些数据集上已经达到 0.9 甚至 0.95 的 mAP score 。
- 然而，一旦将问题背景转化为暗光，检测正确率则直线下降。
- 试想，你都看不清，水平还没你高的模型凭什么能看清？

# 暗光人脸检测

- 机器相比人眼的优势：在低光照条件下能正常分辨明暗。即使图片看起来是一片黑，像素值实际上仍然是有起伏的。
- 不幸的是，现有的人脸检测模型也都在正常光照的图片上被惯坏了，它们在黑暗中也看不清。而且由于缺少数据，重新训练是不太现实的。
- 因此，应对暗光图片的方法也很简单。先用一些方法放大像素间的差别（即提亮），然后再进行检测即可。

# 小组介绍

- 组名：「随便」
- 成绩：DARKFACE 第一，PKUCVFACE 第三
- 组员介绍和分工：

（组长）邱天异 Baseline，Pseudo-Labeling，鲁棒性与可解释性

钱易 Baseline，Per-Block Enhancement，视频

彭博 Baseline，Double Pass，鲁棒性，视频

1 引入

2 尝试的思路

Baseline

Pseudo-Labeling

Per-Block Enhancement

Double Pass

Conclusion

3 鲁棒性与可解释性

本节中，我们会介绍四种思路，包括一个最基础的 Baseline 做法，和三种巧妙的优化。

1 引入

2 尝试的思路
- Baseline
- Pseudo-Labeling
- Per-Block Enhancement
- Double Pass
- Conclusion

3 鲁棒性与可解释性

# Baseline: Main Idea

- 这是最朴实无华的做法。
- 选用一个图片提亮模型，再选用一个人脸识别模型。
  - 提亮模型：DUAL（传统算法），Zero-DCE 或 SCI
  - 人脸识别模型：DSFD 或 MogFace
- 然后……

**方案一** 把两个模型拼接后直接使用

**方案二** 拼接后在暗光图片上进行进一步训练（finetune）

**方案三** 利用正常光照图片，进行迁移学习

- 对于方案二，注意把两个模型在一起训练而非分开。

# Baseline 方案一：直接拼接

- 获取两个模型的代码和预训练参数，直接拼接，并运行得出结果。
  - DUAL + DSFD：$\mathrm{mAP} = 0.37$
  - Zero-DCE + DSFD：$\mathrm{mAP} = 0.38$
  - Zero-DCE + MogFace：$\mathrm{mAP} = 0.43$
- 以上成绩均在 DARKFACE 上测得，下同。

# Baseline 方案一：直接拼接

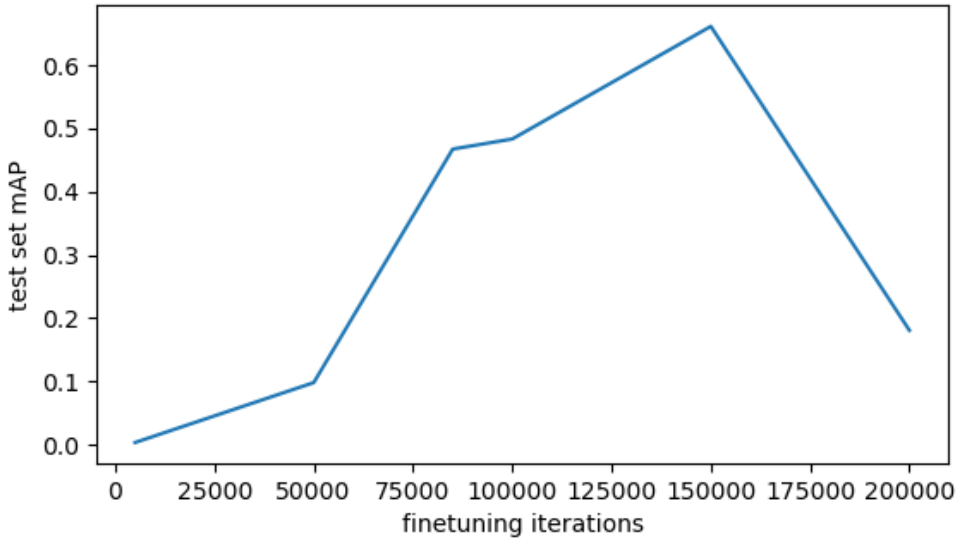
- 下面是提亮后的效果图。
- 提亮过的图片相比原图的能见度高了不少，但还存在很大问题。

