# VB-MITIGATOR: 一个评估和推进视觉偏见缓解的开源框架

Ioannis Sarridis[1,2], Christos Koutlis[1], Symeon Papadopoulos[1], and Christos Diou[2]

[1]Information Technologies Institute, CERTH, Greece
[2]Department of Informatics and Telematics, Harokopio University of Athens, Greece
{ gsarridis,ckoutlis,papadop } @iti.gr, { isarridis,cdiou } @hua.gr

**ABSTRACT**

计算机视觉模型中的偏差仍然是一个重大挑战，通常导致不公平、不可靠和无法推广的 AI 系统。尽管减轻偏差的研究已经加强，但进展仍然因零散的实现和不一致的评估实践而受到阻碍。不同的研究中使用的不一致的数据集和指标使得可重复性变得复杂，难以公平地评估和比较各种方法的有效性。为克服这些限制，我们引入了视觉偏差缓解器（VB-Mitigator），这是一个开源框架，旨在简化视觉偏差缓解技术的发展、评估和比较分析。VB-Mitigator 提供了一个统一的研究环境，包含 12 种已建立的缓解方法和 7 个多样化的基准数据集。VB-Mitigator 的一个关键优势是其可扩展性，允许无缝集成其他方法、数据集、指标和模型。VB-Mitigator 旨在通过为研究社区的发展和评估方法提供基础代码库，加速面向公平意识的计算机视觉模型的研究。为此，我们还推荐最佳评估实践，并在最先进的方法之间提供全面的性能比较。

计算机视觉（CV）系统在各个领域经历了显著的增长和应用。这些进步极大地提高了大量应用中的自动化、效率和准确性。然而，CV 模型中偏见的广泛存在仍然是一个关键且未解决的挑战。这些偏见通常源于不平衡的训练数据集，导致模型学习的是虚假的相关性而非有意义的、可推广的模式。因此，这类模型常常会产生不可靠的预测、降低的推广能力以及维系人工智能（AI）偏见和刻板印象的结果。例如，在偏斜的人口分布上训练的面部识别系统表现出种族偏见，导致有害的现实后果。

尽管研究人员越来越认识到这些问题，并投入了大量努力来制定偏差缓解策略，但该领域在实施和评估实践中存在碎片化，使得公平评估和比较不同缓解方法的效果具有挑战性。总体而言，这种碎片化使得再现性复杂化，减缓了健壮解决方案的发展，并限制了社区有效推进公平意识技术的能力。

为了解决这些挑战，我们引入了视觉偏差缓解器（VB-Mitigator），这是一个开源的 [1] 框架，专门为促进视觉偏差缓解方法的标准化开发、评估和比较分析而创建。VB-Mitigator 提供了一个一致的研究环境，目前支持 12 种成熟的偏差缓解方法，7 个常用的数据集——包括合成数据集、涉及标准保护属性（如性别、年龄或种族）的数据集、背景相关偏差的数据集，以及通用计算机视觉数据集——以及评价指标来全面评估公平性和稳健性。VB-Mitigator 的一个关键优势是其可扩展性。模块化架构能够轻松集成新方法、数据集、评价指标和模型。

VB-Mitigator 的主要目标是通过提供一个全面且可扩展的代码库，促进更公平和更可靠的计算机视觉系统的发展。此外，在这项工作中，我们通过统一和标准化的评估协议，对 VB-Mitigator 中包含的偏差缓解方法进行了广泛的比较评估。

这项工作的主要贡献如下：

- 一个开源框架，旨在标准化和简化视觉偏差减缓方法的开发、评估和比较。
- 一种模块化和可扩展的架构，允许轻松集成新的偏差缓解方法、数据集、评估指标和模型。
- 使用 VB-Mitigator 对 12 种已确立的偏差缓解方法进行了广泛的比较评估。

