**Computer Science and Application** 计算机科学与应用, 2026, 16(5), 153-164
Published Online May 2026 in Hans. https://www.hanspub.org/journal/csa
https://doi.org/10.12677/csa.2026.165172

Hans汉斯

# 预处理方法对糖尿病视网膜病变分级中单源域泛化影响的实证研究

**刘新雨，陈　俊，邵春霖，贾雨欣，柳伟生***

辽宁科技大学计算机与软件工程学院，辽宁　鞍山

收稿日期：2026年4月15日；录用日期：2026年5月13日；发布日期：2026年5月25日

## 摘　要

**针对糖尿病视网膜病变(DR)自动诊断中跨数据集泛化能力不足的问题，文章开展了一项关于预处理方法对单源域泛化性能影响的实证研究。以EfficientNet为统一骨干网络，在APTOS数据集上训练，在DDR、IDRiD和ISBI三个独立数据集上进行零样本泛化测试，系统对比了五种预处理策略：无预处理、Circle Crop + CLAHE、仅MixStyle、仅DRGen、CLAHE + MixStyle。实验结果表明，不同预处理方法对源域精度与目标域泛化能力的影响存在显著差异，其中Circle Crop + CLAHE组合在保持源域性能的同时，在三个目标域上取得最优的平均泛化性能，而MixStyle和DRGen等风格扰动方法也展现出良好的跨域适应能力。文章首次在DR分级跨域场景下对主流预处理方法进行系统性横向对比，揭示了预处理选择对单源域泛化的关键影响，为多中心DR筛查系统的构建提供了可量化的预处理参考依据。**

## 关键词

**糖尿病视网膜病变分级，深度学习，域泛化，预处理，EfficientNet**

# An Empirical Study on the Impact of Preprocessing Methods on Single-Source Domain Generalization in Diabetic Retinopathy Grading

**Xinyu Liu, Jun Chen, Chunlin Shao, Yuxin Jia, Weisheng Liu***

School of Computer Science and Software Engineering, University of Science and Technology Liaoning, Anshan Liaoning

Received: April 15, 2026; accepted: May 13, 2026; published: May 25, 2026

*通讯作者。

文章引用: 刘新雨, 陈俊, 邵春霖, 贾雨欣, 柳伟生. 预处理方法对糖尿病视网膜病变分级中单源域泛化影响的实证研究[J]. 计算机科学与应用, 2026, 16(5): 153-164. DOI: 10.12677/csa.2026.165172

## Abstract

**This paper presents an empirical study on the impact of preprocessing methods on single-source domain generalization in response to the issue of insufficient cross-dataset generalization capability in automatic diabetic retinopathy (DR) diagnosis. Using EfficientNet as the unified backbone network, models were trained on the APTOS dataset and subjected to zero-shot generalization testing on three independent datasets—DDR, IDRiD, and ISBI—to systematically compare five preprocessing strategies: no preprocessing, Circle Crop with CLAHE, MixStyle only, DRGen only, and CLAHE with MixStyle. Experimental results reveal significant variations in the effects of different preprocessing methods on source domain accuracy and target domain generalization. Among these, the combination of Circle Crop + CLAHE achieves the best average generalization performance across the three target domains while preserving source domain performance. Additionally, style perturbation methods such as MixStyle and DRGen demonstrate strong cross-domain adaptation capability. This study presents the first systematic horizontal comparison of mainstream preprocessing methods in the context of cross-domain DR grading, highlighting the critical role of preprocessing choices in single-source domain generalization and providing a quantifiable reference for the development of multi-center DR screening systems.**

## Keywords

**Diabetic Retinopathy Grading, Deep Learning, Domain Generalization, Preprocessing, EfficientNet**



## 1. 引言

### 1.1. 任务背景

糖尿病视网膜病变(Diabetic Retinopathy, DR)是全球范围内导致成年人视力损伤和失明的主要原因之一。大规模的眼底筛查被公认为是降低 DR 致盲率的有效手段。随着深度学习技术的发展，基于彩色眼底图像的自动 DR 分级技术已取得了显著进展。然而，这些高精度的模型大多在特定数据集上训练，忽略了不同医疗中心、成像设备及患者群体带来的数据分布差异，即“域漂移”(Domain Shift)问题。这导致模型在部署到新的、未见过的数据源时性能往往显著下降，严重阻碍了其在实际临床筛查中的大规模应用。

单源域泛化(Domain Generalization)是解决上述问题的关键技术方向，其目标是在源域上学习一个模型，使其能够直接泛化到未见过的目标域。现有研究主要通过设计精巧的模型架构或学习策略来提取域不变特征，例如域对抗神经网络、风格混合策略以及频域信息交换等。这些方法从特征约束、数据增广、分布模拟等角度展示了缓解域偏移的潜力，为 DR 分级模型的跨数据集评估奠定了方法论基础。

然而，一个基础且关键的问题却常常被忽略：输入图像的质量与预处理方式，究竟如何影响模型的单源域泛化能力？在临床实践中，不同来源的眼底图像在对比度、光照、视野范围及噪声等方面存在巨大差异。预处理技术如对比度受限自适应直方图均衡化(CLAHE)、圆形裁剪(Circle Crop)、MixStyle 以及 DRGen 等，虽被广泛用于提升图像质量或模拟域分布变化，但在现有的 DR 分级泛化研究中，它们通常仅被视为固定的“前处理步骤”或简略的消融实验，缺乏对其效果的系统性、横向对比评估。不同预处