# Baseline 方案二: Finetuning

- 获取两个模型的代码和预训练参数，将其拼接；
- 以预训练参数为起点继续训练，取训练过程中的最优模型，并运行得出结果。
  - Zero-DCE + DSFD：$\mathrm{mAP} = 0.57$
  - SCI + DSFD：$\mathrm{mAP} = 0.56$
  - SCI + MogFace：$\mathrm{mAP} = 0.66$ （第一名）

# Baseline 方案二: Finetuning

- 在 finetune 的过程中，测试集（验证集）表现会先增后降。
- 原因：finetune 越久，预训练数据就忘得越多，于是就越容易在数量有限的 finetune 训练集上过拟合。
- 下图为 SCI + MogFace 在 finetune 过程中的训练曲线。



# Baseline 方案二: Finetuning

- 下面是提亮后的效果图。
- 为与方案一对比，选择的是 Zero-DCE + DSFD 。

# Baseline 方案三: 迁移学习（Transfer Learning）

- 前面提到，暗光人脸检测的数据有限，不能支持重新训练模型。那么能否使正常光照的图片为我们所用呢？
- 在这一方向上我们未做很多探索。

1 引入

2 尝试的思路

Baseline

Pseudo-Labeling

Per-Block Enhancement

Double Pass

Conclusion

3 鲁棒性与可解释性

# Pseudo-Labeling: Motivation

- 一般来说我们会在训练集上把模型训好，再去测试集上进行评测。
- 但是这次并不一样，我们不仅手握训练集的图片和标注，还有测试集的图片。
- 期末考试之前，偷到了马上要考的卷子，应该如何利用？

**第一步** 把课程内容学懂（否则看不懂题）

**第二步** 把考题背下来

**第三步** 回头在课本和往年题里找答案，把答案背下来

# Pseudo-Labeling: Implementation

**第一步** 在训练集上训练模型

- 实验中直接使用了已训好的 Baseline 模型

**第二步** 让模型记住测试集上的重要特征

- 如何提取出与人脸最相关的特征，并让模型记住？
- 用已训好的模型（teacher）在测试集上检测人脸，再让待学习的模型（student）拟合这些检测结果。
- 这些检测结果即称为 pseudo-labels（伪标签）。

**第三步** 在训练集里学习那些重要特征的含义

- 把模型在训练集中再次训练。

# Pseudo-Labeling: Implementation

- 二三两步在实现中可以合并：把测试集和训练集按一定比例打乱混合起来，用作训练样本。
- 理想情况下，student 至少可以全盘复制 teacher，所以成绩至少不会差于 teacher 。
- 在实验中，teacher 和 student 都取为 $\mathrm{mAP} = 0.57$ 的 Baseline 模型。

# Pseudo-Labeling: Results

- 100% 训练集（Baseline）：$\mathrm{mAP} = 0.57$
- 80% 训练集 +20% 测试集：$\mathrm{mAP} = 0.52$
- 73% 训练集 +27% 测试集：$\mathrm{mAP} = 0.51$
- 57% 训练集 +43% 测试集：$\mathrm{mAP} = 0.50$

- 测试集比重越大，训练效果越差。
- 可见，伪标签这一方法起到了反效果。

# Pseudo-Labeling: Analysis

失败原因？

- 把 teacher 的预测作为 ground truth 时，丢失了置信度信息。这一问题可以用软标签（soft labels）缓解。
- 在针对测试集优化模型时，验证集就失去了作用。