在接下来的部分中，我们详细介绍了 VB-Mitigator 的架构和功能，讨论了已实现的偏见缓解方法、数据集和评估指标，并展示了全面的实验评估。

[1]https://github.com/mever-team/vb-mitigator

# 1 框架架构

VB-Mitigator 经过精心设计，作为一个坚固且可适应的平台，促进视觉偏见缓解的研究。其架构设计优先考虑模块化、可扩展性和可重复性，允许轻松开发和评估偏见缓解方法。本节深入探讨了赋予 VB-Mitigator 能力的复杂结构和基本特性。

构建一个全面的视觉偏差缓解代码库面临着重大挑战。现有技术差别很大，在训练流程的不同阶段进行干预——从数据加载器中的数据操作和损失函数的调整，到外部偏差检测模型的集成以及复杂的多阶段训练协议的使用。这种情况进一步复杂化，因为度量实现通常是数据集特定的，并且限于单一或双重偏差场景，从而妨碍了通用评估平台的创建。VB-Mitigator 直接通过基于 PyTorch 的抽象和模块化架构来应对这些障碍，选择这个框架是因为它的灵活性和强大的生态系统。通过为数据集、模型、缓解策略和评估指标提供标准化的接口，该框架能够无缝地集成和比较各种方法。

### 1.0.1 数据集

数据集组件（`datasets/`）封装了 PyTorch 数据集类，通过 `__getitem__` 方法返回包含输入图像、目标、偏差（或受保护属性）及样本索引的字典。此外，`builder.py` 模块促进数据集构建，生成包含关键元数据的全面字典，其中包括类别数量、受保护属性列表、子组数量、类别名称、数据子集以及初始数据加载器。该元数据对于模型初始化、指标计算、训练编排和动态数据加载器更新至关重要。

### 1.0.2 缓解措施

在 VB-Mitigator 中，减轻器组件（`mitigators/`）作为核心算法引擎，提供一个灵活和标准化的平台用于偏差减轻算法。`BaseTrainer` 类建立了一个全面的方法实现基础，定义了训练管道每个阶段的函数，包括数据集处理、模型训练、指标计算和记录（图 **??**）。此外，该框架支持方法特定的配置，方便了如偏差伪标签生成等预处理步骤的引入。这种模块化架构允许每个偏差减轻策略仅在其实际干预的管道组件中进行实现，确保了集成和维护的简便性。

模型组件（`models/`）包含在视觉偏差减轻研究中常用的多样化的神经网络架构集合。它支持多种架构，包括为小规模数据集（如 Biased-MNIST [**?**]）设计的轻量级卷积神经网络（CNNs），广泛采用的 CNN 架构如 ResNets [**?**] 和 EfficientNets [**?**]，以及用于更复杂任务的现代视觉变换器（[**?**, **?**]）。此外，它还适应针对特定偏差减轻方法定制的架构，确保了多样化实验设置的灵活性。

### 1.0.3 指标

度量组件（`metrics/`）提供了一套全面的评估指标，用于衡量公平性。鉴于常用的公平评估方法通常采用多种指标（例如，最差组准确性与平均组准确性同时使用），我们的实现支持包含多个测量的度量类。每个度量类进一步配置了两个属性：(i) 指示该指标是基于错误还是更高为佳，以及 (ii) 用于检查点选择的主要评估指标的指定。

### 1.0.4 工具和实用程序

该组件（`tools/`）包含了实验管理的基本工具，确保在框架内实现简化的工作流程。它包括用于实验执行的启动脚本和诸如日志记录和模型检查点等关键功能。日志系统支持 Wandb [2] 和 TensorBoard [3] 进行实时监控，同时也生成易于阅读的详细日志和可用于可视化的结构化 CSV 文件。此外，它负责管理检查点，存储模型在各个阶段的状态，包括最新的 epoch、性能最佳的模型和中间检查点，从而支持实验的恢复。