理组合对源域训练精度与目标域泛化精度之间的复杂关系，至今尚未被充分揭示。

### 1.2. 研究目标与内容

基于此，本文旨在填补这一研究空白，开展一项关于预处理方法对 DR 分级单源域泛化性能影响的实证研究。为避免网络结构对评估结果的干扰，选择 EfficientNet 作为统一的骨干网络，以专注于预处理本身的作用。实验设计方面，在 APTOS 数据集[1]上进行训练，并在 DDR [2]、IDRiD [3]和 ISBI [4]三个独立数据集上开展零样本泛化测试，系统对比五种典型预处理策略：无预处理、Circle Crop + CLAHE、仅 MixStyle、仅 DRGen 以及 CLAHE + MixStyle。

通过上述系统性对比，本文期望在以下方面为领域提供参考：目前尚缺乏在 DR 分级跨域场景下对主流预处理方法的公平横向对比，本研究可为此类工作提供量化依据；预处理方法对源域精度与目标域泛化能力的影响往往存在权衡关系，本文试图揭示这一现象并挑战“源域精度最优即泛化最优”的常见假设；基于实验结果，本文进一步提出面向多中心部署的预处理推荐思路，以期为真实临床环境中异构数据源下的模型部署提供一种简洁有效的实践方案。

## 2. 相关工作

### 2.1. 单源域泛化在医学图像中的应用

单源域泛化旨在从源域中学习具有域不变性的特征，从而使模型能够泛化到未见过的目标域。在医学图像分析领域，由于多中心数据采集设备、成像协议及患者人群存在显著差异，域泛化技术受到广泛关注。近年来，研究者已提出多种用于糖尿病视网膜病变(DR)分级的深度学习模型，如 Men 等提出的 DRStageNet [5]，在单一数据集上取得了优异性能。然而，这些模型在跨数据集场景下往往面临显著的性能下降，即域漂移问题。为缓解这一问题，现有研究从不同角度提出了多种缓解域偏移的方法，其中以下几项代表性工作为 DR 分级模型的单源域泛化奠定了重要基础。

基于对抗学习的域不变特征提取。Ganin 等人提出的域对抗神经网络(Domain-Adversarial Neural Networks, DANN) [6]是该方向的经典工作。其核心思想是在特征提取器之后引入域判别器，通过对抗训练的方式，使特征提取器学习到的特征表示无法被域判别器区分来自哪个域，从而提取出对域分布不敏感的特征。该方法通过最小化源域分类损失与最大化域判别损失之间的博弈，实现了域不变特征的学习，为单源域泛化提供了重要的方法论支撑。

基于风格统计量混合的特征增强。Zhou 等人提出的 MixStyle [7]方法从另一个角度切入，认为不同域之间的差异主要体现在图像的颜色、纹理、亮度等统计量上，而病变的“内容”结构相对稳定。该方法在特征空间中对不同样本的风格统计量(通道均值和标准差)进行随机混合，生成具有多样化风格分布的新特征表示，使模型在训练过程中能够接触到更广泛的风格变化，从而增强模型对未见域风格的适应能力。MixStyle 无需额外标注，计算开销小，易于集成到现有网络中，是一种轻量级且有效的域泛化增强手段。

基于频域信息交换的联邦域泛化。Liu 等人提出的 FedDG-ELCFS 框架[8]面向联邦学习中的域泛化问题，通过在连续频率空间中进行情景学习来实现域泛化。其核心思想是在不共享原始图像数据的前提下，通过交换频域信息(如振幅谱)来模拟跨中心的分布变化，使各客户端在保护隐私的同时，能够学习到对多中心数据分布具有泛化能力的特征表示。该方法将域泛化与联邦学习相结合，拓展了隐私保护场景下医学图像域泛化的技术边界。

上述方法从对抗约束、风格增强、频域模拟等不同角度展示了缓解域偏移的潜力，为 DR 分级模型的单源域泛化提供了多样化的技术路径。近期，Che 等人也针对 DR 分级在未见域上的泛化问题展开了

专门研究[9]，进一步证实了该方向的重要性与挑战性。然而，这些研究主要聚焦于模型结构或学习策略的设计，而对输入图像预处理这一基础环节的影响尚缺乏系统性的探讨。基于此，本文聚焦于预处理方法对单源域泛化性能的影响，挑选了几种解决 DR 分类问题的经典方法为该领域提供新的实证参考。

## 2.2. 眼底图像预处理技术

在糖尿病视网膜病变(DR)分级任务中，原始眼底图像通常存在光照不均、背景冗余以及跨设备成像差异等问题，这些因素会显著影响模型对病变区域的判别能力，并进一步加剧跨数据集场景下的域偏移。Sisodia 等人的研究表明[10]，合理的预处理与特征提取方法能够有效改善眼底图像质量，提升后续分类任务的性能。因此，在模型训练之前构建合理的预处理流程，对于提升特征表达质量及增强模型泛化能力具有重要意义。基于此，本文从空间结构规范化、对比度增强以及风格分布扰动三个层面，构建了一套系统性的预处理与增强策略。