1 引入

2 尝试的思路

Baseline

Pseudo-Labeling

Per-Block Enhancement

Double Pass

Conclusion

3 鲁棒性与可解释性

# Per-Block Enhancement: Motivation

- 上图中，左边很亮，右边很暗。
- 一张图中亮度不同的人脸，可能对识别产生影响。
- 常见的提亮方法无法保证各张人脸亮度相近。

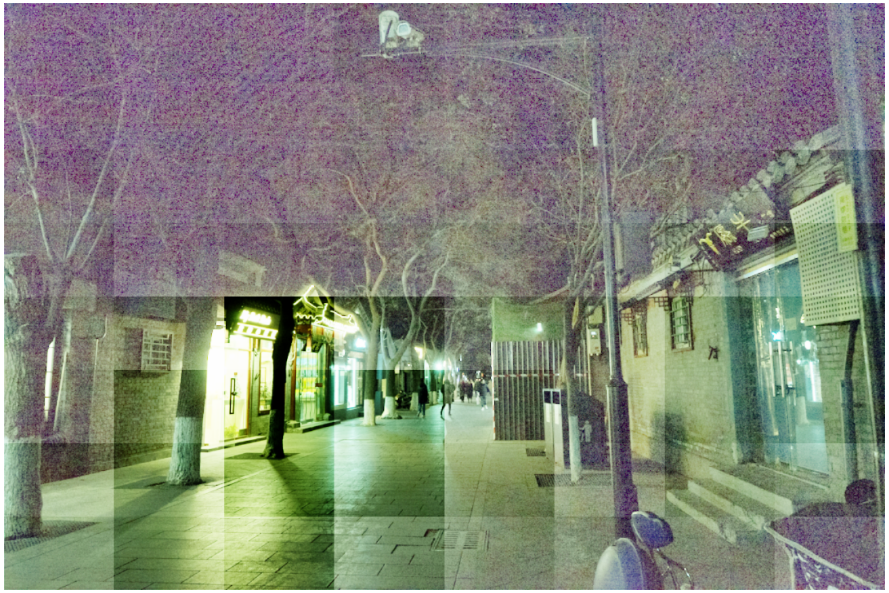
# Per-Block Enhancement: Idea & Implementation

- 分块提亮：把图片划分成几块，先将图片过一次 Zero-DCE，再把每个块提升到同一亮度水平。
- 如果人脸在分块边界，怎么办？
  - 使用四种错开的分块方式，这样人脸至少在其中一个方式中不在边界。
  - 需简单修改模型，使其支持输入四张图片。

# Per-Block Enhancement: Idea & Implementation

如何提亮到同一水平？

- 直方图均衡？
  - 没有梯度，且图片质量不佳。
- 每个像素值改为其 $\alpha$ 次方，令块内总和为定值。
  - 如何求出 $\alpha$ 的取值，同时保证求解过程有梯度？
  - 牛顿迭代。



# Per-Block Enhancement: Idea & Implementation

- 在 Baseline 基础上，Zero-DCE 提亮后加入分块提亮。
- 不同的分块方式会产生不同的图片，因此把 DSFD 的第一个卷积层拆开，分别处理四张图片。
- 最后在 DARKFACE 上 finetune ，或进行迁移训练。

# Per-Block Enhancement: Results

- 训练短暂一段时间后，得到了 0.51624 的 map，是还算不错的结果。

1 引入

2 尝试的思路
- Baseline
- Pseudo-Labeling
- Per-Block Enhancement
- Double Pass
- Conclusion

3 鲁棒性与可解释性

# Double Pass: Motivation

- Zero-DCE 的模型大小只有 313KB ，弱小可怜又无助。
- DSFD 的模型大小高达 189MB ，能不能让它帮 Zero-DCE 一把？
- Zero-DCE 并不知道提亮的目标是要识别人脸，于是对所有位置一视同仁。但如果附带一些额外的信息，或许能让它对可能的人脸位置特别关照。