### 1.0.5 配置和执行脚本

配置文件充当中央控制机制，允许研究人员在数据集、偏差缓解方法和超参数之间无缝切换。考虑到可配置变量的数量众多，我们使用了 YACS [4] 库，它提供了一种结构化且高效的方式来管理配置。该方法在提高实验设置灵活性的同时保持了清晰性。最后，`scripts/` 目录包含了一组旨在简化和自动化跨不同偏差缓解方法和数据集执行实验的 shell 脚本。

VB-Mitigator 的设计强调了几项关键特性，这些特性增强了它作为一个全面的偏差缓解框架的实用性：

[2] https://wandb.ai/site
[3] https://www.tensorflow.org/tensorboard
[4] https://github.com/rbgirshick/yacs

## 2 方法论

### 2.1 初步符号

在这里，我们介绍了本文中贯穿始终使用的符号，以确保所有方法的一致性。设 $f(\boldsymbol{x};\boldsymbol{\theta})$ 表示一个由 $\boldsymbol{\theta}$ 参数化的神经网络，它将输入样本 $\boldsymbol{x} \in \mathcal{X}$ 映射到输出预测 $\hat{y} \in \mathcal{Y}$ 。数据集由 $N$ 个训练样本组成，其中每个样本都与一个目标标签 $y \in \mathcal{Y}$ 相关，并且可能还包括一个引入伪相关性的偏置属性 $a \in \mathcal{A}$ 。输入的学习特征表示记为 $\boldsymbol{z} = h(\boldsymbol{x};\boldsymbol{\theta})$ ，其中 $h$ 表示模型的特征提取组件。标准模型的目标是学习模型参数 $\boldsymbol{\theta}$ ，该参数最小化训练样本的平均损失，形式为：

$$\min_{\boldsymbol{\theta}} \frac{1}{N} \sum_{i=1}^{N} L(f(\boldsymbol{x}_i;\boldsymbol{\theta}), y_i).$$

### 2.2 方法描述

偏见缓解方法可以根据其对显式偏见注释的依赖性大致分类。具体来说，我们区分了偏见标签无感（BLU）方法，这些方法在没有关于造成虚假相关属性的信息的情况下操作，以及偏见标签感知（BLA）方法，这些方法利用这些注释。虽然 BLA 技术在控制环境中经常表现出优越的性能，但 BLU 方法展示了更广泛的适用性，使其更适合偏见注释可能不可用或不可靠的实际场景。VB-Mitigator 包含了两类方法中的几种，具有以下 BLA 方法：GroupDro [?] 、域独立（DI）[?] 、纠缠与解纠缠（EnD）[?] 、偏见平衡（BB）[?] 、和偏见添加（BAdd）[?] ，以及以下 BLU 方法：从失败中学习（LfF）[?] 、谱分离（SD）[?] 、只训练两次（JTT）[?] 、SoftCon [?] 、去偏置备用网络（Debian）[?] 、公平意识表示学习（FLAC 和 FLAC-B）[?] 、以及缓解任何视觉偏见（MAVias）[?] 。这些方法的关键特征见表 1 。下面我们简要介绍所考虑的方法。

Table 1: 综合偏差减缓方法的总结。

| Method | Bias Labels | Bias-Capturing Model | Summary |
|---|---|---|---|
| GroupDRO [?] | ✓ | × | Minimizes worst-case loss across predefined groups. |
| DI [?] | ✓ | ✓ | Uses domain-specific classification heads for domain invariance. |
| EnD [?] | ✓ | × | Disentangles bias representations and entangles class representations. |
| BB [?] | ✓ | × | Balances bias in the logit space using prior bias information. |
| BAdd [?] | ✓ | ✓ | Adds bias-capturing features to training to discourage their use. |
| LfF [?] | × | ✓ | Reweights samples based on bias-conflicting predictions from an auxiliary model. |
| SD [?] | × | × | Regularizes network logits for spectral decoupling and bias robustness. |
| JTT [?] | × | ✓ | Reweights misclassified samples, assuming they are bias-conflicting. |
| SoftCon [?] | × | ✓ | Uses a weighted contrastive loss based on bias-capturing features. |
| Debian [?] | × | ✓ | Alternates training of main and auxiliary models for debiasing. |
| FLAC [?] | × | ✓ | Minimizes mutual information between representations and bias-capturing features. |
| MAVias [?] | × | ✓ | Infers and mitigates visual biases using foundation models and regularization. |