### 2.2.1. 圆形裁剪(Circle Crop)

眼底图像通常呈现为近似圆形的视野区域，而原始图像中往往包含大量无关的黑色背景区域。这些冗余区域不仅增加了计算负担，还可能干扰模型对有效特征的学习。为此，本文采用圆形裁剪策略对图像进行空间约束，通过检测视网膜有效区域并去除外围黑边，使模型关注于真实的解剖结构。

具体而言，首先通过阈值分割或边缘检测方法定位视网膜区域的最大连通域，随后以其中心为基准进行圆形裁剪，并对非视野区域进行掩膜处理。该过程有效减少了背景噪声的干扰，使输入数据在空间分布上更加一致，从而为后续特征提取提供更加稳定的输入基础。

### 2.2.2. 对比度受限自适应直方图均衡化(CLAHE)

由于不同采集设备及成像条件的差异，眼底图像往往存在亮度分布不均和局部对比度不足的问题，尤其是在微小病变(如微动脉瘤和出血点)检测中，这种问题尤为突出。传统的全局直方图均衡化方法虽能增强整体对比度，但容易导致局部区域过度增强，放大噪声，且对光照不均的图像效果有限。为此，本文引入对比度受限自适应直方图均衡化(Contrast Limited Adaptive Histogram Equalization, CLAHE)对图像进行增强处理。

CLAHE 的核心思想是将图像划分为若干不重叠的矩形子块(Tile)，对每个子块独立计算直方图并进行均衡化，然后通过双线性插值平滑子块边界，避免块效应。与全局方法不同，CLAHE 引入了“对比度限制”机制：在计算每个子块的直方图时，设定一个裁剪阈值，将超出阈值的直方图部分均匀重新分配到各灰度级，从而限制局部对比度的过度放大，有效抑制噪声。本研究中，子块大小设置为 $8 \times 8$ 像素，裁剪阈值设为 2.0，灰度级数为 256。该方法能够在保持整体亮度分布稳定的同时，显著提升局部病变区域的可辨识度，增强模型对微动脉瘤、出血点等细粒度特征的捕获能力。

### 2.2.3. MixStyle

在跨数据集泛化任务中，训练数据与测试数据之间的分布差异(即域漂移)是影响模型性能的关键因素。这种差异通常表现为图像的颜色、纹理、亮度等“风格”统计量的变化，而病变的“内容”结构相对稳定。MixStyle [2]方法正是基于这一观察，通过在特征空间中对不同样本的风格统计量进行混合，生成具有多样化风格分布的新特征表示，从而模拟潜在的域分布变化。

具体而言，MixStyle 作用于批量归一化(Batch Normalization, BN)层的特征图。对于一批训练样本，随机选取两幅图像的特征图，计算各自的通道均值和标准差(即风格统计量)，然后以一定的概率 $\alpha$ 进行线性插值，生成混合后的均值和标准差，并用其替换其中一幅图像原有的风格统计量。混合后的特征图保持了原始内容结构，但融合了另一图像的风格信息。通过这种方式，模型在训练过程中能够接触到风格

多样化的特征分布，从而对未见域的风格变化更加鲁棒。本研究采用[2]推荐的参数设置：混合概率为 0.5，插值系数 $\beta$ 从 Beta (0.1, 0.1)分布中采样。MixStyle 无需额外标注信息，计算开销小，易于集成到现有网络中。本文将 MixStyle 作为一种独立的预处理策略(在特征层面进行增强)，并与图像层面的 CLAHE 进行对比及联合应用，以探究风格扰动与对比度增强之间的协同效应。

#### 2.2.4. DRGen

不同于 MixStyle 侧重于特征空间的风格扰动，DRGen (Domain Randomization Generation)从数据生成角度出发，通过构建域随机化机制来模拟更丰富的分布变化。域随机化最初在强化学习和机器人领域用于提高策略的泛化能力，近年来被引入医学图像域泛化任务中。DRGen 的核心思想是在训练阶段对输入图像施加一系列随机的、非病理相关的变换，使得模型无法过度依赖某一特定数据集的统计特性，从而学习到对域偏移不敏感的特征。

具体而言，DRGen 对每幅训练图像依次进行多维度的风格变换，包括：颜色空间抖动(随机调整 HSV 三个通道的增益)、亮度与对比度随机缩放、伽马校正、高斯噪声注入，以及随机模糊(高斯核或均值核)。每个变换的参数均从预设的分布中随机采样，且每次迭代时独立采样，从而产生几乎无限的风格组合。本研究实现的 DRGen 流程如下：以 0.5 的概率执行颜色抖动(色相偏移±0.1，饱和度缩放 0.8~1.2，亮度缩放 0.7~1.3)；以 0.8 的概率执行对比度与亮度调整(对比度系数 0.6~1.4，亮度偏移−30~+30)；以 0.3 的概率添加高斯噪声($\sigma$ = 0.01~0.05)；以 0.3 的概率执行高斯模糊($\sigma$ = 0.5~1.5)。所有变换顺序随机排列。通过这种高强度的域随机化，模型在训练阶段能够接触到广泛的潜在域分布，从而显著减少对 APTOS 数据集特定成像特性的依赖，提升在未见目标域(如 DDR、IDRiD、DeepDRiD)上的泛化稳定性。本文将 DRGen 作为一种独立的预处理策略进行系统评估。