# Double Pass: Idea

- Double Pass: 让图片经过模型两次，并在第二次附加上第一次的结果 mask 。
- 这样做的目的是，用 DSFD 强大的表达能力帮 Zero-DCE 提出原本无法得到的 feature ，并使得人脸所在位置的提亮程度变高。

# Double Pass: Results & Analysis

图 1: Baseline

图 2: Double Pass

- 然而，经过一段时间的训练，Loss 最终降到了与 Baseline 同一水平线之后就停滞不前。
- 观察两次提亮的图片，会发现几乎没有区别。
- 或许 Zero-DCE 已经尽力了。

1 引入

2 尝试的思路

Baseline

Pseudo-Labeling

Per-Block Enhancement

Double Pass

Conclusion

3 鲁棒性与可解释性

- 在这一节中，我们介绍了 Baseline 和基于它的三种优化。
- 实验表明，四者中效果最佳的是 Baseline 。
- 三种优化的优化幅度从 $0\%$ 到 $-13\%$ 不等。

# 参考资料

关于四个模型：


- Chunle Guo et al. "Zero-Reference Deep Curve Estimation for Low-Light Image Enhancement". In: CVPR (2020)
- Long Ma et al. "Toward Fast, Flexible, and Robust Low-Light Image Enhancement". In: CVPR (2022)
- Jian Li et al. "DSFD: Dual Shot Face Detector". In: CVPR (2019)
- Yang Liu et al. "MogFace: Towards a Deeper Appreciation on Face Detection". In: 2021


关于伪标签：


- Dong-Hyun Lee. "Pseudo-Label: The Simple and Efficient Semi-Supervised Learning Method for Deep Neural Networks". In: 2013

1 引入

2 尝试的思路

3 鲁棒性与可解释性

引入和概述

鲁棒性：对抗样本和攻防实验

可解释性：模型识别机制的分析

1 引入

2 尝试的思路

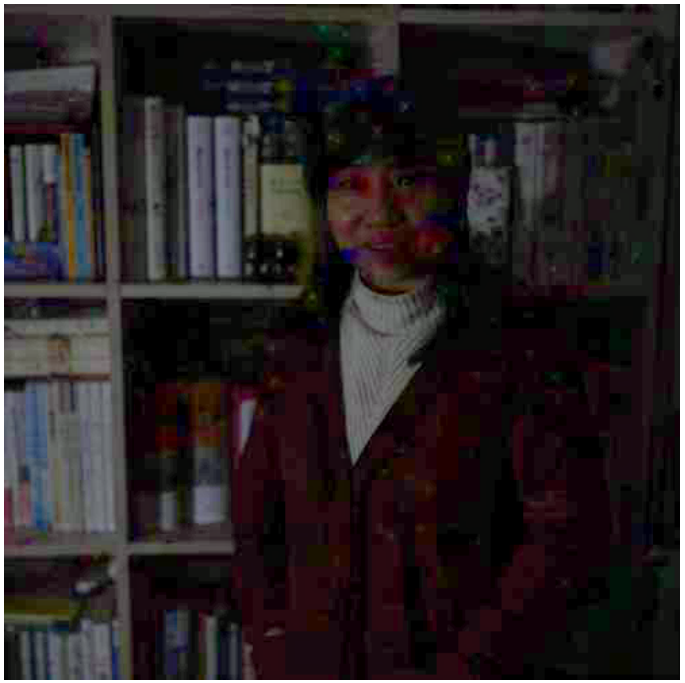
3 鲁棒性与可解释性

引入和概述

鲁棒性：对抗样本和攻防实验

可解释性：模型识别机制的分析

# 对抗样本



- 左图（已提亮）：模型以 $\approx 1$ 的置信度判定其中**有**人脸。
- 右图（已提亮）：模型以 $\approx 1$ 的置信度判定其中**无**人脸。
- 通过对图像进行较小的修改，可以彻底混淆模型的判断。