**群体分布鲁棒优化（GroupDro）。** GroupDro 通过最小化预定义组中的最坏情况损失来解决偏差问题，确保对组级别偏差的稳健性。组被定义为 $\mathcal{Y}$ 和 $\mathcal{A}$ 的组合。其目标是最小化组间的最大损失，定义如下：

$$\min_{\boldsymbol{\theta}} \max_{g \in \mathcal{G}} \frac{1}{N_g} \sum_{i:g_i=g} L(f(\boldsymbol{x}_i;\boldsymbol{\theta}), y_i)$$

其中 $\mathcal{G}$ 是组的集合，$N_g$ 是组 $g$ 中的样本数量，$g_i$ 是样本 $i$ 的组分配。

**领域独立（DI）。** 域独立（DI）旨在通过学习在不同域间不变的表征来减轻偏差。与传统方法不同，DI 使用多个分类头，每个对应一个不同的域。对于属于域 $\alpha$ 的每个输入样本 $\boldsymbol{x}$ ，模型只选择并使用由与域 $\alpha$ 相关的分类头产生的 logits。设 $f_\alpha(\boldsymbol{x};\boldsymbol{\theta})$ 表示与域 $\alpha$ 对应的分类头的输出 logits，其中 $\boldsymbol{\theta}$ 表示模型参数。目标是在最小化域特定损失的同时，促进域不变的表征。这可以表示为:

$$\min_{\boldsymbol{\theta}} \left[ \frac{1}{N} \sum_{i=1}^{N} L(f_{\alpha_i}(\boldsymbol{x}_i;\boldsymbol{\theta}), y_i) \right].$$

**纠缠与解纠缠（EnD）。** EnD 显式地解耦共享同一偏差标签的样本的表征，同时耦合在同一类别下具有不同偏差标签的样本的表征。损失函数为:

$$\min_{\boldsymbol{\theta}} \Big[ L(f(\boldsymbol{x};\boldsymbol{\theta}),y) + \lambda_{\text{EnD},1} L_{dis}(\boldsymbol{z},\alpha) + \lambda_{\text{EnD},2} L_{ent}(\boldsymbol{z},\alpha,y) \Big]$$

其中，$L_{dis}$ 是解耦损失，$L_{ent}$ 是耦合损失，正则化项表示为 $\lambda_{\text{EnD},1}$ 和 $\lambda_{\text{EnD},2}$ 。

**偏差平衡 (BB)**。 BB 从偏斜的分布推断出无偏分布，并使用一种损失函数在 logit 空间中平衡偏差影响，表示为：

$$\min_{\boldsymbol{\theta}} L(f(\boldsymbol{x};\boldsymbol{\theta}) + \boldsymbol{p}, y)$$

其中 $\boldsymbol{p}$ 表示与偏差相关的先验知识。

**偏置加法 (BAdd)**。 BAdd 显式地将偏差添加到训练过程中，并鼓励模型不要学习与该偏差相关的特征。目标函数具有以下形式：

$$\min_{\boldsymbol{\theta}} L(f(\boldsymbol{x};\boldsymbol{\theta}), \boldsymbol{b}, y)$$

，其中 $\boldsymbol{b}$ 表示由捕获偏差的模型得出的偏差特征。

**从失败中学习 (LfF)**。 LfF 采用双模型训练范式，同时训练一个主分类模型和一个辅助偏差预测模型。辅助模型用于预测偏差属性 $\mathcal{A}$ 。通过比较每个小批量内两个模型的损失，LfF 识别出偏差冲突样本。这些样本随后被重新加权，以调整它们对主模型训练的影响。优化目标定义为：

$$\min_{\boldsymbol{\theta}} \frac{1}{N} \sum_{i=1}^{N} w_i L(f(\boldsymbol{x}_i;\boldsymbol{\theta}), y_i).$$

