arXiv:2302.03750v2 [cs.CV] 3 Dec 2023

将卷积核大小与人脸分析卷积神经网络中的泛化偏差联系起来

Hao Liang 莱斯
大学 休斯敦，德克萨斯州，美国
hl106@rice.edu

Josue Ortega Caro
耶鲁大学
纽黑文 , CO, USA
josue.ortegacaro@yale.edu

Vikram Maheshri
休斯敦大学
休斯敦 , TX, USA
vmaheshri@uh.edu

Ankit B. Patel 莱斯
大学 休斯敦，德克萨斯州，美国
ankit.patel@rice.edu

Guha Balakrishnan
莱斯大学 休斯敦，德克萨斯州，美国
guha@rice.edu

摘要

在解释神经网络的算法偏差时，训练数据集偏差无疑是受到最严格审视的因素。相比之下，与神经网络架构相关的超参数在很大程度上被忽视了，尽管众所周知，不同的网络参数化会对学习到的特征产生不同的 隐式偏差 。例如，众所周知，卷积核大小会影响卷积神经网络中学习到的特征的频率内容。本文提出了一个因果框架，用于将架构超参数与分布外算法偏差联系起来。我们的框架采用实验方法：通过对特定超参数进行干预，训练网络的多个版本，并测量在施加特定分布外图像扰动时，这一选择对性能偏差产生的因果效应。在实验中，我们重点测量了卷积核大小与人脸分析分类偏差之间的因果关系，并考察其在不同子群体（种族/性别）中的表现，以及图像高频细节的影响。我们表明，即使只修改卷积神经网络某一层中的卷积核大小，也会显著改变不同数据子群体中学习到的特征的频率内容，从而导致有偏的泛化性能，即使在使用均衡数据集的情况下也是如此。

1. 引言

深度神经网络的算法偏差，即数据分布中不同子组之间的性能差异，通常被归因于其训练数据集中的抽样偏差：数据分布中的某些群体所占比例显著低于或高于其他群体。尽管训练数据的不平衡确实会强烈影响深度网络的算法偏差 [1, 11, 25, 40, 49, 59], 但它只是开发流程中的一个因素。模型设计中其他重要因素的例子包括其参数化方式和训练目标函数。尽管此前已有多项研究 [4, 28, 33, 36, 41, 54] 探讨了目标函数对偏差的影响，并提出在训练过程中加入公平性正则项，但在偏差语境下，网络的参数化方式，即其特定的函数形式，几乎尚未得到研究。神经网络具有多个必须由设计者设定的参数化超参数，包括层数、激活函数，以及卷积神经网络（CNN）情况下的卷积核大小。这些选择中的每一项都会影响模型所学习的特征类型，进而可能影响偏差。在本研究中，我们迈出了第一步，展示如何将卷积神经网络的卷积核大小与其算法偏差建立因果联系。

已知不同的网络参数化会对学习到的特征产生不同的隐式偏差。例如，卷积神经网络往往会学习对输入信号的高空间频率特征敏感的表征[8, 51, 56]。基于傅里叶不确定性原理，这种现象可能归因于卷积核大小 [8] ——较小的卷积核大小会使特征覆盖傅里叶域中更大的范围。这种隐式特征偏差可以通过向测试图像中注入某种高频信号，并观察算法性能下降来揭示 [53]。隐式偏差通常不会损害模型在分布内测试样本上的表现，因为参数已经针对该分布的统计特性进行了良好调优。频率噪声/能量注入对于推动图像分布外（*OOD*）至关重要，从而将隐式偏差与网络的泛化性能联系起来。在这项工作中，我们展示了除这一影响之外，隐式偏差还会影响算法的

网络偏差，使得网络针对一个数据子群体使用的特征，可能与针对另一个数据子群体使用的特征具有显著不同的特性。这反过来可能导致这些群体的分布外性能存在差异。

我们工作的主要贡献是提出一个因果框架，用于关联神经网络的架构超参数与其分布外偏差。我们的框架基于一种实验性流程：每次修改神经网络架构中的一个或多个参数，并研究由此对分布外样本产生的偏差。首先，我们从头开始训练同一网络模型的多个版本，这些版本仅在超参数的选择上有所不同，例如卷积核大小。其次，我们通过向图像注入受控扰动来构建分布外测试集，并获取每个模型对每张图像的预测。扰动类型应根据所研究的隐式特征偏差来确定。例如，在我们的实验中，我们采用对抗攻击和在傅里叶通带中注入能量的方法来探查频率偏差。第三，我们拟合线性回归器，根据超参数选择、图像的扰动程度以及图像的各种属性，预测模型在分布外测试图像上的性能。第四，也是最后一步，我们利用回归系数，衡量该超参数对不同数据子群体的模型性能所产生的因果效应。此分析定量回答了该超参数是否会对不同数据子群体产生不同的因果效应。

虽然我们的框架具有通用性，但实验重点研究了卷积核大小变化所引起的对图像高频细节的敏感性，与人脸分析分类器在不同子群体（种族/性别受保护群体）上的性能之间的因果关系。我们在公开数据集上训练了多个研究级人脸性别分类器，并展示了将卷积核大小从常用范围 3×3 调整为 11×11，即使只调整这些卷积神经网络的第一层，也不仅会改变学习到的特征的频率内容，这种变化在不同种族/性别群体之间还可能存在显著差异。我们通过对测试图像的高频带施加对抗性扰动和能量注入，证实了这一效应。这项工作为进一步审慎研究神经网络设计决策对算法偏差的影响铺平了道路。

2. 相关工作

2.1. 计算机视觉中的公平性

计算机视觉领域的公平性研究，主要聚焦于衡量和缓解计算机视觉模型与数据集可能存在的偏置 [1,11,17,25,39,40,49,59]。偏置可以通过多种指标进行衡量 [9,18,50]，这些指标用于量化算法在不同人口子群体间的性能差异。人脸识别和人脸分析系统由于性质敏感，往往受到最严格的审视 [2,24,59]。这些研究中或许最著名的是“Gender Shades”研究 [6], ，该研究指出，人脸分析系统在特定种族与性别人口群体上存在系统性失效。

众所周知，自然图像数据集存在抽样偏差 [1,38,49], ，也就是说，其属性的联合分布远非随机。例如，众所周知，与男性相比，CelebA 人脸数据集中年轻女性的比例更高 [2]。在此类数据集上训练的模型可能会继承其偏差 [24, 38, 57]。因此，如果在更加均衡的数据集上训练算法，算法公平性问题便能得到显著缓解。人脸数据集尤其受到审视 [2,26,29,30,35], 因为在这些数据上训练的模型可能会在受法律保护的属性方面出现系统性失效 [27]。缓解数据集偏差的多种方法包括收集更多样化的样本 [35], 、使用图像合成来弥补分布差异 [2, 30, 43, 47, 52, 59],以及重采样 [31]。与此不同，我们的工作侧重于理解深度学习模型因参数化决策而产生的偏差，而非数据本身。此外，[10],[37], [42], [3] 和 [45] 提出了相关思路，探寻兼顾公平性的最优超参数，但研究重点并非计算机视觉任务和架构。

2.2. 对抗攻击

对抗攻击会对图像进行扰动，直到给定的网络改变其预测结果，通常通过对图像应用梯度下降来实现。图像因此产生的变化具有高频特征，肉眼难以察觉。这种鲁棒性的缺失引发了许多理论 [5, 13–16, 21, 34, 44, 48]。近期研究表明，最先进的卷积神经网络中常见的对抗样本包含特定于数据集的信息 [51]。此外，这些对抗攻击反映了模型所学习特征的属性 [8], , 也表明模型会因其架构选择而偏向某些特征 [12]。在这项工作中，我们分析攻击所携带的信息如何随不同的架构超参数选择而变化。此外，我们还探究一种新颖的假设：对抗攻击或许能让我们揭示模型特征在不同数据集子群体中捕获的可微信息。

2.3. CNN中的频率偏差

在图像处理中，表示像素位置最常见的方式是在空间域中使用列（x）、行（y）和 z（值）。频率（或傅里叶）域则提供了观察信号的另一种视角，即将其分解为不同频率的正弦波。

近期的一些研究为.....带来了新的见解。

卷积神经网络的行为提供了新的见解。例如，一项研究发现，与人类感知相比，高频成分在卷积神经网络的决策函数和性能中起着显著更大的作用 [51]。高频和低频图像成分的示例请参见图2。另一项研究表明，造成这一现象的一个因素是，卷积神经网络中的卷积运算会引入一种隐式偏差，使其倾向于使用图像中的较高频率 [8]。修正线性单元（ReLU）等非线性激活函数也可能是一个影响因素 [22]。我们基于这些研究的发现，探讨基于频率的特征对不同数据集子群（如性别和种族）的可微算法性能的影响。离散傅里叶变换（DFT）通常用于在图像的空间域和频域之间进行转换。

3. 方法

我们的目标是揭示卷积核大小因学习到的特征特性发生改变而对潜在算法偏差产生的因果效应。为此，我们提出了一个框架（概览见图 1 ）。该框架包含三个关键步骤。首先，我们训练 K 个版本的同一网络架构，这些版本之间唯一的区别是单个超参数的取值。超参数的选择构成了一种因果干预，因此我们采用的是实验性而非观察性的测试方法。接着，我们对一组测试图像施加分布外（OOD）扰动；这些图像标注了多种我们感兴趣的属性，包括将用于偏差分析的“受保护”属性（例如面部图像中的种族和性别）。最后，我们采用线性回归，根据超参数取值、图像属性和分布外扰动程度等协变量，预测与模型预测相关的某种度量。我们将回归系数视为各因素对模型产生因果效应的估计值，并具体比较与受保护属性相对应的系数差异，以评估偏差。我们将在下文介绍这三个步

p在以下各节中介绍我们的框架。

3.1. 架构训练

我们首先训练同一架构的 K 个不同版本，除了对所关注的超参数进行修改外，其结构完全相同。我们在同一训练数据集上训练所有架构。此外，我们从相同的正态分布（即均值和方差均相同）中初始化所有网络的权重和偏置。网络训练完成后，我们会“冻结”其参数，在我们的框架中不再对其进行修改。

3.2. 用于测试数据的分布外扰动

给定各种训练好的架构，我们的目标是放大学习到的特征中的偏差。一种方法是将来自与训练数据相同分布的测试样本输入这些网络，并测量不同受保护属性子群体的性能。这种策略的问题在于，深度神经网络是过参数化的，因此能够近乎完美地拟合任何训练数据分布。因此，即使某个超参数在不同网络之间有所改变，这两个网络在训练数据点上的性能也很可能相似。