# 可解释性

- 神经网络是一个黑箱。
- 虽然我们用训练集为其指定了优化目标，但是：
  - 它有可能过拟合或欠拟合。
  - 训练集所指定的目标，与我们所期望的目标可能并不相同 (specification gaming)。
- 如果不能剖析其内部、了解其判断依据，我们很难信任神经网络的判断。

# 可解释性

一个故事：

- 九十年代美国曾有一个基于神经网络的模型，部署在一些医院里，为肺炎病人选择治疗方案。
- 研究者用可解释性工具分析后，发现该模型学到了一些令人大跌眼镜的规则，例如“肺炎与哮喘并发的患者是低风险的”。
- 究其原因，是因为训练集中只能体现相关性而不是因果性。所幸发现及时，否则后果不堪设想。

ref. B. Christian.
The Alignment Problem: Machine Learning and Human Values.

# Put it all together

- 对于复杂模型，必须关注其鲁棒性、可解释性、安全性、可靠性等。AI 研究也正在这些方向上加速推进。
- 我们的模型完美地符合了复杂、黑箱、不可解释这些特性。在本节中，我们试着打开黑箱，分析模型的鲁棒性和可解释性。

1 引入

2 尝试的思路

3 鲁棒性与可解释性

引入和概述

鲁棒性：对抗样本和攻防实验

可解释性：模型识别机制的分析

# Credit

攻防实验由两个小组合作完成：

- 视觉小班「随便」小组（负责 API 和防御）
- 可信机器学习小班李昂同学，以及他所在的小组（负责攻击）

# 对 Baseline 模型的攻击

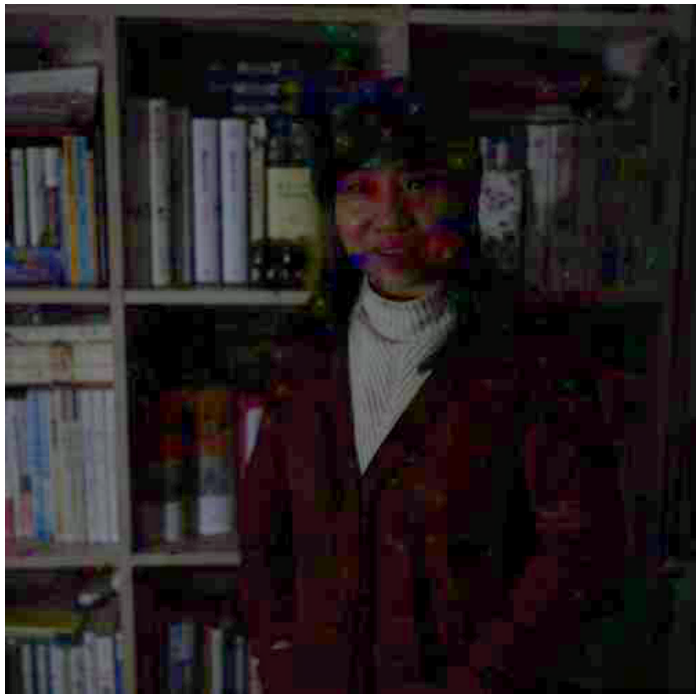


图 3: L − 2 攻击

图 4: L − ∞ 攻击

# 对 Baseline 模型的攻击

- 在攻击中发现，有时仅仅一次迭代就可让模型的置信度大跌 0.47。(总的迭代数在千这个量级)
- 这表明，我们模型的鲁棒性很成问题。

```
query_num = [1357.],loss = [0.88084]
query_num = [1358.],loss = [0.88084]
query_num = [1359.],loss = [0.04321]
query_num = [1360.],loss = [0.04321]
query_num = [1361.],loss = [0.04321]
```