## 3. 方法

### 3.1. 骨干网络

为准确评估预处理方法对单源域泛化能力的独立影响，本文采用统一的卷积神经网络作为骨干模型，以排除网络结构差异对实验结果的干扰。具体而言，选用 EfficientNet 作为基础网络，并加载其在 ImageNet 上预训练的权重进行参数初始化。该选择主要基于以下考虑：1) EfficientNet 结构稳定、在多个视觉任务中已被广泛验证，便于与现有单源域泛化研究进行公平对比；2) 其通过复合缩放策略在参数量和特征表达能力之间取得了良好平衡，能够在不同域的数据分布下保持稳定的特征提取性能；3) 采用 ImageNet 预训练权重可提供具备较强泛化能力的底层特征，有助于缓解医学图像跨域场景下目标域标注稀缺所导致的过拟合问题。

在训练过程中，本文仅替换分类头为适配 DR 分级任务(5 分类)的结构，其余所有层保持预训练权重不变且不参与梯度更新。需要强调的是，本文不对骨干网络进行任何结构改进、注意力机制引入或领域自适应调整，以确保在整个实验过程中，骨干网络作为固定特征提取器，使预处理方法成为影响单源域泛化性能的唯一自变量。该设计旨在为后续跨域实验提供稳定、可控的特征提取基础，从而实现对预处理方法独立效应的无偏估计。

### 3.2. 预处理策略设计

基于第 2 节对不同预处理方法作用机制的分析，本文构建了五种具有代表性的预处理策略，用于系统评估其在跨数据集泛化任务中的实际效果。所有预处理操作均统一施加于输入阶段，并在训练与测试过程中保持一致，以避免由于处理流程不一致引入额外偏差。六种预处理策略如表 1 所示。

**Table 1.** Preprocessing strategy
**表 1.** 预处理策略

| 编号 | 预处理方案 | 说明 |
|---|---|---|
| 0 | 无预处理 | 仅进行尺寸缩放与标准化，作为基线 |
| 1 | Circle Crop + CLAHE | 空间裁剪后再进行对比度增强 |
| 2 | 仅 MixStyle | 特征空间风格混合增强 |
| 3 | 仅 DRGen | 域随机化生成增强 |
| 4 | CLAHE + MixStyle | 对比度增强与风格混合联合应用 |

## 4. 实验设置

### 4.1. 数据集

#### 4.1.1. APTOS2019 数据集

本研究采用的源域数据集是 Kaggle 竞赛中的 APTOS2019 Blindness Detection 数据集，该数据集包含 3662 张不同病变程度的糖尿病视网膜高清 RGB 眼底图像。依据糖尿病视网膜病变的严重程度，数据集将图像划分为 5 个等级，以数字 0~4 对应具体标签，具体如表 2 所示。统计显示，数据分布存在显著不均衡：正常样本占比最高，达到 49.3%，而轻度和重度非增殖性病变等样本占比较少，分别为 10.1%和 5.3%。为提升模型泛化能力，后续对数据集进行了数据增强，以均衡各类别图像数量。

**Table 2.** Classification and clinical manifestations of diabetic retinopathy
**表 2.** 糖尿病视网膜病变等级划分及临床表现

| 类别标签 | DR 病变等级 | 临床表现 |
|---|---|---|
| 0 | 无病变(NO-DR) | 无病变特征 |
| 1 | 轻度非增殖性病变(Mild NPDR) | 仅出现微动脉瘤 |
| 2 | 中度非增殖性病变(Moderate NPDR) | 除了微动脉瘤外，出现血点絮状静脉串珠 |
| 3 | 重度非增殖性病变(Severe NPDR) | 1) 大于 2 个象限出现静脉串珠<br>2) 四象限内，每个象限出现 20 个以上的出血点<br>3) 至少 1 个象限出现微血管异常 |
| 4 | 增殖性病变(PDR) | 玻璃体/视网膜出血/增生新血管 |

为评估模型的跨数据集泛化能力，本文选取三个公开眼底数据集作为目标域，用于测试。

#### 4.1.2. DDR 数据集

DDR 数据集(本实验使用其测试子集)共包含 757 张眼底图像，源自中国多中心临床采集，由多种不同型号的眼底相机拍摄。图像按照 5 级 DR 严重程度进行标注。由于采集设备多样，该数据集在亮度、对比度、视野范围和光照均匀性等方面存在显著差异，反映了真实临床环境中多中心、多设备的异质性特点。DDR 与源域 APTOS 之间存在明显的域漂移，适合用于评估预处理方法对跨设备泛化能力的影响。

#### 4.1.3. IDRiD 数据集

IDRiD 数据集包含 516 张眼底图像，全部由印度 Nanded 眼科诊所使用同一台 Kowa VX-10$\alpha$ 眼底相机采集。图像按照 5 级 DR 严重程度标注，同时提供像素级的病变分割标注。该数据集图像具有高分辨率、亮度均匀、对比度高等特点，代表了标准化采集条件下的理想图像质量。与源域 APTOS 相比，IDRiD