然而，正如以往研究所展示的那样 [53, 56]，分布外（OOD）样本可能呈现出截然不同的情况：与其他模型相比，一些模型的性能会下降，从而暴露出所学习特征之间的差异。因此，我们模型中的一个关键步骤是向测试集中的图像注入一类细微的扰动，使其成为分布外样本。在实验中，我们关注卷积神经网络与频率相关的隐式偏差，因此考虑了神经网络文献中的两类扰动：对抗攻击和频率能量注入。

3.2.1 对抗攻击

我们在实验中考虑两种类型的对抗攻击。第一种，**FGSM** [16]，通过对网络输出相对于输入图像的损失函数进行梯度下降，以增量步骤“轻推”图像，使其朝着能够改变网络预测的方向移动。第二种，**CW攻击** [7]，使用两个独立的损失函数：一个基于梯度的损失函数，用于使分类器改变其预测（与 FGSM 类似）；另一个是正则化项，用于尽可能减小图像变化的幅度。因此，CW攻击的扰动距离（即 l_2 扰动图像与原始图像之差的范数）可作为衡量扰动图像难度的有效指标。补充材料中的图1展示了对同一输入图像进行 CW 攻击和 FGSM攻击的示例，进一步表明，由于正则化的作用，CW 扰动小了一个数量级。

3.2.2 频率能量注入

我们还尝试向特定频率频带注入能量，以更细粒度地建立频率内容与网络特征之间的联系。图 2 描述了我们的处理流程。对于每幅测试图像，我们使用 DFT 获得傅里叶频谱，并放大频谱中位于某个环带上的傅里叶系数的幅度。具体而言，设 $F[\omega_x, \omega_y] = |A|e^{-j\phi}$ 表示图像傅里叶频谱中位置为（ω_x，ω_y）处的一个复系数（分别对应 x 和 y 频率），其半径为 $r = \sqrt{\omega_x^2 + \omega_y^2}$，位于由（$r - r_0$）$^2 \leq \Delta^2$ 定义的环带内。我们将幅度 A 放大 $1 + \delta$ 倍，得到修改后的系数 $F'[\omega_x, \omega_y] = (1 + \delta)\,|A|e^{-j\phi}$。在实验中，我们设定 $\Delta = 2$，而 $\delta = 15$ 和 $r_0 > 0$ 是我们注入能量的频率半径范围 [56]。如果 r_0 较小（较大），我们修改的是图像的低频（高频）

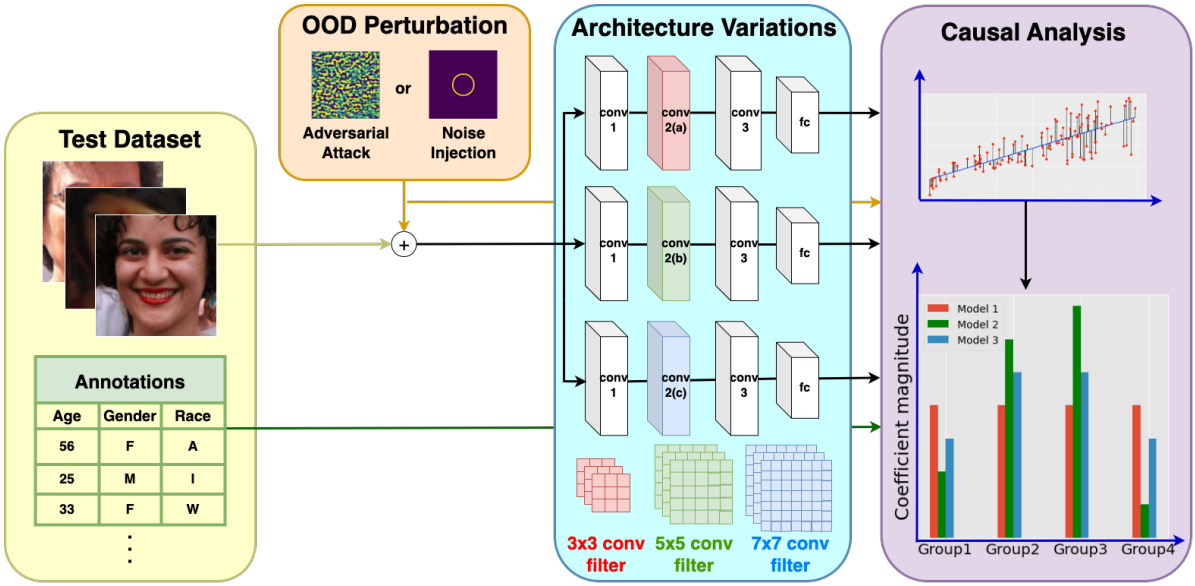

图1. **方法框架**。我们的方法框架由三个部分组成。首先，我们使用分布外（OOD）扰动（本文中使用对抗攻击扰动和频率噪声注入）对所有测试图像进行扰动，生成一个新的测试数据集，其中包含相同的图像，但注入了不同的噪声。其次，我们将新的测试图像输入不同的模型。这些模型的大部分架构设计相同，仅有少数部分不同（例如卷积核大小）。最后，我们收集上一步得到的结果，按敏感属性对其进行划分，并对结果进行因果分析。

成分。最后，我们使用逆离散傅里叶变换重建受扰图像。

3.3. 因果分析

我们将分布外图像测试集输入 K 网络，每张图像得到 K 个预测。我们假设每张图像还附有各种相关语义属性的标注（包括可用于计算算法偏差度量的受保护属性），以及扰动属性（例如能量注入频率）。我们的目标是衡量感兴趣的架构超参数（卷积核大小）对各受保护属性子群体模型性能的因果效应。

为此，我们使用多变量线性回归模型，根据多个自变量预测一个因变量。对于在网络 k中处理的测试图像 i，设 x_k 为相应的超参数，y_{ik} 为网络在图像 i 上的性能度量。那么，我们可以指定以下回归方程：

$$y_{ik} = \beta x_k + \epsilon_{ik}^0, \tag{1}$$

其中，ϵ_{ik} 为误差项。我们关注的系数是β。在假设 $E[x_k \cdot \epsilon_{ik}^0] = 0$成立的情况下，我们可以将 β 解释为网络架构对性能的因果效应。当然，这一独立性假设不太可能成立，因为图像属性（包括分布外扰动值）通常会影响神经网络的性能。

我们可以借助图像属性向量放宽这一假设 $\boldsymbol{Z_i}$，并按如下方式扩展方程（1）：

$$y_{ik} = \beta x_k + \boldsymbol{Z_i}'\boldsymbol{\gamma} + \epsilon_{ik}^1, \tag{2}$$

其中，$\boldsymbol{\gamma}$ 是一个系数向量，而 β 是在较弱假设 $E[x_k \cdot \epsilon_{ik}^1 | \boldsymbol{Z_i}] = 0$下网络架构对性能的因果效应。此外，我们假设架构超参数 x对性能的影响可能因受保护属性而异；受保护属性是$\mathbf{Z_i}$中所有图像属性的一个子集。为考虑这种可能性，我们进一步按如下方式扩展方程（2）：

$$y_{ik} = \boldsymbol{P_i}'\boldsymbol{\beta} x_k + \boldsymbol{Z_i}'\boldsymbol{\gamma} + \epsilon_{ik}^1, \tag{3}$$

其中 $\boldsymbol{P_i}$ 是受保护图像属性的向量，而$\boldsymbol{\beta}$ 现在是系数向量。我们采用一种启发式方法来选择向量 $\boldsymbol{Z_i}$ 和 $\boldsymbol{P_i}$，这也是社会科学中因果推断常用的方法 [55]。首先，我们逐步向 $\boldsymbol{Z_i}$ 中添加控制变量，并检验在不同设定下，我们对 $\boldsymbol{\beta}$ 的估计是否发生变化（使用 F检验，零假设为$\boldsymbol{\beta}$ 的估计值在不同设定下相等）。这是对外生性假设的检验；如果 $\boldsymbol{Z_i}$ 是足够丰富的控制变量向量，能够满足 $E[x_k \cdot \epsilon_{ik}^0 | \boldsymbol{Z_i}] = 0$，那么在给定扩充后的控制变量向量的条件下，该假设也将成立。其次，我们从一个丰富的向量 $\boldsymbol{P_i}$ 开始，以便灵活地估计网络架构对性能的影响。在我们的应用中，我们首先将 $\boldsymbol{P_i}$ 设定为一个完全饱和的虚拟变量向量，对应于所有受保护属性组合

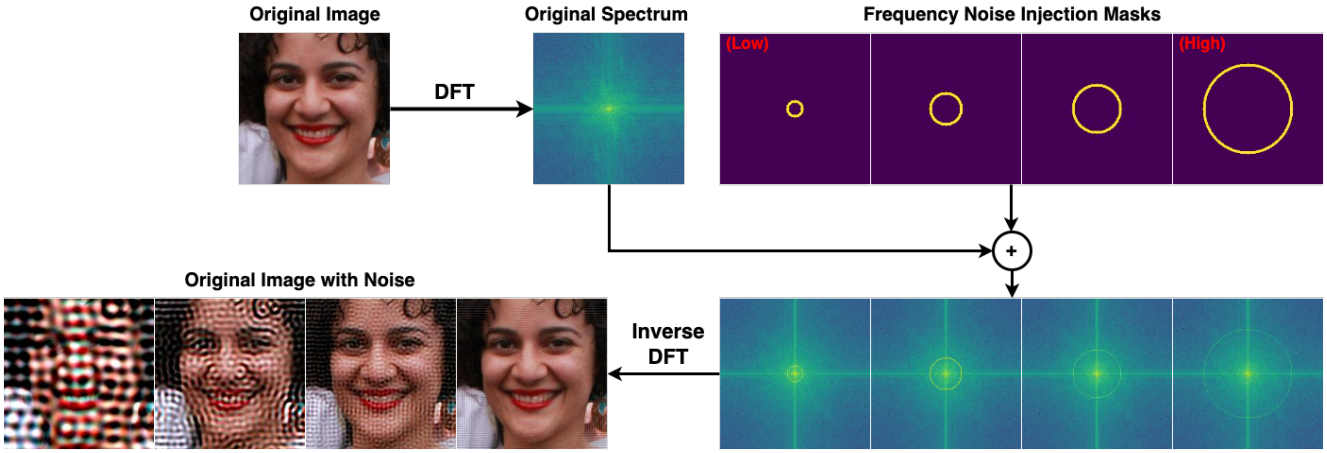

图 2. **频率能量注入扰动示例**。我们对输入图像施加这种扰动，旨在以相同幅度对所有频率进行“抖动”。为此，我们按照以下方式，将随机噪声注入傅里叶频谱中特定半径环带（环带顶行）上的图像傅里叶系数：$(1+\delta)|A|e^{-j\phi}$。