其中 $w_i$ 表示分配给样本 $i$ 的权重。

**光谱解耦 (SD)**。 SD 表明，在网络的对数上添加一个正则化项能够导致谱分离，从而增强网络在虚假相关性上的鲁棒性。目标可以定义为：

$$\min_{\boldsymbol{\theta}} L(f(\boldsymbol{x};\boldsymbol{\theta}), y) + \lambda_{\text{SD}} \|\hat{y}\|^2$$

，其中 $\lambda_{\text{SD}}$ 为正则化权重。与其他 BLU 方法不同，SD 采用了一种通用的方法来缓解偏差，而不是针对任何特定偏差的推理进行操作。

**只需训练两次 (JTT)**。 JTT 专注于从错误分类的样本中学习，因为这些样本往往是与偏差相冲突的样本。它使用重加权方案在训练过程中通过相应地修改数据加载器来优先考虑这些样本。

**软对比 (SoftCon)**。 SoftCon 是一种加权的 SupCon [**?**] 损失，通过由捕获偏差模型导出的特征间的余弦距离加权，鼓励具有相同目标标签的样本之间的特征相似性，其代表引入虚假相关性的属性。这些权重可以表示为：

$$w_{i,j} = 1 - \frac{\boldsymbol{b}_i \boldsymbol{b}_j}{\|\boldsymbol{b}_i\| \|\boldsymbol{b}_j\|}.$$

与 LfF 类似，Debian 使用一个辅助模型来封装与偏差相关的信息。特别地，它采用了一种在训练主网络和辅助网络之间交替的方案来减轻偏差。辅助模型的预测用于给主网络的损失分配权重。然后，与 LfF 类似，目标是：

$$\min_{\boldsymbol{\theta}} \frac{1}{N} \sum_{i=1}^{N} w_i L(f(\boldsymbol{x}_i;\boldsymbol{\theta}), y_i).$$

**公平意识表示学习 (FLAC)**。 FLAC 通过最小化特征与敏感属性之间的依赖性来学习公平表示。其目标是最小化表示与敏感属性之间的互信息，定义为：

$$\min_{\boldsymbol{\theta}} L(f(\boldsymbol{x};\boldsymbol{\theta}), y) + \lambda_{\text{FLAC}} I(\boldsymbol{z}, \alpha)$$

其中 $\lambda_{\text{FLAC}}$ 是一个超参数，$I(\boldsymbol{z},\alpha)$ 是表示与敏感属性之间的互信息，而不访问 $\mathcal{A}$ 标签。为此，FLAC 采用一个偏差捕获模型，训练该模型以提取特征 $\boldsymbol{b}$ 。在训练专门的偏差捕获模型不切实际的情形（例如，偏差未知）下，可以使用带偏差的原始模型，从而得到 FLAC-B 变体。$\boldsymbol{\theta}$ 代表主模型的可学习参数，而 $\boldsymbol{z}$ 和 $\boldsymbol{a}$ 分别是特征和敏感属性的向量表示。

Table 2: 支持的数据集摘要。

| Dataset | Data Type | Bias Type | # Biases | Spurious Correlations |
|---|---|---|---|---|
| Biased-MNIST | digits | foreground color | 1 | 99 % -99.9 % |
| FB-Biased-MNIST | digits | foreground & background color | 2 | 90 % -99 % |
| Biased-UTKFace | faces | demographics (race or age) | 1 | 90 % |
| Biased-CelebA | faces | demographics (gender) | 1 | 90 % |
| Waterbirds | bird species | background scene | 1 | 95 % |
| UrbanCars | car type | background scene & co-occuring object | 2 | 95 % |
| ImageNet9 | general purpose | background & texture | unknown | unknown |