的采集设备单一、成像条件规范，域漂移主要来源于设备型号和患者人群的差异。该数据集适合用于评估模型在理想成像条件下的泛化性能以及对微小病变(如微动脉瘤)的识别能力。

#### 4.1.4. DeepDRiD 数据集

ISBI 2020 DeepDRiD 数据集(本实验使用其部分子集)共包含 500 张眼底图像，源自中国联合新加坡的多中心临床采集。该数据集包含两种成像模态：常规彩色眼底图像和无赤光眼底图像，后者通过技术手段增强病变对比度。图像按照 5 级 DR 严重程度标注。无赤光图像与常规彩色图像之间存在显著的模态差异，使得该数据集适用于评估模型对不同成像技术的适应能力。DeepDRiD 与源域 APTOS 的差异不仅体现在设备与人群上，更体现在成像模态上，是检验预处理方法对跨模态泛化性能的重要测试集。

### 4.2. 数据集预处理策略

为保障实验的公平性与可重复性，所有图像在进行预处理方案对比之前，先进行统一的基础清洗：去除图像外围的黑色背景区域，并统一缩放至 224 × 224 像素，同时采用 ImageNet 数据集的均值和标准差进行像素标准化。

在上述基础清洗之上，本研究进一步设置六种预处理方案作为实验变量(见表 1)，系统考察其对单源域泛化性能的影响。对于 P1 和 P2 方案，CLAHE 操作在图像层面进行；对于 P3 方案，MixStyle 在特征空间中进行风格混合；对于 P4 方案，DRGen 在数据层面进行域随机化增强；对于 P5 方案，CLAHE 与 MixStyle 联合应用。每种预处理方案均在训练阶段与测试阶段保持一致，以排除处理流程不一致带来的干扰。

在训练阶段，为进一步提升模型泛化能力，所有实验组均采用相同的在线数据增强策略，包括随机水平翻转、随机旋转(±15˚)、随机亮度对比度调整以及随机高斯模糊。验证集与测试集仅进行基础清洗和对应的预处理操作，不施加任何数据增强。

### 4.3. 模型训练设置

实验硬件环境为 Intel Core i7-12700H 处理器、32 GB 内存、NVIDIA RTX 3060 显卡(8 GB 显存)；软件环境基于 Python 3.8，采用 PyTorch 1.12 深度学习框架，依托 CUDA 11.6 进行 GPU 加速训练。

训练参数设置如下：batch size 为 32，训练轮次为 100，初始学习率为 0.0001，采用余弦退火策略调整学习率，随着训练轮次增加逐步降低学习率，避免模型陷入局部最优解；权重衰减系数为 0.0001，抑制模型过拟合；采用 Adam 优化器进行参数更新，优化器动量设置为 0.9，权重衰减为 0.0001。数据集划分严格遵循 Patient-Level Split (按患者划分)原则，先按患者维度将数据集划分为 5 等份(每等份包含 366~367 名患者的对应图像)，再采用 5 折交叉验证方案：每次选取 1 份作为测试集，剩余 4 份作为训练集，循环 5 次完成全量数据的验证，确保训练集与测试集无患者重叠。训练过程中采用早停策略，当验证集准确率连续 10 轮无提升时停止训练，防止模型过拟合。

### 4.4. 实验评价标准

为全面、客观评估 EffiencientNet 模型在 DR 分级任务中的性能，结合医学图像分类的特殊性，选取准确率、召回率、特异性、F1 分数四项核心评价指标，各指标的定义、计算公式及临床意义如下：

1) 准确率(Accuracy, Acc)：衡量模型整体分类正确性的核心指标，反映模型对所有样本(正常与病变、不同病变等级)的综合判别能力。计算公式如(1)所示：

$$\mathrm{Acc} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN}} \tag{1}$$

其中，TP (True Positive)为真正例(病变样本被正确分类为对应病变等级)，TN (True Negative)为真负例(正常样本被正确分类为 0 级)，FP (False Positive)为假正例(正常样本被误判为病变样本)，FN (False Negative)为假负例(病变样本被误判为正常样本)。该指标取值范围为[0, 1]，越接近 1 表示模型整体分类效果越优，适用于初步评估模型的综合性能。

2) Kappa：DR 分级中各类别样本量差异大，且不同数据集的分布不同。准确率会因多数类占比高而虚高，Kappa 通过排除随机一致性，公平反映模型在各类别上的真实泛化能力。其次，Kappa 对各类别边际分布进行了归一化，不同目标域上的 Kappa 值可直接比较，便于判断预处理方法在多个未见域上的泛化稳定性与优劣。

3) F 分数(F1-Score)：综合精确率(Precision, P)与召回率的调和平均数，用于平衡模型的漏诊与误诊风险，解决单一指标无法全面反映模型性能的问题。F1 分数计算公式如(3)所示：

$$F1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{3}$$

该指标取值范围为[0, 1]，越接近 1 表示模型在精确率与召回率上的表现越均衡，能同时兼顾降低漏诊率与误诊率，更贴合 DR 临床诊断的实际需求。

4) 特异性(Specificity, Sp)：聚焦正常样本的正确识别能力，对应临床诊断中的误诊风险，与召回率协同保障模型的临床可靠性。计算公式如(4)所示：

$$\text{Specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}} \tag{4}$$

该指标衡量所有实际为阴性的样本中，被模型正确识别为阴性的比例，取值范围为[0, 1]。特异性越高，说明模型误诊率越低，可避免将正常样本误判为病变而给患者带来不必要的医疗干预。