（例如，白人男性、白人女性等），并估计 β。然后，我们使用两两 F 检验，检验 β 的各个元素是否彼此相等。如果我们无法拒绝系数相等的假设，就无法拒绝网络架构对性能的影响在这两个群体之间存在差异这一假设。

β 编码了超参数值 x 和受保护属性对 $\boldsymbol{P_i}$ 输出 y 的联合因果效应。具体而言，β_g 表示当特征 y 为“True”（设为 1）时，x发生单位变化所导致的 g 的预期变化。在我们的实验中，我们比较了对应不同受保护属性子组的 β_g 值（参见补充材料中的表 2 和图 6）。

4. 实验与结果

我们使用两个常用数据集 Fairface[23]和 UTKFace [58]，评估了基于人脸图像进行性别分类的任务。Fairface 中不同种族/性别群体的样本数量大致相等，包含 86,744 个训练样本和 10,954 个测试样本。FairFace 包含 7 个种族群体（“东亚裔”“白人”“拉丁裔西班牙人”“东南亚裔”“黑人”“印度裔”“中东裔”）和 2 个性别群体（“男性”和“女性”）的标签。UTKFace 包含 20,000 个训练样本和 3,705 个测试样本，但种族群体分布不均衡。它包含 5 个种族群体（“白人”“黑人”“亚裔”“印度裔”“其他”）和 2 个性别群体（“男性”和“女性”）的标签。我们移除了“其他”类别的人脸，因为这些人脸不具备一致的特征。为减轻训练过程中的抽样偏差影响，我们根据各个种族群体的样本数量采用了逆向抽样。训练详情见补充材料的 A 节。

我们的实验展示了使用 ResNet-34 [19] 基础架构得到的结果，但使用另外两种流行的网络 DenseNet [20] 和 VGG-16 [46] 也获得了类似的结果。关于这两个模型的结果，请参阅补充材料中的 D、E 节。

我们关注的架构超参数是卷积核大小。我们考虑了两种不同的情形：仅改变第一层的卷积核大小，以及同时改变所有层的卷积核大小。有趣的是，这两种情形得出的结果相似，因此我们将后一种情形的结果放在补充材料的图6中。我们将首层卷积核大小（FLKS）设在 $[3, 11]$, 范围内，该范围涵盖了文献中几乎所有卷积神经网络常用的这一超参数取值。我们通过从方差设为 0。02 的正态分布中抽样，随机初始化所有模型的权重和偏置。对于每种网络和卷积核大小取值，我们训练了 3 个独立模型，并报告平均结果，以减轻随机初始化因素的影响。

我们报告了不同种族群体在非分布外测试图像上的网络准确率，详见补充材料中的表1，以表明它们在两个数据集上都取得了较高的准确率。由于训练和测试图像均来自同一分布，性能不会因 FLKS 而显著变化。现在，我们分别展示第3.2节所述两种分布外扰动的结果：对抗攻击和频率能量注入。

4.1. 对抗攻击

本节展示使用 CW 对抗攻击得到的结果。使用 FGSM 得到的结果也类似（见补充材料中的图8）。

4.1.1 分析傅里叶频谱

我们首先在图 3 中，按种族/性别群体和 FLKS 对对抗扰动图像进行划分，并可视化其平均傅里叶频谱幅值（Fairface 的结果）；UTKFace 的结果见补充材料中的图 7。两个数据集的结果呈现出相似趋势。首先，随着 FLKS 增大，

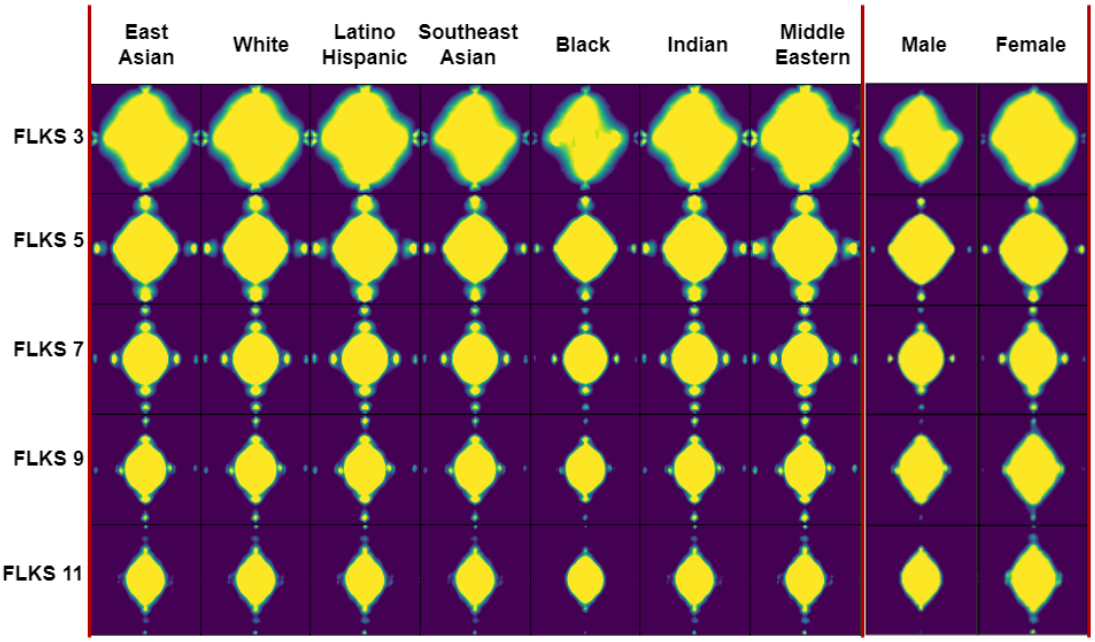

图3. **Fairface 中按种族和性别划分的对抗扰动图像平均频谱**。每一行代表一个首层卷积核大小（FLKS）不同的模型。随着 FLKS 增大，频谱会更加集中于低频。与其他种族群体的频谱相比，黑人种族组的频谱在高频处的能量始终较低。男性频谱的高频信息也少于女性频谱。这些结果表明，FLKS 的变化会使网络产生不同的特征偏差，且这些偏差也因受保护属性子群体而异。类似的 UTKFace 数据集频谱见补充材料中的图 7。

频谱能量越来越集中于低频（更靠近中心）。其次，在保持 FLKS 值不变的情况下，我们观察到，与其他种族群体的频谱相比，黑人群体的频谱始终包含更少的高频能量。与女性群体相比，男性群体也呈现出这一结果。随着 FLKS 增大，不同子组之间的差异逐渐缩小，这与先前研究的发现一致，即较低的 FLKS 会导致更高的隐式频率偏差 [8]。

为了定量评估扰动频谱的差异，我们还计算了 $f_{0.5}$ 这一指标，称为“半功率频率”，即信号功率有一半所处频率以下的频率。 $f_{0.5}$ 是衡量频谱能量集中程度的稳健指标。图4上方展示了 $f_{0.5}$ 频谱的得分，证实了图3中观察到的趋势。更多详情请参阅图4的图注。

4.1.2 扰动距离

接下来，我们在图 *4*下图中展示了不同种族群体和模型在对抗攻击下的平均扰动距离。扰动距离 d_p 可以通过简单地计算原始测试图像 I 与扰动图像 I' 之间的 L2 范数来求得： $d_p(I,\ I') = ||I' - I||_2$，它量化了扰动图像与原始图像之间的接近或偏离程度。距离越大，说明需要付出更多“努力”才能更难欺骗模型，同时也反映了模型对其他分布外扰动的鲁棒性 [32]。

结果表明，对于所有种族群体，扰动距离（及其方差）都会随着 FLKS 的增大而增加。因此，较大的 FLKS 会使模型更难受到对抗攻击，这可能是因为此类模型将更多能量集中于低频图像信息（见图 3），因而更加稳健。此外，我们看到 d_p 黑人群体的该值显著低于其他群体。更多详情请参见图 4 的图注。

4.1.3 因果分析

接下来，我们定量分析种族和性别对扰动距离的因果关系 d_p，应用第 3.3 节介绍的因果分析框架。具体而言，根据公式 3，我们将 y 设为 d_p，并将$\mathbf{P_i}$ 和 $\mathbf{Z_i}$ 均设为包含对应于所有种族/性别组合的“虚拟变量”。我们使用 Fairface 数据集进行此项分析，所采用的种族群体包括东亚裔、白人、拉丁裔西班牙人、东南亚裔、印度裔和黑人，性别群体包括男性和女性。

我们使用*Python* 的 *statsmodel* 软件包运行该回归，数值 β 和 γ 的结果见补充材料中的表 2。关于 β 的结果也显示在图6(a) 中。根据结果可以明显看出，黑人和印度裔的系数显著高于

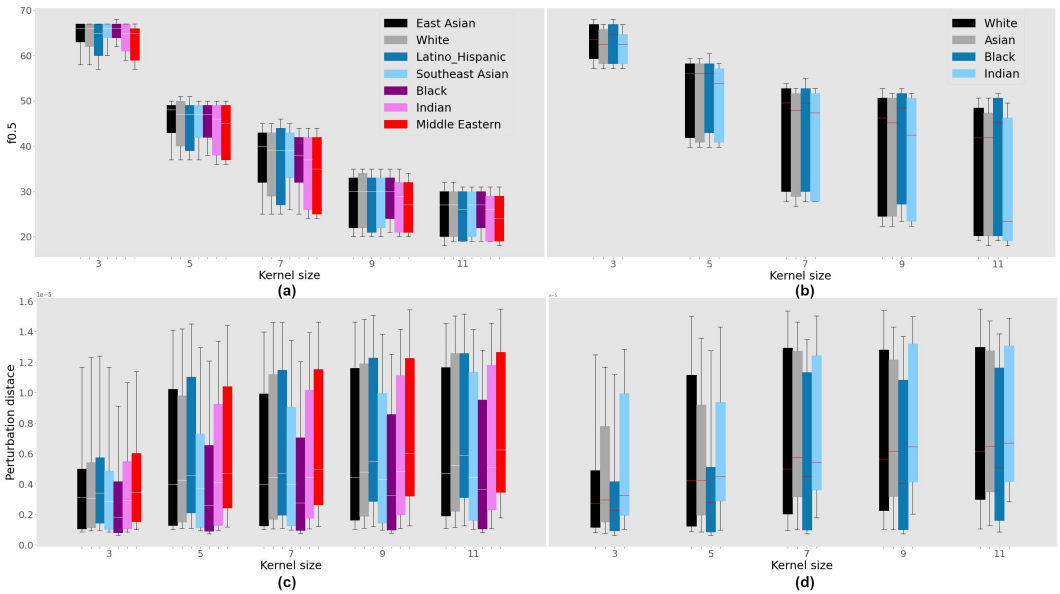

图 4. $f_{0.5}$ **对抗扰动频谱和对抗扰动距离的测量**每个箱线图显示不同受保护属性组的中位数得分（箱体中的白色/红色线）和 15% − 85% 置信区间。横轴表示模型的第一层卷积核大小（FLKS）。**（a）和（b）**分别显示使用 Fairface 和 UTKFace 时对抗扰动的 $f_{0.5}$ 测量结果。随着 FLKS 增大，所有人口统计组的 $f_{0.5}$ 得分均会下降，这表明对于较大的 FLKS，对抗攻击较少关注图像的高频信息。**（c）和（d）**分别显示使用 Fairface 和 UTKFace 时各个种族群体的对抗扰动距离，其中距离就是扰动图像的 l_2 范数。随着 FLKS 增大，所有人口统计组的扰动距离通常也会增大。此外，对于每个 FLKS 值，黑人群体的扰动距离始终显著低于其他人口统计组。