**减轻任何视觉偏见 (MAVias)**。 MAVias 采用两阶段的方法。首先，它通过利用基础模型来为输入图像生成描述性的标签，并评估它们与目标类别的相关性，从而推断出潜在的视觉偏差。然后，这些潜在的偏差通过视觉-语言模型进行编码，并在训练过程中作为正则化引入，以阻止模型学习到虚假的相关性。其最小化目标定义为：

$$\min_{\boldsymbol{\theta},\boldsymbol{\phi}} L(f(\boldsymbol{x};\boldsymbol{\theta}),y) + \lambda_{\text{MAVias},1} L_{reg}(\boldsymbol{x},\boldsymbol{b},\lambda_{\text{MAVias},2})$$

其中，$\boldsymbol{\phi}$ 表示将偏差嵌入映射到视觉空间的投影层参数，$L_{reg}$ 是正则化项，$\lambda_{\text{MAVias},1}$ 和 $\lambda_{\text{MAVias},2}$ 是超参数。$\boldsymbol{\theta}$ 表示主模型的可学习参数。

## 3 数据集

VB-Mitigator 支持多样化的数据集，以评估在各种情境下的偏见缓解技术。这些数据集包括具有控制偏见的合成数据、在既有基准中手动注入的偏见，以及通用的计算机视觉数据集，从而促进对比方法的全面评估。Biased-MNIST [?] 是 MNIST 数据集的一个修改版，引入了背景颜色与数字标签的相关性。该数据集提供不同强度的偏见，常用于评估偏见缓解性能的 99 % 、99.5 % 、99.7 % 和 99.9 % 的共现水平。

**FB 偏置的 MNIST** 在 Biased-MNIST 的基础上，FB-Biased MNIST [?] 通过同时将前景和背景颜色与数字标签相关联，引入了多个偏见。此数据集建议的共现水平为 90 % 、95 % 和 99 % ，提供了更具挑战性的基准。

**偏见的 UTKFace** UTKFace 数据集 [?] ，由带有年龄、性别和种族注释的面部图像组成，是 Biased-UTKFace [?] 的基础。在此偏见变体中，性别被指定为目标变量，而种族或年龄则作为偏倚属性，与目标呈现 90 % 的共现。此数据集在面部属性分类中研究人口统计偏见方面具有重要作用。

**偏见-CelebA** 利用大型的 CelebA 数据集 [?] ，该数据集提供多样的面部属性注释，Biased-CelebA [?] 重点关注与性别相关的偏见。在这里，金发或浓妆是目标属性，性别表现出 90 % 的共现，能够研究特定属性的性别偏见。

**水鸟**。 Waterbirds 数据集 [?] 促进了对鸟类与其背景环境之间虚假相关性的研究。其特点是包含陆鸟和水鸟在对应陆地或水域背景下的图像，呈现了一种鸟类与背景的 95 % 共现性，作为评估与情境相关的偏差缓解的基准。

**都市汽车** UrbanCars [?] 是一个汽车图像数据集，旨在探索物体识别中的偏差。它检查了汽车类型分类中的偏差，其中背景（乡村或城市）和同时出现的物体以 95 % 的共现率与目标引入偏差。

**ImageNet-9**。 ImageNet-9 [?] 是 ImageNet [?] 的一个子集，专注于九个物体类别。它用于研究大规模物体识别中的与背景相关的偏差，提供了一个更加复杂和现实的评估场景，其中偏差是未知的。

在视觉偏差缓解中使用的数据集的进展反映了复杂性日益增加，这与该领域研究的发展相呼应。早期的研究主要集中在单一属性偏差上，通常使用合成或简化的数据集，如 Biased-MNIST。近期的研究则转向探索多属性偏差，如 FB-Biased MNIST 和 UrbanCars，这些数据集让许多现有方法难以维持性能。此外，像 ImageNet-9 这样的通用 CV 数据集的引入标志着解决现实场景中可能出现的多种交互偏差挑战的转变。表 2 概述了所考虑的数据集。