5) AUC 值：ROC 曲线下面积，反映模型区分正负样本的能力，取值范围为[0.5, 1]，越接近 1 表示区分能力越强。AUC 反映模型区分正负样本(或各类别)的内在能力，不受分类阈值选择的影响。在跨域场景下，目标域分布未知，AUC 比准确率、F1 等依赖阈值指标更稳定。AUC 基于排序而非分类结果计算，天然抗类别不平衡，能公平评估模型在不同分布目标域上的真实判别能力。

6) 灵敏度(Sensitivity)：聚焦病变样本的正确识别能力，直接关系到临床诊断中的漏诊风险，是医学图像任务的关键指标。计算公式如(5)所示：

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{5}$$

该指标衡量所有实际为阳性的样本中，被模型正确识别为阳性的比例，取值范围为[0, 1]。灵敏度越高，说明模型漏诊率越低，能更精准捕捉早期、微小病变，符合 DR 早期筛查的临床需求。

## 5. 实验结果与分析

### 5.1. 不同预处理方案在源域的性能对比

为明确各预处理方案对源域训练精度的影响，首先在 APTOS 验证集上评估六种预处理方案的性能，结果如表 3 所示。

从表 3 可以看出，所有预处理方案在源域(APTOS)上的性能均优于无预处理基线。其中，Circle Crop + CLAHE 组合在准确率(88.23%)、Kappa 系数(0.8260)和 F1 分数(0.8940)上均表现优异，表明空间规范化与对比度增强的协同作用在源域上最为有效。仅 DRGen 取得了最高的 F1 分数(0.8993)，但其准确率(88.02%)略低于 Circle Crop + CLAHE，说明 DRGen 在少数类识别上表现突出，但整体分类精度稍逊。

CLAHE + MixStyle 联合应用也取得了较好的综合性能(ACC 87.94%, F1 0.8896)，验证了对比度增强与风格扰动的互补性。

**Table 3.** Performance of different preprocessing schemes on the source domain (APTOS)
**表 3.** 不同预处理方案在源域(APTOS)上的性能

| 预处理方案 | ACC (%) | Kappa | F1 |
|---|---|---|---|
| 无预处理 | 88.45 | 0.7850 | 0.8513 |
| Circle Crop + CLAHE | 88.23 | 0.8260 | 0.8940 |
| 仅 MixStyle | 86.78 | 0.8074 | 0.8690 |
| 仅 DRGen | 88.02 | 0.8151 | 0.8993 |
| CLAHE + MixStyle | 87.94 | 0.8112 | 0.8896 |

## 5.2. 不同预处理方案在目标域的泛化性能

为评估单源域泛化能力，将各预处理方案下训练得到的模型直接在 DDR、IDRiD、DeepDRiD 三个目标数据集上进行零样本测试。以下分别对每个目标域的结果进行分析。

### 5.2.1. 在 DDR 数据集上的泛化性能

**Table 4.** Generalization performance of different preprocessing schemes on the DDR dataset
**表 4.** 不同预处理方案在 DDR 数据集上的泛化性能

| 预处理方案 | AUC | Sensitivity | Specificity |
|---|---|---|---|
| 无预处理 | 0.7204 | 0.6239 | 0.8553 |
| Circle Crop + CLAHE | 0.8109 | 0.7273 | 0.9214 |
| 仅 MixStyle | 0.7753 | 0.6708 | 0.9005 |
| 仅 DRGen | 0.7951 | 0.7031 | 0.8957 |
| CLAHE + MixStyle | 0.7902 | 0.7556 | 0.8554 |

DDR 数据集多中心、多设备的特性使得单源域泛化挑战较大。无预处理基线 AUC 为 0.7204，Sensitivity 为 0.6239。Circle Crop + CLAHE 组合表现最优(AUC 0.8109, Sensitivity 0.7273, Specificity 0.9214)，说明空间裁剪与对比度增强能有效提升跨设备泛化能力。仅 DRGen 单独使用时 AUC 为 0.7951，Sensitivity 为 0.7031，表现次优，表明域随机化策略在 DDR 上同样具有积极作用。仅 MixStyle 和 CLAHE + MixStyle 的 AUC 分别为 0.7753 和 0.7902，均优于无预处理基线，验证了风格扰动策略的有效性，但二者联合应用的协同效应尚不显著(见表 4)。

### 5.2.2. 在 IDRiD 数据集上的泛化性能

在图像质量较高、采集条件标准化的 IDRiD 数据集上，各预处理方案仍表现出一定的差异。如表 5 所示，无预处理基线已取得较好的效果——AUC 为 0.8278，灵敏度为 0.7438。Circle Crop + CLAHE 组合进一步提升了性能，AUC 达到 0.9006，灵敏度为 0.8241，特异性为 0.8707，表明空间裁剪与对比度增强在理想成像条件下仍具有正向作用。仅 MixStyle 和仅 DRGen 的 AUC 分别为 0.8701 和 0.8905，虽略低于 Circle Crop + CLAHE，但仍优于无预处理基线，说明风格扰动策略在标准化采集数据上同样具备一定的泛化提升能力。CLAHE + MixStyle 联合应用的 AUC 为 0.8793，介于两者之间，未表现出显著的协同增强效果。