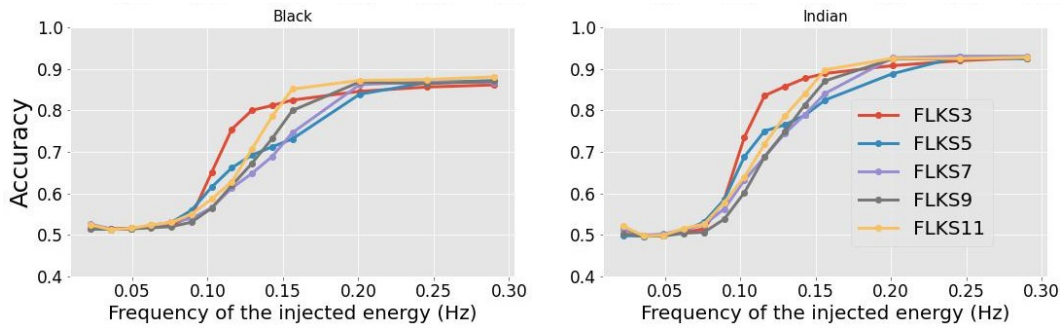

图5。**频率能量注入结果**。我们分别展示了针对黑人和印度裔种族群体、采用不同 FLKS 的模型性能（其他种族群体的结果请参见补充材料中的图9）。在每张图中，横轴表示注入能量的频率，纵轴表示不同模型的准确率。显然，所有模型都会受到低至中频能量注入的影响，而对中高频噪声具有鲁棒性。很难直接判断哪个群体受到的影响更大，因此需要进行定量分析。

其他种族群体，这表明卷积核大小对这两个群体扰动距离的影响显著更大。白人、黑人和印度裔女性的β值高于各自对应的男性群体。

4.2. 频率能量注入

接下来，我们按照第3.2.2 节所述，对测试图像施加基于频率的分布外扰动，并在图5中展示结果。所有模型/组的准确率

受低至中频扰动（$0.02--0.20$（Hz））的影响更大，而受中至高频扰动的影响较小。FLKS为3时，受（$0.08-0.15$（Hz））范围内频率注入的影响较小。不过，总体而言，仅凭这些图很难提炼出显著趋势。

4.2.1 因果分析

与第4.1.3节类似，我们现在进行回归分析，以衡量卷积核大小、能量注入频率和受保护属性子群体对模型误差的影响

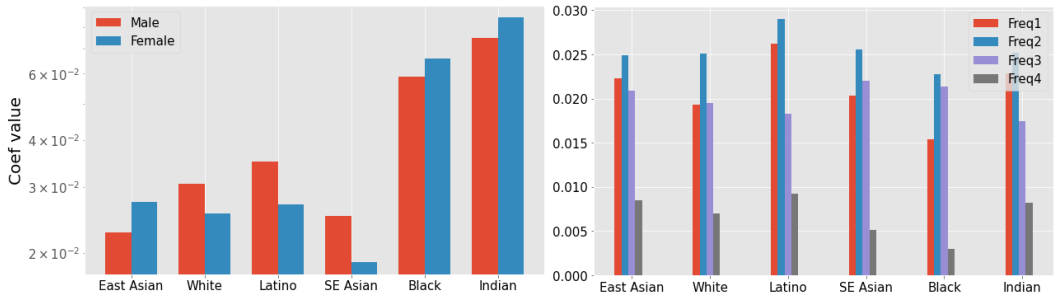

图6. **回归系数值（针对）β在不同种族群体中的情况。(a)** β第4.1.3 节中回归分析的值，该分析探讨了卷积核大小与对抗扰动距离之间的关系；相关内容也见补充材料中的表 2。受保护群体之间存在显著差异，例如，黑人和印度裔群体的值显著高于其他群体。 **(b)** β第4.2.1节中回归分析的值。黑人对应的系数始终低于其他种族群体。此外，同一种族群体内不同频率对应的系数值也存在显著差异。

（逐图像计算）。使用公式 3，我们将 y 设为图像的错误率，并将 P_i 和 Z_i 都设为包含种族／频率（注入能量）组合对应的“虚拟变量”。我们使用四个频率子群体：{(0.05, 0.07)、(0.09, 0.11)、(0.13, 0.15)、(0.17,0.19)(Hz)}，为方便起见，分别将其标记为 1、2、3、4。图 6(b) 展示了系数 β 的取值。显然，频率组 4 的系数显著小于其他频率组的系数，这表明在相对高频的注入条件下，卷积核大小的变化对分布外样本性能的影响较小。频率组 1 中黑人群体的系数也显著小于其他组。随着频率升高，这一差距逐渐缩小。

5. 讨论与结论

图3、4、5、6 首先表明，较小的卷积核大小会使卷积神经网络偏向高频特征，而增大卷积核大小能够减轻这种偏差。我们还发现，这种频率偏差在不同种族/性别子群体之间存在显著差异。在两个数据集上训练的所有模型，对黑人男性子群体的高频特征关注都较少。尽管特征不同并不一定意味着测试样本上存在性能偏差，但我们的结果表明，这些差异确实会导致分布外（OOD）样本上的性能偏差。我们观察到，两种不同类型的分布外图像扰动算子都会出现这种情况：对抗攻击和频域能量注入。

不同人口子群体会具有不同的图像特征。例如，黑人群体的肤色可能比其他种族群体更深，而女性平均而言比男性有更多头发。因此，不同种族的图像经网络处理后呈现出一些差异并不令人意外。然而，我们的结果表明，存在更为显著的差异：网络用于做出决策的图像特征在不同群体间具有根本不同的频率特性。这种差异也可能导致性能偏差，具体取决于模型所面对的分布外数据类型。我们的两种分布外扰动在概念上清晰且有充分依据，但并不对应任何真实现象。下一步，一个值得探讨的方向是将特征的频率偏差与模型在散粒噪声、雾和运动模糊等真实世界的分布外伪影上的性能差异联系起来 [56]。

我们的研究存在若干局限。由于我们只关注一个特定应用，因此无法对卷积神经网络在所有应用中的卷积核大小特性得出普遍结论。下一步的重要工作是针对更广泛的应用领域开展评估。我们还将因果分析限定在少数几个关键变量上。不过，因果分析通常依赖于“不存在隐藏混杂因素”假设。全面考察图像因素将有助于计算出更精确的因果效应。

本研究聚焦于网络的卷积核大小，因为以往研究已明确揭示这一超参数与频率内容之间的联系 [8]。然而，我们的框架不受超参数性质的限制。事实上，该研究领域的下一步工作包括对更全面的一组网络超参数进行类似分析，例如激活函数、网络层数、权重初始化策略，乃至高层设计（如残差连接、Transformer 模块）。我们将本研究视为迈向重要目标的第一步：理解神经网络设计选择如何影响偏差，进而影响这些系统在我们社会中的公平性。