评估视觉偏差缓释技术的有效性需要仔细考虑适当的性能度量。虽然标准准确率，定义为：

$$\text{Acc} = \frac{|\{\boldsymbol{x} \in \mathcal{X} : \hat{y} = y\}|}{|\mathcal{X}|},$$

是一个基础的度量，其效用依赖于具体上下文。即使偏差在数据集内均匀分布，模型可能对不同的偏差属性表现出不同的敏感性，从而使在平衡测试集（如 Biased-MNIST 和 FB-Biased-MNIST）上的准确率不足以捕获缓释方法的细微行为。另一方面，在偏差未知并通过增强（例如在 ImageNet9 中移除混淆背景信息）使测试集去偏时，准确率是一个合适的度量。

此外，偏见冲突准确率（BCA）通常用于 Biased-CelebA 和 Biased-UTKFace 数据集。BCA 定义为 $\text{BCA} = \text{Acc}(\mathcal{X}_{BC})$，其中 $\mathcal{X}_{BC} = \{\boldsymbol{x} \in \mathcal{X} : a \text{ conflicts with } y\}$，旨在关注数据中代表性不足的群体。尽管该指标提供了对偏见冲突样本的性能见解，但它无法泛化以考虑多个相互作用的偏见。

鉴于这些指标的局限性，我们主张采用最差组准确率（WGA）和平均准确率（AvgAcc）。WGA 定义为 $\text{WGA} = \min_{g\in\mathcal{G}} \text{Acc}(g)$，AvgAcc 定义为 $\text{AvgAcc} = \frac{1}{|\mathcal{G}|}\sum_{g\in\mathcal{G}} \text{Acc}(g)$，它们提供了一种更稳健的评估框架，因为它们能够有效捕捉子群体间的表现差异，使其适用于具有复杂偏差分布和多种偏差属性的数据集。

## 4 实验

本节对 VB-Mitigator 中的方法进行了比较分析，并在 Biased-CelebA，Waterbirds，UrbanCars 和 ImageNet9 上进行了评估。这些数据集的选择是为了在多样的偏差情境中提供具有代表性的评估，分别包含人口统计、背景场景、多属性和未知偏差。

### 4.1 评估协议

鉴于 WGA 和 AvgAcc 适用于具有明确定义偏差的数据集，如第 **??** 节所述，这些指标被用于对 Biased-CelebA，Waterbirds 和 UrbanCars 的模型进行评估。对于 ImageNet9，我们利用其七个官方测试集变体的准确性来进行评估，这有助于更全面地评估模型对非目标对象特征的依赖。这些变体是：

在下面，我们概述了所有方法共同的数据预处理、模型架构和超参数，除非另有说明。

我们使用了 Biased-CelebA 数据集，将 “金发” 作为目标属性，“性别” 作为虚假相关。图像被调整为 $224 \times 224$ 像素，并使用 ImageNet 统计数据进行归一化。采用 ResNet18 架构，使用批量大小为 128 训练了 10 个周期。Adam 优化器被使用，初始学习率为 0.001 ，并在第 3 和第 6 轮时乘以 0.1 进行减小。施加了权重衰减 0.0001 。

**水鸟**。 图像被调整为 $256 \times 256$ 像素，中心裁剪至 $224 \times 224$ 像素，并使用 ImageNet 统计数据进行规范化。使用 ResNet50 模型进行训练，训练了 100 个周期，批量大小为 64。采用随机梯度下降（SGD）优化器，学习率为 0.001 ，权重衰减为 0.0001 。

**城市汽车** 图像被调整为 $256 \times 256$ 像素大小，中心裁剪为 $224 \times 224$ 像素，并进行了随机旋转（最多 45 度）和水平翻转。使用 ImageNet 统计数据进行归一化。使用 ResNet50 架构，采用 SGD 方法训练了 150 个周期，批量大小为 64，权重衰减为 0.0001 。