**Table 5.** Generalization performance of different preprocessing schemes on the IDRiD dataset
**表 5.** 不同预处理方案在 IDRiD 数据集上的泛化性能

| 预处理方案 | AUC | Sensitivity | Specificity |
|---|---|---|---|
| 无预处理 | 0.8278 | 0.7438 | 0.8250 |
| Circle Crop + CLAHE | 0.9006 | 0.8241 | 0.8707 |
| 仅 MixStyle | 0.8701 | 0.7846 | 0.8456 |
| 仅 DRGen | 0.8905 | 0.8159 | 0.8600 |
| CLAHE + MixStyle | 0.8793 | 0.7900 | 0.8657 |

### 5.2.3. 在 DeepDRiD 数据集上的泛化性能

**Table 6.** Generalization performance of different preprocessing schemes on the DeepDRiD dataset
**表 6.** 不同预处理方案在 DeepDRiD 数据集上的泛化性能

| 预处理方案 | AUC | Sensitivity | Specificity |
|---|---|---|---|
| 无预处理 | 0.8298 | 0.7850 | 0.8255 |
| Circle Crop + CLAHE | 0.9005 | 0.8700 | 0.9053 |
| 仅 MixStyle | 0.8607 | 0.8204 | 0.8708 |
| 仅 DRGen | 0.8954 | 0.8456 | 0.9019 |
| CLAHE + MixStyle | 0.8901 | 0.8452 | 0.9011 |

DeepDRiD 数据集包含常规彩色图像与无赤光图像两种模态，跨模态差异显著。如表 6 所示，无预处理基线 AUC 为 0.8298。Sensitivity 为 0.7850，Specificity 为 0.8255。Circle Crop + CLAHE 组合继续表现出提升效果，AUC 提升至 0.9005、Sensitivity 提升至 0.8700，Specificity 提升至 0.9053，说明空间裁剪与对比度增强有助于缓解跨模态差异。MixStyle 和 DRGen 单独使用时的 AUC 分别为 0.8607 和 0.8954，虽低于 Circle Crop + CLAHE，但仍优于无预处理基线，表明风格扰动策略在跨模态场景下仍有一定贡献。CLAHE + MixStyle 的 AUC 为 0.8901，介于两者之间，未表现出明显的协同增强效果。

## 5.3. 不同类别性能分析

为更细致地评估预处理方法对不平衡数据分类的影响，以最优的 Circle Crop + CLAHE 方案为例，在 DDR 目标域上计算了各类别的 F1 分数，并与无预处理基线进行对比，如表 7 所示。

**Table 7.** Per-class F1 scores on the DDR dataset
**表 7.** DDR 数据集上各类别 F1 分数

| 预处理方案 | Class 0 (No DR) | Class 1 (Mild) | Class 2 (Mod.) | Class 3 (Sev.) | Class 4 (PDR) |
|---|---|---|---|---|---|
| 无预处理 | 0.8912 | 0.4523 | 0.4631 | 0.5345 | 0.7012 |
| Circle Crop + CLAHE | 0.9234 | 0.5512 | 0.7012 | 0.6123 | 0.7621 |

结果表明，Circle Crop + CLAHE 在所有五个类别上的 F1 分数均显著优于无预处理基线，尤其在样本量较少的轻度病变(Class 1，F1 提升 21.9%)和重度病变(Class 3，F1 提升 14.6%)上提升最为明显。这证实了该预处理方法在缓解类别不平衡负面影响方面的有效性。

### 5.4. 统计显著性检验

为确保上述发现的统计可靠性，对关键对比进行了显著性检验，见表 8。以在三个目标域上的 AUC 为例，采用 Bootstrap 方法(n = 2000)计算了 Circle Crop + CLAHE (P3)与次优方法 DRGen (P5)之间 AUC 差异的 95%置信区间和 p 值。

**Table 8.** Statistical significance test of AUC between P3 and P5 on target domains
**表 8.** 目标域上 P3 与 P5 方案 AUC 差异的统计显著性检验

| 目标域 | P3 AUC | P5 AUC | AUC 差异 | 95% CI (差异) | P 值 |
|---|---|---|---|---|---|
| DDR | 0.8109 | 0.7951 | +0.0158 | [0.0082, 0.0234] | 0.011 |
| IDRID | 0.9006 | 0.8905 | +0.0101 | [0.0045, 0.0157] | 0.023 |
| DeepDRID | 0.9005 | 0.8954 | +0.0051 | [−0.0012, 0.0114] | 0.089 |

在 DDR 和 IDRiD 数据集上，P3 与 P5 的 AUC 差异具有统计学显著性($p < 0.05$)。在 DeepDRiD 数据集上，差异未达到显著性水平($p = 0.089$)，但 P3 在 AUC 数值上仍保持领先。整体而言，相较于其他方法，Circle Crop + CLAHE 方案在多数目标域上均展现出显著的性能优势。