参考文献

[1] Vitor Albiero、Krishnapriya Ks、Kushal Vangara、Kai Zhang、Michael C King 和 Kevin W Bowyer。人脸识别准确率中的性别不平等分析。载于IEEE/CVF计算机视觉应用冬季会议研讨会论文集，第 81–89 页，2020年。1，2[2] Guha Balakrishnan、Yuanjun Xiong、Wei Xia 和 Pietro Perona。迈向人脸分析算法偏差的因果基准评测。载于《基于深度学习的人脸分析》，第 327–359 页。施普林格，2021年。2[3] Ioana Baldini、Dennis Wei、Karthikeyan Natesan Ramamurthy、Mikhail Yurochkin 和 Moninder Singh。公平性可能因人而异：有害文本分类中预训练语言模型的公平性。arXiv 预印本 arXiv:2108.01250，2021年。2[4] Alex Beutel、Jili n Chen、Tulsee Doshi、Hai Qian、Li Wei、Yi Wu、Lukasz Heldt、Zhe Zhao、Lichan Hong、Ed H. Chi 和 Cristos Goodrow。通过成对比较实现推荐排序公平性。载于 ACM 知识发现与数据挖掘国际会议（KDD），2019年。1[5] S´ebastien Bubeck、Eric Price 和 Ilya Razenshteyn。计算约束下的对抗样本。arXiv 预印本 arXiv:1805.10204，2018年。2[6] Joy Buolamwini 和 Timnit Gebru。Gender Shades：商业性别分类中的交叉性准确率差异。载于公平性、问责制与透明度会议，2018年。2[7] Nicholas Carlini 和 David Wagner。迈向神经网络鲁棒性的评估。载于2017年 IEEE安全与隐私研讨会（SP），第 39–57 页。IEEE，2017年。3[8] Josue Ortega Caro、Yilong Ju、Ryan Pyle、Sourav Dey、Wieland Brendel、Fabio Anselmi 和 Ankit Patel。局部卷积会导致对高频对抗样本的隐式偏差。arXiv 预印本 arXiv:2006 11440，2020年 1，2，3，6，8 . . [9] Sam Corbett-Davies 和 Sharad Goel。公平性的衡量与误测：对公平机器学习的批判性综述。arXiv 预印本 arXiv:1808.00023，2018年。2[10] Andr´e F Cruz、Pedro Saleiro、Catarina Bel´em、Carlos Soares 和 Pedro Bizarro。通过超参数优化促进公平性。载于2021年 IEEE数据挖掘国际会议（ICDM），第 1036–1041 页。IEEE，2021年。2[11] Mengnan Du、Fan Yang、Na Zou 和 Xia Hu。深度学习中的公平性：计算视角。IEEE Intelligent Systems，36(4)：25–34，2020年。1，2[12] Fartash Faghri、Sven Gowal、Cristina Vasconcelos、David J Fleet、Fabian Pedregosa 和 Nicolas Le Roux。弥合对抗鲁棒性与优化偏差之间的差距。arXiv 预印本 arXiv:2102.08868，2021年。2[13] Alhussein Fawzi、Hamza Fawzi 和 Omar Fawzi。任意分类器的对抗脆弱性。载于《神经信息处理系统进展》，第 1178–1187 页，2018年。2[14] Nic Ford、Justin Gilmer、Nicolas Carlini 和 Dogus Cubuk。对抗样本是噪声中测试误差的自然结果。arXiv 预印本 arXiv:1901.10513，2019年。2 [15] Justin Gilmer、Luke Metz、Fartash Faghri、Samuel S Schoen- holz、Maithra Raghu、Martin Wattenberg 和 Ian Goodfel- low。对抗球体。arXiv预印本 arXiv:1801 02774，2018 年 2 . . [16] Ian J Goodfellow 、Jonathon Shlens 和 Christian Szegedy。解释并利用对抗样本。arXiv预印本 arXiv:1412.6572，2014年。2，3[17] Patrick Grother、Mei Ngan 和 Kayee Hanaoka。人脸识别供应商测试（fvrt）：第3部分，人口统计学影响。美国国家标准与技术研究院，马里兰州盖瑟斯堡，2019年。2[18] Moritz Hardt、Eric Price 和 Nati Srebro。监督学习中的机会平等。神经信息处理系统进展，第29卷，2016年。2 [19] 何恺明、张祥雨、任少卿和孙剑。用于图像识别的深度残差学习。载于《IEEE计算机视觉与模式识别会议论文集》，第770–778页，2016年。5[20] 黄高、刘壮、Laurens Van Der Maaten 和 Kil- ian Q Weinberger。密集连接卷积网络。载于《IEEE计算机视觉与模式识别会议论文集》，第4700–4708页，2017年。5[21] Andrew Ilyas、Shibani Santurkar、Dimitris Tsipras、Logan Engstrom、Brandon Tran 和 Aleksander Madry。对抗样本不是缺陷，而是特征。载于《神经信息处理系统进展》，第125 136页，2019年 2 – . [22] Nikos Karantzas、Emma Besier、Josue Ortega Caro、Xaq Pitkow、Andreas S Tolias、Ankit B Patel 和 Fabio Anselmi。通过傅里叶掩码理解人工神经网络的鲁棒性与泛化。《人工智能前沿》，第5卷，2022年。3[23] Kimmo K¨arkk¨äinen 和 Jungseock Joo。Fairface：用于平衡种族、性别和年龄的人脸属性数据集。arXiv预印本 arXiv:1908.04913，2019 年。5[24] Kimmo Karkkainen 和 Jungseock Joo。Fairface：用于偏差测量与缓解、平衡种族、性别和年龄的人脸属性数据集。载于《IEEE/CVF计算机视觉应用冬季会议论文集》，第1548–1558页，2021年。2[25] Michael Kearns、Seth Neel、Aaron Roth 和 Zhiwei Steven Wu。防止公平性划分操纵：面向子群体公平性的审计与学习。载于《国际机器学习会议》，第2564–2572页。PMLR，2018年。1，2[26] Brendan F Klare、Mark J Burge、Joshua C Klontz、Richard W Vorder Bruegge 和 Anil K Jain。人脸识别性能：人口统计信息的作用。《IEEE信息取证与安全汇刊》，7(6):1789–1801，2012年。2[27] Jon Kleinberg、Jens Ludwig、Sendhil Mullainathan 和 Cass R Sunstein。算法时代的歧视。《法律分析期刊》，10:113–174，2018年。2[28] Matth¨aus Kleindessner、Samira Samadi、Muhammad Bilal Zafar、Krishnaram Kenthapadi 和 Chris Russell。序数回归的成对公平性。国际人工智能与统计会议（AISTATS），2022年。1[29] Adam Kortylewski、Bernhard Egger、Andreas Schneider、Thomas Gerig、Andreas Morel-Forster 和 Thomas Vetter。

实证分析数据集偏差对深度人脸识别系统的影响。载于《IEEE计算机视觉与模式识别会议研讨会论文集》，2018年。2[30] Adam Kortylewski、Bernhard Egger、Andreas Schneider、Thomas Gerig、Andreas Morel-Forster 和 Thomas Vet- ter。利用合成数据分析并减少数据集偏差对人脸识别造成的损害。载于《IEEE/CVF计算机视觉与模式识别会议研讨会论文集》，2019年。2[31] Yi Li 和 Nuno Vasconcelos。Repair：通过数据集重采样消除表征偏差。载于《IEEE计算机视觉与模式识别会议》（CVPR），2019年。2[32] Zhe Li、Josue Ortega Caro、Evgenia Rusak、Wieland Bren- del、Matthias Bethge、Fabio Anselmi、Ankit B Patel、An- dreas S Tolias 和 Xaq Pitkow。鲁棒深度学习物体识别模型依赖自然图像中的低频信息。bioRxiv，2022年。6[33] Michael Lohaus、Michael Perrot 和 Ulrike von Luxburg。放松过度，难以实现公平。载于《国际机器学习会议》（ICML），2020年。1[34] Saeed Mahloujifar、Dimitrios I Diochnos 和 Mohammad Mahmoody。鲁棒学习中的集中现象之祸：由测度集中导致的规避与投毒攻击。载于《AAAI人工智能会议论文集》，第33卷，第4536–4543页，2019年。2[35] Michele Merler、Nalini Ratha、Rogerio S Feris 和 John R Smith。人脸多样性。arXiv预印本 arXiv:1901 10436，2019年 2 . . [36] Manisha Padala 和 Sujit Gujar。FNNC：通过神经网络实现公平性。载于《国际人工智能联合会议组织》（IJCAI），2020年。1[37] Valerio Perrone、Michele Donini、Muhammad Bilal Zafar、Robin Schmucker、Krishnaram Kenthapadi 和 C´edric Ar- chambeau。公平的贝叶斯优化。载于《2021年AAAI/ACM人工智能、伦理与社会会议论文集》，第854–863页，2021年。2[38] Jean Ponce、Tamara L Berg、Mark Everingham、David A Forsyth、Martial Hebert、Svetlana Lazebnik、Marcin Marsza- lek、Cordelia Schmid、Bryan C Russell、Antonio Torralba 等。物体识别中的数据集问题。载于《迈向类别级物体识别》，第29–48页。施普林格，2006年。2[39] Novi Quadrianto、Viktoriia Shar manska 和 Oliver Thomas。在数据域中发现公平表征。载于《IEEE/CVF计算机视觉与模式识别会议论文集》，第8227–8236页，2019年。2[40] Vikram V Ramaswamy、Sunnie SY Kim 和 Olga Rus- sakovsky。通过潜在空间去偏实现公平属性分类。载于《IEEE/CVF计算机视觉与模式识别会议论文集》，第9301 9310页，2021年 1，2 – . [41] Laurent Risser、Quentin Vincenot 和 Jean-Michel Loubes。使用Wasserstein-2正则化应对神经网络分类器中的算法偏差，2020年。1[42] Kit T Rodolfa、Hemank Lamba 和 Rayid Ghani。对公平性与准确率权衡可忽略不计的实证观察，

面向公共政策的机器学习。《自然机器智能》，3(10):896–904，2021年。2[43] Prasanna Sattigeri、Samuel C Hoffman、Vijil Chenthama-rakshan 和 Kush R Varshney。公平性生成对抗网络：利用生成对抗网络生成具有公平性属性的数据集。《IBM研究与开发期刊》，63(4/5):3–1，2019年。2[44] Ludwig Schmidt、Shibani Santurkar、Dimitris Tsipras、Kunal Talwar 和 Aleksander Madry。对抗鲁棒泛化需要更多数据。载于《神经信息处理系统进展》，第5014–5026页，2018年。2[45]Thibault Sellam、Steve Yadlowsky、Jason Wei、Naomi Saphra、Alexander D'Amour、Tal Linzen、Jasmijn Bastings、Iulia Turc、Jacob Eisenstein、Dipanjan Das 等。多版本 BERT：用于鲁棒性分析的 BERT 复现。arXiv 预印本 arXiv:2106.16163，2021年。2[46] Karen Simonyan 和 Andrew Zisserman。用于大规模图像识别的超深卷积网络。arXiv 预印本 arXiv:1409.1556，2014年。5[47] Chandan Singh、Guha Balakrishnan 和 Pietro Perona。利用生成对抗网络进行匹配样本选择，以缓解属性混杂。arXiv 预印本 arXiv:

2103 13455, 2021 2 。[48]

Thomas Tanay 和 Lewis Griffin。对抗样本现象的边界倾斜视角。arXiv 预印本 arXiv:1608.07690，2016年。2[49] Antonio Torralba 和 Alexei A Efros。审视数据集偏差的无偏视角。载于 CVPR 2011，第1521–1528页。IEEE，2011年。1, 2[50] Sriram Vasudevan 和 Krishnaram Kenthapadi。LIFT：用于衡量机器学习应用公平性的可扩展框架。载于第29届 ACM信息与知识管理国际会议论文集，第2773 2780页，2020年 2 –。[51] Haohan Wang、Xindi Wu、Zeyi Huang 和 Eric P Xing。高频分量有助于解释卷积神经网络的泛化。载于IEEE/CVF计算机视觉与模式识别会议论文集，第8684–8694页，2020年。1, 2, 3[52] Zeyu Wang、Klint Qinami、Ioannis Karakozis、Kyle Genova、Prem Nair、Kenji Hata 和 Olga Russakovsky。迈向视觉识别中的公平性：缓解偏差的有效策略。载于IEEE计算机视觉与模式识别会议（CVPR），2020年。2[53] Zeyu Wang、Klint Qinami、Ioannis Christos Karakozis、Kyle Genova、Prem Nair、Kenji Hata 和 Olga Russakovsky。迈向视觉识别中的公平性：缓解偏差的有效策略。载于IEEE/CVF计算机视觉与模式识别会议论文集，第8919–8928页，2020年。1, 3[54] Michael Wick、Swetasudha Panda 和 Jean-Baptiste Tristan。揭示公平性：重新审视权衡。载于神经信息处理系统（NeurIPS），2019年。1[55] Jeffrey M Wooldridge。《横截面数据与面板数据的计量经济学分析》。麻省理工学院出版社，

2010 4 。Dong Yin、Raphael [56] Gontijo Lopes、Jonathon Shlens、Ekin D. Cubuk 和 Justin Gilmer。计算机视觉中模型鲁棒性的傅里叶视角。Curran Associates Inc.，美国纽约州雷德胡克，2019年。1, 3, 8