**ImageNet9**。 图像被调整到 $256 \times 256$ 像素尺寸，中心裁剪到 $224 \times 224$ 像素，并使用 ImageNet 的统计数据进行归一化。使用 ResNet50 模型进行训练，共 30 个 epoch，批量大小为 64。使用 SGD 优化器进行训练，初始学习率为 0.001 ，并在第 25 个 epoch 时减少到 0.5 的一个因子。

根据各自原始出版物中推荐的数值配置了方法特定的超参数。对于 GroupDro，使用了 0.01 的鲁棒步长。在 SoftCon 中，交叉熵损失被加权为 0.01。对于 Biased-CelebA、Waterbirds、UrbanCars 和 ImageNet9，参数 $\lambda_{\text{FLAC}}$ 分别设置为 30,000, 10,000, 10,000 和 100。对于 JTT，对抗偏见的样本被加权提升了 100 倍，并采用学习率 0.00001 和权重衰减 1。对于 MAVias，参数 $(\lambda_{\text{MAVias},1}, \lambda_{\text{MAVias},2})$ 分别在 Biased-CelebA、Waterbirds、UrbanCars 和 ImageNet9 中设为 (0.01, 0.5), (0.05, 0.6), (0.01, 0.4) 和 (0.001, 0.7)。对于 SD，参数 $\lambda_{\text{SD}}$ 设置为 0.1。对于 EnD，参数 $\lambda_{\text{EnD},1}$ 和 $\lambda_{\text{EnD},2}$ 都设为 1。实验在 NVIDIA A100 GPU 上针对 5 个随机种子重复进行。本节展示了在不同数据集中的偏见缓解方法的比较评估，涵盖了不同的偏见情境。正如表格 **??** 所示，BLA 方法（如 DI 和 BAdd）通常在所有数据集中表现出较优的 WGA，展示了利用显性偏见信息的有效性。然而，GroupDRO、EnD 和 BB 的性能表现不稳定，特别是在 UrbanCars 数据集上，它们表现不佳。这是由于 UrbanCars 的多属性偏见结构，与这些方法为单一属性偏见设计的方式形成对比。BLU 方法虽然通常比 BLA 方法表现逊色，但显示出相似的趋势。值得注意的是，MAVias 在各个数据集上表现稳定，而 SoftCon 训练不稳定，可能是由于其对辅助模型的强依赖性。

对于 ImageNet9，由于偏差本质上是未知的，因此只有 BLU 方法适用。请注意，这里我们使用 FLAC 的 BLU 变体，记为 FLAC-B，使用原始模型作为偏差捕捉组件。如表 **??** 所示，SD 和 MAVias 在 ImageNet9 的各种测试集配置中展示出强大的泛化能力。相反，SoftCon 再次无法收敛。

## 5 局限性和伦理考虑

尽管 VB-Mitigator 的架构旨在促进现有视觉偏差缓解方法的无缝整合，但需要承认未来的方法可能会带来无法预见的整合挑战。此外，需要强调的是，偏差缓解是一个开放的研究问题。目前由 VB-Mitigator 支持的方法虽然有效，但并不能保证创建出完全公平的模型。

## 6 结论

本文介绍了 VB-Mitigator，这是一个开源框架，旨在应对计算机视觉模型中偏见这一重大挑战。实施和评估实践的碎片化阻碍了偏见消除的进展。VB-Mitigator 通过提供一个标准化平台来解决这一问题，推动可重复性和公平比较。该框架作为研究社区的基础代码库，使得对公平性意识方法的开发和评估更加高效。未来的工作将专注于通过新的方法和数据集扩展 VB-Mitigator。一个关键的方向是整合利用基础模型来推导潜在偏见的方法 [**?**, **?**] 。这些注释将允许在尚未探索偏见的广泛通用计算机视觉数据集中评估公平性，从而增强框架在解决现实世界场景中偏见问题上的影响。

## 7

致谢 本研究得到了欧盟地平线欧洲项目 MAMMOth（资助号 101070285）、ELIAS（资助号 101120237）和 ELLIOT（资助号 101214398）的支持。

## References