# 防御思路一：去噪

- 你想给我加噪？
- 我把你的噪声去掉！

# 防御思路一：去噪

- 采用 BM3D 算法，对传进来的图片进行去噪后再做提亮和检测。
- 我们特意采用了 BM3D 这一传统算法，而非基于神经网络的去噪算法。这是因为神经网络鲁棒性差，易受攻击。
- 攻击算法采用的是逐渐调整加噪方式、逐渐增加加噪程度的思路。只要我们能消去少量的增量噪点，就能大大干扰攻击者的优化进程。

# 防御思路一：去噪

- 防御效果：起到了反作用。
- 在不同图片、不同攻击方法的四次实验中，该思路分别让攻击者的开销发生了以下变化：

  **第一次** $511 \to 81$
  **第二次** $341 \to 237$
  **第三次** $1359 \to 1461$
  **第四次** $1359 \to 1359$

- 猜测的原因：BM3D 在去噪的同时会对图片有所扰动。攻击程序可以利用这一点，让 BM3D 的扰动为它所用。

# 防御思路二：加噪

- 你想给我加噪？
- 我先把噪给你加上！

# 防御思路二：加噪

- 每次给输入的图随机加噪再喂进模型。
- 这样，最后输出的置信度里就会有一个随机扰动，从而可以干扰攻击者的优化进程。
- 为了保证一致性，用来生成噪声的随机种子取为输入的图的哈希值。

# 防御思路二：加噪

- 每次给输入的图随机加噪再喂进模型。
- 这样，最后输出的置信度里就会有一个随机扰动，从而可以干扰攻击者的优化进程。
- 为了保证一致性，用来生成噪声的随机种子取为输入的图的哈希值。

# 防御思路二：加噪

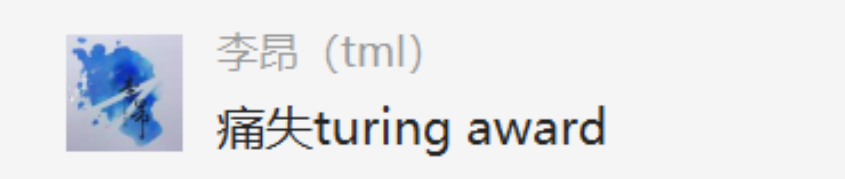
- 效果拔群！
- 在不同图片、不同攻击方法的四次实验中，该思路分别让攻击者的开销发生了以下变化：

  **第一次** $511 \to$ Fail
  **第二次** $341 \to$ Fail
  **第三次** $1359 \to$ Fail
  **第四次** $1359 \to$ Fail

# 防御思路二：加噪

事后我们找到了一篇一模一样的论文：

- Zeyu Qin et al. "Theoretical Study of Random Noise Defense against Query-Based Black-Box Attacks". In: NeurIPS. 2021



# 攻防回顾

- 我们实验了两种防御方法，其中一种效果优秀。
- 我们注意到，两种防御方法都在一定程度上降低了模型在正常数据上的表现。这说明，提高模型的鲁棒性，有时需要以性能的小幅下降为代价。但考虑到鲁棒性的重要，这大概是一个值得付出的代价。

1 引入

2 尝试的思路

3 鲁棒性与可解释性

引入和概述

鲁棒性：对抗样本和攻防实验

可解释性：模型识别机制的分析

# 分析工具：LIME

思路：

- 在给定样本的邻域，用一个简单模型来近似待分析的复杂模型。
- 然后转而对这个简单模型进行分析。

# 识别机制的多样性：寻找人脸和裸露肌肤

# 识别机制的多样性：寻找人脸和裸露肌肤

# 识别机制的多样性：寻找人体

# 识别机制的多样性：寻找背景边缘

# 失败模式的分析：寻找背景边缘

# 参考资料

关于基于加噪的防御：

- Zeyu Qin et al. “Theoretical Study of Random Noise Defense against Query-Based Black-Box Attacks”. In: NeurIPS. 2021

部分相关研究的概述：

- Dan Hendrycks et al. “Unsolved Problems in ML Safety”. In: ArXiv. 2021

*Thanks!*