[57] Rich Zemel、Yu Wu、Kevin Swersky、Toni Pitassi 和 Cynthia Dwork。公平表征学习。载于 国际机器学习会议，第325–333页。PMLR，2013年。 2[58] Zhifei Zhang、Yang Song 和 Hairong Qi。通过条件对抗自编码器实现年龄增长/回归。载于*IEEE*计算机视觉与模式识别会议（*CVPR*）。IEEE，2017年。 5[59]Dominik Zietlow、Michael Lohaus、Guha Balakrishnan、Matth¨aus Kleindessner、Francesco Locatello、Bernhard Sch¨olkopf 和 Chris Russell。计算机视觉中的降格：公平深度分类器中的帕累托低效。载于*IEEE/CVF*计算机视觉与模式识别会议论文集，第10410–10421页，2022年。 1，2

补充材料

A. 实验设置

A.1. 训练详情

模型采用多任务方式进行训练。训练期间，模型需要预测训练样本的种族和性别标签，而在推理时，我们只使用模型预测的性别标签。训练超参数方面，我们使用了批次大小为 128，初始学习率为 $1e^{-3}$，并在第 13 和第 17 个训练轮次时将其降低 10 倍。所有模型均训练了 21 个训练轮次。我们使用了 Adam 优化器。

A.2. 计算资源

所有实验均使用 NVIDIA Tesla A100 GPU 运行。每次运行训练一个模型需要 1 GPU小时。对 FairFace 测试数据集应用 FGSM攻击耗时 ~ 1，000 秒，对 FairFace 测试数据集应用 CW攻击耗时 ~ 6， 000 秒。

B. 模型性能

我们在表格 1 中报告了各网络在非分布外测试图像上针对不同种族群体的准确率，以说明它们在两个数据集上都取得了相当高的准确率。由于训练和测试图像均来自同一分布，性能不会随 FLKS 显著变化。

表1。**Fairface 和 UTKFace 数据集上未受扰动的（非分布外）图像的模型性能**。我们报告了训练好的模型在按种族群体划分的测试集上的性能。每个数值都是 3 个不同训练模型的平均值。第一列标明了数据集名称和模型的第一层卷积核尺寸（FLKS）。Fairface 有 7 个标注的种族群体，UTKFace 有 4 个。由于测试图像和训练图像属于同一分布，因此性能相对稳定，随卷积核大小变化不大。

数据集（FLKS）	总体	白人	黑人	东亚裔	印度裔	东南亚裔	拉丁裔	中东人
Fairface (3)	0.947	0.950	0.894	0.942	0.945	0.894	0.957	0.977
Fairface（5）	0.949	0.947	0.896	0.939	0.957	0.896	0.957	0.980
Fairface（7）	0.946	0.943	0.895	0.947	0.956	0.895	0.963	0.978
Fairface（9）	0.947	0.946	0.895	0.937	0.951	0.895	0.960	0.979
Fairface（11）	0.946	0.949	0.892	0.937	0.949	0.885	0.967	0.979
UTKFace（3）	0.929	0.949	0.905	0.931	0.942	/	/	/
UTKFace（5）	0.935	0.951	0.901	0.940	0.951	/	/	/
UTKFace（7）	0.934	0.955	0.905	0.939	0.953	/	/	/
UTKFace（9）	0.937	0.955	0.910	0.941	0.955	/	/	/
UTKFace（11）	0.936	0.950	0.901	0.943	0.955	/	/	/

C. 对抗攻击示例

我们在补充材料中展示了同一输入图像的 CW 攻击和 FGSM 攻击示例，这进一步表明，由于正则化的作用，CW 扰动小了一个数量级。

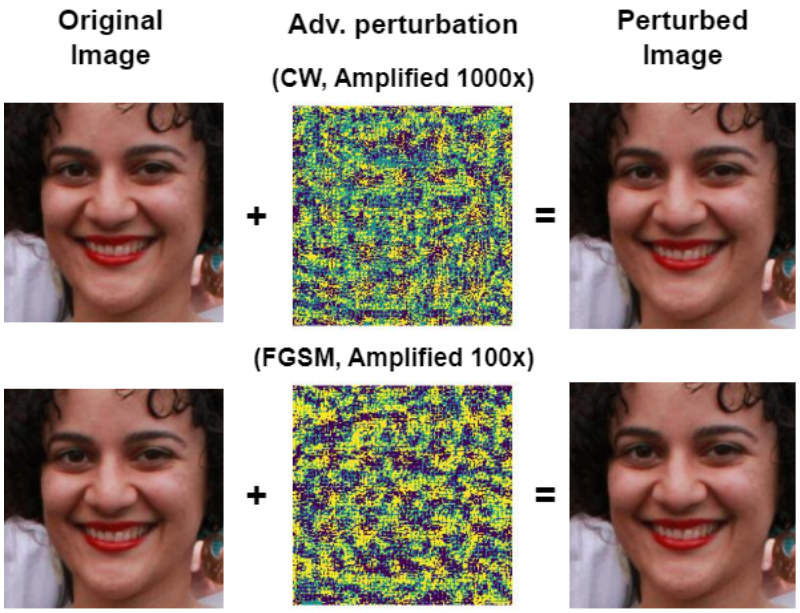

图 7. **对抗攻击扰动示例**。通过向测试图像（左）添加微小的类噪声扰动（中间图像，为便于观察放大了 100/1000 倍），目标神经网络会输出错误预测。然而，扰动图像（右）在人眼看来与原始图像没有可察觉的差异。我们的实验使用了 CW 和 FGSM 攻击。

D. DenseNet121 的结果

为了进一步检验我们框架和结论的鲁棒性与普适性，我们还在 DenseNet121——另一种流行的人脸分析模型上进行了测试。我们还将第一个卷积核大小从 $\{3, 5, 7, , 9, 11\}$进行了变化。图 8展示了平均频谱，其对应的扰动距离 & $f_{0.5}$ 分数见图 9。

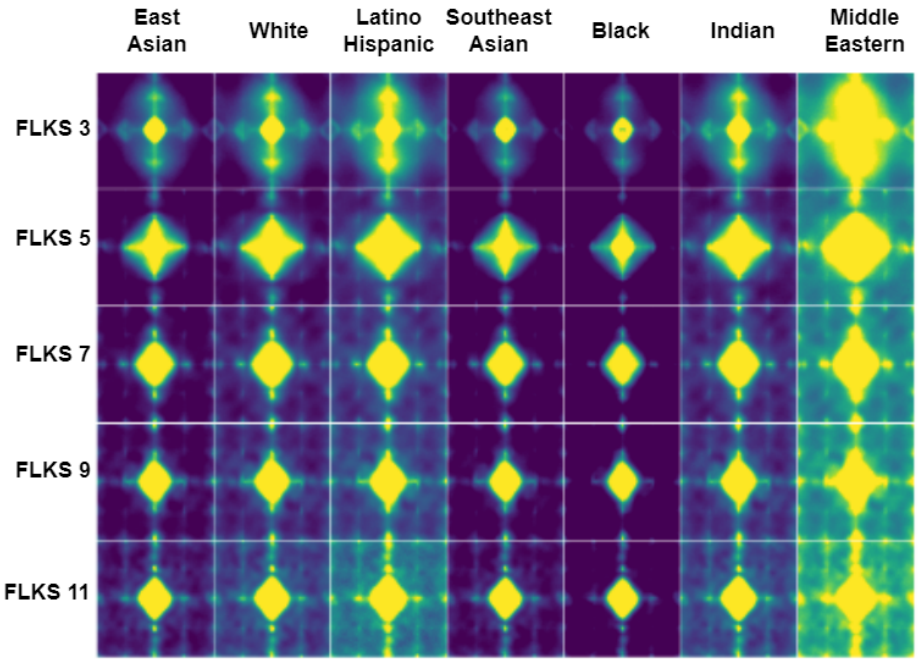

图8. **使用 DenseNet121 在 Fairface 上进行扰动谱可视化。** 与正文图3中的结果类似，每行代表一个第一层卷积核大小（FLKS）不同的模型，每列对应一个受保护属性组。我们观察到与上述另外两项结果类似的趋势：总体而言，随着 FLKS 增大，扰动会将关注点转向低频信息；黑人对应的扰动的高频关注度始终低于其他种族群体。

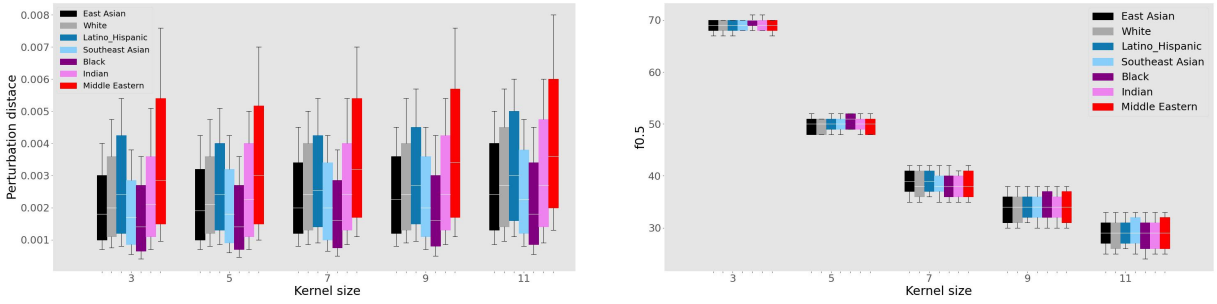

图9. **扰动频谱 $f_{0.5}$ 与 DenseNet121 的扰动距离。** 我们按照第4.1节和图4中所述的相同方式，将扰动和 $f_{0.5}$ 分数可视化。我们观察到类似的趋势：对于所有人口统计组，$f_{0.5}$ 都会随着 FLKS 增大而显著下降；同时，随着 FLKS 增大，所有人口统计组的扰动距离总体上也会增加。

E. Vgg16 上的结果

为了进一步检验我们框架和结论的鲁棒性与普适性，我们还在 Vgg16——另一种流行的人脸分析模型上进行了测试。我们还将第一个卷积核大小从 $\{3, 5, 7, , 9, 11\}$进行了变化。图 10展示了平均频谱，其对应的扰动距离 & $f_{0.5}$ 分数见图 11。