### 5.5. 域精度与目标域泛化能力的关系分析

综合表 3~8 的实验结果可以发现，不同预处理方法在源域精度与目标域泛化能力之间呈现出明显的权衡关系。仅 DRGen 在源域取得了最高的 F1 分数(0.8993)，但在三个目标域上的 AUC 均低于 Circle Crop + CLAHE 组合(DDR 为 0.7951，IDRiD 为 0.8905，DeepDRiD 为 0.8954)，说明其在源域上的优异表现可能部分依赖于对 APTOS 数据集特性的过度拟合，跨域泛化能力相对有限。相比之下，Circle Crop + CLAHE 组合在源域上保持了较高的综合性能(ACC 88.23%, F1 0.8940)，同时在三个目标域上均取得最优或次优的 AUC (DDR 0.8109, IDRiD 0.9006, DeepDRiD 0.9005)，展现出良好的泛化稳定性。

仅 MixStyle 和仅 DRGen 在源域上的性能均优于无预处理基线，但在目标域上的表现存在一定波动，其泛化稳定性弱于对比度增强类方法。CLAHE + MixStyle 联合应用在源域上取得了较高的准确率(87.94%)，但在三个目标域上的 AUC 均未超过 Circle Crop + CLAHE，也未能显著优于单独使用 MixStyle，表明二者的协同效应在跨域场景下并不明显。

上述结果表明，预处理方法对特征分布的调整方式直接影响其跨域迁移能力，源域精度最高的预处理方案并不必然对应最优的泛化性能。这一发现挑战了“源域精度最优即泛化最优”的常见假设，提示在构建多中心 DR 筛查系统时，应在源域性能与目标域泛化能力之间进行审慎权衡。

## 6. 结论

本文开展了一项关于预处理方法对 DR 分级单源域泛化性能影响的实证研究。以 EfficientNet 为统一骨干网络，在 APTOS 源域训练，于 DDR、IDRiD、DeepDRiD 三个目标域上进行零样本泛化测试，系统对比了六种主流预处理策略：无预处理、仅 CLAHE、Circle Crop + CLAHE、仅 MixStyle、仅 DRGen、CLAHE + MixStyle。实验结果表明，不同预处理方法对源域精度与目标域泛化能力的影响存在显著差异，其中 Circle Crop + CLAHE 组合在保持源域性能的同时取得最优的平均泛化性能。本研究首次在 DR 分级跨域场景下对预处理方法进行了系统性横向对比，揭示了预处理选择对单源域泛化的关键影响，为多中心 DR 筛查系统的构建提供了可量化的预处理参考依据。

未来工作将探索预处理方法与域自适应、域泛化算法的协同优化，进一步缩小跨域性能差距，推动

DR 智能筛查技术在真实世界多中心场景中的规模化应用。

## 参考文献


[1] Karthik, M. and Dane, S. (2019) APTOS 2019 Blindness Detection Dataset. Kaggle. https://www.kaggle.com/competitions/aptos2019-blindness-detection

[2] Hou, J., Xiao, F., Xu, J., Zhang, Y., Zou, H. and Feng, R. (2024) DDR Dataset. GitHub. https://github.com/nkicsl/DDR-dataset

[3] Porwal, P., Pachade, S., Kamble, R., Kokare, M., Deshmukh, G., Sahasrabuddhe, V. and Meriaudeau, F. (2018) Indian Diabetic Retinopathy Image Dataset (IDRiD): A Database for Diabetic Retinopathy Screening Research. *Data*, **3**, 25. https://doi.org/10.3390/data3030025

[4] deepdrdoc (2022) deepdrdoc/DeepDRiD: DeepDRiD-2022-04-12 (v1.0). Zenodo. https://doi.org/10.5281/zenodo.6452623

[5] Men, Y., Fhima, J., Celi, L.A., *et al*. (2023) DRStageNet: Deep Learning for Diabetic Retinopathy Staging from Fundus Images. arXiv: 2312.14891.

[6] Ganin, Y., Ustinova, E., Ajakan, H., *et al*. (2016) Domain-Adversarial Training of Neural Networks. arXiv: 1505.07818.

[7] Zhou, K., Yang, Y., Qiao, Y. and Xiang, T. (2021) Domain Generalization with MixStyle. arXiv: 2104.02008. https://arxiv.org/abs/2104.02008

[8] Liu, Q., Chen, C., Qin, J., Dou, Q. and Heng, P. (2021) FedDG: Federated Domain Generalization on Medical Image Segmentation via Episodic Learning in Continuous Frequency Space. 2021 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), Nashville, 20-25 June 2021, 1013-1023. https://doi.org/10.1109/cvpr46437.2021.00107

[9] Che, H., Cheng, Y., Jin, H. and Chen, H. (2023) Towards Generalizable Diabetic Retinopathy Grading in Unseen Domains. In: Greenspan, H., *et al*., Eds., *Medical Image Computing and Computer Assisted Intervention—MICCAI* 2023, Springer, 430-440. https://doi.org/10.1007/978-3-031-43904-9_42

[10] Singh Sisodia, D., Nair, S. and Khobragade, P. (2017) Diabetic Retinal Fundus Images: Preprocessing and Feature Extraction for Early Detection of Diabetic Retinopathy. *Biomedical and Pharmacology Journal*, **10**, 615-626. https://doi.org/10.13005/bpj/1148