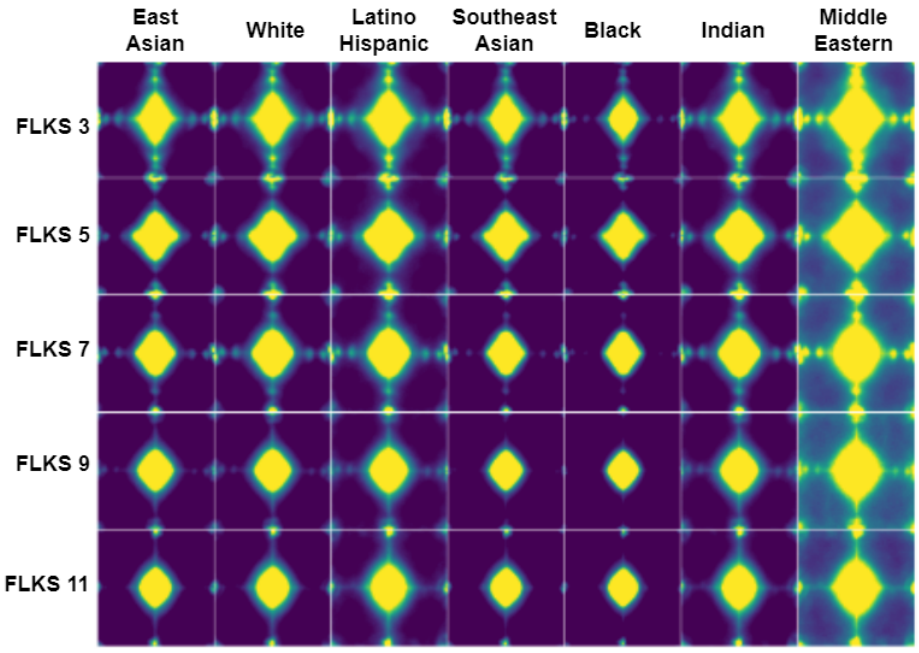

图10. **使用 Vgg16 在 Fairface 上进行的扰动谱可视化**。 与正文图3中的结果类似，每行代表卷积核大小（FLKS）不同的模型，每列对应一个受保护属性组。我们观察到的趋势与上述另外两项结果相似：总体而言，随着 FLKS 增大，扰动会将关注点转向低频信息；与其他种族群体相比，针对黑人的扰动始终具有较低的高频关注度。

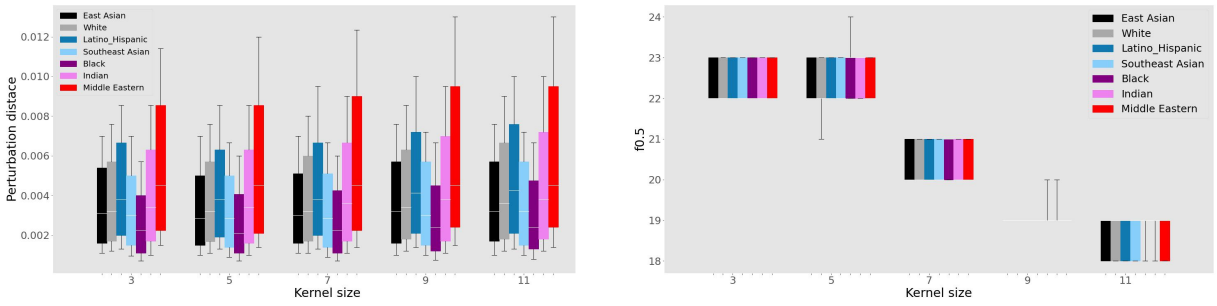

图11. **扰动频谱 $f_{0.5}$ 与 Vgg16 的扰动距离**。 我们按照第4.1节和图4中所述的相同方式，对扰动和 $f_{0.5}$ 分数进行可视化。我们观察到类似的趋势：对于所有人口统计组，随着 FLKS 增大， $f_{0.5}$ 会显著下降；同时，随着 FLKS 增大，所有人口统计组的扰动距离通常也会增加。

F. 回归结果

我们报告第 4.1.3 节所述回归的系数 $\boldsymbol{\beta}$（左）和 $\boldsymbol{\gamma}$（右）。我们还报告系数值的标准差，并根据 $t = \frac{coefvalue}{stderr}$ 计算其 t 值，以及 $P > |t|$。$P \leq 0.05$表示该值具有显著性。————

表 2. **扰动距离的回归结果**。β 系数名称采用下标，分别对应种族（E：东亚裔，W：白人，B：黑人，I：印度裔，L：拉丁裔，S：东南亚裔）和性别（M：男性，F：女性）。不同群体的 $\boldsymbol{\beta}$ 值之间存在一些显著差异，例如，黑人亚裔女性的值比东亚裔女性高 $\sim 100\%$。有关 $\boldsymbol{\beta}$的图表，请参阅主论文中的图6。

系数名称	系数值	标准误	t	$P > \|t\|$	系数名称	系数值	标准误	t	$P > \|t\|$
$\boldsymbol{\beta}_{EM}$	0.0227	0.003	6.930	0.000	$\boldsymbol{\gamma}_{EM}$	0.4251	0.025	17.308	0.000
$\boldsymbol{\beta}_{EF}$	0.0274	0.003	8.323	0.000	$\boldsymbol{\gamma}_{EF}$	0.4232	0.025	17.065	0.000
$\boldsymbol{\beta}_{WM}$	0.0306	0.003	11.392	0.000	$\boldsymbol{\gamma}_{WM}$	0.4453	0.020	21.969	0.000
$\boldsymbol{\beta}_{WF}$	0.0254	0.003	8.614	0.000	$\boldsymbol{\gamma}_{WF}$	0.4302	0.022	19.361	0.000
$\boldsymbol{\beta}_{LM}$	0.0351	0.003	11.143	0.000	$\boldsymbol{\gamma}_{LM}$	0.2088	0.012	17.566	0.000
$\boldsymbol{\beta}_{LF}$	0.0269	0.003	8.521	0.000	$\boldsymbol{\gamma}_{LF}$	0.2425	0.012	20.304	0.000
$\boldsymbol{\beta}_{SM}$	0.0251	0.003	7.552	0.000	$\boldsymbol{\gamma}_{BM}$	0.1856	0.013	14.783	0.000
$\boldsymbol{\beta}_{SF}$	0.0189	0.004	5.982	0.000	$\boldsymbol{\gamma}_{BF}$	0.2163	0.014	15.281	0.000
$\boldsymbol{\beta}_{BM}$	0.0589	0.001	48.770	0.000	$\boldsymbol{\gamma}_{BM}$	0.2088	0.012	17.505	0.000
$\boldsymbol{\beta}_{BF}$	0.0659	0.001	48.588	0.000	$\boldsymbol{\gamma}_{BF}$	0.2425	0.012	20.234	0.000
$\boldsymbol{\beta}_{IM}$	0.0748	0.001	61.160	0.000	$\boldsymbol{\gamma}_{BM}$	0.1856	0.013	14.783	0.000
$\boldsymbol{\beta}_{IF}$	0.0846	0.001	65.294	0.000	$\boldsymbol{\gamma}_{BF}$	0.2163	0.014	15.821	0.000

G. 改变所有卷积核大小的结果

我们还测试了修改所有卷积层的卷积核大小时框架的表现。结果见图 12。总体而言，我们发现，与仅修改第一个卷积核大小相比，修改所有卷积层的卷积核大小并不会产生显著差异。更多详情请参阅图注。

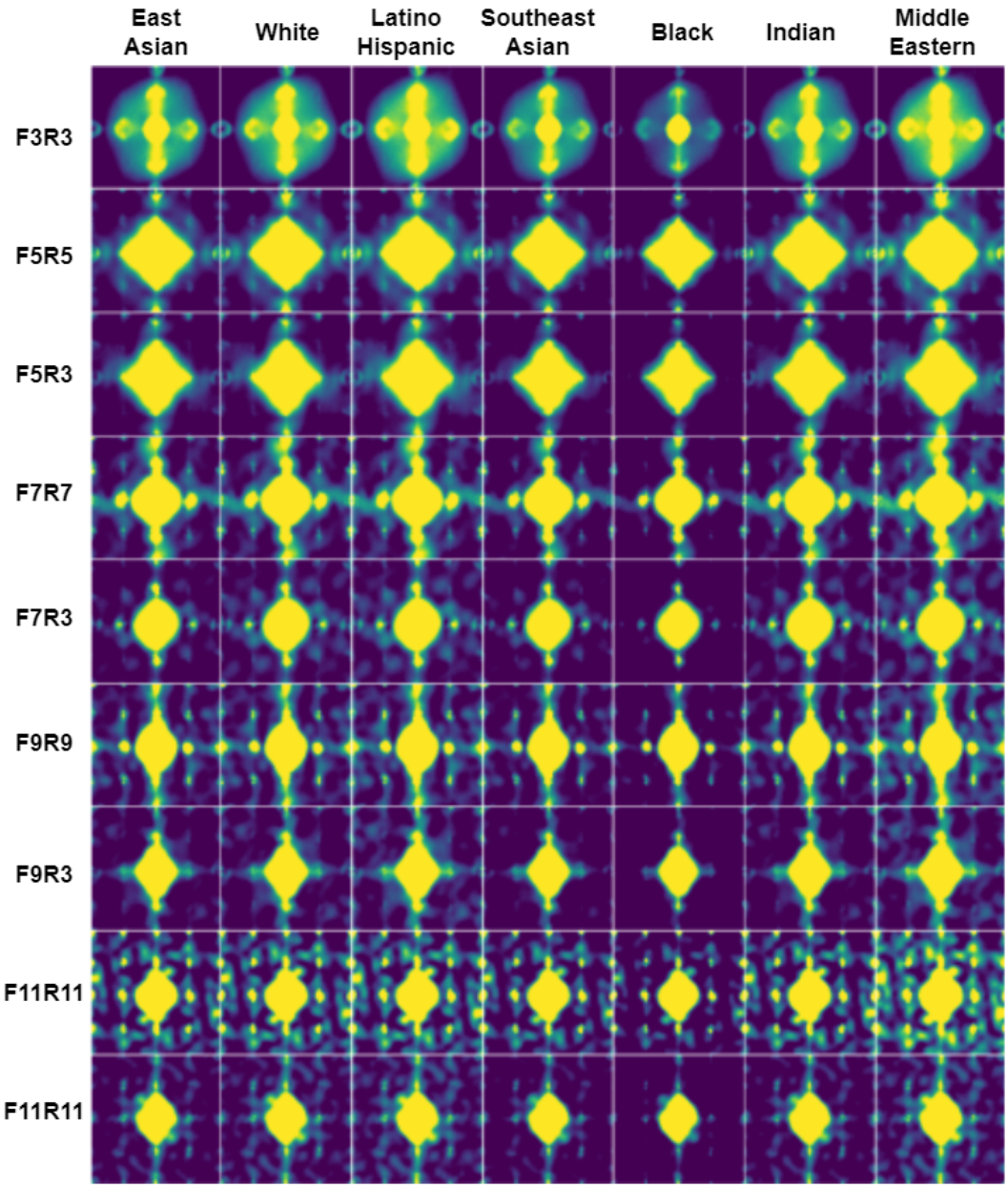

图12. **在 Fairface 上修改所有卷积层卷积核大小的扰动谱可视化**。每一行代表一种采用不同架构选择的模型，例如，“F3R3”表示模型的第一层卷积核大小为 3，其余层的卷积核大小也为 3；每一列对应一个受保护属性组。我们发现，与仅 修改第一层卷积核大小相比，修改所有卷积层的卷积核大小并没有带来显著差异。

H. 对在 UTKFace 上训练的模型应用 CW攻击的结果

我们还在另一个广受欢迎的人脸图像数据集 UTKFace 上进行了实验，并在图13中报告了结果。有关 Fairface 的结果和分析，请参阅正文中的图 3。

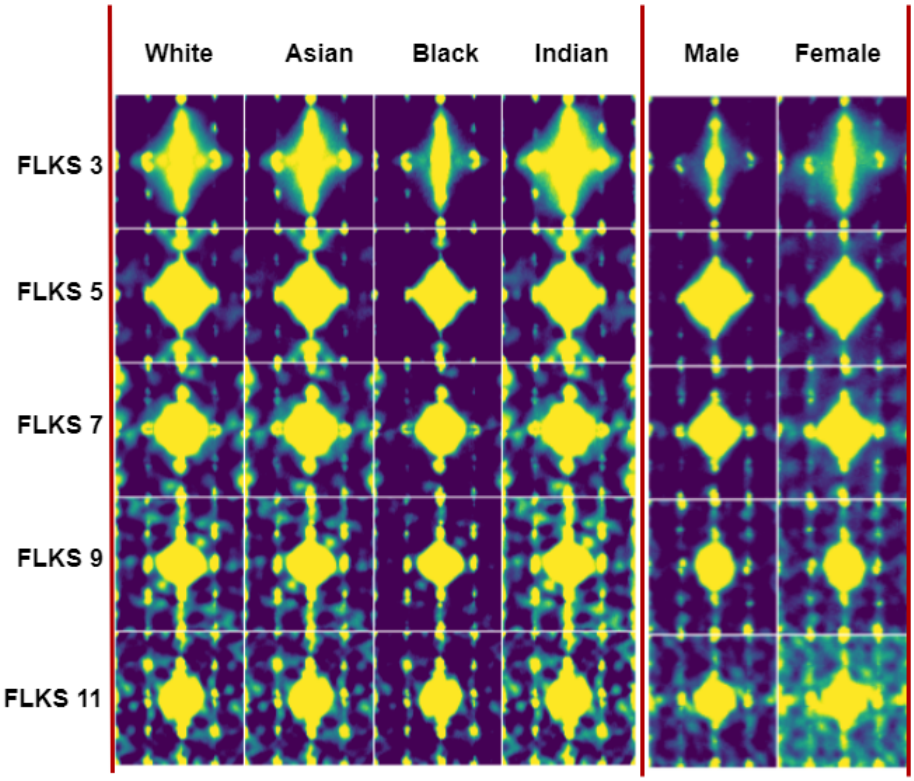

图13. **按种族和性别划分的 UTKFace 对抗扰动图像平均频谱**。我们在这些频谱中看到的趋势，与主论文图3中 Fairface 的频谱趋势相似。

I. 对模型应用 FGSM 的结果

我们还测试了这样一种情况：对所有在 Fairface 上训练的模型都应用 FGSM攻击。结果见图15。其趋势与此前所有频谱可视化结果基本一致。

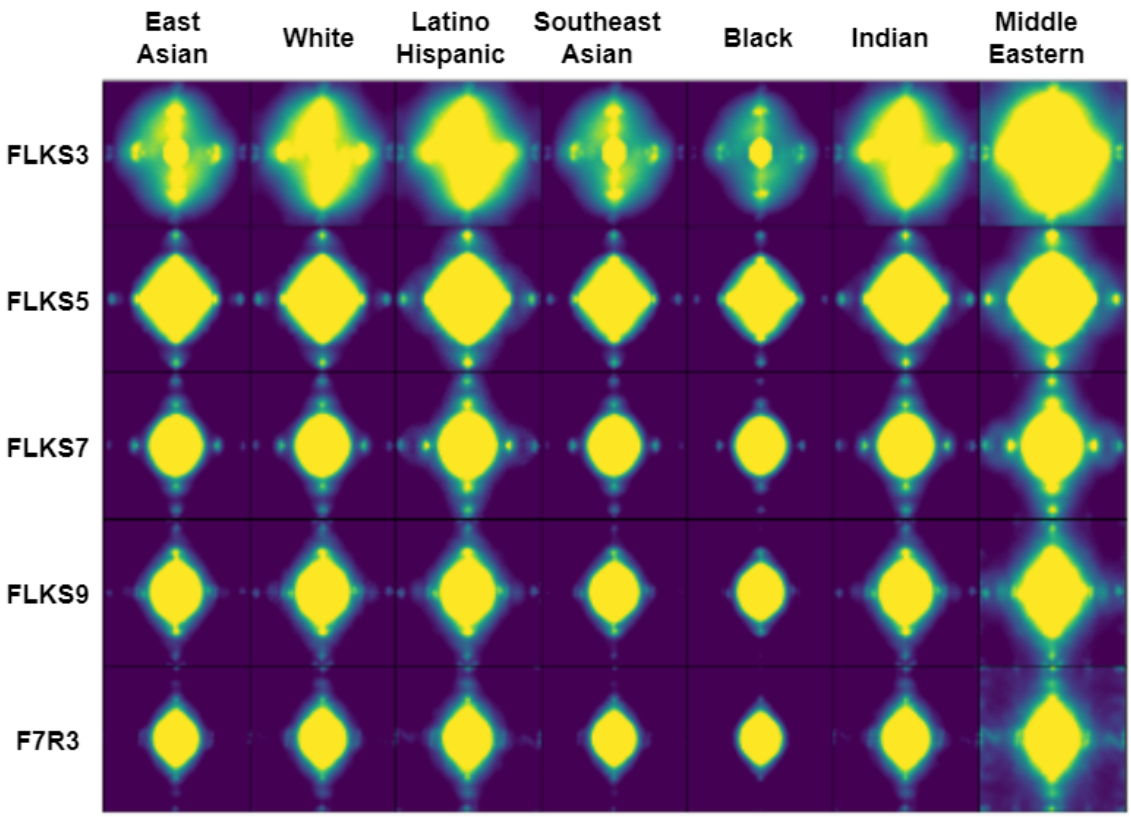

图14. **使用 FGSM 对 Fairface 进行扰动谱可视化**。 与图4中的结果类似，每一行代表一个采用不同第一层卷积核大小（FLKS）的模型，每一列对应一个受保护属性组。我们观察到类似的趋势：总体而言，随着 FLKS 增大，扰动的关注点会转向低频信息；黑人群体的扰动始终比其他种族群体具有更低的高频关注度。

J. 频率能量注入结果

与主论文中的图5相同, 我们分别展示了所有种族群体中采用不同 FLKS 的模型性能。

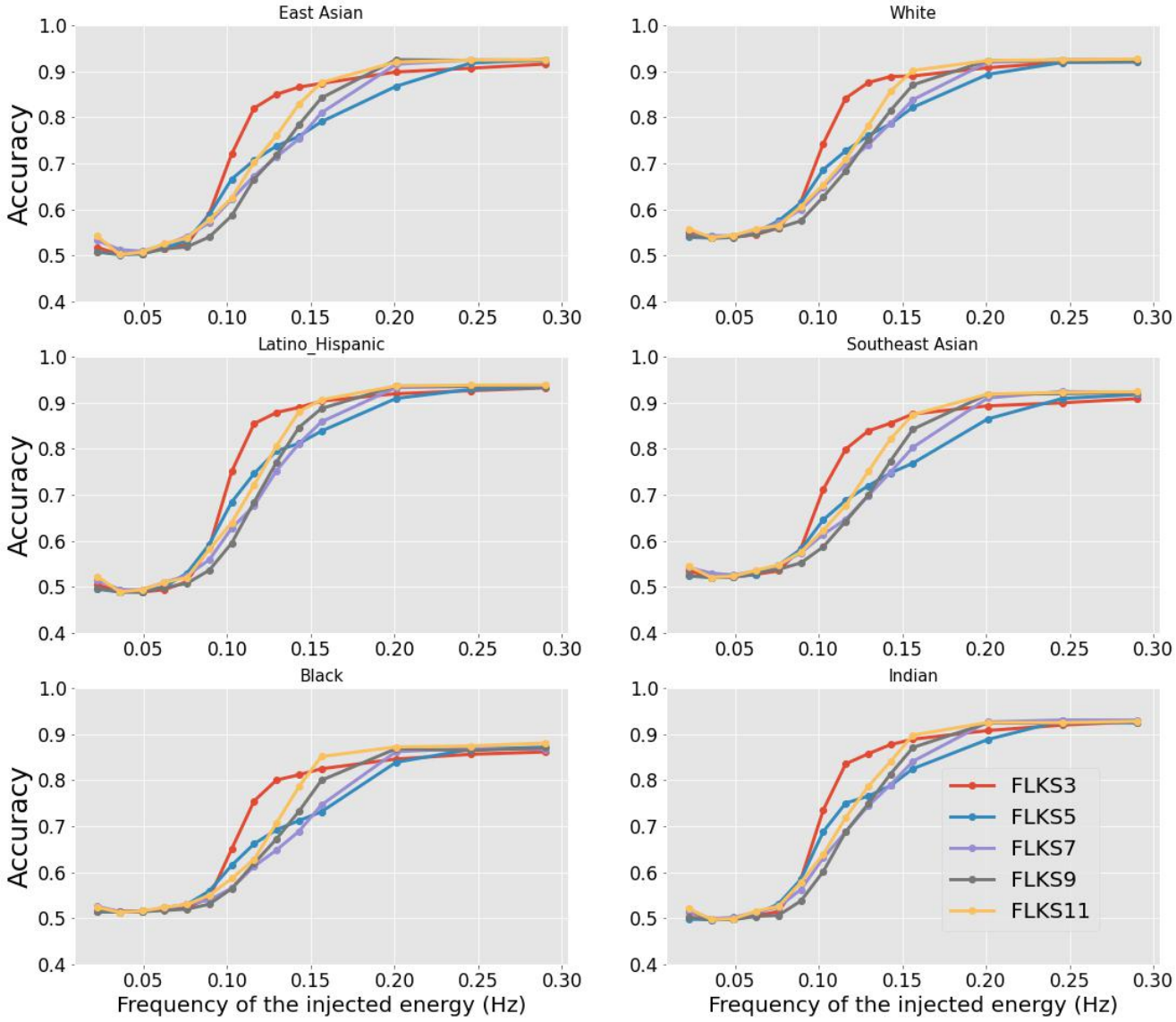

图15。**频率能量注入结果**。每个子图中，横轴表示注入能量的频率，纵轴表示不同模型的准确率。显然，所有模型都会受到低至中频能量注入的影响，而对中高频噪声则具有鲁棒性。很难直接判断哪个群体受到的影响比其他群体更大，因此需要进行定量分析。